كلية العلوم و التقنيات فاس
ⵜⴰⵖⵉⵡⴰⵏⵜ ⵏ ⵜⵎⴰⵙⵙⴰⵏⵉⵏ ⴷ ⵜⵙⵏⵉⴳⵉⵜⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⵎⴰⴷ ⴱⵏ ⵄⴱⴷⵓⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

Centre d'Etudes Doctorales : Sciences et Techniques de l'Ingénieur

N° d'ordre : 35/2019

THESE DE DOCTORAT

Présentée par

M. : Karim DAHDOUH

Spécialité : Informatique

Sujet de la thèse : Intégration du cloud computing en se basant sur les nouvelles technologies big data dans un environnement informatique pour l'apprentissage humain

Thèse présentée et soutenue le 9 juillet 2019 devant le jury composé de

Nom Prénom	**Titre**	**Etablissement**	
Salaheddine ALAMI KAMMOURI	PES	Faculté des Sciences et Techniques de Fès	Président
Brahim AKSASSE	PES	Faculté des Sciences et Techniques d'Errachidia	Rapporteur
Badraddine AGHOUTANE	PH	Faculté Polydisciplinaire d'Errachidia	Rapporteur
Abderrahim BENABBOU	PH	Faculté des Sciences et Techniques de Fès	Rapporteur
Aziz DEROUICH	PH	Ecole Supérieure de Technologie de Fès	Examinateur
Abdelali IBRIZ	PH	Ecole Supérieure de Technologie de Fès	Examinateur
Adil KENZI	PH	Ecole Nationale des Sciences Appliquées de Fès	Examinateur
Ahmed DAKKAK	P.E.S	Faculté Polydisciplinaire de TAZA	Directeurs de thèse
Lahcen OUGHDIR	P.H	Ecole Nationale des Sciences Appliquées de Fès	

Laboratoire d'accueil : LSI (laboratoire des sciences de l'ingénieur)-FP-Taza

Etablissement : Faculté des Sciences et Techniques de Fès

Centre d'Etudes Doctorales Sciences et Techniques de l'Ingénieur

Faculté des Sciences et Techniques – Route d'Immouzer, B.P. 2202 – Fès – Maroc

Tél.: (+212) 535 60 80 14; Fax : (+212) 535 60 82 14; Site web : www.fst-usmba.ac.ma

Remerciements

Les travaux de recherche présentés dans ce rapport de thèse de Doctorat ont été effectués sous la direction du professeur Ahmed DAKKAK et Lahcen Oughdir, et ont été préparés au sein du laboratoire des sciences de l'ingénieur, Faculté des Sciences et Techniques, Université Sidi Mohamed Ben Abdellah de Fès.

Je remercie en premier lieu Dieu le tout puissant de m'avoir donné la puissance, le courage, la patience, et la volonté pour achever ce travail.

J'ai l'honneur et le plaisir d'exprimer ma profonde gratitude et mes sincères remerciements à mon directeur de thèse le Professeur Ahmed DAKKAK d'avoir accepté de diriger les travaux de recherche de cette thèse. Je tiens à le remercier tout particulièrement pour son suivi, ses encouragements, son ouverture d'esprit, sa disponibilité, son soutien constant et bien sûr ses qualités scientifiques exceptionnelles. Je le remercie également pour sa modestie, son amabilité et ses qualités humaines remarquables. Qu'il trouve ici le témoignage de mon estime et de ma profonde reconnaissance pour ses conseils pertinents et ses recommandations précieuses qui m'ont permis de mener à bien les travaux de cette thèse. Il m'est très difficile de lui exprimer en quelques mots mes remerciements. Je lui suis très reconnaissant de la confiance qu'il m'a accordée pendant des années.

Je tiens à remercier du fond du cœur mon encadrant, le Professeur Lahcen OUGHDIR de m'avoir encadré et encouragé durant toute la période de préparation de cette thèse. Je tiens à le remercier vivement pour son suivi rigoureux, ses recommandations précieuses, sa gentillesse et ses grandes qualités humaines. Je lui exprime également mes remerciements les plus vifs pour ses compétences scientifiques qu'il a mises à ma disposition pour mener à bien ces travaux de recherches. Qu'il trouve ici l'expression de mon estime et ma profonde reconnaissance pour sa disponibilité et ses conseils avisés qui m'ont permis de progresser et de persévérer tout au long de cette thèse.

Mes remerciements les plus sincères s'adressent ensuite aux membres du jury qui me font l'honneur de participer à la soutenance. Je remercie sincèrement Messieurs les professeurs Brahim AKSASSE de la Faculté des Sciences et Techniques d'Errachidia, Badraddine AGHOUTANE de la Faculté Polydisciplinaire d'Errachidia, ainsi que le professeur Abderrahim

BENABBOU de la Faculté des Sciences et Techniques de Fès d’avoir accepté d'évaluer mes travaux en qualité de rapporteurs. Mes remerciements s'adressent également à Messieurs les professeurs Abdelali IBRIZ de l’Ecole Supérieure de Technologie de Fès, Aziz DEROUICH de d’Ecole Supérieure de Technologie de Fès, et Adil KENZI de Ecole Nationale des Sciences Appliquées de Fès, pour leur participation au jury en tant qu'examinateurs de cette thèse. Également, je remercie vivement Monsieur Salaheddine ALAMI KAMMOURI, professeur de la Faculté des Sciences et Techniques de Fès, pour avoir accepté d’être le président du jury de cette thèse.

Je tiens à remercier tous les membres de ma famille qui m’ont toujours soutenu et encouragé. Sans eux, je ne serais pas qui je suis aujourd’hui, sans eux l’atteinte de mes objectifs n’aurait pas été possible. Je remercie, également, mon frère Rida DAHDOUH pour ses suggestions, sa relecture critique et sa révision attentive de ce rapport.

Enfin, j’adresse mes remerciements les plus chaleureux à tous ceux qui ont contribué, de loin ou de près, à l’aboutissement de cette thèse de doctorat.

Résumé

Dans un monde en pleine transformation digitale, où les nouvelles technologies de l'information et de la communication ne cessent d'évoluer, le défi actuel des Environnements Informatiques pour l'Apprentissage Humain (EIAH) est de trouver la façon adéquate permettant l'intégration et l'exploitation de la puissance de ces technologies. En fait, ces environnements sont confrontés à de nombreux challenges, surtout la demande accrue de l'apprentissage, l'énorme croissance du nombre des apprenants, l'hétérogénéité des ressources mises à disposition ainsi que les besoins du traitement et de l'analyse des données produites par les systèmes de l'enseignement à distance, ce qui dépasse les limites des infrastructures traditionnelles et des systèmes de gestion des bases de données relationnelles.

Dans cette thèse, nous présentons un certain nombre de solutions dédiées aux EIAH en termes d'approches, d'architectures, de méthodologies et d'applications autour de deux grands paradigmes, à savoir le cloud et le big data. Un premier volet de ce travail est consacré à la présentation d'une approche d'intégration de l'e-learning dans l'environnement cloud. Elle vise à améliorer la qualité des plates-formes d'apprentissage en ligne en se basant sur les services fournis par le cloud accessibles via internet. Cette approche permet aux universités et aux organismes de formation de migrer leurs systèmes e-learning vers l'informatique en nuage afin de bénéficier d'une élasticité illimitée, d'une disponibilité permanente, d'une mutualisation de ressources, et d'un mode de paiement à l'usage. Le deuxième volet de cette thèse est réservé à la proposition d'un modèle d'e-learning incorporant à la fois les technologies émergentes de l'écosystème big data et les services à la demande du cloud. Il s'agit d'introduire l'informatique distribuée et parallèle du big data en vue d'apporter des solutions robustes aux exigences du calcul intensif, de l'analyse prédictive, et du stockage massif des données d'apprentissage. Pour ce faire, une méthodologie est présentée et appliquée qui décrit la démarche du mécanisme d'intégration. De plus, ce volet aborde également le déploiement d'une architecture distribuée de l'e-learning combinant plusieurs outils récents du big data et se basant sur une stratégie de la décentralisation des données et la parallélisation des traitements sur un cluster de nœuds.

Pour valider notre modèle fondé sur l'intégration des technologies et des méthodes du big data dans le contexte d'EIAH, un système de recommandation de cours a été mis en œuvre pour démontrer l'efficacité et l'utilité de notre contribution. Il s'agit d'un moteur de recommandation agissant comme une unité de prédiction capable de s'adapter à un apprenant donné en anticipant ses prochaines actions à travers la recommandation des ressources pédagogiques pertinentes qui répondent au mieux à ses préférences et à ses intérêts. L'objectif est donc de fournir à l'apprenant un schéma éducatif et pédagogique personnalisé en fonction de son profil. Pour ce faire, notre système de recommandation fait appel aux techniques avancées d'apprentissage automatique, notamment celle des règles d'association permettant l'extraction des connaissances via l'analyse des traces d'apprentissage. La réalisation de ce système a été faite en utilisant les données historiques des activités des apprenants collectées à partir de la plate-forme ESTenLigne de l'Ecole Supérieure de Technologie de Fès. Deux versions de systèmes de recommandation ont été développées dans ce travail. La première version de ce système se base sur une architecture simple et utilise dans son implémentation l'algorithme Apriori et le langage R. La deuxième version s'appuie sur une architecture totalement distribuée qui consiste à mettre en place un système de recommandation de cours à grande échelle. Elle a été développée et testée en utilisant l'algorithme parallèle FP-growth, et le Framework Apache Spark. Le déploiement de cette version est fait par l'intermédiaire de l'infrastructure répartie de cluster Hadoop.

Mots clés : EIAH, e-learning, apprentissage en ligne, cloud computing, big data, méthodes d'apprentissage automatique, technique des règles d'association, architecture distribuée, calcul parallèle, système de recommandation de cours, algorithme Apriori, FP-growth, langage R, Apache Spark, Hadoop.

Abstract

In a world in full digital transformation, where new information and communication technologies are constantly evolving, the current challenge of Computing Environments for Human Learning (CEHL) is to find the right way to integrate and harnessing the power of these technologies. In fact, these environments face many challenges, especially the increased demand for learning, the huge growth in the number of learners, the heterogeneity of available resources as well as the needs for processing and analyzing data produced by distance education systems, which goes beyond the limits of traditional infrastructures and relational database management systems.

In this thesis, we present a number of solutions dedicated to CEHL in terms of approaches, architectures, methodologies and systems around the two big paradigms, namely cloud computing and big data. The first part of this work is dedicated to the presentation of an approach to integrate e-learning in the cloud environment. It aims to improve the quality of e-learning platforms based on the services provided by the cloud accessible via the internet. This approach allows universities and training organizations to migrate their e-learning systems to cloud in order to benefit from an unlimited elasticity, a permanent availability, a pooling of resources, and a pay per use mode. The second part of this thesis is dedicated to the proposal of an e-learning model incorporating both emerging technologies of the big data ecosystem and on-demand services of the cloud. It aims to introduce distributed and parallel computing of big data in order to provide robust solutions to the requirements of intensive processing, predictive analysis, and massive storage of learning data. To do this, a methodology is presented and applied which describes the integration mechanism. In addition, this part also addresses the deployment of a distributed e-learning architecture combining several recent tools of the big data and based on a strategy of data decentralization and the parallelization of the treatments on a cluster of nodes.

To validate our model based on the integration of big data technologies and methods in the context of CEHL, a course recommender system has been implemented to show the effectiveness and usefulness of our contribution. It is about a recommendation engine acting as a predictive unit able to adapt to a given learner by anticipating its next actions through the

suggestion of relevant pedagogical resources that best meet its preferences and interests. The aim is, therefore, to provide the learner with a personalized educational and pedagogical plan according to his profile. For this purpose, our recommender system uses advanced machine learning techniques, especially association rules method for extracting knowledge through the analysis of learning traces. The realization of this system was made using the historical data of learner's activities collected from the ESTenLigne platform of the Higher School of Technology of Fez. Two versions of recommender systems have been developed in this work. The first version of this system is based on a simple architecture and uses in its implementation the Apriori algorithm and the R language. The second version relies on a totally distributed architecture which consists in setting up a large-scale course recommender system. It was developed and tested using the FP-growth parallel algorithm, and the Apache Spark Framework. The deployment of this version is done through the Hadoop distributed cluster infrastructure.

Keywords: CEHL, e-learning, online learning, cloud computing, big data, machine learning methods, association rules technique, distributed architecture, parallel computing, course recommender system, Apriori, FP-growth algorithm, R language, Apache Spark, Hadoop.

Sommaire

Remerciements **1**
Résumé **3**
Abstract **5**
Sommaire **7**
Liste des figures **10**
Liste des tableaux **13**
Liste des Abréviations **14**
Introduction Générale **16**
1. Contexte de la thèse 17
2. Motivation 18
3. Problématique 19
4. Objectifs et contributions 20
5. Organisation du rapport 22
I. CHAPITRE I : Environnements Informatiques pour l'Apprentissage Humain **25**
I.1 Introduction 26
I.2 Etat de l'art : évolution et définitions 26
I.2.1 Environnements informatiques pour l'apprentissage humain 26
I.2.2 Brève histoire des EIAH 28
I.2.3 E-learning 34
I.2.4 Formation Ouverte et A Distance (FOAD) 35
I.2.5 Les tendances de l'e-learning 36
I.3 Etat de l'art : travaux existants 39
I.3.1 Architecture de l'e-learning basée sur le cloud 39
I.3.2 Infrastructure du cloud computing au service de l'e-learning 42
I.3.3 Adaptation du contenu e-learning basée sur le cloud 46
I.3.4 Framework d'intégration du big data dans l'e-learning 50
I.3.5 Méthodologie d'intégration du big data dans l'e-learning 52
I.3.6 Modèle d'adaptation du big data et du cloud computing avec l'e-learning 54
1.1 Conclusion 56
II. CHAPITRE II : Cloud Computing et Big Data **57**
II Cloud Computing 58

II.1.1 Introduction 58
II.1.2 Brève Histoire 59
II.1.3 Définitions 60
II.1.4 Caractéristiques 61
II.1.5 Concept du cluster 62
II.1.6 Services 63
II.1.7 Modèles de déploiement 65
II.1.8 Technologies 67
II.1.9 Avantages 70
II.1.10 Défis 72
II.1.11 Principaux acteurs 73
II.1.12 Conclusion 78
II.2 Big Data 79
II.2.1 Introduction 79
II.2.2 Courte Histoire 79
II.2.3 Définition 80
II.2.4 Domaine du big data 82
II.2.5 Caractéristiques du big data 83
II.2.6 Types de données en big data 85
II.2.7 Comparaison entre SGBDR et big data 86
II.2.8 Les technologies big data 87
II.2.9 Les bases de données NoSQL 104
II.2.10 Méthodes et techniques du machine learning 108
II.2.11 Conclusion 120
III. CHAPITRE III : Intégration du cloud et du big data dans un EIAH 121
III.1 Approche d'intégration de l'e-learning dans l'environnement cloud 122
III.1.1 Introduction 122
III.1.2 Modèle de l'e-learning traditionnel 122
III.1.3 Modèle de l'e-learning basé sur le cloud 125
III.1.4 Mode interactif de l'e-learning en mode SaaS 131
III.1.5 Comparaison entre le modèle traditionnel de déploiement et celui du cloud 132
III.1.6 Avantages et bénéfices du modèle e-learning basé sur le cloud 135
III.1.7 Valeur ajoutée du cloud pour les plates-formes e-learning 139

III.1.8 Contraintes et limites du cloud 141
III.1.9 Conclusion 142
III.2 Modèle d'e-learning basé sur le cloud et le big data 143
III.2.1 Introduction 143
III.2.2 Modèle e-learning basé sur le cloud et le big data 145
III.2.3 Méthodologie d'intégration des technologies big data dans l'e-learning 149
III.2.4 Architecture distribuée du big data pour les systèmes e-learning 153
III.2.5 Infrastructure répartie du cluster Hadoop 155
III.2.6 Bénéfices et avantages de l'intégration du big data dans les EIAH 157
III.2.7 Conclusion 161
IV. CHAPITRE IV : Mise en œuvre des technologies big data dans un EIAH 162
IV.1 Système de recommandation de cours 163
IV.1.1 Introduction 163
IV.1.2 Méthode 167
IV.1.3 Algorithmes 171
IV.1.4 Application et résultats 177
IV.1.5 Conclusion 195
IV.2 Système de recommandation de cours à grande échelle 197
IV.2.1 Introduction 197
IV.2.2 Algorithme FP-growth Parallèle 198
IV.2.3 Technologies mises en œuvre 199
IV.2.4 Architecture répartie Hadoop 201
IV.2.5 Préparation et configuration de l'environnement big data 205
IV.2.6 Application distribuée Spark 206
IV.2.7 Résultats 208
IV.2.8 Performance du système de recommandation de cours 210
IV.2.9 Outils de monitoring ou supervision 212
IV.2.10 Conclusion 219
Conclusion générale 220
Références 223
Travaux scientifiques 232
Prix de recherche 235
Certificat 235

Liste des figures

Figure I.1 Supports apprentissage ASSIMIL 1970-80 30
Figure I.2: Les ordinateurs Thomson MO5 et T07 dédié à l'apprentissage (1980) 30
Figure I.3: Brève histoire de l'enseignement à distance 32
Figure I.4: Architecture traditionnelle d'un STI 33
Figure I.5: M-learning est un prolongement de l'e-learning 37
Figure I.6: Caractéristiques du MOOC 39
Figure I.7: Intégration du cloud computing dans l'éducation 40
Figure I.8: Architecture du could e-learning[23] 41
Figure I.9: Infrastructure du cloud privé pour e-learning 43
Figure I.10: Modèle conceptuel de l'infrastructure réseau du cloud 45
Figure I.11: Diagramme d'activité du service de réservation des ressources cloud 46
Figure I.12: Adaptation du contenu intelligent d'un système e-learning dans le cloud. 47
Figure I.13: Environnement de services des médias éducatifs basés sur le cloud 48
Figure I.14: Les processus de création et de distribution du contenu d'apprentissage 49
Figure I.15: Différents types de périphériques pris en charge par l'afficheur de contenu 50
Figure I.16: Framework big data pour un apprentissage intelligent 51
Figure I.17: Identification des sources big data d'un système e-learning 53
Figure I.18: Extraction, collecte et agrégation des données massives à l'aide du Framework Flume 54
Figure I.19: Un modèle big data pour l'e-learning dans un environnement cloud 55
Figure II.1: Evolution du cloud computing 59
Figure II.2: Infrastructure simplifiée d'un cluster 62
Figure II.3: Diagramme de séquence d'allocation d'une machine virtuelle sur un cluster cloud 63
Figure II.4: Services du cloud 63
Figure II.5: composants des différents services du cloud 64
Figure II.6: Modèles de déploiement du cloud 66
Figure II.7: Architecture simplifiée de la technologie de la virtualisation 68
Figure II.8: La Console de management web d'Amazon 74
Figure II.9: La croissance mondiale du volume de données entre 2006 et 2020 [63] 81
Figure II.10: Domaine du big data 82
Figure II.11: Caractéristiques du big data ou 5V 84
Figure II.12: Types de données en big data 85
Figure II.13: Architecture de Hadoop Distributed File System (HDFS) 91
Figure II.14: Le processus de fonctionnement de MapReduce 93
Figure II.15: Architecture de MapReduce 93

Figure II.16: Les composants du Framework Spark 95
Figure II.17:Architecture du Framework Apache Spark 97
Figure II.18: Architecture de Hadoop Yarn 98
Figure II.19: Architecture du service Apache Flume 100
Figure II.20: Architecture d'Apache Sqoop 101
Figure II.21: Apache Ambari interface 103
Figure II.22: phases historiques des systèmes de gestion des bases de données 104
Figure II.23: Les modèles de bases de données NoSQL 105
Figure II.24: Technologies big data 108
Figure II.25: Apprentissage automatique 109
Figure II.26: Arbre de décision 112
Figure II.27: Régression Linéaire 114
Figure II.28: Régression Logistique 115
Figure II.29: Modèle de Time Serie Analysis 118
Figure II.30: Exemple du Deep Learning 119
Figure III.1 : Architecture traditionnelle de l'e-learning 123
Figure III.2: Architecture N-Tier 124
Figure III.3: Approche de l'e-learning basé sur le cloud 127
Figure III.4: Couche d'infrastructure en tant que service 128
Figure III.5: Mode interactif de l'e-learning en mode SaaS 132
Figure III.6: Approche d'intégration des systèmes d'apprentissage en ligne, du big data et du cloud 147
Figure III.7: Méthodologie d'intégration du big data et des systèmes d'apprentissage en ligne 150
Figure III.8: Acquisition des données dans un système e-learning 151
Figure III.9: Architecture du big data pour un système d'apprentissage en ligne 154
Figure III.10: L'infrastructure distribuée d'un cluster Hadoop 156
Figure IV.1: Différents types de systèmes de recommandation [134] 164
Figure IV.2: Filtrage collaboratif [138] 168
Figure IV.3 : Algorithme Apriori 172
Figure IV.4: Organigramme de l'algorithme Apriori [144] 173
Figure IV.5: Base des données des transactions & items fréquents ordonnés 175
Figure IV.6: Construction de la structure FP-tree 176
Figure IV.7: Mécanisme de recommandation des cours 178
Figure IV.8: Architecture du système de recommandation de cours 179
Figure IV.9: Phase de préparation des données 180
Figure IV.10: Diagramme de classes du système de recommandation de cours 182
Figure IV.11: Flow chart du système de recommandation de cours 183
Figure IV.12: Résultat de l'algorithme Apriori 188
Figure IV.13: Distribution de fréquence relative aux inscriptions aux cours 190
Figure IV.14: Performance de FP-growth vs Apriori 191

Figure IV.15: Temps d'exécution de FP-growth vs Apriori en fonction du nombre des transactions 192
Figure IV.16:Méthode de visualisation des règles d'association : Scatter Plot 193
Figure IV.17: Méthode de visualisation des règles d'association : graph 193
Figure IV.18: Méthode de visualisation des règles d'association : Grouped Matrix 195
Figure IV.19: Technologies Big Data 199
Figure IV.20: Architecture du système de recommandation de cours à grande échelle 202
Figure IV.21: Architecture du cluster Hadoop 204
Figure IV.22: Apache Yarn Web UI : Configuration des nœuds du cluster 206
Figure IV.23: Script d'exécution du système de recommandation de cours 208
Figure IV.24: Etude comparative des performances entre Spark, Weka et RSStudio 211
Figure IV.25: Spark Web UI 213
Figure IV.26: Vue chronologique des événements Spark 213
Figure IV.27: Direct Acyclic Graph 214
Figure IV.28: Yan Web UI 215
Figure IV.29: Interface web de HDFS 216
Figure IV.30: Interface web Ganglia 217
Figure IV.31: Interface mobile Ganglia 217
Figure IV.32: Les nœuds du cluster Hadoop supervisés par Ganglia 218
Figure IV.33: Supervision du réseau via Ganglia 218
Figure IV.34: Interface récapitulative d'un nœud via le système Ganglia 219

Liste des tableaux

Tableau II-1: Avantages et inconvénients des services du cloud 65
Tableau II-2: Comparaison entre le big data et les SGBDR 87
Tableau III-1: Comparaison entre le modèle traditionnel de déploiement et celui du cloud . 133
Tableau IV-1: Résultats de l'algorithme Apriori 189
Tableau IV-2: Performance de FP-growth vs. Apriori 191
Tableau IV-3: Configuration des nœuds du cluster Hadoop 205
Tableau IV-4: Résultats de l'algorithme PFP-growth: Itemset fréquents 208
Tableau IV-5: Résultats PFP-growth: règles d'association 209
Tableau IV-6: Résultats de prédiction du système de recommandation de cours 210

Liste des Abréviations

A

API : Application Programming Interface

B

BDD : Base De Donnée**s**

BI : Business Intelligence

C

CMS : Content Management System

D

DUT: Diplôme Universitaire de Technologie

E

EAO : Enseignement Assisté par Ordinateur

EIAH : Environnements Informatiques pour l'Apprentissage Humain

EIAO : Environnement Interactif d'Apprentissage avec Ordinateur

E-learning : Electronic learning

ETL : Extract, Transform and Load

G

GPS: Global Positioning System

M

MOOC : Massive Open Online Course

L

LMS : Learning Management System

N

NoSQL : Not only SQL

P

PaaS : Platform as a Service

R

RDD : Resilient Distributed Dataset

REST : Representational State Transfer

S

SOAP : Simple Object Access Protocol

SQL : Structured Query Language

STI : Systèmes Tutoriels Intelligents

SaaS : Software as a service

SGBDR : Système de Gestion de Bases de Données Relationnelles

H

HDFS : Hadoop Distributed File System

HTML : HyperText Markup Language

HTTP : HyperText Transfer Protocol

I

IDE : Integrated Development Environment

IaaS : Infrastructure as a Service

IHM : Interface Homme-Machine

J

JDBC : Java Database Connectivity

JSON : JavaScript Object Notation

U

UML : Unified Modeling Language

UNESCO : United Nations Educational, Scientific and Cultural Organization

X

XML : Extensible Markup Language

Y

YARN : Yet Another Resource Negotiator

Introduction Générale

Contexte de la thèse

Motivation de la thèse

Problématique de la thèse

Objectifs de la thèse

Organisation du rapport

1. Contexte de la thèse

De tout temps, les Environnements Informatiques pour l'Apprentissage Humain (EIAH) évoluent d'un rythme exponentiel. Depuis que, l'ordinateur, dans un premier temps, l'internet, ensuite, le cloud computing et le big data, actuellement, ont fait leur apparition pour un usage éducatif. Par conséquent, les universités sont demandées de changer leur infrastructure pour pouvoir intégrer et exploiter la puissance de ces technologies, surtout dans le contexte de l'enseignement à distance.

Le but principal de l'EIAH est de favoriser l'apprentissage chez l'apprenant en informatisant des situations pédagogiques. Cependant ces EIAH n'ont jamais été parfaits. Ils demandent constamment des ajustements et des efforts d'amélioration parce que nous sommes, tout simplement, confrontés à de nombreuses réalités changeantes et des mutations technologiques très rapides. Alors, les systèmes, les structures, et les processus d'e-learning doivent rester souples pour pouvoir être ajustés à ces changements. Actuellement, les EIAH font face à plusieurs défis tels que l'adaptation des ressources pédagogiques à chaque situation d'apprentissage, la valorisation des données d'apprentissage, et le traitement des quantités massives et parfois en temps réel des interactions des apprenants avec le système e-learning. En vue d'apporter les meilleures réponses, nous mettons à disposition des bonnes pratiques informatiques, à savoir : le cloud computing, le big data.

Notre travail se situe dans le cadre des travaux de recherche sur les EIAH et s'intéresse à l'intégration continue des nouvelles technologies de l'information et de la communication (TIC) dans l'enseignement à distance. Plus particulièrement, ce projet de thèse s'inscrit dans le contexte du développement et de l'amélioration continue des EIAH au sein des organismes de formation à travers la mise en œuvre d'une plate-forme d'apprentissage en ligne fiable, robuste, et adaptative répondant au mieux aux besoins évolutifs et aux choix de l'apprenant en prenant en considération les dimensions de mutualisation ressources, de disponibilité, de qualité du service d'apprentissage et de meilleur rapport coût-efficacité également.

Pour ce faire, il s'est avéré indispensable d'incorporer les innovations technologiques modernes dans les EAIH. Ainsi, nous choisissons d'utiliser les services fournis par le cloud computing, comme un modèle informatique prometteur, permettant l'externalisation sur des serveurs à distance d'un fournisseur cloud des ressources matérielles, logicielles et plates-

formes de manière à les rendre accessibles sur internet sous forme de services tout en facilitant leur acquisition et en optimisant leur exploitation. D'ailleurs, nous utilisons les dernières avancées technologiques à travers la mise en place des technologies de pointe de l'écosystème big data dans les systèmes d'apprentissage en ligne. Le big data offre, en plus des puissances de calcul massivement parallèle et des capacités de stockage distribué, des méthodes et des algorithmes sophistiqués dédiés à l'apprentissage automatique pour traiter et extraire des connaissances à partir des divers types de données produites par les systèmes e-learning, y compris les informations du profil d'apprenant, ses activités, ses préférences, ses résultats, etc. Ce travail mis en place un moteur de recommandation, l'application la plus populaire du monde big data, afin d'aider et de guider l'apprenant à identifier et sélectionner facilement les ressources pédagogiques les plus pertinentes en lui offrant un système intelligent capable de générer un catalogue de cours adaptés à ses intérêts et à son niveau cognitif via l'analyse des données historiques des activités d'apprentissage.

2. Motivation

Ces dernières années, le monde informatique connaît un développement technologique considérable sur les plans d'acquisition des ressources et de la gestion et la manipulation des données. Ce développement est marqué à la fois par l'évolution du modèle du cloud computing et l'avènement de la nouvelle génération des technologies big data. L'ensemble de ces technologies change complétement les pratiques informatiques en introduisant des services à la demande, des approches distribuées et des architectures scalables pour l'exploitation et l'analyse répartie des données.

La motivation principale de nos travaux de recherche est d'améliorer et d'enrichir les EAIH via l'intégration des dernières innovations technologiques. De ce fait, notre thèse est consacrée à la migration des systèmes e-learning vers l'environnement cloud pour minimiser leur coût de déploiement de maintenance. D'ailleurs, elle est destinée à incorporer et exploiter des capacités de stockage réparti, des puissances de calcul intensif et d'analyse de haute performance offertes par des technologies émergentes du big data afin de mettre en place des solutions distribuées et parallèles dédiées aux plates-formes d'apprentissage en ligne.

Etant donné que nous sommes extrêmement motivés par la nécessité de posséder un cadre conceptuel solide pour que les universités et les organismes de formation en ligne puissent

adopter le cloud computing et le big data d'une manière efficace, nous proposons un certain nombre d'approches qui permettent à ces technologies prometteuses de trouver leur place dans le domaine d'EIAH. D'ailleurs, vue l'importance d'un système adaptatif pour ce type d'environnement, ce travail met surtout l'accent sur le besoin d'appliquer des technologies big data aux données produites par les EIAH pour développer un système de recommandation capable d'adapter les cours au profil et aux préférences de chaque apprenant.

3. Problématique

Actuellement, le domaine des EAIH connait un développement remarquable et les systèmes d'apprentissage en ligne ne cessent d'évoluer, grâce au développement rapide des technologies de l'information et de la communication qui ont transformé la manière d'enseigner et d'apprendre. Cependant, la plupart des formes éducatives existantes sont devenues inadaptées aux exigences du progrès technologique et au développement de d'enseignement à distance. Ces derniers doivent être en mesure de rattraper les changements d'apprentissage, surtout avec l'apparition des nouvelles approches et des modèles informatiques, la forte demande d'apprentissage et l'énorme croissance du nombre d'apprenants. Par conséquent, les institutions qui offrent l'apprentissage en ligne devraient s'adapter aux attentes et aux besoins de l'apprenant en termes de personnalisation des cours, d'adaptation des activités pédagogiques, et de qualité du service d'apprentissage en général.

En fait, les dimensions des systèmes e-learning se croissent d'un rythme exceptionnel et les défis en ce qui concerne l'acquisition et la gestion automatisée des ressources, l'optimisation de leur utilisation, et la puissance de calcul et de stockage mettent en évidence la nécessité d'utiliser une plate-forme répondant aux exigences évolutives des systèmes d'apprentissage en ligne ainsi qu'au contrôle des coûts.

En fait, les EIAH sont confrontés à de nombreux challenges, notamment la diversité des supports d'apprentissage mises à disposition, l'adaptation des ressources pédagogiques aux apprenants, ainsi que les problèmes liés à la complexité du traitement et de l'analyse du volume important des données produites par les systèmes d'apprentissage à distance. Ces ensembles de données, leur vitesse de génération, leur valorisation, l'hétérogénéité des types et formats de ces données, ainsi que la contrainte liée aux limites des approches classiques de gestion de l'information sont autant de facteurs qui poussent les spécialistes et les chercheurs

à découvrir et cerner de nouvelles façons ou parfois des façons complémentaires permettant de répondre aux contraintes du traitement massif et de l'analyse en temps réel des bases de données géantes. Ces facteurs ont abouti aujourd'hui à l'avènement de solutions innovantes fondées sur des approches distribuées et parallèles via des technologies big data. Ces technologies, qui se développent très rapidement, sont conçues dans le but de surmonter les défis de gestion des données volumineuses et des calculs intensifs.

4. Objectifs et contributions

Les environnements informatiques pour l'apprentissage humain dédiés à l'apprentissage en ligne présentent plusieurs défis, notamment ceux liés à la performance du service d'apprentissage, à la vitesse de traitement et à la capacité limitée des ressources disponibles. D'un point de vue économique, les systèmes e-learning existants, hébergés sur des serveurs web classiques, sont souvent situés à l'intérieur des établissements d'enseignement ou des universités, ce qui nécessite un taux d'investissement très élevé pour leur déploiement et leur maintenance. Ces plates-formes font également face aux problèmes de l'optimisation d'utilisation des ressources et de la pérennité des données.

Pour répondre à ces challenges, nous présentons un certain nombre de travaux de recherche basées sur les dernières révolutions numériques. En effet, cette thèse a pour objectif de proposer, dans un premier temps, des solutions conceptuelles en termes de modèles, d'architectures, et de méthodologies. Dans un deuxième temps, elle vise à concevoir, développer et déployer des solutions technologiques via la mise en application des Frameworks, outils et techniques du big data dans le domaine de l'e-learning.

Premièrement, pour bien cerner les objectifs de cette thèse, nous avons effectué une étude approfondie sur les travaux existants dans le champ de recherche d'EIAH. Nous avons constaté qu'il n'y a pas une approche concrète pour intégrer le paradigme du cloud computing dans ce type d'environnement. Une telle approche permet de bien situer les EAIH au niveau de l'infrastructure du cloud afin de faciliter le mécanisme d'intégration. D'où vient l'idée de proposer un modèle en couche visant à incorporer les services de l'informatique en nuage dans les plates-formes d'apprentissage en ligne en permettant aux institutions d'enseignement à distance de migrer leurs plates-formes d'apprentissage traditionnelles vers un modèle de cloud. Dans ce contexte, le cloud vise à répondre à la forte demande d'apprentissage en ligne

de la part des universités et des différents organismes de formation qui ont besoin de développer, d'exécuter et de déployer des applications performantes et à moindre coût. Il a également pour but de permettre aux professionnels de l'e-learning de faire face à la multiplicité des appareils (desktop, smartphones, tablettes), de systèmes d'exploitation, de langages de programmation et de Frameworks de développement.

Deuxièmement, pour faire face à l'explosion exponentielle du volume des données générées par les plates-formes e-learning, surtout l'énorme croissance de nombre d'apprenants, la diversité des contenus éducatifs et l'hétérogénéité des ressources mises à disposition. Nous avons opté pour l'utilisation des capacités de stockage réparti et des puissances de calcul massivement parallèle de la nouvelle génération des technologies big data qui s'appuient sur une architecture scalable et entièrement distribuée capable de supporter une montée en charge importante. Pour ce faire, un modèle d'intégration a été proposé en vue de profiter du traitement à haute performance, de l'analyse avancée et de l'intelligence de l'apprentissage machine offerte par les outils big data. Ces derniers vont permettre à un système e-learning d'évoluer et de s'adapter tout seul à des situations différentes en fonction du profil de l'apprenant et de ses interactions. En effet, grâce à l'intégration du big data dans les EIAH, un système d'apprentissage a désormais la capacité de prendre des décisions et de faire des prédictions à travers des modèles et des algorithmes de Machine Learning, qui fait partie intégrante du big data. Comme mise en œuvre des technologies big data dans le champ d'enseignement à distance, nous avons mis en place un système e-learning adaptatif sous forme d'une application répartie en se basant sur des techniques sophistiquées d'apprentissage automatique appliquées aux données historiques des activités des apprenants. Dans ce sens, nous avons choisi de développer un système de recommandation de cours en nous appuyant sur l'approche du filtrage collaboratif et les traces d'apprentissage, afin de prédire une liste de cours appropriés pour chaque apprenant en fonction de son profil, de son style d'apprentissage et de son niveau cognitif. L'opérationnalisation de ce système est faite à l'aide des techniques des règles d'association, des algorithmes Apriori et FP-growth, et de l'environnement R. De plus, afin d'améliorer les performances de traitement et de stockage, cette thèse présente une version à grande échelle du moteur de recommandation de cours en implémentant les technologies open sources récentes du monde big data, surtout la plate-forme Hadoop et le Framework Apache Spark.

5. Organisation du rapport

Cette thèse s'articule autour de quatre chapitres, à savoir : les EIAH, le cloud computing et le big data, et les deux derniers chapitres de ce travail seront consacrés à nos contributions en termes d'approches, d'architectures et d'implémentations. Le contenu de ce rapport de thèse sera organisé comme suit :

Le premier chapitre présente les environnements informatiques pour l'apprentissage humain pour introduire les axes de recherche principaux dans ce domaine. Nous commencerons par donner un bref aperçu sur l'état de l'art afin de situer notre recherche et de bien orienter nos activités scientifiques dans le contexte des EAIH. Il traite les concepts constituants le fondement de nos travaux, y compris les définitions, l'évolution des EIAH et ses origines depuis la première génération de l'enseignement par correspondance en passant par les médias audiovisuels (télévision et radio), l'informatique et le multimédia et télécommunication, ainsi que les différentes formes de l'enseignement à distance basées sur les technologies de l'information et de la communication et les tendances actuelles de l'e-learning et leurs caractéristiques. De plus, ce chapitre exposera les travaux de recherche relatifs à l'intégration des services du cloud computing et à l'utilisation des technologies big data dans les d'EIAH.

Le deuxième chapitre comporte deux parties. La première partie est consacrée à la présentation du concept du cloud computing. Dans ce chapitre, nous nous focaliserons sur les caractéristiques et les services de l'informatique en nuage. Ensuite, nous nous intéresserons aux différents modèles de déploiement du cloud. Puis, l'accent sera mis sur les technologies innovantes qui ont contribué à l'avènement de ce nouveau paradigme. Un aperçu sur les avantages et les défis du cloud sera également présenté. Pour clôturer cette partie, nous aborderons les principaux leaders ou fournisseurs des services cloud. Dans la seconde partie de ce travail, nous présenterons l'écosystème big data en décrivant son histoire, ses caractéristiques, ainsi que les différents types de données supportées par ce dernier. Ensuite, cette partie fera une comparaison sommaire entre les systèmes de gestion des bases de données relationnelles et les systèmes de gestion des données distribuées comme le mouvement NoSQL et le système HDFS. Enfin, cette partie introduira les techniques avancées de l'apprentissage automatique et leurs algorithmes associés.

Le troisième chapitre est réservé à ce que nous avons pu apporter comme innovation dans le domaine d'EIAH en termes de modèles et de méthodologies qui vont permettre au cloud et au big data de trouver leur place au sein de l'enseignement à distance. La première partie de ce chapitre concerne la première contribution qui consiste à proposer un modèle permettant l'intégration des systèmes e-learning dans l'environnement du cloud computing. Dans cette partie, nous présenterons également une petite comparaison entre le modèle traditionnel de d'hébergement et celui du cloud. Ensuite, nous discuterons les principaux bénéfices et les contraintes de cette intégration. La dernière partie du chapitre sera destinée à la présentation d'une nouvelle approche de l'e-learning basée sur l'écosystème big data en incorporant les services offerts par le cloud. D'ailleurs, nous décrirons la méthodologie proposée détaillant la démarche à suivre et les principales étapes du processus d'intégration. De plus, cette partie proposera une architecture distribuée en vue de manipuler des données provenant des systèmes e-learning parallèlement sur une infrastructure de clusters capables de répartir la charge sur plusieurs nœuds. Nous conclurons par quelques avantages de l'exploitation des puissances de calcul, des capacités de stockage et des méthodes d'analyse avancées des technologies big data dans les plates-formes d'apprentissage en ligne.

Le quatrième chapitre de ce travail de recherche concerne la mise en œuvre des technologies big data dans les plates-formes d'apprentissage en ligne basées sur le LMS Moodle. La première partie de ce chapitre présentera la conception et le développement d'un système de recommandation de cours en implémentant des techniques d'apprentissage automatique, qui fait partie intégrante du paradigme big data, afin d'aider l'apprenant à sélectionner les ressources pédagogiques les plus pertinentes en fonction de ses intérêts à travers l'analyse des traces d'apprentissage. Ensuite, nous aborderons en détail la base de données, l'architecture, et les outils de développement employées dans l'élaboration du moteur de recommandation proposé. Enfin nous analyserons les résultats obtenus et nous les interpréterons via des techniques avancées de visualisation de données fournies par l'environnement RStudio. Dans la seconde partie nous nous concentrerons sur la mise en place d'un système de recommandation de cours à grande échelle. Il s'agit d'utiliser une approche entièrement distribuée en s'appuyant sur des dernières technologies open source du big data. La réalisation de notre solution de recommandation se fait à l'aide de la plate-forme Apache Hadoop. Cette partie présente également la configuration et la préparation d'environnement de cluster mis en

place dans le déploiement du système de recommandation développé. Ensuite, nous présenterons les résultats obtenus qui montrent les performances du choix d'utilisation du big data par rapport aux méthodes classiques de traitement d'information. Ce chapitre se termine par une présentation de quelques outils de monitoring utilisés pour superviser l'état de cluster en permettant de mesurer les performances et la consommation de ressources.

Pour finir, nous présenterons une conclusion générale et les perspectives pour la suite de ce travail de recherche.

I. CHAPITRE I : Environnements Informatiques pour l'Apprentissage Humain

Introduction

Etat de l'art : évolution et définitions

- Environnements informatiques pour l'apprentissage humain (EIAH)
- Brève histoire des EIAH
- E-learning
- Formation Ouverte et A Distance (FOAD)
- Les tendances de l'e-learning

Etat de l'art : travaux existants

- Architecture de l'e-learning basée sur le cloud
- Infrastructure du cloud computing au service de l'e-learning
- Adaptation du contenu e-learning basée sur le cloud
- Framework d'intégration du big data dans l'e-learning
- Méthodologie d'intégration du big data dans l'e-learning
- Modèle d'adaptation du big data et du cloud computing avec l'e-learning

I.1 Introduction

Depuis l'avènement de l'internet et l'apparition des technologies du web, les environnements informatiques pour l'apprentissage humain ne cessent de se développer en permettant la création d'un nouveau style d'apprentissage à distance. Ces environnements ont changé la manière d'enseignement et d'apprendre tout en éliminant les barrières de temps et de lieu de formation. Ils ont pour but de construire des connaissances chez l'apprenant par l'intérmediaire des situations pédagogiques informatisées sous forme de cours, d'exercices, ou d'autres activités.

Dans ce chapitre, nous aborderons les grandes lignes de l'état de l'art des EIAH. Ensuite, nous traiterons le concept des environnements informatiques pour l'apprentissage humain, ainsi que celui de l'e-learning. Dans ce chapitre l'accent sera mis sur les définitions des différentes formes de l'apprentissage en ligne en s'appuyant sur les technologies de l'information et de la communication. Nous présenterons également les travaux réalisés dans le cadre de l'intégration du paradigme big data et les services du cloud computing dans le domaine d'enseignement à distance.

I.2 Etat de l'art : évolution et définitions

I.2.1 Environnements informatiques pour l'apprentissage humain

Nous allons présenter dans cette section les grandes périodes qui ont marqué l'histoire de l'enseignement à distance, ainsi que la formalisation des concepts de ce domaine. Pour bien comprendre cet état d'art, nous allons adapter une perspective historique qui montre l'évolution des systèmes informatiques dédiés à l'apprentissage.

Une longue histoire derrière ce qu'on appelle actuellement l'E-learning. Il s'agit de l'évolution des formes de l'éducation via des progrès technologiques au fil du temps. Ce qui a encouragé l'émergence d'une diversité de termes sur les sujets liés à l'enseignement à travers les technologies de l'information et de la communication.

Dans ce contexte, un environnement informatique pour l'apprentissage humain est un ensemble d'outils, de systèmes et de plates-formes d'apprentissage reposant sur l'utilisation des nouvelles technologies de l'information et de la communication pour permettre à l'apprenant de réaliser son parcours de formation en se libérant des contraintes de temps et de distance. Par conséquent, l'apprenant peut suivre ses cours à tout moment, n'importe où et à

partir de n'importe quel appareil. De plus, l'instructeur ou l'enseignant n'est pas présent physiquement sur le lieu de formation et le processus d'apprentissage se déroule généralement de manière asynchrone. Beaucoup de termes ont été utilisés pour décrire ce mode d'enseignement à distance :

- EIAH (Environnements Informatiques pour l'apprentissage humain) ;
- E-learning (Apprentissage électronique) ;
- Apprentissage en ligne ;
- EAO (Enseignement Assisté par Ordinateur) ;
- TLS (Systèmes Tutoriels Intelligents)
- EIAO (Environnement Interactif d'Apprentissage avec Ordinateur)
- EAD (Education ou Enseignement A distance) ;
- FAD (Formation A Distance) ;
- FOAD (Formation Ouverte A Distance) ;
- ENT (Environnement Numérique de Travail) ;
- LMS (Système de gestion de l'apprentissage);
- M-learning (Apprentissage Mobile);
- MOOC (Massive Open Online Courses);
- …

Grâce aux nouvelles technologies de l'information et de la communication (NTIC), les établissements d'enseignement, universités, et spécialistes de ce domaine peuvent être capables de cibler les lacunes d'apprentissage et d'apporter des réponses à plusieurs problèmes. Également, ces technologies offrent différentes solutions afin d'améliorer les plates-formes d'apprentissage en ligne en termes de qualité de service éducatif, de performance, d'optimisation de stockage et traitement, de l'adaptabilité et la personnalisation des ressources pédagogiques aux choix et aux préférences de chaque apprenant.

L'informatique est intégrée dans tous les domaines, y compris l'enseignement. En fait, les plates-formes e-learning sont principalement des applications informatiques dédiées à l'éducation[1]. La communauté francophone, qui s'intéresse au développement de l'enseignement à travers les technologies informatiques, convient d'appeler tous systèmes,

plates-formes ou logiciels destinés à l'apprentissage par Environnements Informatiques pour Apprentissage Humain (EIAH).

I.2.2 Brève histoire des EIAH

L'enseignement à distance a été développé essentiellement pour rendre l'enseignement accessible à distance à tous ceux qui ne peuvent pas suivre l'enseignement présentiel, que ce soit pour des raisons d'éloignement géographique, de manque de ressources financières, ou par manque de temps.

Depuis l'apparition des langages de programmation, Plusieurs systèmes d'apprentissage ont été créés. Ceux-ci visent à informatiser l'activité pédagogique afin de faciliter aux apprenants l'accès à une formation sans besoin d'être présents physiquement.

Les termes utilisés pour désigner e-learning sont trop variés. Donc pour une meilleure compréhension de l'évolution de ce mode d'enseignement à distance, il faut faire un rappel des grandes étapes de son histoire qui comprend trois grandes générations.

I.2.2.1 Première génération : l'enseignement par correspondance

L'enseignement à distance a vu le jour la première fois sous forme de cours par correspondance en Angleterre au milieu du 19ème siècle (1840). C'étaient surtout des cours destinés aux adultes n'ayant pas pu suivre l'enseignement secondaire ou supérieur. Ces cours ont été élaborés au profit des citoyens britanniques résidant très loin de l'université qui ne pouvaient pas se déplacer pour suivre des cours en présentiel.

Aux Etats-Unis, le mouvement des services de cours par correspondance débutera en 1873 à l'université Illinois Wesleyan de Bloomington. Deux ans plus tard, des professeurs de certaines universités américaines fondèrent l'université par correspondance (Correspondence University) d'Ithaca. En fait, William Rainey Hayer qui fut considéré comme le « père » de l'enseignement par correspondance aux Etats-Unis. Il a créé des cours par correspondance d'hébreu. Puis, il a organisé un département des études à domicile (Chicago Extension Department) lorsqu'il devint président de la nouvelle Université de Chicago en 1892. Après la réussite de l'exemple de l'université de Chicago, l'enseignement par correspondance s'étendit rapidement à d'autres établissements d'enseignement partout aux Etats-Unis.

L'enseignement par correspondance connait un grand développement sur le continent européen. En France, la première école privée d'enseignement par correspondance appelée

Eyrolles a vu le jour en 1877. En 1939, La France a créé une structure nationale d'enseignement par correspondance appelée Centre National d'Enseignement à Distance (CNED)[1]. Le CNED a pour mission de dispenser des formations à distance destinées aux apprenants de la maternelle, de l'enseignement supérieur en formation initiale ou en formation professionnelle continue, et pour les préparations aux concours de recrutement, etc. Aux Pays-Bas, on peut situer les débuts de cette nouvelle forme d'enseignement en 1890. En Belgique, la création de l'Institut d'Avenir dédié à l'enseignement par correspondance remonte à 1904. En Grande Bretagne, l'établissement Foulks Lynch (1884), le Skerry College (1880), l'University Correspondence College de Cambridge a vu je jour en 1887 pour soutenir ce type d'apprentissage. Nous pouvons constater que d'enseignement par correspondance en Europe a connu un grand progrès dans la seconde moitié du XIXème siècle[2].

I.2.2.2 Deuxième génération : Les médias audiovisuels (télévision, radio, cassettes)

Les médias audiovisuels ont contribué largement à l'essor de l'apprentissage à distance. En fait, la deuxième génération de l'enseignement à distance est caractérisée par l'utilisation de la radio (1920) et la télévision (1950). A ce moment, les télédistributeurs, les télévisions, et l'audioconférence devenaient de plus en plus populaires par plusieurs universités et jouaient un rôle pédagogique important.

A partir de 1920, la radio a commencé d'être utilisée par quelques universitaires en Europe pour radiodiffuser des programmes éducatifs. En 1929, la société ASIMMIL[2] est créée dans le but de proposer aux apprenants des contenus pédagogiques dans un format papier, puis seront complétés par des séances d'apprentissage diffusées par la radio.

En 1969, le gouvernement britannique créé l'Open University, la première Université ouverte utilisant divers types de médias (surtout la radio et la télévision) dans les cours à distance et permettant à un plus grand nombre d'apprenants d'achever leurs études dans différents niveaux d'enseignement. Elle représente l'exemple type de l'EAD de la seconde génération car elle rationalise, industrialise et planifie en divisant les fonctions d'enseignant en différents rôles confiés à des personnes différentes : experts académiques auteurs de cours, pédagogues

[1] http://www.cned.fr/
[2] http://fr.assimil.com/

et techniciens réalisateurs de cours, tuteurs facilitant les progrès des apprenants et évaluateurs[3].

Dans les années 1970-1980, la méthode d'apprentissage de langue ASSIMIL passe vers un système d'apprentissage par cassettes (Figure I.1)[4].



Figure I.1: Supports apprentissage ASSIMIL 1970-80

I.2.2.3 Troisième génération : L'informatique, le multimédia et la télécommunication

Dans les années 1970-80, l'enseignement à distance connait une nouvelle mutation, grâces au développement des technologies destinées à l'éducation, à l'utilisation de télécommunications, et à l'intégration d'informatique dans les établissements d'enseignement.

Les premiers programmes éducatifs d'enseignement assisté par ordinateur ont émergé dans les années 1970. En 1980, la France a lancé un plan gouvernemental, appelé "Informatique pour tous", visant à utiliser les applications informatiques pour renforcer l'apprentissage numérique chez les apprenants. Pour ce faire, des ordinateurs Thomson MO5 et T07 ont été installés dans les écoles (Figure I.2).Par la suite, les universités ont continué à utiliser les programmes informatiques et les différents moyens de communication à des fins éducatifs pour véhiculer les connaissances à distance [4].



Figure I.2: Les ordinateurs Thomson MO5 et T07 dédié à l'apprentissage (1980)

L'évolution des logiciels d'apprentissage et des plates-formes éducatifs s'est accompagnée du développement de nouveaux supports informatiques : des interfaces homme-machine, des outils multimédia comme la carte son 1989, des systèmes de stockage (CD-Rom, disque dur, etc.) permettant l'accès et le stockage des données. Elle s'est accompagnée également du développement de technologies de télécommunication (réseau informatique) permettant l'accès aux ressources à distance.

La télécommunication a connu une grande révolution dans les années 1990, surtout avec l'arrivé de l'internet. Depuis, les technologies du web ont été utilisées dans tous les domaines, y compris l'éducation pour améliorer la qualité de l'enseignement à distance, faciliter l'accès aux ressources éducatives et offrir un apprentissage plus interactif. L'avènement de ces médias interactifs a changé les pratiques d'apprentissage à distance, notamment en termes d'interaction, de communication et de travail collaboratif. En effet, le mode d'interaction devient très riche et l'apprenant peut communiquer avec son tuteur via des emails, des forums, du chat, des blogs, etc. L'apprenant peut, aussi, exploiter les ressources pédagogiques sur internet et de s'auto-évaluer sur le web.

Cet ensemble d'innovations technologiques a participé à la diffusion des programmes d'éducation numérique et a bouleversé les pratiques traditionnelles d'enseignement, et a contribué grandement à l'amélioration du système éducatif. En particulier, Il a permis de diminuer le coût, d'élargir l'accès à l'éducation en éliminant les contraintes de temps et d'endroit [5].

Dans ce contexte, les organismes internationaux y compris l'Unesco, l'Unicef et la Banque mondiale, dans leur rapport relatif au programme d'évaluation économique des technologies éducatives, ont conclu que l'utilisation des technologies modernes augmentait la qualité de formation et permettait une offre éducative économiquement pertinente. Ils ont également encouragé l'usage du numérique dans les écoles et les universités.

L'histoire de l'apprentissage via les technologies informatiques est marquée par plusieurs phases dans lesquelles beaucoup d'applications et plates-formes web et mobile ont été élaborées. Le schéma suivant (Figure I.3) décrit l'histoire de développement des systèmes d'apprentissage au fil du temps.

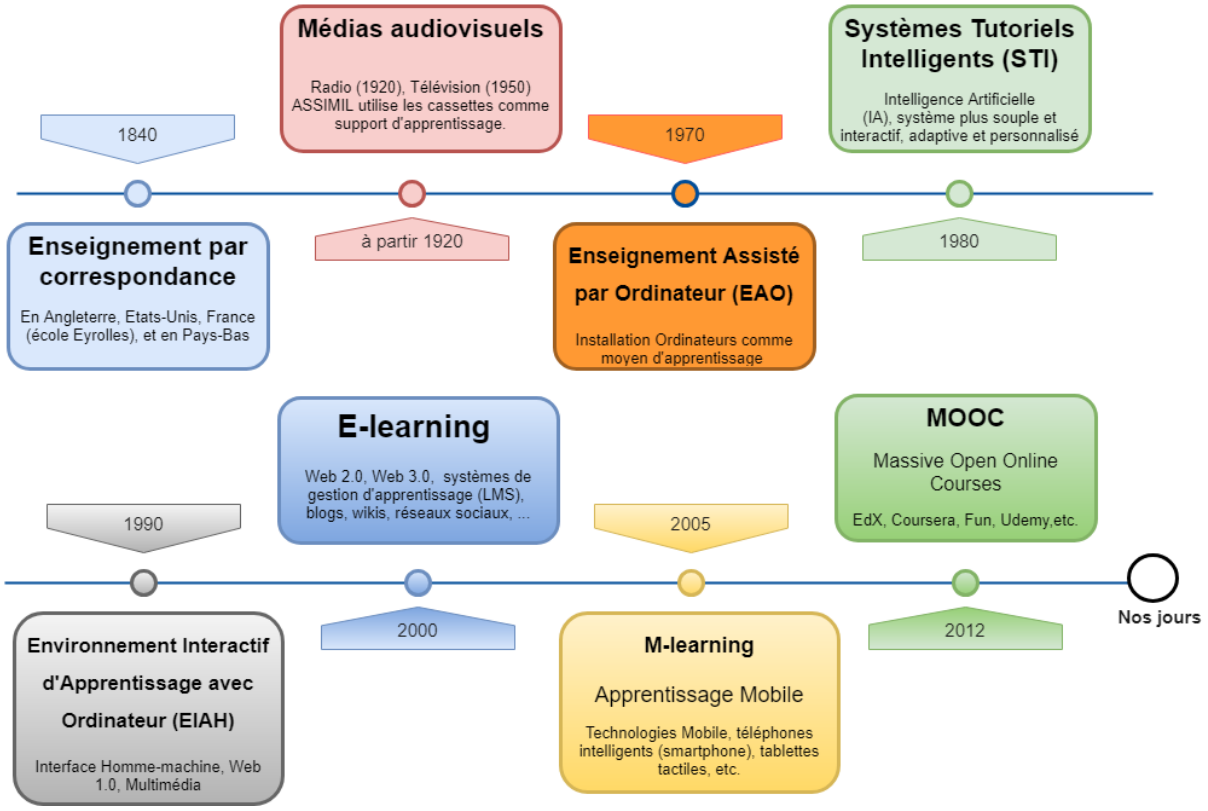


Figure I.3: Brève histoire de l'enseignement à distance

I.2.2.4 Enseignement Assisté par Ordinateur (EAO)

L'évolution technologique a donnée naissance à de nouveaux modes d'apprentissage. L'enseignement à distance a connu une forte demande grâce notamment à l'apparition de nouveaux moyens de communication. Les années 1970-1980 ont été marquées par le développement de l'Enseignement Assisté par Ordinateur (EAO), en anglais Computer based training (CBT), avec la création des programmes informatiques destinés à l'éducation.

I.2.2.5 Systèmes Tutoriels Intelligents (STI)

Par la suite, en 1980, on a vu apparaître des Systèmes Tutoriels Intelligents (STI), en anglais "Intelligent Tutoring Systems", comme successeur de l'enseignement assisté par ordinateur (EAO). Les STI sont une convergence de plusieurs disciplines (intelligence artificielle, éducation, psychologie cognitive) pour fournir un système plus souple et interactif. Ces systèmes ont pour but de placer l'utilisateur au centre du processus d'apprentissage par l'adaptation des ressources pédagogiques aux besoins, au style d'apprentissage, et au niveau cognitif de chaque apprenant. Il s'agit de logiciels de formation (en ligne ou pas) dotés d'une intelligence artificielle capable de s'adapter à l'apprenant et d'offrir un enseignement

personnalisé, et peuvent interagir dynamiquement avec l'utilisateur en lui proposant des contenus éducatifs en fonction du mode d'apprentissage le plus adéquat. Plusieurs STI ont vu le jour et ont fait leurs preuves, parmi les premiers systèmes tutoriels intelligents développés dans les années 1980, on trouve SCHOLAR pour enseigner la géographie en Amérique de sud, BANDAID pour le langage de programmation Basic, et SOPHIE pour l'électronique[6].

Principalement, les STI sont dotés de trois modèles: le modèle du domaine, de la connaissance pédagogiques, et de l'apprenant [7]. L'architecture simplifiée représentant ces différents composants est illustrée dans la Figure I.4.

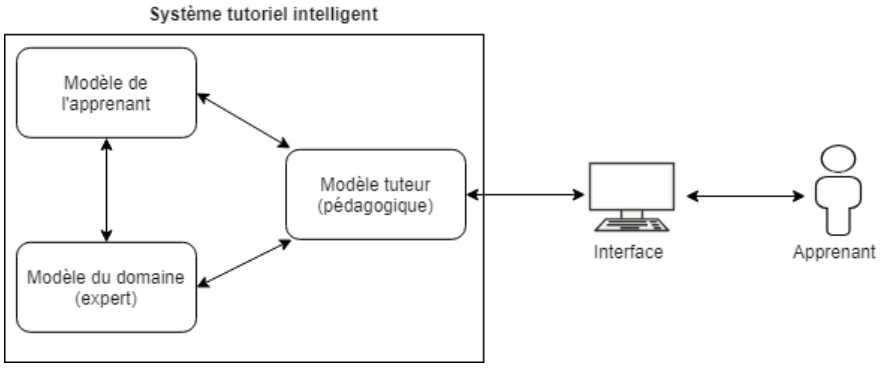


Figure I.4: Architecture traditionnelle d'un STI

En bref, Le STI est composé de trois modèles[8]:

Le modèle tuteurs ou modèle pédagogique : contient les stratégies pédagogiques du système. Autrement, il précise les tâches à exécuter par l'apprenant en fonction du profil de ce dernier. Il planifie les activités à faire par l'apprenant, lui fournit l'aide appropriée, et détermine quand et comment intervenir. En effet, le modèle pédagogique sert à gérer les interactions du système avec l'apprenant. Suite à ces interactions, le système intelligent propose des problèmes à l'apprenant, en le conduisant pour trouver la solution, en lui apportant l'aide lors qu'il en a besoin[9].

Le modèle de l'apprenant : permet de fournir au système d'apprentissage les informations pertinentes sur l'apprenant telles que ses compétences, son style d'apprentissage, ses préférences, etc. Il sert à construire un diagnostic qui pourra utiliser par le système tutoriel pour prendre la bonne décision selon le type de soutien (explications, questionnement, etc.) à proposer à l'apprenant. Dans ce sens, l'état de connaissance de l'apprenant est représenté comme étant un sous-ensemble de celle de l'expert[10].

Modèle du domaine ou modèle de l'expert : le tutoriel intelligent doit se comporter comme un expert du domaine enseigné. Pour cela, il a besoin d'une base de connaissances que le système tente à enseigner à l'apprenant ainsi que de mécanismes liés à leur exploitation. Ces connaissances permettent au tutoriel de raisonner et de tirer des conclusions. Pour être en mesure de cela, le système doit contenir les savoirs spécifiques au domaine ainsi que les techniques de raisonnement liés à la résolution des problèmes[11].

I.2.2.6 Environnement Interactif d'Apprentissage avec Ordinateur (EIAO)

L'évolution technologique (interface Homme-machine, web 1.0, HTML, Multimédia) a permis l'émergence, dans les années 1990-2000, d'un nouveau mode d'apprentissage appelé EIAO. Les EIAO ont désignés initialement "Enseignement Intelligemment Assisté par Ordinateur" et plus tard "Environnement(s) Interactif(s) d'Apprentissage avec Ordinateur", en anglais "Intelligent Computer Aided Instruction" (ICAT)[12]. Ils ont permis une grande flexibilité des EAO en introduisant une forte interaction entre l'apprenant et le système d'apprentissage.

Les EIAO sont essentiellement conçus dans but de résoudre le problème d'enseignement à l'aide d'outils informatiques. Ces systèmes permettent de proposer une solution à travers l'analyse des interactions causales entre divers éléments du problème, par exemple : "Pourquoi pleut-il ? Le climat de telle région, sera-t-il favorable à la culture du riz et pourquoi ?". En fait, les EIAO combinent plusieurs champs de recherches : l'informatique, la psychologie, la didactique, et les sciences d'éducation[13].

I.2.3 E-learning

A partir des années 2000, l'e-learning a vu le jour suite au progrès technologique. D'abord, l'e-learning a commencé d'être utilisé par les universitaires et dans les milieux scolaires. Ce terme est employé pour désigner la méthode d'apprentissage à travers les sites web éducatifs, les bases de données de ressources pédagogiques, et les forums et groupes par internet. L'essor extraordinaire du développement de technologies éducatives ont conduit à l'émergence de l'e-learning partout dans le monde. Dans ce contexte, les systèmes de gestion d'apprentissage LMS (Learning Mangement System) sont devenus de plus en plus populaire. Les LMS ont pour but principal de gérer l'ensemble des ressources éducatives et d'administrer les plates-formes d'apprentissage en ligne [14].

Il existe une multitude de définitions de l'e-learning, certaines se concentrent sur le moyen technologique pour éliminer la contrainte de distance, d'autres insistent sur les méthodes d'enseignement et les aspects pédagogiques.

En fait, l'e-learning se compose de deux termes. Premièrement, le "e" qui est l'abréviation d'électronique. C'est ce "e" qui est également utilisé dans les termes e-tourism, e-commerce, etc.. Deuxièmement, le terme "learning" ou apprentissage qui reflète l'acquisition personnelle des connaissances, des compétences, des attitudes, etc.

L'e-learning est traduit par l'apprentissage électronique. Avec l'avènement de la technologie internet, la notion "électronique" est remplacée par le terme "en ligne" pour constituer ce qu'on appelle actuellement "l'apprentissage en ligne".

L'e-learning est défini par L'Union Européenne (2001) comme "l'utilisation des nouvelles technologies multimédias de l'Internet pour améliorer la qualité de l'apprentissage en facilitant d'une part l'accès à des ressources et à des services, d'autre part les échanges et la collaboration à distance".

Il s'agit d'un style d'apprentissage qui repose sur la mise à disposition de contenus pédagogiques via un support électronique (internet, intranet, extranet, télévision interactive, etc.), dans le but de favoriser l'apprentissage et soutenir les différentes formes d'enseignement.

L'apprentissage en ligne n'inclut pas les supports numériques physiques de type DVD, clé USB, support papier, parce qu'ils ne sont pas disponibles en ligne.

De manière générale, L'e-learning ou enseignement électronique est une discipline où les théories de l'enseignement et les technologies de l'information et de la communication sont combinées pour permettre à l'apprenant de réaliser son cursus de formation par l'intermédiaire d'un réseau de communication. Ce qui ne remplacera pas l'enseignement présentiel mais améliore grandement l'efficacité de l'éducation.

I.2.4 Formation Ouverte et A Distance (FOAD)

Le terme Formation ouverte et à distance (FOAD), en anglais Open and Distance Learning (ODL). Il s'agit d'une formation entièrement à distance qui se déroule totalement

par internet. En effet, la FOAD permet à chacun d'apprendre de façon autonome, à son propre rythme, et de suivre des cours en toute liberté.

La FOAD ajoute la notion d'ouverture à l'opération d'apprentissage. Cela fait référence à une liberté d'accès aux contenus éducatifs mis à disposition de l'apprenant, une souplesse de formation qui s'adapte au rythme et à la disponibilité de l'apprenant. Dans le mode de formation ouverte, il n'est y a aucune restriction d'accès pour l'apprenant concernant la possibilité de choisir librement les cours et le temps de formation.

On trouve plusieurs définition pour FOAD. Le collectif de Chasseneuil [15] a défini la FOAD comme: "un dispositif organisé qui prend en compte la singularité des personnes dans leurs dimensions individuelle et collective et repose sur des situations d'apprentissage complémentaires et plurielles en termes de temps, de lieux, de médiations pédagogiques humaines et technologiques, et de ressources."

La FOAD peut aussi être définie[16] comme "un concept d'apprentissage en construction qui fait appel à une modularisation ('learning objects', objets ou grains d'apprentissage) des contenus didactiques, la modélisation des parcours et des activités, la prise en compte des styles cognitifs des apprenants, la mise à jour possible des contenus en temps réel, et l'ouverture vers des bases de données extérieures".

I.2.5 Les tendances de l'e-learning

Tout comme les technologies internet qui ne cessent d'évoluer, l'e-learning change également sans cesse pour proposer des moyens et des méthodes plus développés et plus adaptés aux besoins de d'apprenant. Actuellement, on constate le développement des nouvelles formes d'apprentissage en ligne, notamment :

I.2.5.1 Mobile learning ou M-learning

Les technologies mobiles évoluent constamment. Actuellement, tout un éventail d'appareils utilise ces technologies y compris les téléphones mobiles ou smartphones, les tablettes électroniques et les consoles des jeux portables. En effet, les téléphones intelligents font désormais partie de la vie quotidienne de milliards de personnes partout dans le monde. Selon L'UNESCO[17], "L'apprentissage mobile fait appel à la technologie mobile, seule ou combinée à d'autres technologies de l'information et de la communication (TIC), pour permettre d'apprendre en tout lieu et à tout moment. L'apprentissage mobile est une des

facettes de l'application des TIC dans l'éducation. Il repose sur une technologie plus abordable et plus facile à utiliser que l'ordinateur."

Sharples [18] a identifié trois directions de la définition de l'apprentissage mobile. La première définition : est centrée essentiellement sur les technologies mobiles. Dans ce sens, le m-learning peut être vu comme un prolongement de l'e-learning sur des technologies mobiles (Figure I.5).

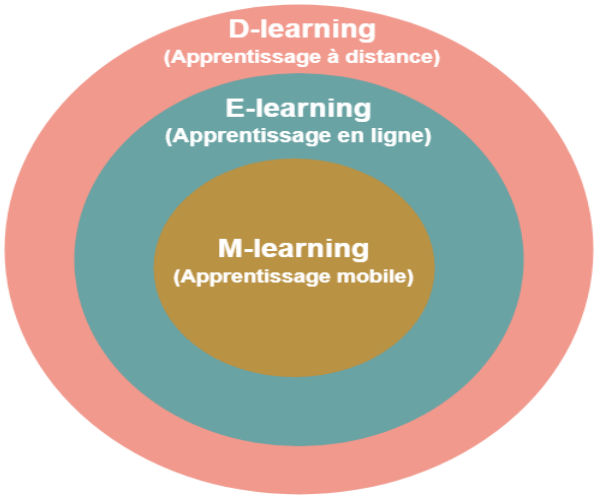


Figure I.5: M-learning est un prolongement de l'e-learning

La deuxième définition : se concentre sur la possibilité de l'apprentissage à l'extérieur des murs d'établissement de formation. Cette définition élimine la contrainte de l'endroit d'apprentissage pour être omniprésent. Par exemple, un voyage d'études ou une sortie scolaire. Dans ce contexte, Selon O'Malley et al. [19], "L'apprentissage mobile a lieu lorsque l'apprenant ne se trouve pas à un endroit déterminé, ou lorsqu'il profite des possibilités d'apprentissage offertes par les technologies mobiles".

La troisième direction : se focalise essentiellement sur la localisation de l'apprenant dans un contexte spécifique grâces aux technologies de capture de localisation tels que le système GPS, les réseaux sans fil, etc. afin d'analyser ce qui se passe autour de l'apprenant. En collectant les informations contextuelles, l'apprenant peut apprendre les concepts liés à son environnement[20].

I.2.5.2 Massive Online Open Courses (MOOC)

Le MOOC est un terme à la mode abondamment utilisé de nos jours pour décrire un nouveau modèle d'enseignement à distance qui a pour principal objectif de fournir des contenus d'apprentissage en ligne ouverts à tous, chose qui va changer fondamentalement le domaine éducatif.

Depuis 2012, nous assistions à l'apparition des MOOC. En effet, plusieurs plates-formes MOOC ont été développées dans ce domaine (Coursera, edX, FUN, etc.). Suite au succès éclatant des MOOC, ils deviennent la solution la plus commode aux internautes pour satisfaire leurs besoins d'apprentissage sans avoir besoin de se déplacer afin de suivre en présentiel ses cours et ses formations. En effet, beaucoup d'universités du monde entier, surtout américaines et francophones, se sont orientées vers l'utilisation des MOOC, comme alternative, en particulier dans le milieu d'enseignement supérieur.

L'acronyme MOOC, ou Massive Open Online Course, désigne en français, Cours en Ligne Ouvert et Massif. Il représente une nouvelle forme d'apprentissage, ouvert et à distance, par l'intermédiaire de l'internet. Il regroupe ainsi un ensemble de technologies informatiques et de méthodes pédagogiques pour permettre à l'apprenant, indépendamment de son niveau d'apprentissage, d'apprendre et d'acquérir des nouvelles connaissances et d'affiner ses compétences dans pratiquement tous les champs disciplinaires.

Le M de Massif signifie que le cours peut accueillir un nombre illimité des apprenants. Le O de Open décrit l'ouverture du cours à tout le public, indépendamment de l'origine de l'apprenant, de son niveau d'études, ou de n'importe quel critère. Le O de Online désigne que le cours est accessible en ligne. En dernier lieu, le C de Course ou Cours représente un ensemble cohérent de ressources pédagogiques, de stratégies, de méthodes d'enseignement et d'acteurs interagissant dans un contexte donné pour atteindre un objectif.

la Figure I.6 ci-après résume les caractéristiques du MOOC, à travers un schéma synthétique de l'ensemble des composantes de ce nouveau mode d'apprentissage en ligne. Qui sont "Cours", "En Ligne", "Ouvert" et "Massif".

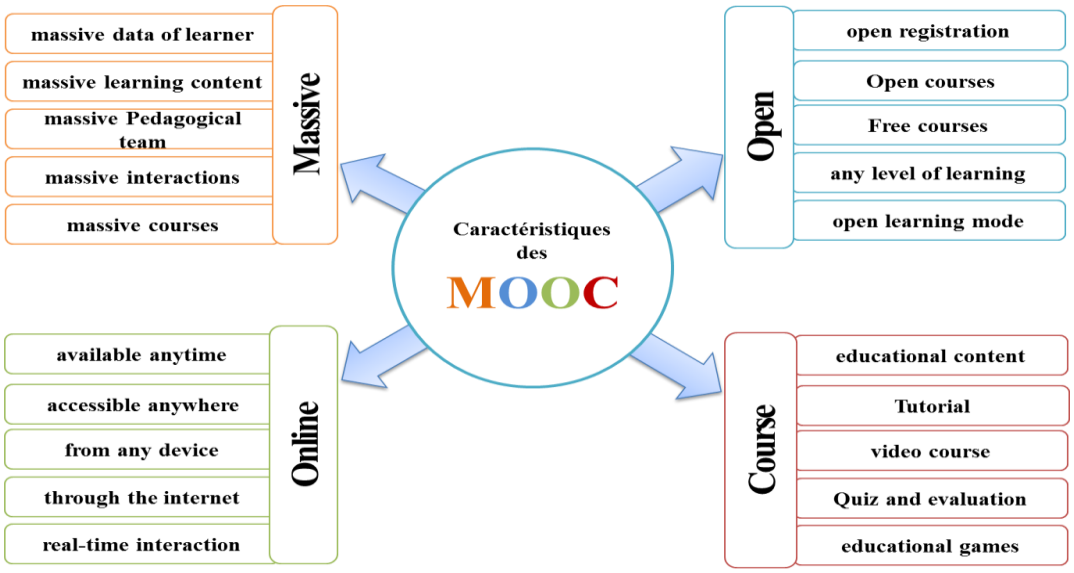


Figure I.6: Caractéristiques du MOOC

I.2.5.3 Digital learning

Le digital learning fait référence à la présence du numérique dans l'opération d'apprentissage, que cet apprentissage soit présentiel ou à distance (synchrone ou asynchrone). Il s'agit de l'usage des nouvelles technologies et des supports numériques dans la formation de manière globale. Ce mode d'apprentissage inclus l'e-learning, m-learning, etc.[5].

Parmi les technologies les plus fréquemment utilisées, le PowerPoint (96 %), les ressources numériques telles que les vidéos (68 %), le tableau blanc interactif (27 %), les espaces collaboratifs (26 %), les cours numérique (25 %), et les jeux numériques (16 %) [21].

I.3 Etat de l'art : travaux existants

Ce chapitre présente la description de plusieurs travaux sur le cloud computing et le big data en relation avec le domaine e-learning. Il explore les recherches qui ont été faites par des académiques et chercheurs pour incorporer ces trois grands paradigmes en proposant des architectures, des approches et des cas d'utilisation.

I.3.1 Architecture de l'e-learning basée sur le cloud

Le travail présenté par Anwar et al. [22] décrit l'architecture du système e-learning basée sur le cloud computing. Il vise à combiner les systèmes d'apprentissage en ligne et le modèle cloud computing en incorporant les composants de chaque paradigme dans un seul système.

Dans la Figure I.7, un nouveau concept est mis en évidence dans le domaine de l'éducation. Il s'agit du cloud computing dont le but d'améliorer l'évolutivité, la flexibilité et la disponibilité des systèmes d'apprentissage en ligne.

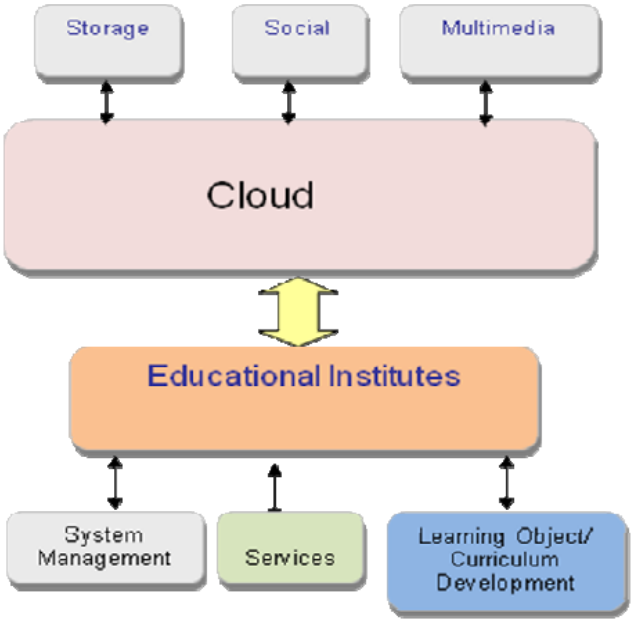


Figure I.7: Intégration du cloud computing dans l'éducation

Les auteurs ont évalué le modèle traditionnel d'e-learning, avec ses avancés et ses enjeux, et la possibilité de déplacer le système e-learning en ligne hors des écoles ou des universités, sur une infrastructure de cloud computing. Les institutions seront responsables du processus d'apprentissage, de la gestion et la diffusion de contenu, et le fournisseur du cloud s'occupe de la construction, de la maintenance, du développement et de la gestion et l'administration de l'architecture technique. Les services du cloud, exploités, sont facturés en fonction des ressources utilisées selon le nombre d'apprenants supporté par le système e-learning et la quantité des contenus pédagogiques.

L'architecture de l'e-learning basée sur le cloud est une architecture en couche comportant les éléments suivants (Figure I.8):

- Couche d'infrastructure : un pool d'hôte physique dynamique et évolutif,
- Couche de ressources logicielles : offre une interface unifiée pour les développeurs d'e-learning,

- Couche de gestion des ressources : permet un couplage des ressources logicielles et matérielles,
- Couche de service : contient trois niveaux de services (logiciel en tant que service, plate-forme en tant que service et infrastructure en tant que service),
- Couche d'application : fournit la production de contenu, le déploiement de supports d'apprentissage, l'apprentissage collaboratif, et l'évaluation et les fonctionnalités de gestion.



Figure I.8: Architecture du could e-learning[23]

Premièrement, La couche d'infrastructure se compose d'une infrastructure d'information et de ressources pédagogiques. L'infrastructure d'information comprend l'Internet / l'Intranet, les logiciels, le système de gestion de l'information et certains logiciels et matériels communs. Cette couche est située dans le niveau le plus bas de l'architecture proposée, et comprend la puissance de calcul de base comme la mémoire RAM, CPU et disque dur. L'utilisation de la technologie de virtualisation assure une gestion efficace et une utilisation optimisée des ressources serveur en termes de calcul, de stockage, et de réseau. Ces ressources virtuelles sont exploitées facilement par la plate-forme logicielle supérieure. Le pool d'hôtes physiques est dynamique et évolutif, c'est-à-dire qu'un nouvel hôte physique peut être ajouté afin d'améliorer la puissance de calcul des services du cloud computing.

Deuxièmement, la couche de ressources logicielles est principalement constituée le système d'exploitation et le middleware. Grâce à la technologie middleware[3], une variété de ressources logicielles sont intégrées pour fournir une interface unifiée pour les développeurs de logiciels, de sorte qu'ils peuvent facilement créer, tester, déployer leurs applications en utilisant les ressources logicielles disponibles sur cloud.

La couche de gestion des ressources vise à gérer les ressources matérielles pour qu'elles soient utilisées par les applications des utilisateurs. Elle effectue l'allocation de la mémoire RAM, de l'espace de stockage et du temps de calcul. Grâce à l'intégration de la virtualisation et de la stratégie de planification du cloud computing, l'utilisation des ressources devient plus souple et sur demande, ainsi la distribution du stockage et le traitement en parallèle offrent des services de haute performance.

La couche de service comprend trois niveaux de services : SaaS (logiciel en tant que service), PaaS (plate-forme en tant que service), IaaS (infrastructure en tant que service). Dans le mode de fonctionnement du cloud, l'utilisateur n'a besoin ni d'acheter les logiciels et/ou les matériels, ni de les maintenir et de les mettre à niveau. Car c'est le fournisseur du cloud qui s'occupe des tâches d'installation et de maintenances de l'infrastructure physique. Le client cloud doit simplement payer les ressources consommées.

La couche d'application est constituée par des applications spécifiques servant à l'intégration des ressources pédagogiques dans le modèle du cloud computing, notamment, les outils de création des cours interactifs et le partage des contenus éducatifs. Ainsi, l'apprentissage devient plus interactif et plus efficace que l'enseignement traditionnel. Cette couche s'occupe principalement de la production des contenus, du déploiement, et de la gestion des ressources pédagogiques[24].

I.3.2 Infrastructure du cloud computing au service de l'e-learning

La recherche de [25] a pour objectif principal de concevoir un environnement informatique évolutif, fiable et sécurisé offrant l'apprentissage en ligne via les modèles de services du

[3] Middleware: est un logiciel tiers permettant la création d'un réseau d'échange d'informations entre plusieurs applications ou système informatiques (source: https://fr.wikipedia.org/wiki/Middleware#cite_note-1).

cloud. Le travail utilise le cloud privé pour le déploiement de l'e-learning au sein des universités. Un modèle cloud privé permet aux établissements d'enseignement d'avoir un contrôle complet de la gestion des services, de la sécurité des données, des applications et des ressources.

La Figure I.9 illustre le modèle conceptuel des services d'apprentissage en ligne dans l'infrastructure du cloud privé au sein d'un établissement d'enseignement.

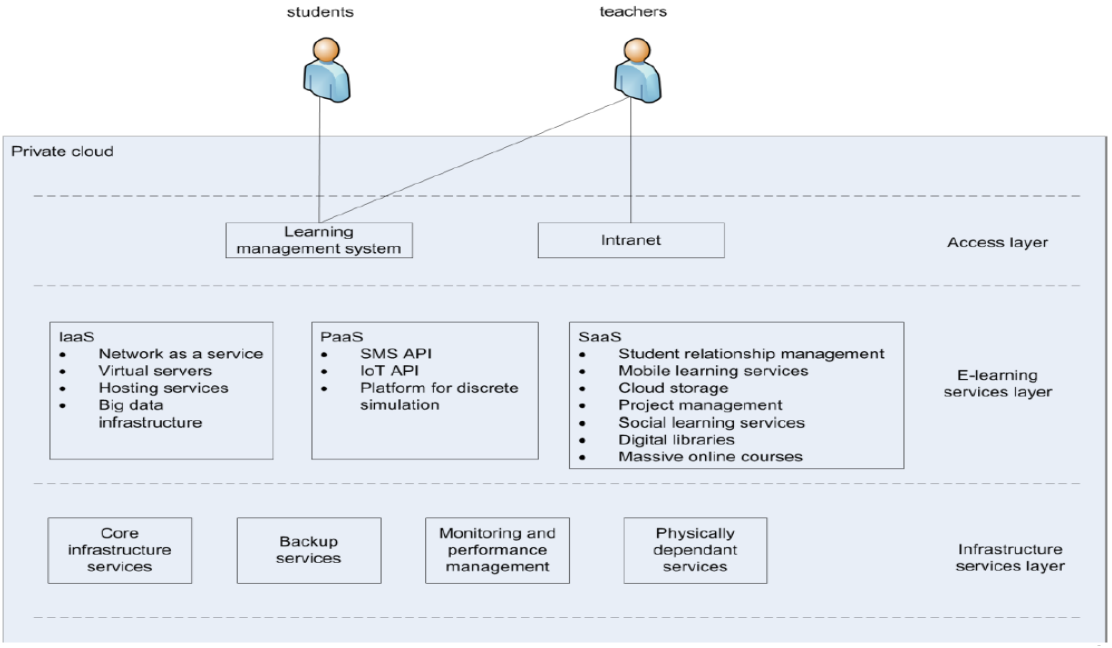


Figure I.9: Infrastructure du cloud privé pour e-learning

Les services d'infrastructure peuvent être regroupés comme suit :

1) *Core infrastructure services* (services d'infrastructure de base) : sont des services critiques qui doivent rester opérationnels pour permettre le bon fonctionnement du système d'apprentissage en ligne.

2) *Monitoring and performance management* (Services de surveillance et de gestion de performance) : Il s'agit des services de surveillance réseau, des ressources matérielles et des services. Ce sont des services optionnels permettant des fonctionnalités supplémentaires, mais ils assurent une détection et une correction rapides des problèmes de fonctionnement de l'infrastructure.

3) *Backup services* (services de sauvegarde) : ces services sont responsables de la création de copies de sauvegarde du système. Leur rôle est récupérer et de restaurer les données en cas de panne.

4) *Physical dependant services* : permettent l'administration, le contrôle d'accès et la gestion des ressources physique ou virtuel.

La couche de services e-learning est conçue pour fournir un environnement complet d'apprentissage. Elle supporte les trois types de services ci-après :

- Infrastructure en tant que service (IaaS) : offre aux enseignants la possibilité de concevoir une infrastructure spécifique pour chaque cours. Ce type de services permet de fournir aux étudiants des machines virtuelles préconfigurées et prêtes à utiliser. De plus, l'IaaS peut être utilisé pour fournir aux étudiants un environnement informatique adapté à leurs propres projets.
- Plateforme en tant que service (PaaS) : offre aux étudiants un environnement de développement intégré pour créer des logiciels en tant que services dans différents langages de programmation (web, mobile et desktop). Le modèle PaaS est adéquat aux cours de type développement. Par exemple, les applications du big data, ou les projets « Internet des objets » peuvent être enseignés avec les plates-formes PaaS.
- Logiciel en tant que service (SaaS) : est parfaitement adapté aux cours où les étudiants doivent apprendre à utiliser des logiciels spécifiques (Word, Excel, etc.). Ce modèle peut également être utilisé pour fournir des outils sophistiqués visant à améliorer et faciliter le processus d'apprentissage, tels que les systèmes de gestion de l'apprentissage, les systèmes de collaboration, les outils de gestion de projets, les bibliothèques numériques et bien d'autres.

Le modèle du réseau hiérarchique, décrit dans Figure I.10, est destiné à la conception de l'infrastructure réseau d'un cloud éducatif. Cette conception implique la division du réseau en couches de distribution et d'accès. Chaque couche fournit des fonctions bien déterminées. Le concept du réseau local virtuel (VLAN) est utilisé pour construire cette infrastructure cloud, car il est important d'isoler le trafic réseau pour chaque client. L'utilisation de réseaux locaux virtuels permet d'administrer le réseau de manière simple, flexible et peu coûteuse. Cette approche permet une certaine souplesse dans le développement de l'e-learning dans le cloud,

car les étudiants peuvent profiter d'un réseau, d'une infrastructure, de plates-formes de développement et de logiciels en tant que service, en fonction des spécificités de chaque cours.

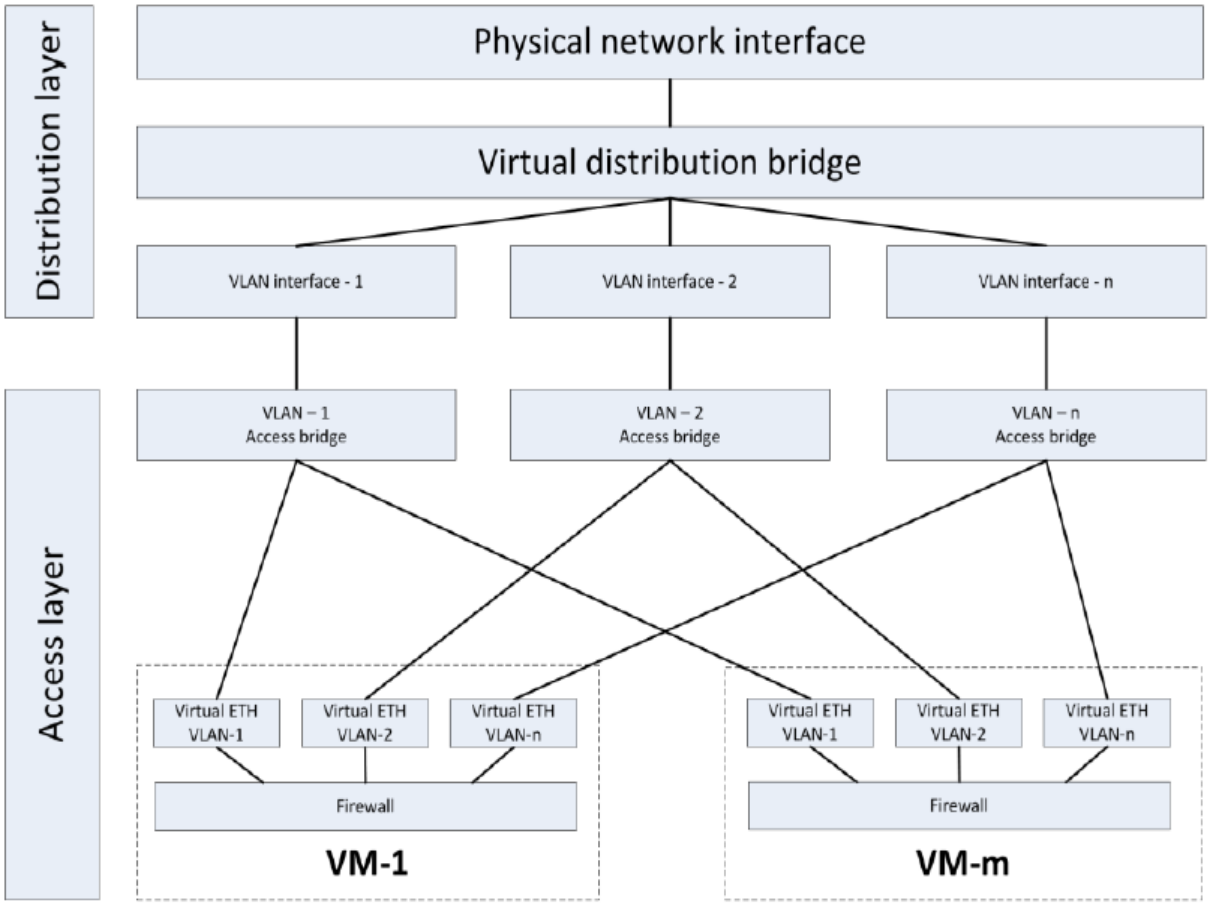


Figure I.10: Modèle conceptuel de l'infrastructure réseau du cloud

La construction de l'infrastructure cloud est réalisée par la solution OpenStack. Il s'agit d'une plate-forme open source pour déployer des infrastructures cloud computing. Elle principalement dédiée pour le cloud privé et public. Elle possède une architecture modulaire composée de plusieurs outils capables de créer, de gérer, et de pilôter un grand nombre de pools de ressources virtuelles (serveur, réseau et stockage) réparties dans les Datacenters via des interfaces web simples et prêtes à exploiter.

Ainsi, dans un premier temps, un enseignant définit des machines virtuelles (VM) en fonction des besoins de chaque cours et l'effectif des apprenants, par la suite, les étudiants peuvent

consulter ces machines qui sont disponibles pour chaque cours auquel ils s'inscrivent, et réserver les ressources nécessaires. Les activités des enseignants et des apprenants sont gérées via l'interface LMS Moodle, tandis que l'allocation des ressources cloud sont gérées par la plate-forme OpenStack. Cette approche permet aux établissements d'enseignement d'utiliser efficacement leurs ressources cloud, et aussi d'ajouter aisément des nouvelles ressources grâce à la grande élasticité de l'architecture cloud. Un exemple du processus de réservation de ressources cloud est présenté dans la Figure I.11, qui illustre la gestion de l'allocation des machines virtuelles en fonction du besoin des apprenants.

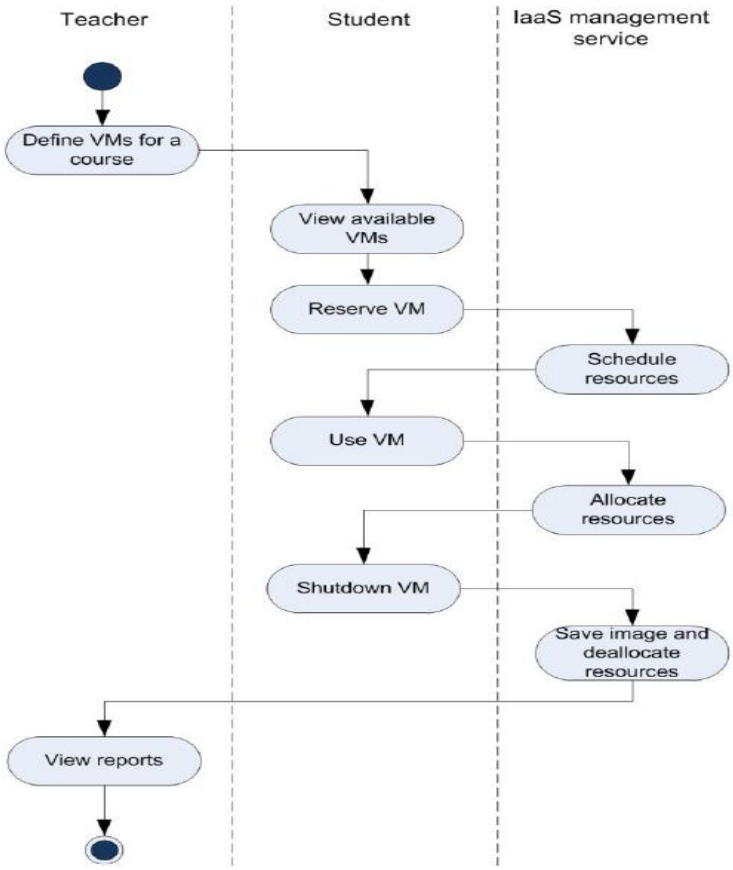


Figure I.11: Diagramme d'activité du service de réservation des ressources cloud

I.3.3 Adaptation du contenu e-learning basée sur le cloud

Dans le travail de Jeong et al[26], l'accent est mis sur les services du contenu d'une plate-forme e-learning basés sur le cloud en vue de fournir et de partager des formes améliorées des

contenus éducatifs. Ces derniers peuvent être du texte, des images, des vidéos et des objets 3D, des scènes de réalité virtuelle[4] (VR), etc. Ce travail a proposé un système d'apprentissage intelligent, orienté contenu utilisant les services du cloud et intégrant un certain nombre de fonctionnalités. La Figure I.12 décrit l'architecture générale des services du contenu intelligent d'un système e-learning dans le cloud.

Les enseignants dans les établissements d'enseignement sont responsables de la production de diverses formes de contenu d'apprentissage en utilisant le système auteur offerte par le cloud. Ces contenus sont gérés dans un format commun et compatible avec toutes les plates-formes y compris les ordinateurs personnels, les ordinateurs portables, les tablettes, les téléviseurs intelligents et les Smartphones. Le système fournit également un outil de visualisation de contenu permettant d'afficher le contenu éducatif, un moteur d'inférence et un système de sécurité.

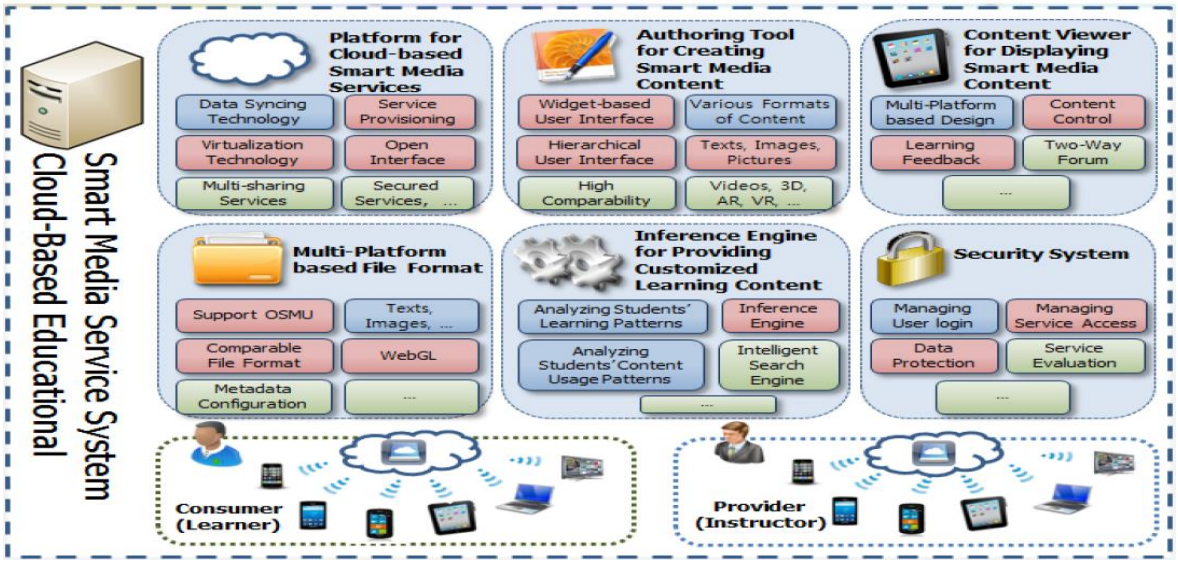


Figure I.12: Adaptation du contenu intelligent d'un système e-learning dans le cloud.

[4] « réalité virtuelle » (ou réalité simulée par ordinateur) renvoie typiquement à une technologie informatique qui simule la présence physique d'un utilisateur dans un environnement artificiellement généré par des logiciels. La réalité virtuelle crée un environnement avec lequel l'utilisateur peut interagir. La réalité virtuelle reproduit donc artificiellement une expérience sensorielle, qui peut inclure la vue, le toucher, l'ouïe et l'odorat (visuelle, sonore ou haptique). Source : https://fr.wikipedia.org/wiki/Réalité_virtuelle

Le système proposé comprend six fonctions principales requises pour le déploiement des contenus éducatifs intelligents basés sur le cloud :

1) Plate-forme cloud (Figure I.13) : Il s'agit d'établir et de configurer un cloud privé fournissant une infrastructure pour la réalisation d'un environnement complet du multimédia éducatif à travers la mise en place de plusieurs technologies tels que la synchronisation, la virtualisation. Elle exploite les trois services du cloud : IaaS, PaaS, et SaaS.

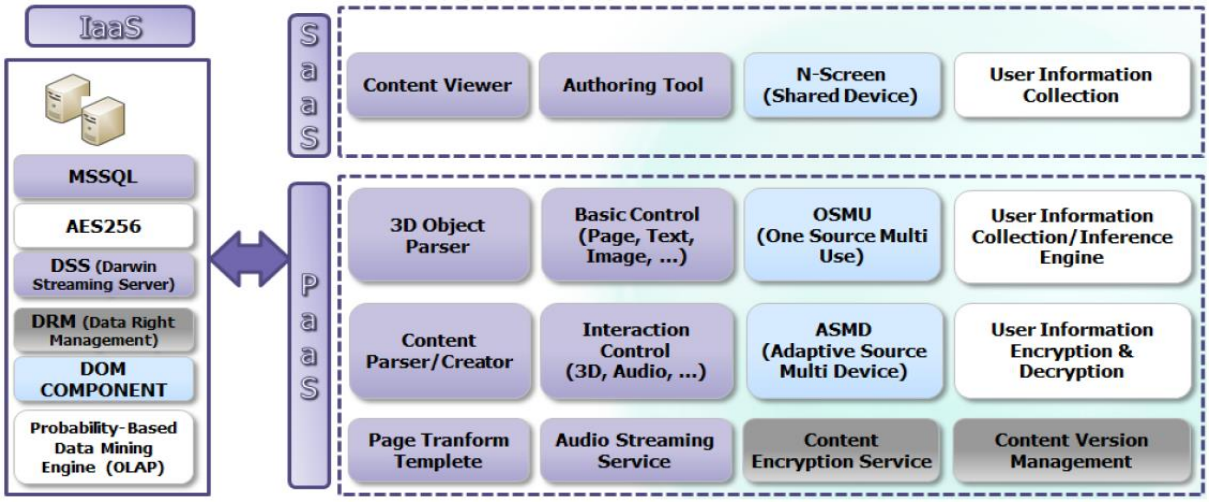


Figure I.13: Environnement de services des médias éducatifs basés sur le cloud

2) Format de fichier commun : permet de fournir du contenu multimédia supportable par différents types de terminaux en utilisant un format de document XML, HTML5, eXtensible 3-Dimensional (X3D) et JavaScript.
3) Outil de création (Authoring Tool) : le système fournit un outil de création permettant aux enseignants de créer divers types de contenu multimédia intelligent. La Figure I.14 montre les processus de création du contenu qui est divisé en trois phases : la création, la configuration et l'édition, et la distribution et le partage du contenu d'apprentissage.

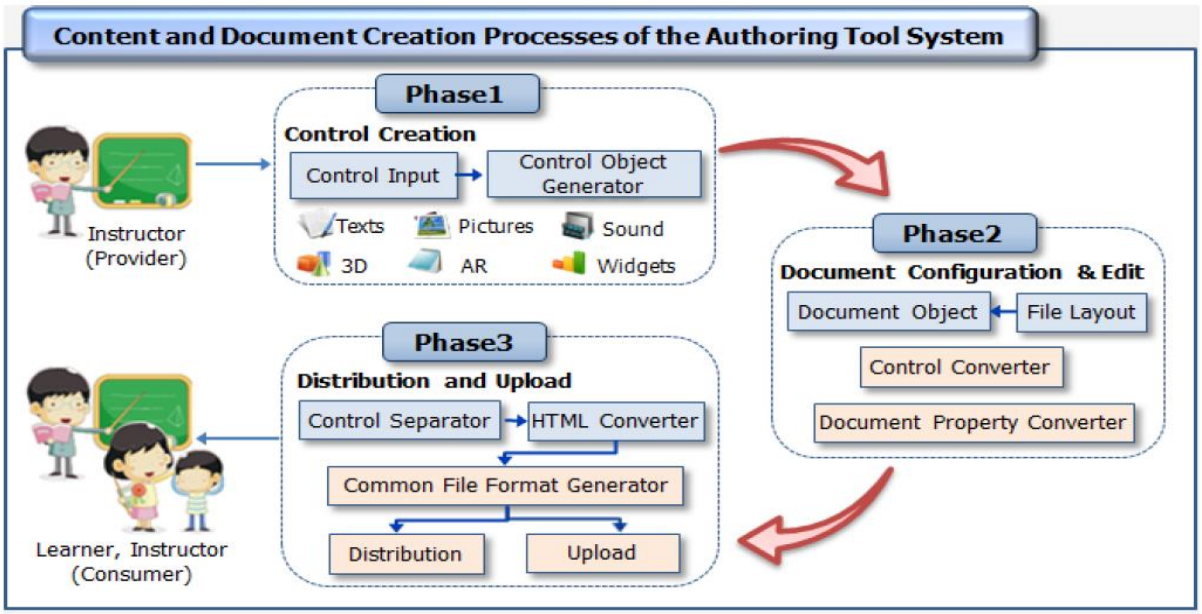


Figure I.14: Les processus de création et de distribution du contenu d'apprentissage

- Dans la phase 1, les utilisateurs peuvent effectuer des opérations de création de contenu sous diverses formes (pages web, blogs, document, etc.).
- Dans la phase 2, les utilisateurs peuvent définir le template et configurer d'autres paramètres.
- Au cours de la phase 3, le système de création convertit le document nouvellement créé au format de fichier commun, puis télécharge le fichier sur le serveur pour le distribuer aux périphériques clients.

4) Affichage de contenu (content viewer) : permet d'afficher des médias sur plusieurs plates-formes. La Figure I.15 présente un certain nombre de périphériques pris en charge par l'afficheur de contenu. Le contenu créé par l'outil de création (système auteur) est enregistré dans le format de fichier commun XML. Le système de visualisation affiche le contenu sur chaque périphérique en convertissant les données XML en HTML.

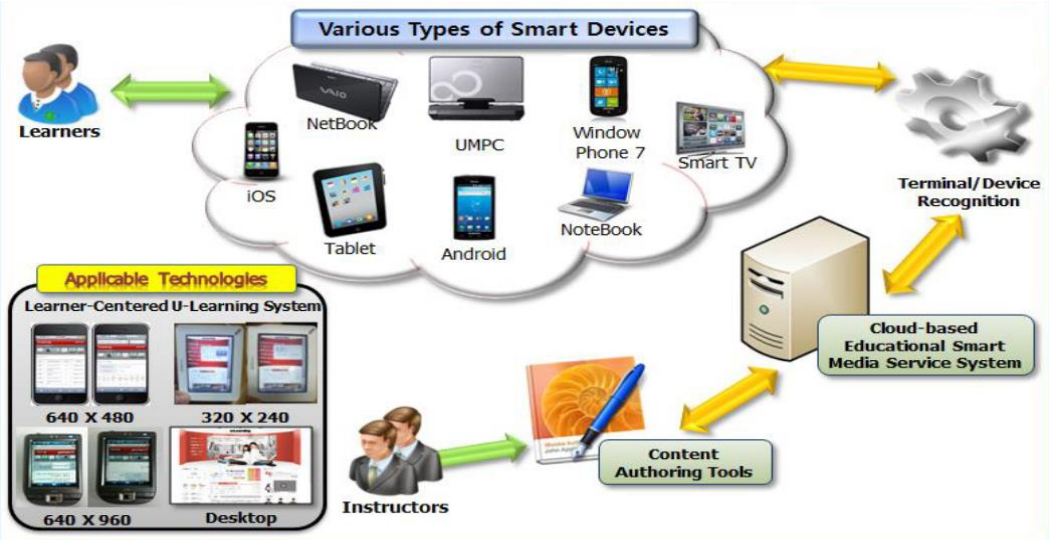


Figure I.15: Différents types de périphériques pris en charge par l'afficheur de contenu

5) Moteur d'inférence (inference engine) : ce moteur a pour but de fournir aux étudiants un contenu d'apprentissage personnalisé en analysant leurs préférences et leurs styles d'apprentissage. Afin de fournir des contenus pédagogiques adaptés au profil d'apprenant, le système d'apprentissage basé sur le cloud collecte toutes les informations liées à l'apprenant donné et les enregistre dans la base de données. Ces informations seront analysées, par la suite, pour produire les contenus les plus appropriés à chaque apprenant en fonction des choix et préférences de ce dernier.
6) Système de sécurité : est mis en place pour permettre non seulement de chiffrer et sécuriser les données et de contrôler l'accès des utilisateurs privilégiés, mais aussi pour protéger et résoudre les problèmes du réseau dans le cloud. L'algorithme AES (Advanced Encryption Standard), comme un standard de cryptage établi par le NIST [27], est utilisé pour la sécurité des contenus éducatifs.

I.3.4 Framework d'intégration du big data dans l'e-learning

L'intégration du big data dans les systèmes e-learning est l'une des pistes principales de ce travail de recherche qui vise à exploiter ces technologies innovantes pour résoudre certains problèmes des systèmes d'apprentissage en ligne (recommandation, prédiction, motivation). Dans ce contexte, le travail [28] aborde l'incorporation de l'écosystème big data dans le système e-learning en proposant un Framework en couches pour offrir un service d'apprentissage de qualité.

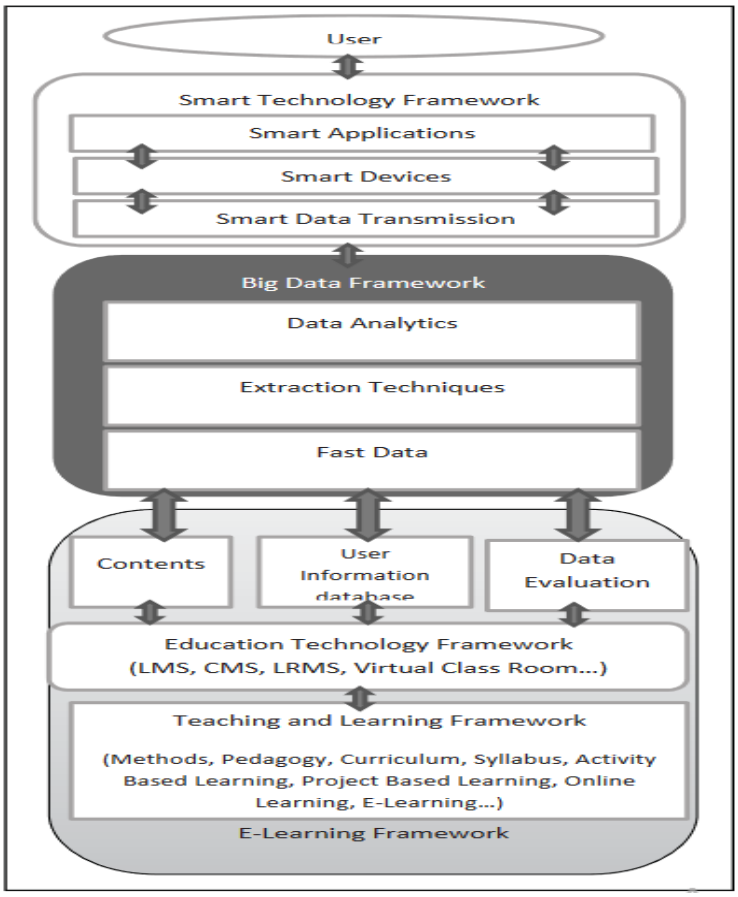


Figure I.16: Framework big data pour un apprentissage intelligent

Ce Framework facilite l'intégration de l'e-learning et du big data en bénéficiant des techniques d'analyse avancées de ce dernier. Il se constitue de trois couches (Figure I.16) :

- La couche e-learning constitue la couche inférieure du Framework proposé. Elle comprend des méthodes pédagogiques d'enseignement, des contenus éducatifs, et un ensemble de technologies dédiées à l'éducation tels que LMS, VCL (Virtual Class Room). Cette couche contient également des informations sur les apprenants, les données d'évaluation et de tests, etc. Ces données jouent un rôle très important pour identifier les styles d'apprentissage de l'apprenant et ses connaissances en vue d'adapter l'apprentissage au profil de chaque utilisateur.
- La couche big data : les données collectées à partir de la plate-forme e-learning sont passées à travers cette couche d'extraction afin d'extraire et de préparer ces données qui seront transmises, par la suite, à la couche d'analyse pour le traitement et l'analyse.

- Les technologies big data avancées forment la couche supérieure du Framework. Cette couche s'occupe au traitement et à l'analyse des données collectées et permet des algorithmes d'analyse prédictive, ainsi que la représentation des résultats.

I.3.5 Méthodologie d'intégration du big data dans l'e-learning

Le travail [29] vise à montrer comment le big data peut aider à résoudre certaines problèmes d'enseignement à distance en exploitant ses technologies dans l'analyse des contenus d'apprentissage afin d'assurer l'efficacité et la fiabilité des systèmes e-learning. Il utilise un certain nombre de techniques et outils big data, à savoir :

- Hadoop[5]: Il s'agit d'une infrastructure distribuée pour le stockage et le traitement de données massives à grande échelle.
- MapReduce[6]: Un Framework pour effectuer des calculs en parallèles sur plusieurs nœuds de cluster.
- HDFS[7] (Hadoop Distributed File System) : Un système de fichier distribué d'Apache Hadoop. Il est tolérant aux pannes et offre un accès à haut débit aux données des applications.

Il s'agit de proposer une méthodologie afin d'incorporer les technologies big data dans le contexte d'EIAH. Cette méthodologie comprend quatre étapes :

1. Sources du big data (Figure I.17) : consiste à identifier les sources des données big data et de déterminer les sources probables des données éducatives. Ces données peuvent être de divers types tels que des bases de données e-learning, des e-mails, des réseaux sociaux, etc.

[5] http://hadoop.apache.org/
[6] https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
[7] https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html



Figure I.17: Identification des sources big data d'un système e-learning

2. Extraction des données: Cette étape concerne la collecte des données à partir des différentes sources. Pour cela, Apache Flume[8] est utilisé comme Framework distribué qui est tolérant aux pannes afin d'agréger et de collecter efficacement des grandes quantités de données. Pour transmettre ces données massives via Flume, il est nécessaire de définir un agent regroupant un ensemble de composants tels que la source, les canaux (Channel), et les Sink (intercepteurs), qui sont chargés pour effectuer les tâches suivantes (Figure I.18) :
   1) Récupérer les données à partir d'une source.
   2) Transiter ces données sur un chemin
   3) Filtrer et réaliser des opérations sur ces données
   4) Les envoyer vers une destination

[8] https://flume.apache.org/

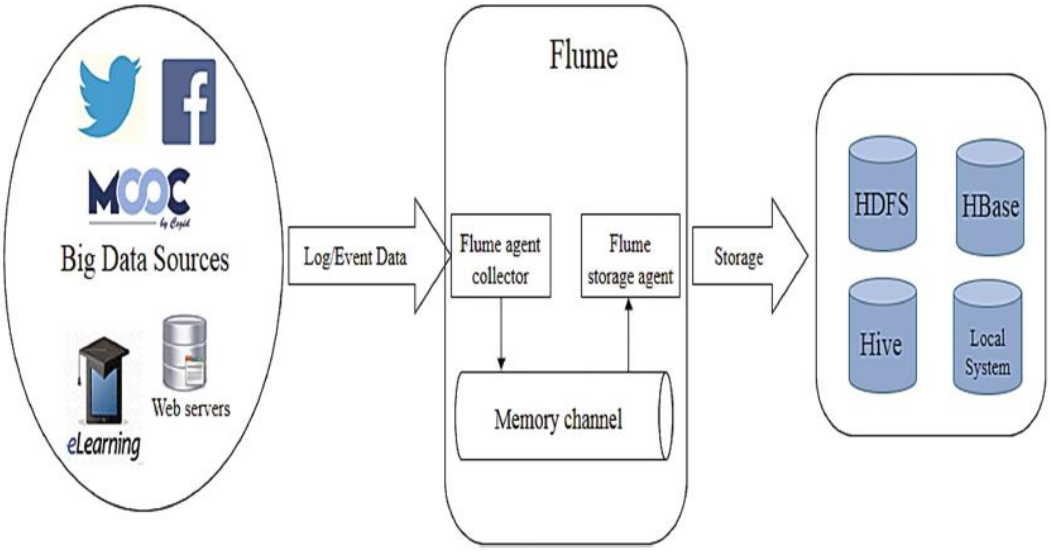


Figure I.18: Extraction, collecte et agrégation des données massives à l'aide du Framework Flume

3. Traitement du big data : Après avoir identifié les sources et faire la collecte de données, il est temps de les traiter. La première phase consiste à choisir le langage de programmation le plus approprié capable manipuler les différents types de données collectés. Par la suite vient le choix d'un algorithme adéquat permettant de retourner les résultats sollicités par l'utilisateur.
4. Visualisation des données : a pour objectif de présenter les données sous forme visuelle interactive. Cette étape cherche à rendre accessible et compréhensible le résultat de traitement et d'analyse des données. En d'autres termes, elle transforme un ensemble de données complexes en des représentations graphiques (graphes, histogrammes, etc.) simples et faciles à apprendre.

I.3.6 Modèle d'adaptation du big data et du cloud computing avec l'e-learning

Le travail de recherche [30] aborde l'utilisation du big data dans le contexte académique. Il propose un modèle pour adapter les technologies big data avec les plates-formes e-learning. Cela vise à intégrer les LMS (Learning Management System) existants déjà dans les universités avec le Framework Hadoop déployé en tant que SaaS. Pour pouvoir exploiter les données, d'un système LMS classique, qui reposent sur un SGBDR, Hadoop utilise deux méthodes pour cet objectif. La première méthode consiste en la migration de la base de données relationnelle LMS vers un système de fichier distribué HDFS (Hadoop Distributed File System). Le deuxième mécanisme consiste à transférer les données structurées de LMS

dans un Data Warehouse. Il existe plusieurs outils pour effectuer ces opérations, tels que Apache Sqoop ou ETL (Extract, Transform, Load) et des interfaces JDBC (Java Database Connectivity), qui sont capables de collecter les données provenant d'une source externe. Après l'intégration des données du LMS dans le Framework Hadoop, il est possible d'appliquer des conversions et du filtrage sur ces données, puis d'effectuer des analyses avancées, par exemple, prédire le comportement de l'apprenant, analyser les activités de ce dernier sur les réseaux sociaux, etc. Ces analyses peuvent contribuer grandement à l'élaboration d'un système d'apprentissage adaptatif et personnalisé à chaque situation d'apprentissage.

Il s'agit d'une architecture en trois couches associant le cloud hybride, un ensemble de solutions big data et les données d'apprentissage en ligne gérées par un LMS. La Figure I.19 montre le modèle big data en cloud pour les plates-formes e-learning des universités en Roumanie.

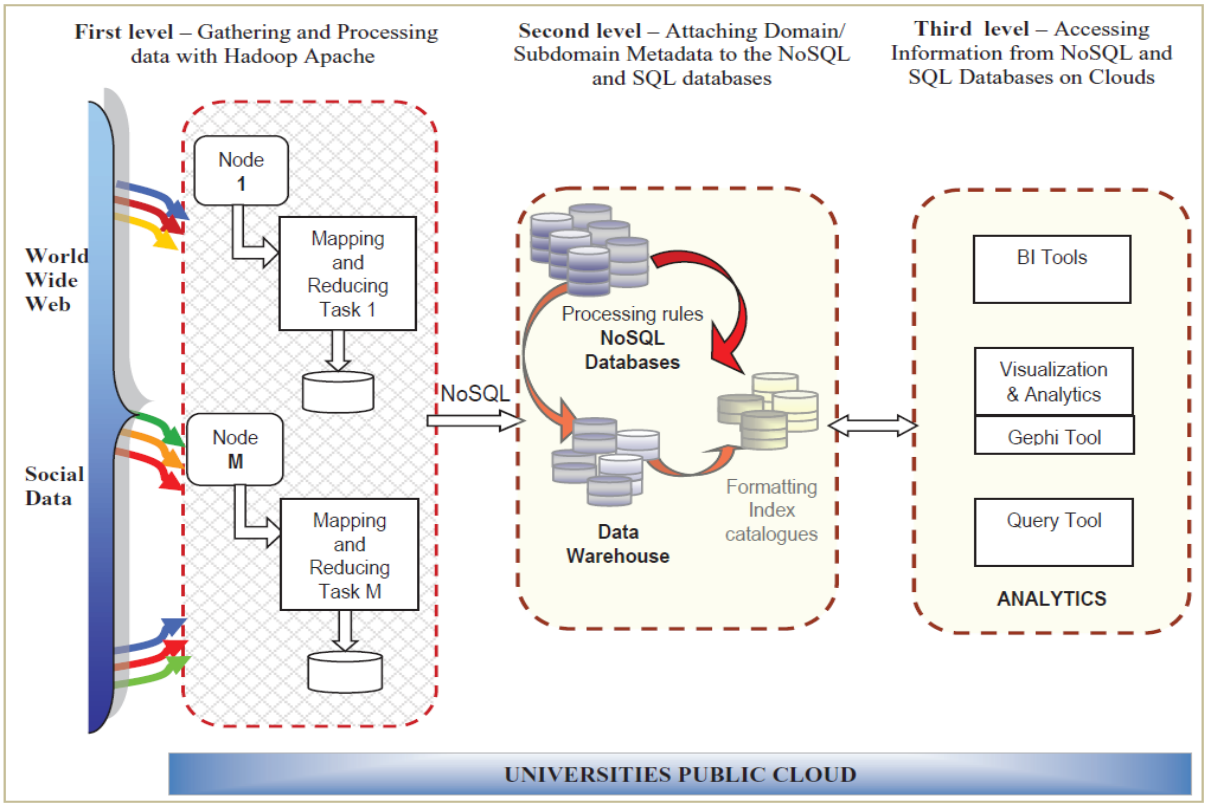


Figure I.19: Un modèle big data pour l'e-learning dans un environnement cloud

Le premier niveau est destiné à collecter tout type de données en se basant sur divers outils logiciels tels qu'Apache Flume (données non structurées), Apache Sqoop (données structurées), et les traiter à l'aide du Framework MapReduce sur un cluster Hadoop. Ce niveau utilise une infrastructure cloud adaptée à ce type de traitement. Le deuxième niveau a pour objectif de réaliser une classification des données stockées dans les nœuds du cluster Hadoop en utilisant des dictionnaires de données et de métadonnées. À cette phase les données sont indexées et cataloguées. En fin, le troisième niveau rassemble plusieurs technologies pour accéder aux données, y compris, des outils BI, des applications pour faire des requêtes sur les données et pour visualiser des résultats d'analyse.

1.1 Conclusion

Dans ce chapitre, nous avons fait un tour sur l'évolution des environnements informatiques pour l'apprentissage humain, à partir de leur apparition jusqu'au aujourd'hui, en décrivant les grandes périodes qui ont marqué leur histoire. Ensuite, nous avons défini ces environnements dans leurs diverses formes, tels que l'apprentissage à distance, l'e-learning, m-learning, et les MOOCs. D'ailleurs, une partie de ce chapitre a été consacrée aux travaux de recherche qui ont été réalisés dans le but d'intégrer les trois paradigmes : le cloud, le big data et l'apprentissage en ligne.

II. CHAPITRE II : Cloud Computing et Big Data

Cloud Computing

- Brève Histoire
- Définitions
- Caractéristiques
- Concept du cluster
- Services
- Modèles de déploiement
- Les technologies
- Avantages
- Défis
- Principaux acteurs

Big Data

- Courte Histoire
- Définition
- Domaine du big data
- Caractéristiques du big data
- Types de données en big data
- Comparaison entre SGBDR et big data
- Les technologies big data
- Bases de données NoSQL
- Méthodes et techniques du machine learning (apprentissage automatique)

II Cloud Computing

II.1.1 Introduction

Le terme cloud computing est apparu au début des années 2002, lorsqu'Amazon a proposé de louer à ses clients une partie des capacités de son parc informatique pour stocker des données. Le cloud offre aux utilisateurs des services numériques en utilisant les standards Internet et les technologies de virtualisation. Ces services numériques peuvent être de différentes natures : services matériels ou d'infrastructure (espace de stockage, puissance de calcul etc.); services de plateforme (environnements de développement, systèmes de gestion de base de données et serveurs d'application, etc.); services logiciels (applications et logicielles utilitaires pour les entreprises et les particuliers). Les services du cloud peuvent être déployés soit sur un environnement partagé et mutualisé entre tous les utilisateurs (cloud public), soit sur un environnement dédié à un certain nombre de client (cloud privé), soit sur un environnement partagé entre une communauté bien définie d'utilisateurs (cloud communautaire) [31].

Plusieurs facteurs sont à l'origine de l'apparition du cloud computing. Tout d'abord, le développement rapide des technologies internet a permis d'externaliser le stockage et le traitement de données hors les murs des entreprises. En outre, le cloud computing a bénéficié de la diminution des coûts de stockage, de l'apparition de la fibre optique et la technologie de virtualisation. En effet, la virtualisation est l'une des évolutions majeures ayant favorisé l'émergence du cloud computing.

Depuis son avènement dans le marché, le cloud n'a cessé d'évoluer et d'être utilisé en raison de ses avantages non seulement financiers, mais aussi organisationnels et sécuritaires. Plusieurs pays y compris les Etats-Unis ou Royaume-Unis, veulent migrer leurs données vers le modèle cloud en externalisant une partie de leur infrastructure sur le cloud. Dans le domaine d'enseignement, les projets cloud sont nombreux. Au Maroc, l'IMIST-CNRST[9], a mis en place, en collaboration avec le ministère de l'enseignement supérieur de la recherche scientifique et de la formation des cadres, et la société IBM, le premier cloud académique

[9] https://www.imist.ma/

national[10]. Cette plate-forme est lancée dans le cadre du programme IBM MEA (Middle East and Africa) University. Elle propose des cours sur les dernières technologies de l'industrie informatique comme le big data, cyber sécurité, etc.

II.1.2 Brève Histoire

En 1970, la technologie majeure des premiers systèmes d'information était le Mainframe. Il s'agit d'un ordinateur central de grande puissance de traitement. L'année 1980 est marquée par le développement de l'architecture client-serveur qui a été la première étape vers la distribution de traitement et de stockage des données.

Par la suite, en 1990, l'internet est né. Il a changé la façon d'externalisation des données qui deviennent accessibles par le grand public. En 2000, les technologies internet dominent toutes les architectures de communication informatique, à travers le concept de SOA (Service Oriented Architecture) principalement basé sur l'usage des web services (http/SOAP/XML).

En parallèle au développement rapide des technologies internet, se développe une autre technologie innovante sur le plan matériel, il s'agit de la virtualisation. Cette dernière permet de faire fonctionner, plusieurs serveurs virtuels sur un même serveur physique, comme s'ils fonctionnaient sur des machines distinctes. L'idée de virtualisation consisté à mutualiser les capacités du serveur afin de rationaliser et d'assurer une gestion efficace et moins cher des ressources physiques [32].

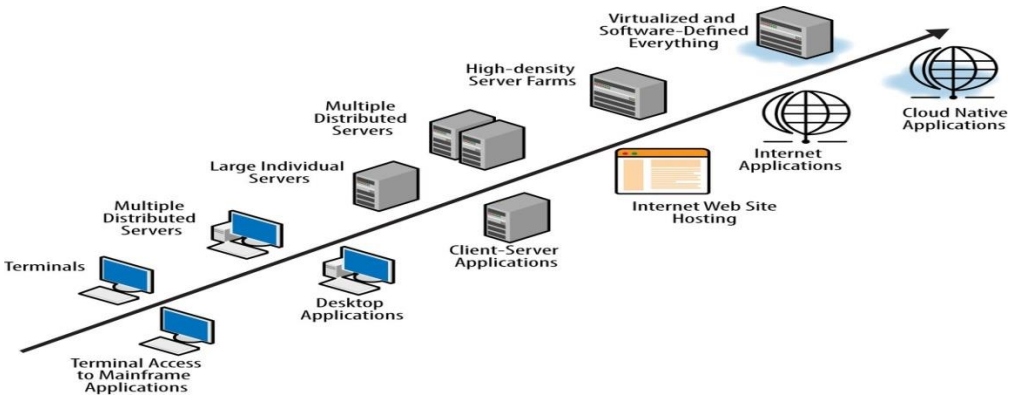


Figure II.1: Evolution du cloud computing[11]

[10] https://cloud.imist.ma/index.php

Par la suite, le concept du cloud computing a commencé à apparaître. En 2002, le géant du web Amazon inventa le terme cloud computing pour décrire son offre de services en termes de ressources d'infrastructure. En effet, Il a investi dans des clusters d'ordinateurs immenses pour supporter les charges importantes des commandes faites sur son site e-commerce[12] dans la période de Noël. Mais, ces clusters restent inutilisés pendant le reste de l'année. Amazon a pensé donc à ouvrir toutes ces ressources inutilisées aux entreprises pour être louées à la demande en proposant à ses clients ses outils S3 (Simple Storage Service) et EC2 (Elastic Compute Cloud). Par la suite, plusieurs leaders des services ont lancé leurs offres cloud, tels que IBM, Microsoft, etc. [33].

II.1.3 Définitions

Le cloud computing, traduit par "Informatique en nuage", est un terme à la mode, abondamment utilisé de nos jours, pour décrire un éventail de technologies. En général, il est considéré comme un nouveau paradigme pour organiser et gérer les ressources TIC, mais, en réalité, est un terme assez vague possédant plusieurs définitions.

The National Institute of Standards and Technology (NIST) a défini le cloud computing comme : “un modèle pratique, à la demande et universel, permettant d'établir un accès par le réseau à un réservoir partagé de ressources informatiques configurables (réseaux, serveurs, stockage, applications et services) qui peuvent être rapidement mobilisées et mises à disposition en minimisant les efforts de gestion et de contact avec le fournisseur de services” [34].

Microsoft Azure [35] donne la définition suivante : “le cloud computing est un ensemble de services informatiques (serveurs, stockage, bases de données, composants réseau, logiciels, outils d’analyse, etc.) fournis via internet. Les sociétés qui proposent ces services informatiques sont appelées fournisseurs de services cloud. Elles facturent en général ces services en fonction de l’utilisation, un peu comme la facture d’eau ou d’électricité.”

[11] https://twitter.com/antgrasso/status/982747363575058432
[12] https://www.amazon.com/

La plupart des définitions s'intéressent aux éléments suivants : l'élasticité des services, la mutualisation des ressources, la technologie de virtualisation, le paiement à l'usage, l'accès léger ou encore l'universalité d'accès.

II.1.4 Caractéristiques

Libre-service à la demande (self-service) : les services du cloud mettent en œuvre des solutions en libre-service. Les utilisateurs peuvent automatiquement provisionner leurs propres besoins en termes de ressources informatiques sans intervention humaine à travers des interfaces web interactives.

Accès léger : comme l'indique le terme cloud ou nuage, l'accès aux ressources est accessible n'importe où, n'importe quand et à partir de n'importe quel terminal. En effet, l'utilisateur a la possibilité d'accéder au cloud par l'intermédiaire d'un navigateur à partir de son ordinateur, son téléphone intelligent, sa tablette ou tout autre appareil informatique.

Élasticité rapide : le cloud a la capacité de s'étendre et de se réduire les ressources (machines virtuelles, stockage, etc.) à la demande. Si la demande de l'entreprise en ressources informatiques augmente, le fournisseur de services en nuage peut augmenter son approvisionnement en ressources de façon entièrement automatisée, sans aucune intervention manuelle de sa part. Pour le consommateur, les ressources informatiques mises à disposition par le fournisseur semblent théoriquement illimitées.

Paiement à l'usage (pay as you go) : dans le modèle cloud, le paiement se fait en fonction des services utilisés (ressources allouées et nombre des utilisateurs, …). Les services du cloud permettent au client de payer uniquement pour les ressources qu'il consomme, tout comme l'eau, le gaz ou l'électricité. Il suffit de se brancher pour en bénéficier, et la facturation est calculée en fonction de la durée et de la quantité de ressources exploitées.

Mutualisation des ressources : Mise en commun de ressources informatiques d'un fournisseur destinées à plusieurs utilisateurs. Il s'agit d'un modèle de service partagé dans lequel différentes ressources physiques et virtuelles sont affectées et réaffectées dynamiquement en fonction de la demande du client. En effet, la stratégie de mutualisation est tellement efficace puisqu'elle permet la réduction des coûts (énergie, loyer, ressources humaines, etc.)

II.1.5 Concept du cluster

Les clusters constituent l'élément de base dans l'infrastructure du cloud. Il s'agit d'un groupement de machines appelé grappe de serveurs fonctionnant comme un seul système. Le but principal du cloud est de construire un nuage de clusters permettant d'interconnecter un ensemble de nœuds sur un réseau dédié. Pour gérer cet ensemble de clusters, les fournisseurs déploient la technologie de virtualisation pour optimiser l'utilisation des ressources telles que l'espace disque, la RAM, ou encore le CPU.

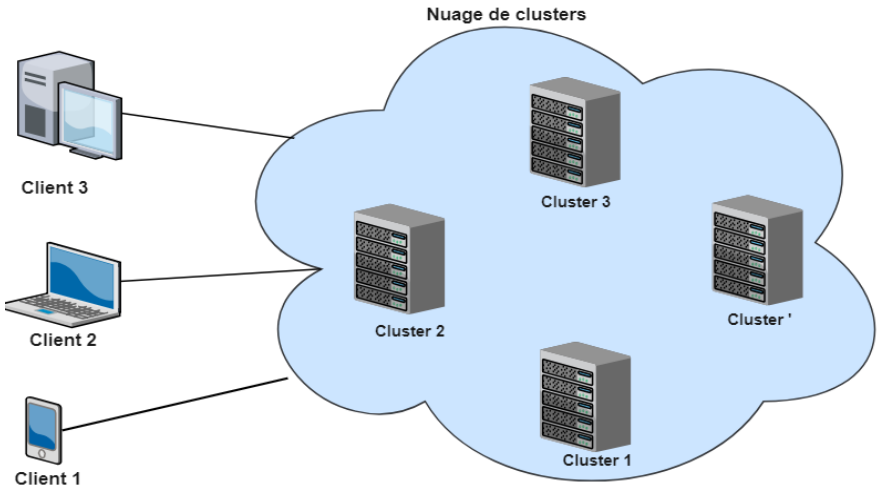


Figure II.2: Infrastructure simplifiée d'un cluster

L'infrastructure simplifiée du cloud est décrite dans la Figure II.2. Elle constitue d'un ensemble de clusters et dont chacun est composé d'un certain nombre de nœuds. Ces nœuds sont gérés par un serveur dédié appelé master qui distribue le traitement sur l'ensemble des machines en fonction des ressources demandées par les utilisateurs ce qui permet d'assurer un équilibrage de charge entre différents nœuds d'un cluster.

Lorsqu'un utilisateur demande la création d'une machine virtuelle, le gestionnaire des ressources cluster crée une instance du système d'exploitation désiré (Unix, Windows) en obtenant les ressources (disque, CPU, et RAM) nécessaires via les nœuds. La Figure II.3 montre le diagramme de séquence UML décrivant le scénario de réservation d'une ressource (machine virtuelle) sur le cluster. Pour permettre aux différents nœuds de communiquer entre eux et avec le gestionnaire de nœuds ou master, le cloud utilise le protocole de communication sécurisé SSH afin d'effectuer l'échange de données entre l'ensemble des neouds d'un cluster.

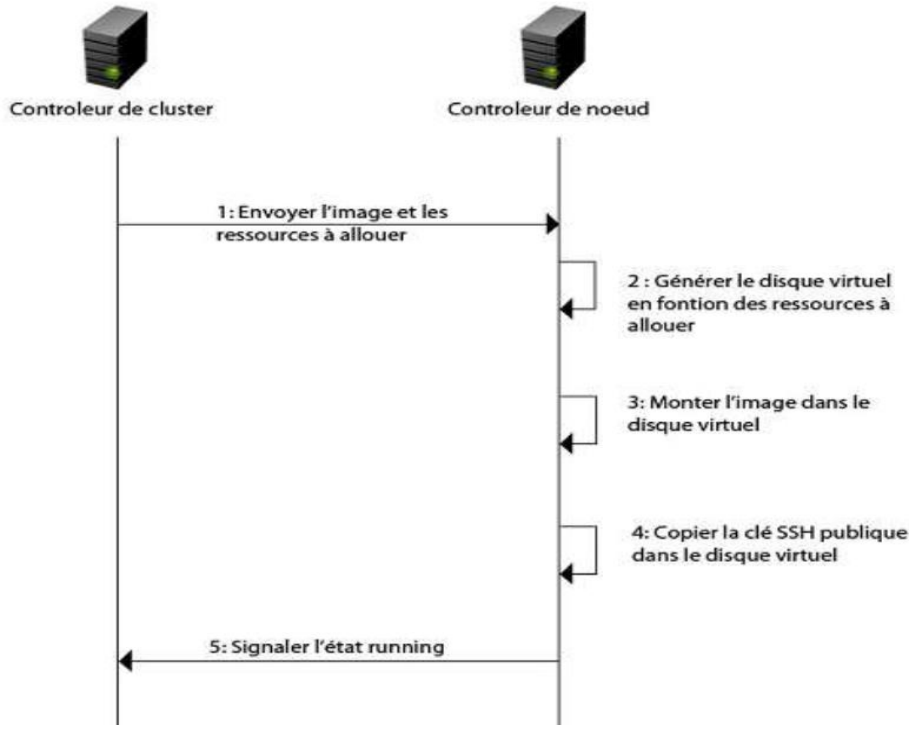


Figure II.3: Diagramme de séquence d'allocation d'une machine virtuelle sur un cluster cloud

II.1.6 Services

Grâce au cloud, les données sont disponibles sur internet, et on n'a pas besoin forcément d'avoir un ordinateur physique pour y accéder. En fait, il existe plusieurs services de cloud différents. Ces services sont de trois types: du matériel, des plates-formes et des logiciels [31]. La Figure II.4 donne un schéma synthétique de ces services.

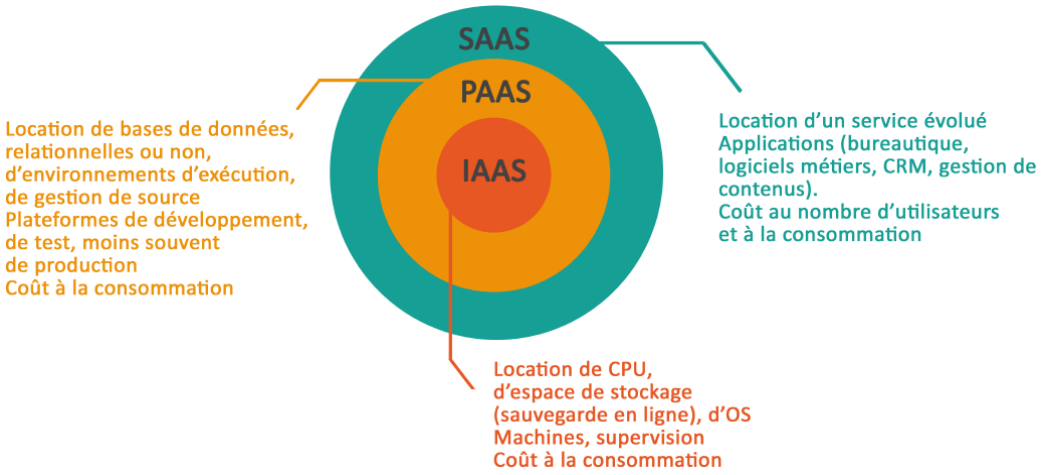


Figure II.4: Services du cloud

En termes de services mises à disposition, le modèle cloud peut servir à de nombreuses finalités, comme illustré dans le schéma ci-dessous (Figure II.5)

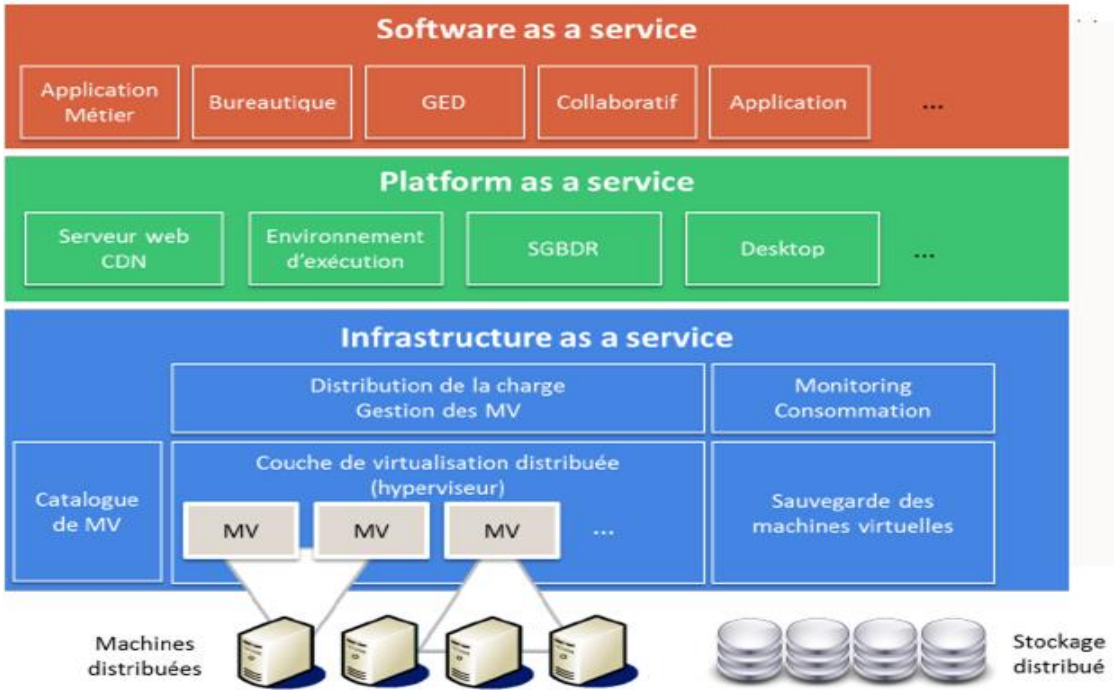


Figure II.5: composants des différents services du cloud

II.1.6.1 SaaS

Le premier et le plus connu est le SaaS, en anglais Software As A Service. Il s'agit d'une application fournie sous forme de service. Les utilisateurs y accèdent via une connexion au réseau internet. Ces logiciels prêts à l'emploi ne nécessitent pas de maintenance. Les mises à jour sont effectuées régulièrement par le fournisseur du cloud et les clients utilisent la même version d'un logiciel. En effet, le SaaS permet à un établissement d'enseignement ou entreprise de ne plus installer des logiciels sur ses propres machines mais de s'abonner à des logiciels en ligne et de payer les frais d'abonnement. Les exemples les plus connus sont, Google Apps ou DropBox.

II.1.6.2 PaaS

Le deuxième type des services du cloud est le PaaS, en anglais Platform As A Service. Une PaaS est une plate-forme applicative et un environnement d'exploitation et de suivi pour permettre la gestion d'applications dans le cloud. Le principe de PaaS consiste à louer des logiciels intermédiaires comme systèmes d'exploitation, serveurs d'applications, systèmes de gestion des bases de données (SGBD) et un environnement de développement et d'exécution

(IDE) pour développer, tester, et exécuter des applications en mode SaaS. Depuis plusieurs années, Amazon, Google et d'autres proposent des solutions PaaS. Les exemples les plus connus sont Heroku, OpenShift, ou encore clever cloud.

II.1.6.3 IaaS

Le troisième type du cloud est le IaaS, en anglais Infrastructure As A Service. Il consiste en une infrastructure souple et flexible des équipements informatiques (serveurs, stockage, équipements réseaux etc.). L'IaaS donne accès à la location des machines, d'espaces de stockage, et de capacités de calcul. Il offre un ensemble de ressources virtuelles, évolutives et robustes disponible, à la demande, sur internet. Les exemples les plus connus sont Amazon web services ou encore OpenStack.

Les avantages et les inconvénients liés à chaque service du cloud sont listés dans la Table II.1 :

Tableau II-1: Avantages et inconvénients des services du cloud

Service	**Avantages**	**Inconvénient**
SaaS	• Pas d'installation • Plus de licence • Mise à jour automatique	• Logiciel limité • Dépendance du prestataire
PaaS	• Environnement hétérogène • Pas d'installation • Fiabilité • Disponibilité	• Limitation des langages supportés • Pas de personnalisation dans la configuration des machines virtuelles
IaaS	• Administration • Personnalisation • Flexibilité d'utilisation • Élasticité	• Perte de contrôle (sécurité, confidentialité)

II.1.7 Modèles de déploiement

Le cloud computing est un environnement complet des ressources virtuelles dans lequel l'accès se fait, simplement, via une connexion internet. Le principe du cloud consiste à

mutualiser des ressources d'infrastructures et de services afin d'optimiser leur utilisation. Selon l'institut national des normes et de la technologie [34], les quatre modèles de déploiement du cloud les plus connus sont les suivants, (voir la Figure II.6) .

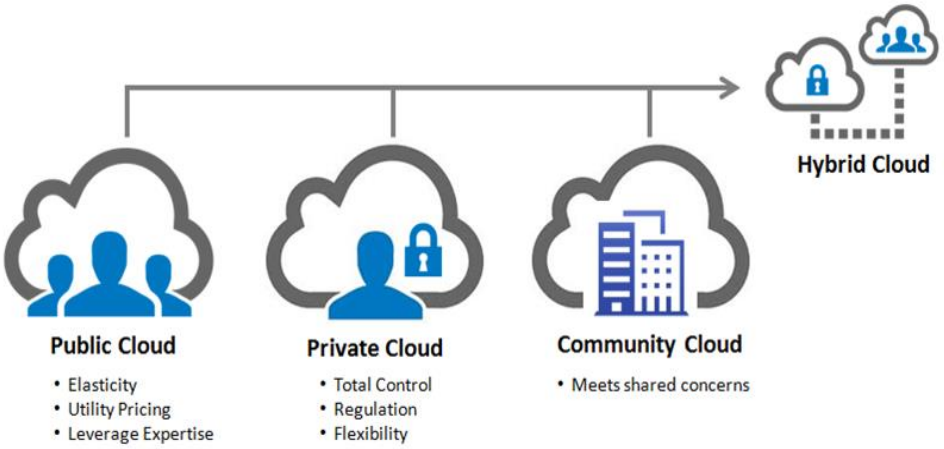


Figure II.6: Modèles de déploiement du cloud

II.1.7.1 Cloud public

Dans un cloud public les fournisseurs de ce dernier offrent leurs ressources comme services pour une utilisation ouverte par le grand public via internet. Ce modèle de déploiement peut être possédé, géré et exploité par une entreprise, une université, une organisation ou une combinaison entre eux. Il offre plusieurs avantages, y compris la mise en place rapide et peu coûteuse, et l'adaptation à l'augmentation des besoins des clients. Ces derniers n'ont pas un contrôle total sur la sécurité de leurs données, ou sur la configuration d'infrastructure de leurs serveurs.

II.1.7.2 Cloud privé

Le modèle du cloud privé est fourni pour une utilisation exclusive par une seule organisation comprenant plusieurs utilisateurs. Il peut être possédé, géré et exploité par une organisation ou un tiers. L'infrastructure du cloud privé peut exister dans les locaux de l'organisation. Le cloud privé offre un contrôle beaucoup plus complet sur la disponibilité des données et sur la sécurité des réseaux. L'inconvénient de ce modèle de déploiement est que son infrastructure est similaire aux serveurs d'hébergement traditionnels (coûts d'investissement dans les matériels et la maintenance).

II.1.7.3 Cloud communautaire

L'infrastructure cloud est fournie pour une utilisation exclusive par une communauté spécifique d'utilisateurs ayant des préoccupations communes (Université, politique, etc.). Son infrastructure est gérée par un ou plusieurs organismes de la communauté ou un tiers. Ce modèle rassemble les avantages du cloud public en termes de paiement en fonction de consommation, et ceux du cloud privé comme la sécurité et la confidentialité des données.

II.1.7.4 Cloud Hybride

Le modèle de déploiement hybride est composé de deux ou plusieurs infrastructures distinctes du cloud. En d'autres termes, il s'agit de la mise en commun des ressources du public, du privé, et/ou du communautaire. Dans l'architecture hybride du cloud, une partie des services (ressources critiques) est fournie par le cloud privé tandis que le reste des services s'exécute en mode cloud public. Le cloud hybride offre une grande flexibilité.

II.1.8 Technologies

Le cloud computing rassemble plusieurs technologies modernes dans un même écosystème. Parmi ces technologies, on trouve :

- Au niveau matériel : la virtualisation et les processeurs multicœurs
- Au niveau calcul distribué : MapReduce et Spark
- Au niveau technologies d'internet : les architectures orientées services, les web services, et le web 2.0.

Dans ce rapport de thèse, nous concentrons, d'une part, sur la technologie de virtualisation car elle constitue le cœur de l'infrastructure cloud permettant la mise en place des services IaaS; d'autre part, nous focalisons sur les technologies d'internet parce qu'elles donnent la possibilité de livrer les services cloud aux clients via des plates-formes web [36].

II.1.8.1 Virtualisation

Le terme virtualisation recouvre l'ensemble des techniques matérielles et ou logiciels qui permettent de faire fonctionner et cohabiter plusieurs environnements logiques indépendants (systèmes d'exploitation) sur une seule machine. Le système d'exploitation ainsi virtualisé est appelé système invité (guest system) et le système d'exploitation servant d'accueil s'appelle le système hôte (host system). La machine physique utilisée est appelée machine hôte tandis que celle émulée au système invité est la machine virtuelle. Dans une architecture de

virtualisation, la machine virtuelle peut être considérée comme un logiciel au même titre qu'un navigateur fournissant au système invité des ressources matérielles similaires à celles d'une machines physique. Il existe plusieurs catégories de virtualisation utilisant chacune des technologies différentes. La Figure II.7 montre une architecture simplifiée de la virtualisation [37].

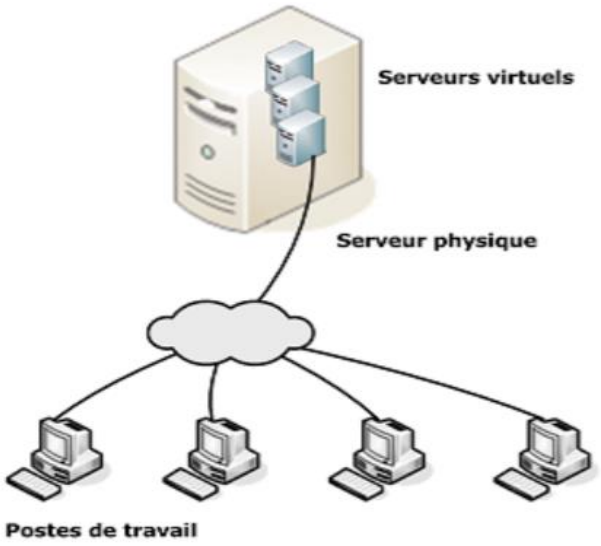


Figure II.7: Architecture simplifiée de la technologie de la virtualisation

II.1.8.2 Services web

Il s'agit d'une technologie permettant la communication et l'échange de données entre des applications dans des environnements distribués indépendamment des plates-formes et des langages de programmation qu'ils utilisent. Un web service fournit un ensemble de fonctionnalités délivrées via internet ou intranet en utilisant le protocole de transfert HTTP(S). La communication entre les services web est basée sur le principe de demandes et réponses, effectuées par le biais d'un format commun tel que XML et JSON [38]. Les solutions de cloud computing se basent sur les services Web à la fois pour l'exposition des fonctionnalités du système sur internet, et aussi pour l'interaction entre les différents composants internes du cloud en se basant sur les protocoles de communication. Il existe deux grandes familles de services web :

- Les services de type RCP

  RPC (Remote Procedure Call) est un protocole réseau permettant de faire des appels de procédures sur un ordinateur distant. Il s'agit d'une spécification et un ensemble de

codes qui permettent à des processus s'exécutant dans des environnements différents de faire des appels de méthodes à travers un réseau. Ce protocole est utilisé dans le modèle client-serveur pour assurer la communication entre le client et le serveur en utilisant le protocole HTTP pour le transfert des données et la norme XML pour formater et structurer ces données. Ce protocole permet d'appeler une fonction sur un serveur distant à partir de n'importe quel système (Windows, Mac OS X, Linux) et avec n'importe quel langage de programmation. Le serveur lui-même peut fonctionner sur n'importe quel système et peut être programmé dans n'importe quel langage. [39].

- Les services de type SOAP

  SOAP, en anglais Simple Object Access Protocol, est un protocole de type RCP orienté object conçu dans le but d'échanger des données dans un environnement distribué. Il permet la transmission des données entre deux objets distants. En effet, il autorise à un objet d'invoquer une méthode d'un autre objet physiquement situé dans un autre serveur. Pour formater les massages, sa syntaxe est fondée sur XML et ses commandes sont envoyées sur internet, le plus souvent, par l'intermédiaire du protocole HTTP, et parfois, via le protocole SMTP ou POP. SOAP représente un protocole plus performant par rapport au protocole RCP, mais il est plus complexe.

  L'architecture SOAP est constituée de deux éléments :

  - Une enveloppe : contenant les informations sur l'objet à transmettre afin de permettre son acheminement et son traitement.
  - Un modèle de données : définissant le format du message à transmettre. Il contient les données échangées entre le client et le service sous la forme d'un fragment de document XML.

- Les services de type REST

  Representational State Transfert ou REST est un style d'architecture, qui repose sur le protocole http, permettant l'accès à une ressource distante via URI (Uniform Resource Identifier) pour procéder à diverses opérations supportées par HTTP. La principale différence qui réside entre le REST et d'autre web service est qu'il ne s'agit pas d'un standard mais d'un style architectural, il est également plus rapide que SOAP ou RCP. Les applications respectant l'architecture REST sont appelées RESTful.

II.1.8.3 Web 2.0

Le Web 2.0 représente une avancée dans l'utilisation des technologies du Web. Il s'agit d'un ensemble d'outils, de langages, de normes et de protocoles permettant les interactions entre les utilisateurs (réseaux sociaux) ou entre des applications par l'échange de données via des protocoles normalisés (web service). Il est basé sur plusieurs technologies telles que le HTML5, le CSS4, XML, Javascript, AJAX, etc.

Le Web2.0 est considéré comme un service dynamique, interactif, et coopératif. Il permet [40] :

- La création : la production de contenus par les utilisateurs via des réseaux sociaux en utilisant des applications et des outils dédiés : Wikis, Blogs, plates-formes dynamiques, etc.
- La décentralisation : la création peut se faire par des personnes éloignées, physiquement distribuées.
- La compilation de contenus : génération de contenus à partir d'autres contenus via des applications respectant des normes et des protocoles permettant la recherche et les échanges des données.

Les services du cloud sont par nature des applications web dynamiques qui fournissent aux utilisateurs les ressources demandées à travers des plates-formes mises à disposition par du fournisseur cloud.

II.1.9 Avantages

Le cloud computing est associé à une multitude d'avantages et de bénéfices qui motivent les professionnels et les grands organismes de s'orienter vers ce nouveau modèle d'externalisation des ressources informatiques. Parmi les avantages du cloud, on trouve, entre autres :

- Flexibilité

  Dans le mode cloud, les utilisateurs bénéficient d'une grande agilité et peuvent facilement dimensionner les services alloués en fonction de leurs besoins. Ils peuvent personnaliser leurs serveurs, le stockage et les applications utilisés à tout moment. En fait, le cloud est accessible partout où il y a une connexion internet à travers des interfaces web afin de faciliter la demande et l'administration des ressources à l'aide d'un ordinateur, ou d'un Smartphone tout simplement.

- Efficacité

  Grâce aux services du cloud, les entreprises sont désormais capables de mettre en place leurs systèmes d'information rapidement sans s'inquiéter des coûts trop élevés et du temps d'installation et de maintenance. En plus, le coud propose un mode de tarification en fonction de la consommation des ressources, ce qui permet aux entreprises d'économiser le budget des serveurs et d'autres équipements.

- Optimisation

  Le cloud est basé sur le principe de mutualisation de ressources, c'est-à-dire la distribution du stockage et du traitement des charges de travail sur tous les nœuds d'un cluster donnée, ce qui permet de minimiser le temps de réponse et d'optimiser l'utilisation des réseaux, des applications et de toute autre ressource afin de fournir un service de qualité partageable entre différents utilisateurs à travers les outils et les technologies de virtualisation.

- Mise à niveau

  Le prestataire cloud offre aux utilisateurs la capacité de la mise à niveau automatique des versions des logiciels et des APIs pour assurer la compatibilité des services cloud avec les dernières technologies informatiques. Cela permet de libérer les entreprises des tâches fastidieuses de maintenance des versions des différents langages de programmation ou des logiciels installés sur les serveurs.

- Elasticité

  Elle fait référence à l'évolutivité de l'environnement cloud à la taille de chaque organisme. En fait, le cloud s'adapte très rapidement aux besoins de chaque client selon le nombre d'utilisateur et la quantité des données stockées et manipulées, etc.

- Collaboration

  Le cloud mis en place de nombreux outils de collaboration et des moyens de communication pour faciliter le partage des ressources en ligne via des solutions de messagerie, de réseaux sociaux, etc. Ainsi, Les employés utilisent une messagerie instantanée pour se connecter et échanger les données avec leurs équipes. L'un des exemples des outils de collaboration fournies par Google est Hangouts permettant le partage des fichiers et des documents par l'intermédiaire de Google Drive.

- Accessibilité

Les services du cloud sont accessibles partout, sur tous les supports, là où il y a une connexion internet, à travers des interfaces web afin de faciliter la demande et l'administration des ressources à travers différents appareils.

- Sauvegarde

  Le cloud est désormais capable de proposer la sauvegarde automatique et rapide des données des clients. Ces données sont stockées dans une infrastructure performante qui respecte les normes de sécurité mondiale. Il assure la configuration de la sauvegarde, du cryptage de l'information et la récupération en cas de panne.

II.1.10 Défis

Le cloud constitue un changement radical dans la façon d'externaliser les services informatiques d'une entreprise. Bien que les avantages de l'entreprise prête pour le cloud peuvent s'avérer considérables, l'adoption du cloud a de nombreux risques, parmi lesquels, on trouve :

- Sécurité

  La sécurité est un défi majeur, surtout dans les environnements publics ou partagés, où le fournisseur de services cloud doit s'assurer que la protection des données et la conformité aux règles sont garanties. La sécurité et l'efficacité des échanges de données dans le cloud, de même que la sécurité de la connectivité des applications, sont des sujets de préoccupation majeurs. Également, la gestion des images est primordiale, car ces images constituent une solution rapide de déploiement afin de contourner les problèmes d'installation. Dans ce contexte, les fournisseurs cloud doivent se doter des moyens pour organiser, sécuriser, gérer et déployer les images de manière évolutive sur les différentes plateformes virtualisées. Après le déploiement, Ils ont besoin d'outils pour gérer les images virtuelles, notamment pour assurer la surveillance, la mise à jour, le suivi et la maintenance.

- Normalisation

  La normalisation est un aspect important du cloud car elle a la capacité de réduire les variations de mise en œuvre des services cloud. La normalisation des processus permet d'unifier les couches, les protocoles et les normes utilisées par les fournisseurs cloud. Tout comme l'adoption de l'HTML comme langage standard a accéléré l'adoption et

la diffusion d'internet, il faut aussi normaliser l'informatique en nuage pour que son adoption se généralise[41].

- Accès

  Dans le cas où l'utilisateur n'a pas de connexion internet, ou une connexion insuffisante, alors il ne peut pas accéder à ses données hébergées sur la plate-forme cloud. Pour résoudre ce problème, il faut permettre le travail sur une application locale qui synchronise ensuite les données avec le cloud dès que l'utilisateur a un accès internet. Le problème du traitement des données en local se pose donc à nouveau [42].

II.1.11 Principaux acteurs

Le cloud fait référence à l'utilisation des ressources informatiques virtualisées à la demande sous forme de services facturés en fonction de consommation. En fait, il existe plusieurs acteurs du cloud et chaque fournisseur a sa propre offre.

- **Amazon**

Depuis 2006, Amazon Web Services (AWS), a mis à disposition du client un cloud public. Il propose de nombreux services en ligne répartis en diverses catégories telles que le stockage, la puissance de calcul, le réseau, l'analyse des données des services applicatifs, la gestion système, les solutions du big data et les outils pour l'internet des objets, etc. Les services les plus connus sont Amazon Elastic Compute Cloud (EC2) et Amazon Simple Storage Service (S3) [43].

Amazon EC2 offre un environnement virtuel riche et facile à accéder et à gérer. Il s'agit d'un service web fournissant une puissance de calcul performante et redimensionnable dans le cloud. Il permet à travers des interfaces web de démarrer des instances avec une variété de systèmes d'exploitation, de gérer les droits d'accès et d'augmenter ou de diminuer rapidement leurs capacités en fonction de besoins de traitement. Il est destiné aux développeurs pour créer des applications sans se soucier des problèmes d'installation et de configuration de l'environnement du développement. En effet, Amazon EC2 permet de choisir facilement la configuration de la mémoire, du CPU et du stockage d'instance, ainsi que la taille de la partition de démarrage, qui répondent le mieux aux besoins du choix du système d'exploitation et d'applications [44].

Les machines virtuelles d'Amazon sont préconfigurées et prêtes à l'emploi avec une multitude de choix des systèmes d'exploitation comme Ubuntu, OpenSolaris, RedHat, Windows Server, etc. Les ressources de chaque machine virtuelle sont configurables depuis la console de gestion web d'Amazon (Figure II.8) [45].

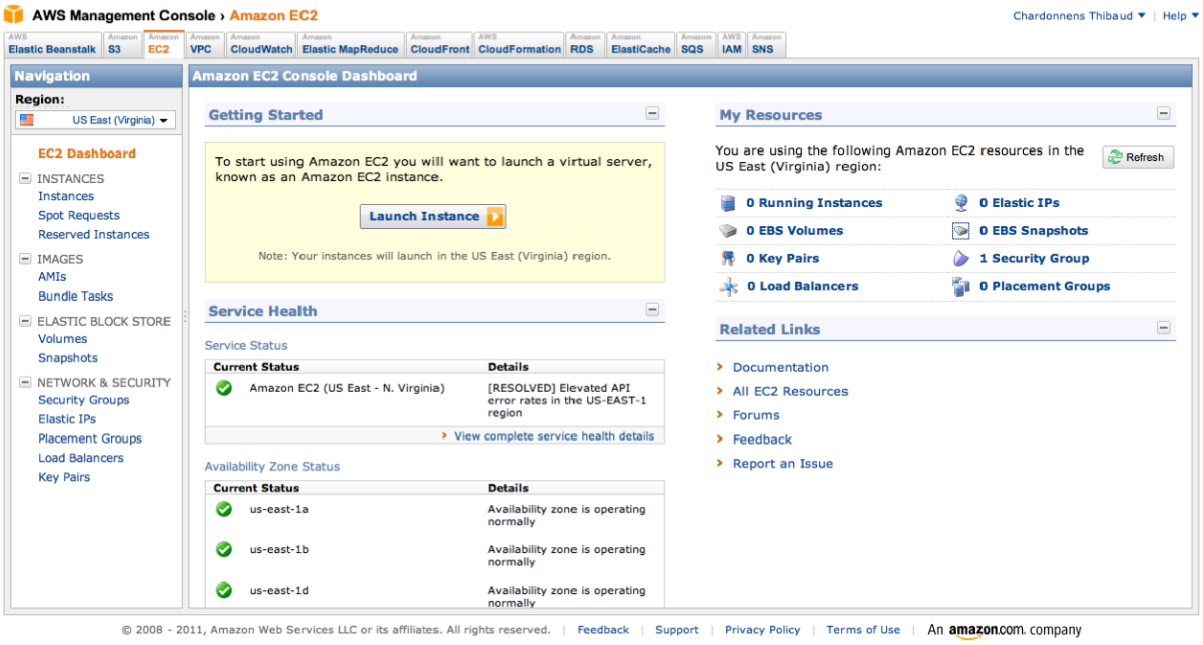


Figure II.8: La Console de management web d'Amazon

Amazon offre également des services de type PaaS tels que Amazon S3, l'acronyme d'Amazon Simple Storage. Amazon S3 est un service web dédié pour le stockage, la sauvegarde et l'archivage des données en ligne. Il est conçu pour stocker et récupérer des grandes quantités de données, à tout moment, et n'importe où sur le web [46].

Pour les traitements du big data, Amazon fournit un service appelé Elastic MapReduce (EMR) [47]. Amazon EMR est un service Web s'appuyant sur une infrastructure Hadoop qui permet de traiter de manière simple, rapide et rentable des grandes quantités de données sur des instances Amazon EC2 dynamiquement évolutives. Ce service peut également exécuter d'autres Frameworks big data récents tels que Hbase, Apache Spark, Flink, Pig, et Mahout. Il s'appuie sur le service S3 pour le stockage des données après le traitement en parallèle sur des cluster Hadoop. Amazon EMR supporte un large éventail de cas d'utilisation du big data de façon sure et fiable, à savoir: les algorithmes de l'apprentissage automatique (machine

learning), les transformations de données, l'indexation web, et l'analyse d'historique des fichiers journaux.

Amazon DynamoDB [48] est proposé comme une solution des bases de données NoSQL (Not Only Sql). NoSQL est une famille de bases des données non relationnelles. Il est destiné au stockage et à l'archivage rapide supportant d'énormes volumes de données. Jusqu'à présent, Amazon est le plus grand fournisseur des services cloud. Pour garder sa position de leader, Amazon propose des services variés entre stockage distribué (S3), traitement en parallèle (MapReduce et Spark), des bases données NoSQL, et d'autres technologies prometteuses, ce que peu de concurrents offrent.

- **Google**

Google Cloud Platform [49] est une plate-forme qui rassemble les différents services cloud proposés par Google. Il regroupe une vaste gamme de service, y compris le calcul, le stockage, le réseau, la sécurité, les analyses de données massives, les techniques de machine learning, et le développement et déploiement des applications sur les serveurs google accessible via internet. Les trois services de Google Cloud Platform les plus populaires sont :

- Google Compute Engine : est une infrastructure en tant que service (SaaS) permettant aux utilisateurs de louer et de lancer des instances des machines virtuelles sur les serveurs distants de Google à la demande.
- Le Google App Engine : il s'agit d'une plate-forme en tant que service (PaaS). Sa première version est sortie en avril 2008. Il offre aux développeurs un environnement complet pour la création et l'hébergement des applications Web et mobiles modernes. App Engine supporte de nombreux langages de programmation comme Java, Node.js, Go, Phyton, PHP, et Ruby.
- Google Cloud Storage : est une plate-forme de stockage et est considérée comme un système distribué visant à sauvegarder les données volumineuses et hétérogènes. Google offre aussi autant de possibilités de stockage et de bases des données, comme le Cloud Datastore et Bigtable.

- **Salesforce**

Salesforce est un pionnier de services SaaS basé à San Francisco aux États-Unis et spécialisé dans la gestion des relations avec les clients ou CRM (customer relationship management). Il offre une multitude d'options pour gérer et contrôler les ventes, le marketing, et entrer en contact avec les clients. Les solutions de Salesforce peuvent être divisées en trois grandes familles: Sales Cloud, Chatter Collaboration Cloud et Service Cloud [50].

- Sales Cloud est un ensemble d'outils déployés sur le cloud, disponibles, en plus, en plusieurs langues et accessibles en ligne depuis un ordinateur ou un appareil mobile. Il permet à l'utilisateur à tout moment de consulter les activités de ses clients, de gérer les compagnes du marketing et de visualiser la performance de vente de son entreprise.
- Chatter Collaboration Cloud Créé en 2010, il s'agit une plate-forme de collaboration en temps-réel sous forme d'un réseau social pour une entreprise.
- Service Cloud est une plate-forme cloud rassemblant une nouvelle génération de moyens de collaboration, de canaux de communication via des réseaux sociaux ou des services d'e-mail et de chat.
- **Microsoft**

Comme les autres leaders du cloud. Microsoft permet d'offrir des ressources virtuelles du cloud sous forme de services à la demande de types logiciel (SaaS), plate-forme (PaaS) et infrastructure (IaaS).

Au niveau SaaS, Microsoft propose une suite de logiciels en tant que service permettant aux utilisateurs de se connecter, via internet, à des applications et de les utiliser sans la moindre configuration ou installation préalable. Les exemples les plus connus de ce type de services sont des outils de messagerie comme Outlook [51], une version web des applications de bureautique comme Microsoft Office 365 [52, p. 365] offrant les mêmes fonctionnalités que les logiciels Microsoft Office classique (Excel, Word, etc.). Plusieurs utilisateurs donc peuvent partager et éditer un document et de travailler en collaboration entre eux.

Au niveau PaaS, en 2010, Microsoft a lancé son offre de cloud appelé Microsoft Azure [53]. Il s'agit d'un service PaaS phare de Microsoft. Microsoft Azure permet aux développeurs à la fois d'utiliser des technologies open-source et des technologies propriétaires de Microsoft, aussi dans différents langages de programmation dont .NET, Java, PHP, Python pour créer,

déployer et gérer des applications en ligne. Également, Microsoft met en place Azure SQL Database destiné pour la migration des bases de données relationnelles existantes vers la plate-forme cloud. Ainsi, il offre une interface d'administration simple afin de faciliter la création, la gestion et d'assurer la sécurité d'une base de données SQL server.

Au niveau IaaS, Microsoft offre une infrastructure évolutive en tant que service capable de gérer les ressources serveurs (CPU, stockage, et mémoire) qui répondent aux besoins de chaque client en permettant à ce dernier de se concentrer sur les règles métiers de l'entreprise au lieu de gaspiller du temps et de l'argent à acheter, gérer, maintenir et sécuriser l'infrastructure physique. IaaS de Microsoft fournit des machines virtuelles préconfigurées et préinstallées supportant divers systèmes d'exploitation tels que Unix, Windows Server, SAP ou Oracle. L'offre IaaS de Microsoft comprend le service Virtual Network pour créer un réseau virtuel permettant aux utilisateurs la gestion et la sécurité réseau [54].

- **IBM**

IBM offre un ensemble de services du cloud couvrant six domaines : le développement des applications, l'analyse des données, le stockage, la gestion de serveurs, la collaboration, et les processus métiers. Le portefeuille d'IBM en matière des services cloud est très riche et il permet de répondre à tous les besoins des entreprises et il comprend :

- Services SaaS: Le fournisseur IBM propose plus de 140 offres d'applications sous forme de SaaS pour satisfaire les besoins des entreprises. Ces logiciels peuvent être de divers types: CRM, collaboration et email, sécurité, etc.
- Services PaaS: IBM Bluemix est une véritable plate-forme en tant que services permettant de concevoir, créer, et de déployer rapidement des applications avec une grande variété de choix de langage de programmation. Elle propose également un large catalogue d'API et des services de type serveurs d'application et/ou des bases de données [55].
- services IaaS: SoftLayer [56] est une offre de l'infrastructure cloud en tant que service qui comprend des ressources informatiques virtuelles telles que des serveurs de stockage et des équipements réseaux. Il y a aussi CMS (Cloud Managed Services) d'IBM, qui offre une infrastructure agile, sécurisée et prête à l'emploi permettant des bonnes performances et de la haute disponibilité.

En général, Il n'existe pas un fournisseur cloud idéal pouvant s’adapter à toutes les entreprises. Il faut prendre en considération plusieurs paramètres tels que le coût, la qualité, la fiabilité pour décider quelle offre du cloud à choisir. Pour cela, chaque organisme désirant de migrer son système d'information vers le cloud doit faire une étude approfondie de ses besoins, de son budget et des exigences non fonctionnelles d'évolutivité, de fiabilité et de haute performance.

II.1.12 Conclusion

Dans ce chapitre, nous avons abordé le concept du cloud computing. Nous avons commencé par présenter l'informatique en nuage et ses caractéristiques. Un bref aperçu de l'histoire de l'évolution de ce paradigme nous aide à comprendre son principe de fonctionnement et à avoir une idée sur ses technologies et ses opportunités. De plus, nous avons mis l'accent sur les services du cloud avant de discuter leurs modèles de déploiement. Ensuite, nous avons entamé la liste des technologies clés de ce modèle, y compris la virtualisation, les technologies et services web. D'ailleurs, ce chapitre présente les avantages, les défis, et les principaux acteurs de l’écosystème cloud.

II.2 Big Data

II.2.1 Introduction

Pour faire face à l'explosion du volume des données, un nouveau domaine technologique a vu le jour : le big data comme un paradigme qui a révolutionné les pratiques et les architectures classiques de traitement et de stockage de données. Les géants du web, tels que Google, Amazon, Facebook, ont été les tous premiers acteurs à déployer ce type de technologie. Ensuite, les solutions big data suscitent plus l'intérêt de différentes entreprises, qui sont destinées à mettre en place des technologies big data pour résoudre les défis liés à la gestion des bases de données massives. Dans ce chapitre, nous présentons le concept du big data et ses technologies innovantes. Ainsi, nous mettons l'accent sur l'architecture de stockage distribué et les Frameworks de traitement en parallèle fournis par l'écosystème big data. Ensuite, nous abordons la nouvelle génération des systèmes de gestion des bases de données NoSQL. Vers la fin, nous présentons les méthodes d'apprentissage machine et leurs algorithmes.

II.2.2 Courte Histoire

Le Big Data correspond au premier moment de l'histoire au cours duquel la capacité de l'humanité à produire de l'information a dépassé sa capacité de la stocker et de la restituer. Nous présentons dans cette section l'histoire récente du big data en tant que technologie de gestion de l'information.

Le terme big data est apparu pour la première fois en octobre 1997 selon les archives de la bibliothèque numérique IEEE, dans un article scientifique publié [57] par les deux chercheurs de NASA: Michael Cox, et David Ellsworth, sur les défis technologiques à relever pour visualiser les grands ensembles de données et la difficulté des systèmes à pouvoir traiter les volumes massives des données.

En 2003, Google publie un premier papier sur le Google File System[58], et révèle ainsi les premiers secrets du succès de son moteur de recherche. En 2004, Google a développé MapReduce comme un Framework de calculs parallèles et distribués de données volumineuses. Une année plus tard, Doug Cutting et Michael Cafarella, à l'époque employés chez Yahoo et inspirés par les travaux de Google, créent Nutch Search Engine, qui deviendra aujourd'hui Apache Hadoop[59].

En 2012, Barack Obama, le président des États-Unis d'Amérique annonçait la mise à disposition de 200 millions de dollars pour un fond de recherche sur la thématique du big data. Depuis cette date plusieurs recherches ont été menées par des informaticiens venus de Facebook, Google et autres géants du web[60]. Par la suite, les technologies big data connaissaient beaucoup de développement en termes de stockages (HDFS, NOSQL), de traitement (Apache Spark, Storm) et d'analyse (machine leaning, deep learning).

II.2.3 Définition

Le big data est un terme à la mode, abondamment utilisé de nos jours, pour décrire une large palette de technologies. Il vise à proposer une alternative aux solutions traditionnelles de bases de données et d'analyse (serveur SQL, plateforme de Business Intelligence, etc.).

En général, le terme big data signifie gégadonnées, grosses quantités de données ou encore données massives. Il désigne un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment traiter. En effet, il est estimé que 90 % des données récoltées depuis le début de l'humanité ont été générées ces deux dernières années[61].

L'augmentation exponentielle (Figure II.9) du volume des données aboutit aujourd'hui à chercher de nouvelles méthodes, architectures, et outils pour traiter et analyser les données décisionnelles et même celles en temps réel. Nous procréons environ 2,5 trillions d'octets de données tous les jours. Ce sont des informations provenant de partout : messages que nous nous envoyons, vidéos que nous publions, informations climatiques, signaux GPS, enregistrements transactionnels d'achats en ligne et bien d'autres encore. Les géants du Web, au premier rang desquels Google, Amazon, Facebook, ont été les tous premiers à déployer ce type de technologie [62].

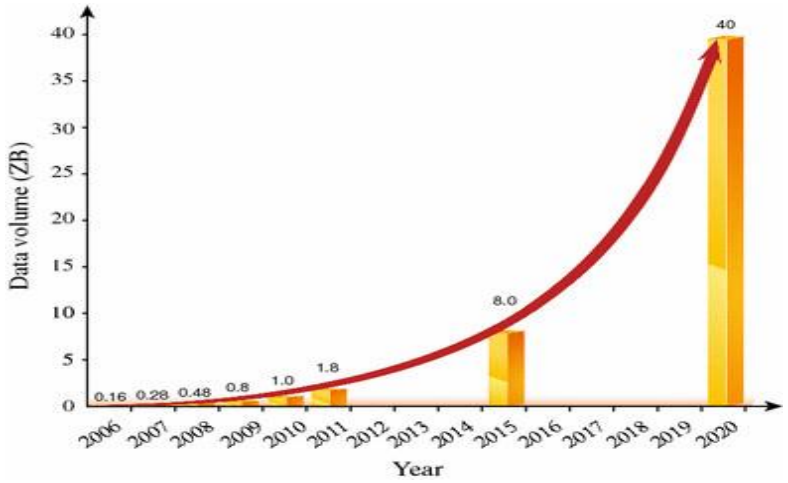


Figure II.9: La croissance mondiale du volume de données entre 2006 et 2020 [63]

En fait, le terme big data désigne non seulement un stockage de données très important, mais aussi des besoins de traitement sur ces volumes de données. Il s'agit, en plus de stocker les données sur un système de fichier distribué, de pouvoir effectuer des traitements sur ces données de façon également distribuée et de pouvoir en restituer les résultats.

International Data Corporation (IDC) est un leader des solutions big data. Selon lui, “Les Technologies Big Data décrivent une nouvelle génération de technologies et d’architectures, conçues pour extraire économiquement la valeur à partir de très grands volumes d'une grande variété de données, en permettant une grande vitesse de capture, de découverte et/ou d'analyse” [64].

Gartner dans son rapport [65] a donné la définition suivante: “le big data regroupe des données présentant une grande variété, arrivant en volumes croissants, à grande vitesse. C’est ce que l’on appelle les 3V”. En d'autres termes, le big data est composé des données complexes et très diversifiées. Ces données volumineuses sont générées à grande vitesse que les systèmes actuels de bases de données deviennent incapables de les gérer. Alors, les outils big data peuvent apporter des solutions efficaces pour surmonter ces défis.

Le NIST suggère que, “On parle du big data lorsque le volume de données, la vitesse d'acquisition, ou la représentation des données limite la capacité d'effectuer une analyse efficace en utilisant des modèles relationnels classiques ou nécessite l'utilisation d’une échelle horizontale significative pour un traitement efficace” [66]. Particulièrement, le big data peut être catégorisé en science de données (data science) et technologies du big data. La science de

données est “l'étude des techniques couvrant l'acquisition, le conditionnement, l'évaluation et l’exploitation des données”, alors que les technologies big data sont « des bibliothèques logicielles, des Frameworks, des outils avec leurs algorithmes associés qui permettent des traitements distribués et l'analyse des problèmes big data entre les groupements (clusters) de machines [64] ».

II.2.4 Domaine du big data

Le big data est un domaine pluridisciplinaire rassemblant plusieurs disciplines (Figure II.10), qui sont[67]:

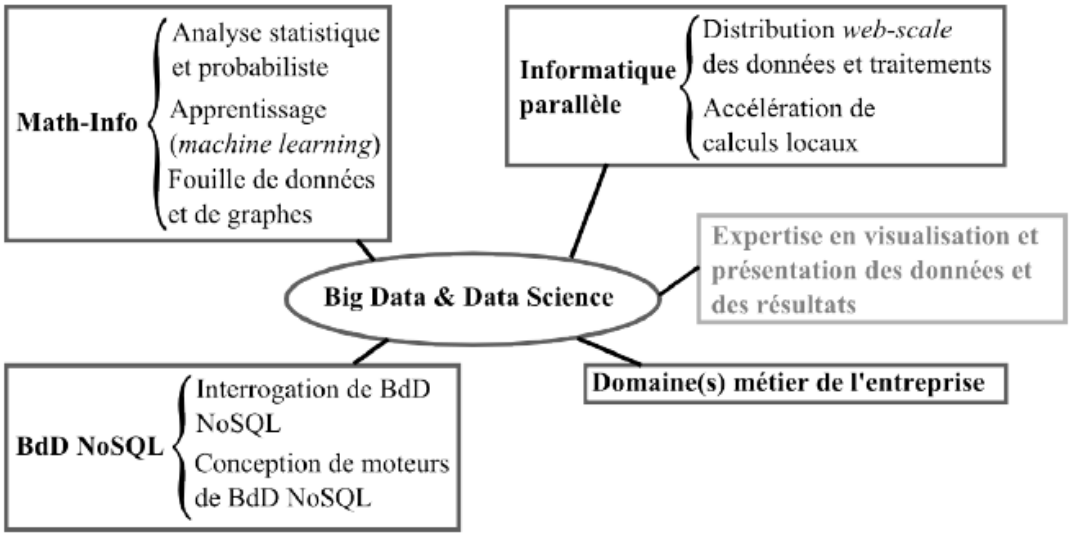


Figure II.10: Domaine du big data

- Les mathématiques constituent la base des technologies big data, y compris, les statistiques et les théorèmes probabilistes sur lesquelles se basent les algorithmes de l'apprentissage machine. Cette branche du big data, appelée souvent le data science, forme le cœur mathématique du big data.
- L'informatique distribuée est la deuxième discipline permettant le stockage et la manipulation de données à grande échelle. Il s'agit d’un ensemble d’algorithmes et de programmes capables de répartir le traitement et l'analyse de données sur plusieurs machines ou serveurs distants par l’intermédiaire des protocoles réseau. Cette approche permet le traitement des quantités de données très volumineuses en un temps presque réel. Les technologies implémentant cette approche sont nombreuses telles que MapReduce et Apache Spark. Le big data utilise une architecture souple fondée

sur les calculs parallèles à haute performance pour l'analyse de données et l'apprentissage automatique nécessitant des puissances de traitement énormes via des clusters de machines.

- Les bases de données NoSQL (Not Only SQL) est un élément essentiel du big data. Elles sont considérées comme des bases de données de nouvelle génération, car elles sont généralement non relationnelles, distribuées et open source. Elles permettent de stocker, gérer et manipuler des données structurées et complexes. Certaines de ce type de bases de données ont été conçues en particulier pour être flexibles et évolutives en supportant des volumes importants de données hétérogènes.
- Le big data comprend également d'autres disciplines informatiques, y compris, la visualisation des données, la Business Intelligence (BI), la data mining ou apprentissage automatique, etc. Les domaines d'application des solutions big data sont très divers comme l'e-commerce, l'énergie, éducation, recherche scientifique, etc. Par conséquent, la connaissance du domaine d'activité de l'entreprise est nécessaire pour les spécialistes du big data, appelés data scientists (spécialistes de données), pour bien exploiter et interpréter les données et prendre les bonnes décisions.

II.2.5 Caractéristiques du big data

Les caractéristiques du big data sont en nombre de cinq, ou le modèle 5V. Les V font référence à cinq éléments clés qui sont: le volume, la variété, la vélocité, la véracité et la valeur (Figure II.11).

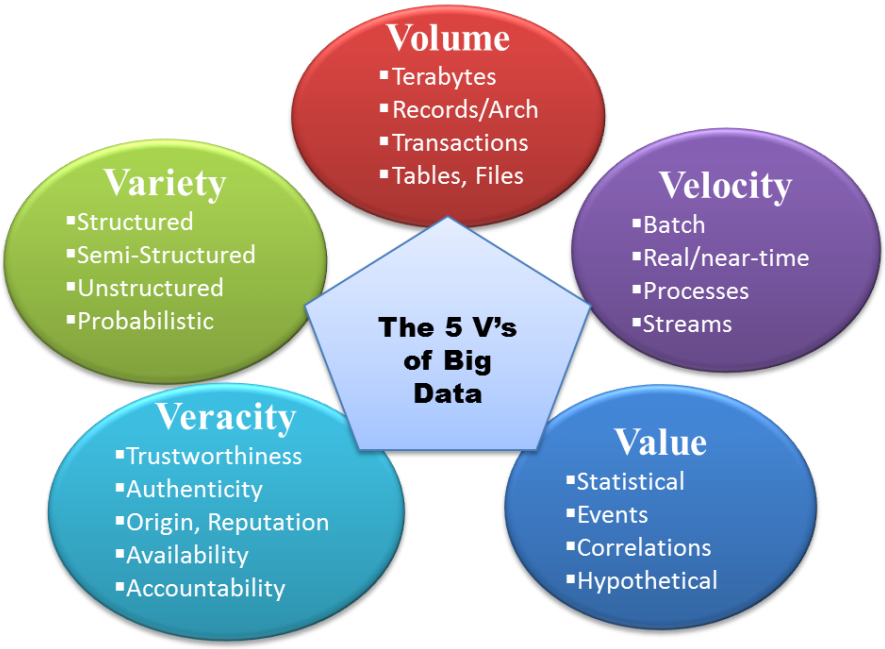


Figure II.11: Caractéristiques du big data ou 5V

II.2.5.1 Volume

Le volume désigne la quantité très volumineuse des données à stocker, traiter, analyser et diffuser par des outils et technologies big data. En effet, le volume de données, créé et géré par les entreprises, est en constante augmentation. Aujourd'hui, les données se comptent en téraoctets (e^{12} octets), voire en pétaoctets (e^{15} octets), exaoctets (e^{18} octets).

II.2.5.2 Variété

La variété fait référence à la diversité des formats et des types de données. En fait les technologies big data peuvent traiter des données hétérogènes provenant de diverses sources. Le format classique est celui de la base de données relationnelle, dans laquelle le donnée est stockée selon un schéma rigide et organisé, par exemple une table. Mais actuellement, plus de 80 % des données générées par les entreprises sont de type semi-structuré et non-structuré, par exemple, le texte, l'image, la vidéo, la voix, etc. Pour cela, le big data offre la capacité de réunir toutes ces données et de les analyser.

II.2.5.3 Vélocité

La vélocité est l'un des points forts du paradigme big data qui fait référence à l'aspect dynamique et temporel de la génération des données. Ces données sont produites en flots continus et doivent être utilisées immédiatement afin de prendre les bonnes décisions. Il s'agit de pouvoir traiter et analyser les données collectées en temps réel. Mais pour cela, il faut une

puissance de calcul et des outils d'analyse très performants. Dans ce contexte, on peut faire appel à l'infrastructure du cloud offerte en tant que services (IaaS).

II.2.5.4 Véracité

La véracité fait référence à la qualité et validité des données. Il s'agit de la faible fiabilité et au désordre qui règnent dans la donnée. Celle-ci manque trop souvent d'exactitude et de précision, ce qui la rend peu contrôlable. Elle désigne aussi la confiance que l'on peut avoir dans les données. Un décideur sur trois ne fait pas confiance aux données sur lesquelles il se base pour prendre ses décisions [68]. Alors l'une des missions du big data est d'apporter un peu d'ordre à tout cela non pas en organisant la donnée, mais plutôt en organisant son accès et en permettant d'y associer les analyses appropiées qui correspondent aux besoins des utilisateurs [69].

II.2.5.5 Valeur

La valeur est le dernier "V" à prendre en compte quand on parle du big data. Il fait référence à la capacité du big data à tirer la valeur à partir des masses énormes de données hétérogènes. C'est bien d'avoir accès aux volumes importants de données mais encore faut-il les transformer en valeur. En effet, il faut que la nouvelle génération des technologies distribuées (Hadoop, Spark, ...), de SGBD NoSQL, et des méthodes avancées, servent à quelque chose d'utile et d'utilisable pour les entreprises et les universités.

II.2.6 Types de données en big data

En général, il y a trois types de données à prendre en compte, structurés, non structurés et semi-structurés. Parmi ceux-ci, les deux derniers sont nouveaux dans le Big Data. La Figure II.12 montre les différents types de données.

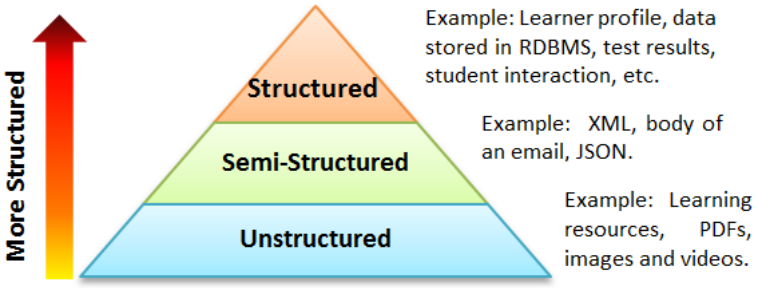


Figure II.12: Types de données en big data

Premièrement, Les données structurées sont celles dont l'ensemble des valeurs possibles est déterminé et connu à l'avance. Elles sont organisées dans un format rigide qui permet d'y accéder et de les manipuler très facilement. Les données structurées sont souvent gérées par des systèmes de gestion des bases de données relationnelles (SGBDR) sous forme de tables, et elles sont interrogées à l'aide d'un langage de requête tel que SQL ou PL/SQL.

Deuxièmement, les données non structurées sont l'opposé des données structurées. Ce sont des données qui ne disposent pas d'un modèle de données et ne sont pas organisées d'une manière prédéfinie. Ce type de données peut avoir plusieurs formats tels que les documents texte, les images, les fichiers audio, les vidéos et les publications sur les réseaux sociaux, etc. Elles sont à la fois complexes et volumineuses et dont les bases de données traditionnelles ne peuvent pas les manipulées ou interrogées.

Outre les données structurées et non structurées, il existe également une troisième catégorie, appelée données semi-structurées. Les données semi-structurées sont des informations qui ne sont pas stockées dans un jeu de données structuré, mais leur structure contenant des balises facilitant leur manipulation et leur analyse. Des exemples de données semi-structurées peuvent inclure des fichiers XML, des e-mails textes et des documents JSON.

II.2.7 Comparaison entre SGBDR et big data

Les SGBDR sont des systèmes de bases de données relationnelles pour la gestion et l'organisation de données structurées dans des tables reliées entre elles par des relations. Le succès de ces systèmes de gestion de données provient de son intelligence et de ses capacités à répondre aux défis du moment, en permettant le stockage et le traitement des données à une époque où les capacités matérielles étaient limitées. Une base de données relationnelle est idéale pour garantir d'excellentes performances et l'intégrité des données sur un serveur dont les ressources sont par définitions limitées[70].

Les SGBDR ont la capacité de traiter des giga-octets de données, mais lorsque le volume de données devient très important et d'une grande variété, les SGBDR ne peuvent pas répondre aux besoins de traitement et d'analyse de ces données massives. Dans ce contexte, le big data est conçu en particulier pour résoudre ces problèmes en offrant une large gamme des solutions technologiques et des outils puissants, en termes de stockage réparti, de traitement en parallèle et d'analyse avancée des données volumineuses provenant de diverses sources.

Les systèmes big data sont basés sur une architecture distribuée afin de répartir les données et les traitements. Dans ce type d'architecture les données sont stockées sur plusieurs machines ou serveurs distants à travers des systèmes de fichiers distribués tels que HDFS [71] ou via des bases de données NoSQL comme Cassandra [72] et MongoDB, etc. Généralement l'infrastructure du cloud computing, fournie sous forme de services, est utilisée pour assurer la gestion efficace des clusters et garantir la disponibilité et la haute performance. De plus, ces systèmes permettent le traitement et l'analyse des données hétérogènes qui ne peuvent pas être manipulées avec les solutions de gestion des bases de données relationnelles. La Table III.1 donne une comparaison synthétique entre RDMBS et big data.

Tableau II-2: Comparaison entre le big data et les SGBDR

	Traditional RDBMS	**Big Data**
Volume	GB	TB, PB, EB, ZB
Types de données	Structurées	Semi-structurées ou non structurées
Architecture	Centralisée	Entièrement distribuée
Intégration	facile	Difficile
Stockage	SGBDR (MySQL, PostgreSQL, Oracle)	HDFS, NoSQL
Modèle	Relationnel	Non Relationnel
Schéma	Statique (défini au départ)	Dynamique

En premier lieu, concernant la taille de données manipulées, les systèmes de gestion des bases de données relationnelles ont une capacité de stockage limitée (GB) par rapport aux solutions du big data qui peuvent conserver des TeraBit, PetaBit, ExaBit voire même ZettaBit (10^{21} bit). À propos du type de données stockées et traitées, le SGBDR peut travailler seulement avec des données structurées sous forme de tables, alors que la génération de NoSQL et des systèmes distribués du big data supportent tous les types de données (structurées, semi-structurées et non-structurées). En revanche, l'intégration des données est très facile pour les SGDBR contre le big data qui demande beaucoup de temps et d'efforts dans le chargement et la préparation des données.

II.2.8 Les technologies big data

Toute la puissance du big data repose sur ses technologies de pointe. Ces technologies couvrent tous les aspects de la gestion de l'information, y compris ceux qui sont destinées au stockage et la conservation distribuées des données, aussi ceux qui sont dédiées au traitement

parallèle en lots ou en temps réel des quantités massives des données hétérogènes, et plein d'autres outils de collecte, d'analyse avancées et de visualisation des résultats.

II.2.8.1 Hadoop

Hadoop [59] a vu le jour suite à la limite des approches classiques pour répondre aux défis des traitements et d'analyses des volumes importants de données produits par les entreprises. Tout est commencé en 2004 lorsque Google publie son article “MapReduce” [73] comme un modèle de traitement réparti sur de grands clusters. Par la suite, en 2005, la solution Hadoop a été mise en place, inspirée de la publication de Google. Une version open source de Hadoop a été publiée en 2007. Hadoop est composé principalement d'un système de fichier distribué (HDFS) permettant le stockage des données et de MapReduce en tant que Framework de programmation en parallèle. La technologie Hadoop est déjà utilisée par plusieurs entreprises et fournisseurs cloud tels que Yahoo [74], Facebook [75], Amazon Web Sevices [47] et Microsoft Azure [76].

Apache Hadoop constitue le cœur des traitements big data. Il s'agit d'un ensemble de composants logiciels open-source pour effectuer un calcul distribué performant et évolutif. L'idée fondamentale de Hadoop consiste à élaborer un environnement informatique distribué. Elle vise à fournir une infrastructure de clusters regroupant nombreux ordinateurs, à travers le réseau, capables d'effectuer des traitements à grande échelle sur des quantités massives des données. Il est le plus utilisé actuellement dans le développement des solutions big data parce qu'il offre un environnement complet de Frameworks et outils très puissants. Les modules du projet Hadoop sont au nombre de quatre.

- Hadoop Common: les composants de base qui prennent en charge les autres modules Hadoop.
- Hadoop HDFS: Un système de fichiers distribué qui fournit un accès à haut débit aux données d'application.
- Hadoop Yarn: un Framework pour la planification des tâches et la gestion des ressources de cluster
- Hadoop MapReduce: Un modèle de programmation basé sur YARN pour le traitement parallèle de grands ensembles de données.

L'idée fondamentale de Hadoop est de permettre le traitement distribué des grands ensembles de données sur des clusters de machines ou de serveurs en utilisant des modèles de programmation simples. Hadoop est conçu pour évoluer à partir des serveurs uniques vers des milliers de machines, chacune offrant un calcul et un stockage local. Plutôt que de compter sur le matériel pour fournir une haute disponibilité, la bibliothèque logicielle elle-même est conçue pour détecter et gérer les défaillances au niveau de la couche application, délivrant ainsi un service hautement disponible sur un cluster d'ordinateurs.

II.2.8.2 Distributions de Hadoop

Apache fait l'objet de plusieurs distributions offrant une solution stable et plus complète des packages avec des fonctionnalités sophistiquées facilitant l'installation et l'administration des projets big data. Ci-dessous une liste non exhaustive des entreprises qui proposent leur propre distribution de Hadoop.

1) HortonWorks [77] est une entreprise de développement des projets big data autour de Hadoop. Il a été fondé par une équipe de Yahoo en 2011 à Santa Clara. HortonWorks est un grand contributeur du Framework Hadoop et l'éditeur d'Hortonworks Data Platform (HDP). HDP est une plate-forme basée sur Hadoop qui comprend nombreuses technologies, y compris HDFS, Yarn, Apache Hive, Pign HBase, Zookeeper. Elle est destinée pour stocker, manipuler et analyser d'importantes quantités de données collectées depuis diverses sources. En plus, HortonWorks utilise Ambari pour faciliter l'installation et la gestion automatique des nœuds sur le cluster Hadoop.
2) Cloudera [78] est à l'origine une entreprise créée en 2008 par des experts provenant de Google, Facebook, Yahoo et Oracle. Elle est spécialisée dans l'implémentation des distributions Hadoop destinées aux entreprises comme une plate-forme big data. En plus les fonctionnalités de bases de Hadoop, Cloudera offre des logiciels propriétaires pour assurer la bonne gestion des clusters. Il y a une version open source de Cloudera, avec des limitations, destinée aux utilisateurs normaux à des fins de test et de recherche.
3) MapR [79] est une plateforme big data regroupant les Frameworks d'Apache Hadoop et Spark. Elle a été fondée en 2009 pour but de développer des projets big data de haute performance. MapR a été récompensé de la meilleure solution big data pour

l'année 2017 pour le “Internet of Things and Companies Driving Innovation” [80]. MapR propose son propre système de fichiers distribués (MapR FS) et une version modifiée de MapReduce (MapR).

4) HP Vertica Analytics [81]: Cette distribution a été développée par HP qui fournit une grande gamme de solutions. Son produit principal est la plate-forme Vertcia Analytics, conçue pour gérer et analyser des volumes massifs de données, y compris les pétaoctets d'informations structurées et non structurées. Il est basé sur Hadoop et permet de bonne performance dans le traitement des requêtes client via SQL Analytics.
5) Amazon [47] : Amazon Web Services est une filiale de l'e-commerce Amazon qui loue des services de cloud computing. Parmi ces services, il y a de nombreuses offres du big data, tels que Elastic MapReduce basé sur Hadoop, Kinesis Analytics pour analyser les données massives en temps réel. Il propose également des solutions pour les bases de données non relationnelles (DynamoDB) et distribuées (HBase).

   Google, le premier leader commençant la recherche dans le domaine du big data, continue de développer ses plates-formes d'analyse, de traitement et de stockage des données réparties. Il propose une plateforme d'analyse rapide, basée sur le cloud, appelée BigQuery. Google propose également Dataflow, un service de traitement des données en temps réel et Dataproc [82] permettant d’exécuter l’écosystème Apache Spark et Apache Hadoop sur Google Cloud Platform.
6) VMware s’est orienté vers la virtualisation de Hadoop sur vSphere en offrant des choix techniques pour exécuter des applications Hadoop et Spark sur des machines virtuelles. Il s'agit d'une extension big data qui permet de provisionner, créer, gérer ou surveiller entièrement des clusters basés sur hadoop. Il prend en charge plusieurs distributions de Hadoop, y compris Hortonworks, Cloudera, MapR.
7) SAP [83] est une multinationale développant des systèmes d'information de gestion des entreprises. Elle utilise son outil big data SAP VORA, qui intègre Hadoop, pouvant exécuter et analyser jusqu’à 80 téraoctets de données provenant de diverses sources. SAP VORA permet des analyses avancées comme l'analyse prédictive des données issues de graphes et les données en temps réel.

II.2.8.3 HDFS

Hadoop Distributed File System (HDFS) [71] est un système de fichiers distribué doté d'une architecture master/worker. Son développement est inspiré du système de fichier GFS (Google File System). HDFS est hautement tolérant aux pannes et il est conçu pour être déployé sur du matériel de base (serveurs standards). HDFS fournit un accès haut débit aux données d'application. Il est souhaitable pour les applications gérant des grands ensembles de données. Il crée une abstraction de ressources disques en permettant la gestion de plusieurs ressources physiques distribuées comme s'il y a d'un disque dur unique.

Généralement, HDFS est constitué de deux composants, nommés Namenode et Datanode. Un cluster Hadoop entièrement distribué comprend un Namenode et nombreux Datanodes. Les données stockées dans les Datanodes sont distribués à travers le cluster sous forme de fichiers structurées en blocs. La taille du bloc peut être paramétrée via le fichier de configuration hdfs-site.xml (le paramètre dfs.blocksize), par défaut, la taille du bloc est 128 MO.

La Figure II.13 montre l'architecture du HDFS décrivant ses principaux composants.

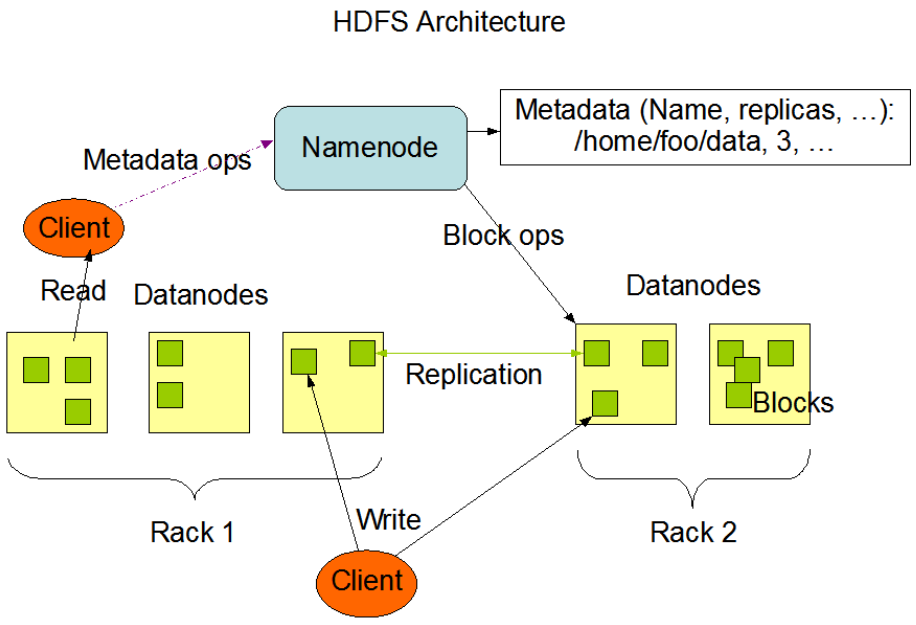


Figure II.13: Architecture de Hadoop Distributed File System (HDFS)

Le Namenode est un composant master permettant la gestion du système de nommage des fichiers et répertoires gérés par le cluster Hadoop. HDFS stocke les données dans des blocs et Namenode conserve et gère les emplacements des fichiers / répertoires. Namenode suit la

trace où les données sont stockées et lorsqu'un client interroge ces données, Namenode fournit les emplacements des blocs de données demandées. De même, lorsqu'un client souhaite enregistrer de nouvelles données, le Namenode fournit le chemin du bloc disponible auquel les données peuvent être stockées.

Datanodes stockent les données dans Apache Hadoop. Alors que le Namenode fournit aux clients les emplacements des blocs à partir desquels les données peuvent être lues et où les nouvelles données pouvent être stockées, les Datanodes lisent et écrivent les requêtes des utilisateurs sur des fichiers HDFS. Les Datanodes effectuent également les tâches de création, de réplication et de suppression de bloc sous les instructions du Namenode.

Le système de fichier distribué de Hadoop offre plusieurs avantages, à savoir :

- Tolérant aux pannes: HDFS a le potentiel de détecter automatiquement la défaillance, surtout les pannes matérielles du disque dur et les réparer rapidement. Il est résistant aux erreurs et gère efficacement les exceptions.
- Evolutivité (Scalability): l'évolutivité fait référence à la possibilité de supporter une montée en charge horizontale en ajoutant des serveurs à un cluster pour faire face aux demandes accrue en termes de ressources machine. C'est une stratégie très efficace utilisée actuellement par les fournisseurs du cloud.
- Simplicité: Il est très simple de mettre en place le système HDFS en assurant la compatibilité avec différents systèmes d'exploitation.

II.2.8.4 MapReduce

MapReduce[73] est un modèle de programmation massivement parallèle adapté au traitement distribué des données gigantesques. Il s'agit d'un Framework de Google basé sur la décomposition d'un calcul important en plusieurs tâches plus petites, qui seront distribuées sur un cluster de machines pour être traitées et qui produiront chacune une partie du résultat final. Principalement, le fonctionnement du MapReduce s'articule sur deux étapes : Map, où les problèmes sont découpés en sous-problèmes et distribués sur plusieurs nœuds, les sous problèmes sont ensuite traités par différents nœuds d'un cluster à travers la fonction Map (clé, valeur). Par la suite vient l'étape Reduce qui associe toutes les valeurs correspondantes à la même clé à une unique paire (clé, valeur) puis déduit le résultat final. La Figure II.14 décrit le processus pour calculer le nombre des mots à l'aide de MapReduce.

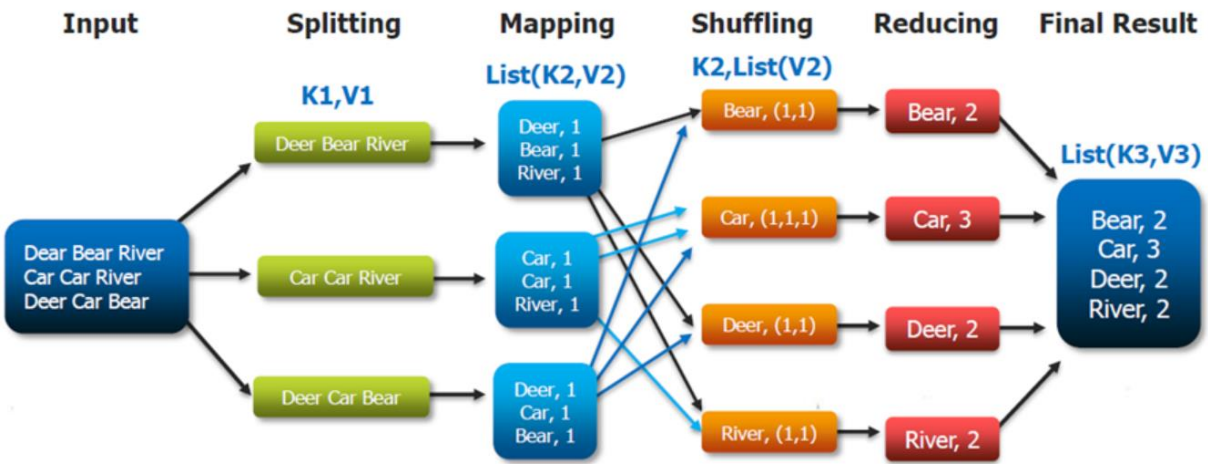


Figure II.14: Le processus de fonctionnement de MapReduce[13]

Dans l'architecture Hadoop, le nœud maître de MapReduce est appelé le "JobTracker", et les nœuds esclaves sont appelés les "TaskTracker". Le JobTracker s'occupe de la coordination des différents calculs en ordonnançant les tâches qui s'exécutent sur les TaskTrackers. Ces derniers exécutent les tâches et envoient un état d'avancement au JobTracker, qui contient un registre de suivi de tous les calculs en cours. Si une tâche échoue, le JobTracker peut la relancer sur un autre TaskTracker (Figure II.15).

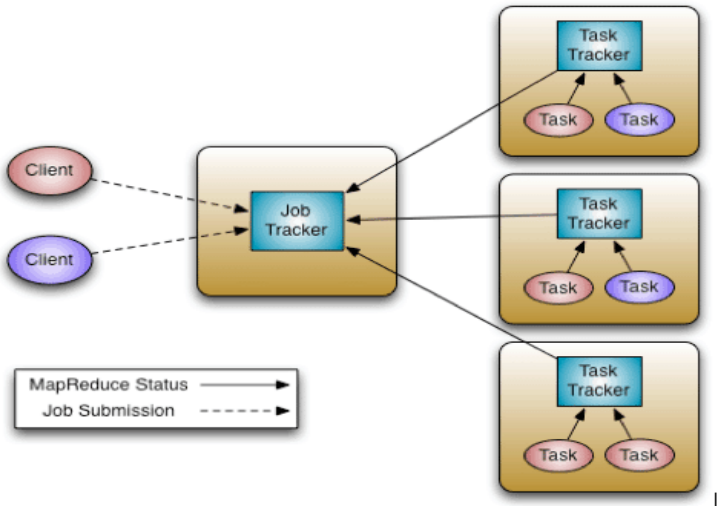


Figure II.15: Architecture de MapReduce

[13] https://wikis.nyu.edu/display/NYUHPC/Big+Data+Tutorial+1%3A+MapReduce

En plus, ce Framework offre plein de possibilités de programmation, telles que les jointures, la fusion, le tri, etc. MapReduce est une solution très efficace pour les calculs en parallèle à passe unique. Mais, il n'est pas souhaitable pour les applications qui nécessitent des traitements et des algorithmes récursifs. En effet, chaque phase de traitement constitue des fonctions Map et Reduce; et les données résultant de chaque étape sont stockées sur un système de fichier distribué (HDFS) avant de passer à la phase suivante. Cette approche est peu rapide et demande beaucoup de temps à cause du stockage et de la réplication des résultats sur disque. Pour résoudre ce problème, Plusieurs autres alternatives big data ont été développées en se basant sur le traitement en mémoire, c'est le cas du Framework Apache Spark.

II.2.8.5 Apache Spark

Apache Spark [84] est un Framework d'analyse unifié pour le traitement distribué big data. Il a originellement été développé par l'Université UC Berkeley en 2009 et a été passé Open Source sous forme de projet Apache en 2010. La puissance de Spark réside dans sa capacité à réaliser des calculs en mémoire (In-memory) ce qui lui permet d'être plus rapide 100 fois que MapReduce qui utilise le disque. Spark peut être déployé en mode standalone, Hadoop Yarn, Mesos ou Kubernetes. Il supporte plusieurs systèmes de gestion de données tels que HDFS, HBase, Hive, Cassandra, Amazon S3 et plusieurs SGBD (Oracle, Mysql).

II.2.8.5.1 Présentation

Spark est un vaste écosystème d'outils et d'APIs. Il est composé de deux niveaux: des APIs de bas niveau ou de base et un ensemble riche d'APIs et de fonctionnalités supplémentaires. Généralement, le composant Spark Core est responsable des opérations d'entrée/sortie, de la planification, de la répartition, de l'ordonnancement, de la surveillance des tâches sur les clusters d'ordinateurs distribuées, la connexion aux systèmes de stockage externes, la récupération en cas de pannes et la gestion efficace de la mémoire (RDD). Alors que Spark APIs fournit une collection très riche de bibliothèques permettant de prendre en charge une variété de cas d'utilisation et d'analyses avancées d'apprentissage automatique. Parmi ses composants (Figure II.16) on trouve Spark SQL, Streaming, MLlib, GraphX. En outre, il supporte plusieurs langages de programmation, y compris Scala, Java, Python, R [85].

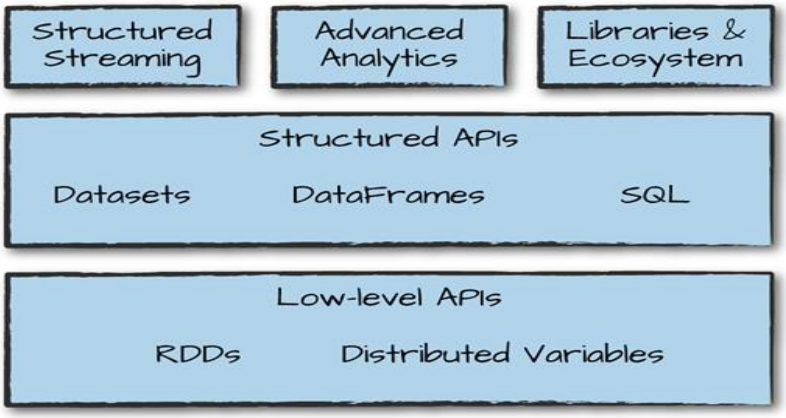


Figure II.16: Les composants du Framework Spark

- Spark SQL: est un module Spark dédié pour la manipulation des données structurées. Il est conçu pour l'objectif d'exploiter des données massives via une API de type JDBC offrant de nombreux interfaces de connexion pour accéder aux diverses bases de données (Oracle , MariaDB, PostgreSQL, …). En plus, il permet d'exécuter des requêtes SQL pour extraire, transformer et charger les données sous différents formats. En fait, SQL est issu des bases de données relationnelles tandis que dans Spark, il peut être utilisé pour traiter des données hétérogènes.
- Spark Streaming: est une extension de Spark Core API destinée pour le traitement en temps-réel des données en flux. Il s'appuie sur un mode de traitement micro-Batch et fondé sur l'ancienne version DStream supportant des flux continus de données. Il utilise le RDD (Resilient Distributed Dataset) comme une abstraction de données capable de gérer et de réutiliser efficacement les données dans un cluster de nœuds.
- MLlib: est la bibliothèque de Machine Learning de Spark. Son objectif est de rendre l'apprentissage automatique pratique et évolutif. Elle contient la plupart des algorithmes d'apprentissage automatique et de data mining, y compris la Classification, le Clustering, la Regression, le Collaborative Filtering et le Frequent Pattern Mining, en plus des méthodes statistiques comme le calcul de corrélation et les tests d'hypothèses.
- GraphX: Il s'agit d'un nouveau composant dans Spark. Il permet le traitement des données issues des graphes. Graphx étend les RDDs de Spark Core en introduisant RDD dédié au graphe, appelé Resilient Distributed Dataset Graph, un multi-graphe ayant des propriétés attachées aux nœuds et aux arcs. En outre, GraphX fournit un

ensemble très riche d'algorithmes (PageRank, Shortest Path) permettant de faciliter l'analyse des graphes.

- SparkR: il s'agit d'un package R qui fournit une interface légère supportant le langage d'analyse statistique R tout en permettant de la manipulation reparties des données et l'utilisation des algorithmes de machine learning. Dans sa version 2, SparkR fournit une implémentation distribuée de DataFrame qui offre plein d'opérations telles que la sélection, le filtrage et l'agrégation sur des grands ensembles de données.

Par ailleurs, Apache Spark permet aux programmeurs de développer des pipelines des traitements complexes en plusieurs phases en s'appuyant sur le pattern DAG (Directed Acyclic Graph). Pour optimiser l'exécution d'un job (travail), Spark fait appel à un planificateur de bas niveau nommé Catalyst Optimizer qui décide comment le code utilisateur doit être exécuté et établit un plan pour le faire. Enfin, spark exécute le code et envoie le résultat à l'utilisateur.

II.2.8.5.2 Architecture de Spark

Apache Spark se constitue d'un certain nombre d'éléments qui se communiquent entre eux pour traiter, en parallèle, le travail soumis par un client. En fait, Une application Spark s'exécute comme un ensemble de processus indépendants sur un cluster de nœuds, coordonnés par un objet SparkContext du programme principal, appelé Driver Program. En général, les applications Apache Spark se composent de deux composants essentiels: le processus pilote (***driver process)*** et les processus exécuteurs (***executor processes***). Le processus pilote stocke les informations pertinentes sur SparkSession pendant le cycle de vie de Spark Application, il exécute également la fonction principale main (), distribuant et ordonnançant le travail en parallèle entre les exécuteurs (workers) à travers le cluster. D'autre part, les processus exécuteurs sont chargés du traitement des tâches (Tasks) qui leur est assigné par le processus pilote et de la soumission de l'état du d'avancement d'exécution au processus pilote. La Figure II.17 donne un schéma synthétique de l'architecture du Framework Apache Spark.

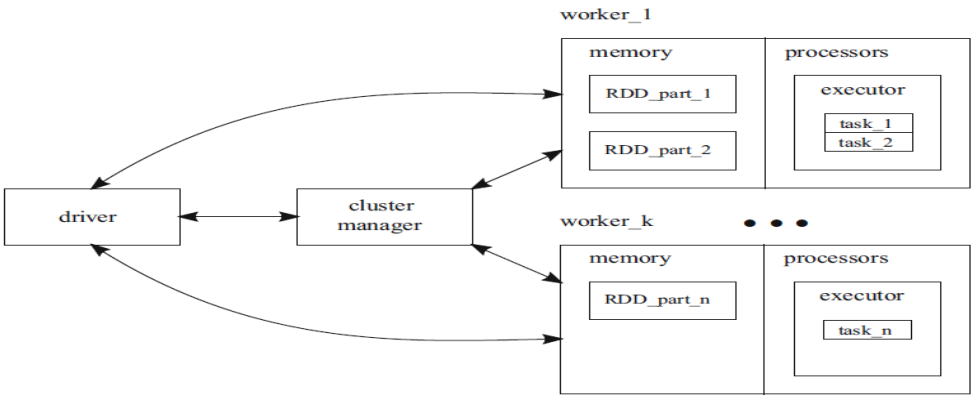


Figure II.17:Architecture du Framework Apache Spark[14]

Spark a besoin d'un gestionnaire de ressources tel que Yarn, Mesos ou son gestionnaire de cluster autonome (Standalone), qui contrôle les machines physiques et alloue les ressources requises aux applications Spark pour effectuer les calculs répartis sur le cluster. Dans le cas d'un gestionnaire de ressources Yarn, une instance de l'objet SparkSession se connecte au Yarn Cluster Manager, pour l'allocation des ressources nécessaires. Une fois connecté, Spark lance des exécuteurs sur les nœuds workers qui démarrent les processus du traitement et du stockage des données. Il envoie ensuite le code de l'application (jar) aux exécuteurs. SparkSession distribue, ensuite, les tâches aux exécuteurs pour qu'ils les traitent.

II.2.8.5.3 Les avantages du Framework Apache Spark

Apache Spark présent autant d'avantages par rapport à d'autres Frameworks du traitement en parallèle des données massives. D'abord, Spark est un écosystème complet et unifié qui répond à un nombre très important des problèmes du big data et couvre nombreux cas d'utilisation. Spark a le potentiel de prendre en charge divers jeux de données que ce soit leur nature (text, csv, logs), ou leur type (Batch et Streaming). En plus, Il supporte une large gamme de langages de programmation pour permettre aux développeurs de créer des applications distribuées en Scala, Java, Python, R, SQL. Ces applications peuvent être exécutées sur différents gestionnaire de cluster (Standalone, Hadoop Yarn, Messos, cloud). En outre, Spark s'appuie sur le traitement en mémoire ce qui le rendre très efficace en termes de temps d'exécution et de performances de calcul. En effet, Spark peut être cent fois plus rapide

[14] https://spark.apache.org/docs/latest/img/cluster-overview.png

que MapReduce et ceci revient essentiellement à deux raisons majeures : d'une part, Spark permet de stocker les données intermédiaires en mémoire et d'autre part, parce qu'il possède un moteur d'optimisation logique et autre physique capables d'optimiser largement le temps de calcul. Également, Spark est évolutif, puisqu'on peut facilement augmenter le nombre des nœuds d'un cluster pour améliorer les performances du traitement des données.

II.2.8.6 Apache Yarn

YARN, en anglais Yet Another Resource Negotiator, est le successeur de MapReduce (version 2). Il s'agit d'un Framework dédié à la planification des tâches et la gestion des ressources du cluster. L'idée principale de Yarn est de séparer la gestion des ressources et le modèle du calcul. Une telle séparation permet au Framework de prendre en charge un certain nombre d'autres Frameworks du traitement distribué tels que Spark[84], Giraph[86], Storm[87].

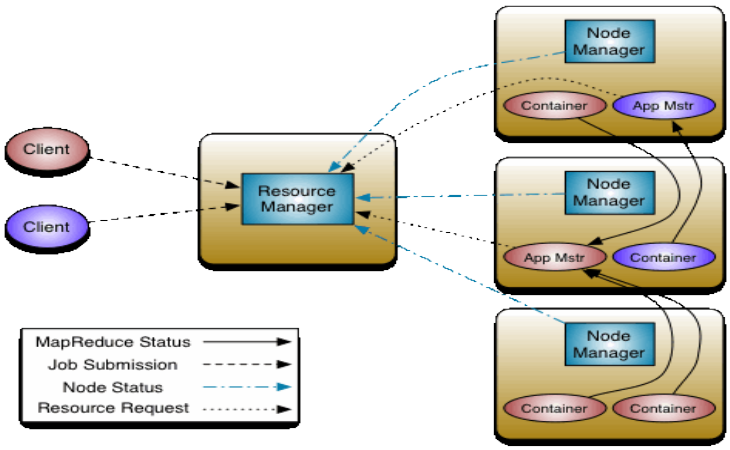


Figure II.18: Architecture de Hadoop Yarn[15]

Dans l'architecture YARN (Figure II.18), il existe un maître global nommé ResourceManager (RM) et un ensemble d'esclaves appelés NodeManagers (NM), qui forment un système générique de gestion des applications de manière distribuée. Le RM est un ordonnanceur responsable du suivi et de l'arbitrage des ressources disponibles entre les applications tout en optimisant l'utilisation des ressources cluster. En revanche, le NM est responsable du

[15] Image disponible sur le lien : https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/yarn_architecture.gif

lancement des tâches et de la surveillance de l'utilisation des ressources par les nœuds workers. De plus, il existe un autre composant spécifique du Framework appelé ApplicationMaster. Il est chargé de négocier les ressources de la RM et des nœuds esclaves avec le NM pour exécuter et surveiller la progression des tâches. Par ailleurs, toutes les ressources de YARN sont demandées sous forme de conteneur (container), qui incorpore des éléments tels que la mémoire, le processeur, le disque, le réseau, etc. Ainsi, il organise les applications soumises par l'utilisateur en files d'attente pour garantir la gestion des demandes de ressources. Ces containers sont supervisés par les NodeManagers qui se trouvent au niveau de chaque nœud du cluster ce qui permet une amélioration significative d'utilisation du cluster en allouant dynamiquement des ressources en fonction des besoins requis par les applications pour pouvoir être exécutées.

II.2.8.7 Apache Flume

Flume est un service fiable et distribué. Il est délivré par la fondation Apache et destiné spécifiquement à la collecte, l'agrégation et le transfert des grandes quantités de données logs. Il possède une architecture robuste, flexible et tolérante aux pannes, et basée sur les flux de données en continu. Apache Flume fait partie des solutions open source big data de Hadoop.

Ci-dessous certains concepts de l'architecture Apache Flume (Figure II.19):

- Agent: représente le composant principal du service Flume utilisant la machine virtuelle java pour exécuter les tâches soumises par le client.
- Source: responsable de la collecte des données à partir de la source et de leur transmission sur le canal (Channel)
- Sink: collecte les données à partir du canal et les écrit sur le système de stockage comme HDFS et HBase .
- Channel: Connecte la source avec le Sink. Son fonctionnement ressemble à celui d'une fille d'attente.
- Event: il s'agit de la représentation des bases de données dans l'architecture Apache Flume. Il permet de fournit un accès aux données circulées par le système.

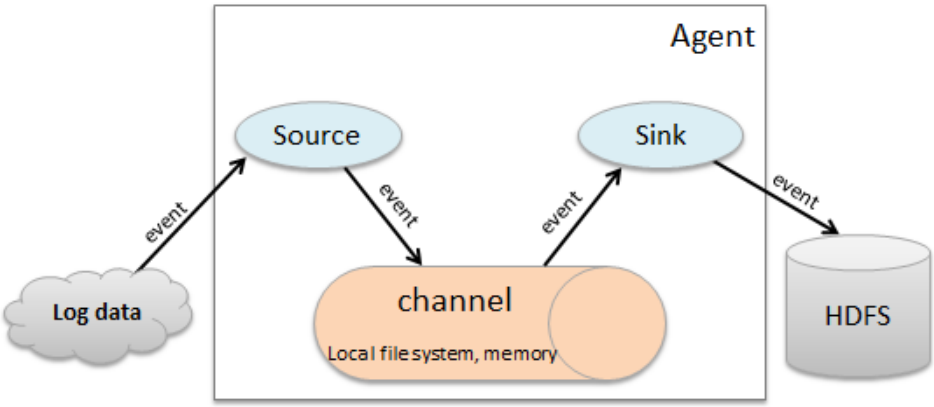


Figure II.19: Architecture du service Apache Flume

Flume possède une architecture très simple. En effet, Il se compose de trois composants: Source, Channel et Sink, tous hébergés par un agent Flume. La source du canal capture les unités des flux de données, appelées événements (Events), à partir d'une source de données externe (fichier journal, syslog) ou depuis des données d'un autre Sink. Ensuite, il stocke les événements reçus dans le canal (Channel) pour garantir la non-perte de ces données en cas de panne matérielle ou redémarrage de l'agent. Enfin, le Sink déplace les événements du canal en lot et les écrit sur un système de fichiers distribué Hadoop ou un autre système de stockage (HBase, S3, Cassandra, MongoDB).

II.2.8.8 Apache Sqoop

Le terme Sqoop [88] est le résult de concaténation de «Sq» de SQL et de «oop» du paradigme Hadoop. Il s'agit d'un outil permettant de transférer efficacement et rapidement des données volumineuses (bulk data) entre Apache Hadoop et des sources de données structurées telles que des bases de données relationnelles. Il est capable d'importer des données depuis un entrepôt de données structuré, y compris des systèmes de gestion des bases de données vers HDFS ou un autre système de stockage de données NoSQL telles que HBase et Cassandra. Il utilise MapReduce pour charger les données, ce qui permet un fonctionnement en parallèle sur un cluster et assure la tolérance aux pannes. Sqoop peut aussi être utilisé pour extraire des données du HDFS et de les exporter vers des bases de données relationnelles ou des entrepôts de données.

Sqoop fournit un ensemble de connecteurs supportant différents systèmes de gestion des bases de données relationnelles (MySQL, Oracle, PostgreSQL, SQL Server) qui permettent une connexion facile et une haute interopérabilité avec les systèmes externes. La Figure II.20 illustre l'architecture simplifiée du Sqoop.

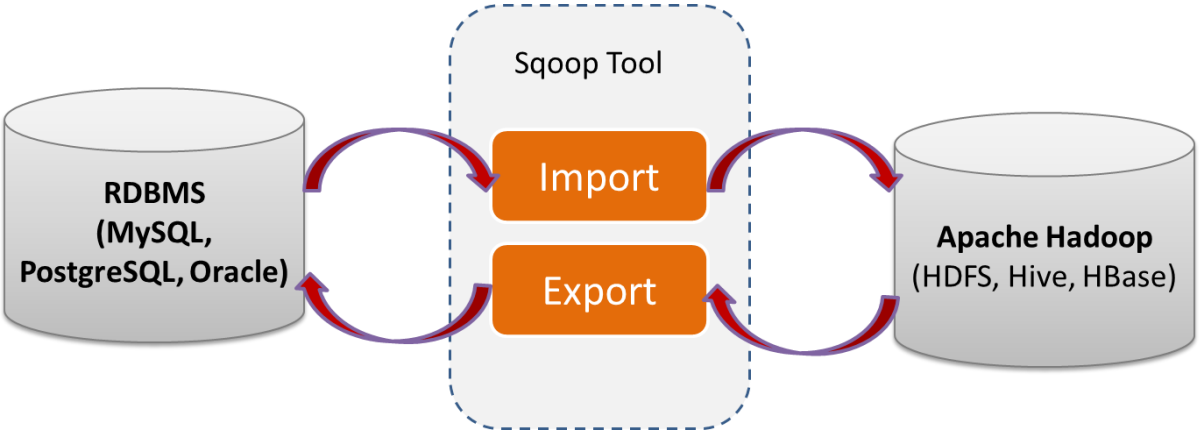


Figure II.20: Architecture d'Apache Sqoop

Pour importer les données à partir d'une base de données MySQL vers Sqoop, il suffit d'utiliser la commande suivante (paramètre --m représente le nombre de processus Map).

Script Sqoop
Sqoop import --connect jdbc:mysql://192.168.56.101:3306/estonline --username root --password **** --table mdl-usr --m 4

II.2.8.9 Zookeeper

Apache ZooKeeper est un service open source de la fondation Apache Software. Il s'agit d'un logiciel de gestion de configuration et de coordination pour les systèmes distribués. ZooKeeper fait partie des produits big data offerts par Hadoop. Il permet d'assurer la synchronisation distribuée et la haute disponibilité des données grâces à ses services redondants. Le stockage des données se fait selon un modèle hiérarchique respectant la structure arborescente. Un client Zookeeper peut communiquer facilement avec un nœud, et peut switcher vers un autre nœud si le premier ne répond pas. Ces nœuds sont appelés “Znodes” et chaque nœud est identifié par un chemin et un parent à l'exception de la racine. Les clients peuvent lire et écrire sur les nœuds et avoir, ainsi, un service de configuration

partagé. Cette solution a été utilisée par de nombreuses entreprises sur le web comme Yahoo[89], eBay[90] ainsi que des moteurs de recherches tel que Solr [91].

II.2.8.10 Oozie

Oozie est un système d'ordonnancement des flux de travail dédié aux technologies Hadoop. Il est développé comme une application web utilisant le conteneur des servlets Java et distribué sous la licence Apache 2. Il permet de gérer et de coordonner les jobs (travaux) écrits en différents langages tels que MapReduce, Pig et Hive. Oozie a la capacité de combiner plusieurs jobs dans une seule unité logique.

Il s'agit, en effet, d'un système de gestion des workflows qui relient les jobs. Les workflows dans Oozie sont regroupés en deux catégories:

- Oozie Workflow : sont des programmes constitués en séquence d'actions à exécuter, appelés DAGs (Directed Acyclical Graphs)
- Oozie Coordinator : sont des tâches récurrentes des flux de travail déclenchés par le temps et par la disponibilité des données.

Dans le cas d'une application créée avec MapReduce, Oozie prend en charge la planification et le lancement des actions, mais c'est MapReduce qui exécute les tâches.

II.2.8.11 Pig

Pig est une plate-forme d'analyse des grands ensembles de donnée créée par Yahoo. Il possède un langage de programmation de haut niveau appelé Pig Latin pour la création des programmes MapReduce sous l'environnement Hadoop. Ce langage fournit une syntaxe abstraite similaire à SQL et capable de transformer le code source Pig en tâches MapReduce. Pig donne la possibilité aux développeurs de programmer leurs applications MapReduce en utilisant des langages de programmation autres que Java, comme Python, Ruby et JavaScript, via les fonctions UDF (User Defined Functions) afin de faciliter la programmation parallèle dans Hadoop.

II.2.8.12 Hive

Apache Hive est un système open source d'entrepôt des données. Il permet de manipuler et d'analyser des données massives stockées dans le système des fichiers distribués (HDFS) de Hadoop ou un autre système compatible tels qu'AmazonS3. Il met à disposition son propre langage de type SQL, appelé HiveQL, pour interroger et analyser des grands volumes

d'informations. Hive prend en charge la traduction des requêtes SQL en tâches MapReduce s'exécutant sur un cluster Hadoop.

II.2.8.13 Ambari

Apache Hadoop fournit un environnement complet pour construire des applications du traitement des données à grande échelle. Cependant, la gestion des clusters des nœuds est souvent difficile et demande beaucoup du temps et d'effort. Pour faire face à ce problème, la fondation apache a développé Ambari pour faciliter le contrôle des tâches de configuration du cluster. Il s'agit d'une interface web permettant de simplifier la surveillance et d'optimiser les clusters Hadoop. En effet, Ambari fournit un assistant, étape par étape, pour l'installation des services Hadoop sur un nombre quelconque d'hôtes. En plus, il offre un tableau de bord affichant l'état d'un cluster [92]. Ambari est utilisé par des grandes entreprises telles que Hortonworks, eBay, IBM, Expedia et Samsung[93]. La Figure II.21 présente l'interface du logiciel Ambari qui décrit un ensemble de services (HDFS, Yarn, MapReduce, Tez, Hive, Sqoop, etc.), en plus elle affiche les informations concernant la capacité du disque, la mémoire utilisée, la bande passante consommée, et plein d'autres options :

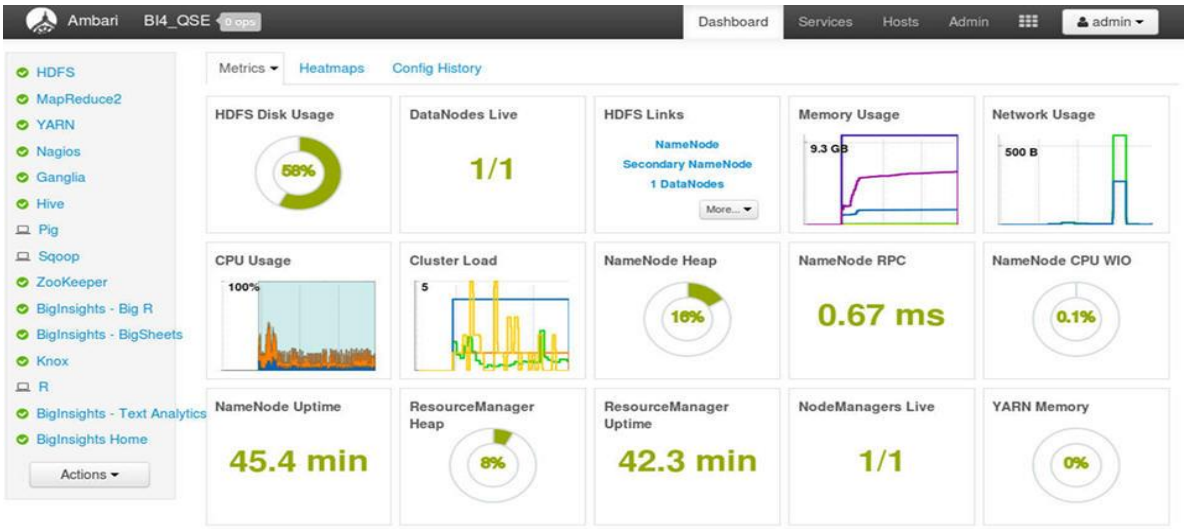


Figure II.21: Apache Ambari interface[16]

[16] Image disponible sur le lien : https://www.ibm.com/support/knowledgecenter/en/SSPT3X_4.2.0/com.ibm.swg.im.infosphere.biginsights.admin.doc/doc/admin_ambari_metrics.html

II.2.9 Les bases de données NoSQL

Pour répondre aux besoins accrus en termes du stockage massif et du traitement à grande échelle des jeux de données volumineuses, le mouvement NoSQL a fait son apparition autour des années 2010. NoSQL signifie «Not Only SQL» et non pas No SQL, c'est-à-dire, pas seulement SQL. Actuellement, les bases de données NoSQL sont considérées comme des bases de données de nouvelle génération, car elles sont généralement non relationnelles, distribuées, et open source. La Figure II.22 illustre les grandes phases historiques qui ont marqué le développement des systèmes de gestion de l'information.

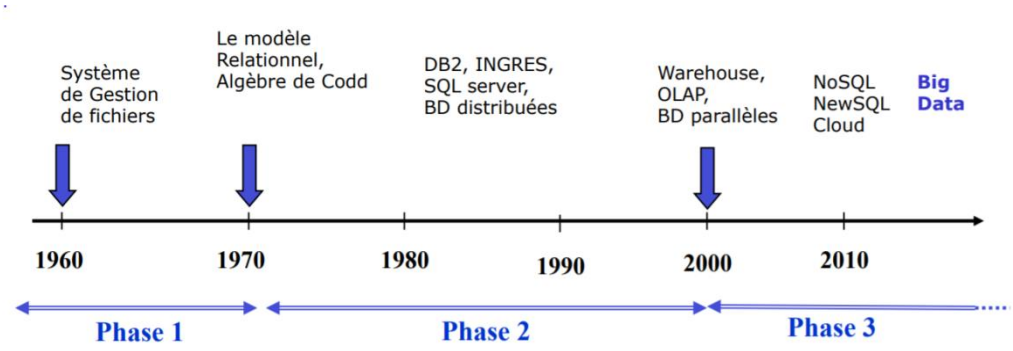


Figure II.22: phases historiques des systèmes de gestion des bases de données

NoSQL désigne une large catégorie des systèmes de gestion des bases de données conçu dans le but de contourner les limites des bases de données relationnelles. Les bases de données NoSQL n'ont pas vocation à s'opposer ou de s'éloigner du langage SQL, mais du modèle relationnel. Ce sont les implémentations courantes des bases de données relationnelles qui comportent des limitations que les moteurs NoSQL cherchent à dépasser. On devrait donc appeler ce mouvement «NoRelational » plutôt que NoSQL. Elles sont principalement caractérisées par [70]:

- **Schéma implicite :** Dans les moteurs relationnels, le schéma des données doit être prédéfini explicitement au niveau du serveur. Par contre, dans la quasi-totalité des moteurs NoSQL, le serveur des bases de données ne contrôle pas le schéma des données et c'est à l'application cliente de structurer correctement ce qu'elle veut manipuler dans la base des données. Mais il y a aujourd'hui un contre-exemple :

Cassandra, qui, à partir de sa version 2 sortie en 2014, impose un schéma fort, prédéfini au niveau du serveur.

- **Absence de relation :** Dans les systèmes de gestion des bases de données NoSQL, il n'y a pas de relations entre les tables. A l'inverse d'un moteur relationnel comme son nom l'indique.
- **Solution libre :** Beaucoup des bases de données sont des logiciels libres, et participent au mouvement du libre. Le logiciel libre a maintenant prouvé ses qualités. Une grande partie des logiciels trouvés sur internet tournent sur des environnements totalement open source (Unix). HBase et Cassandra sont des serveurs de bases de données totalement libres.

Le NoSQL regroupe de nombreuses bases de données récentes qui peuvent être classées en fonction du schéma ou de la structure des données qu'ils manipulent. Généralement, on distingue quatre modèles des bases de données NoSQL (Figure II.23).

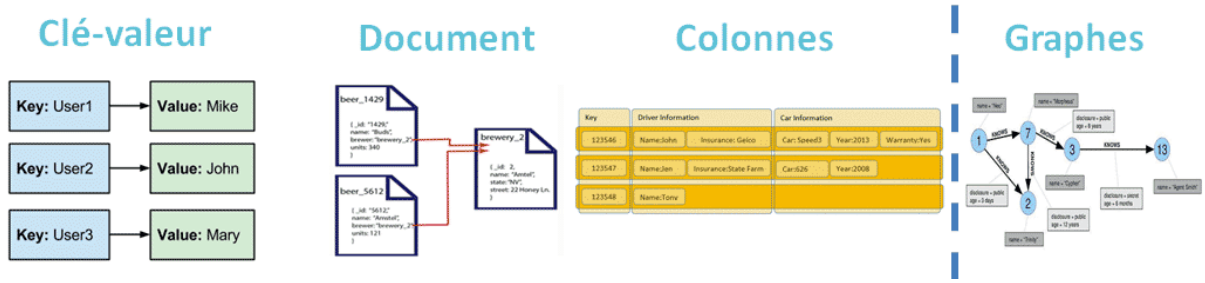


Figure II.23: Les modèles de bases de données NoSQL

II.2.9.1 Modèle clé-valeur

Comme tous les moteurs NoSQL, les bases de données clé-valeur se basent sur le principe de stockage d'une valeur associée à une clé unique. En effet, ce type des bases de données est assimilé à une collection de paires clé-valeur, dont l'accès se fait exclusivement par la clé.

Comme dans le cas de la recherche d'un livre à travers son numéro international normalisé (ISBN). Dans cet exemple, le numéro ISBN est la clé et toutes les autres informations concernant le livre constituent la valeur. En fait, la clé est obligatoirement connue et peut donc être recherchée. Par contre, la valeur est un ensemble d'informations sans signification qui doivent être interprétées.

Les bases de données clé-valeur son principalement faites pour le stockage temporaire et ne permettent que 4 opérations :

- La création : créer un nouveau couple (clé, valeur). La valeur peut être n'importe quel objet.
- La lecture : lire un objet en connaissant sa clé.
- La modification : mettre à jour l'objet associé à une clé.
- La suppression : supprimer un objet connaissant sa clé.

Elles sont beaucoup utilisées en tant que cache, pour conserver les sessions d'un site web et plus généralement pour toutes les données que l'on souhaite conserver pendant un certain laps de temps. L'exemple le plus connu est la gestion du panier d'achat (Amazon).

II.2.9.2 Modèle orienté documents

Les moteurs orientés documents sont une évolution de la base des données clé-valeur dans laquelle chaque clé associe un document dont la structure reste libre. Le format de sérialisation et d'échange des données le plus populaire est JSON (JavaScript Object Notation). Il permet d'exprimer un document structuré qui comporte des données simples, mais aussi des données composées comme des listes, sous une forme hiérarchique. Il est idéal pour représenter des données structurées ou semi-structurées. Les exemples des bases de données orientées documents les plus connues sont, MongoDB et CouchDB.

Dans l'approche NoSQL orientée documents, le schéma des données n'est pas établi à l'avance, chaque document peut avoir sa propre structure, on parle de « dynamic schema ». En plus, l'enregistrement des valeurs dans un document n'est soumis à aucun contrôle du système. Cependant, il est préférable d'avoir une structure minimale commune afin de faciliter la manipulation des données [94].

II.2.9.3 Modèle orienté colonnes

Ce modèle est inspiré par Google Big Table. Plusieurs moteurs NoSQL implémentent une structure proche de la table. Contrairement aux moteurs orientés documents, les données sont représentées en ligne séparées par des colonnes. Chaque ligne est identifiée uniquement par une clé, ce qu'on appelle dans le modèle relationnel la clé primaire. Plus utilisés pour les volumétries importantes, Cassandra et HBase appartiennent à cette catégorie des bases de données.

Dans le modèle relationnel, la structure des données est déterminée en avance par le schéma des relations et le nombre des colonnes est limité et similaire pour tous les lignes. Alors que le modèle orienté colonnes se caractérise par son schéma flexible qui peut varier d'un enregistrement à un autre. Cette flexibilité des bases de données orientées colonnes permet de gérer l'absence de certaines colonnes entre les différentes lignes de la table. Une base de données orientée colonnes est un ensemble de tables qui sont définies ligne par ligne, mais dont le stockage est organisé verticalement par des familles de colonnes. Une famille des colonnes peut contenir un très grand nombre de colonnes. Chaque famille des colonnes a une clé-valeur permettant l'accès et la manipulation des données[94].

II.2.9.4 Modèle orienté graphes

Les bases de données orientées Graphe est un type des bases de données se basant sur la théorie des graphes et utilisant des nœuds et arrêtes pour stocker et manipuler les données. Un nœud est représenté par un identifiant unique et un ensemble de propriétés sous forme clé-valeur tandis que l'arrête représente une relation entre un couple de données. Ce type des bases de données est beaucoup utilisé pour l'exploitation des données issues des réseaux sociaux comme Facebook et Twitter. En effet, Twitter a créé son propre base de données orientée graphe, appelée FlockDB [95] pour gérer des larges réseaux Graph.

Il existe 3 familles de systèmes de gestion des bases de données orientées graphe :

- Les graphes basés en mémoire sont les plus rapides dans la vitesse du traitement des données mais sont limités par la taille de mémoire disponible et sont volatiles.
- Les graphes basés sur disque sont moins performants. Ils sont limités par la taille des volumes du disque local ou du serveur distant.
- Les graphes en cluster gèrent les plus grosses quantités de données et ils se basent sur une infrastructure distribuée de cluster des machines.

La Figure II.24 donne une idée claire sur l'écosystème technologique du big data. Ces technologies peuvent servir à plusieurs cas d'utilisation, y compris l'intégration des données (Apache Flume, Apache Sqoop), le stockage (HDFS, NoSQL database), le traitement et analyse (MapReduce, Spark), et la représentation (outils de visualisation, Applications BI).

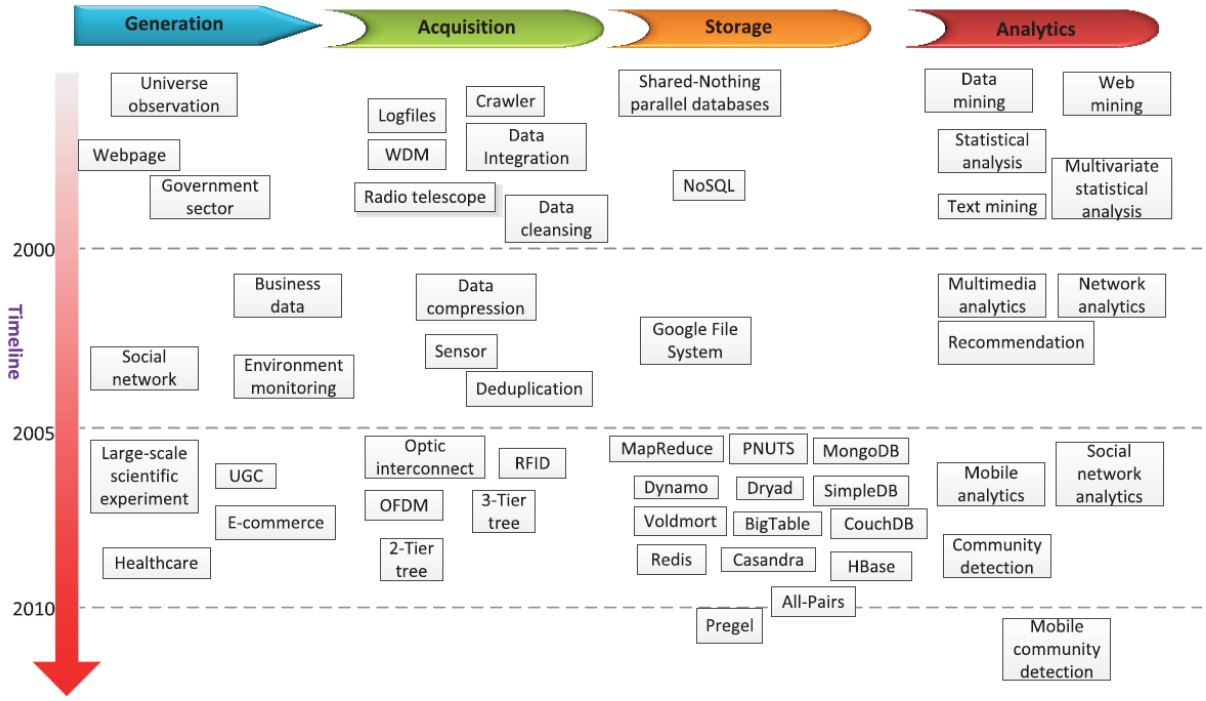


Figure II.24: Technologies big data

II.2.10 Méthodes et techniques du machine learning

Cette section est consacrée aux techniques d'analyse des données massives en présentant les différentes méthodes de l'apprentissage machine. Nous commençons par définir le concept, ensuite nous présentons ses méthodes et ses algorithmes. Plus spécifiquement, nous abordons les méthodes d'apprentissage ayant une implémentation fournit par le Framework Apache Spark qui vise à distribuer l'apprentissage automatique vers une infrastructure décentralisée de cluster.

II.2.10.1 Définition

Le Machine Learning, ou encore l'apprentissage automatique/machine en français, est un champ de recherche de l'intelligence artificielle. Il est dédié à la conception, l'analyse, et l'implémentation des algorithmes permettant à une machine d'évoluer automatiquement sans intervention d'un être humain. Ce concept a plusieurs définitions [96]. Arthur Samuel [97], un des pionniers de l'intelligence artificielle, a défini le machine learning comme le champ d'étude visant à donner la capacité à une machine d'apprendre sans être explicitement programmée.

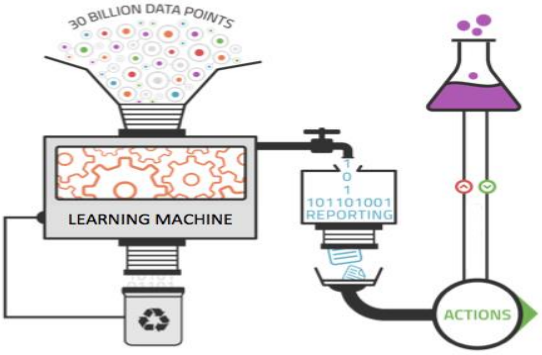


Figure II.25: Apprentissage automatique

Il s'agit d'une science permettant de découvrir des patterns et des relations entre des variables et d'effectuer des prédictions à partir des données en se basant sur des méthodes mathématiques et statistiques. Le machine learning est très efficace quand le volume des données à analyser et très grand parce qu'il permet de trouver des corrélations et des relations entre les données afin de prendre des bonnes décisions. Un exemple simple de l'utilisation du machine learning et le filtre anti spam dans une boite de messagerie. Dans ce cas, la machine va toute seule analyser les emails entrants et les classer grâce aux algorithmes d'apprentissage automatique.

II.2.10.2 Algorithmes supervisés et non supervisés

Les méthodes d'apprentissage automatique sont destinées aux divers cas d'utilisation. On distingue essentiellement deux grandes catégories:

Les méthodes supervisées et non supervisées. Une méthode supervisée (supervised learning) permet d'extraire le résultat à partir d'un ensemble de données contenant des couples entrée-sortie. Le but de l'apprentissage supervisé est que l'algorithme puisse «apprendre» en comparant la sortie réelle avec les sorties «enseignées» pour trouver des erreurs et modifier le modèle en conséquence. Ces couples sont déjà connus pour guider l'algorithme dans sa voie d'apprentissage en lui fournissant des exemples ou des indications par un expert. Ces indications peuvent être sous forme de réponses de type Oui/Non ou Accepté/refusé, etc. Les algorithmes de l'apprentissage supervisé cherchent à trouver une liaison entre les entrées et les sorties par l'intermédiaire d'une fonction de prédiction.

D'autre part, dans les méthodes non supervisées, il n'existe pas la notion d'entrée-sortie. En effet, le système ne dispose que d'exemples, et de résultats qui ne sont pas prédéterminés, et l'algorithme doit découvrir par lui-même la structure en fonction des données. Dans ce cas, l'algorithme cherche à classer les données en groupes, chaque groupe doit contenir des données similaires. Pour les données différentes, elles doivent être organisées dans des groupes distincts. L'apprentissage ne se fait pas à partir des indications préalablement fournies par un expert, mais à partir des corrélations trouvées entres les données [98].

II.2.10.3 Techniques et méthodes

II.2.10.3.1 Clustering

Le clustering, partitionnement des données en français, est une méthode de classification non supervisée. Il s'agit d'une méthode d'analyse des données qui vise à partitionner les données en différents groupes homogènes. En effet, les éléments de chaque groupe partagent des caractéristiques communes en s'appuyant sur des critères de similitude.

Contrairement à la classification, il n'y a pas de connaissances a priori sur les classes dans lesquelles nous regroupons les éléments. La séparation des instances dans les différents groupes s'effectue en se basant sur le calcul de similarité entre les éléments [99]. En général, il y a plusieurs méthodes de clustering, comme le clustering par partitionnement, le clustering basé sur la densité, etc.

Le premier type de clustering est celui par partitionnement. K-means [100] , k-moyennes en français, est la méthode de partitionnement la plus populaire. Elle est notamment utilisée pour diviser des observations en k partitions, appelées clusters. Cependant, cette méthode souffre de quelques lacunes, principalement, elle ne donne pas le résultat optimum. Ainsi, la performance de K-means et son efficacité, comme la méthode de détection des anomalies, dépendent de la sélection aléatoire du nombre des groupes initiaux. Par conséquence, un "mauvais choix" de ce nombre influencera négativement sur la détection de vraies anomalies et provoquera la génération de faux résultats.

En plus, la méthode de clustering basé sur la densité permet de résoudre le problème de la dépendance du choix aléatoire du nombre des groupes initiaux. Les méthodes basées sur la densité cherchent à découvrir des clusters pour lesquels tous les objets sont en relation avec un même objet central. Dans ce sens, l'algorithme DBSCAN (density-based spatial clustering

of applications with noise), a été proposé en 1996 par Martin Ester [101]. Il s'agit d'un algorithme fondé sur la densité dans la mesure qui s'appuie sur la densité estimée des clusters pour effectuer le partitionnement des groupes en régions denses. Néanmoins, la méthode basée sur la densité a une grande influence sur la capacité de détection. Elle engendre des partitions sous-optimales et considère beaucoup d'instances comme des bruits.

II.2.10.3.2 Classification

La classification est l'un des problèmes du domaine de l'apprentissage machine. Les champs d'application des algorithmes de classification sont très variés, y compris le traitement d'image, la reconnaissance d'objets, etc. La classification est une méthode d'apprentissage supervisé. La notion de supervisé signifie que les exemples sont étiquetés (la classe est connue), c-à-d. étiqueter chaque donnée en l'associant à une classe. L'objectif de la classification est essentiellement de définir des règles permettant de grouper des éléments ou objets dans des classes à partir des données d'apprentissage (variables qualitatives ou quantitatives). Il existe de nombreuses méthodes de classification. Dans ce rapport, nous allons nous concentrer sur les algorithmes les plus utilisés.

II.2.10.3.3 Naïve Bayes

Classificateurs bayésiens naïfs est une famille de classificateurs de type probabiliste fondée sur le théorème de Bayes (1763) [102]. Il s'agit d'une technique de l'apprentissage automatique supervisé. Cette méthode de classification permet, dans un premier temps, de construire un modèle d'apprentissage basé sur un jeu de données d'entraînement. Pendant la phase d'apprentissage, l'algorithme élabore ses règles de classification, pour les appliquer dans un second temps à la classification d'un jeu de données de test afin de générer finalement des prédictions. La caractéristique principale du classificateur Bayésien est qu'il émet une hypothèse forte a priori inadaptée aux cas pratiques. En d'autres termes, un classificateur bayésien naïf suppose que l'existence d'une caractéristique déterminant l'appartenance à une classe est indépendante de l'existence d'autres caractéristiques. Par exemple, un animal sera considéré avec une grande probabilité comme un canard s'il a des ailes. Naïve bayes est utilisé dans de nombreux domaines tels que le diagnostic des maladies et le filtrage des e-mail (spam).

Le théorème de Bayes est formulé mathématiquement par l'équation suivante (1) [102]:

$$P(C \mid A) = \frac{P(A \cap C)}{P(A)} = \frac{P(A \mid C)P(C)}{P(A)} \quad (1)$$



Le principe du théorème de Bayes est basé sur la probabilité conditionnelle : Si A et B sont deux évènements, la probabilité que A se produise sachant B réalisé est égale à la probabilité que les deux évènements surviennent au cours de la même épreuve.

II.2.10.3.4 Decision Tree

Decision Tree [103], en français, l'arbre de décision est une méthode très populaire dans le domaine de l'apprentissage machine ou data mining. Elle est utilisée pour résoudre quelques problèmes de classification et de régression. Généralement, Decision Tree est une technique d'aide à la décision en représentant les résultats possibles d'un ensemble de choix sous forme graphique d'un arbre aisément interprétable. Les différentes décisions possibles sont situées aux extrémités des branches, et sont atteints en fonction des décisions prises à chaque étape [104].

Les arbres de décision sont exploités dans plusieurs domaines, y compris l'informatique décisionnelle, ou l'exploration des données et le diagnostic médical en permettant aux médecins d'interpréter les raisons de la maladie (Figure II.26).

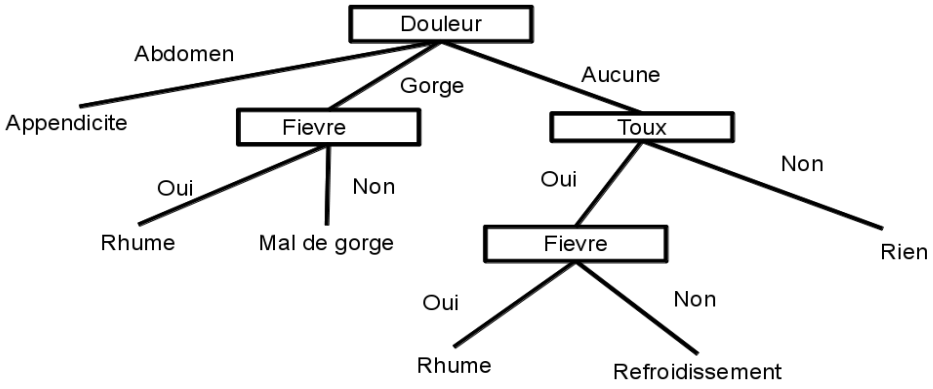


Figure II.26: Arbre de décision[17]

[17] Image disponible sur : http://www.up2.fr/M1/td/TD10_2.html

Il y a deux types d'arbres: Arbres de classification (Classification Trees) et arbres de régression (Regression Trees).

i. Arbres de classification: sont utilisés pour segmenter les observations en groupes plus homogènes en prédisant l'appartenance d'éléments (observations, individus) à une classe (attribuer des étiquettes). Ils s'appliquent généralement aux résultats de nature binaire ou catégorique.
ii. Arbres de régression: Leur but est de créer un modèle qui prédit la valeur d'une variable-cible depuis la valeur de plusieurs variables d'entrée. Cette technique peut être appliquée à des résultats continus (les dépenses ou le revenu personnel).

Les valeurs d'entrée peuvent être continues ou discrètes. Les modèles d'arbres de décision génèrent un arbre décrivant le processus de décision sous forme d'une série d'instructions logiques "si-alors". Les feuilles de l'arbre renvoient les étiquettes de classe et, dans certains cas, ils renvoient également les scores de probabilité.

II.2.10.3.5 Régression

La régression fait partie des méthodes du machine learning les plus connues. Elle se base sur un ensemble de méthodes statistiques pour analyser la relation d'une variable par rapport à une ou plusieurs autres variables. La technique de régression permet de trouver un modèle mathématique en fonction des données d'entrainement.

Les algorithmes de régression peuvent prendre plusieurs formes en fonction du problème abordé et du type des données d'entrée. En effet, il existe plusieurs types de régressions, y compris la régression linéaire, logistique, simple, multiple, etc. La régression linéaire est le modèle le plus simple. Elle consiste à trouver la meilleure droite qui s'approche le plus des données d'apprentissage (Figure II.27). La fonction de prédiction sera donc une droite [105].

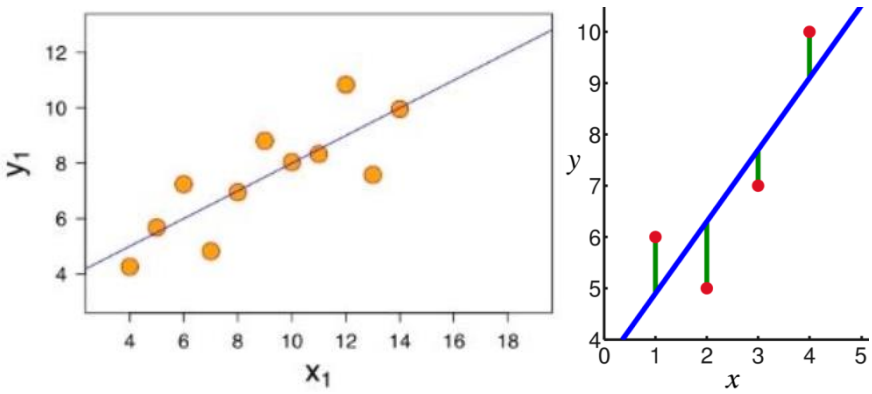


Figure II.27: Régression Linéaire

Le modèle mathématique de la régression linéaire est défini par l'équation suivante (2) [106]:

$$Y = f(X) = \alpha * X_1 + \beta \quad (2)$$

Comme pour la régression linéaire, le but de la régression logistique (logistic regression) est de caractériser les relations entre une variable dépendante et une seule (régression simple) ou plusieurs variables prises en compte simultanément (régression multiple). Il s'agit donc d'un modèle permettant de relier la variable dépendante (Y) à des variables explicatives (X_1, X_2, X_3, … X_n).

Toutefois, le modèle de la régression linéaire s'applique uniquement dans le cas où la variable Y est quantitative (Surface, Poids) et ne permet pas de mesurer la dépendance entre les variables qualitatives notamment binaires où Y s'exprime en termes de oui/non. Il est donc nécessaire d'utiliser un modèle adapté permettant de relier les variables explicatives à la variable qualitative (Y) à prédire. En effet, le modèle de régression logistique peut s'appliquer lorsque la variable à expliquer (Y) est qualitative (vrai/faux, accepter/refuser) [107].

La formule mathématique de la régression linéaire est montrée dans l'équation (3):

$$ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_{p-1} \quad (3)$$

Dans le cas d'une régression simple, la fonction logistique a une forme de sigmoïde (Figure II.28)

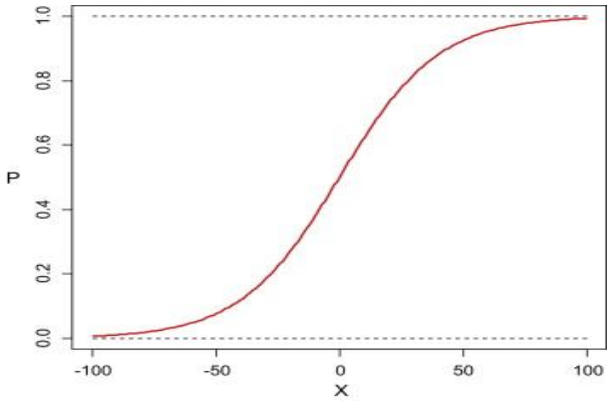


Figure II.28: Régression Logistique

La Figure II.28 montre la probabilité de réalisation d'un évènement (p) en fonction des valeurs prises par une variable (X) variant de (–100 à 100) selon la fonction logistique. La courbe représentant p en fonction de X a une forme sigmoïde et varie de façon croissante entre 0 et 1.

II.2.10.3.6 Règles d'association (Association Rules)

Association rules est une méthode d'apprentissage automatique non supervisé qui est largement utilisée dans de nombreux domaines, notamment les moteurs de recommandation, l'analyse de la transaction en ligne et le flux de clics sur les pages Web. Son but est de découvrir des relations intéressantes entre des variables dans de grands ensembles de données.

La technique de règles d'association a été évoquée par Rakesh Agrawal et son équipe en 1993 [108]. Elle était utilisée, la première fois, pour découvrir des similitudes entre des produits achetés par des clients en déterminant les éléments qui apparaissent fréquemment ensemble dans une base des données des transactions de vente d'un supermarché. L'application la plus connue de cette technique et celle de l'analyse du panier de vente «Market Basket Analysis». Elle consiste à analyser les achats où chaque transaction est constituée d'une liste d'articles qui sont analysés dont le but de déterminer l'existence des règles intéressantes de types X=>Y. En effet, Les règles d'association sont des relations d'implication conditionnelles, entre les attributs de la base des données, de la forme: "Si un client achète le produit A (lait), alors il est possible qu'il va acheter le produit B (pain) aussi".

Association rules utilise de nombreux critères pour mesurer la qualité des règles d'association générées lors de l'analyse des données. Cette méthode se base principalement sur deux mesures : le support et la confidence (4):

$$Support = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions} = P(A \cap B) \quad (4)$$

$$Confidence = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P(A \cap B)}{P(A)}$$

Le support correspond à la probabilité que les éléments A et B existent simultanément dans la base des données des transactions, et la confidence représente la probabilité conditionnelle d'avoir B parmi les transactions contenant l'élément A. Par ailleurs, le support est utilisé pour extraire la liste des éléments fréquents, et la confidence mesure la force d'une association ou la relation entre des variables.

On trouve plusieurs algorithmes qui implémentent cette méthode d'apprentissage, parmi lesquels, Apriori, Eclat, FP-growth.

1) Apriori [109]: se base sur la génération des ensembles d'éléments (itemset) candidats et se divise en deux étapes. En premier lieu, il s'agit de trouver les éléments fréquents (frequent itemsets) qui doivent apparaitre un nombre de fois égal ou supérieure au support minimum prédéfini (étape 1). Ensuite, il génère les règles d'association les plus intéressantes et utiles en s'appuyant sur le résultat de l'étape 1. Ces règles doivent satisfaire la confidence minimum spécifiée par l'utilisateur.
2) Eclat [110]: est un algorithme de recherche d'associations, considéré comme une alternative d'Apriori, en offrant plusieurs améliorations. L'idée principale d'Eclat est d'utiliser un codage dit "vertical" du jeu des données utilisant les intersections d'ensembles.
3) FP-growth (frequent pattern growth) [111]: possède une structure d'arbre (FP-tree) pour manipuler et stocker une forme compressée d'une base des données. il adopte une stratégie de découpage pour faciliter les tâches d'exploration des données massives. FP-growth utilise la méthode "pattern fragment growth" pour trouver les relations entre les objets de la base des données. Cet algorithme est considéré le plus performant et le plus rapide, par rapport aux autres algorithmes implémentant la technique de

règles d'association, en évitant le processus de génération et de test d'éléments candidats de l'algorithme Apriori.

II.2.10.3.7 Time Series Analysis

Analyse des séries temporelles ou chronologiques (Time Series Analysis) [112], est une suite de valeurs numériques représentant l'évolution d'une quantité de données au cours du temps. De telles suites de variables aléatoires peuvent être exprimées mathématiquement afin d'en analyser le comportement pour comprendre son évolution et pour en prévoir son comportement futur. Son modèle mathématique s'appuie le plus souvent sur des concepts de probabilité et de statistique [113]. Cette méthode est exploitée dans plusieurs contextes, y compris le domaine de la statistique, la finance, et la géophysique, dont le but principal est la prédiction des événements futurs. Dans le contexte du traitement du signal et de la communication, elle est utilisée pour la détection du signal. Tandis que dans le contexte de data mining et machine learning, cette technique est utilisée pour le groupement (clustering), la classification, ainsi que la prévision.

Les principaux objectifs de Time Series Analysis sont:

1. Comprendre la structure sous-jacente d'une série chronologique.
2. Ajuster un modèle mathématique et ensuite procéder à la prédiction du futur.

Les périodes du temps sont généralement espacées et les observations peuvent être soit univariées, soit multivariées. Les séries temporelles univariées sont celles où une seule variable est mesurée dans le temps, alors que les séries temporelles multivariées sont celles où plusieurs variables sont mesurées simultanément.

L'analyse des séries chronologiques est utile pour savoir des informations pertinentes sur les prochains évènements. Il y a quatre types principaux:

1) Trend: un mouvement à long terme concernant les séries chronologiques pouvant être à la hausse, à la baisse ou stationnaires;
2) Cyclical: un modèle généralement observé sur deux ans ou plus et qui est provoqué par des circonstances qui se répètent périodiquement;
3) Seasonal: variations dans l'année qui dépendent généralement de la météo, des habitudes des clients, etc.

4) Irregular : événements aléatoires ayant une influence imprévisible sur la série chronologique

Le diagramme (Figure II.29) montre un exemple de Time Series Analysis qui vise à prévoir les ventes futures en se basant sur l'analyse de l'historique des activités de vente.

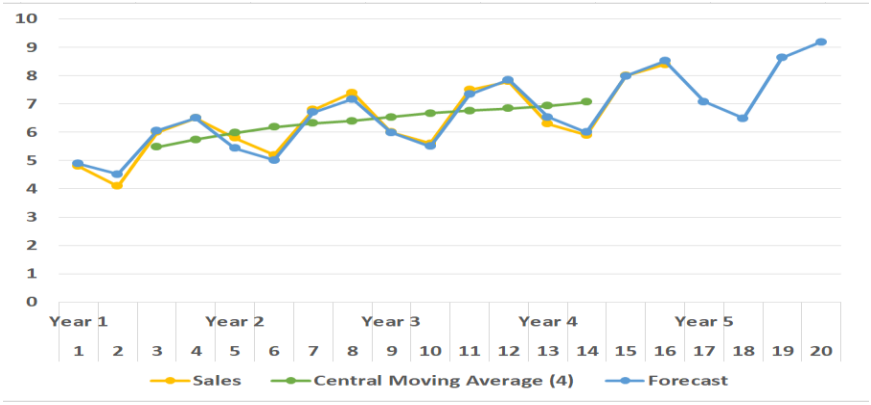


Figure II.29: Modèle de Time Serie Analysis[18]

II.2.10.3.8 Text analysis

Text Analysis ou text mining, en français traduit par l'analyse de texte, est un ensemble de techniques d'apprentissage et d'intelligence artificielle dédiées au domaine de la linguistique et de la sémantique. Il permet de former des informations à partir d'un document textuel en utilisant des algorithmes informatiques avancés pour faire face au volume massif provenant d'une multitude de contenus disponibles en ligne [114].

L'exploration du texte implique généralement le processus de structuration du texte d'entrée en effectuant plusieurs opérations: analyse, ajout de certaines caractéristiques et suppression d'autres, puis l'insertion ultérieure dans une base de données, et enfin l'évaluation et l'interprétation. Les tâches classiques d'exploration du texte comprennent la catégorisation, la classification, l'extraction des concepts, l'analyse des sentiments, la synthèse des documents, etc. Il s'agit d'une discipline très utilisée dans plusieurs domaines tels que le marketing, la science politique, et les réseaux sociaux [115].

[18] L'exemple et les données de ventes sont accessibles sur le lien : http://analyticsmodels.blogspot.com/2017/08/70-time-series-analysis.html

II.2.10.3.9 Deep learning

Le deep learning ou apprentissage profond est un sous-domaine de l'apprentissage automatique qui vise à utiliser les méthodes de l'intelligence artificielle pour imiter le fonctionnement du cerveau humain. Il s'appuie sur les techniques des réseaux des neurones artificiels afin de reconstruire des connaissances à partir des grandes quantités de données. Ces réseaux sont composés des centaines des couches de neurones dont chacune recevant et analysant les données de la couche précédente. En effet, un réseau neuronal est une sorte de logiciel cérébral ou de machine virtuelle composée de milliers d'unités (des neurones) qui effectuent des calculs. Plus précisément, des unités logiques relient les données d'entrées et de sorties entre elles par l'intermédiaire d'un réseau complexe fonctionnant comme un cerveau et qui est capable de prendre des décisions en se basant sur des données d'entrainement [116].

Comme l'illustre la Figure II.30, par exemple, via un processus d'autoapprentissage, la machine est capable d'identifier un chat sur une photo. Dans ce cas, chaque couche du réseau des neurones correspond à un aspect particulier de l'image [117].

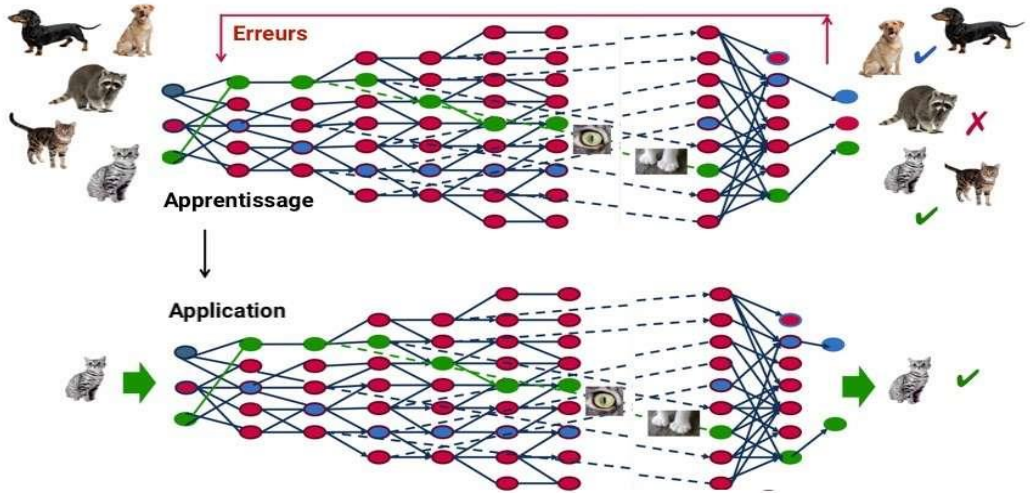


Figure II.30: Exemple du Deep Learning

Le deep learning est utilisé dans plusieurs domaines: la reconnaissance de l'image et de la parole, les voitures autonomes, le diagnostic médical, la détection des fraudes et les robots intelligents. Également, les géants du web, tel que Google, ont commencé à exploiter ses techniques afin d'améliorer la qualité de leurs services.

II.2.11 Conclusion

Dans ce chapitre, nous avons présenté le paradigme big data en commençant par définir ce concept et ces caractéristiques. Ensuite, Nous avons fourni une comparaison entre les systèmes classiques de gestion des bases de données relationnelles et ceux du big data. Ce chapitre a décrit également l'ensemble des Frameworks open source les plus utilisés par les entreprises pour construire des applications supportant le stockage et le traitement des volumétries massives des données. Par la suite, Nous avons introduit les solutions du mouvement NoSQL conçues pour contourner les limites des SGBDR en fournissant des bases de données distribuées et évolutives. A la fin du chapitre, Nous avons présenté les méthodes les plus populaires de l'apprentissage automatique utilisées pour extraire les informations pertinentes à partir des grosses quantités de données. Dans le chapitre suivant, nous allons présenter nos contributions qui consistent à intégrer le cloud computing et la nouvelle génération des technologies big data dans le contexte des environnements informatiques pour l'apprentissage humain dédié à l'e-learning.

III. CHAPITRE III : Intégration du cloud et du big data dans un EIAH

Approche d'intégration de l'e-learning dans l'environnement cloud

- Modèle de l'e-learning traditionnel
- Modèle de l'e-learning basé sur le cloud
- Mode interactif de l'e-learning en mode SaaS
- Comparaison entre le modèle traditionnel de déploiement et celui du cloud
- Avantages et bénéfices du modèle e-learning basé sur le cloud
- Valeur ajoutée du cloud pour les plates-formes d'apprentissage en ligne
- Contraintes et limites du cloud

Modèle d'e-learning basé sur le cloud et le big data

- Modèle e-learning basé sur le cloud et le big data
- Méthodologie d'intégration des technologies big data dans l'e-learning
- Architecture distribuée du big data pour les systèmes e-learning
- Infrastructure répartie du cluster Hadoop
- Bénéfices et avantages de l'intégration du big data dans les EIAH

III

III.1 Approche d'intégration de l'e-learning dans l'environnement cloud

III.1.1 Introduction

De nos jours, les systèmes d'apprentissage en ligne ont connu une grande évolution. En effet, les technologies de l'information et de la communication ont changé la manière d'enseigner et d'apprendre. Cependant, la plupart des formes éducatives sont devenues inadaptées aux exigences du progrès technologique et au développement de l'éducation. Pour cela, les systèmes e-learning doivent être en mesure de rattraper les changements d'apprentissage, qui ne cessent d'augmenter, en particulier avec l'énorme croissance du nombre d'apprenants, des contenus éducatifs et des ressources mises à disposition. Par conséquent, les institutions qui offrent l'apprentissage en ligne devraient s'adapter aux attentes et besoins d'apprenants. En outre, les dimensions des systèmes e-learning se croissent d'un rythme exceptionnel et les défis en ce qui concerne l'acquisition des ressources et d'optimisation de leur usage, aussi le contrôle du budget mettent en évidence la nécessité d'utiliser une plate-forme répondant aux exigences évolutives et au contrôle des coûts.

Cependant, dans le mode traditionnel de l'enseignement à distance basé sur les serveurs web calssiques, les plates-formes d'apprentissage en ligne sont situées à l'intérieure des établissements ou des universités ce qui provoque beaucoup de problèmes tels que le taux d'investissement élevé ou le coût de déploiement et de maintenance des systèmes e-learning mais sans gains en rendement et sans potentiel de développement. Dans ce contexte, le cloud représente un formidable accélérateur de transformation digitale dans les universités et les établissements d'enseignement. Alors, nous proposons d'incorporer et d'exploiter les services de ce dernier qui s'avère un modèle prometteur en raison de son évolutivité permanente, de sa disponibilité, de sa gestion efficace et de son utilisation économique des ressources.

Dans ce qui suit nous proposons une approche visant à améliorer la qualité des services des systèmes d'apprentissage en ligne en se basant sur l'environnement du cloud. Ensuite, Nous discutons les avantages et les bénéfices attendus de l'intégration de l'e-learning et du cloud.

III.1.2 Modèle de l'e-learning traditionnel

Les environnements informatiques pour l'apprentissage humain ont pour but de favoriser l'apprentissage chez l'apprenant en lui permettant de suivre son parcours de formation en ligne

et d'apprendre à son propre rythme et sans contrainte de temps ni de lieu de formation. Dans ce sens, l'e-learning peut être défini comme un processus basé sur internet qui utilise les technologies web pour concevoir, mettre en œuvre et gérer l'apprentissage. Il ne remplace pas l'enseignement présentiel, mais il améliore grandement l'efficacité et la qualité de l'éducation. Comme l'e-learning a beaucoup d'avantages tels que la flexibilité, la diversité, l'ouverture et l'autonomie, il devient un moyen incontournable pour acquérir des nouvelles connaissances et maîtriser des compétences dans différents champs disciplinaires.

L'architecture simplifiée du système d'apprentissage en ligne est montrée dans la Figure III.1.

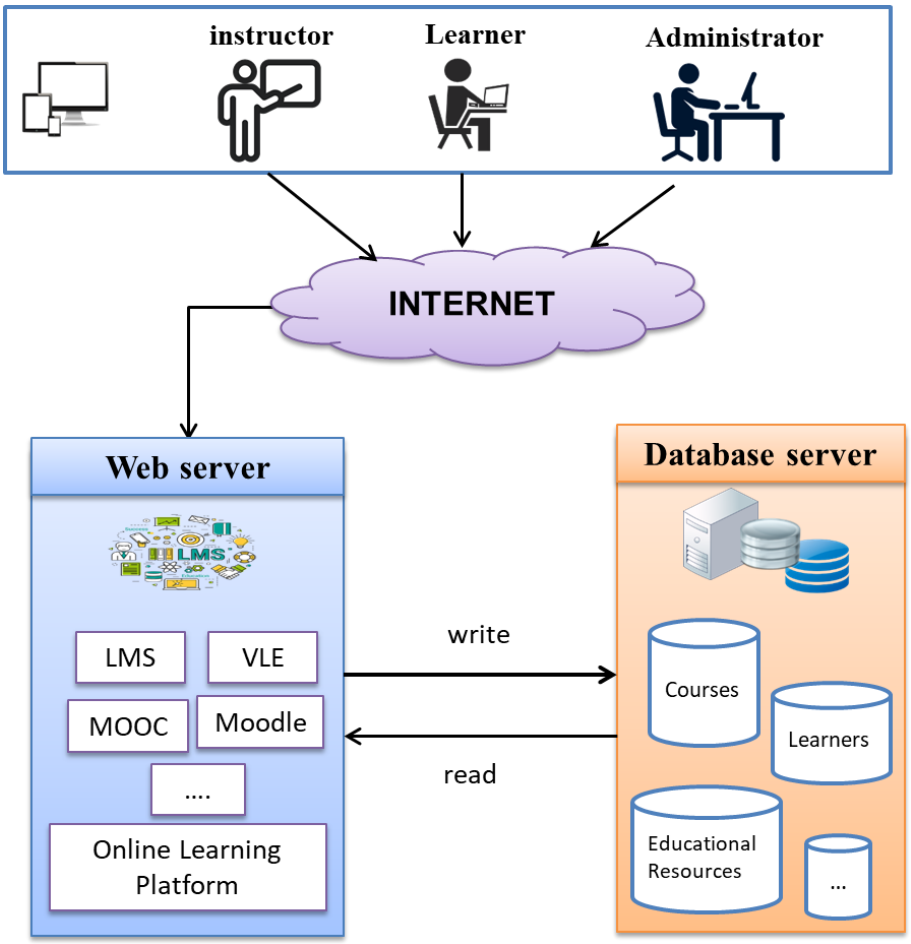


Figure III.1 : Architecture traditionnelle de l'e-learning

Cette architecture montre les différents rôles dans le processus d'apprentissage en ligne. D'ailleurs, les systèmes d'apprentissage en ligne traditionnels se composent de trois éléments. Premièrement, un apprenant qui se connecte au système e-learning pour apprendre. L'enseignant et l'équipe pédagogique préparent les cours et les mettent en ligne sur le serveur,

et parfois ils interviennent pour faire des modifications ou des améliorations au niveau d'un module d'un cours donné. Deuxièmement, il y a la plateforme d'apprentissage en ligne déployée sur le serveur Web. Cette plate-forme peut être développée à travers différents systèmes de gestion de d'apprentissage (LMS). Parmi les LMS les plus connus, il y a BlackBoard[118], Edx[119], Coursera[120], et Moodle[121], etc. Ce dernier est largement utilisé dans les universités marocaines, y compris notre université Sidi Mohamed Ben Abdellah. Enfin, il s'agit d'un système de gestion des bases de données relationnelle (Mysql, PostgreSQL, Oracle, etc.) qui permet le stockage des informations sur des profils, des activités, des évaluations, des ressources pédagogiques, des interactions d'apprenants, etc.

Cette architecture est principalement conçue pour faire la séparation entre les différentes couches de système d'apprentissage. La couche de présentation correspond à l'affichage via IHM, par exemple une interface web, bureau ou mobile. Puis la couche de traitement permettant de contrôler et de gérer la logique applicative. Finalement, la couche d'accès aux données chargée de lire et d'écrire les données manipulées par la couche de traitement. C'est une architecture N-tiers ou multi-tiers dans laquelle chaque niveau est spécialisé dans des tâches précises.

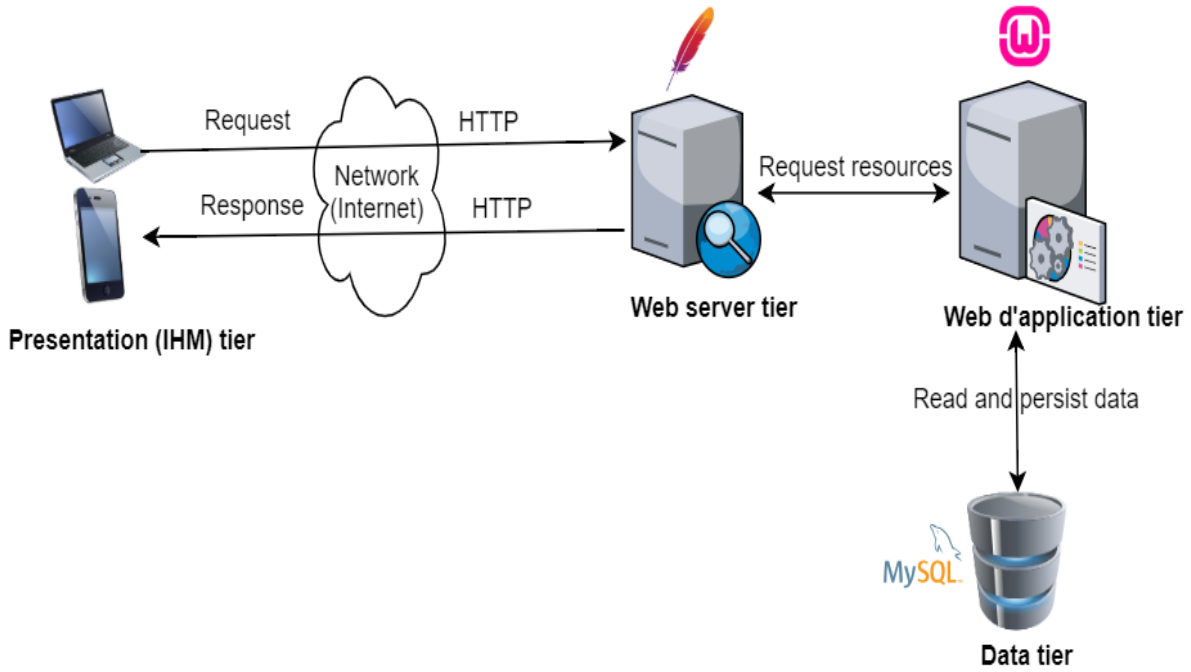


Figure III.2: Architecture N-Tier

Il s'agit d'une architecture composée des éléments suivants (Figure III.2):

- L'utilisateur : qui demande des ressources à travers les requêtes de type HTTP. Dans le cas de l'enseignement à distance, l'utilisateur peut avoir le profil de l'apprenant ou de l'étudiant, comme peut être l'enseignant, l'équipe pédagogique, l'administrateur du système ou toute autre personne qui participe dans le processus éducatif, à condition d'avoir un compte avec un login et un mot de passe lui permettant de se connecter à la plate-forme d'apprentissage.
- Le serveur web : qui intercepte la demande de l'utilisateur, cherche la ressource demandée en faisant appel au serveur d'application, puis renvoie le résultat au client.
- Le serveur d'application : permet d'exécuter ou d'interpréter la logique métier d'une application et d'envoyer le résultat au serveur web. Plusieurs serveurs d'application ont été développés, tels que Wamp pour interpréter les applications basées sur le langage de programmation PHP, Apache Tomcat pour déployer des plates-formes JEE, aussi IIS dédié aux technologies Microsoft comme ASP.NET.
- Le serveur des bases de données (SGBD) : qui est chargé de fournir au serveur d'application les données dont il a besoin. Son rôle principal est de manipuler et de gérer les données produites par un système e-learning. Actuellement, le modèle des bases de données relationnelles est le plus implémenté dans le cas des environnements d'apprentissage en ligne. Ce type de SGBD utilise le plus souvent le langage SQL pour lire, écrire, et chercher des données. Parmi ces systèmes en trouve : MySQL, Oracle, PostgreSQL, etc.

L'e-learning est, en effet, une discipline qui combine les théories de l'enseignement, les méthodes pédagogiques, et les technologies informatiques pour offrir un environnement virtuel d'apprentissage. Il s'agit d'un domaine qui exploite divers technologies et outils informatiques à des fins éducatives dans le but de promouvoir l'enseignement autonome et à distance.

III.1.3 Modèle de l'e-learning basé sur le cloud

Nous vivons aujourd'hui dans un changement technologique majeur dont le cloud fait partie. Ce dernier représente une nouvelle forme d'acquisition des services informatiques sur demande par l'intermédiaire d'internet. Il repose sur l'externalisation et la mutualisation des

ressources matérielles, plates-formes, et logiciels afin de rationaliser leur utilisation. Depuis son apparition, il est de plus en plus utilisé par divers secteurs, y compris celui d'apprentissage en ligne. Dans ce sens, l'e-learning est une discipline qui résulte du mariage de l'enseignement et des technologies de l'information et de la communication. De ce fait, les plates-formes de l'e-learning doivent rester en amélioration continue en incorporant les derniers changements technologiques, qui ne cessent d'évoluer, afin d'assurer un meilleur service éducatif et de répondre au mieux aux besoins accrus d'apprentissage de la part des universités et des organisme de formation.

E-learning basé sur le cloud est une migration du domaine d'apprentissage en ligne vers l'environnement cloud, qui constitue une infrastructure future des EIAH, y compris toutes les ressources matérielles et logicielles requises pour l'e-learning. Une fois que ces ressources informatiques sont virtualisées, elles peuvent être offertes sous forme de services aux établissements d'enseignement et aux universités pour les utiliser. L'intégration de l'e-learning dans le cloud est une mise à jour de la technologie, des concepts et des architectures, en donnant de nouvelles applications, des outils et des méthodes pour les professionnels de l'e-learning en vue de fournir un système d'apprentissage personnalisé et performant capable de s'adapter aux spécificités et aux exigences de chaque établissement d'enseignement. Certains LMS [122] font complétement confiance au cloud et ne proposent que des plates-formes e-learning en mode cloud pour assurer un service de gestion d'apprentissage agile et plus facile à exploiter.

Afin d'intégrer l'e-learning dans le cloud, nous avons opté pour une stratégie numérique qui s'appuie sur l'externalisation des ressources du stockage, du traitement, et de l'infrastructure réseau des systèmes de gestion d'apprentissage (LMS) sur la plate-forme cloud en utilisant ses différents services. Pour faire cette intégration, tout d'abord, nous avons fait une étude approfondie du rôle et des fonctionnalités de chaque composant du cloud afin de situer les systèmes e-learning dans cet environnement. Ensuite, nous avons proposé une approche en couche (Figure III.3) décrivant l'infrastructure, les plates-formes, et les logiciels permettant la disponibilité et la haute performance du système d'apprentissage en ligne.

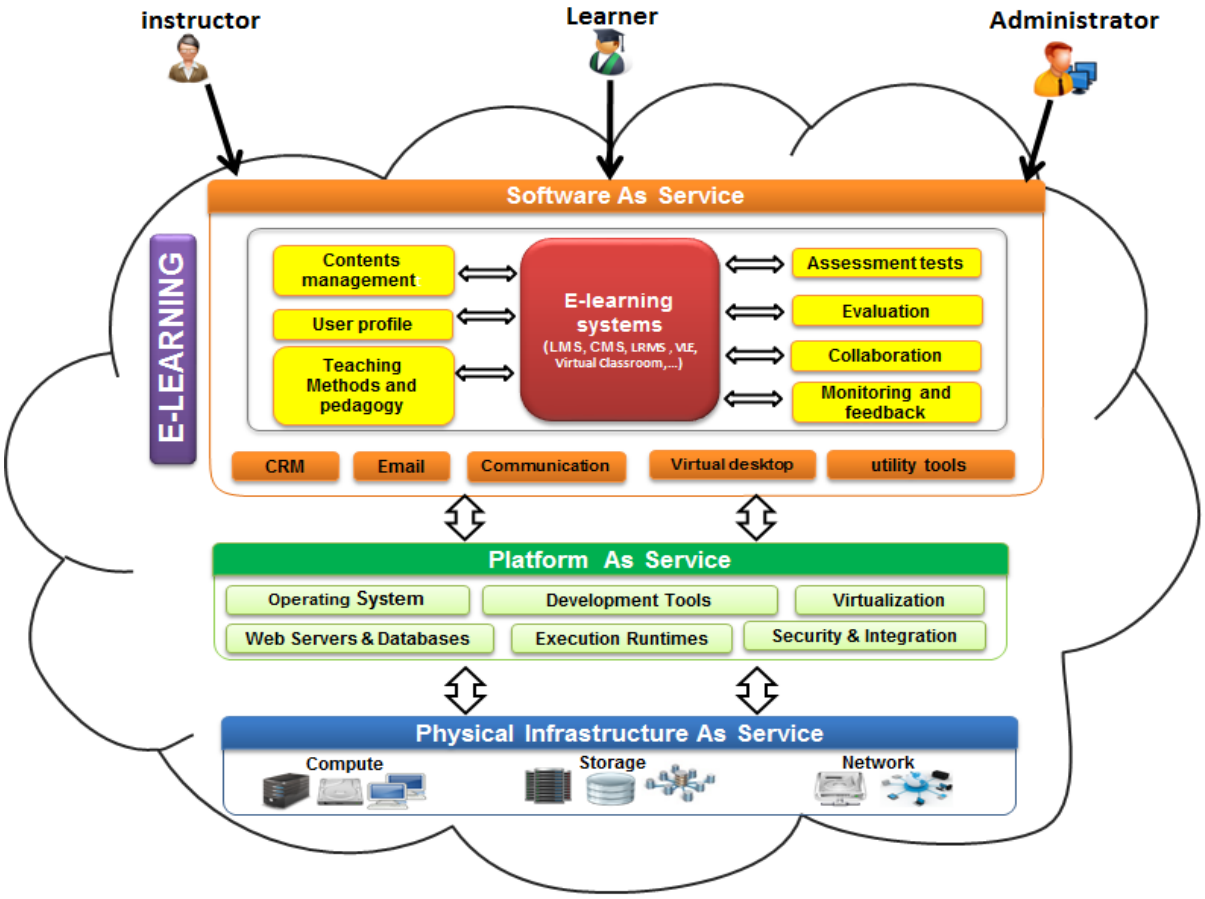


Figure III.3: Approche de l'e-learning basé sur le cloud

L'approche proposée pour l'e-learning en mode cloud peut être divisée en trois couches. Premièrement, la couche d'infrastructure en tant que service contenant un ensemble d'hôtes physiques dynamiques et évolutifs. Deuxièmement, on trouve la couche de plateforme en tant que service qui fournit des systèmes d'exploitation et d'autres outils assurant le développement, le déploiement, l'exécution, la gestion, et le contrôle des applications sur le cloud. La troisième couche est le logiciel en tant que service fournissant les fonctionnalités de création des contenus éducatifs, de diffusion des ressources pédagogiques, de tests en ligne, d'apprentissage collaboratif, d'évaluation et de gestion du processus d'apprentissage.

III.1.3.1 Couche d'infrastructure en tant que service

Cette couche est située dans le niveau le plus bas des services du cloud. Elle permet de fournir les ressources virtuelles suffisantes au bon fonctionnement d'un système e-learning, notamment les capacités de stockage, la puissance de calcul (CPU), et la mémoire. En effet, elle regroupe les services d'infrastructure et elle a la capacité de fournir l'infrastructure réseau

afin de traiter, transmettre, et conserver les données de manière répartie sur des serveurs distribués du fournisseur cloud. Grâce à l'utilisation de la technologie de virtualisation, les serveurs de calcul, de disques durs et toutes les autres ressources physiques peuvent être exploitées par un certain nombre d'applications en fonction des besoins et des exigences de chaque application. Cette couche sera appelée par des plates-formes et des logiciels utilisateurs et son rôle consiste à allouer et mettre en place des ressources informatiques nécessaires permettant de tourner les applications du client. Les ressources physiques fournies par cette couche sont dynamiques et évolutives. Un nouvel hôte physique peut être ajouté afin d'améliorer les performances des systèmes e-learning. Dans ce type d'architecture, où les ressources informatiques sont accessibles sous forme de services, l'administrateur système de la plate-forme d'apprentissage ne s'occupe ni de l'installation, ni de la configuration ou de la maintenance de l'architecture technique des serveurs cloud. Il est seulement responsable du contrôle des machines virtuelles crées, des espaces de stockage alloués et des composants de réseaux utilisés. La Figure III.4 illustre plus clairement les éléments qui composent cette couche[23].

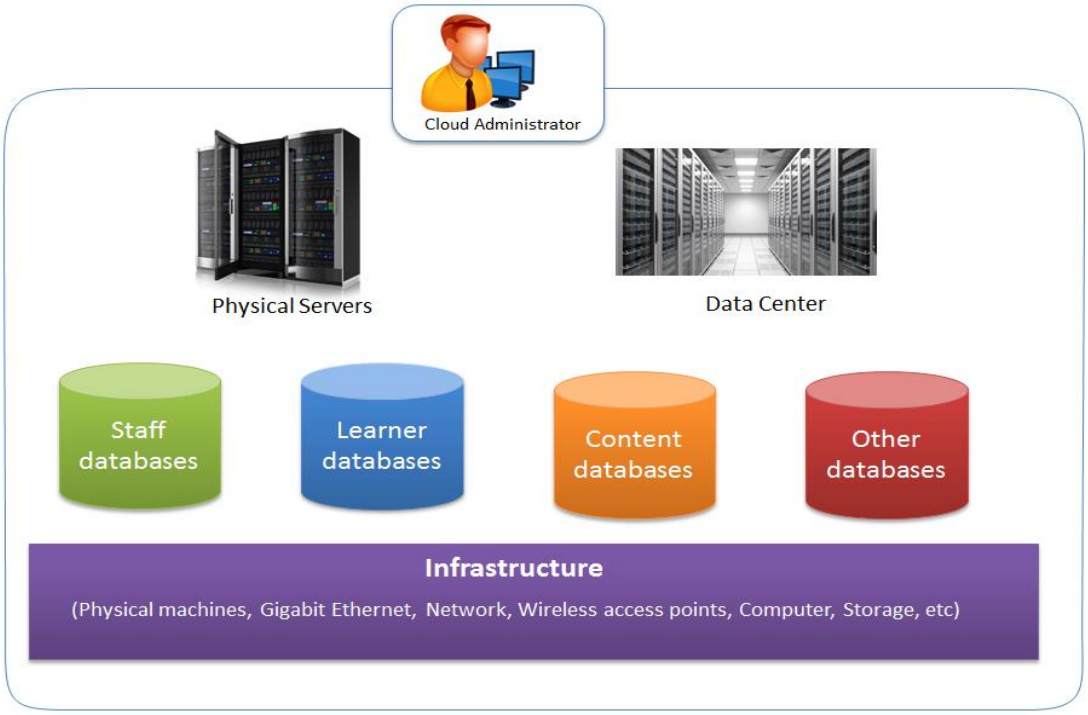


Figure III.4: Couche d'infrastructure en tant que service

Les professionnels de l'e-learning peuvent louer des espaces dans un Datacenter ou centre de données virtuel pour avoir la mémoire de stocker des cours et des ressources pédagogiques. Ces Datacenters peuvent contenir diverses bases de données pour le stockage des informations sur les profils des apprenants, leurs activités, leurs interactions avec leurs collègues ou avec leurs enseignants, etc. Parmi les composants de cette couche on trouve également des serveurs de haute performance pour garantir un temps de traitement optimal et répondre aux demandes accrues des apprenants. En utilisant les services d'infrastructure du cloud, les établissements, qui proposent l'apprentissage en ligne, peuvent bénéficier des services robustes dotés d'une capacité de calcul et de stockage presque illimitée, d'une disponibilité permanente des services d'apprentissage, d'une grande flexibilité d'évolution et d'un meilleur rapport coût-efficacité.

III.1.3.2 Couche de plate-forme en tant que service

Les plates-formes en tant que service représentent le deuxième niveau de notre approche. Cette couche regroupe des services de type plate-forme. Ce sont des logiciels intermédiaires qui servent à la création, la compilation ou l'interprétation, l'exécution, le déploiement et la gestion des applications des utilisateurs et les rendent accessibles sur internet sous forme de services logiciels. Elle comprend des systèmes d'exploitations (Windows, UNIX, etc.), des outils de développement et d'exécution, tels que codeanywhere IDE [123], codenvy[124] et des éditeurs comme koding[125], des serveurs d'application, qui sont préconfigurés et prêts à être utilisés via un navigateur web (Chrome, Mozilla, Opéra, Safari, etc.). Cette couche est composée également des technologies de virtualisation, y compris KVM [126], Xen[127], etc. En plus, la couche des services de plates-formes contient les serveurs Web, les systèmes de gestion des bases de données relationnelles et la nouvelle génération des systèmes NoSQL. Par ailleurs, Elle offre des fonctionnalités de sécurité pour protéger les données des utilisateurs contre des attaques et des virus malveillants. Généralement, cette couche fournit des logiciels et des plates-formes intermédiaires pour développer, exécuter, contrôler, surveiller, et sécuriser des applications de la couche supérieure (logiciels des utilisateurs).

Les utilisateurs de ces services sont le plus souvent des développeurs des logiciels qui utilisent des API spécifiques ou des langages de programmation pour créer, tester, déployer et gérer leurs applications sur la plate-forme cloud, ainsi que des administrateurs systèmes qui assument la tâche de mettre les supports d'apprentissage en ligne et de faire la correction des

bugs et des problèmes techniques. Grâce au cloud, ces développeurs ne doivent plus se soucier de migrer des LMS vers une nouvelle version ou à installer des outils de développement sur leurs propres machines ou des systèmes d'exploitation sur leurs serveurs. Car, c'est le fournisseur cloud qui s'occupe de ces tâches pour rendre disponible tout ce dont ils ont besoin. Il suffit juste de se connecter pour bénéficier d'un environnement complet des outils, des serveurs web, des bases de données, des APIs, et des Frameworks (Spring, JSF, Symfony, Node.JS, .NET). Les professionnels de l'e-learning ne gèrent que leurs propres systèmes d'apprentissage hébergés sur les serveurs du coud.

III.1.3.3 Couche de logiciel en tant que service

La troisième et dernière couche de l'approche proposée est la couche de logiciel en tant que service. Celle-ci regroupe les services des logiciels destinés aux utilisateurs finaux. Cette couche est celle avec laquelle l'apprenant et le tuteur ou enseignant sont les plus susceptibles d'interagir durant le processus d'apprentissage. Toute application ou logiciel client déployés sur les serveurs distants du cloud auxquels on peut accéder via Internet est considérée comme un SaaS. Dans le cas de l'enseignement à distance, le système de gestion d'apprentissage en mode cloud est une application qui prend la forme d'un service SaaS.

La couche de logiciel en tant que service se compose des applications spécifiques de l'intégration et du partage des ressources pédagogiques dans le modèle du cloud, telles que des cours et des contenus éducatifs interactifs. C'est dans cette couche où se situer les systèmes e-learning avec tous leurs composants. Cette couche comprend également la production du contenu de formation, des outils, des systèmes de collaboration entre les apprenants et les tuteurs ou formateurs, des modèles d'apprenants, des méthodes éducatives et stratégies pédagogiques, ainsi que des systèmes de gestion du contenu (CMS). Par ailleurs, c'est dans cette couche où se trouve le système de gestion de l'apprentissage (LMS) qui représente le cœur du système e-learning et joue le premier rôle dans la création, et la gestion des pages web, des blogs, des forums, des modules, etc.

Les enseignants peuvent diffuser des nouveaux contenus pédagogiques tels que des cours, des exercices, et des quizz en accédant à l'espace d'administration consacré aux formateurs où ils peuvent facilement publier des activités aux apprenants avec la facilité d'ajuster et d'adapter toute autre information en fonction du progrès et du résultat d'évaluation des apprenants. Cette couche contient aussi des bibliothèques numériques et des outils de contrôle et de

gestion des projets et des travaux des apprenants. Le système e-learning en mode SaaS peut communiquer avec d'autres plates-formes comme les réseaux sociaux et les services de messagerie ce qui permet aux participants de discuter leurs problèmes et de partager leurs connaissances avec leur entourage.

Par ailleurs, la solution e-learning en tant que service cloud est plus efficace pour les établissements et les universités souhaitant diffuser rapidement des contenus éducatifs, parce que le choix du cloud réduit grandement le temps du déploiement et de la mise en place d'une plate-forme d'apprentissage en ligne. En plus, la plupart des enseignants et des formateurs n'ont pas des compétences informatiques suffisantes, et ne veulent pas se soucier des pannes matérielles ou techniques ou même des problèmes de mise à jour des logiciels. Alors, l'objectif des services SaaS, en premier temps, est d'écarter les problèmes de dysfonctionnement du système e-learning en permettant aux professionnels de l'e-learning de se libérer des aspects techniques et de se concentrer sur tout ce qui métier tels que la préparation et l'élaboration des contenus d'apprentissage et des activités pédagogiques au profit des apprenants.

III.1.4 Mode interactif de l'e-learning en mode SaaS

Dans le modèle classique d'apprentissage, les enseignants attribuent des tâches d'enseignement, animent régulièrement des conférences ou forment les apprenants. Les apprenants suivent le processus d'apprentissage en ligne et contribuent aux sessions d'apprentissage coopératif, ou accomplissent les tâches des enseignants. Mais dans l'e-learning en mode cloud, les enseignants répondent également aux questions des apprenants en un temps record et offrent un apprentissage adaptatif en fonction des besoins pédagogiques de l'apprenant.

En outre, les équipes pédagogiques peuvent également utiliser le multimédia pour améliorer le support d'enseignement. Les participants à la formation élaborent leurs propres plans d'apprentissage en déterminant les méthodes éducatives de manière autonome. Ces derniers interagissent avec le système e-learning pour apprendre, finissent leurs tests et établissent des statistiques sur les résultats de chaque module. Les enseignants encouragent également les apprenants à coopérer pour mener à bien des tâches d'apprentissage simples ou des projets complexes basés sur des groupes. Grâce à l'apprentissage coopératif via des outils de collaboration du cloud, les apprenants peuvent acquérir des nouvelles connaissances, leur

esprit d'équipe et leurs capacités d'expression et de communication seront également renforcées, leurs compétences relationnelles avec les autres seront améliorées. Ainsi, l'apprentissage sera plus interactif et individualisé, ce qui permet de fournir un service d'apprentissage de qualité. Le mode interactif de l'e-learning basé sur le cloud est illustré dans la Figure III.5.

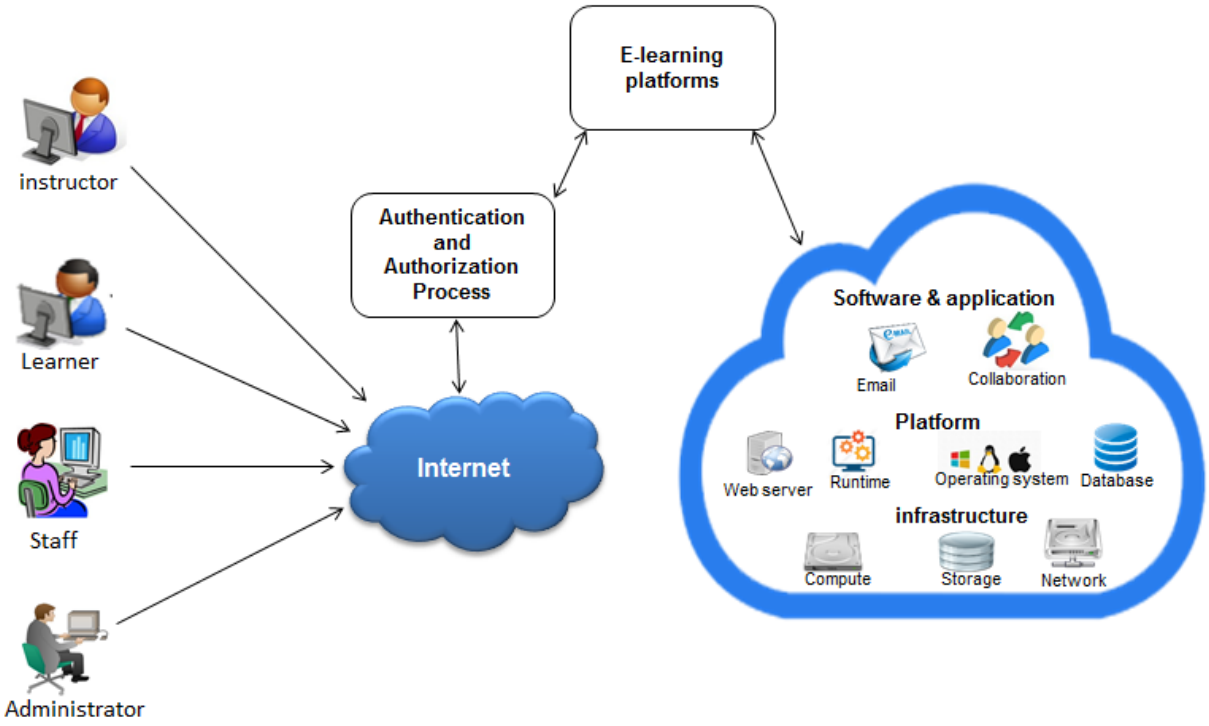


Figure III.5: Mode interactif de l'e-learning en mode SaaS

Comme il est montré dans la Figure III.5 tous les services du cloud, y compris l'infrastructure, les plates-formes et les logiciels se communiquent entre eux pour offrir un service e-learning performant, évolutif et adaptatif aux besoins d'apprentissage.

III.1.5 Comparaison entre le modèle traditionnel de déploiement et celui du cloud

Le développement remarquable du cloud ces dernières années suscite de plus en plus l'intérêt des utilisateurs d'internet, des organismes et des universités qui cherchent à profiter de ces services et de ces outils. Principalement, grâce à sa facilité d'utilisation et d'accès à travers des plates-formes disponibles en ligne, via le Web, en mode service. L'utilisation des services cloud est à la demande et le paiement est en fonction de la consommation des ressources.

Pour comprendre la différence entre les deux modes de déploiement, le modèle traditionnel d'hébergement et celui basé sur cloud, il faut identifier les caractéristiques de chaque mode. La table III.1 donne une comparaison synthétique des deux modèles.

Tableau III-1: Comparaison entre le modèle traditionnel de déploiement et celui du cloud

Modèle / **Caractéristiques**	**Traditionnel**	**Cloud**
Acquisition des ressources	Acheter des ressources informatiques physiques (serveurs, stockage,…) Construire l'architecture technique	S'abonner aux services en ligne Architecture technique est inclue
Économique	Payer le coût des ressources matérielles Payer les frais des tâches d'administration	Paiement en fonction d'utilisation Réduction des fonctions d'administration et de maintenance
Accès à des ressources	Accès au sein du réseau local locaux d'entreprise ou université	Accès universel Compatibilité avec tous les appareils
Technique	Statique	Dynamique, flexible, évolutif, et élastique
Déploiement	Hébergement coûteux et difficile Demande beaucoup de temps et d'effort	Déploiement rapide et facile Gestion automatique des dépendances et des versions

Les cinq caractéristiques clés qui différencient les deux modèles sont regroupées par niveau comme suit:

Au niveau du modèle d'acquisition des ressources : dans le mode d'infrastructure classique, les établissements d'apprentissage à distance ou les universités doivent acheter des ressources matérielles et construire l'architecture technique telle que l'installation et la configuration des serveurs physiques. Alors que dans une infrastructure basée sur le Cloud, les entreprises s'abonnent juste aux services de la plate-forme web du fournisseur cloud, et l'architecture technique est y incluse par défaut.

Au niveau économique : dans le mode modèle traditionnel, le client doit payer pour l'achat des matériels et pour les tâches d'administration, par contre dans le modèle du cloud on paye seulement pour l'utilisation ou la consommation des ressources, et les fonctions d'administration et de maintenance sont réduits ce qui permet aux universités de gagner en temps et en budget.

Au niveau d'accès à des ressources : lorsqu'on utilise les serveurs traditionnels l'accès à des données est seulement autorisé à l'intérieur du réseau local de l'université, au moment où dans le modèle cloud les données sont accessibles par n'importe qui, n'importe quand, et à partir de de n'importe quel terminal via internet.

Au niveau du modèle technique : pour l'architecture traditionnelle le modèle technique est statique alors que dans l'environnement cloud, il est flexible, évolutif, et élastique. Les services cloud mettent en place des machines virtuelles et des serveurs web et des serveurs d'application avec des mises à jour continues et automatiques. En plus, le cloud permet une mutualisation des ressources et une adaptation dynamique des capacités des serveurs en fonction des variations de la charge.

Au niveau du modèle du déploiement : lorsqu'on utilise l'infrastructure traditionnelle le déploiement des applications est coûteux et demande beaucoup du temps, tandis que dans le modèle basé sur le cloud, le déploiement est très rapide. Par ailleurs, Le cloud facilite l'hébergement des applications modernes en permettant la gestion des dépendances et des versions et une meilleure interopérabilité avec d'autres systèmes comme les SGBDR, les outils de messagerie, etc.

III.1.6 Avantages et bénéfices du modèle e-learning basé sur le cloud

L'e-learning en mode logiciel en tant que service (SaaS) est une plate-forme d'apprentissage en ligne déployée sur l'infrastructure cloud. Il vise essentiellement à proposer, organiser, suivre des cours ou des formations à distance. Ce nouveau mode d'hébergement a la capacité d'automatiser tout le processus d'apprentissage, y compris la création, l'édition, et la diffusion des contenus éducatifs, en plus des outils de collaboration entre apprenants et tuteurs, etc. Il est diffèrent des logiciels classiques d'apprentissage qui demandent des efforts de maintenance, des frais d'installation et de maintenance[128].

La migration d'un système e-learning à partir des serveurs web traditionnels vers l'environnement du cloud présente une large gamme de possibilités et d'outils innovants pouvant améliorer grandement le service d'apprentissage en ligne en offrant un système d'enseignement à distance performant, évolutif et adapté à la taille de tous les établissements d'enseignement souhaitant externaliser tout ou partie de leur infrastructure informatique. En résumé, l'implémentation des technologies du cloud a plusieurs avantages et bénéfices pour les environnements informatiques pour l'apprentissage humain grâce au concept de la décentralisation des espaces de stockage et de la mutualisation des ressources. Parmi ces avantages, on trouve notamment :

- **Puissante capacité de calcul et de stockage** : l'architecture E-learning basée sur le cloud localise les données dans un grand nombre (des dizaines de milliers) d'ordinateurs distribués pour permettre une puissante capacité de calcul et un énorme espace de stockage des données offerts sous forme de services capable des stocker des volumes importants de données et de supporter une montée en charge importante. En utilisant, l'infrastructure cloud, les systèmes d'apprentissage sont désormais capables de réponde aux demandes des apprenants, des enseignants et des équipes pédagogiques en un laps de temps record.
- **Réduction des coûts** : le modèle du coud donne une grande efficacité économique. Dans le cas d'hébergement en mode SaaS, les frais d'installation et de configuration sont très faibles. En plus, le déploiement nécessite le minimum de ressources de stockage, de CPU et de mémoire. En fait, grâce aux services cloud, les professionnels de l'e-learning peuvent déployer et exécuter un système e-learning à partir de n'importe quel périphérique (téléphone portable, PC, ou tablette) possédant une

connexion internet. Les données produites par ces systèmes sont accessibles via le cloud, et les établissements d'enseignements n'ont pas besoin de dépenser leur budget dans l'achat et d'installation des équipements de stockage et de traitement (disques durs, RAM & processeurs) sur des serveurs locaux, en revanche ils peuvent s'abonner à un forfait d'un fournisseur cloud et payer les ressources consommées. En effet, Le mode de facturation du cloud se base sur le principe du paiement à l'usage, en fonction de la consommation des services. Alors, les responsables du système e-learning doivent uniquement rembourser un abonnement mensuel en fonction du nombre d'apprenants ce qui permet de maitriser le budget des organismes de formation à distance.

- **Haute disponibilité** : grâce à l'intégration du stockage de masse et de la puissance de calcul de haute performance, la migration d'e-learning sur l'environnement cloud permet de fournir une meilleure qualité de service d'apprentissage. En effet, l'écosystème cloud est très tolérant aux pannes techniques et gère efficacement les défaillances matérielles des serveurs à l'aide du mécanisme de redondance des serveurs et la répartition des bases de données dans différents datacenters. Chaque fournisseur a ses propres techniques et stratégies pour détecter automatiquement la défaillance d'un nœud et l'exclure, sans affecter le fonctionnement normal du système, ce qui permet de garantir une fiabilité et une disponibilité totale et continue des plates-formes d'apprentissage en ligne dans n'importe quelle condition.
- **Haute sécurité :** dans le modèle de l'e-learning en mode SaaS, les données sont stockées de manière intensive. En s'appuyant sur un ou plusieurs centres de données, les responsables e-learning gèrent les données, allouent les ressources, déploient le système d'apprentissage en ligne. En revanche, le fournisseur cloud s'occupe du contrôle de la sécurité des applications hébergées sur le PaaS via des stratégies efficientes en installant des logiciels antivirus et des Pare-feu sur les serveurs pour protéger les données des apprenants et les ressources pédagogiques contre les menaces et les programmes indésirables des hackers. Par ailleurs, dans le mode de l'e-learning en tant que service, les données sont plus fiables et plus sécurisées que dans le mode d'hébergement traditionnel sur des serveurs dédiés. En plus, la plate-forme cloud attribuent aux administrateurs d'e-learning des rôles leurs permettant de contrôler

l'accès aux données et restreindre l'accès aux utilisateurs qui n'ont pas de privilèges suffisants.

- **La virtualisation :** représente l'élément clé des technologies du cloud. En fait, elle constitue la caractéristique la plus importante de ce type d'architecture. La virtualisation est un mécanisme qui consiste à faire fonctionner plusieurs systèmes, serveurs ou applications, sur un même serveur physique. Dans le cas de l'enseignement à distance, la plate-forme e-learning est hébergée, gérée, contrôlée et sauvegardée par l'intermédiaire d'un outil de virtualisation. Ainsi, les organismes de formations ne sont plus obligés d'acheter un nouveau serveur physique pour chaque application, de même plusieurs instances d'un système d'exploitation (Windows, Unix, etc.) peuvent être lancées sur une même machine. La technique de virtualisation permet une meilleure élasticité en fournissant des services évolutifs capables de supporter de très grandes montées en charge.
- **Mutualisation des ressources :** Il s'agit du partage des ressources non seulement entre différentes applications (virtualisation), mais également entre plusieurs clients. Cette mutualisation fait des systèmes e-learning en mode cloud un service très efficace qui garantit aux professionnels d'apprentissage en ligne le bon fonctionnement de leurs plates-formes. Cette technique permet l'optimisation des capacités de calcul, de stockage et de traitement des serveurs virtuels du cloud. Ces ressources allouées peuvent être libérées lorsqu'elles deviennent inactives, pour être utilisées par d'autres utilisateurs ce qui les rend moins cher et rapidement réutilisable.
- **Mises à jour logicielles instantanées et automatiques :** les logiciels et les Framework sont automatiquement mis à jour dans le cloud. En effet, le fournisseur des services du cloud s'occupe des tâches d'installation et de maintenance de l'environnement du développement et du déploiement pour que les utilisateurs puissent obtenir les mises à jour instantanément. Le modèle de l'e-learning basé le cloud, permet une mise à jour permanente des composants du système de gestion d'apprentissage (LMS) et de gestion des contenus (CMS). Le fournisseur cloud s'occupe également des sauvegardes automatiques des données des utilisateurs. Il offre également des mises à jour des APIs, des langages de programmation, et des versions des logiciels installés sur leurs serveurs. De cette façon les professionnels d'e-learning

peut bénéficier des dernières avancées technologiques immédiatement et régulièrement.

- **Compatibilité avec tous les formats des documents :** comme certains formats de fichiers et documents ne s'ouvrent pas correctement sur certains ordinateurs ou téléphones mobiles à cause de de la non compatibilité du format du document consulté, ce qui nécessite l'installation d'un logiciel supplémentaire. En utilisant l'environnement cloud, les enseignants et les apprenants ne doivent pas s'inquiéter de ce genre de problèmes puisque les applications d'apprentissage en ligne basées sur le cloud ouvrent toutes les extensions de documents à l'aide des logiciels en ligne fournis par le cloud. Un exemple de ce type de logiciel est celui de Google Docs ou de Dropbox qui permettent de visualiser, éditer, stocker, et partager les documents Word, PowerPoint, et Excel facilement sur le cloud par l'intermédiaire uniquement d'un navigateur web.
- **Avantages pour les étudiants :** les étudiants tirent davantage parti de l'apprentissage basé sur le cloud, car ils peuvent améliorer leurs compétences en suivant des cours, en passant des examens en ligne, en obtenant les retours de leurs tests et en envoyant leurs travaux et projets à leurs enseignants n'importe quand, n'importe où, et à partir de n'importe quel appareil. Par ailleurs, via des outils de collaboration du cloud, ils peuvent apprendre en échangeant les idées et les concepts d'un module de formation sur des forums et des réseaux sociaux et participer à la réalisation des activités pédagogiques en groupe ce qui leur permet de renforcer leur esprit d'équipe et la culture de partage des connaissances.
- **Avantages pour les enseignants :** les enseignants tirent également de nombreux avantages de l'apprentissage en ligne dans le cloud. Ils peuvent préparer des tests en ligne pour les étudiants, et suivre en temps réel le progrès de leurs apprenants de n'importe quel endroit où ils se situent géographiquement. Ils peuvent également profiter des techniques avancées de l'apprentissage machine proposés par le cloud pour identifier les besoins d'apprentissage les plus demandées et déterminer les modules auxquels ils peuvent apporter des améliorations. En plus, ils ont accès à des nombreux services cloud favorisant le travail collaboratif et la participation des étudiants dans des activités et des projets en ligne. L'écosystème cloud permet aussi aux enseignants et équipes pédagogiques de se concentrer sur les méthodes

pédagogiques et de se libérer des aspects technologiques qui posent souvent beaucoup de problèmes.

- **Mise en œuvre rapide :** Dans l'e-learning en mode SaaS, le temps de déploiement d'un système d'apprentissage est significativement réduit et l'hébergement est beaucoup plus rapide par rapport à celui des serveurs classiques. Car, en mode cloud, l'infrastructure physique et les serveurs web et d'application sont déjà installés et prêtes à utiliser. Ainsi, le cloud assure une connexion rapide en fournissant divers APIs permettant de se communiquer facilement avec différents SGBD. Également, La gestion des dépendances et l'interopérabilité avec d'autres systèmes informatiques sont confiées aux services PaaS. Le cloud s'occupe de nombreuses tâches techniques ce qui permet un déploiement en un temps record d'un LMS sur un environnement cloud. Généralement, le cloud fournit un environnement complet permettant le déploiement rapide des plates-formes e-learning sans se soucier des contraintes de compatibilité ou de mise à jour des logiciels et serveurs. De surcroît, la plupart des solutions cloud supportent l'hébergement des applications écrites en utilisant différents langages de programmation (PHP, Java, Phyton, etc.). Cet environnement préconfiguré minimise beaucoup le temps et l'effort de la mise en place d'un système e-learning.

Le principal avantage de l'approche proposée, est qu'elle vise à fournir un accès facile à des logiciels coûteux fonctionnant sur des processeurs de hautes performances pour les étudiants des établissements ruraux qui ne disposent pas de budget d'investissement pour implémenter l'installation et effectuer la maintenance de l'infrastructure physique nécessaire. L'e-learning en mode SaaS est une meilleure solution pour les universités ayant un manque dans les ressources informatiques, car il permet d'éviter les dépenses d'achat, et de configuration des serveurs, des espaces de stockage, et de l'infrastructure réseau. En bref, le modèle du cloud constitue un véritable accélérateur de la transformation digitale pour les organismes de formation souhaitant avoir des services numériques de qualité.

III.1.7 Valeur ajoutée du cloud pour les plates-formes e-learning

L'architecture cloud exploite la puissance des serveurs répartis du fournisseur cloud qui sont interconnectés entre eux par des protocoles de communication réseau pour fournir un environnement complet capable de s'adapter dynamiquement aux besoins et aux exigences de toute application. Elle offre une infrastructure plus flexible grâce aux technologies de

virtualisation et de mutualisation des services informatiques. Grâce au cloud, il est désormais possible d'accéder, de gérer, et de déployer des plates-formes d'apprentissage en ligne. Pour ce faire, Il suffit juste d'une simple connexion internet et d'un abonnement en fonction de l'utilisation. En effet, Le cloud rend les contenus éducatifs accessibles partout en donnant une autonomie d'apprentissage aux apprenants, et une simplicité de gestion aux équipes pédagogiques. En plus, Il permet une grande flexibilité dans la mise en place des projets e-learning. Actuellement, via une simple interface web, les administrateurs des systèmes d'apprentissage en ligne ont la possibilité d'allouer et de contrôler toutes les ressources nécessaires au bon fonctionnement de ces plates-formes, y compris, l'infrastructure réseau virtuelles, les espaces de stockages, et le temps de calcul. Par ailleurs, il s'occupe de la mise à jour automatique, de la sécurité et de la pérennité des données. Les services PaaS proposent également un ensemble de systèmes supplémentaires comme des outils de workflow et d'autres outils d'analyse avancée, mais aussi des APIs sophistiqués visant à aider les équipes pédagogiques à créer des applications de collaboration et de partage pour répondre au mieux aux besoins des utilisateurs.

Par ailleurs, le PaaS permet aux enseignants et équipes padégogiques de se concentrer sur la création d'excellentes ressources pédagogiques de qualité, sans devoir se soucier de l'installation et de la configuration de l'architecture technique ou du lancement des systèmes d'exploitation. L'environnement de développement et les serveurs web virtuels sont facilement accessibles via le cloud, ce qui permet d'éviter des tâches complexes et onéreuses d'administration et de maintenance en interne. Par conséquent, le système e-learning est hébergé plus rapidement et à des coûts raisonnables, sans aucun risque de retard dû à des problèmes système ou d'infrastructure.

De nos jours, le cloud est la solution informatique la plus efficace non seulement pour les universités et les établissements d'enseignements, mais aussi pour toute sorte d'entreprise quelle que soit leur taille. Le plus important pour ce nouveau paradigme digital est surtout son mode du paiement qui se base sur une facturation en fonction de la consommation. Ainsi, le cloud permet de garantir une infrastructure IT de haute performance pour déployer et exécuter un système d'apprentissage en ligne dans des bonnes conditions en assurant le rapport coût-qualité, ce qui permet d'améliorer grandement la disponibilité, la fiabilité, et la robustesse du service d'apprentissage.

III.1.8 Contraintes et limites du cloud

Le cloud est une incontestable solution informatique pour toute entreprise ou établissement d'enseignement quelle que soit sa forme, sa taille (petite, moyen, grande), et ses ressources. Malgré les avantages du cloud, il y a un certain nombre d'inconvénients et des limites qui poussent quelques clients à ne pas faire totalement confiance au choix de migration de leurs infrastructures sur la plate-forme cloud.

III.1.8.1 Sécurité

Le grand problème du cloud est celui de la sécurité. Les cas de piratage informatique se multiplient malgré l'installation des mesures de sécurité comme Pare-feu (Firewall) et/ou antivirus sur les serveurs des fournisseurs cloud. Parmi ces attaques on trouve le vol des données des géants du web (Dailymotion ou de Yahoo) et des grandes organisations telles que Panama Papers [129] et Wikileaks[130]. Le risque de sécurité réside, et les menaces de sécurité et le cyber sécurité se diversifient. En effet, parfois les hackers attaquent les utilisateurs au lieu des serveurs. Une fois ils s'acquièrent les identifiants (login et mot de passe) d'un utilisateur, ils peuvent accéder aux serveurs et lancer programmes malveillants pour récupérer les données dont ils ont besoin. Par conséquent, les clients restent très prudents à propos de la sécurité et des accès non autorisés à leurs données confidentielles. En hébergeant des données personnelles sur des serveurs distants, l'utilisateur risque de perdre le contrôle sur ses données notamment en cas de bug ou de piratage.

III.1.8.2 Confidentialité

Lorsqu'on fait le choix d'externalisation du système e-learning sur le cloud, les données deviennent disponibles sur des serveurs distants dans des locaux du cloud, ce qui signifie qu'elles sont en dehors de l'université. Dans ce sens, les lois du pays en matière de la protection des données ne restent plus valables, et les données d'apprentissage subissent à la réglementation du pays du fournisseur cloud. Ce qui peut poser beaucoup de problèmes surtout dans le cas des données sensibles. En outre, ces données peuvent être inaccessibles pour des situations où il y a des problèmes de communication réseau et l'utilisateur peut avoir des difficultés à accéder à ses services. Ainsi, l'absence d'un service pendant une durée peut signifier la perte de confiance du client ainsi que du budget.

III.1.9 Conclusion

Dans ce chapitre, nous avons proposé une approche pour intégrer les services du cloud computing dans les systèmes e-learning. La mise en œuvre de l'apprentissage en ligne est généralement faite à l'intérieur d'un établissement d'enseignement ce qui rend le déploiement de de ce type de plates-formes très coûteux en termes d'infrastructure et de ressources humaines pour l'installation et la maintenance des serveurs physiques et des logiciels. Dans ce sens, le cloud computing s'avère un modèle prometteur en raison de son élasticité, sa disponibilité, sa stratégie de mutualisation de ressources, et son mode de facturation en fonction de la consommation. Avec la mise en œuvre d'un système e-learning dans un environnement cloud, les établissements d'enseignement ou les universités n'ont plus à payer pour l'achat des matériels (disque durs, serveurs, ...) et les tâches d'administration, ils paient uniquement les frais d'abonnement. Ce chapitre a également analysé le concept du cloud computing et a décrit son architecture en combinant les caractéristiques de l'apprentissage en ligne. Par la suite, nous avons mené une étude comparative entre le modèle traditionnel de déploiement et le celui du cloud computing. En fin nous avons discuté les avantages attendus de l'approche proposée, ainsi que les contraintes de la migration des systèmes e-learning vers l'architecture du cloud.

III.2 Modèle d'e-learning basé sur le cloud et le big data

III.2.1 Introduction

Actuellement, le monde de l'e-learning connait un grand changement, surtout avec l'apparition des nouvelles technologies de l'information et de la communication et avec la forte demande d'apprentissage, l'énorme croissance du nombre des apprenants, et la diversité des contenus éducatifs et des ressources mises à disposition. Par conséquent les dimensions des systèmes d'apprentissage en ligne se croissent d'un rythme exceptionnel. En fait, les défis en ce qui concerne la gestion des données en termes de stockage, de traitement, d'analyse et d'optimisation, nécessitent de mettre en place une plate-forme capable de répondre aux besoins de chaque apprenant.

Pour faire face à ces défis, les chercheurs ont orienté leurs études vers la recherche des nouvelles façons et stratégies basées sur des approches distribuées et parallèles pour réponde aux exigences du traitement à grande échelle et d'analyse en temps réel des bases de données géantes. Dans ce contexte, le big data s'avère un paradigme prometteur en raison de son évolutivité permanente et de ses opportunités qui offrent aux professionnels de l'e-learning en matière de stockage distribué, de calcul en parallèle, et d'analyse avancée des données massives et hétérogènes.

Le principe du paradigme big data consiste à répartir le traitement intensif et le stockage massif des données sur plusieurs nœuds d'un cluster afin d'accélérer leur manipulation. Pour ce faire, un certain nombre de systèmes distribués [59], [73], [84] ont été développés permettant l'exécution des applications simultanément sur plusieurs machines de calcul. Cependant, ces approches distribuées impliquent un temps d'échange et de transfert des données entre les nœuds. Pour résoudre ce problème, les technologies big data ont proposé des mécanismes d'optimisation afin de réduire au maximum le temps de calcul et la latence de communication. Nous citons à titre d'exemple, le Framework Apache Spark qui offre un système d'optimisation très robuste appelé Catalyst Optimizer permettant d'établir le plan d'exécution optimal avant de lancer le traitement d'une tâche sur un cluster.

Par ailleurs, les technologies big data demandent des puissances de traitement et des espaces de stockage énormes. Ce qui nécessite des grands ordinateurs dotés des ressources processeur, mémoire, et espace disque offrant des performances considérables. Dans ce sens, on peut

faire appel aux services du cloud. Lorsqu'on parle de big data, bien entendu on parle également de big computer, c'est tout simplement le cloud qui offre des ressources informatiques de calcul théoriquement illimitées sous forme des services. En effet, le cloud et le big data sont deux éléments indissociables. Le cloud offre des puissances de calcul des espaces de stockage énormes pour exécuter les algorithmes en parallèle de machine learning. Donc, pas de big data sans cloud.

En fait, le big data désigne non seulement un stockage des données très important, mais aussi des besoins de traitement sur ces volumes de données. Il s'agit, en plus de stocker les données sur un système de fichier distribué (HDFS) ou sur une base de données répraties NoSQL afin de pouvoir effectuer des traitements sur ces données de façon également distribuée et de pouvoir en restituer les résultats.

Le big data, dans le cadre de l'e-learning, concerne les données créées par les apprenants lorsqu'ils suivent un cours en ligne ou un module de formation. Par exemple, si un apprenant interagit avec un système d'apprentissage lors d'une session d'apprentissage en ligne, ses progrès, ses résultats d'évaluation, son partage social, ses commentaires, ses interventions pédagogiques et toute autre donnée produite pendant une formation à distance est considérée comme big data. Ces données peuvent être analysées pour offrir aux universités et aux professionnels de l'e-learning la possibilité de déterminer comment l'apprenant acquiert l'information, à quel rythme et d'identifier les problèmes qui peuvent exister dans une stratégie d'apprentissage. En outre, grâce à l'analyse avancée des techniques d'apprentissage automatique, on peut prédire les prochaines actions d'un apprenant et lui proposer les dispositifs d'enseignement les plus adaptés à son niveau cognitif et son style d'apprentissage. Dans le contexte de l'apprentissage en ligne, les techniques du big data peuvent servir à développer des systèmes de prédictions sophistiqués dans le but d'orienter et d'aider les apprenants à trouver les cours les plus pertinents en leur suggérant les ressources pédagogiques les plus adaptées à leurs préférences.

Ce chapitre sera consacré à la présentation de notre contribution dans le domaine de l'EIAH, en particulier, nous présenterons notre modèle visant à intégrer la nouvelle génération des technologies big data dans le contexte d'apprentissage en ligne. Nous commencerons par exposer notre approche intégrant les Frameworks du big data, les systèmes d'apprentissage en

ligne et les services du cloud. En outre, nous proposerons une méthodologie big data adaptée au contexte d'EIAH. Elle décrit les différentes étapes du processus de la manipulation des données d'apprentissage en ligne dans un environnement big data distribué. Par la suite, nous proposerons l'architecture de l'e-learning basé sur le paradigme big data. Ce chapitre abordera également les avantages et les bénéfices du big data pour les professionnels de l'e-learning.

III.2.2 Modèle e-learning basé sur le cloud et le big data

Le cloud computing est devenu une solution incontournable pour les entreprises de toutes tailles pour surmonter les défis d'acquisition, d'installation et de maintenance des ressources informatiques qui demandent beaucoup d'investissement en termes des compétences techniques et du budget. Il consiste à externaliser les ressources d'infrastructure, de plate-forme et des logiciels sur des serveurs virtuels distants de la plate-forme cloud qui sont offerts sous forme de services accessibles via internet. En fait, le cloud a le potentiel d'offrir des capacités de stockage et des puissances de calcul adaptés aux quantités énormes des traitements requises par les algorithmes parallèles du big data.

Les universités et établissements d'enseignement peuvent également profiter des avantages de ces deux nouveaux paradigmes. En effet, les EIAH dédiés à l'apprentissage en ligne, peuvent exploiter les méthodes d'analyse avancées pour exploiter les données produites par un EIAH, afin de développer des applications sophistiquées en se basant sur l'infrastructure cloud. De cette manière, les professionnels de l'e-learning seront capables de fournir un service d'enseignement capable de répond aux besoins accrus d'apprentissage et d'adapter les décisions pédagogiques aux compétences individuelles de chaque apprenant.

En fait, nous ne pouvons pas discuter l'intégration du big data et des plates-formes de l'e-learning sans considérer le cloud, parce qu'il fournit les ressources nécessaires au déploiement des technologies et des outils big data, ainsi que des systèmes de gestion d'apprentissage (LMS). En plus, le cloud offre un environnement préconfiguré et prêt à être utilisé, et qui incorpore les technologies de traitement des données massives. Cet environnement permet aux professionnels de l'e-learning de se concentrer sur l'aspect métier et de se libérer des aspects techniques d'installation des matérielles et de mise à jour des logiciels. En fait, les organismes de formation qui ne disposent pas de l'infrastructure nécessaire, peuvent désormais s'orienter vers les offres du cloud pour gérer facilement leurs données et profiter des calculs de haute performance et d'analyse avancée et en temps réel. En combinant les

services du cloud et les technologies de traitement massivement parallèle du big data, les systèmes e-learning vont grandement améliorer leur performance et renforcer leur disponibilité afin de fournir un service d'apprentissage de haute qualité.

Le big data et le cloud sont devenus des éléments clés de tout système d'information, y compris les systèmes e-learning. De ce fait, leur intégration est une nécessité majeure pour se libérer des enjeux d'acquisition et d'infrastructure et pour tirer parti des volumes importants de données générées par un tel système via des modèles analytiques très riches et des algorithmes performants de machine learning. L'identification des informations pertinentes à partir des données d'apprentissage constitue un grand défi, surtout avec l'augmentation constante des quantités de données produites chaque jour par les plates-formes d'apprentissage en ligne. Pour arriver à résoudre cette problématique, le big data introduit des techniques d'intelligence artificielle et des méthodes de machine learning sous forme d'APIs, de Frameworks et d'outils très puissants et simples à mettre en place. De telles technologies permettent de préparer et d'analyser, de manière distribuée, des grosses quantités de données afin de prendre la meilleure décision et d'aider les professionnels de l'e-learning pour être capables d'améliorer en permanence leurs stratégies d'apprentissage pour qu'elles soient adaptées aux besoins des apprenants.

Afin de réaliser une intégration souple de ces trois paradigmes nous proposons une approche en couche combinant les systèmes d'apprentissage en ligne, le big data et le cloud. Ce dernier constitue le noyau qui sert à déployer, exécuter, stocker des données, et faire fonctionner tous les autres composants de notre modèle. C'est une approche de quatre niveaux comme l'illustre la Figure III.6.

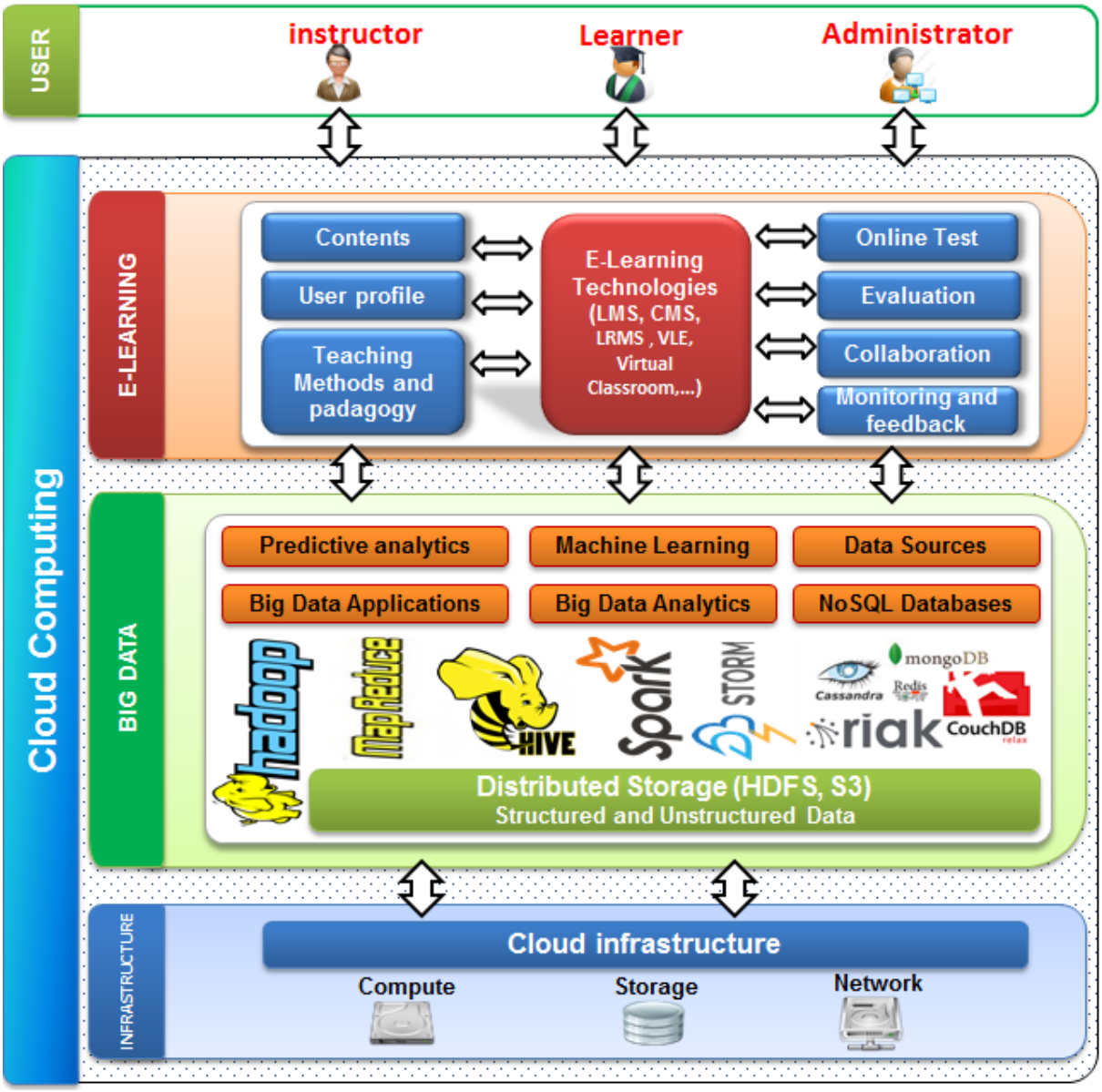


Figure III.6: Approche d'intégration des systèmes d'apprentissage en ligne, du big data et du cloud

L'infrastructure est la première couche qui constitue le niveau le plus bas de l'approche proposée. La couche infrastructure est construite avec des ressources de calcul, de stockage et de réseau qui sont virtualisées et offertes sous forme de services par le cloud. Ce dernier est chargé de fournir des ressources informatiques virtuelles et les technologies big data nécessaires pour offrir aux systèmes e-learning un environnement d'exécution favorable. Les ressources de cette couche sont dynamiques et évolutives. Si, par exemple, lors de l'analyse d'un volume massif de données, une application demande des capacités énorme de temps de

calcul ou d'espace disque, l'infrastructure cloud va s'étendre automatiquement pour allouer les ressources requises par cette application. Ce mécanisme permet une grande flexibilité par rapport à l'approche classique basée sur l'infrastructure traditionnelle d'hébergement dans laquelle les ressources du serveur sont limitées. Le grand avantage de cette couche est qu'elle offre une infrastructure scalable, élastique et tolérante aux pannes.

La deuxième couche de cette approche est celle de l'écosystème big data. La couche big data comprend des technologies de stockage décentralisé et réparti des données volumineuses, de calcul en massivement parallèle, d'analyse avancée, d'optimisation et de visualisation des résultats de traitement. Elle regroupe diverses outils du big data qui peuvent être classées en trois catégories:

- **Systèmes de fichiers distribués** : qui stockent les données dans plusieurs nœuds d'un cluster d'une manière répliquée pour assurer la redondance et la haute disponibilité. HDFS est le système le plus connu comme solution open source de stockage et de gestion distribuée des données.
- **Bases de données NoSQL** : qui représentent la nouvelle génération des systèmes de gestion des données. Elles sont développées dans le but de dépasser les limites des SGBDR en termes de la complexité et des types et formats des données supportées et manipulées. Parmi les bases de données distribuées, il y a CouchDB (orientée document), Cassandra et Hbase (orientée colonne), qui n'imposent pas des règles strictes de schéma comme dans le cas du modèle relationnel.
- **Infrastructure de traitement distribué et d'analyse prédictive** : permet le calcul en parallèle de grands ensembles de données à travers des clusters de machines. L'exemple open source le plus connu des systèmes répartis est celui de Hadoop développé par la fondation Apache. Il s'agit également des outils implémentant des méthodes mathématiques appliqués à l'informatique pour l'analyse des bases de données géantes via des modèles prédictifs de l'apprentissage machine. En plus, Apache Spark est un Framework très performant dédié pour la conception et la création des applications à grande échelle permettant l'analyse prédictive en mémoire.

Le système d'apprentissage en ligne constitue le troisième niveau. Il s'agit de la couche applicative représentée par le système d'apprentissage en ligne contenant, en particulier, des

outils de gestion d'apprentissage (LMS), des systèmes de gestion de contenu (CMS), des environnements d'apprentissage virtuel (VLE), etc. Les informations de cette couche peuvent être des données sous forme de contenu éducatif, d'informations sur le profil d'apprenant ou d'enseignant, des inscriptions aux cours, etc. Ces données jouent un rôle très important et seront utiles pour produire des ressources pédagogiques personnalisées en adaptant les supports d'apprentissages aux besoins de chaque apprenant afin d'offrir une plate-forme d'apprentissage plus appropriée. Pour effectuer ce mécanisme d'adaptation, le système e-learning doit faire appel aux technologies de la couche inférieure (big data) afin d'exploiter des modèles prédictifs avancés en appliquant des algorithmes parallèles de machine learning sur les données d'apprentissage.

La quatrième couche de cette architecture est la couche utilisateur. Généralement, un système e-learning est un système composé de matériel, de logiciel, et d'utilisateur. Le matériel comprend tous les périphériques réseau et informatiques fournit par les services IaaS du cloud. La plate-forme e-learning est le logiciel. Les utilisateurs sont des acteurs qui utilisent le système pour communiquer, stocker et traiter des informations. On distingue trois types d'acteurs. Ce sont l'apprenant, l'enseignant ou tuteur, et l'administrateur du système. En outre, un quatrième acteur peut être ajouté. C'est le data scientist qui est chargé de la gestion et du contrôle de l'environnement distribué big data, ainsi que de l'élaboration et du déploiement des modèles analytiques et de l'implémentation des techniques du machine mining sur des plates-formes réparties big data.

III.2.3 Méthodologie d'intégration des technologies big data dans l'e-learning

Généralement, dans les différents domaines, il est primordial d'avoir une méthodologie claire avant d'aborder la phase de réalisation d'un projet. Une telle méthodologie permet de décrire les lignes directrices et la démarche à suivre pour mener à bien et atteindre l'objectif visé. De même, dans les projets big data, nous devons adresser une méthodologie adaptée au contexte de l'apprentissage en ligne. Pour ce faire, nous identifions les principales étapes du processus big data permettant, à partir des sources de données issues des plateformes d'apprentissage en ligne, de créer quelque chose de nouveau ayant une valeur ajoutée afin d'aider les professionnels de l'enseignement à distance à prendre des bonnes décisions. Cela comprend l'acquisition et la découverte, la préparation, la modélisation, le traitement et la visualisation des résultats de l'analyse des données. Lorsque les données sont efficacement capturées,

traitées et analysées, les professionnels de l'apprentissage en ligne peuvent avoir une compréhension complète de leurs apprenants, ressources pédagogiques, résultats d'évaluation, etc. De cette façon, ils seront capables d'offrir des activités pédagogiques plus personnalisées, de créer des stratégies d'apprentissage plus innovantes, d'augmenter le niveau de satisfaction de l'apprenant, et d'améliorer la qualité des contenus éducatifs.

Nous pouvons résumer notre méthodologie d'utilisation du big data, dans le contexte d'un environnement informatique d'apprentissage humain dédié à l'e-learning, dans la Figure III.7 qui illustre la séquence des différentes étapes respecté pour réussir un projet big data.

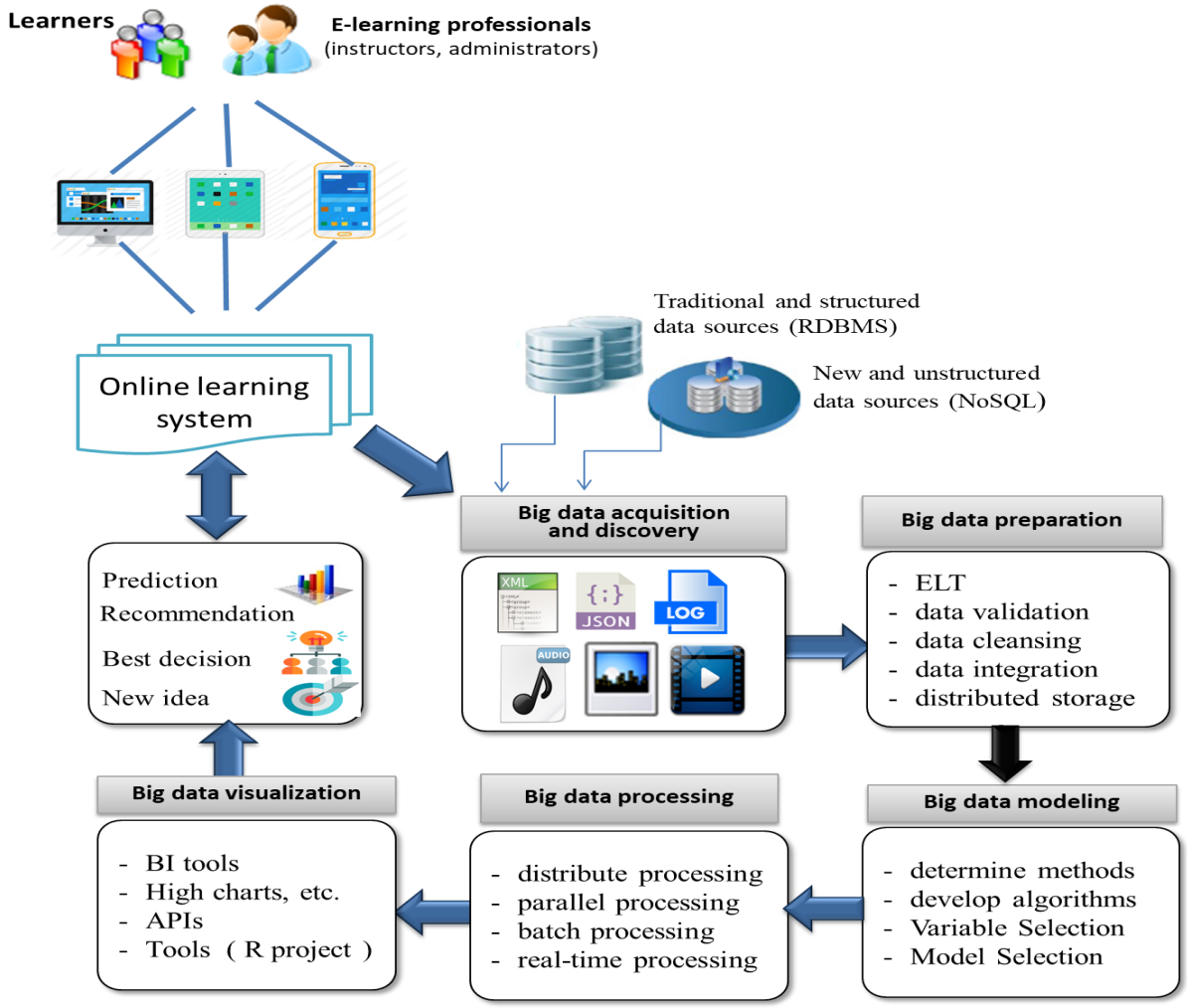


Figure III.7: Méthodologie d'intégration du big data et des systèmes d'apprentissage en ligne

En fait, afin de gérer le volume important des données générées par les plateformes d'apprentissage en ligne, les données elles-mêmes doivent passer par une série de cinq étapes :

III.2.3.1 Phase de découverte et d'acquisition

Il s'agit d'un processus de collecte et de découverte des données produites par les environnements d'enseignement à distance, qui peuvent être des informations sur l'apprenant (profil, connaissances, compétences, etc.), des ressources pédagogiques incluant tous les formats (texte, image, vidéo, page web, etc.), des interactions et des comportements des apprenants avec le système d'apprentissage. Elles proviennent également des activités de collaboration des apprenants avec leurs collègues ou avec leurs enseignants via des réseaux sociaux, des forums et des wikis. La Figure III.8 donne une liste non exhaustive des données générées par un systèmes e-learning.

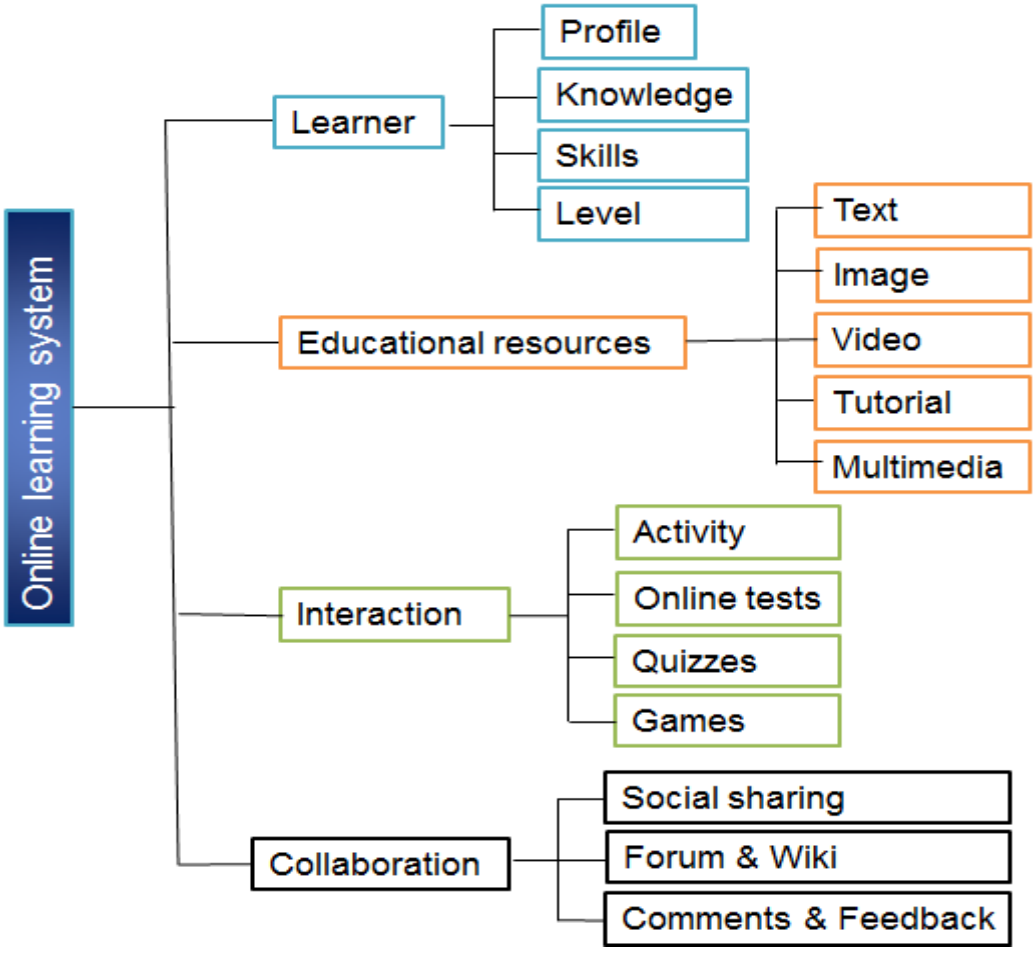


Figure III.8: Acquisition des données dans un système e-learning

III.2.3.2 Phase de préparation

La deuxième phase de notre méthodologie est celle de la préparation des données qui vient juste après la phase d'acquisition et de découverte. Il s'agit d'un ensemble cohérent d'opérations permettant d'extraire, de charger et de transformer (ELT) les données

hétérogènes. En effet, cette phase comprend l'intégration des données générées par les plates-formes e-learning, de les préparer et de les transférer pour être stockées dans un système de fichier distribué ou une base de données NoSQL telle que Cassandra, HBase, etc. La préparation des données est une étape cruciale dans le processus d'analyse, car c'est à ce niveau qu'on doit filtrer les données collectées pour écarter les données inutiles ou bruitées comme la redondance et de ne garder que celles qui sont pertinentes et de bonne qualité, chose qui sera employée par la suite comme entrée du modèle analytique.

III.2.3.3 Phase de modélisation

Elle constitue la troisième phase, et représente la dernière étape avant l'exécution du modèle analytique. Elle vise à spécifier une méthode adéquate pour tirer profit des grands ensembles de données. Pour ce faire, il faut choisir le bon modèle à appliquer sur ces données. En effet, au cours de cette phase, nous devons identifier un bon modèle analytique, déterminer une méthode appropriée à utiliser et choisir ou développer un algorithme adéquat pour implémenter la méthode sélectionnée. Dans ce sens, les technologies big data offrent une implémentation à grande échelle de plusieurs techniques de machine learning ou data mining, y compris : classification, clustering, association rules, regression, collaborative filtering. Le choix d'un modèle dépend du type, de la qualité, et de la taille du volume des données à analyser et des ressources de calcul disponibles. Elle consiste également à explorer les données pour savoir plus sur les relations entre les variables, puis à sélectionner les variables clés les plus appropriés à un tel modèle.

III.2.3.4 Phase de traitement

Le traitement des données constitue la quatrième phase de notre méthodologie. En fait, le big data s'appuye sur une approche de calcul en parallèle pour faire face aux besoins intensifs de traitement et à l'augmentation du volume de données. Les technologies dédiées à la manipulation des données sont transformées chronologiquement en traitement par lots, traitement en temps réel et calcul hybride. Le traitement par lots est une solution au problème du volume, le traitement en temps réel traite le problème de la vitesse et le calcul hybride convient aux deux problèmes. Les systèmes de calcul reparti sont largement développés, principalement pour prendre en charge l'analyse des données gigantesques, notamment Frameworks Hadoop MapReduce[73] et Apache Spark[84].

III.2.3.5 Phase de visualisation des données

Cette phase vise à visualiser et à communiquer clairement et efficacement les résultats de l'analyse des données d'apprentissage en utilisant des outils avancées et très riches permettant de synthétiser les informations du traitement. Cela contient souvent des tableaux, sous forme de représentations graphiques, telles que des courbes, des barres, des secteurs, et des histogrammes. Il s'agit également de décrire un ensemble de techniques, de logiciels et d'utilitaires destinés à aider les professionnels de l'apprentissage en ligne à avoir une vision claire des données énormes générées par les apprenants.

III.2.4 Architecture distribuée du big data pour les systèmes e-learning

Depuis l'avènement de l'internet, des architectures informatiques ont connu un fort développement. Les systèmes big data s'appuient sur une architecture distribuée dans laquelle les ressources de calcul, de stockage, et de mémoire ne se trouvent plus dans un seul serveur, mais ils sont situés dans plusieurs ordinateurs distants interconnectés entre eux au moyen des protocoles réseaux. Ces différentes machines sont configurées pour fonctionner comme une seule unité dotée des puissances de calcul immenses et des capacités de stockage gigantesque. Cette architecture permet la répartition de la charge du travail d'une application sur plusieurs nœuds d'un cluster pour partager ses ressources.

L'architecture proposée du système e-learning basé sur les technologies big data est illustrée dans la Figure III.9. Elle consiste à utiliser la plate-forme Hadoop qui représente le noyau de l'infrastructure des solutions big data open source, et intègre un système de fichiers distribués (HDFS) et plusieurs outils tels qu'Apache Sqoop et Flume. La figure III.9 montre également comment ces technologies peuvent être combinées et communiquer en échangeant les données lors de l'exécution d'une application big data. Ce mécanisme permet d'obtenir les meilleurs résultats par rapport à l'approche de calcul classique centralisé. L'approche de décentralisation du traitement intensif sur des unités de calcul réparties permet d'améliorer grandement le temps de réponse en supportant une montée en charge importante.

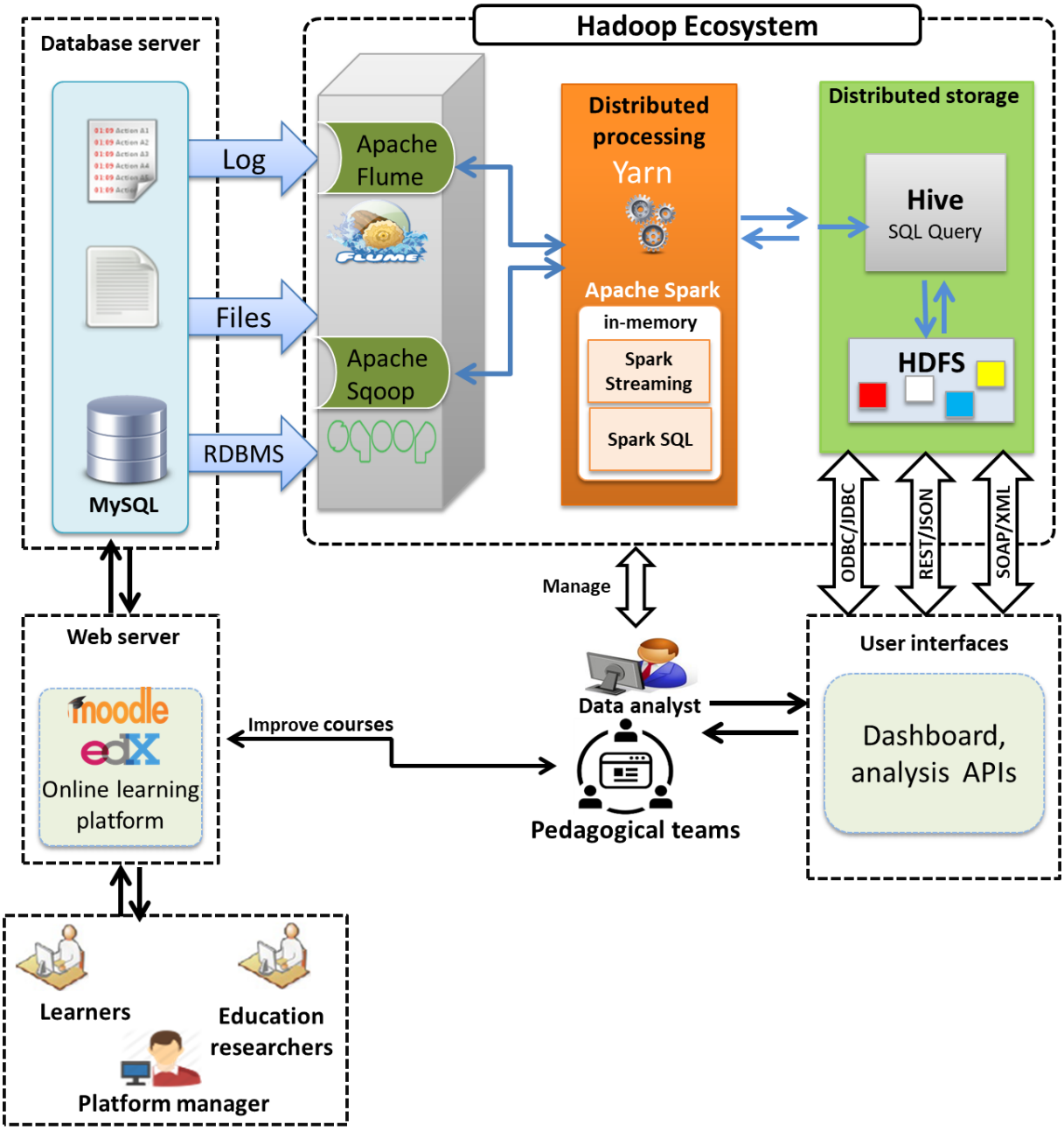


Figure III.9: Architecture du big data pour un système d'apprentissage en ligne

Le processus d'intégration de l'environnement big data et des systèmes d'apprentissage en ligne, peut être résumé comme suit :

Au début, nous devons charger les données produites par Moodle, EdX, Blackboard ou toute autre plate-forme d'enseignement en ligne. Ces données sont de trois types : elles peuvent être des données structurées, des données semi-structurées (fichiers journaux et des documents XML) et des données non structurées (PDF, audio, vidéo et image). Pour cela, nous avons

utilisé des outils big data facilitant le processus de chargement. En effet, nous avons employé Apache Flume pour charger les données à partir des fichiers journaux. Pour les données stockées dans la base des données relationnelle, nous avons choisi de travailler avec Apache Sqoop pour transférer (importer et exporter) les données structurées entre Apache Hadoop et les bases de données relationnelles.

Ensuite, les données chargées par Flume ou Sqoop sont traitées de manière distribuée à l'aide du Framework Spark et l'infrastructure Hadoop. Une fois soumise au cluster, une application en parallèle a besoin des ressources de calcul et de stockage suffisantes pour s'exécuter. Pour ce faire, nous avons choisi d'exploiter un outil big data très performant dédié à la gestion et à la surveillance des traitements sur des groupements de machines ou cluster. Il s'agit d'un gestionnaire de ressources, appelé Yarn, qui s'occupe de la planification et du contrôle des Jobs d'une manière répartie en allouant les ressources nécessaires pour le traitement de l'ensemble des tâches d'une application. En fait, Nous avons deux scénarios. Premièrement, les données en streaming (flux continu de données) sont traitées avec la bibliothèque Spark Streaming. Deuxièmement, concernant les données stockées dans une base de données relationnelle, elles doivent être gérées et manipulées avec le module Spark SQL, spécialement conçu pour le traitement des données structurées. Ainsi, Spark, tout comme un Framework de calcul distribué, se connectera à Hadoop HDFS pour stocker les données sur un cluster des machines. Afin de faciliter la lecture et l'écriture sur HDFS, nous avons utilisé Apache Hive comme data warehouse (entrepôt de données).

Enfin, grâces aux techniques de visualisation des résultats et de traitement des données d'apprentissage, les professionnels de l’e-learning peuvent avoir une vision claire sur toutes les données générées par les apprenants et de leurs interactions avec le système d'apprentissage. Ils pourront ainsi identifier les besoins des apprenants et proposer les ressources pédagogiques les mieux adaptées à chaque situation d’apprentissage. De plus, grâce à l’analyse des données massives, les équipes pédagogiques seront capables de prendre de bonnes décisions, d’améliorer leurs stratégies d’apprentissage, et de prédire le comportement futur des apprenants.

III.2.5 Infrastructure répartie du cluster Hadoop

Afin de donner une vision plus claire de notre architecture, la Figure III.10 illustre l'infrastructure d'un petit cluster. Cette infrastructure qui est entièrement distribuée est mise en

place par l'intermédiaire de la technologie Hadoop. En effet, nous avons installé un cluster Hadoop composé de trois nœuds connectés entre eux par un réseau virtuel. Chaque machine du cluster a deux rôles. Dans un premier temps, elle effectue et gère les traitements via le composant NodeManger, ensuite, elle stocke les données traitées via le composant DataNode.

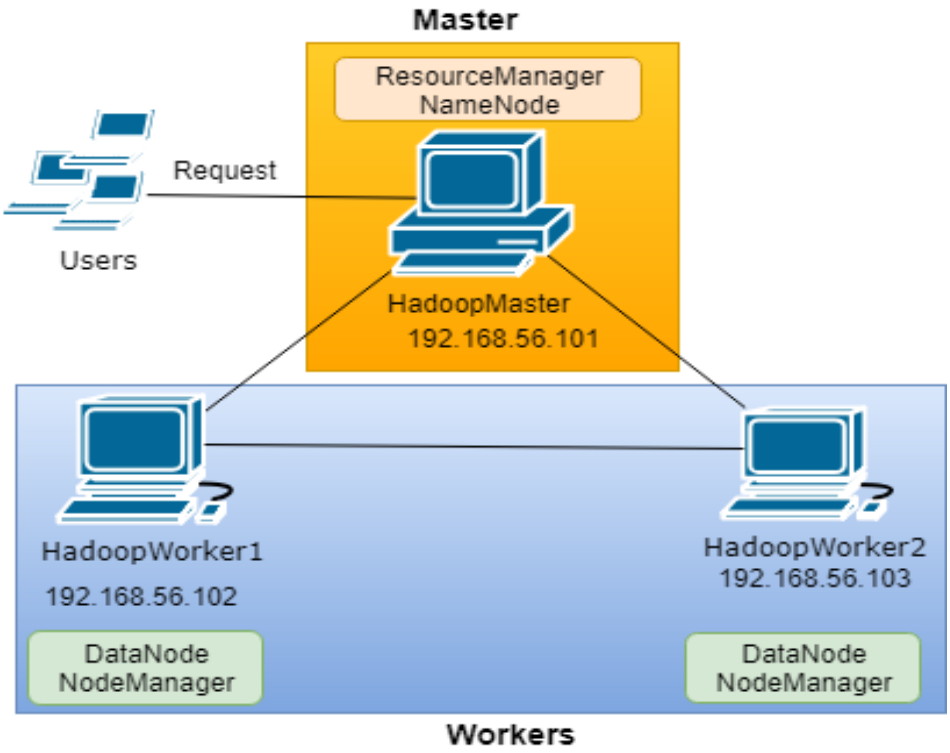


Figure III.10: L'infrastructure distribuée d'un cluster Hadoop

La configuration de Hadoop multi-nœuds est faite dans un environnement réparti composé de trois nœuds. La machine ayant l'adresse IP : 192.168.56.101 est configurée pour fonctionner en tant que master Hadoop, qui contient deux composants Hadoop, à savoir ResourceManager et NameNode. Le nœud master exécute les processus Hadoop pour gérer et coordonner les tâches et les services du cluster. Dans le cas du manque des ressources matérielles, ce nœud peut être configurée à la fois en tant que master et en tant que worker. En effet, la machine virtuelle identifiée par l'adresse IP 192.168.56.101 fonctionne comme un master et worker en même temps. Les autres machines du cluster sont des workers. Les nœuds workers sont chargés de l'exécution des processus ou des tâches élémentaires de l'application parallèle big data. Ils fournissent aussi des ressources au cluster pour effectuer le traitement des données assignées par le master. Comme le montre la Figure III.10, ils ont respectivement les adresses

IP 192.168.56.102 et 192.168.56.103 et se communiquent entre eux et avec le master via le protocole de communication réseau SSH.

L'architecture proposée est basée sur les bibliothèques logicielles de Hadoop. En effet, si nous voulons augmenter les capacités de calcul et/ou de stockage, il faut simplement ajouter un nouvel hôte au cluster. De ce fait, le Framework Hadoop offre une infrastructure évolutive très élastique. Il est principalement conçu pour passer d'une machine à des dizaines, centaines, voire des milliers de machines. Cet environnement distribué de cluster big data a été construit en utilisant la version « Apache Hadoop 3.1.0 ». Pour le traitement et l'analyse des données, Spark 2.3.0 est choisi comme Framework de calcul en parallèle permettant une vitesse d'exécution 10 fois plus rapide que le modèle MapReduce en effectuant des traitements en mémoire. Pour la gestion des ressources et la planification des tâches, nous avons configuré Hadoop pour exécuter un Job Spark parallèlement à l'aide du gestionnaire des ressources YARN. Ce dernier possède deux composants essentiels. Premièrement, le ResourceManager qui est installé coté master et qui a pour rôle d'intercepter les requêtes d'un utilisateur, d'allouer les ressources nécessaires, et de distribuer le traitement des tâches élémentaires sur différents nœuds workers. Deuxièmement, sur chaque nœud worker, on trouve le NodeManager ayant comme but de surveiller et de rapporter l'avancement de l'exécution des tâches au master.

Cette architecture réparie a fait l'objet d'implémentation dans le chapitre consacré au développement d'une application big data à grande échelle. Notre proposition d'une application répartie consiste à développer un système de recommandation de cours afin d'aider les apprenants à trouver les ressources pédagogiques les plus appropriés à leurs profils et leurs préférences. Cette application a été exécutée dans un environnement virtuel, entièrement distribué, qui combine une panoplie de technologies big data dans le but fournir un moteur de recommandation évolutif, performant, et capable de gérer des quantités massives de données d'apprentissage et de supporter une montée en charge importante dans un temps record.

III.2.6 Bénéfices et avantages de l'intégration du big data dans les EIAH

De nombreuses raisons sont derrière le choix de l'adoption du big data dans le domaine de l'éducation en générale, et dans l'enseignement à distance en particulier. En premier lieu, cela permet aux professionnels de l'e-learning de personnaliser leurs plates-formes d'apprentissage

en ligne en proposant des cours et des ressources pédagogiques plus adaptés au profile et aux préférences de chaque apprenant. Lorsque les données provenant d'un système e-learning sont capturées, traitées et analysées efficacement, les enseignants et les équipes pédagogiques sont en mesure d'obtenir une meilleure compréhension de leurs apprenants, des contenus pédagogiques, des résultats d'évaluation, etc., ce qui peut conduire à des stratégies d'enseignement plus pertinentes, augmenter le niveau de satisfaction des apprenants, améliorer la qualité des supports d'apprentissage, réduire les coûts et offrir un service éducatif de haute qualité. En outre, grâce aux riches technologies de l'écosystème big data, les professionnels de l'e-learning ont désormais plein de possibilités pour concevoir des cours plus personnalisés et plus ciblés en prenant en compte les exigences et les besoins de chaque situation d'apprentissage tout en analysant le comportement individuel et collectif des apprenants et de leurs interactions.

De surcroît, les outils du big data offrent des implémentations à grande échelle des modèles et des techniques analytiques prédictifs. En se basant sur ces techniques, nous pouvons interpréter en temps presque réel les différentes données des apprenants afin d'anticiper leurs prochaines actions. Par exemple, en utilisant cette nouvelle génération de technologies, on sera capable de développer rapidement un système de recommandation permettant de prédire les cours les plus appropriés à chaque apprenant.

Le big data a la capacité de changer la façon dont les universités offre leurs formations par internet. Il est aussi capable d'améliorer considérablement le futur de l'e-learning. En effet, l'environnement du big data offre plusieurs avantages aux professionnels de l'e-learning :

Stockage distribué

Les systèmes d'apprentissage en ligne utilisent très souvent des contenus multimédias (vidéos, images, sons, etc.) pour faire face à cette variété de données volumineuses, le big data peut effectivement stocker ces volumes importants, supporter et manipuler cette grande variété de données. Pour y parvenir, le big data propose une large panoplie de solutions technologiques innovantes incarnées par la nouvelle famille des systèmes de gestion des bases de données NoSQL. Par ailleurs, il offre plusieurs systèmes de stockage réparti, le plus populaire est celui de HDFS qui permet de gérer des données hétérogènes issues de

différentes sources (profil de l'apprenant, stratégie d'apprentissage, ressources pédagogiques, fichier journal, etc.).

Traitement en parallèle

Il existe beaucoup d'outils puissants du big data, y compris Hadoop MapReduce, Apache Spark, permettant le calcul en parallèle sur un environnement complétement distribué. Ces technologies offrent la possibilité d'exécuter des traitements par lots en accélérant la vitesse de manipulation des grosses quantités de données d'apprentissage grâce à l'approche de décentralisation du calcul sur des big computer (clusters des ordinateurs). Également, à l'aide des outils du big data qui permettent un traitement massivement parallèle et des méthodes d'analyse avancées, les professionnels de l'e-learning peuvent désormais évaluer en temps réel l'efficacité de leurs stratégies d'apprentissage et identifier les besoins de chaque apprenant pour offrir des ressources pédagogiques adaptées à chaque situation d'apprentissage.

Feedback rapide

Actuellement, grâce aux technologies big data les systèmes de l'e-learning peuvent fournir des feedbacks rapides aux questions des apprenants à l'aide des forums interactifs, outils des discussions avec les enseignants, etc. Les responsables de formations peuvent collecter, traiter et analyser les données d'apprentissage afin d'identifier les lacunes de chaque apprenant. En fait, les techniques avancées du big data permettent de cibler les modules ou les cours qui demandent des changements et des modifications en proposant un enseignement plus adaptatif. Par exemple, si plusieurs apprenants prennent trop de temps pour terminer un module particulier, cela signifie probablement que ce module a besoin d'être amélioré afin de le rendre plus adapté aux exigences d'apprentissage. En outre, si l'analyse des données recueillies indique que la majorité des étudiants ont des difficultés avec un module bien déterminé. Dans ce cas, le système d'apprentissage apporte immédiatement des modifications nécessaires pour y remédier ou proposer automatiquement les éléments d'aide afin de faciliter la complexité d'une difficulté donnée[131].

Suivi des apprenants

A l'ère du big data et de l'apprentissage automatique, les systèmes e-learning deviennent plus intelligents en offrant aux professionnels de l'apprentissage en ligne la possibilité de suivre les interactions des apprenants tout au long de leurs parcours d'apprentissage. En effet, ils peuvent cerner les difficultés, analyser les résultats, savoir les préférences et le temps consacré pour compléter chaque cours, etc. En plus, le big data offre des méthodes et des techniques avancées capables d'analyser le comportement d'un groupe d'apprenants, en se basant sur l'historique des activités des cours suivis, pour trouver des similitudes dans leurs comportements ce qui servira à faire des prédictions et de proposer des recommandations en termes de cours et de ressources pédagogiques les plus appropriées pour chaque apprenant. De ce fait, les apprenants bénéficient davantage des scénarios pédagogiques très riches et d'un enseignement interactif de qualité leur permettant de maitriser leurs connaissances et d'améliorer leurs compétences.

Par ailleurs, les enseignants peuvent préparer et corriger des tests en ligne, profiter des outils de collaboration en ligne en vue d'impliquer des apprenants dans des travaux pédagogiques en équipe et de les aider à réaliser des projets en groupe. En plus, le big data favorise l'apprentissage participatif en créant un environnement d'échange et de partage des connaissances entre des participants d'un cours. Ces outils se diversifient entre forum, réseaux sociaux, messagerie instantanée, visioconférence, etc. De cette manière, en utilisant les technologies big data on peut assurer une meilleure interaction entre les différents acteurs au cours du processus d'apprentissage.

Offre des cours personnalisés

Le fait de savoir comment l'apprenant acquiert les cours, à quelle vitesse et de déterminer les problèmes qui peuvent existés au sein d'une stratégie d'apprentissage, permet aux enseignants et aux équipes pédagogiques de concevoir des cours plus personnalisés et plus adaptés au profile de chaque apprenant. Les modules peuvent ainsi être préparés pour répondre au mieux aux besoins des apprenants. Cette adaptation des contenus éducatifs au style d'apprentissage et au niveau cognitif de l'apprenant permettre aux établissements d'enseignement et/ou aux universités d'améliorer grandement l'efficience de leurs plates-formes d'apprentissage et d'offrir une formation en ligne de haute qualité.

Comprendre le processus d'apprentissage

Le big data propose des technologies innovantes pour avoir une compréhension approfondie du processus d'apprentissage. En utilisant le big data, les professionnels de l'e-learning peuvent savoir la manière dont les apprenants interagissent avec les cours proposés, à quelle heure, et de comprendre leurs comportements avec le système d'apprentissage. Le big data a le potentiel de classer, filtrer et interpréter les données des interactions entre l'apprenant et le système d'apprentissage, ce qui permet de sélectionner les ressources éducatives les plus convenables pour chaque situation d'apprentissage. Ainsi, par l'intermédiaire des méthodes et des modèles prédictifs de l'apprentissage machine, on peut anticiper les prochaines actions des apprenants en leur recommandant des activités pédagogiques plus appropriées à leurs profils et à leurs niveaux cognitifs.

III.2.7 Conclusion

Ce chapitre vise à intégrer la nouvelle génération des technologies de l'information et de la communication, en particulier l'écosystème big data, dans des environnements informatiques d'apprentissage humain dédiés à l'apprentissage en ligne. En plus, nous avons proposé un modèle de l'e-learning basé sur les services du cloud en incorporant les systèmes du big data. Puis, nous avons décrit la méthodologie d'intégration des technologies big data dans les plates-formes apprentissage en ligne. Ensuite, nous avons proposé une architecture e-learning distribuée combinant diverses solutions technologiques du traitement en parallèle et du stockage décentralisé en se basant sur un environnement entièrement réparti. Par ailleurs, ce chapitre a présenté les avantages et les bénéfices de l'intégration des technologies de pointe du big data dans l'e-learning en termes du traitement intensif et de l'analyse prédictive des données.

Dans ce qui suit, nous nous nous intéresserons à la mise en application des technologies big data dans le contexte des EIAH en utilisant les données réelles d'une plate-forme e-learning opérationnelle de l'Ecole supérieur de Technologie (EST) de Fès. Plus spécifiquement, nous allons travailler avec un cluster de nœuds configurés à l'aide la plate-forme open source Hadoop, et le Framework Spark sera utilisé comme un modèle de développement à grande échelle d'un système intelligent de recommandation des cours.

IV. CHAPITRE IV : Mise en œuvre des technologies big data dans un EIAH

Système de recommandation de cours

- Méthode
- Algorithmes
- Application et résultats

Système de recommandation de cours à grande échelle

- Algorithme FP-growth Parallèle
- Technologies mises en œuvre
- Architecture répartie Hadoop
- Préparation et configuration de l'environnement big data
- Application distribuée Spark
- Résultats
- Performance du système de recommandation de cours
- Outils de monitoring ou supervision

IV

IV.1 Système de recommandation de cours

IV.1.1 Introduction

Pour aider l'esprit humain dans son processus de sélection et de filtrage d'information, des systèmes de recommandation ont vu le jour dans la dernière décennie du vingtième siècle. Par la suite, avec l'essor de l'internet et ses applications, il y a eu un engouement pour les systèmes de recommandation qui se sont développés dans de multiples domaines d'applications, comme l'e-commerce, le tourisme, les articles d'actualités, les Films (Netflix [132]), etc. Ou chaque domaine présente différents problèmes qui requièrent différentes solutions. En fait, un système de recommandation a pour but de proposer à un utilisateur les éléments les plus susceptibles de l'intéresser. En d'autres termes, ce type de système tente de suggérer à un utilisateur des objets qui répondent au mieux à ses préférences et ses besoins. Il peut être représenté comme une unité de prédiction du comportement d'un utilisateur en anticipant ses prochaines actions.

Généralement, une recommandation peut être comparée à un dialogue entre une personne experte d'un domaine et une autre désireuse d'acquérir des informations dans ce domaine. Dans notre cas, un système de recommandation de cours va pouvoir, en fonction des préférences d'un de ses apprenants, proposer une liste des ressources pédagogiques à ce dernier. Pour ce faire, un système de recommandation doit avoir une base des connaissances contenant différentes informations sur des items ou des éléments à recommander. Après, vient le rôle du système de recommandation qui consiste à analyser et à effectuer des associations entre des items suivant différents critères caractérisés par le profil d'apprenant. Il serait alors en mesure de suggérer des activités pédagogiques à un apprenant, qui ont été suivies par d'autres apprenants similaires. De cette manière, un système de recommandation peut être vu comme un moyen efficace pour aider l'apprenant à choisir et à cibler un certain nombre de cours parmi une masse de ressources très importante où il n'a pas assez de connaissances de ces derniers[133].

Plusieurs facteurs entrent en considération afin de construire un moteur de recommandation : les données à recommander, les informations disponibles et bien évidemment l'objectif visé. Généralement, il y trois approches pour construire un système de recommandation : le filtrage

basé sur le contenu, le filtrage collaboratif et le modèle hybride combinant les deux premiers type de filtrage. La Figure VI.1 monte un schéma synthétique des différents types de systèmes de recommandation.

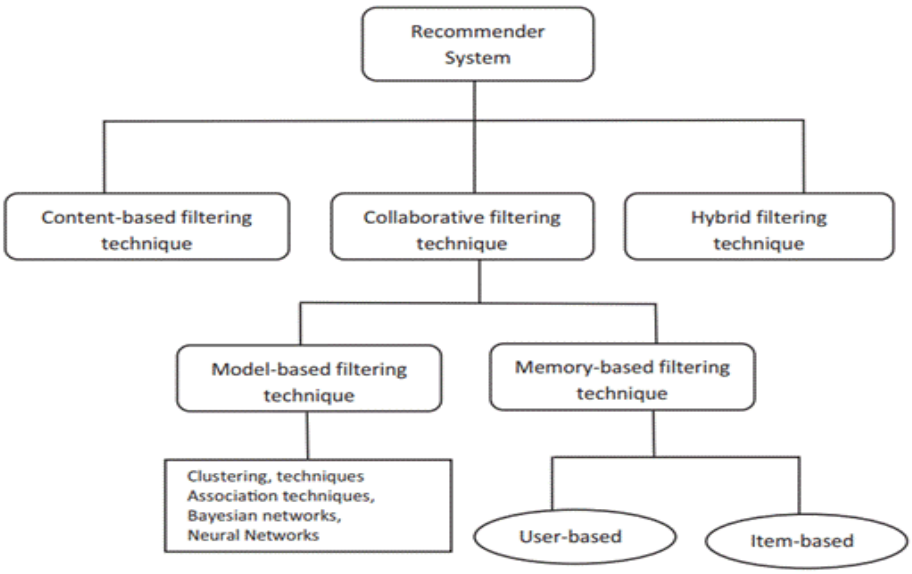


Figure IV.1: Différents types de systèmes de recommandation [134]

Dans un système de recommandation basé sur le contenu, le choix des items à recommander s'appuie sur une comparaison des thèmes et des caractéristiques abordées dans ces items par rapport à celles définies dans le profil utilisateur. Le mécanisme de recommandation de ce type de systèmes est effectué en identifiant les ressources similaires à celles appréciées par un utilisateur en fonction de leur contenu. Cette technique de recommandation consiste à analyser le contenu des ressources à recommander ou des descriptions de celles-ci afin de déterminer quelles ressources sont susceptibles d'être les plus utiles ou intéressantes pour un utilisateur donné. On distingue deux types de recommandation basée sur le contenu : recommandation basée sur les mots clefs et celle basée sur l'analyse sémantique.

Contrairement à l'approche de filtrage basée sur le contenu qui s'intéresse à un seul utilisateur, les recommandations par filtrage collaboratif utilisent les préférences de plusieurs utilisateurs. Cette méthode essaye de former un groupe d'utilisateurs ayant les mêmes préférences. Dans un système de recommandation collaboratif, l'idée est non plus de s'intéresser spécifiquement au nouvel item qui serait susceptible de plaire à l'utilisateur mais de regarder quels items ont appréciés les utilisateurs proches de l'utilisateur courant. Ainsi, il

n'y a aucune analyse du sujet ou du contenu des objets à recommander. Ce type de système est très efficace au cas où le contenu des objets est complexe ou difficilement définissable (exp. ressources pédagogiques, films).

La troisième approche permettant de développer des systèmes de recommandation est l'approche hybride. Celle-ci peut être définie comme étant une combinaison de deux ou plusieurs méthodes différentes. Par exemple, un tel système peut utiliser à la fois des techniques de filtrage collaboratif et celles du filtrage basé sur le contenu, etc. Le point fort de la stratégie hybride est sa capacité à présenter les avantages des approches qui la compose tout en limitant leurs inconvénients. L'utilisation des approches hybrides permet d'améliorer la pertinence des résultats des systèmes de filtrage en palliant certaines limites de types des filtrages présentés précédemment. Par exemple, le démarrage à froid au niveau des items des approches collaboratives, peut bénéficier des avantages des techniques du filtrage basé sur le contenu afin de construire des profils initiaux pour les utilisateurs qui viennent de rejoindre un système pour la première fois.

En fait, les techniques du filtrage que ce soit collaboratif ou basé sur le contenu ne prennent pas en compte un bon nombre d'informations utiles pour expliquer le jugement qu'effectue l'utilisateur. Ceci requiert la prise en compte de différentes sources d'informations sur l'utilisateur (ses préférences, son comportement, son réseau social, ...) et leurs corrélations avec le contenu des documents qui lui sont proposés. Ainsi, nous avons assisté, ces dix dernières années, à la naissance de plusieurs nouvelles familles de systèmes de filtrage d'information, comme les systèmes basés sur la localisation géographique, les systèmes de filtrage d'information à base d'ontologies, les systèmes d'annotations collaboratives, les systèmes de filtrage d'informations à base de réseaux sociaux, etc.

Notre système de recommandation de cours se base essentiellement sur les techniques de l'apprentissage automatique. En effet, ce dernier est employé dans le but de vulgariser un système adaptatif capable d'analyser des situations d'apprentissage ou des comportements des apprenants afin d'en tirer des recommandations. Ce système est ainsi capable d'apprendre par lui-même et d'évoluer pour proposer des ressources pédagogiques adaptées aux attentes des apprenants. Les méthodes d'apprentissage automatique permettent donc de créer un système qui s'adapte aux divers contextes, tels que les jeux vidéo, la santé numérique [135]. Notre

objectif est de concevoir un système adaptatif pour un environnement d'apprentissage en ligne. Ce système consiste à adapter des décisions pédagogiques aux préférences et besoins individuels de chaque apprenant. Pour ce faire, notre système de recommandation fait appel à des techniques d'apprentissage automatique afin d'extraire des connaissances via l'analyse des races d'apprentissage. L'objectif est donc de fournir à l'apprenant un schéma éducatif et pédagogique personnalisé en fonction de son profil, ses connaissances ou encore ses aptitudes.

Dans nos travaux de recherche, nous avons choisi d'utiliser l'approche de filtrage collaboratif dans le développement du système de recommandation de cours. Le choix d'une méthode de recommandation dépend, dans un premier lieu, des données disponibles sur les éléments à recommander et les profils d'utilisateurs. Un système de recommandation doit, avant tout, s'adapter aux données sur lesquelles il va s'appuyer pour proposer des recommandations pertinentes à un utilisateur. En fait, le filtrage collaboratif est à la base de la recommandation, les méthodes de filtrage par le contenu étant plutôt liées aux systèmes de recherche d'informations utilisés par des moteurs de recherche comme Google. En outre, Les systèmes de recommandation utilisant le filtrage par le contenu sont assez rares et très anciens. En fait, la plupart des systèmes de recommandation actuels utilisent le filtrage collaboratif. Citons par exemple le géant du e-commerce Amazon [136] qui propose un système de recommandation d'un large panel de produits (livres, matériel électronique, etc.) sur son site internet par l'intermédiaire des méthodes collaboratives. Par ailleurs, un système de recommandation basé sur le contenu a un niveau de complexité plus élevé par rapport à la technique de filtrage collaboratif. Cette complexité dépend principalement des problèmes liés à l'extraction et l'analyse des données sémantiques sur les items. En effet, s'il s'avère difficile de dégager automatiquement quelques propriétés sur les items, les algorithmes de recommandation par le contenu deviennent inutiles. Par exemple, la qualité et la diversité des données multimédia (image, vidéo, audio ou page web) d'une plate-forme d'apprentissage peuvent représenter une information importante pour certains apprenants, cette information est difficile à extraire automatiquement à partir d'un cours en ligne. En outre le filtrage collaboratif est basé sur l'analyse de l'historique des interactions d'une communauté d'utilisateurs avec l'interface d'un système quelconque. Pour produire des recommandations il suffit juste de chercher la similitude entre des actions et interactions homme-machine, du point de vue préférences,

choix et comportements d'utilisateurs, afin de produire des items pertinents adaptés au profil et besoins de chaque utilisateur. Dans le cas d'un environnement e-learning, la technique du filtrage collaboratif consiste à analyser les traces des activités d'apprentissage des apprenants, puis suggérer des cours appropriés pour chaque apprenant en fonction du profil, du style d'apprentissage et de niveau cognitif de celui-ci.

IV.1.2 Méthode

Dans cette section, nous nous concentrons sur les concepts de base des règles d'association utilisées pour développer le système de recommandation pour un environnement d'apprentissage en ligne. Nous présentons également l'algorithme FP-growth, considéré comme l'un des algorithmes les plus couramment utilisés et le plus performant pour extraire les ensembles d'items fréquents, sans génération des éléments candidats, à partir des inscriptions historiques des apprenants au cours du processus d'apprentissage.

IV.1.2.1 Filtrage Collaboratif (Collaborative Filtering)

Le filtrage collaboratif [137] est une approche très populaire de l'apprentissage machine. Il est couramment utilisé dans le développement des systèmes de recommandation. Ces systèmes ont pour objectif de proposer à l'utilisateur des objets (cours, films, etc.) qui répondent aux mieux à ses intérêts et à ses préférences. En d'autres termes, Il consiste à prédire les items (éléments) que des utilisateurs vont aimer ou apprécier dans le futur. En effet, l'approche du filtrage collaboratif repose sur l'analyse et l'exploitation des données historiques d'un ensemble d'utilisateurs. Par exemple si un étudiant est intéressé par le cours "systèmes d'exploitation", et la plupart des étudiants qui se sont inscrits à ce cours suivent aussi le cours "Unix", alors il est très probable que cet étudiant aime aussi de suivre cet item et il est recommandé de lui suggérer le cours "Unix".

Dans un système de recommandation collaboratif, l'idée est de regarder quels items ont été appréciés par les utilisateurs qui ont auparavant fait des choix similaires à ceux de l'utilisateur courant (Figure VI.2). Cette technique cherche à trouver des ressemblances dans les comportements des utilisateurs. Ainsi, il n'y a aucune analyse du sujet ou du contenu des objets à recommander. Ce type de système est très efficace au cas où le contenu des objets est complexe et difficile à analyser (ex. ressources pédagogiques).

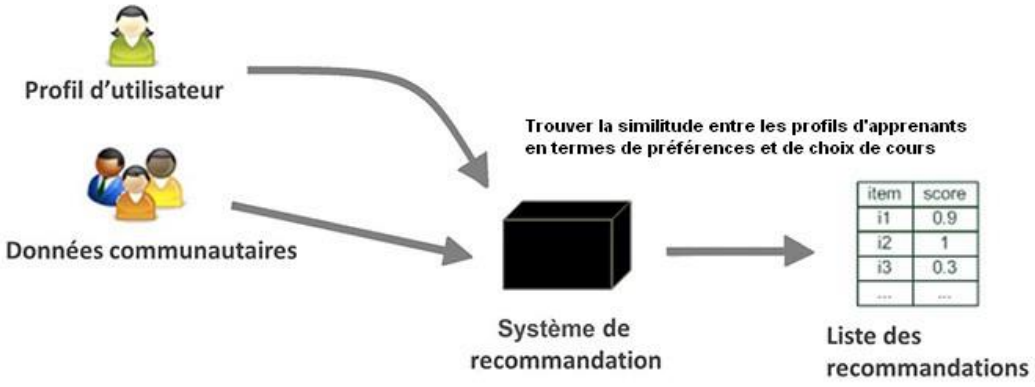


Figure IV.2: Filtrage collaboratif [138]

Dans le contexte de l'e-learning, le filtrage collaboratif se fonde sur la recherche des similitudes entre les inscriptions des apprenants dans des cours. Pour pouvoir proposer des recommandations a un apprenant donné, un système de recommandation a besoin dans un premier temps d'identifier cet apprenant en analysant ses actions et ses interactions avec le système d'apprentissage. Cette approche s'appuie sur l'analyse des traces des utilisateurs et constitue le principe sur lequel nous développons, dans ce chapitre, notre système de recommandation de cours. En fait, notre solution de recommandation compare les profils des apprenants ainsi que les cours qu'ils ont suivi afin de leur proposer des documents proches de leurs attentes.

En général, les systèmes de recommandation collaboratifs comprennent trois étapes principales; premièrement, collecter des données à partir d'une base de données contenant une volumétrie importante de transactions des utilisateurs. Ensuite, trouvez des similitudes entre les comportements des utilisateurs et les items les plus fréquents. Enfin, recommander un catalogue des cours les plus appropriés en fonction du profil et des connaissances de chaque utilisateur.

Le principe du système de recommandation de cours est de produire un assistant ou un guide d'apprentissage en ligne pour un apprenant en ciblant dans un premier temps son choix des ressources pédagogiques et en lui recommandant dans un second temps des cours relatifs à son profil. La base de connaissances est construite à partir d'informations suivantes :

1. Les informations personnelles relatives à un utilisateur (Id, nom, prénom, ...)
2. L'historique de la liste des activités pédagogiques de l'ensemble des apprenants.

On distingue deux grandes familles de recommandation collaborative. D'abord, on trouve des systèmes basés sur l'approche passive dans laquelle l'apprenant n'est pas demandé à évaluer en donnant une note ou son avis à une ressource ou un item sélectionné. Contrairement au premier type de recommandation collaborative, il y a des systèmes qui s'appuient sur l'approche active où on a besoin de la contribution des utilisateurs en les interrogeant sur la qualité d'un élément choisi.

Un système de recommandation collaboratif offre plusieurs avantages :

- Il permet des prédictions précises en utilisant les informations historiques de l'ensemble des utilisateurs d'un système.
- Il peut être appliqué à un très grand nombre d'applications
- Il supporte des volumes massifs de données en fournissant des résultats en temps réel.
- Il a une grande flexibilité et agilité d'adaptation à tous les secteurs.

Par ailleurs, le filtrage collaboratif est une technique simple à utiliser et à implémenter et qui permet de concevoir des systèmes personnalisables capables de s'adapter aux divers contextes (e-commerce, e-learning, etc.)

IV.1.2.2 Association Rules

Dans le domaine de l'apprentissage machine ou automatique, Association Rules, traduit par règles d'association, est une méthode d'apprentissage non supervisée, largement utilisée dans de nombreux domaines, notamment les moteurs de recommandation, l'analyse de la transactions d'achat en ligne et l'analyse du flux de clics sur plusieurs pages Web [139]. Son objectif est de découvrir des relations entre des variables dans un ensemble de données, qu'on appellera transactions, sous la forme de règles d'association intéressantes. En d'autres termes, cette méthode consiste à détecter des associations entre des données stockées dans une base de données géante. La puissance de la méthode de règles d'association réside dans sa capacité à extraire des structures cachées dans une quantité massive de données. Il s'agit d'un ensemble de techniques exploratoires puissantes largement employées dans de nombreux secteurs mais également à des fins de recherche scientifique. L'application la plus populaire utilisant les règles d'association est celle qui concerne l'analyse des habitudes de consommation. Dans le domaine d'e-commerce, cette technique permettent de savoir la relation entre les choix des

clients, comme les clients qui ont acheté le produit A ont également acheté avec une grande propbabilité le produit B ou le produit C.

Le terme règles d'association est souvent appelé application d'analyse du panier de consommation. Ce terme a été utilisé la première fois en 1993 par Agrawal et al. [108], [109] afin de trouver des relations utiles entre les produits dans une base de données volumineuse des transactions client où chaque transaction comprend des articles achetés par un client.

Soit $C = \{c_1, c_2, c_3, \ldots, c_n\}$ l'ensemble d'attributs binaires, appelés item ou éléments. Ces derniers sont représentés dans le contexte e-learning par des cours ou ressources pédagogique. Et considérant $L = \{l_1, l_2, l_3, \ldots, l_n\}$ un ensemble d'utilisateurs ou d'apprenants. Soit $T = \{t_1, t_2, t_3, \ldots, t_n\}$ une base de données des transactions. Chaque transaction t est représentée comme un vecteur binaire, avec t [k] = 1 si la transaction t contient l'item c_i sinon t[k] = 0. Un apprenant l_i s'inscrit aux k cours, où k est un sous-ensemble des items sélectionnés à partir de l'ensemble des cours disponibles dans la base de données C. Une règle d'association est définie sous la forme d'une implication $X \Longrightarrow Y$ où $X, Y \subseteq C$. et $X \cap Y = \emptyset$ (X et Y sont des items disjoints). Cette implication se traduit par « si X alors Y » comme dans l'exemple : *si le cours X est suivi par l'apprenant l_i, le cours Y est probablement suivi par ce dernier avec une probabilité élevée.* L'ensemble d'attributs X est appelé antécédent ou côté gauche, left-hand-side (LHS) de la règle. D'autre part, l'ensemble d'attributs Y est appelé conséquence ou côté droit, right-hand-side (RHS) de la règle[140].

Généralement, la technique des règles d'association produit un nombre important des relations, mais pour sélectionner des règles intéressantes dans l'ensemble des relations générées, elle possède deux critères importants pour déterminer la qualité d'une règle en mesurant sa force, à savoir : les seuils minimaux de support et de confidence.

Le support est le pourcentage (%) de transactions contenant l'ensemble d'items X, tandis que la confidence est définie comme le pourcentage (%) de transactions contenant X, qui contiennent également Y. La définition formelle de la confidence est la suivante : conf ($X \Longrightarrow Y$) = (supp ($X \cup Y$)) / (supp (X)). Par conséquent, une règle d'association $X \Longrightarrow Y$ forte devrait satisfaire : supp ($X \cup Y$) $\geq \sigma$ et conf ($X \Longrightarrow Y$) $\geq \delta$, où σ et δ représentent respectivement le seuil minimal de support et de confidence.

Dans le cadre de nos recherches, nous avons appliqué la technique des règles d'association dans le contexte d'un environnement informatique pour l'apprentissage humain dédié à l'e-learning. Par conséquent, une transaction dans notre cas est représentée par le profil d'apprenant. De même, les items sont remplacés par l'ensemble des ressources disponibles dans la base de données. Une transaction est représentée par des inscriptions d'un apprenant dans un certain nombre de cours durant son cursus d'apprentissage. Nous pouvons donc définir le support et la confidence comme suit :

$$supp(X \Longrightarrow Y) = \frac{\text{nombre d'apprenants inscrits aux X et Y}}{\text{nombre total d'inscriptions d'apprenants}} \quad (5)$$

$$conf(X \Longrightarrow Y) = \frac{\text{nombre d'apprenants inscrits aux X et Y}}{\text{nombre d'apprenants inscrits aux X}} \quad (6)$$

Le premier critère est le support qui est définie comme le pourcentage de transactions contenant X (5). Cela signifie que le support indique la fréquence à laquelle l'itemset X apparaît dans la base de données. Par ailleurs, la confidence est définie comme la proportion des transactions contenant X et Y en même temps (6). Dans le processus de génération des règles d'association, une règle doit obligatoirement satisfaire respectivement le seuil minimal de support et confidence, sinon elle va être écartée de la liste des règles fortes et intéressantes.

IV.1.3 Algorithmes

La recherche des règles d'association intéressantes est un thème privilégié dans le domaine de recherche d'extraction des connaissances à partir des données. Dans ce contexte, plusieurs algorithmes ont été conçus pour permettre de découvrir, à partir d'un ensemble de transactions des règles qui expriment la possibilité d'association entre différents items. Parmi ces algorithmes, nous trouve Apriori[108], FP-growth[111] et Eclat[110].

Nous présentons ici les deux algorithmes les plus puissants pour la génération des sous-ensembles d'items (itemset) fréquents. Le premier est Apriori qui très coûteux en termes d'utilisation des ressources de calcul et d'accès à la base des données. En revanche, le deuxième algorithme, appelé FP-growth, est plus rapide et optimise le coût d'accès.

IV.1.3.1 Apriori

Le problème de l'extraction des règles d'association est abordé pour la première fois par l'algorithme Apriori [108], [109]. Il existe plusieurs algorithmes pour implémenter une

méthode de règles d'association afin d'extraire des relations intéressantes entre des variables ou des éléments dans une base de données volumineuse, en recherchant des ensembles d'éléments plus fréquents. Apriori est l'algorithme le plus populaire. Il a été implémenté dans plusieurs langages, notamment R, Java, C/C++, Python, etc.

Son mécanisme de génération des règles d'association est très simple. Dans un premier temps, il identifie les éléments fréquents dans la base des données et les étend à des ensembles d'éléments plus grands tant que ces ensembles d'éléments apparaissent fréquemment dans les transactions et satisfont la condition du support minimal [141]. L'algorithme Apriori est décrit dans la Figure IV.3.

Modèle d'algorithme Apriori

k-itemsets	An itemset having *k* items
C_k	Set of candidates k-itemsets (potentially large itemsets), which each member of this set has two fields: itemset and support count.
L_k	Set of frequent items of size *k*.
sup_{min}	User-specified minimum support
$conf_{min}$	User-specified minimum confidence

1) $L_1 = \{\text{large l-itemsets}\}$
2) **for** (k=2 ; L_{k-1}!=∅; k++) **do begin**
3) $C_k =$ Apriori-gen(L_{k-1}); // generate new candidates from L_{k-1} [142]
4) **for each** transaction $t \in D$ **do begin**
5) $C_t = subset(C_k, t)$; // candidates contained in t
6) **for each** candidate $c \in C_t$ **do begin**
7) c.count++; //increment the count of all candidates that are contained in the transaction t
8) end
9) $L_k = \{c \in C_k \mid c.count \geq sup_{min}\}$ //candidates in C_k with minimum support
10) end

return $\cup_k \mathbf{L_k}$

Figure IV.3 : Algorithme Apriori

L'algorithme Apriori permet de découvrir les sous-ensembles d'items fréquents en partant de ceux dont la longueur est 1 et en augmentant la longueur au fur et à mesure. Cet algorithme est fondé sur la propriété des sous-ensembles d'items fréquents. En effet, la propriété clé de cet algorithme est la suivante : « Si un ensemble d'items (cours) est fréquent, tous ses sous-ensembles d'éléments seront fréquents ». La Figure IV.4 décrit le modèle de l'algorithme Apriori. Au début, l'algorithme compte simplement les occurrences d'éléments afin de déterminer l'ensemble global d'éléments de longueur L. La deuxième étape consiste en ces phases : Premièrement, les grands ensembles d'éléments L_{k-1} sont utilisés par la fonction *Apriori-gen* pour générer les ensembles d'éléments candidats C_k. Ensuite, la base des données est analysée et le support des éléments candidats dans C_k est calculé, c'est-à-dire incrémenter le nombre de tous les éléments candidats contenus dans une transaction donnée t_i. Ensuite, le support de chaque candidat k-itemset est comparé à sup_{min}. Le k-itemset est fréquent seulement s'il a un support supérieur au support minimum spécifié par l'utilisateur sup_{min}. L'étape suivante consiste à identifier toutes les règles ayant la mesure de confidence supérieure ou égale à $conf_{min}$, la confidence minimale spécifiée par l'utilisateur, c'est-à-dire que l'algorithme supprime toutes les règles ne satisfaisant pas la condition de $conf_{min}$ [12], [143] .

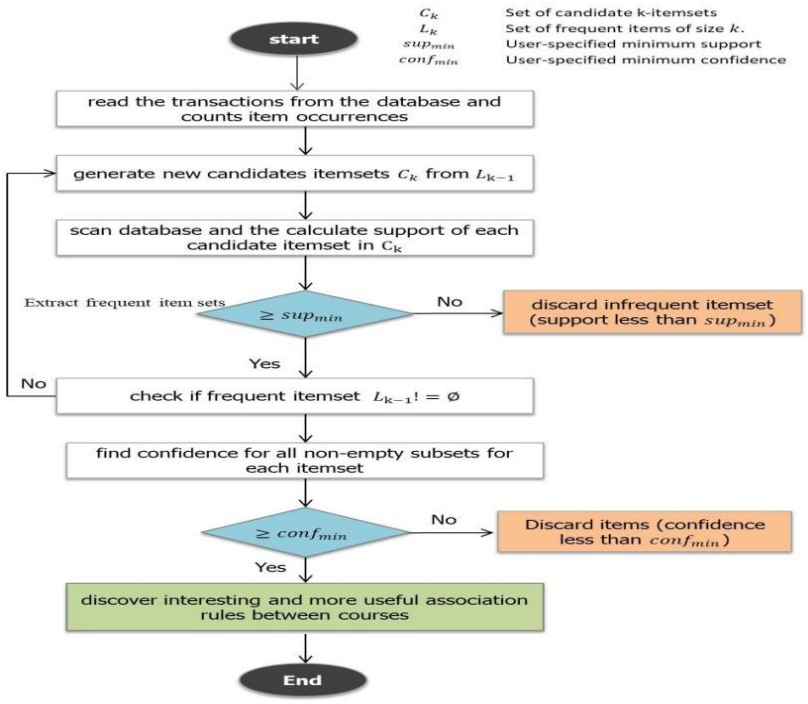


Figure IV.4: Organigramme de l'algorithme Apriori [144]

La figure IV.4 montre l'organigramme de l'algorithme et son fonctionnement. Cela consiste à compter le nombre d'occurrences des items. Ensuite, le processus de génération des jeux d'éléments candidats vérifie le support de chaque sous-ensemble d'éléments candidats, ainsi la confidence minimale. Enfin, on utilise les ensembles d'éléments fréquents pour extraire les règles d'association les plus fortes parmi différents éléments des bases de données des transactions.

Cependant, cet algorithme présente 4 inconvénients majeurs :

- L'inconvénient majeur de cet algorithme est le nombre considérable de scans de la base des données
- Le coût de l'extraction des itemsets fréquents, notamment pour les contextes denses, est exponentiel
- Le nombre des règles d'association générées peut être excessivement élevé
- Les règles d'association produites sont en grande partie redondantes ou inintéressantes.

IV.1.3.2 FP-growth

Pour pallier les limites et inconvénients de l'algorithme Apriori, l'algorithme FP-growth [111] a été développé pour permettre de réduire le nombre de scans de la base des données et d'améliorer les résultats en termes de vitesse de calcul et de qualité des règles d'association extraites.

L'algorithme FP-growth se base sur la stratégie "diviser pour régner" qui consiste à décomposer un problème en sous-problèmes. Tout d'abord, il compresse les itemsets fréquents représentés dans la base des données à l'aide d'une structure de données compacte appelée FP-Tree (frequent-pattern tree) dont les branches contiennent les associations possibles des items. Chaque association peut être divisée en fragments (pattern fragment) qui constituent les itemsets fréquents. La méthode FP-growth transforme le problème de la recherche de l'itemset fréquent le plus long par la recherche du plus petit et sa concaténation avec le suffixe correspondant (le dernier itemset fréquent de la branche aboutissant à l'item considéré). Ceci permet de réduire le coût de la recherche.

FP-growth (Frequent Pattern Growth) est un algorithme très performant pour extraire les éléments les plus fréquents à partir de grands ensembles de données en permettant une

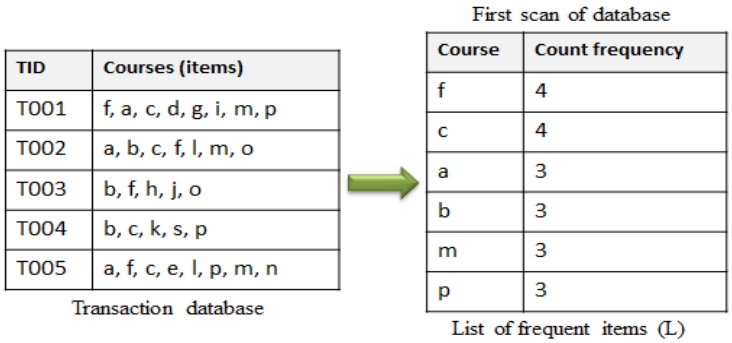
découverte très rapide des règles d'association sans génération des éléments candidats ce qui demande plus mémoire et de temps processeur.

En fait, la génération et le test des candidats nécessitent plusieurs analyses de la base des données. En utilisant FP-growth, le nombre de scans des bases de données est réduit à deux. Le premier scan a pour objectif de compter le support de chaque item, les items non fréquents sont supprimés, tandis que les items fréquents sont triés par ordre décroissant de support, sous la forme d'une liste d'éléments fréquents (L), comme l'illustre la Figure IV.5. Ensuite, dans le deuxième scan l'algorithme construit la structure FP-tree avec la création de la racine et l'insertion des différents nœuds. Ces opérations constituent la première étape de l'algorithme. D'autre part, la deuxième étape vise à extraire les ensembles d'éléments fréquents de l'arborescence FP construite.

La construction de l'arbre FP-tree s'effectue suivant les étapes ci-dessous[145]:

1. Calculer le support minimal.
2. Calculer chacune des occurrences de chaque item
3. Établir un critère de priorité pour ces items.
4. Faire le tri des items en fonction de leur priorité.
5. Créer le nœud racine.
6. À partir de chaque nœud père insérer les enfants en partant du nœud racine
7. Valider la structure de l'arbre FP-growth

Considérons la base des données des transactions, suivante, qui contient 5 transactions et 16 articles ou cours. Supposons que le support minimal égale 3.



TID	Courses (items)
T001	f, a, c, d, g, i, m, p
T002	a, b, c, f, l, m, o
T003	b, f, h, j, o
T004	b, c, k, s, p
T005	a, f, c, e, l, p, m, n

Transaction database

First scan of database

Course	Count frequency
f	4
c	4
a	3
b	3
m	3
p	3

List of frequent items (L)

Figure IV.5: Base des données des transactions & items fréquents ordonnés

Dans un premier temps, l'algorithme FP-growth analyse pour la première fois l'ensemble des données des transactions afin de compter le support de chaque élément et de rechercher les éléments fréquents sous forme de liste L. Les éléments sont triés selon l'ordre décroissant.

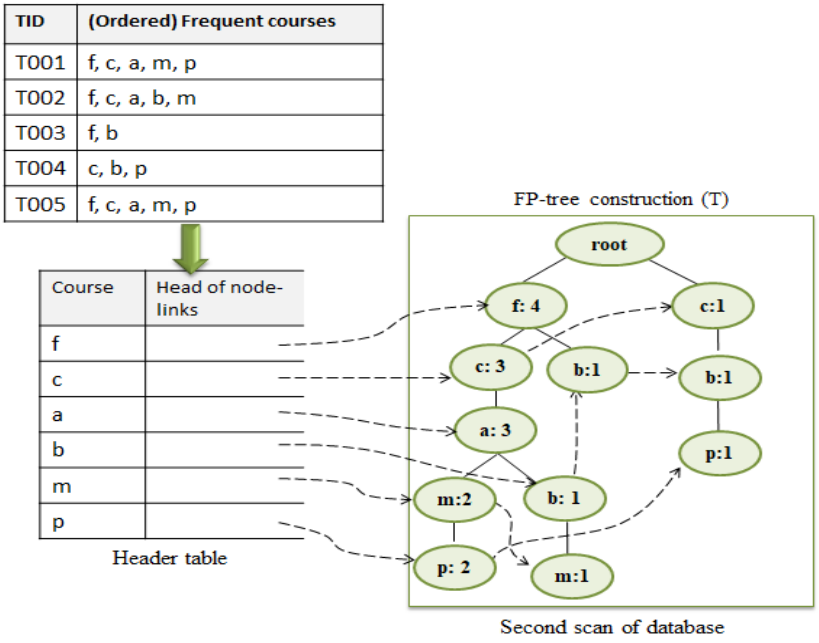


Figure IV.6: Construction de la structure FP-tree

Ensuite, les éléments fréquents de toutes les transactions sont réordonnés en fonction de l'ordre de la liste (L). Puis, FP-growth analyse la base des données pour la deuxième fois et construit FP-tree. Après avoir créé la racine nulle de l'arbre. Il lit la première transaction puis construit la première branche {(f: 1), (c: 1), (a: 1), (m: 1), (p: 1)} de l'arbre (le numéro suivant l'item indique le support). La deuxième transaction T002 partage un préfixe commun (f, c, a) avec le chemin de la première transaction (T001). Ainsi, l'algorithme FP-growth incrémente de 1 le support de chaque nœud appartenant au préfixe, c'est-à-dire que les comptes du premier chemin deviennent {(f: 2), (c: 2), (a: 2), (m: 1), (p: 1)}. Et un nouveau nœud (b: 1) est inséré dans l'arbre et lié en tant qu'enfant du nœud (a: 2); un autre nouveau nœud (m: 1) est également créé et initialisée avec la valeur 1 de support, lequel a pour parent le nœud (b: 1). Ce mécanisme est répété de manière récursive pour chaque transaction du jeu de données jusqu'à ce que toutes les transactions soient scannées et mappées sur un chemin de la structure FP-tree [111]. A la fin du traitement de toutes les transactions, la structure finale de FP-tree est représentée dans la Figure IV.6.

En général, un arbre FP-tree est composé d'une racine est d'un certain nombre de nœuds. Un nœud est composé par le nom de l'item, le nombre d'occurrences de cet item et un lien inter-nœud vers les autres occurrences du même élément figurant dans d'autres séquences de transactions. Une table d'entête pointe sur la première occurrence de chaque item. La deuxième étape de l'algorithme de FP-growth consiste à extraire des modèles fréquents en utilisant la structure FP-tree. Le mécanisme de l'extraction des règles d'association est décrit en détail dans l'article de Jiawei Han et al [111].

IV.1.4 Application et résultats

IV.1.4.1 Projet ESTenligne

Dans le cadre de la mise en application des technologies big data dans un environnement d'enseignement à distance, nous avons choisi une plate-forme d'apprentissage en ligne opérationnelle. Celle-ci s'inscrit dans le cadre du projet ESTenLigne[19], un système e-learning basé sur le LMS Moodle, visant à offrir des cours au profit des étudiants de niveau DUT.

Le projet ESTenligne est le fruit de plusieurs années d'expériences pour le développement de l'e-learning au sein de l'université Sidi Mohamed Ben Abdellah de Fès. Il est lancé depuis 2012 par le réseau des EST du Maroc, qui prévoit le développement de l'enseignement à distance basé sur les nouvelles technologies de l'information et de la communication au sein des universités à travers la mise en œuvre d'une plate-forme pour l'apprentissage en ligne, ouverte, adaptée et évolutive, et en tenant compte des dimensions d'échange, de partage et de mutualisation des ressources pédagogiques [146], [147].

Plusieurs activités de formation ont été réalisées dans le cadre de ce projet. Nous citons à titre d'exemple, la formation d'experts en e-learning dans le cadre du projet Coselearn I, le projet CVM/USMBA, et la formation des enseignants à travers la coopération Franco-marocaine EST[20] et IUT[21]. Il y a, également, des travaux de recherche qui ont été effectués autour de ce projet telle que l'analyse de l'usage des ressources pédagogiques où l'objectif était d'analyser l'usage dans certains cours comme l'algorithmique[148] par exemple. Cependant, dans cette

[19] http://elearn.est-usmba.ac.ma/
[20] http://www.est-usmba.ac.ma/
[21] http://www.iut.fr

plate-forme on n'a jamais travaillé sur les systèmes de recommandation malgré leur rôle primordial comme support d'aide pour les étudiants. En fait, Les étudiants rencontrent beaucoup de difficultés d'accompagnement et se trouvent perdus devant la diversité des ressources pédagogiques et le grand nombre de cours disponibles. Surtout que la plupart des apprenants dans cette plate-forme sont des nouveaux bacheliers qui viennent d'intégrer l'enseignement supérieur et qui ont besoin d'un système pour les aider à cibler les cours pertinents à suivre pendant leur cursus d'apprentissage.

D'où vient la nécessité d'adapter la formation aux besoins des étudiants de l'EST pour répondre aux attentes et aux exigences d'apprentissage. Ce travail vise, alors, à développer un système de recommandation de cours afin de favoriser l'apprentissage chez l'apprenant en lui offrant une solution intelligente capable de proposer les cours les plus appropriés tout en utilisant les données historiques des activités des apprenants.

IV.1.4.2 Architecture du système de recommandation de cours

Le mécanisme de construction de notre système de recommandation est très simple. Il est basé sur l'analyse de l'historique des activités d'inscription des apprenants aux cours. Pour cela, il a besoin de l'ensemble des items (cours) disponibles dans la base des données de la plate-forme ESTenLigne pour être capable de générer un catalogue des cours pertinents tout en analysant les traces d'apprentissage des apprenants. La Figure IV.7 donne un schéma simplifié décrivant le processus de recommandation des cours.

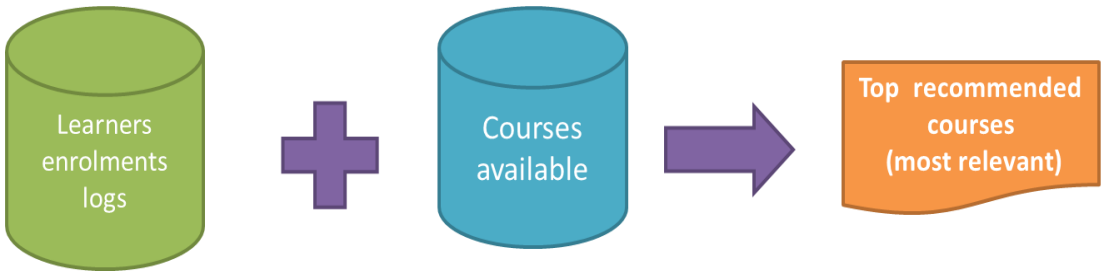


Figure IV.7: Mécanisme de recommandation des cours

D'ailleurs, l'architecture proposée se compose de plusieurs systèmes. Généralement, elle comprend les composants suivants : plate-forme d'apprentissage en ligne et serveur de base des données MySQL qui stocke et gère les données produites par les interactions des apprenants avec le système e-learning, notamment leurs inscriptions aux cours, leurs

comportements, leurs préférences, etc. Par ailleurs, l'environnement RStudio est utilisé pour développer, exécuter et tester notre système de recommandation (Figure IV.8).

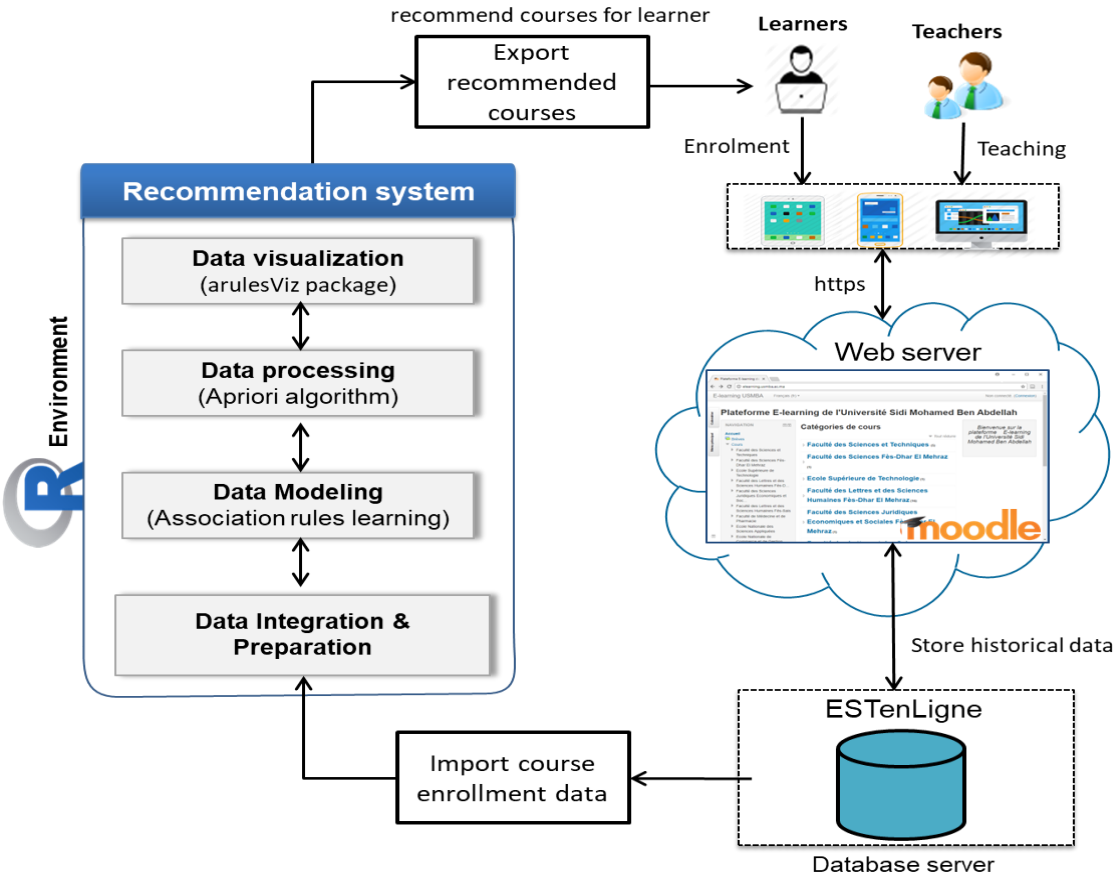


Figure IV.8: Architecture du système de recommandation de cours

En fait, le mécanisme de génération des recommandations est constitué de plusieurs phases allant de la collecte et l'acquisition des données jusqu'à la visualisation et l'interprétation des résultats, en passant par la phase de modélisation, de traitement ou d'analyse. Ce processus repose sur les cinq phases suivantes :

Phase d'intégration et de préparation des données: C'est la phase qui vient juste après l'importation des données d'apprentissage. Il s'agit de la phase la plus difficile et la plus longue au cours du processus de recommandation. Elle consiste à sélectionner les items et les apprenants de la base des données des transactions qui seront utilisées par la suite dans la recherche des relations entre ces variables. Pour cela, le jeu de données doit être filtré de telle sorte à regrouper les cours suivis par chaque utilisateur sous forme d'un vecteur $v = (U, C)$ où U est ensemble d'objets (apprenants), et C est un ensemble d'attributs (cours) ou cours. Un

exemple décrivant la préparation de données constituées de N apprenants, chacun identifié par son ID, et M cours est représenté dans la Figure IV.9.

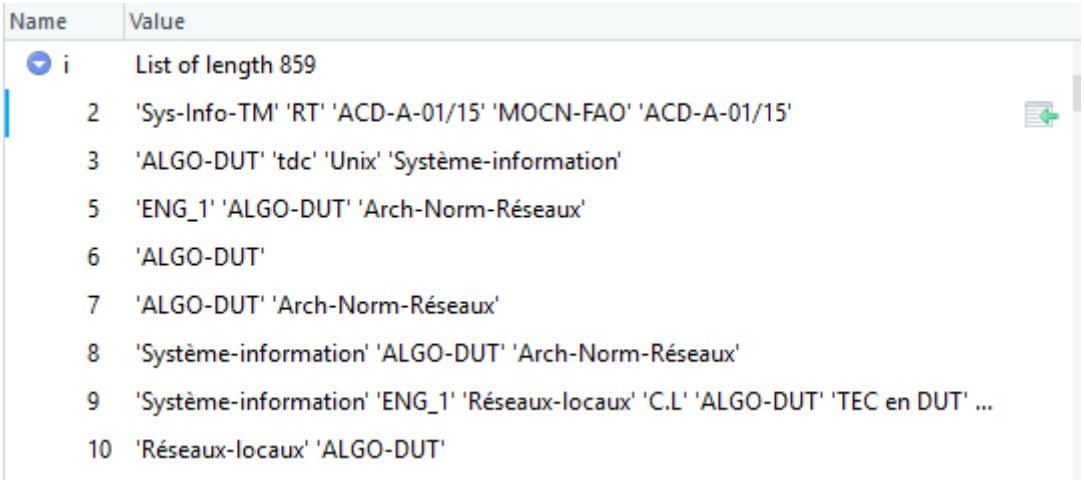


Figure IV.9: Phase de préparation des données

Cette phase de préparation est nécessaire afin qu'il soit possible d'appliquer un algorithme d'extraction des connaissances sur des données collectées. Elle vise à concentrer la recherche des règles d'association sur les données utiles afin de minimiser le temps de génération des recommandations.

Phase de modélisation : A ce niveau, nous identifions le bon modèle à appliquer sur ces données préparées. Dans notre cas nous avons choisi de travailler avec le modèle Association Rules, car nous nous sommes intéressés à la recherche des relations fortes entres les inscriptions des apprenants aux cours, en vue d'extraire les règles d'association puis de recommandation. En outre, cette méthode nous permet de prédire le comportement d'un apprenant donné en anticipant ses prochaines actions et ses activités pédagogiques. Cette phase consiste également à déterminer un algorithme adéquat pour implémenter le modèle analytique choisi. Pour la mise en place, nous avons décidé, dans un premier temps, de construire notre système de recommandation de cours à l'aide de l'algorithme Apriori.

Phase de traitement : C'est à ce niveau qu'on exécute notre modèle analytique. Elle se compose de deux étapes principales :

- Extraction des ensembles fréquents d'attributs : Cela consiste à trouver tous les ensembles d'itemset fréquents dans la base des données des transactions. Un itemset

est fréquent si son support, qui correspond au nombre d'occurrences d'itemset, est supérieur ou égal au seuil minimal du support prédéfini par l'utilisateur.

- Génération des règles d'association : Au cours de cette étape, la liste d'itemset dégagée pendant la phase précédente est utilisée pour générer des règles d'association. Afin de sélectionner uniquement les règles les plus intéressantes, seules celles possédant une confidence supérieure ou égale sont prises en compte et considérées comme des règles fortes.

Phase de visualisation: Par la suite, nous pouvons visualiser les résultats à l'aide des techniques de visualisation arulesViz [149] pour mieux comprendre les résultats d'analyse. En effet, arulesViz est une extension de l'environnement R qui offre plusieurs techniques de visualisation permettant de faciliter la lecture et l'interprétation des résultats d'analyse.

Le résultat de notre système de recommandation est le catalogue des cours recommandés, correspondant aux intérêts et aux compétences individuelles de chaque apprenant. Les étudiants peuvent, ainsi, parcourir ces cours et commencer à apprendre des nouvelles ressources pédagogiques plus adaptées à leurs profils et leurs préférences.

IV.1.4.3 Structure de la base des données

Le projet ESTenLigne est un système d'apprentissage en ligne basé sur le LMS Moodle. Il s'agit d'un système de gestion d'apprentissage open source. Il utilise une base de données relationnelle contenant environ 250 tables. Nous nous sommes concentrés uniquement sur l'historique des inscriptions d'apprenants aux cours. En particulier, nous nous sommes intéressés à quatre tables qui contiennent les informations dont nous avons besoin pour mettre en œuvre notre moteur de recommandation. Le diagramme de classes illustré dans la Figure IV.10 montre la structure, les attributs et les relations entre les différentes entités de la base de données de la plate-forme ESTenLigne.

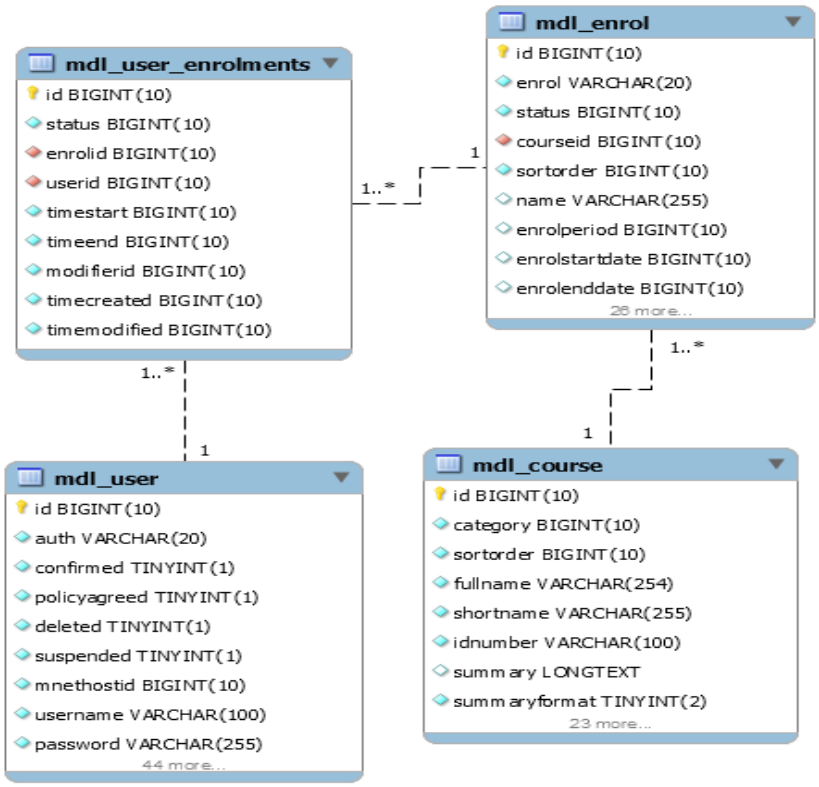


Figure IV.10: Diagramme de classes du système de recommandation de cours

Ce diagramme de classes décrit les quatre tables, importantes, utilisées lors de la phase de collecte des données sur les inscriptions des apprenants. Tout d'abord, On trouve *mdl_user* qui donne des informations sur le profil de l'apprenant. Deuxièmement, toutes les inscriptions des utilisateurs sont enregistrées dans la table *mdl_user_enrolments*. Troisièmement, la table *mdl_enrol* contient des données sur les inscriptions aux cours. En effet, pour un cours donné, il peut y avoir différentes dates de début et de fin d'inscription. Certains étudiants peuvent avoir besoin d'un cours pendant une période donnée, en revanche d'autres étudiants veulent s'inscrire au même cours pendant une période différente. Quatrièmement, la table *mdl_course* stocke tous les détails des cours disponibles dans la base des données de la plate-forme d'apprentissage. Elle contient les noms des cours, la catégorie, le nom complet, le nom abrégé, le résumé, l'heure de création, l'heure de modification, etc. [150]. Ces quatre tables sont liées entre elles par des associations. Par exemple, l'association qui relie *mdl_user* et *mdl_user_enrolments* ayant respectivement les cardinalités (1..1) et (1..*) ce qui signifie qu'un apprenant donnée peut effectuer plusieurs inscriptions différentes, et une inscription ne peut être associée qu'un seul apprenant.

IV.1.4.4 Flow chart du système de recommandation de cours

Le Flow chart montre l'aspect dynamique de notre système. En effet, il décrit une série d'actions ou d'activités effectuées par différents acteurs en schématisant le déroulement du processus de génération des recommandations. La Figure IV.11 illustre le comportement du système de recommandation durant son cycle de vie et également le flux de traitement des données.

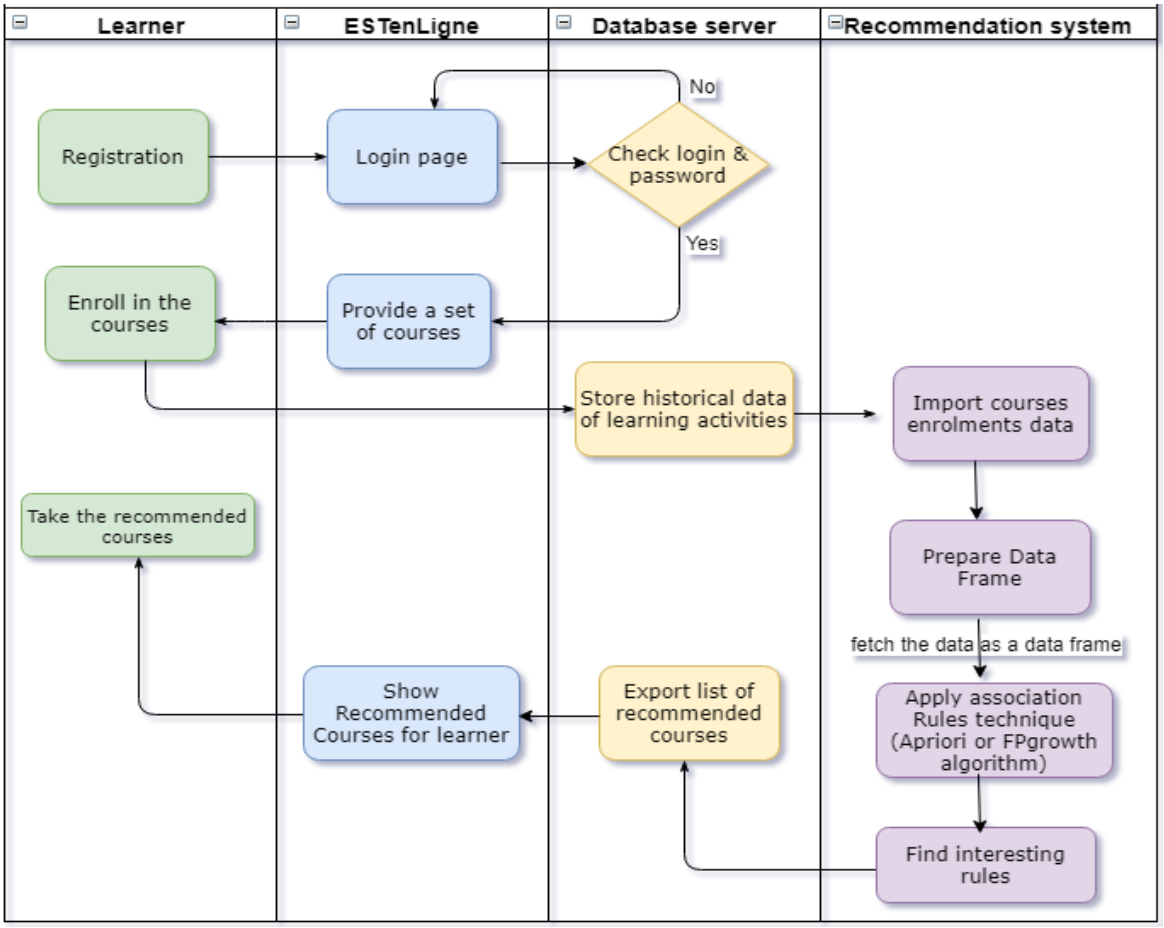


Figure IV.11: Flow chart du système de recommandation de cours

Au début, un apprenant doit créer un compte via la page d'accueil pour pouvoir se connecter au système d'apprentissage. Après la connexion à la plateforme ESTenLigne, le nom d'utilisateur (login) et le mot de passe d'apprenant sont vérifiés. Si la combinaison login et mot de passe est valide, l'apprenant peut accéder à un ensemble de cours disponibles sur cette plate-forme.

Après avoir parcouru une liste des cours disponibles sur la plate-forme ESTenLigne, les apprenants choisissent les cours qui les intéressent. En effet, les données historiques

d'inscription des apprenants, pendant leurs parcours d'apprentissage, sont stockées au niveau de la base des données ESTenLigne. Ensuite, nous collectons des données appropriées à partir des traces d'apprentissage. Par la suite, nous préparons ces données sous forme de Data Frame (structure de données) de façon à ce que les inscriptions aux cours appartenant à chaque apprenant soient regroupées dans un seul vecteur sous la forme d'un tableau d'inscriptions.

Plus tard, nous utilisons l'environnement R pour appliquer les techniques des règles d'association. Lors de la phase d'exécution de notre modèle, Apriori, dans un premier temps, cherche à trouver les cours les plus fréquents, c'est-à-dire les cours sollicités par un grand nombre d'apprenants. Dans un second temps, il extrait le catalogue de recommandation des cours en fonction du profil et du niveau cognitif de chaque apprenant.

L'étape suivante consiste à exporter la liste des cours recommandés vers la base des données de LMS Moodle. Enfin, nous pouvons afficher le résultat du système de recommandation à l'apprenant en lui suggérant des cours plus pertinents afin de le guider durant son cursus d'apprentissage.

IV.1.4.5 Architecture technique et outils de développement

Notre solution de recommandation de cours implémente une combinaison de plusieurs technologies comme décrit ci-après :

IV.1.4.5.1 Langage R

R[151] est un langage de programmation open source utilisé pour le traitement et l'analyse statistique avancée des données. Il s'agit d'une version améliorée et libre du langage interprété S[152]. La première version de R a été développée par les chercheurs de l'université d'Auckland en Nouvelle-Zélande en 1993. L'idée derrière R est de créé un projet collaboratif où chacun pourrait ajouter ses propres packages (fonctionnalités) et les mettre à disposition des autres utilisateurs. R fourni une large palette d'algorithmes avancés pour l'extraction des connaissances (modélisation linéaire et non linéaire, régression, règles d'association, tests des statistiques classiques, classification, clustering) et de techniques graphiques (chart, histogramme, ...). En effet, l'un des points forts de R réside dans sa capacité à produire des graphiques.

IV.1.4.5.2 RStudio

Afin de tester et de valider le système de recommandation de cours, nous avons choisi RStudio comme environnement de développement intégré (IDE) gratuit, libre et multiplateforme facilitant et accélérant le développement avec le langage R. C'est un outil très puissant créé à des fins d'analyse statistique des données volumineuses. Il fournit un ensemble de packages permettant de répondre à de multiples cas d'utilisation et implémente nombreux algorithmes de l'apprentissage automatique, y compris Apriori pour extraire des règles d'association. Il fournit également des options graphiques flexibles permettant différentes techniques de visualisation des données. RStudio est un environnement de développement complet disponible en deux versions. La première version est de type Desktop pour exécuter le code source en locale, alors que RStudio Server est la deuxième version permettant l'accès via un navigateur web. La puissance de cet IDE est sa compatibilité avec la plupart des plates-formes.

IV.1.4.5.3 Eclipse Oxygen

Eclipse[153] est un environnement de développement intégré libre, multi-langages (C/C++, Python, PHP, Ruby,...) et multi-plates-formes permettant de configurer, développer, tester, valider des logiciels. Eclipse IDE est principalement écrit en Java (à l'aide de la bibliothèque graphique SWT). La spécificité d'Eclipse IDE vient du fait que son architecture est totalement développée autour de la notion de plugins ou d'extensions. En plus, il s'agit d'un IDE très riche qui permet tous types de développement, y compris d'applications desktop, web et mobile. Figurant parmi les grandes réussites de l'Open source, Eclipse est devenu un standard du marché de développement des logiciels, intégré par de grands éditeurs des solutions informatiques et sociétés de services.

IV.1.4.5.4 Maven

Maven[154] est un outil pour la gestion et l'automatisation de production des projets logiciels basés sur le langage de programmation Java. Son développement est fait principalement par la fondation Apache. Maven permet de standardiser la forme d'un projet ainsi que son utilisation. Il permet également de gérer les dépendances d'un projet, d'effectuer une livraison (packaging) complète et automatique, et de faciliter le déploiement des applications. Il utilise un paradigme connu sous le nom de Project Object Model (POM) afin de décrire un projet

logiciel, ses dépendances avec des modules externes et l'ordre à suivre pour sa production. Il est livré avec un grand nombre de tâches prédéfinies, comme la compilation du code Java ou encore sa modularisation.

IV.1.4.5.5 MySQL

MySQL est un système de gestion des bases de données relationnelles (SGBDR). Il fait partie des logiciels de gestion des base de données les plus utilisés au monde, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle[155], PostgreSQL[156], et Microsoft SQL Server[157]. C'est un serveur des bases de données relationnelles SQL développé pour offrir des performances élevées en lecture, ce qui signifie qu'il est orienté davantage vers le service des données déjà en place que vers celui des mises à jour fréquentes. Il est multithread et multiutilisateur.

IV.1.4.5.6 Scala

Scala[158] est un langage de programmation multi-paradigme développé à l'École polytechnique fédérale de Lausanne en Suisse. Il est considéré comme une évolution du langage Java et hérite toutes les fonctionnalités offertes par ce dernier en matière de programmation orientée objet comme l'héritage, le polymorphisme, etc. Scala est conçu pour programmer le code source des applications dans une forme concise et élégante. En plus, Scala intègre la possibilité de la programmation fonctionnelle. Ce langage a fait sa preuve dans plusieurs secteurs et il est adopté par Twitter, Siemens, et Sony. De surcroît, la fondation Apache a utilisé Scala comme langage natif pour développer son Framework Spark, le leader des solutions big data.

IV.1.4.5.7 Oracle VM VirtualBox

VirtualBox[159] est une solution de virtualisation libre publiée par Oracle. Elle permet de faire tourner sur une seule machine (système hôte) plusieurs machines virtuelles dans lesquelles on peut installer des systèmes d'exploitation variés (systèmes invités). L'avantage de VirtualBox est sa capacité à supporter plusieurs systèmes d'exploitation, y compris Linux[160], Mac OS X[161], Solaris[162, p. 11], et Windows[163]. VirtualBox offre un ensemble complet d'outils permet l'accès aux ressources de la machine virtuelle telles que le disque virtuel, la mémoire et le réseau via des interfaces modernes, simples et ergonomiques.

IV.1.4.6 Expériences et résultats

Pour évaluer l'efficacité du système de recommandation de cours présenté dans ce projet de thèse, nous avons mené de nombreuses expériences sur 1218 étudiants de l'Ecole Supérieure de Technologie de Fès (ESTF). 859 parmi ces étudiants suivent au moins un cours et 359 sont des nouveaux inscrits à la plate-forme ESTenLigne et n'ont pas encore choisi des cours à apprendre. Nous nous sommes donc concentrés uniquement sur les étudiants inscrits aux cours. Le système développé recommande les cours pour les étudiants en sélectionnant les ressources pédagogiques les plus pertinentes parmi 153 cours en total. Nous n'avons considéré donc que l'historique des cours choisis par les étudiants.

Avant de lancer l'exécution du modèle des règles d'association, il faut que les données d'apprentissage soient structurées pour qu'elles soient traitées par l'algorithme Apriori. Pour préparer ces données, On regroupe les cours suivis par chaque apprenant sous forme d'un vecteur d'inscriptions, et on supprime les données inutiles. Ensuite, nous transformons les données collectées en objet optimisé (S4) pour faciliter leur manipulation par l'algorithme Apriori. Pour ce faire, nous avons développé une requête SQL afin d'extraire la liste des cours suivis par tous les apprenants à partir de la base des données de la plate-forme ESTenLigne. En effet, nous nous sommes concentrés essentiellement sur quatre tables (voir la section 'Structure de la base des données'). Il s'agit de la table *mdl_user*, *mdl_user_enrolments*, *mdl_enrol* et *mdl_course*. Afin d'effectuer cette tâche efficacement, nous avons intégré le package RMySQL dans l'environnement RStudio pour connecter ce dernier avec la base de données MySQL du LMS Moodle. Après avoir exécuté la requête SQL à l'aide de l'outil RMySQL, nous avons obtenu une collection de cours classées par l'identifiant de l'apprenant (user_id) chose qui servira, par la suite, comme données d'entrée de notre model.

IV.1.4.7 Résultats Apriori

L'étape suivante consiste à lancer l'exécution de l'algorithme Apriori en utilisant les données capturées et préparées dans la phase précédente. Pour ce faire, nous avons utilisé le package arules [24] qui implémente l'algorithme Apriori. Ainsi, dans un premier temps, nous avons spécifié le seuil minimal de support et de confidence, respectivement, pour pouvoir extraire les relations les plus fortes entre les inscriptions des apprenants aux cours. Le nombre de règles d'association intéressantes générées change en fonction de la valeur du support de la confidence spécifié par le data scientist, mais aussi de la taille de la base des données à

analyser. Afin de trouver des corrélations plus fortes, nous avons utilisé un seuil minimal de support de 5% et nous avons fixé un seuil de confidence minimal de 50%. Finalement, nous avons pu exécuter l'algorithme Apriori comme illustré dans la Figure IV.12.

Résultats d'Apriori algorithme

```
Apriori

Parameter specification:

 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext

 0.5   0.1 1 none FALSE  TRUE  5 0.05   1  10  rules FALSE

Algorithmic control:

 filter tree heap memopt load sort verbose

    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 42


set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[144 item(s), 859 transaction(s)] done [0.00s].

sorting and recoding items ... [15 item(s)] done [0.00s].

creating transaction tree ... done [0.00s].

checking subsets of size 1 2 3 done [0.00s].

writing ... [23 rule(s)] done [0.00s].

creating S4 object  ... done [0.00s].
```

Figure IV.12: Résultat de l'algorithme Apriori

Le résultat d'exécution de l'algorithme Apriori est un ensemble de 23 règles qui satisfont les conditions de support et de confidence minimales. Les 10 règles les plus intéressantes ordonnées selon la mesure de confidence sont illustrées dans Table IV.1.

Tableau IV-1: Résultats de l'algorithme Apriori

Rule	lhs	rhs	Confidence	Support
[1]	{11,46}	{45}	1.0000000	0.06519208
[2]	{11,45}	{46}	0.9824561	0.06519208
[3]	{46}	{45}	0.9593496	0.13736903
[4]	{15,46}	{45}	0.9387755	0.05355064
[5]	{15,45}	{46}	0.9387755	0.05355064
[6]	{45}	{46}	0.9365079	0.13736903
[7]	{6,7}	{18}	0.8518519	0.05355064
[8]	{6,18}	{7}	0.8363636	0.05355064
[9]	{7}	{18}	0.8024691	0.07566938
[10]	{7,18}	{6}	0.7076923	0.05355064

Selon les résultats obtenus de support et de confidence, et comme le montrent les résultats expérimentaux de la Table VI.1, l'association entre les cours {11 et 46} et {45} possède la plus grande confidence (1.000), aussi l'association entre les cours {7 et 18} et {6} a la plus faible confidence (0.707). En se basant les valeurs calculées de support et de confidence, il apparait clairement quels sont les cours les plus favorisés par les apprenants et nous pouvons, même, prédire les cours les plus adéquats pour chaque situation d'apprentissage. Par exemple, pour la première règle d'association {11, 46} => {45}, notre système recommande donc le cours 45 aux étudiants qui s'inscrivent aux cours {11 et 46}. Selon la Table VI.1, la confidence de la règle 1 est de 100%, car au niveau de la base des données de la plate-forme ESTenLigne, on trouve que 56 étudiants se sont inscrits aux cours {11 et 46}, et 56 d'entre eux se sont inscrits également au cours {45}. De même, pour la deuxième règle d'association, Il y a 57 étudiants qui se sont inscrits aux cours {11 et 45}, et 56 parmi eux ont consulté le cours {46}. L'efficacité de la règle 2 est donc de 98%. Le système suggère donc le cours {46} aux étudiants inscrits aux cours {11 et 45}. En ce qui concerne la règle 3, 123 apprenants se sont inscrits au cours {46}, et une forte proportion d'entre eux (118 apprenants) se sont inscrits également au cours {45}, la confidence de la règle 3 est donc 95%, le système propose le cours {45} aux étudiants inscrits au cours {46}, etc. Pour les 10 règles principales,

nous notons que les valeurs de confidence se situent entre 0,707 (70,7%) et 1,00 (100%), ce qui prouve que nous avons obtenu des résultats encourageants. Nous pouvons donc en conclure que le système de recommandation de cours, que nous avons proposé, fournit des contenus éducatifs plus pertinents et plus adaptés aux préférences et aux connaissances de chaque apprenant.

La Figure IV.13 montre la fréquence des éléments de manière visuelle, nous avons utilisé la fonction itemFrequencyPlot[164] du package arules[165] qui permet d'afficher la répartition relative aux inscriptions aux cours les plus fréquents dans notre base de données.

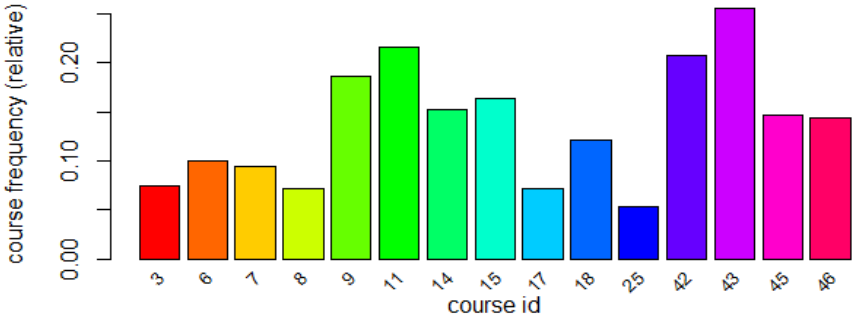


Figure IV.13: Distribution de fréquence relative aux inscriptions aux cours

Cet histogramme se limite seulement aux 15 cours les plus fréquents. On remarque que le cours le plus fréquent est celui ayant l'id 43, avec 219 inscriptions sur 2553 transactions dans la base des données ESTenLigne.

IV.1.4.8 Amélioration des performances du système de recommandation de cours

Afin d'améliorer les performances de notre système de recommandation, il existe un algorithme alternatif, plus rapide qu'Apriori, permettant d'obtenir de meilleurs résultats en termes de temps d'exécution et d'optimisation des ressources. En effet, Apriori nécessite un nombre très élevé de scans de la base des données des transactions, c'est-à-dire que la transaction est analysée à chaque génération d'éléments candidats, alors que lorsqu'on utilise FP-growth, le nombre de lectures de la base de données est réduit à deux. Tout d'abord, FP-growth analyse la base des données pour déterminer les ensembles d'éléments fréquents dans un ordre décroissant en fonction de leurs supports, et supprimer les ensembles d'éléments non fréquents. Le deuxième scan est effectué pour la construction de FP-tree[111].

Pour montrer la performance des deux versions de notre système, il suffit de mesurer le résultat d'exécution des deux algorithmes. La Table IV.2 montre la différence de temps

d'exécution entre l'algorithme FP-growth par rapport à Apriori en utilisant des valeurs variées du support allant de 0,5% à 3%.

Tableau IV-2: Performance de FP-growth vs. Apriori

Support	0.005	0.01	0.015	0.02	0.025	0.03
FP-	0.07	0.06	0.05	0.05	0.04	0.04
Apriori	0.09	0.08	0.07	0.07	0.06	0.06

En vue de faciliter la compréhension des mesures de performance en termes de temps d'exécution, les résultats de la Table VI.2 sont représentés dans la Figure IV.14.

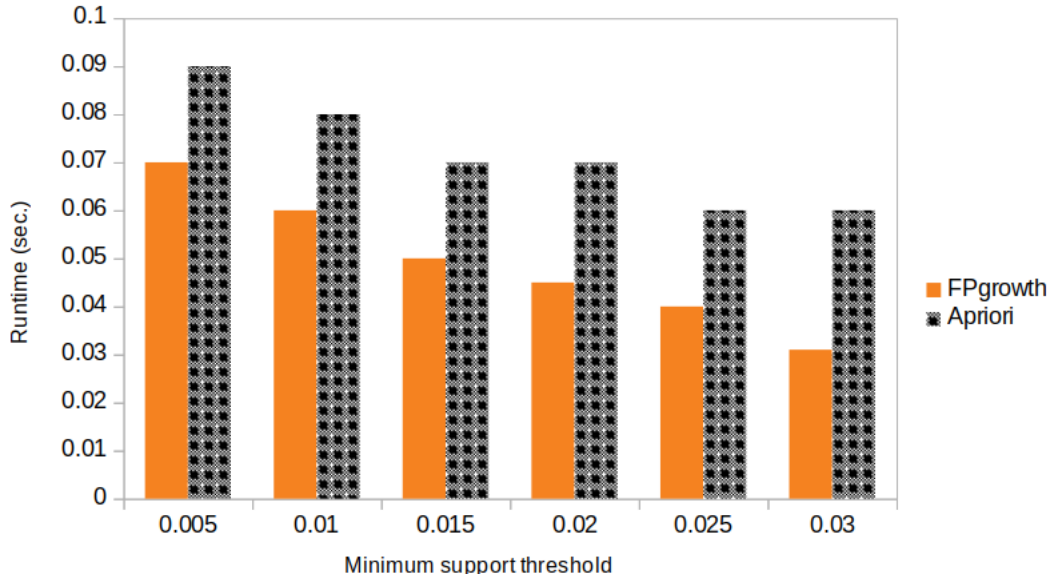


Figure IV.14: Performance de FP-growth vs Apriori

Les résultats des expériences montrent que la performance de la vitesse d'exécution de l'algorithme FP-growth est bien meilleure que celle d'Apriori. En effet, Apriori a besoin de plus de temps pour générer des éléments fréquents par rapport à FP-growth. Lorsque la valeur de support minimal diminue, les nombres fréquents augmentent. Par conséquent, le temps d'exécution d'Apriori croît d'une manière exponentielle. Par exemple, pour le seuil de support minimal égal à 2,5%, l'exécution de l'algorithme Apriori passe 0,06 seconde pour trouver des règles intéressantes alors que FP-growth n'a besoin que de 0,04 seconde pour chercher les cours fréquents et extraire des recommandations à partir de ces derniers. Parce que, Apriori requiert un nombre considérable de lectures de la base des données. En d'autres termes, les transactions sont scannées à chaque génération d'éléments candidats, tandis que FP-growth

demande uniquement deux scans afin d'extraire les connaissances à partir des traces des inscriptions des apprenants aux cours.

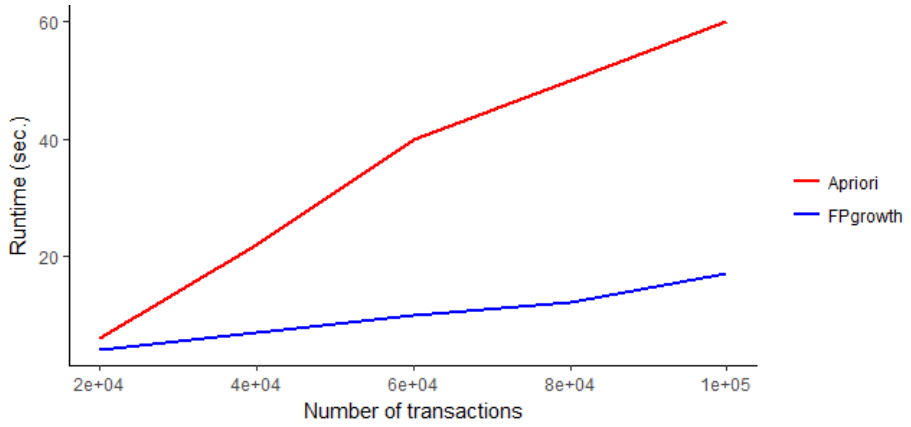


Figure IV.15: Temps d'exécution de FP-growth vs Apriori en fonction du nombre des transactions

La Figure IV.15 expose l'impact de l'augmentation de la taille des bases de données des transactions sur les performances des deux algorithmes. Quand le nombre de transactions augmente, l'exécution d'Apriori devient beaucoup plus lente par rapport à celle de FP-growth. Parce qu'Apriori génère plus d'items candidats et doit analyser plusieurs fois la base des données. Quand il s'agit d'une base de données volumineuse, le calcul devient coûteux.

En résumé, les résultats d'exécution obtenus en se basant sur le calcul du temps d'exécution que ce soit pour le support (Figure IV.14) ou pour le nombre de transactions dans la base des données (Figure IV.15), il est clair que notre système de recommandation utilisant FP-growth est plus rapide et donne des meilleurs résultats que lorsque nous choisissons de travailler avec l'algorithme Apriori.

IV.1.4.9 Visualisation des résultats de recommandation

La visualisation des règles d'association pose un défi, surtout lorsqu'il s'agit d'un grand jeu de données. En vue de faciliter la compréhension des résultats d'analyse des données historiques des apprenants et d'améliorer la prise de décision. Cette phase consiste en la visualisation des règles d'association découvertes et leur interprétation afin d'en déduire des relations intéressantes entre les variables.

Afin de visualiser plus clairement les résultats des expérimentations menées, nous avons utilisé une extension de l'environnement R, appelée arulesViz[149] package. Ce dernier offre

plusieurs techniques permettant une visualisation graphique et interactive des règles d'association.

La Figure IV.16, ci-dessous, donne une vue globale sur la distribution, selon le support et confidence, des règles d'association obtenues par notre système de recommandation.

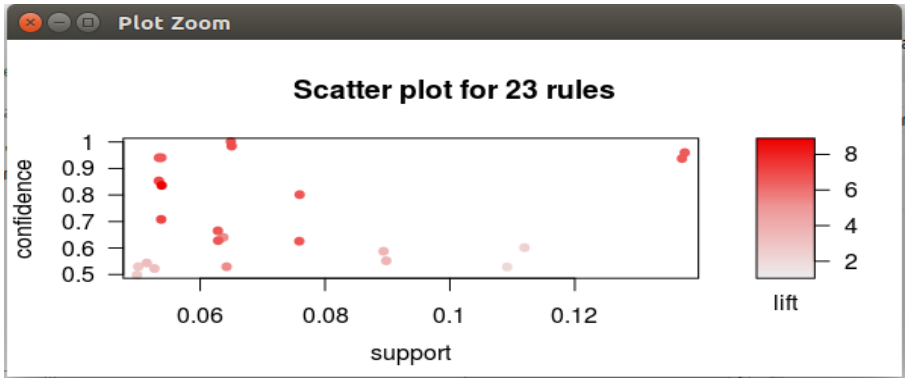


Figure IV.16:Méthode de visualisation des règles d'association : Scatter Plot

Scatter plot est une méthode de visualisation des règles d'association qui consiste à utiliser un nuage de points utilisant le support et la confidence sur les axes. De plus, une troisième mesure (lift) est utilisée comme couleur des points. On peut également représenter graphiquement les règles en utilisant les techniques de visualisation qui s'appuyent sur des graphes représentant des règles d'association dans des sommets.

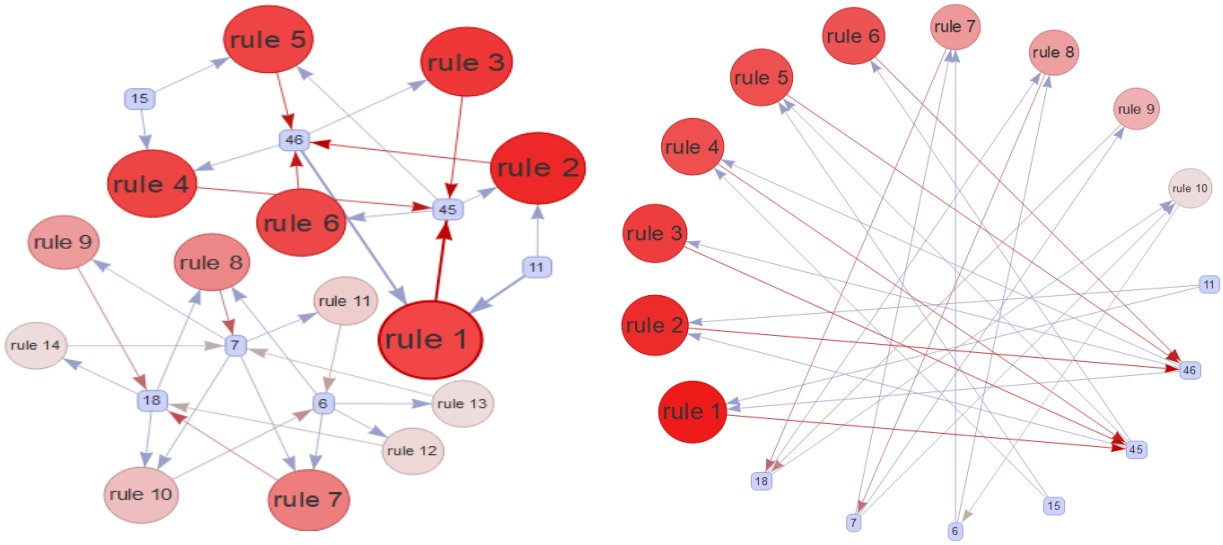
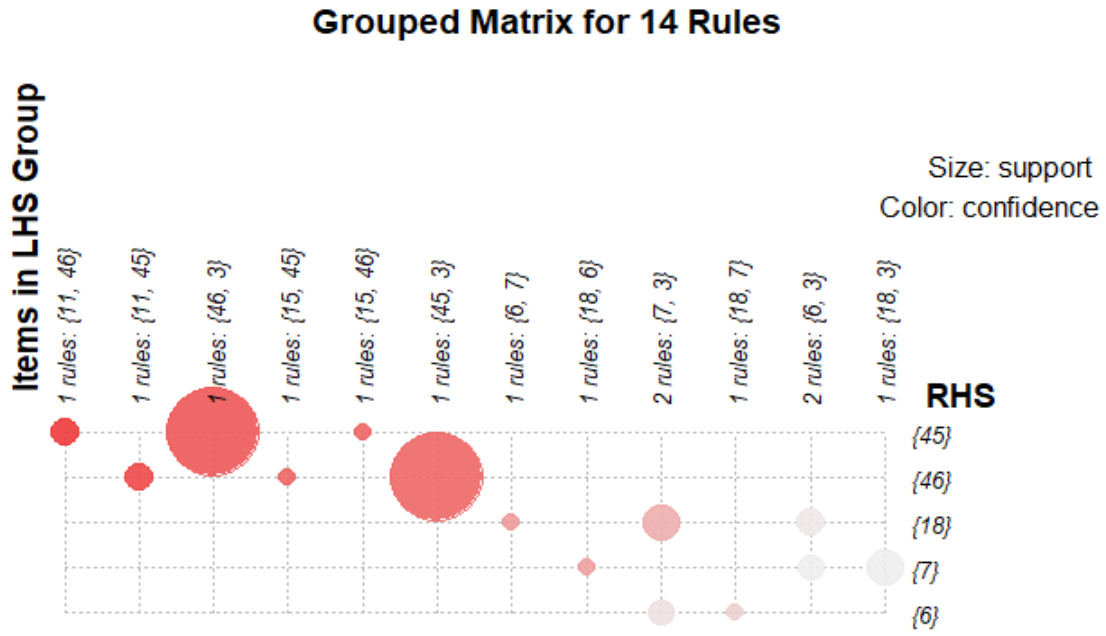


Figure IV.17: Méthode de visualisation des règles d'association : graph

La Figure IV.17 illustre très clairement les 10 règles les plus fortes utilisant des techniques basées sur des graphes. En effet, Elle visualise les règles d'association à l'aide de sommets et d'arêtes, où les sommets représentent généralement un élément (cours) ou un identifiant de la règle, et les arêtes indiquent une relation entre des règles. La règle 1 du graphe montre que, les étudiants qui s'inscrivent au cours {11 et 46} s'inscrivent également au cours 45.

Une autre technique, permettant de représenter les règles fortes, est Grouped Matrix (Figure IV.18). Elle conçue pour visualiser des grands ensembles de règles d'association. En effet, cette technique nous donne la possibilité de sélectionner et de faire du zoom afin de bien cibler les relations intéressantes parmi un grand ensemble de règles extraites à partir des traces d'apprentissage.



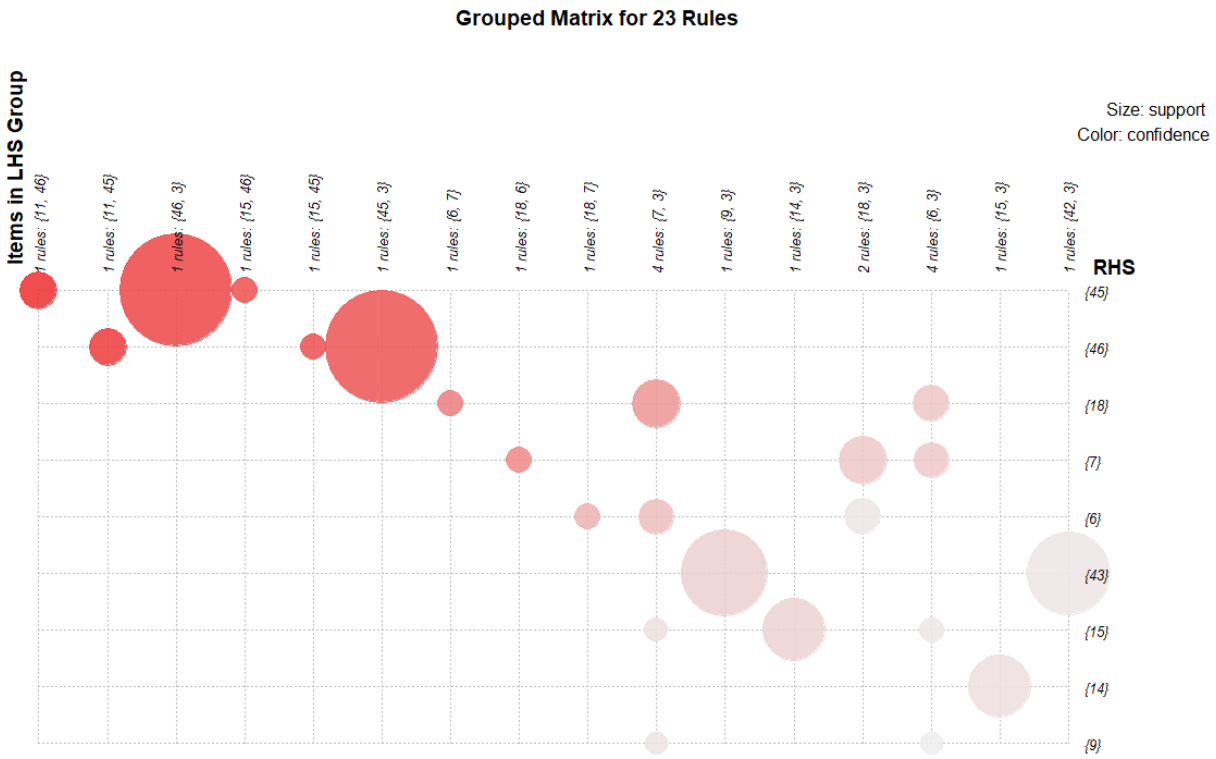


Figure IV.18: Méthode de visualisation des règles d'association : Grouped Matrix

IV.1.5 Conclusion

Dans ce chapitre nous avons conçu et développé un système de recommandation de cours pour un environnement informatique pour l'apprentissage humain dédié à l'e-learning. L'objectif est d'aider les étudiants à trouver les cours les plus adaptés à leurs besoins et leurs préférences en recherchant toutes les relations intéressantes entre les inscriptions aux cours et les transactions de la base des données. Pour mettre en place notre moteur de recommandation nous avons choisi de travailler avec le langage R. Ce dernier fournit une implémentation de plusieurs méthodes d'apprentissage automatique comme la technique des règles d'association. La collecte des données historiques sur les activités des apprenants a fait usage de la plate-forme d'apprentissage ESTenLigne. Notre solution de recommandation proposée, a été construite en utilisant le logiciel open source RStudio pour la découverte, la préparation, la modélisation, le traitement et la visualisation des données via le package arulesViz. Ainsi, l'implémentation de notre système a été faite, dans un premier temps, en utilisant l'algorithme Apriori pour extraire les connaissances à partir des historiques des inscriptions des apprenants aux cours. Les résultats expérimentaux ont montré l'efficacité du système de recommandation développé dans le but de fournir un service d'apprentissage intelligent et capable de guider et

d'orienter un apprenant durant son parcours de formation en lui suggérant des ressources pédagogiques plus adéquates et appropriées à son profil et son niveau cognitif. Par la suite, nous avons remplacé L'algorithme Apriori par FP-growth. Ce dernier donne d’excellents résultats en termes de temps d'exécution et d'utilisation des ressources machines, ce qui permet d'améliorer grandement l'efficacité des recommandations et la performance du moteur de recommandation proposé.

IV.2 Système de recommandation de cours à grande échelle

IV.2.1 Introduction

Dans les environnements informatiques pour l'apprentissage humain dédiés à l'apprentissage en ligne, la quantité croissante des données produites par un système e-learning rend plus difficile à trouver ce dont l'apprenant a vraiment besoin. Une telle plate-forme est également confrontée à de nombreux défis, tels que la diversité des cours et le grand nombre d'apprenants. Pour faire face à ces challenges, les systèmes de recommandation sont conçus pour aider ces apprenants à identifier facilement les cours les plus appropriés sur ces plateformes. Un système de recommandation a pour objectif de fournir à un apprenant donné les ressources pédagogiques les plus pertinentes en fonction de son profil et de son niveau cognitif. En effet, un système de recommandation tente d'apporter des réponses aux besoins d'apprentissage d'un apprenant en lui suggérant les activités pédagogiques les plus susceptibles de lui plaire, ce qui permet de réduire significativement le temps de recherche et de le guider au cours de son cursus d'apprentissage.

Dans cette partie, nous allons migrer notre système de recommandation d'une architecture standard s'exécutant sur un seul ordinateur vers une infrastructure totalement distribuée composé de plusieurs machines à l'aide de la nouvelle génération des technologies big data. En fait, nous allons développer et déployer notre système de recommandation en utilisant l'écosystème big data, notamment les Frameworks Hadoop et Spark. Ce dernier, se base sur une approche tout à fait différente qui consiste à répartir le traitement et les données sur plusieurs nœuds, chacun effectue des tâches spécifiques afin d'exécuter, parallèlement, un algorithme sur différentes machines d'un cluster. Ce mécanisme de distribution du calcul et du stockage permet de résoudre le problème de la limitation des ressources disponibles tout en accélérant la vitesse d'exécution et la performance du système proposé.

Après avoir présenté notre système de recommandation dans un environnement centralisé, composé d'une seule unité physique de calcul et de stockage, nous allons décrire, dans plus de détails, l'infrastructure distribuée du paradigme big data. Dans ce qui suit, nous allons présenter le système de recommandation de cours à grande échelle que nous avons développé. Ainsi, nous commencerons par une petite description de la version distribuée de l'algorithme FP-growth fourni par le Framework Apache Spark. Par la suite, nous exposerons l'infrastructure distribuée constituée d'un cluster Hadoop afin de permettre l'exécuter du

travail (job) de notre solution de recommandation sur plusieurs machines simultanément. La section suivante est consacrée aux résultats expérimentaux de notre système. Ensuite, nous présenterons des outils performants dédiés pour la surveillance et le pilotage des ressources de l'infrastructure répartie du cluster Hadoop.

IV.2.2 Algorithme FP-growth Parallèle

Il existe plusieurs implémentations de l'algorithme FP-growth, la meilleure est celle qui implémente une version parallèle de cet algorithme, appelée Parallel FP-growth (PFP).

Le PFP[166] est basé sur un nouveau schéma de distribution du calcul. En effet, il répartit le travail sur un cluster de nœuds à l'aide du modèle MapReduce. Ainsi, il est plus évolutif et rapide que son implémentation traditionnelle basée sur une seule machine. La bibliothèque Spark MLlib (spark.ml) [167] fournit une implémentation en mémoire de l'algorithme PFP, ce qui donne de bonnes performances en accélérant la vitesse de découverte des règles d'association tout en s'appuyant sur une approche de décentralisation du calcul. Le PFP fourni par Spark est disponible dans les quatre langages de programmation les plus populaires : Scala, Java, Python, et R. En effet, MLlib est une libraire très riche, elle implémente les principaux algorithmes d'apprentissage automatique ainsi que des méthodes statistiques. Outre les paramètres par défaut de FP-growth classique (*minSupport* et *minConfidence*), la bibliothèque *spark.ml* utilise également le paramètre *numPartitions* qui spécifie le nombre de partitions utilisées pour fractionner un travail donné en plusieurs tâches élémentaires traitées sur un cluster de nœuds. En effet, l'idée de base étant de subdiviser les données et les taches de façon à ce qu'on puisse, exécuter facilement les algorithmes correspondants d'une manière concurrente et dans des unités physiques de calcul différentes. L'avantage de cet algorithme est qu'il offre de meilleures performances en termes de temps d'exécution et d'évolutivité.

L'algorithme FP-growth parallèle comprend trois méthodes :

1) **freqItemsets** : il s'agit d'une méthode qui permet de trouver des itemsets fréquents dans la base des données des transactions.
2) **associationRules** : permet d'extraire les règles d'association les plus fortes qui satisfont le seuil minimal de confidence (minConfidence). Ces règles sont générées sous forme de DataFrame ("antecedent" [Array], "conséquent" [Array], "confidence" [Double]).

3) **Transform** : examine chaque transaction et fait la comparaison des éléments avec les antécédents de chaque règle d'association. Si une transaction contient tous les antécédents d'une règle d'association, la règle sera considérée comme applicable et ces conséquences seront ajoutées au résultat de la prédiction. Ensuite, cette méthode récapitule les conséquences de toutes les règles applicables en tant que prédiction et présente le résultat à l'utilisateur. Les résultats de recommandation sont affichés dans des colonnes de prédiction qui listent un catalogue des éléments les plus susceptible de plaire un utilisateur donné.

IV.2.3 Technologies mises en œuvre

Cette section introduit toutes les technologies big data utilisées ainsi que leurs différents composants. Chaque technologie a un rôle bien déterminé et participe dans le processus d'extraction des connaissances en proposant une liste des cours fortement recommandés selon les besoins et les intérêts de chaque apprenant. Ces technologies sont déployées sur une infrastructure de cluster des nœuds qui sont interconnectés via des protocoles réseaux pour pouvoir communiquer et échanger les données lors de l'analyse des traces d'apprentissage. Ces technologies peuvent être organisées en trois niveaux. La Figure IV.19 décrit les différents Frameworks big data utilisés afin d'élaborer notre système de recommandation de cours à grande échelle.

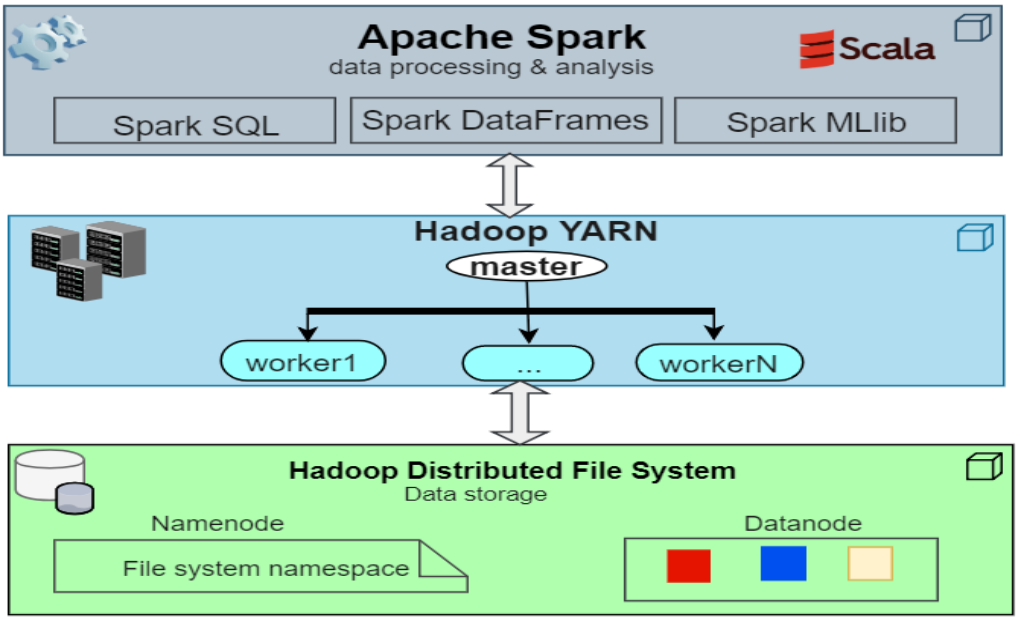


Figure IV.19: Technologies Big Data

L'ensemble des technologies implémentées peuvent être regroupées dans 3 couches:

En premier lieu, il y a la couche du stockage distribué des données. Nous avons choisi d'employer de système de fichier distribué (HDFS) de Hadoop. En fait, HDFS est un système de fichiers tolérant aux pannes et capable de gérer des données réparties sur des grands clusters. Il est doté d'une architecture master/worker (maître-esclave) et conçu pour être déployé sur des infrastructures standards. HDFS fournit un accès haute performance aux quantités volumineuses de données, il est, de même, adapté aux applications manipulant de grands volumes de données. Il crée une abstraction des ressources du disque dur pour permettre la gestion du stockage physique distribué de plusieurs nœuds comme s'il n'y avait qu'un seul espace de stockage. Dans l'architecture HDFS, les données sont gérées à travers le cluster, dans différents Datanodes, par les workers sous la forme de fichiers structurés en blocs. Les emplacements de ces blocs et de l'espace de nommage des fichiers et des répertoires sont conservés dans le composant Namenode dans le nœud master.

Dans le deuxième niveau, on trouve Yarn en tant que gestionnaire de ressources du cluster. Il s'agit d'un module de Hadoop dédié à la planification et l'exécution des tâches, parallèlement, sur un certain nombre d'ordinateurs. Il est, également, responsable de la gestion des ressources du disque, de la mémoire, du processeur et du réseau. L'idée principale de Yarn est de séparer la gestion des ressources du modèle de calcul. Pour pouvoir lancer une application Spark, il faut connecter ce dernier à un gestionnaire de cluster Yarn. En effet, Yarn va s'occuper de louer les ressources nécessaires et de répartir les tâches élémentaires sur différentes unités de calcul des nœuds workers.

Finalement, la couche supérieure représentée par le Framework Spark, est responsable de la manipulation et de l'analyse des données. Ce dernier est utilisé pour appliquer la technique des règles d'association aux données d'apprentissage des apprenants, colletées à partir du projet ESTenLigne. Pour implémenter notre système de recommandation de cours, Scala a été choisi comme langage de développement. L'avantage de Spark est sa capacité à supporter plusieurs langages de programmation tels que Scala, Java, Python et R. Ce Framework fournit de nombreuses bibliothèques. Dans notre cas d'utilisation, nous nous sommes concentrés sur trois composants uniquement : Spark SQL, Spark DataFrames, et Spark MLlib.

- Spark SQL permet d'établir la connexion à la base des données du LMS moodle et d'exécuter des requêtes SQL.
- Spark DataFrames dédié pour le traitement des données structurées.
- Spark MLlib implémente plusieurs algorithmes d'apprentissage automatique, y compris l'algorithme parallèle FP-growth.

IV.2.4 Architecture répartie Hadoop

En général, notre approche consiste à générer des recommandations en analysant les traces des activités d'apprentissage qui représente la base de données du processus de personnalisation des ressources pédagogiques fournit aux apprenants. L'entrée du système est donc constituée de l'historique des inscriptions des apprenants aux cours, importées de la base des données de la plate-forme e-learning.

En fait, le processus de recommandation comprend les étapes suivantes :

Au début, nous devons charger les données produites par les interactions des apprenants avec la plateforme ESTenLigne. Ensuite, ces données, chargées par la bibliothèque Spark SQL, sont traitées de manière distribuée à l'aide du Framework Spark qui s'exécute sur un cluster Hadoop et utilise le gestionnaire des ressources Yarn. En effet, Apache Spark fournit une bibliothèque particulière dédiée aux techniques d'apprentissage automatique, appelée MLlib. Cette bibliothèque propose une implémentation de l'algorithme parallèle FP-growth dans le langage Scala. Par la suite, les données préparées sont analysées, à l'aide de l'algorithme FP-growth de la bibliothèque Spark MLlib, pour profiter d'une implémentation à grande échelle des techniques de des règles d'association. Puis, Spark se connecte à Hadoop HDFS pour stocker les données d'une manière répartie sur un cluster de machines. Ensuite, le système de recommandation génère le catalogue des cours les plus pertinents pour chaque apprenant. Enfin, les résultats du moteur de recommandation peuvent être présentés à l'utilisateur afin de le guider et de lui suggérer les ressources pédagogiques les plus adaptées à ses intérêts. Ainsi, l'apprenant peut parcourir la liste des cours recommandés par notre système et commencer à apprendre ceux qui conviennent à son niveau cognitif et ses préférences. La Figure IV.20 décrit cette architecture en détails.

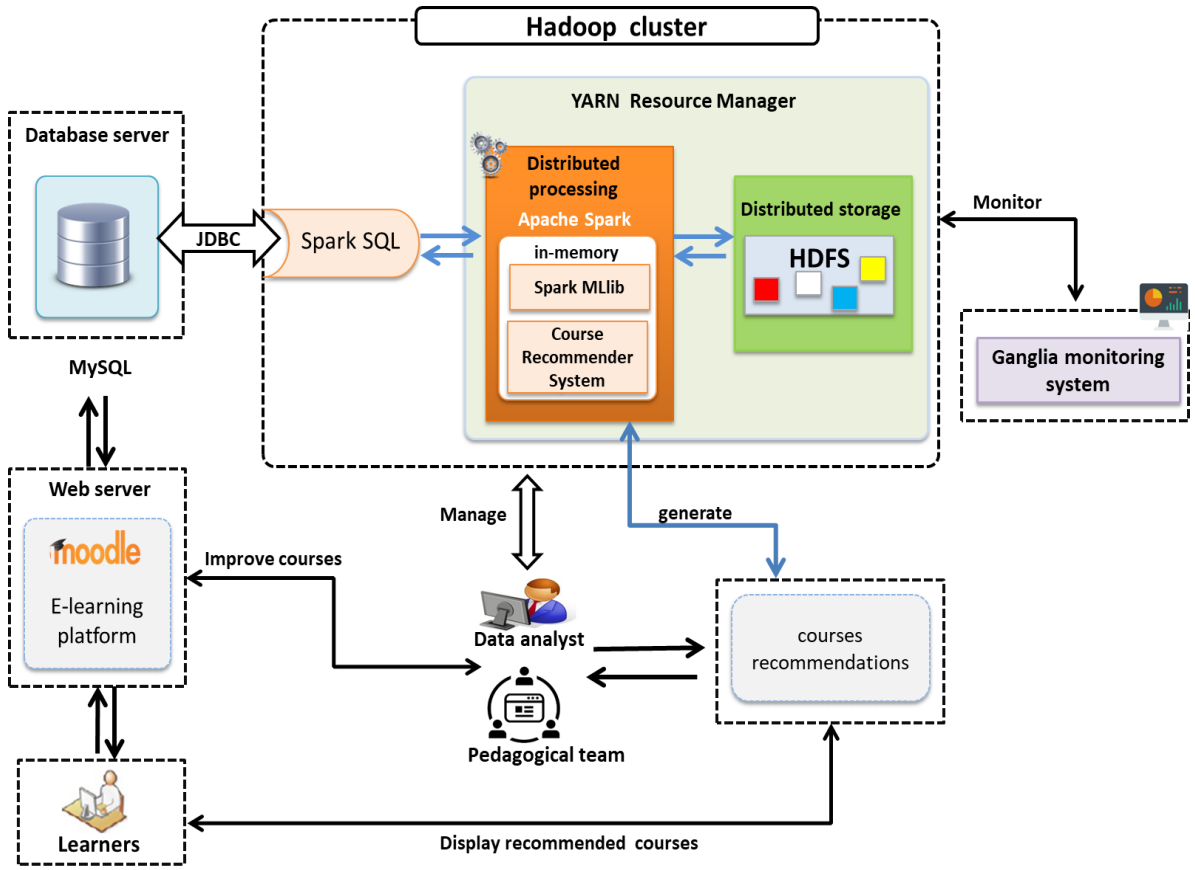


Figure IV.20: Architecture du système de recommandation de cours à grande échelle

Au niveau technologique, notre architecture utilise une combinaison de plusieurs technologies de l'écosystème big data qui communiquent entre elles pour charger, préparer, traiter et stocker les données historiques des cours consultés par l'apprenant. En réalité, il s'agit d'une architecture implémentant, principalement, 3 Framework :

a) **Apache Spark :** un moteur unifié pour le calcul distribué des données collectées à partir de la plate-forme ESTenLigne. Il utilise des interfaces de l'API Spark SQL pour se connecter à la base des données du système e-learning.

b) **Hadoop Yarn** : un gestionnaire des ressources du cluster permettant la planification et l'exécution du système de recommandation sur les nœuds d'un cluster.

c) **Hadoop HDFS** : un système par défaut des fichiers distribués du logiciel Hadoop. Il est responsable du stockage et de la manipulation répartie des données sur un cluster.

L'architecture illustrée dans la Figure IV.21 décrit le mécanisme de fonctionnement du système proposé. Il s'agit d'un environnement distribué composé d'un cluster d'ordinateurs.

En général, Spark est utilisé comme Framework de calcul en parallèle pour développer le moteur de recommandation des cours. Il consiste en un certain nombre d'éléments qui fonctionnent simultanément pour effectuer le travail soumis par un utilisateur sous la forme d'un fichier archivé (jar). L'utilité de ce dernier est de regrouper les différents fichiers relatifs (le code source, connecteurs BDD, dépendances, …) à notre système de recommandation en un seul fichier pour le déployer plus rapidement.

En fait, l'application Spark s'exécute sous la forme d'un ensemble de processus indépendants sur un cluster des noeuds coordonnés par l'objet SparkContext du programme pilote /principal. En outre, les applications Apache Spark se composent de trois composants essentiels : driver process, Spark Master et executor processes. Le driver process est un processus pilote qui conserve les informations pertinentes sur SparkSession pendant toute la durée de vie de l'application Spark. Il exécute également la fonction main (), effectue les transformations et les actions et soumet le travail au Spark Master. Le deuxième composant de cette architecture est Spark Master, qui crée les tâches et les distribue aux nœuds workers. Spark Master assure également la coordination entre les étapes de travail à travers les processus exécuteurs du cluster situés au niveau de chaque nœud. D'autre part, les processus des exécuteurs sont responsables de l'exécution en parallèle du travail en exécutant des tâches sur plusieurs threads et en retournant l'état et les résultats du traitement des données au Master. Lorsque le processus pilote a besoin de ressources pour exécuter un travail, il demande au master d'allouer les ressources requises à partir des différents nœuds du cluster(workers). Ensuite, chaque nœud crée des processus exécuteurs pour effectuer le traitement des tâches. Les nœuds Workers, notamment processus exécuteurs, lisent et écrivent des données depuis et sur le système de fichiers distribués Hadoop (HDFS). Ils stockent également les données transformées en mémoire comme les jeux de données résilients distribués (RDD). Finalement, le processus pilote appelle les processus exécuteurs pour exécuter le travail et restituer les résultats des opérations.

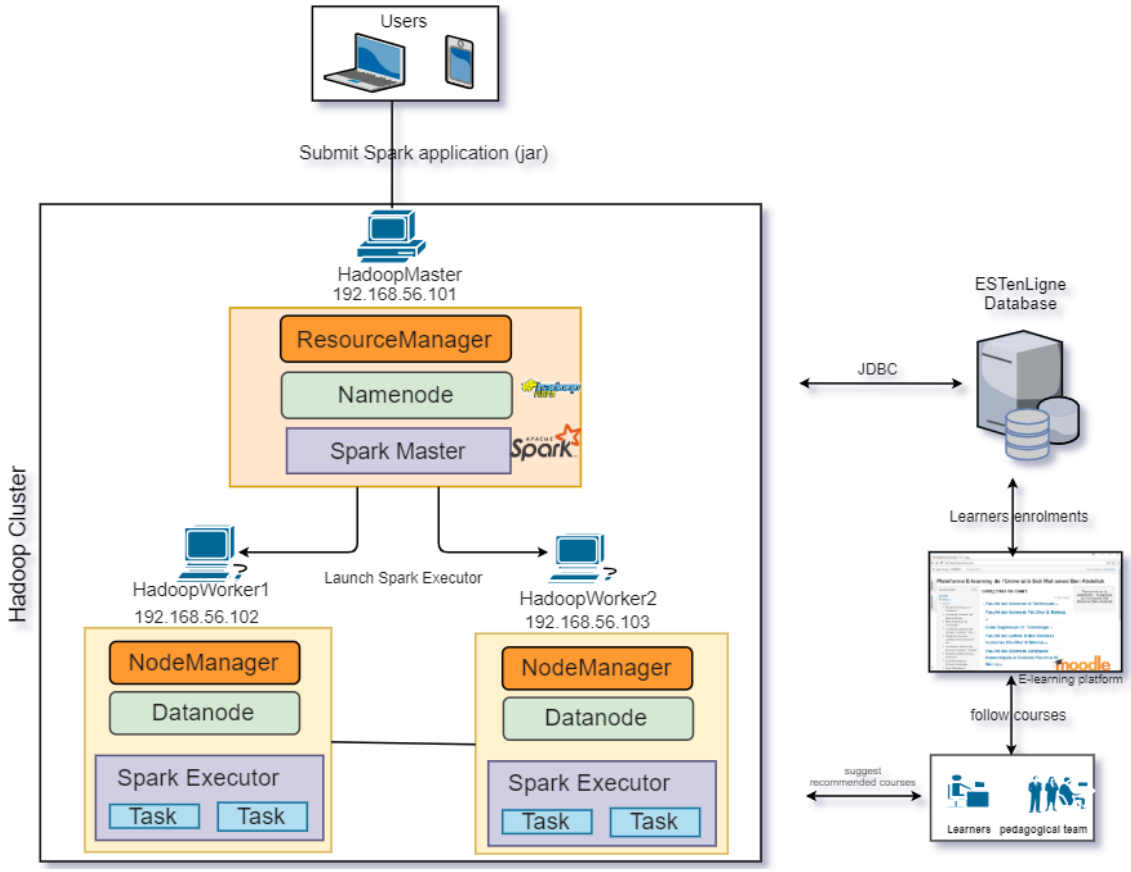


Figure IV.21: Architecture du cluster Hadoop

L'architecture du système proposé consiste en un cluster multi-nœuds. En fait, une machine du cluster est définie en tant que nœud master. Le master a deux composants Namenode pour le stockage distribué (HDFS) et ResourceManager pour le gestionnaire des ressources. Les autres nœuds du cluster sont configurés comme workers (esclaves). Chaque nœud worker est composé de Datanode et NodeManager. Le Datanode exécute les opérations demandées par Namenode. En ce qui concerne le composant NodeManager, il suit les instructions de ResourceManager et gère les ressources disponibles sur un nœud donné. Un autre composant utilisé pour construire cette architecture est le système des fichiers distribués Hadoop qui vise à gérer les données d'apprentissage de notre moteur de recommandation. Ainsi, on trouve Namenode qui est situé au niveau du nœud master. Il conserve le système de nommage des fichiers et répertoires gérés par un cluster Apache Hadoop. En effet, HDFS stocke les données dans des blocs et NameNode conserve et gère les emplacements des blocs de fichiers/répertoires. En d'autres termes, lorsqu'un client souhaite récupérer des données, NameNode fournit les emplacements des blocs de données demandées. De même, lorsqu'un

client veut stocker des nouvelles données, le NameNode retourne l'endroit où ces données peuvent être stockées, alors que, Datanode lit et écrit ces données physiquement dans des fichiers sous forme de blocs. Il est également responsable de la création, de la suppression, et de la réplication des blocs de données sous le contrôle du NameNode.

IV.2.5 Préparation et configuration de l'environnement big data

Vue l'indisponibilité ou le coût élevé de l'installation d'une infrastructure physiques des clusters big data, l'utilisation des outils de virtualisation pour une architecture distribuée représente une solution alternative pour configurer, développer, tester et valider les algorithmes parallèles de l'apprentissage machine sur un environnement informatique dédié au calcul réparti à grande échelle.

Afin de construire notre environnement big data, nous avons préparé un cluster composé de trois machines virtuelles. Comme solution de virtualisation nous avons utilisé la solution libre et populaire VirtualBox. Avec les ressources informatiques disponibles, nous avons créé trois machines virtuelles en installant le système d'exploitation Ubuntu 18.04.1 LTS sur chaque nœud du cluster Hadoop. Ces nœuds sont connectés les uns aux autres à l'aide d'un réseau virtuel privé. La capacité et la configuration de toutes les machines virtuelles sont décrites dans la Table IV.3.

Tableau IV-3: Configuration des nœuds du cluster Hadoop

Machine	Network	Cores	Memory	Disk
Master	192.168.56.101 (master.domain.com)	8 Core i5 (7th Gen)	8 GB	32 GB
Worker1	192.168.56.102 (worker1.domain.com)	8 Core i5 (7th Gen)	8 GB	32 GB
Worker2	192.168.56.103 (worker2.domain.com)	8 Core i5 (7th Gen)	8 GB	32 GB

De plus, la Figure IV.22 montre la configuration détaillée de notre cluster Hadoop via l'interface utilisateur Web du gestionnaire des ressources Hadoop Yarn. Cette interface donne des informations utiles sur l'état du cluster. Autrement dit, elle affiche les nœuds actifs, les nœuds en panne, les machines arrêtées, la mémoire totale disponible, etc.

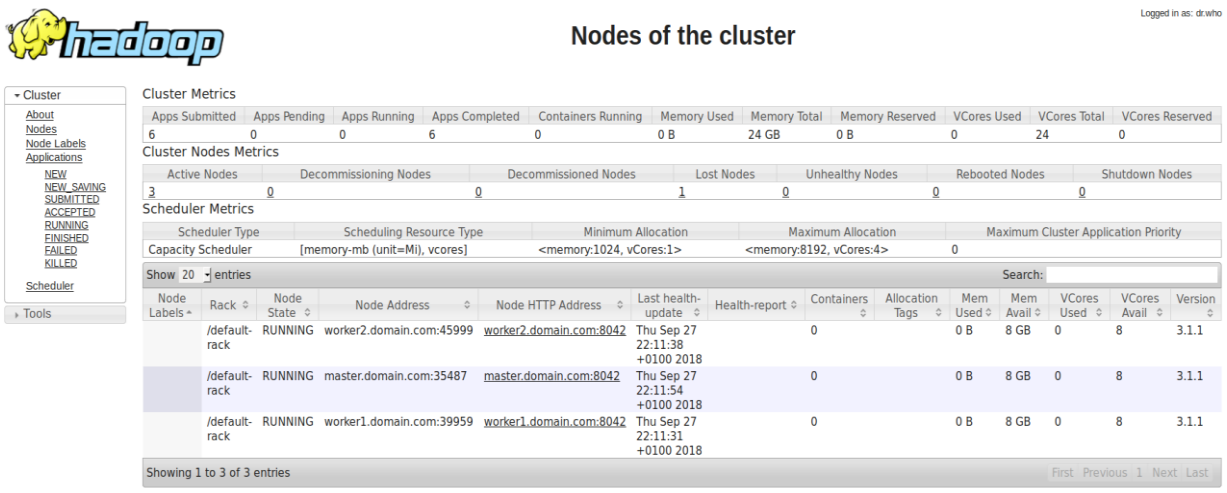


Figure IV.22: Apache Yarn Web UI : Configuration des nœuds du cluster

Après avoir terminé la configuration du réseau virtuel des clusters, nous avons décompressé et installé la plate-forme Hadoop 3.1.1 et le Framework Spark 2.3.1 au niveau du nœud master du cluster. Ensuite, nous avons transféré le dossier d'installation des deux Frameworks dans les nœuds workers à l'aide du protocole de communication sécurisé SSH. De même, nous avons employé la puissance de la fonction Secure Copy (SCP) afin d'obtenir la même copie d'Apache Hadoop et Spark sur les différents nœuds de notre cluster. La version Java 1.8.0_181 a été installée sur chaque nœud et nous avons configuré une SSH sans mot de passe entre les nœuds pour que le nœud master Hadoop puisse se connecter, démarrer, arrêter et exécuter des tâches dans les différents workers.

IV.2.6 Application distribuée Spark

IV.2.6.1 Développement

Dans Spark, une application doit être configurée pour s'exécuter sur un cluster Spark responsable de l'exécution du travail soumis par le data scientist. Dans le développement du système de recommandation de cours, le cluster Spark est remplacé par le cluster Hadoop. D'autre part, le code source du programme est écrit en scala, le langage de programmation natif de Spark Framework. De plus, l'application est organisée et gérée à l'aide du gestionnaire de dépendance Maven en tant qu'outil de construction des projets (build) open source de la fondation Apache. Notre système de recommandation a été développé en utilisant l'IDE Eclipse 4.7.0 (Oxygen) fournissant le plug-in Scala 2.12.6.

Pour mettre en œuvre l'algorithme FP-growth parallèle (PFP), l'application Spark développée avait besoin de deux API Spark : Spark SQL (org.apache.spark.sql) pour se connecter à la source des données et MLlib (org.apache.spark.ml) pour appliquer les techniques d'analyse avancées et des algorithmes d'apprentissage automatique. Avant de lancer l'exécution de notre modèle analytique, Il fallait préparer les données requises en collectant des données à partir de la plate-forme ESTenLigne. Pour cela, nous avons développé une requête SQL afin d'extraire la liste des cours suivis par tous les apprenants. En effet, nous nous sommes concentrés sur quatre tables : *mdl_user*, *mdl_user_enrolments*, *mdl_enrol* et *mdl_course*. En plus, nous avons utilisé la bibliothèque Spark SQL qui permet d'exécuter facilement des instructions SQL. Ensuite, nous avons dû classer l'historique des cours par les IDs des étudiants. En effet, les cours individuels suivis par chaque apprenant sont regroupés dans un seul vecteur sous forme de tableau des cours qui représentent les données d'entrée de l'algorithme PFP-growth. Il est important de diviser le jeu des données préparées en une base des données d'apprentissage (70%) et une autre pour le test (30%) pour évaluer les performances du notre système. Les données d'apprentissage sont utilisées pour construire le modèle des règles d'association afin d'identifier les éléments fréquents, tandis que la base des données de test vise à évaluer le modèle en comparant les éléments d'entrée par rapport à toutes les règles d'association et à récapituler les conséquences en tant que prédictions ou recommandations.

IV.2.6.2 Déploiement

Lancer Spark en cluster signifie la distribution du Job afin d'être exécuter sur un cluster d'ordinateurs distants. Pour ce faire, l'application Spark doit être packagée à travers l'IDE eclipse qui permet de créer un package d'assemblage (assembly package) contenant le code source et toutes les dépendances de notre système de recommandation. En pratique ce package prend la forme d'un fichier 'jar'. Pour exécuter le package jar, nous avons exploité la puissance du script spark-submit à partir de la ligne de commande, ce qui permet de soumettre des tâches à un cluster distant composé de trois nœuds interconnectés entre eux par l'intermédiaire d'un réseau virtuel. En fait, spark-submit est responsable du lancement et de l'exécution du code source de d'application spark sur l'infrastructure Hadoop gérée par le gestionnaire des ressources Yarn.

Script d'exécution du système de recommandation de cours
bin/spark-submit --master yarn --deploy-mode cluster --class org.karim.spark.sparkmaven.FpGrowth --executor-memory 2G --num-executors 12 'PFP-growth-jar-with-dependencies.jar'

Figure IV.23: Script d'exécution du système de recommandation de cours

La Figure IV.23 décrit en détail le script d'exécution de notre système de recommandation sur le cluster Hadoop. Comme nous pouvons remarquer, nous avons créé 12 exécuteurs répartis sur les différents nœuds du cluster, et la taille de la mémoire allouée à chaque exécuteur est 2Go. Autrement dit, chaque machine du cluster contient 4 exécuteurs et 8 Go de mémoire en total.

IV.2.7 Résultats

Pour exécuter l'algorithme PFP-growth, nous avons dû spécifier les seuils minimales de support et de confidence afin de trouver les corrélations les plus fortes entre les inscriptions aux cours dans les traces des activités d'apprentissage. Le nombre des règles d'association intéressantes change en fonction de la valeur du support, de la confidence et de la taille de la base de données. Ainsi, nous avons utilisé le seuil minimal de support de 5% et nous avons fixé 60% comme seuil minimal de confidence. En fait, le système de recommandation de cours génère trois types de résultats qui satisfont aux critères de support et de confidence spécifiés. Premièrement, il trouve la liste des cours fréquents dans la base des données du système e-learning en s'appuyant sur le calcul du support, comme indiqué dans la Table IV.4 qui montre les 10 ensembles des cours (itemsets) les plus fréquents.

Tableau IV-4: Résultats de l'algorithme PFP-growth: Itemset fréquents

Items	Fréquence
[43]	167
[11]	128
[9]	125
[42]	123
[14]	93
[15]	88
[45]	87
[46]	84

[9, 43]	80
[46, 45]	80

Deuxièmement, à base du calcul de la confidence de chaque itemset, notre solution de recommandation garde uniquement la liste des itemsets qui satisfont la condition du seuil minimal de confidence. Cette liste représente l'ensemble des règles d'association fortes extraites à partir de l'historique des inscriptions des apprenants. Les 10 principales règles intéressantes, ordonnées selon la mesure de confiance, sont illustrées dans la Table IV.5.

Tableau IV-5: Résultats PFP-growth: règles d'association

Rule	Antécédent	Conséquent	Confidence
[1]	[46, 11]	[45]	1.000
[2]	[45, 11]	[46]	0.975
[3]	[46]	[45]	0.952
[4]	[45]	[46]	0.919
[5]	[7, 6]	[18]	0.868
[6]	[7]	[18]	0.818
[7]	[6, 18]	[7]	0.785
[8]	[7, 18]	[6]	0.733
[9]	[6]	[18]	0.711
[10]	[7]	[6]	0.690

Selon les résultats obtenus dans la Table IV.5, la règle d'association entre les cours {11 et 46} et {45} a la plus grande confidence. De même, la règle d'association entre les cours {7} et {6} a la plus basse confidence. À la base des valeurs calculées et de support et confidence, nous pouvons clairement déterminer les cours les plus susceptibles d'être suivis par les apprenants, et les plus pertinents à leurs recommander. Par exemple, la règle 1 {11, 46} => {45} a la plus grande confidence, c'est-à-dire la plus forte. Notre système suggère donc le cours 45 aux étudiants qui se sont déjà inscrits aux cours {11 et 46}. Pour les 10 règles d'association les plus fortes, nous remarquons que les valeurs de confidence se situent entre 0,69 (69%) et 1,00 (100%), ce qui démontre que nous avons obtenu de bons résultats. Nous pouvons donc en conclure que le solution de recommandation développée propose des ressources pédagogiques pertinentes en recommandant des cours adaptés au profil et au niveau cognitif de chaque étudiant afin de le guider au cours de son cursus d'apprentissage à prendre les bonnes

décisions, et de mettre à sa disposition des contenus éducatifs plus ciblés et plus appropriés à ses intérêts et ses besoins.

Après avoir établi le modèle de recommandation à l'aide de l'ensemble des données d'apprentissage (training dataset) (70%), il peut également être utilisé pour prédire le résultat de recommandation des cours aux apprenants à travers la base des données de test (test dataset) (30%). Le système de recommandation génère un catalogue des cours pour chaque profil d'apprenant. La Table IV.6 montre les 10 principales prédictions sur un total de 48 prédictions.

Tableau IV-6: Résultats de prédiction du système de recommandation de cours

Learner ID	Items (cours)	Prediction (recommandation)
541	[46, 6, 11, 45]	[18, 7]
720	[45, 46, 15]	[14]
19	[6]	[18, 7]
277	[18, 14, 6, 9,	[43]
287	[17, 43, 42, 9,	[6]
155	[15, 17, 42, 9]	[14, 43]
1157	[6]	[18, 7]
184	[40, 6, 43, 42]	[18, 7]
274	[7, 6, 15, 4, 3,	[43]
766	[15, 45, 46]	[14]

Les résultats de la prédiction du système de recommandation que nous avons développé, donne pour chaque apprenant identifié par ID, la liste des cours auxquels il participe et un catalogue des prédictions des cours. Par exemple, notre système de recommandations des cours suggère les cours 18 et 7 à l'apprenant (id = 541) qui est déjà inscrit aux cours 46, 6, 11 et 45. Également, il recommande le cours 14 à l'apprenant (id = 720) qui suit les cours 45, 46 et 15 cours, etc.

IV.2.8 Performance du système de recommandation de cours

Dans cette section, Nous établissons une analyse comparative du temps d'exécution du système de recommandation de cours en se basant sur trois environnements différents, à savoir : Apache Spark, R et Weka afin de démontrer l'efficacité et l'évolutivité de notre solution à grande échelle. Pour cela, nous avons augmenté le nombre des transactions afin

d'obtenir une taille de base des données suffisamment grande. Ensuite, nous avons mesuré la vitesse de calcul de l'algorithme FP-growth en utilisant la bibliothèque Spark MLlib de Spark par rapport au même algorithme sous le logiciel Weka. Nous avons mesuré également les performances de notre système à l'aide du logiciel RStudio. En effet, nous avons comparé la durée d'exécution de l'algorithme FP-growth dans le cluster Hadoop à celle d'une infrastructure standard et centralisée composée d'un seul ordinateur (Weka et RStudio). Les résultats sont représentés dans la Figure IV.24 qui illustre une représentation graphique en courbes des temps d'exécution obtenus dans différents environnements.

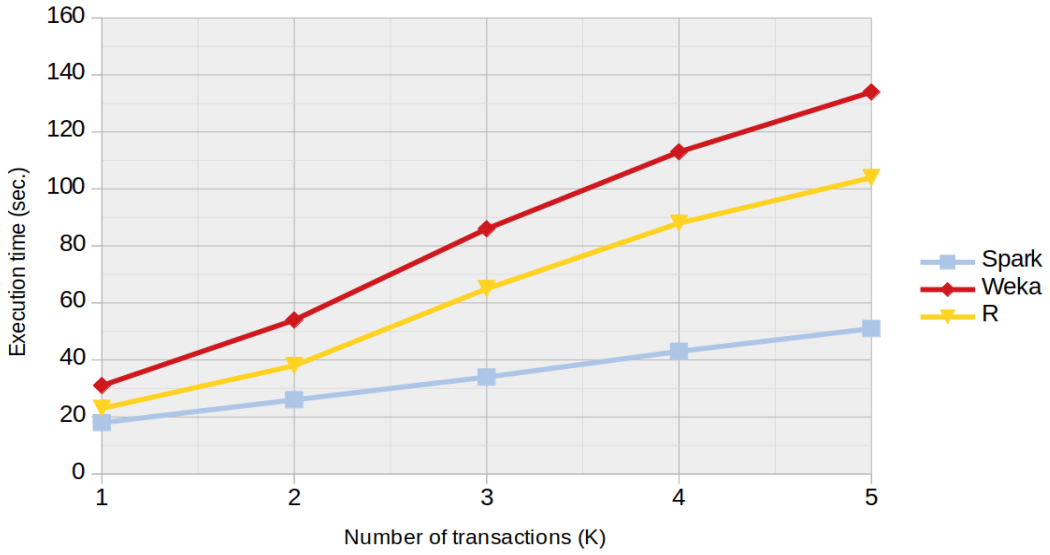


Figure IV.24: Etude comparative des performances entre Spark, Weka et RSStudio

Comme on peut le constater à partir du graphique en courbes, le modèle des règles d'association implémenté en utilisant le Framework Apache Spark est plus rapide que celui l'outil Weka et l'environnement RStudio. En utilisant Spark, il ne faut que 34 secondes pour traiter 3 000 transactions, alors que Weka prend 86 secondes et RStudio 65 secondes. L'algorithme FP-growth parallèle de Spark Mllib atteint donc une bonne performance et excellente évolutivité grâce au mécanisme du calcul, réparti sur les nœuds distribués du cluster. SparkMLlib traite les données plus rapidement car il divise le travail en plusieurs tâches élémentaires exécutées sur des nœuds workers, ce qui accélère le temps nécessaire pour extraire les prédictions des cours recommandés. À base de l'analyse ci-dessus, nous pouvons conclure que Spark représente la meilleure plate-forme pour implémenter le système de recommandation de cours développé dans ce travail de recherche.

IV.2.9 Outils de monitoring ou supervision

Le monitoring ou supervision est une stratégie de surveillance permettant de contrôler et de vérifier le bon fonctionnement d'une infrastructure informatique. Il vise à faciliter la gestion et la maintenance préventive des services et des ressources des serveurs physiques ou virtuels. Les outils de monitoring ont pour but de mesurer la disponibilité et la performance des services afin d'anticiper les dysfonctionnements et des pannes matérielles ou logiciels. Dans notre cas, il s'agit d'une infrastructure de cluster composé d'un certain nombre de nœuds interconnectés par un réseau privé. Dans un premier temps, nous allons présenter la supervision d'exécution du système de recommandation proposé, qui se déroule en plusieurs étapes. Ensuite, nous allons aborder la surveillance de l'utilisation des ressources de notre environnement de cluster, telles que de la bande passante, la mémoire, le CPU et le stockage.

IV.2.9.1 Spark UI

Plusieurs outils de supervision sont utilisés afin de nous aider à visualiser la progression de l'exécution des travaux soumis par un data scientist, et à comprendre comment l'application Spark s'exécute sur une architecture de cluster Hadoop. Spark UI[168] est un outil de monitoring qui permet de suivre le processus d'exécution du traitement réparti d'un Job sur une infrastructure distribuée. Il fournit des informations importantes sur l'exécution de l'application Spark à travers des interfaces intuitives et responsives, notamment les travaux (Jobs), les étapes (Stages) et les tâches (Tasks). En outre, il affiche des variables d'environnement et permet la visualisation DAG (Graphe Acyclique Direct) de l'application Spark une fois terminée. Pour visualiser le comportement d'un Job, on utilise l'adresse http://master:18080 sur un navigateur Web.

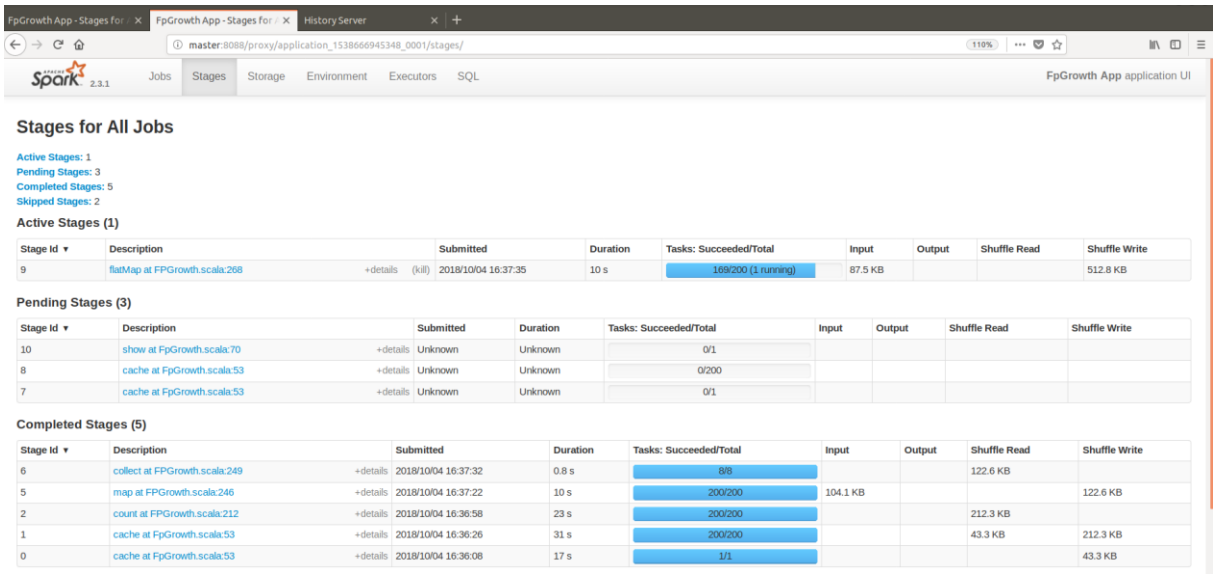


Figure IV.25: Spark Web UI

Comme nous pouvons le voir sur la Figure IV.25, une étape peut avoir l'une des trois niveaux : active, en attente et terminée. De plus, nous avons la possibilité d'afficher une vue chronologie des événements Spark. La Figure IV.26 montre l'ordre de déroulement, dans le temps, des opérations telles que la création exécuteurs et le lancement des tâches d'analyse des données stockées en mémoire sous forme d'abstraction RDD .

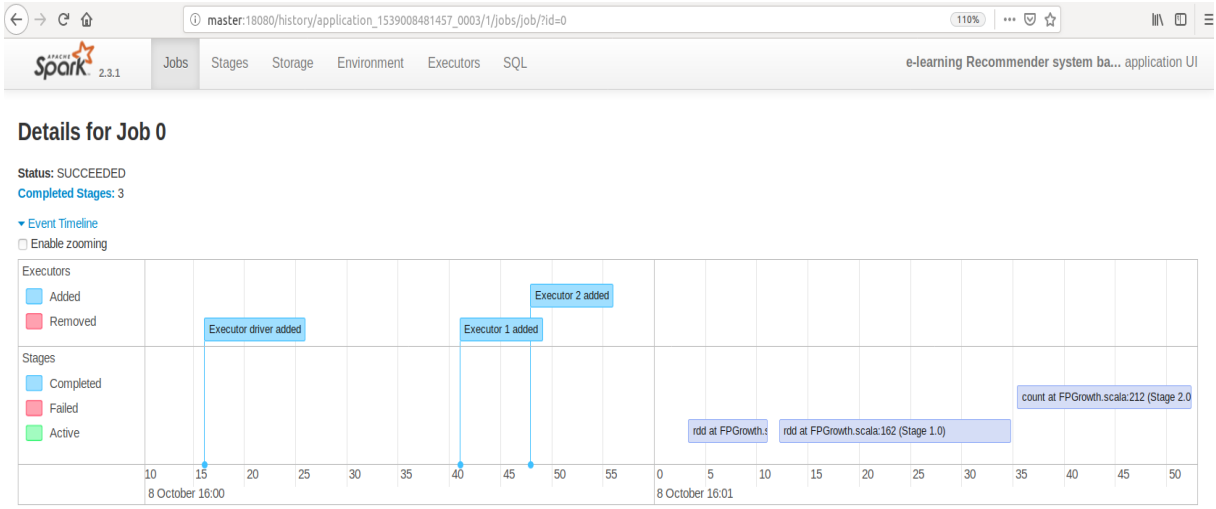


Figure IV.26: Vue chronologique des événements Spark

Les techniques de visualisation fournies par Spark UI permettent également de montrer le DAG d'exécution pour chaque job. Dans Spark, un job est l'exécution d'un travail soumis (chaine de traitement) dans un environnement distribué. Il est généralement découpé en

plusieurs étapes (Stages) dont chacune peut avoir un ensemble de tâches. Un job est organisé en DAG (graphe acyclique direct), qui ressemble à ce qui suit (Figure IV.27).

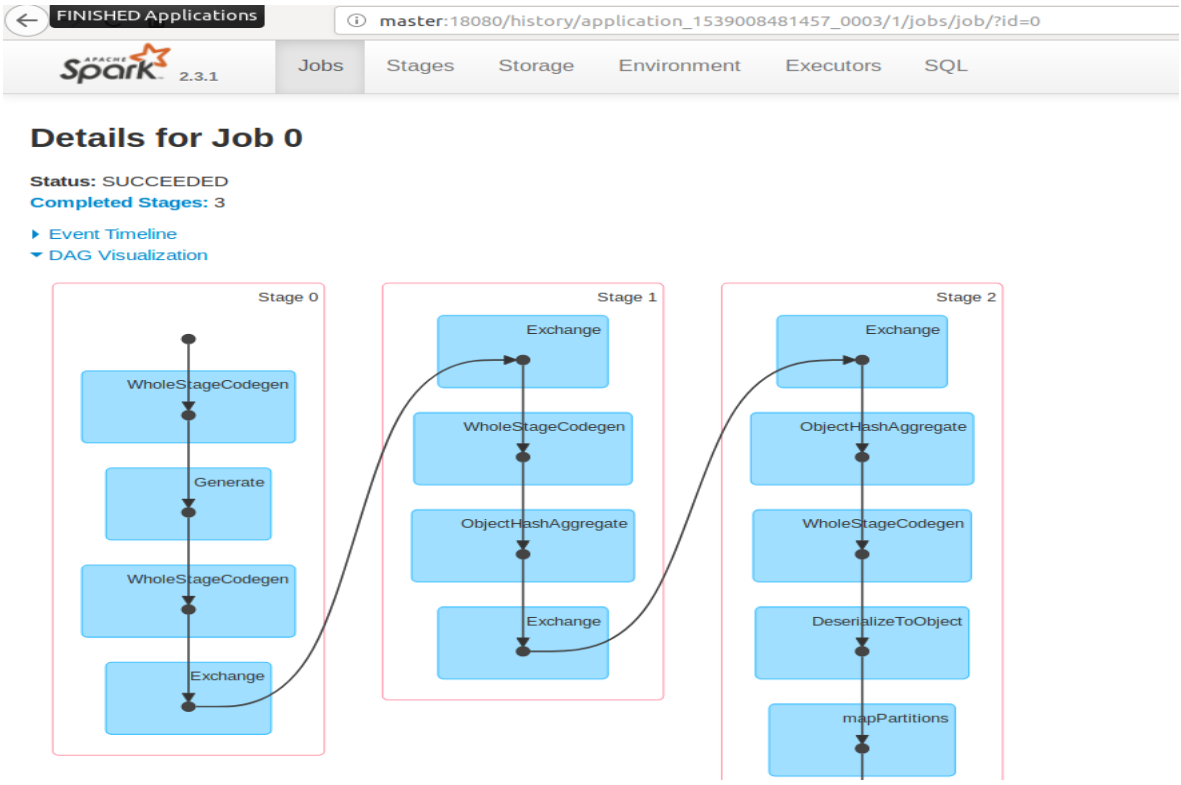


Figure IV.27: Direct Acyclic Graph

IV.2.9.2 Yarn UI

Hadoop fournit une interface utilisateur Web pour HDFS et une autre pour Yarn. La progression d'un Job peut être surveillée via l'interface utilisateur Web [169] (Figure IV.28) dans un mode entièrement distribué de cluster, disponible à l'adresse http://master:8088.

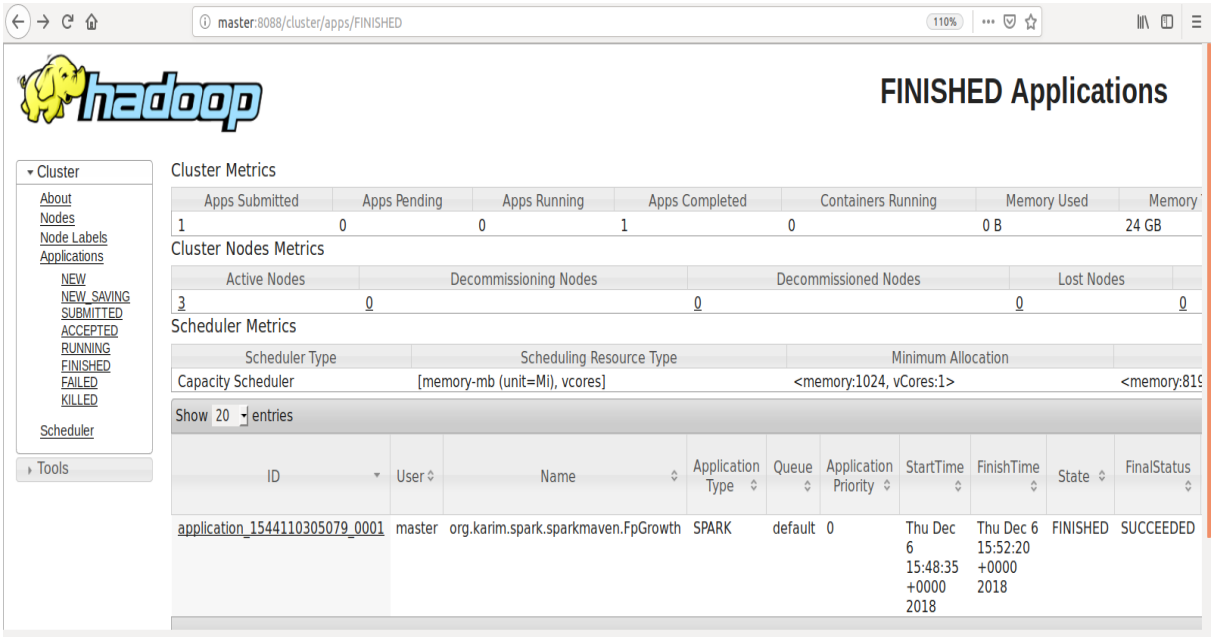


Figure IV.28: Yan Web UI

Nous pouvons observer l'état de chaque Job soumis, s'il est terminé avec succès ou non. Ainsi, Yarn Web UI donne la possibilité d'arrêter un Job pendant son exécution. En plus, s'il y a des tâches qui ne s'exécutent pas correctement, elles seront répertoriées dans la section des tâches échouées (Failed). De même il a un volet consacré aux Jobs exécutés avec succès (Finished).

IV.2.9.3 Interface web de HDFS

Pour faciliter la supervision et le contrôle du système HDFS, Hadoop offre une interface web simple à utiliser permettant de consulter l'état des nœuds d'un cluster. La Figure IV.29 donne un aperçu de l'interface web de HDFS accessible depuis l'URL : http://master:9870/

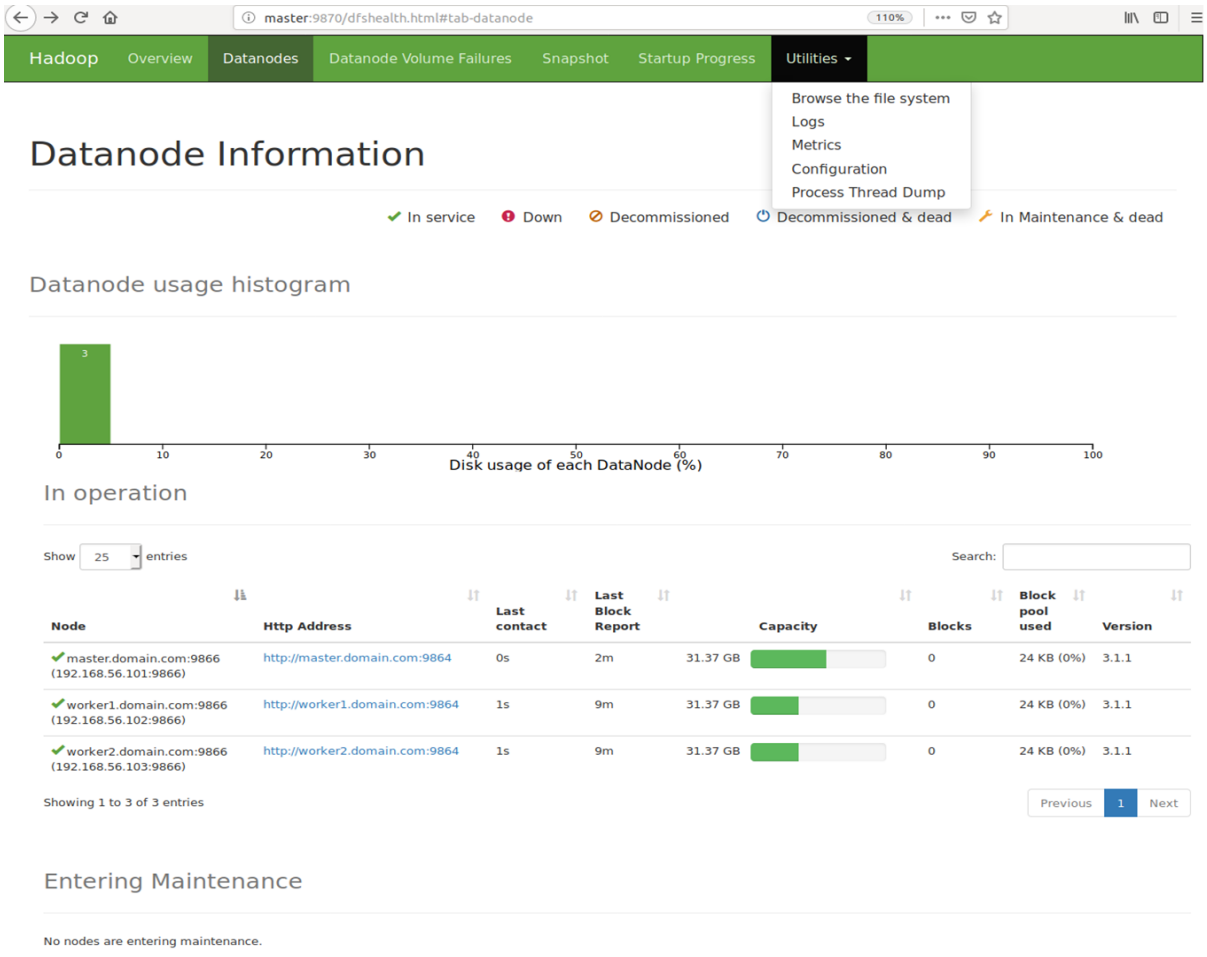


Figure IV.29: Interface web de HDFS

Cette interface web permet d'afficher plein d'informations concernant les trois nœuds opérationnels (In service). Au cas où il y a des nœuds en pannes, elle les affiche dans le volet dédié (Down), la même chose pour les nœuds en état de maintenance (In Maintenanc & dead). En plus,l'interface HDFS fournit des détails pertinents relatifs à la capacité de stockage utilisée et libre de chaque nœud, ainsi que l'url pour y accéder.

IV.2.9.4 Ganglia monitoring

Ganglia[73] est un outil open source, distribué et évolutif permettant de surveiller des grands environnements informatiques tels que des clusters et grids[22]. Pour l'infrastructure big data, il s'agit d'une solution très puissante permettant de mesurer les performances et la

[22] grids : traduit en français par Grille, est un ensemble de cluster ou serveurs hétérogènes et délocalisées

consommation des ressources comme l'utilisation du processeur, de la mémoire et du stockage de chaque nœud du cluster. Il permet également de contrôler et visualiser le trafic d'un réseau, comme l'utilisation de la bande passante ou le volume des données transportées via le réseau. De surcroît, il fait de manière claire et efficace une synthèse de l'utilisations des ressources au cours de la dernière heure, jour, semaine, mois ou année. Les Figure IV.30 (version Web) et IV.31 (version mobile) montrent la page principale de l'interface Ganglia. Elle donne un aperçu interactif des mesures de la consommation du CPU, de la mémoire et du réseau de trois nœuds de notre cluster Hadoop.

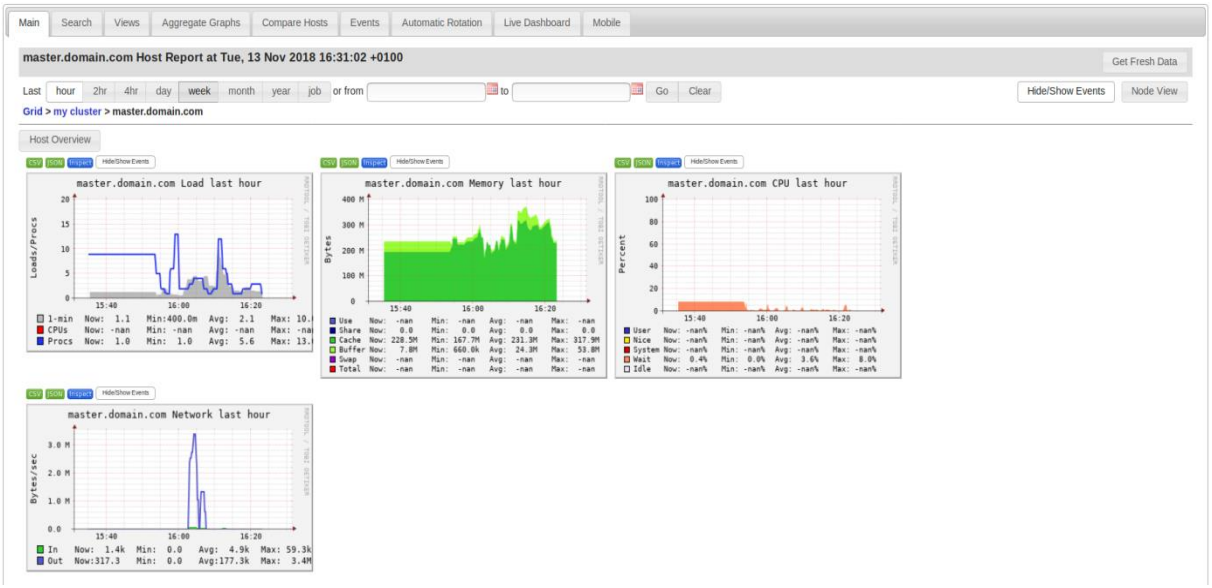


Figure IV.30: Interface web Ganglia

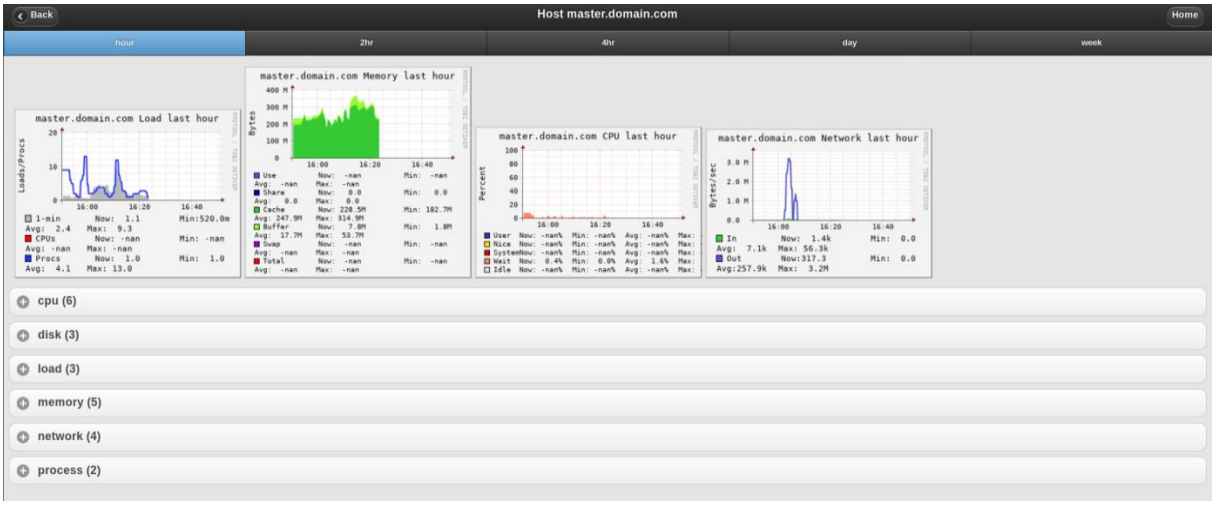


Figure IV.31: Interface mobile Ganglia

Pour pouvoir visualiser les performances de chaque machine du cluster Hadoop, il faut, tout d'abord, choisir un nœud parmi la liste des choix de tous les nœuds détectés sur le cluster Hadoop, comme il est montré dans la Figure IV.32.



Figure IV.32: Les nœuds du cluster Hadoop supervisés par Ganglia

Puis, On peut accéder à l'interface Ganglia consacré à ce nœud et consulter les performances des ressources de la machine sélectionnée. Ainsi, on peut par exemple afficher la quantité des données transférées sur un réseau, comme il est illustré dans la Figure IV.33.



Figure IV.33: Supervision du réseau via Ganglia

Ganglia permet de visualier la bande passante de l'hôte sélectionné, y compris les octets ou les paquets envoyés et reçus. En fait, cet outil fournit des statistiques sur le nombre d'octets

reçus et envoyés par le nœud master. Il fournit, même, des informations importantes sur la taille des paquets.

D'ailleurs, comme le montre la Figure IV.34, Ganglia offre une vue récapitulative de l'ensemble des informations essentielles sur les ressources matérielles et logicielles d'une machine dans une interface simple et résumée,.

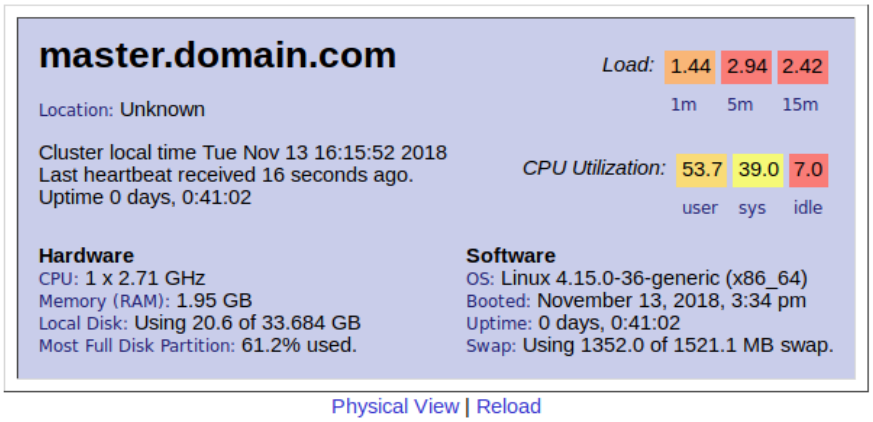


Figure IV.34: Interface récapitulative d'un nœud via le système Ganglia

IV.2.10 Conclusion

Dans ce chapitre, nous avons présenté le système de recommandation de cours à grande échelle et son implémentation dans une architecture entièrement distribuée. Le but étant de recommander à un apprenant un catalogue des cours plus adaptés et plus personnalisés selon le profil et le niveau cognitif de chaque apprenant. L'infrastructure du cluster Hadoop a été mise en place pour implémenter et déployer notre système de recommandation qui a été développé et testé en utilisant le Framework open source Apache Spark permettant le traitement intensif en mémoire. L'implémentation du moteur de recommandation des cours est faite à l'aide de l'algorithme parallèle FP-growth fourni par la bibliothèque d'apprentissage automatique Spark MLlib. Les résultats obtenus montrent que, l'approche de décentralisation du traitement et de l'analyse des données du big data offre des meilleures performances tout en réduisant le coût et le temps du calcul par rapport aux logiciels classiques de data mining (Weka) ou d'outils statistiques (RStudio). D'ailleurs, le système présenté peut être appliqué et facilement intégré à d'autres environnements d'apprentissage tels que MOOC (Massive Open Online Courses). En outre, nous avons déployé un certain nombre d'outils du monitoring afin de superviser le déroulement d'exécution des traitements répartis et de contrôler et maintenir l'état de l'environnement distribué du cluster Hadoop.

Conclusion générale

Le travail de recherche présenté dans cette thèse apporte des solutions conceptuelles, techniques et technologiques visant l'intégration continue des nouvelles technologies de l'information et de la communication dans l'enseignement à distance (NTIC). Il s'intéresse, plus spécifiquement, à l'intégration des paradigmes du cloud computing et du big data dans les systèmes e-learning en vue d'offrir un service d'apprentissage de meilleure qualité. Cette conclusion est l'occasion pour rappeler et résumer nos différentes contributions dans le champ de recherche des EIAH et de présenter quelques perspectives pour les travaux de recherche futurs.

Au début, pour situer notre recherche dans la littérature, nous avons commencé par une présentation de l'évolution des EIAH, à partir de leur apparition jusqu'au aujourd'hui tout en décrivant les grandes périodes qui ont marqué leur histoire. Ensuite, nous avons défini ces environnements dans leurs diverses formes, y compris l'apprentissage à distance, l'e-learning, le m-learning et les MOOCs. Nous avons présenté également les différents travaux de recherche qui ont été effectués dans ce domaine.

Pour intégrer les systèmes de l'e-learning dans le cloud, nous avons mené une étude approfondie autour du modèle cloud et du l'écosystème big data. Puis, nous avons proposé une nouvelle approche pour migrer l'infrastructure traditionnelle des systèmes d'apprentissage en ligne vers l'architecture flexible de cloud. Cette intégration permet aux universités et établissements d'enseignement de bénéficier d'une large palette des avantages offerts par le cloud tels que la disponibilité permanente, la mutualisation de ressources, le paiement en fonction de la consommation, etc. Comme deuxième axe de recherche de cette thèse, nous avons proposé également un modèle e-learning incorporant les technologies big data en se basant sur les services du cloud. Ce modèle comprend quatre technologies big data : les systèmes de fichiers distribués, les bases de données NoSQL, les outils et Frameworks de traitement distribué et d'analyse prédictive, en plus des services d'infrastructure, de plate-forme et logiciels fournis par le cloud. Ainsi, nous avons décrit une méthodologie big data adaptée au contexte d'EIAH permettant, à partir des sources de données produites par les plateformes d'apprentissage en ligne, de créer de la valeur via des techniques d'extraction de

connaissances afin d'aider les professionnels de l'enseignement à distance à prendre des bonnes décisions. Nous présenté la démarche à suivre pour bien manipuler les données d'apprentissage, allant de l'acquisition, la découverte, la préparation, la modélisation, le traitement, jusqu'à la visualisation et l'interprétation des résultats de l'analyse.

Dans cette thèse, notre idée de base était de remplacer l'approche classique centralisée par une autre entièrement distribuée. Dans ce sens, nous avons proposé une architecture distribuée pour exploiter et valoriser les données provenant des systèmes de l'e-learning en combinant plusieurs technologies big data, telles que HDFS, Sqoop, Flume et Yarn, qui communiquent entre eux via des protocoles réseau, en échangeant des données lors de l'exécution d'une application distribuée dans un environnement de cluster Hadoop. Cette architecture permet de répartir la charge du travail et les données sur plusieurs nœuds d'un cluster dans le but de réaliser un partage des ressources. L'approche de décentralisation du traitement intensif sur des unités de calcul réparties permet d'obtenir les meilleurs résultats et d'analyser un volume de données très important.

Pour mettre en œuvre les technologies big data dans les systèmes e-learning, nous avons conçu et développé un système de recommandation de cours en vue d'offrir un service e-learning adaptatif, qui consiste à adapter les ressources pédagogiques aux préférences et besoins individuels de chaque apprenant. Pour implémenter notre moteur de recommandation des cours, nous avons fait appel aux techniques d'apprentissage automatique, en particulier la technique des règles d'association permettant de découvrir toutes les relations intéressantes à partir des données de traces d'apprentissage des apprenants. La réalisation d'un tel système a été effectuée en utilisant la base des données de la plate-forme ESTenLigne de l'Ecole Supérieure des Technologies de Fès. Les résultats du système développé ont démontré la bonne qualité et la pertinence des cours recommandés.

Afin d'améliorer les performances de notre système de recommandation de cours, nous avons développé une version à grande échelle de ce système à l'aide du big data. Pour déployer ce dernier, nous avons utilisé l'infrastructure distribuée de la plate-forme open source Hadoop et la puissance du traitement en mémoire et d'analyse avancée du Framework Spark. L'application Spark implémentée est basée sur une approche tout à fait différente qui consiste, en effet, à répartir le traitement et les données sur plusieurs nœuds dont chacun effectue des

tâches spécifiques afin d'exécuter l'algorithme FP-growth parallèlement sur différentes machines d'un cluster. Le mécanisme de distribution du calcul et du stockage permet de résoudre le problème de limitation des ressources disponibles tout en accélérant la vitesse d'exécution et le coût de traitement. Les résultats obtenus montrent l'efficacité de l'approche de décentralisation du traitement et de l'analyse des données, car cette approche donne des meilleures performances par rapport aux logiciels classiques de data mining ou d'outils statistiques.

Le travail présenté ici a montré certains avantages de l'intégration des Frameworks Spark et Hadoop pour améliorer les environnements d'apprentissage en ligne. Cependant, il reste encore des technologies big data et des modèles d'apprentissage automatique à mettre en œuvre. Comme continuité des travaux de recherche présentés dans cette thèse, nous envisageons de lancer des activités de recherche sur les thématiques suivantes :

- La mise en œuvre de la technologie Spark Streaming dans le domaine de l'e-learning pour bénéficier de l'analyse en temps réel des données d'apprentissage. De cette façon, les professionnels de l'apprentissage en ligne peuvent évaluer rapidement l'efficacité des stratégies d'apprentissage et prendre des bonnes décisions pédagogiques.
- L'implémentation d'autres méthodes du big data dédiées à la prévision du comportement des apprenants et de leurs activités.
- L'adaptation et la personnalisation de notre système de recommandation de cours avec les MOOCs.
- Le développement d'un système de recommandation en se basant sur les techniques du filtrage collaboratif actif de type Ranking systems qui utilise des évaluations ou des avis des apprenants sur les cours pour générer des recommandations.

Références

[1] E. Bruillard, *Les Machines à enseigner*. Paris: Hermès, 1997.

[2] F. Winckel et D. Amaury, « L'enseignement à distance en mutation : diagnostic et perspectives en Communauté française de Belgique », p. 2, 2003.

[3] « Distance education and industrial production: a comparative interpretation in outline (1967) », *Taylor & Francis*. [En ligne]. Disponible sur: https://www.taylorfrancis.com/. [Consulté le: 21-déc-2018].

[4] M. Jacques, « Le recours aux formations à distance (e.learning) dans la formation professionnelle des salariés », 2014.

[5] HAS, « E-learning - Guide de conception de formation ouverte et à distance (FOAD) dans le monde de la santé », *Haute Autorité de santé*, 2015.

[6] « L'intelligence artificielle en éducation ». [En ligne]. Disponible sur: https://ledidacticien.com/2016/09/02/ia-en-education/. [Consulté le: 21-déc-2018].

[7] W. J. Clancey, « Model construction operators. Artificial lntelligence », p. 53, 1992.

[8] A. N. T. Ange, « Développement d'un système tutoriel intelligent pour l'apprentissage du raisonnement logique », Université du Québec à Montréal, 2016.

[9] P. Mendelsohn et P. Dillenbourg, « Le Développement de l'Enseignement Intelligement Assisté par Ordinateur (13-Dec-1995) », 1991. [En ligne]. Disponible sur: http://tecfa.unige.ch/tecfa/publicat/mendel-papers/apslf92.html. [Consulté le: 21-août-2018].

[10] J. Self, « Cognitive Diagnosis for Tutoring Systems », *Actes de European Conference on Artificial Intelligence (ECAI)*, p. 699-703, 1992.

[11] M. Swartz, « Intelligent Tutoring Systems for Foreign Language Learning », *Springer*, 1992. [En ligne]. Disponible sur: https://www.springer.com/gp/book/9783642772047. [Consulté le: 21-déc-2018].

[12] M. ARON, R. GRA, et J. NICAUD, Deuxièmes journées EIAO de Cachan, Ecole normale supérieure, janvier 1991. Ecole normale supérieure, 1991.

[13] R. Zaafrani, « Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère », Lyon II, 2002.

[14] « Les principes clés du E-Learning », *Le blog du Digital Learning*, 14-mars-2014. .

[15] Collectif De Chasseneuil, Philippe Carré, Carine Degallaix, *Accompagner des formations ouvertes*, Conférence de consensus. L'Harmattan, 2001.

[16] J. Frayssinhes, « La Révolution Numérique : Les nouveaux paradigmes de l'apprentissage des adultes | Implications philosophiques ». .

[17] M. West et S. Vosloo, « Principes directeurs de l'UNESCO pour l'apprentissage mobile », p. 41, 2013.

[18] M. Sharples, « Big Issues in Mobile Learning - How can we address the conflicts between personal informal learning and traditional classroom education », United Kingdom, 2006.

[19] C. O'Malley et al., Guidelines for learning/teaching/tutoring in a mobile environment. 2005.

[20] F. Soualah-Alila, « CAMLearn: Une Architecture de Système de Recommandation Sémantique Sensible au Contexte. Application au Domaine du M-Learning. », p. 265.

[21] J. Pietri, « L'apport du e-learning dans la formation initiale et continue en odontologie », Université Nice-Sophia Antipolis, 2017.

[22] A. H. Masud et X. Huang, « An E-learning System Architecture based on Cloud Computing », World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering, p. 5, 2012.
[23] A. H. Masud et X. Huang, « An E-learning System Architecture based on Cloud Computing », vol. 6, n° 2, p. 5, 2012.
[24] X. Hu et Z. D. J. Zhang, « Medical Informatization Based on Cloud Computing Concepts and Techniques », *Journal Of. Medical Informatics*, 2010.
[25] B. Radenkovic, M. Despotovic-Zrakic, Z. Bogdanovic, V. Vujin, et D. Barac, « Harnessing cloud computing infrastructure for e-learning services », *Facta universitatis - series: Electronics and Energetics*, vol. 27, n° 3, p. 339-357, 2014.
[26] J.-S. Jeong, M. Kim, et K.-H. Yoo, « A Content Oriented Smart Education System based on Cloud Computing », *International Journal of Multimedia and Ubiquitous Engineering*, vol. 8, n° 6, p. 313-328, nov. 2013.
[27] « FIPS 197, Advanced Encryption Standard (AES) », p. 51.
[28] P. K. Udupi, P. Malali, et H. Noronha, « Big data integration for transition from e-learning to smart learning framework », in *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, Muscat, Oman, 2016, p. 1-4.
[29] M. Birjali, A. Beni-Hssane, et M. Erritali, « Learning with Big Data Technology: The Future of Education », in *Proceedings of the Third International Afro-European Conference for Industrial Advancement — AECIA 2016*, vol. 565, A. Abraham, A. Haqiq, A. Ella Hassanien, V. Snasel, et A. M. Alimi, Éd. Cham: Springer International Publishing, 2018, p. 209-217.
[30] B. Logica et R. Magdalena, « Using Big Data in the Academic Environment », *Procedia Economics and Finance*, vol. 33, p. 277-286, 2015.
[31] S. Van Gaver, A. Fiorentino, M. Slitni, T. Bedouin, et A. Alehyane, « Un nuage bien réel, bonnes pratiques pour la mise en place d'un cloud dans l'enseignement supérieur », 2015.
[32] T. BELARBI, « Le Cloud Computing et la naissance d'Openstack », *Blog Technique Sopra Steria*, 13-mars-2017. [En ligne]. Disponible sur: https://blogtech.soprasteria.com/2017/03/13/le-cloud-computing-et-la-naissance-dopenstack/. [Consulté le: 06-sept-2018].
[33] MOHAMAD, « Autonomie, sécurité et QoS de bout en bout dans un environnement de Cloud Computing ».
[34] A. P. Mell (NIST) et A. T. Grance (NIST), « The NIST Definition of Cloud Computing ». [En ligne]. Disponible sur: https://csrc.nist.gov/publications/detail/sp/800-145/final. [Consulté le: 27-août-2018].
[35] « Qu'est-ce que le cloud computing ? Microsoft Azure ». [En ligne]. Disponible sur: https://azure.microsoft.com/fr-fr/overview/what-is-cloud-computing/. [Consulté le: 27-août-2018].
[36] B. Hammi, « Vers une détection à la source des activités malveillantes dans les clouds publics : application aux attaques de déni de service », 2015.
[37] « Les technologies du cloud computing ». .
[38] « Service web », *Wikipédia*. 08-août-2018.
[39] « XML-RPC », *Wikipédia*. 09-avr-2016.
[40] H. Fadili, « WEB 2.0 & WEB 3.0 popularisation de la création et de la promotion culturelle et scientifique : cas des Wikis Opportunité pour Tamazight », p. 12.
[41] « Regards sur l'informatique en nuage – Conclusions de 110 projets de mise en œuvre ».

[42] « Le Cloud Computing : définition, avantage et inconvénients », *Renaud Venet*, 26-janv-2011. [En ligne]. Disponible sur: http://www.renaudvenet.com/cloud-computing-avantages-et-inconvenients-2011-01-26.html. [Consulté le: 09-sept-2018].
[43] « Amazon Web Services », *Wikipédia*. 21-août-2018.
[44] « AWS | Amazon EC2 – Service d'hébergement cloud évolutif », *Amazon Web Services, Inc.* [En ligne]. Disponible sur: https://aws.amazon.com/fr/ec2/. [Consulté le: 10-sept-2018].
[45] T. Chardonnens, « Les enjeux du Cloud Computing en entreprise », Université de Fribourg, Suisse, 2012.
[46] « Amazon Simple Storage Service - Manuel du développeur », p. 698.
[47] « Amazon EMR – Amazon Web Services », *Amazon Web Services, Inc.* [En ligne]. Disponible sur: https://aws.amazon.com/fr/emr/. [Consulté le: 28-oct-2018].
[48] « Amazon DynamoDB – Service de base de données NoSQL dans le cloud », *Amazon Web Services, Inc.* [En ligne]. Disponible sur: https://aws.amazon.com/fr/dynamodb/. [Consulté le: 01-janv-2019].
[49] « Produits et services - Google Cloud ». [En ligne]. Disponible sur: https://cloud.google.com/products/?hl=fr. [Consulté le: 01-janv-2019].
[50] « Salesforce CRM - Gestion de la Relation Client (GRC) », *Salesforce.com*. [En ligne]. Disponible sur: https://www.salesforce.com/fr/. [Consulté le: 01-janv-2019].
[51] « Outlook.com - Microsoft free personal email ». [En ligne]. Disponible sur: https://outlook.live.com/owa/. [Consulté le: 12-sept-2018].
[52] « Microsoft Office 365 ». [En ligne]. Disponible sur: https://www.office.com/. [Consulté le: 12-sept-2018].
[53] « Services et plateforme de cloud computing Microsoft Azure ». [En ligne]. Disponible sur: https://azure.microsoft.com/fr-fr/. [Consulté le: 12-sept-2018].
[54] « Infrastructure en tant que service (IaaS) Microsoft Azure ». [En ligne]. Disponible sur: https://azure.microsoft.com/fr-fr/overview/what-is-azure/iaas/. [Consulté le: 01-janv-2019].
[55] « Catalogue - IBM Cloud », 01-mai-2016. [En ligne]. Disponible sur: https://console.bluemix.net/catalog/. [Consulté le: 13-sept-2018].
[56] « SoftLayer | Cloud Servers, Storage, Big Data, & More IAAS Solutions ». [En ligne]. Disponible sur: http://www.softlayer.com/fr. [Consulté le: 13-sept-2018].
[57] M. Cox et D. Ellsworth, « Application-controlled demand paging for out-of-core visualization », in *Proceedings. Visualization '97 (Cat. No. 97CB36155)*, Phoenix, AZ, USA, 1997, p. 235-244,.
[58] S. Ghemawat, H. Gobioff, et S.-T. Leung, « The Google File System », 2003.
[59] Hadoop, « Apache Hadoop », 2019. [En ligne]. Disponible sur: http://hadoop.apache.org/. [Consulté le: 23-sept-2018].
[60] S. DAVIAUD, « Guide du big data l'annuaire de référence à destination des utilisateurs ». Cloudera, 2015.
[61] L. Echos, « Big data et intelligence artificielle : un peu de décryptage pour y voir plus clair », *lesechos.fr*, 21-juill-2016. [En ligne]. Disponible sur: http://archives.lesechos.fr/archives/cercle/2016/07/21/cercle_159119.htm. [Consulté le: 19-oct-2018].
[62] « Définition : Qu'est-ce que le Big Data ? », *LeBigData.fr*. [En ligne]. Disponible sur: https://www.lebigdata.fr/definition-big-data. [Consulté le: 22-sept-2018].

[63] H. Guo, L. Wang, F. Chen, et D. Liang, « Scientific big data and Digital Earth », *Chin. Sci. Bull.*, p. 9, 2014.
[64] D. A. Chakhari, « La digitalisation est une guerre mondiale armez-vous », p. 70, 2015.
[65] Gartner, « Big data: The next frontier for innovation, competition, and productivity », McKinsey Global Institute, 2011.
[66] NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, « NIST Big Data Interoperability Framework: Volume 1, Definitions », National Institute of Standards and Technology, NIST SP 1500-1, oct. 2015.
[67] S. Vialle, « Big Data : Informatique pour les données et calculs massifs ». CentraleSupélec, 2018.
[68] « Développement informatique, développement spécifique informatique - Cynapsys ». [En ligne]. Disponible sur: http://www.cynapsys.de/content/big-data-le-nouvel-art-de-traitement-des-donn%C3%A9es-massives. [Consulté le: 20-sept-2018].
[69] Y. Grandmontagne, « Les 5 V du Big Data - IT Social | Média des Enjeux IT & Business, Innovation et Leadership », 2014. .
[70] R. Bruchez, Les bases de données NoSQL et le BigData: Comprendre et mettre en oeuvre, 2ème édition. EYROLLES, 2015.
[71] « The Hadoop Distributed File System (HDFS) », 2018. [En ligne]. Disponible sur: https://hadoop.apache.org/docs/r3.1.1/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html. [Consulté le: 28-sept-2018].
[72] « Apache Cassandra ». [En ligne]. Disponible sur: http://cassandra.apache.org/. [Consulté le: 25-oct-2018].
[73] J. Dean et S. Ghemawat, « MapReduce: simplified data processing on large clusters », *Communications of the ACM*, vol. 51, n° 1, p. 107, janv. 2008.
[74] C. Metz, « How Yahoo Spawned Hadoop, the Future of Big Data », *Wired*, 18-oct-2011.
[75] « Hadoop ». [En ligne]. Disponible sur: https://www.facebook.com/notes/facebook-engineering/hadoop/16121578919/. [Consulté le: 28-oct-2018].
[76] JasonWHowell, « What are HDInsight and the Hadoop and Spark technology stack? - Azure ». [En ligne]. Disponible sur: https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-introduction. [Consulté le: 28-oct-2018].
[77] « Data Management Platform, Solutions and Big Data Analysis », *Hortonworks*. [En ligne]. Disponible sur: https://fr.hortonworks.com/. [Consulté le: 29-sept-2018].
[78] « Machine Learning | Analytics | Cloud - Cloudera ». [En ligne]. Disponible sur: https://www.cloudera.com/. [Consulté le: 29-sept-2018].
[79] « Industry's Leading Data Platform for AI and Analytics ». [En ligne]. Disponible sur: https://mapr.com/. [Consulté le: 28-oct-2018].
[80] « MapR Awarded Best Big Data Product for Internet of Things and Companies Driving Innovation ». [En ligne]. Disponible sur: https://mapr.com/company/press-releases/mapr-awarded-best-big-data-for-iot-and-innovation/. [Consulté le: 28-oct-2018].
[81] « Big Data Analytics on Premise, in the Cloud, or on Hadoop », *Vertica*. [En ligne]. Disponible sur: https://www.vertica.com/. [Consulté le: 30-sept-2018].
[82] « Cloud Dataproc – Apache Hadoop et Apache Spark en mode cloud natif », *Google Cloud*. [En ligne]. Disponible sur: https://cloud.google.com/dataproc/?hl=fr. [Consulté le: 28-oct-2018].
[83] « Bridging two worlds : Integration of SAP and Hadoop Ecosystems | SAP Blogs ». [En ligne]. Disponible sur: https://blogs.sap.com/2017/07/19/bridging-two-worlds-integration-of-sap-and-hadoop-ecosystems/. [Consulté le: 28-oct-2018].

[84] « Apache Spark™ - Unified Analytics Engine for Big Data », 2019. [En ligne]. Disponible sur: https://spark.apache.org/. [Consulté le: 24-sept-2018].
[85] B. Chambers et M. Zaharia, « Spark: The Definitive Guide », p. 601.
[86] « Giraph - Welcome To Apache Giraph! » [En ligne]. Disponible sur: http://giraph.apache.org/. [Consulté le: 27-sept-2018].
[87] « Apache Storm ». [En ligne]. Disponible sur: http://storm.apache.org/. [Consulté le: 27-sept-2018].
[88] « Sqoop User Guide (v1.4.7) ». [En ligne]. Disponible sur: http://sqoop.apache.org/docs/1.4.7/SqoopUserGuide.html. [Consulté le: 02-oct-2018].
[89] « Yahoo », *Yahoo*. [En ligne]. Disponible sur: http://fr.yahoo.com. [Consulté le: 02-oct-2018].
[90] « Electronics, Cars, Fashion, Collectibles, Coupons and More », *eBay*. [En ligne]. Disponible sur: https://www.ebay.com. [Consulté le: 02-oct-2018].
[91] « Apache Solr - ». [En ligne]. Disponible sur: http://lucene.apache.org/solr/. [Consulté le: 02-oct-2018].
[92] « Apache Software Foundation - Ambari », 2018. [En ligne]. Disponible sur: https://ambari.apache.org/. [Consulté le: 27-oct-2018].
[93] « Apache Ambari », *Wikipedia*, 12-sept-2018. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Apache_Ambari&oldid=859242637. [Consulté le: 27-oct-2018].
[94] M. E. Malki, « Modélisation NoSQL des entrepôts de données multidimensionnelles massives », p. 148.
[95] « File: README — Documentation for flockdb (0.3.2) ». [En ligne]. Disponible sur: https://www.rubydoc.info/gems/flockdb/0.3.2. [Consulté le: 05-oct-2018].
[96] « Apprentissage automatique », *Wikipédia*. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Apprentissage_automatique&oldid=151937464. [Consulté le: 07-oct-2018].
[97] A. L. Samuel, « Some Studies in Machine Learning Using the Game of Checkers », p. 20, 1956.
[98] É. Biernat et M. Lutz, Data science fondamentaux et études de cas: machine learning avec Python et R. Paris: Eyrolles, 2016.
[99] P. Cuxac et J.-C. Lamirel, « Clustering incrémental et méthodes de détection de nouveauté: application à l'analyse intelligente d'informations évoluant au cours du temps », p. 27-28.
[100] J. Macqueen, « SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS », *MULTIVARIATE OBSERVATIONS*, p. 17, 1967.
[101] M. Ester, H.-P. Kriegel, J. Sander, et X. Xu, « A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise », in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996, p. 226–231.
[102] T. Bayes et R. Price, « An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S », 1763.
[103] « Decision tree learning », *Wikipedia*, 10-oct-2018. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=863324897. [Consulté le: 10-oct-2018].

[104] « Arbre de décision », *Wikipédia*. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Arbre_de_d%C3%A9cision&oldid=151158919. [Consulté le: 10-oct-2018].
[105] « L'apprentissage supervisé - Machine Learning », *Mr. Mint : Apprendre le Machine Learning de A à Z*, 19-févr-2017. [En ligne]. Disponible sur: https://mrmint.fr/apprentissage-supervise-machine-learning. [Consulté le: 11-oct-2018].
[106] « Linear regression », *Wikipedia*, 10-oct-2018. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=863434309. [Consulté le: 11-oct-2018].
[107] M. E. Sanharawi et F. Naudet, « Comprendre la régression logistique », */data/revues/01815512/v36i8/S0181551213002490/*, mai 2013.
[108] R. Agrawal, T. Imielinski, A. Swami, H. Road, et S. Jose, « Mining Association Rules between Sets of Items in Large Databases », p. 10, 1993.
[109] R. Agrawal et R. Srikant, « Fast Algorithms for Mining Association Rules », p. 13, 1994.
[110] M. J. Zaki, « Scalable algorithms for association mining », *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, n° 3, p. 372-390, juin 2000.
[111] J. H. J. Pei, « Mining Frequent Patterns without Candidate Generation », p. 12, 2000.
[112] « Time series », *Wikipedia*, 28-sept-2018. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Time_series&oldid=861564221. [Consulté le: 14-oct-2018].
[113] « Série temporelle », *Wikipédia*. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=S%C3%A9rie_temporelle&oldid=148770890. [Consulté le: 13-oct-2018].
[114] E. Mercier, « Text mining : quand le texte devient donnée », *Introduction aux humanités numériques*. .
[115] « Text mining », *Wikipedia*, 14-oct-2018. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Text_mining&oldid=863978865. [Consulté le: 19-oct-2018].
[116] « Apprentissage automatique et deep learning », *STEMMER IMAGING*. [En ligne]. Disponible sur: https://www.stemmer-imaging.fr/fr/conseil-techniqu/apprentissage-automatique-et-apprentissage-profond/. [Consulté le: 19-oct-2018].
[117] « Définition | Deep Learning - Apprentissage profond | Futura Tech ». [En ligne]. Disponible sur: https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/. [Consulté le: 19-oct-2018].
[118] « Blackboard | Education Technology & Services ». [En ligne]. Disponible sur: https://www.blackboard.com/index.html. [Consulté le: 19-févr-2019].
[119] « edX | Online courses from the world's best universities ». [En ligne]. Disponible sur: https://www.edx.org/?g_acctid=926-195-8061&g_campaign=new-gs-row-brand-core-broad-free&g_campaignid=1669870784&g_adgroupid=73172705028&g_adid=323329644428&g_keyword=free%20edx&g_keywordid=kwd-297055126689&g_network=g&utm_source=adwords&utm_campaign=1669870784&utm_medium=73172705028&utm_term=free%20edx&gclid=CjwKCAiA767jBRBqEiwAGdAOrwQUafJ2aWsID180Vp4jNWOQc7gANP1rlnIWI7pLhuWbP02zPdvFVhoCYysQAvD_BwE. [Consulté le: 19-févr-2019].

[120] « Coursera | Online Courses & Credentials by Top Educators. Join for Free ». [En ligne]. Disponible sur: https://www.coursera.org/. [Consulté le: 19-févr-2019].
[121] « Moodle - Open-source learning platform | Moodle.org ». [En ligne]. Disponible sur: https://moodle.org/. [Consulté le: 14-déc-2018].
[122] « The Ultimate List Of Cloud-Based Learning Management Systems », *eLearning Industry*, 18-mai-2013. [En ligne]. Disponible sur: https://elearningindustry.com/the-ultimate-list-of-cloud-based-learning-management-systems. [Consulté le: 06-nov-2018].
[123] « Codeanywhere ». [En ligne]. Disponible sur: https://codeanywhere.com/. [Consulté le: 07-nov-2018].
[124] « Codenvy ». [En ligne]. Disponible sur: https://codenvy.io/site/login?redirect_url=https%3A%2F%2Fcodenvy.io%2F. [Consulté le: 07-nov-2018].
[125] « Modern Dev Environment Delivered · Koding », *Modern Dev Environment Delivered · Koding*. [En ligne]. Disponible sur: https://www.koding.com/. [Consulté le: 07-nov-2018].
[126] « Kernel-based Virtual Machine ». [En ligne]. Disponible sur: https://www.linux-kvm.org/page/Main_Page. [Consulté le: 07-nov-2018].
[127] « The Xen hypervisor ». [En ligne]. Disponible sur: https://www.xenproject.org/about/history.html. [Consulté le: 07-nov-2018].
[128] M. Dufresne, « 8 Avantages d'une Plateforme e-Learning en SaaS », 2016. [En ligne]. Disponible sur: https://www.skillbuilderlms.com/fr/plateforme-e-learning-saas-avantages/. [Consulté le: 31-août-2018].
[129] « Panama Papers », *Wikipedia*, 14-nov-2018. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Panama_Papers&oldid=868806850. [Consulté le: 16-nov-2018].
[130] « WikiLeaks' piracy and American security - Chicago Tribune ». [En ligne]. Disponible sur: https://www.chicagotribune.com/news/opinion/editorials/ct-wikileaks-cia-hack-assange-nsa-edit-0309-jm-20170308-story.html. [Consulté le: 16-nov-2018].
[131] « Big data dans le eLearning : l'avenir du secteur du eLearning », *eLearning Industry France*, 02-août-2016. [En ligne]. Disponible sur: https://elearningindustry.fr/big-data-dans-le-elearning. [Consulté le: 21-sept-2018].
[132] « Netflix - Site web pour regarder des films en ligne ». [En ligne]. Disponible sur: https://www.netflix.com. [Consulté le: 30-nov-2018].
[133] N. Bechet, « Etat de l'art sur les Systèmes de Recommandation », p. 23.
[134] F. O. Isinkaye, Y. O. Folajimi, et B. A. Ojokoh, « Recommendation systems: Principles, methods and evaluation », *Egyptian Informatics Journal*, vol. 16, n° 3, p. 261-273, nov. 2015.
[135] « L'adaptive Learning et/ou le Big Data au service de l'apprenant ». [En ligne]. Disponible sur: https://www.linkedin.com/pulse/ladaptive-learning-etou-le-big-data-au-service-de-yann-lescurat/. [Consulté le: 18-déc-2018].
[136] « Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more ». [En ligne]. Disponible sur: https://www.amazon.com/. [Consulté le: 30-sept-2018].
[137] « Collaborative filtering », *Wikipedia*. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Collaborative_filtering&oldid=863509411. [Consulté le: 14-oct-2018].

[138] « Les systèmes de recommandation : une catégorisation ». [En ligne]. Disponible sur: https://interstices.info/les-systemes-de-recommandation-categorisation/. [Consulté le: 05-déc-2018].
[139] D. T. Larose, « Data Mining and Predictive Analytics », p. 827, 2015.
[140] M. P. Hrzenjak, M. Matetic, et M. B. Bakaric, « Mining association rules in learning management systems », in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2015, p. 986-991.
[141] T. A. Kumbhare et S. V. Chobe, « An overview of association rule mining algorithms », *International Journal of Computer Science and Information Technologies*, vol. 5, n° 1, p. 927–930, 2014.
[142] R. Agrawal et R. Srikant, « Fast Algorithms for Mining Association Rules », 1994.
[143] R. Agrawal et R. Srikant, « Fast Algorithms for Mining Association Rules in Datamining », *Proceedings of the 20th VLDB Conference Santiago, Chile*, vol. 2, p. 13–24, 1994.
[144] M. Mittal, J. Singh, A. Aggarwal, K. Kumari, et M. Yadav, « Ordering Policy for Imperfect Quality Itemsets Using Cross Selling Effects », *International Journal of Modeling and Optimization*, vol. 4, n° 1, p. 25-30, 2014.
[145] G. Jatteau, « MÉMOIRE PRÉSENTÉ COMME EXIGENCE PARTIELLE DE LA MAÎTRISE EN INFORMATIQUE », p. 83.
[146] A. Ibriz, « C.D.I.O : Une Démarche Innovante de Conduite de Projet Elearning », *2ème Congrès International du Génie Industriel et du Management des Systèmes (CIGIMS)*, p. 14, 2015.
[147] A. Ibriz et A. Safouane, « L'Innovation Pédagogique dans les EST du Maroc : Le Model et la Conduite d'un cas réussi à travers le Projet ESTenLigne », p. 6, 2014.
[148] M. Benslimane, K. Ouazzani, M. Tmimi, et M. Berrada, « Proposal of an Approach of Online Course Design and Implementation: A Case Study of an Algorithmic Course », vol. 7, p. 7, 2016.
[149] M. Hahsler et S. Chelluboina, « Visualizing Association Rules: Introduction to the R-extension Package arulesViz », p. 24.
[150] A. Garg, « SCORM BASED LEARNING MANAGEMENT SYSTEM FOR ONLINE TRAINING », p. 51.
[151] « R: The R Project for Statistical Computing ». [En ligne]. Disponible sur: https://www.r-project.org/. [Consulté le: 16-déc-2018].
[152] « S (programming language) », *Wikipedia*. 30-août-2018.
[153] E. F. Inc, « The Platform for Open Innovation and Collaboration - The Eclipse Foundation ». [En ligne]. Disponible sur: https://www.eclipse.org/. [Consulté le: 17-déc-2018].
[154] « Apache Maven ». [En ligne]. Disponible sur: https://maven.apache.org/. [Consulté le: 17-déc-2018].
[155] « Oracle Database ». [En ligne]. Disponible sur: https://www.oracle.com/database/products-a-z.html. [Consulté le: 17-déc-2018].
[156] « PostgreSQL open source database ». [En ligne]. Disponible sur: https://www.postgresql.org/. [Consulté le: 17-déc-2018].
[157] « Microsoft SQL Server », *Microsoft SQL Server - US (English)*. [En ligne]. Disponible sur: https://www.microsoft.com/en-us/sql-server/sql-server-2017. [Consulté le: 17-déc-2018].

[158] « The Scala Programming Language ». [En ligne]. Disponible sur: https://www.scala-lang.org/. [Consulté le: 18-déc-2018].
[159] « Oracle VM VirtualBox ». [En ligne]. Disponible sur: https://www.virtualbox.org/. [Consulté le: 23-déc-2018].
[160] « The Linux Foundation – Supporting Open Source Ecosystems », *The Linux Foundation*. [En ligne]. Disponible sur: https://www.linuxfoundation.org/. [Consulté le: 24-déc-2018].
[161] « Apple - MAC OS X operating system ». [En ligne]. Disponible sur: https://support.apple.com/downloads/macos. [Consulté le: 24-déc-2018].
[162] « Oracle Solaris 11 ». [En ligne]. Disponible sur: https://www.oracle.com/solaris/solaris11/. [Consulté le: 24-déc-2018].
[163] « Official Site for Microsoft Windows ». [En ligne]. Disponible sur: http://windows.com. [Consulté le: 24-déc-2018].
[164] « itemFrequencyPlot function | R Documentation ». [En ligne]. Disponible sur: https://www.rdocumentation.org/packages/arules/versions/1.6-2/topics/itemFrequencyPlot. [Consulté le: 19-déc-2018].
[165] M. Hahsler, C. Buchta, B. Gruen, K. Hornik, I. Johnson, et C. Borgelt, *arules: Mining Association Rules and Frequent Itemsets*. 2018.
[166] H. Li, Y. Wang, D. Zhang, M. Zhang, et E. Y. Chang, « Pfp: parallel fp-growth for query recommendation », in *Proceedings of the 2008 ACM conference on Recommender systems - RecSys '08*, Lausanne, Switzerland, 2008, p. 107.
[167] « MLlib: Main Guide - Spark 2.4.0 Documentation ». [En ligne]. Disponible sur: https://spark.apache.org/docs/latest/ml-guide.html. [Consulté le: 21-déc-2018].
[168] « Monitoring and Instrumentation - Spark 2.3.1 Documentation ». [En ligne]. Disponible sur: https://spark.apache.org/docs/latest/monitoring.html. [Consulté le: 24-oct-2018].
[169] « Apache Hadoop 2.9.2 – Hadoop: YARN-UI V2 ». [En ligne]. Disponible sur: https://hadoop.apache.org/docs/r2.9.2/hadoop-yarn/hadoop-yarn-site/YarnUI2.html. [Consulté le: 28-déc-2018].
[170] « Ganglia Monitoring System ». [En ligne]. Disponible sur: http://ganglia.sourceforge.net/. [Consulté le: 21-nov-2018].

Travaux scientifiques

Publications dans des Journaux Internationaux :

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Abdelali Ibriz, “Building an e-learning recommender system using Association Rules techniques and R environment”, IJIST The International Journal of Information Science & Technology, Special Issue: “E-Learning Systems and Applications”, https://innove.org/ijist/index.php/ijist/article/view/51/35, ISSN: 2550-5114, VOL 3, NO 2 (2019), pp 11–18, Submission date: 20 Dec 2018, Accepted: 20 Feb 2019, Published Online: 27 March 2019.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir and Abdelali Ibriz, “Large-scale e-learning recommender system based on Spark and Hadoop”, Journal of Big Data, DOI: https://doi.org/10.1186/s40537-019-0169-4, ISSN: 2196-1115 (electronic version), Vol 6, Issue 1, Received: 30 October 2018, Accepted: 3 January 2019, Published: 8 January 2019.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Fayçal Messaoudi, “Big data for online learning systems”, Journal Education and Information Technologies (2018), 23: 2783. DOI: https://doi.org/10.1007/s10639-018-9741-3, ISSN: 1360-2357 (print version) ISSN: 1573-7608 (electronic version), Vol 23, Issue 6, pp 2783–2800, Received: 9 February 2018, Accepted: 16 May 2018, Published Online: 30 May 2018.

Karim Dahdouh, Ahmed Dakak, Lahcen Oughdir, “Integration of the Cloud Environment in E-learning Systems”, Transactions on Machine Learning and Artificial Intelligence, DOI: http://dx.doi.org/10.14738/tmlai.54.2017, ISSN: 2054-7390, Vol 5, No 4 (2017).

Chapitre

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Abdelali Ibriz, Chapter title: “Association Rules Mining Method of Big Data for E-Learning Recommendation Engine”. Book title: “Advanced Intelligent Systems for Sustainable Development (AI2SD’2018)”, DOI 978-3-030-11928-7_43, https://www.springer.com/gp/book/9783030119270, Springer, 2019.

Articles soumis aux journaux internationaux

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, “Big data: A distributed storage and processing for online learning systems”, International Journal of Computational Intelligence Studies, https://www.inderscience.com/jhome.php?jcode=ijcistudies, submission date: 14 Feb 2018, Editor's accepted version: 30 May 2018,status: In Production.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Abdelali Ibriz, Chapter title: “Improving online education using big data technologies”, Book title: “The Role of Technology in Education”, IntechOpen Open Access Publisher, https://www.intechopen.com, submission date: 27 January, 2019.

Communications dans des congrès et des conférences nationaux et internationaux

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, “Intégration de l'environnement Cloud dans les systèmes e-learning”, The 2nd Scientific Day on Computer Science, Optimization and Systems’ Modelization (CSOSM’17), March 9, 2017, Faculty of Sciences, Kenitra, Morocco.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, “Approach of using and harnessing Cloud Computing services in E-learning systems”, The First International Conference on Affective computing, Machine Learning and Intelligent Systems (ACMLIS'17), May 18-20,2017, Tetouan ,Morocco, pp 91-92, http://scholarpublishing.org/index.php/TMLAI/issue/view/165.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, “An approach for using Big Data technologies in e-learning systems”, An approach for using Big Data technologies in e-learning systems 06 Juillet 2017, Kenitra, Maroc.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, “Integrating Big Data technologies in a dynamic environment EIAH dedicated to e-learning systems based on cloud infrastructure”, Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems (ICCWCS'17), Article No. 20 , November 14 - 16, 2017, Larache, Morocco, Association for Computing Machinery, https://dl.acm.org/citation.cfm?id=3167506, ACM ISBN: 978-1-4503-5306-9, DOI: 10.1145/3167486.3167506.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Fayçal Messaoudi, “A new architecture for E-learning systems based on Big Data technologies in Cloud Computing”, International

Symposium on Computer Sciences and Applications (ISCSA2017), October 26-28, 2017, FST, Errachidia, Morocco.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, “Les environnements informatiques pour l’apprentissage humain à l’ère du big data : massive open online course (MOOC)”, 3ème Rencontre Nationale des Jeunes Chercheurs en Sciences de l’Ingénieur (RNJC’2017), Novembre 29-30, 2017, FST, Fès, Morocco.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Abdelali Ibriz, “Association Rules Mining Method of Big Data for E-Learning Recommendation Engine”. The International Conference on Advanced Intelligent Systems for Sustainable Development applied to Agriculture, Energy, Health, Environment, Industry and Economy. July 12–14, 2018, FST, Tangier, Morocco. Advances in Intelligent Systems and Computing, vol 915. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-11928-7_43. Online ISBN: 978-3-030-11928-7, Print on Demand(PoD) ISBN: 978-3-030-11927-0, DOI: https://doi.org/10.1007/978-3-030-11928-7_43. Published 07 March 2019.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Abdelali Ibriz, “Smart Courses recommender system for online learning platform”, The 2nd IEEE Conference on E-Learning and innovation in Education (ELED'18), IEEE 5th International Congress on Information Science and Technology (CiSt), October 21-27, 2018, Marrakech, Morocco, https://ieeexplore.ieee.org/document/8596516, Electronic ISBN: 978-1-5386-4385-3, Print on Demand(PoD) ISBN: 978-1-5386-4386-0, Electronic ISSN: 2327-1884, DOI: 10.1109/CIST.2018.8596516.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, “E-learning in the era of big data: An integration methodology”, The international conference on pedAgogical aPproaches and E-Learning (APEL'2018), November 28-30, 2018, EST, Fez, Morocco.

Karim Dahdouh, Ahmed Dakkak, Lahcen Oughdir, Abdelali Ibriz, “Building an adaptive e-learning system using machine learning techniques”, The international conference on pedAgogical aPproaches and E-Learning (APEL'2018), November 28-30, 2018, EST, Fez, Morocco.

Prix de recherche

Troisième Prix de l'édition 2018 des Doctoriales organisée à la cité de l'innovation de Fès du 19 au 21 juin 2018 par l'université Sidi Mohamed Ben Abdellah en collaboration avec l'université Mohamed Premier d'Oujda, sous le thème "Smart E-learning".

Certificat

Certificat obtenu lors de la deuxième édition de la session de formation "Data Science & Big Data Analytics" organisée par le CED "Sciences et Techniques de l'Ingénieur" de la Faculté des Sciences et Techniques de Fès (FSTF) en collaboration avec EMC Academic Alliance.

Karim Dahdouh, EMC Education Services, EMC Academic Associate, Data Science and Big Data Analytics, March 10, 2018, Verify at: www.certmetrics.com/emc/public/verification.aspx, Verification Code: 5BGSD3NT2MFQ14CX.