UNIVERSITE SULTAN MOULAY SLIMANE

Faculté des Sciences et Techniques

Béni-Mellal

Formation Doctorale : **Mathématiques et Physique Appliquée**

THÈSE

Présentée par

AMINE ABDELLAH

Pour l'obtention du grade de

DOCTEUR

Spécialité ***: INFORMATIQUE***

Option ***: NFORMATIQUE***

Analyse et Conception d'un Système d'Information Décisionnel basé sur APOGEE : Cas relatif à l'Enseignement Supérieur

Soutenue le Samedi 16 Mars 2019 à 11h devant la commission d'examen :

MOHAMED FAKIR	Professeur à la Faculté des Sciences et Techniques, Béni Mellal, Maroc	Président
HICHAM MONCIF	Professeur à la Faculté Polydisciplinaire, Béni Mellal, Maroc	Rapporteur
ALI RACHIDI	Professeur à l'Ecole Nationale du Commerce et Gestion, Agadir, Maroc	Rapporteur
MOHAMED BASLAM	Professeur à la Faculté des Sciences et Techniques, Béni Mellal, Maroc	Rapporteur
JILALI ANTARI	Professeur à la Faculté Polydisciplinaire, Taroudant, Maroc	Examinateur
NOUREDDINE ASKOUR	Professeur à la Faculté des Sciences et Techniques, Béni Mellal, Maroc	Directeur de thèse
BELAID BOUIKHALENE	Professeur à la Faculté Polydisciplinaire, Béni Mellal, Maroc	Co-directeur de thèse

Résumé

Suites aux recommandations des reformes de l'enseignement supérieur au Maroc, l'université abrite différents acteurs qui ont recours à des systèmes de ressources documentaires, des systèmes de production d'information et des systèmes de recherche d'information. L'intérêt que nous portons pour la construction d'un entrepôt de données et des bases métiers associées permet de faire évoluer un système d'information en un système d'information stratégique basé sur la source d'information APOGEE. La thématique traitée dans cette thèse a pour objectif de développer un système d'information de prise de décisions lié à l'environnement numérique du travail universitaire. Nous proposons de modeler les données dans une université publique pour transformer un système d'information dans un système d'information de prise de décisions, qui est basé sur les bases de données d'affaires orientées vers les acteurs. Dans cette optique nous présentons un système universitaire d'aide à la décision baptisé **SDUM**« Système décisionnel pour l'université marocaine ».

Notre système est une solution qui permet d'accéder facilement aux données, générer des rapports représentatifs, et améliorer fortement le processus de prise de décision.

Notre système décisionnel **SDUM** est constitué de différentes phases à savoir :

- La phase d'alimentation qui permet d'intervenir des processus ETL qui se chargeront de récupérer toutes les données nécessaires depuis les différentes sources de stockage.
- La phase de modélisation et de stockage ou les données sont stockées sous une forme adaptée pour les analyses qui sera effectué.
- La phase de réorganisation qui permet de réorganiser les données entreposées en les stockant dans des magasins de données visant à supporter efficacement les processus d'interrogation et d'analyse.
- La phase d'analyse et d'exploitation ou les utilisateurs finaux interviennent et analysent les informations qui leurs sont fournies.

Mots-clés

Systèmes d'information stratégiques, base de données décisionnelle, entrepôt de données, datamining, modèle multidimensionnel, modèle de classification d'utilisateur, Visual Studio, Pentaho, KNIME, segmentation des étudiants, rôle d'association, classification.

Abstract

A university shelters various actors who resort to systems of documentary resources, systems of information production, and systems of information research. The importance that we give to the building of a data warehouse and to the associated business databases allows us to develop an information system into a strategic information system. The theme dealt with in this thesis consists in developing a decision-making information system, which is linked to the digital environment of University work.

We suggest modeling the data in the university to transform an information system into a decision-making information system, which is based on actor-oriented business databases. Decision-making information is a system that allows university decision-makers to have powerful and relevant information analysis tools to help them make the right decision at the right time.

In this perspective, we present a university decision support system called **SDUM** «Decision system for the Moroccan university»

Key words

Strategic Information Systems, Decisional data, Data Warehouse, Data Mining, multidimensional model, User Classification Model, Visual Studio, Pentaho, KNIME, Student Segmentation, Association Role, Classification.

Table des matières

Liste des Abréviations

ED	Entrepôt de données
MD	Magasin de données
SI	: Système d'information
OLAP	: On-Line Analytical Processing
OLTP	: On-Line Transaction Processing
DW	: Data Warehouse
AG	: Algorithme Génétique
SGBD(R)	: Système de Gestion de Base de données (Relationnel)
MOLAP	: Multidimensionnel OLAP
ROLAP	: Relational OLAP
OOLAP	: Object OLAP
DOLAP	: Desktop OLAP
BI	: Business Intelligence
DM	: Data Mining
ETL	: Extract Transform Load
ERP	: Enterprise Resource Planning
AMUE	: L'agence de mutualisation des universités et des établissements
OMG	: Object Management Group
JMS	: Java Message Service
SSAS	: SQL Server Analysis Services
SSIS	: SQL Server Intégrations Services

Liste des figures

Liste des tableaux

Remerciements

Je remercie toutes les personnes avec lesquelles j'ai eu le plaisir de collaborer et qui ont pu par la même occasion de m'aider durant toute la durée de ce projet.

Je tiens à adresser mes remerciements et à exprimer ma profonde admiration et gratitude aux personnes qui m'ont apporté leur aide et qui ont ainsi contribué à l'élaboration, et favorisé l'aboutissement de ce projet à :

Mr. Pr Noureddine ASKOUR, mon encadrant, qui a fait preuve de disponibilité à chaque fois que j'avais besoin de son soutien. Son encadrement et ses conseils ont été d'un appui considérable.

Toute ma profonde gratitude à Monsieur Pr. Belaid BOUIKHALENE, mon Co encadrant dans ce projet, pour l'aide et le temps qu'il m'a consacré à ma disposition.
Sans l'environnement de recherche qu'il a su créer, je n'aurais pas pu me lancer dans la préparation de cette thèse. Je lui exprime ici ma profonde reconnaissance et le remercie vivement pour ces années de soutien, pour ses précieux conseils, et pour sa manière très simple de toujours trouver les mots d'encouragement qui ne manquaient pas de raviver ma motivation.
Je tiens à remercier aussi les membres du jury :

Pr. Mohamed Fakir
Pr. Hicham Moncif
Pr. Ali Rachidi
Pr. Mohamed Baslam
Pr. JilaliAntari,

pour avoir accepté de juger notre modeste travail.
Enfin, sans oublier mon père **M'hammed AMINE** et ma mère **Malika HADDAD**, ma petite famille et nos amis, pour leur soutien inconditionnel.

Introduction Générale

Contexte et problématique :

Les technologies de l'information nous génèrent une multitude de données comme jamais auparavant. Le problème n'est donc plus tant d'acquérir une masse de données, mais de l'exploiter. Pour cela il faut collecter des données de qualité, les normaliser, les classer, les agréger, et les analyser, pour les exploiter afin de prendre la bonne décision.
Dans ce but, il est nécessaire de mettre en place un système décisionnel permettant de présenter de manière simple les chiffres recueillis pour mettre en lumière la conjoncture actuelle et indiquer implicitement la voie à suivre.
Les différents processus métiers de l'université étant tous implémentés dans son système d'information, une quantité importante de données est produite chaque jour donnant lieu à l'information cartes utile mais difficile à exploiter. L'informatique décisionnelle devient ainsi un enjeumajeur pour assurer la pérennité de l'université en exploitant son capital intellectuel et en contribuant à la création de valeur.
La gestion des données est un critère essentiel pour toute université il se peut que ça soit une gestion de stock, gestion des données, gestion des ressources humaines…

Le problème se pose au niveau de l'administration, surtout lorsque le nombre des étudiants, des professeurs et des matières augment et ça devient de plus en plus délicat de traiter toutes ces données manuellement, et stocker par la suite dans l'archive de l'université. C'est ici que l'informatisation des données devient un élément nécessaire, et le traitement automatique améliore énormément le rendement de l'établissement que ça soit au niveau du gain du temps ou au niveau des ressources humaines choisi pour effectuer les tâches.

L'évolution rapide et la complexité des données, nécessite le travail avec la technologie du business intelligence pour pouvoir obtenir une vision synthétique.

Cette thèse est composée de trois éléments essentiels à savoir :

- L'analyse et l'implémentation d'un système d'information de prise de décisions lié à l'environnement numérique du travail universitaire SDUM «Système décisionnel pour l'université marocaine ».
- Le choix d'un bon outil d'extraction des données pour la réalisation de notre système.
- Analyse de données en se basant sur des algorithmes de Data Mining.

Plan du mémoire

Cette thèse s'articule autour des points suivants :

- Le premier chapitre présente l'état de l'art associé à cette thèse en décrivant l'étude de cas de notre travail. Ce chapitre commence par présenter un aperçu général sur l'informatique décisionnelle. Il se poursuit sur les magasins de données en présentant la manipulation multidimensionnelle associée. Le chapitre se termine sur l'analyse de différentes techniques du DataMining
- Le deuxième chapitre donne une étude comparative de deux outils d'aide à la décision à savoir l'open source Pentaho et le propriétaire Sqlserver. Nous mettons en évidence les Problèmes liés à chaque outil.
- Le troisième chapitre présente la démarche d'analyse et de conception d'un Système décisionnel pour l'université marocaine (**SDUM**)
- Le quatrième chapitre traite une analyse avancée du système décisionnel pour l'université marocaine (**SDUM**)
- La thèse se termine par une conclusion générale dans laquelle nous traçons un bilan de nos travaux et nous présentons ensuite des perspectives qui nous permettraient d'améliorer ce qui a été proposé.

Travaux publiés

Les travaux de recherche de ce travail ont fait l'objet des publications suivantes :

a) Publications dans des journaux :

Abdellah Amine, Rachid Ait Daoud, Belaid Bouikhalene,
"Development of a Decision-Making System for Sultan Moulay Slimane University in Beni Mellal, Morocco".
Indonesian Journal of Electrical Engineering and Computer Science Vol. 2, No. 2, May 2016, pp. 469 ~ 477 DOI: 10.11591/ijeecs.v2.i2.pp469-477

Abdellah Amine, Rachid Ait Daoud, Belaid Bouikhalene,
"Efficiency Comparaison and Evaluation between Two ETL Extraction Tools".
Indonesian Journal of Electrical Engineering and Computer Science Vol. 3, No. 1, July 2016, pp. 174 ~ 181 DOI: 10.11591/ijeecs.v3.i1.pp174-181.

Rachid Aitdaoud, Abdellah Amine, Belaid Bouikhalene, Rachid Lbibb,
"Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model: The Case of Online Stores in Morocco".
World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol: 9, No: 8, 2015.

Rachid Aitdaoud, Abdellah Amine, Belaid Bouikhalene, Rachid Lbibb
"Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context".
International Journal of Electrical and Computer Engineering Vol. 8, No. 4, 2018

b)Communications dans des congrès internationaux :

Abdellah Amine*, Rachid Ait Daoud, Belaid Bouikhalene,
"Application of data mining to the classification of students in the University of Beni Mellal, Morocco".

CBI'18 The 4th International Conference on Business Intelligence, April 25 - 27, 2018, Beni Mellal, Morocco

Rachid Aitdaoud, Abdellah Amine, Belaid Bouikhalene, Rachid Lbibb.
"Combining RFM Model a Clustering Technique for Customer Value Analysis of a Company Selling Online".
12th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2015 November 17-20, 2015, Marrakech, Morocco.

I- Systèmes décisionnels

I.1 Introduction

La complexité et le volume des données évoluent rapidement, telles que les données médicales, socio-économiques, démographiques et environnementales, qui obligent l'ensemble des responsables de gérer toujours plus d'informations permettant l'utilisation de ces données à des fins d'analyse et de prise de décision.

Ils ont besoin d'outils et de modèles pour la mise en place de systèmes décisionnels comportant des données évolutives.

La prise de décision est devenue dès les années 90une activité primordiale nécessitant la mise en place de systèmes dédies efficaces [1].

Dans ce chapitre, nous donnons un aperçu sur les notions élémentaires des systèmes décisionnels et sur les outils ETL. Par la suite nous présentons la modélisation des entrepôts de données, des magasins de données et la manipulation de données multidimensionnelles, Ensuite, nous basculons vers l'architecture et les composantes de base d'un entrepôt de données ainsi que ses étapes de construction. Enfin, nous abordons les différentes techniques du DataMining

I.2 Systèmes d'aide à la décision

Une des préoccupations essentielles pour les dirigeants et décideurs d'une entreprise réside dans l'exploitation des informations contenues dans les systèmes opérationnels pour une meilleure prise de décision. De nos jours, les bases de données des entreprises contiennent d'extraordinaires quantités de données, dont la taille se compte parfois en téraoctets [2] et les décideurs ont besoin d'une vision synthétique et globale des informations circulant dans leur organisation afin de guider et adapter leur prise de décision. Pour réaliser facilement ce processus, ils emploient des systèmes d'aide à la décision. Ces outils permettent aux utilisateurs finaux d'avoir une vision globale sur les activités d'une entreprise par un accès rapide et interactif à un ensemble de vues des données organisées pour refléter l'aspect multidimensionnel des données de l'entreprise [3].

L'informatique décisionnelle (Decision Support System ou Business Intelligence) désigne les méthodes, les outils et les moyens qui permettent de collecter, consolider, modéliser les données d'une entreprise afin d'offrir une aide à la décision et de permettre au corps exécutif d'une entreprise d'avoir une vue d'ensemble de l'activité [4].

I.3 Outils ETL

- L'outil ETL se décompose en trois phases :
    - La phase d'Extraction consiste à collecter les données en provenance d'une ou plusieurs sources.
    - La phase de transformation consiste à reformater et à transformer les données.
    - La phase de chargement (loading) consiste à transférer les données vers la DataWarehouse, le Data Store.

Pour analyser les données, il est indispensable de les rassembler en un seul endroit. Or, les données d'une entreprise se trouvent dans de multiples endroits et, souvent, elles ne sont pas cohérentes. Afin de rassembler et de nettoyer les données, nous utilisons les techniques de type ETL.

I.3.1 Définition

Un ETL est un acronyme pour ExtractTransform and Load, en français, Extraire, transformer et loader. Il s'agit d'un logiciel qui se place au début de la chaine de production des datas. Son rôle est de préparer les datas avant leur intégration dans un entrepôt de données (datawarehouse).

I.3.2 Fonctionnement

Les transformations confiées à un ETL sont souvent simples, mais elles peuvent dans certains cas inclure des traitements procéduraux, de véritables programmes spécifiques.

Un ETL permet d'éviter la réalisation de programmes batch répétitifs, souvent semblables, dont il faudra également assurer la maintenance. Le principe est que l'intégration d'un nouveau flux de données ne requiert aucun développement, et s'opère par une simple configuration interactive : on choisit les éléments de données dans le référentiel source, on indique les transformations simples qu'ils doivent subir, et on précise la destination de la donnée dans le datawarehouse [5].

Un ETL prend en charge différentes sources de données, en entrée et en sortie. Les principales :

- SGBD relationnels [6],
- les flux XML [7],
- fichiers à formats fixes ou avec séparateurs (CSV).

Une fois qu'un flux d'extraction-transformation-chargement a été défini, il est généralement déclenché de manière régulière, ceci sous le contrôle d'un outil de planification de tâches, ou bien d'ordonnancement.

Un ETL traite généralement des flux de point à point, c'est à dire entre une source unique et une destination unique.

I.3.3 Quelques ETL :

Open Source Propriétaire

- Apatar, CloverETL, Pentaho Data Integration, Scriptella, Talend Open Studio, GeoKettle, HPCC Systems, Jedox, GeoKettle, Jasper ETL.
- Les outils ETL open source ont une connectivité limitée, Une capacité à tester et un seuil de complexité qui nécessite l'interrogation des traitements données supplémentaires.

- Les ETL open source sont certes gratuits. Par contre, leurs coûts cachés en matière d'intégration, pour des fonctionnalités additionnelles, empêchent de budgétiser véritablement leur déploiement.

Version payante

- Datastage 7.1, Kettle etTalend, BODI (Business Objects Data Integrator), ODI (Oracle Data Integrator), SSIS.
  Les ETL propriétaires permettent de mettre en œuvre très rapidement les traitements d'intégration des données.

I.4 Les indicateurs

Un indicateur [8] est un outil d'évaluation qui nous permet de mesurer objectivement une situation et de suivre les progrès d'une organisation dans l'atteinte de ses objectifs stratégiques.
Un bon indicateur est un indicateur :
-**S**ignificatif : a une raison d'être (l'objectif stratégique auquel il se rattache).
-**M**esurable : on dispose de tous les éléments pour le calculer.
-**A**ctionnable : on dispose des leviers d'actions pour agir sur son niveau.
-**R**esponsabilisé : un acteur responsable de son niveau est clairement désigné.

-**T**emporellement défini : a une périodicité de production et de suivi définie.
La BI est souvent utilisée pour analyser le passé, également utilisée dans le reporting afin d'avoir une vue en temps réel d'une activité. Cela permet de savoir à tout moment où nous en sommes et grâce à des techniques scientifiques et informatique, il est possible d'obtenir des prévisions sur le futur.

I.5 La modélisation des entrepôts de données

D'après InmonW.H[9] [10] les entrepôts de données constituent une solution adéquate pour construire un système décisionnel.

Un entrepôt de données est défini comme étant « une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse ».

Cette définition met l'accent sur les caractéristiques suivantes :

Intégrées : Les données alimentant l'entrepôt proviennent de sources multiples et hétérogènes. Les données des systèmes de production doivent être converties, reformatées et nettoyées de façon à avoir une seule vision globale dans l'entrepôt.

Orientées sujet : Les données s'organisent par sujets ou thèmes, contrairement auxdonnées des systèmes de production généralement organisées par processus fonctionnel.

L'intérêt de cette organisation est de disposer de l'ensemble des informations utiles sur un

sujet, le plus souvent transversal aux structures fonctionnelles et organisationnelles del'entreprise.

Variant selon le temps : Toutes les données d'un entrepôt sont identifiées par des périodes temporelles spécifiques. On parle d'historisation des données [11].

Non volatiles : Les données des systèmes opérationnels sont constamment manipulées, modifiées ; elles sont mises à jour à chaque nouvelle transaction.

De son côté, Ralph Kimball [12] a fourni une définition plus simple d'un entrepôt de données, mais qui n'en est pas moins précise : « un entrepôt de données est une copie des données transactionnelles d'une entreprise structurée de manière spécifique pour l'interrogation et

l'analyse. ».

Par contre la dernière approche c'est l'approche Hybride "Mixte [13].

Comme son nom l'indique, est un mix entre les deux approches déjà citées. On commence par concevoir un modèle de données de l'entreprise en même temps que les modèles spécifiques. Puis on crée un modèle normalisé d'entreprise de haut niveau ainsi générer les modèles des premiers Datamart qui seront chargés avec les données atomiques en utilisant un staging area temporaire.

Dans la Figure suivante, nous schématisons l'architecture des systèmes décisionnels [14]

Qui se base sur le modèle de Ralph Kimball, basé sur une schématisation en étoile.

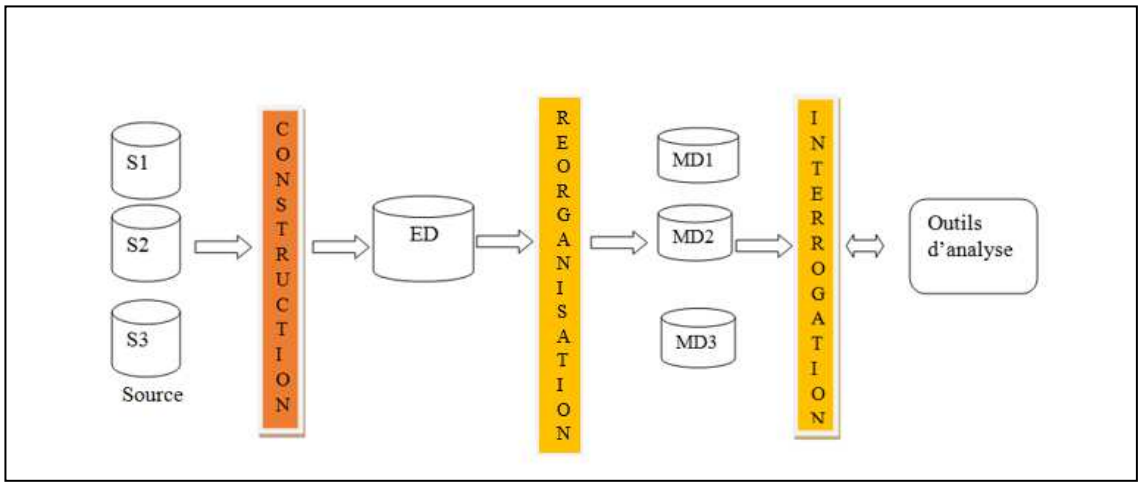


Figure 1:ENTREPOT ET MAGASINS DE DONNEES

La construction consiste à extraire les données pertinentes pour la prise de décision

et à les recopier dans l'entrepôt de données. Celui-ci constitue une collection centralisée de

données matérialisées et historiques[15]. Le modèle de l'entrepôt doitsupporter des structures complexes[16] et supporter l'évolution desdonnées [17][18][19][20]

La réorganisation permet de réorganiser les données entreposées en les stockantdans des magasins de données visant à supporter efficacement les processusd'interrogation et d'analyse [21].

L'interrogation consiste à utiliser les données des magasins pour prendre desdécisions. La représentation des données doit faciliter leur compréhension et leurmanipulation par les utilisateurs finaux.

I.6 Concepts de la modélisation dimensionnelle

Tous les cours sur la modélisation des bases de données évoquent le modèle entité-relation (ou relationnel) et les formes normales. Ces types de modèles ont été conçus pour donner aux bases de données les caractéristiques suivantes :

- Une facilité de mise à jour des données.
- Une garantie contre les incohérences des données.
- Limiter la redondance des données (et ainsi limiter la taille de la base).

Avec l'apparition des entrepôts de données, qui sont des systèmes décisionnels basés sur le concept multidimensionnel, tous les utilisateurs peuvent désormais accéder à des grandes quantités d'information.

La modélisation multidimensionnelle propose d'analyser des indicateurs numériques (par exemple chiffre d'affaires, nombre d'individus etc.) dans un contexte précisé par le croisement de plusieurs dimensions d'analyse (par exemple temps, géographie, organisation, produits, etc.), généralement présentées sous forme d'arbres hiérarchiques.

I.7 Modélisation des magasins de données

Un magasin de données est un extrait de l'entrepôt conforme à des besoins d'analyse particuliers et organisé selon un modèle adapté aux outils d'analyse et d'interrogation décisionnelle. Les magasins de données sont généralement stockés au sein d'une base de données multidimensionnelle (BDM) et emploient généralement une modélisation dite en étoile ou en flocon pour représenter les sujets et les axes d'analyse des données. Au niveau logique, la base de données multidimensionnelle hébergeant le magasin de données est souvent structurée avec une technologie relationnelle dite « relational-OLAP » (ROLAP) ou multidimensionnelle : « multidimensionnel-OLAP » (MOLAP) [22][23].

Par la suite, le terme magasin de données classique désignera un magasin de données modélisé selon un schéma en étoile ou flocons et bâti sur une architecture ROLAP, MOLAP

- **Module en étoile**

Un schéma en étoile [24] est une structure dimensionnelle qui représente une seule table de faits entourée par un seul cercle de dimensions. Toute dimension à niveaux multiples est aplatie en une seule dimension. Le schéma en étoile est conçu pour répondre à des requêtes inhérentes à la structure dimension-fait.

- **Module en flocon de neige**

Le schéma en flocon de neige est une extension du schéma en étoile. Dans un schéma en étoile, les informations associées à une hiérarchie de dimension, sont représentées dans une seule table, même si les différents niveaux de la hiérarchie ont des propriétés différentes. Le schéma en flocon est le résultat de la décomposition d'une ou plusieurs dimensions en plusieurs niveaux formant une hiérarchie. Les tables de dimensions sont ainsi éclatées en plusieurs tables, ce qui peut être vu comme une normalisation des tables de dimensions. La table de faits reste inchangée

- **Module en constellation**

Les schémas en constellations sont des schémas où plusieurs modèles dimensionnels se partagent les mêmes dimensions, c'est-à-dire les tables de faits ont des tables de dimensions en commun. Les tables de dimensions partagées par plusieurs tables de fait doivent être exactement les mêmes [25].

- **Les tables de faits**

Chaque entrepôt de données inclut une ou plusieurs tables de faits [26]. Centrale par rapport à un schéma en étoile ou en flocons, une table de faits capture les données qui mesurent les opérations de l'équipe. Les tables de faits contiennent habituellement de grands nombres de lignes, en particulier lorsqu'elles contiennent une ou plusieurs années d'historique pour un grand projet d'équipe

- **Les tables de dimension**

Elles contiennent les données brutes non calculées et des attributs sous forme de descriptions textuelles permettant de qualifier l'activité [27].

- **Les tables d'agrégats**

D'une manière générale, le mot agrégat désigne l'action d'agréger, de regrouper des éléments. En BI les Agrégats [28] sont les résultats calculés des données contenues dans une table de faits.

- **Axes d'analyse**

Dimension ou axe [29] est une table qui contient les axes d'analyse selon lesquels on veut étudier des données observables (les faits) qui, soumises à une analyse multidimensionnelle, fournissent aux utilisateurs des renseignements nécessaires à la prise de décision. Elle peut s'agir des clients ou des produits d'une entreprise, d'une période de temps comme un exercice financier, des activités menées au sein d'une société, etc.

I.8 Manipulation de données Multidimensionnelles

I.8.1 Les outils OLAP

Les outils OLAP [30] (On Line Analytical Process) reposent sur une base de données multidimensionnelle, destinée à exploiter rapidement les dimensions d'une population de données.

La plupart des solutions OLAP [31] reposent sur un même principe : restructurer et stocker dans un format multidimensionnel les données issues de fichiers plats ou de bases relationnelles. Ce format multidimensionnel, connu également sous le nom d'hypercube, organise les données le long de dimensions. Ainsi, les utilisateurs analysent les données suivant les axes propres à leur métier.

Ce type d'analyse multidimensionnelle nécessite à la fois l'accès à un grand volume de données et des moyens adaptés pour les analyser selon différents points de vue. Ceci inclut la capacité à discerner des relations nouvelles ou non prévues entre les variables, la capacité à identifier les paramètres nécessaires à manier un volume important de données pour créer un nombre illimité de dimensions et pour spécifier des expressions et conditions inter dimensions. Ces dimensions représentent les chemins de consolidation.

OLAP concerne de ce fait au moins autant le monde des serveurs, voire des structures de stockage, que celui des outils.

Afin de formaliser le concept OLAP, fin 1993, à la demande de Arbor Software, Edgar F. Codd publie un article intitulé « *Providing OLAP to User Analysts* » aux Etats Unis, dans lequel il définit 12 règles que tout système de pilotage multidimensionnel devrait respecter.

« *Ce qu'il y a d'agréable avec ces outils OLAP* », explique EricKlusman, de CantorFitzgerald LP, *"c'est que je suis en mesure de distribuer les données aux utilisateurs sans les obliger à apprendre des complexes formules de programmation, d'interrogation ou même à ce qu'ils aient à programmer leurs tableurs"*. D'une façon générale, tous affirment que l'on peut interfacer de nombreux outils d'utilisateurs avec des bases de données multidimensionnelles sans qu'il soit nécessaire de consentir de lourds efforts de formation ou des interventions importantes du service informatique.

Règle 1 : Vue Conceptuelle Multidimensionnelle des données

Ces modèles permettent des manipulations simples : rotation, pivot ou vues par tranche, analyse de type permutations d'axes *(slice and dice)* ou en cascade *(drillanywhere).*

Règle 2 : Le système est transparent pour l'utilisateur

Cette transparence se traduit pour l'utilisateur par un complément à ses outilshabituels garantissant ainsi sa productivité et sa compétence. Elle s'appuie sur une architecture ouverte permettant à l'utilisateur d'implanter le système OLAP sans affecter les fonctionnalités du système central.

Par ailleurs, l'utilisateur ne doit pas être concerné par l'intégration des données dans OLAP provenant d'un environnement homogène ou hétérogène.

Règle 3 : Accessibilité à toutes les données utiles à la décision

Le système OLAP doit donner accès aux données nécessaires aux analysesdemandées. Les outils OLAP doivent avoir leur propre schéma logique de stockage des données physiques hétérogènes, doivent accéder aux données et réaliser n'importe quelle conversion afin de présenter à l'utilisateur une vue simple et cohérente. Ils doivent aussi savoir de quel type de systèmes proviennent les données.

Règle 4 : La performance demeure stable quel que soit le volume de données

L'augmentation du nombre de dimensions ou du volume de la base de données ne doit pas entraîner de dégradation visible par l'utilisateur.

Règle 5 : L'architecture est Client / Serveur

La plupart des données pour OLAP sont stockées sur des gros systèmes et sont accessibles via des PC. Il est donc nécessaire que les produits OLAP soient capables de travailler dans un environnement Client/serveur.

Règle 6 : Toutes les dimensions sont équivalentes en structure et en calcul

Toutes les dimensions doivent être équivalentes en structure et en calcul. Ilne doit exister qu'une seule structure logique pour toutes les dimensions. Toute fonction qui s'applique à une dimension doit être aussi capable de s'appliquer à une autre dimension.

Règle 7 : Le système gère dynamiquement les Matrices Creuses

Le schéma physique des outils OLAP doit s'adapter entièrement au modèle d'analyse spécifique créé pour optimiser la gestion des matrices creuses. En effet, dans une analyse à la fois sur les produits et les régions, tous les produits ne sont pas vendus dans toutes les régions.

Règle 8 : L'accès possible à plusieurs utilisateurs simultanément

Les outils OLAP doivent supporter les accès concurrents, garantir l'intégrité et la sécurité afin que plusieurs utilisateurs accèdent au même modèle d'analyse.

Règle 9 : Il n'y a pas d'opérations restrictives sur les dimensions. Les opérations doivent pouvoir s'effectuer sur toutes les dimensions et ne doivent pas faire intervenir l'utilisateur pour définir un calcul hiérarchique.

Règle 10 : Manipulation intuitive des données

Toute manipulation doit être accomplie via une action directe sur les cellules du modèle sans utiliser de menus ou des chemins multiples à travers l'interface utilisateur.

Règle 11 : Souplesse de Création de Rapports

La création des rapports dans les outils OLAP doit permettre aux utilisateurs de présenter comme ils le désirent des données synthétiques ou des résultats en fonction de l'orientation du modèle.

Règle 12 : Nombre illimité de dimensions et de niveaux d'agrégation

Tout outil OLAP doit gérer au moins 15 à 20 dimensions.

I.8.2 Les outils MOLAP

OLAP sera l'outil à privilégier pour les données quantitatives si leur structuration a priori est naturelle (cas rencontré fréquemment pour les applications financières ou commerciales), alors que le requêteur sera idéal pour les données qualitatives et pour toute analyse impromptue nécessitant l'autonomie de l'utilisateur (cas rencontré fréquemment pour le marketing ou la gestion du personnel). Si les besoins sont à combiner, il faudra choisir entre la richesse fonctionnelle apportée par plusieurs outils interfacés ou l'homogénéité des outils intégrés.

Deux versions d'OLAP s'affrontent actuellement. Les outils MOLAP [32] (Multidimensional OLAP) d'une part qui s'appuient sur une base de données multidimensionnelle. Les outils ROLAP [33] (Relational OLAP) d'autre part, qui représente leur équivalent sur une base de données relationnelle.

MOLAP est conçue exclusivement pour l'analyse multidimensionnelle, avec un mode de stockage optimisé par rapport aux chemins d'accès prédéfinis. Ainsi, toute valeur d'indicateur associée à l'axe temps sera pré-calculée au chargement pour toutes ses valeurs hebdomadaires, mensuelles, etc.

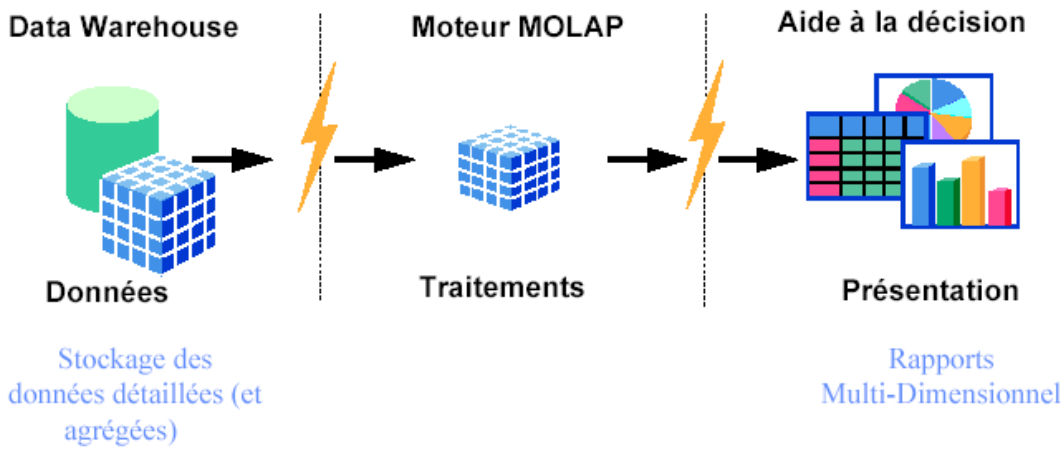


Figure 2: Traitement avec MOLAP

MOLAP agrège tout par défaut. Plus le volume de données à gérer est important, plus les principes d'agrégations implicites proposés par MOLAP sont pénalisants dans la phase de chargement de la base, tant en terme de performances que de volume. La limite fréquemment évoquée pour MOLAP étant de quelques giga octets.
MOLAP surpasse ROLAP pour des fonctionnalités avancées comme la prévision ou la mise à jour des données pour la simulation. Cependant, ces différences s'expliquent par une plus grande maturité en faveur de MOLAP, concept qui date de près de vingt ans.

MOLAP est incompatible avec d'autres modes d'accès aux données. Si MOLAP doit cohabiter avec d'autres techniques d'accès aux données (par requêteur, par data mining, etc.), deux bases de données doivent cohabiter. En effet, MOLAP repose sur un moteur spécialisé, qui stocke les données dans un format tabulaire propriétaire (cube). Pour accéder aux données de ce cube, on ne peut pas utiliser le langage de requête standard SQL, il faut utiliser une API spécifique.
Le marché des bases MOLAP étant plus réduit, il est plus difficile pour les éditeurs qui le représentent d'investir sur de telles évolutions.

I.8.3 Les outils ROLAP

Les outils ROLAP superposent au-dessus des SGBD/R bidimensionnels un modèle qui représente les données dans un format multidimensionnel. Ces produits diminuent sensiblement le coût lié à la mise en œuvre d'un serveur de base de données multidimensionnelle supplémentaire. Au travers des métadonnées, ils permettent de transformer l'analyse multidimensionnelle demandée par l'utilisateur en requêtes SQL. Pour

cela, ces outils s'appuient pour la plupart sur une modélisation particulière des données, distinguant les axes d'analyse et les faits à observer. On parlera notamment de modèle en étoile et de modèle en flocon ou encore des techniques de définition physique d'agrégations.
Ceci oblige à définir le modèle en fonction de l'outil à utiliser et des analyses à mener mais est un gage de performance et de cohérence lors de l'utilisation de ce type de produits.
Cette contrainte exige un travail important des équipes informatiques et donc enlève beaucoup à l'intérêt d'utiliser un SGBD Relationnel comme support de stockage pour l'analyse multidimensionnelle.

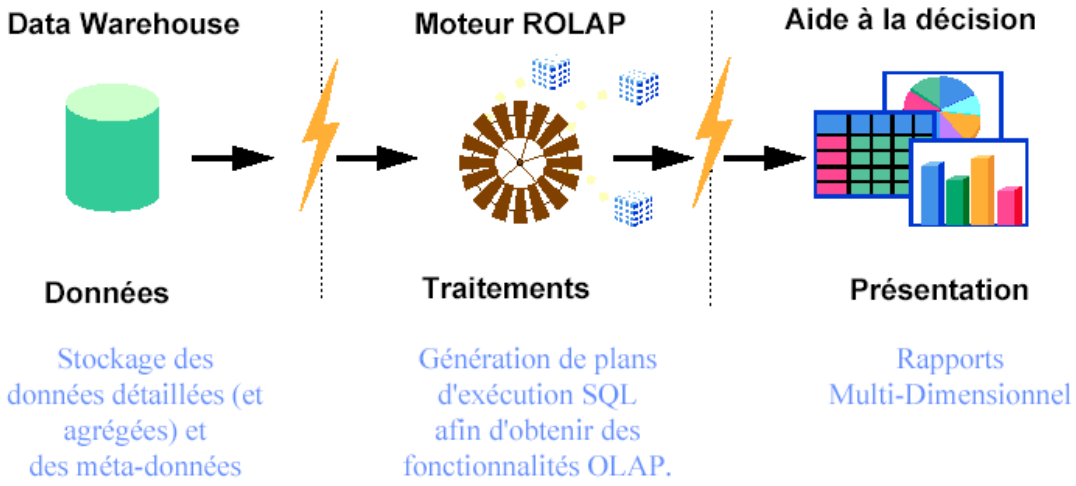


Figure 3: Traitement avec ROLAP

Les outils ROLAP proposent le plus souvent un composant serveur, pour optimiser les performances lors de la navigation dans les données ou pour les calculs complexes. Avec ROLAP, il est déconseillé d'accéder en direct à des bases de données de production pour faire des analyses sérieuses, pour des raisons de performances.
ROLAP n'agrège rien, mais tire parti des agrégats s'ils existent. De ce fait ROLAP est plus lourd à administrer que MOLAP, puisqu'il demande de créer explicitement certains agrégats.
Certains éditeurs, comme Informix avec Métacube ou Oracle avec Discoverer 2000, pallient cependant à cette faiblesse avec des outils d'administration aptes à conseiller pour une politique d'agrégation adéquate. ROLAP est donc mieux adapté aux gros volumes.
En s'appuyant sur les bases relationnelles, référence du marché, ROLAP tire parti des évolutions de celles-ci (adaptation aux architectures hardware sophistiquées, extensions objets, etc.).

I.8.4 Les outils HOLAP

Le HOLAP (Hybrid OLAP) [34] est une architecture hétérogène composé de tout ou partie des architectures précitées. Les données sont stockées dans des tables relationnelles et les données agrégées sont stockées dans des cubes. Les requêtes vont chercher les données dans les tables et les cubes.

I.8.5 Les outils DOLAP

Le DOLAP [35][36] (Desktop OLAP) décrit une catégorie de produits qui ne sont pas nécessairement connectés à un serveur. Ils peuvent s'exécuter sur un client avec la possibilité d'utiliser une source de données sous la forme d'un « Data Cube» construit et stocké localement sur une machine utilisateur.

I.9 Présentation du datamining

Le terme de Data Mining [37] est souvent employé pour désigner l'ensemble des outils permettant à l'utilisateur d'accéder aux données de l'entreprise, de les analyser. Nous restreindrons ici le terme de Data Mining aux outils ayant pour objet de générer des informations **riches** à partir des données de l'entreprise, notamment des données **historiques**, de découvrir des **modèles** implicites dans les données. Ils peuvent permettre par exemple à un magasin de dégager des profils de client et des achats types et de prévoir ainsi les ventes futures. Il permet d'augmenter la valeur des données contenues dans le DWH.

Les outils d'aide à la décision, qu'ils soient relationnels ou OLAP, laissent l'initiative à l'utilisateur, qui choisit les éléments qu'il veut observer ou analyser. Au contraire, dans le cas du Data Mining, le **système à l'initiative** et découvre lui-même les associations entre données, sans que l'utilisateur ait à lui dire de rechercher plutôt dans telle ou telle direction ou à poser des hypothèses. Il est alors possible de prédire l'avenir, par exemple le comportement d'un client, et de détecter, dans le passé, les données inusuelles, exceptionnelles.

Ces outils ne sont plus destinés aux seuls experts statisticiens mais doivent pouvoir être employés par des utilisateurs connaissant leur métier et voulant l'analyser, l'explorer. Seul un utilisateur connaissant le métier peut déterminer si les modèles, les règles, les tendances trouvées par l'outil sont pertinents, intéressantes et utiles à l'entreprise. Ces utilisateurs n'ont donc pas obligatoirement un bagage statistique important. L'outil doit donc soit être ergonomique, facile à utiliser et rendant transparentes toutes les formules mathématiques et

termes techniques utilisés, soit permettre de construire une application «clé en main», rendant à l'utilisateur transparentes toutes les techniques utilisées.

On pourrait définir le Data Mining comme une démarche ayant pour objet de découvrir des relations et des faits, à la fois nouveaux et significatifs, sur de grands ensembles de données.

On devrait ajouter que la pertinence et l'intérêt du Data Mining sont conditionnés par les enjeux attachés à la démarche entreprise, qui doit être guidée par des objectifs directeurs clairement explicités ("améliorer la performance commerciale", "mieux cibler les prospects", "fidéliser la clientèle", "mieux comprendre les performances de production"...).Le succès du concept de Data warehouse et le nombre croissant de bases de données décisionnelles disponibles dans les entreprises, dynamise fortement l'offre Data Mining.

Le terme de Data Mining signifie littéralement **forage de données.** Comme dans tout forage, son but est de pouvoir extraire un élément : la connaissance. Ces concepts s'appuient sur le constat qu'il existe au sein de chaque entreprise des informations cachées dans le gisement de données. Ils permettent, grâce à un certain nombre de techniques spécifiques, de faire apparaître des connaissances.

Nous appellerons Data Mining l'ensemble des techniques qui permettent de transformer les données en connaissances.

L'exploration se fait sur l'initiative du système, par un utilisateur métier, et son but est de remplir l'une des tâches suivantes : classification, estimation, prédiction, regroupement par similitudes, segmentation (ou clusterisation), description et, dans une moindre mesure, l'optimisation. [38]

I.9.1 Data Mining et la Recherche Opérationnelle

La recherche opérationnelle n'est pas assimilée aux techniques de Data Mining. Son objectif est l'optimisation et la recherche prouvée de la meilleure solution, ce qui n'est pas le cas du Data Mining [39] :

- Son champ d'application est plus large,
- On ne recherche pas la meilleure solution prouvée mais à faire le mieux possible,
- Enfin un outil de Data Mining appliqué à un même ensemble de données ne donne pas toujours les mêmes résultats, contrairement à la recherche opérationnelle.

I.9.2 Statistiques et Data Mining

On pourrait croire que les techniques de Data Mining [40] viennent en remplacement des statistiques. En fait, il n'en est rien et elles sont omniprésentes. On les utilise :

- Pour faire une analyse préalable,
- Pour estimer ou alimenter les valeurs manquantes,
- Pendant le processus pour évaluer la qualité des estimations,
- Après le processus pour mesurer les actions entreprises et faire un bilan.

Statistiques et Data Mining sont tout à fait complémentaires.

I.9.3 Recherche des connaissances

KDD est un processus non trivial, semi-automatique et itératif, qui sert à l'identification, au sein des données, de patterns valides, nouveaux, utiles et bien compréhensibles. Il passe par plusieurs étapes, du traitement et préparation des données, jusqu'à l'interprétation des résultats, en passant par la méthode de recherche des connaissances « le datamining ». En fait, le datamining se réfère à une étape particulière dans le processus KDD pour l'application des algorithmes spécifiques pour extraction et recherche des motifs à partir des données. Ce processus est présenté dans la figure

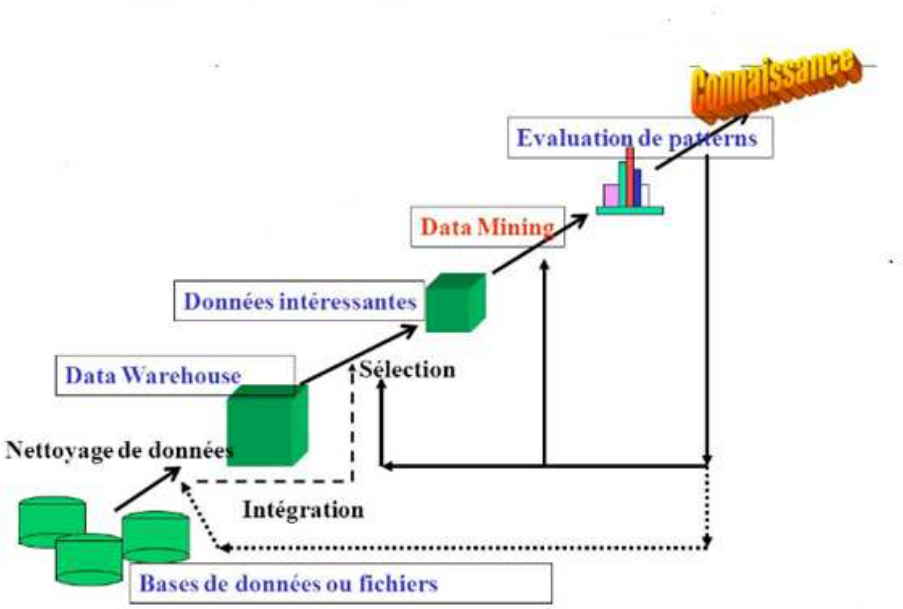


Figure 4: Processus KDD

I.9.4 Développement d'une compréhension du domaine

Dans cette étape, on répond aux plusieurs questions. A savoir :

• Quels sont les obstacles du domaine.

• Que nécessite l'automatisation et qui est manuellement transformable ?

• Quels sont les enjeux ?

• Quels sont les normes de performance ?

• Est-ce qu'on va utiliser la classification, la visualisation, l'exploration ou quelque chose d'autre.

I.9.5 Création des données cible et la sélection des données

Dans cette étape le but c'est de sélectionner et créer un ensemble des données sur lequel la découverte sera effectuée, en fonction des objectifs, puis, on prépare les données qui seront utilisées pour la découverte de connaissances, qui sont : les données disponibles, les données supplémentaires si nécessaires et l'intégration de toutes les données pour la découverte des connaissances cachée dans le data set (ensemble de données).

I.9.6 Nettoyage des données et prétraitement

Cette étape comprend le nettoyage et la purification des données, comme la gestion des valeurs manquantes, et la suppression des valeurs incorrectes. Tout en essayant de corriger les fautes afin de trouver une stratégie stricte, on peut utiliser des méthodes statistiques complexes ou des algorithmes de data mining.

I.9.7 Réduction des données et transformation

Cette étape chargée de traitement des données de l'étape précédente pour la mise en œuvre du data mining, afin de générer un ensemble important des données de meilleure qualité. Dans cette phase on s'intéresse particulièrement à

- la réduction de dimension et extraction de résumés de données
- la transformation des attributs : on applique la discrétisation à des variables continues et on numérise les variables nominales et puis on invente des nouveaux variables.

I.10 Les tâches du Data Mining

Contrairement aux idées reçues, le Data Mining n'est pas le remède miracle capable de résoudre toutes les difficultés ou besoins de l'entreprise. Cependant, une multitude de

problèmes d'ordre intellectuel, économique ou commercial peuvent être regroupés, dans leur formalisation, dans l'une des tâches suivantes :

Classification,Estimation,Prédiction,Groupement par similitudes,Segmentation (ou clusterisation),Description,Optimisation.

Afin de lever toute ambiguïté sur des termes qui peuvent paraître similaires, il semble raisonnable de les définir.

I.10.1 La classification

La classification se fait naturellement depuis déjà bien longtemps pour comprendre et communiquer notre vision du monde

« La classification consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini. » [41]

Dans le cadre informatique, les éléments sont représentés par un enregistrement et le résultat de la classification viendra alimenter un champ supplémentaire.

La classification permet de créer des classes d'individus (terme à prendre dans son acception statistique). Celles-ci sont discrètes : homme / femme, oui / non, rouge / vert / bleu, ...

Les techniques les plus appropriées à la classification sont :

- Les arbres de décision,
- Le raisonnement basé sur la mémoire,
- Eventuellement l'analyse des liens.

Pour dépasser la limite de la classification binaire dans des modèles de réponse, Charles et al. ont appliqué ada-boos [42] aux réseaux bayésiens [43] [44] tels que les algorithmes pour la programmation efficace afin de classer les clients selon leur aptitude à répondre aux offres

I.10.2 L'estimation

Contrairement à la classification, le résultat d'une estimation [45] permet d'obtenir une variable continue. Celle-ci est obtenue par une ou plusieurs fonctions combinant les données en entrée. Le résultat d'une estimation permet de procéder aux classifications grâce à un

barème. Par exemple, on peut estimer le revenu d'un ménage selon divers critères (type de véhicule et nombre, profession ou catégorie socioprofessionnelle, type d'habitation, etc.).

Il sera ensuite possible de définir des tranches de revenus pour classifier les individus.

Un des intérêts de l'estimation est de pouvoir ordonner les résultats pour ne retenir si on le désire que les n meilleures valeurs. Cette technique sera souvent utilisée en marketing, combinée à d'autres, pour proposer des offres aux meilleurs clients potentiels. Enfin, il est facile de mesurer la position d'un élément dans sa classe si celui-ci a été estimé, ce qui peut être particulièrement important pour les cas limitrophes.

La technique la plus appropriée à l'estimation est : le réseau de neurones.

I.10.3 La prédiction

La prédiction [46] ressemble à la classification et à l'estimation mais dans une échelle temporelle différente. Tout comme les tâches précédentes, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé. La seule méthode pour mesurer la qualité de la prédiction est d'attendre !

Les techniques les plus appropriées à la prédiction sont :

- L'analyse du panier de la ménagère
- Le raisonnement basé sur la mémoire
- Les arbres de décision
- les réseaux de neurones

I.10.4 L'analyse des clusters

L'analyse des clusters consiste à segmenter une population hétérogène en sous populations homogènes. Contrairement à la classification, les sous populations ne sont pas préétablis. La technique la plus appropriée à la clusterisation est l'analyse des clusters [47]

I.10.5 La description

C'est souvent l'une des premières tâches demandées à un outil de Data Mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent une exploitation supplémentaire en vue de fournir des explications.

I.10.6 L’optimisation

Pour résoudre de nombreux problèmes, il est courant pour chaque solution potentielle d’y associer une fonction d’évaluation. Le but de l’optimisation est de maximiser ou minimiser cette fonction. Quelques spécialistes considèrent que ce type de problème ne relève pas du Data Mining. La technique la plus appropriée à l’optimisation est le réseau de neurones

I.11 Conclusion

Ce premier chapitre présente un état de l’art des notions et concepts fondamentaux concernant les entrepôts de données à travers sa définition, son architecture, ses modèles de données utilisant le paradigme multidimensionnel (MOLAP, ROLAP et HOLAP), et son cycle de développement

Cette étude nous permet de cerner les principaux concepts clé des EDs dont on aura besoin pour tenter de répondre à notre problématique de conception d’un ED

Pour l’alimentation du Data Warehouse on est besoin de choisir un bon outil d’extraction des données Dans la suite on s’intéresse à une comparaison entre deux outils d’extractions des données << open source PentahoBI et sql Server Integration Service SSIS>> afin de choisir le meilleur pour exploiter notre modèle.

II. Microsoft SSIS et Pentaho Kettle : Une étude comparative pour les entrepôts de données

II.1 Introduction

Il existe de nombreux documents de recherche qui fournissent une analyse comparative des principaux outils ETL du marché, tels que [48][49]. Ils analysent en profondeur les fonctionnalités et les capacités offertes par ces outils.

Les outils d'extraction, de transformation et de chargement (ETL) intègrent des schémas hétérogènes,

Extraire, transformer, nettoyer, valider, filtrer et charger des données provenant de différentes sources dans un entrepôt de données. Le processus ETL et les outils associés peuvent être utilisés dans un nombre très varié dans des situations où les données doivent être nettoyées et déplacées dans une base de données. [50-52]

Dans la perspective de faire un éventail de prise de décision à bord d'une université publique, nous présentons une comparaison entre deux outils d'extraction ETL à partir d'une base de données de production contenant des informations sur les étudiants. Pour la mise en œuvre, nous utilisons les outils Pentaho et Sql Server

Le processus ETL et les outils associés peuvent être utilisés dans un grand nombre de situations où les données doivent être nettoyées et déplacées entre les sources de données. Le sujet d'ETL ne couvre que l'utilisation de deux outils ETL (Microsoft SQL Server Integration Services SSIS et PentahoKettle).

Différentes approches de conception, d'optimisation et d'automatisation des processus ETL ont été proposées ces dernières années. Dans cette section, nous examinons brièvement ces différentes approches [53]. Certains des principaux fournisseurs d'intégration de données sont : IBM, Informatica, Oracle, Microsoft, Talend, Pentaho, Information Builders, etc.

Différentes variétés d'approches pour l'intégration de l'outil ETL dans l'entrepôt de données ont été proposées. Shaker H. Ali ElSappagh [54] essaie de naviguer à travers les efforts déployés pour conceptualiser les abréviations pour ETL, DW, DM, OLAP, traitement analytique de ligne ionique. Un entrepôt de données donne un ensemble de valeurs numériques qui sont basées sur un ensemble de valeurs d'entrée dans les dimensions du formulaire [55]. Jian L, BihuaX [56], a conquis les points faibles de l'architecture traditionnelle Extract, Transform and Load et a traité une architecture à trois couches basée

sur des métadonnées. Cela a permis de développer un processus ETL plus flexible, polyvalent et efficace et enfin, ils ont conçu et mis en œuvre un nouvel outil ETL pour l'entrepôt de données de forage.

Dans ce chapitre, on se contentera au deux des outils les plus utilisés et qui peuvent répondre aux exigences des entreprises : SSIS, Pentaho.

II.2 Microsoft SQL Server Integration Services : SSIS

SSIS est fourni avec Microsoft SQL Server et nécessite une licence SQL Server pour l'utiliser. Microsoft offre également une suite complète de business intelligence. De plus, SSIS peut être utilisé avec un certain nombre de serveurs de base de données via les pilotes OLE et ADO.NET. Microsoft n'offre pas le code source dans le cadre du produit, ce qui signifie que le développeur ne peut pas apporter des modifications au produit en fonction des besoins du projet. En outre, il n'y a aucune avenue pour un développeur de contribuer à la version de future du produit autre que demander la fonctionnalité à Microsoft. [57]

Développent Studio est l'équivalent de Microsoft Visual Studio 2008 complété par des types supplémentaires de la partie de prise de décision spécifique au projet de SQL Server. Business Intelligence Development Studio (BIDS) est le principal environnement pour le développement de solutions de gestion des employés, y compris les projets Analysis Services, Integration Services et Reporting Services. Chaque type de projet fournit des modèles pour créer des objets requis pour l'intelligence d'affaires et une variété de concepteurs, d'outils et d'assistants pour manipuler ces objets.

L'outil SSIS a des points forts qui se résument :

- Développement rapide de scripts d'import / export
- Richesse des outils (tâches, connecteurs, transformateurs…)
- Visualisation des flux de données et utilisation de points d'arrêts lors de l'exécution dans Visual Studio facilitant la phase de Debug
- Facilité de déploiement et d'utilisation.
- SSIS peut gérer des données provenant de sources de données hétérogènes dans un même package. Nous disons que les sources de données peuvent être diverses, y compris des adaptateurs personnalisés ou scriptés.

- SSIS consomme des données difficiles comme les services FTP, HTTP, d'analyse, etc.…
- Parfaitement intégré à Microsoft Visual Studio et SQL Server.
- Utilisez la destination SQL Server au lieu de OLE DB; ce qui vous permet de charger des données en SQL plus rapidement.
- Idéal pour les transformations complexes, les opérations en plusieurs étapes, l'agrégation de données provenant de différents types ou sources de données et la gestion structurée des exceptions.
- Les données peuvent être chargées en parallèle vers de nombreuses destinations variées.

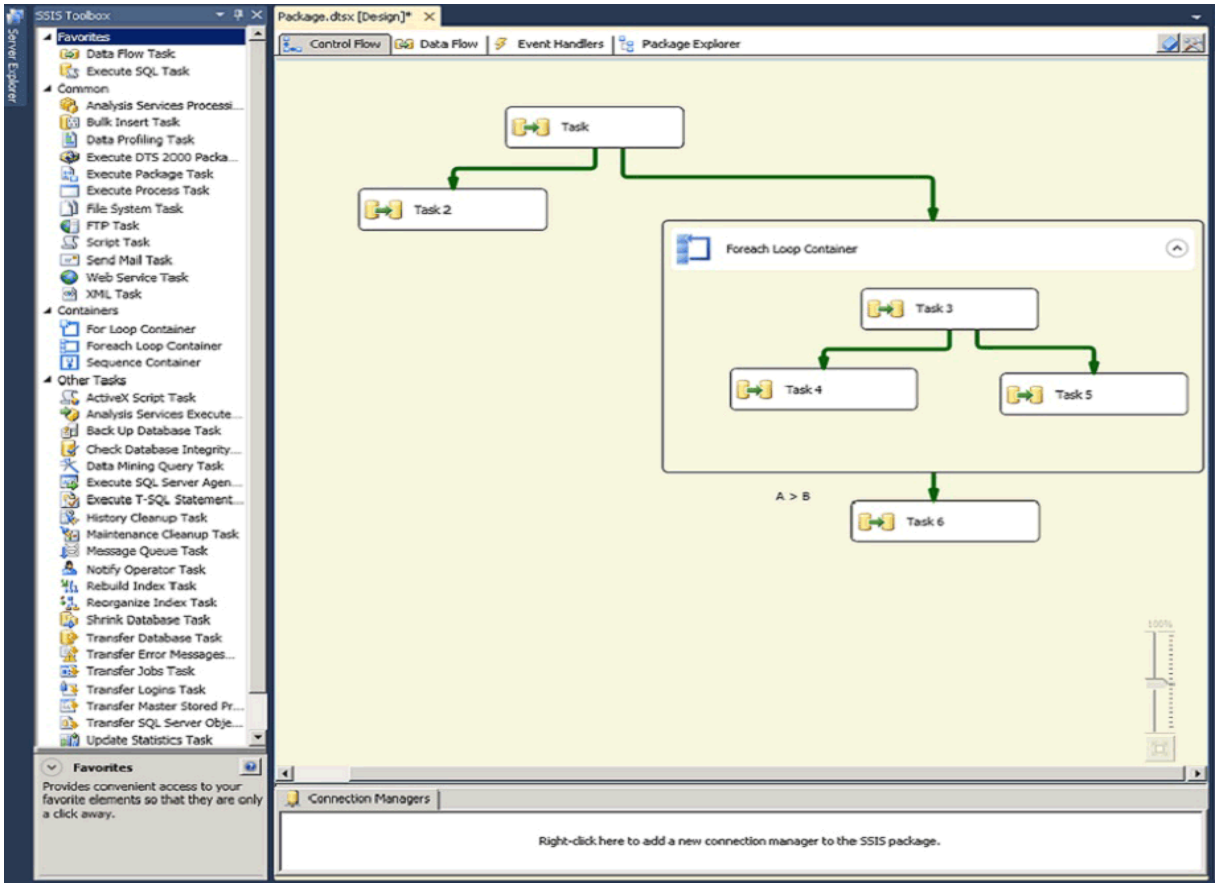


Figure 5: Capture d'écran de SSIS BIDS

Un projet BIDS comprend des sources de données, des vues de source de données et des packages SSIS.

La création d'un flux de contrôle comprend les tâches suivantes :

- ajout des conteneurs qui implémentent les flux de travail répétitifs dans un package ou divisent un flux de contrôle en sous-ensembles ;

- ajout des tâches qui prennent en charge les flux de données, préparent les données, réalisent les fonctions de flux de travail et de Business Intelligence et implémentent le script ;
- connexion des conteneurs et des tâches à l'aide de contraintes de précédence pour former un flux de contrôle ordonné.
- Ajout de gestionnaires de connexions.

Un exemple d'utilisation de différents flux de contrôle est présenté à la Figure 6. Les flux de contrôle permettent l'exécution de diverses tâches telles que SQL, VB, VC # scripting ainsi que FTP, Send Mail et Web [58].

Les flux de contrôle peuvent appeler Flux de données ou autres flux de contrôle. La relation entre les flux de données et les flux de contrôle peut être résume dans la Figure 7.

Les flux de données sont des transformations de données réelles. Les flux de données sont divisés en sources, Transformations et Destinations. Les flux de données source fournissent des composants pour accéder à une source de donnée

À l'aide de sources ADO.net, Excel, fichier plat, OLE DB et XML. Les transformations Flux de données ont 29 tâches de transformation à choisir, y compris FuzzyLookups, et regroupement Fuzzy, en changeant lentement les dimensions avec plusieurs fonctions agrégées.

Les objets de transformation SSIS sont très configurables et favorisent les appels de chaînes pour les chaînes de manipulations [59].

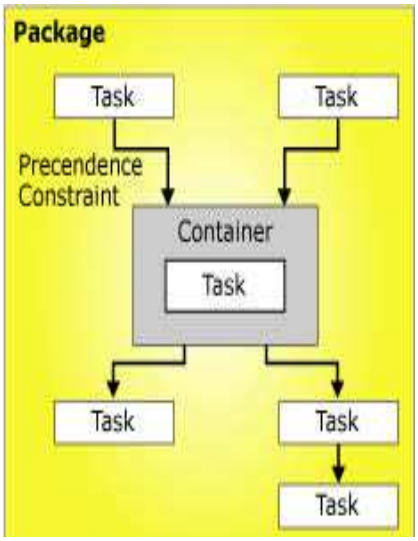


Figure 6: Exemple de tâches de flux de contrôle

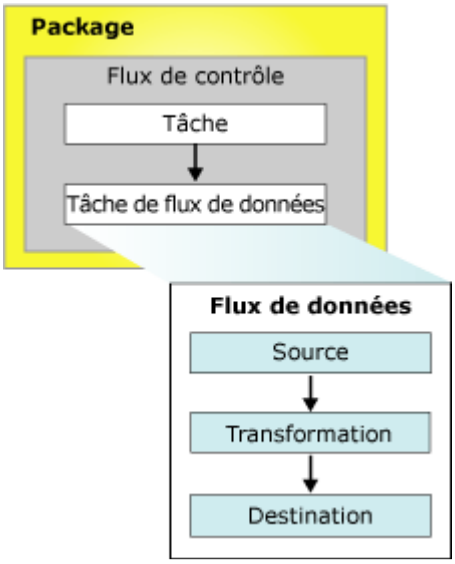


Figure 7: SSIS Contrôle et tâches de flux de données

II.3 Présentation de l'intégration de données Pentaho

Pentaho Data Integration (PDI), connue sous le nom de Kettle, est une ETL à source ouverte qui permet de concevoir et mettre en œuvre la gestion et la transformation des données. C'est un outil complet avec des fonctionnalités avancées telles que le «clustering» du traitement ETL. Ces fonctionnalités sont disponibles à partir de la version open source de PDI et ne sont trouvées que dans les versions commerciales des concurrents ETLS [60].

Pentaho Data Integration fournit une interface graphique "Spoon" (basée sur SWT), à partir de laquelle vous pouvez créer deux types de traitement : les transformations et les tâches (travaux). Les écrits et les transformations sont stockés dans un méta langue qui peut être soit Stockés au format XML ou dans une base de données. [61]

II.3.1 Interface et capacités de développement de Kettle

Pentaho Kett le comprend quatre programmes distincts

1. **Spoon** : l'environnement de développement de Pentaho qui est utilisé pour concevoir et coder les transformations. [62]
2. **Pan** : pour exécuter des fichiers de transformation XML créés par Spoon ou à partir d'une base de données [63]
3. **Kitchen** : pour exécuter des fichiers de transformation XML créés par Spoon ou d'un référentiel de base de données qui est programmé pour fonctionner en mode batch.[64]

4. **Chef** introduit une autre notion : La tâche (ou Task en anglais). Une tâche est une organisation qui permet d'automatiser des tâches complexes de transformations. En effet, l'exécution de chaque entrée ne démarre que si l'entrée précédente a été terminée. De plus, on peut être le résultat de chaque entrée. A-t-elle été exécutée avec succès ? Une entrée peut être une transformation ou des transformations spéciales comme la récupération de fichiers par FTP ou l'exécution de fichier shell

Parce que la majorité des efforts de développement auront lieu dans le programme Spoon, on va se concentrer à cette interface utilisateur qui est basée sur l'IDE Eclipse basé sur Java comme le montre la Figure 8. L'outil est organisé sous trois perspectives de visionnage :

- Intégration des données : permet la conception de processus et de tâches ETL
- Modèle : Permet de concevoir des modèles de métadonnées OLAP
- Visualise : permet de tester les modèles de métadonnées OLAP

Spoon utilise une interface avec une longue liste de fonctions intégrées. Dans Spoon le travail d'intégration de données consiste à une conception de travail et de transformation. Un modèle conceptuel de la structure du travail de Spoon peut être vu dans la figure 8. Les capacités de script de travail de Spoon sont robustes et incluent trois options de script : JavaScript, Shell et SQL Scripting. Les fonctions de transfert sont disponibles, y compris FTP, SFTP et SSH. Les composantes du transformation de Spoon sont très spécifiques et comprennent : 38 types de sources d'entrée de données, y compris un fichier texte de formats commun, applications de bureau, tables de base de données et une interface directe à SAP ainsi que d'autres applications, 26 transformations individuelles et 15 méthodes de recherche, y compris le système d'opérations de fichiers, requêtes de base de données et appels de service Web. Les transformations comprennent sept options de script telles que les classes SQL, JavaScript et Java. Spoon permet également aux développeurs pour développer ou utiliser des plugins existants pour ajouter des fonctionnalités. [65]

II.4 Comparaison entre PDI et SSIS

II.4.1 Une comparaison des activités ETL par outil

Voici un tableau montrant les transformations intégrées fournies par les deux outils.

Transformation Catégorie	**SQL SSIS**	**PentahoKettle**
Niveau de ligne : Fonction qui peut être appliqué localement à une seule rangée.	- Table des caractères - Copier la colonne - Conversion de données - Colonne dérivée - Composant de script - Commande OLE DB - Transformation du cache - Changement des dimensions - Autres filtres (non nul, sélections, etc	- Add Checksum - Add Constants - Add Sequence - Add Value fields changing sequence - Add XML - Calculator - Number Range - Replace in String - Select Values - Set field Value to a constant - Split Fields - String Operations - Strings Cut - Value Mapper - ETL Metadata injection - Filter Rows, Last Row, Java Filter, - Scripting, Java, JavaScript, SQL.
Mérou unaire : Transformer un ensemble de lignes à un une seule rangée.	- Agrégat - Pivot	- - Row Flattener - Unique Rows - Unique Rows (HashSet) - Analytic Query - Group by - Memory Group by - UnivariateStatistics
Séparateur Unaire : Diviser une seule ligne à un ensemble de lignes.	Non pivot - Groupement confus	- -Row Normaliser - Split Fields to Rows - Clone Row
Unary Holistic: Perform a transformation to the entiredataset	- Sort - PercentageSampling - RowSampling	- - Sort Rows - XSL Transformation - Change file encoding - SampleRows- SampleRows
Binary or N-ary: Combine many inputs into one output.	Union Like - Union All - Merge Join-like	Join-like - Get ID from Slave Server - Row Denormaliser - Set Field Value

	- MergeJoin - Importation de colonne - Consultation Confuse - Extraction de Terme - Consultation de Terme	- Append streams - Database Join - Database Lookup -HTTP Post, client, REST, Stream, SOAP Lookup. - Sorted Merge - Merge Join - Merge Rows (diff) Union Like - Join Rows
Routers: Locally decide for each row, which of the many outputs it should be sent to..	-Conditional Split - Multicast	Process Files - Switch/Case - Dynamic SQL Row - Mapping (input, output, sub transformation)

Tableau 1: les transformations intégrées fournies par SSIS et PDI

II.4.2 Comparaison des fonctionnalités entre PDI et SSIS

Dans cette section, nous allons faire une étude comparative des fonctionnalités des deux outils d'extraction, en particulier l'intégration de données Pentaho et Microsoft SQL Server Intégration Services.

- **Accès aux données relationnelles**

Caractéristiques	PDI	SSIS
Lecture de table complète	✓	✓
Lecture de vue complète	✓	✓
Appel de procédures stockées	✓	✓
Ajoue de clause where/orderby	✓	✓
Exécution de requète	✓	✓
Outil de création de requète	✓	✓

Lecture/écriture de types complexes de données	✓	✓
CSV	✓	✓
Fixed / Limité	✓	✓
XML	✓	✓
Excel	✓	✓
Validité des fichiers plats	x	✓
Validités des fichiers XML	✓	✓
SAP	lecture	✓
ERP	✓	✓
Web Services	✓	✓
Cubes OLAP	✓	✓
Divers	LDAP	RSS, LDAP, POP

Tableau 2: Accès aux données

Pour l'accès aux données relationnelles, des fichiers plats et les applications de connecteurs, PDI et SSIS sont de bonnes solutions pour ces fonctions. Les deux outils permet d'analyser les données de sources diverses, de déterminer les transformations nécessaires, d'exécuter des accumulations, des effacements de données, les corrections automatiques d'erreurs, etc.
Mais pour la validation des fichiers plats, l'outil SSIS est plus robuste que PDI.

- **Déclenchement des processus**

Caractéristiques	***PDI***	***SSIS***
CORBA	x	✓
XML RPC	x	✓
JMS	x	x
MOMS	x	✓
Répertoire	✓	✓

POP	✓	✓

Tableau 3: Déclenchement des processus par message et par type de polling

Nous notons pour le processus de déclenchement selon le message, l'outil PDI n'est pas approprié pour cette procédure, tandis que pour la détente par le type de polling les deux outils sont robustes

L'oracle est la seule base de données qui supporte JMS natalement en forme d'Oracle Avancé.

- **Traitement des données**

Caractéristiques	PDI	SSIS
Fonctions de transformations des dates et des notes	✓	✓
Fonctions statistiques de qualités	x	✓
Permet le transcodage par une table de référence	x	✓
Jointures hétérogènes	x	✓
Modes de jointure supportées	externe	✓
Gestion des requêtes imbriquées	x	✓
Possibilité de traitements par un langague de programmation	✓	✓
Ajout de nouvelles transformations et processus métiers	✓	✓
Mapping graphique	✓	✓
Drag and Drop	✓	✓
Représentation graphique des flux	✓	✓
Visualisation des données en cours de développement	x	✓
Outil d'analyse d'impact	✓	✓
Outils de debugging	✓	✓
Génération de documentation technique	x	✓
Génération de documentation fonctionnelle	x	✓

Consultation de la documentation à travers le web	x	✓
Gestion des erreurs d'intégration	Pour certains étapes	✓

Tableau 4: Traitement des données

Les deux outils fournissent un mécanisme de question directement dans le SQL qui permet de faire tous les modes de jointure.

Pour le traitement des données, les deux outils ne sont pas compatibles pour les transformations et les calculs par défaut, on leur recommande pour les transformations manuelles à part la génération de documents techniques et fonctionnels.

- **Développement avancé et mise en production**

Caractéristiques	PDI	SSIS
Présence d'une API	✓	✓
Intégration de fonctions externes	✓	✓
Mécanisme de reprise sur incident	x	x
Paramétrage des buffers/indexes/caches	✓	✓
Gestion du développement en équipe	✓	✓
Versioning	x	✓
Compilation des traitements	x	oui pour C#
Type de mise en production	ligne de commandes Windows ou Unix	Ligne de commandes Windows
Visualisation de l'histoire des mises en production	x	x

Tableau 5 : Développement avancé et mise en production

Il a été trouvé que les deux outils ne sont pas compatibles pour le mécanisme de rétablissement sur l'incident et pour la visualisation d'histoire dans la production, mais généralement ils sont utilisés pour les autres propriétés du développement avancé et du déploiement de mise de production.

- **Administration et Gestion de la sécurité**

Caractéristiques	PDI	SSIS
Console d'administration	✓	✓
Gestion automatisée des logs	✓	✓
Génération de log spécifique	x	✓
Interfaçage avec des outils de supervision	x	✓
Outil de planification des traitements intégré	x	✓
Utilisation des droits d'un annuaire	x	x
Type de sécurité	Sécurité du SGBD qui contient le référentiel	✓
Sécurité sur la création de scénario	✓	✓
Sécurité sur la majorité de scénario	✓	✓
Sécurité sur l'accés aux métadonnées	✓	✓
Sécurité sur la console d'administration	✓	✓
Sécurité sur le lancement manuel des tâches	✓	✓

Tableau 6:.Administration et Gestion de la sécurité

Nous notons que le PDI n'est pas compatible pour la génération de journal spécifique, l'interfaçage avec les Outils de Surveillance, la planification de traitement intégré et pour la sécurité du système de gestion de données qui ne contient pas le dépôt.

II.4.3 Comparatif des temps de traitements

- **Méthodologie de réalisation des tests**

Les performances des temps de traitements sont un critère important dans le choix d'un ETL.
Les résultats des tests qui sont donnés dans les paragraphes suivants correspondent à des cas simples et ne peuvent en aucun cas préjuger des performances réelles en environnement de production.
Seuls des tests poussés sur des traitements d'intégration réels peuvent permettre de qualifier définitivement l'ETL choisi.

Les versions utilisées des 2 ETL sont les suivantes :

• Pentaho Data Integration v3.0.1

• SQL Server 2012

Pour chaque test, les résultats présentés sont issu à partir des essais réalisés dans des conditions identiques.

- **Test n°1**

1. Extraction des données d'un fichier Excel
2. Chargement des données dans un autre fichier Excel
3. Le fichier d'entrée comporte 5 champs typés :
4. COD_IND [NUMBER] (Code de l'étudiant)
5. COD_NNE_IND [NUMBER] (ID National de l'étudiant)
6. DATE_NAI_IND [DATE] (Date Naissance de l'étudiant)
7. LIB_NOM_PAT_IND [String] (Nom Familial de l'étudiant)
8. LIB_PR_IND [String] (Prénom de l'étudiant)

- **Modélisation dans Pentaho Data Integration (PDI)**

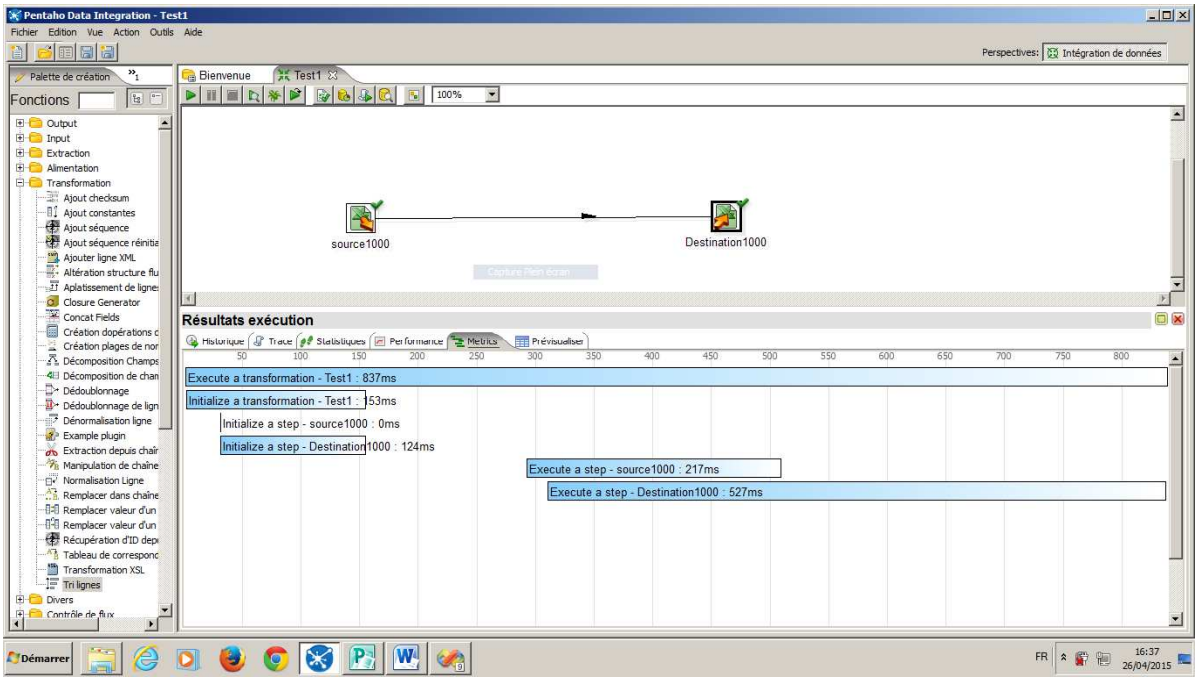


Figure 8: Extraction de 1000 lignes avec PDI

- **Modélisation dans SQL Server Intégration Services**

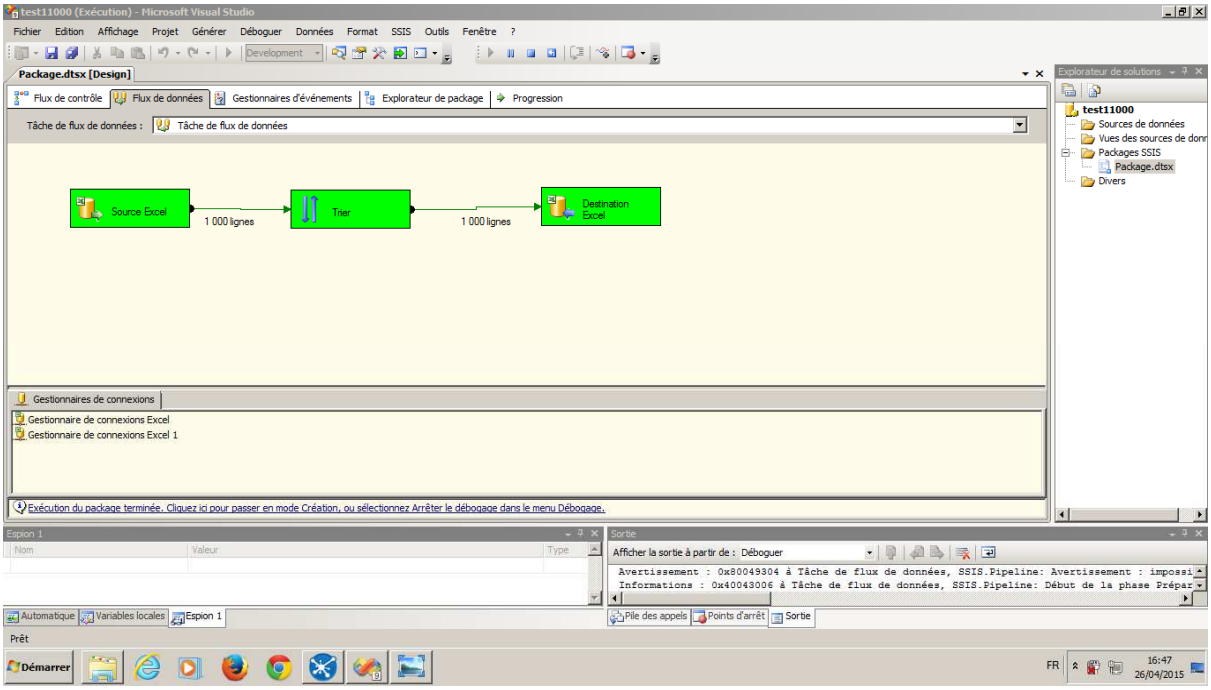


Figure 9 : Extraction de 1000 lignes avec SSIS

Nous avons effectué le même travail pour 5000 lignes et 10000 lignes

- **Temps de traitement pour les deux outils**

Nombres de lignes	PDI	SSIS
1000	837	655
5000	1384	1014
10000	3009	1513

Tableau 7: Temps de traitement pour SSIS et PDI

- **Résultat des tests**

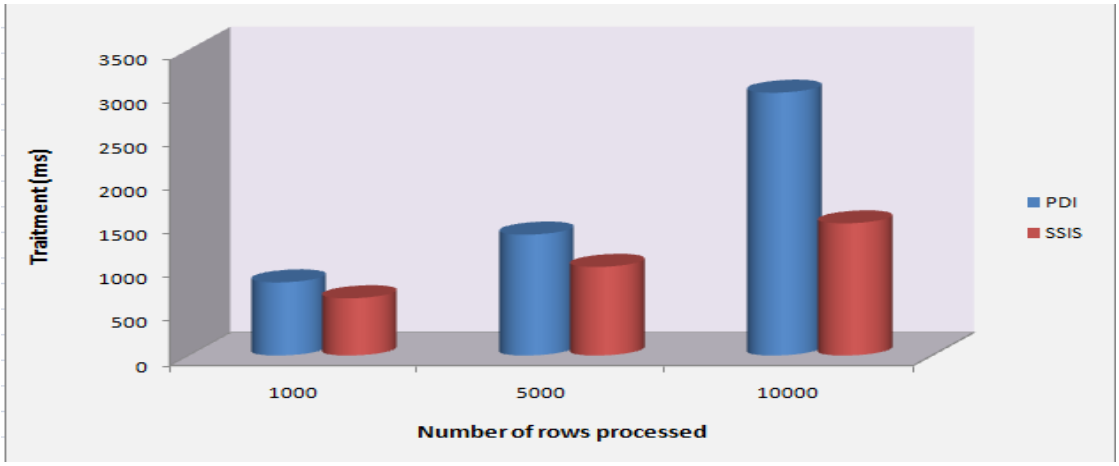


Figure 10: Comparaison des résultats obtenus pour les deux outils

On remarque que pour les données volumineuses le SSIS nécessite presque la moitié du temps que le PDI a besoin.

- **Test n°2**

1. Extraction des données d'un fichier Excel
2. Chargement des données dans un fichier XML
3. Le fichier d'entrée comporte 5 champs typés :
4. COD_IND [NUMBER] (Code de l'étudiant)
5. COD_NNE_IND [NUMBER] (ID National de l'étudiant)
6. DATE_NAI_IND [DATE] (Date Naissance de l'étudiant)
7. LIB_NOM_PAT_IND [String] (Nom Familialde l'étudiant)
8. LIB_PR_IND [String] (Prénom de l'étudiant)

- **Modélisation dans Pentaho Data Integration (PDI)**

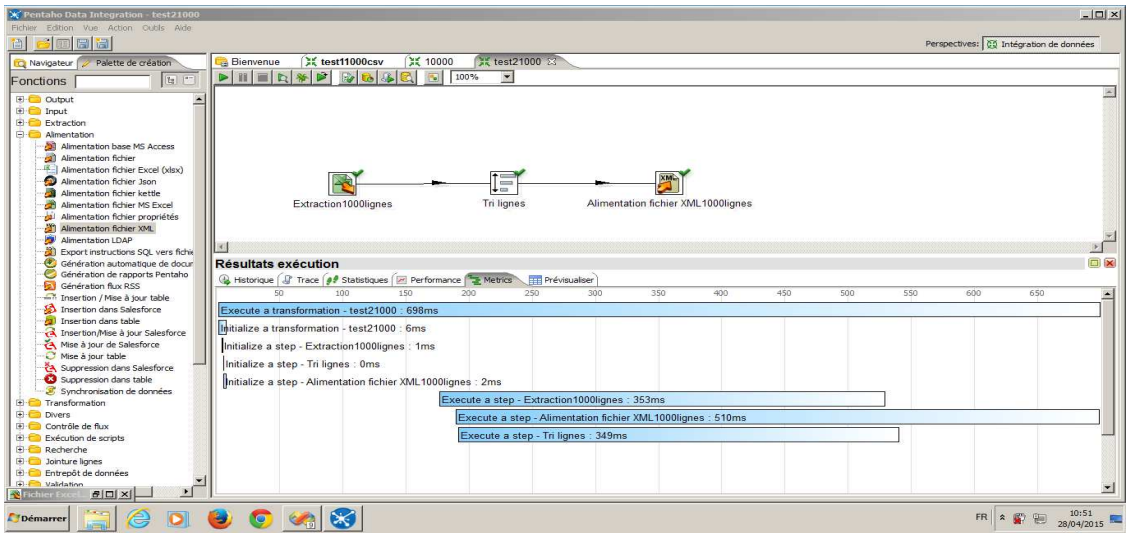


Figure 11: Extraction de 1000 lignes avec PDI

Nous avons effectué le même travail pour 5000 lignes et 10000 lignes

- **Modélisation dans SQL Server Intégration Services**

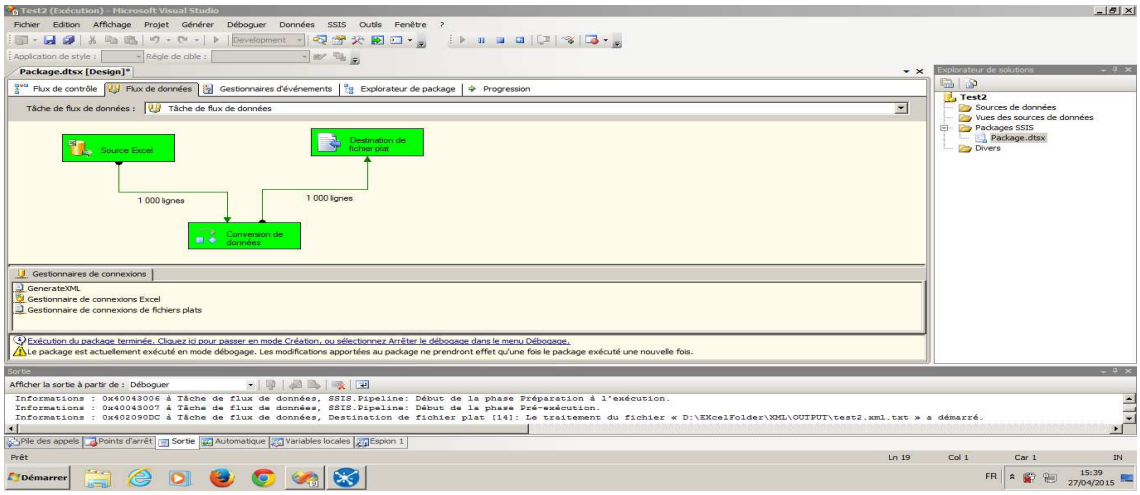


Figure 12: Extraction de 1000 lignes avec SSIS

- **Temps de traitement pour les deux outils**

Nombres de lignes	PDI	SSIS
1000	353	374
5000	2042	546
10000	3561	811

Tableau 8: Temps de traitement pour SSIS et PDI

- **Résultat des tests**

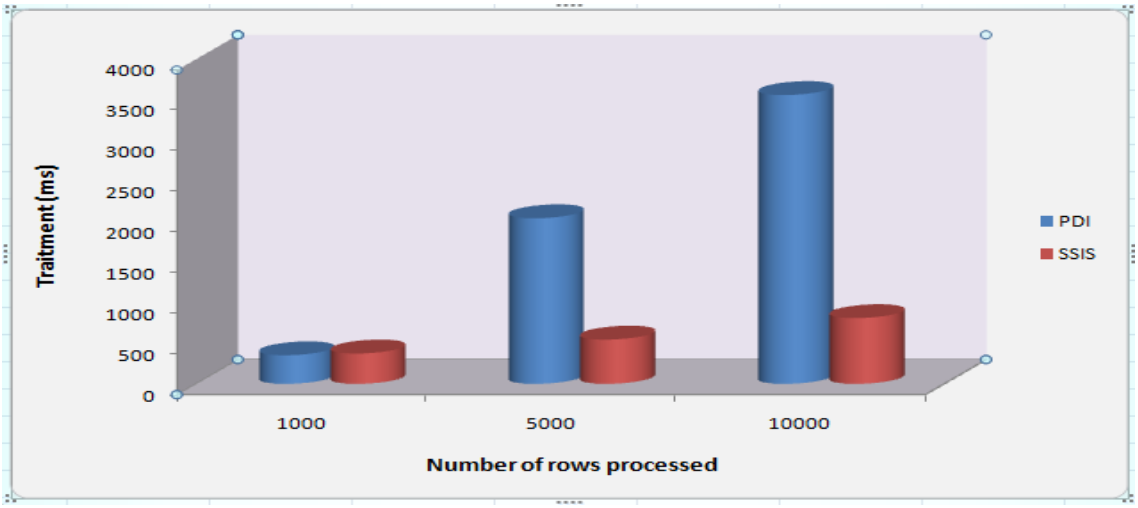


Figure 13: Comparaison des résultats obtenus pour les deux outils

Pour l’extraction des 10000 lignes le SSIS est plus robuste que le PDI, ce dernier a besoin plus de temps pour la phase d’extraction des données surtout pour le chargement des fichiers plats.

II.5 Conclusion

La performance du traitement du temps est un critère important dans le choix d'un ETL, mais à partir de ces résultats, nous ne pouvons pas préjuger de la performance réelle dans un environnement de production, puisque le temps d'exécution varie suivant la typologie des traitements.

A l'issue de l'ensemble de notre étude comparative, voici ce que nous pouvons conclure :

SSIS et Pentaho Data Integration sont des ETL ayant leurs propres spécificités, et sont donc dece fait plus complémentaires que « concurrents »

Ce sont des alternatives réelles à des ETL propriétaires comme Informatica Power Center, Oracle Warehouse ; Builder, Cognos DecisionStream...

les conclusions suivantes peuvent être tirées :

- ETL avec SSIS est une alternative séduisante et devrait être considérée sérieusement lors du démarrage d’un projet BI
- Reporting avec SSRS est avancé concernant la réalisation rapide de rapports et la possibilité de créer des requêtes type tableaux croisés.

Pour conclure, la maturité de l’outil SSIS dans le domaine de la Business Intelligence permet de le considérer comme une réelle alternative pour la réalisation des projets décisionnels.

Dans le chapitre suivant on va travailler avec la technologie SSIS du Microsoft afin de suivre les démarches d’analyse et de conception d’un système décisionnel pour l’université marocaine(SDUM)

III . Démarche d’analyse et de conception d’un Système décisionnel pour l’université marocaine (SDUM)

III.1 Introduction

L’enseignement supérieur constitue un des piliers de développement du Maroc. Il est porteur des projets d’innovation qui produisent des compétences et des qualifications des jeunes et de la production du savoir. La performance globale de l’enseignement supérieur nécessite l’utilisation des outils de prédiction et d’aide à la décision. Par conséquent toutes les réformes de l’université préconisent l’utilisation des technologies de l’information pour le développement de ses briques métiers, ressources humaines, patrimoine, finance, la qualité de sa gouvernance ainsi que l’efficacité de ses modèles pédagogique. En l’occurrence les systèmes d’aide à la décision. La branche informatique dédiée à ces systèmes est notée l’informatique décisionnelle ID ou encore la Business Intelligence. [66]

III.2 Etude de l’existant

L’Apogée est une application pour l'organisation et la gestion des enseignants et des étudiants, apporte des réponses précises en matière de clarification de l'offre de formation, d'amélioration de l'accueil des étudiants, de gestion de la scolarité et de pilotage de l'établissement.

Le logiciel Apogée implanté dans toutes les universités marocaines. APOGEE est basée sur une base de données oracle, ces données sont une suite des tableaux d’Excel sous forme des données brutes afin de choisir l’outil convenable pour faire une extraction des données qui aide à la réalisation et la construction d’un entrepôt de données.

III.3 Structure d’APOGEE

Le progiciel APOGEE est conçu pour la gestion des inscriptions administratives, inscriptions pédagogiques (rattachement à un diplôme, une année, à un ensemble de modules), des examens (relevé de notes etc.), pour l’aide aux jurys de semestre et d'année (aide à la délibération) et à la production des diplômes (procès-verbaux, annexe au diplôme, etc.).APOGEE est structuré en « modules » :

- Inscription administrative. – création automatique des formulaires d’inscription remplis par les étudiants lors de leurs réinscriptions.

- Dossier "Étudiant : – Données administratives des étudiants : cursus, inscription pédagogique, adresse...
- Stage : gestion des conventions de stage
- Contrôle des connaissances : saisie des barèmes, coefficients et règles de calcul de notes et de résultat.
- Résultat. — Saisies des notes, calcul automatique des notes, classement des étudiants, calcul des moyennes (compte tenu des "points de jury" attribués en fin de semestre. Impression de documents : procès-verbaux de notes, relevés de notes, diplômes...
- Inscription pédagogique : Inscription des étudiants dans les cours, soit de manière individuelle (étudiant par étudiant), soit "en masse".
- Référentiel : stockage des données techniques. Par exemple, les données "Utilisateur" (droits d'accès des utilisateurs) ou les données "Environnement" (paramétrage des données : liste des composantes de l'université, décrets des diplômes etc.)

III.4 Critique de l'existant

APOGEE est une application bien structurée mais elle possède des limites qui se résument dans les points suivants :

- **Données non consolidées :** les données sont éparpillées sur différents supports ce qui rend les choses difficile de les traiter.
- **Insécurité :**n'importe quelle personne ayant accès au poste de travail ou avec de minime connaissance du monde l'informatique peut accéder aux données.
- **Complexité du traitement :** La nature et la diversité des données rendent leurs traitements complexes.
- **Non précision des indicateurs et des axes d'analyse des rapports générés** *:* avec un tel processus de collecte d'informations les indicateurs définis ne pourront fournir des rapports d'une grande précision.
- **Un système de Reportingmoins détaillé :**les données sont traitées manuellement par le responsable en personne. Cette solution est opérationnelle mais demeure moins performante qu'une base de données décisionnelle.

III.5 Objectif du SDUM

L'objectif de ce projet consiste à développer les tableaux de bords en se basant sur des maquettes de présentation qui illustrent l'évolution d'un thème de gestion et sa projection sur un axe donné tout en respectant une mesure définie sachant que la projection peut se faire sur un axe comme elle peut se faire sur plusieurs selon les exigences de l'utilisateur ou du décideur ,et elle met à disposition de l'utilisateur en général et du décideur d'une manière directe l'ensemble des éléments nécessaires à l'élaboration d'une synthèse, qui sert à mettre à jour les connaissances de tout utilisateur et de visualiser la situation de l'université

III.6 Modèle conceptuel de données

Le modèle conceptuel des données (**MCD**) [67] a pour but d'écrire de façon formelle les données qui seront utilisées par le système d'information. Il s'agit donc d'une représentation des données, facilement compréhensible, permettant de décrire le système d'information à l'aide d'entités. Pour générer le MCD on a utilisé le logiciel **« POWER AMC »**, c'est un logiciel de conception de la base de donnée, la figure suivante vous montre le MCD effectué

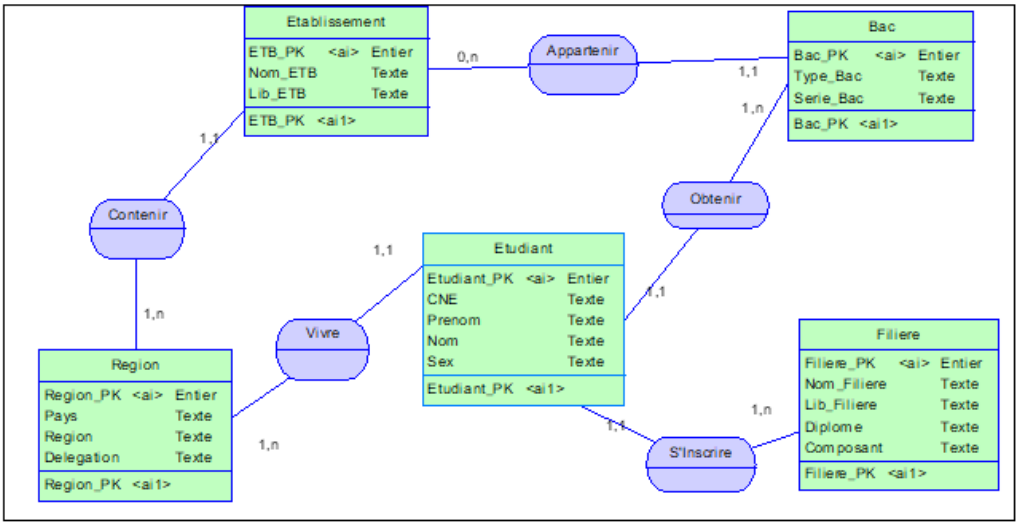


Figure 14: Modèle Conceptuel des Données

III.7 Modèle Logique de Données (MLD)

Un Modèle Logique des Données, est un modèle construit à l'aide d'un MCD entenant compte du type du SGBD utilisé, En tenant compte des paramètres suivants, La prise en compte du SGBD Machine, La configuration des données, le rapprochement des données vers la machine. Une base de données définie entièrement par un seul schéma logique :

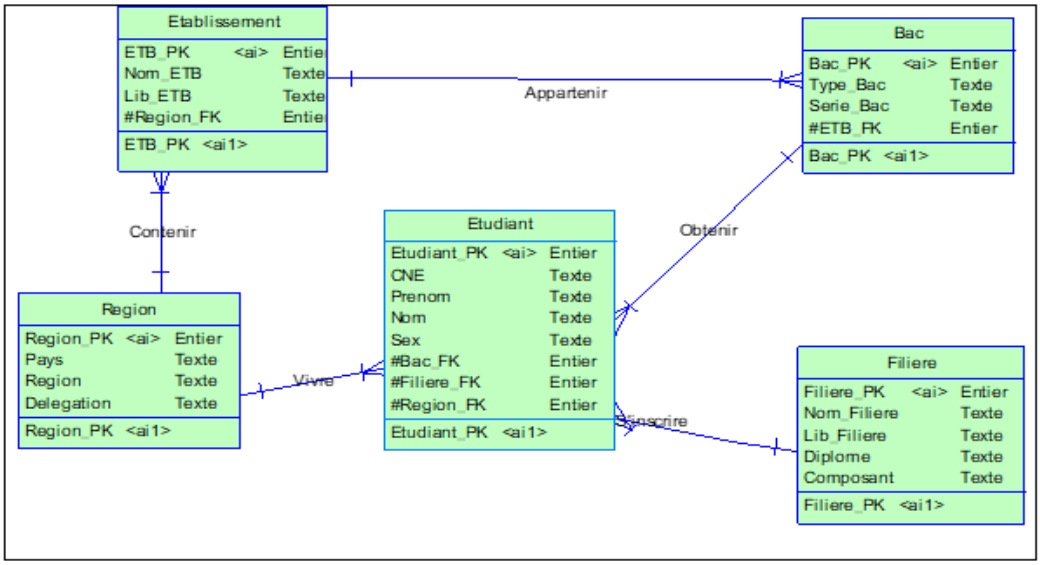


Figure 15: Modèle Logique de données

III.8 Présentation de notre environnement décisionnel

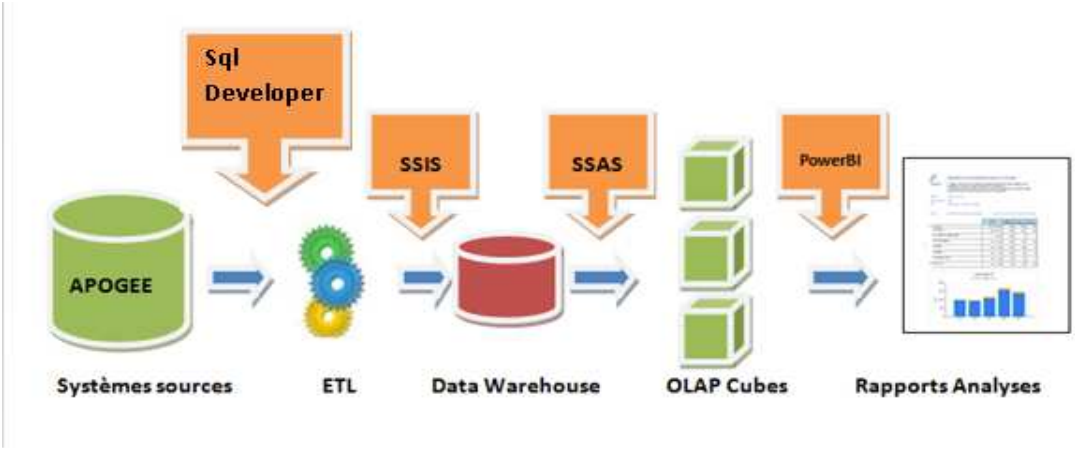


Figure 16: Schéma général de notre architecture Business Intelligence

Pour analyser les données, il est indispensable de les rassembler en un seul endroit. Or, les données d'une entreprise se trouvent dans de multiples fichiers et, souvent, elles ne sont pas cohérentes. Afin de rassembler et de nettoyer les données, nous utilisons un logiciel de type ETL.

III.8.1 L'outil choisi d'intégration des données : SQL SERVER

La première étape consiste à extraire les tables de la base de données APOGEE à l'aide de l'outil sqldeveloper.

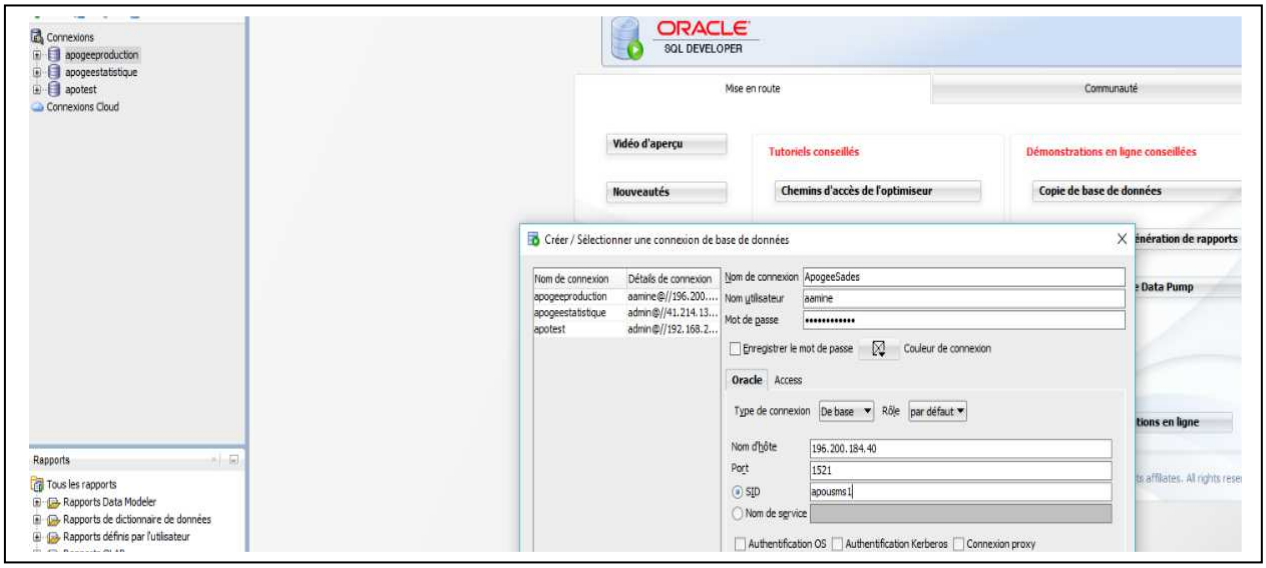


Figure 17: Connexion au serveur APOGEE

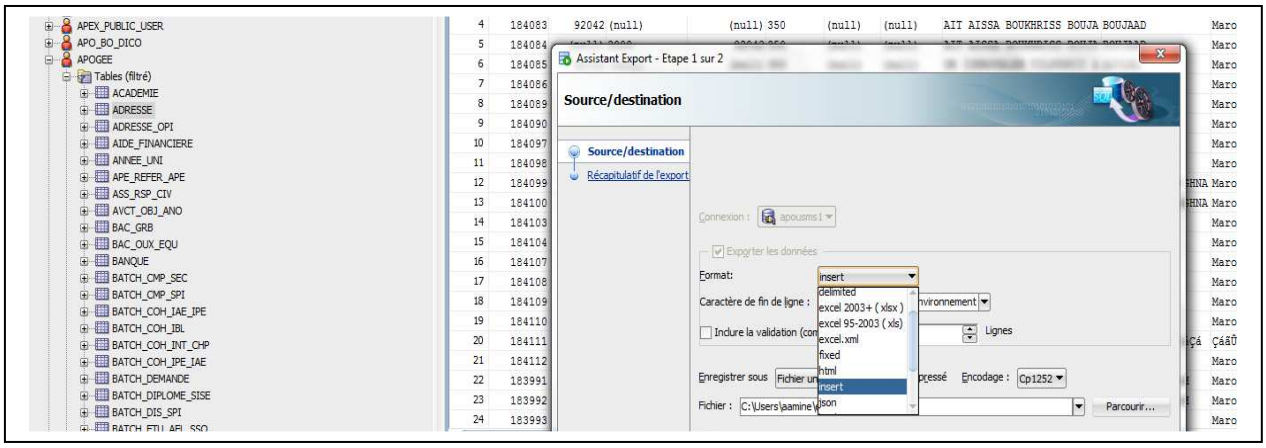


Figure 18:Extraction des tables de la base de données APOGEE

III.8.2 SQL Server

Microsoft SQL Server est un système de gestion de base de données(SGBD) en langage SQL incorporant entre autres un SGBDR (SGBD relationnel») [68] développé et commercialisé par la société Microsoft. Il fonctionne sous les OS Windows et Linux (depuis mars 2016).

La souplesse de la gestion des schémas SQL est telle qu'il est possible de transférer un objet d'un schéma à l'autre par le simple biais d'une commande ALTER SCHEMA.

Les propriétaires sont distincts des schémas et il est possible de transférer la propriété d'une base, d'un schéma ou d'un objet d'un utilisateur SQL à l'autre, par le biais de la commande ALTER AUTHORIZATION.

III.8.3 Modélisation de l'entrepôt de données

III.8.3.1 Règles de gestion de quelques indicateurs demandés

Le système à développer doit permettre aux décideurs d'avoir une idée claire sur :

Indicateurs	Règle de gestion
Effectif des étudiants dans les semestres et diplômes/par filière/par année	Nombre globale des étudiants inscrits par année/étape/ établissement
Nombre de filières nouvellement crées/par établissement/ par année	Nombre de filière dans les établissements
Effectif des Lauréats par établissements/par année	(Nombre d'inscrits/Nombre de diplôme)*100
Taux d'abondant par établissement/par année	Nombre d'inscrits – ((Nombre d'inscrits/Nombre de diplôme)*100)

Tableau 9: Représentation des indicateurs

III.8.3.2 Liaisons entre les tables de la source

Il s'agit de tables suivantes :

Table Individu
Table Inscription_Administrative_Etape
Table Composante
Table Année
Table Diplôme
Table Résultat

III.8.3.3 Schéma correspondantes

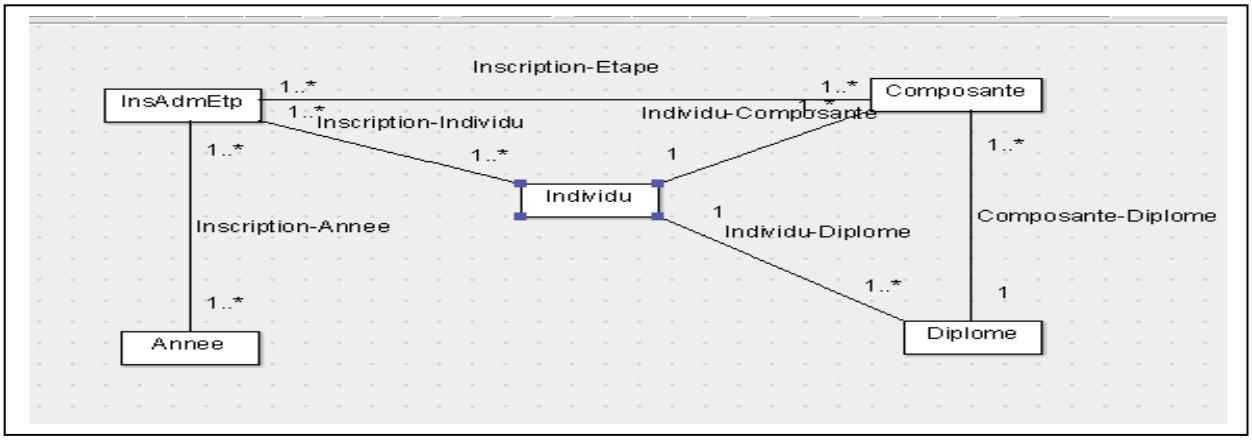


Figure 19: Schéma correspondante pour Effectif des étudiants

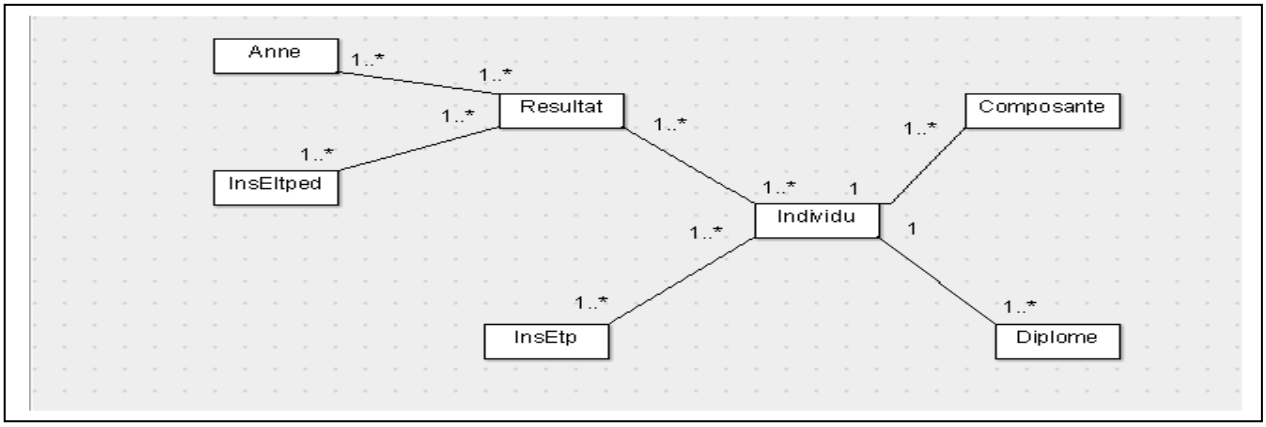


Figure 20: Schéma correspondante pour l'indicateur Taux des lauréats

III.8.3.4 Importation des données

Après l'extraction des données da l'application Apogée, on va les importer à l'aide de l'outil Sql Server.

Dans l'explorateur du SqlServer d'objets on va créer trois bases de données selon le schéma suivant :

Création de la base data warehouse(DWH)

Source de données ⟶ ODS ⟶ DWH

Source de données ⟶ Vue de données ⟶ Cube

ODS : Un operationnel data store est une base de données conçue pour centraliser les données issues de sources hétérogènes afin de faciliter les opérations d'analyse et de reporting

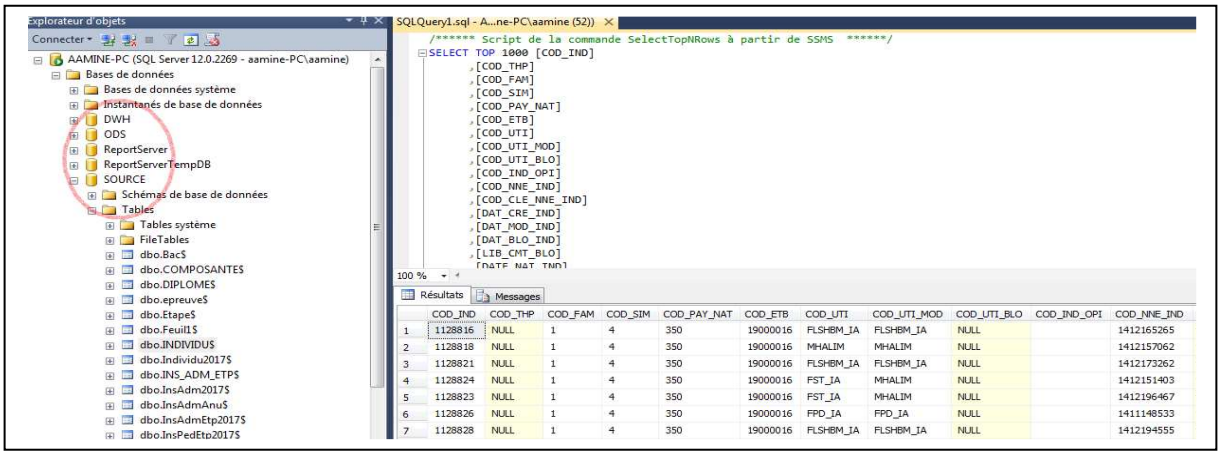


Figure 21: Création de trois bases

III.8.3.5 Différentes étapes de l'importation des données

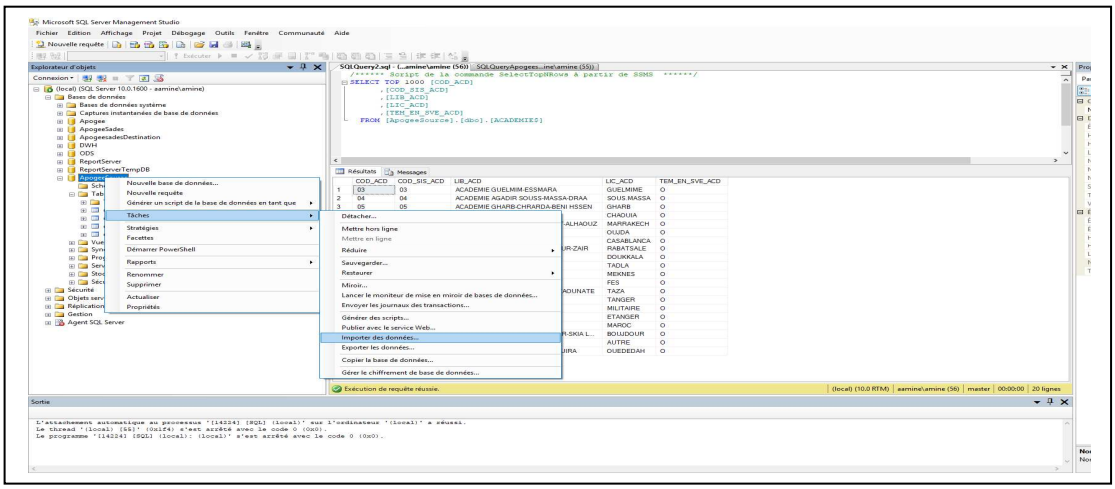


Figure 22: Importation des données source

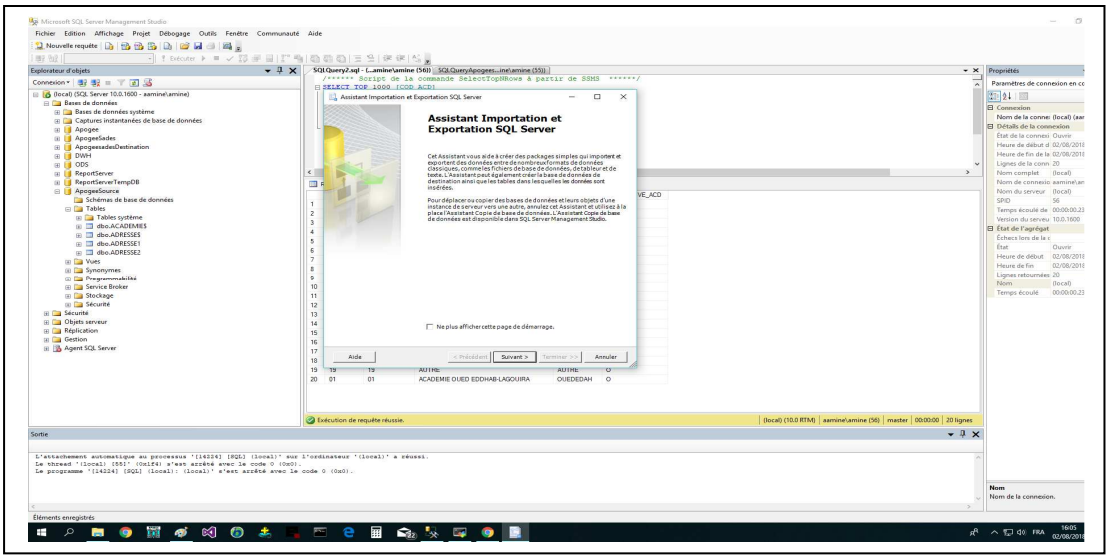


Figure 23: Assistant de l'importation

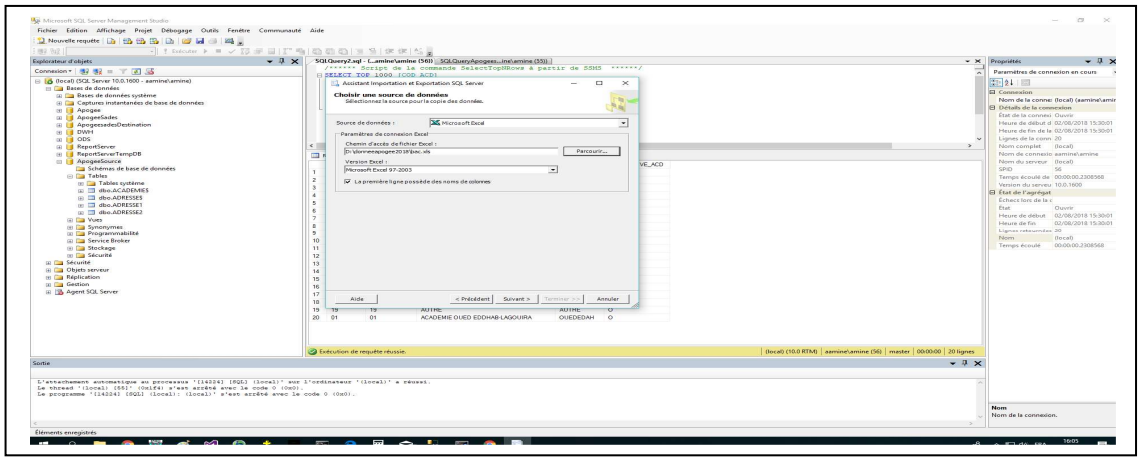


Figure 24: Choix de la source de données

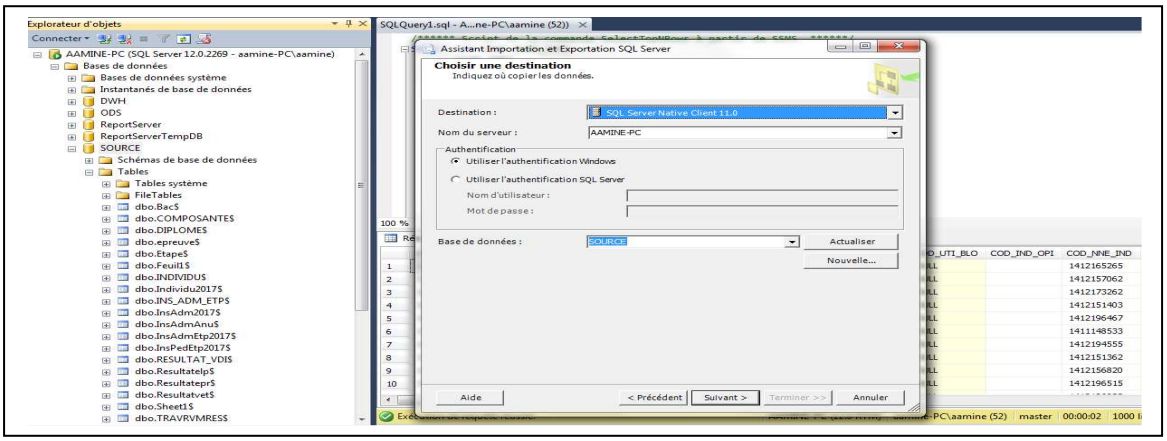


Figure 25: Choix de la destination

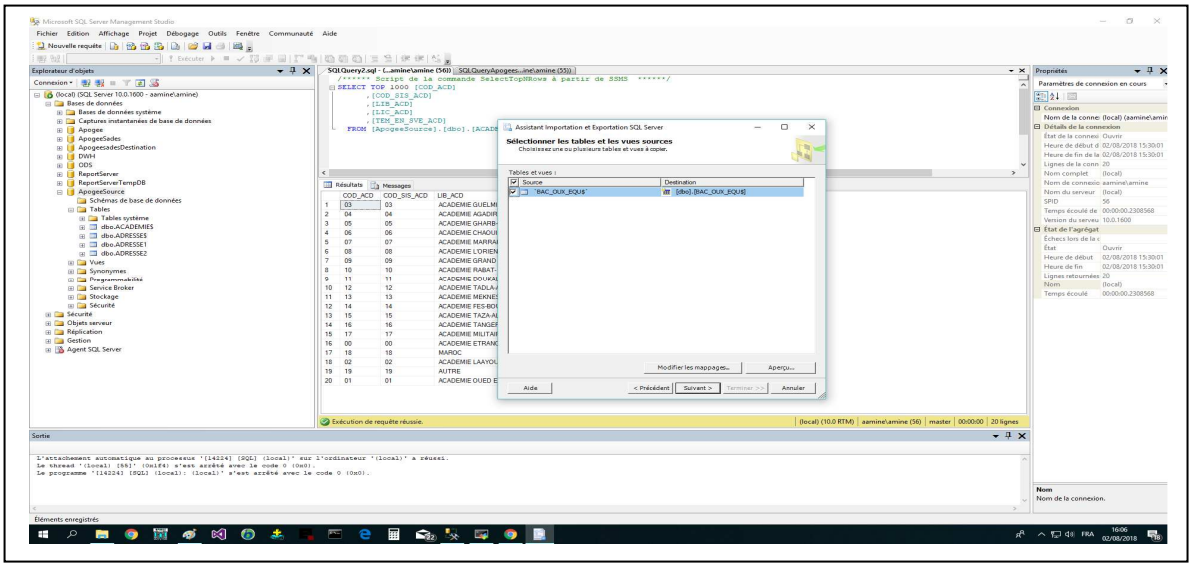


Figure 26:Sélection des tables et des vues sources

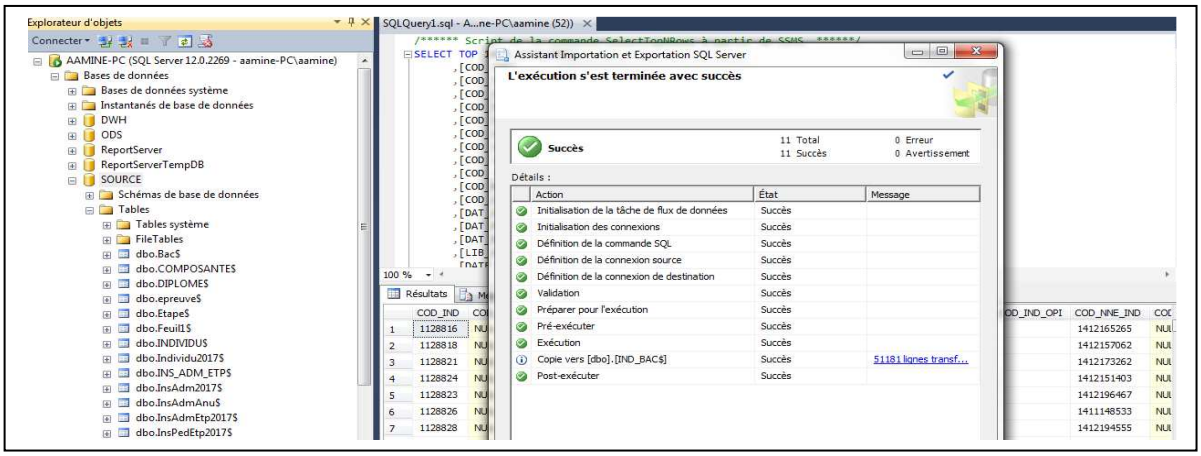


Figure 27: exécution finale

III.8.3.6 Intégration des données source

Dans cette partie nous allons vous présenter des éditions écran sur quelques scénarios du projet

Les données provenant de la technologie SQL Server 2014 seront intégrées dans l'outil Microsoft Visual Studio en vue de les charger dans des TABLES ODS via les opérations ELT appliquées sur les interfaces.

Apparences sur les étapes suivies :

III.9 Partie SSIS

Pour la connexion avec le Sql Server on est obligé d'établir trois connexions liées aux

différentes bases qu'on a construites dans le Sql Server.

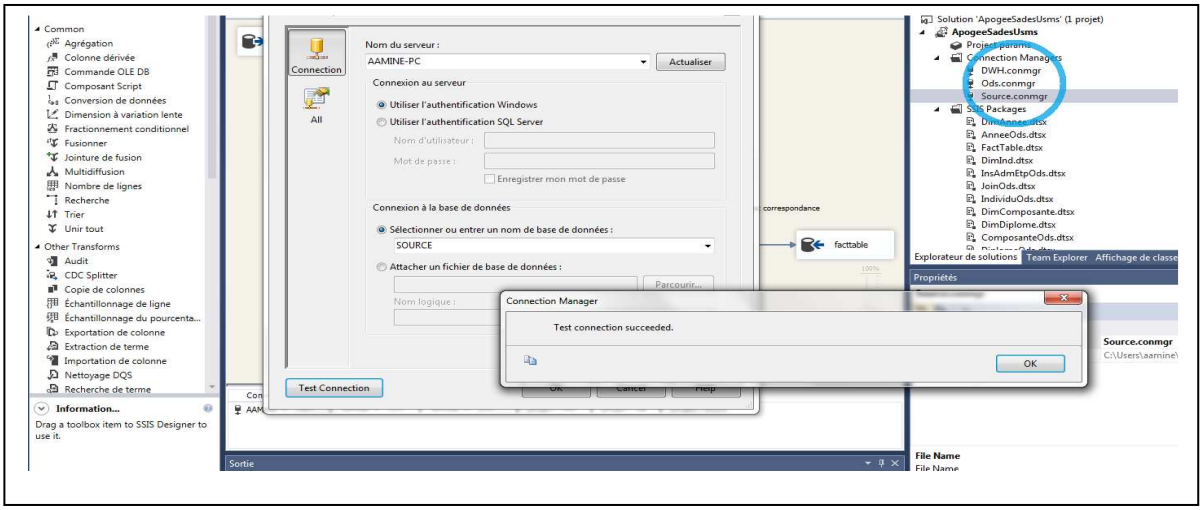


Figure 28: Connexion Manager

La création des TABLES ODS s'effectue au niveau du Visual Studio et à l'aide de la connexion réalisée les tables seront chargées au niveau du SqlSever.

Le chargement de ces tables s'effectue suivant les étapes suivantes :

Dans l'interface Flux de Contrôle on a glissé deux taches une de flux de données qui sert à charger la TABLE ODS correspondant et l'autre de flux d'exécution de requête SQL qui sert à vider cette table pendants le deuxième chargement selon la requête

TRUNCATE TABLE [ODS]. [DBO]. [NOM TABLE].

On va faire l'exemple de la table Composante, et c'est la même chose pour les autres tables (Diplôme, Individu, Année, Inscription_Administrative_Etape)

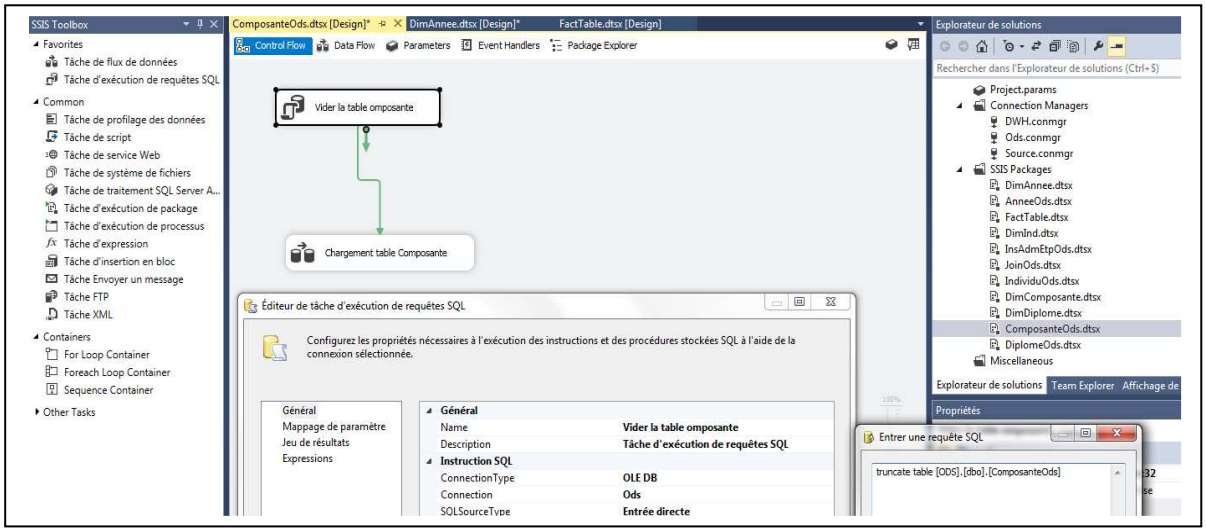


Figure 29: Création de la table ODS_COMPOSANTE

En cliquant deux fois sur la tâche de flux de données on obtient l'interface Flux de Données qui contient la boite d'outil correspondante.

A partir de cette boite on va glisser deux composantes une Source ADO NET et une Destination ADO NET

En cliquant Sur cette source on obtient l'interface dans lequel on va interroger les données de la source dans le Sql Server via la connexion Source.

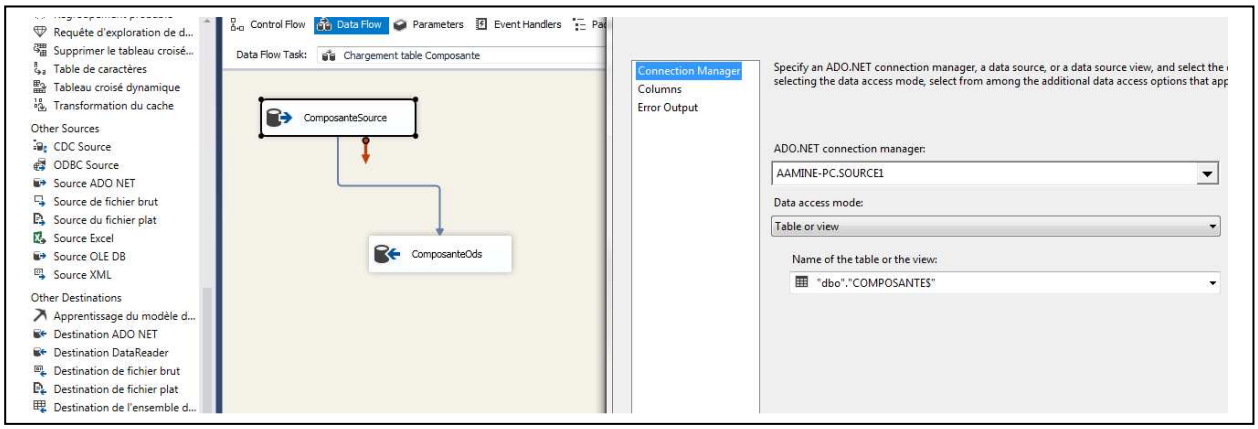


Figure 30: Source ADO NET

Après la liaison de cette source avec la destination, une interface sera l'objet clé de tout chargement des données d'une source vers une cible, et dans laquelle s'effectuant les opérations ELT ;

Voyons l'illustration ci-dessous de l'ensemble des interfaces utilisées du projet :

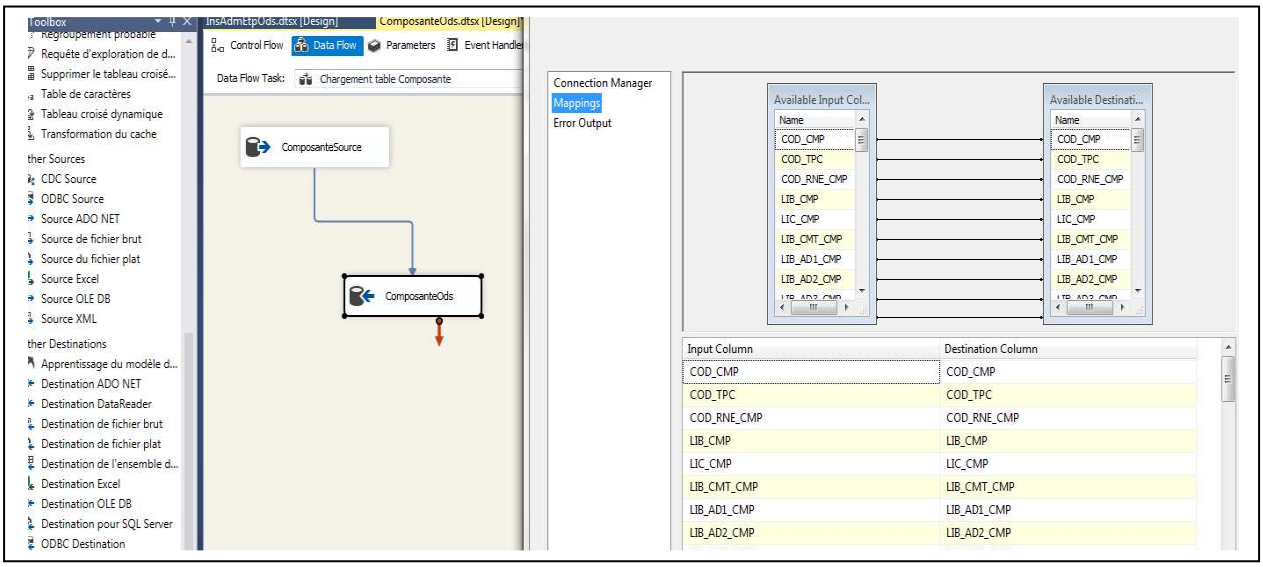


Figure 31: Correspondance via des interfaces :Mapping

Après l'exécution de cette requête le chargement s'effectue dans les deux outils.

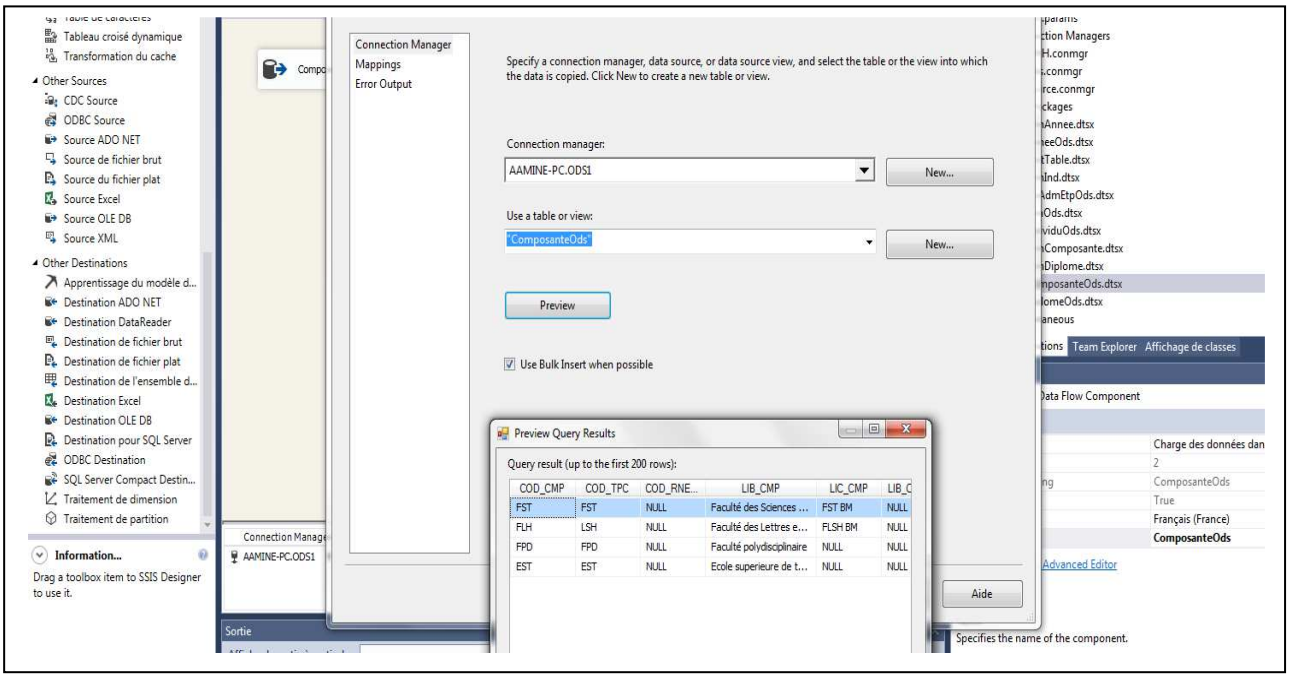


Figure 32: Chargement dans Microsoft Visual Studio

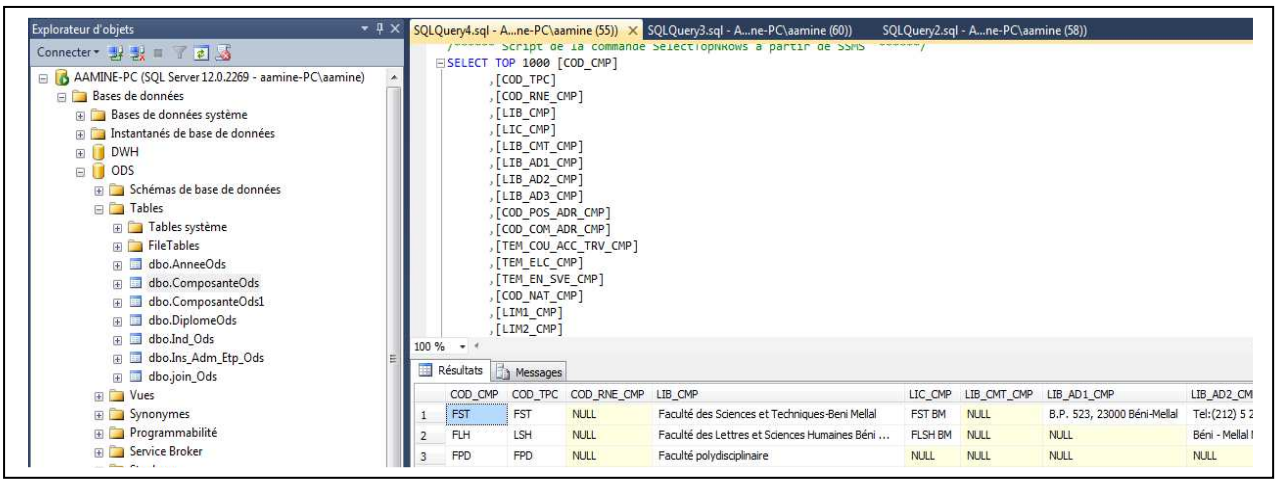


Figure 34: Chargement dans Sql Server

III.9.1 Réalisation du schéma correspondante

Pour réaliser l'indicateur<< nombre d'inscrits par établissement/filière>> on est obligé de faire une jointure entre les tables Année, Composante, Diplôme, Individu avec la table Ins_Adm_Etp dans la partie ODS

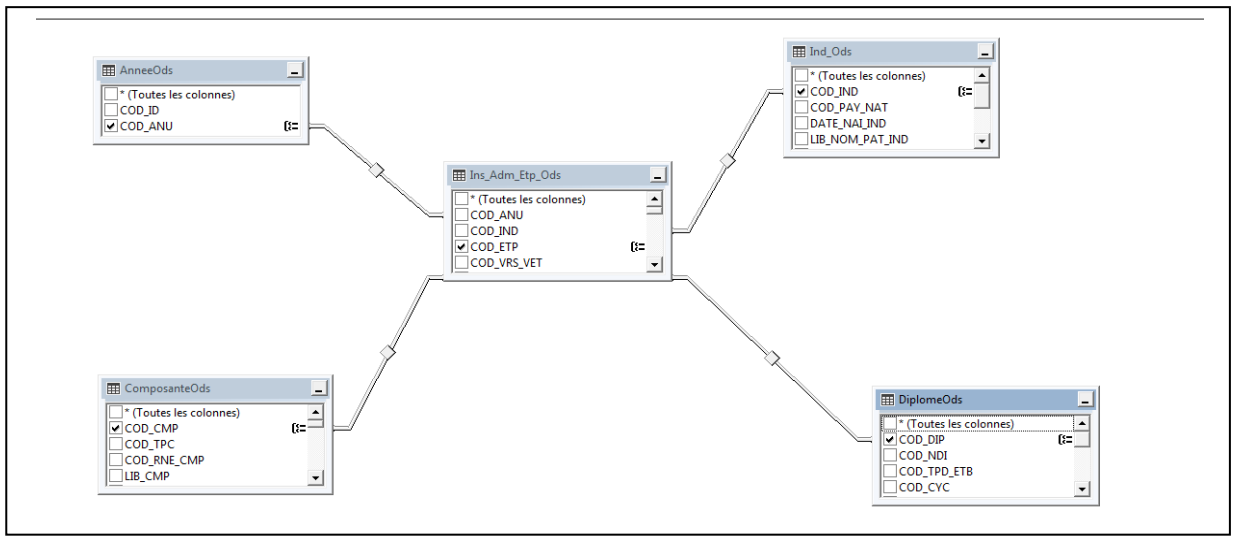


Figure 35: Jointure de tables correspondantes

La jointure se fait suivant la requête qui va définir le nombre des étudiants inscrits

Dans les établissements par année/Composante/Diplôme et d'autres mesures.

```
SELECT     COUNT(Ind_Ods.COD_IND) AS nbrglobal, Ind_Ods.COD_IND, Ins_Adm_Etp_Ods.COD_ETP, Ins_Adm_Etp_Ods.ETA_IAE, ComposanteOds.COD_CMP,
             DiplomeOds.COD_DIP, Ins_Adm_Etp_Ods.NBR_INS_CYC, Ins_Adm_Etp_Ods.NBR_INS_ETP, Ins_Adm_Etp_Ods.NBR_INS_DIP, AnneeOds.COD_ANU
FROM       ComposanteOds INNER JOIN
             Ins_Adm_Etp_Ods INNER JOIN
             Ind_Ods ON Ins_Adm_Etp_Ods.COD_IND = Ind_Ods.COD_IND INNER JOIN
             DiplomeOds ON Ins_Adm_Etp_Ods.COD_DIP = DiplomeOds.COD_DIP ON ComposanteOds.COD_CMP = Ins_Adm_Etp_Ods.COD_CMP INNER JOIN
             AnneeOds ON Ins_Adm_Etp_Ods.COD_ANU = AnneeOds.COD_ANU
GROUP BY Ind_Ods.COD_IND, Ins_Adm_Etp_Ods.COD_ETP, Ins_Adm_Etp_Ods.ETA_IAE, ComposanteOds.COD_CMP, DiplomeOds.COD_DIP, Ins_Adm_Etp_Ods.NBR_INS_CYC,
             Ins_Adm_Etp_Ods.NBR_INS_ETP, Ins_Adm_Etp_Ods.NBR_INS_DIP, AnneeOds.COD_ANU
```

Figure 36: Requête correspondante

III.9.2 Choix des dimensions

Dimension ou axe est une table qui contient les axes d'analyse selon lesquels on veut étudier des données observables (les faits) qui, soumises à une analyse multidimensionnelle, fournissent aux utilisateurs des renseignements nécessaires

Une dimension correspond à un axe d'analyse de la problématique du projet, elle est nécessairement un facteur dont dépend l'aspect que l'objectif du projet essaye de mesurer.
A chaque dimension est attribuée une table. Il existe autant de tables dimensions que de dimensions. Chaque table de dimension contient les attributs de la dimension en question plus une clé primaire indépendante de ces attributs.

III.9.3 Liste des tables de dimensions

Le modèle retenu à une perspective globale et multidimensionnelle, il est représenté par un schéma en étoile des tables de dimensions et des tables de faits réduite et particularisée.

Ces dernières sont organisées en dimensions représentant les axes de recherche suivants

Dim_Individu

Dim_Composante

Dim_Diplôme

Dim_Année

L'analyse technique a permis d'identifier les tables de dimensions suivantes :

Dimension	Attribut	Type	Description
Dim_Individu	COD_IND	int	Clé primaire
	COD_PAY_NAT	nvarchar(255)	Code du Pays
	DATE_NAI_IND	nvarchar(255)	Date de Naissance
	LIB_NOM_PAT_IND	nvarchar(255)	Nom
	LIB_PR1_IND	nvarchar(255)	Prénom
	COD_ETU	float	Code APOGEE
	COD_SEX_ETU	nvarchar(255)	Code sexe de l'étudiant
	LIB_VIL_NAI_ETU	nvarchar(255)	Ville de Naissance
	COD_DEP_PAY_NAI	nvarchar(255)	Code Pays de Naissance
Dim_Diplôme	COD_DIP	NNVARCHAR(255)	Code Diplôme(Clé primaire)
	COD_NDI	NVARCHAR(255)	Code Nature Diplôme
	COD_TPD_ETB	NVARCHAR(255)	Code Type Diplôme Etablissement
	COD_CYC	NVARCHAR(255)	Code SISE Cycle
	COD_NIM	NVARCHAR(255)	Code Niveau Interministériel
	COD_ETB	NVARCHAR(255)	(COPIED)Code National de l''Etablissement Principal qui de
	COD_SDS	NVARCHAR(255)	Code Secteur Disciplinaire SISE
	LIB_DIP	NVARCHAR(255)	Libelle Long Diplôme
	LIC_DIP	NVARCHAR(255)	Libelle Court Diplôme
	NBR_MAX_INSC_DEUG	NVARCHAR(255)	Nombre Maximum d''Inscriptions en DEUG Autorisé
	TEM_COU_ACC_TRV_DIP	NVARCHAR(255)	Témoin Couverture Accident Travail
	TEM_OUV_DRT_SSO_DIP	NVARCHAR(255)	Témoin Ouvre Droit Sécurité Sociale
	LIB_DIP_ARB	NVARCHAR(255)	Libelle Long Diplôme en arabe
	LIC_DIP_ARB	NVARCHAR(255)	Libelle Court Diplôme en arabe
	COD_PER		Code personne de coordinateur de la filière
Dim_Composante	COD_CMP	NNVARCHAR(255)	Code Composante (Clé primaire)
	COD_TPC	NVARCHAR(255)	Code Type de Composante
	LIB_CMP	NVARCHAR(255)	Libelle Long Composante
	LIC_CMP	NVARCHAR(255)	Libelle Court Composante
	COD_NAT_CMP	NNVARCHAR(255)	Code Nature de la Composante
	LIB_VIL_CMP	NVARCHAR(255)	Ville de la composante

Tableau 10: description des tables de dimension

La création des tables de dimensions se fait de la même manière que les tables ODS,

sauf qui' il y a une étape pour la finalisation de ces tables.

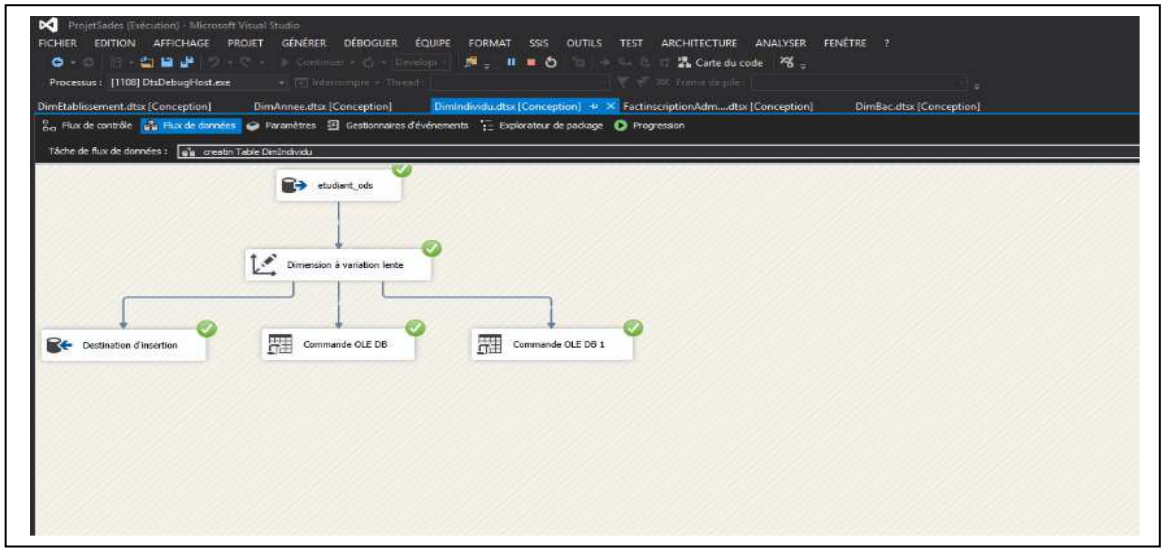
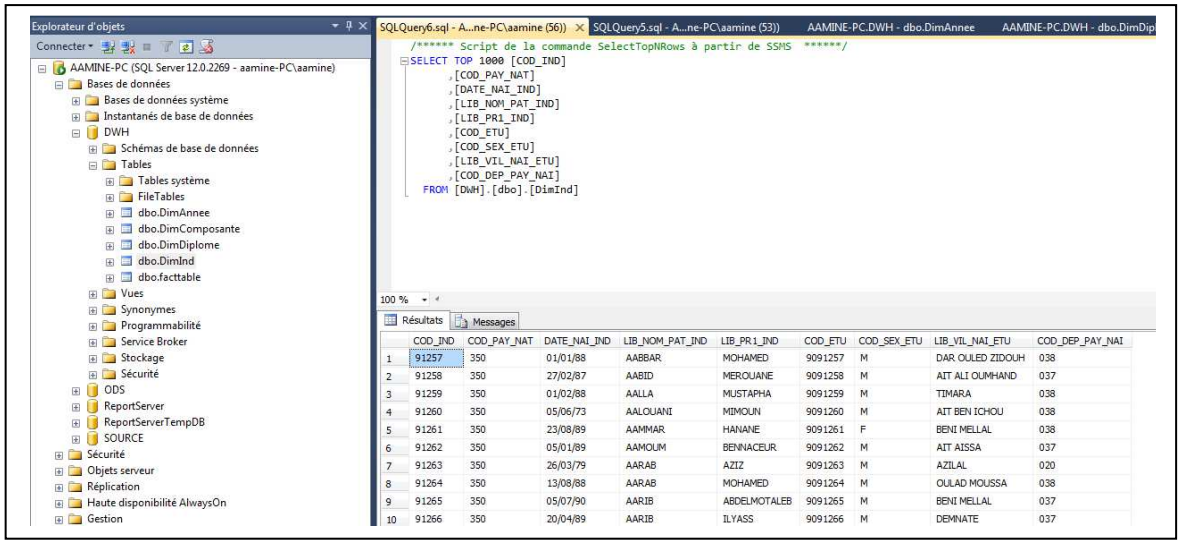


Figure 37: Etape finale pour la création de la table de Dimension



	COD_IND	COD_PAY_NAT	DATE_NAI_IND	LIB_NOM_PAT_IND	LIB_PR1_IND	COD_ETU	COD_SEX_ETU	LIB_VIL_NAI_ETU	COD_DEP_PAY_NAI
1	91257	350	01/01/88	AABBAR	MOHAMED	9091257	M	DAR OULED ZIDOUH	038
2	91258	350	27/02/87	AABID	MEROUANE	9091258	M	AIT ALI OUMHAND	037
3	91259	350	01/02/88	AALLA	MUSTAPHA	9091259	M	TIMARA	038
4	91260	350	05/06/73	AALOUANI	MIMOUN	9091260	M	AIT BEN ICHOU	038
5	91261	350	23/08/89	AAMMAR	HANANE	9091261	F	BENI MELLAL	038
6	91262	350	05/01/89	AAMOUM	BENNACEUR	9091262	M	AIT AISSA	037
7	91263	350	26/03/79	AARAB	AZIZ	9091263	M	AZILAL	020
8	91264	350	13/08/88	AARAB	MOHAMED	9091264	M	OULAD MOUSSA	038
9	91265	350	05/07/90	AARIB	ABDELMOTALEB	9091265	M	BENI MELLAL	037
10	91266	350	20/04/89	AARIB	ILYASS	9091266	M	DEMNATE	037

Figure 38: Chargement de la table de Dimension

III.8.3 La table de faits

Les tables de faits représentent des associations dont l'existence d'une occurrence dépend de l'existence des occurrences correspondantes dans les tables dimensionnelles, c'est-à-dire la table de fait contient l'ensemble des mesures correspondant aux informations de l'activité à analyser .Une table de faits contient les valeurs numériques de ce qu'on désire mesurer.

Une table de fait contient les clés associées aux dimensions. Il s'agit des clés étrangères dans la table de faits.

Une table de fait contient plus d'enregistrements qu'une table de dimension.

Les informations dans une table de fait sont caractérisées par :

- Elles sont numériques et sont utilisées pour faire des SUM, AVG...
- Les données doivent être additives ou semi additives

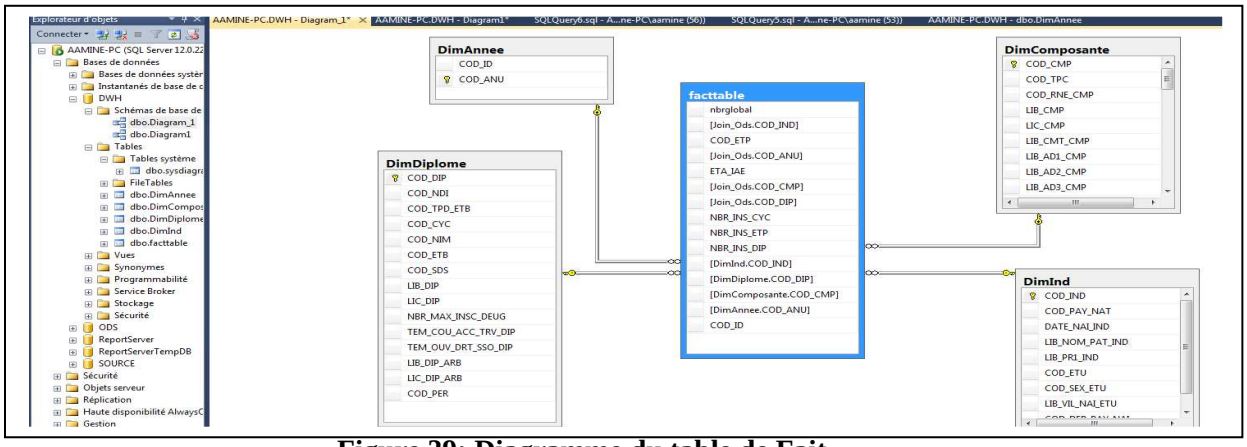


Figure 39: Diagramme du table de Fait

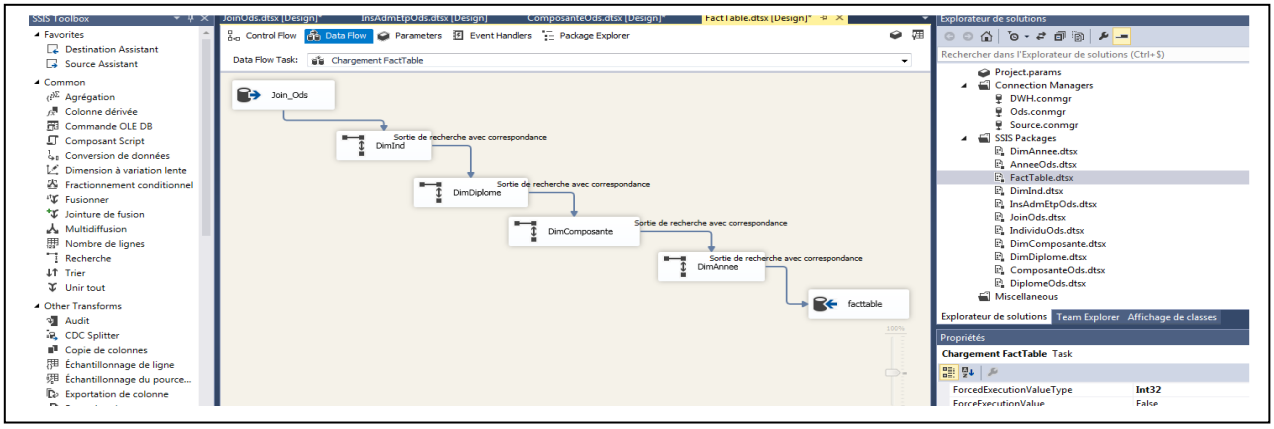


Figure 40: Conception du table de Fait

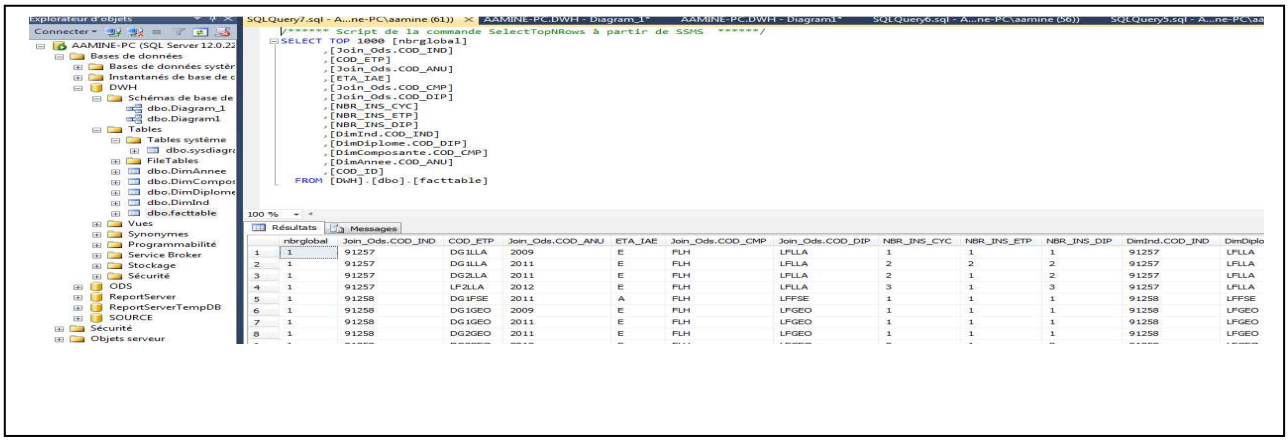


Figure 41: Chargement du table de Fait

III.10 Partie SSAS

III.10.1 Création de cubes dans SQL Server Analysis Services (SSAS)

La partie SSAS vient après la finalisation de la partie SSIS, et dans laquelle on crée le cube associé aux différents dimensions.

III.10.2 Création des sources de données

Dans un modèle multidimensionnel Analysis Services, un objet de source de données représente une connexion à la source de données depuis laquelle vous traitez (ou importez) des données. Un modèle multidimensionnel doit contenir au moins un objet de source de données

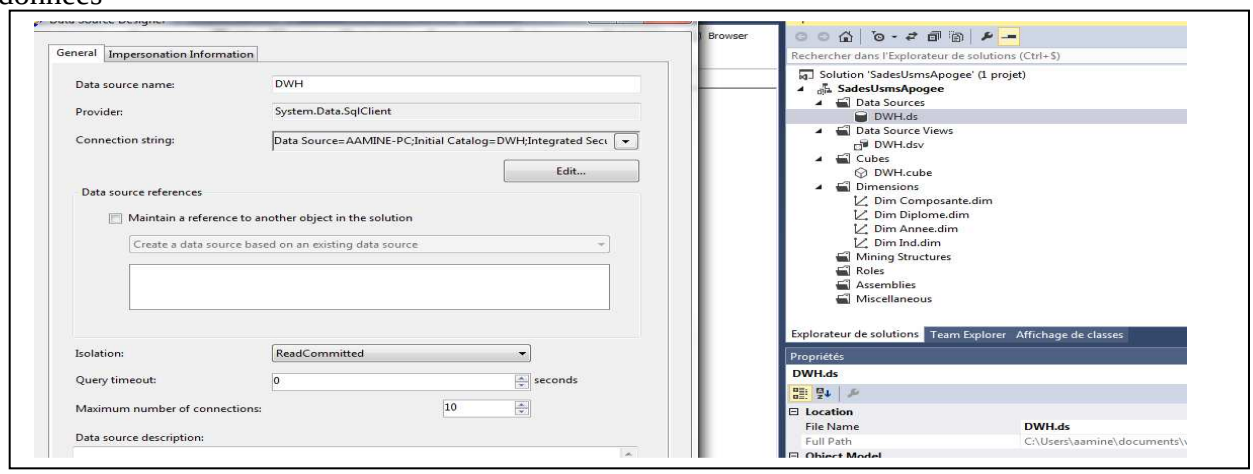


Figure 42: Source de données :

III.10.3 Création d'une vue de sources de données

Une vue de source de données contient le modèle logique du schéma utilisé par les objets multidimensionnels de base de données Analysis Services, autrement dit des cubes, des dimensions et des structures d'exploration de données.

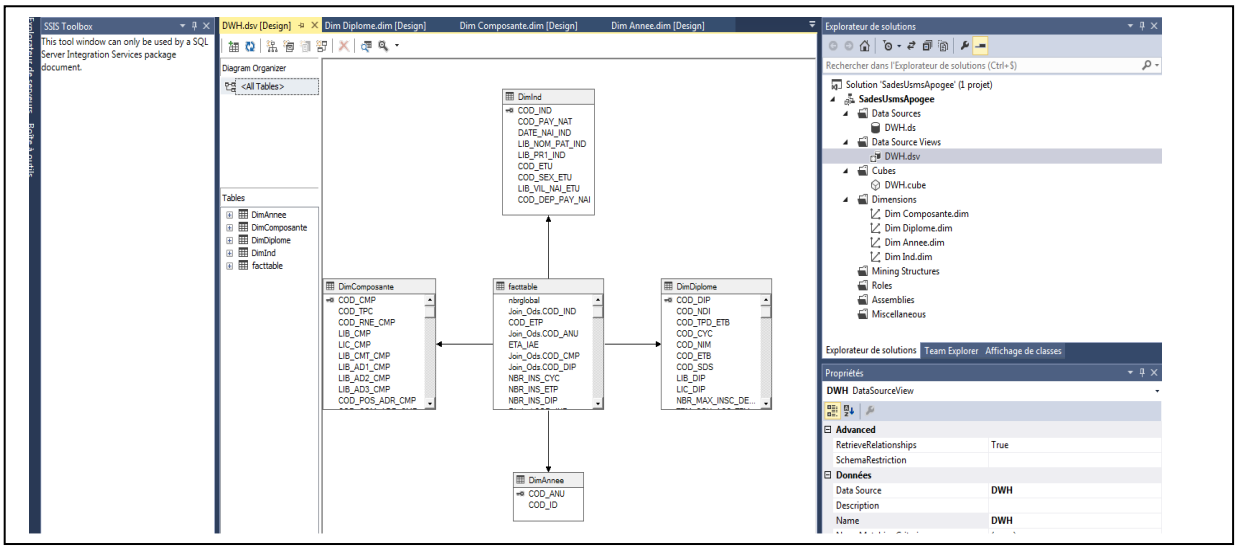


Figure 43: Création d'une vue de source de données

III.10.4 Création du cube

Dans les cubes OLAP, les données (ou mesures) sont classées par dimensions. Les cubes OLAP sont souvent pré-synthétisés entre les dimensions, ceci afin d'accélérer considérablement l'interrogation par rapport aux bases de données relationnelles.

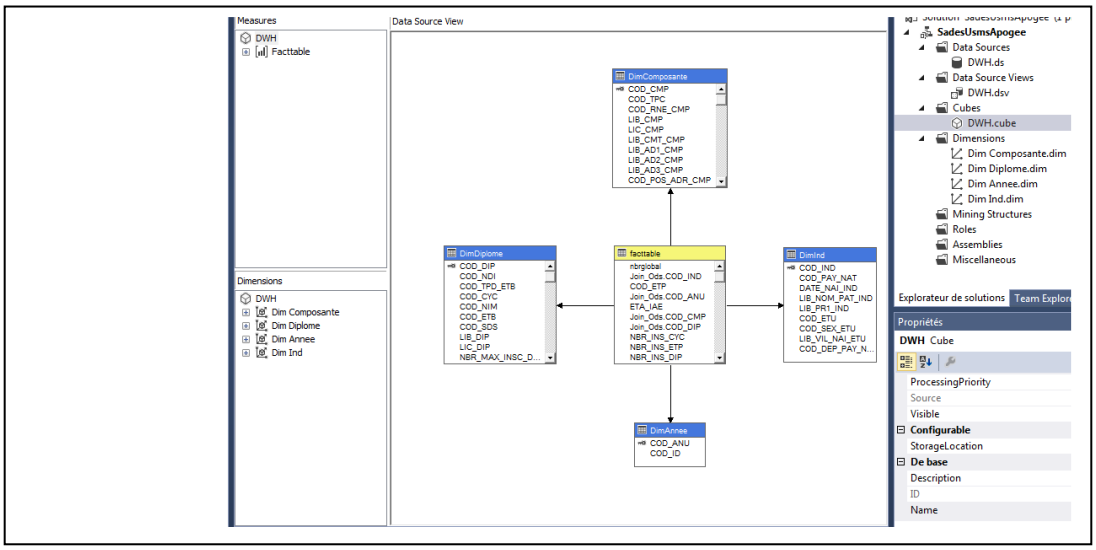


Figure 44: Création du Cube

III.11 Partie Reporting

III.11.1 Tableaux de bords avec le Power BI

Power BI est une solution de Business Intelligence développée par Microsoft pour permettre aux entreprises d'agréger, d'analyser et de visualiser les données en provenance de sources multiples

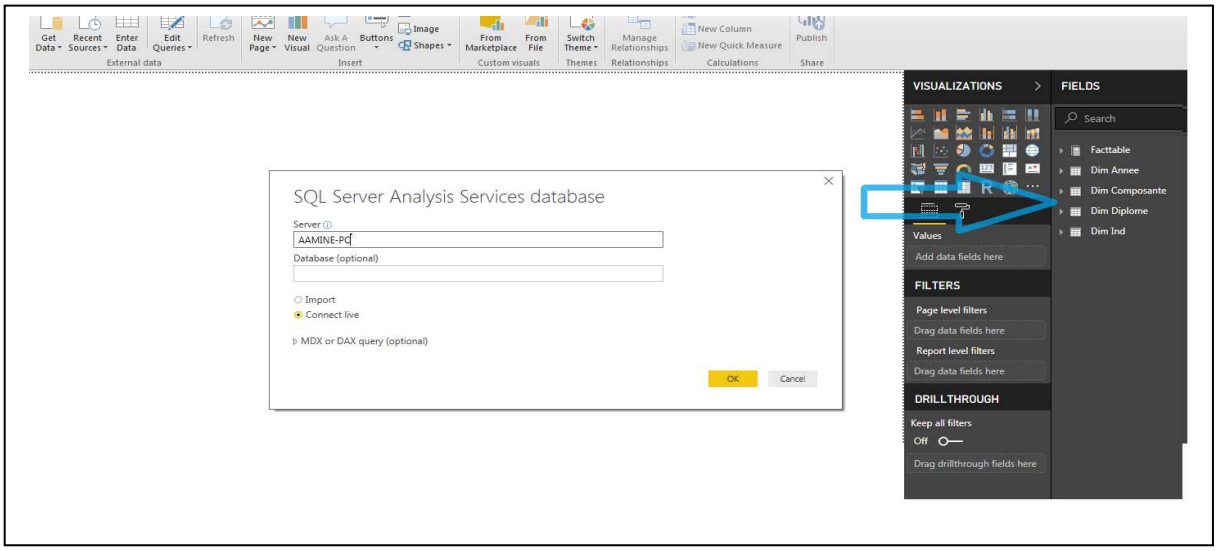


Figure 45: Connexion Power BI avec SSAS

La connexion du Power BI avec le SSAS permet de visualiser les différents attributs des tables de dimensions avec l'indicateur calculé dans la table de fait.

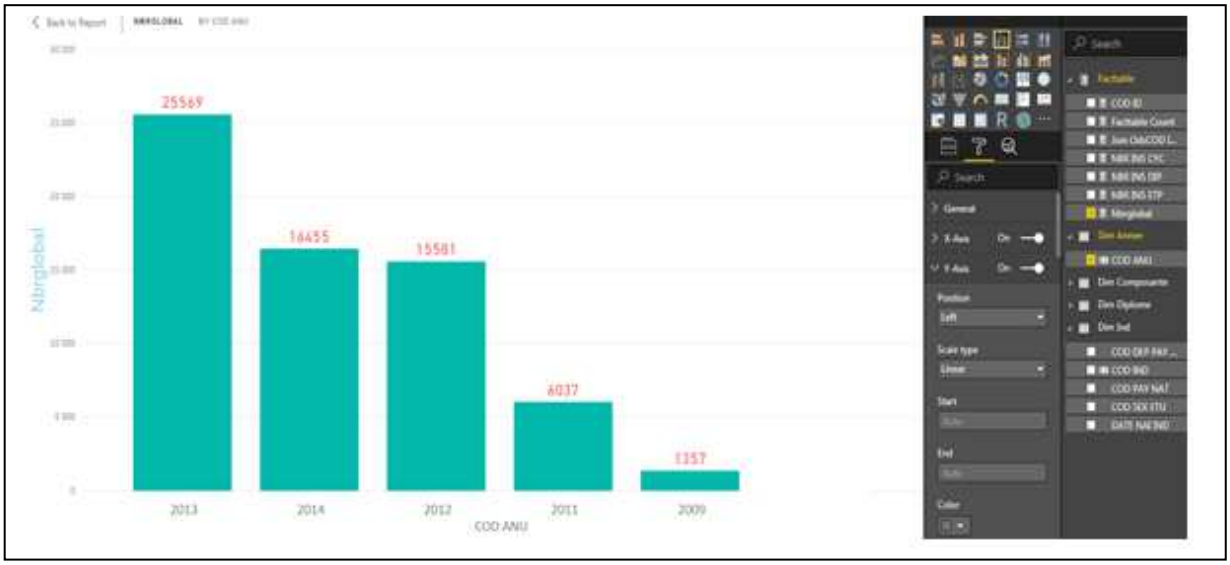


Figure 46: Nombre Globale des étudiants par année

Les statistiques présentées dans la figure 46 montrent le nombre global des étudiants par année (de l'année 2009 jusqu'au l'année 2014) évolue d'une façon rapide.

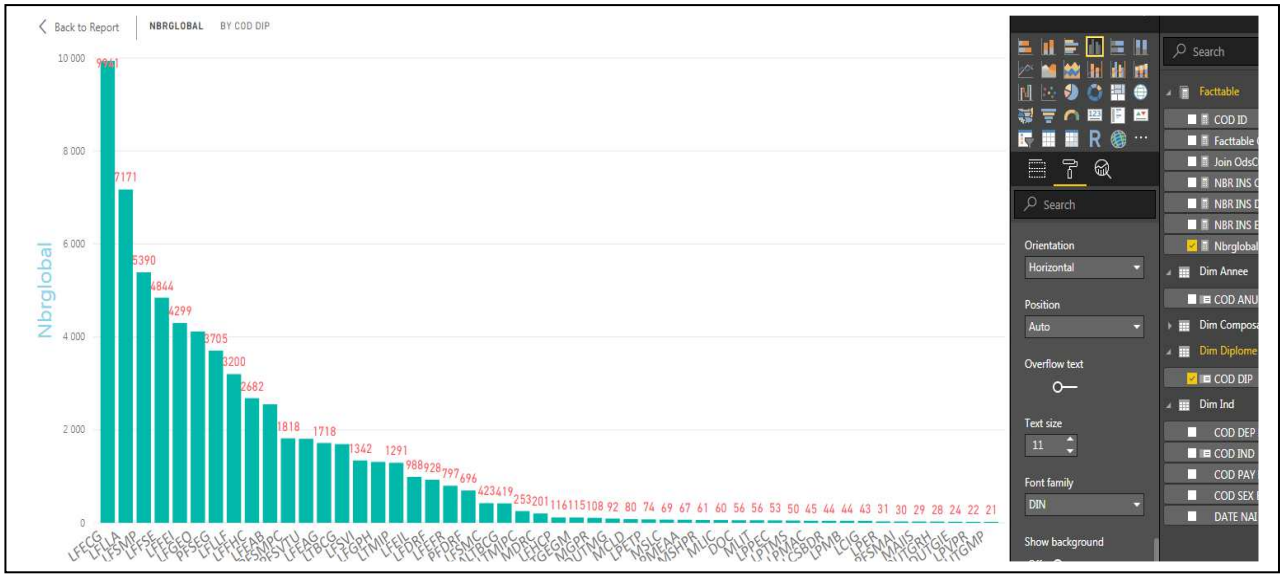


Figure 47: Nombre Globale des étudiants par Diplôme

Les statistiques présentées dans la figure 47 montrent le nombre global des étudiants par diplôme.

La majorité des étudiants sont inscrits dans la licence fondamentale d'économie et de gestion.

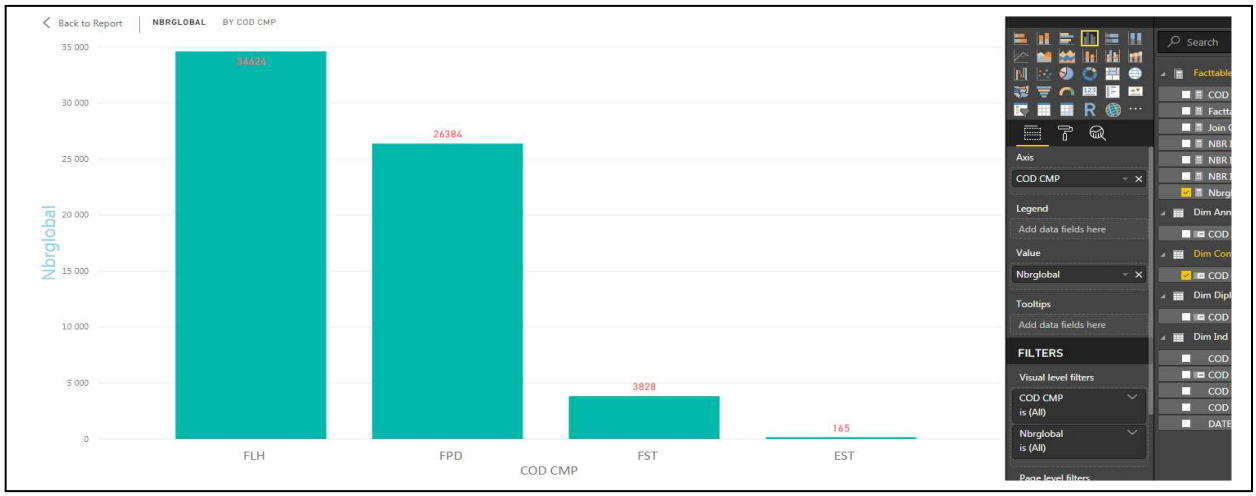


Figure 48: Nombre Globale des étudiants par Composante

Les statistiques présentées dans la figure 48 montrent le nombre global des étudiants par établissement.

La majorité des étudiants sont inscrits à la faculté des lettres et des sciences humaines.

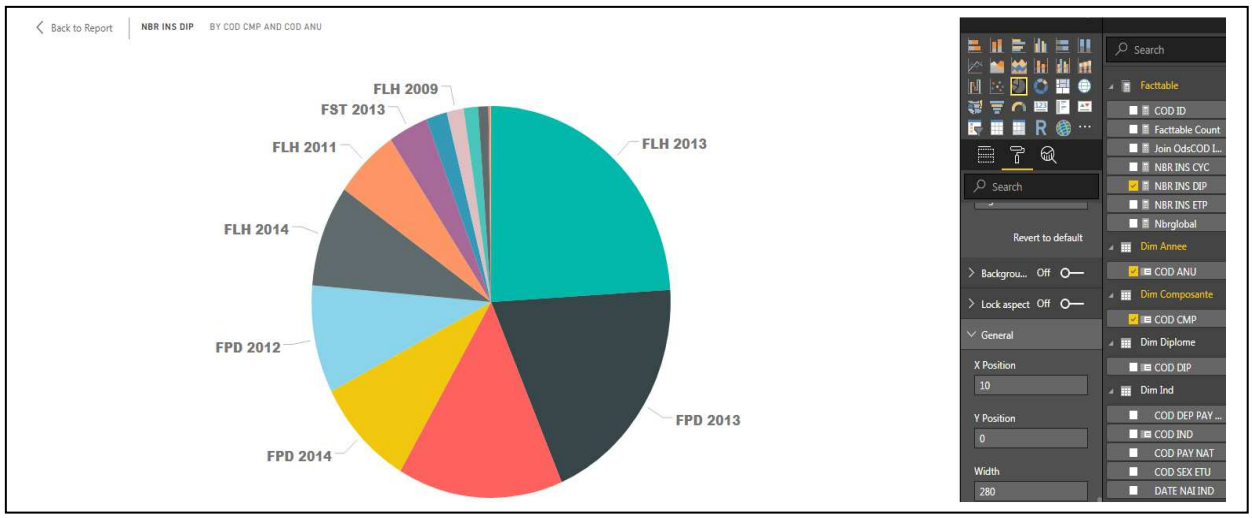


Figure 49 : Nombre des étudiants inscrits par Diplôme par Année par composante

Le résultat présenté dans la figure 49 englobe le nombre des étudiants inscrits par diplôme, par année et par établissement.

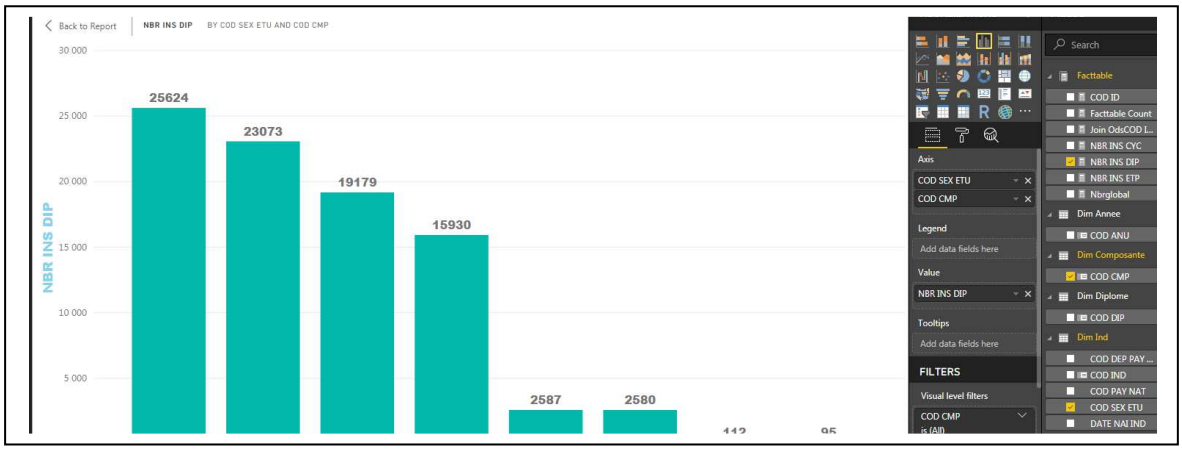


Figure 50: Nombre des étudiants inscrits par Diplôme par composante par Sexe

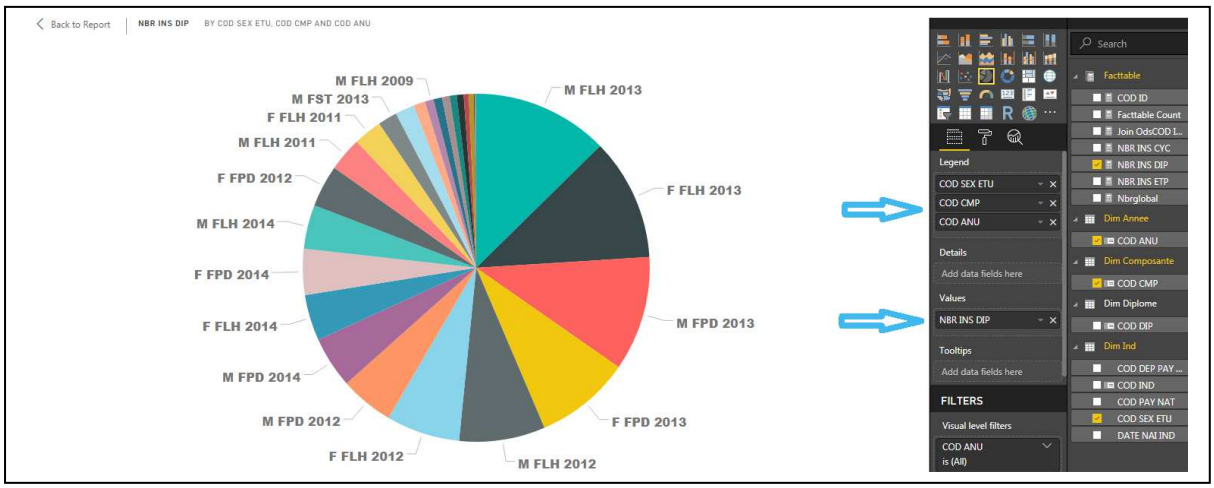


Figure 51: Nombre des étudiants inscrits par Diplôme/ composante/ Sexe/ Année

Les statistiques présentées dans les figures 50 et 51 montrent que le nombre global des étudiants inscrits par diplôme, par établissement, par sexe et par année augmente d'une façon rapide.

Figure 52: Surface de plusieurs attributs

Pour la réalisation de plusieurs attributs, la figure 52 englobe cette fonctionnalité. A partir de cette interface on peut réaliser plusieurs statistiques avec des indicateurs différents.

III.12 Conclusion

La mise en place d'une solution de BI consiste à adopter une démarche prédictive. Tout l'enjeu réside dans la possibilité de fournir la bonne information, au bon moment et à la bonne personne

Déployer une solution de BI apporte de nombreux avantages :

- Solution intuitive de présentation de rapports
- N'importe quel type de données (Excel, ERP, etc.) est exploité
- Gain de temps et analyse plus juste et plus puissante
- Meilleur accès à l'information, analyse à tout moment et n'importe où, toujours à jour
- Amélioration des performances et des stratégies globales puisque accès aux bonnes informations
- Amélioration de l'efficacité de l'entreprise grâce à une meilleure collaboration

- Développement de la stratégie, définition des objectifs, surveillance des performances, analyse de groupe et prise de décisions avisées qui prennent en charge la stratégie globale de l'entreprise

- Présenter de manière structurée et cohérente les informations
- Analyser les données de l'entreprise
- Faciliter la prise de décision grâce à des indicateurs pertinents
- Consolider l'ensemble des données, achats, ventes, comptabilité, clients, etc.
- Automatiser le processus de décision en se basant sur les mêmes indicateurs pour toute l'entreprise
- Améliorer la visibilité sur les chiffres, les écarts, les anomalies
- Anticiper et prévoir les tendances

Comme expliqué, un entrepôt de données ne met pas l'information à disposition de l'utilisateur basique. Pour cela, il faut adjoindre à l'entrepôt un Data mart[69], dont le rôle est de retraiter l'information pour la rendre exploitable par un corps de métier de l'entreprise. A chaque métier son Data Mart : l'information, pour être transformée en ressources, doit être retravaillée sous forme d'agrégats pour être compréhensible de l'utilisateur. De plus Le DM est une base de données moins coûteuse que le DW et plus légère puisque destinée à quelques utilisateurs d'un département

Dans le chapitre suivant on va faire une analyse avancée sur le projet SDUM basée sur les algorithmes du Datamining suivie d'une application basée sur le langage JAVA afin de réaliser des tableaux de bords qui permet de faire un suivi de la qualité du service globalement offerte aux utilisateurs de l'université afin de faciliter le pilotage d'une ou plusieurs activités dans le cadre d'une démarche de progrès ainsi d'avoir une vision immédiate et instantanée d'une situation donnée dans le temps.

VI . Analyse avancée du SDUM

VI.1 Introduction

Le développement avancé du projet SDUM est basée sur quelques techniques du data mining qui permet de préciser l'influence des critères comme la filière, la moyen, la région, le sexe sur le résultat d'étudiant ou bien la relation existante entre ces critères et le résultat

Le résultat est un variable cible parmi plusieurs variables qu'on va les exploiter au futur

Ce travail permettra de prendre les décisions pertinentes pour classifier les étudiants selon ces critères afin de réaliser une application basée sur JAVA comme outil de reporting.

VI.2 Comparaison entre Datawarehouse et Datamining

	Data Warehouse	**Data Mart**
Cible utilisateur	Toute l'université	Chaque département
Implication du service informatique	Elevée	Faible ou moyen
Base de données d'entreprise	SQL type serveur	SQL milieu de gamme, bases multidimensionnelles
Modèles de données	A l'échelle de l'université	Département
Champ applicatif	Multi sujets, neutre	Quelques sujets, spécifique
Sources de données	Multiples	Quelques-unes
Stockage	Base de données	Plusieurs bases distribuées
Taille	Centaine de GO et plus	Une à 2 dizaines de GO
Matériel	Unix	NT, petit serveur Unix

Tableau 11: Comparaison entre Datawarehouse et Datamining

VI.3 Mise en place

Construire un ou plusieurs Data Marts [70][71] départementaux au lieu d'un entrepôt de données central permet de valider rapidement le concept d'informatique décisionnelle. Mais construire des Data Marts n'est pas sans risques :

- En effet, dans les universités, des Data Marts isolés peuvent proliférer. Ces universités risquent de retomber dans le piège d'une architecture composée de multiples systèmes décisionnels incohérents, contenant des informations redondantes. Cela coûte plus cher et c'est plus complexe à gérer qu'un entrepôt de données centralisé.
- Les Data Marts résolvent les problèmes de performance des gros entrepôts de données. Mais ils font régresser vers le vieux problème des îlots isolés. Les universités vont devoir affronter des problèmes techniques complexes et coûteux pour remettre en cohérence les ensembles.

- Fédérer des Data Marts ou les faire évoluer vers une structure centralisée n'est pas facile.
- On peut se poser la question s'il est préférable de bâtir un gros et unique entrepôt de données ou bien de concevoir un réservoir plus modeste, nourri par les données d'un seul département. Il est intéressant de commencer par un Data Mart, à condition de respecter certaines règles :
- Impliquer les utilisateurs.
- Ne pas construire de multiples Data Marts isolés.
- Bannir les redondances.

Les DataMarts sont facile à déployer que les entrepôts de données. Les Data Marts peuvent évoluer facilement vers un entrepôt de données. Les différents Data Marts indépendants peuvent être dynamiquement couplés pour se métamorphoser en entrepôt de données [72][73].

VI.4 Les algorithmes du datamining

Face à la limite de la classification binaire dans les modèles de réponse, Charles et al dans ont appliqué ada-boost aux réseaux bayésiens comme des algorithmes d'apprentissages efficaces pour classifier les clients par aptitude de réponse aux offres.

Les algorithmes d'induction des arbres de décisions sont présentés dans le tableau suivant :

Nom de l'algorithme	Développeur	Année
CHAID	Kass	1980
CART	Breiman, et al	1984
ID3	Quinlan	1986
C4.5	Quinlan	1993
SLIQ	Agrawal, et al	1996
SPRINT	Agrawal, et al	1996

Tableau 12: Algorithmes de classification

Apriori a été le premier et avant tout algorithme et constitue la base de la plupart des algorithmes connus dans ce type de problème. En cas de très grands ensembles d'entrée, Apriori souffre de deux problèmes répétés de balayage d'E / S et le cout de calcul élevé. Agrawal et al. [74][75] a proposé l'algorithme hybride AprioriTid. Park et al a proposé une

optimisation, appelé DHP (hachage direct et élagage) destiné à restreindre le nombre d'itemsets candidats. Brin et al présentent l'algorithme DIC qui sépare la base de données dans des intervalles d'une taille fixe de manière à réduire le nombre de passages à travers la base de données. L'algorithme par ZakiÉclat [76] est considéré comme l'archétype dans la première profondeur de manière génération de motifs fréquents. Algorithme FP-Growth par Han et al [77]. est le plus célèbre et largement utilisé. Il s'agit d'un premier algorithme de profondeur. Yu-Chiang Li et al.ont évalué l'importance des itemsets pour l'extraction de règles d'association à partir de bases de données. Ils ont proposé un algorithme, Enhanced FSM (MESF), ce qui réduit efficacement la complexité du temps de l'étape de jointure. C. Hidbera présenté un nouvel algorithme nommé CARMA (Continuous Association règle Algorithme Mining), qui est utilisé pour calculer les grands itemsets.

Les algorithmes d'induction des règles associatives son présenté dans le tableau13 :

Nom de l'algorithme	Développeur	Année
APRIORI	Agrawal, et al	1993
FP-GROWTH	Han, et al	2000
ECLAT	Zaki	2000
SSDM	Escovar, et al	2005
KDCI	Orlando, et al	2003

Tableau 13: Les algorithmes de règles d'association

La technique de clustering permet d'identifier les groupes d'individus ayant des caractéristiques similaires. On peut l'utiliser pour distinguer les segments des étudiants par sexe/faculté/région... Pour atteindre ce but on a utilisé les algorithmes k-means [78-82].

VI.5 Application de quelques algorithmes du data mining à la classification des étudiants dans l'université de Beni Mellal, Maroc

VI.5.1 Introduction

L'évolution des étudiants dans l'université est de nos jours une réalité à laquelle elle doit faire face. Notamment, elle doit permettre à ses responsables de déceler les informations pertinentes afin de prendre les bonnes décisions dans les .plus brefs délais.

On prend comme exemple l'évolution du nombre d'étudiants inscrits dans la faculté polydisciplinaire. Les Statistiques fournies par l'application APOGEE

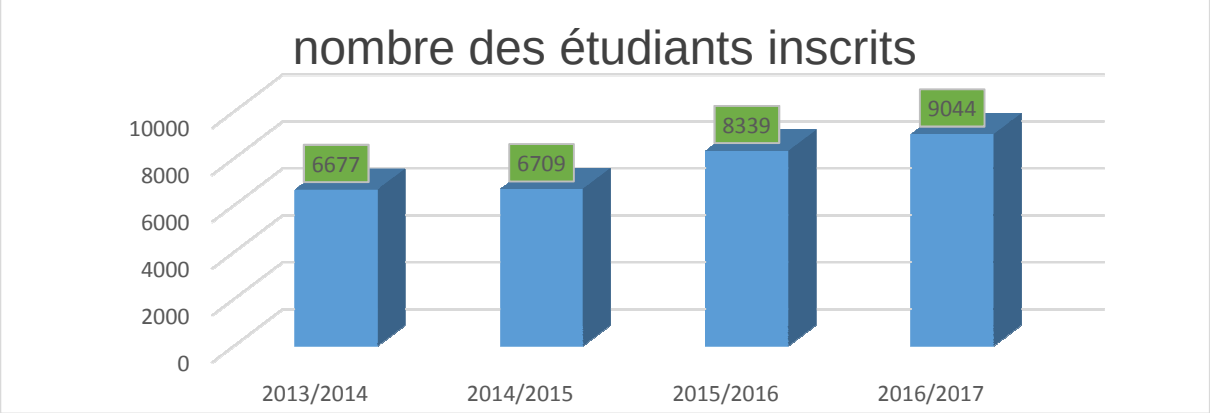


Figure 53:nombre des étudiants inscrits à la polydisciplinaire

VI.5.2 Problématique

La croissance du nombre d'étudiants est la clé du succès pour le développement d'une université. Différentes techniques d'analyse des données s'appliquent à la classification des étudiants dans une université. Nous appliquons certaines techniques d'exploration de données, ce qui nous permet de préciser l'influence de critères tels que la moyenne, la région, le sexe sur le résultat comme variable cible de l'étudiant afin d'analyser la relation existante entre ces critères et le résultat. Le but est de prendre les décisions pertinentes pour classer les étudiants selon ces critères.

VI.5.3 Solution

Pour atteindre ce but décrit dans la problématique, nous adoptons des techniques de datamining pour classifier des étudiants selon leurs résultats, pour les segmenter selon la similitude de leur sexe, des niveaux et des régions et savoir les étudiants qui sont forts dans les modules de leur programme d'étude.

VI.5.4 Définitions

Item : Un item est tout objet, article, attribut, littéral, appartenant à un ensemble fini d'éléments distincts I = {x1 ; X2 …Xn}, par exemple l'ensemble des modules {M1, M2, M3, M4, M5, M6, M7, M8}.

Itemset : On appelle ItemSet tout sous-ensemble d'items de I. Un itemset constitué de K items sera appelé un K Itemset, par exemple l'ensemble des modules validés pour un étudiant {m1, m2, m5}.

Transaction : Une transaction est un itemset identifie par un identificateur unique TID. L'ensemble de toutes les transactions tIDs sera désigné par l'ensemble T.

Support : Le support d'un itemset X, noté supp(X) est la proportion de transactions de D contenant X.

$$\text{Support(X)} = \frac{Count(X)}{Nbre\_Transactions}$$

Confiance : C ($\boxed{X \cup Y}$)= Supp ($\boxed{X \cup Y}$) / Supp (X)

Une règle d'association est valides si :

S$\boxed{(X \cup Y)}$>=MinSupp

C ($\boxed{X \cup Y}$) >= MinCon

Avec minSupp et minConf sont les seuils pour le support et la confiance.

Dans le domaine de Datamining la recherche et la génération des motifs fréquents dans une grande base de données, joue un rôle très important.

Exemple : en extrait des règles d'association concernant les modules que les étudiants sont forts de leur filière.

Considérons la base de données D figurant dans le tableau ci-dessous et supposons que le support minimum (minSupp) est de 2 transactions.

Etudiants (transactions)	Module1	Module2	Module3	Module4
Etud1	16	13	10	6
Etud2	12	8	12	7
Etud3	14	17	13	8
Etud4	15	4	14	9
Etud5	6	13	11	3
Etud6	7	9	7	14

Tableau 14:Les Transactions des étudiants

En la transforme à la base formelle (le module validé « 1 » et non validé « 0 » :

Etudiants (transactions)	Module1	Module2	Module3	Module4
Etud1	1	1	1	0
Etud2	1	0	1	0
Etud3	1	1	1	0
Etud4	1	0	1	0
Etud5	0	1	1	0
Etud6	0	0	0	1

Tableau 15: La transformation de base formelle pour la table15

Paramètres : Fixer un degré d'exigence sur les règles à extraire

>> Support min. (ex. 2 transactions)

>> Confiance min. (ex. 75%)

L'idée est surtout de contrôler (limiter) le nombre de règles produites.

Démarche : Construction en deux temps

>> Recherche des itemsets fréquents (support >= support min.)

>> À partir des itemsets fréquents, produire les règles (conf. >= conf. min.)

Quelques définitions :

>> Item = produit

>>Itemset = ensemble de modules (ex. {m1, m3})

>>sup (itemset) = nombre de transactions d'apparition simultanée des modules (ex. sup {m1, m3} = 4)

>>card (itemset) = nombre de produits dans l'ensemble (ex. card {m1, m3} = 2)

VI.5.5 Recherche des règles d'association

Sup Min=2, Items=Modules

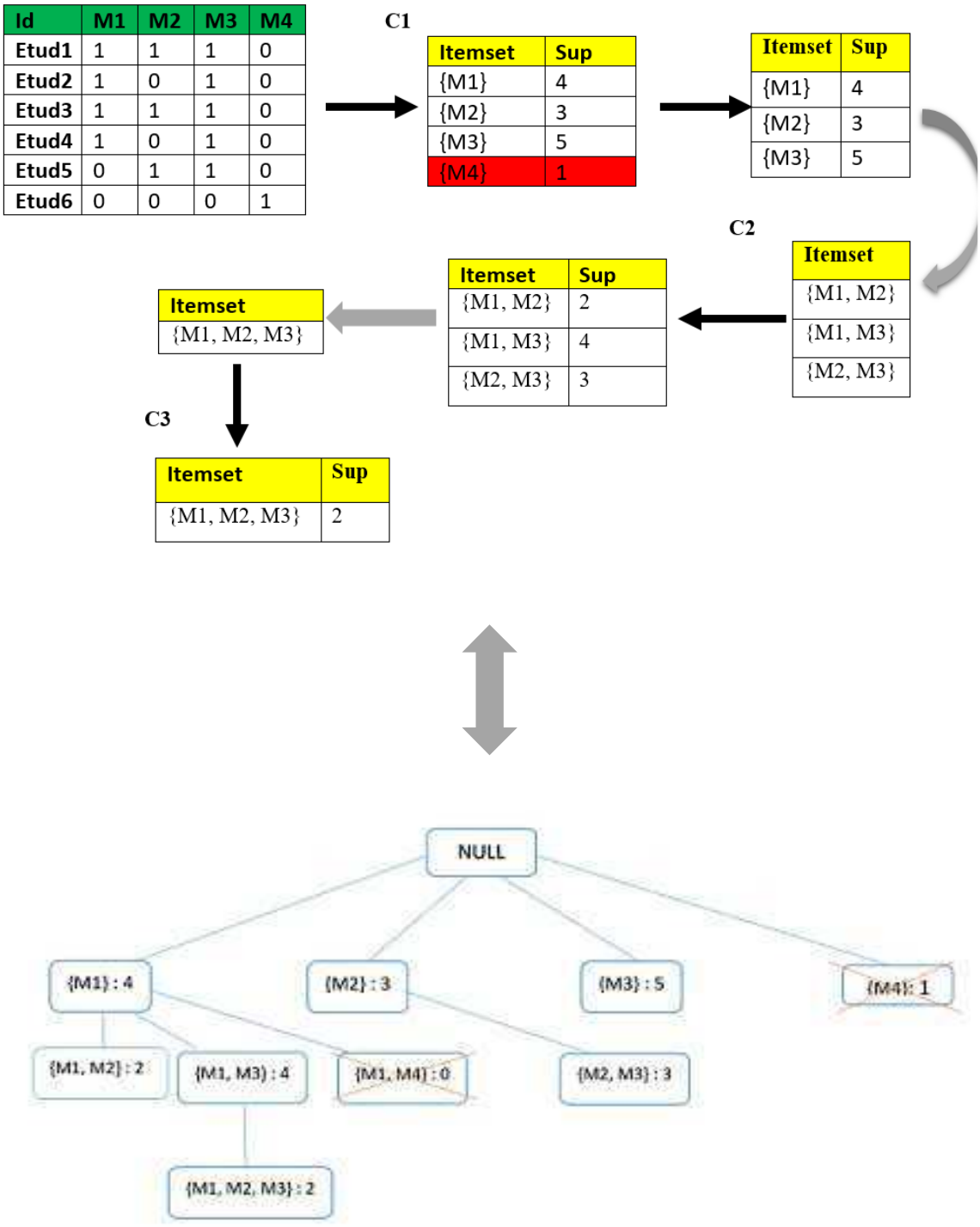
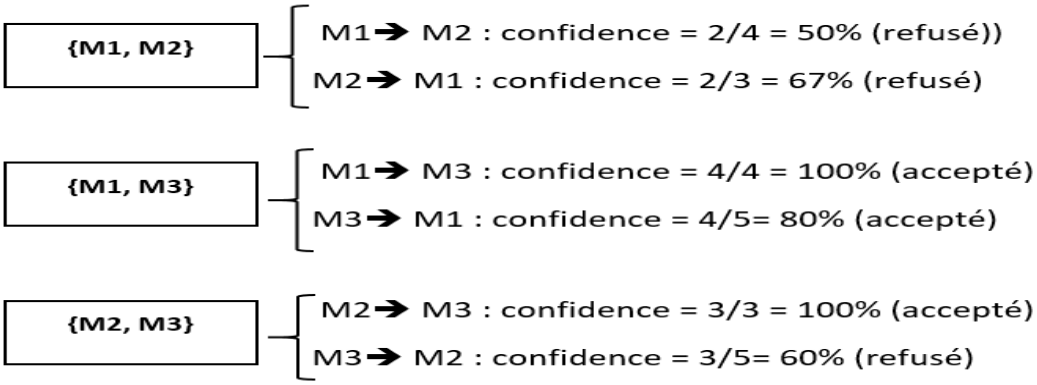


Figure 54: Recherche des règles d'association

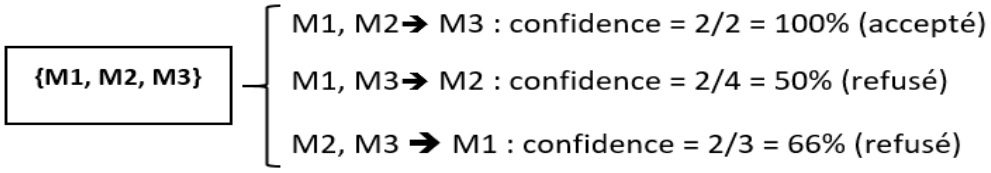
Pour le module 4 son support (supp (m4) <minsupp=2), et aussi pour la règle (m1et m4).

- Recherche des règles d'association pour les itemsets de card=2 :

Il faut tester toutes les combinaisons : 2 tests par itemset



- Recherche des règles d'association pour les itemsets de card= 3 :



VI.5.6 Interprétation :

Donc on s'interprète les règles d'association suivantes :

- ✓ les étudiant sont validés le module 1, ils sont validés le module 3,
- ✓ les étudiant sont validés le module 3, ils sont validés le module 1
- ✓ les étudiant sont validés le module 2, ils sont validés le module 3
- ✓ les étudiant sont validés le module 1 et 2, ils sont validés le module 3

VI.5.7 Segmentation des étudiants

La segmentation [83] est une technique qui permet de découper une université déterminée et d'isoler les étudiants qui se distinguent par un comportement identique, alors bien comprendre on va donner un exemple afin de l'étudier comme un cas pratique :

- La démarche : segmentation
- L’université.
- Le groupe non identifié : les étudiants

Au sein de la faculté on ne peut pas mettre le doigt sur tous les étudiants au même temps, car l’université n’est jamais une entité homogène.

Les étudiants sont appelés "segments d’université”, ils se diffèrent selon leurs sexes à savoir mention bac, filières, régions etc. ... Chacune de ces variables peut être utilisée pour segmenter les étudiants.

L’université cherche donc à délimiter des catégories des étudiants différents (des segments) qui peuvent être comme des cibles et les atteindre en utilisant une stratégie spécifique.

Par exemple, les étudiants peuvent être groupés selon la similitude de sexe, la catégorie ou la zone(le domaine) de résidence.

VI.5.8 Classification :

Classificateurs sont des modèles de calcul généraux pour l'attribution d’un résultat à une entrée. Les entrées peuvent être des vecteurs de caractéristiques ou les articles étant classés ou des données sur les relations entre les éléments. Le résultat est une classification spécifique à un domaine comme validation /Echec des étudiants. Classificateurs peuvent être mises en œuvre à l'aide de nombreuses stratégies d’apprentissages automatiques différents, y compris les arbres de décision, les réseaux de neurones, et les réseaux bayésiens.

En raison, de clarté de résultats on a utilisé les arbres de décisions qui offrent une sortie claire et facile à interpréter.

- Classification des étudiants par résultat **:**

Le principe de la classification basée sur la valeur du résultat est d’avoir des attributs déterminant l’étudiant (filière, moyen, région …) comme des attributs non cible et une classe cible qui est dans notre cas le résultat d’étudiant. Ce processus nous permet de préciser l’influence de ces critères sur le résultat d’étudiant ou bien la relation existante entre ces critères et le résultat.

Classification des étudiants par sexe :

Ce critère utilisé pour déterminer le sexe d’étudiant pour bien préciser

L’influence de ces critères sur le sexe d’étudiant.

- Classification des étudiants par filière.

- Classification des étudiants par note.
- Classification des étudiants par région
- etc.

VI.5.9 Conclusion

La mise en place d'un data mart peut être perçue comme un élément d'amélioration des prises de décision au niveau de l'entreprise. Nos travaux nous ont permis de faire une étude théorique sur les entrepôts de données ou data warehouse, les data marts, la modélisation dimensionnelle et le concept OLAP.

Cette étude était indispensable pour la mise place d'un data mart concernant la classification des étudiants dans l'université.

La mise en oeuvre d'un data mart nécessite la connaissance et la maitrise des concepts et méthodes étudiés. Aussi, comme tout projet informatique, pour que le projet réussisse, il faut :

- Une maitrise du sujet et une bonne organisation de projet ;
- Une forte implication des futurs utilisateurs ;

Au regard des processus multiples qui peuvent exister dans l'université, l'approche par les data marts pour la mise en place d'un système global d'aide à la décision de l'université peut s'avérer une solution efficace. Cela revient à l'application de la théorie du « Diviser pour régner».

Dans ce chapitre on a présenté le datamining, en expliquant le processus d'extraction de connaissance et citant ses différentes tâches et leurs algorithmes. Donc on va essayer d'appliquer et évaluer ces algorithmes dans l'entrepôt de données pour la réalisation de notre application.

VI.6 Phases Expérimentales

VI.6.1 Technologie et outils utilisés

a. NetBeans

NetBeans est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Development and Distribution License)

En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XXMML, Ruby, PHP et HTML. Il comprend toutes les

caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, éditeur graphique d'interfaces et de pages Web). Conçu en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java). Un environnement Java Development Kit JDK est requis pour les développements en Java. NetBeans constitue par ailleurs une plateforme
qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L''IDE NetBeans s'appuie sur cette plateforme. L'IDE NetBeans s'enrichit à l'aide de plugins

b. EasyPHP

EasyPHP est un environnement de travail packagé comprenant le serveur web Apache, le système de gestion de bases de données MySQL et le support du langage PHP.
Il est fourni avec phpMyAdmin, une interface permettant de manipuler très facilement les enregistrements de la base de données.
Les rôles de ces trois composantes sont les suivant :

Les rôles de ces quatre composants sont les suivants :

- Apacheest le serveur web « frontal » : il est « devant » tous les autres et répond directement aux requêtes du client web (navigateur) ;
- MySQL est un système de gestion de bases de données (SGBD). Il permet de stocker et d'organiser des données ;
- le langage de script **PHP** permet la génération de pages web dynamiques et la communication avec le serveur MySQL.

Tous les composants peuvent être situés :

- sur une même machine ;
- sur deux machines, généralement Apache et le langage de script d'un côté et MySQL de l'autre ;
- sur de nombreuses machines pour assurer la haute disponibilité.

c. PHP

PHP est un langage de programmation qui s'intègre dans les pages HTML. Il permet entre autres de rendre automatiques des tâches répétitives, notamment grâce à la communication avec une base de données (utilisation la plus courante de PHP).

Lorsqu'une page PHP est exécutée par le serveur, alors celui-ci renvois généralement au client (aux visiteurs du site) une page web qui peut contenir du HTML, XHTML, CSS, JavaScript ...

d. MySQL

MySQL est une **base de données** relationnelle très employée sur le Web, souvent en association avec *PHP* (langage) et *Apache* (serveur web). MySql fonctionne indifféremment sur tous les systèmes d'exploitation (*Windows*, *Linux*, *Mac OS* notamment).

Le principe d'une base de données relationnelle est d'enregistrer les informations dans des **tables**, qui représentent des regroupements de données par **sujets** (table des clients, table des fournisseurs, table des produits, par exemple). Les tables sont reliées entre elles par des **relations**.

Le langage **SQL** (acronyme de *Structured Query Language*) est un langage universellement reconnu par MySql et les autres bases de données et permettant d'interroger et de modifier le contenu d'une base de données. Les autres bases de données utilisées en informatique sont essentiellement *Microsoft SQL Server* et *Oracle*.

VI.6.3 Représentation de l'application

VI.6.3.1 Relations entre les tables

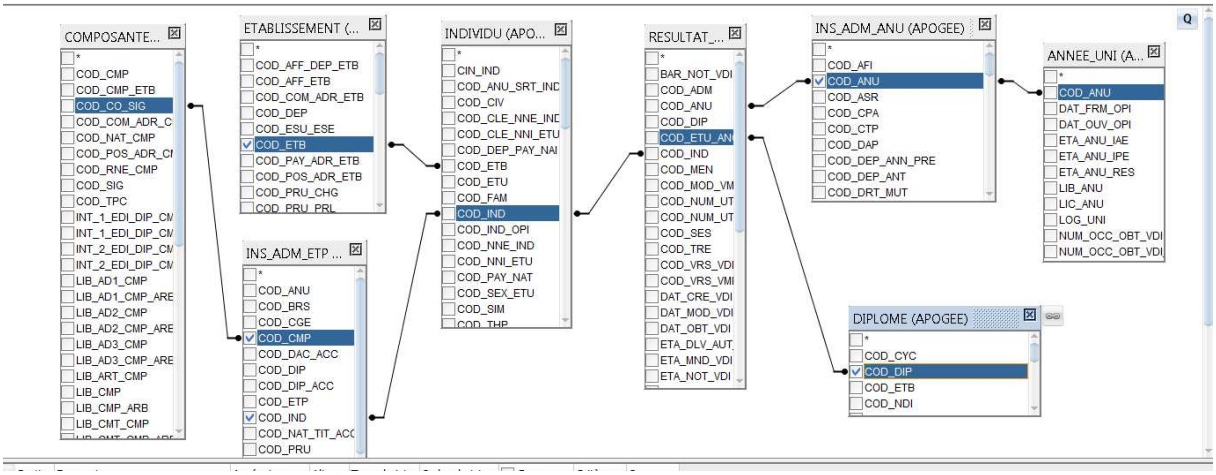


Figure 55:Modélisation des données

L'application est divisée en trois parties :

- Page d'accueil
- Partie Datamining
- Partie des statistiques

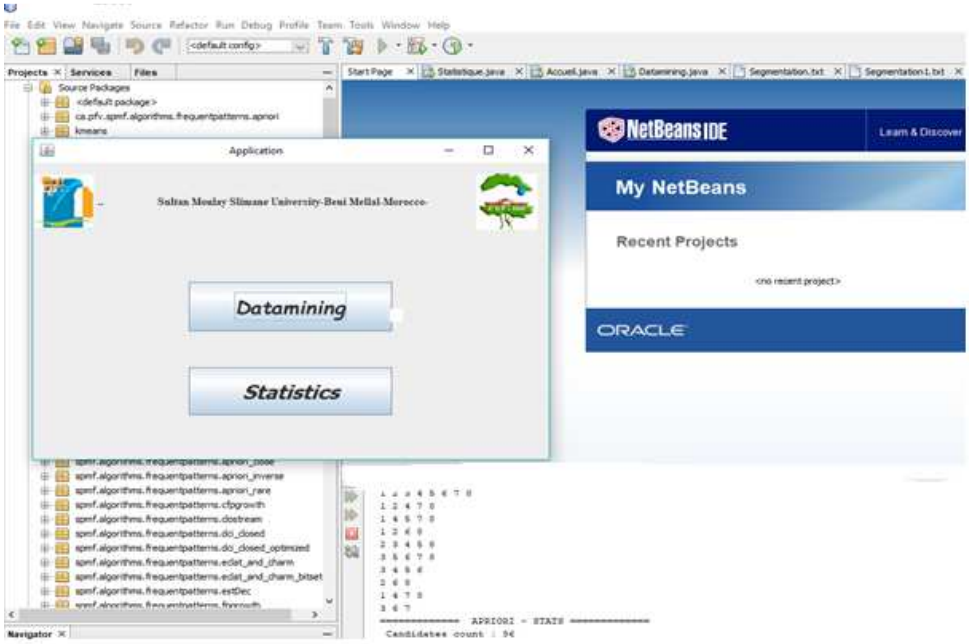


Figure 56: Interface de l'application

a) Partie Datamining de l'application

Dans la partie Data Ming, trois techniques ont été utilisés :

- La classification des étudiants,
- La segmentation des étudiants,
- Les règles d'associations.

La majorité du travail était concentré sur la classification des étudiants par leurs résultats ; Il existe deux statuts des résultats : valide et non valide.

Les attributs choisis pour la classification :

Région, composante, sexe, bac distinction, la discipline d'étude, l'année.

L'attribut cible : Résultats des étudiants

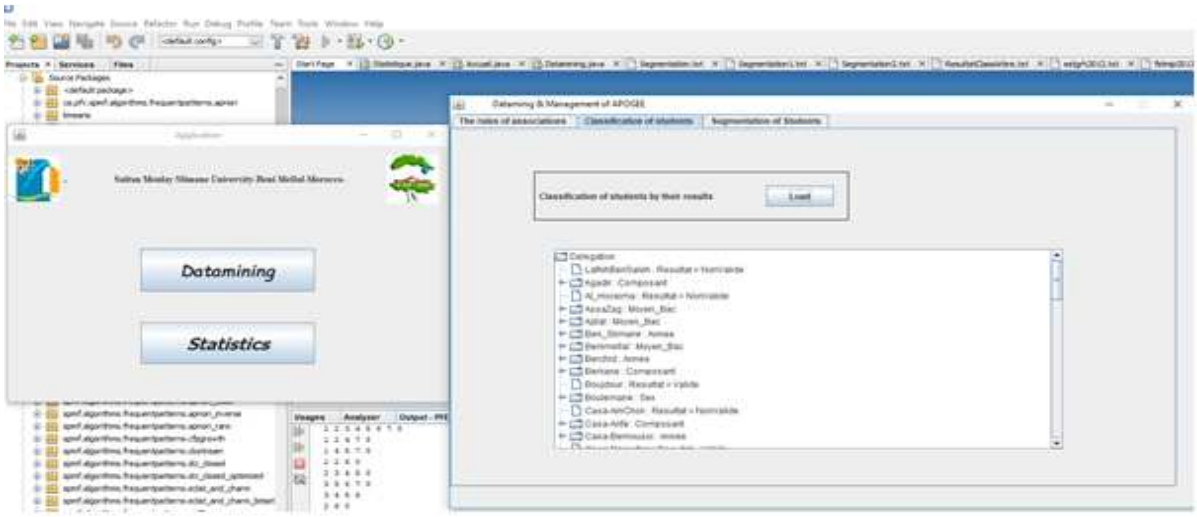


Figure 57: La classification des étudiants

b) Partie Statistique

Dans la partie statistique on a développé quatre rubriques :

- Les inscriptions par région et par mention bac
- La validation par année et par établissement
- Les inscriptions par sexe
- La bourse par région

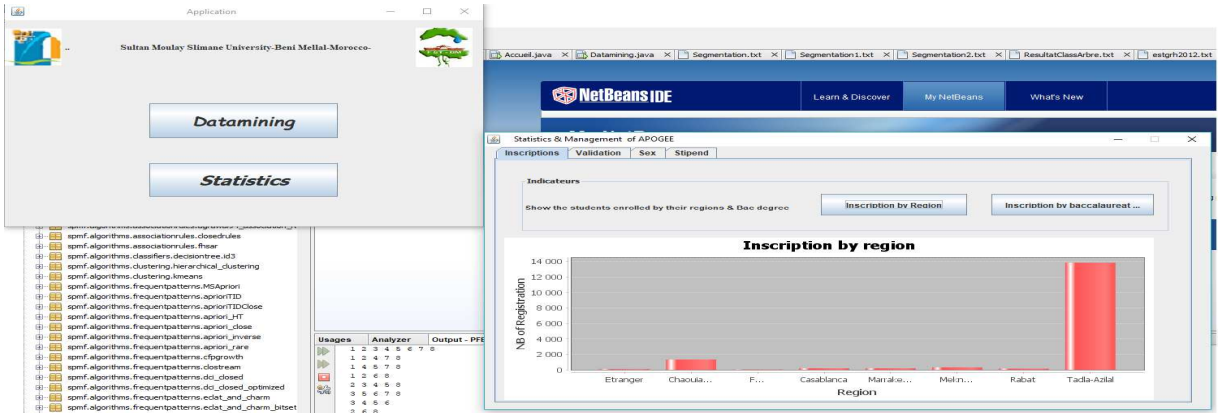


Figure 58: Inscription par région

Les statistiques présentées par la figure montrent que la majorité des étudiants inscrits, sont d'origine de la région Tadla Azilal

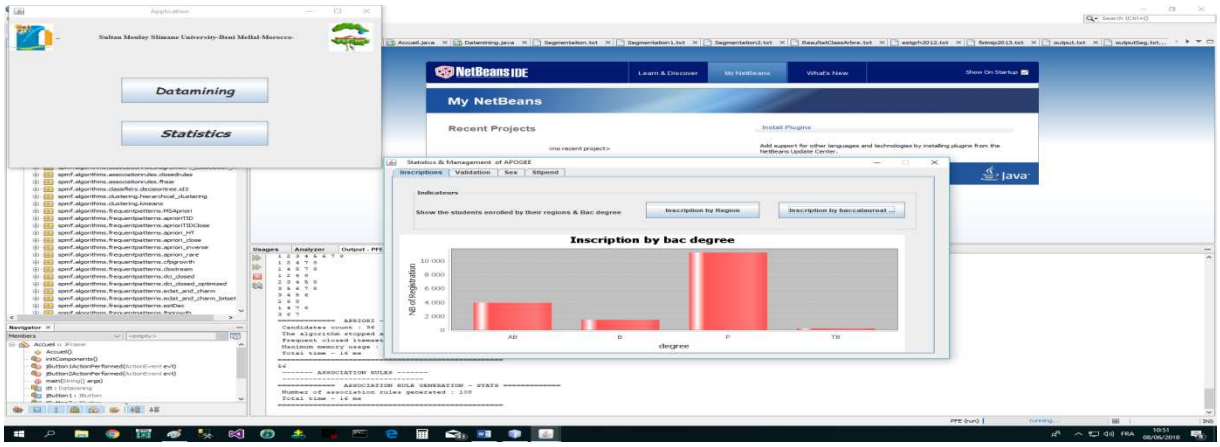


Figure 59: Inscription par mention bac

La majorité des étudiants qui sont inscrits dans les établissements universitaires ont un une mention de bac passable.

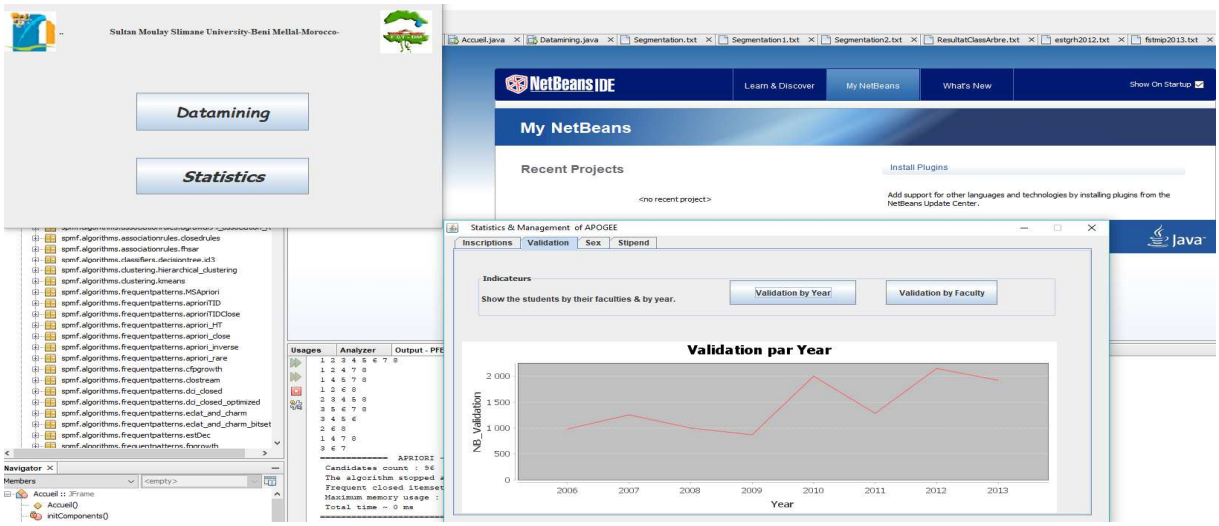


Figure 60: Validation par année

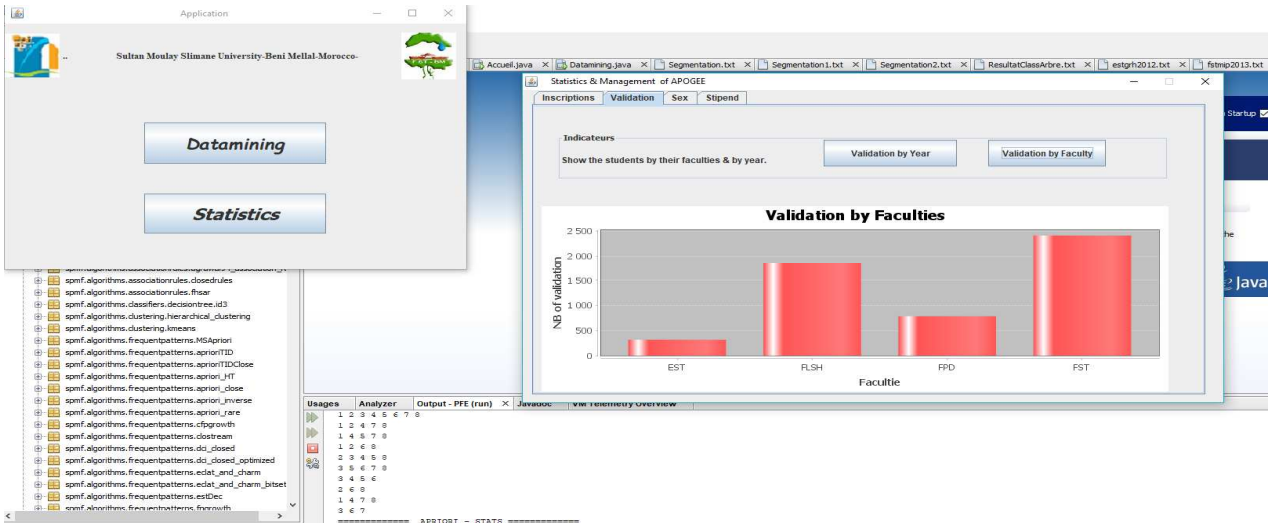


Figure 61: Validation par établissement

Les figures (60,61) montrent que la majorité des étudiants ont validé pendant l'année 2012 dans la faculté des sciences et techniques.

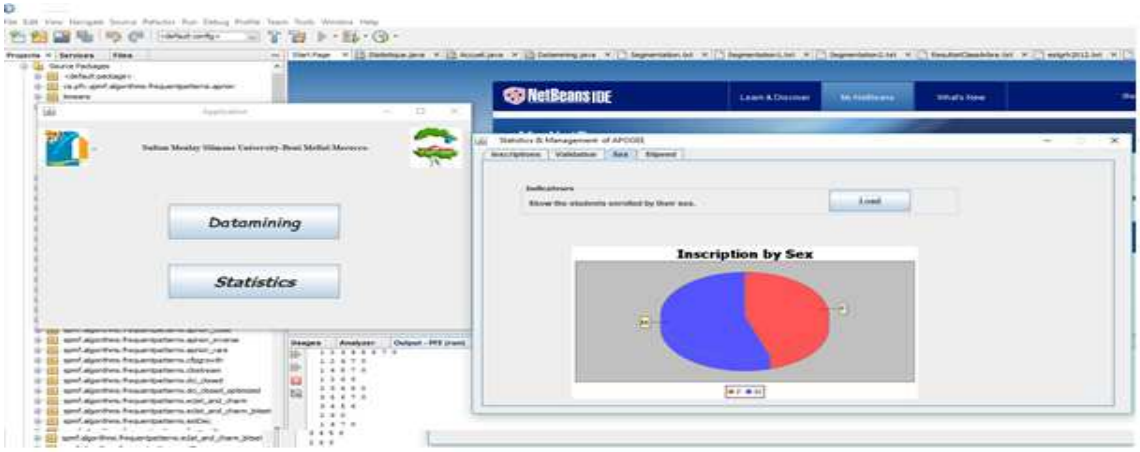


Figure 62: Inscription par Sexe

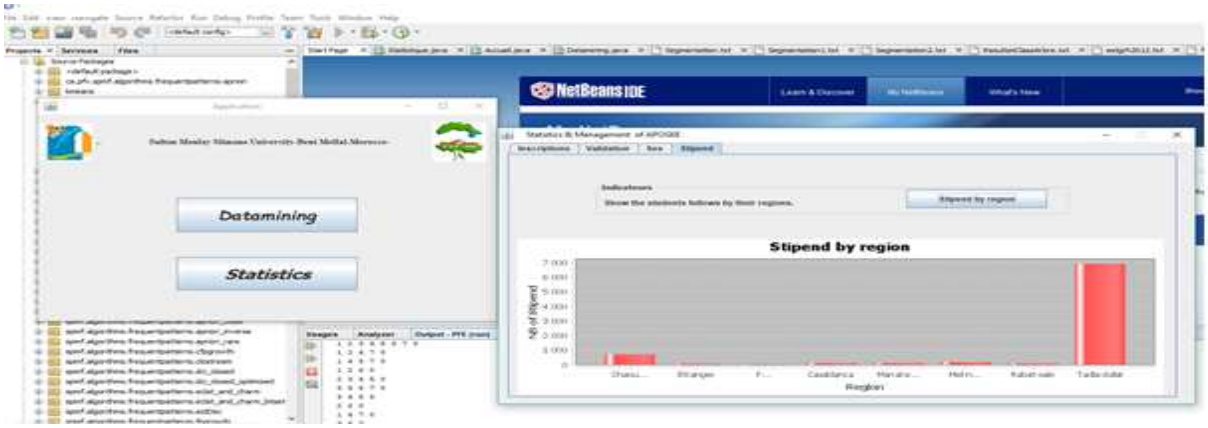


Figure 63: Bourse par région

La majorité des étudiants qui sont inscrits dans les établissements universitaires et qui ont issues de la région Tadla Azilal ont un une bourse.

VI.7 Comparaison des résultats avec le logiciel KNIME (the Konstanz Information Miner)

Knime. Est un logiciel open source utilisé par plus de 15000 utilisateurs dans le monde provenant de différents milieux (universitaires, recherche, petites et grandes entreprises) dans différents secteurs (banque, pharmacie, tourisme, ...) possède de solides atouts :

- Sa facilité d'utilisation et son interface graphique le rendent accessible aux non-initiés en fouille de données ;
- Il peut lire de très nombreux formats de données ;
- Il comporte de très nombreuses solutions pour prétraiter, analyser et visualiser des données et des résultats d'analyses ;
- La communauté des utilisateurs est très active et peut contribuer à ajouter de nouvelles fonctionnalités au logiciel.

VI.7.1 Importation des données

Après avoir démarré KNIME, un nouveau 'workflow' est créé (menu FILE / NEW). Nous le nommons « Classification ». Nous insérons le composant Excel Reader dans l'espace de travail. Nous le paramétrons (menu contextuel « Configure ») de manière à charger le fichier «classificationresultats.xlsx».

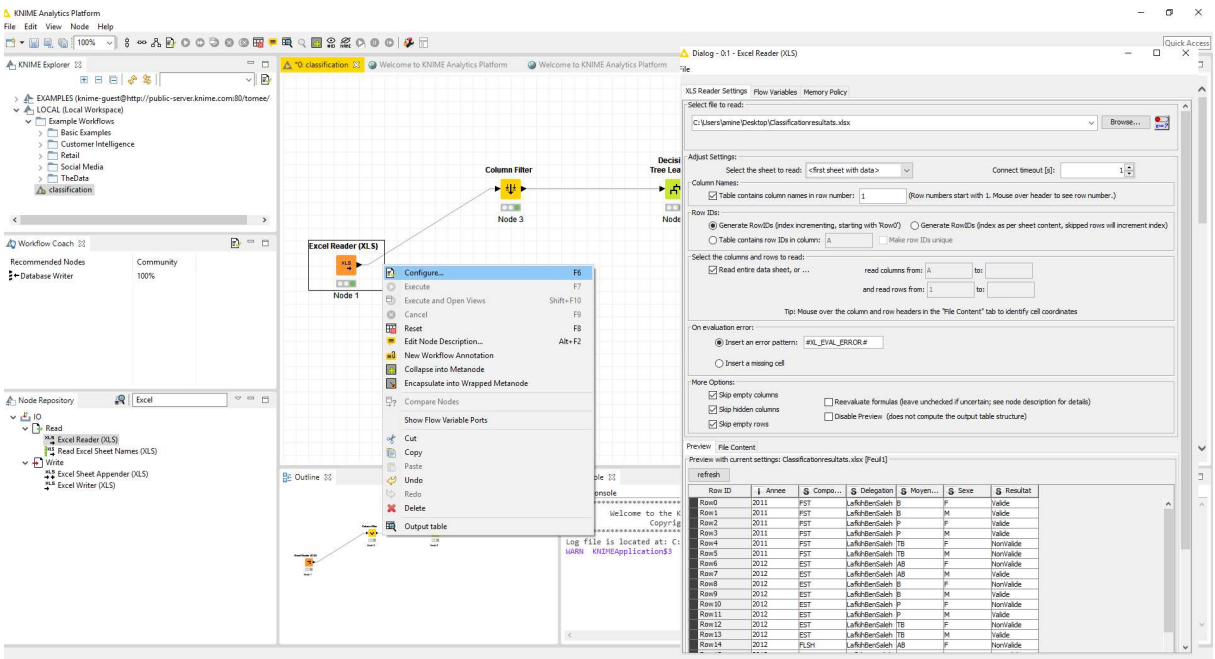


Figure 64: Importation des données par KNIME

Nous actionnons le menu contextuel « Execute ». Le témoin lumineux passe au vert si l'opération est couronnée de succès.

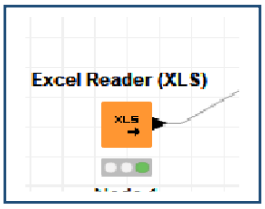


Figure 65: Le témoin lumineux

Nous utilisons le composant « ColumnFilter » pour sélectionner les attributs choisis pour la classification

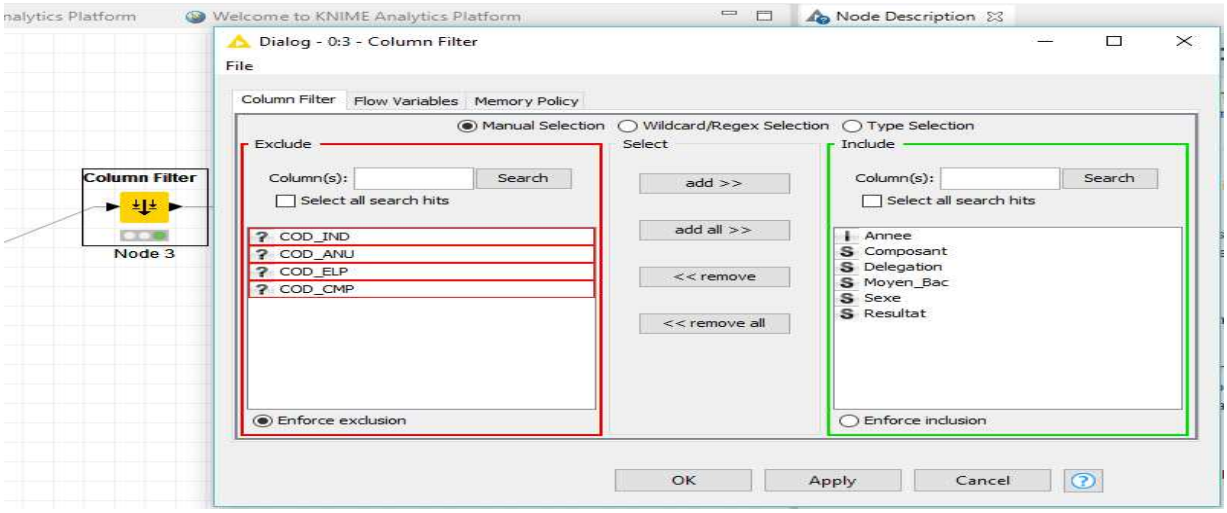


Figure 66: Le choix des attributs

De nombreux composants ont été placés dans l'espace de travail, voici la chaîne de traitements à ce stade de notre analyse.

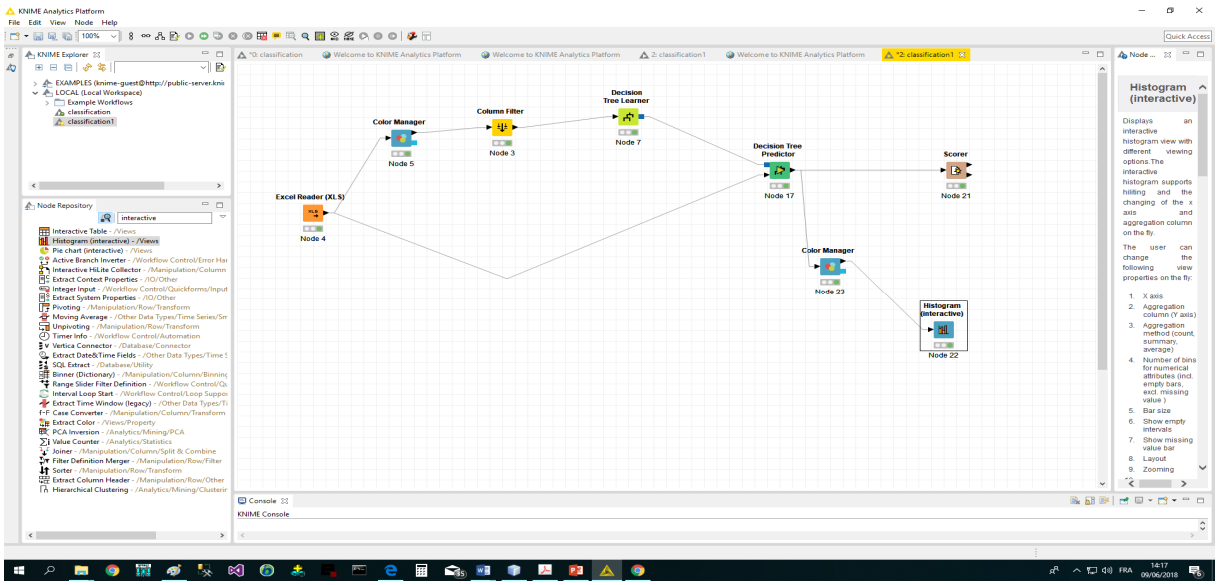


Figure 67: Espace de travail

On a choisi le composant « Color Manager Filter » pour séparer les étudiants validés et non validés par différents couleurs.

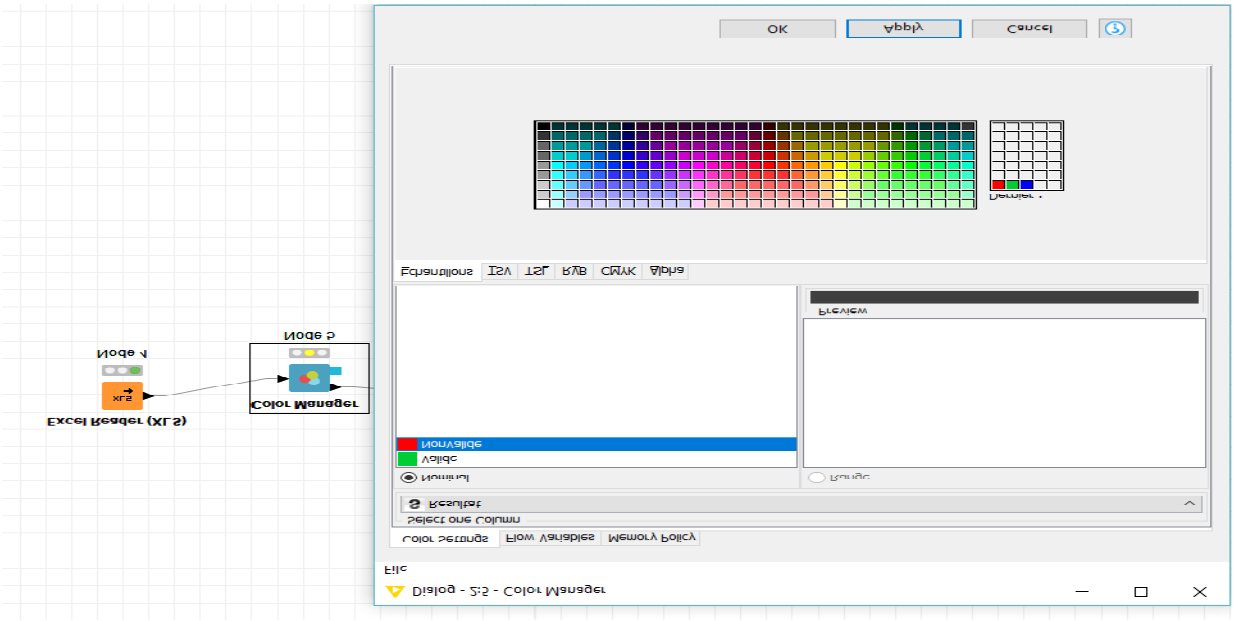


Figure 68: Le choix des couleurs

VI.7.3 Construction du classifieur – Arbre de décision

La variable cible « Résultat » est spécifiée lors du paramétrage de l'outil.

a. Résultats

Plusieurs fenêtres sont générées à l'issue des calculs. Dans l'onglet « VIEW : DecisionTree View », nous avons accès à l'arbre de décision. La représentation graphique n'est pas très lisible compte tenu de la taille de l'arbre.

b. Interprétation :

Par exemple les étudiants qui viennent de la région de Settat, inscrits à la faculté des Sciences et Techniques et qui sont réussies leur bac en 2011 avec la mention bien, ont validé le Module X.

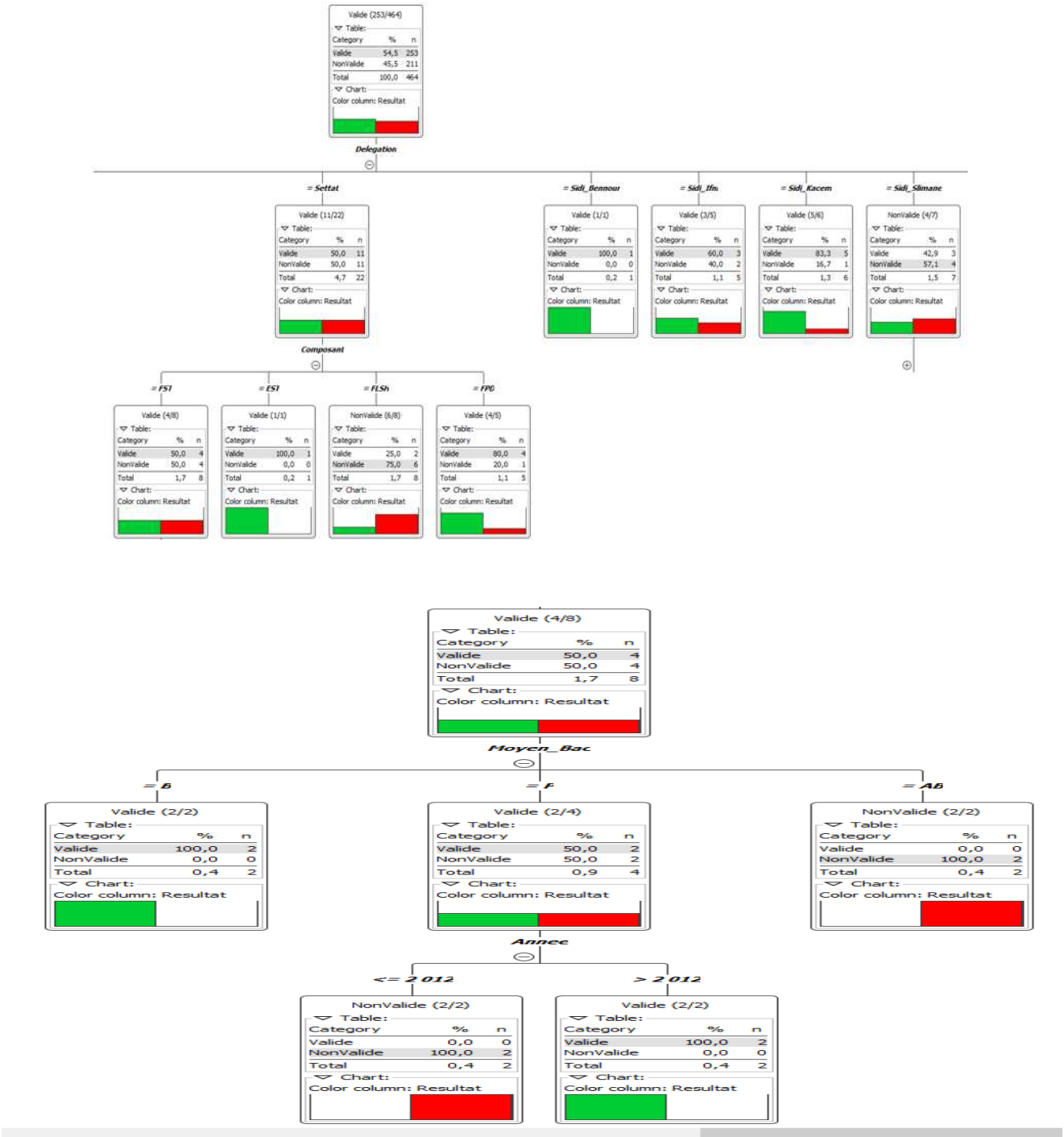


Figure 69: construction de l'arbre de décision

Nous préférerons l'affichage textuel sous la configuration de « DecisionTreeView».

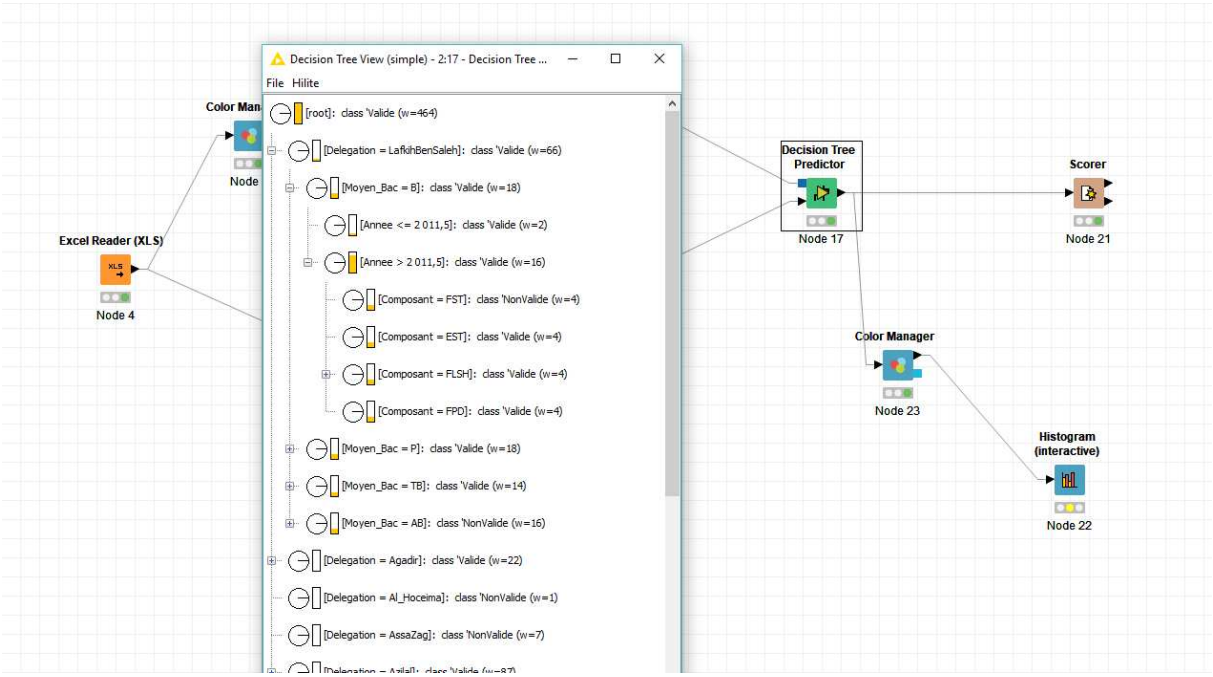


Figure 70: l'affichage textuel de l'arbre de décision

c. Résultats des étudiants validés par année

En cliquant sue la composante « Scorer » nous avons accès à la matrice de confusion et au taux d'erreur en validation croisée. Si nous appliquons le modèle sur notre cas on trouve les résultats présents en dessous.

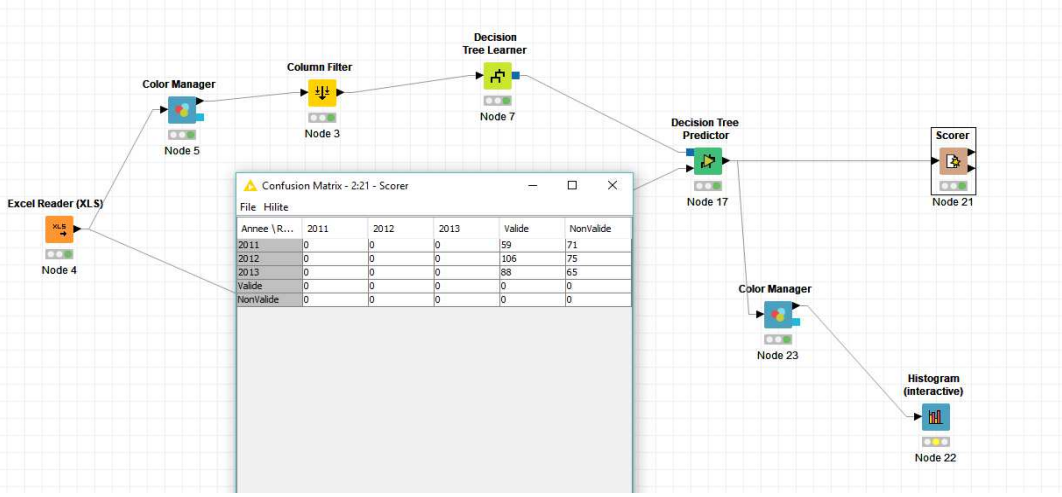


Figure 71: Résultats des étudiants par année

VI.8 Conclusion

Dans ce chapitre on a réalisé une application sous l'environnement de programmation JAVA, afin d'évaluer les algorithmes de datamining sur des datasets qu'on a extrait au niveau de l'entrepôt de données. Ainsi qu'on a présenté les statistiques qui décrit l'entrepôt de données à l'aide des graphes et diagrammes.

Les résultats préliminaires sont encourageants est nous laisse espérer différentes perspectives, à savoir l'application de cette approche sur d'autre acteurs de système universitaire, l'utilisation des différents ETL open source et la comparaison entre eux, et à développer encore plus se travail à travers d'autres méthodes d'analyse et de Datamining.

Pour la classification des étudiants j'ai travaillé avec deux outils différents Netbeans et KNIME afin de conclure que le logiciel KNIME est un bon outil pour la partiereporting.

Je trouve que ce dernier est plus simple d'utilisation et fait un effort méritoire sur les outils de management des données (préparation, recodage, etc.). Je note en tous les cas que nous, utilisateurs (chercheurs, étudiants, analystes, chargés d'études, etc.), sommes les principaux bénéficiaires de cette saine émulation qui pousse ces sociétés à mettre en ligne des versions gratuites de leurs logiciels.

Conclusion générale

L'évolution des activités au sein de l'université évolue de façon organique ainsi que l'augmentation rapide des étudiants exigent la nécessité de mettre en place un Système Décisionnel pour la gestion des étudiants

Pour réussir la mise en place d'un système décisionnel, il faut comme le cas de tout projet de système d'information procéder par une phase de Planification, ensuite une phase d'architecture, après une phase d'implémentation et enfin une phase de déploiement.

Dans la première étape de ce travail, on a présenté les constituants d'un système décisionnel à savoir les processus ETL, le Datawarehouse (stockage de données) et le Datamining (la restitution et l'analyse de données), et on a montré comment il est possible d'exploiter les bases de données existantes afin de concevoir un entrepôt de données qui intègre dès la conception les besoins des utilisateurs finals. Les bases métiers, orientées acteurs, permettent de suggérer des prises de décision. Puis on a comparé deux outils d'extraction de données à savoir le SSIS et Kettle, et on a conclu que les deux outils sont des solutions robustes pour effectuer l'ETL dans un entrepôt de données à trois niveaux.

SSIS met l'accent sur la configuration plutôt que sur le codage. Cependant, en raison de la quantité limitée d'objets de transformation disponibles, le codage sera nécessaire pour traiter des données complexes.

La force de SSIS vient de son flux de contrôle, de son flux de données et de son architecture pilotée par les événements. Cela permet une flexibilité pour le développeur de modifier le code source.

PentahoKettle propose plus d'objets de transformation plus simples. Il inclut beaucoup d'options pour accéder à des données externes telles qu'une interface SAP, Google Analytiqueset aux services Web. Il peut être utilisé sur Windows ou Linux

La force de Kettle provient de la possibilité d'utiliser des scripts Shell, JavaScript, définis par l'utilisateuret la possibilité de modifier le code source pour répondre aux besoins du projet. Le chargement est efficace avec de nombreuses options de chargement en masse pour les bases de données principales

Le travail accompli dans la seconde partie a porté sur la réalisation d'une application sous l'environnement de programmation JAVA, afin d'évaluer les algorithmes de datamining sur des données issues de l'application APOGEE Puis on à présenter les statistiques qui décrit l'entrepôt de données à l'aide des graphes et des diagrammes.

La conception et la réalisation d'un Datawarehouse comportaient en particulier l'analyse des besoins et la modélisation du Datawarehouse mais aussi le développement de l'alimentation et la création des états de reporting et des tableaux de bords. L'objectif était de développer un système d'information décisionnel relatif à l'environnement numérique de travail au
Sein de l'université publique.

Perspectives

Ce travail ouvre la voie vers différentes perspectives**,** comme nous venons de le voir, chaque domaine d'application de la fouille de données génère des problèmes spécifiques en raison de ses caractéristiques structurelles Les travaux de recherche sur la fouille de ces données complexes proposent des solutions intégrant au mieux ces caractéristiques. Nous avons contribué par la proposition de nouveaux algorithmes ou de nouvelles approches générales

Il reste encore des points à traiter à savoir :

Fusionner les projets de conception et réalisation des systèmes d'information de production (Opérationnel) avec les systèmes décisionnels, pour une meilleure intégration et une vision plus globale.

Définir des fourchettes (au point de vue Taille du Data Warehouse) pour décider à quelle niveau il faut normaliser ou dénormaliser.

Le travail présenté dans cette thèse peut être poursuivi dans plusieurs directions. Nous évaluons ces perspectives en considérant deux plans : celui de l'approfondissement et de la continuité de la recherche réalisée et celui de l'élargissement du domaine de recherche.

Le SID est un système critique pour la réactivité d'une organisation. Il serait donc pertinent d'évaluer la qualité des schémas produits à partir de critères définis en fonction des exigences de l'organisation (tactiques, stratégiques et systèmes) mais aussi en fonction de son environnement. Ces propositions permettraient de confronter les modèles des exigences à partir du poids des exigences pour les décideurs mais aussi à partir de paramètres externes qui contraignent la réactivité d'une organisation.

BIBLIOGRAPHIE ET WEBLIOGRAPHIE

[1] W.H. Inmon, “Building the Data Warehouse”, John Wiley & Sons, ISBN 0471-14161-5, 1994.

[2]G. El Helou, C. Abou Khalil. “Data mining: technique d‟extraction des connaissances ». Rapport de recherche. Université Panthéon-Assas, Paris 2, 2004.

[3] George Colliat, “OLAP, relational, and multidimensionaldatabasesystems”, ACM SIGMOD Record, vol.25(3), ACM Press, p. 64–69, septembre 1996.

[4] Kimball R., Ross M., "The Data Warehouse Toolkit", Wiley, New York,deuxièmeédition, 2002.

[5] Thiéry Odile, Cours ‘ ETL ‘ Les recherches avancées en systèmes d’informations stratégiques, Présentation PowerPoint, M2 MIAGE (Méthodes Informatiques Appliquées à la Gestion des Entreprises), PLG Nancy, 2007.

[6] Atkinson M.P., Bancilhon F., DeWitt D.J., Dittrich K.R., Maier D.,Zdonik S.B., ‘TheObject-Oriented Database System Manifesto’. In Proceedings of the International Conference on Deductive and Object-Oriented Databases - DOOD'89, Kyoto (Japan), December 1989.

[7] XML for Analysis (XMLA), SQL Server 2005 Books Online (Septembre 2007), Microsoft TechNet,
[En ligne],http://technet.microsoft.com/enus/library/ms187178.aspx.

[8]Villenga B., Van de Velde, Schreiber G., Akkermans H., ExpertiseBmodel définitiondocument, KADS Project document, UniversityBof Amsterdam, 1993.

[9]Widom J., “Research problems in data warehousing”, Dans International Conference on Information and Knowledge Management (CIKM95), Baltimore, Maryland, USA, p. 25-30, Novembre 1995.

[10]Inmon W. H., "Building the Data Warehouse". John Wiley& Sons, deuxième édition, ISBN 04771-14161-5, 1996.

[11]Olivier Teste, Modélisation et Manipulation d'Entrepôts de Données Complexes et Historisées, thèse de doctorat, Université Paul Sabatier, Toulouse 3 (France), décembre 2000.

[12]Ralph Kimball, The data warehouse toolkit: Practical Techniques for Building Dimensional Data Warehouses, John Wiley and Sons,ISBN : 0-471-153370, 1996, 2ème ed.: Ralph Kimball, Margaery Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, John Wiley & Sons, 2002.

[13]Michael Akinde, Michael H. Böhlen, DamianosChatziantoniou, Johann Gamperconstrained multi-dimensional aggregation
Information Systems, Volume 36, Issue 2, April 2011, Pages 341-358
[14]Ravat F., Teste O., "An Object Data Warehousing Approach: a Web Site Repository". Dans 4th East-European Conference on Advances in Databases and tel-Information Systems (ADBIS-DASFAA'00), Prague, Czech Republic, p. 128-137,Septembre 2000.
[15]Baril X., Bellahsène Z., "Selection of Materialized Views: A Cost-Based Approach". Lecture Notes in Computer Science, Springer-Verlag Heidelberg ISSN: 0302-9743 Computer Science, Vol. 2681, p. 665-680, 2003.
[16]Pedersen T.B., Jensen C.S., "Research Issues in Clinical Data Warehousing". Dans 10th International Conference on Scientific and Statistical Database Management (SSDBM'98), Capri, Italie, p.43-52, juillet 1998.
[17] Pedersen T.B., Jensen C.S. "Multidimensional Data Modeling for Complex Data". Dans 15th International Conference on Data Engineering (ICDE'99), Sydney, Australie, p. 363-345, mars1999
[18] Yang J., Widom J., "Temporel View Self-Maintenance in a warehousing Environment", Dans 7th International Conference on Extending Database Technology (EBDT'00), Konstanz, Allemagne, pp. 395-412, Mars 2000.
[19] Bellahsene Z., "Schema Evolution in Data Warehouses". Knowledge andInformation Systems, Springer-Verlag London Ltd, Vol. 4, N.3, p. 283-304, juillet 2002.
[20] Mendelzon A. O., Vaisman A.A., "Time in MultidimensionalDatabases". Dans "Multidimensional Databases: Problems and Solutions". Idea GroupInc., IGP/INFOSCI/IRM Press, Hershey, PA - USA, p 166-199, juin 2003
[21] Ravat F., Teste O., Zurfluh G., "Modélisation dimensionnelle des systems décisionnels". Dans Revue extraction des connaissances et apprentissage (ECA), Vol. 1, N. 1-2, p. 201-212, 2001.
[22]Surajit Chaudhuri, UmeshwarDayal, "An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD Record, vol.26(1), ACM Press, p. 65–74, mars 1997.
[23] Franck Ravat, Olivier Teste, Ronan Tournier, Gilles Zurfluh, "Algebraic and graphic languages for OLAP manipulations", Intl. Journal of Data Warehousing and Mining (ijDWM), Idea Group Publishing (IGP), juin 2007 (à paraître).

[24] LadjelBellatreche. « Utilisation des index et de la fragmentation dans la conception logique et physique d"un entrepôt de données » .Thèse pour obtenir le grade de Docteur en Informatique, université de Clermond Ferrand II, 2000.

[25] Moody D. L., Kortink M. A.R. "From Enterprise Models to Dimensional Models:A Methodology for Data Warehouse and Data Mart Design". Dans 2nd International Workshop on Design and Management of Data Warehouses (DMDW'2000) Stockholm, Suède, papier 5, juin 2000.

[26] M. Golfarelli, D. Maio, S. Rizzi, "The Dimensional Fact Model: A Conceptual Model for Data Warehouses ", International Journal of Cooperative Information Systems, 7(2-3), pp.215-247, 1998

[27] R Tournier, ≪Vers un langage de manipulation graphique des bases multidimensionnelles ≫, Memoire D.E.A. 2IL, Universite Paul Sabatier, Toulouse III, Juin 2004.

[28] J. Gray, A. Bosworth, A. Layman, H. Pirahesh, ≪Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total ≫, 12th International Conference on Data Engineering (ICDE'96), IEEE Computer Society, pp.152-159, New Orleans (Louisiana, USA), Mars 1996.

[29] Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., Paraboschi, S., "Designing data marts for data warehouses ", ACM Trans. Softw. Eng. Methodol., 10(4), pp.452‑483, 2001.

[30]Nguyen T. B., TJOA A. M., Wagner R.R. "An Object Oriented Multidimensional Data Model for OLAP". Dans 1st International Conference on Web-Age Information Management (WAIM'00), Shanghai, Chine, Springer LNCS 1846, 2000, p. 69-82, juin 2000

[31] Tsois A., Karayannidis N., Sellis T., "MAC: Conceptual Data Modeling for OLAP". Dans International Workshop on Design and Management of Data Warehouses (DMDW'2001) Interlaken, Switzerland, juin 2001.

[32] Gardarin G., "Bases de données", Eyrolles, Janvier 1999

[33] Prat N., Akoka J., "From UML to ROLAP multidimensional databasesusing a pivot model". Dans 18ème journées Bases de données avancées (BDA02), Evry, France, p. 171-195, octobre 2002.

[34] Vassiliadis P., "ModellingMultidimensionalDatabases Dans 10th International Conference on Scientific and StatisticalDatabase Management (SSDBM'98), Capri, Italie, p. 53-62, juillet 1998.

[35] Abelló A., Samos J., and Saltor F. "Implementing Operations to Navigate Semantic Star Schemas". Dans 6th International Workshop on Data Warehousing and OLAP (DOLAP'03). New Orleans, USA, ACM, p. 56-62, novembre 2003.

[36] Hahn K., Sapia C., Blaschka M., "Automatically Generating OLAP Schemata from Conceptual Graphical Models". Dans 3rd International Workshop on Data Warehousing and OLAP (DOLAP'00, in connection with CIKM), Washington, USA, novembre 2000, p. 9-16.

[37] Samtani S., Mohania M. K., Kumar V., Kambayashi Y., "Recent Advances and Research Problems in Data Warehousing". Dans Workshops on Data Warehousing and Data Mining, Mobile Data Access, and Collaborative Work Support andSpatio-Temporal Data Management (ER '98), Singapore, Springer LNCS 1552, p. 81-92, novembre 1998.

[38] Drug Safety, Principles of Data Mining, July 2007, Volume 30, Issue 7, pp 621–622

[39] Hand DJ, Manila H, Smyth P. Principles of data mining. Cambridge (MA): The MIT Press, 2001

[40] Hand DJ, Blunt G, Kelly MG, et al. Data mining for fun and profit. Stat Sci 2000; 15(2): 111–31

[41] Michael .J Berry, Gordon Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, 1997

[42] Freund, Y., &Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Proceedings of the Thirteenth international Conference on Machine Learning, pp. 148 - 156

[43] C.Elkan Boosting and naive Bayesian learning. Tech. rep. CS97-557, Department of Computer Science and Engineering University of California, San Diego

[44] Langley, P., Iba, W., & Thompson, K. (1992). an analysis of Bayesian classifiers. In Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 223 – 228

[45]Karel Dejaeger ; WouterVerbeke ; David Martens ; Bart Baesens,Data Mining Techniques for Software Effort Estimation: A Comparative Study, IEEE Transactions on Software Engineering (Volume: 38, Issue: 2, March-April 2012)

[46] GökhanYavaş, DimitriosKatsaros, ÖzgürUlusoy, YannisManolopoulos, A data mining approach for location prediction in mobile environments, Data & Knowledge Engineering, Volume 54, Issue 2, August 2005, Pages 121-146

[47] Rapha¨el Couturier∗, Régis Gras∗∗, traitement de données avec l'analyse implicative, ∗Ecole polytechnique de l'université de Nantes

[48] Thoo E, Friedman T, Beyer Mark A. Magic Quadrant for Data Integration Tools. Gartner RAS Core Research Note G. 2013.

[49] Pall AS, Khaira JS. A comparative review of extraction, transformation and loading tools. *Database Systems Journal BOARD*. 2013; 4(2): 42-51.

[50] Kimball R, Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publishing, Inc. 2004.

[51] Kabiri A, Wadjinny F, Chiadmi D. Towards a Framework for Conceptual Modelling of ETL Processes. Proceedings of The first international conference on Innovative Computing Technology (INCT 2011), Communications in Computer and Information Science. Heidelberg. 2011; 241: 146-160.

[52] Vassiliadis P, Simitsis A. Extraction-Transformation-loading, (ETL) Processes in Data Warehouse Environments. In Encyclopedia of Database Technologies and Applications, Idea Group. 2005.

[53] Jovanovic P, Theodorou V, Abelló A, Nakuçi E. Data generator for evaluating ETL process Quality (Science direct). In Press. 2016.

[54] Shaker H. AliEl-Sappagh, Abdeltawab M. AhmedHendawi,Ali HamedElBastawissy "A proposed model for data warehouse ETL processes"Journal of King Saud University - Computer and Information Sciences, Volume 23, Issue 2, July 2011, Pages 91-104

[55] Simitsis A, Vassiliadis P, Sellis T. State space optimization of ETL workflow. *IEEE Transactions on Knowledge and Data Engineering*. 2005; 17(10): 1404-1419.

[56] Jian L, Bihua X. *ETL tool research and implementation based on drilling data warehouse*. Seventh International Conference on Fuzzy Systems and Knowledge Discovery. 2010; 6: 2567-2569.

[57] Coutaud R, Jehl F, Harel P. *Editors*. SQL Server 2012 Integration Services (SSIS) Broché. France: ENI Édition. 2012.

[58] Zacklad M., Barbaud X.,Vers une application du Web Socio Sémantique pour la réalisation d'un système d'information destiné aux réseaux de santé**,** Second séminaire francophone du Web Sémantique Médical WSM'2004, Rouen, France, 9 mars 2004.

[59] 2012, 07/05/12). *SQL Server Integration Services*. Available: http://msdn.microsoft.com/en-us/library/ms141026.aspx

[60] J. Foley. (2008, Startup Of The Week: Pentaho Offers Opens Source BI Alternative. *InformationWeek*. Available: http://www.informationweek.com/news/206904925

[61] (2012, 07/05/2012). *Pentaho, Powerful Analytics Made Easy*. Available: http://www.pentaho.com/

[62] Christian VIGOUROUX , <<Pentaho - Mise en place d'une solution Open Source de Business Intelligence>>, 11 avril 2011

[63] Maria Carina Roldan, <<Learning Pentaho Data Integration 8 CE - Third Edition:
An end-to-end guide to exploring, transforming, and integrating your data across .multiple sources>>.

[64]María Carina Roldán, <<Pentaho Data Integration Quick Start Guide: Create ETL processes using Pentaho>>, 30 août 2018

[65]Manoj R Patil,FerisThia<< Pentaho for Big Data Analytics Kindle Edition>>
PacktPublishing, November 25, 2013.

[66] ZDnet.fr, Dossier - Thema Business Intelligence, Dossier source de nombreux articles et données, [En ligne], http://www.zdnet.fr/themas/business-intelligence/

[67] GhozziFaiza. CONCEPTION ET MANIPULATION DE BASES DE DONNEES DIMENSIONNELLES À CONTRAINTES. Informatique [cs]. Université Paul Sabatier - Toulouse III, 2004. Français.

[68] Atkinson M.P., Bancilhon F., DeWitt D.J., Dittrich K.R., Maier D.,Zdonik S.B., 'The Object-Oriented Database System Manifesto'. In Proceedings of the International Conference on Deductive and Object-Oriented Databases - DOOD'89, Kyoto (Japan), December 1989.

[69] Rashmi Chhabra,PayalPahwa, Data Mart Designing and Integration Approaches, IJCSMC, Vol. 3, Issue. 4, April 2014, pg.74 – 79

[70]. S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007

[71] Fayyad U. M., Piatetsky-Shapiro G., Smyth P. « From datamining to knowledge discovery: an overview », in Advances in knowledge discovery and datamining p. 1–34. (AAAI/MIT Press), 1996.

[72] Stéphane TUFFERY, 2007. Datamining et statistique décisionnelle :
L'intelligence des données.

[73] Georges El Helou et Charbel Abou Khalil Data Mining Techniques d'extraction des connaissances, 2004

[74] R.Agrawal, T.Imielinski, and A.Swami, 1993. "Mining association rulesbetween sets of items in large databases", in proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207-216

[75] Rakesh Agrawal, and Ramakrishnan Srikant, 1994. "Fast Algorithms for Mining Association Rules", In Proceedings of the 20th Int. Conf. Very Large Databases, pp. 487-499.

[75] M.J. Zaki, May/June 2000. “Scalable algorithms for association mining”.IEEE Transactions on Knowledge and Data Engineering, 12(3):372–390.

[76] Han, J., J. Pei, and Y. Yin, 2000. “Mining Frequent Patterns without Candidate Generation” in ACM SIGMOD Int'l Conference on Management of Data.

[77] Kaufman, L. and Rousseeuw, P.J. (1987), Clustering by means ofMedoids, in Statistical Data Analysis Based on the –Norm and RelatedMethods, edited by Y. Dodge, North-Holland, 405–416.

[78] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57

[79] Gustafson D.E., Kessel W.C. Fuzzy Clustering with a Fuzzy Covariance Matrix, IEEE Conference. On Adaptive Processes, (17) 1978, No. 1, 761-766

[80] Babuska R., Veen P.J. Improved covariance estimation for Gustafson-Kessel clustering. IEEE Conference. On Fuzzy Systems, (2)2002, 1081-1085

[81] Badrul M. Sarwar†∗, George Karypis‡, Joseph Konstan†, and John Riedl† Recommender Systems for Large-scale E-Commerce: ScalableNeighborhood Formation Using Clustering

[82] Rakesh Agrawal, and Ramakrishnan Srikant, 1994. “Fast Algorithms for Mining Association Rules”, In Proceedings of the 20th Int. Conf. Very Large Data

[83] Salvador, S., & Chan, P. (2004). Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation. Algorithms.Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, pp. 576-584°.