REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Génie Logiciel

Par :

CHELIK Nassima
OULD ALI Nadia

Sur le thème

Application des méthodes de data Mining dans le but d'analyser l'échec/succès des étudiants

---

Soutenu publiquement le / 07 / 2019 à Tiaret devant le jury composé de :

Mr CHIKHAOUI Ahmed	Dr.	Université de Tiaret	Président
Mr TALBI Omar	Dr.	Université de Tiaret	Encadreur
Mr OUARED Abdelkader	Dr.	Université de Tiaret	Co-Encadreur
Mr BENAOUDA Habib	MAA	Université de Tiaret	Examinateur

Dédicace

A mes chers parents qui m'ont donné un magnifique modèle de labeur et de persévérance que Dieu les Protège et puisse-t-il les rendre fiers de moi.

A Mes frères et sœurs qui m'ont beaucoup aidé et soutenu et qui ont toujours cru en moi.

A toute ma famille sans exceptions.

A tous mes amis(e) et collègues.

A tous ceux qui m'ont aidé.

A tout le personnel du département informatique.

Je dédie cet humble travail. ***Nadia***

Je dédie ce modeste travail à :

MES CHERS PARENTS Sources de mes joies, secrets de ma force.

Merci pour tous vos sacrifices pour que vos enfants grandissent et prospèrent, merci de trimer sans relâche, malgré les péripéties de la vie au bien être de vos enfants, merci d'être tout simplement mes parents, c'est à vous que je dois cette réussite, et je suis fière de vous l'offrir.

Mes chères sœurs et surtout Halima et Bouchra.

Mon cher frère Abdallah et son fils Abdelkader.

Mes amies en particulier Nadia, Imen, Khadidja et Messouda.

Nassima.

Remerciements

Nous remercions DIEU tout puissant de nous avoir donné la force, la santé, le courage et la patience de pouvoir accomplir ce travail.

Un grand merci à toutes nos familles surtout nos parents pour leur encouragement et leur suivi avec patience du déroulement de notre projet.

Nos remerciements s'étendent à nos encadreurs **Mr.Talebi Omar** et **Mr.Ouared Abdelkader** pour leur judicieux conseils, leur patience, leur disponibilité et surtout pour leur confiance qu'ils nous ont toujours témoignée.

Nos sincères remerciements s'adressent à **Mr Yacine** pour sa disponibilité et ces efforts qui nous ont aidés à réaliser notre travail, aussi à tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Nous tenons à remercier chacun des membres du jury pour nous avoir fait l'honneur d'examiner et dévaluer notre travail et de l'enrichir par leurs propositions.

Résumé

De nos jours, les nouvelles technologies de l'information permettent de concevoir des systèmes d'information et de la communication (NTIC) particulièrement performants et novateurs. Avec l'apparition des entrepôts de données, tous les utilisateurs peuvent désormais accéder facilement à l'information stratégique, ce qui permet aux décideurs d'être plus réactifs dans la prise de décision.

Le conseil pédagogique de notre département d'informatique procède aux délibérations chaque fin d'année. Ce conseil ne dispose que des notes des étudiants pour évaluer leurs parcours. Ceci reste insuffisant pour une meilleure compréhension des causes ayant conduit à de tels résultats. Le conseil pédagogique se retrouve ainsi devant des échecs et des réussites des étudiants qu'il n'arrive pas à prévoir, encore moins, à en expliquer les raisons. C'est dans cette optique que les acteurs de la formation aspirent à la mise en place d'un système qui leur procurerait les informations nécessaires et fiables, les aidant ainsi à prendre dans les meilleurs délais les décisions les plus appropriées.

Le but recherché dans ce PFE est d'aller vers la mise en place d'un système s'inscrivant dans le cadre de l'informatique décisionnelle et du Data mining appliqués aux données des étudiants qui suivent un cursus de formation LMD du domaine Mathématiques et Informatique (MI) au sein de notre département d'informatique de l'université de Tiaret. Ce système sera construit autour d'une base de données dédiée totalement aux décisionnel, un *Data Warehouse*. Ce dernier pourra répondre aux besoins d'analyse des décideurs par le bais des techniques de Data Mining afin de repérer les facteurs d'échecs et de réussites des étudiants

Mots-clés :
Entrepôts de données, OnLine Analytical Processing, Datamining, Prise de décision, Cause réussite/échec.

Abstract

Nowadays, new information technologies make it possible to design Information and communication systems (NTICs) particularly powerful and innovative.With the emergence of data warehouses, all users can now easily access strategic information, which allows decision-makers to be more responsive in deciding.

The educational council of our IT department conducts the deliberations each end of the year. This board only has student grades to evaluate their background. This remains insufficient for a better understanding of the causes that led to such results. The pedagogical council is thus faced with failures and successes of students that it can not predict, neither explain its reasons. It is in this perspective that the training actors aspire to the establishment of a system that would provide them with the necessary and reliable information, thus helping them to take the most appropriate decisions as soon as possible.

The goal sought in this PFE is to move towards the establishment of a system within the framework of business intelligence and data mining applied to the data of students who follow a course of LMD training in the field of Mathematics and Computer Science (MI) in our computer science department at the University of Tiaret. This system will be built around a database dedicated to decision-making, a Data Warehouse. The latter can meet the analysis needs of decision-makers through Data Mining techniques to identify the factors of failure and success of students.

Keywords :

Dat Warehouse, OnLine Analytical Processing, Datamining, Responsive in deciding, Factors failure/success.

Table de matière

Table des figures

Liste des tableaux

Introduction Générale

Introduction générale

De nos jours, la donnée est devenue l'un des principaux actifs de l'entreprise quel que soit son secteur d'activité. L'université ne fait pas exception à cette règle; en effet, les informations extraites des données issues des environnements d'apprentissage en ligne ou de systèmes d'informations de l'université constituent une clé de performance. Ces informations vont lui permettre d'améliorer la qualité de la formation et par voie de conséquence satisfaire aux besoins et exigences de son environnement socio-économique.

Cependant, cette masse d'informations gigantesque est distribuée sur des systèmes opérationnels. Ces derniers sont peu adaptés pour servir de support à la prise de décision. La question qui se pose est : Comment procéder à une intégration décisionnelle des données éparpillées et hétérogènes se trouvant dans des bases de données opérationnelles?
L'informatique décisionnelle fournit tous les moyens et les procédures possibles d'accès, de traitement, de transformation et d'exploitation des données au moment opportun.

C'est dans ce contexte que l'université se doit d'acquérir un système décisionnel qui puisse répondre à ces exigences, c'est le cas des outils de *Business Intelligence* (BI). Ceci va permettre aux décideurs de disposer d'informations pertinentes et d'outils d'analyse puissants pour les aider à prendre les bonnes décisions au bon moment. Destiné

Contexte de travail

Le travail mené dans le cadre de ce PFE relève du domaine de l'informatique décisionnelle et du *Data mining* appliqué aux données des étudiants qui suivent un cursus de formation LMD[1] du domaine Mathématiques et Informatique (MI) au sein de notre département d'informatique de l'université de Tiaret.

Problématique

Le département d'Informatique de l'université de Tiaret, comme tout autre département universitaire, a comme objectif une formation de qualité. Cette dernière est reflétée par l'obtention de bons résultats de la part des étudiants.

Au cours de l'année universitaire notre département, par le biais de ces équipes de formation pédagogiques, tient des réunions pour statuer sur l'évolution du parcours des étudiants. En fin d'année ces équipes pédagogiques procèdent aux délibérations.

Notre constat est qu'au fil des années les résultats obtenus par les étudiants s'avèrent insatisfaisants. Il nous parait donc nécessaire de procéder à des analyses sur toutes les données pertinentes relatives au cursus de formation des étudiants qui sans doute nous permettrons de

[1] LMD : Licence Master Doctorat

repérer : (1) les facteurs d'échecs des étudiants afin d'y remédier et (2) les facteurs de leur réussite afin de les multiplier. Comment y parvenir?

Cette question de recherche se décline en sous-questions :

- Quelle sont les types de données à prendre en compte dans l'analyse?
- Quel mécanisme adopter pour récupérer ces données?
- Quel serait la meilleure méthode à appliquer pour analyser ces données?
- Quelle serait la meilleure façon de présenter les résultats du processus de façon à offrir une vue synthétique et compréhensible par les utilisateurs?

Objectifs

Ce travail vise trois objectifs. Le premier est la mise en place d'un entrepôt de données afin de centraliser et d'historiser les données relatives au cursus des étudiants. Le second objectif est de nous permettre, à travers cet entrepôt de données, la production de rapports et de tableaux de bord. Ces outils de *Reporting* ne sont pas, à proprement parler, des instruments d'aide à la décision, mais, lorsqu'ils sont utilisés de manière appropriée, ils peuvent fournir une précieuse vue d'ensemble aux acteurs de la formation. Notre troisième objectif est d'appliquer sur cet entrepôt de données des techniques de *Data Mining* dans le but de faire de la prédiction et de la recommandation.

Nous répondrons à ces questions à différentes phases de notre conception du système final.

Méthodologie de travail

La Figure 1 montre la méthodologie suivie pour atteindre nos résultats. Nous suivons la méthodologie de conception pour mener des recherches comme décrit dans le génie logicielle et l'informatique [1]. Le cycle de conception comprend des étapes majeures (voir la figure 1): (1) sensibilisation de problème de recherche (2), fondements théoriques, (3) étude de l'existant, (4) conception de la solution, (5) mise en œuvre de la solution et (6) conclusion. Le cycle de conception commence par la sensibilisation du problème et des lacunes de l'existant. Le problème est formulé et accompagné par des questions de recherche. Les suggestions pour une solution à la problématique dans lequel notre recherche prend l'ancrage à partir de l'état de l'art afin de trouver la voie de notre contribution originale. Dans l'étude de la littérature, nous examinons la littérature des travaux en tenant compte des projets académiques. Dans la conception de la solution, nous proposons l'approche adoptée ainsi que notre modélisation. Dans la phase de mise en œuvre, nous démontrons la faisabilité et la praticabilité de notre proposition et les conclusions seront établies.

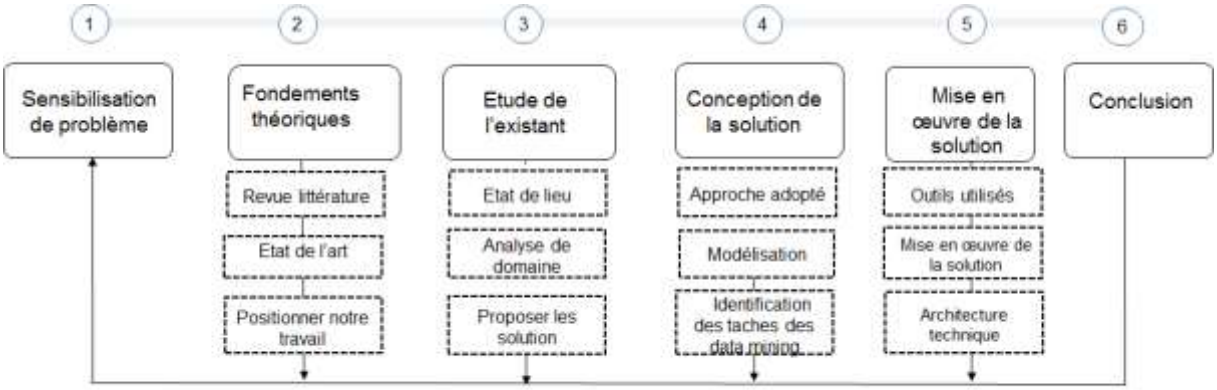


Figure 1 : Méthodologie adoptée dans notre mémoire.

Plan du mémoire

Notre mémoire est organisé en quatre chapitres (voir la figure2). Le premier chapitre est composé de trois sections. Nous commencerons ce chapitre par des définitions et concepts sur le BI, dans la deuxième section nous présenterons un état de l'art sur des travaux académiques et industriels similaires. Nous clorons ce chapitre par un tableau récapitulatif qui va nous permettre de positionner notre projet.

Le deuxième chapitre sera consacré à une étude détaillée de l'existant. Ce chapitre est composé de trois sections. Dans la première section nous dresserons un état des lieux du système actuel. Dans la deuxième section nous ferons une analyse du domaine d'étude afin de (1) dégager les besoins des acteurs de la formation et (2) faire faire ressortir certaines anomalies. Ceci nous permettra, dans la troisième section, de proposer notre solution future. Cette dernière comprendra un premier volet suggestions et un deuxième volet où nous donnerons une vision du système projeté. Nous expliquerons également comment nous y parviendrons.

Dans le troisième chapitre, nous détaillerons les différentes conceptions de notre solution. Ce chapitre est composé de deux sections. Dans la première section nous ferons la conception de l'entrepôt de données. Cette conception se décline en la conception de modèles multidimensionnels, la création de cubes *On Line Analytical Processing* (OLAP) et leurs représentations graphiques. Dans la deuxième section nous présenterons l'application des techniques de *Data Mining* sur cet entrepôt de données.

Dans le quatrième chapitre nous présentons la mise en œuvre de notre solution. Ce chapitre est composé de trois sections. Dans la première section nous décrivons les différents outils et technologies que nous allons utiliser. Dans la deuxième section nous présenterons la mise en œuvre de la solution. Dans la dernière section nous présenterons l'architecture technique, quelque capture de l'application et notre politique de sécurité.

Pour finir une conclusion générale et les perspectives relatives à notre travail sont présentées

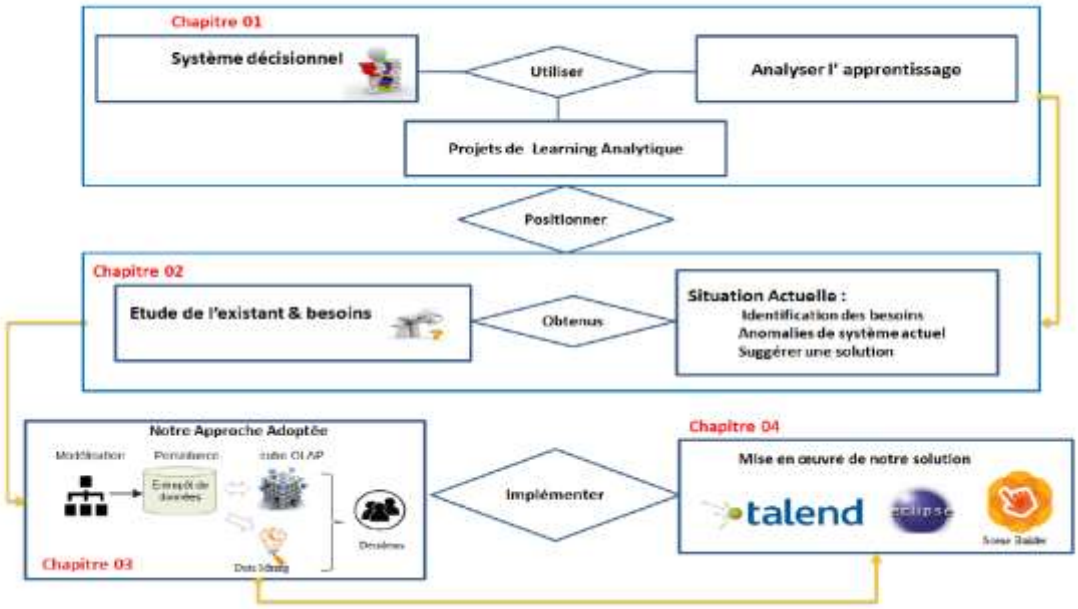


Figure 2 : Organisation de notre mémoire

Première partie

Synthèse bibliographique

&

Étude de l’existant

Chapitre 01 *:*
Généralités sur le système décisionnel

Introduction

L'informatique décisionnelle a pour objectif de créer de l'information à partir des données internes ou externes de l'entreprise, et aider les décideurs de l'entreprise dans leurs prises de décisions. De plus en plus la Business Intelligence est devenue l'une des préoccupations majeures au sein des directions des systèmes d'informations des grandes entreprises, En effet, dans un monde concurrentiel, la BI représente pour les entreprises une opportunité d'optimisation du pilotage de leurs activités, et d'anticipation des évolutions du marché en proposant des mécanismes d'aide à la décision.

Dans ce chapitre, nous allons aborder le concept de la Business Intelligence et des systèmes décisionnels puis nous présenterons un état de l'art sur des travaux académiques et industriels similaires. Nous clorons ce chapitre par un tableau récapitulatif qui va nous permettre de positionner notre projet.

1.1. Concept

Dans cette section, nous explorerons quelques concepts liés au système décisionnel.

1.1.1. Les systèmes d'informations décisionnelles

Les systèmes opérationnels génèrent quotidiennement, par les opérations classiques (insérer, modifier, supprimer, sélectionner), un grand nombre d'informations détaillées. Ces dernières ne sont pas exploitables à des fins d'analyse (erreurs de saisie, valeurs non homogènes, ressources différentes,…), De ces limites est apparu un nouveau système d'information conçu spécialement pour l'aide à la décision.

1.1.1.1. Définition

« *Le système d'information décisionnel (SID) est un ensemble de données organisées de façon spécifique (entrepôts de données), facilement accessible à la prise de décision, ou encore une représentation intelligente de ces données à travers d'outils spécialisés. La finalité d'un système décisionnel est le pilotage de l'entreprise* » [2].

1.1.1.2. Comparaison entre un système décisionnel et un système transactionnel

Le concept d'un entrepôt de données est apparu lors de l'existence de différence entre les systèmes transactionnels en ligne (OLTP) et les systèmes informationnels, dont certaines de ces différences fondamentales sont listées à travers le Tab1.1.Mais d'autres méthodes et

techniques de conception et d'implémentation d'ED ont vu le jour, l'une de ces techniques est le modèle dimensionnel de Kim Bail apparue en 1996 [3].
Le tableau suivant montre les différents points entre les systèmes transactionnels et décisionnels :

Systèmes transactionnels	Systèmes décisionnel
Les données sont courantes et orientées applications	Les données sont relatives à l'historique et orientées thème et sujets
Gros volumes de données à gérer.	Petits volumes de données à gérer.
Données en lecture seule.	Données en lecture - Écriture.
Nombre d'utilisateur restreint (décideurs, analystes).	Utilisé par toute l'entreprise.
Des requêtes relativement simples du point de vue informatique.	Les requêtes appliquées sont beaucoup plus complexes.
Temps de réponse très réduits.	Ils disposent de plus de temps pour les exécuter.
Les opérations sont des insertions et mises à jour courtes et rapides.	Le rafraîchissement des données est périodique et relativement long à exécuter.

Tableau 1 : Tableau comparatif entre le transactionnel et le décisionnel*[4] [5].

Ces différences font ressortir la nécessité de mettre en place un système répondant aux besoins décisionnels. Ce système n'est rien d'autre que le « ***Data Warehouse*** ».

1.1.2. Le Data Warehouse

Dans cette section, nous présentons la notion de data warehouse avec un schéma global de l'architecture présentés dans la figure 2.

1.1.2.1. Qu'est-ce qu'un Data Warehouse ?

Plusieurs définitions ont été données pour le concept d'entrepôt de donnée (ED). Nous retenons la définition de W.H. Inmon, considéré comme le père des ED qui décrit un ED dans son ouvrage de référence "*Building the Data Warehouse*" [6], comme "*une collection de données orientées sujet, intégrées, non volatiles et évolutif dans le temps, organisées pour supporter un processus d'aide à la décision*". Cette définition englobe les termes clés suivants [7] :

- **Orienté sujet**

Les données du *data warehouse* sont organisées par thèmes ou par sujets (par exemple : production, vente, marketing, etc.).

- **Intégré**

Le Data Warehouse va intégrer des données en provenance de différentes sources. Cette intégration permet d'éliminer l'ensemble des conflits (syntaxiques et sémantiques entre les données des sources) afin d'avoir une représentation uniforme et cohérente des données lors de leur chargement au niveau de l'ED. Selon W.H. Inmon [6] de toutes les caractéristiques d'un ED, l'intégration est l'un des aspects les plus importants.

- **Non volatiles**

Les données d'un ED sont généralement utilisées en mode consultation. Elles peuvent être interrogées mais ne sont ni modifiées, ni supprimées (sauf dans les cas de rafraichissement de l'ED). Ceci permet de conserver la traçabilité des informations afin de pouvoir effectuer des analyses sur une longue période.

- **Évolutif dans le temps**

La conservation de différentes valeurs d'une donnée est très importante dans un système décisionnel, cela permet les comparaisons et le suivi de l'évolution des valeurs dans le temps, tandis que dans un système opérationnel la valeur d'une donnée est simplement mise à jour.

- **Organisées pour supporter un processus d'aide à la décision**

Les données provenant des sources doivent être agrégées afin de faciliter leur analyse. Ces données peuvent être consultées à travers des outils (requêtes, outils OLAP, outils de fouille de données, outils de statistiques, etc.) permettant leur manipulation et leur analyse.

1.1.2.2. Architecture d'un entrepôt de données

Le processus de construction d'un ED, comme illustré dans la Figure suivante peut être structuré en cinq axes (Data sources, Back-and tier, Data warehouse tier, OLAP tier et Front-end tier). Ces composantes seront détaillées ci-dessous [8]:

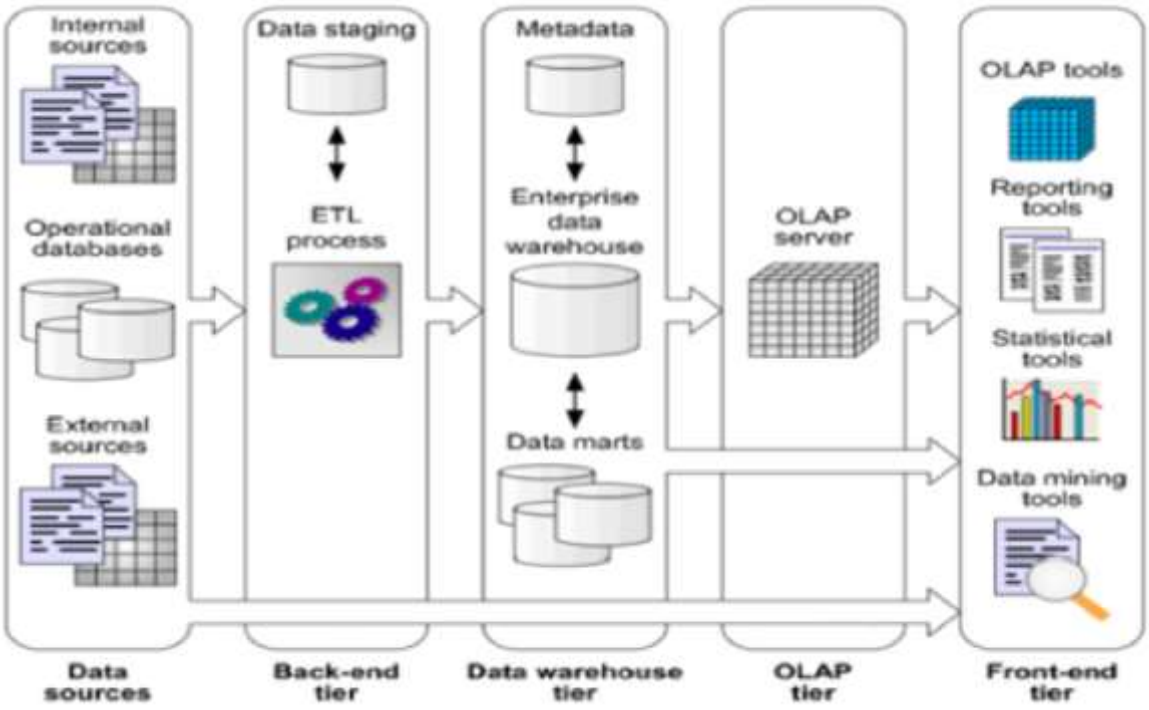


Figure 3 : Architecture d'un entrepôt de données [9].

- **Data sources**

L'ED stocke des données provenant de différentes sources d'informations hétérogènes et distribuées. Ces sources peuvent être des BD, des fichiers de données, des sources externes à l'entreprise, etc.

- **Back-end tier (niveau d'arrière-plan)**

Représente la zone de construction qui contient l'ensemble d'outils et techniques utilisés lors du processus de préparation des données, avant leurs chargements au niveau de l'entrepôt, qui sont : (1) processus ETL et (2) *data staging*.

1 processus ETL (Extraction, Transformation et Chargement)

Utilisé pour alimenter le *Data Warehouse* tier. Comme son nom l'indique, l'ETL comprend les activités suivantes:

- **L'extraction**

Est la première étape du processus d'apport de données à l'entrepôt. Elle consiste à recueillir des données hétérogènes provenant de multiples sources : base de données opérationnelles et des fichiers de différents formats; lire et les interpréter en vue de manipulation.

- **La transformation**

Une fois que les données sont extraites du système source, elles subissent une série de traitements destinée à les transformer en informations présentables pour les adapter à un usage décisionnel. Cette procédure comporte plusieurs aspects : le nettoyage, qui consiste à sélectionner et à épurer les données pour éliminer toute erreur et réconcilier les différences sémantiques entre ces données; l'intégration, qui concilie les données provenant de

différentes sources de données, au niveau schéma et données ; et l'agrégation, qui résume les données obtenues à partir de sources de données en fonction du niveau de détail, ou la granularité, de l'entrepôt de données.

- **Le chargement**

Alimente l'entrepôt de données avec les données transformées en respectant les contraintes du SGBD cible. Cela inclut également le rafraîchissement de l'entrepôt de données à savoir, la propagation des mises à jour dans l'entrepôt de données à partir des sources de données à une fréquence spécifiée afin de fournir des données à jour pour le processus de prise de décision.

2 Data staging

Le processus ETL nécessite généralement une zone de préparation de données (datastaging), comme illustré dans la figure ci-dessus. Le data staging représente le chantier de l'entrepôt de données. C'est là que les données sont chargées, nettoyées, combinées, archivées, puis rapidement exportées vers l'entrepôt. L'objectif de cette zone est l'obtention de données prêtes à être chargées sur un serveur de présentation (un moteur OLAP ou un SGBDR) [10].

- **Data warehouse tier (niveau entrepôt de données)**

Se compose d'un entrepôt de données d'entreprise et/ou de plusieurs magasins de données (data marts), et un catalogue de métadonnées stockant des informations sur l'entrepôt de données et son contenu.

L'entrepôt de données est un entrepôt centralisé qui englobe tous les domaines fonctionnels ou départementaux dans une organisation. Un ED peut comporter plusieurs magasins de données. Ces derniers sont des extraits de l'ED consacré à un type d'utilisateurs et répondant à un besoin spécifique. Ils sont dédiés aux analyses décisionnelles de type OLAP. Les métadonnées désignaient les données relatives aux données. Elles décrivent la signification des données et comment les données sont structurées et stockées dans un système informatique.

- **OLAP tier**

Un serveur OLAP permet d'accéder à l'entrepôt, il convertit les requêtes des clients en requêtes d'accès à l'ED et fournit des vues multidimensionnelles des données à des outils d'aide à la décision.

- **Front-end tier**

Est la partie publique de l'entrepôt de données. Il comporte les outils clients d'analyse et de visualisation des données. Les outils typiques sont les suivants (les outils OLAP, les outils de reporting, les outils statistiques, les outils de fouille de données) :

- **Les outils OLAP :** Permettent l'exploration et la manipulation des données de l'entrepôt afin de trouver des modèles ou des tendances importantes pour l'organisation. Ils facilitent la formulation de requêtes complexes qui peuvent

impliquer de grandes quantités de données. Ces requêtes sont appelées des requêtes ad-hoc.

- **Les outils de reporting** : consiste à collecter des données à partir de différentes sources et les restituer sous forme de rapport afin qu'elles soient prêtes à être analysées et que l'audience finale puisse à la fois voir et comprendre les données, et surtout prendre des décisions.

- **Les outils statistiques** : sont utilisés pour analyser les données du cube en utilisant des méthodes statistiques.

- **Les outils de fouille de données (data mining) :** permettent d'extraire des modèles d'une base de données historisée pour décrire le comportement actuel et/ou de prédire le comportement futur d'un procédé [10].

1.1.2.3. La modélisation multidimensionnelle

Les applications d'aide à la décision à base d'ED utilisent des processus d'analyse en ligne de données OLAP répondant à des besoins d'analyse de l'information. Pour ce type d'environnement OLAP, une nouvelle approche de modélisation a été proposée : la modélisation multidimensionnelle. Popularisée par Ralph Kimball dans les années 90, cette modélisation est aujourd'hui reconnue comme la modélisation la plus appropriée aux besoins d'analyse et de prise de décision [11]. Un modèle multidimensionnel renferme les deux concepts fondamentaux de fait et de dimension.

- Un fait représente le sujet d'analyse. Il est composé d'un ensemble de mesure qui représente les déférentes valeurs de l'activité analysée. Où les mesures doivent être valorisées de manière continue et elles peuvent être additives (additionnable suivant toutes les dimensions); semi-additives (additionnable suivant certaines dimensions) et non additives (fait non additionnable quel que soit la dimension). Ainsi elle se constitue d'un ensemble de clés qui sont des clés étrangères associées aux dimensions [7].

- Une dimension représente un contexte d'analyse d'un fait. Concept essentiel des bases de données multidimensionnelles, la dimension est le critère suivant lequel on souhaite évaluer, quantifier et qualifier le fait, elle peut être utilisée pour la sélection de données selon le niveau de précision désiré. Aussi, elle peut être affinée, décomposée en hiérarchies, afin de permettre à l'utilisateur d'examiner ses indicateurs à différents niveaux de détail.

1.1.2.4. Le concept OLAP

Les données opérationnelles constituent la source principale d'un système d'information décisionnel. Les systèmes décisionnels complets reposent sur la technologie OLAP, conçue pour répondre aux besoins d'analyse des applications de gestion.

1.1.2.4.1. L'implémentation d'un cube multidimensionnel [12]

Il existe deux manières principales pour construire un système basé sur un modèle multidimensionnel, selon la manière dont le cube est stocké : l'approche MOLAP et l'approche ROLAP.

Remarque

Le Datawarehouse est le cœur, l'ossature du système d'information décisionnel. Évidemment Data Warehouse et Data Mining sont deux choses très différentes. Data Warehouse est usuellement le point le départ de Data Mining. Data Warehouse et Data Mining sont des parties du processus Extraction de Connaissance à partir de Données (ECD).

1.1.3. Data Mining

Dans cette partie, nous présentons la définition de Data Mining puis nous expliquons le processus Extraction de Connaissance à partir de Données (ECD), l'objectif de datamining et nous concluons ses taches et techniques.

1.1.3.1. Présentation du Data Mining [13]

Le terme de Data Mining signifie littéralement forage de données. Comme dans tout forage, son but est d'extraire des connaissances à partir de données. Les données peuvent être stockées dans des entrepôts « *data warehouse* », dans des bases de données distribuées ou sur Internet. Il est apparu au début des années 90. Celui-ci n'est pas le fruit du hasard mais le résultat de la combinaison de nombreux facteurs à la fois technologiques, économiques et même sociopolitiques. Le data mining peut être vu comme une nécessité imposée par le besoin des entreprises de valoriser les données qu'elles accumulent dans leurs bases.
L'ECD, par le biais du *data mining*, est alors vue comme une ingénierie pour extraire des connaissances à partir de données.

1.1.3.2. Processus ECD

Extraction de Connaissance à partir de Données (ECD) désigne tout le cycle de découverte d'informations ou de connaissances dans les bases de données. Il regroupe donc toutes les opérations à effectuer pour extraire de l'information de ces données. L'ECD est un processus complexe qui se déroule suivant une suite d'opérations. Peut être vu comme un processus en six étapes [14] :

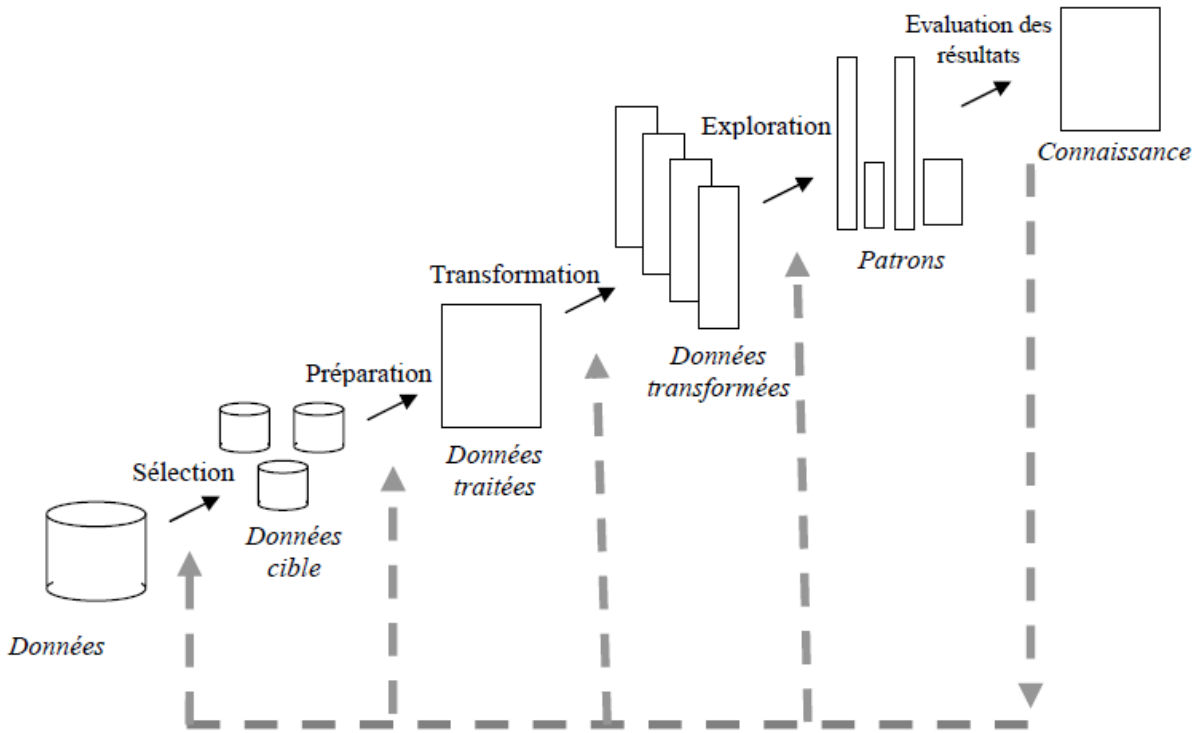


Figure 4 : Processus ECD [15].

- **La sélection**

Consiste à choisir les données qui sont en accord avec les objectifs fixés. Généralement cette étape demande souvent la présence d'un expert qui épure les données pour ne garder que celles qui décrivent au mieux la problématique à résoudre (sélectionner les données pertinentes).

- **La préparation**

Consiste à nettoyer les données avant de les utiliser et cela en éliminant les bruits, en traitant des données manquantes et en vérifiant la cohérence des données.

- **La transformation**

Elle transforme et prépare les données pour la phase suivante qui est la fouille de données. Cette transformation se fait en regroupant des données, en en les normalisant et en les transformant d'un format à un autre.

- **L'exploration**

Il s'agit de la phase la plus souvent décrite Sous le terme de data mining, c'est le cœur du processus ECD, elle donne naissance à un modèle qui se doit d'être : rapide à créer et à utiliser, compréhensible par l'utilisateur, fiable et évolutif. Il existe plusieurs méthodes de fouille de données (Règles d'associations, arbre de décision, réseaux de neurones …).

- **L'évaluation des résultats**

C'est la dernière phase du processus ECD. Elle consiste à évaluer les modèles générés dans la phase précédente. Elle se fait grâce à une série de tests sur un échantillon prévu pour ça. Si les résultats retournés ne sont pas satisfaisants, il faut réitérer le processus.

- **La représentation de la connaissance**

La connaissance découverte par le processus ECD est représenté à l'utilisateur (client) à travers des techniques de visualisation et de représentation comme : les tables, des arbres, des graphique, des courbes, des règles, etc.

1.1.3.3. Les objectifs du datamining [16]:

Nous évoquons dans cette section, trois intérêts du data mining :

- **Expliquer**

Le datamining pourra tenter d'expliquer un événement ou un incident indiscernable. Par la consultation des informations contenues dans l'entrepôt de données de l'organisation.

- **Confirmer**

Le datamining aidera à confirmer un comportement ou une hypothèse. Dans le cas où le décisionnaire aurait un doute concernant une hypothèse, le datamining pourra tenter de confirmer cette hypothèse en la vérifiant en appliquant des méthodes statistiques ou d'intelligence artificielle.

- **Explorer**

Enfin, le datamining peut explorer les données pour découvrir un lien «inconnu». Quand le décisionnaire n'as pas d'hypothèse ou d'idée sur un fait précis, il peut demander au système de proposer des associations ou des corrélations qui pourront aboutir à une explication.

1.1.3.4. Tâches réalisées en Data Mining [17] [18]:

Le data mining est un ensemble de techniques complémentaires dédiées à différentes tâches. Ces technique sont partagées, principalement, entre la classification automatique (supervisée et non supervisée) et la recherche d'associations.

1.1.3.5. Techniques du Data Mining:

Les techniques de data mining diffèrent en fonction des besoins de l'utilisateur, entre autres la tâche à effectuer. Chacune des tâches citées ci-dessus regroupe une multitude d'algorithmes pour construire le modèle auquel elle est associée. Selon [19], les dix algorithmes les plus populaires dans le domaine de data mining sont, dans l'ordre: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, k-NN, NaiveBayes et CART.

1.2. État de l'art

D'après la section précédente, nous avons récencé que les technologies de l'informatique décisionnelle fournissant tous les moyens et les procédures possibles d'accès, de traitement, de transformation et d'exploitation des données au moment voulu. Mais reste à savoir

maintenant s'il existe des travaux qui touchent notre problématique ou bien d'inspirer leurs idées.

Au fil de notre recherche nous avons constaté que les projets se catégorisent en deux sortes : (1) les projets académique, (2) les projets industriel. Dans notre situation la deuxième sorte de projets ne nous s'intéresse pas car ils nous s'emblait comme des boite noirs. Etant donné que notre projet est dans un cadre d'étude supérieure et de la recherche scientifique nous somme intéressé par les projets académique mais ces projets relie au *Learning Analytic*(LA).

1.2.1. Présentation de *Learning Analytics*

Siemens définit les *Learning analytics* comme « la mesure, la collection, l'analyse et l'interprétation des traces des apprenants et de leurs contextes, pour comprendre et optimiser l'apprentissage et les environnements dans lesquels il se produit. » [20].

1.2.1.1. Les *LEARNING ANALYTICS* pour quoi faire ?

Le schéma de Gartner[2] (La figure 5) permet de synthétiser ce que font les Learning Analytics en quatre catégories [21] :

- **Les *Analytics* descriptifs**, qui répondent à la question « que c'est- il passé ? ».

Ainsi, à partir de l'analyse des traces disponibles, l'objectif peut être de connaître le positionnement relatif d'un étudiant par rapport à ses pairs, en fonction de critères prédéfinis.

- **Les *Analytics* de diagnostic**, qui répondent à la question « pourquoi est-ce arrivé ? ». Il s'agit de déterminer des facteurs explicatifs des comportements observés, par exemple quels facteurs peuvent expliquer l'échec à un module d'un étudiant.

- **Les *Analytics* prédictifs**, qui répondent à la question « que va-t-il se passer ? ». Il s'agit d'anticiper le futur, en considérant ce que l'on connaît du passé. C'est par exemple le cas de la prédiction de l'échec d'un étudiant à un module.

- **Les *Analytics* prescriptifs**, qui répondent à la question « que faut- il faire pour que cela se produise ? ». Un exemple est celui des systèmes de recommandation qui fournissent des préconisations sur les ressources à consulter, les actions à entreprendre ou les tâches à accomplir, afin d'atteindre un objectif d'apprentissage prédéfini ou estimé.

[2] **Gartner Group** : est une société mondiale de recherche et de conseil fournissant des informations, des conseils et des outils aux leaders des secteurs de l'informatique, de la finance, des ressources humaines

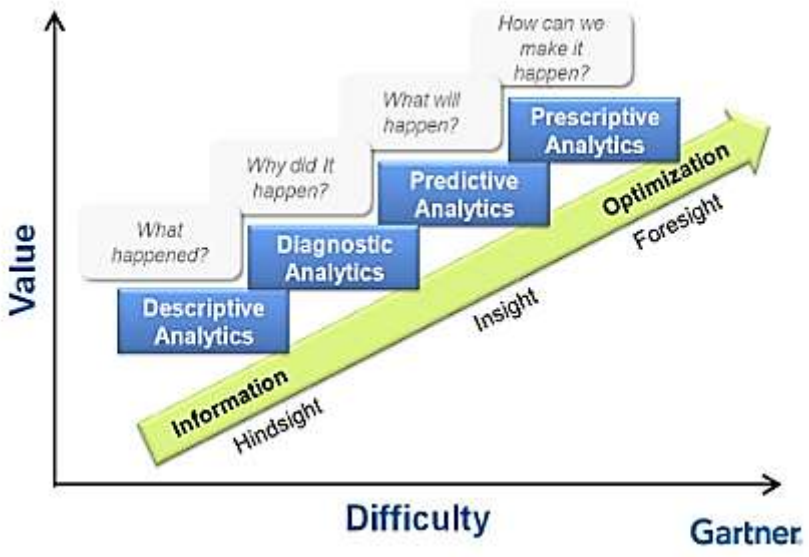


Figure 5 : Le cadran magique de Gartner.

1.2.2. Émergence de communautés scientifiques Internationales

Au cœur de la recherche sur les Learning Analytics, deux communautés développent des méthodes, des approches et objectifs sensiblement différents. Ces deux communautés sont :

- **2007 : IEDMS** (International Educational Data Mining Society)
- **2011 : SoLAR** (Society for Learning Analytics Research).

1.2.2.1. Différence entre EDM et SoLAR [22]

Pour les membres d'EDM, la conception d'algorithmes vise ainsi à donner au logiciel la capacité de prédire les résultats d'un apprenant et de personnaliser sa stratégie d'apprentissage.

Dans le sillon de SoLAR en revanche, modélisation et visualisation des données sont transmises aux acteurs de l'apprentissage (apprenant, personnels d'éducation, enseignants, etc.). Dans le premier cas, il s'agit donc « de réduire progressivement le système d'apprentissage à ses composantes principales, en modélisant séparément les apprenants, les tuteurs, les domaines enseignés », tandis que les chercheurs de la communauté SoLAR privilégieraient une approche systémique qui viserait à rendre les acteurs plus autonomes.

En d'autres termes, concluent les auteurs : les résultats des recherches se trouvent réinvestis dans la machine, tandis que ceux de SoLAR sont restitués aux acteurs, notamment sous la forme de visualisation, de tableaux de bord, etc.

1.2.2.2. Caractéristique

Caractéristiques	EDM	SoLAR
Méthode privilégiée	Fouille de données automatique.	Outil d'aide à la décision: découverte des données par les acteurs *via* des outils d'analyse et de visualisation
Approche	Réduire les systèmes d'apprentissage à ses composantes principales. Modélisation séparée des apprenants, des tuteurs, du domaine enseigné. Repose sur les théories du DM.	Systémique: appréhender la situation pédagogique comme un ensemble
Objectif	Concevoir des systèmes automatiques	Donner de l'autonomie aux acteurs (enseignants, apprenants…)

Tableau 2 : Caractéristique d'EDM et SoLAR.

1.2.2.3. Alors pourquoi cette émergence ?

Il existe quatre facteurs [23]:

- **Le volume des données**

Mis à disposition des chercheurs a très rapidement progressé par la multiplication des dispositifs d'apprentissage en ligne (LMS[3] pour *Learning Management System*, MOOCs[4] pour *Massive Open Online Courses*) permettant de capter les interactions des utilisateurs ; soit par croisement avec des données académiques.

- **Données structurés**

Les travaux se succèdent en termes d'interopérabilité des données venant de plusieurs environnements d'apprentissage.

- **Capacités de calcul**

Des smartphones d'aujourd'hui dé passent celles des ordinateurs d'il y a 10 ans.

[3] **LMS** : En technologies de l'information et de la communication, un learning management system ou learning support system est un logiciel qui accompagne et gère un processus d'apprentissage ou un parcours pédagogique.
[4] **MOOCs** : est un type ouvert de formation à distance capable d'accueillir un grand nombre de participants.

- **Nouveau outil d'analyse sans programmation**

Enfin de nouveaux frameworks comme Apache Hadoop permettent de gérer des données à la mesure du web ; et de nombreux outils d'analyse, adaptés de la *Business intelligence* à l'éducation, permettent de mener des recherches sans être nécessairement avancé en programmation ou en sciences statistiques. Rapid Miner, KEEL, Weka incluent ce type d'algorithmes.

1.2.3. Travaux de recherche des deux Communautés

Dans ce qui suit, nous sélectionnant les travaux de recherches similaires qui touchent notre objectif de près ou de loin à savoir leurs objectifs, les techniques utilisés et les résultats obtenues. Dans les deux communautés EDM et SoLAR, des Modèles prédictifs et des méthodes variées sont mises en œuvre.

1.2.3.1. Prédire la progression de l'apprenant [23]

La prédiction du parcours (comportemental ou cognitif) d'un apprenant est l'un des plus anciens problèmes des Environnements Informatiques pour l'Apprentissage Humain (EIAH), des systèmes experts aux MOOCs. Avec l'acquisition en temps réel de données d'apprentissage toujours plus fines et massives, les chercheurs ont développé des techniques d'analyse issues du *data mining* afin de :

- classer des profils, les orienter en fonction des capacités estimées.
- adapter les contenus.
- déployer des stratégies d'engagement.
- Lutter contre le décrochage.

Ces analyses sont menées à différents niveaux de granularité, de la simple interaction en temps réel d'un individu (microgenèse) aux apprentissages d'une cohorte sur une période donnée.

1.2.3.2. Donner à voir l'apprentissage [24]

A l'origine des Learning Analytics, les Visual Data Analytics (Analyse de données visuelles) reposent sur deux principes essentiels : donner aux acteurs de la communauté éducative un pouvoir de décision et permettre aux apprenants de devenir acteur de leur apprentissage.

Les Tableaux de bord numériques (*Dashboards*) Offrent aux humains une interprétation visuelle de larges ensembles de données pour, en quatre étapes, découvrir, interroger, comprendre les modèles portés par ces données et, éventuellement modifier ses représentations.

1.3. Synthèse

Dans cette section nous voulons présenter un tableau récapitulatif des travaux similaire à savoir leur objectifs, le data set et les techniques utilisées.

Référence	Objectif	Données (data set)	Techniques utilisées
ZIMMERMANN *et al.* 2015	Evaluer la puissance de prédiction des résultats scolaires et leur agrégation, comme indicateurs de performance. Il a pris en compte 81 variables pour une population de 171 étudiants	Sexe, âge, notes des 3 années de licence, notes de master, nombres de crédits validés par an, temps pour obtenir une UE…etc.	A posteriori. Modèle de régression
KNOWLES 2015	Dans 1000 écoles du Wisconsin, un système évalue la probabilité de passage pour chacun des 225 000 collégiens	Cohorte des 12-13 ans en 2005. Les variables portent sur les résultats, l'assiduité, le comportement, la mobilité entre écoles,	Le résultat de ce système est de fournir une probabilité de passage pour chaque apprenant avec un classement (bas, modéré, haut). Les étudiants reçoivent une catégorie de risque pour 4 sous domaines : scolarité, présence, comportement, mobilité. Le système identifie 65% des échecs et des retards dans l'avancement avant l'entrée au lycée (*high school*) avec de faibles taux de fausses alarmes
Broisinet al. 2017	Permettre à l'étudiant d'analyser ses activités en temps réel mais aussi *a posteriori.* fournir un tutorat intelligent	Plateforme Lab4CE: environnement web pour des Télé-TP Toutes les interactions sont tracées: action soumise, correction, connexion	Visualisation d'indicateurs**.**
VAN LEEUWEN 2015	Assister en temps réel les enseignants dans l'évaluation d'apprenants engagés sur une tâche collaborative	Evaluation de 5 groupes par 14 enseignants	Cette étude permet de comprendre comment les enseignants interagissent avec des visualisations de données d'apprentissage

Tableau 3 : Aperçu sur les projets existants[5]

[5] - LABARTHE, Hugues et LUENGO, Vanda. L'analytique des apprentissages numériques. 2016. Thèse de doctorat. LIP6-Laboratoire d'Informatique de Paris 6.

1.3.1. Positionnement de notre travail

Nous avons recensé quelques solutions qui touchent de près ou de loin notre thématique, cette analyse de ces projets industriels et académiques ça nous permet de tirer les fonctionnalités et l'architecture derrière la solution. Cependant que ces solutions souvent sont des boites noires (*blackbox*) avec la difficulté de la personnalisation et la modification, ainsi ces projets sont souvent contextuels.

Dans notre travail on s'intéresse à la mise place de genre de projet dans notre faculté des mathématique et de de l'informatique de l'UIK sachant que l'utilisation des projets existants n'est pas faisable et nécessite de ressortir avec justification les points clés de la réussite de ce type de projet dans notre faculté. Les points qui vont contribuer à la réussite de ce type de projet sont liés à diverses dimensions :

La motivation des décideurs concernés par ce type de système (par ex. travail de sensibilisation, questionnaire, sondage). Alquiet [2001] a recensé quelques erreurs à éviter pour mettre en place un système décisionnel comme :

- Résistance aux changements
- Habitude de travail
- Perte de pouvoir
- Organisation hiérarchique
- Difficulté d'intégration
- Une personne non orientée vers la technologie

L'identification des sources de données d'apprentissage liées à notre système (par ex. LMS Moodle, BD de scolarité).

Le choix de l'infrastructure et l'architecture déploiement qui peuvent être utilisée dans notre faculté (architecture distribuée, parallèle, centralisée, Cloud etc.).

Conclusion

Le travail d'ingénierie des besoins dans le tel système informatique est le cœur de développement logiciel. De ce fait dans le chapitre suivant nous allons aborder un état de lieu qui décrit la situation actuelle.

Chapitre 02 :
Etude de l'existant et de besoins

Introduction

Avant d'entamer la phase de collecte des besoins et de réflexion à une éventuelle solution aux problématiques posées, il convient d'étudier d'abord le système existant. Ceci permettra de mieux le comprendre et d'en déceler les limites, ainsi que les motivations.

De ce fait, nous avons choisi de structurer ce chapitre de la manière suivante : nous commençons par dresser un état des lieux du système actuel. Puis, nous allons effectuer une analyse du domaine d'étude afin de dégager les besoins des acteurs de la formation et faire ressortir certaines anomalies avant de conclure par la solution future. Cette dernière comprendra un premier volet suggestions et un deuxième volet où nous donnerons une vision du système projeté.

2.1. Etat des lieux

Notre département d'informatique dispose d'un système informatique dont les principales fonctionnalités sont les suivantes :

- La gestion des inscriptions des étudiants.
- La gestion des notes (résultats des unités, résultats semestriels).
- L'établissement des différents documents (relevé de notes, attestation d'inscription, statistiques …).
- La gestion des diplômes.

Ce système bénéficie également des services de la plateforme MOODLE de l'université, qui permet le dépôt des rapports, le téléchargement des cours, la proposition des sujets de PFEs (Projet de Fin d'Etude) et leur validation.

Le nombre d'étudiants que compte le département d'informatique est assez élevé. Les deux figures ci-dessous donnent un aperçu sur l'effectif des étudiants inscrits au cours de l'année 2016-2017.

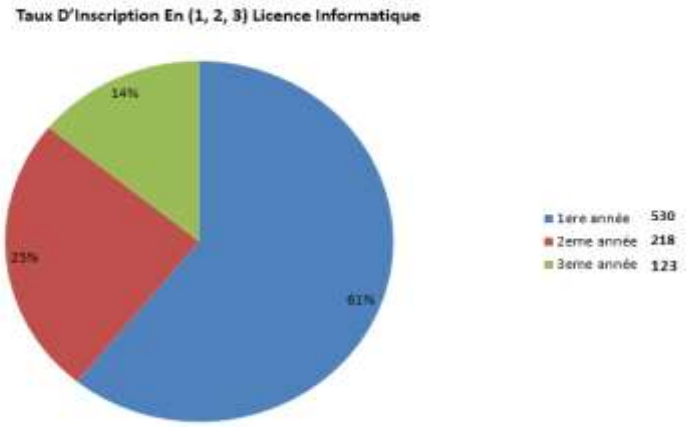


Figure 6 : Statistiques finales de la Faculté des Mathématiques et de L'Informatique en 2016(Licence).

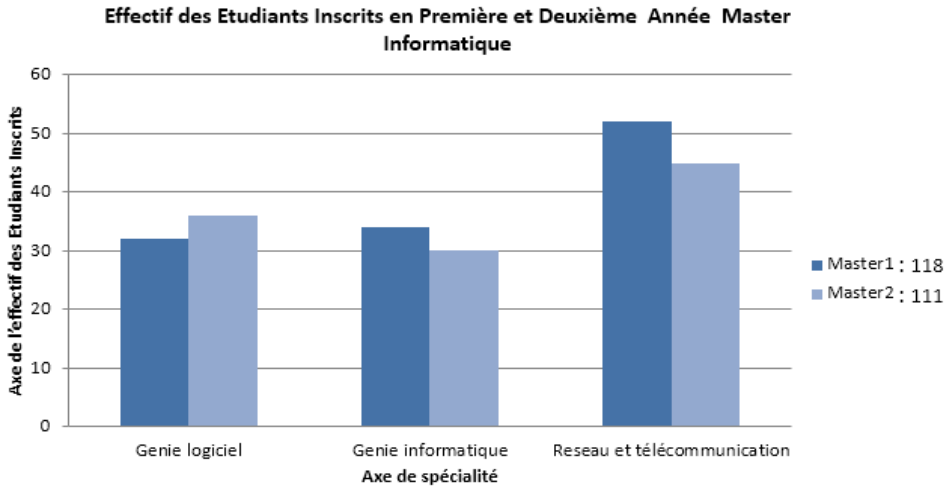


Figure 7 : Statistiques finales de la Faculté des Mathématiques et de L'Informatique en 2016(Master).

2.2. Domaine d'étude

Nous détaillons dans cette section l'analyse du domaine afin de ressortir les besoins des acteurs de la formation ainsi que les anomalies constatés.

2.2.1. Analyse de domaine

Les chances de succès d'un projet se trouvent considérablement accrues par la bonne compréhension des besoins des utilisateurs). Recueillir le besoin des utilisateurs est une tâche délicate. Nous commençons tout d'abord par identifier les acteurs du système. Nous collectons par la suite leurs besoins lors de nos rencontres. Nous devons leur parler de ce qu'ils font, des raisons pour lesquelles ils le font, de la manière dont ils prennent leurs décisions et comment espèrent-ils prendre leurs décisions à l'avenir. A cet effet deux techniques sont utilisés : (1) les entretiens (2) les réunions.

2.2.1.1. Présentation des acteurs du système actuel

Les utilisateurs finaux de notre de future système sont les décideurs chargés par la qualité de la formation de l'enseignement supérieur (tuteur, etc.). La figure suivante montre les différents acteurs.

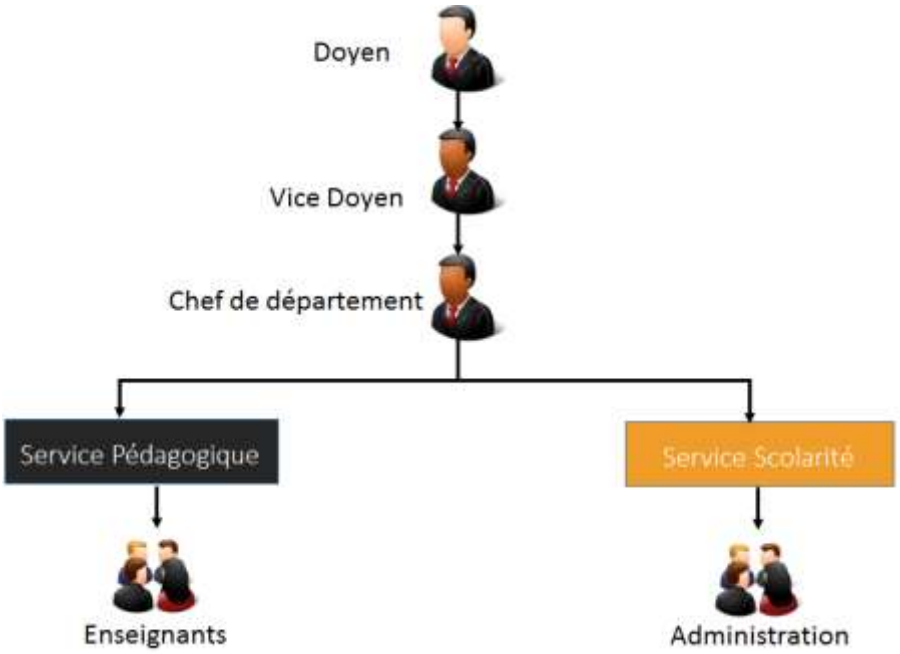


Figure 8 : Acteurs du système actuel.

2.2.1.2. Expression des besoins par les acteurs de da formation

Nos investigations auprès du service de scolarité, de l'équipe pédagogique et du chef de département nous ont permis de faire ressortir les besoins suivants :

- **B1** : Quels sont les facteurs liés à l'échec et la réussite des étudiants du département?
- **B2** : Pourquoi et comment le taux de réussite a baissé ?
- **B3** : Est-il possible d'analyser les données accumulées pendant des années pour extraire des connaissances cachées liées au contexte de notre étude ?
- **B4** : Comment unifier et centraliser les données pour avoir une analyse en ligne ?

Figure 9 : Illustration des problèmes rencontrés durant la réunion.

2.2.1.3. Liste des anomalies constatées

Bien qu'il soit impossible de dresser la liste de toutes les difficultés et les anomalies relevant du système actuel, nous allons essayer de citer celles que nous avons constatées sur le terrain de notre étude (voir la figure10). A partir de ces dernières nous ferons ressortir les corrections à apporter.

Les problèmes soulevés dans l'étude du domaine sont d'ordre informationnel ou organisationnel. Ceux d'ordre informationnel sont essentiellement les principaux. Nous citerons.

1. Ordre informationnel

- **Anomalie N°1 : Les données de scolarité ne sont pas historisées:**

Causes	Le traitement est orienté transactionnel
	Absence de réflexion sur l'aspect décisionnel
Conséquences	L'analyse de donnée est partielle
	Pas de prise de décision à partir de données de scolarité.

Tableau 4 : Anomalie 01

- **Anomalie N°2 : Difficulté d'élaborer des rapports statistiques**

Causes	Les données sont distribuées
	La tâche de repoorting est manuelle
	Les requêtes statistiques sont complexe (jointure).
Conséquences	Retards enregistrés dans l'établissement des statistiques
	Possibilité d'avoir des erreurs.
	Risque de perte d'information
	Perte de temps

Tableau 5 : Anomalie 02

- **Anomalie N°3 : Les rapports établis sont partiels**

Causes	Manque de données
	Nature de donnée : donnée interne/interne, format papier.
Conséquences	Difficulté de repondre aux attentes des décideurs
	Incapacité d'avoir une vue sur la situation globale

Tableau 6 : Anomalie 03

- **Anomalie N°4 : Difficulté d'analyser les causes d'échecs**

Causes	Diverses sources de données
	Multiplicité de format de stockage
Conséquences	Organisation non apprenante
	L'atteindre du niveau de maturité reste difficile

Tableau 7 : Anomalie 04

2. Ordre technique

- **Anomalie N°5 : Difficulté de la visualisation des données statistiques :**
  Causes

Causes	Manque d'outils d'analyse et de visualisation
	La capacité des experts est limitée
Conséquences	La découverte de la connaissance cachée est impossible

Tableau 8 : Anomalie 05

3. Ordre organisationnel

- **Anomalie N°6 : Le département se focalise sur un système opérationnel qui s'avère limité**

Causes	Processus manuelle, semi-automatique
	Système informel
Conséquences	Mise en place d'un système de BI est difficile

Tableau 9 : Anomalie 06

- **Anomalie N°7 : Les décideurs ne sont pas ambitieux**

Causes	Résistance aux changements
	Ne sont pas orienté vers les TIC
	Perte de pouvoir
Conséquences	Réussir un projet de BI reste difficile

Tableau 10 : Anomalie 07

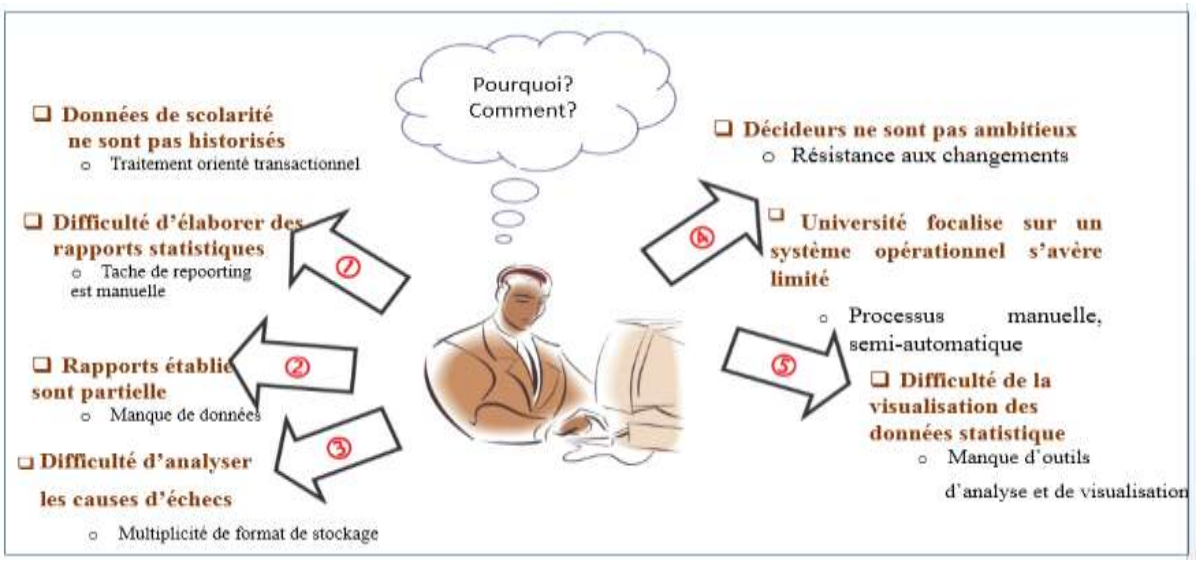


Figure 10 : Les anomalies rencontrées

2.3. Solution future

Dans cette section, nous décrivons nos améliorations, puis nous présentons une vue global sur notre future système.

2.3.1. Suggestion

Après avoir étudié les processus du système actuel et suite au diagnostic effectué, nous sommes arrivés à la déduction que la principale limite de ce système est l'absence d'une méthodologie d'exploitation des données dans le processus existant. Afin de remédier à ces limites, nous proposons les quelques suggestions suivantes :

- Automatiser les tâches manuelles liées à la visualisation et le *reporting* afin d'avoir un accès rapide pour la prise de décision. Par exemple l'informaticien chargé pour l'administration de la base de données de scolarité passe un temps énorme dans la sélection des classifications des meilleurs étudiants pour le stage de bourse à l'engranger.

- Représenter les données d'une manière uniforme et cohérence: les données répliquées et distribuées sur plusieurs services peuvent être communiquées d'une manière incohérente par exemple conflit de nommage (par ex. id etudiant, id student, رقم الطالب) format de donnée JMMAAAA.

- Mettre en place un système décisionnel pour les différents acteurs de la formation. Ce système va permettre d'explorer et de visualiser les données accumulées dans le service de scolarité.

- Mettre en place des outils de fouille de données qui génèrent des connaissances cachées dans les données de la scolarité.

2.3.2. Système projeté

Comme illustré par la figure , le système que nous projetons devrait permettre aux acteurs de la formation, à travers un tableau de bord, de visualiser des *reportings* et de disposer d'informations statistiques sur l'évolution du parcours des étudiants afin de les analyser. En outre il pourrait également leur fournir des prédictions et des recommandations. Pour ce faire, notre système disposerait d'une BDD historisée où seront stockées toutes les données pertinentes relatives au cursus de formation des étudiants.

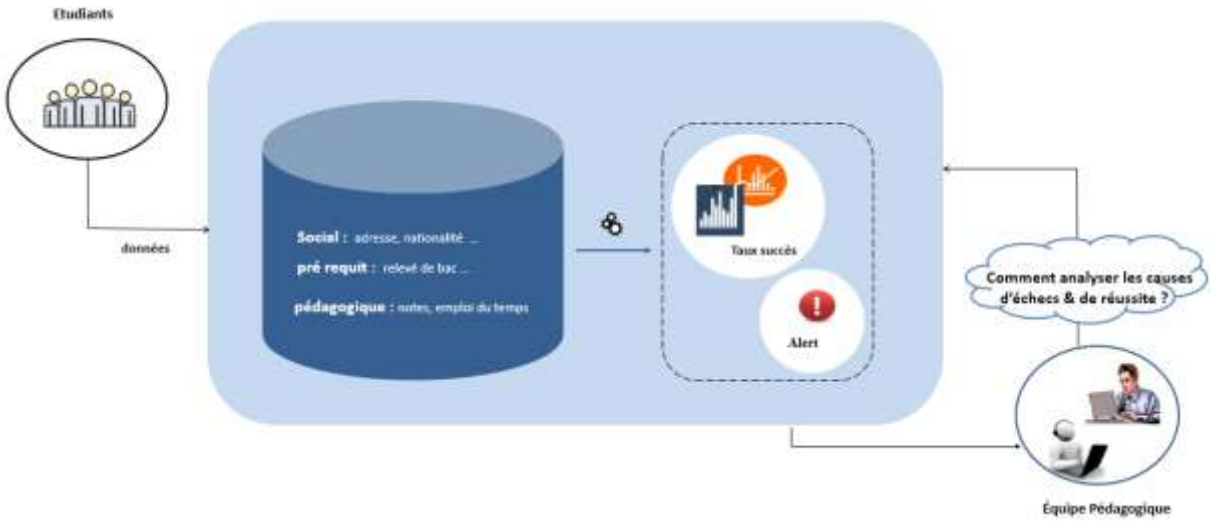


Figure 11 : Système projeté

2.3.3. Détail de la solution

Dans ce qui suit, nous allons présenter la solution envisagée pour la mise en place de notre projet. Cette dernière sera réalisée conformément aux étapes suivantes :

- Obtention des données sources dans un format Excel, Access, PDF
- Extraction, Transformation et Chargement de ces données grâce à un ETL vers une Base de Données Relationnelle (BDR)
- Transformation de la BDR en cube multidimensionnel
- Exploration des données à l'aide des outils OLAP, *reporting* et datamining.

La figure ci-dessous, présente l'architecture de cette solution.

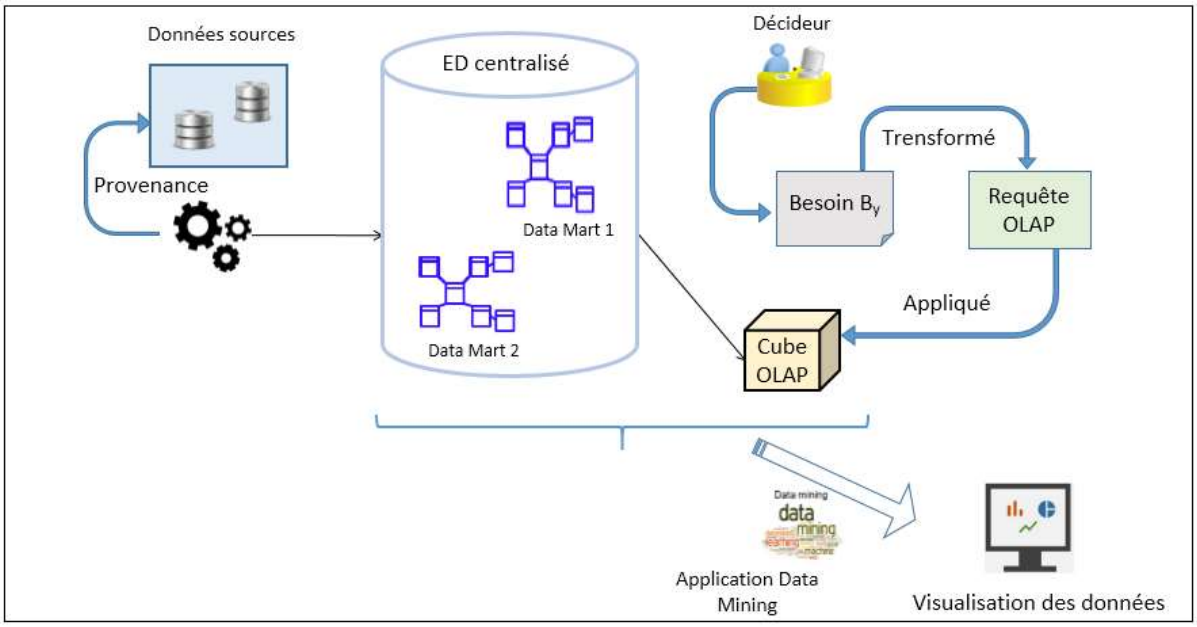


Figure 12 : Solution détaillé.

2.3.3.1. Identification des données sources

Avant d'aller plus loin, nous devons identifier les sources de données sur lesquels opère notre système, le reste des phases de notre solution sera présenté dans le chapitre suivant.

Dans le cadre de ce projet de fin d'études, nos sources de données sont des données hétérogènes, à savoir: (1) **structurés** (base de données de scolarité), (2) **semi-structurés** (relevé de notes du baccalauréat, emplois du temps) et (3) **non structurés** (canevas).

Remarque : Le canevas sera défini par la suite.

1. **Donnée structurées**

- **BDDS[6] (base de données)** : est une partie de l'application SEEE7, Elle contient les données de notre université (département) et est constituée de plusieurs rubriques. Chacune de ces rubriques contient des informations spécifiques. Cette base de données se caractérise par un SGBD « Access » qui reçoit des demandes de manipulation du contenu et effectue les opérations nécessaires.

La figure 13 illustre l'organisation des données de notre base de données source.

[6] BDDS : Base de données de notre scolarité.
[7] SEEE : Application de Suivi des étudiants et des enseignements supérieurs

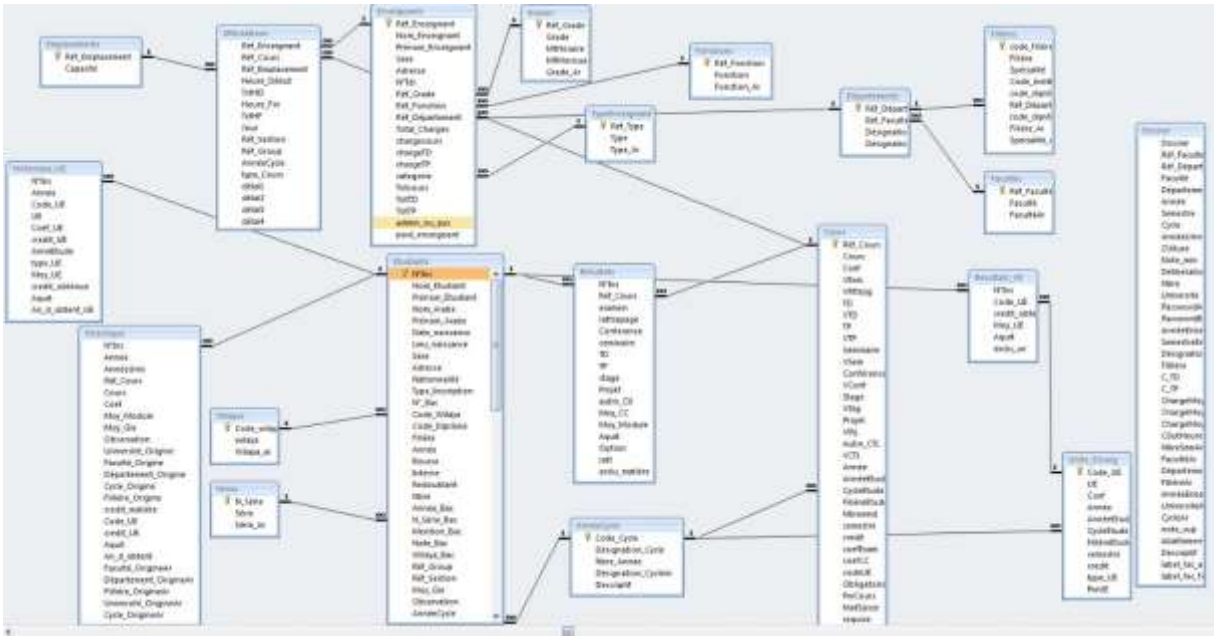


Figure 13 : Schéma relationnel de la base de données.

2. Données semi-structurées (Excel) :

- **Relevé de notes du baccalauréat :** les nouveaux bacheliers qui s'inscrivent à UIK déposent au niveau du secrétariat relevés de de notes du baccalauréat. Ces derniers contiennent leurs propres informations, aux notes de l'ensemble des matières où ils ont été examinés et la moyenne du baccalauréat.
- **Résidence :** les étudiants venant de différentes régions doivent avoir les hébergements dans les résidences universitaires suivantes : Assia Kebir et Habib Khalil.
- **Emplois du temps :** est un calendrier de travail, où figurent à la fois le temps, l'affectation des enseignants, les jours et les horaires de travail. Les enseignements sont organisés selon des séances de cours, de TD ou de TP d'une durée de 1h30. Pour chaque séance il est spécifié la matière enseignée qui fait partie d'une unité d'Enseignement, la salle ou l'amphithéâtre, l'enseignant, le groupe d'étudiants concerné, le jour et l'heure.

3. Données non structurées (PDF):

- **Canevas :** c'est une fiche qui sert à organiser l'ensemble des formations au sein des universités. Elle précise pour chaque enseignement le contexte de la formation, ses objectifs et les moyens humains et matériels disponibles. Il s'agit également d'une Fiche d'organisation semestrielle des enseignements et des unités d'enseignement où sont portés les programmes détaillés pour chaque matière.

2.3.3.2.Points bloquants

Lors de l'étude préliminaire des données un certain nombre de difficulté ont été identifiés et répertoriés, nous les synthétisons dans les points suivants :

- **Etude des données** : nous avons eu une difficulté à étudier la base de données à cause de l'absence des référentiels et des dictionnaires de données. Nous avions eu donc souvent recours aux utilisateurs et gestionnaires de la base pour comprendre le sens des données et la localisation des informations nécessaires à exploiter.

- **Absence de qualité de données :** la présence de données fausses dès leur saisie, le manque de données :

- **Inaccessibilité aux données sources**.

2.4. Conduite de projet

Afin d'assurer le bon déroulement et le suivi du projet, nous avons suivi une démarche de conduite de projet. Cette démarche est détaillée à l'annexe.

Conclusion

Comprendre l'existant permet non seulement de voir les manques qui y existent, mais inspire aussi à la définition d'une nouvelle solution.

Au terme de cette phase, nous nous sommes imprégnés du contexte du projet en termes de discernement des limites du système actuel. Nous avons présenté la solution informatique qui s'y adapte et qui répond au mieux aux objectifs définis précédemment.

Dans le chapitre suivant, nous entamerons l'étude conceptuelle qui nous permettra de concevoir un système d'information décisionnel.

Deuxième partie

Conception
&
Mise en œuvre de notre solution

Chapitre 03 :
Conception de notre solution

Introduction
Nous allons passer dans ce chapitre à la conception et la mise en œuvre de notre système décisionnelle. Et avant de commencer, nous allons étudier les différentes stratégies et approches de conception du schéma de l'entrepôt qui existent pour choisir celles qui conviennent à nos besoins et les données disponibles afin de proposer une modélisation multidimensionnelle exploitable par les outils d'analyse de données. Et nous conclurons par étudier les technique de data mining utilisés.

3.1. La conception de l'entrepôt de données
Dans cette section, nous présentons la méthodologie suivie durant notre travail ainsi que la stratégie adopté pour la conception de l'ED.

3.1.1. Méthodologie suivie
Dans les systèmes transactionnels les concepteurs procèdent par l'approche de Merise ou UML. Mais pour les systèmes décisionnels il y'a d'autres approches spécifiques car ils ont des structures spécifiques afin de répondre à des fins différentes. L'une de ces approches est celle qui est proposée par Ralph Kimball, et elle est basée sur la gestion de projet par prototypage et propose une bonne gestion pour ce type de projet. La figure suivante représente le cycle de vie du logiciel qui montre toutes les étapes de sa conception, de son développement et de sa maintenance d'une manière enchainée. Il s'assure que chaque étape est réussie avant de passer à la suivante. Le but de ce découpage par étapes est d'en contrôler les erreurs afin de minimiser les coûts.

3.1.2. Stratégie de la conception de l'entrepôt
Dans cette section, nous décrivons notre approche adoptée.

3.1.2.1. Approche guidée par les données sources (Approche top-down)
Cette approche s'appelle aussi l'approche orientée données, elle ignore les besoins d'analyse a priori. Elle concerne en particulier les travaux sur l'automatisation de la conception de schéma. Cette approche consiste à construire le schéma de l'entrepôt à partir de ceux des sources de données et suppose que le schéma qui sera construit pourra répondre à tous les besoins d'analyse. Elle est supportée généralement pour les Data Mart (magasins de données), où le nombre des activités et des sujets est restreint on peut donc facilement déterminer l'ensemble des indicateurs à mesurer et déduire les axes des mesures (les dimensions).

3.1.2.2. Approche guidée par les besoins d'analyse (Approche bottom-up)
Cette approche propose de définir le schéma de l'entrepôt en fonction des besoins d'analyse et suppose que les données disponibles permettront la mise en œuvre d'un tel schéma. Parmi les approches orientées besoins d'analyse, on distingue aussi deux sous approches :

- **Approche orientée buts** : Elle suppose que le schéma de l'entrepôt est défini selon les objectifs d'analyse de l'organisation. Ainsi que tous les utilisateurs de l'organisation concernée ont des besoins d'analyses similaires et la même chose pour l'exploitation de l'entrepôt de données. Autrement dit, tous les utilisateurs ont la même vision analytique de l'entrepôt de données.

- **Approche orientée utilisateurs :** on trouve dans ce cas-là que les utilisateurs sont interrogés afin de collecter l'ensemble de leurs besoins d'analyse avant de construire l'entrepôt de données. Ce qui permet de garantir l'acceptation du par les utilisateurs. Cependant la durée de vie de ce schéma peut être courte, car le schéma dépend beaucoup des besoins des personnes impliquées dans le processus de développement de l'entrepôt.

3.1.2.3. Approche mixte (Approche hybride)

L'approche mixte considère à la fois les besoins d'analyse et les données pour la construction du schéma. L'idée dans cette approche est de construire des schémas candidats à partir des données. Ainsi, le schéma construit constitue une réponse aux besoins réels d'analyse et il est également possible de le mettre en œuvre avec les sources de données.

3.1.3. Approche choisie pour la conception de l'entrepôt

Dans notre cas, il est clair qu'on peut bénéficier des avantages des approches précédentes. C'est-à-dire qu'on va construire notre schéma de l'entrepôt de données en se basant sur les données des résultats de délibérations. Car grâce aux données détaillées on peut conclure facilement les sujets et les processus à analyser, de plus à partir des données on peut déterminer l'ensemble des indicateurs ou les mesures qui expriment d'une manière globale les résultats obtenus par les étudiants. On a travaillé en collaboration avec la direction des études qui ont la grande part dans la détermination et la limitation du nombre d'indicateurs selon l'objectif de notre travail afin de présenter une proposition d'un point de vue informatique, pour montrer aux décideurs de l'UIK que le système décisionnel peut être une solution scientifique et automatique pour l'extraction de connaissance, et un outil performant pour l'amélioration de l'ensemble des procédures de gestion de cette opération annuelle (Délibération).

3.1.4. Conception du modèle multidimensionnel

3.1.4.1. Introduction

Après Avoir vu les différentes approches et différentes méthodologies nous passons à la construction du Data Warehouse, cette dernière passe par trois étapes :

- L'étude préalable qui nécessite une connaissance détaillée des données source.

- La conception du modèle dimensionnel de données qui représente l'entrepôt conceptuellement et logiquement.

- La mise en œuvre de l'architecture, par une suite de trois sous étapes:

  - Construction de l'entrepôt, la base de l'entrepôt.

  - Construction des cubes OLAP, la base multidimensionnelle.

  - Construction de la zone d'alimentation, qui reprend à un niveau plus précis l'examen des données, le choix des méthodes et des dates auxquelles les données entreront dans l'entrepôt.

Selon Ralph Kimball: La conception logique d'un *Data Warehouse* passe par quatre étapes (processus) :

- ***Choix du processus d'activité à modéliser*** *:* Un processus d'activité est un processus opérationnel important pour l'organisation, étayé par une ou plusieurs applications à partir desquelles des données peuvent être collectées au profit de l'entrepôt de données, comme exemple : le processus les résultats de délibération.

- ***Choix du grain du processus d'activité*** : le grain est niveau de détail fondamental, atomique, des données figurant dans la table de faits pour le processus. Il est impossible de passer à l'étape trois sans avoir préalablement défini le grain.

- ***Choix des dimensions applicables à chaque table de*** **faits** : Le choix d'une dimension s'accompagne par la définition de tous les attributs textuels qui garniront la table de dimension.

- ***Choix des mesures que contiendra chaque enregistrement de la table de faits*** Qui sont généralement des quantités numériques additives telles que le note moyenne…etc.

3.1.4.2. Identification des processus

Lors de l'identification, nous cherchons les processus concernés par notre étude. Cela est le résultat d'une étude approfondie des données sources qu'on va l'explorer. Le processus clé à analyser est : Résultat de délibérations des notes des étudiants de l'UIK, les Note_module et Note_UE.

3.1.4.3. Les indicateurs d'analyse

Il est clair que l'utilisateur de notre système doit avoir une vision transversale des objectifs déclinés en plans d'action et indicateurs décisionnels, implémentés au niveau processus. Ces indicateurs permettent de :

- Mesurer les résultats de délibérations.

- Contrôler, mesurer, maitriser et améliorer les processus de gestion de délibérations des notes des étudiants de l'UIK.

Et voici les principaux indicateurs que nous avons extraits avec les utilisateurs:

- Nombre d'étudiant par année, par sexe, série du bac, par section
- Moyenne des moyennes obtenues par les étudiants de l'UIK par année, par sexe, par série du bac, par section.
- La note moyenne par année, par sexe, par section, par groupe et par module.
- Note maximale des notes par année, par sexe, par section, par groupe et par module.
- Note minimale des notes par année, par sexe, par section, par groupe et par module.
- Nombre des étudiants acquit par année universitaire, année_étude et unité enseigné.

3.1.4.4. Modélisation multidimensionnelle

Rappelons que la modélisation des entrepôts de données se base sur deux concepts fondamentaux : le concept de fait et le concept de dimension. Un fait représente un sujet d'analyse, caractérisé par une ou plusieurs mesures, qui ne sont autres que des indicateurs décrivant le sujet d'analyse.

Ce fait est analysé selon des axes d'observation qui constituent également ses descripteurs. « Donc on peut dire qu'un entrepôt de données présente une modélisation multidimensionnelle puisqu'elle répond à l'objectif d'analyser des faits en fonction de dimensions qui constituent les différents axes d'observation des mesures ».

1. Modélisation multidimensionnelle de l'activité «Fait_délibération»

Dans cette section nous présentons le processus de cette activité, son gain, ses dimensions et les faits mesurés

- **Le processus d'activité**

Après une étude détaillée des données sources qu'on veut explorer, on est arrivé à choisir un sujet important c'est : Résultat de délibération, Les décideurs ont besoin d'évaluer les résultats réalisés par les étudiants de notre université « UIK ».

- **Le grain du processus d'activité**

Lors de l'analyse du résultat de délibération, il faut la répartition des résultats des étudiants par rapport à l'ensemble des axes qui seront cités dans ce qui va suivre afin de mesurer l'effet de chaque dimension sur ce résultat.

- **Les dimensions**

Les dimensions qu'on a choisies dans notre projet pour ce processus sont :

La dimension «Dim_année_universitaire » : Selon Ralph Kimball, la dimension temps est : « La seule dimension qui figure systématiquement dans tout entrepôt de données, car en pratique tout entrepôt de données est une série temporelle. Le temps est le plus souvent la première dimension dans le classement sous-jacent de la base de données » [Kim, 96].

Pour notre cas il est évident d'utiliser le temps car il référence les résultats de délibération. De plus l'année du bac est un des principaux axes d'agrégation pour naviguer dans les cellules d'un cube OLAP.

La dimension «Dim_année_étude » : Cette dimension va contenir les cycles d'étude qui existent à l'UIK (Master, licence), les spécialités (Génie informatique, Genie logiciel et Réseau de télécommunication) et l'année lui-même 1ere année, 2eme année , 3eme année. Qui nous aide à savoir les taux d'échecs et taux de réussites par promotion.

La dimension «Dim_Sexe» : Cette dimension va contenir bien sûr deux enregistrements (un pour désigner le sexe féminin et l'autre pour le sexe masculin), cette dimension est importante car ce critère est généralement utilisé dans tous les établissements d'enseignement de notre pays.

La dimension «Dim_Série_bac» : Les étudiants qui peuvent accéder à l'UIK sont les étudiants qui ont un bac de série soit sciences exactes, soit sciences de la nature et de la vie ou un bac technique (génie électrique, génie mécanique). Pour voir l'influence de la série du bac sur le rendement des étudiants durant la période d'étude à l'UIK.

La dimension « Dim_Adresse » : Cette dimension va contenir les adresses des étudiants.

- **Les Faits mesurés**

Le fait modélise le sujet de l'analyse. Un fait est formé de mesures correspondant aux informations de l'activité analysée. Dans notre cas nous avons enregistrés des indicateurs qui donnent brièvement la description des résultats de délibération et qui sont :

.**Mesures**	**Calcul**	**Type mesures**	Description
Nbr_étudiants	Count	Semi-additif	Nbr des étudiants inscrit
Nbr_ad_ses1			Nbr étudiants admis normal
Nbr_ad_ses2			Nbr étudiants admis par rattrapage
Nbr_dettes			Nbr étudiants admis par dettes
Nbr_admis	Sum(Nbr_ad_ses1+ Nbr_ad_ses2+ Nbr_dettes)	Non-additif	Nbr étudiants admis
Nbr_ajourné	Count	Semi-additif	Nbr étudiants doublants
Max_Moy_annuel	MAX (moy_annuel)		Moyenne annuelle
Nbr_aquit_s1	Count		Nbr étudiants réussits dans le 1er semestre
Nbr_aquit_s2	Count		Nbr étudiants réussits dans le 2eme semestre
Max_Moy_s1	MAX (moy_s1)		Moyenne maximale du 1er semestre
Max_Moy_s2	MAX (moy_s2)		Moyenne maximale du 2eme semestre

Tableau 11 : Faits mesurés de l'activité «Fait_délibération»

Après avoir défini les faits et les dimensions, notre schéma conceptuel de l'activité « Fait_délibération » se présente comme suit :

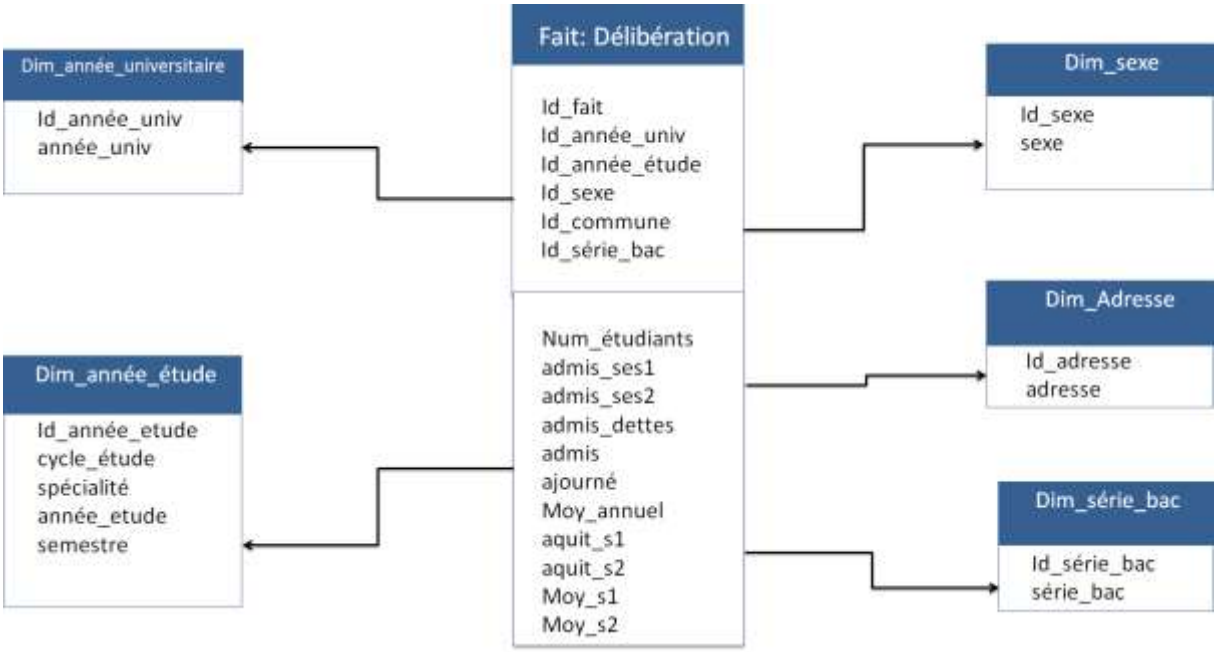


Figure 14 : Modèle multidimensionnel en étoile de l'activité « Délibération »

2 . *Modélisation multidimensionnelle de l'activité «Fait_Note_Module»*

Dans cette section, nous détaillons le processus de cette activité, le gain du processus, ses dimensions et les faits mesurés.

- **Le processus d'activité**

D'après l'activité précédente, nous sommes arrivé à la conclusion que les mesures qui expriment les notes module par (Moy_module, Max_moy_examen, Max_moy_td, Max_moy_tp) par rapport aux enseignant, section et module nécessitent de les regrouper dans une autre table de faits « fait_Note_Module ».

- **Le grain du processus d'activité**

Pour l'analyse les notes des étudiants par module, il faut savoir la moyenne des notes, maximum et minimum des notes pour chaque année par modules, enseignant et par section.

- **Les dimensions**

Pour cette table de faits nous avons déterminé les dimensions suivantes:
Dim_année_universitaire, Dim_année_étude, Dim_module, Dim_enseignant, Dim_section. Dim_année_universitaire, Dim_année_étude sont étudiées dans l'activité « Fait_délibération», sauf ces dimension Dim_module, Dim_enseignant, Dim_section.

La dimension « Dim_module » : Elle comporte l'ensemble des modules étudiée à l'UIK. L'étude des notes nécessite la connaissance des modules.

La dimension « Dim_enseignant » : Cette dernière comporte l'ensemble des enseignants qui enseigné à l'UIK.

La dimension « Dim_section » : Cette dimension va contenir les sections qui existent ainsi que les groupes associés. On a intégré la section pour pouvoir avoir le taux d'échecs et réussites par section.

- **Les faits mesurés**

Selon nos besoins d'étude de cette activité nous avons choisis les mesures suivantes :

Mesures	Calcul	Type mesures	Description
Moy_ module	AVG (moy_module)	Semi-additif	
Nbr_ratr_module	Count		Nbr des étudiants ayant subi un rattrapage
Nbr_aquit_mod			Nbr étudiants réussits aux modules
Nbr_aquit_examen			Nbr étudiants réussits aux examens
Nbr_aquit_td	Count	Semi-additif	Nbr étudiants réussits aux TD_s
Nbr_aquit_tp	Count	Semi-additif	Nbr étudiants réussits aux TP_s
Max_moy_examen	Max (moy_examen)		Moyenne maximale de l'examen
Max_moy_td	MAX (moy_td)		Moyennne maximale au TD
Max_moy_tp	MAX (moy_tp)		Moyenne maximale au TP

Tableau 12 : Faits mesurés de l'activité «Fait_Note_Module»

Ces mesures nous permettent de savoir le niveau global des étudiants par module. Et voilà le schéma conceptuel de l'activité

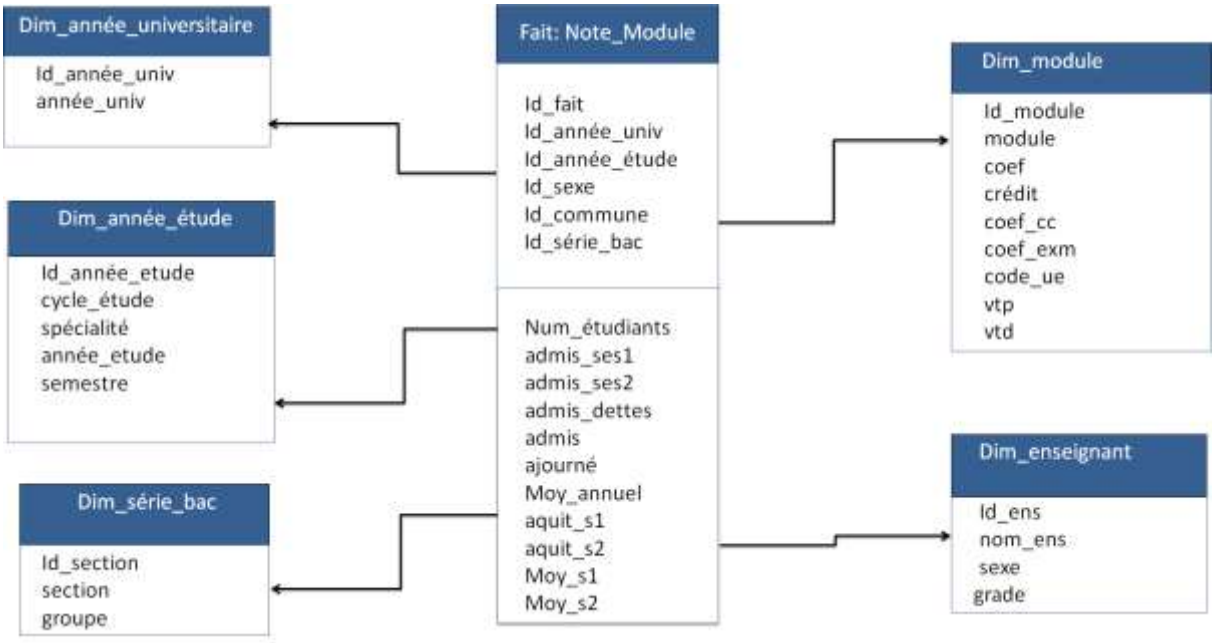


Figure 15 : Modèle multidimensionnel en étoile de l'activité «Fait_Note_Module»

2. Modélisation multidimensionnelle de l'activité «Fait_Note_UE»

- **Processus d'activité :**

Comme nous l'avons vu précédemment le nombre des étudiants aquit par unité enseigné n'est pas concerné par les dimensions précédentes.

- **Le grain du processus d'activité**

Quand on fait une analyse des résultats d'échecs des étudiants par rapport aux unités, il est important de savoir le nombre des étudiants pour chaque année par unité enseigné et année universitaire.

- **Les dimensions**

Toutes les dimensions sont étudiées dans les activités précédentes sauf la dimension Dim_unité_enseigné.

La dimension « Dim_unité_enseigné» : cette dimension va contenir les unités enseigné à l'UIK.

- **Les faits mesurés**

Mesures	Calcul	Type mesures	Description
Nbr_aquit_ue	Count	additif	Nbr des étudiants réussits par unité

Tableau 13 : Faits mesurés de l'activité «Fait_Note_UE»

Et voilà le schéma conceptuel de l'activité

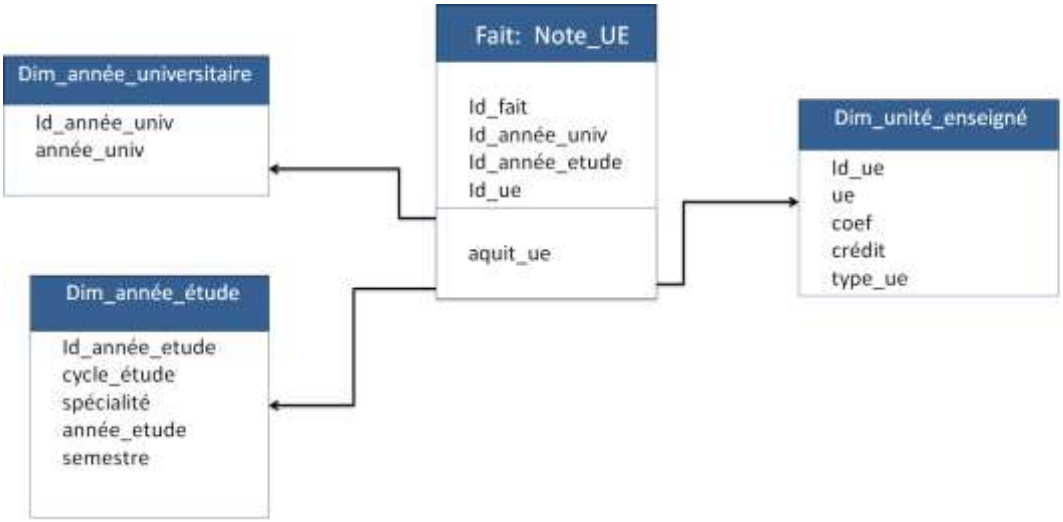


Figure 16 : Modèle multidimensionnel en étoile de l'activité «Fait_Note_UE»

2.1.5. Conception de cube OLAP

Nous proposons également d'enrichir la phase d'exploitation de l'entrepôt de besoins par des techniques utilisées pour l'OLAP sur les données, en fournissant : un langage de requêtes OLAP dédié aux besoins qui peut être complété par des outils d'interrogation destinés aux décideurs et gestionnaires, la fouille des besoins entreposés, des techniques de visualisation des rapports.

- Énumérer les besoins B_j concernés par les causes d'échec et de réussite par les décideurs.

- Convertir les besoins du format source (langage naturel) vers le format MDX (requête OLAP) en se basant sur le schéma cible de l'entrepôt (cube). Cette partie illustrée dans la figure 17.

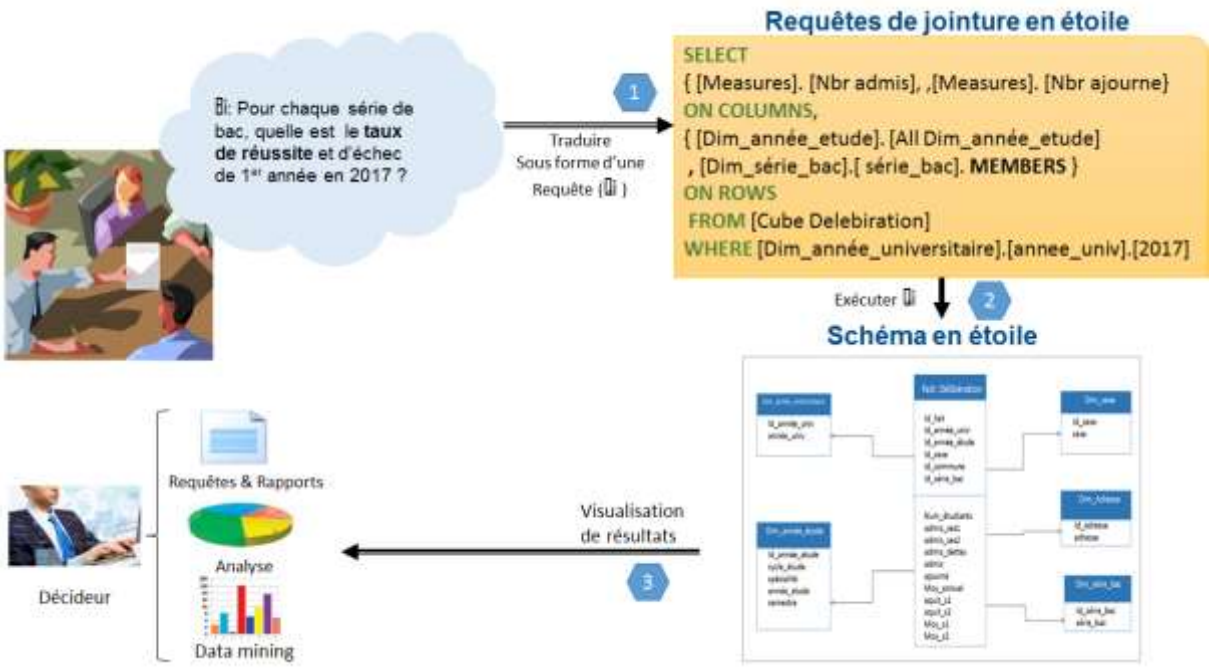


Figure 17 : L'expression des besoins Langage naturel vers Requête OLAP.

L'expression des besoins peut se faire selon plusieurs langages de modélisation, qui s'étendent des langages complètement informels comme la langue naturelle, jusqu'aux langages SQL.

Notre approche permet aux utilisateurs d'effectuer des analyses sur les résultats de délibération obtenu à l'aide de requêtes OLAP. Généralement, ces requêtes sont implémentées à l'aide de MDX.

Les Faits qui correspondent à un test donné sont appelés observations. Chaque observation représente un point dans l'espace multidimensionnel formé par les dimensions. La figure 18 donne un exemple de découpe appliquée à un cube de données stockant des données des résultats d'évaluation en prenant en compte trois dimensions : *année_universitaire, année_étude, serie_bac*

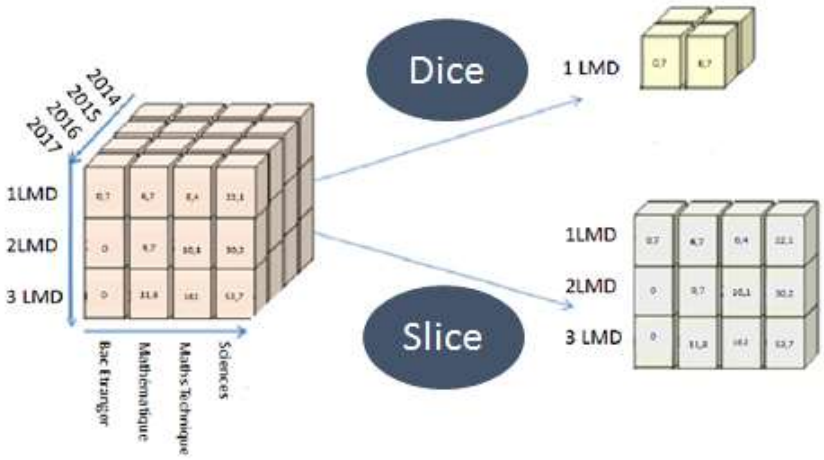


Figure 18 : Processus OLAP slice et dice.

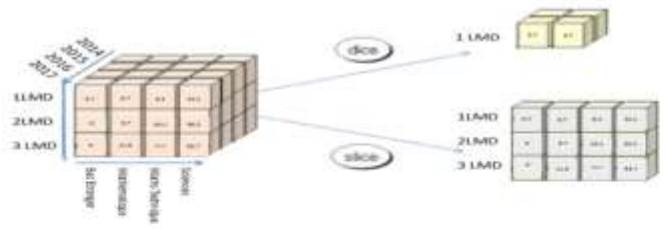
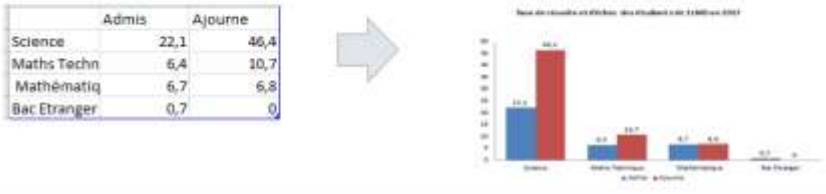
Le tableau suivant présente une correspondance entre les besoins des décideurs en langage naturel et les requêtes OLAP.

Besoin en langage naturel (BI)	Requête OLAP (Qi)
Évolution de réussite des étudiants de 2010 à 2017	**SELECT** {[Dim_année_universitaire]. [annee_univ]. **MEMBERS**, [Measures]. [Nbr étudiants]} **ON COLUMNS,** {[Dim_année_étude]. [All Dim_année_étude]} **ON ROWS** FROM [Cube Délibération]
Taux de redoublement et de réussite en 2016	**SELECT** {[Measures]. [Nbr ajourne], [Measures]. [Nbr admis]} **ON COLUMNS**, {[Dim_année_étude]. [All Dim_année_étude] } **ON ROWS** **FROM** [Cube Délibération] **WHERE** ([Dim_année_universitaire]. [annee_univ]. [2016])
Quels sont les meilleurs étudiants de licence de la faculté de l'informatique	**SELECT** {[Dim_année_universitaire]. [annee_univ]. **MEMBERS**, [Measures]. [Max Moy_annuel]} **ON COLUMNS**, **FROM** [Cube Délibération] **WHERE** ([Dim_année_étude]. [cycle étude]. [Licence])
Taux de réussites et d'échec Garçons/filles par année d'étude	**SELECT** {[Dim_année_universitaire]. [annee_univ]. **MEMBERS**, [Measures]. [Nbr admis]} **ON COLUMNS,** {[Dim_année_étude]. [All Dim_année_étude] **,** [Dim_sexe]. [sexe]. **MEMBERS**} **ON ROWS** **FROM** [Cube Délibération]
Quel est le nombre des étudiants réussit par semestre	**SELECT** {[Dim_année_universitaire]. [annee_univ]. **MEMBERS**, [Measures]. [Nbr aquit s1], [Measures]. [Nbr aquit s2]} **ON COLUMNS,** {[Dim_année_étude]. [All Dim_année_étude] } **ON ROWS** **FROM** [Cube Délibération]

Tableau 14 : Requête en langage naturel et OLAP

2.1.5.1.Scénario d'usage

La figure suivante montre un scénario de processus d'analyse OLAP.

Step	Example
Step 0: Besoin	Un utlisateur veut analyser le taux de réussite et d'échec des etudiants de première année licence par série de bac,
Step 1: Choix de dimension d'analyse : « Checklist Approach »	✓ Dim_anné_universitaire ✓ Dim_série_bac ✓ Dim_année_étude ✗ Dim_sexe ✗ Dim_adresse
Step 2 : Choix de métrique « Checklist Approach »	✓ Admis(AVG)
Step 3 : Requête OLAP	SQL script to extarct result of a manifest SQL Script generation SELECT { [Measures]. [Nbr admis], ,[Measures]. [Nbr ajourne} ON COLUMNS, { [Dim_année_etude]. [All Dim_année_etude] , [Dim_série_bac].[série_bac]. **MEMBERS** } ON ROWS FROM [Cube Delebiration] WHERE [Dim_année_universitaire].[annee_univ].[2017]
Step 4: Extraction de cube	
Step 5: Visualisation de résultat	

	Admis	Ajourne
Science	22,1	46,4
Maths Techn	6,4	10,7
Mathématiq	6,7	6,8
Bac Etranger	0,7	0

Figure 19 : Exemple de processus d'analyse OLAP.

2.2. Application des techniques de *Data Mining*

Le datamining répond à de nombreuses applications en divers domaines et cela permettant de tirer profit de disponibilité croissante de données localisées et de leurs richesses potentielles ; c'est le cas de l'analyse de réussite et l'échec d'un étudiant.

3.2.1. Processus d'extraction

Nous présentons dans cette section, l'enchaînement des phases de processus d'extraction proposé dans le chapitre1, adapté au cas étudié.

- Prévision de la réussite
- Analyse module

3.2.1.1. Intervenant internes et externes

Avant de présenter les différentes étapes du processus comme indiqué dans la figure 20, nous définissons les trois types d'intervenants au cours de ce processus comme suit :

- **Le décideur** : son rôle est de poser le problème, définir l'objectif et valider les résultats obtenus

- **L'analyste** : c'est une personne qui va superviser tout le processus, son rôle est de préparer les données à étudier, créer ou sélectionner un index de jointure, exécuter l'algorithme de classification et valider les résultats obtenus.

- **L'expert du domaine** : c'est une personne qui a des acquis dans l'analyse, ainsi il permettra d'identifier les variables à étudier et fourni des informations utiles à la recherche.

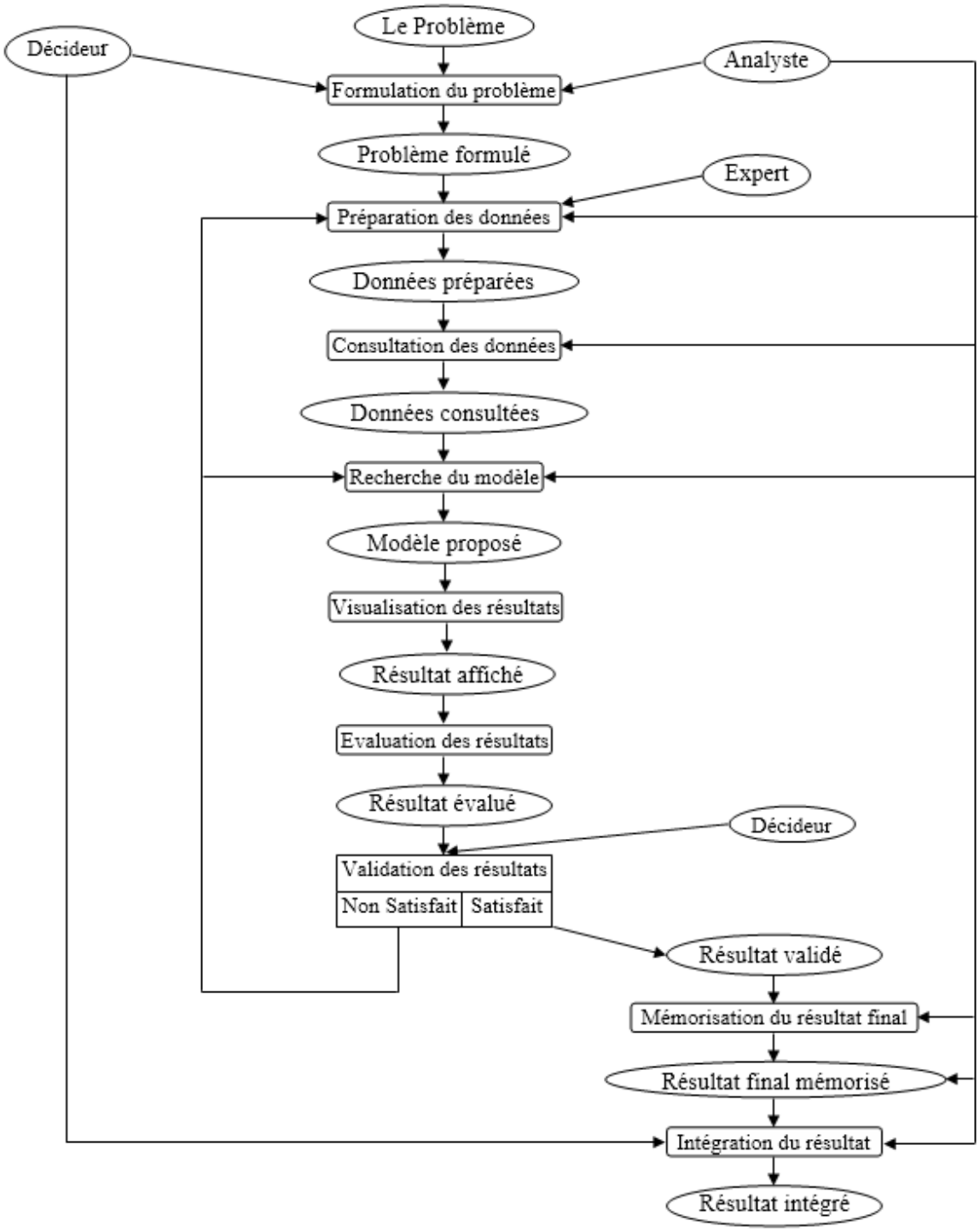


Figure 20 : Schéma général du processus proposé

3.2.1.2.Analyse du problème

Notre travail s'appuiera sur les deux techniques de Datamining :

- La méthode KNN pour la classification des étudiants.
- Les Règles d'association en se basant sur l'algorithme a priori

L'étude porte sur les données réelles de l'UIK qui constitueront la base de données de notre système.

- ***Prévision la réussite***

Dans cette section, nous allons présenter la formation du problème de prévision (entrées, *sorites et objectif*) et l'algorithme de résolution de datamining utilisé. L'intérêt de cette tâche de datamining est de remédié l'échec massif de ses étudiants de première et deuxième année. D'après l'enquête menée l''échec concerne essentiellement les deux premières années.

- **Description du problème**

Etant donné que un décideur qui veut prévoir la réussite d'un nouveau étudiant, un étudiant au cours du semestre ou bien pour la nouvelle année (ex. du 1ere à la 2 eme année). Ce dernier est caractérisé par un ensemble de caractéristiques qui sont : l'âge, sexe, nationalité, note bac …etc. Il faut qu'il attribue chaque classe (échec ou réussite) à un nouveau étudiant. Le problème de prévision de réussite d'un étudiant donné est un problème de classification binaire.

- **Formalisation du problème**

Le problème de classification es basé sur le jeu de données DS, comportant N étudiants étiquetées (dont la classe est connue). Chaque étudiant E est caractérisé, par P attributs

(S_i, $note_{ij}$) et par sa classe Ci ϵ C, C étant l'ensemble des classes (Échec, Réussite).

Entrées :

DS < (E_i, C_i), (E_i, C_i) … (E_n, C_n) > sachant que:

Ci= {Échec, Réussite}.

E_i <S_i,$Note_i$> tel que :

S_i = {genre, age, nationalité…} un ensemble de caractéristiques

$Note_i$ = {$note_{i1}$, $note_{i2}$,…, $note_{ik}$}

Avec $note_{ik}$ est la note de module m_k de l'étudiant E_i.

Pour prédire la classe d'un nouveau étudiant (Ê), nous s'appuyons sur l'ensemble DS précédent : Ê<Ŝ, Note,?>

Sorties :

C= (Échec, Réussite).

- **Rechercher les données**

La recherche des données consiste dans un premier temps à obtenir des données en accord avec les objectifs que l'on s'impose. Afin de sélectionner les attributs à retenir pour l'analyse des données des étudiants, nous sommes inspirés dans la littérature, des types d'attributs utilisés au sein des domaines d'enseignement et d'apprentissages de l'université. Ces données sont intégrées dans notre entrepôt de données.

- **Sélectionner les données pertinentes :** Nous présentons les attributs que nous avons retenus [25]. Les attributs pertinents sont : nationalité, sexe, genre,

année étude, wilaya de bac, note bac, série bac, type de bac, résidence, notes (mathématique, physique, français, anglais).

- **Nettoyer les données :** Le nettoyage de de données consiste à :

  - Corrections des doublons, des erreurs de saisie : Les doublons peuvent se révéler gênants parce qu'ils vont donner plus d'importance aux valeurs répétées. Une erreur de saisie pourra à l'inverse occulter une répétition.

  - Contrôle sur l'intégrité des domaines de valeurs : Un contrôle sur les domaines des valeurs permet de retrouver des valeurs aberrantes.

  - Détection des informations manquantes : C'est le terme utilisé pour désigner le cas où des champs ne contiennent aucune donnée.

- **Transformation des données :**

Une étape très dépendante du choix de l'algorithme de fouilles de données utilisés. Une fois les variables sont pertinentes et les données sont fiables, une transformation éventuelle s'impose pour les préparer au travail d'analyse. Il s'agit d'intervenir sur les valeurs des variables pour qu'elles soient mieux exploitables par les algorithmes de traitement.

- **Recherche des modèles :**

Une approche traditionnelle pour découvrir ou expliquer un phénomène est d'établir un modèle. Ce dernier se mesure selon les critères suivants :

  - Rapide à créer.

  - Rapide à utiliser.

  - Compréhensible pour l'utilisateur.

  - Les performances sont bonnes; Le modèle est fiable.

  - Les performances ne se dégradent pas dans le temps.

- Il évolue facilement.

- **Mesure de similarité**

Les mesures de similarité utilisées sont les corrélations, les distances et la méthode de quotient (voir la figure 21). Dans notre cas nous avons choisi la distance euclidienne.

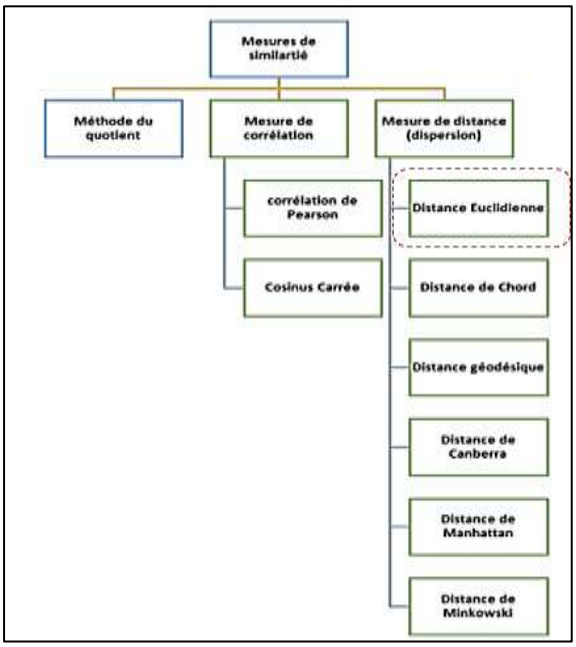


Figure 21 : Aperçu des mesures de similarité utilisées dans l'extraction des liens

- **Choix d'algorithme :**

Parmi les algorithmes de classification supervisée les plus populaires dans la littérature [26], on trouve : les arbres de décision, l'algorithme *k-NN* et les réseaux de neurones.

Nous avons choisi l'algorithme *k-NN* (*K-Nearest Neighbors*, où k plus proches voisins) qui est une méthode de classification supervisée non paramétrique puisque aucune estimation de paramètres n'est nécessaire comme pour la régression linéaire. On dispose de données d'apprentissage (training data) pour lesquelles chaque observation dispose d'une classe y. Si le problème est à 2 classes, y est binaire. L'idée de l'algorithme des **KNN** est pour une nouvelle observation $(u_1, u_2, \ldots, u_p)$ de prédire les k observations lui étant les plus similaires dans les données d'apprentissage. Quand on parle de voisin**,** cela implique la notion de distance ou de dissimilarité.

La distance la plus populaire est la distance euclidienne : g

$$D\left(\left(x_1, x_2, \ldots, x_p\right), \left(u_1, u_2, \ldots, u_p\right)\right) = \sqrt{\left(x_1 - u_1\right)^2 + \left(x_2 - u_2\right)^2 + \ldots + \left(x_p - u_p\right)^2}$$

Le principe de cette méthode est:

- Trouver les k plus proches observations.

- Utiliser une règle de décision à la majorité pour classer une nouvelle observation.

• Évaluer et valider les résultats

Après l'exécution de l'apprentissage, le résultat est visualisé sous forme graphique. L'évaluation des résultats consiste à mesurer la pertinence du résultat trouvé; et cela en mesurant l'écart entre les résultats réels et les résultats du modèle.

• Extraire la connaissance

Dans cette phase, le décideur donne ces appréciations et il décide :

- soit de continuer l'exploitation pour améliorer d'avantages les résultats.
- soit d'arrêter la recherche si la solution est satisfaisante.

➢ Analyse module

L'analyse de modules a pour objectif d'identifier des relations d'association et des corrélations entre modules. Ce résultat aide le décideur à identifier les modules qui ont contribué dans l'échec et la réussite des étudiants.

• Description du problème

Dans notre étude on se propose de répondre à la question suivante: Est-il possible de trouver des règles d'association entre modules afin de détecter ceux qui contribuent à la réussite ou l'échec d'un module donné.

Étant donné l'exemple suivant : nous voulons savoir si un étudiant qui réussit dans les deux modules d'analyse et d'algèbre, il peut réussir dans le module de l'algorithme. L'extraction de règles d'association consiste à extraire les règles dont le support et la confiance sont au moins égaux à des seuils minimaux de support et de confiance définis par l'utilisateur.

• Formaliser le problème

Cette phase consiste à sélectionner les données (attributs et objets) de la base de données utiles à l'extraction des règles d'association et transformer ces données en un contexte d'extraction. Ce contexte, ou jeu de données.

Entrée :

Étant donnée le jeu de données DS :

Tel que : DS $<T_1, T_2, \ldots, Tn>$

$T_i <m_1,m_2,\ldots,m_3>$ tel que m_i est les module aquit par étudiant.

Soit les données suivantes représentant les modules aquit par un ensemble des étudiants :

Id_étudiant	Modules
1	m_1, m_2, m_4
2	m_2, m_3, m_5
3	m_1, m_2, m_3, m_5
4	m_2, m_5

Tableau 15 : Exemple de donnée

Appliquez l'algorithme A priori, pour rechercher les règles associatives entre les modules , en fixant le support minimum.

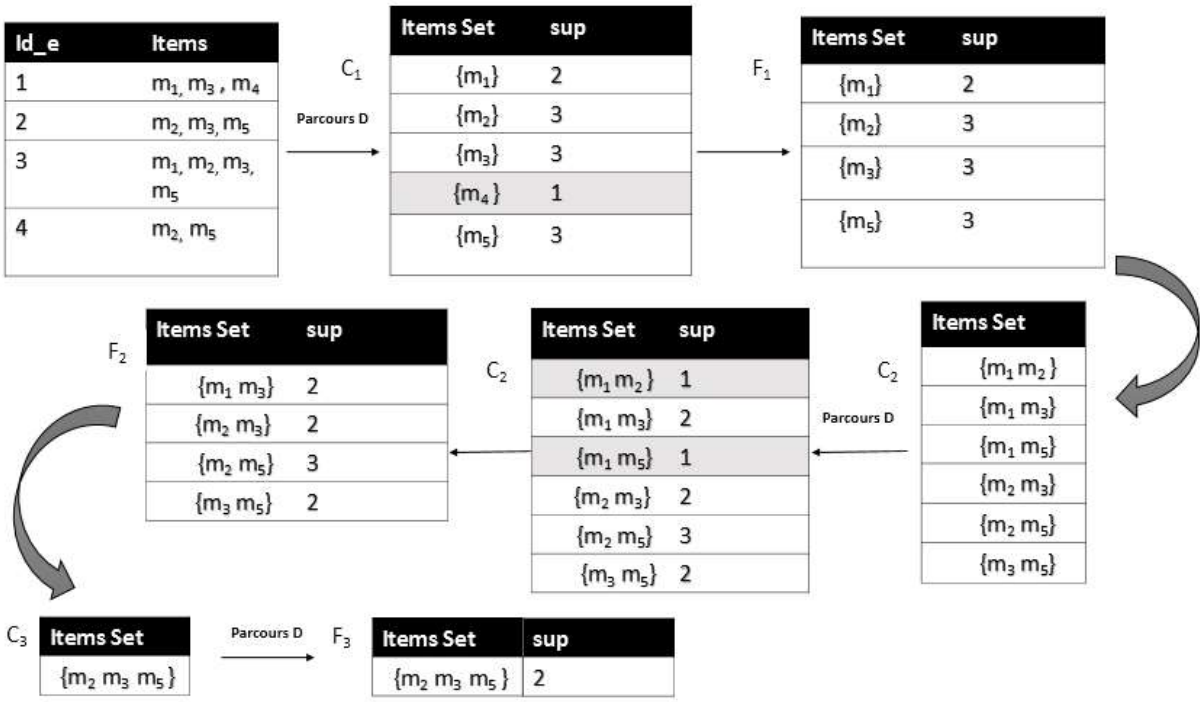


Figure 22 : Exemple A priori algorithme avec le support minimum =2

Sortie :

Ressortir un ensemble de règles Ri,

$R_i = \{m_1 \rightarrow m_2 , m_2, m_3 \rightarrow m_4 \ldots.\}$

telles que m_i soit un sous-ensemble de T_i ($m_i \subseteq T_i$).

- **Choix d'algorithme**

La Rechercher des règles associatives consiste à extraire des items fréquents par des techniques de datamining (par ex. **APRIORI, Close, OCD, Partition, DIC°).** Dans notre cas on applique l'algorithme APRIORI.

Algorithme APRIORI se déroule comme suit :

- Génération d'ensembles d'items.
- Calcul des fréquences des ensembles d'items.
- On garde les ensembles d'items avec un support minimum: les ensembles d'items fréquents.

- **Interprétation des résultats**

Cette phase consiste en la visualisation par l'utilisateur des règles d'association extraites du contexte et leur interprétation afin d'en déduire des connaissances utiles pour l'amélioration de l'activité concernée.

Conclusion

Nous avons vu dans ce chapitre les différentes approches de modélisation et nous avons choisi une approche qui convient à nos besoins. Puis nous sommes passés à la modélisation multidimensionnelle de l'entrepôt après avoir défini les concepts de fait et dimension ainsi que la conception de notre cube OLAP. En fin nous avons entamé la partie data mining.

Nous décrivons dans le prochain chapitre l'architecture technique de notre solution, nous citons les différents outils et technologies que nous avons utilisées, ainsi nous présenterons les quelques captures d'écran.

Chapitre 04 :
L'implémentation et la mise en œuvre de notre application

Introduction
Dans ce dernier chapitre nous allons aborder la phase d'implémentation qui expose les techniques détaillées du système conçu. En premier lieu nous allons décrire les ressources utilisés, ainsi que la mise en œuvre de la solution. Ensuite nous allons définir l'architecture technique de notre système et nous finirons par exposer l'aspect sécuritaire de notre solution.

4.1. Ressource utilisées
Dans cette section, nous allons présenter les différents outils que nous avons utilisés pour la réalisation de notre solution

- **ETL : Talend Open Studio**

Talend Open Studio est le leader incontestable dans les outils d'intégration de données open source. Ce logiciel a été développé en 2005 par la société Talend dont le siège social se trouve actuellement aux États-Unis.
L'une des grandes forces de cet ETL est sa capacité à pouvoir se connecter à quasiment toutes les sources de données. Il comporte près de 250 composants manipulables à travers une interface graphique job designer et permet de créer des routines personnalisées avec les langages de programmation PERL et JAVA.

Parmi les avantages de Talend, nous trouvons :

- Une interface ergonomique et intuitive.
- La possibilité de créer des composants personnalisés.
- Une documentation riche et régulièrement mise à jour.
- La possibilité de manipuler les jobs et composants graphiquement.
- Une forte présence de sa communauté sur internet.
- Une portabilité et puissances.

- **SGBD : My SQL**

MySQL est l'un des systèmes de gestion de base de données libres les plus utilisés au monde, autant par le grand public que par des professionnels.
MySQL est un serveur de bases de données relationnelles SQL développées dans un souci de performances élevées en lecture, multi-thread et multi-utilisateurs, il fonctionne sur de

nombreux systèmes d'exploitation différents incluant Linux, Mac OS X, Solaris, Unix, et Windows.

.

- **Scene Builder**

Scene Builder est un outil de présentation visuel qui permet aux utilisateurs de concevoir rapidement des interfaces utilisateur d'applications JavaFX, sans codage. Les utilisateurs peuvent faire glisser et déposer des composants de l'interface utilisateur dans une zone de travail, modifier leurs propriétés, appliquer des feuilles de style et le code FXML de la présentation qu'ils créent est automatiquement généré en arrière-plan.
Le résultat est un fichier FXML qui peut ensuite être combiné à un projet Java en liant l'interface utilisateur à la logique de l'application

- **Environnement de développement Eclipse IDE Indigo» :**

Eclipse IDE (Integrated Development Environment) est un environnement de développement libre permettant de créer des programmes dans de nombreux langages de programmation (Java, C++, PHP…).
La caractéristique essentielle d'Eclipse est l'extensibilité de l'environnement. Plus que de se focaliser sur un environnement de développement Java, les concepteurs d'Eclipse se sont efforcés avant tout de créer un socle applicatif sur lequel viennent se greffer des modules et/ou plugins. La version Eclipse IDE pour JAVA EE 6 support : un éditeur graphique JSP/ JSF/ HTML, les outils de gestion de base de données, et support la plupart des serveurs d'application.

- **Mondrian SchemaWorkbench**

Le Mondrian SchemaWorkbench est une interface de conception qui vous permet de créer et de tester Mondrian OLAP schémas de cube visuellement. Le moteur Mondrian traite MDX requêtes avec le ROLAP (Relational OLAP) schémas.
Ces fichiers de schéma sont des modèles de métadonnées XML qui sont créés dans une structure spécifique utilisé par le moteur Mondrian. Ces modèles XML peuvent être considérés comme des structures cubiques qui utilisent FACT et DIMENSION tableaux trouvés dans votre SGBDR existantes. Il ne nécessite pas qu'un cube physique est construit ou maintenu; seulement que le modèle de métadonnées est créé.

- **Weka**

Est un logiciel open source développé par l'université Wakaito en Nouvelles Zélande. Il possède plusieurs algorithmes de traitement de données de filtrage de classification, d'apprentissage et de visualisation.
Weka possède soixante-quinze algorithmes dont les réseaux de neurones, l'arbre de décision et la régression logistique.

4.1.1. Étude comparative des outils BI Open source

Le choix de ces outils technologiques est basé sur une étude faite par **Gartner, Inc** ; où il avait recensé les outils de BI les plus utilisé. La figure ci-dessous montre leur position par rapport aux autres outils existants sur le marché.

Figure 23 : Positionnement des outils de visualisation dans le marché selon le groupe Gartner

Talend met la qualité de données au cœur de la chaîne de valeur des données Cela permet à chacun de contribuer à la qualité de données plus facilement.
Selon **Gartner, Inc** « *Talend a pleinement adopté les nouveaux technologie et propose aujourd'hui une solution idéale pour les organisations cherchant à mettre en place une approche orientée données». [27]*

4.2.Mise en œuvre de la solution

4.2.1. Présentation des étapes de conception de notre entrepôt de données

La réalisation de notre entrepôt de données s'effectue en 2 étapes essentielles comme suit:

4.2.1.1.Création de l'entrepôt de donnée

- **Création de connexions aux bases de données**

Nous procédons à établir la connexion au base de donnée préalablement créée intitulé « donne_source » avec MySQL.

Cette connexion est réalisée de la manière suivante :

Lors du lancement d'ETL, on a le panneau «Repository » qui nous permet de créer une connexion à travers « Connexions aux bases de données » qui se trouve dans Le répertoire « Métadonnées », cette étape est montrée dans la figure ci-dessus :

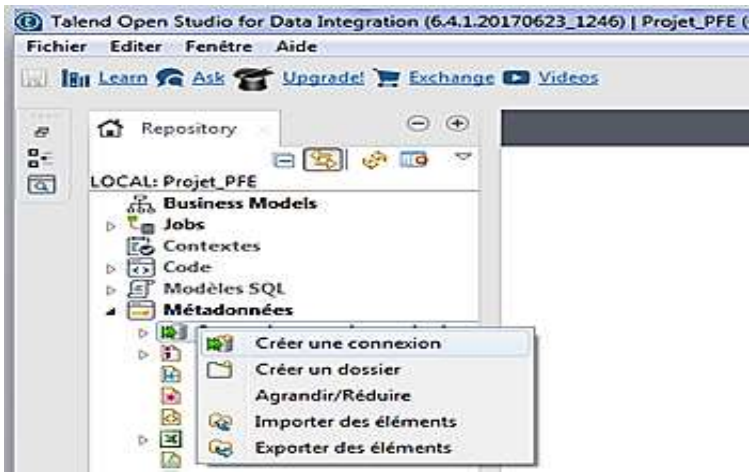


Figure 24 : Création de connexion

Les deux figures suivantes (figure25) et (figure 26) montrent la suite de configuration qui se réalise en deux étapes

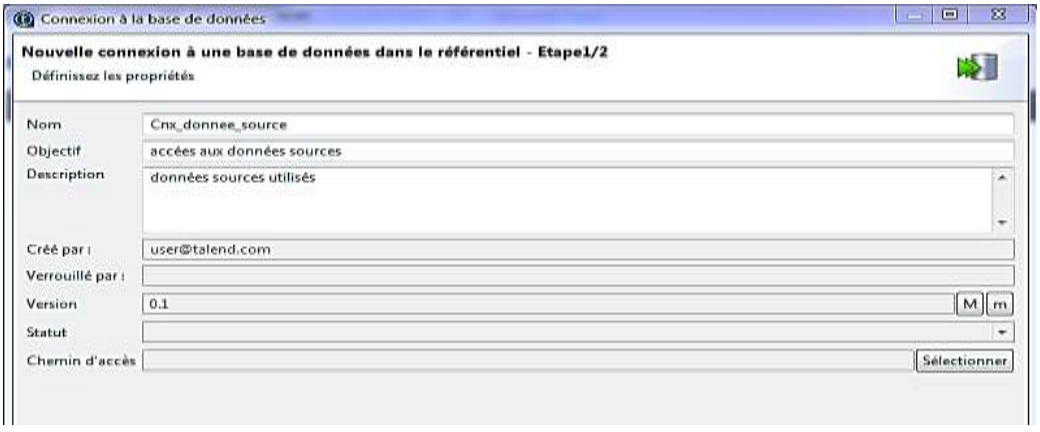


Figure 25 : Configuration de propriétés de la connexion

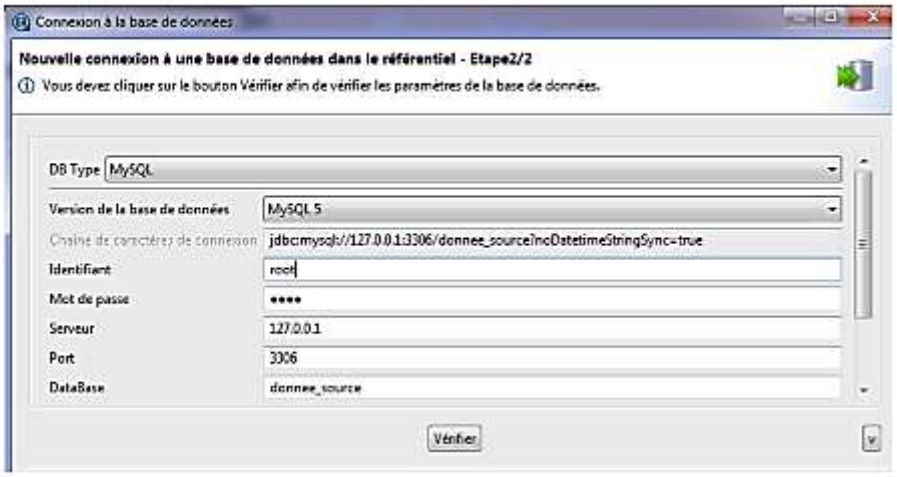


Figure 26 : Configuration générale des paramètres de la base de données.

- **Alimentation de la base de données**

Dans cette phase, nous passons à l'alimentation des tables qui est l'étape la plus critique et la plus importante pour l'alimentation de notre entrepôt.
L'extraction des données, leurs transformations ainsi que leurs chargements sont réalisés de la manière suivante :

Dans le panneau « repository », le répertoire job donne la possibilité de faire une extraction de n'importe qu'elle type de fichier, dans notre cas sont des fichiers EXCEL. Pour cela, nous allons créer une transformation avec Talend. Cette transformation est très simple et ne contient que deux étapes. La première pour l'extraction des données du fichier Excel et la seconde pour l'insertion en base de données grâce au composant de chargement *Tmap*[8]. Ce qui donne ceci :

[8] **Tmap** : Composant permet aux développeurs de mapper aisément les données issues de base de données vers MYSQL.

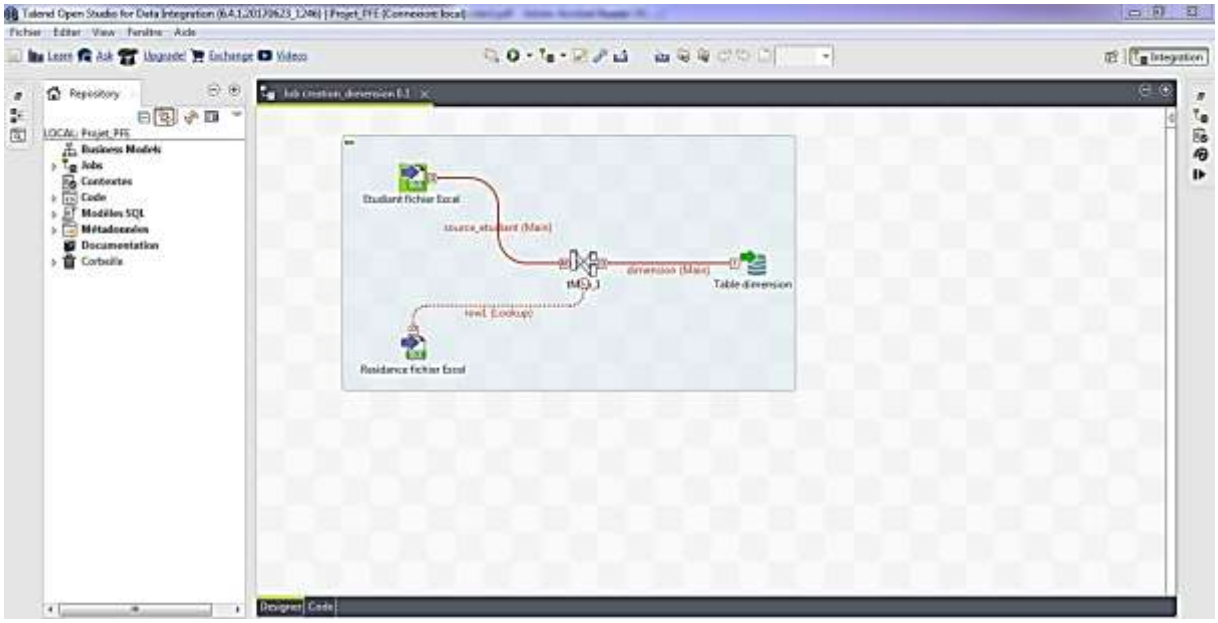


Figure 27 : Extraction, transformation et chargement de dimension.

La figure ci-dessus montre l'extraction et la transformation de dimension cette étape se répétera pour les autres dimensions.

La figure suivante illustre la création et l'alimentation de la table des faits « **fait_délibération »** à partir des dimensions.

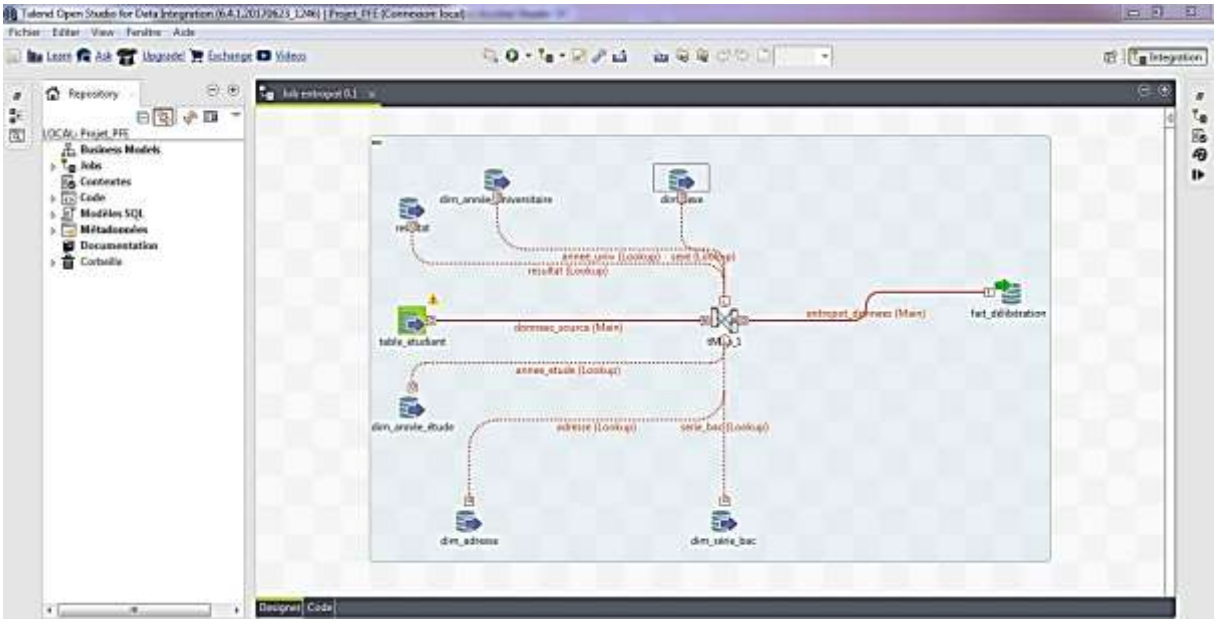


Figure 28 : Création de la table des faits « fait_délibération »

La figure ci-dessous montre :

- Une jointure entre les tables d'entrée en faisant simplement glisser le champ concerné de la table principale vers le champ équivalent de la table de référence.

- Déposez le contenu des tables source vers la table cible

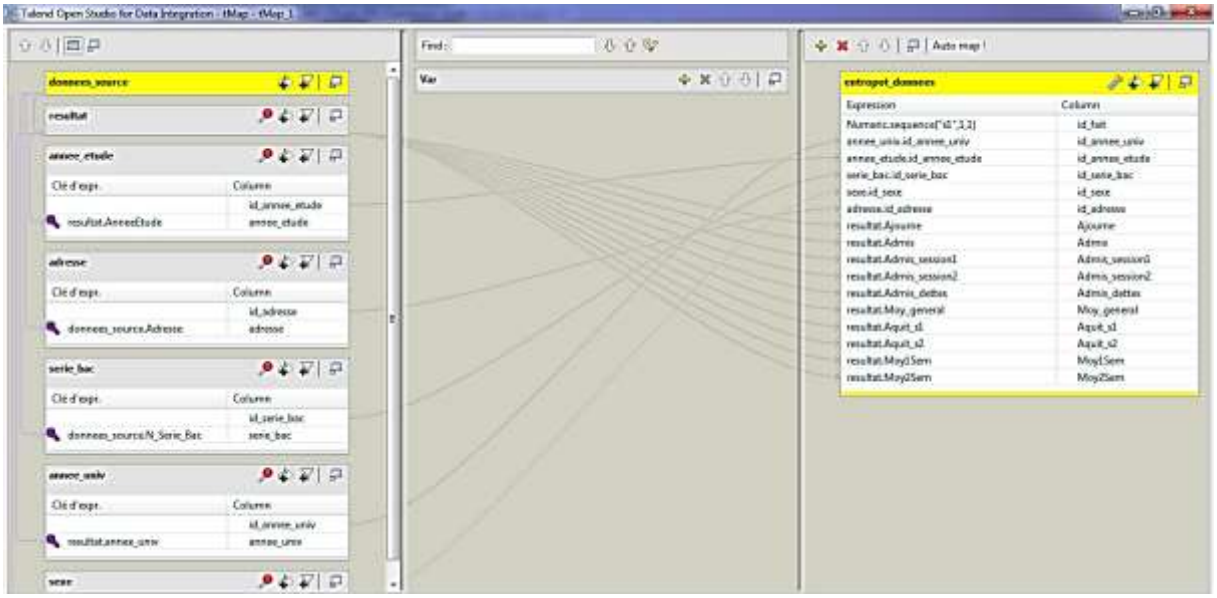


Figure 29 : Mapping de données avec filtre et jointure.

4.2.1.2. Création du cube

La création d'un cube au sens Mondrian est tout simplement la rédaction d'un fichier XML. Ce fichier permet de lier les informations du cube que nous souhaitons faire apparaître et l'entrepôt de donnée. Pour faciliter la création de ce fichier, nous allons utiliser l'outil *Schéma Workbench* qui offre une interface graphique pour effectuer cette tâche. Après avoir configuré la connexion à l'entrepôt de donnée, nous pouvons commencer à créer notre cube.

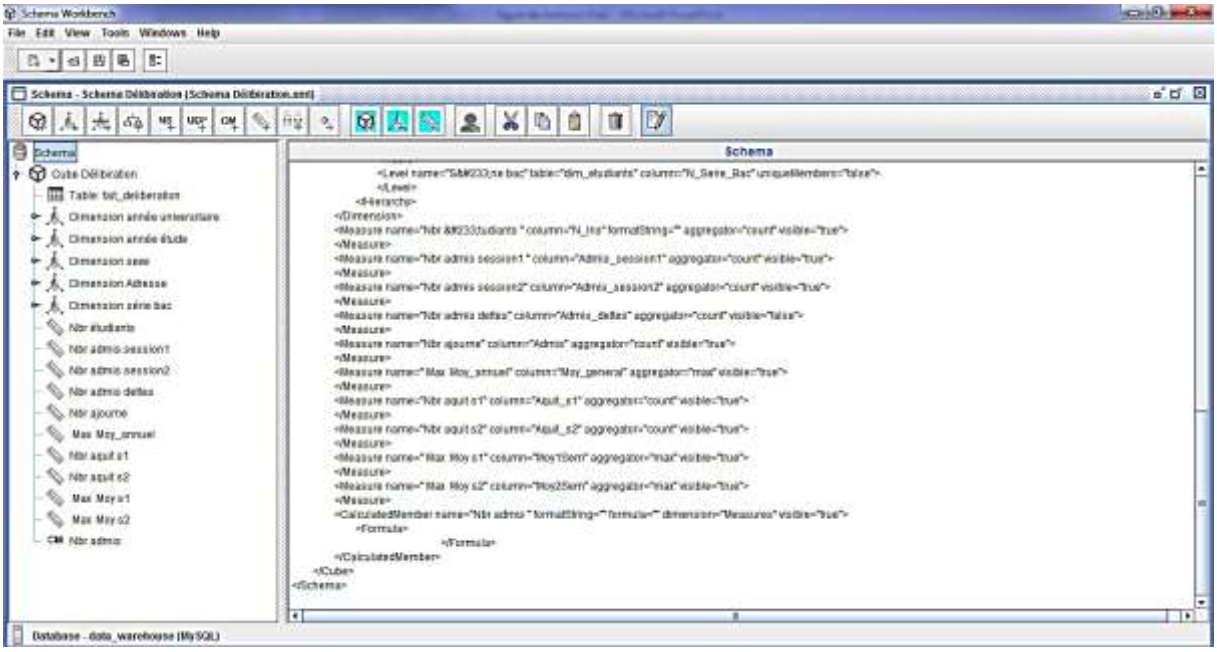


Figure 30 : Création de notre cube.

La figure ci-dessus montre que pour la création de cube se déroule comme suit :

- De la table de fait « fait_delibértion »
- Des dimensions « année universitaire » « année étude» «sexe» «Adresse» « série bac ».
- De hiérarchies dimension et property pour chaque dimension.
- De mesures dans notre cas « Nbr étudiants » « Nbr admis session1 » « Nbr admis session2 » « Nbr admis dettes » « Nbr ajourné » « Max moy_annuel » « Nbr aquit s1 » « Nbr aquit s2 » « Max moy s1 » « Max moy s2 » « Nbr admis »

Cette étape se répète pour les autres cubes (cube_note_module) et (cube_note_ue).

4.3. L'architecture technique

La réalisation de notre solution passe par cinq étapes principales :

- Identification des données sources.
- L'implémentation de l'ETL.
- Stockage du Data Warehouse.
- L'implémentation des modèles de Data mining

- La visualisation des résultats de Data mining sur un tableau de bord

A travers la figure suivante, nous présentons l'architecture technique de notre solution qui illustre les outils et technologies utilisés.

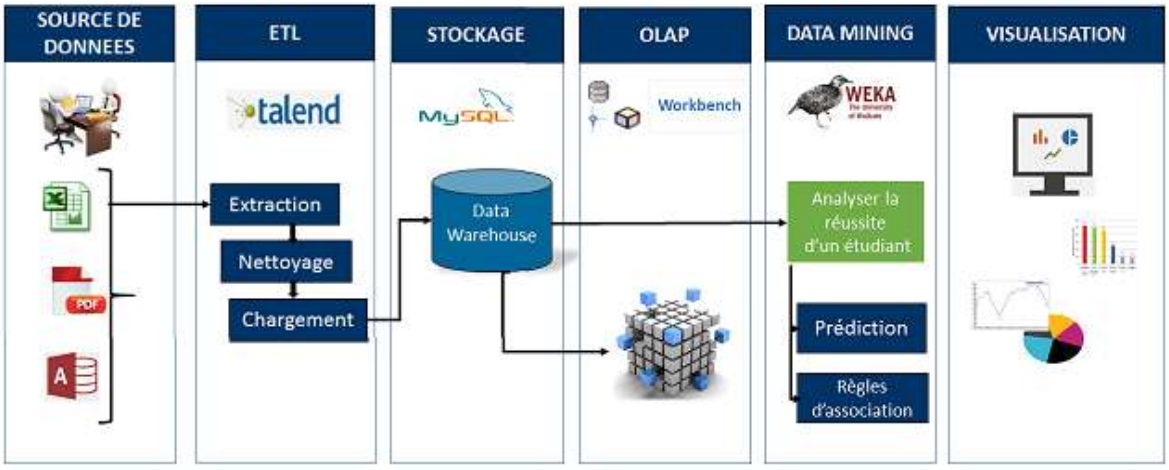


Figure 31: Architecture de notre solution.

4.3.1. Capture d'écran

Pour mieux expliquer l'usage de notre système, nous décrivons dans ce qui suit les fonctionnalités à travers les différentes interfaces.

- **Interface Accueil**

La figure suivante illustre l'interface principale de notre outil assistant nommé **MINERS** RoS dédie pour l'équipe pédagogique.

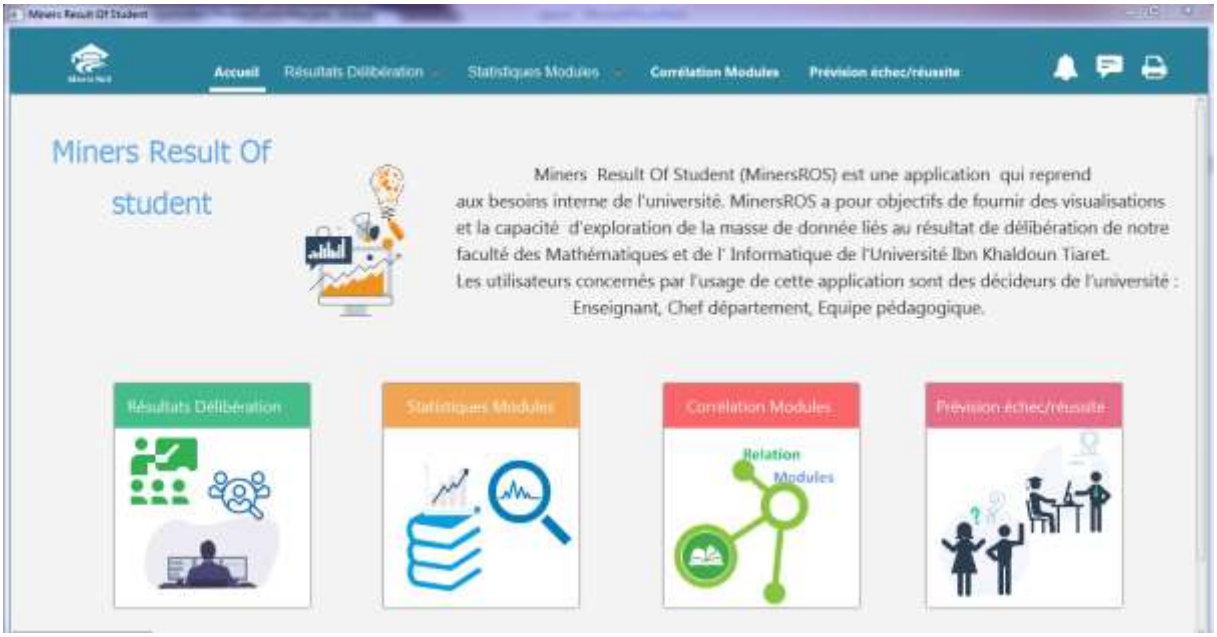


Figure 32 : Interface Accueil.

- **Interface des résultats de délibérations (volet 1)**

L'interface des résultats de délibération montre aux décideurs des *reporting* sous forme graphique (par ex Histogramme). Ces résultats est visualisées en fonction de certains critères d'analyse comme *cycle étude et année universitaire.* La courbe montrée dans la capture d'écran permet d'illustrer l'évolution des pourcentages des étudiants admis dans années différentes.

Notre outil affiche des meilleurs de classe par année d'étude. Cette information synthétisée est exploitée par les décideurs pour sélectionner les étudiants concernés par une formation de doctorat à l'étranger sachant que ce type d'information est difficile à le restituer à partir des données de production.

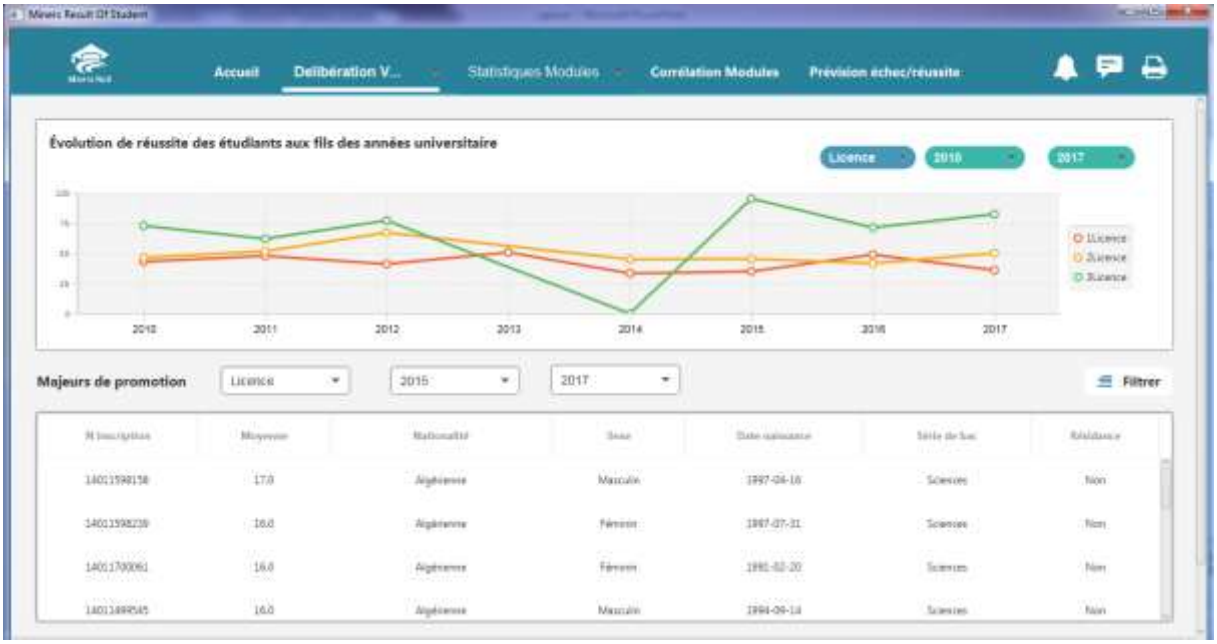


Figure 33 : Interface des résultats de délibérations (volet 1)

- **Interface des résultats de délibérations (volet 2)**

Le volet 2 de module résultats de délibérations affiches les différents forme graphique qui montre la situation relatives aux résultats des évaluations des étudiants à un moment données. Le décideur peut visualiser dans n'importe quel moment les nombre admis/ajourné par niveau d'étude et année d'étude, le taux de réussite par sexe, par année universitaire et par semestre avec l'effectif des étudiants relatif.

Figure 34 : Interface des résultats de délibérations (volet 2)

- **Interface statistiques modules (volet 1)**

Cette interface le taux réussite des unités de modules (découverte, transversale, méthodologie et fondamentale), par années d'études et par semestre dans des années différentes. Nous croyons que ces informations visuelles aident de près ou de loin la décideuse à analyser des facteurs qui causent l'échec et la réussite des étudiants.

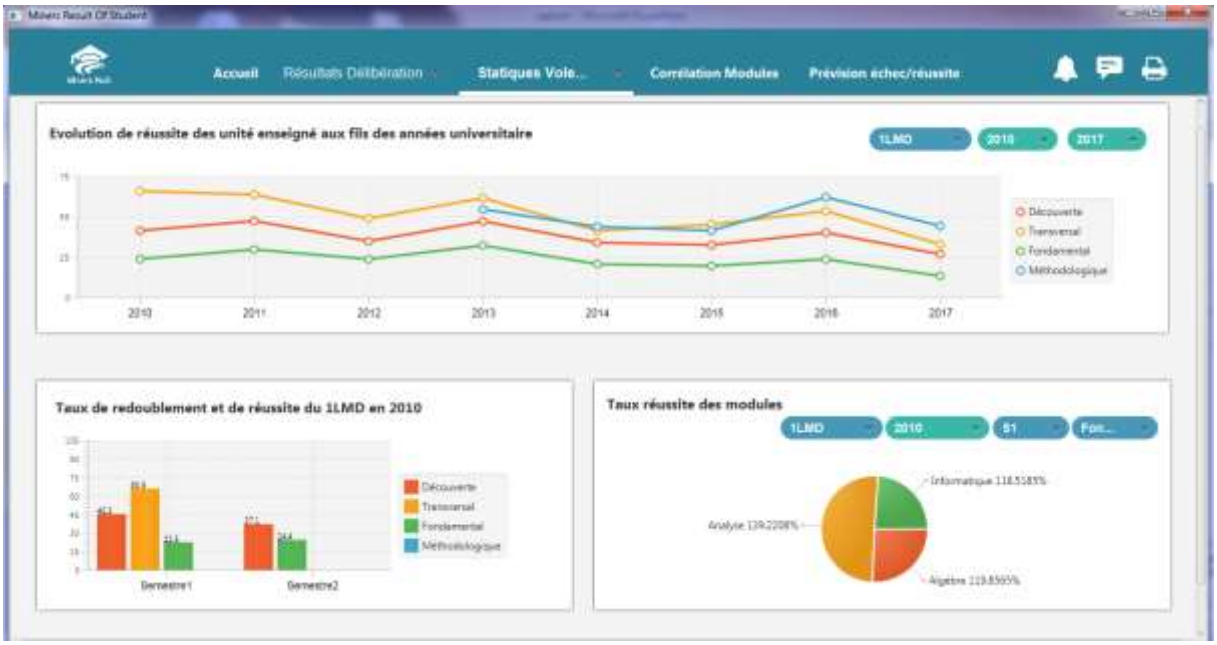


Figure 35 : Interface statistiques modules (volet 1)

- **Interface statistiques modules (volet 2)**

Cette interface fait un zoome sur les résultats des modules selon différents mode d'évaluation TD, TP er examen et le taux réussite par section et groupe.

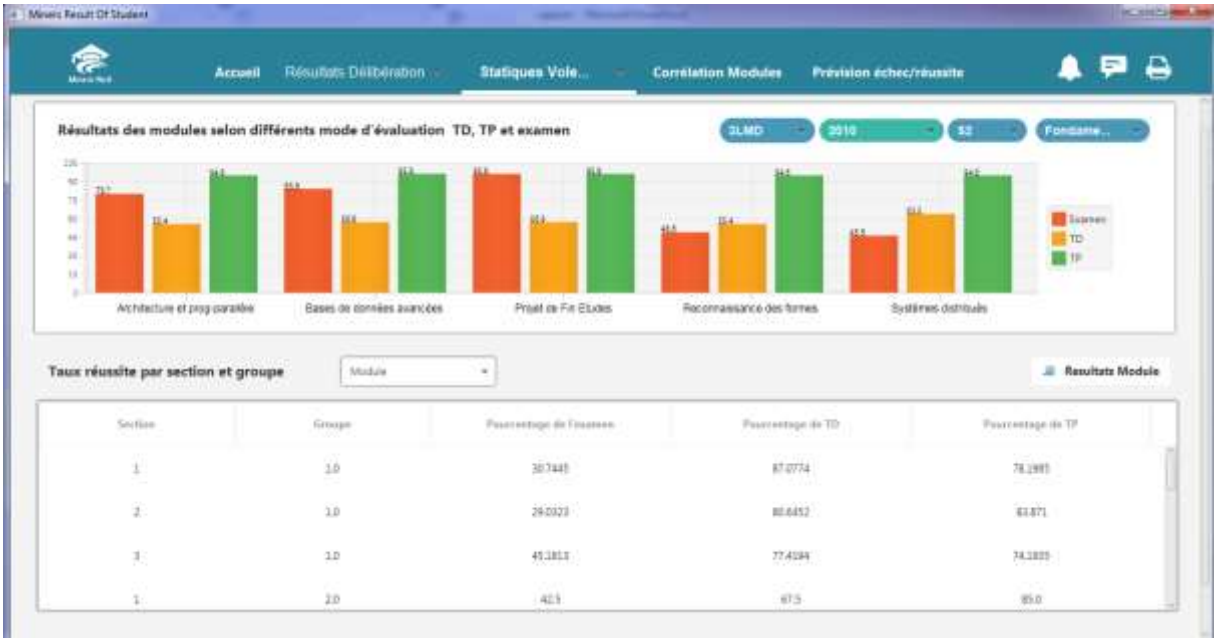


Figure 36 : Interface statistiques modules (volet 2)

- **Interface Corrélation Module :**

Cette interface montre l'implémentation de l'algorithme a priori pour identifier les règles d'association entre les différents modules selon un critère fixé par le décideur (support). Le détail de cette implémentation est montré dans le chapitre III. Ce type de connaissance peut monter aux décideurs que l'échec dans un module donné à cause des lacunes dans des modules fournisseurs dans des années précédentes.

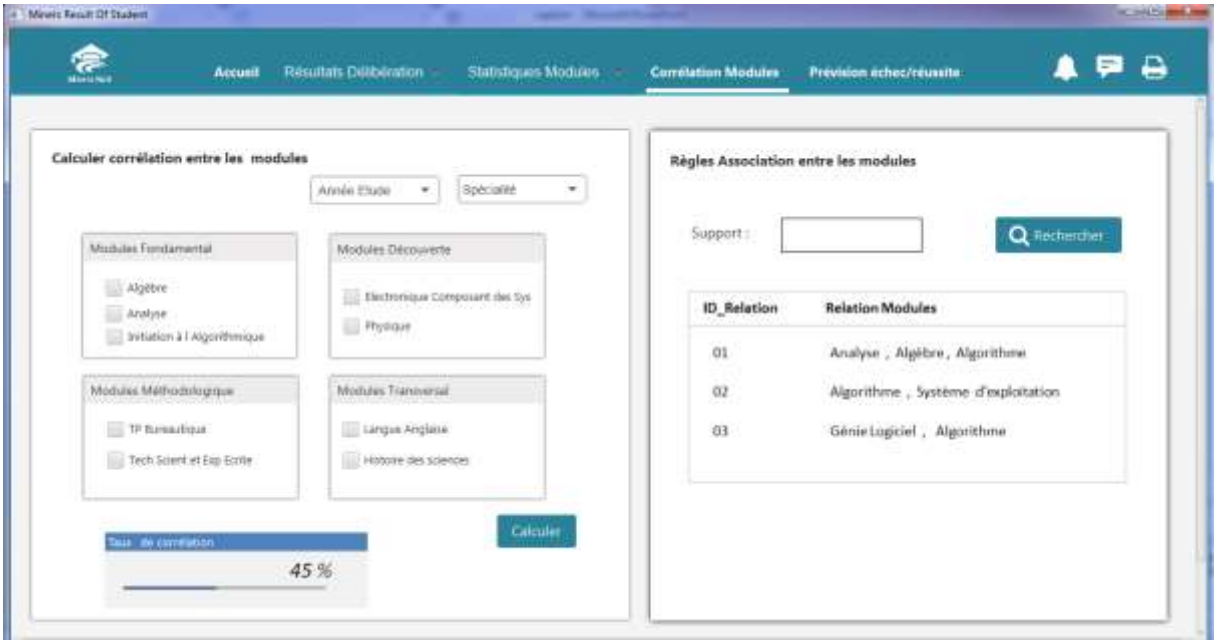


Figure 37: Interface Corrélation Module

- **Interface Prévision échec/réussite :**

Un autre usage important de notre entrepôt de données est la classification d'une nouvelle instance (un nouveau étudiants) en se basant sur l'historique de notes persistées dans l'entrepôt de données. L'algorithme KNN permet de classer cette nouvelle instance dans la classes majoritaire (Réussite / Echec) des instances similaires.

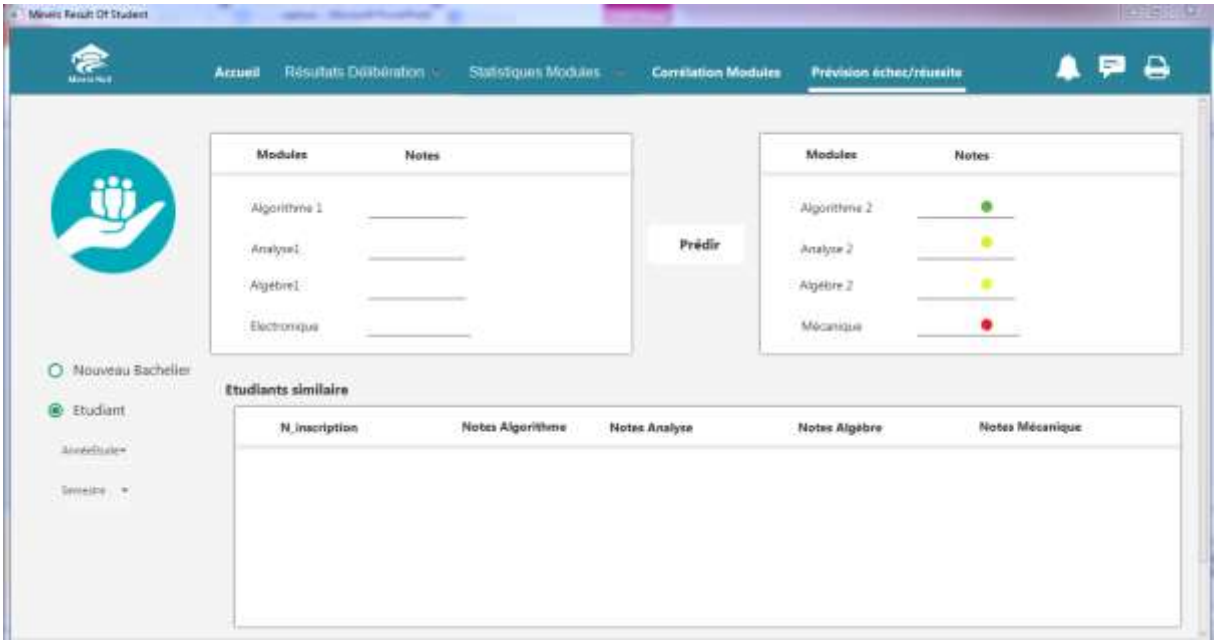


Figure 38 : Interface Prévision échec/réussite

4.3.2. Sécurité de la solution

La Sécurité de la solution représente la destruction et la perte des données de majeures menaces pour les systèmes informatiques. Ce qui nous oblige de suivre une politique de sécurité rigoureuse afin de se protéger contre les accès non autorisés. Notre politique est basée sur :

- La sécurité physique

La sécurité physique consiste à faire prendre des mesures de sécurité face aux dégâts matériels par des sauvegardes périodiques des données dans une autre machine secondaire ou des périphériques de stockage externe.

- La sécurité logique

- La sécurité logique consiste à protéger les données contre les dégâts provoquées par des logiciels tiers (les accès non autorisé, les virus) par l'installation des antivirus puissants des pare-feu contre les virus et les l'utilisation des programmes qui nuirait nos données.

- L'utilisation d'un mécanisme d'autorisation et d'authentification permettant de sécuriser l'accès des utilisateurs.

- L'utilisation des Data Mart.
- Sauvegarde hebdomadaire du DW.
- Mécanismes de gestion des erreurs et des pannes ETL

Conclusion

A travers ce chapitre nous avons abordé la mise en œuvre de notre solution. Nous avons commencé par voir les ressources utilisées, puis présenter l'architecture du système, ensuite nous avons présenté les captures de notre système et nous avons fini par une présentation de notre politique de sécurité.

Conclusion Générale

Conclusion générale

L'informatique décisionnelle est un domaine en pleine évolution pour l'aide à la prise de décisions aux seins des entreprises. Elle est devenue un besoin capital, car la simple logique de produire pour répondre à une demande, ne suffit plus pour pérenniser l'activité de celle-ci, de ce fait sont apparus les entrepôts de données.

A travers ce mémoire, nous avons montré les étapes de mise en place et de conception d'un entrepôt de données, à l'aide d'outils open source qui facilitera la prise de décisions aux seins de l'UIK et permettre aux décideurs tout comme aux utilisateurs de sélectionner et filtrer des données utiles à la prise de décision.

Notre projet consiste à concevoir et réaliser une application *DataMining* pour analyser les causes d'échecs et de réussite des étudiants afin de faciliter la lecture de ce mémoire nous avons organisé le travail en quatre parties

Dans la première partie, nous avons présenté des notions et des conceptsnous sur l'informatique décisionnelle puis nous avons vu un aperçu sur les travaux similaires afin de positionner notre travail. Nous avons commencé la deuxième partie par un état de lieu sur l'existant avec une analyse de domaine pour ressortir les besoins des acteurs de la formation, faire ressortir certaines anomalies. Puis nous avons conclu par proposer notre solution future.

Ensuite nous avons entamé la troisième partie par la conception de l'entrepôt de données en suivant l'approche du cycle de vie dimensionnel initié par Ralph Kimball l'un des pionniers dans le domaine des entrepôts de données. Par la suite, nous avons conçu l'entrepôt de données ainsi que le cube OLAP, et nous avons conclu par étudier les technique de data mining.

Enfin dans la dernière partie nous avons présenté l'architecture de déploiement ainsi nous avons cités les différentes technologies utilisées ainsi que la politique de notre sécurité.

Comme une appréciation personnelle, nous tenons à préciser qu'au cours de la réalisation de ce projet nous avons énormément appris à réaliser un outil concret en mettant en pratique nos connaissances acquises durant notre cursus à l'UIK et aussi ça nous a permis d'explorer et d'apprendre de nouvelles technologies web et outils d'analyse et de développement. En résumé, nous pouvons dire que ça été une agréable expérience qui nous a aidé à découvrir notre potentiel et à nous préparer pour continuer nos chemins vers ce qui est de mieux.

Ce travail ouvre plusieurs perspectives, nous pouvons citer :

- Considérer d'autres données sources comme les traces numériques des LMS (par ex. Moodle) et les données non structurées.
- Implémenter d'autres algorithmes de datamining comme la *clustering* de l'ensemble des étudiants selon des critères différents. Cette classification est exploitée pour affecter automatiquement les spécialistes de master de notre faculté : GL, RT et GI.
- Étendre cet entrepôt de données dans l'ère de Big Data.

Bibliographie

[1] R. Elio et al. « About computing science research methodology ». In : (2011) (cf. p. 10).
[2] J.Marie GOUARNE Le projet décisionnel, Enjeux, Modèles architecture du Data warehouse, Eyrolles 1998.
[3] Dugré, Mathieu. *Conception et réalisation d'un entrepôt de données: intégration à un système existant et étape nécessaire vers le forage de données*. Diss. Université du Québec à Trois-Rivières, 2004.
[4] GAM EL GOLLI I. Ingénierie des Exigences pour les Systèmes d'Information Décisionnels: Concepts, Modèles et Processus La méthode CADWE. Thèse de doctorat en informatique. Paris I : Université Paris I – Panthéon – Sorbonne, 2008
[6] W. H. Inmon. Building the data warehouse. J. Wiley, 2002.
[7] Khouri, S. "Modélisation conceptuelle à base ontologique d'un entrepôt de données." *Mémoire du magistère, Ecole doctorale des Sciences et Technologies de l'Information et de la Communication, Oued-Smar Alger* (2009)
[8] R. Kimball, M.Ross, «The Data Warehouse Toolkit 2nd Ed», Wiley Computer Publishing 2002.
[10] Ralph Kimball, Laura Reeves, Margy Ross etWarren Thornthwaite, Concevoir et déployer un data warehouse : Guide de conduit de projet, édition Eyrolles, octobre 2000.
[11] Adamson, Christopher. *Mastering data warehouse aggregates: solutions for star schema performance*. John Wiley & Sons, 2012.
[12] WEHRLE, Pascal. Modèle multidimensionnel et OLAP sur architecture de grille. Thèse de doctorat LIRIS - Laboratoire d'Informatique en Image et Systèmes d'information, 2009.
[13] Zighed & Rakotomalala, « Extraction des Connaissances à partir des Données (ECD) », Techniques de l'‟Ingénieur, 2002
[14] U. Fayyad, G. Piatetsky-Shapiro et P. Smyth, Data Mining to Knowledge Discovery in Databases, 1996.
[15] Bruno Agard, Andrew Kusiak, « *Exploration des bases de données industrielles à l'aide du Data Mining – Perspectives* », 9ème colloque national AIP PRIMECA, Avril 2005.
[16] Paolo GIUDICI «Applied Data Mining Statistical Methods for business and Industry» Publishing WILEY 2002.
[17] C. Bouveyron et S. Girard, Classification supervisé et non supervisé des données de grande dimension, 2009.
[18] M. Campedel, Classification supervisé, 2005.
[19] XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. ‘‘*The Top Ten Algorithms in Data Mining*’’. Ouvrage. Editions Springer-Verlag. 2007.
[20] Djoudi, M., Luengo, V., El Kechai, H., François, J., Cerisier, E. M., Cherigny, F., ... & Beust, P. (2018). Thème 2: Learning Analytics Axe 6: Terminologie du Learning Analytics.
[23] Baker, R. et Siemens, G. (2014). Educational Data Mining and Learning Analytics. Dans K. Sawyer (dir.), *Cambridge Handbook of the Learning Sciences: 2nd Edition* (p. 253-274), New York, NY : Cambridge University Press

[24] LABARTHE, Hugues et LUENGO, Vanda. L'analytique des apprentissages numériques. 2016. Thèse de doctorat. LIP6-Laboratoire d'Informatique de Paris 6.
[26] E-G Talbi. ‘‘*Fouille de données (Data Mining) - Un tour d’horizon*’’. Présentation en ligne. Laboratoire d’informatiue fondamentale de Lille (LIFL).

Webographie

Webographie

[5] GRIM Y. Passez en mode BI [en ligne]. Disponible sur : www.developpez.com

[9] http://www.univ-bejaia.dz/dspace/handle/123456789/1096

[21] accessible à http://blogs.gartner.com/matthew-davis/top-10-moments-from-gartners-supply-chain-executive-conference

[22] Daniel Peraya, « Les *Learning Analytic*s en question », *Distances et médiations des savoirs* [En ligne] URL : http://journals.openedition.org/dms/3485

[25] http://www.cue-aquitaine.fr/docs/poleetudes/Reussite-dans-enseignement-superieur

[27] Gartner https://fr.talend.com/about-us/communiques-de-presse/talend-positioned-as-a-leader-in-gartner-magic-quadrant-for-data-integration/

ANNEXE

Annexe

Gestion de projet

1. Organisation du projet

Ce projet de fin d'étude a été suivi par l'encadrant pédagogique. Notre objectif principal est d'effectuer un travail complet de master issu de l'universitaire Ibn Khaldoun. Pour assurer le suivi de notre projet, nous envoyons chaque semaine un compte rendu pour lister les tâches effectuées, les actions à réaliser, les comptes rendus des réunions effectuées. Le rôle de l'encadrant intervient dans la phase d'exploration des concepts théoriques, les orientations durant les principales phases du projet et dans la dernière phase de rédaction du rapport final du projet. Nous nous sommes basés sur une méthode de travail qui nécessite le contact hebdomadaire avec nos encadreurs. Ce contact se traduit par les différents types des réunions.

Le tableau suivant représente le plan de communication utilisé lors du projet :

Type de contacte	Durée	Participant	Objectif
Mails hebdomadaires		Étudiants Co-Encadreur	échange d'information
Réunion de discussion	20mn à 1h	Étudiants Co-Encadreur	Résoudre des problèmes
Réunion de présentation du travail	30mn à 2h	Étudiants Co-Encadreur Encadreur	Présentation du travail effectué
Réunion de correction et validation	40mn à 2h	Étudiants Co-Encadreur Encadreur	Validation du, travail effectuée

Tableau 16 : Le contact hebdomadaire avec l'encadreur

1.1.Réunions

Dans cette partie, nous présentons un exemplaire de compte rendu des réunions que nous avons effectué avec les experts du domine. Le but des réunions effectuées étant de conduire une étude du système actuel et récolter les besoins des utilisateurs finaux.

Annexe

Compte Rendu - Réunion N°1 avec l'expert

1. Informations sur la réunion

Date : /05/2019

Heure : 10 :00 :00

Lieu : Tiaret

Service : Centre d'Orientation

Interviewé(s) : Mr. Kedroussi Mohamed (directeur du centre)

Rapporteurs : Chelik Nassima et Ould Ali Nadia

Objectifs :

- Récupérer les données
- Analyser l'existant du projet

2. Compte rendu

2.1. Points de l'ordre du jour

Questions sur les besoins et l'existant du service orientation, les questions posées sont présentées ci-dessous.

- **Q1** : Quels sont vos sources principales de données ?
- **Q2** : Quels sont les données sur lesquelles vous travaillez ?
- **Q3** : Quels sont les séries du baccalauréat qui vous intéressent ?
- **Q4** : Quels sont les années scolaire du baccalauréat existantes ?
- **Q5 :** Pouvez-vous nous donner les nouveaux bacheliers de 2010 à 2017 inscrits en informatique ?

2.2. Discussion

Les réponses apportées par Mr. Kedroussi, aux questions précédentes sont présentées ci-dessous.

R1 : Nous avons principalement une source de données qui est la base de données, les données extraites, de ce dernier, sont sous format de fichiers Excel.

R2 : Nous travaillons principalement sur les données des élèves (nouveaux bacheliers).

R3 : Toutes les séries du baccalauréat des élèves nous intéressent.

R4 : Nous disposons que de quatre années (2014 à 2017).

R5 : Nous pouvons vous fournir l'ensemble des résultats, il vous appartient de les trier.

Annexe

Les livrables que nous allons énumérer dans cette section ont sont été partagés avec nos encadreurs pour assurer le suivi de notre projet.

1.2.1. Documentation

Tout au long du projet, nous avons réalisé des livrables intermédiaires qui ont permis aux encadrants d'être à jour avec les différentes fonctionnalités proposées, les livrables intermédiaires sont ceux suivants :

- le plan des chapitres
- La conception du système
-

1.2.2. Rapport de PFE

Il permet de mieux comprendre le raisonnement, le fonctionnement du système, son implémentation ainsi que l'évaluation des performances de l'architecture finale conçue.

2. Planning prévisionnel

Les étapes principales du projet sont celles suivantes.

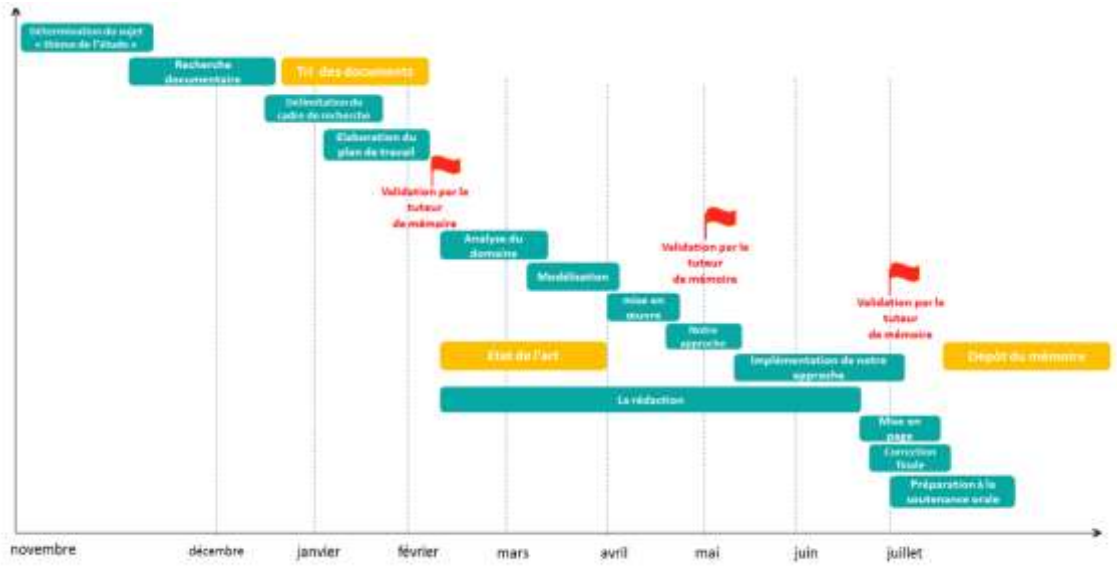


Figure 39 : Diagramme de Gantt de notre thèse.