

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA
RECHERCHE SCIENTIFIQUE

UNIVERSITÉ AMAR TELIDJI DE LAGHOUAT
FACULTÉ DES SCIENCES
DÉPARTEMENT D'INFORMATIQUE

Mémoire de MASTER

Domaine : Mathématiques et Informatique
Filière : Informatique
Option : Systèmes d'Information et de Décision

Présenté par : Sarah BOUDOUH & Hadjer GUELLOUMA

**Conception et mise en œuvre d'un entrepôt
de données pour l'analyse du cursus
universitaire des étudiants**

Soutenu publiquement le...../...../..... Devant le jury composé de :

Mme. B. Rerrouche	MA(A)	Présidente
Mr. Y. Guelloma	MC(B)	Examineur
Mr. B. Ziani	MC(A)	Encadreur

Résumé

La qualité de la formation dans le secteur de l'enseignement supérieur constitue un enjeu fondamental pour la société. L'analyse des résultats obtenus par les étudiants durant leur cursus universitaire est l'un des principaux déterminants de l'évaluation de la qualité de l'enseignement. Le travail présenté dans ce mémoire a pour objectif de poser quelques jalons pour identifier les facteurs susceptibles d'influencer le parcours des étudiants à l'université Algérienne.

Au contraire des études empiriques traditionnelles basées sur les questionnaires, notre démarche s'appuie sur une technique d'entrepôt de données. L'objectif principal est la conception d'un entrepôt de données pour pouvoir intégrer toutes les données relatives au parcours universitaire des étudiants dans un système qui servira de support à des outils d'analyse. L'entrepôt de ces données constitue une étape importante pour offrir aux décideurs de nouvelles connaissances utiles identifiant des facteurs explicatifs de la réussite et de l'échec chez les étudiants. Ceci va les aider à mettre en place des mesures améliorant la qualité de la formation et favorisant par conséquent la réussite des étudiants.

Notre démarche se compose de deux phases de réalisation. La première est une phase conceptuelle pour la structuration et la modélisation des données d'intérêt sous la forme d'une constellation de faits et de dimensions. La deuxième phase est technique et concerne la création et l'alimentation de l'entrepôt grâce à un processus ETL (Extract Transform Load) pour l'extraction, la transformation et chargement des données. Les données sources sont issues de l'archive du département d'informatique de l'université Amar Télidji de Laghouat. Enfin, pour valider notre proposition, nous avons présenté les résultats de quelques requêtes analytiques OLAP.

Mots clés : Systèmes décisionnels, entrepôt de données, modélisation multidimensionnelle, schéma en étoile, requêtes OLAP, cursus universitaire.

Abstract

The imperative for countries to improve employment skills calls for quality teaching within educational institutions. The analysis of the results obtained by the students during their university course is one of the main determinants of the evaluation of the quality of teaching. The work presented in this thesis aims to set some milestones to identify the factors likely to influence the course of students at the Algerian University.

Choosing reliable and quantifiable indicators to assess the quality of one's teaching and the efficiency of teaching initiatives remains challenging. Unlike traditional empirical studies based on questionnaires, our approach is based on a data warehousing technique. The main objective is the design of a data warehouse to be able to integrate all the data relating to the university course of the students in a system which will be used to support analytical tools. Warehousing this data is an important step in providing decision-makers with new, useful knowledge that identifies factors explaining student success and failure. This will help making decisions that improve the quality of training and consequently promote student success.

Our approach consists in a two phases process. The first is a conceptual phase for structuring and modeling the data of interest in the form of a constellation of facts and dimensions. The second phase is technical and concerns ETL (Extract-Transform-Load) process. The data source are extracted from the archive of the computer science department of Amar Telidji University in Laghouat. Finally, to validate our proposal, we presented the results of some analytical OLAP queries.

Keywords : Decision systems, data warehouse, multidimensional modeling, star schema, OLAP queries, university course.

Remerciements

Nos remerciements s'adressent en premier lieu à **Allah** qui nous a donné le courage, la force et les moyens pour réussir et donner le meilleur de nous-mêmes.

Nous voudrions ensuite remercier notre encadreur Benameur ZIANI pour la confiance qu'il nous a accordé en acceptant d'encadrer ce travail. Nous lui sommes reconnaissantes de nous faire bénéficier tout au long de ce travail de sa compétence, son encouragement et de ses remarques très constructives. Merci pour sa grande disponibilité pour ses relectures et ses corrections qui ont permis à notre document de devenir ce qu'il est.

Nous tenons à remercier également Mme Badra Kerrouche pour sa disponibilité et ses remarques pertinentes pour la partie conceptuelle de notre projet.

Nous sommes très honorées à remercier les membres du Jury pour avoir accepté d'évaluer et d'enrichir ce travail. Nous vous témoignons notre grand respect et notre reconnaissance infinie.

Merci à tous les enseignants qui ont contribué de près ou de loin à notre formation. Enfin, nous adressons nos plus sincères remerciements à nos familles, nos parents, nos frères et sœurs et tous nos proches et amis, qui nous ont accompagné, aidé, soutenu et encouragé tout au long de la réalisation de ce mémoire.

Dédicaces

Je dédie ce modeste travail :

En premier lieu à mes chers parents pour leur soutien, leurs sacrifices, leurs prières, et leur amour tout au long de mon parcours.

À mon cher frère, pour ses encouragements constants et son soutien moral.

À mes chères tantes et cousines.

À ma seule et meilleure amie, et mon binôme Hadjer.

Merci d'être toujours là pour moi.

S.S. BOUDOUH

Je dédie ce modeste travail :

En premier lieu à mon père qui n'a jamais cessé de m'encourager pour terminer ma formation et de me soutenir.

À ma très chère mère, qui a m'as comblé avec ses prières, ses sacrifices tout au long de mon parcours.

À mes chères sœurs, et mes chers frères.

À ma seule et meilleure amie, et mon binôme Sarah.

H. GUELLOUMA

Table des matières

1	Introduction générale	8
2	Notions de base sur les entrepôts de données	10
2.1	Introduction	10
2.2	Un peu d'histoire	10
2.3	Définition d'un entrepôt de données	12
2.4	Architecture d'un entrepôt de données	13
2.4.1	Extraction des données	14
2.4.2	Transformation des données	14
2.4.3	Chargement des données dans l'entrepôt	14
2.5	Concepts de Fait et Dimension	14
2.5.1	Concept de Fait	14
2.5.2	Concept de dimension	15
2.6	Modélisation multidimensionnelle	15
2.6.1	Schéma en étoile	16
2.6.2	Schéma en flocon de neige	16
2.6.3	Schéma en constellation de faits	17
2.7	Implémentation du modèle multidimensionnel	17
2.7.1	ROLAP (Relational OLAP)	19
2.7.2	MOLAP (Multidimensional OLAP)	19
2.7.3	HOLAP (Hybrid OLAP)	19
2.8	Analyse et restitution des données	19
3	Étude conceptuelle	21
3.1	Introduction	21
3.2	Brève présentation du système LMD	21
3.3	Système opérationnel actuel	22
3.4	Notre conception	23
3.4.1	Les faits	24
3.4.2	Les dimensions	24
3.4.3	Schéma en constellation	25
4	Implémentation et validation	27
4.1	Introduction	27
4.2	Tout dépend des données !	27
4.3	Création des tables de l'entrepôt	28
4.4	Processus ETL	28

4.4.1	Données sources	28
4.4.2	Transformation des données sources	30
4.4.3	Chargement des données transformées	32
4.5	Exploitation de l'entrepôt : résultats préliminaires	32
4.5.1	Quelques aspects d'analyse	32
4.5.2	Discussion des résultats	33
5	Conclusions et perspectives	35

Table des figures

2.1	Des données aux connaissances.	11
2.2	Principe d'un entrepôt de données.	11
2.3	Architecture d'un entrepôt de données.	13
2.4	Exemple de fait.	15
2.5	Exemple de dimensions.	15
2.6	Exemple d'un un schéma en étoile.	16
2.7	Exemple d'un un schéma en flocon de neige.	17
2.8	Exemple d'un un schéma en constellation.	18
3.1	Les formations dans le système LMD.	21
3.2	Poids des UE dans le système LMD.	22
3.3	Schéma en constellation de faits de notre conception.	25
4.1	Script de création de la table Coursus-Module.	29
4.2	Extrait des données brutes (Étudiants).	29
4.3	Extrait des données brutes (Résultats).	30
4.4	Extrait des données brutes (Modules).	31
4.5	Scripts de chargement des données dans les tables des faits.	32

Liste des tableaux

2.1	Différences entre les modèles OLTP et OLAP.	12
4.1	Compensation des modules par UE.	33
4.2	Parcours vs. Sexe.	33
4.3	Parcours vs. Mention Bac.	33

Chapitre 1

Introduction générale

La croissance sans cesse et la profusion des capacités de stockage ont contribué à l'explosion du volume de données recueillies dans les entreprises. La gestion et l'exploitation de ces gros volumes de données à des fins d'analyse deviennent depuis plusieurs années un problème crucial pour les analystes et les gestionnaires.

Afin de répondre à ce problème, de nouvelles méthodes et technologies connues sous le terme d'informatique décisionnelle, ont vu le jour. L'informatique décisionnelle est un concept apparu dans les années 90 visant le traitement et la valorisation des données à des fins de compréhension et de la prise de décision.

Les entrepôts de données constituent le cœur des systèmes décisionnels. Ils intègrent les données en provenance des différents systèmes transactionnels, souvent répartis et hétérogènes. Les entrepôts de données ont pour objectif de transformer les données collectées en informations utiles appelées connaissances qui sont utilisées pour prendre des décisions stratégiques.

A travers le processus d'analyse en ligne (On-Line Analytical Processing - OLAP), de tels systèmes servent à établir des requêtes, des rapports et des analyses permettant d'améliorer les performances décisionnelles de l'entreprise. Concrètement, ils sont utilisés pour faire des statistiques, des calculs prévisionnels, des analyses des tendances et donc pour établir des stratégies de développement de l'entreprise.

Depuis leur introduction, les entrepôts de données ont suscité beaucoup d'intérêt et deviennent rapidement une partie intégrante des organismes dans différents secteurs tels que la grande distribution, les banques, les assurances, l'industrie de l'automobile et les institutions médicales[1, 2].

Le secteur de l'enseignement supérieur représente sans aucun doute l'un des importants piliers de la société avec ses répercussions sur le développement de tous les secteurs. Les questions liées à la qualité de la formation dans ce secteur constituent un enjeu majeur pour l'ensemble des responsables dans le domaine. L'analyse des résultats obtenus par les étudiants durant leur cursus universitaire est l'un des principaux déterminants pour l'évaluation de la qualité de l'enseignement.

Malgré la popularité et la prolifération des entrepôts de données pour la prise de décision dans le monde des affaires, le secteur de l'enseignement supérieur en Algérie semble être encore à la marge de cette technologie. En effet, les services pédagogiques concernés sont dotés de systèmes transactionnels dédiés qui se limitent à la gestion de la saisie des notes et l'édition des procès verbaux de délibérations.

Le constat dégagé est qu'il n'existe, à notre connaissance, aucun système décisionnel

visant l'analyse des données collectées durant le cursus universitaire des étudiants. De plus, les études académiques qui ont abordé le problème de l'évaluation des systèmes pédagogiques dans tous les niveaux, du primaire à l'université, sont souvent des études empiriques basées sur des questionnaires[3, 4, 5, 6, 7]. De telles études ne peuvent être fiables parce que, d'un côté, elles ne sont pas basées sur des données réelles. De l'autre côté, les réponses aux questionnaires sont souvent subjectives et ne peuvent être considérées comme une base solide pour la prise de décision.

Ceci nous a motivé au problème de l'analyse des données émanant du parcours des étudiants universitaires. C'est un sujet qui constitue un cadre applicatif particulièrement intéressant pour notre formation en Master sur les SYSTÈMES D'INFORMATION DÉCISIONNELS (SID).

La masse de données collectée, peu ou pas exploitée par les systèmes pédagogiques transactionnels classiques, représente une source précieuse d'informations pour l'évaluation du parcours pédagogique des étudiants. En effet, elle permet de décrire le comportement des étudiants avec davantage de précisions. Toutefois, l'exploitation de ces données nécessite des systèmes et des outils adaptés pour en permettre une analyse optimisée par les décideurs.

Le travail présenté dans ce mémoire aborde les problèmes associés à ce domaine d'application. L'objectif principal est de procéder à l'intégration de toutes ces données dans un entrepôt qui servira de support à des outils d'analyse. Ceci va permettre aux décideurs de faire un suivi et des analyses utiles afin d'en tirer des actions à mener pour améliorer la qualité de la formation et favoriser par conséquent la réussite des étudiants.

Ce manuscrit est articulé autour de quatre chapitres :

- Le **premier chapitre** constitue une introduction à notre travail de mémoire. L'objectif de ce chapitre est de présenter la problématique abordée et la motivation derrière le choix du sujet de ce mémoire.
 - Le **second chapitre** est consacré à la présentation des notions de base sur les entrepôts de données. Nous décrivons en particulier l'architecture et la modélisation multidimensionnelle des entrepôts de données.
 - Le **troisième chapitre** est consacré à la première phase de notre projet. Il décrit notre vision conceptuelle pour la structuration et la modélisation des données d'intérêt sous la forme d'une constellation de faits et de dimensions.
 - Le **quatrième chapitre** présente la deuxième phase de notre projet. C'est une phase technique et concerne la création et l'alimentation de l'entrepôt grâce à un processus ETL (Extract Transform Load) pour l'extraction, la transformation et chargement des données. Ce chapitre comporte également une partie expérimentale pour valider notre approche.
 - Le manuscrit s'achève par une conclusion et quelques perspectives envisagées.
- .
-

Chapitre 2

Notions de base sur les entrepôts de données

2.1 Introduction

Ce chapitre introduit les notions de base sur les entrepôts de données. Nous présentons dans un premier temps les définitions relatives à ce contexte. Nous abordons, ensuite, les termes et les concepts clés liés à l'architecture et la modélisation d'un entrepôt de données.

2.2 Un peu d'histoire

Les systèmes de bases de données ont considérablement évolué depuis leur apparition dans les années 70 du dernier siècle. De nos jours ils sont utilisés partout. Il s'agit notamment des données de ventes, de médecine, des finances, d'Internet et des domaines sociaux, représentant toutes les activités de notre société. Cette omniprésence des bases de données a contribué à l'explosion du volume de données recueillies dans les entreprises. Cependant, les données collectées ne sont pas nécessairement exploitées par la suite.

En effet, les systèmes de gestion de bases de données (SGBD) sont basés sur le modèle OLTP (On Line Transaction Processing) qui leur permet la prise en charge du traitement quotidien d'une organisation. Le but principal de tels systèmes est de stocker, d'organiser et d'interroger les données opérationnelles de l'entreprise[8, 9]. Ils ont prouvé leur performance dans l'automatisation de différentes tâches transactionnelles telles que la gestion de la paye du personnel, la gestion des stocks de produits, la gestion de la comptabilité..etc. Cependant, ces systèmes sont particulièrement adaptés au traitement de petits volumes de données. De plus, ils ne permettent pas d'effectuer efficacement les opérations d'analyse de données[10].

Si les données ne sont plus, comme dans le passé, une ressource rare, c'est leur analyse qui permet d'en extraire des indicateurs d'aide à la décision, communément appelés **connaissances**. Cependant, le passage des données vers des **connaissances** n'est pas une tâche facile et les entreprises sont souvent caractérisées comme étant riches en **données** mais pauvres en **connaissances** [11]. La figure 2.1 illustre le passage des données aux connaissances synonyme d'acquérir de la compétence. Il est important de préciser la distinction entre les termes donnée, information et connaissance [12] :

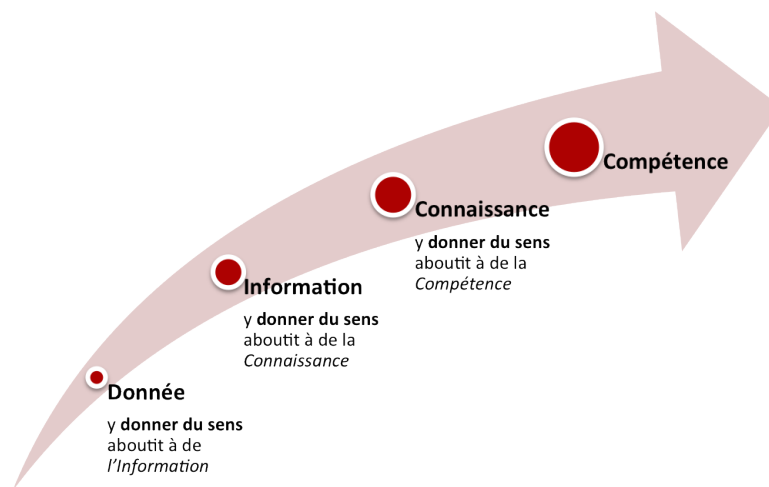


FIGURE 2.1 – Des données aux connaissances

- **Les données** sont les résultats de l'observation. Par exemple le nombre d'unités achetées pour un produit donné,
- **Les informations** sont les résultats de l'analyse des données. Par exemple, la répartition géographique de tous les produits achetés,
- **Les connaissances** définissent les informations utiles pour la prise de décision. Par exemple, la découverte du mauvais emplacement de certains magasins de vente.

La détention de la connaissance est l'une des plus importantes tâches de l'entreprise moderne. Elle permet aux organisations de mieux comprendre leur environnement et donc à prédire les évolutions futures afin de s'y adapter et être le plus efficace et le plus compétitif possible.

Ainsi, pour une bonne prise de décision, les gestionnaires ont besoin de connaissances appropriées, au moment opportun. Ceci devient désormais possible car le flot continu et croissant des données collectées peut maintenant être stocké et exploité grâce aux nouvelles techniques d'Entrepôts de Données (Data Warehouse en Anglais).

Les entrepôts de données sont donc des systèmes spécifiques entièrement dédiés à des utilisateurs particuliers qui sont les **décideurs** (Figure 2.2). Ces derniers ont besoin d'en faire ressortir des tendances (connaissances) leur permettant de prendre des décisions stratégiques.

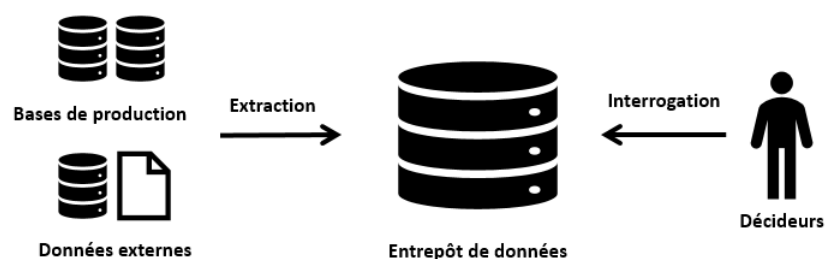


FIGURE 2.2 – Principe d'un entrepôt de données

Reposant sur le système OLAP (On Line Analytical Processing), les entrepôts de données ont apporté une solution adéquate et efficace au problème de la collecte de données massives [13]. Un entrepôt de données permet de stocker, sur de longues périodes, des données sur les activités d'un organisme, ce qui permet de les étudier ultérieurement. Par exemple, l'historisation des ventes d'un produit dans une entreprise permet d'étudier son évolution dans le temps afin mettre au point des stratégies commerciales ou des campagnes de promotion. Toutefois, le problème majeur de tels systèmes est leur complexité dans la mise en œuvre des données qui nécessite beaucoup du temps de réalisation. Le tableau 2.1 récapitule les principales différences entre les modèles OLTP et OLAP [14].

	OLTP	OLAP
Utilisation	SGBD	Entrepôt de données
Opération typique	Mise à jour	Analyse
Type d'accès	Lecture écriture	Lecture
Niveau d'analyse	Elémentaire	Global
Quantité d'informations échangées	Faible	Importante
Taille de la base	Faible (quelques Go)	Importante (plusieurs To)
Ancienneté des données	Récente	Historique

TABLE 2.1 – Différences entre les modèles OLTP et OLAP [14]

2.3 Définition d'un entrepôt de données

Le concept d'entrepôt de données a été formalisé pour la première fois par W. Inmon [15] comme étant une *"collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support du processus d'aide à la décision"*. Cette définition met l'accent sur les caractéristiques suivantes [15] :

- **Orientées sujet** : Les données des entrepôts sont organisées par thème plutôt que par application. Par exemple, une chaîne de magasins d'alimentation organise les données de son entrepôt par rapport aux ventes qui ont été réalisées par produit et par magasin, au cours d'un certain temps.
- **Intégrées** : Les données provenant des différentes sources doivent être intégrées, avant leur stockage dans l'entrepôt de données. L'intégration (mise en correspondance des formats, par exemple), permet d'avoir une cohérence de l'information.
- **Non volatiles** : A la différence d'une base de données classique supportant des requêtes transactionnelles de type OLTP (On-Line Transaction Processing), un entrepôt de données est conçu pour supporter des requêtes de type OLAP (On-Line Analytical Processing). L'interrogation est l'opération la plus utilisée dans le contexte d'entrepôt de données où la mise à jour consiste seulement à alimenter l'entrepôt. Les données de l'entrepôt sont ainsi permanentes et ne peuvent pas être modifiées. Le rafraîchissement de l'entrepôt consiste à ajouter de nouvelles données, sans modifier ou perdre celles qui existent. On y intègre seulement de nouvelles données datées qui viennent s'ajouter aux précédentes.

- **Historiées** : La prise en compte de l'évolution des données est essentielle pour la prise de décision. Par exemple, on utilise des techniques de prédiction en s'appuyant sur les évolutions passées pour prévoir les évolutions futures.

2.4 Architecture d'un entrepôt de données

La figure 2.3 illustre une représentation simplifiée de l'architecture d'un entrepôt de données[10]. Au niveau des traitements de données, on peut observer que cette architecture est caractérisée par deux grandes phases qui sont (1) l'intégration des données (processus ETL) et (2) l'analyse des données (reporting).

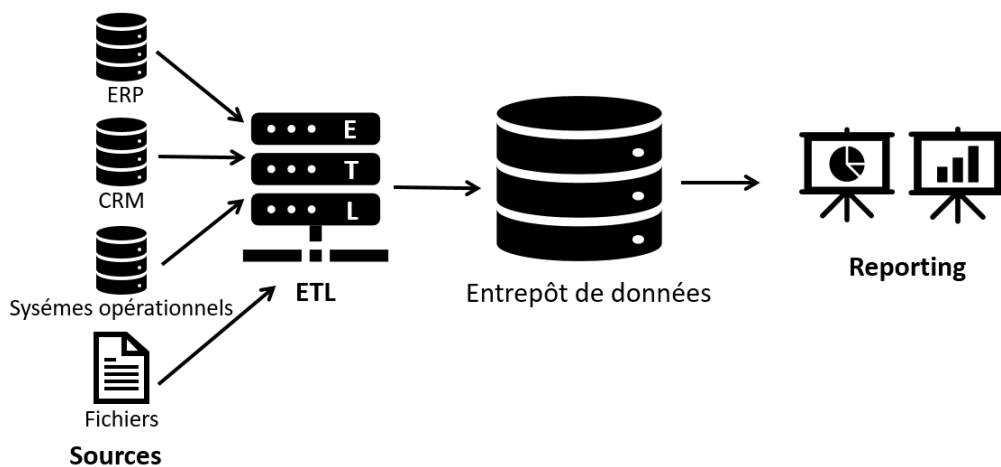


FIGURE 2.3 – Architecture simplifiée d'un entrepôt de données

Avant de procéder à l'intégration et l'analyse des données, il faut identifier les sources des données de l'entrepôt. Rappelons que les sources de données sont souvent hétérogènes car elles ont été définies indépendamment les unes des autres. Elles ont par conséquent des schémas et des formats différents. A titre d'exemple, la figure 2.3 indique qu'il peut s'agir des données issues d'un système de gestion des clients (en Anglais CRM : customer relationship management), ou d'une base de données assurant la gestion de l'ensemble de l'activité d'une entreprise (gestion comptable, gestion commerciale, gestion des stocks). Cette dernière est connue par l'acronyme ERP (Entreprise Resource Planning). Par conséquent, les données peuvent être de plusieurs types : structurées (données relationnelles, données objet), semi-structurées (HTML, XML, graphes) ou même non structurées (texte, images, son).

Une fois les sources de données identifiées, on peut procéder au processus d'intégration de données dans l'entrepôt. Ce processus est connu sous l'acronyme ETL (Extract-Transform-Load) et est composé de trois étapes[10, 15] :

1. Extraction des données à partir de leurs sources,
2. Transformation des données,
3. Chargement des données dans l'entrepôt.

La réalisation de l'ETL est une étape très importante et très complexe. Elle constitue environ 70% d'un projet de mise en place d'un système décisionnel[16].

2.4.1 Extraction des données

Cette étape est responsable de l'identification des données dans les diverses sources et d'en extraire les informations d'intérêt par rapport aux analyses envisagées. La difficulté principale dans cette étape est que certaines données peuvent être manquantes. Ce phénomène se manifeste lorsque les valeurs de certaines données n'ont pas pu être observées, elles ont été perdues ou elles n'ont pas été enregistrées. Dans d'autres situations, l'archive pour les données n'est pas du tout disponible. Ceci constitue une problématique sur la qualité des données utilisées dans le processus d'analyse[17]. En effet, le manque de certaines données peut mener à une perte de précision et induire à de fausses conclusions analytiques.

2.4.2 Transformation des données

Comme noté auparavant, les données d'intérêt identifiées dans la phase précédente sont généralement issues de sources différentes et donc très hétérogènes. La phase de transformation consiste à la corrections des inconsistances et la mise au format des données pour les rendre prêtes au processus d'analyse. A titre d'exemple, on peut procéder dans cette phase à la suppression des doublons ou à la correction des conflits de structures : différentes appellations pour un même objet. Dans la même optique, le traitement des valeurs manquantes ou nulles est une tâche importante à réaliser. On peut procéder également à des opérations de tri, de filtrage, d'agrégation, de regroupement ou encore de jointure pour pallier les différences dans les sources et générer ainsi des données cohérentes pour l'entrepôt.

2.4.3 Chargement des données dans l'entrepôt

C'est la dernière étape qui permet de charger les données dans l'entrepôt. Deux types de chargement sont appliqués aux données : un chargement initial et un chargement périodique (rafraîchissement) qui correspond à l'ajout des nouvelles données à l'entrepôt existant afin de garantir des données actualisées pour le processus de prise de décision. Considérant le volume des données traitées, l'opération de chargement, et tout particulièrement le chargement initial, risque d'être assez longue. Il est donc nécessaire de mettre en place une stratégie pour assurer de bonnes conditions à leur réalisation. On peut procéder à une alimentation progressive de l'entrepôt pour éviter de faire transiter des gigaoctets, voir plus, de données vers les tables de l'entrepôt.

2.5 Concepts de Fait et Dimension

Les systèmes d'entrepôts de données reposent sur le paradigme multidimensionnel qui est basé sur deux concepts clés : **le fait** et **la dimension**[10, 15].

2.5.1 Concept de Fait

Le fait est un thème ou sujet à analyser. La table des faits regroupe donc un ensemble de données observables appelés **mesures** que l'on souhaite analyser selon plusieurs points de vues. Les mesures sont souvent des valeurs numériques permettant de réaliser des opérations d'agrégation : addition, minimum, maximum, moyenne...etc.

- **Exemple :** Considérons le fait **Vente** représenté par la figure 2.4. Le thème à analyser est alors l'activité des **ventes** réalisées par l'entreprise. Les analyses sont effectuées en observant les mesures **Quantité** et **Montant**.

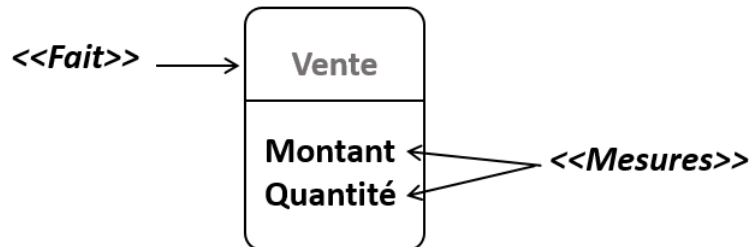


FIGURE 2.4 – Exemple de fait

2.5.2 Concept de dimension

Une **dimension** est un axe d'analyse d'un fait. Les dimensions désignent alors le contexte d'analyse du fait. Une table de dimension contient des attributs, appelés **paramètres**, qui représentent différents niveaux de détail selon lesquels sont visualisées les mesures d'un fait.

- **Exemple :** Le fait **Vente** de l'exemple précédent (figure 2.4) peut être analysé selon différentes perspectives. La figure 2.5 illustre quatre dimensions d'analyse : **Temps**, **Magasin**, **Produit** et **Client**. L'analyse des ventes selon la dimension **Produit**, par exemple, implique les paramètres **Marque** et **Catégorie**.

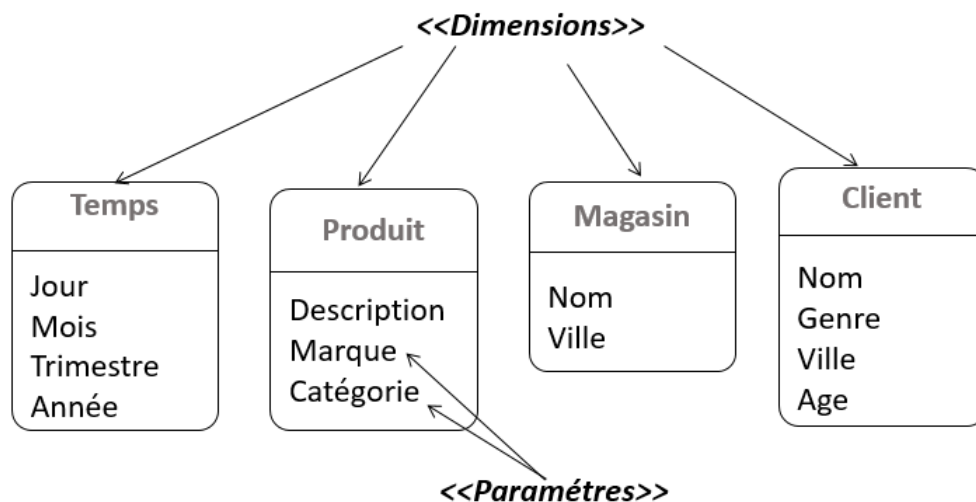


FIGURE 2.5 – Exemple de dimensions

2.6 Modélisation multidimensionnelle

Afin de permettre des analyse selon différents points de vues, les faits et les dimensions sont reliés par des clés. Ainsi, en plus des attributs représentant les mesures d'acti-

tivité, la table des faits contient les attributs clés étrangères qui la relient vers chacune des tables de dimension associée à un axe d'analyse. Chaque table de dimension est dotée d'une clé primaire permettant de réaliser des jointures avec la table des faits. Ces relations sont organisées selon trois types de schéma : schéma en étoile, schéma en flocon et schéma en constellation.

2.6.1 Schéma en étoile

C'est un schéma relationnel dans lequel une table centrale contenant les faits analysés référence des tables de dimensions. Chaque dimension est décrite par une seule table dont les attributs représentent les divers niveaux de granularité. Un exemple d'un schéma en étoile est illustré dans la figure 2.6. La modélisation en étoile a l'avant-

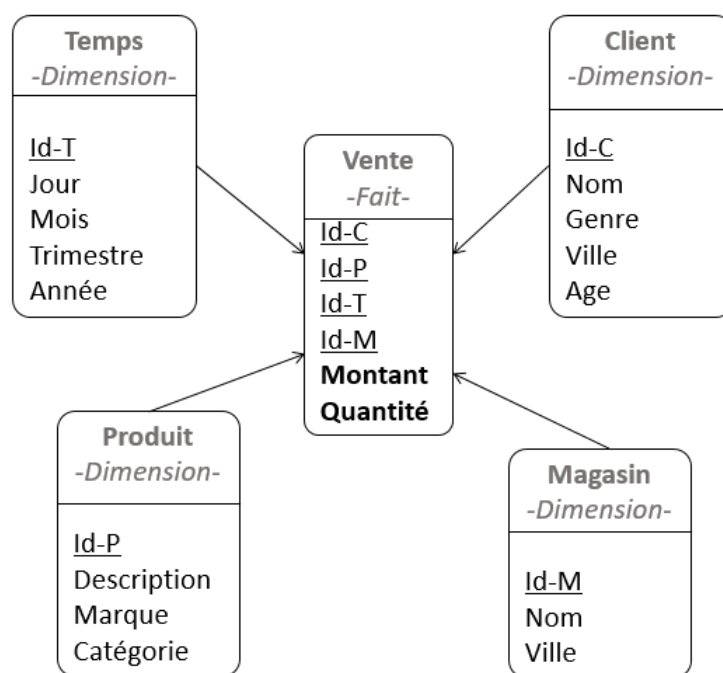


FIGURE 2.6 – Exemple d'un schéma en étoile

tage d'être simple où la table des faits est reliée à chacune des tables de dimensions par une relation, une jointure. Cela simplifie considérablement le traitement des requêtes. En revanche, la modélisation en étoile a un gros inconvénient qui est la redondance des données ce qui entraîne une dé-normalisation. À cause de cette redondance, la modélisation en étoile utilise plus d'espace de stockage que les autres modèles. Cela augmente également le risque d'atteinte à l'intégrité du système[18].

2.6.2 Schéma en flocon de neige

C'est une variété du schéma en étoile où chaque dimension est décrite par une succession de tables représentant les diverses granularités ce qui donne naissance à des hiérarchies au sein des dimensions. Un exemple d'un schéma en étoile est illustré dans la figure 2.7. Ce modèle minimise le volume des tables à joindre lors de l'exécution

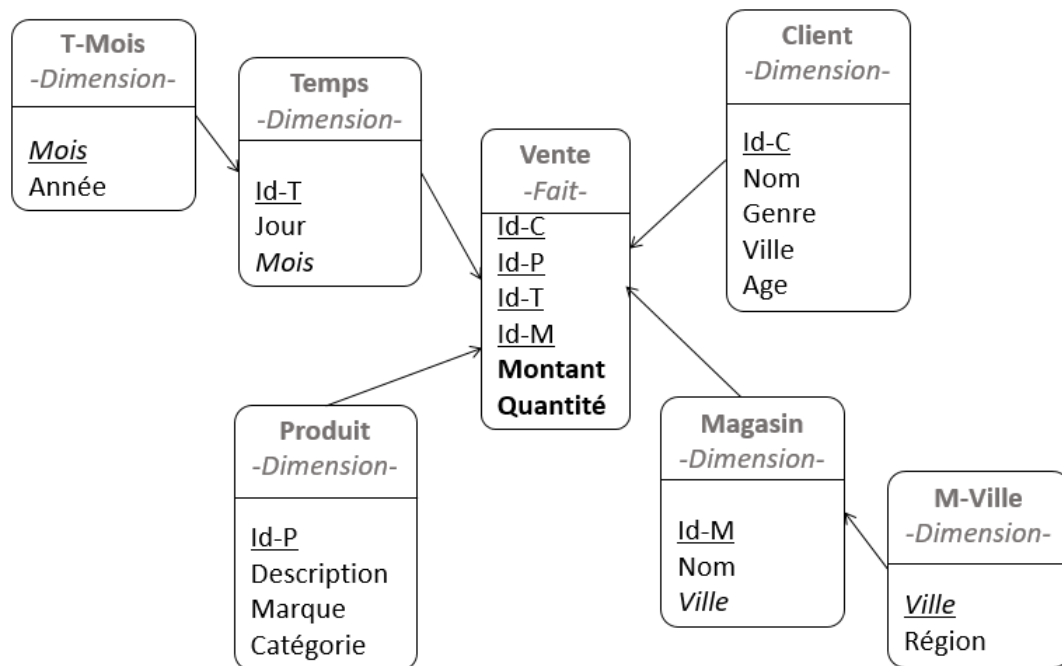


FIGURE 2.7 – Exemple d'un schéma en flocon de neige

des requêtes par rapport au schéma en étoile. par contre, ce modèle est plus complexe que le schéma en étoile car comportant un nombre plus élevé de tables[18].

2.6.3 Schéma en constellation de faits

Ce schéma est une extension du schéma en étoile où plusieurs thèmes, et non pas un seul, sont à analyser. Un schéma en constellation comprend donc plusieurs tables de faits reliés à un ensemble de dimensions qui peuvent être partagées. L'avantage pour ce type de modélisation est de pouvoir réaliser différentes analyses en évitant de définir les dimensions plusieurs fois. Un exemple d'un schéma en constellation est illustré dans la figure 2.8. Ce modèle a les mêmes avantages et inconvénients que le schéma en étoile[18].

2.7 Implémentation du modèle multidimensionnel

L'implémentation du modèle multidimensionnel sur un SGBD réel peut se faire selon trois principales solutions[10, 15] :

1. **ROLAP** (Relational On- Line Analytical Processing),
2. **MOLAP** (Multidimensional On-Line Analytical Processing),
3. **HOLAP** (Hybrid On-Line Analytical Processing).

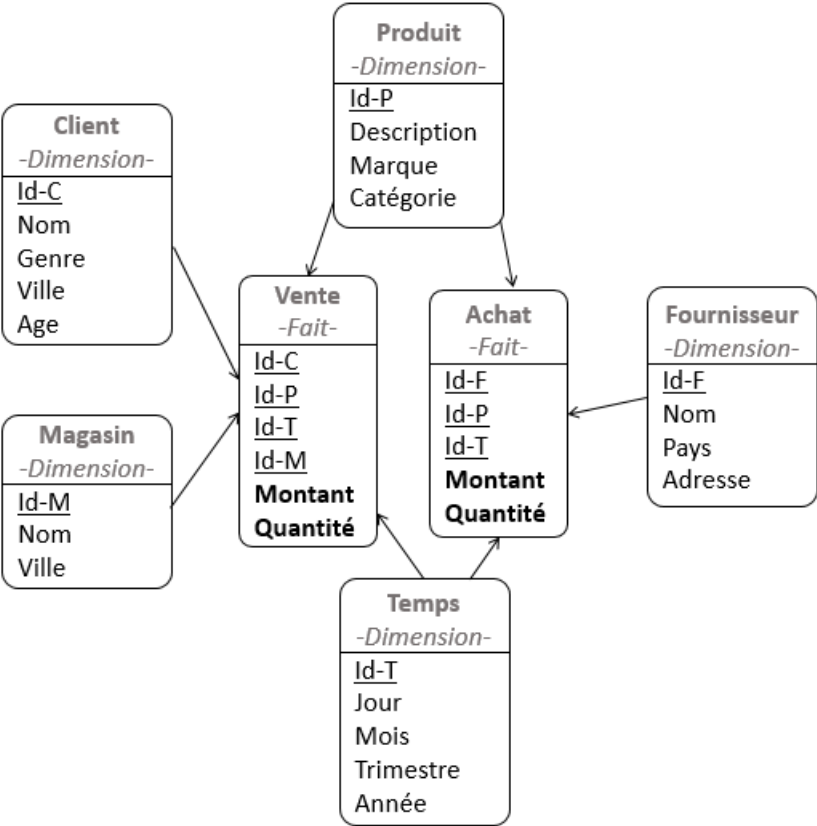


FIGURE 2.8 – Exemple d’un un schéma en constellation

2.7.1 ROLAP (Relational OLAP)

C'est l'approche la plus couramment utilisée pour implanter un entrepôt de données. Elle utilise un SGBD relationnel pour le stockage et le traitement des données[10, 15]. Ainsi, les entreprises peuvent directement exploiter leurs SGBDR déjà en place. La force principale du modèle ROLAP est qu'il s'appuie sur la maturité de la technologie relationnelle et du langage d'interrogation SQL. Cependant elle peut conduire à des temps de réponses très élevés ce qui constitue sa faiblesse majeure. La solution pour ce problème passe par l'application de différentes stratégies d'optimisation telles que l'indexation, la matérialisation de vues et la fragmentation[19].

2.7.2 MOLAP (Multidimensional OLAP)

Cette approche est basée sur un SGBD dédié appelé SGBD multidimensionnel. Le stockage des données ne se fait pas dans des bases de données relationnelle, mais dans des formats propriétaires appelés cubes de données[10, 15]. Le calcul des agrégats dans un tel modèle est rapide. Cependant, les bases de données MOLAP sont limitées dans leur capacité à gérer d'importants volumes de données. De plus, elles ne disposent pas d'un langage d'interrogation puissant d'où la nécessité de redéfinir des opérateurs spécifiques pour la manipulation de structures des cubes de données[1, 13].

2.7.3 HOLAP (Hybrid OLAP)

Comme leur nom l'indique, les systèmes HOLAP tentent de combiner les avantages des systèmes ROLAP et MOLAP en faisant cohabiter au sein du même système les technologies ROLAP et MOLAP. Pour cela, ils procèdent à stocker les données détaillées de l'entrepôt dans un SGBD ROLAP. Les données agrégées de l'entrepôt sont stockées dans un SGBD MOLAP. La limite de cette approche est la difficulté de la mise à jour et de la gestion du modèle hybride. De plus, la séparation des données doit être transparente à l'utilisateur final

2.8 Analyse et restitution des données

Les données d'un entrepôt sont créées pour être exploitées et valorisées. L'analyse et la restitution des données désignent l'ensemble des moyens et techniques fournis aux utilisateurs leur permettant l'exploration et la manipulation des données de l'entrepôt[10, 15]. Basiquement, l'exploration des données est réalisée à travers des requêtes analytiques appelées requêtes OLAP facilitant l'analyse interactive et la synthèse des données pour soutenir les prises de décision.

Un système d'analyse et restitution des données est composé de trois éléments[20] :

- Une base de données multidimensionnelle qui correspond à l'entrepôt de données,
- Un serveur OLAP qui analyse et traduit les requêtes OLAP,
- Un client OLAP qui permet de créer les requêtes et afficher les résultats de leurs exécution.

Rappelons que l'entrepôt de données est généralement destiné aux décideurs de l'entreprise souvent non spécialistes en informatique. Ainsi, les outils d'exploitation et visualisation des données doivent être adaptés. Le rôle du décideur est ici central puisque c'est

lui qui détient une connaissance du domaine d'analyse et il est en mesure de savoir si les informations restituées sont intéressantes ou non.

Selon l'objectif visé, un outil d'analyse des données peut être aussi simple qu'une interface de requêtes analytiques ou aussi complexe qu'une application de fouille de données (data mining). Toutefois, les fonctionnalités nécessaires d'un outil d'analyse sont :

- Affichage de grandes quantités de données
 - Présence de multiples types d'affichage : graphiques, couleurs, ...etc.
 - Présence d'une interface pour l'exploration : comparaison des données avec changement des modes de visualisation.
-

Chapitre 3

Étude conceptuelle

3.1 Introduction

Dans ce chapitre nous mettons en relief l'analyse et critique du système actuel. Nous spécifions par la suite les besoins analytiques et nous justifions l'intérêt de leur utilisation pour la représentation des données du système proposé. Nous présentons plus particulièrement la définition des faits et des dimensions et nous finissons par l'élaboration du modèle conceptuel retenu pour notre application.

3.2 Brève présentation du système LMD

Le système LMD est un système de formation pour l'enseignement supérieur basé sur trois niveaux d'études[21] (figure 3.1).



FIGURE 3.1 – Les formations dans le système LMD

Les formations sont organisées en semestre et en unités d'enseignement (UE). Un semestre dure en moyenne de 14 à 16 semaines d'études[21].

- La **Licence** est découpée en 06 semestres avec un projet de fin d'études dans le dernier semestre,
- Le **Master** est découpé en 04 semestres avec un projet de fin d'études dans le dernier semestre,
- Le **Doctorat** est découpée en 06 semestres d'études et de recherche (cours, stages, séminaires, projets) avec une thèse de fin d'études.

Chaque formation comporte quatre grandes catégories d'UE :

- Les UE **fondamentales** (UEF) : correspondant aux enseignements que tous les étudiants doivent suivre et valider,
- Les UE **méthodologique** (UEM) : permettant à l'étudiant d'acquérir l'autonomie dans le travail,

- Les UE **découverte** (UED) : permettant l’approfondissement, l’orientation, les passerelles, la professionnalisation,
- Les UE **transversales** (UET) : enseignements destinés à donner des outils aux étudiants : langues, Informatique.

Chaque semestre est affecté de 30 crédits répartis entre plusieurs unités d’enseignements. Un crédit correspond à la charge de travail de l’étudiant (cours, stages, mémoires, travail personnel). La validation des acquis de l’étudiant dans les différents modules correspond à l’attribution de crédits nécessaires. La figure 3.2 illustre le poids des unités d’enseignements exprimé en terme de pourcentage de crédits affectés pour chaque unité.



FIGURE 3.2 – Poids des UE dans le système LMD

Dans ce travail, nous nous sommes plus intéressés par le parcours des étudiants en graduation Licence-Master. Les performances des étudiants sont évaluées au niveau des modules et des UE comme suit[21] :

- Un module est acquis si la note obtenue dans ce module (moyenne entre l’examen et le contrôle continu) est égale ou supérieure à $\frac{10}{20}$.
- Une UE est acquise pour tout étudiant ayant acquis tout les modules qui la composent. Elle peut aussi être acquise par compensation si la moyenne de l’ensemble des modules qui la constituent, pondérées de leurs coefficients respectifs, est égale ou supérieure à $\frac{10}{20}$.
- De la même manière un semestre est considéré comme acquis pour tout étudiant ayant obtenu l’ensemble des UE qui le composent. Il peut également être acquis par compensation entre les différentes UE qui le composent avec une moyenne égale ou supérieure à $\frac{10}{20}$.
- Finalement, l’année universitaire est acquise si les deux semestre qui la composent sont acquis ou par compensation entre les deux semestres.

3.3 Système opérationnel actuel

Un des principaux problèmes rencontré durant la réalisation de ce travail est l’accès à un historique des données pour effectuer des comparaisons et des analyses. A ce jour, les services pédagogiques à l’université Algérienne sont dotés de systèmes transactionnels opérant sur les tâches élémentaires de gestion. Le logiciel **SEES** (Suivi des Etudiants et Enseignement Supérieurs)[22] est connu pour être recommandé par le ministère pour la gestion pédagogique pendant les dernières années. **PROGRES** est un autre outil qui commence à prendre la relève[23].

Ces outils gèrent essentiellement les opérations liées à la saisie des notes et l’élaboration des différents PVs. Ils ne permettent pas en revanche le suivi de la performance des étudiants. Une enquête au niveau du département d’informatique de l’université Amar Têlidji de Laghouat nous a permis d’identifier les principales fonctions réalisées :

- La gestion de la procédure d’inscription-réinscription des étudiants,
- L’élaboration des listes de groupes pour les travaux dirigés (TDs) et travaux pratiques (Tps),
- La gestion des délibérations : saisie des notes des examens et des Tds et TPs pour les différents modules.
- Calcul des moyennes par module, unité et semestre.
- Établissement des procès verbaux de délibérations et édition des relevés des notes.

Le constat dégagé est que les systèmes utilisés ne permettent aucune tâche d’analyse. Les responsables pédagogiques (décideurs) aux différents niveaux (département, faculté, vice-rectorat) ne disposent pas de tout l’archive nécessaire à une étude analytique. Certaines données sur le parcours des étudiants sont disponibles au niveau du département, d’autres au niveau de la faculté, faute d’aucune stratégie globale d’archivage et de diffusion de l’information. De plus, une bonne partie des données relatives aux anciennes promotions n’est pas disponible en format numériques.

L’importance des données générées par les systèmes opérationnels utilisés nous a motivé à attaquer le problème d’exploitation de ces données. Ces dernières peuvent être intégrées, historisées et organisées pour aider les responsables aux différents niveaux dans la prise de décision. D’où la proposition de la mise en place d’un entrepôt de données dédié à l’analyse des données relatives au parcours universitaire des étudiants.

3.4 Notre conception

Intuitivement, un entrepôt de données est conçu pour répondre à un ensemble de besoins d’analyse recensés auprès des décideurs. Par conséquent, notre approche de conception est naturellement guidée par les besoins analytiques du futur système. En d’autres termes, il s’agit, en premier lieu, de mettre en valeur **ce qu’on veut analyser**. Ceci va permettre de définir les **faits** (avec les mesures) de l’analyse. Par la suite, il faut identifier **les points de vue de l’analyse** afin de dégager les **dimensions** de l’analyse.

Ainsi nous avons mené une enquête au niveau du département de l’informatique à l’université Amar Télidji de Laghouat. Nous avons demandé aux responsables de spécifier leurs attentes d’analyse. En voici quelques exemples :

- Existe-il une relation entre la scolarité antérieure (le type de baccalauréat, lycée d’origine...) d’un étudiant et ses résultats à l’université ?
 - Les étudiants qui réussissent leurs cursus sans rachats, sans dettes, ni rattrapage, détiennent quel type de baccalauréat ? sont de quel sexe ?
 - Quel est l’impact des absences dans le cours, les TDs et Tps ?
 - Les étudiants ont ils le même comportement vis à vis des modules des unités fondamentales et ceux des unités de découverte ?
 - Quel est le profil des étudiants qui réussissent facilement les unités fondamentales ?
 - Pour quel semestre les résultats sont les plus mauvais ? pourquoi ?
 - Quels sont les profils et les grades des enseignants responsables des modules avec le meilleures (mauvaises) moyennes ?
-

- Quels est le profil des étudiants abandons ? Au niveau de quels semestres on enregistre le plus d'abandons ?
- Et plus généralement : Est ce que le cheminement d'un étudiant est fortement caractérisée par certains facteurs explicatifs de sa réussite et de son échec ?

L'objectif principal par la suite est de traduire de tels besoins dans un schéma conceptuel décrivant les informations nécessaires au système décisionnel futur.

3.4.1 Les faits

Une analyse approfondie des attentes des responsables nous a permis d'identifier deux faits à analyser à savoir :

- Les résultats par **module** (analyse détaillée),
- Les résultats par **semestre** (analyse moins détaillée).

Naturellement, les **mesures** pour l'analyse orientée module sont la note obtenue à l'**examen** et la note du **contrôle continu**. Notons que la note du contrôle continu pour les modules ayant simultanément un TP et un TD est la moyenne entre les deux notes. Pour l'analyse orientée semestre, une seule mesure est considérée et concerne la **moyenne** du semestre.

3.4.2 Les dimensions

Les dimensions expriment les différents points de vue de l'analyse. Dans notre cas nous avons identifié les analyses-types suivantes :

- Analyser les résultats en fonction des profils des étudiants,
- Analyser les résultats en fonction des profils des enseignants,
- Analyser les résultats en fonction des modules,
- Analyser les résultats en fonction des semestres.

Ceci a conduit à la considération des dimensions suivantes :

- **Étudiant** : cette dimension permet d'analyser les résultats par rapport au profil d'un étudiant (la scolarité antérieure, le sexe, l'âge, le type (interne, externe), ...etc.).
- **Module** : cette dimension permet d'analyser les résultats des étudiants par rapport au module en question (le type (unité fondamentale ou découverte), taux d'absences (beaucoup, moyen, pas d'absences), module acquis après rattrapage ou en dette...etc).
- **Enseignant** : cette dimension permet d'analyser les résultats par rapport au profil des enseignants (le grade, la spécialité, le type (permanent/vacataire), ancienneté ..etc).
- **Semestre** : cette dimension est temporelle et permet d'analyser les résultats des étudiants par rapport au semestre considéré (numéro du semestre, année, semestre acquis ou non..etc).

La description détaillée des différentes tables est la suivante :

1. Les tables de dimensions

- **Étudiant** { Id-Etudiant (Chaîne), Date-naissance (Date), Sexe (M,F), État-Social (Bon, Moyen Mauvais), Interne (Booléen), Année-Inscription (numérique), Filière (Chaîne), Série-Bac (Chaîne), Mention-Bac (Chaîne), Année-Bac (numérique), Lycée d'origine (Chaîne), Orientation (Booléen), Abondant (Booléen) }.
- **Module** { Id-Module (Chaîne), Code-Module (Chaîne), Type-Unité (Chaîne), Absence-TD-TP (Beaucoup, peu, pas), Absence-Cours (Beaucoup, peu, pas), Rattrapage (Booléen), Dette(Booléen) }
- **Enseignant** { Id-Enseignant (Chaîne), Grade (P, MCA, MCB, MAA, MAB), Spécialité (Chaîne), Type ("P", "V"), Année-recrutement (numérique) }.
- **Semestre** { Id-Semestre (Chaîne), Num-Semestre (Chaîne), Année-Scolaire (numérique), Rachat(Booléen), Redoublant(Booléen) }.

2. Les tables de faits

- **Cursus-Module** { Id-Etudiant (Chaîne), Id-EnsTD/TP(Id-Enseignant) (Chaîne), Id-EnsC(Id-Enseignant) (Chaîne), Id-Module (Chaîne) ,Id-semester (Chaîne), Note-Examen (Numérique), Note-CC (Numérique). }.
- **Cursus-Semestre** { Id-Etudiant (Chaîne), Id-Semestre, Moy-Semestre (Numérique). }.

3.4.3 Schéma en constellation

Le schéma en constellation de faits et de dimensions relatif à notre conception est illustré dans la figure3.3.

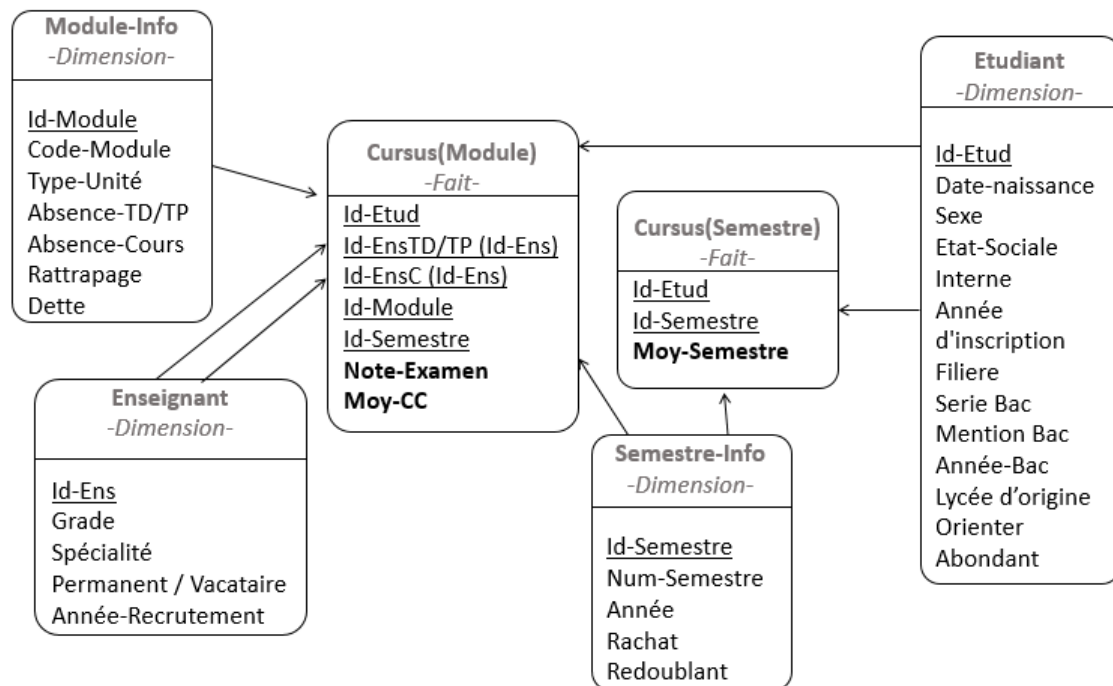


FIGURE 3.3 – Schéma en constellation de faits de notre conception

La table des faits **Cursus-Module** est liée à quatre tables de dimensions alors que la table des faits **Cursus-Semestre** est liée à deux tables de dimensions seulement. Ceci est dû à notre perception des différentes analyses envisagées.

L'objectif avec la table des faits **Cursus-Module** est permettre une analyse détaillée des résultats obtenus dans les examens et les contrôles continus. En considérant isolément les différentes dimensions, notre conception permet de chercher si les résultats obtenus sont liés (corrélés) au (1) profil de l'étudiant, ou bien (2) à la nature du module, ou bien (3) au profil de l'enseignant, ou encore (4) à la nature du semestre.

Notre conception permet des analyses plus complexes en considérant simultanément deux, trois ou quatre tables de dimensions. Ceci va permettre d'identifier des corrélations entre, par exemple, la note obtenue, la nature du module en question et le profil de l'enseignant chargé du module. Des requêtes type dans ce cas peuvent être :

- Quel est le nombre moyen des étudiants redoublants par module des unités fondamentales entre les années 2000 et 2010 ?
- Quels sont les profils des étudiants ayant plus de dettes pendant les premiers semestres des années 2014 et 2015 ?
- Étant donné le profil d'un étudiant, quel est l'état prévu de sa performance dans un semestre donné S_i ?.

D'un autre côté, les analyses envisagées avec la table des faits **Cursus-Semestre** sont moins détaillées. La mesure analysée est une agrégation de la performance d'un étudiant durant tout le semestre. Ceci va certainement faciliter la recherche de certaines corrélations **Étudiant-Semestre** sans être obligé à recalculer la moyenne du semestre. On peut alors avoir facilement des statistiques sur le nombre de redoublants (rachetés) pour un semestre donné ainsi que les profils des étudiants concernés.

Dans les deux cas de figure, les statistiques élaborées sont porteuses de connaissances que les décideurs doivent interpréter et exploiter.

Chapitre 4

Implémentation et validation

4.1 Introduction

Ce chapitre a pour objectif de présenter les phases d'implémentation et de la validation de notre conception. Nous décrivons en premier lieu les problèmes rencontrés lors de l'identification des données sources. Nous présentons par la suite les choix techniques retenus et les outils utilisés pour la transformation et le chargement des données disponibles. Finalement, un ensemble de requêtes analytiques est exécuté pour illustrer la mise en œuvre des données intégrées dans l'entrepôt. L'ensemble des réalisations est présenté sous forme de prises d'écran commentées pour faciliter la compréhension des différentes étapes de la mise en place de notre entrepôt de données.

4.2 Tout dépend des données !

Tout d'abord, nous avons opté pour une implémentation **ROLAP** pour notre schéma conceptuel. La raison est simple : c'est le modèle le plus répandu qui est techniquement supporté par tous les SGBD disponible en open source. De plus, l'implémentation **ROLAP** est basée essentiellement sur le formalisme relationnel et le langage SQL dont nous nous sommes familiarisés.

Pour le traitement des données sources, nous nous sommes naturellement dirigé vers deux types de solutions techniques :

- L'utilisation d'une plate-forme open source complète couvrant la partie ETL (intégration de données),
- L'utilisation d'une solution "maison", par scripts, en se basant sur nos propres compétences en terme de création et d'interrogation des bases de données en utilisant un SGBDR open source.

Nous avons commencé par explorer la première solution sachant qu'elle nécessite un temps d'adaptation et donc une prise en main assez difficile. Nous avons ainsi procédé à l'installation sur nos machines de deux outils parmi les plus utilisés à savoir Talend¹ et JasperETL².

1. <https://fr.talend.com/products/talend-open-studio/>

2. <https://community.jaspersoft.com/project/jaspersoft-etl>

Cependant, nous nous sommes rapidement rendu compte que de telles solutions sont très gourmandes en espace mémoire et en temps d'exécution et nécessitent en conséquence des serveurs dédiés. De plus les circonstances qui ont perturbé le bon déroulement du deuxième semestre nous ont empêchés d'explorer toutes les possibilités qu'elles pouvaient nous offrir. Les solutions ETL ont été installées sur un Pentium Core i3 avec 4Go de mémoire vive sous Linux Ubuntu 18.04.

Pour éviter une perte de temps due aux maladresses techniques, nous nous sommes dirigés vers la deuxième solution. Un autre facteur déterminant qui a tranché pour cette solution concerne le volume et le format des données disponibles. Il concerne notamment la problématique de disponibilité des données brutes et élaborées sur l'historique des étudiants. Rappelons que pour la mise en fonctionnement de notre conception, nous avons considéré les données sources issues du cas réel relatif à l'historique des étudiants du département d'informatique de l'université de Laghouat.

Idéalement, nous espérons à récupérer au moins les données relatives à une promotion pour pouvoir analyser son évolution de la première année jusqu'au Master 2. Cependant, les seules données qu'on a pu récupérer concernent :

- La 2^{ème} année Licence et le Master 1 pour l'année 2017/2018, et
- La 3^{ème} année Licence et le Master 2 pour l'année 2018/2019.

Le volume assez réduit des données disponibles ainsi que leur format ne nécessitent pas l'installation d'une solution ETL complète ce qui nous a motivé pour opter à la deuxième solution. Favorablement, les données sources ne sont pas très hétérogènes et elles sont toutes disponibles sous forme de fichiers plats Excel. Grâce à la compatibilité logicielle, il est facile d'importer, de transformer et de charger des données provenant d'Excel.

4.3 Création des tables de l'entrepôt

La première étape pour la mise en œuvre de notre solution consiste à la création des tables qui constituent notre entrepôt. Pour cela nous avons utilisé l'outil **MySQL** qui est un système de gestion de bases de données relationnelles utilisant le langage SQL. C'est un outil open source développé et supporté par Oracle³. Il fait partie des logiciels de gestion de base de données les plus populaires au monde[24]. Il offre en particulier les fonctions nécessaires pour la création des tables, la recherche des données (en langage SQL) et la gestion des utilisateurs. La figure 4.1 illustre le script de création de la table des faits **Cursus-Module**. Des scripts similaires sont utilisés pour créer le reste des tables.

4.4 Processus ETL

4.4.1 Données sources

Les défis majeurs pour la mise en œuvre de l'entrepôt sont liés à la disponibilité et la qualité des données sources. L'ensemble des données brutes disponibles est constitué de trois fichiers au format Excel :

- Le premier fichier contient des informations générales relatives aux étudiants : Numéro d'inscription, date de naissance, sexe, nationalité, numéro du baccalauréat...etc. Un extrait de ce fichier est illustré dans la figure 4.2.

3. <https://www.mysql.com/>

```

Command Prompt - mysql -u Sarah -p

MariaDB [db]> CREATE TABLE `cursusm` (
  -> `ID_etud` varchar(100) NOT NULL,
  -> `ID_ensTD` int(11) NOT NULL,
  -> `ID_ensC` int(11) NOT NULL,
  -> `ID_module` int(11) NOT NULL,
  -> `ID_semestre` int(11) NOT NULL,
  -> `Note_examen` double NOT NULL,
  -> `Moy_CC` double NOT NULL,
  -> PRIMARY KEY (ID_etud,ID_ensTD,ID_ensC,ID_module,ID_semestre),
  -> FOREIGN KEY (ID_etud) REFERENCES etudiant(ID_etud),
  -> FOREIGN KEY (ID_ensTD) REFERENCES enseignant(ID_ens),
  -> FOREIGN KEY (ID_ensC) REFERENCES enseignant(ID_ens),
  -> FOREIGN KEY (ID_module) REFERENCES module(ID_module),
  -> FOREIGN KEY (ID_semestre) REFERENCES semestre(ID_semestre)
  -> ) ENGINE=InnoDB DEFAULT CHARSET=utf8;
Query OK, 0 rows affected (2.268 sec)

```

FIGURE 4.1 – Script de création de la table Coursus-Module

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	
N°Ins	te_naissan	Sexe	CP	tionnale	_Inscrip	N°_Bac	e_W	Dir	Filière	Année	Bourse	Interne	double	Nbre	inée	Série	Mention_Ba	Note_Balla	aya	lf	Grof	S
10039-1-03	20-Nov-84	أ		جزائرية		18/000000					FALSE	FALSE	FALSE	0	2018	B01	مقبول	10	03		2	
15/39025406	24-Mar-98	ذ		جزائرية		16/25698					FALSE	FALSE	FALSE	0	2016	G05	مقبول	10	17		2	
15D03017	1-Jan-97	ذ		جزائرية		17/000000					FALSE	FALSE	FALSE	0	2017	B01	مقبول	10	03		2	
161631001283	16-May-98	F		جزائرية		31001283					FALSE	FALSE	FALSE	0	2016	G05	Passable	11.17	03		1	
161631055738	1-Jan-99	ذ		جزائرية		17/000000					FALSE	FALSE	FALSE	0	2017	B01	مقبول	10	03		1	
161639000723	14-Feb-99	ذ		جزائرية		39000723					FALSE	FALSE	FALSE	0	2016	G05	Bien	14.22	03		1	
161639000852	6-Apr-98	ذ		جزائرية		39000852					FALSE	FALSE	FALSE	0	2016	G05	قريب من الجيد	13.82	03		1	
161639000976	20-Jun-97	ذ		جزائرية		16/300					FALSE	FALSE	FALSE	0	2016	G05	مقبول	10	03		1	
161639000986	16-May-98	ذ		جزائرية		39000986					FALSE	FALSE	FALSE	0	2016	G05	قريب من الجيد	12.18	03		1	
161639001063	14-Feb-99	M		جزائرية		39001063					FALSE	FALSE	FALSE	0	2016	G05	مقبول	10.76	03		1	
161639001113	31-Jan-99	ذ		جزائرية		39001113					FALSE	FALSE	FALSE	0	2016	G05	قريب من الجيد	12.12	03		2	
etudiant 2018-2019		resultat 2018-2019			cours		resultat 2017-2018			etudiant 2017-2018			⊕		⋮							

FIGURE 4.2 – Extrait des données brutes (Étudiants)

- Le deuxième fichier contient des informations relatives aux résultats académiques des étudiants : Référence du module, note d'examen, moyenne du contrôle continu...etc. Un extrait de ce fichier est illustré dans la figure 4.3.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
N°Ins	Réf_Cours	examen	rattrapage	Conference	seminaire	TD	TP	stage	Projet	autre_Ctl	Moy_CC	y_Mod	Aquit	Option	ratt
17173901375	D111	8	7			9					0	8.33	FALSE	FALSE	FALSE
161639001960	D111	10				10					2	10.00	FALSE	FALSE	FALSE
161739013641	D111	10	7			9					0	9.67	FALSE	FALSE	FALSE
207	D111										0	0.00	FALSE	FALSE	FALSE
250	D111										0	0.00	FALSE	FALSE	FALSE
386	D111										0	0.00	FALSE	FALSE	FALSE
6360	D111										0	0.00	FALSE	FALSE	FALSE
6363	D111										0	0.00	FALSE	FALSE	FALSE
6367	D111										0	0.00	FALSE	FALSE	FALSE
6370	D111										0	0.00	FALSE	FALSE	FALSE
6378	D111										0	0.00	FALSE	FALSE	FALSE
6386	D111										0	0.00	FALSE	FALSE	FALSE
6396	D111										0	0.00	FALSE	FALSE	FALSE
6397	D111										0	0.00	FALSE	FALSE	FALSE
6407	D111										0	0.00	FALSE	FALSE	FALSE
6410	D111										0	0.00	FALSE	FALSE	FALSE
6413	D111										0	0.00	FALSE	FALSE	FALSE
6422	D111										0	0.00	FALSE	FALSE	FALSE
6434	D111										0	0.00	FALSE	FALSE	FALSE
6445	D111										0	0.00	FALSE	FALSE	FALSE
6446	D111										0	0.00	FALSE	FALSE	FALSE
6448	D111										0	0.00	FALSE	FALSE	FALSE

FIGURE 4.3 – Extrait des données brutes (Résultats)

- Le troisième fichier contient des informations relatives aux modules enseignés : Référence du module, désignation, coefficient, volume horaire...etc. Un extrait de ce fichier est illustré dans la figure 4.4.

4.4.2 Transformation des données sources

Comme indiqué auparavant, ces données sont issues d'outils utilisés pour la gestion pédagogique des étudiants. Elles contiennent ainsi des informations importantes relatives aux résultats (moyenne des examens, des contrôles continus et des semestres) obtenus par les étudiants. En revanche, d'autres données utiles à la mise en œuvre de notre conception sont inexistantes :

- Pour la table **Étudiant** : pas d'information sur l'état social. De même que pour l'état de son orientation initiale. Pas d'information indiquant si sa filière d'étude correspond à sa fiche de vœux,
- Pour la table **Module** : aucune trace des états des absences dans les cours ou les TDs, TP,
- Pour la table **Enseignant** : absence de données sur le grade et l'année de recrutement de l'enseignant. De même pour son statut (permanent, vacataire),
- Pour la table **Semestre** : pas d'information si le semestre est acquis avec rachat ou non. Souvent, quand un étudiant est racheté, on procède à lui attribuer une moyenne de $\frac{10}{20}$ sans mentionner qu'il est racheté.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Réf_Cours	Cours	Coef	VExm	VRttrpg	TD	VTD	TP	VTP	Seminaire	VSem	Conférence	VConf	Stage	VStg	Projet
ALG3	Algèbre3	2	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ALG4	Algèbre4	2	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ALGO	Algorithmique et s.de	5	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ALGO2	Algorithmique et S.de données	5	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANA3	Analyse 3	3	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANA4	Analyse 4	3	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANAC	Analyse complexe	2	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANAN	Analyse numérique	1	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANAN2	Analyse numérique 2	3	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANG3	Anglais3	1	FALSE	FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANG4	Anglais 4	1	FALSE	FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
ANN1	Analyse numérique 1	3	FALSE	FALSE 1		FALSE 0		FALSE 0		FALSE 0		FALSE 0		FALSE 0	
	Application														

FIGURE 4.4 – Extrait des données brutes (Modules)

Toutefois, la problématique des données manquantes restent très répandues dans mise en œuvre des entrepôts de données. Dans notre cas ceci est expliqué par le fait que les données relatives au parcours des étudiants ne sont pas enregistrées dans une perspective décisionnelle. De plus, les données sont produits à plusieurs niveaux avec des périmètres de responsabilité différents.

Les données au niveau du département concernent particulièrement l'aspect pédagogique. Au niveau de la faculté et du vice-rectorat, les données sont d'ordre administratif lié à la gestion des diplômes. Les données relatives aux enseignants quant à elles sont centralisées au niveau du service du personnel de la faculté.

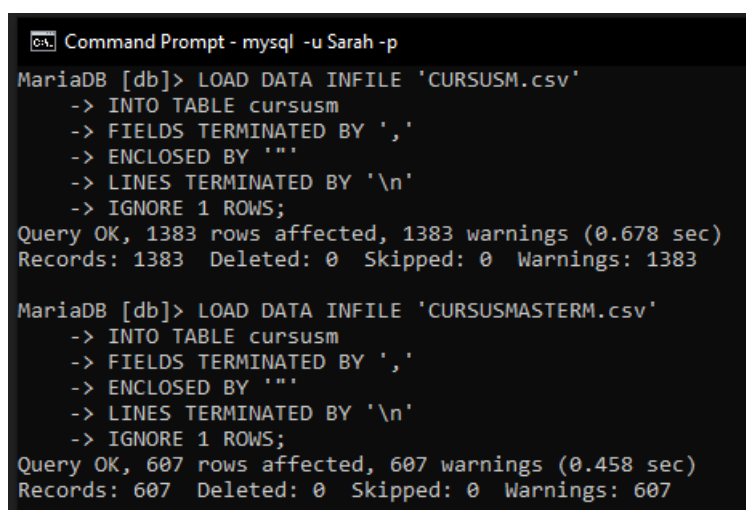
Faute d'absence d'une stratégie globale d'archivage et d'analyse, de précieuses données ont été perdues ou simplement non enregistrées. Par conséquent, nous étions contraints à traiter manuellement les données non disponibles. Les principaux traitements réalisés concernent :

- La récupération uniquement les champs dont on a besoin pour nos tables à partir des fichiers Excel sources. Toutes les données ne sont pas intégrées, seules celles pertinentes et ont été conservées.
- La normalisation de la langue.
- La suppression des données relatives aux étudiants qui se sont inscrits sans vraiment étudier.
- La suppression des champs vides.
- L'unification de la note de l'examen en prenant le maximum entre la note de rattrapage et la note de l'examen.
- La saisie des informations manquantes comme par exemple pour le champs : 'Rachat', 'Rattrapage', 'Redoublant', et 'Abondant'.

D'un autre côté, nous avons élaboré un fichier Excel contenant les informations relatives aux enseignants. Tous les traitements ont été réalisés dans des feuilles de calcul grâce au logiciel Excel de la suite Office 2017.

4.4.3 Chargement des données transformées

Après avoir créé les tables de l'entrepôt et transformer les données sources, il est temps de charger les données vers leurs destinations. Si l'outil Excel nous a largement facilité le traitement des données sources, il ne nous permet pas d'importer les données traitées directement dans les tables MySQL. Heureusement, la commande **Sql Load** permet d'importer des données en format CSV directement dans une table. Nous avons ainsi transformé nos données en format csv et les charger facilement dans les tables de l'entrepôt. La figure 4.5 illustre les scripts relatifs au chargement des données dans les deux tables des faits. Des scripts similaires sont utilisés pour charger le reste des tables.



```
Command Prompt - mysql -u Sarah -p
MariaDB [db]> LOAD DATA INFILE 'CURSUSM.csv'
-> INTO TABLE cursusm
-> FIELDS TERMINATED BY ','
-> ENCLOSED BY '"'
-> LINES TERMINATED BY '\n'
-> IGNORE 1 ROWS;
Query OK, 1383 rows affected, 1383 warnings (0.678 sec)
Records: 1383 Deleted: 0 Skipped: 0 Warnings: 1383

MariaDB [db]> LOAD DATA INFILE 'CURSUSMASTERM.csv'
-> INTO TABLE cursusm
-> FIELDS TERMINATED BY ','
-> ENCLOSED BY '"'
-> LINES TERMINATED BY '\n'
-> IGNORE 1 ROWS;
Query OK, 607 rows affected, 607 warnings (0.458 sec)
Records: 607 Deleted: 0 Skipped: 0 Warnings: 607
```

FIGURE 4.5 – Scripts de chargement des données dans les tables des faits

4.5 Exploitation de l'entrepôt : résultats préliminaires

4.5.1 Quelques aspects d'analyse

Les données de l'entrepôt sont générées pour être exploitées. Une méthode communément utilisée consiste à interroger l'entrepôt à travers des requêtes analytiques OLAP. Ces dernières permettent à un utilisateur de consulter et d'extraire facilement des données et les comparer de différentes façons dans une perspective de prise de décision. Les requêtes OLAP sont souvent des requêtes complexes vis à vis du nombre des dimensions et des prédicats impliquées. Nous présentons dans ce qui suit les résultats de quelques requêtes illustrant le comportement des étudiants en fonction de certains facteurs liés à notre conception. Les objectifs de ces requêtes étaient :

- Analyser la compensation de la note de l'examen par celle du contrôle continu pour un semestre donné, par rapport à aux grades des enseignants du cours et du contrôle continu et leur ancienneté et par rapport au type de l'unité d'enseignement.
- Déterminer le pourcentage des étudiants qui ont eu leurs modules sans rattrapage dans un semestre donné.

- Analyser les performances des étudiants par sexe. En particulier, déterminer le pourcentage des étudiants qui ont réussi un parcours sans faute et ceux qui ont passé au moins une fois un examen de rattrapage, avec des dettes,
- Analyser les performances des étudiants par rapport à la mention obtenue au Bac.

Les résultats des différentes analyses sont présentés dans les tableaux 4.1 4.2 et 4.3

UE	Module	Sans compensation	Avec compensation
Fondamentale	F2111	12.5%	57.14%
	F2112	34.55%	26.32%
	F2113	25.45%	35.71%
Découverte	F2121	29.82%	52.94%
	F2122	05.56%	66.67%
	F2123	62%	32.26%

TABLE 4.1 – Compensation des modules par UE

Sexe	Nombre	Parcours avec échec (%)
Féminin	41 (38.68%)	32.88%
Masculin	65 (61.32%)	67.12%

TABLE 4.2 – Parcours vs. Sexe

Mention Bac	Nombre	Parcours avec échec(%)	Abndons(%)
Passable	67	63.01%	62.5%
A.Bien	34	32.88%	37.5%
Bien	5	04.11%	00.00%

TABLE 4.3 – Parcours vs. Mention Bac

4.5.2 Discussion des résultats

Les résultats préliminaires obtenus montrent que :

- Environ 50.61% des étudiants ont compensé des modules appartenant à des unités d'enseignement fondamentales grâce à la note du contrôle continu. Par contre le pourcentage des étudiants qui ont compensés des modules appartenant à des unités d'enseignement de découverte est environ 39.72%. Cette différence peut être expliquée par le fait que les étudiants éprouvent plus de difficultés pour les modules des unités fondamentales. D'autre part, ceci explique en partie que les étudiants qui valident les unités d'enseignement fondamentales valident également les unités d'enseignement de découverte. Nous avons noté également que pour ce cas de figure les enseignants chargé du cours assurent en même temps les travaux des contrôles continus.
- Le pourcentage des étudiants avec un parcours sans fautes (sans rattrapage, ni dettes, ni redoublement) est d'environ 67.12% chez le sexe féminin et de 32.88%

chez le sexe masculin. Ces écarts peuvent témoigner de préférences et d'intérêts différents selon le sexe.

- Le pourcentage des étudiants avec la mention **Passable** au Bac qui ont abandonnés est d'environ 62.5%. Le même pourcentage est d'environ 37.5% et 0.0% pour les mentions **A. Bien** et **Bien** respectivement. Le premier constat qui peut être réalisé concerne les taux élevés des abandons pour les mentions **Passable** et **A. Bien**. Ceci explique en partie l'évaluation au niveau de l'examen du baccalauréat ainsi que l'orientation à l'entrée à l'université. D'autre part, le pourcentage des étudiants dont le parcours est marqué par au moins un échec (rattrapage, dettes, rachat) est de 63.01%, 32.88% et 4.11% pour les mentions **Passable**, **A. Bien** et **Bien** respectivement. On pourrait s'attendre à ce que les bacheliers avec mention réussissent facilement leur parcours. Les résultats montrent qu'au contraire, même certains étudiants avec la mention **Bien** ont du avoir des difficultés pour valider leurs cursus.
-

Chapitre 5

Conclusions et perspectives

Les entrepôts de données occupent désormais une place prépondérante au sein des entreprises et des organisations pour permettre des analyses dédiées à la prise de décisions. Dans ce mémoire nous nous sommes intéressés à la conception et la mise en œuvre d'un entrepôt de données pour l'analyse du cursus universitaire des étudiants. Cette solution répond à une nécessité dans le domaine de l'enseignement supérieur qui ne doit resté à la marge de cette technologie.

Les données sociologiques et académiques relatives au parcours universitaire des étudiants sont de véritables mines d'informations. L'entreposage de ces données constitue une étape importante pour offrir aux décideurs des connaissances utiles identifiant des facteurs explicatifs de la réussite et de l'échec chez les étudiants.

Nous avons présenté dans un premier lieu les motivations derrière le sujet abordé. Nous avons expliqué par la suite les concepts et les définitions relatifs à l'architecture et la modélisation des entrepôts de données. En se basant sur un ensemble de besoins analytiques des décideurs, nous avons fini par illustré notre modélisation en constellation des faits. En raison des contraintes de temps et des données disponibles notre mise en œuvre de l'entrepôt proposé s'est portée sur une partie de l'archive du département d'informatique de l'université Amar Télidji de Laghouat.

Ce projet a été une opportunité pour nous afin de mettre en valeur et d'approfondir les connaissances et les compétences acquises au cours de notre formation en Master sur les systèmes d'information décisionnels. Au cours du projet nous étions amenés à jouer le rôle du concepteur, du développeur, du testeur et de l'utilisateur. Ainsi, l'expérience s'avère très positive et nous sommes convaincus que le travail effectué nous servira à l'avenir dans notre vie professionnelle.

A l'issue de ce travail, le premier constat dégagé est que la conception et la mise en œuvre d'un entrepôt de données est une tâche complexe. La partie du projet la plus fastidieuse et consommatrice en temps était la partie ETL. Nous étions en particulier confronté à la problématique de disponibilité des données brutes et élaborées sur l'historique des étudiants.

Bien que testée sur un ensemble réduit de données disponibles au niveau du département d'informatique de l'université Amar Télidji de Laghouat, notre approche a été conçue pour être générique et utilisée avec toutes les données de n'importe quelle université. Comme un projet de mise en œuvre d'un entrepôt de données peut prendre des années et nécessite un rafraîchissement périodique, nous envisageons les perspectives suivantes :

1. L'exploitation réelle de l'entrepôt proposé nécessite des solutions technologiques adéquates. Ainsi, il est recommandé d'installer un serveur OLAP avec un minimum de plusieurs gigaoctets de RAM et plusieurs téraoctets d'espace de stockage. Des solutions logicielles open source gratuites telles que Mondrian ou Oracle OLAP sont disponibles et permettent d'effectuer une analyse multidimensionnelle de données avec des temps de réponse optimisés. En parallèle, il faut utiliser des outils tels que Kibana ¹ qui facilitent l'extraction et la visualisation des données avec une interface conviviale et facile pour les décideurs.
2. Étendre le jeu de données de manière à couvrir la totalité de l'université sur un archive de plusieurs années. Le rafraîchissement de l'entrepôt pourra être réalisé chaque fin de semestre. Dans ce cas de figure le processus ETL nécessite l'installation d'une plate forme professionnelles open source telles que Talend ² qui est un leader de l'intégration de données.
3. L'accent sera mis sur deux types de données : les données relatives au système classique et celles relatives au système LMD. On pourra ainsi penser à avoir deux data marts permettant la comparaison des deux systèmes, une question souvent posée dans le secteur de l'enseignement supérieur.
4. Une autre perspective intéressante à explorer consiste à implémenter la solution proposée avec Hadoop ³ sous forme de base de données NoSql. Hadoop est un logiciel open source de la fondation Apache. C'est un environnement logiciel dédié au stockage et au traitement de larges volumes de données.
5. Enfin, l'entreposage réalisé permet une meilleure exploitation des données avec l'utilisation des méthodes de la fouille de données (Data Mining) telles que les motifs fréquents et les règles d'association.

1. <https://www.elastic.co/fr/kibana>

2. <https://fr.talend.com/>

3. <https://hadoop.apache.org/>

Bibliographie

- [1] María Trinidad. *Entrepôts de données pour l'aide à la décision médicale : Conception et expérimentation*. PhD thesis, Université Joseph Fourier, 2005. Thèse de doctorat.
- [2] Fanta BOUBA. *Système d'information décisionnel sur les interactions environnement-santé*. PhD thesis, Université Pierre et Marie Curie, 2015. Thèse de doctorat.
- [3] Mohamed GUIDOUME. *Représentations et pratiques enseignantes dans une approche par compétences du FLE au secondaire algérien : Cas des PES de la Wilaya de Tiaret*. PhD thesis, Université d'Oran, 2010. Thèse de doctorat.
- [4] A. AMRANI. Mérites et limites du système de formation LMD par référence aux programmes de formation. *AL-IJTIHED Revue des études juridiques & économiques*, 2012.
- [5] Messaoud AICHE. *Analyse des pratiques pédagogiques d'enseignement du projet architectural au département d'architecture et d'urbanisme de Constantine à la lumière des méthodes pédagogiques actives : Pour un meilleur développement des compétences des étudiants de fin de cycle*. PhD thesis, Université Mentouri - Canstantine, 2006. Thèse de doctorat.
- [6] Riad MESSOUAR. *L'apport du cursus universitaire dans la formation des enseignants du FLE en Algérie*. PhD thesis, Université Mohamed Lamine DEBAGHINE SETIF 2, 2016. Mémoire de Magister.
- [7] Nadjet HOUHAMDI. Les pratiques évaluatives sommatives par rapport aux attentes des nouveaux programmes : Cas des enseignants du cycle primaire. Master's thesis, Université Larbi Ben M'Hidi , Oum El Bouaghi, 2017. Mémoire de Master.
- [8] R.A. Mata-Toledo and P.K. Cushman. *Introduction aux bases de données relationnelles*. Série Schaum. Dunod, 2002.
- [9] Jean-Luc Hainaut. *Bases de données : Concepts, utilisation et développement*. InfoSup. Dunod, 4ème édition, 2018.
- [10] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit : The Complete Guide to Dimensional Modeling*. John Wiley & Sons, Inc., USA, 3rd edition, 2013.
- [11] J. Han and M. Kamber. *Data Mining : Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.
- [12] D. Kayser. *La représentation des connaissances*. Collection Informatique. Hermès, 1997.
- [13] R. Benmessaoud. *Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes*. PhD thesis, Université Lumière Lyon 2, 2006.

-
- [14] Yoann Pitarch. *Résumé de Flots de Données : Motifs, Cubes et Hiérarchies*. PhD thesis, Université Montpellier 2, 2011. Thèse de doctorat.
- [15] W. H. Inmon. *Building the Data Warehouse*. QED Information Sciences, Inc., Wellesley, MA, USA, 1992.
- [16] Ines BENKRAIEM. Modélisation et intégration de données de capteurs/compteurs du SGE. Master's thesis, IRIT -Toulouse, 2017. Rapport de stage - Master 2 -.
- [17] Matthieu Leblanc. Entrepôt de données : de l'alimentation des données au décisionnel de l'établissement. Master's thesis, Conservatoire National des Arts et des Métiers - Paris, 2012. Mémoire d'Ingénieur.
- [18] Neosoft. La modélisation, de l'étoile à la constellation. [en ligne] <https://www.neosoft-solutions.fr/la-modelisation-de-letoile-a-la-constellation/>; dernière consultation : 10 mai 2020.
- [19] Benameur ZIANI. *Techniques de fouille de données pour la formalisation et l'optimisation de la conception physique des entrepôts de données relationnels*. PhD thesis, Université de Laghouat, 2016. Thèse de doctorat.
- [20] Ali HASSAN. *Modélisation des Bases de Données Multidimensionnelles : Analyse par Fonctions d'Agrégation Multiples*. PhD thesis, Université Toulouse 1 Capitole, 2014. Thèse de doctorat.
- [21] Université de Boumerdes. Guide pratique du système LMD. [en ligne] https://www.univ-boumerdes.dz/arrete/guide_pratique_systeme_LMD_en_francais.pdf; dernière consultation : 28 avril 2020.
- [22] MESRS. Système SEES. [en ligne] http://www.univ-soukahrass.dz/Foire_Aux_Questions_SEES.pdf; dernière consultation : 28 avril 2020.
- [23] MESRS. Système PROGRES. [en ligne] http://www.univ-tebessa.dz/vrg/fichiers/Manuel_PROGRES.pdf; dernière consultation : 28 avril 2020.
- [24] Wikipédia. Site web français de wikipédia. [en ligne] <https://fr.wikipedia.org/wiki/MySQL>; dernière consultation : 10 mai 2020.
-