RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ 8 MAI 1945 - GUELMA -

FACULTÉ DES MATHÉMATIQUES, D'INFORMATIQUE ET DES SCIENCES DE LA MATIÈRE

Département d'Informatique

Mémoire de Fin d'études Master

Filière : Informatique

Option : Science et technologie de l'information et de la communication

Thème

Utilisation des Techniques de Fouille de Données dans un Système de Business Intelligence

---

Présenté par :

CHETTIBI Radja Sara

Membres du jury :

N	NOM	Qualité
1	Mme. Djakdjakha Lynda	Président
2	Mr. Khebizi Ali	Encadreur
3	Mr. Derdar Salah	Examinateur

Juin 2023

REMERCIEMENTS

Tout d'abord, je remercie Dieu de m'avoir donné le courage et la patience dont j'avais besoin durant cette langue année.

Je tiens également à remercier mes parents et tous les membres de ma famille qui sont restés à mes cotés pendant mes études, et qui n'ont cessé de m'apporter un soutien moral et matériel, d'autant plus que j'ai terminé ce mémoire.

Mes sincères remerciements vont droit au monsieur **Khebizi Ali** étant un directeur de mon mémoire, m'a inspiré à réaliser ce modeste travail et qui a été d'une grande utilité avec ses directions et n'a pas hésité à consacrer son temps pour me guider. Merci de votre dévouement, sans vous ce travail n'aurait jamais vu le jour.

Je remercie **les MEMBRES DE JURY**, c'est un grand honneur pour moi d'accepter de siéger mon projet de fin d'études.

Je tiens à remercier tous ceux qui ont contribué de près ou de loin au bon déroulement de ce travail.

Merci a toutes et tous

RÉSUMÉ

Dans ce projet de fin d'études, nous proposons de concevoir et d'implémenter un système d'informatique décisionnel qui intègre différentes techniques de fouille de données.

Contrairement aux système conventionnels basés sur une vision mono-perspective qui se focalise sur un seul aspects, nous suggérons une intégration qui utilise plusieurs technique à la fois.

Nous admettons que l'entrepôt de données est déjà conçu et qu'il est alimenté en données historiques et nous spécifions quatre techniques de fouille de données, à savoir : l'analyse de données, intelligence artificielle, OLAP et statistiques.

L'approche proposée a été implémentée et expérimentée sur un jeu d'essai réel et les premiers résultats d'expérimentation sont très prometteur.

Mots Clés : Business Intelligence « B.I », Data Warehouse « Entrepôt de données », Data Analytics, OLAP, data mining, ETL.

ABSTRACT

In this graduate project, we propose to design and implement a decision-making IT system that integrates different data mining techniques.
Unlike conventional systems based on a mono-perspective vision that focuses on one aspect, we suggest an integration that uses multiple techniques at once.
We admit that the data warehouse is already designed and it is powered by historical data and we specify four data mining techniques, namely : data analysis, artificial intelligence, OLAP and statistics.
The proposed approach has been implemented and experimented on a real trial game and the first experimental results are very promising.
Keywords : Business Intelligence « B.I », Data Warehouse « Entrepôt de données », Data Analytics, OLAP, data mining, ETL.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SYMBOLES

<**ADD**> <*Analyse de données*>

<**ACP**> <*Analyse en Composantes Principales*>

<**AFC**> <*Analyse factorielle des correspondances* >

<**BI**> <*Business Intelligence*>

<**BDD**> <*Base de données*>

<**CRM**> <*Customer Relationship Management*>

<**EDD**> <*Entrepôt de données*>

<**ERP**> <*Entreprise Ressource Planning*>

<**ETL**> <*Extraction Transformation Chargement*>

<**FDD**> < *Fouille de données*>

<**IA**> <*Intelligence artificielle*>

<**KPI**> <*Key Indicator Performance*>

<**MOLAP**> <*Multidimensional On-Line Analytical Processing*>

<**OLAP**> <*Online Analytical Processing*>

<**OLTP**> <*Online Transactional Processing*>

<**ROLAP**> <*Relational On-Line Analytical Processing*>

INTRODUCTION GÉNÉRALE

Dans les systèmes économiques actuels, les données sont devenues un atout précieux. Les entreprises génèrent et collectent une quantité massive de données provenant de diverses sources telles que les transactions, les interactions clients, les médias sociaux, les appareils connectés, et bien d'autres encore. Cependant, sans une compréhension approfondie de ces données, elles restent souvent sous-exploitées et ne fournissent pas les informations nécessaires pour prendre des décisions éclairées.
C'est là que l'informatique décisionnels ou la Business Intelligence (BI) entre en jeu.
La BI est un domaine qui vise à transformer les données brutes en informations significatives, afin de soutenir la prise de décision stratégique dans les organisations. L'une des principales composantes de la BI est la fouille de données, également connue sous le nom de Data Mining. Il s'agit d'un processus qui permet d'extraire des connaissances précieuses et exploitables à partir de grandes quantités de données stockées dans les systèmes informatique des organisations.
L'objectif de ce mémoire est d'explorer l'utilisation des techniques de fouille de données dans un système de business intelligence. Nous examinerons les différentes étapes du processus de fouille de données, telles que la collecte des données, la préparation des données, l'exploration des données, la modélisation et l'évaluation des résultats. Nous discuterons également des différentes techniques de fouille de données

qui peuvent être appliquées dans le contexte de la BI ou bien les techniques statistiques et même celles relatives aux analyse de données, telles que la classification, la prédiction, le regroupement et l'association. L'objectif de ce travail et d'enrichir les systèmes BI conventionnels qui utilisent qu'une seule option de fouille de données par des techniques supplémentaires et variées.

En effet, l'utilisation de techniques de fouille de données dans un système de BI offre de nombreux avantages potentiels. Cela permet aux entreprises d'identifier des tendances et des modèles cachés dans leurs données, de découvrir de nouvelles opportunités commerciales, d'améliorer l'efficacité opérationnelle, de prendre des décisions plus éclairées et de rester compétitives sur le marché. Cependant, l'adoption de ces techniques soulève également des défis tels que la qualité des données, la confidentialité et la sécurité, ainsi que l'interprétation et la validation des résultats.

Au cours de ce mémoire, nous analyserons en détail les opportunités et les défis associés à l'utilisation de techniques de fouille de données dans un système de business intelligence. Nous examinerons également des cas d'étude réels de différentes industries pour illustrer l'application pratique de ces techniques. En fin de compte, ce travail vise à fournir des recommandations et des lignes directrices pour une utilisation efficace et réussie de la fouille de données dans le contexte de la BI.

Le manuscrit est structuré comme suit.

Dans le chapitre 1, nous introduisons le domaine de la BI, et nous donnons toutes les définitions et concepts utiles à la compréhension du mémoire.

Le chapitre 2 est réservé au techniques de fouille de données. On y présente les approches les plus utilisées à savoir : intelligence artificielle, analyse de données, statistique et l'OLAP.

Dans le chapitre 3, un état de l'art des travaux connexes ayant abordé la question de l'intégration des techniques de fouille de données dans une solution est exposé. Les avantages et les inconvénients de chaque technique sont mise en évidence.

Le chapitre 4 constitue notre contribution par les propositions d'une nouvelle approche combinée intégrant les différentes techniques dans un système BI.

Le dernier chapitre est réservé a l'implémentation et l'expérimentation de l'approche proposée.

Première partie

Etat de l'art

CHAPITRE 1

PRÉSENTATION DE L'INFORMATIQUE DÉCISIONNELLE (OU BUSINESS INTELLIGENCE)

1.1 Introduction

Dans ce premier chapitre, nous présentons les définitions et concepts de base qui seront utiles à la compréhension de la suite du rapport.
Nous nous intéressons essentiellement aux notions de BI, d'entrepôt de données et les architectures qui leur sont afférentes.

1.2 Définition de la Business Intelligence

Définition 1.1 Business intelligence (informatique décisionnelle en français) désigne ensemble des moyens, outils et méthodes qui permettent de collecter, intégrer, diffuser et restituer l'information en vue d'offrir une aide à la décision [27].
Définition 1.2 BI est une procédure technologique d'analyse des données et de présentation des informations destinée à aider les gestionnaires, les cadres et les autres utilisateurs finaux à prendre des décisions commerciales avisées [1].
Ce type d'application utilise généralement le traitement par lots pour rassembler les

données provenant de plusieurs sources hétérogènes et les stocker dans un entrepôt de données (ou data warehouse DWH).

1.2.1 Architecture d'une solution BI

Un système BI est composée de quatre grandes catégories d'outils : l'ETL, data warehousing, data mining et de reporting, nous le représentons dans le schéma de la figure 1.1 ci-dessous.

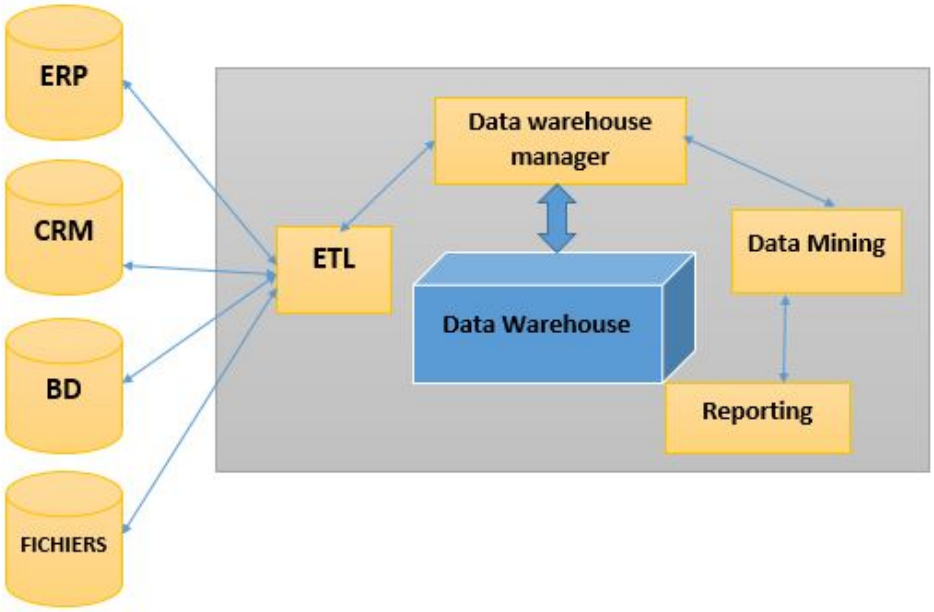


FIGURE 1.1 – Schéma général d'une solution Business Intelligence

1.2.2 Le rôle de la BI

— La Business Intelligence (BI) permet aux entreprises d'améliorer leurs performances en analysant leurs données.

— La BI permet aux entreprises de prendre des décisions plus éclairées et plus informées en fournissant des connaissances pertinentes et à jour sur les performances passées, présentes et futures.

— Peut aider les entreprises à améliorer leur efficacité opérationnelle, à réduire les coûts et à améliorer leurs marges bénéficiaires. Elle peut également aider les entreprises à mieux comprendre leurs clients et à prendre des décisions plus stratégiques [27].

1.2.3 Les étapes du processus BI

Un processus BI passe par quatres phases qui les suivantes.

— La phase de collecte de données ou d'alimentation

— La phase d'intégration

— La phase d'organisation

— La phase réstitution

Le fonctionnement de ces phases est expliqué ci dessous.

1.2.4 Fonctionnement du processus BI

Le fonctionnement de ces phases est expliqué ci dessous.

a) **La phase de collecte (ou d'alimentation)** : Cette première phase consiste à collecter, nettoyer et consolider les données de l'entreprise issues de différentes sources. Cela en utilisant les outils d'ETL adéquats.

b) **La phase d'intégration** : Les données résultantes de la première phase seront ensuite stockées dans une base spécialisée. Nommé DWH ou dans un data mart (qui est une version plus réduite du DWH). Dans l'objectif de les préparer au rôle final dit analyse décisionnelle [27].

c) **La phase de diffusion (ou organisation)** : La phase d'organisation en business intelligence est une étape cruciale qui consiste à structurer les données collectées

afin de les rendre exploitables. Cette phase implique la mise en place d'un système de stockage et de gestion des données, ainsi que la définition des règles et des normes pour leur utilisation. L'objectif principal de cette phase est de garantir la qualité et l'intégrité des données, ce qui permettra aux utilisateurs finaux d'obtenir des informations fiables et pertinentes utile à la prise de décision [27].

d) **La phase de réstitution** :Une fois les données collectées, nettoyées, stockées, et accessibles elles peuvent être analysées pour faire ressortir des prévisions ; ou des estimations futures en utilisant les outils du data mining. Selon les besoins, différents types d'outils d'extraction et d'exploitation seront utilisés, tels que :

- OLAP pour les analyses multidimensionnelles, notamment analyser les données.
- Le Datamining pour rechercher des corrélations.
- Les tableaux de bord présentant les indicateurs clés de l'activité à l'entreprise.
- Le Reporting pour communiquer la performance via des présentations graphiques et ergonomiques.

1.3 Les sorties d'une solution BI

Une solution de business intelligence peut fournir une variété de sorties, notamment :

a) **Rapports et tableaux de bord** : Les rapports et tableaux de bord sont des documents qui résument les données et les informations d'une organisation. Ils peuvent être utilisés pour surveiller les performances, identifier les tendances et prendre des décisions éclairées [27].

b) **Présentations** : Les présentations sont des documents qui peuvent être utilisés pour communiquer des informations à un public cible. Elles peuvent inclure des

graphiques, des diagrammes et d'autres visualisations pour aider à illustrer les points clés.

c) **Alertes** : Les alertes sont des notifications envoyées aux utilisateurs lorsque certains seuils sont atteints ou dépassés. Elles peuvent être configurées pour envoyer des notifications par courrier électronique ou SMS afin que les responsables puissent prendre rapidement des mesures correctives si nécessaire.
Par exemple, le responsable du service approvisionnement est alerté par message sonore a chaque rupture du stock d'un produit particulier.

d) **Dashboards mobiles** : Les dashboards mobiles sont conçus pour être accessibles depuis un appareil mobile, ce qui permet aux utilisateurs d'accéder aux données et aux informations en temps réel où qu'ils soient.

e) **Applications métiers** : Les applications métiers sont conçues pour aider les entreprises à prendre des décisions plus éclairées en leur fournissant une vue complète de leurs données et informations commerciales. Ces applications peuvent inclure des outils tels que le suivi du temps, la gestion de projet, la gestion de la relation client (CRM) et la gestion financière.
Après avoir exposer le principe de fonctionnement d'une solution BI, nous présentons ci-dessous les outils ETL.

1.4 Définition des outils ETL

Un logiciel ETL (extract, transform, load) permet de s'assurer que le système est connecté à une ou plusieurs sources de données et que les données sont extraites, transformées et chargées dans l'entrepôt. Il sert de socle au système d'information et constitue un outil clé de l'Entrepôt de Données du système d'information d'une solution BI .

Définition 1.3 ETL est un logiciel permettant d'effectuer des synchronisations massives d'informations entre bases de données. Il commence par l'extraction des données des bases de données de production. Puis, leur transformation pour effectuer

des calculs, pour les enrichir avec des données externes et enfin, le chargement des données dans les différentes applications décisionnelles [2].
La figure 1.2 suivante, illustre le séquencement des trois opérations assurées par un outil ETL.

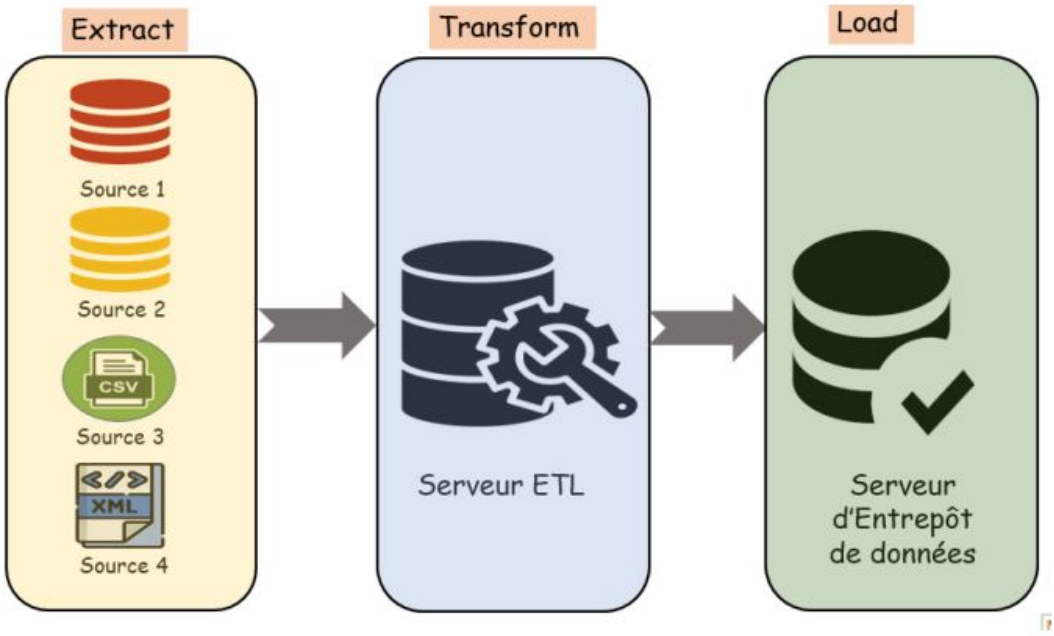


FIGURE 1.2 – Les trois opération d'outil ETL d'après [14]

Ces trois opérations sont examinées en détails dans ce qui suit.

a) **Extractioon des données** : L'ETL se charge de récupérer toutes les données nécessaires depuis les différentes sources de stockage (SGBD, CRM, ERP....).

b) **Transformation des données** : Consiste à appliquer certaines règles de transformations aux données pour les nettoyer, les intégrer et les agréger.

c) **Chargement des données** :L'ETL insère les données dans l'entrepôt de données (Data Warehouse).

1.4.1 Principe de fonctionnement des outils ETL

Les outils ETL assurent l'extraction des données des différentes sources, puis opèrent leurs transformations en des formats plus adéquats et enfin, ils les stockent dans l'entrepôt de données.
La figure 1.3, ci-dessous illustre le principe général de fonctionnement d'un processus ETL. Comme il est observé dans la figure, le mécanisme ETL est un processus incrémental qui passe par plusieurs opérations complémentaires, dont l'explication détaillée est donnée ci-dessous.
Les trois premières opérations constituent l'étape d'extraction, les trois suivantes l'étape de transformation et les trois dernières forment l'étape de chargement.

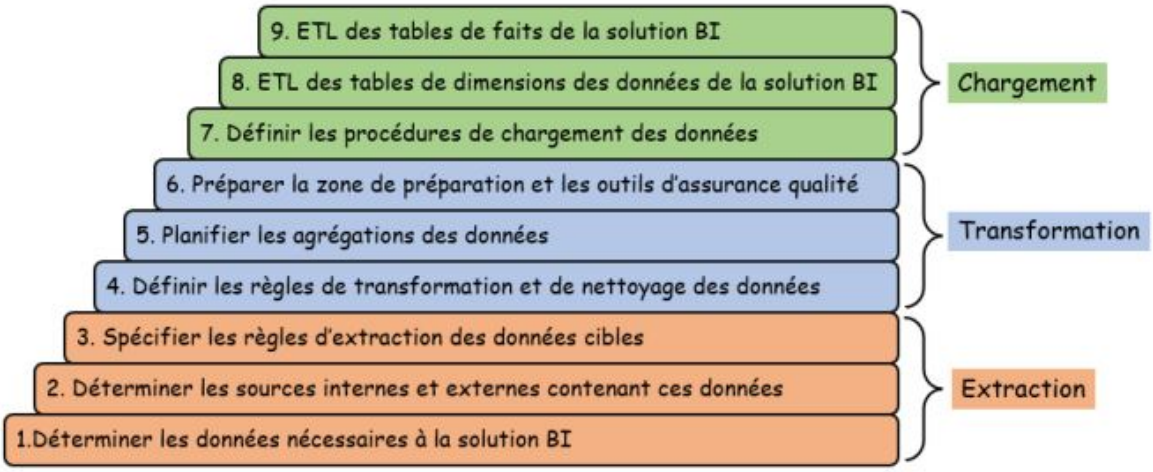


FIGURE 1.3 – Enchainement général des opérations du processus ETL d'après [14]

La section suivante est dédiée à l'exposé de certaines solutions BI existantes dans le domaine.

1.4.2 Les solution ETL commerciales

a) **Talend** : Est une entreprise de logiciels française fondée en 2005. Les produits Talend sont conçus pour aider les entreprises à intégrer, gérer et analyser leurs

données à travers des processus automatisés et l'intégration d'applications, la gestion des données et l'analyse prédictive [3].

b) **Pentaho** : Est une plateforme d'intégration et d'analyse de données open source développée par Hitachi Vantara, une filiale de Hitachi Ltd. Lancée en 2004, elle offre des solutions complètes pour le traitement des données, l'analyse prédictive et la visualisation.

c) **Spagobi** : Est une plateforme open source de BI et d'analyse des données développée par Engineering Group. Lancée en 2006, elle offre des outils pour l'analyse et la visualisation des données, le reporting, le monitoring et la gestion des processus métiers.

d) **Clover ETL** : Est un logiciel d'intégration et de transformation de données (ETL) développé par le fournisseur de logiciels CloverDX. Lancé en 2002, CloverETL est conçu pour aider les entreprises à gérer leurs données et à les transformer en informations exploitables. Il offre une variété d'outils pour extraire, transformer et charger des données à partir de sources diverses, y compris des bases de données relationnelles, des fichiers texte et des systèmes d'entreprise. Il est observé qu'une grande panoplie de système existe sur le marché, nous nous sommes limités à ceux qui sont très populaires.

1.5 Les entrepôts de données

En plus des outils ETL, les entrepôts de données constituent la deuxième composante fondamentale de toute solution BI.
Les EDD sont examinés en détails dans cette section.

1.5.1 Définition d'entrepôts de données

Un Entrepôt de données ou (data warhouse en anlgais) est une base de données utilisée pour collecter et stocker des informations provenant d'autres base de données.

On peut percevoir un entrepôt de données comme une base de données relationnelle conçue pour l'interrogation et l'analyse de données, la prise de décision et les activités de type informatique décsionnelle, plutôt que de traiter des transactions ou d'autres utilisations traditionnelles des bases de données [4].

1.5.2 Fonctionnement des entrepôts de données

Un entrepôt de données fonctionne comme un référentiel central des données. Les informations proviennent d'une ou plusieurs sources de données, telles que des systémes transactionnels ou d'autres bases de données relationnelles. Les données peuvent être de différents format, tels que les données structurées, semi-structurées ou non structurées. Une fois intégrés dans l'entrepôt, elles sont traitées et transformées et prêtes pour toute utilisation future. En effet, les utilisateurs peuvent ensuite y accéder à l'aide d'outils d'informatique décisionnelle (requêtes OLAP : Online analytical Processing), de clients SQL ou simplement via de feuilles de calcul. De plus, les entrepôts de données rendent possible l'exploration de données et leur analyse. Ce processus implique de rechercher les tendances des consommateurs et client par le biais d'extraction des modèles à partir de l'analyse des données. Les décidreus s'appuient sur les modeles élaborer à partir de l'entrepôt pour augmenter les ventes et les revenus de l'entreprise [4].
Comme exemple commercial du fonctionnement d'un entrepôt de données, les entreprises peuvent utiliser un entrepôt de données pour collecter, stocker et analyser les données des clients afin de mieux comprendre leurs habitudes d'achat et leurs préférences. En analysant ces informations, les entreprises peuvent créer des campagnes marketing ciblées qui sont plus susceptibles d'attirer l'attention des clients et de générer plus de ventes.

1.5.3 La modélisation d'un EDD

Dans un EDD, nous distinguons deux types de tables. C'est le mode d'interconnexion de ses tables qui caractérise une modélisation. On trouve donc :

a) **Les tables de dimensions** : Elles sont utilisées pour présenter les données que l'on souhaite stocker dans le DWH, chaque table de dimensions peut avoir plusieurs attributs.

b) **Les tables de faits** : Elles contiennent les données que l'on souhaite voir dans des rapports d'analyse. Une table de faits se présente sous la forme d'un ensemble de colonnes stockant des valeurs dites mesures, et des clés étrangères (identifiant) qui sont généralement les clés primaires associées aux tables de dimensions.

c) **Mesure** : les indicateurs d'analyse sont représentés sous forme de mesures, extraites par une ou plusieurs dimensions.

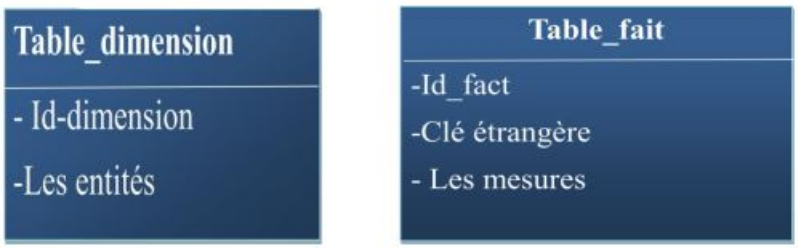


FIGURE 1.4 – Représentation de table de fait et table de dimension

Les modélisations possibles pour organiser les données stockées dans un DWH ; qui sont construites à partir des tables mentionnées ci-dessus sont :

a) **La modélisation en étoile** : Ce modèle tire son nom de sa configuration, en effet il se forme d'un objet central qui est la table des faits reliée à un ensemble de tables de dimensions. Présenté dans la figure 1.5 ci-dessous :

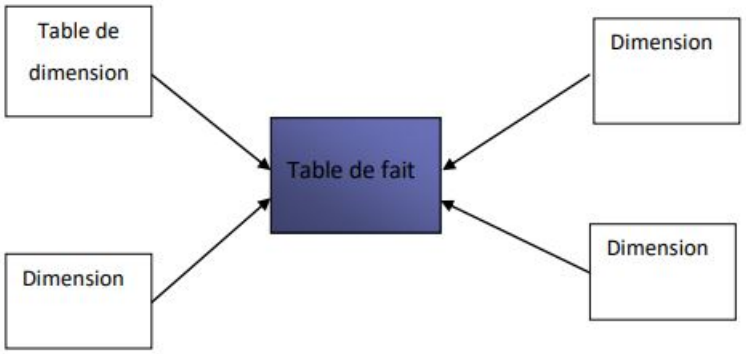


FIGURE 1.5 – Modélisation en étoile

b) **La modélisation en flocons** : Ce modèle est plus complexe de celui en étoile du fait qu'il contient beaucoup plus de tables. Le principe était qu'il peut exister des hiérarchies de dimensions,qui sont reliées à la table de faits.

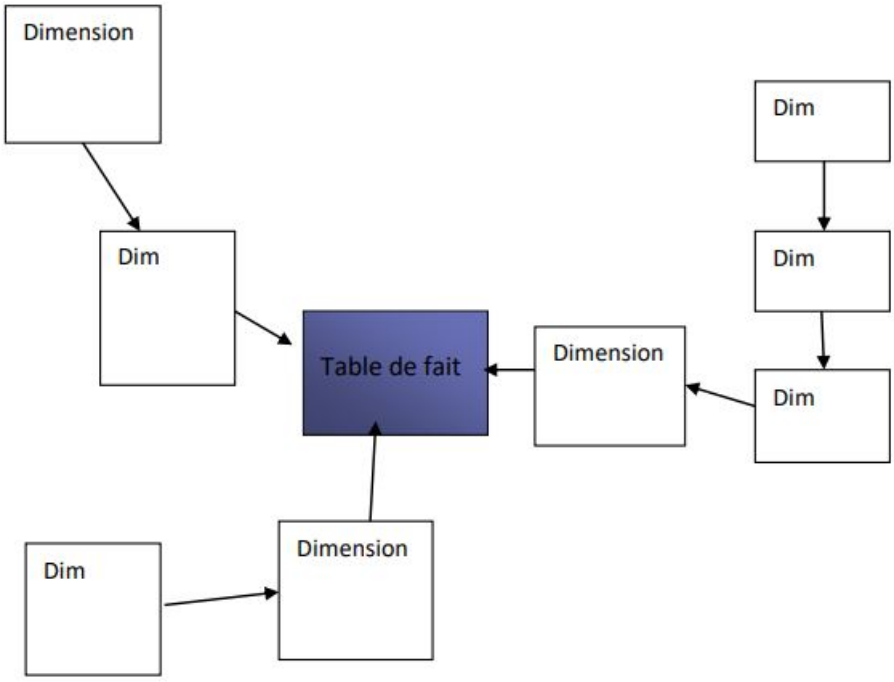


FIGURE 1.6 – Modélisation en flocons

Y a d'autre modélisation et configuration peuvent exister, telle que le modèle en constellation.

1.5.4 Magasin de données ou datamart

Magasin de données ou datamart est un sous-ensemble logique d'un data warhouse.Il est généralement exploité en entreprise pour restituer des informations ciblées sur un métier spécifique.
De nombreuses entreprises utilisent des systèmes de traitement des données en ligne pour accroître leur efficacité et garantir la précision de leurs processus. L'OLTP et l'OLAP sont deux de ces systèmes et offrent aux entreprises diverses fonctionnalités liées au traitement des données en ligne.Nous voulons définir les deux concepts et c'est quoi les différences entre ces deux systèmes .

1.6 Exploitation des EDD

Cette section est réservée à l'exposé des outils OLAP utilisés pour exploiter l'EDD.

1.6.1 Les outils OLAP

Définition 1.5 (OnLine Analytical Processing), ou traitement analytique en ligne permet aux utilisateurs d'effectuer des analyses rapides et efficaces sur de grandes quantités de données [18].
Définition 1.6 OLAP est l'acronyme de Oline Analytical Processing, il désigne les bases de données multidimensionnelles, appelées aussi cubes ou hypercubes, destinées à l'anayse, permet aussi aux utilisateurs d'accéder à des données sommaires plus rapidement et plus facilement [18].

1.6.2 Composants d'un système OLAP

Un système OLAP est composé de plusieurs éléments. Il y a une vue du haut niveau du système, qui comprend une source de données. Il y a également un serveur OLAP et un client. La source de données est constituée des données à analyser. Les données de la source sont transférées ou copiées dans le serveur OLAP où elles sont

organisées et préparées pour fournir des temps d'interrogation courts. Le backend d'un système OLAP est le serveur OLAP. C'est lui qui effectue tout le travail (selon le modèle du système), et où sont stockées les données auxquelles on accède activement. Le client est l'interface utilisateur du serveur OLAP. Le client c'est ce qui est utilisé pour visualiser et manipuler les données dans la base de données. Il peut être aussi un tableur intégrant les fonctions OLAP telles que le pivotement et le forage. Mais aussi, un client peut être un visualiseur de rapports spécialisé mais simple, ou encore une application personnalisée conçue pour une manipulation plus complexe des données.

1.6.3 Types d'OLAP

Il existe généralement trois types d'OLAP à savoir : ROLAP, MOLAP et HOLAP.

a) **ROLAP** : (Relational OLAP) constitue le SGBDR étendu avec mappage de données multidimensionnelles au fonctionnement relationnel standard.

b) **MOLAP** : (Multidimensional OLAP) consiste en une implémentation dans les données multidimensionnelles.

c) **HOLAP** : (Hybrid OLAP) est une approche hybride de la solution où les totaux agrégés sont stockées dans une base de données multidimensionnelle, tandis que les données détaillées sont stockées dans une base de données relationnelle.

Souvent le concept OLAP est confronté à celui d'OLTP. C'est pourquoi, on expose dans la prochaines section des outils OLTP après nous dressons une comparaison entre OLAP et OLTP.

1.7 Les outils OLTP

Définition 1.7 L'OLTP (OnLine Transaction Processing), ou traitement des transactions en ligne, est une méthode permettant d'effectuer des transactions en temps réel à l'aide d'une base de données en ligne qui se met automatiquement à jour au fur et à mesure des transactions [18].

Définition 1.7 Le traitement des transactions en ligne désigne une catégorie de systèmes qui facilitent et gèrent les applications orientées transaction, généralement pour le traitement des transactions de saisie et d'extraction de données. Il a principalement été utilisé pour faire le traitement dans lequel le système répond immédiatement aux demandes de l'utilisateur.
C'est l'exemple d'un guichet automatique pour une banque qui constitue une application de traitement de transactions commerciales.

1.7.1 Avantages d'OLTP

— Simplicité et l'efficacité.

— Il fournit une base concrète pour une organisation stable en raison de la modification en temps voulu de toutes les transactions.

— Il rend les transactions beaucoup plus faciles pour les clients en leur permettant d'effectuer les paiements selon leur choix.

1.7.2 Applications d'OLTP

Les applications OLTP sont souvent utilisées pour saisir de nouvelles données ou mettre à jour données existantes. Un système de saisie des commandes est un exemple typique d'application OLTP.
Nous dressons, dans ce qui suit une comparaison entre les systèmes OLAP et OLTP.

1.7.3 Comparaison OLAP | OLTP

OLTP	OLAP
très rapide	Rapide
Fermés : on ne laisse pas la place à l'imprévu dans les OLTP, les utilisateurs sont guidés dans le processus	Simplicités : les environnements d'un SID doivent permettre d'accéder le plus simplement possible aux données
Transactionnels : OLTP fonctionne selon les principes de transaction	Non transactionnels : l'utilisateur doit pouvoir démarrer l'analyse, puis revenir en arrière
Fragmentés : ou dispersé. Sauf le cas des ERP	Centralisés : toutes les données sont regroupées dans une seule source
Grand public : destinés à toute personne	Petit public : sauf les décideurs
Faible niveau de demande d'analyse	Niveau élevé de demande analytique

En plus des outils OLTP et OLAP qui sont utilisés pour exploiter des EDD, il existe d'autres outils que nous allons exposer ci après.

1.8 Autres outils d'exploitation des EDD

1.8.1 Les outils de reporting

Les outils de reporting sont des outils qui permettent aux entreprises de collecter, analyser et présenter des données. Ils peuvent être utilisés pour créer des rapports, des tableaux de bord et des graphiques qui peuvent aider les entreprises à prendre des décisions éclairées.
Les outils de reporting peuvent inclure :

a) **Des tableaux de bord interactifs** : Sont une collection d'indicateurs clés qui sont regroupés sur une seule page et mis à jour en temps réel. Ils peuvent être utilisés pour surveiller les performances d'une entreprise et identifier les tendances.

b) **Des graphiques** : Les graphiques sont un moyen visuel d'illustrer les données et les tendances. Les types courants de graphiques comprennent les diagrammes à barres, les diagrammes circulaires, les lignes et les histogrammes.

c) **Des rapports** : Les rapports sont une collection organisée de données qui peuvent être utilisés pour analyser le rendement d'une entreprise ou pour comparer différents aspects d'une activité commerciale.

d) **Des indicateurs clés de performance (KPI)** : Les KPI sont des mesures spécifiques qui peuvent être utilisées pour évaluer la performance d'une entreprise ou d'un processus particulier.

e) **Des alertes** : Les alertes sont un moyen pratique de surveiller en temps réel certains indicateurs clés afin que vous puissiez prendre rapidement des mesures si nécessaire.

1.8.2 Les outils de fouille de données (data mining)

Les outils de data mining sont des logiciels qui aident à extraire, analyser et interpréter les données. Ils peuvent être utilisés pour créer des modèles prédictifs, identifier les tendances et mieux comprendre le comportement des consommateurs.
Vu leur importance dans le contexte de notre sujet, et étant donné leur diversité, nous les présentons dans le prochain chapitre.

1.9 Conclusion

Dans ce chapitre, nous avons exposé les définitions et concepts de base de la BI, d'entrepôt de données et les architectures qui leur sont afférentes. Le prochain chapitre sera consacré à la présentation des différentes techniques utilisées pour exploiter

les données stockées dans un EDD afin d'extraire des connaissances utiles à la prise de décision.

CHAPITRE 2

LES TECHNIQUES DE FOUILLE DE DONNÉES

2.1 Introduction

Ce chapitre est consacré à la présentation des différentes techniques utilisées pour exploiter les données stockées dans un EDD afin d'extraire des connaissances utiles à la prise de décision.
Nous commençons par exposer la définition du domaine de fouille de données.

2.2 Définition de la fouille de données

Fouille de données (ou Data mining en anglais), se concentre sur l'extraction de connaissances cachées et non triviales à partir de grandes quantités de données. Fouille de données peut être utilisée comme une technique de soutien à la BI.
Il existe de nombreuses techniques d'exploration de données qui ont été développées et utilisées dans des projets d'exploration de données, notamment l'association, la classification, le regroupement, l'arbre de décision, la prédiction et les réseaux neuronaux, etc. Chaque technique possède ses propres règles et méthodes, en fonction du type de problème à résoudre.

A titre d'exemple, un magasin peut utiliser la fouille de données pour analyser ses ventes et déterminer quels produits sont les plus populaires auprès des clients.
Dans les sections suivantes, nous allons examiner en détails ces techniques d'extraction de données.
La suite de ce chapitre est réservée à la présentation des différentes technique de FDD.

2.3 Aperçu des techniques de fouille de données

Les techniques de fouille de données peuvent être utilisées pour enrichir les systèmes de Business Intelligence en fournissant des valeurs (connaissances) supplémentaires à partir de grandes quantités de données.
Nous pouvons classer ces techniques en quatre grandes familles :

— Les techniques statistiques.

— Les techniques de base de données.

— Les techniques d'analyse de données.

— Les techniques de l'intelligence artificielle.

Pour des raisons de clarté de mémoire, ces différentes techniques sont examinées en détails ci-dessous, dans des sections séparées.

2.4 Les techniques statistiques

Les techniques statistiques sont une forme de fouille de données qui permet d'analyser et de comprendre les données. Elles peuvent être utilisées pour découvrir des tendances, des corrélations et des relations entre différents ensembles de données. Les techniques statistiques sont largement utilisées dans le domaine des sciences sociales, de la finance, du marketing et de la recherche scientifique.
Voici quelques exemples courants de techniques statistiques qui peuvent êtres déployées dans le contexte d'une solution BI :

2.4.1 La moyenne

C'est une mesure descriptive qui permet d'obtenir une valeur représentative à partir d'un ensemble de données. Par exemple, si vous avez un ensemble de 10 nombres (1, 2, 3, 4, 5, 6, 7, 8, 9 et 10), la moyenne serait 5.5 (la somme des nombres divisée par le nombre total).
En commercial, la moyenne des ventes et la somme des ventes divisée par le nombre des vendeurs.

2.4.2 La médiane

La médiane est une autre mesure descriptive qui permet d'obtenir une valeur représentative à partir d'un ensemble de données. Elle est calculée en triant les données par ordre croissant ou décroissant et en trouvant le nombre au milieu du jeu de données. Dans l'exemple ci-dessus (1-10), la médiane serait 5 (le nombre au milieu).

2.4.3 Le mode

C'est une autre mesure descriptive qui permet d'obtenir une valeur représentative à partir d'un ensemble de données. Il s'agit du nombre qui apparaît le plus souvent dans un jeu de données. Dans l'exemple ci-dessus (2, 4, 6, 4, 8, 4, 10) le mode est 4 car il apparaît trois fois, plus souvent que les autres nombres.

2.4.4 Le maximum

Le maximum est une mesure statistique qui permet de déterminer la plus grande valeur d'un ensemble de données. Par exemple, si on a cinq nombres (1, 2, 3, 4 et 5), le maximum serait 5.

2.5 Les requêtes d'agrégations

Les bases de données offrent des requêtes d'agrégations pour calculer et trier les données de l'EDD. Voici un exemple de fichier de base de données contenant des informations de ventes d'une entreprise commerciale :

Code_produit	Nom_produit	Catégorie	Prix unitaire	Quantité vendue
1	Stylo	Papeterie	1.50	100
2	Cahier	Papeterie	3.00	50
3	Crayon	Papeterie	0.75	200
4	Clé USB	Électronique	10.00	20
5	Souris sans fil	Électronique	15.00	30

TABLE 2.1 – Table des ventes d'un magasin

2.5.1 La Somme

La commande Somme est utilisée pour additionner les valeurs d'un champ spécifique dans une table.
Maintenant, appliquons la requête SQL **SUM** sur le tableau ci dessus :
SELECT SUM('Prix unitaire' * 'Quantité vendue') AS 'Total des ventes' FROM Ventes ;
Résultat de la requête :

Total des ventes
605.00

2.5.2 Regroupement des données «*Grouped by*»

La commande Group By est utilisée pour regrouper les enregistrements d'une table selon un ou plusieurs champs.
Maintenant, appliquons la requête SQL **GROUP BY** sur le tableau des Ventes précèdent :

SELECT Catégorie, SUM (Quantité vendue)

AS 'Total des ventes' FROM Ventes

GROUP BY Catégorie ;

Résultat de la requête :

Catégorie	Total des ventes
Papeterie	350.00
Électronique	50.00

Explication : On a une table qui contient des informations sur les ventes de produits, on a utiliser la fonction Group By pour regrouper les enregistrements par catégorie de chaque produit et calculer le total des ventes pour chaque produit.

2.5.3 Tri des données *«Sorted by»*

La requête **Sorted By** est utilisée pour trier les enregistrements d'une table selon un ou plusieurs champs.

SELECT * FROM Ventes ORDER BY Prix unitaire DESC ;

Résultat de la requête :

Code_produit	Nom_produit	Prix unitaire	Quantité vendue
5	Souris sans fil	15.00	30
4	Clé USB	10.00	20
2	Cahier	3.00	50
1	Stylo	1.50	100
3	Crayon	0.75	200

Explication : On une table qui contient des informations sur les ventes de magasin, on a utiliser la requête **Sorted By** pour trier les enregistrements par prix et afficher le produit le plus cher en premier.

Après avoir exposé les techniques de base de données, nous présentons ci-après les techniques d'analyse de données.

2.6 Les techniques d'analyse de données

Il existe plusieurs techniques, mous examinons les importantes.

2.6.1 Analyse factorielle des correspondances (AFC)

Cette technique est utilisée pour explorer les relations entre plusieurs variables qualitatives. Par exemple, une AFC peut être utilisée pour étudier les préférences alimentaires des consommateurs en examinant leurs choix alimentaires par rapport à différents types de produits alimentaires [26].

2.6.2 Analyse factorielle discriminante (AFD)

Cette technique est utilisée pour distinguer entre différents groupes de données basés sur des variables qualitatives et quantitatives. Par exemple, une AFD peut être utilisée pour déterminer si un groupe de personnes avec un certain type de caractéristiques est plus susceptible d'acheter un produit que les autres groupes [26].

2.6.3 Analyse en composantes principales (ACP)

L'ACP utilisée pour réduire le nombre de variables dans un jeu de données et expliquer la variabilité totale des données avec le moins grand nombre possible de variables explicatives. Par exemple, une ACP peut être utilisée pour réduire le nombre de variables dans un jeu de données sur les habitudes alimentaires des consommateurs et expliquer la variabilité totale des données avec seulement quelques variables explicatives clés [26].

2.6.4 La régression

C'est une méthode statistique qui permet d'étudier la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle permet de prédire la valeur d'une variable à partir des valeurs des autres variables. Par exemple, on peut

utiliser la régression pour étudier le lien entre le niveau de revenu et le niveau d'éducation. On peut alors prédire le niveau de revenu en fonction du niveau d'éducation. A présent, nous allons présenter les techniques de l'IA pour l'exploration de données.

2.7 Les techniques de l'intelligence artificielle

2.7.1 La recherche de motifs

La recherche de motifs (ou patterns) est un aspect important du data mining. Les algorithmes de data mining cherchent à identifier des relations cachées, des tendances, des modèles récurrents et d'autres patterns dans les données. Les résultats de cette analyse peuvent être utilisés pour prédire les tendances futures, améliorer les processus commerciaux, optimiser les ressources et prendre des décisions plus informées [25].

Exemple, un chercheur peut utiliser cette technique pour découvrir des motifs dans les données de vente d'un magasin. Il peut rechercher des tendances telles que les jours où les ventes sont plus élevées, les produits qui se vendent le mieux et les périodes de l'année où les ventes sont plus fortes.

2.7.2 Les règles d'association

L'association est une technique qui consiste à identifier les relations entre les items dans un ensemble de données. Les algorithmes d'association recherchent des relations fréquentes entre les items, qui peuvent être utilisées pour générer des règles d'association. Ces règles indiquent quels items sont souvent achetés ensemble et peuvent être utilisées pour prédire les habitudes d'achat des consommateurs, optimiser les campagnes de marketing, etc.

Exemple, une entreprise peut associer sa marque à un produit populaire, comme une boisson énergisante, afin de créer une association positive entre les deux. Cela peut aider à améliorer la notoriété de la marque et à attirer plus de clients.

2.7.3 Les arbres de décision

L'arbre de décision est une technique de fouille de données qui permet d'analyser des ensembles de données complexes et d'en extraire des informations utiles. Il s'agit d'une méthode graphique qui utilise un arbre pour représenter les différentes possibilités et leurs conséquences [25].

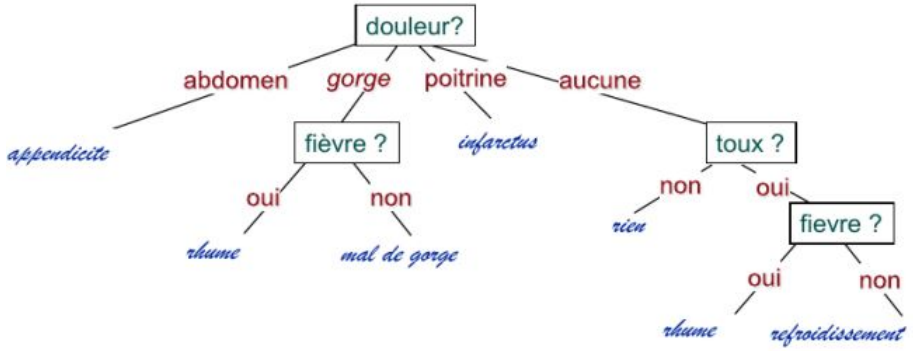


FIGURE 2.1 – Arbre de décision pour le diagnostic d'une douleur d'après [26]

2.7.4 Les techniques de Prédiction

La prédiction est une une technique qui utilise des modèles statistiques et d'apprentissage automatique pour prédire des résultats futurs.
Par exemple, un détaillant peut utiliser l'analyse prédictive pour prédire les tendances des ventes à l'aide de données historiques sur les ventes, les prix et le comportement des consommateurs.

2.7.5 La classification

On distingue deux types de classification :

a) **Classification Supervisée** : Est une forme d'apprentissage automatique où un algorithme apprend à classer des données en fonction de leurs caractéristiques.

L'algorithme est entraîné sur des données étiquetées, ce qui signifie que chaque exemple a une étiquette qui indique à quelle catégorie il appartient.
Exemple, un algorithme de classification supervisée pourrait être entraîné pour reconnaître les différents types de fruits en fonction de leurs couleurs et formes.

b) **Classification Non Supervisée** : La classification non supervisée est une forme d'apprentissage automatique où un algorithme apprend à classer des données sans étiquettes prédéfinies. L'algorithme analyse les caractéristiques des données et détermine comment les groupes sont liés entre eux.
Exemple, un algorithme de classification non supervisée pourrait être utilisé pour regrouper des clients en fonction de leurs habitudes d'achat et de leur comportement.

2.7.6 Les techniques de clustering

La technique de clustering est utilisée pour regrouper des données en groupes similaires. Le clustering est l'ensemble de clusters qui sont produits par une analyse de cluster. Dans ce contexte, une variété de techniques de clustering différentes peuvent être utilisées pour produire différents clusterings sur le même ensemble de données. Le partitionnement est effectué par un algorithme de clustering plutôt que par des humains. Par conséquent, le regroupement est utile car il peut aider à trouver des groupes qui étaient jusque-là inconnus dans les données [26].
Exemple, un détaillant peut utiliser la technique de clustering pour regrouper ses clients en différents groupes selon leurs habitudes d'achat. Les groupes peuvent être basés sur des caractéristiques telles que le type de produits achetés, le montant dépensé ou le nombre d'achats effectués. Une fois les groupes créés, le détaillant peut alors cibler ses campagnes marketing et offres spéciales à chaque groupe afin d'améliorer l'efficacité et l'efficience de ses efforts marketing.

2.8 Les domaines d'application du data mining en informatique décisionnnelle

Vue leur importance, les techniques de FDD sont largement utilisées en informatique décisionnelle. On peut citer notamment, les domaines phares suivants :

2.8.1 Prévision des ventes

Les techniques de fouille de données peuvent être très utiles dans la prévision des ventes, car elles permettent d'analyser les données historiques de ventes et d'identifier les modèles et les tendances qui peuvent aider à prédire les ventes futures. En utilisant ces techniques de fouille de données, les entreprises peuvent prédire les ventes futures avec une plus grande précision et prendre des décisions plus éclairées concernant la gestion des stocks, la planification de la production et les stratégies de marketing. Cela peut aider à améliorer l'efficacité opérationnelle et à stimuler la croissance de l'entreprise.

2.8.2 Détection et prévention des fraudes

L'analyse et la détection des fraudes est un enjeu important pour les entreprises, car les fraudes peuvent avoir un impact financier et réputationnel important. Dans le cadre d'une solution Business Intelligence (BI), l'analyse de données peut aider à identifier les transactions suspectes et à détecter les fraudes potentielles.

2.8.3 Analyse du panier de consommation

L'analyse du panier de consommation est une technique d'analyse de données qui permet d'identifier les produits ou services qui sont souvent achetés ensemble par les clients. Cette technique peut être utilisée dans le cadre d'une solution Business Intelligence (BI) pour mieux comprendre le comportement d'achat des clients et pour optimiser les stratégies de vente et de marketing.

2.8.4 Analyse des médias sociaux

L'analyse des médias sociaux est une application importante des techniques de fouille de données. Les médias sociaux sont une source riche de données non structurées, qui peuvent être exploitées pour mieux comprendre les opinions, les sentiments et les comportements des utilisateurs.

2.9 Conclusion

Dans ce chapitre, nous avons exposé les différentes techniques de fouille de données et nous les avons illustrées par des exemples réels.
Ces techniques constituent un moyen incontournable pour l'exploration des données de l'EDD.
Le prochain chapitre sera consacré à une analyse des travaux existants pour l'intégration des techniques de fouille de données dans solution BI.

CHAPITRE 3
PROBLÉMATIQUE ET TRAVAUX CONNEXES

3.1 Introduction

Les techniques de fouille de données ont été largement utilisées dans différents domaines, tout en exploitant des structures de données spécifiques, telles que les bases de données, les arbres et structures de données complexes (piles, files,....ect). D'autre part, le domaine de la BI s'est appuyé fondamentalement sur les techniques OLAP en exploitant le données stockées dans L'EDD.
Dans ce sujet, nous essayerons de faire la jonction entre les deux domaines, à savoir : la BI et FDD en vue de faciliter la prise de décision en entreprise . Dans l'ensemble, l'intégration des techniques de fouille de données dans une solution business intelligence peut fournir aux organisations un ensemble d'outils puissants pour découvrir des informations précieuses à partir de leurs données. En utilisant ces informations, les organisations peuvent prendre des décisions plus éclairées, améliorer l'efficacité opérationnelle, améliorer l'expérience client et gérer les risques plus efficacement.
Ce chapitre est dédié à la présentation de la problématique, des motivations et enfin on examinera les travaux connexes qui ont abordé ce sujet.

3.2 Problématique

L'exploitation de technique de fouille de données contribuera considérablement à l'enrichissement des systèmes décisionnels, par la fourniture des informations utiles dans le cadre de la BI. En effet, les techniques statistiques, celles des bases de données, d'analyse de données et aussi d'intillegence artificielle pourront être exploitées afin de produire différents indicateurs clés de performance qui peuvent servir d'éléments de base pour renforcer tout système décisionnel.
Notre objectif est de savoir comment intégrer efficacement les techniques de fouille de données dans un système décisionnel pour améliorer la prise de décision. Il s'agit d'exploiter les données stockées dans un entrepôt de données en utilisant des algorithmes et des techniques adéquates pour identifier des motifs spécifiques qui peuvent servir d'éléments clés de performance dans tout système décisionnel. L'objectif est donc de concevoir un système décisionnel qui utilise les techniques de fouille de données pour extraire des motifs d'intérêt et produire des éléments d'appréciation utiles à la prise de décision en entreprise.

3.3 Motivations

L'intégration des techniques de fouille de données dans une solution business intelligence est de plus en plus importante à mesure que les organisations collectent et stockent des quantités de données toujours plus importantes. Les techniques de fouille de données, telles que exposées dans le chapitre 2, peuvent aider les organisations à analyser ces données et à découvrir des informations précieuses qui peuvent éclairer la prise de décision et stimuler la croissance de l'entreprise.
Les avantages dont peuvent bénificier les entreprises par l'utilisation des techniques de fouille de données dans le contexte dans solution BI se résument aux aspects suivants :

a) **Prise de décision plus éclairée** : Les techniques de fouille de données peuvent aider à identifier des modèles, des tendances et des relations dans les données stockées dans l'EDD et qui pourraient ne pas être évidents par des méthodes d'analyse traditionnelles. En découvrant ces informations, les organisations peuvent prendre des décisions plus éclairées concernant l'allocation des ressources, le développement de produits, les stratégies de marketing, etc.

b) **Amélioration de l'efficacité opérationnelle** : Les techniques de fouille de données peuvent aider les organisations à identifier des zones d'inefficacité ou de gaspillage dans leurs opérations. En analysant les données de l'EDD et relatives aux ventes de production et d'autres métriques clés, les organisations peuvent identifier des opportunités pour rationaliser les processus, réduire les coûts et améliorer l'efficacité globale.

c) **Amélioration de l'expérience client** : Les techniques de fouille de données peuvent aider les organisations à analyser les données clients pour mieux comprendre leurs besoins, leurs préférences et leurs comportements. En adaptant les produits, les services et les efforts de marketing à des segments de clients spécifiques, les organisations peuvent améliorer l'expérience client globale et stimuler la fidélité.

d) **Meilleure gestion des risques** : Les techniques de fouille de données peuvent aider les organisations à identifier les risques et les menaces potentiels pour leurs opérations. En analysant les données de comportement client, les transactions financières et d'autres métriques clés, les organisations peuvent identifier la fraude potentielle ou d'autres types de risques et prendre des mesures pour les atténuer.

e) **Elaborer des recommandations pour la personnalisation des clients :** Pour élaborer ces recommandations, les entreprises utilisent les techniques de segmentation de marché pour identifier les groupes de clients ayant des besoins similaires. Ensuite, elles peuvent utiliser les modèles de prédiction pour comprendre le comportement d'achat de chaque segment de client et pour prédire

les produits et services qui conviendraient le mieux à chaque client et aider les entreprises à élaborer des recommandations personnalisées pour répondre aux besoins des clients. Cela peut améliorer la satisfaction des clients, augmenter les ventes et la fidélité des clients, et renforcer la position concurrentielle de l'entreprise.

f) **Détection des risques et des anomalies :** Les techniques de fouille de données utilisent des algorithmes et des modèles statistiques pour analyser les données et détecter les risques et les anomalies. Les méthodes courantes incluent l'analyse de la variance, la régression, l'analyse de cluster, l'analyse en composantes principales, la classification et la détection d'anomalies. Ces techniques peuvent être utilisées pour identifier des schémas inattendus ou des comportements anormaux dans les données, ce qui peut indiquer un risque potentiel ou une anomalie. La détection d'anomalies peut également être utilisée pour identifier des points de données qui se situent en dehors de la norme et qui peuvent indiquer un risque ou une anomalie potentielle.

Guidé par ces motivations, dans ce mémoire, nous proposons une solution pour l'intégration des techniques de FD dans une solution BI.
Dans la section suivante, une analyse des travaux connexes qui met la lumière sur ce qui a été fait, aussi bien dans le domaine de la recherche académique que dans le domaine industriel est exposée.

3.4 Travaux connexes

Nous abordons les travaux connexes selon deux aspects : celui de recherche académique aussi-bien que celui du domaine industriel.

3.4.1 Dans le domaine de la recherche académique

Plusieurs travaux de recherche académique ont abordé la problématiquede l'intégration des techniques de fouille de données dans les solutions BI. On peut les classer en quatres catégories principales.

— Les techniques statistiques dans BI.

— Les techniques de base de données dans BI.

— Les techniques de l'intelligence artificielle dans BI.

— Les techniques de l'analyse de données dans BI.

■ Dans [16], les auteurs proposent une approches qui combine des techniques de Business Intelligence et d'analyse de données pour extraire des informations significatives à partir de données volumineuses. Plus spécifiquement, l'article présente une méthodologie en quatre étapes pour l'analyse de données en entreprise, qui comprend :

— La collecte de données à partir de sources multiples, y compris les données structurées et non structurées.

— Le nettoyage et la préparation des données pour l'analyse, y compris l'utilisation de techniques statistiques pour la normalisation, la transformation et la sélection des variables.

— L'analyse des données à l'aide de techniques de Business Intelligence et d'analyse statistique, telles que les analyses de corrélation, de régression et de clustering.

— La présentation des résultats de l'analyse de données sous forme de rapports, de tableaux de bord et de visualisations graphiques pour faciliter la prise de décision.

En utilisant cette approche, l'article propose que les organisations peuvent mieux comprendre leurs données, identifier des tendances et des modèles cachés, et améliorer la prise de décision en utilisant des informations factuelles basées sur les données.

■ Dans [21], les auteurs proposent une approche basée sur l'utilisation de la BI et des méthodes de fouille de données pour améliorer la gestion des données dans le secteur des télécommunications. Les techniques de FDD que les auteurs

proposent comprennent notamment l'analyse de clustering, l'analyse de classification, l'analyse de régression. Les auteurs soulignent l'importance de l'intégration des méthodes de fouille de données dans les systèmes de gestion des entreprises de télécommunications pour assurer une utilisation efficace et efficiente de ces outils.

Malgé l'éfficacité de l'approche, elle reste limite au domaine spécifique des télécommunications. En plus cette approche intégre uniquement les techniques d'IA.

- Dans [22] propose une approche pour utiliser des techniques de fouille de données et d'analyse commerciale (business analytics) dans le domaine de BI. L'objectif principal de cette approche est d'exploiter les données disponibles dans les organisations pour prendre des décisions éclairées, améliorer la performance commerciale et obtenir un avantage concurrentiel. L'article décrit probablement différentes techniques de fouille de données et d'analyse commerciale qui peuvent être utilisées pour extraire des informations précieuses à partir des données commerciales, telles que l'exploration de données, la modélisation prédictive, l'analyse des tendances, la segmentation de la clientèle, etc. L'article aborder divers domaines commerciaux où la business intelligence et l'analyse de données peuvent être appliquées, tels que la finance, le marketing, la gestion des ressources humaines, la logistique, la gestion de la chaîne d'approvisionnement, la gestion de la relation client, ou d'autres domaines pertinents où l'utilisation de techniques de fouille de données incluent l'arbre de décision, la régression, les algorithmes de clustering, les réseaux de neurones, les algorithmes d'association, les méthodes de classification, l'analyse de texte, et d'autres méthodes de modélisation et d'exploration de données.

  Malgé l'éfficacité de l'approche, elle intégre uniquement les techniques d'IA et se baser uniquement sur des données simulées ou des scénarios hypothétiques, sans inclure d'études de cas réelles ou de données réelles pour valider les résultats obtenus.

- Dans [24], les auteurs proposent une approche pour le développement de BI basée sur l'intégration de données et l'exploration de données. Dans leur approche, il proposent d'utiliser des techniques d'intégration de données pour rassembler des données hétérogènes provenant de différentes sources, telles que des bases de données internes, des données externes, des données en ligne, etc. Ils mettent en avant l'importance de la qualité des données et de la gestion des méta-données dans le processus d'intégration des données. Il ont utilisé des requêtes de base de données pour interroger et récupérer des données à partir des sources intégrées afin de les analyser et d'en extraire des informations utiles. Dans [23], les auteurs exposent les différentes méthodes de data mining et d'analyse commerciale qui peuvent être utilisées pour extraire des informations utiles à partir de grandes quantités de données d'entreprise. Ils mettent en évidence comment ces techniques peuvent être appliquées dans divers domaines de l'entreprise, tels que la banque ou la finance, le marketing ou le commerce de détail, l'assurance, l'analyse de données biomédicales et d'ADN, ainsi que l'industrie des télécommunications, pour prendre des décisions éclairées et basées sur des données. Les auteurs présentent également des exemples concrets d'utilisation de ces techniques dans des études de cas réels pour illustrer leur application pratique. Ils mettent en évidence les avantages potentiels de l'utilisation de ces techniques, tels que l'amélioration de l'efficacité opérationnelle, l'optimisation des processus d'affaires, l'identification de nouvelles opportunités commerciales, la gestion des risques, etc.
- Le travail proposé dans l'article [19], expose une méthodologie détaillée pour évaluer la BI des systèmes d'entreprise. Cela comprend l'identification des critères clés d'évaluation, la définition des indicateurs de performance clés (KPI - Key Performance Indicators) pertinents pour BI, ainsi que la mise en œuvre d'une approche pour mesurer ces indicateurs. L'utilisation de techniques d'analyse des données, telles que l'exploration de données, l'analyse statistique, l'analyse de tendances, l'analyse des modèles, etc., pour extraire des informations

significatives et des tendances à partir des données collectées.

- L'article [17] porte sur l'utilisation d'un système de veille stratégique qui fait appel à des techniques statistiques pour repérer les fraudes dans les données relatives aux paiements des demandes de remboursement des médicament par la sécurité sociale (Medicaid). Les auteurs proposent une approche basée sur la Business Intelligence, qui consiste à utiliser des techniques d'analyse de données, de modélisation statistique et de visualisation pour identifier les schémas de fraude potentiels dans les demandes de remboursement Medicaid. Les auteurs mettent en avant l'importance de la Business Intelligence comme outil pour améliorer la détection de la fraude dans le domaine de la santé, en utilisant des données disponibles dans les systèmes de gestion des réclamations Medicaid.
- Les auteurs dans [20] proposent une méthodologie empirique pour étudier l'impact de la BI sur la performance financière des start-ups. Plus précisément, les auteurs ont collecté des données auprès de 245 start-ups en Chine en utilisant une enquête par questionnaire en ligne. Les données collectées comprenaient des mesures de l'utilisation de BI, de la performance financière et des caractéristiques de l'entreprise. Les données ont ensuite été analysées à l'aide de techniques statistiques, notamment une analyse de régression.

Catégorie	Réf	Principale contribution	Avantages	Inconvénients
IA,ADD	[16]	Met en avant l'adoption généralisée de la BI comme une approche clé pour analyser les données et générer des informations exploitables afin d'améliorer la prise de décision et les performances commerciales.	Amélioration des performances commerciales.	Complexité de l'implémentation.
IA	[21]	L'utilisation de la BI et des méthodes de fouille de données pour améliorer la gestion des données dans le secteur des télécommunications	Améliorer la gestion des données dans le secteur des télécommunications	Limité au domaine des télécommunications.
IA	[22]	Exploiter les données disponibles dans les organisations avec différentes techniques de fouille de données et d'analyse commerciale à partir des données commerciales	Prendre des décisions éclairées, améliorer la performance commerciale et obtenir un avantage concurrentiel.	Intégre uniquement les techniques d'IA.
BDD	[24]	L'intégration de données et l'exploration de données pour développement de la BI.	Concentre sur le développement de BI	Approche gńéraliste et manque de détails sur les techniques utilisées.

TABLE 3.1 – Synthèse des travaux de recherche académique

Catégorie	Réf	Principale contribution	Avantages	Inconvénients
STAT,ADD	[19]	Méthodologie et un cadre d'évaluation pour mesurer et évaluer la capacité des systèmes d'entreprise.	Mesurer et évaluer la capacité des systèmes d'entreprise.	Manque d'utilisation de techniques (uniquement 2).
STAT	[17]	Une méthodologie de détection de la fraude dans les demandes de remboursement Medicaid en utilisant des concepts de Business Intelligence.	Aider les analystes à prendre des décisions éclairées dans la gestion des demandes de remboursement Medicaid.	Limité au domaine de la santé et remboursement des frais.
STAT	[20]	Examiner l'impact de l'utilisation de BI sur la performance financière des start-ups.	Prendre des décisions éclairées en matière d'investissement dans la BI pour leur entreprise.	limité uniquement les techniques statistiques.

TABLE 3.2 – Synthèse des travaux de recherche académique

3.4.2 Dans le domaine de l'industrie du logiciel

Plusieurs logiciels et outils ont été proposés par différent firmes pour prendre en charge le problème de l'intégration des techniques de fouille de données dans un système BI.
Distingue notamment :

A) **Talend :** Talend est une plateforme d'intégration de données open source, qui permet de simplifier et d'automatiser les tâches d'intégration de données, de transformation, de nettoyage et de gestion des données, Voici quelques détails sur le fonctionnement de Talend open source :

1. **Architecture :** Talend utilise une architecture modulaire basée sur des composants (appelés "talend job" ou "job") qui sont assemblés en workflows pour répondre aux besoins spécifiques des projets d'intégration de données. Les composants Talend sont réutilisables et couvrent une large gamme de fonctions pour l'intégration de données.
2. **Connectivité :** Talend prend en charge une grande variété de connecteurs de données pour se connecter à différents systèmes de sources de données, y compris les bases de données relationnelles, les systèmes de fichiers, les services web, les services de cloud computing et bien plus encore.
3. **Transformation de données :** Talend permet de transformer les données pour répondre aux besoins de l'entreprise. Les transformations peuvent être effectuées à l'aide d'une interface graphique intuitive pour la création de workflows, d'expressions et de fonctions pour la manipulation de données.
4. **Déploiement :** Les jobs de Talend peuvent être déployés sur différents environnements, y compris les serveurs locaux, les serveurs distants et les environnements cloud. Talend prend également en charge les outils de gestion de version pour gérer les versions des jobs et les mises à jour de code.

5. **Supervision :** Talend fournit des fonctionnalités de supervision pour surveiller les performances des jobs, les erreurs, les alertes et les métriques de qualité des données en temps réel.

Talend open source est une solution complète d'intégration de données qui offre des fonctionnalités avancées pour l'intégration de données, le traitement en temps réel, la qualité des données et la gouvernance des données. En tant que plate-forme open source, Talend permet aux entreprises de bénéficier d'une solution d'intégration de données puissante et évolutive sans les coûts élevés associés aux licences logicielles propriétaires [3].

B) **Pentaho :** Pentaho est une plate-forme open source de la BI qui permet aux entreprises de collecter, de stocker, de traiter et de visualiser des données[5]. Voici quelques détails sur le fonctionnement de Pentaho :

1. **Collecte de données :** Pentaho prend en charge de nombreux types de sources de données, notamment les bases de données relationnelles, les systèmes de fichiers, les applications, les services web, les services de cloud computing et les réseaux sociaux.
2. **Intégration de données :** Pentaho permet d'extraire, de transformer et de charger (ETL) les données à partir de différentes sources de données. Les données peuvent être nettoyées, transformées et agrégées à l'aide de la fonctionnalité ETL.
3. **Stockage de données :** Pentaho prend en charge plusieurs bases de données, notamment MySQL, Oracle, Microsoft SQL Server, PostgreSQL et Hadoop, pour stocker les données collectées et intégrées.
4. **Analyse de données :** Pentaho fournit des fonctionnalités d'analyse de données, notamment des rapports, des tableaux de bord, des graphiques et des analyses multidimensionnelles. Les utilisateurs peuvent créer des visualisations personnalisées pour analyser les données en fonction de leurs besoins.

5. **Sécurité :** Pentaho fournit des fonctionnalités de sécurité pour protéger les données et les ressources. Les utilisateurs peuvent définir des rôles et des permissions pour les utilisateurs et les groupes.
6. **Déploiement :** Pentaho peut être déployé sur site ou dans le cloud. Les utilisateurs peuvent également déployer des solutions Pentaho sur des appareils mobiles.

Pentaho est une solution complète de BI qui offre des fonctionnalités avancées pour la collecte, l'intégration, le stockage et l'analyse de données. En tant que plate-forme open source, Pentaho permet aux entreprises de bénéficier d'une solution BI puissante et évolutive sans les coûts élevés associés aux licences logicielles propriétaires [6].

C) **SQL Server Integration Services (SSIS) :** Est une plate-forme d'intégration de données propriétaire de Microsoft qui permet aux entreprises de collecter, transformer et charger (ETL) les données à partir de différentes sources de données. Voici quelques détails sur le fonctionnement de SSIS :

1. **Collecte de données :** SSIS prend en charge de nombreux types de sources de données, notamment les bases de données relationnelles, les fichiers plats, les services web, les applications, les services de cloud computing et les réseaux sociaux.
2. **Intégration de données :** SSIS permet d'extraire, de transformer et de charger les données à partir de différentes sources de données. Les données peuvent être nettoyées, transformées et agrégées à l'aide de la fonctionnalité ETL.
3. **Stockage de données :** SSIS prend en charge plusieurs bases de données, notamment SQL Server, Oracle, MySQL, PostgreSQL et d'autres bases de données ODBC, pour stocker les données collectées et intégrées.
4. **Analyse de données :** SSIS permet aux utilisateurs de créer des rapports, des tableaux de bord et des analyses de données en utilisant les outils de

Business Intelligence (BI) de Microsoft, tels que Power BI et SQL Server Analysis Services.

5. **Sécurité :** SSIS fournit des fonctionnalités de sécurité pour protéger les données et les ressources. Les utilisateurs peuvent définir des rôles et des permissions pour les utilisateurs et les groupes.
6. **Déploiement :** SSIS peut être déployé sur site ou dans le cloud en utilisant Azure Data Factory, qui est une plateforme de traitement de données cloud de Microsoft [7].

D) **SAS (Statistical Analysis System) :** Est un logiciel de traitement de données et d'analyse statistique utilisé par les entreprises pour effectuer des tâches telles que la gestion de données, la modélisation statistique, l'analyse de données et la création de rapports. Voici comment fonctionne SAS :

1. **Importation de données :** La première étape consiste à importer les données dans SAS à partir de différentes sources, telles que des fichiers CSV, des fichiers Excel, des bases de données relationnelles ou des fichiers texte. Les données sont importées à l'aide de la procédure IMPORT de SAS.
2. **Nettoyage des données :** Une fois que les données sont importées, elles peuvent nécessiter un nettoyage pour éliminer les erreurs, les doublons ou les données manquantes. SAS dispose de nombreuses fonctions et procédures pour nettoyer les données, telles que la fonction de traitement des chaînes de caractères (TRIM), la fonction de suppression des doublons (SORT) et la procédure de traitement des données manquantes (MEANS).
3. **Transformation de données :** Les données peuvent nécessiter une transformation pour les préparer à l'analyse statistique. Les transformations peuvent inclure le tri, la fusion de données, la conversion de types de données, la création de variables dérivées et la normalisation de données. SAS dispose de nombreuses procédures pour effectuer des transformations de données,

telles que la procédure de tri (SORT), la procédure de fusion (MERGE) et la procédure de création de variables dérivées (DATA STEP).

4. **Analyse statistique :** Une fois que les données ont été nettoyées et transformées, elles peuvent être analysées à l'aide de techniques statistiques telles que la régression, l'analyse de variance, l'analyse de séries chronologiques, la modélisation prédictive et l'analyse de données textuelles. SAS dispose de nombreuses procédures pour effectuer des analyses statistiques, telles que la procédure de régression (REG), la procédure d'analyse de variance (ANOVA), la procédure d'analyse de séries chronologiques (ARIMA), la procédure de modélisation prédictive (HPFORECAST) et la procédure d'analyse de données textuelles (TEXT MINING).
5. **Création de rapports :** Une fois que l'analyse statistique est terminée, les résultats peuvent être présentés sous forme de tableaux, de graphiques et de rapports. SAS dispose de nombreuses procédures pour créer des rapports, telles que la procédure de création de tableaux (TABULATE), la procédure de création de graphiques (GPLOT) et la procédure de création de rapports (REPORT) [8].

E) **Hadoop :** Hadoop est un système de stockage et de traitement distribué de données open source conçu pour stocker et traiter de grandes quantités de données à travers de multiples machines en cluster. Il repose sur deux composants principaux : Hadoop Distributed File System (HDFS) et MapReduce. Il utilise HDFS pour stocker les données de manière fiable et distribuée, MapReduce pour traiter les données en parallèle, YARN pour gérer les ressources de traitement et des outils pour interroger les données stockées dans HDFS [9].

F) **Power BI :** Power BI est une plateforme de Business Intelligence de Microsoft, qui permet aux utilisateurs de collecter, analyser et visualiser des données provenant de différentes sources. Voici un aperçu du fonctionnement de Power BI :

1. **Collecte de données :** Power BI peut se connecter à différentes sources de données telles que des bases de données, des fichiers Excel, des fichiers

CSV, des services cloud (tels que Microsoft Dynamics 365, Salesforce, etc.), des applications Web et bien plus encore. Les utilisateurs peuvent également utiliser des connecteurs tiers pour connecter Power BI à des sources de données supplémentaires.

2. **Préparation des données :** Les utilisateurs peuvent nettoyer, transformer et enrichir les données à l'aide de Power Query, un outil d'extraction et de transformation de données inclus dans Power BI. Cela permet de préparer les données pour l'analyse ultérieure.
3. **Modélisation des données :** Les utilisateurs peuvent créer un modèle de données à l'aide de Power BI Desktop, qui permet de créer des relations entre différentes tables de données, d'ajouter des mesures et des colonnes calculées, et de créer des hiérarchies pour faciliter l'analyse.
4. **Création de visualisations :** Les utilisateurs peuvent créer des visualisations en faisant glisser des champs sur une page de rapport. Power BI offre une variété de visualisations, telles que des graphiques, des tableaux croisés dynamiques, des cartes et bien plus encore.
5. **Création de tableaux de bord :** Les utilisateurs peuvent créer des tableaux de bord à partir des visualisations créées et les partager avec d'autres utilisateurs de l'organisation.
6. **Partage des rapports :** Les utilisateurs peuvent publier des rapports et des tableaux de bord sur Power BI Service, la plate-forme en ligne de Power BI, et les partager avec des collègues de travail ou des clients.
7. **Collaboration :** Les utilisateurs peuvent collaborer sur les rapports et les tableaux de bord en permettant à d'autres utilisateurs de visualiser, d'éditer ou de commenter les rapports. Les utilisateurs peuvent également partager les rapports à l'aide de SharePoint ou d'autres applications Microsoft.

En somme, Power BI permet aux utilisateurs de collecter, analyser et visualiser des données de manière simple et intuitive, offrant ainsi une meilleure compréhension des données et une prise de décision plus éclairée [10].

3.4.3 Synthèse des outils logiciels

Outil	Réf	Principe	Avantages	Inconvénients
Talend	[3]	Plateforme d'intégration de données open source basée sur une architecture orientée services (SOA).	Solution open source, facilité d'utilisation, l'extensibilité et la gestion centralisée.	La courbe d'apprentissage, la performance, le support payant et la dépendance technique.
Pentaho	[5]	Pentaho une suite logicielle open-source de BI se compose de plusieurs modules	flexible et facile à utiliser, complémentarité	L'installation et la configuration peuvent être difficiles pour les utilisateurs débutants, ne pas être aussi puissant que d'autres outils de BI.
SSIS	[7]	plateforme ETL pour l'intégration de données. Utilise des flux de données pour déplacer des données entre les sources et les destinations.	Plate-forme ETL solide, il offre une haute performance, une personnalisation avancée et des fonctionnalités de surveillance.	La complexité, les coûts et les limites de la solution.

TABLE 3.3 – Synthèse pour les outils industriels

Outil	Réf	Principe	Avantages	Inconvénients
Hadoop	[9]	Plate-forme open-source hautement évolutive pour le stockage, le traitement et l'analyse de grands volumes de données distribuées.	Hadoop offre une flexibilité, une tolérance aux pannes et une réduction des coûts de stockage et de traitement de données par rapport à des solutions propriétaires.	La complexité, la latence et la nécessité de compétences techniques.
Power BI	[10]	La collecte, la transformation et la visualisation des données.	Connexions à diverses sources de données,Intégration avec d'autres outils Microsoft, Facilité d'utilisation.	Coût, limitations de données, Dépendance aux outils Microsoft.

TABLE 3.4 – Synthèse pour les outils industriels

3.5 Conclusion

Dans ce chapitre, nous avons fait un examen, plus en moins, appronfondis des différents travaux qui ont traité le problème de l'intégration des techniques de FDD dans une solution BI. En plus, nous avons fait ressortir les insuffisances des travaux existants et les limites des logiciels du marché.
En partant des insuffisances constatées, nous proposons dans le prochain chapitre notre contribution pour remédier à ces limites et nous proposons une solution au problème relevé en début du chapitre.

Deuxième partie

Contribution

CHAPITRE 4
CONCEPTION DE L'APPROCHE

4.1 Introduction

Dans le chapitre précédent nous avons exposé notre probématique et nous l'avons positionnée par rapport aux travaux connexes. Dans ce chapitre nous apportons notre contribution pour l'intégration des techniques de fouille de données dans une perspective de Business Intelligence. Notre objectif est de construire un système qui permet de réaliser les activités suivantes :

— Obtenir des informations à partir de diverses sources de données.

— Améliorer la prise de décision par les exploitation des données collectées.

— Accroître l'efficacité opérationnelle de l'organisation.

Nous commençons le chapitre par la présentation du fondement de notre approche.

4.2 Fondement de l'approche

Comme il a été montré dans l'analyse de l'état de l'art sur les travaux relatifs à l'utilisation des techniques de fouille de données dans le contexte d'une solution BI,

chaque solution prend en compte **une perspective (axe) spécifique** pour l'exploitation des données de l'EDD.
En effet, plusieurs travaux font usage des techniques statistiques, tels que le calcul de variation, la recherche de la corrélation entre les attributs. D'autres travaux exploitent les techniques de l'IA, telles que : la classification, le clustering, la prédiction pour analyser les données. Aussi, il a été constaté que le déploiement des outils OLAP pour l'exploration des données de l'EDD permettent d'extraire certains patterns d'intérêt. D'autre travaux utilisent les techniques d'ADD, telles que l'analyse factorielle des correspondances AFC, AFD, ACP et les techniques de régression, les domaines d'application sont très variés touchent à la détection des risques et des fraudes, interaction avec les clients, planification urbaine, soins de santé, etc.
Malgré leurs avantages, chacune des approches précédentes demeures restreinte et ne tolère l'utilisation que d'une seule technique à la fois.
Notre contribution consiste, justement à offrir aux utilisateurs du système BI la possibilité de choisir la technique la plus appropriée à son contexte et même de faire usage de plusieurs techniques à la fois pour un même ensemble de données.
En effet, nous visons à construire un système personnalisable qui offre un ensemble de techniques de fouille de données et l'utilisateur qui pourra sélectionner la technique la plus adéquate en fonction de ses besoins.
La section suivante montre l'architecture globale de notre solution.

4.3 Architecture du système

Pour la conception de notre solution, nous avons élaborer un système dénommé **DMinBI**, dont l'architecture est exposée dans la figure 4.1 suivante.

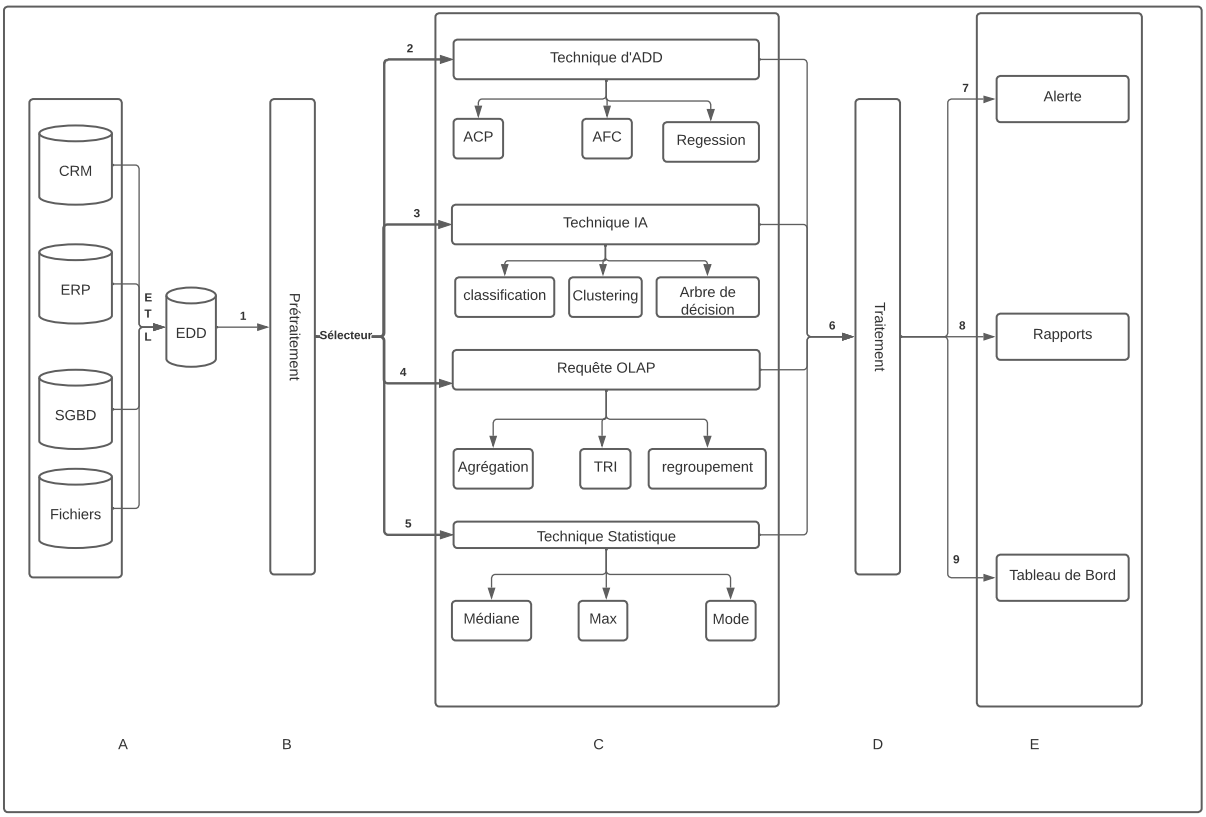


FIGURE 4.1 – Architecture du système proposé

Ci-après, nous décrivons les composants de l'architecture proposée et les interactions entre les composants.

4.3.1 Les composants du système DMinBI

Comme il est observé dans la figure 4.1, DMinBI est composé des quatre modules principaux suivants :

A) Comme la thématique d'intégration des données a été déjà réalisée dans les projets de fin d'études 2021 et 2022 [15][14]. Par conséquent, nous considérons que l'EDD est déjà construit et nous nous focalisons sur les étapes suivantes du système :

Entrepôt de données (EDD) : l'EDD construit à l'aide de techniques ETL (Extract, Transform, Load) pour extraire, transformer et charger les données à partir de différentes sources, telles que des bases de données opérationnelles, des fichiers plats, des applications, des systèmes de suivi de transactions, des réseaux sociaux dans un format cohérent. Les données sont souvent nettoyées et transformées avant d'être chargées dans l'entrepôt de données pour assurer la qualité et la cohérence des données.

B) **Pré-traitement :** Les données extraites peuvent nécessiter un nettoyage, une normalisation, une agrégation ou d'autres transformations pour les préparer à l'analyse. Après on peut appliquer les différents techniques et les algorithmes de FD. À titre d'exemple, cette phase procède à la suppression des données redondantes et à l'élimination des données manquantes.

C) Après avoir nettoyé les données (Module B), dans le module C les techniques de fouille de données peuvent être appliquées pour découvrir des modèles, des tendances et des relations cachées dans les données. Le système offre plusieurs techniques de FDD notamment l'ADD, IA, BDD et les techniques statistiques, et l'utilisateur aura la possibilité de choisir une technique parmi celles qui sont offertes.

D) **Traitement :** Une fois l'utilisateur a fait son choix, nous pouvons procéder au traitement sur les données et appliquer la ou les techniques de fouille de données appropriées.

E) **L**e module E (dernière phase) qui permet de représenter le résultat obtenu après l'utilisation des techniques de fouille de données . Les résultats proposés par la solution BI peuvent varier en fonction des objectifs de l'entreprise et des besoins des utilisateurs, mais voici quelques exemples de résultats courants : es rapports, des tableaux de bord aussi que des alertes.

Dans ce qui suit, on va présenter les définitions de chaque technique et on va commencer par les techniques d'analyse de données. Dans cette catégorie, nous allons

s'intéresser aux techniques d'analyse factorielle des correspondances (AFC) et l'analyse en composantes principales (ACP).

a) **Analyse factorielle des correspondances (AFC) :** L'approche statistique de l'analyse des données connue sous le nom d'analyse factorielle des correspondances (AFC) lit les données contenues dans un espace multidimensionnel en abaissant la dimension de cet espace tout en conservant un maximum de la connaissance présente dans le point de départ. Son but est de traiter les données à partir d'un tableau d'informations appelé contingence ou dépendance qui est défini par des facteurs qualitatifs et mis en relation selon une méthode naturelle ou expérimentale plus ou moins connue. Contrairement à l'ACP, elle permet de fournir un espace de représentation commun aux variables et aux individus. En d'autres termes, elle sert à identifier et à hiérarchiser l'ensemble des relations entre les lignes et les colonnes du tableau. En outre, elle est utilisée pour examiner la corrélation entre deux variables qualitatives [26].

b) **Analyse en Composantes Principales (ACP) :** L'ACP également connue sous le nom de PCA (Principal Component Analysis) en anglais, est une technique largement utilisée dans différents domaines tels que la statistique, l'apprentissage automatique, l'analyse des données et la reconnaissance de formes. Elle permet de transformer un grand nombre de variables en un nombre plus restreint de variables, appelées composantes principales, tout en préservant au maximum l'information contenue dans les données d'origine. L'objectif principal de l'ACP est de trouver des combinaisons linéaires des variables d'origine qui captent le plus de variance dans les données. Les premières composantes principales sont celles qui expliquent la plus grande partie de la variance totale, tandis que les composantes suivantes expliquent progressivement moins de variance. Les composantes principales sont orthogonales entre elles, ce qui signifie qu'elles sont indépendantes les unes des autres.

Maintenant, on va présenter les trois techniques de l'IA que nous avons intégré dans notre approche.

a) **K-means :** K-means, traduit littéralement par "k-moyennes", est un algorithme de regroupement non supervisé utilisé pour partitionner un ensemble de données en k clusters distincts. L'objectif de l'algorithme k-means est de minimiser la variance intra-cluster, c'est-à-dire de regrouper les points de données similaires dans le même cluster tout en maintenant une séparation maximale entre les clusters.

   L'algorithme k-means fonctionne de la manière suivante :

   — Définir le nombre k de clusters et initialiser aléatoirement les centres de ces clusters.

   — Attribuer chaque point de données au cluster dont le centre est le plus proche, en utilisant une mesure de distance telle que la distance euclidienne.

   — Mettre à jour les centres des clusters en calculant la moyenne des points de données qui y sont assignés.

   — Mettre à jour les centres des clusters en calculant la moyenne des points de données qui y sont assignés.

   À la fin de l'algorithme, les points de données seront répartis en k clusters, où chaque cluster est caractérisé par son centre, également appelé centroid.

b) **Agglomératif :** L'agglomératif, dans le contexte de l'apprentissage automatique, fait référence à une approche de regroupement (clustering) des données. L'algorithme de regroupement agglomératif est utilisé pour former des groupes (clusters) en regroupant itérativement des points de données similaires. Plus précisément, l'algorithme agglomératif commence par considérer chaque point de données comme un cluster individuel. Ensuite, il fusionne les clusters les plus similaires en un seul cluster à chaque étape, jusqu'à ce qu'un seul cluster global soit formé. La mesure de similarité utilisée peut être basée sur des critères tels que la distance euclidienne, la similarité cosinus, etc.

   L'algorithme agglomératif utilise une matrice de similarité ou une matrice de

distance pour prendre des décisions de fusion. Il utilise une approche ascendante (bottom-up), où les points de données individuels sont fusionnés pour former des clusters plus grands. Cela permet de créer une structure hiérarchique de clusters, où chaque étape de fusion est enregistrée. L'un des avantages de l'algorithme agglomératif est sa simplicité conceptuelle et sa capacité à gérer des ensembles de données de grande taille. Cependant, il peut être plus coûteux en termes de temps de calcul que d'autres méthodes de regroupement, notamment lorsque le nombre de points de données est élevé [26].

c) **Minibatch K-means :** C'est une variante de l'algorithme K-means qui vise à accélérer le processus de regroupement en utilisant des mini-lots d'échantillons au lieu d'utiliser l'ensemble complet de données à chaque itération. L'algorithme Minibatch K-means fonctionne de manière similaire à l'algorithme K-means, mais au lieu de mettre à jour les centres de clusters à chaque itération en utilisant tous les points de données, il utilise un sous-ensemble aléatoire (mini-lot) des données pour effectuer les mises à jour. Cela permet de réduire considérablement le coût de calcul, ce qui peut être bénéfique lorsque l'ensemble de données est volumineux.

Les étapes de l'algorithme Minibatch K-means sont les suivantes :

— Définir le nombre k de clusters et initialiser aléatoirement les centres de ces clusters.

— Sélectionner un sous-ensemble aléatoire (mini-lot) de points de données à partir de l'ensemble complet.

— Attribuer chaque point du mini-lot au cluster dont le centre est le plus proche, en utilisant une mesure de distance telle que la distance euclidienne.

— Mettre à jour partiellement les centres des clusters en calculant la moyenne des points de données assignés au mini-lot.

— Répéter les étapes 2 à 4 pour un certain nombre d'itérations ou jusqu'à ce qu'un certain critère de convergence soit atteint.

Enfin, nous exposons les différentes techniques OLAP intégrées dans notre approche. Mais, avant de commencer, nous rappelons la définition et l'intérêt des techniques OLAP.

a) **OLAP :** (Online Analytical Processing) est une méthode de traitement et d'analyse de données multidimensionnelles qui permet aux utilisateurs d'explorer et de comprendre les données de manière interactive. Contrairement aux bases de données relationnelles, qui sont organisées en tables à deux dimensions, les données OLAP sont organisées en dimensions multiples, souvent appelées "cubes". Chaque dimension représente un aspect différent des données, tel que la géographie, le temps, les produits ou les clients. Les utilisateurs peuvent sélectionner et combiner des dimensions pour obtenir des vues personnalisées des données. Les cubes OLAP permettent également aux utilisateurs de réaliser des agrégations de données complexes et de générer des rapports analytiques interactifs. Les utilisateurs peuvent explorer les données en effectuant des sélections, des filtrages et des regroupements pour obtenir des résultats en temps réel. Les données peuvent être visualisées sous forme de graphiques, de tableaux croisés, de listes déroulantes, de diagrammes de Gant et d'autres types de visualisations.

Les sorties sont :

a) **Alerte :** Une alerte est une notification automatique qui est déclenchée lorsqu'un événement prédéfini se produit dans les données, et un outil clé dans une solution BI, permettant aux utilisateurs de surveiller les données en temps réel et de réagir rapidement aux événements importants. Les alertes sont généralement définies par les utilisateurs et peuvent être personnalisées en fonction de leurs besoins et de leurs préférences. Par exemple, une entreprise peut définir une alerte pour être notifiée lorsque les ventes d'un produit dépassent un certain seuil, lorsque les stocks d'un produit sont faibles, ou lorsque les indicateurs de performance clés (KPI) ne répondent pas aux objectifs de l'entreprise.

b) **Rapports :** Les rapports et les tableaux de bord sont des éléments clés d'une solution BI. Ils permettent aux utilisateurs de visualiser les données de manière

claire et concise, en utilisant des graphiques, des tableaux et des indicateurs de performance clés (KPI). Les rapports peuvent être générés automatiquement à partir des données de l'entreprise ou personnalisés en fonction des besoins de l'utilisateur.

c) **Tableau de bord(dashboard) :** Est une vue synthétique et visuelle des données clés de l'entreprise. Un tableau de bord permet de présenter les indicateurs de performance clés (KPI) et les métriques en temps réel, sous forme de graphiques, de tableaux, de cartes ou d'autres types de visualisations. Les tableau de bord sont personnalisables et peuvent être adaptés aux besoins de différents utilisateurs et départements. Par exemple, un dashboard peut être conçu pour les ventes, un autre pour la gestion de la chaîne d'approvisionnement, et un troisième pour les finances. Chaque dashboard est conçu pour fournir des informations précises et pertinentes pour un objectif spécifique.

Maintenant, nous abordons les interactions entre les composants précédents.

4.3.2 Les interactions entre les composants

Les interactions entre les composants de l'architecture proposée sont indiquées par des flèches numérotées, comme le montre la figure4.1.
Nous expliquons ci-dessous la signification de chaque interaction.

1) Lancement du module de pré-traitement sur les données collectées dans l'EDD.
2) Activer le sélecteur qui permettra à l'utilisateur de choisir la technique ADD.
3) Activer le sélecteur qui permettra à l'utilisateur de choisir la technique IA.
4) Activer le sélecteur qui permettra à l'utilisateur de choisir les requêtes OLAP.
5) Activer le sélecteur qui permettra à l'utilisateur de choisir la technique statistique.
6) Lancement du programme adéquat et répondant au choix spécifié par l'utilisateur dans la phase précédente.

7) Récupérer les résultats du programme de traitement et choisir comme sortie une visualisation sous forme d'alerte.

8) Récupérer les résultats du programme de traitement et choisir comme sortie une visualisation sous forme de rapport.

9) Récupérer les résultats du programme de traitement et choisir comme sortie une visualisation sous forme de tableau de bord.

4.4 Fonctionnement de la solution

L'exploitation de l'architecture proposée est un processus incrémental, qui en partant de l'EDD, permet d'aboutir à des résultats présentés en sortie du système.
En effet, les étapes de l'exploitation de la solution sont les suivantes :

— Une fois l'EDD est construit l'utilisateur commence par déclencher le processus de pré-traitement qui permet de nettoyer les données stockées. A la fin de cette opération les données épurées sont prêtes pour les actions de FD.

— Sélection de la technique de FD parmi celle proposée(techniques d'analyse de données, intelligence artificielle, requête OLAP, et statistique), et chaque de ces techniques elle a une méthode qui on va l'appliquer

— Après que l'utilisateur a fait son choix, nous pouvons lancer le traitement adéquats sur les données.

— La dernière phase est le résultat qui est une vue globale et structurée des données de l'entreprise, qui permet aux utilisateurs de prendre des décisions plus éclairées et plus rapides. Les résultats d'une solution BI peuvent varier en fonction des objectifs de l'entreprise et des besoins des utilisateurs, mais voici quelques exemples de résultats courants : Des rapports, des tableaux de bord et des alertes.

4.5 Scénario illustratif de l'utilisation de DMinBI

Dans cette section, nous allons montrer la faisabilité de notre approche avec un scénario réel.

Notre exemple est relatif a une entreprise commerciale avec différentes sources de données est une chaîne de supermarchés. Les supermarchés recueillent des données sur les ventes, les promotions, les stocks, les prix, les fournisseurs, etc. Ils peuvent également collecter des données sur les habitudes d'achat de leurs clients grâce aux programmes de fidélité.

Le système d'information de l'entreprise qui enregistre les ventes en magasin, en ligne et client mobile. Les données sont collectées à partir de trois sources principales :

a) **Vente en magasin :** L'entreprise dispose d'un système de point de vente dans ses magasins physiques où les clients peuvent acheter des produits. Ces données incluent les informations sur les produits achetés, le prix, le magasin, la date, etc.

b) **Vente en ligne :** L'entreprise dispose également d'un site de commerce électronique où les clients peuvent acheter des produits en ligne. Ces données incluent les informations sur les produits achetés, le prix, la date, l'emplacement d'expédition, etc.

c) **Client mobile :** L'entreprise a une application mobile pour permettre aux clients de naviguer, acheter et interagir avec la marque. Ces données incluent les interactions clients telles que la navigation, la recherche de produits, l'ajout au panier, etc.

Notre objectif est d'utiliser ces données pour améliorer notre offre de produits et nos stratégies marketing. On va appliqué les différents techniques de fouille de données (techniques statistiques, BDD, IA et les techniques d'ADD) pour améliorer le fonctionnement et le rendement de l'entreprise.

Supposons que cette entreprise souhaite classer ses clients en fonction de leur disposition à acheter des produits de catégorie A, B ou C. Admettons que les sources de données de l'entreprise offrent des informations supplémentaires qui sont :

— L'âge du client.

— Le revenu du client.

— Le nombre de fois où le client a acheté des produits de catégorie A, B ou C.

— Le montant total dépensé par le client sur l'ensemble des produits.

L'entreprise souhaite utiliser ces données pour classer chaque client en fonction de sa capacité à acheter des produits de catégorie A, B ou C.
Pour répondre à ce besoin, on peut aussi utiliser la méthode de classification basée sur l'analyse discriminante. Cette méthode consiste à trouver une fonction qui permet de distinguer les clients qui ont tendance à acheter des produits de catégorie A de ceux qui ont tendance à acheter des produits de catégorie B ou C.
Pour répondre à ce besoin, on peut utiliser les techniques statistiques suivantes :

- ☐ La moyenne et la variance : pour chaque variable (par exemple, l'âge ou le revenu), la moyenne et la variance sont calculées pour chaque classe (catégorie A, B ou C).
- ☐ La matrice de covariance : cette matrice représente les corrélations entre les différentes variables pour chaque classe.
- ☐ La fonction discriminante : cette fonction est utilisée pour prédire la probabilité qu'un client appartienne à chaque classe en fonction de ses caractéristiques. La fonction discriminante est généralement construite à partir de la matrice de covariance et des moyennes et variances pour chaque classe.

En général, l'analyse discriminante implique des calculs de matrices et de vecteurs, ainsi que des estimations de paramètres statistiques tels que les moyennes et les variances. Les logiciels de statistiques comme R, Python ou SAS ont des fonctions intégrées pour effectuer des analyses discriminantes et d'autres techniques de classification.
Enfin, nous pouvons faire recours à des requêtes OLAP.
Voici une requête de base de données pour extraire les ventes de la semaine dernière par catégorie de produit :

```
SELECT Category, SUM(Quantity), SUM(Sales)
FROM Sales
WHERE Date BETWEEN '2023-04-17' AND '2023-04-23'
GROUP BY Category
ORDER BY SUM(Sales) DESC;
```

4.6 Conclusion

Dans ce chapitre, nous avons proposé une approche cohérente qui intègre plusieurs techniques de FDD pour consolider une solution BI standard.
L'architecture et le fonctionnement de la solution proposée ont été largement expliqués et illustrés par des scénarios réels.
Dans le prochain chapitre, nous allons aborder l'implémentation et l'expérimentation de notre approche.

CHAPITRE 5

IMPLÉMENTATION ET EXPÉRIMENTATION

5.1 Introduction

Dans ce chapitre, nous allons mettre en oeuvre l'approche proposée pour l'intégration des techniques de fouille de données dans système BI.
Nous commençons par présenter les outils logiciels utilisés par la réalisation de notre système, puis nous exposons un aperçu sur les données expérimentale exploitées pour tester notre approche. Après, les différentes fonctionnalités du système sont illustrées et expliquées. Enfin, nous terminerons le chapitre par une conclusion.

5.2 Les environnements et outils logiciels utilisés

Pour l'implémention de notre approche, nous avons utilisé les outils logiciels suivants.

5.2.1 Python

Python est un langage de programmation de haut niveau, interprété et polyvalent. Il a été créé par Guido van Rossum et publié pour la première fois en 1991. Il met

l'accent sur la lisibilité du code et la simplicité, visant à fournir une syntaxe claire et concise qui permet aux programmeurs d'exprimer leurs idées avec moins de lignes de code que dans d'autres langages de programmation. Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plate-forme [11].
L'une des caractéristiques clés de Python est son utilisation de l'indentation pour définir les blocs de code, plutôt que de se reposer sur des accolades ou des mots-clés. Cette approche unique permet de maintenir un style de codage cohérent et visuellement attrayant. Python prend en charge plusieurs paradigmes de programmation, notamment la programmation procédurale, orientée objet et fonctionnelle.
Python dispose d'une vaste bibliothèque standard qui fournit de nombreux modules et fonctions, offrant des solutions pour différentes tâches et permettant aux développeurs d'écrire du code de manière plus efficace. De plus, il existe un vaste écosystème de bibliothèques et de frameworks tiers disponibles, ce qui rend Python adapté à différents domaines tels que le développement web, le calcul scientifique, l'analyse de données, l'intelligence artificielle, l'apprentissage automatique, et bien plus encore.
Le langage a gagné en popularité en raison de sa simplicité, de sa polyvalence et de sa communauté solide qui contribue à son développement et à sa maintenance. La facilité d'utilisation de Python, combinée à ses nombreuses bibliothèques et frameworks, en fait un choix préféré tant pour les débutants que pour les développeurs expérimentés [11].

5.2.2 Spyder

Spyder est un logiciel open-source et multiplateforme spécifiquement conçu pour les développeurs Python. Il fournit un environnement de développement intégré (IDE) puissant et convivial pour écrire, tester et déboguer du code Python.

5.2.3 Sklearn

Scikit-learn (également connu sous le nom de sklearn) est une bibliothèque open-source très populaire pour l'apprentissage automatique (machine learning) en Python. Elle est conçue pour offrir des outils simples et efficaces pour effectuer des tâches courantes d'apprentissage automatique, telles que la classification, la régression, le clustering et la sélection de modèles. Scikit-learn est construit sur les bibliothèques NumPy [12], SciPy et matplotlib, ce qui lui permet de tirer parti de leurs fonctionnalités pour le traitement numérique, l'optimisation, les opérations matricielle et la visualisation de données [13].

5.3 Présentation des données d'expérimentation

Pour collecter les données d'expérimentation, nous avons contacter des entreprises au niveau de la wilaya de Guelma, telles que : CAB Benamor, supérette OASIS et Sonelgaz. Mais, ces entreprises ont donnée des réponses négatives.
Vu l'absence de données réelles spécifiques à une organisation au niveau de la wilaya de Guelma, nous avons opté pour les données stockées sur le web et qui sont relatives à la gestion des Prix des logements en Californie (Les prix moyens des maisons pour les districts de Californie dérivés du recensement de 1990), les données contiennent des informations du recensement de la Californie de 1990. Donc, bien qu'il ne vous aide peut-être pas à prédire les prix actuels des logements comme le Zillow Zestimate, il fournit un ensemble de données d'introduction accessible pour enseigner aux gens les bases de l'apprentissage automatique. Pour l'expérimentation de notre approche, nous avons utilisé le jeu de données de test dont les caractéristiques sont les suivantes.

— La taille : 295 ko.

— Nombre d'enregistrement : 3000.

— Les attributs manipulé sont : longitude, latitude, housing median age, total rooms, totalbedrooms, population, households, median income, median house value.

5.4 Enchaînement général de l'application

Le prototype DMinBI développé n'est accessible qu'après la phase d'authentification.

Mire d'accueil de l'application.

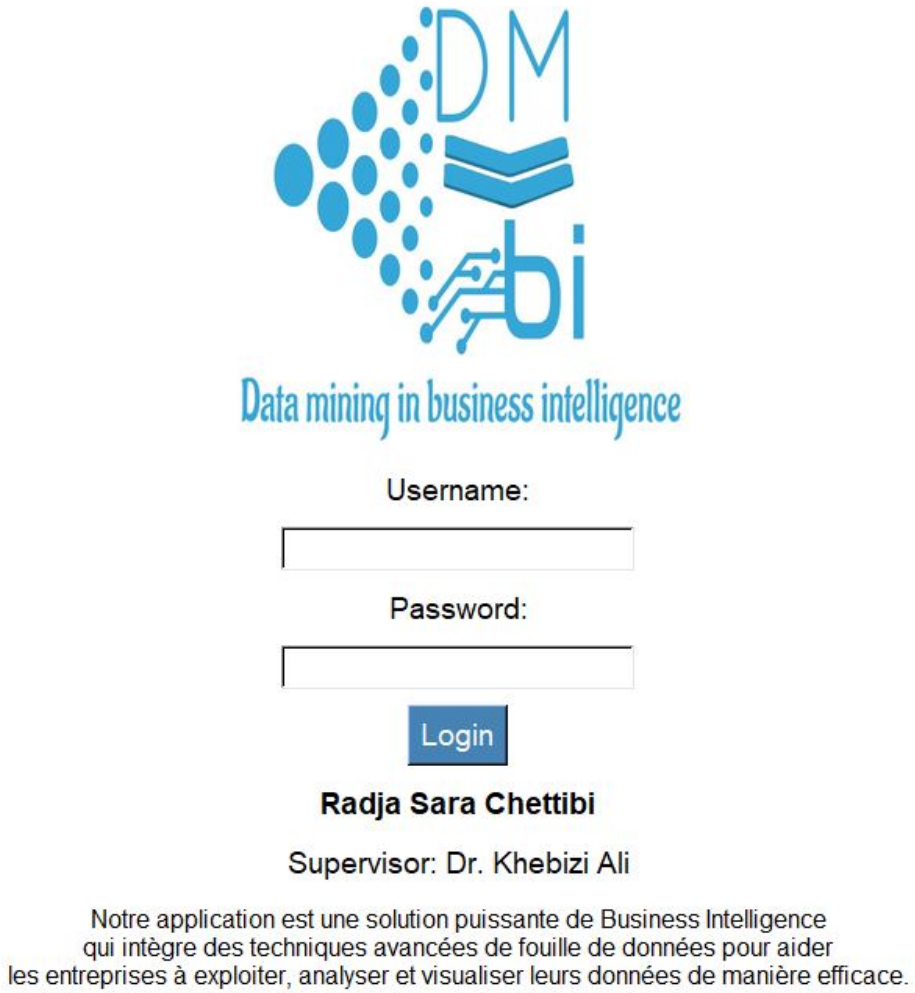


FIGURE 5.1 – Authentification pour l'ouverture de session

Après avoir réussi cette étape, l'utilisateur peut commencer l'exploitation du système par le choix de l'une des 6 options offertes, à savoir :

- Load data, permet de gérer les données.

- La deuxième option est la visualisation, qui nous fait afficher notre jeu de données.
- Le bouton de nettoyage, assure la gestion des valeurs nulles (NULL) de notre jeu de données.
- Maintenant 4 ème option, assure les techniques de l'IA (clustering, K-means...).
- La cinquième option qui représente les techniques ADD.
- La dernière option garantit un ensemble de techniques statiques.

Comme il est illustré dans la figure suivante.

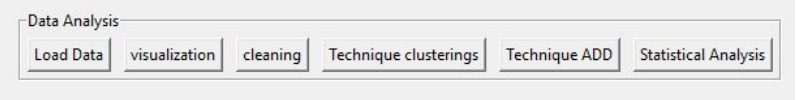


FIGURE 5.2 – Les six options offertes

Ces options sont détaillées ci dessous.

5.4.1 Gestion des données (Load data)

Cette fonctionnalité assure la sélection et chargement de données à partir d'un emplacement sur le disque dur, tel que illustré dans la figure 5.3.

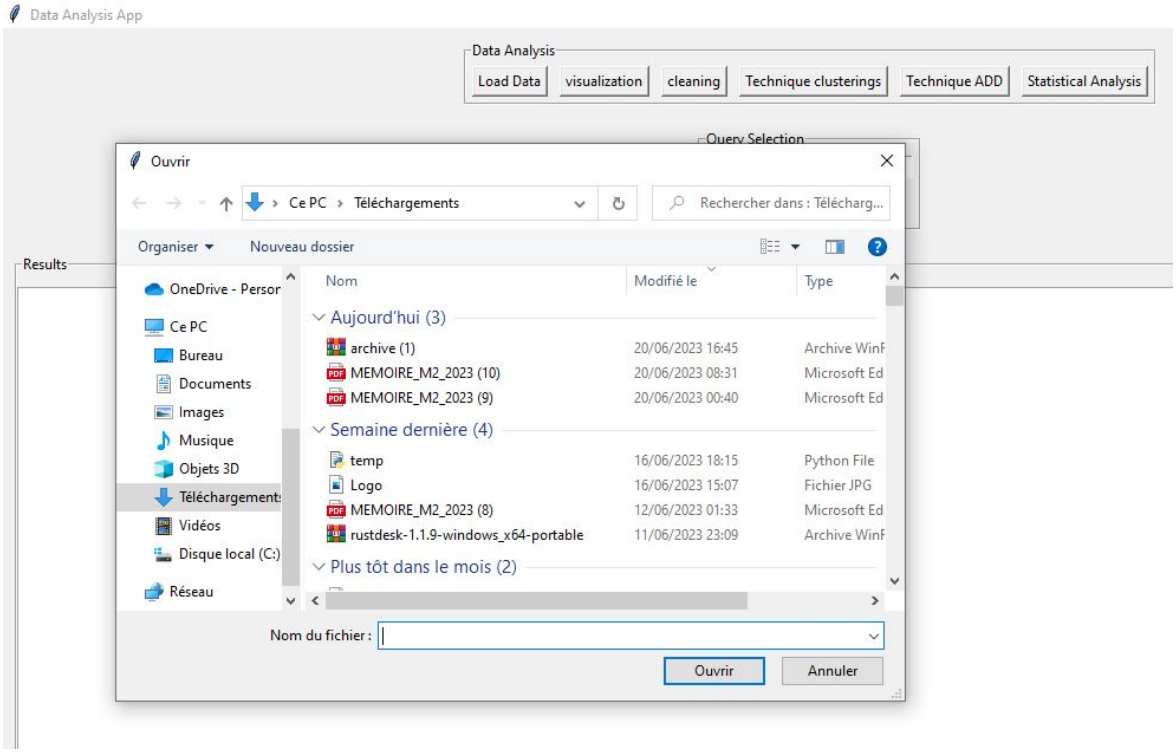


FIGURE 5.3 – Chargement le jeu de données

5.4.2 Visualisation

La fonctionnalité de la visualisation, nous permettons d'afficher notre jeu de données. tel que illustré dans la figure 5.4.

Data Cleaning Results

your data:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
0	-122.05	37.37	27.0	3885.0	661.0	1537.0	606.0	6.6085	344700.0
1	-118.30	34.26	43.0	1510.0	310.0	809.0	277.0	3.5990	176500.0
2	-117.81	33.78	27.0	3589.0	507.0	1484.0	495.0	5.7934	270500.0
3	-118.36	33.82	28.0	67.0	15.0	49.0	11.0	6.1359	330000.0
4	-119.67	36.33	19.0	1241.0	244.0	850.0	237.0	2.9375	81700.0
5	-119.56	36.51	37.0	1018.0	213.0	663.0	204.0	1.6635	67000.0
6	-121.43	38.63	43.0	1009.0	225.0	604.0	218.0	1.6641	67000.0
7	-120.65	35.48	19.0	2310.0	471.0	1341.0	441.0	3.2250	166900.0
8	-122.84	38.40	15.0	3080.0	617.0	1446.0	599.0	3.6696	194400.0
9	-118.02	34.08	31.0	2402.0	632.0	2830.0	603.0	2.3333	164200.0
10	-118.24	33.98	45.0	972.0	249.0	1288.0	261.0	2.2054	125000.0
11	-119.12	35.85	37.0	736.0	166.0	564.0	138.0	2.4167	58300.0
12	-121.93	37.25	36.0	1089.0	182.0	535.0	170.0	4.6900	252600.0
13	-117.03	32.97	16.0	3936.0	694.0	1935.0	659.0	4.5625	231200.0
14	-117.97	33.73	27.0	2097.0	325.0	1217.0	331.0	5.7121	222500.0
15	-117.99	33.81	42.0	161.0	40.0	157.0	50.0	2.2000	153100.0
16	-120.81	37.53	15.0	570.0	123.0	189.0	107.0	1.8750	181300.0
17	-121.20	38.69	26.0	3077.0	607.0	1603.0	595.0	2.7174	137500.0
18	-118.88	34.21	26.0	1590.0	196.0	654.0	199.0	6.5851	300000.0
19	-122.59	38.01	35.0	8814.0	1307.0	3450.0	1258.0	6.1724	414300.0
20	-122.15	37.75	40.0	1445.0	256.0	849.0	255.0	3.8913	126300.0
21	-121.37	38.68	36.0	1775.0	296.0	937.0	305.0	3.1786	83400.0
22	-118.16	34.07	47.0	2994.0	543.0	1651.0	561.0	3.8644	241500.0
23	-122.20	37.79	45.0	2021.0	528.0	1410.0	480.0	2.7788	115400.0
24	-117.28	33.28	13.0	6131.0	1040.0	4049.0	940.0	3.8156	150700.0
25	-118.03	34.16	36.0	1401.0	218.0	667.0	225.0	7.1615	484700.0
26	-122.42	37.76	52.0	3587.0	1030.0	2259.0	979.0	2.5403	250000.0
27	-118.39	33.99	32.0	2612.0	418.0	1030.0	402.0	6.6030	369200.0
28	-118.45	34.07	19.0	4845.0	1609.0	3751.0	1539.0	1.5830	350000.0
29	-118.48	34.01	30.0	3078.0	954.0	1561.0	901.0	3.4852	425000.0
30	-119.35	36.33	14.0	1195.0	220.0	568.0	229.0	3.1486	105600.0
31	-118.30	33.91	34.0	1617.0	493.0	1530.0	500.0	2.6182	172600.0
32	-121.13	39.31	17.0	3442.0	705.0	1693.0	619.0	2.8102	128900.0
33	-118.08	34.55	5.0	16181.0	2971.0	8152.0	2651.0	4.5237	141800.0
34	-118.32	33.94	38.0	1067.0	170.0	499.0	169.0	4.6389	183800.0
35	-118.11	34.00	33.0	2886.0	726.0	2650.0	728.0	2.6250	178700.0
36	-122.53	37.97	52.0	1560.0	451.0	700.0	419.0	2.5125	270800.0
37	-118.02	33.92	34.0	1478.0	251.0	956.0	277.0	5.5238	185300.0
38	-118.05	33.93	31.0	894.0	203.0	883.0	190.0	3.6771	141500.0
39	-119.01	34.23	11.0	5785.0	1035.0	2760.0	985.0	4.6930	232200.0

FIGURE 5.4 – Visualisation de jeu de données

5.4.3 Nettoyage de données

Cette option assure la gestion des valeurs nulles (NULL) de notre jeu de données.

5.4.4 Les techniques de l'intelligence artificielle

La figure 5.5, montre le menu de techniques d'IA.

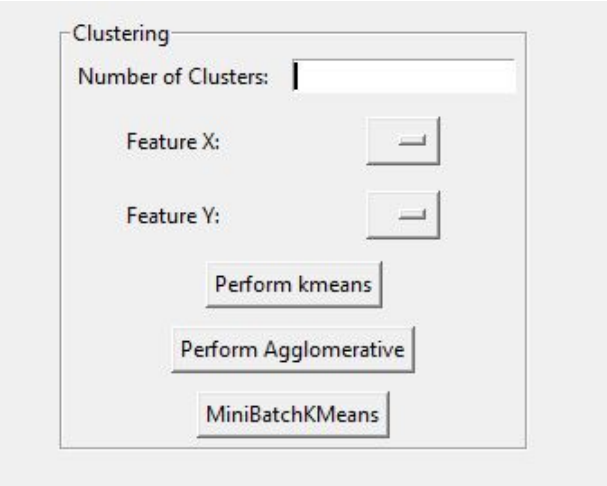


FIGURE 5.5 – Les techniques de IA

Lorsque nous cliquons dessus une autre fenêtre apparaît. Dans cette fenêtre nous pouvons choisir le nombre K de clusters que nous voulons faire et les 2 colonnes que nous utilisons pour dessiner le nuage de points et aussi les techniques (kmeans,miniBatch..). La figure 5.6 représente le résultat de Kmeans appliqués avec K=3 et total rooms et population comme x, y, donc en haut on voit les 3 centres des clusters et en bas le nuage de points.

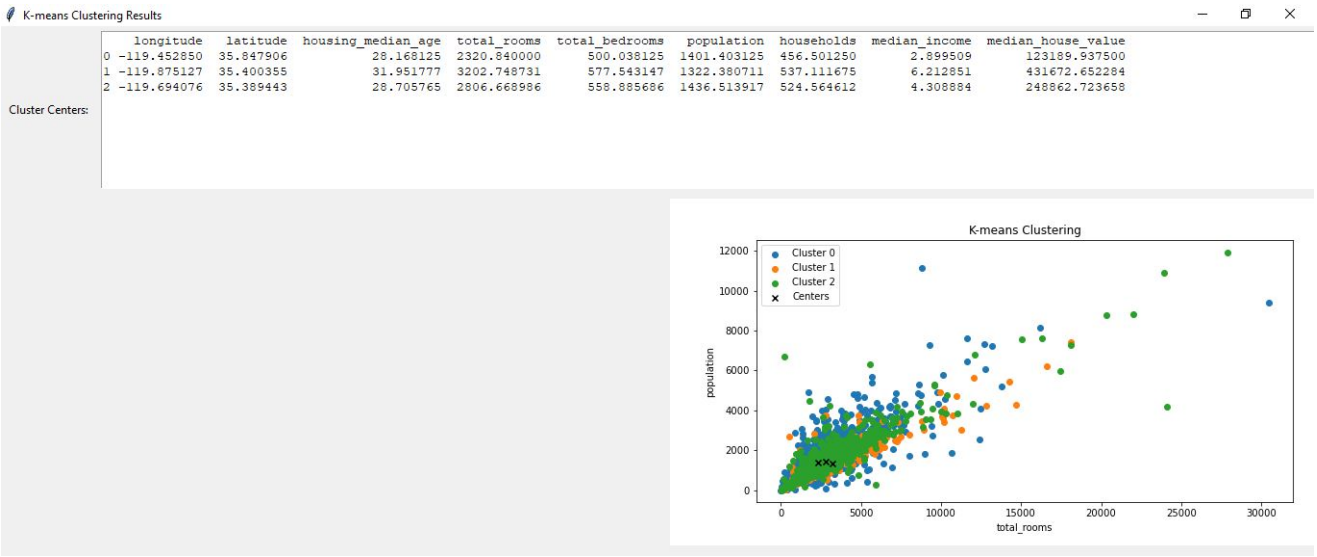


FIGURE 5.6 – Le résultat de Kmeans

5.4.5 Les techniques d'analyse de données

Nous pouvons voir lorsque nous cliquons dessus une boîte de message est apparue, la figure 5.7 montre le menu de technique ADD, contient deux méthodes ADD la première est ACP et la deuxième est AFC.

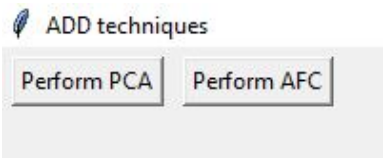


FIGURE 5.7 – Menu des techniques ADD offertes par DMinBI

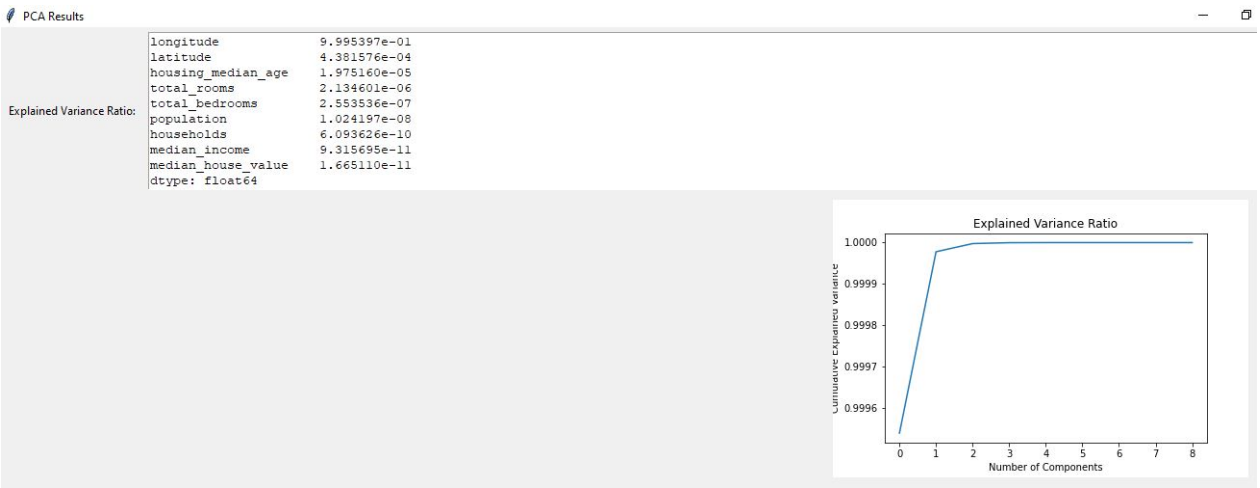


FIGURE 5.8 – Résultat de l'ACP

La figure 5.8 affiche la variance de toutes les colonnes de notre jeu de données. En ACP, la variance de chaque composante principale est une mesure importante de sa signification ou de sa contribution à la variation globale des données. Les composantes principales sont des combinaisons linéaires des variables d'origine, et elles sont classées selon la quantité de variance qu'elles expliquent. La première composante principale explique la plus grande quantité de variance, suivie de la deuxième composante principale, et ainsi de suite.

5.4.6 Les techniques statistiques

La dernière option garantit un ensemble de techniques statiques, la figure 5.9 montre le résultat.

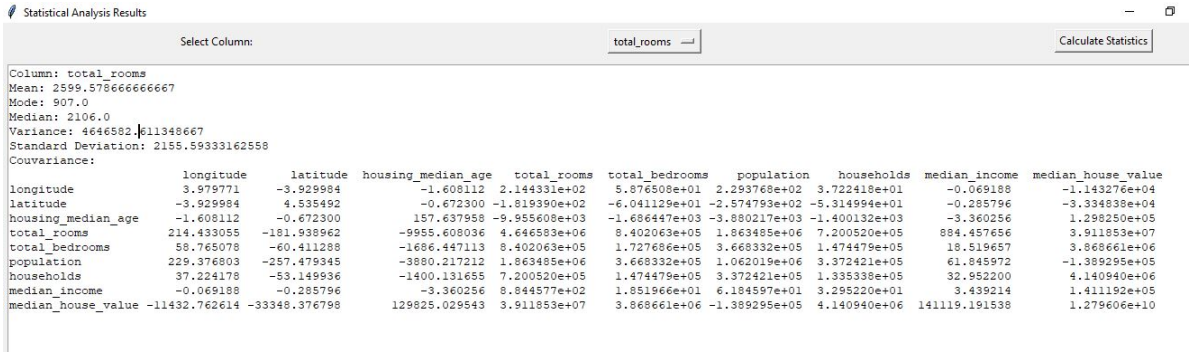
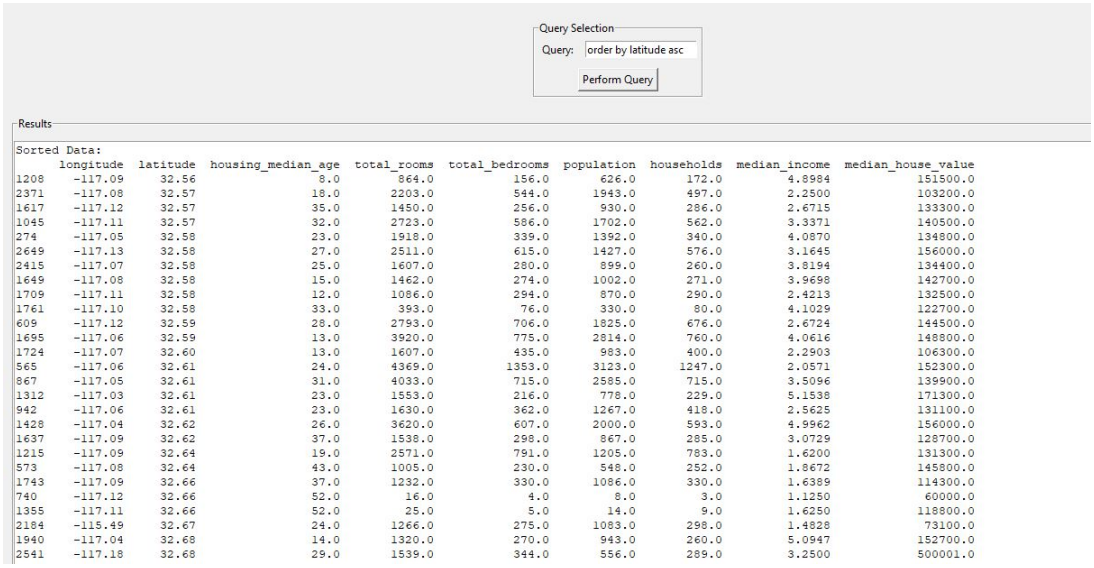


FIGURE 5.9 – Résultat de techniques statistiques

En plus, nous pouvons voir que nous pouvons calculer la moyenne, le mode, la médiane..etc, pour chaque colonne de notre jeu de données.

5.4.7 Les requêtes OLAP

La dernière option se trouve au bas de la première interface, elle aide à écrire une requête pour obtenir des lignes spécifiques de notre jeu de données. la figure 5.10, montre un exemple d'une requête.



	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
1208	-117.09	32.56	8.0	864.0	156.0	626.0	172.0	4.8984	151500.0
2371	-117.08	32.57	18.0	2203.0	544.0	1943.0	497.0	2.2500	103200.0
1617	-117.12	32.57	35.0	1450.0	256.0	930.0	286.0	2.6715	133300.0
1045	-117.11	32.57	32.0	2723.0	586.0	1702.0	562.0	3.3371	140500.0
274	-117.05	32.58	23.0	1918.0	339.0	1392.0	340.0	4.0870	134800.0
2649	-117.13	32.58	27.0	2511.0	615.0	1427.0	576.0	3.1645	156000.0
2415	-117.07	32.58	25.0	1607.0	280.0	899.0	260.0	3.8194	134400.0
1649	-117.08	32.58	15.0	1462.0	274.0	1002.0	271.0	3.9698	142700.0
1709	-117.11	32.58	12.0	1086.0	294.0	870.0	290.0	2.4213	132500.0
1761	-117.10	32.58	33.0	393.0	76.0	330.0	80.0	4.1029	122700.0
609	-117.12	32.59	28.0	2793.0	706.0	1825.0	676.0	2.6724	144500.0
1695	-117.06	32.59	13.0	3920.0	775.0	2814.0	760.0	4.0616	148800.0
1724	-117.07	32.60	13.0	1607.0	435.0	983.0	400.0	2.2903	106300.0
565	-117.06	32.61	24.0	4369.0	1353.0	3123.0	1247.0	2.0571	152300.0
867	-117.05	32.61	31.0	4033.0	715.0	2585.0	715.0	3.5096	139900.0
1312	-117.03	32.61	23.0	1553.0	216.0	778.0	229.0	5.1538	171300.0
942	-117.06	32.61	23.0	1630.0	362.0	1267.0	418.0	2.5625	131100.0
1428	-117.04	32.62	26.0	3620.0	607.0	2000.0	593.0	4.9962	156000.0
1637	-117.09	32.62	37.0	1538.0	298.0	867.0	285.0	3.0729	128700.0
1215	-117.09	32.64	19.0	2571.0	791.0	1205.0	783.0	1.6200	131300.0
573	-117.08	32.64	43.0	1005.0	230.0	548.0	252.0	1.8672	145800.0
1743	-117.09	32.66	37.0	1232.0	330.0	1086.0	330.0	1.6389	114300.0
740	-117.12	32.66	52.0	16.0	4.0	8.0	3.0	1.1250	60000.0
1355	-117.11	32.66	52.0	25.0	5.0	14.0	9.0	1.6250	118800.0
2184	-115.49	32.67	24.0	1266.0	275.0	1083.0	298.0	1.4828	73100.0
1940	-117.04	32.68	14.0	1320.0	270.0	943.0	260.0	5.0947	152700.0
2541	-117.18	32.68	29.0	1539.0	344.0	556.0	289.0	3.2500	500001.0

FIGURE 5.10 – Résultat requête OLAP

5.5 Conclusion

Le dernier chapitre de notre projet a été dédié à l'implémentation de l'application qui mettra en pratique la solution conceptuelle élaborée dans le chapitre précédent. Cette implémentation est en parfaite adéquation avec ce qui a été prévu au début de notre projet et qui n'est autre que la proposition d'une solution pour l'intégration des techniques de fouille de données dans un système BI.

CONCLUSION GÉNÉRALE

La fouille de données dans un système de business intelligence offre un potentiel énorme pour les organisations souhaitant tirer parti de leurs données pour prendre des décisions éclairées et rester compétitives sur le marché.
Ce projet de fin d'études de master a exploré les opportunités et les défis associés à l'utilisation de techniques de fouille de données dans le contexte de la BI, en mettant l'accent sur les différentes étapes du processus de fouille de données, les techniques applicables et les exemples de cas d'étude.
En conclusion, nous avons constaté que l'utilisation de techniques de fouille de données permet aux entreprises de découvrir des tendances, des modèles et des informations cachées dans leurs données, ce qui peut conduire à de nouvelles opportunités commerciales et à une amélioration de l'efficacité opérationnelle. Les techniques de classification, de prédiction, de regroupement et d'association se sont révélées particulièrement utiles dans le cadre de la BI, offrant des outils puissants pour la segmentation des clients, la prévision des ventes, l'optimisation des processus et la détection de fraudes, entre autres.
Sur le plan personnel, la conduite de ce projet m'a permet de capitaliser les compétences suivantes :

— Maîtrise du domaine de la BI.

— Exploration et approfondissement des connaissances sur les techniques de fouille de données.

— Amélioration de mes compétences en python.

— Prise en main et approfondissement de mon expérience dans l'utilisation de l'environnement de composition de texte latex.

Enfin, bien que nous avons échoué par contrainte de temps, à développé notre PFE dans le cadre de l'arrêté 1275 relatif à la création de startup, nous espérons concrétiser ce projet dans le cadre de la vie professionnelle.

BIBLIOGRAPHIE

[1] https://www.oracle.com/fr/database/business-intelligence-definition.html.

[2] https://www.journaldunet.fr/web-tech/guide-du-big-data/1198305-etl-outils-definition-traduction/.

[3] https://www.talend.com/.

[4] https://www.lebigdata.fr/data-warehouse-entrepot-donnees-definition.

[5] https://www.hitachivantara.com/fr-fr/products/data-management-analytics/pentaho-platform.html/.

[6] https://help.hitachivantara.com/Documentation/Pentaho/9.4/.

[7] https://docs.microsoft.com/en-us/sql/integration-services/.

[8] https://www.sas.com/.

[9] https://hadoop.apache.org/.

[10] https://learn.microsoft.com/en-us/power-bi/.

[11] https://docs.python.org/fr/3/tutorial/.

[12] https://numpy.org/doc/stable/.

[13] https://scikit-learn.org/stable/.

[14] LILIA BOUCENA. « Une nouvelle approche d'intégration des données des processus métiers basée sur la technologie ETL ». In : (2022).

[15] BOCHRA BOUZIANE. « Exploitation des bases de données graphes pour le stockage et l'interrogation des données des processus métiers ». In : (2021).

[16] Hsinchun CHEN, Roger HL CHIANG et Veda C STOREY. « Business intelligence and analytics : From big data to big impact ». In : *MIS quarterly* (2012), p. 1165-1188.

[17] Leanndra COPELAND et al. « Applying business intelligence concepts to Medicaid claim fraud detection ». In : *Journal of Information Systems Applied Research* 5.1 (2012), p. 51.

[18] MASIKA MUYISA DORCAS et TITULAIRE DU COURS. « DIFFERENCE ENTRE OLTP ET OLAP EN BUSINESS INTELLIGENCE ». In : ().

[19] MJSRM GHAZANFARI, M JAFARI et S ROUHANI. « A tool to evaluate the business intelligence of enterprise systems ». In : *Scientia Iranica* 18.6 (2011), p. 1579-1590.

[20] Zhi-xiong HUANG, KS SAVITA et Jiang ZHONG-JIE. « The Business Intelligence impact on the financial performance of start-ups ». In : *Information Processing & Management* 59.1 (2022), p. 102761.

[21] Dorina KABAKCHIEVA. « Business Intelligence Applications and Data Mining Methods in Telecommunications : A Literature Review ». In : (2009).

[22] Brojo Kishore MISHRA et al. « Business intelligence using data mining techniques and business analytics ». In : *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*. IEEE. 2016, p. 84-89.

[23] Brojo Kishore MISHRA et al. « Business intelligence using data mining techniques and business analytics ». In : *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*. IEEE. 2016, p. 84-89.

[24] Bogdan NEDELCU et al. « Business intelligence systems ». In : *Database Systems Journal* 4.4 (2013), p. 12-20.

[25] Abdullahi Sidow OSMAN. « Data mining techniques ». In : (2019).

[26] *Ïpolycopié de cours : DATA MINING", author=Dr. Brahim Farou*.

[27] *Ïpolycopié pédagogique :Business Intelligence (Informatique Décisionnelle) Cours ", author=DR. Khebizi Ali*.