

# Dédicaces

Toutes les lettres et tous les mots ne peuvent exprimer de la gratitude, du respect et de l'appréciation à ceux qui m'ont entouré. Je m'engage dans ce travail :

## **À mes chères parents Hamda & Samira**

Qui m'ont toujours poussé, soutenu et motivé en tout, Puissent-ils trouver dans ce travail l'expression de ma gratitude et une récompense pour tous les sacrifices consentis pendant toutes ces années.

## **À mes Chères frères Saif & Ramez**

Merci pour vos encouragements continus et votre soutien spirituel. Je vous dédie ce travail en témoignage de l'amour et de l'affection que je porte pour vous.

## **À ma grande Famille et mes chers amis**

Je vous souhaite plein de succès et de bonheur dans votre vie professionnelle et privée. Je vous dédie ce modeste travail résultat de mon travail acharné et de mes années d'apprentissage.

*Jasser Hadhri*

# Remerciements

*Je remercie Dieu de m'avoir donné la santé et la volonté de terminer ce projet de fin d'études. Je suis très heureux de garder ces mots, pour exprimer ma gratitude à ceux qui ont contribué directement ou indirectement à la réalisation du projet et m'ont apporté de l'aide et de soutien tout au long du stage.*

*Tout d'abord, je tiens à exprimer mon plus grand respect et à remercier Monsieur **Walid AZZOUZI**, mon superviseur au sein de l'entreprise d'accueil **Neoledge**, pour son aide et ses conseils pendant tout le processus de développement de ce projet.*

*Je présente aussi mes remerciements à Madame **Nehla DEBBABI**, mon encadrante à **ESPRIT** pour sa disponibilité, ses conseils précieux ainsi que pour son aide.*

*Aussi, je voudrais adresser mes salutations aux membres du jury et les remercier pour l'honneur qu'ils nous ont apporté en acceptant de juger le présent travail. En espérant qu'ils trouvent dans ce rapport les qualités de motivation et de clarté qu'ils espèrent.*

*Je tiens également à remercier tous les membres de **Neoledge** pour leur accueil chaleureux, leurs conseils avisés et leur bonne énergie partagée.*

*Enfin, je suis profondément reconnaissant à tous les cadres professionnels à **ESPRIT** qui nous fournissent une excellente formation, pour leurs suggestions et contributions, afin que notre carrière commence sur des bases solides.*

# Table des matières

<b>Introduction générale</b>	<b>1</b>
<b>1 Contexte général</b>	<b>3</b>
1.1 Introduction . . . . .	4
1.2 Présentation du projet . . . . .	4
1.2.1 Cadre général du projet . . . . .	4
1.2.2 Présentation de l'organisme d'accueil . . . . .	4
1.3 Problématique . . . . .	8
1.4 Solution proposée . . . . .	8
1.5 Méthodologie de travail . . . . .	9
1.5.1 Cycle de vie de Ralph Kimball . . . . .	9
1.5.2 Méthodologie de GIMSI . . . . .	11
1.5.3 Choix de la méthodologie . . . . .	13
1.6 Conclusion . . . . .	13
<b>2 Définition des besoins métiers et planification du projet</b>	<b>14</b>
2.1 Introduction . . . . .	15
2.2 Planification du projet . . . . .	15
2.3 Identification des acteurs . . . . .	16
2.4 Spécification des besoins . . . . .	16
2.4.1 Besoins fonctionnels . . . . .	16
2.4.2 Besoins non fonctionnels . . . . .	17
2.5 Identification et description des cas d'utilisation du système . . . . .	17
2.6 Conclusion . . . . .	21
<b>3 Environnement technique</b>	<b>22</b>
3.1 Introduction . . . . .	23
3.2 Architecture technique du système . . . . .	23
3.3 Environnement technique du travail . . . . .	25
3.3.1 SQL Server . . . . .	25

---

3.3.2	MSBI . . . . .	26
3.3.3	POWER BI . . . . .	27
3.3.4	Python . . . . .	27
3.3.5	VueJs . . . . .	28
3.4	Conclusion . . . . .	28
<b>4</b>	<b>Intégration des données</b>	<b>29</b>
4.1	Modélisation multidimensionnelle . . . . .	30
4.2	Approche utilisée . . . . .	31
4.2.1	Approche de Ralph Kimball (ou approche Bottom-up) . . . . .	31
4.2.2	Approche Bill Inmon . . . . .	31
4.3	Entrepôt de données . . . . .	32
4.3.1	Modèle en étoile . . . . .	33
4.3.2	Modèle en flocon . . . . .	33
4.3.3	Modèle en constellation . . . . .	34
4.3.4	Détermination des tables de dimensions . . . . .	34
4.3.5	Identification des tables de faits . . . . .	37
4.3.6	Conception du modèle . . . . .	38
4.4	Développement et préparation des données . . . . .	39
4.4.1	Etape d'alimentation de Data Staging Area . . . . .	40
4.4.2	Etape de modélisation . . . . .	41
4.5	Conclusion . . . . .	44
<b>5</b>	<b>Data Mining et web scraping</b>	<b>45</b>
5.1	Introduction . . . . .	46
5.2	Data Mining . . . . .	46
5.2.1	Règles d'associations . . . . .	46
5.2.2	segmentation des clients avec RFM . . . . .	50
5.2.3	Série Temporelle . . . . .	53
5.3	Web scraping . . . . .	61
5.3.1	Extraction de données . . . . .	61
5.3.2	Analyse de données . . . . .	62

---

5.4 Conclusion . . . . .	64
<b>6 Réalisation</b>	<b>65</b>
6.1 Introduction . . . . .	66
6.2 Représentation des applications utilisateurs . . . . .	66
6.2.1 Maquette : Vente dashboard . . . . .	67
6.2.2 Maquette : Perte et casse dashboard . . . . .	67
6.2.3 Maquette : Détails de vente . . . . .	68
6.2.4 Maquette : Suivi de vente . . . . .	68
6.2.5 Maquette : Performances . . . . .	69
6.3 Construction de l'application utilisateur . . . . .	69
6.3.1 Vente Dashboard . . . . .	69
6.3.2 Perte et casse Dashboard . . . . .	70
6.3.3 Détails de vente dashboard . . . . .	71
6.3.4 Suivi de vente Dashboard . . . . .	72
6.4 Performance Dashboard . . . . .	72
6.5 Conclusion . . . . .	73
<b>Conclusion générale</b>	<b>74</b>

# Table des figures

1.1	Logo de NeoLedge[2]	5
1.2	Méthodologie de Ralph Kimball	10
2.1	Planification du projet et identificaiton des besoins	15
2.2	Diagramme de Gantt	16
2.3	Diagramme de cas d'utilistion général	18
3.1	Branche technique du cycle de vie de Ralph Kimball	23
3.2	Architecture technique d'un système de prise de décision	24
3.3	Architecture technique du web scraping	24
3.4	Logo Microsoft SQL server	25
3.5	Architecture MSBI [8]	26
3.6	Architecture POWER BI	27
3.7	Python	28
3.8	VueJS	28
4.1	Branche de modélisation du cycle de vie de Ralph Kimball	30
4.2	L'approche de Kimball [9]	31
4.3	L'approche d'Inmon [10]	32
4.4	Modèle en étoile	33
4.5	Modèle en flocon	34
4.6	Modèle en constellation	34
4.7	Conception de l'entrepôt des données	39
4.8	Restauration de la base de données Backup	41
4.9	Composant Source OLE DB	42
4.10	Connexion à une table	42
4.11	Flux de donnée d'alimentation de Dim-produit	43
4.12	Destination OLE DB	44
5.1	Convertir les données de ventes	48
5.2	Fonction de règles d'associations	49

5.3	Résultats des règles d'associations . . . . .	49
5.4	Supression des données de retour et casse . . . . .	51
5.5	Explorer les valeurs uniques de chaque attribut . . . . .	51
5.6	Dernière date d'achat . . . . .	51
5.7	Calcul du fréquence . . . . .	51
5.8	Calcul du monétaire . . . . .	52
5.9	Table RFM . . . . .	52
5.10	Fonction de calcul de score RFM . . . . .	53
5.11	Total des clients dans chaque segment . . . . .	53
5.12	Convertir les données à une série temporelle . . . . .	54
5.13	Traçage des données . . . . .	55
5.14	Génération des différentes combinaisons de paramètres . . . . .	56
5.15	Itérations des différentes combinaisons des paramètres . . . . .	57
5.16	Diagnostic du modèle . . . . .	58
5.17	Comparaison des valeurs . . . . .	59
5.18	Prédiction de vente d'un produit déterminé . . . . .	60
5.19	Prédiction du Chiffre d'affaire . . . . .	60
5.20	Architecture technique du web scraping . . . . .	61
5.21	Dictionnaire des sentiments . . . . .	63
5.22	Résultat du web scraping et de l'analyse sentimentale . . . . .	63
5.23	Courbe des interactions sur une publication . . . . .	64
6.1	Création des maquettes et développement de la solution . . . . .	66
6.2	Maquette vente dashboard . . . . .	67
6.3	Maquette Perte et casse dashboard . . . . .	67
6.4	Maquette Détails de vente . . . . .	68
6.5	Maquette Suivi de vente . . . . .	68
6.6	Maquette de Performance . . . . .	69
6.7	Vente Dashboard . . . . .	70
6.8	Perte et casse dashboard . . . . .	71
6.9	Détails de vente Dashboard . . . . .	71
6.10	Suivi de vente Dashboard . . . . .	72

Table des figures

---

6.11 Performances Dashboard . . . . .	73
---------------------------------------	----

# Liste des abréviations

- **BI** = Business Intelligence
- **BPM** = Business Process Management
- **ECM** = Entreprise Content Management
- **ERP** = Enterprise Resource Planning
- **ETL** = Extract Transform Load
- **GED** = Gestion électronique des documents
- **ODS** = Operating Data Store
- **OLAP** = On Line Analytical Processing
- **SA** = Staging Area
- **SQL** = Structured Query Language
- **SSAS** = SQL Server Analysis Services
- **SSIS** = SQL Server Integration Services

# Introduction générale

La prise de décision est l'étape la plus importante de la vie d'une entreprise. L'objectif principal des chefs d'entreprise est de veiller sur la prise des décisions opportunes au bénéfice de leur entreprise. Afin de maintenir leur position sur le marché, ils ont aujourd'hui tendance à mettre en place un système de prise de décision pour survivre l'évolution de leurs environnement de compétition cruciale.

Au sein d'une entreprise, une décision est un choix lié aux objectifs fixés, aux enjeux soulevés ou à l'utilisation des ressources.

En effet, les données de Ventes et Marketing ont été toujours les plus reconnues comme des données porteuses d'informations cruciales pour l'entreprise. Les explorer soigneusement et les analyser judicieusement, demeure ainsi une étape primordiale dans l'amélioration de ses connaissances, l'optimisation de sa stratégie de croissance et la fidélisation de ses clients.

Afin d'assurer la pérennité de l'entreprise à court et moyen terme, aussi bien son développement à long terme, il est nécessaire de mettre en place une stratégie de pilotage facilitant la planification. Par conséquent, il est clair que les expériences ne peuvent pas être menées de manière improvisée. Si le pilotage n'est pas réalisé convenablement, il sera difficile de gérer les activités de l'entreprise et il est nécessaire d'avoir une compréhension en temps réel et prospective de la santé de l'entreprise.

Pour garantir une prise de décision sûre, les entreprises utilisent le Business Intelligence. Grâce à cette technique, il est possible de collecter toutes les données de l'entreprise dans une structure de stockage bien déterminée et d'effectuer les traitements nécessaires pour leur donner une signification cohérente et utile, afin de les restituer enfin sous une forme visuelle, pour effectuer une analyse et une prise de décision puissante et facile.

Dans le cadre de ce projet , nous développons une plateforme décision et des tableaux de bord au sein de Neoledge, afin que les managers puissent suivre correctement leurs trafic de vente et leurs stratégie de marketing pour prendre des mesures correctives. .

## Organisation du manuscrit

Le présent manuscrit est composé de 6 chapitres. Le premier chapitre décrit le contexte du projet. Nous commençons par cadre général du travail effectué ainsi que l'organisme d'accueil. Ensuite, nous passons par énoncer la problématique ainsi que la solution proposée et de l méthodologie de travail.

Dans le deuxième chapitre nous présentons l'explication de la plannification des différentes stations ainsi que les besoins de notre projet.

L'architecture de l'application ainsi que les outils utilisés sont détaillées dans le chapitre 3.

Ensuite, nous passons au quatrième chapitre qui est dédié à la présentation de la conception physique de notre entrepôt de données , la phase d'alimentation et de modélisation de notre projet.

Dans le cinquième chapitre , nous présentons la partie d'analyse en utilisant les algorithmes de data mining et le web scraping pour bien avoir une visibilité sur les sentiments des consommateurs.

Enfin, nous présentons dans le sixième chapitre la partie réalisation.

# CONTEXTE GÉNÉRAL

---

## Plan

1	Introduction . . . . .	4
2	Présentation du projet . . . . .	4
3	Problématique . . . . .	8
4	Solution proposée . . . . .	8
5	Méthodologie de travail . . . . .	9
6	Conclusion . . . . .	13

## 1.1 Introduction

Ce chapitre est une mise en situation de l'environnement de notre projet de fin d'étude. Nous aborderons, dans la première partie, la présentation du projet à travers une présentation du cadre général de l'organisme d'accueil. Nous enchaînerons avec une présentation sur l'organisme d'accueil, nous exposerons la problématique et la solution proposée et pour finir, nous ferons le choix de la méthodologie.

## 1.2 Présentation du projet

Nous envisageons dans ce qui suit une présentation du présent projet pour bien le comprendre le cadre et délimiter ce qui est demandé pour passer à l'action et répondre aux spécifications d'une façon concise et efficace.

### 1.2.1 Cadre général du projet

Le présent travail a été réalisé dans le cadre du projet de fin d'études qui conclut la formation d'ingénieur en informatique spécialisée en **Business Intelligence** et **Enterprise Resource Planning** (ERP-BI) à l'**Ecole Supérieure Privée d'Ingénierie et de Technologie (ESPRIT)**. Réalisé au sein de l'entreprise d'accueil Neoledge Tunisie, ce projet intitulé "Elise BI – Ventes Marketing", consiste à la mise en place d'une plateforme décisionnelle permettant de surveiller les activités de vente et marketing d'un client de Neoledge. Celle-ci donne la possibilité aux décideurs de l'organisme client de suivre le trafic de ventes et des actions marketing à travers une interface bien structurée qui répond à tous leurs besoins. Pour mieux comprendre les détails du projet nous commençons par présenter l'organisme d'accueil qui est Neoledge Tunisie, ensuite nous décrivons la problématique et fini par introduire la solution proposée .

### 1.2.2 Présentation de l'organisme d'accueil

Cette partie présente l'entreprise dans laquelle le projet de fin d'études a été effectué, ses domaines d'activités et les innovations qu'elle apporte.

#### 1.2.2.1 Neoledge

Ce projet a été réalisé au sein de Neoledge Moyen-Orient et l'Afrique (Middle-East and Africa MEA) filiale du groupe Archimed[1], ses activités principales consistent à offrir des solutions

de dématérialisation, Capture des flux multicanaux, Gestion Electronique des Documents (GED) , gestion de processus des entreprises (Business Process Management « BPM ») et signature électronique parfaitement adaptées aux besoins de leurs clients . Depuis 1993, les créateurs de NeoLedge et de l'activité GED / gestion de contenu des entreprises (Entreprise Content Management ECM) du groupe Archimed, n'ont jamais perdu de vue leur valeur fondatrice : l'innovation. Portés par elle au quotidien, elle nous incite à rechercher sans cesse de nouvelles solutions, pour accompagner nos clients dans leur transformation digitale et leur recherche de performance. Aujourd'hui, NeoLedge dont le logo est illustré dans la figure 1.1, est fière d'être reconnue comme un éditeur et intégrateur de logiciels de gestion documentaire majeure et à dimension internationale. Etant un partenaire de Microsoft certifié Gold, qui exerce dans le domaine de développement d'applications et des plateformes cloud depuis sa création, elle soutient ses différents clients dans leurs transition numérique. Leurs solutions sont proposées sur le cloud et peuvent également être déployées on-premises (localement). Ils proposent des scénarios de déploiement hybrides, ou encore une intégration complète à Office 365, pour offrir un maximum de flexibilité à leurs clients.



**Figure 1.1:** Logo de NeoLedge[2]

Pour une connaissance plus élargie sur NeoLedge, nous donnerons ci-dessous un aperçu sur son historique.

#### **1.2.2.2 Historique de Neoledge - Groupe Archimed**

Crée en 1993 en France, le Groupe Archimed démarré son activité dans le secteur des bibliothèques et des musées en proposant des solutions multimédias de gestion et de diffusion documentaire. Devenu après 10 ans un éditeur reconnu dans le secteur, la société a très vite élargie son offre au secteur des administrations et des collectivités pour bâtir des applications qui répondent aux préoccupations relatives à la modernisation de l'État, aux Technologies de l'information et de la communication (TIC) et aux sociétés d'information (dématérialisation, gestion du courrier, portail citoyen, etc.). En Juin 2000, Archimed crée sa filiale en Tunisie afin de conquérir le marché du grand Maghreb, Moyen-Orient et de l'Afrique (Middle-East and Africa MEA) . En Juillet 2018, Archimed

choisit de distinguer son activité de Gestion Electronique des Documents (GED) et Enterprise Content Management (ECM) ayant le nom Neoledge, afin d'être mieux identifié sur les marchés en Europe, Afrique et Amérique du Nord.

Le tableau chronologique d'Archimed est représenté par le tableau 1.1 .

<b>1993</b>	<b>Création d'Archimed</b>
<b>2000</b>	<b>Création de la filiale Archimed Tunisie</b>
<b>2008</b>	<b>Création de l'agence de Paris</b>
<b>2010</b>	<b>Rachat d'Opsys, éditeur de logiciels pour les bibliothèques</b>
<b>2015</b>	<b>Lancement du logiciel Syracuse</b>
<b>2018</b>	<b>Filialisation des deux activités GED / ECM et bibliothèques en Neoledge et Archimed</b>

**Tableau 1.1:** Chronologie d'Archimed [1]

Dans ce qui suit, nous présentons les principaux services et activités de Neoledge .

#### **1.2.2.3 Activités et services de Neoledge**

Neoledge aujourd'hui offre ses services pour plus de 250 clients et 2500000 utilisateurs dans plus de 20 pays dans le monde. Son domaine d'activité englobe plusieurs secteurs tels que :

- **Le Gouvernement** (Présidence du gouvernement de Tunisie, Le Ministère de la Défense de France . . . )
- **Les institutions financières Publics et Privés** (La Banque de France , La Banque Centrale de Tunisie, La Banque Centrale de Madagascar, Union Economique et Monétaire Ouest Africaine , Attijari Bank, Assurance Comar . . . )
- **Les établissements publics** (Le Centre National de l'Informatique Tunis , Le Conseil d'Etat , Le Grand Port Maritime de Nantes Saint-Nazaire ..)
- **Smart cities** (Villes de Bayonne , Paris , Saint-Etienne , Lille . . . )
- **Énergie** (Total , Petro-Gabon .. )
- **Industries** ( Notre client , DCNS , InVivo . . . )

Après avoir présenter un aperçu surles activités et services de Neoledge, nous passons maintenant a présenter ce que Neoledge offre comme produits.

#### 1.2.2.4 Produits de Neoledge

L'entreprise Neoledge offre plusieurs produits, parmi lesquels nous avons DocFactory et Elise. Ces derniers représentent deux produits phares, c'est pour cela nous avons choisi de les détailler de plus .

- **Docfactory** : Un logiciel dédié pour la gestion des chaines de traitement des documents. Il est accessible depuis le web. Plus de détails sur ce produit sont donnés dans le tableau 1.2.

<b>DocFactory</b>	<b>Commentaire</b>
Fonctions	Gestion des chaines de traitement des documents
Type d'application	Full web, accessible depuis un simple navigateur
Date de création	2006
Date de première installation	2007
Version actuelle	Version 3.3 – sortie en février 2016

**Tableau 1.2:** Spécifications de DocFactory

- **Elise** : Un logiciel dédié pour la gestion électronique des documents, traitement collaboratif, parapheur électronique... Il est accessible depuis le web. plus de détails sur ce produit sont donnés dans le tableau 1.3.

<b>Elise</b>	<b>Commentaire</b>
Fonctions	Gestion Electronique des Documents, traitement collaboratif, parapheur électronique, etc.
Type d'application	Full web, accessible depuis un simple navigateur
Date de création	2004 (dans le cadre d'un projet avec le ministère de l'intérieur)
Date de première installation	2005
Version actuelle	Version 6 – sortie en Juin 2018

**Tableau 1.3:** Spécifications de Elise

Comme toute entreprise, Neoledge veut élargir les secteurs et les activités qu'elle couvre . Dans ce qui suit on va présenter la problématique que nous allons aborder dans ce projet .

### 1.3 Problématique

Les données présentent aujourd’hui la matière première pour plusieurs tâches industrielles. La gestion et l’exploitation de ces données constituent ainsi une phase primordiale qui doit être prise en compte par les entreprises qui souhaitent maintenir un avantage concurrentiel. Cette phase peut être assurée par la « Business Intelligence » qui est devenue de nos jours l’une des préoccupations principales au sein des Directions des Systèmes d’Information des entreprises. Dans ce contexte, NeoLedge se voit capable de fournir des solutions intelligentes et performantes à ses clients. Par-là, notre client, qui est un groupe tunisien parmi les leaders de l’industrie agroalimentaire pour une plateforme décisionnelle sur l’application ELISE fournie par NeoLedge qui va lui permettre d’avoir une vue structurée de leurs activités de ventes, de prendre des mesures préventives, d’effectuer des analyses prédictives et multidimensionnelles afin de prévoir les aberrations et de prendre des décisions efficaces qui contribuera à l’amélioration des performances de ventes et marketing de l’entreprise. Nous notons que pour des raisons de confidentialité, nous ne dévoilons pas, tout au long de ce manuscrit, toute information dévoilant l’identité de notre client.

### 1.4 Solution proposée

La solution recommandée est de mettre en place une plateforme décisionnelle qui sera développée dans le cadre du système « ELISE » pour les entreprises agroalimentaires, qui permettra à nos clients de suivre les activités de leurs entreprises à travers un ensemble d’indicateurs clairement définis et de les regrouper dans les tableaux de bord. Les principales tâches à réaliser sont les suivants :

- Mise en place d’une plateforme décisionnelle, en suivant une démarche BI, allant de la collecte des données jusqu’à la génération des tableaux de bord interactifs, résumant les trafics de vente, afin de permettre aux gestionnaires de notre client d’avoir une vue d’ensemble, ainsi qu’une bonne maîtrise sur l’état actuel.
- Analyse prédictive sur le trafic de vente grâce à la Data Mining, afin de pouvoir classifier les clients de notre client .
- Analyse sentimentale sur le comportement des consommateurs de notre client grâce aux Web Scraping et le Text Mining, afin de pouvoir collecter les commentaires et les reactions sur les publications de la page Facebook de l’entreprise et analyser les commentaires .
- Développement d’une plateforme web

On sait que la vente touche à plusieurs critères tels que :

- **Les promotions .**
- **Les produits .**
- **Le Chiffre d'affaire .**
- **Les clients .**
- **Les consommateurs .**
- **Les Taxes .**

Pour réussir un projet BI , il faut suivre une bonne méthodologie de travail. C'est pour cela, nous présentons dans ce qui suit les deux méthodologies les plus utilisées en BI, nous les comparons, et nous choisissons celle qui conviendra à notre projet .

## 1.5 Méthodologie de travail

La réalisation optimale d'un projet dans les bonnes conditions exige une méthodologie adoptée au projet en question ainsi qu'un comportement procédural tenant compte des différentes finalités des résultats. Nous commençons par présenter la méthodologie de Ralph kimball

### 1.5.1 Cycle de vie de Ralph Kimball

Le cycle de vie Kimball a été élaboré par des membres du groupe Kimball, et cette méthode a été utilisée avec succès pour développer des projets d'entrepôt de données. Les différentes étapes de cette méthodologie sont :

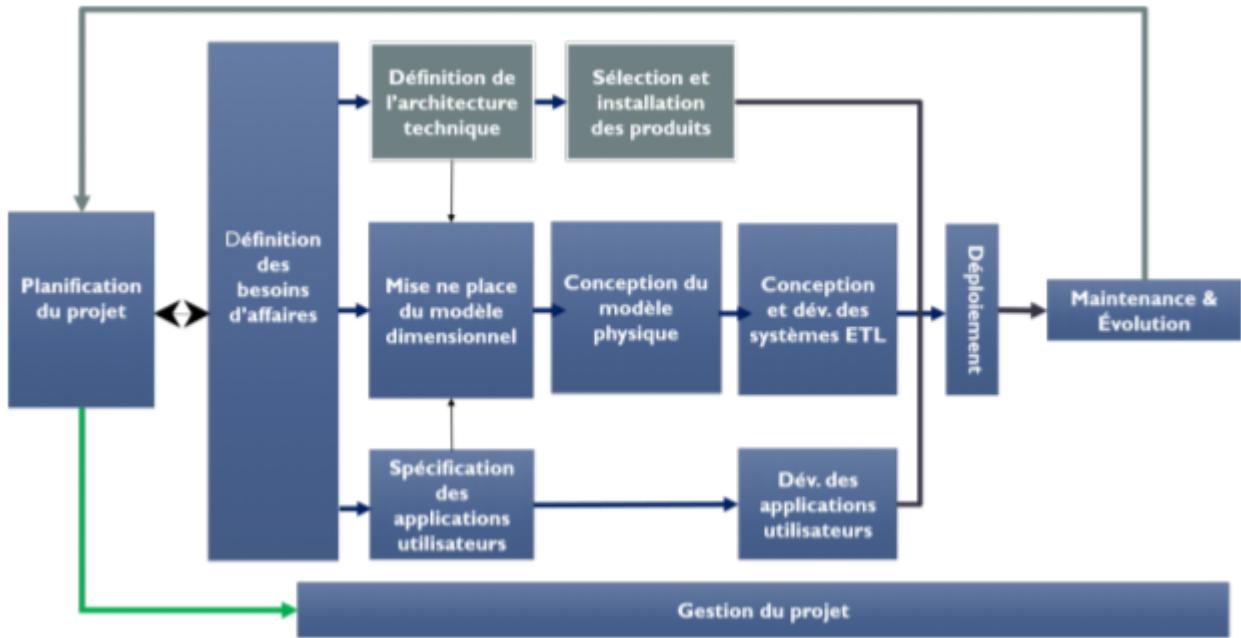


Figure 1.2: Méthodologie de Ralph Kimball

Phases	Etapes
Planification du projet	<ul style="list-style-type: none"> <li>• Définir l'étendue du projet</li> <li>• Définir les besoins en termes de ressources</li> <li>• Classifier les tâches suivant 1 durée et 1 séquence.</li> </ul>
Définition des besoins d'affaires	<ul style="list-style-type: none"> <li>• Savoir les besoins des utilisateurs de l'application ...</li> <li>• Traduction des besoins en facteurs à associer lors de la phase de conception.</li> </ul>
Conception de l'architecture technique	Mettre en œuvre la structure de l'architecture technique en tenant compte des exigences environnementales techniques existantes et de l'orientation technique prévue.

<b>Sélection et installation des outils</b>	Sélectionner et installer les logiciels et les outils indispensables pour notre travail.
<b>Modélisation des données</b>	Faire une analyse détaillée des données des systèmes opérationnels. Réunir l'analyse avec les besoins dégagés dans la phase précédente.
<b>Conception physique</b>	Définir les structures physiques pour la réalisation de la base de données.
<b>Conception, développement et test du système ETL</b>	C'est la phase qui prend le plus de temps, elle comporte : <ul style="list-style-type: none"> <li>• l'extraction .</li> <li>• la transformation.</li> <li>• le chargement des données .</li> </ul>
<b>Développement de l'application BI</b>	Consiste à : <ul style="list-style-type: none"> <li>• modéliser des tableaux de bord, des indicateurs de performance pour chaque utilisateur.</li> <li>• Configurer les outils.</li> <li>• Développer l'application qui sera fournie aux utilisateurs.</li> </ul>
<b>Déploiement</b>	consiste à mettre en place les processus de communication.
Gestion de projet	Cette étape permet de garantir l'enchaînement des activités du cycle de vie dimensionnel en surveillant l'état d'avancement du projet ,identifier les problèmes ainsi que de résoudre les problèmes identifiés.

### 1.5.2 Méthodologie de GIMSI

La méthodologie GIMSI est une méthode collaborative de conception de systèmes de gestion, le système de gestion étant le point central de la gestion des performances de l'entreprise ou de l'entreprise. La méthodologie de GIMSI est structurée en 10 étapes et regroupé en 4 phases qui sont

illustrés dans le tableau 1.4

Phases	Etapes
1. Identification	<ul style="list-style-type: none"><li>● Etape 1 : Environnement de l'entreprise Analyse de l'environnement économique et de la stratégie de l'entreprise afin de définir le périmètre et la portée du projet</li><li>● Etape 2 : Identification de l'entreprise Analyse des structures de l'entreprise pour identifier les processus, activités et acteurs concernés.</li></ul>
2. Conception	<ul style="list-style-type: none"><li>● Etape 3 : Définition des objectifs Sélection des objectifs tactiques de chaque équipe en fonction de la stratégie générale.</li><li>● Etape 4 : Construction du tableau de bord. Définition du tableau de bord de chaque équipe.</li><li>● Etape 5 : Choix des indicateurs Choix des indicateurs en fonction des objectifs choisis, du contexte et des acteurs concernés.</li><li>● Etape 6 : Collecte des informations Identification des informations nécessaires à la construction des indicateurs.</li><li>● Etape 7 : Le système de tableau de bord Construction du système de tableau de bord, contrôle de la cohérence globale.</li></ul>

3. Mise en œuvre	<ul style="list-style-type: none"><li>• Etape 8 : Le choix des progiciels Elaboration de la grille de sélection pour le choix du progiciel adéquat.</li><li>• Etape 9 : Intégration et déploiement Implantation des progiciels, déploiement à l'entreprise.</li></ul>
4. Amélioration permanente	<ul style="list-style-type: none"><li>• Etape 10 : Audit Suivi permanent du système</li></ul>

**Tableau 1.5:** Phases et étapes de la méthode GIMSI

### 1.5.3 Choix de la méthodologie

La raison pour laquelle nous avons choisi d'utiliser le cycle de vie de Kimball est que GIMSI est basé sur le principe de coopération dans la conception du système de direction, il nécessite donc la participation de plusieurs décideurs, ce que notre clients ne le possèdent pas.

## 1.6 Conclusion

Dans ce chapitre , nous avons commencé par présenter le cadre générale du projet. Par la suite, nous sommes passés à la présentation de l'organisme d'accueil Neoledge , la problématique, et les objectifs à atteindre qui constituera le sujet de notre projet. Enfin, après avoir choisi la méthode de Ralph Kimball comme méthode de travail, nous nous dirigeons vers le deuxième chapitre, où, nous aurons plus de visibilité sur l'aspect décisionnel du projet.

# DÉFINITION DES BESOINS MÉTIERS ET

## PLANIFICATION DU PROJET

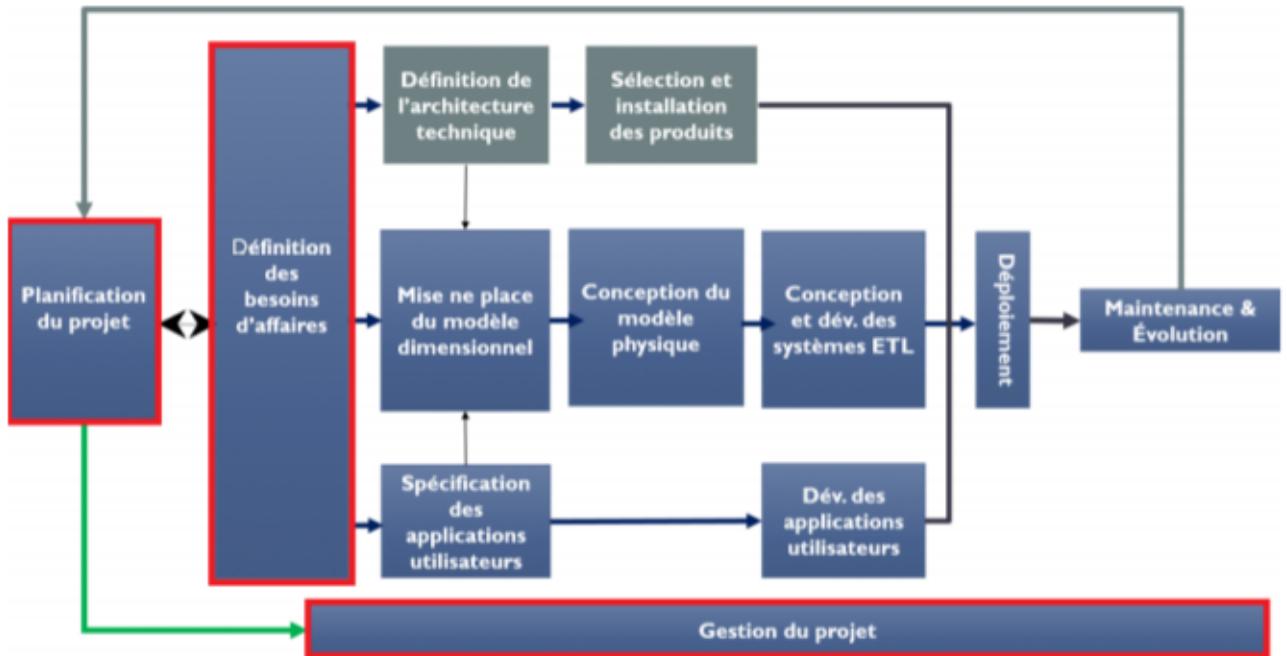
---

### Plan

1	Introduction . . . . .	15
2	Planification du projet . . . . .	15
3	Identification des acteurs . . . . .	16
4	Spécification des besoins . . . . .	16
5	Identification et description des cas d'utilisation du système . . . . .	17
6	Conclusion . . . . .	21

## 2.1 Introduction

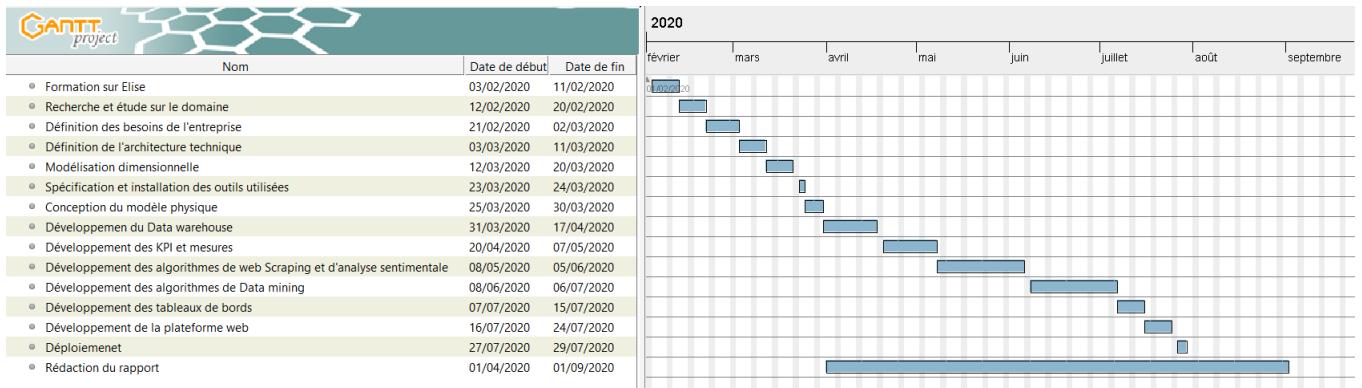
Cette partie est dédiée à la spécification des exigences, qui est la branche fonctionnelle de la méthodologie Ralph Kimball. Tout d'abord, nous commençons par la planification du projet et l'identification des participants, ce qui nous conduira à introduire des exigences fonctionnelles et non fonctionnelles.



**Figure 2.1:** Planification du projet et identificaiton des besoins

## 2.2 Planification du projet

Pour mener à bien ce projet, nous avons conçu un diagramme de Gantt dans lequel nous présentons les différentes tâches. La figure 2.2 illustre cette planification



**Figure 2.2:** Diagramme de Gantt

Après que nous avons présenté la partie plannification du projet, nous passons ensuite à identifier les acteurs de notre solution.

## 2.3 Identification des acteurs

Les utilisateurs de l'application proposée sont ceux qui interagissent directement avec l'application.

Les acteurs de notre application sont :

- Analyste : C'est la mission de l'administrateur qui gère la plateforme. En effet, c'est lui qui peut administrer les fichiers sources, manier la base de données, gérer les données d'insertion et de mise à jour et enfin gérer les tableaux de boards .
- Décideur : Cet utilisateur aura la possibilité de consulter les tableaux de bord et les analyser.

Nous présentons dans ce qui suit les besoins de notre solution.

## 2.4 Spécification des besoins

L'étape suivante consiste à définir diverses exigences fonctionnelles et non-fonctionnelles de notre application, y compris les fonctions fournies par notre solution.

### 2.4.1 Besoins fonctionnels

Afin de répondre aux besoins de notre client, notre solution doit répondre aux besoins suivants :

- Développement d'un entrepôt de données pour stocker les données d'entrée / sortie.
- Calcul des mesures et des indicateurs de performances (Key Performance Indicators KPI).

- Visualisations des données dans des rapports graphiques et des tableaux croisés dynamiques pour le suivi du trafic de vente .
- Développement d'un module de web scraping et d'analyse sentimentale pour savoir les interactions des consommateurs .
- Utilisation d'un ensembles d'algorithmes de data mining pour prédire la vente d'un produit déterminé et le chiffre d'affaire, segmenter les clients et analyser les ventes.

#### 2.4.2 Besoins non fonctionnels

Les exigences non fonctionnelles décrivent toutes les contraintes techniques, ergonomiques et esthétiques que le système doit subir pour atteindre et fonctionner normalement. Concernant notre application, nous avons identifié les exigences suivantes :

- **Convivialité de l'interface graphique** : L'application doit fournir une interface simple et conviviale pour tout type d'utilisateur, car elle peut assurer le premier contact entre l'utilisateur et l'application .
- **Vitesse de traitement** : En effet, compte tenu du volume de transactions quotidien important, il est même nécessaire de rapprocher le plus possible le temps d'exécution du traitement en temps réel.
- **Une solution ouverte et extensible** : l'application peut être améliorée en ajoutant d'autres modules pour garantir flexibilité et évolutivité tout en garantissant l'intégrité des données.

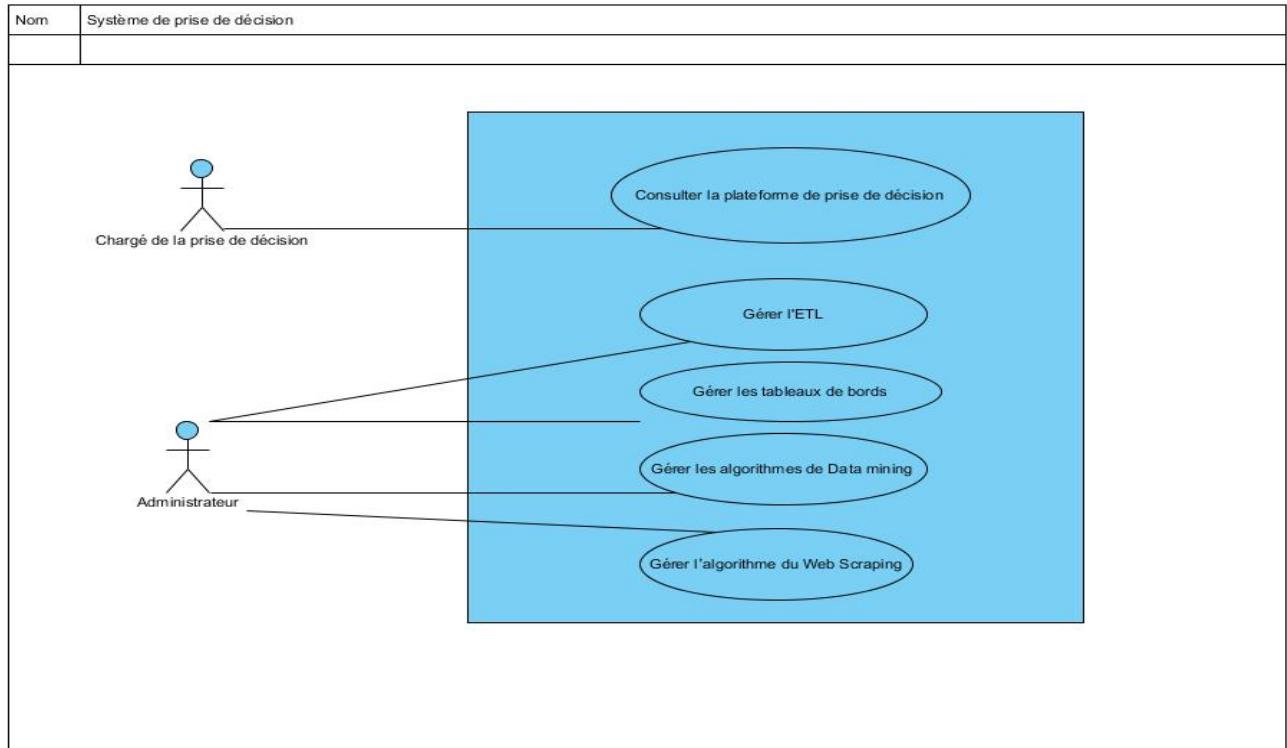
Dans la sections suivante, nous détaillons le cas d'utilisation du système.

### 2.5 Identification et description des cas d'utilisation du système

Dans un système de prise de décision, les acteurs jouent un rôle principal, puisqu'ils sont d'une part les bénéficiaires de l'utilisation du nouveau système, et les clients qu'on doit satisfaire d'une autre part.

Dans notre solution, nous avons le développeur qui gère l'entrepôt de données, les tableaux de bord, les algorithmes d'analyses et le Web Scraping et on a aussi l'utilisateur final variant selon le contexte qui dépend de l'organigramme de la société. Le bénéficiaire du produit final est un client qui utilise le produit GED ELISE. Quel que soit donc le rôle des utilisateurs, ils ont le même cas

d'utilisation, ainsi nous pouvons les fusionner en un seul acteur. Après avoir connus nos besoins fonctionnels de notre solution, nous allons illustrer ses fonctionnalités à travers un diagramme de cas d'utilisation général. En effet, il s'agit d'une description des différentes interactions entre le système et les acteurs qu'ils utilisent.



**Figure 2.3:** Diagramme de cas d'utilisation général

Suite au diagramme ci-dessus nous allons expliquer les cas d'utilisation « consulter les tableaux de bord », « gérer ETL » , « gérer les tableaux de bord » , « gérer les algorithmes des analyses » et « gérer l'algorithme du Web Scraping »

<b>Titre :</b> Consulter la palteforme de prise de décision
<b>Acteur :</b> Chargé de la prise de décision
<b>But :</b> La consultation des données sous forme de graphes et tableaux
<b>Précondition :</b>
<ul style="list-style-type: none"> <li>— S'authentifier</li> <li>— Les données sont générées</li> <li>— Les tableaux de bord sont créés</li> </ul>

**Scénario principale :**

- Lancer le tableau de bord
- Consulter le tableau de bord
- analyser à l'aide des filtres et des sélections
- Lancer les algorithmes d'analyses et de Web Scraping
- Consulter les résultats

**Tableau 2.1:** Description du cas d'utilisation : Consulter les tableaux de bord.

**Titre :** Gérer ETL

**Acteur :** Administrateur

**But :** Manipulation de l'entrepôt de données (SSIS)

**Précondition :**

- S'authentifier.
- Avoir accès aux différentes tables sources d'ELISE.

**Scénario principale :**

- Extraire les données
- Transformer les données
- Charger les données

**Tableau 2.2:** Description du cas d'utilisation : Gérer l'ETL

**Titre :** Gérer les tableaux de bords

**Acteur :** Administrateur

**But :** Manipulation des données sous POWER BI et SSAS

<b>Précondition :</b>
— S'authentifier
— Avoir accès aux données
<b>Scénario principale :</b>
— Choisir les graphes à utiliser
— Calculer les KPI
— Calculer les Mesures
— Glisser les données
— Enregistrer les tableaux de bord

**Tableau 2.3:** Description du cas d'utilisation : gérer les algorithmes des analyses.

<b>Titre :</b> Gérer les algorithmes de Data mining
<b>Acteur :</b> Administrateur
<b>But :</b> Manipulation et analyse de données sous Python
<b>Précondition :</b>
— S'authentifier
— Avoir accès aux données
<b>Scénario principale :</b>
— Choisir la méthode à utiliser
— Choisir les données à utiliser
— Développement du l'algorithme
— Interprétation des résultats

**Tableau 2.4:** Description du cas d'utilisation : Gérer les algorithmes de data mining.

<b>Titre :</b> Gérer l'algorithme du Web Scraping
<b>Acteur :</b> Administrateur
<b>But :</b> Extraction et analyse de données depuis le web
<b>Précondition :</b> <ul style="list-style-type: none"><li>— S'authentifier</li><li>— Avoir une connexion internet</li></ul>
<b>Scénario principale :</b> <ul style="list-style-type: none"><li>— Choisir la page facebook</li><li>— Développement du l'algorithme</li><li>— Extraction du données</li><li>— Nettoyer les données</li><li>— Elaboration du dictionnaire des mots</li><li>— Développement du l'algorithme d'analyse sentimentale</li><li>— Interpréter des résultats</li><li>— Enregistrer l'algorithme</li></ul>

**Tableau 2.5:** Description du cas d'utilisation : Gérer l'algorithme du Web Scraping.

## 2.6 Conclusion

Tout au long de ce chapitre, nous avons présenté la branche fonctionnelle de la méthodologie choisie. Nous avons, également, défini les intervenants dans notre projet pour mieux définir les différents cas d'utilisations.

Ayant une vision claire sur le système qui sera mis en œuvre, nous poursuivons par une analyse de la branche technique de notre méthodologie.

## ENVIRONNEMENT TECHNIQUE

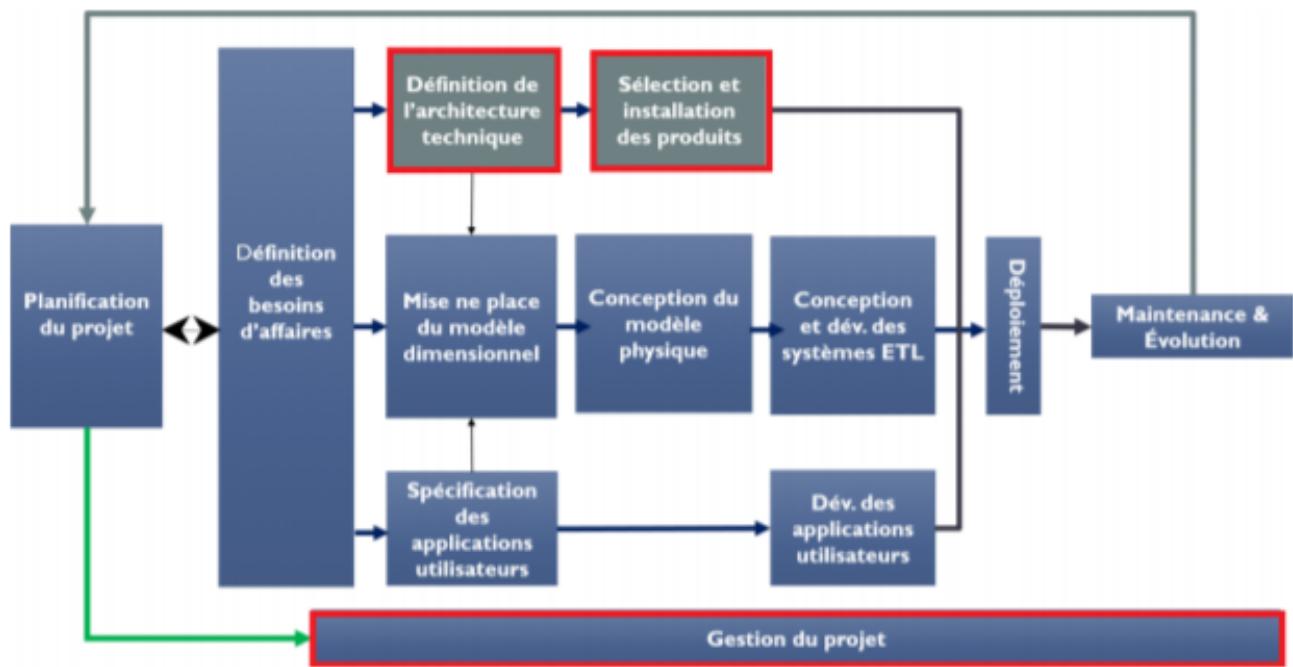
---

### Plan

1	Introduction . . . . .	23
2	Architecture technique du système . . . . .	23
3	Environnement technique du travail . . . . .	25
4	Conclusion . . . . .	28

### 3.1 Introduction

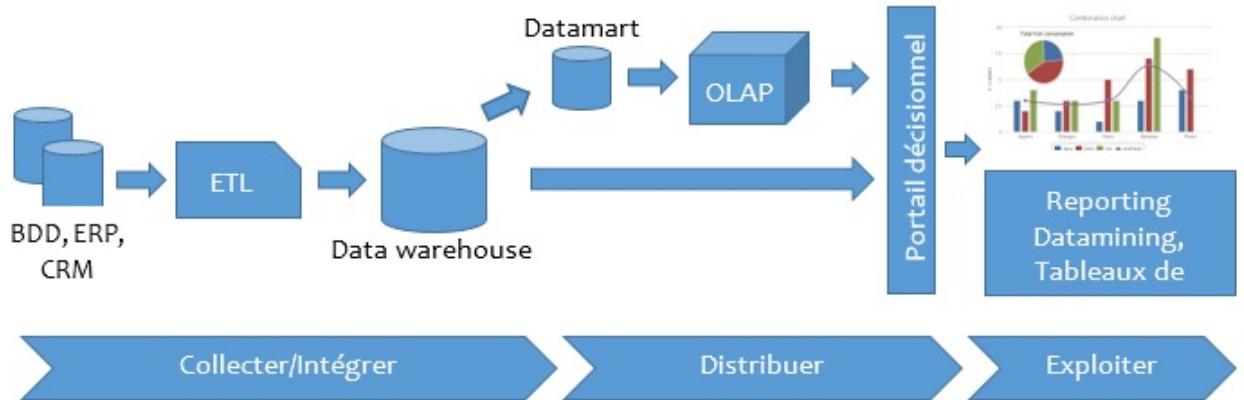
Dans ce chapitre, nous présentons les branches techniques du cycle de vie de Rálph Kimball. Nous expliquerons les différentes techniques utilisées. Commençons par les principaux outils et contraintes techniques. Nous exprimerons les différentes tâches de la phase de conception et leur positionnement dans le cycle de vie multidimensionnel illustré dans la figure ci-dessous. Ensuite, nous dévoilerons l'architecture du système.



**Figure 3.1:** Branche technique du cycle de vie de Ralph Kimball

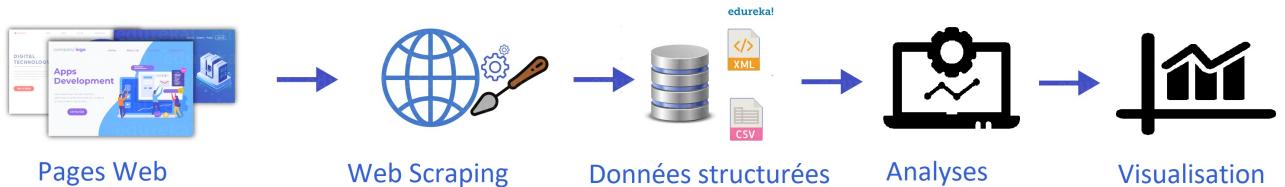
### 3.2 Architecture technique du système

Le système décisionnel de base comprend les étapes suivantes : La collection, la modélistion et la restitution des données. Comme illustré dans la figure 3.2, l'architecture technique d'un système de prise de décision.



**Figure 3.2:** Architecture technique d'un système de prise de décision

La figure 3.3 représent l'architecture technique de web scraping.



**Figure 3.3:** Architecture technique du web scraping

Parmi les données extraites de la base de données de production, nous allons d'abord nous assurer qu'elles sont stockées dans une structure intermédiaire de préparation et de sauvegarde appelée source de données opérationnelle pour assurer l'intégration des données dans le système information. De plus, une seconde structure de stockage sera mise en place pour s'assurer que le second emplacement est la sauvegarde des données converties et nettoyées et la construction de notre modèle multidimensionnel. Par la suite, concernant la restitution de ces données, nous utiliserons POWER BI pour générer un tableau de bord complet pour faciliter la prise de décision et construire un système de prise de décision puissant, analyser le comportement des consommateurs afin d'extraire les données depuis le web (Web scraping), et analyser finalement le trafic de vente de notre client avec Python afin d'assurer un système décisionnel complet.

Après que nous avons présenté l'architecture technique de notre projet, nous passons à décrire l'environnement technique du travail.

### 3.3 Environnement technique du travail

Avant de se lancer dans le projet et de planifier l'architecture technique, la sélection de l'environnement technique des travaux semble être une étape indispensable pour assurer le bon déroulement du projet. Voici l'environnement logiciel sur lequel notre projet est implémenté :

- **SQL Server [3]** : pour le stockge de données
- **MSBI Microsoft business intelligence [4]** : pour gérer l'ETL
- **POWER BI[5]** : pour le reporting
- **Python [6]** : pour la partie Data mining et Web scraping
- **VueJs [7]** : pour le développement de la plateforme web

Nous décrivons dans ce qui suit les outils de notre projet.

#### 3.3.1 SQL Server

SQL Server Management Studio (SSMS) est une application logicielle lancée pour la première fois avec Microsoft SQL Server 2005 et utilisée pour la configuration, l gestion et l administration de tous les composants de Microsoft SQL Server. L'outil comprend à la fois des éditeurs des scripts et des outils graphiques qui fonctionnent avec les objets et les fonctionnalités du serveur. Une caractéristique centrale de SSMS est l'explorateur d'objets, qui permet à l'utilisateur de parcourir, de sélectionner et d'agir sur l'un des objets du serveur. Il a également livré une édition Express séparée qui pouvait être téléchargée gratuitement, mais les versions récentes de SSMS sont entièrement cibles de se connecter et de gérer n'importe quelle instance de SQL Server Express. Microsoft a également intégré la rétrocompatibilité pour les anciennes versions de SQL Server, permettant ainsi à une version plus récente de SSMS de se connecter à des versions plus anciennes d'instances SQL Server.



Figure 3.4: Logo Microsoft SQL server

### 3.3.2 MSBI

MSBI signifie «Microsoft Business Intelligence». Il consiste en des outils permettant de fournir des solutions optimisées pour les requêtes de Business Intelligence et de Data Mining. MSBI utilise Visul studio vec SQL Server, ce qui permet aux utilisateurs d'avoir accès à des informations précises et à jour pour une décision de qualité supérieure. MSBI permet aux utilisateurs de découvrir, d'analyser et de visualiser des données avec l puissante intelligence d'affaires libre-service d'Excel. Il permet également la collaboration et le partage de rapports et de données avec SharePoint.

- Caractéristiques de MSBI : Fournit une «version unique de la vérité» pour prendre des décisions efficaces. Élimine ou réduit les "décisions instinctives". Fournit des réponses rapides et opportunes à l'entreprise, le rendant plus réactif aux tendances commerciales dynamiques. Minimiser le travail manuel et banal sujet aux erreurs. Support robuste pour l'analyse avancée. Prise en charge des données historiques. Prise en charge des données résumées.
- Architecture MSBI

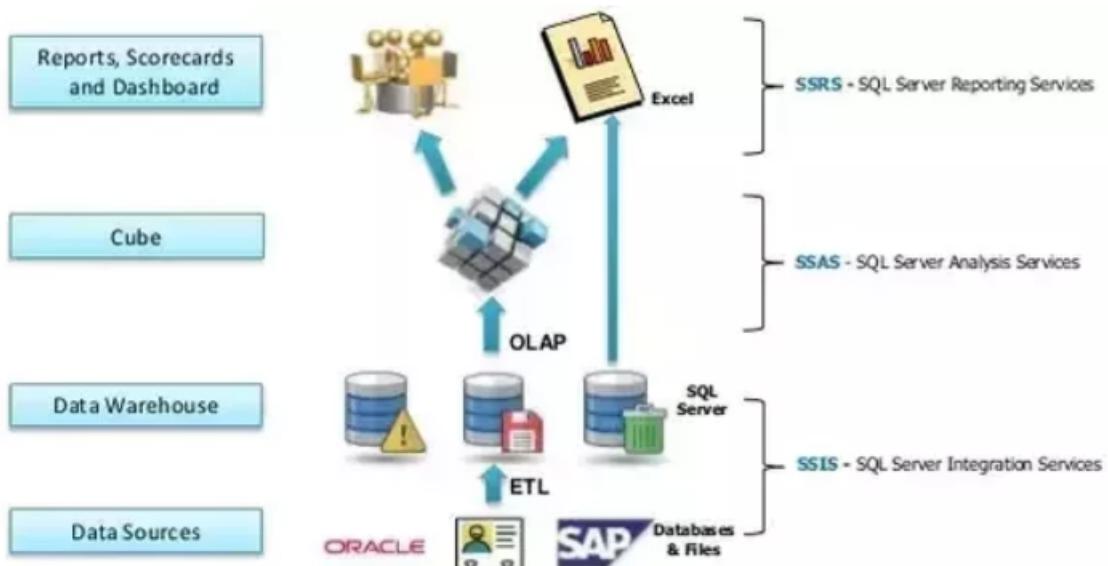


Figure 3.5: Architecture MSBI [8]

MSBI est divisé en 3 catégories :

- SQL Server Intégration Services (SSIS) - Outil d'intégration.
- Services analytiques SQL Server (SSAS) - Outil d'analyse.
- SQL Server Reporting Services (SSRS) - Outil de génération de rapports.

Dans notre projet, nous travaillons sur la partie intégration (SSIS) et la partie d'analyse (SSAS).

### 3.3.3 POWER BI

Power BI est un ensemble de services logiciels, d'applications et de connecteurs qui fonctionnent ensemble pour transformer vos sources de données non liées en informations cohérentes, immersives visuellement et interactives. Vos données peuvent être une feuille de calcul Excel ou un ensemble d'entreposés de données hybrides basés sur le Cloud et sur site. Power BI vous permet de vous connecter facilement à vos sources de données, de visualiser et de découvrir ce qui est important, et de le partager avec qui vous voulez.



**Figure 3.6:** Architecture POWER BI

Power BI peut être simple et rapide - capable de créer des informations rapides à partir d'un tableau Excel ou d'une base de données locale. Mais Power BI est également robuste ce qui le rend adéquat pour le contexte d'entreprise et prêt pour une modélisation poussée et des analyses en temps réel, ainsi que pour un développement personnalisé. Cela peut donc être votre rapport personnel et votre outil de visualisation. Il peut également servir de moteur d'analyse et de décision pour les projets, divisions ou entreprises du groupe.

### 3.3.4 Python

Cette fois, nous ne parlons pas de logiciel, mais de langage de programmation. Python est, parmi les langages de programmation utilisés pour l'analyse prédictive, l'un des plus facile à prendre en main, et bénéficie d'une large communauté de support en ligne. Les possibilités sont ici infinies et Python autorise une analyse prédictive à la pointe de la technologie, offrant ainsi un niveau de qualité bien supérieur à ce qui est habituellement nécessaire dans le monde de

l'entreprise. Vous disposez également de nombreuses bibliothèques proposant des méthodes prêtées à l'emploi. Python est entièrement gratuit tout comme le logiciel Anaconda qui permet d'utiliser une interface plus agréable pour s'y expérimenter, bien que les difficultés inhérentes à la pratique de la programmation demeurent. Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions.



**Figure 3.7:** Python

### 3.3.5 VueJs

Vue.js est un framework JavaScript open source pour la création d'interfaces utilisateur et d'applications Web à page unique. Vue a été créé par Evan You et maintenu par lui et d'autres membres actifs de l'équipe principale travaillant sur le projet et son écosystème.



**Figure 3.8:** VueJS

## 3.4 Conclusion

Dans ce chapitre, nous avons présenté l'architecture technique de notre solution tout en expliquant les différentes étapes suivies. Par la suite, nous définissons l'environnement du travail et les technologies adoptées pour assurer la réalisation du projet.

# INTÉGRATION DES DONNÉES

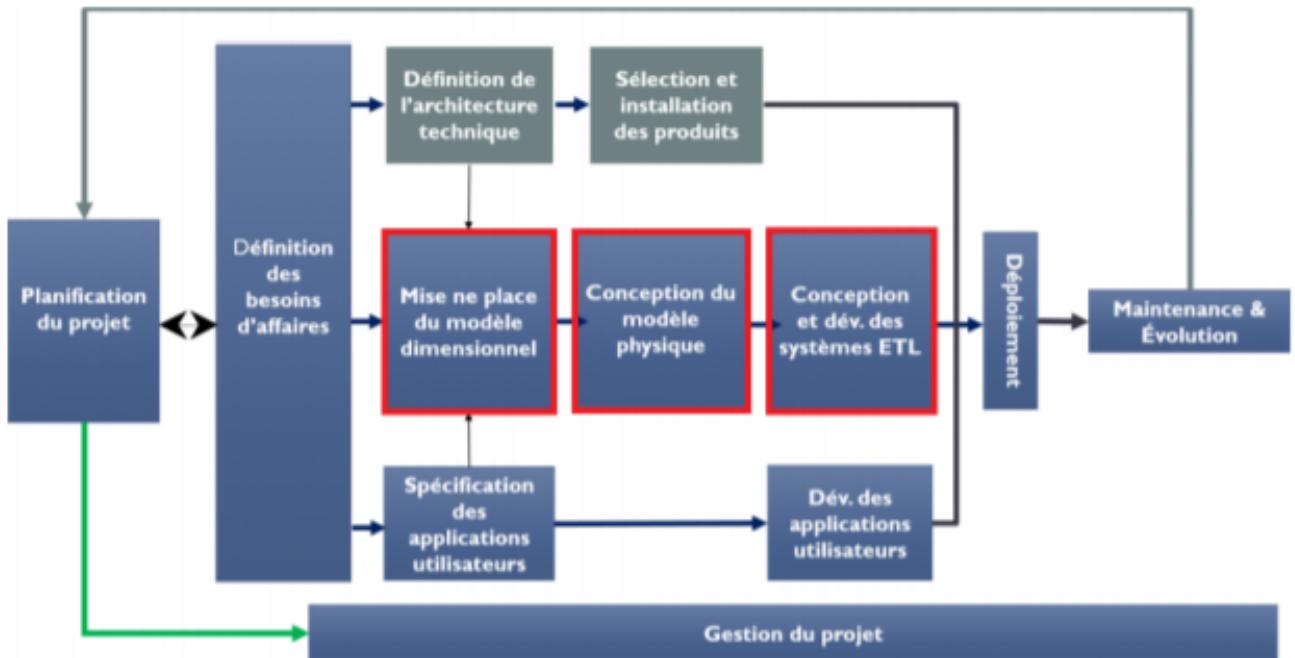
---

## Plan

1	Modélisation multidimensionnelle . . . . .	30
2	Approche utilisée . . . . .	31
3	Entrepôt de données . . . . .	32
4	Développement et préparation des données . . . . .	39
5	Conclusion . . . . .	44

## Introduction

Dans ce chapitre, nous commencerons la partie préparation des données, dans laquelle Correspond à la branche intermédiaire de la méthodologie de Ralph Kimball. La figure suivante illustre l'étape actuelle, la réalisation du modèle dimensionnel, Conception du modèle physique et conception de la zone de préparation des données.



**Figure 4.1:** Branche de modélisation du cycle de vie de Ralph Kimball

Nous commençons par la modélisation multidimensionnelle.

### 4.1 Modélisation multidimensionnelle

Il est nécessaire de concevoir un modèle dimensionnel qui réponde aux exigences du client. La modélisation multidimensionnelle est une approche de conception des données dans un format standardisé et intuitif pour garantir une accessibilité élevée et efficace. Les caractéristiques du modèle multidimensionnel sont :

- Performance d'adaptation aux changements et aux temps d'exécution des requêtes .
- Modèle évolutif qui peut être facilement modifié : ajouter de nouveaux faits, dimensions ou modification du schéma .
- Structure normalisée et prévisible .

- Réduction du nombre de tables et de jointures .

Nous présentons dans ce qui suit l'approche utilisée.

## 4.2 Approche utilisée

Afin de concevoir notre entrepôt de données, nous sommes confrontés à deux méthodes : L'approche top-down de Bill Inmon et l'approche bottom-up de Ralph Kimball. Voici un aperçu des deux méthodes, et le choix de la méthode la plus appropriée pour notre projet.

### 4.2.1 Approche de Ralph Kimball (ou approche Bottom-up)

Ralph Kimball choisit de créer un magasin de données qui répondra aux concepts définis clairement par les départements ou succursales de l'entreprise. Plus tard ces magasins Les données seront fusionnées pour former un entrepôt de données.

Le schéma de la méthode est illustré dans la figure 4.2 :

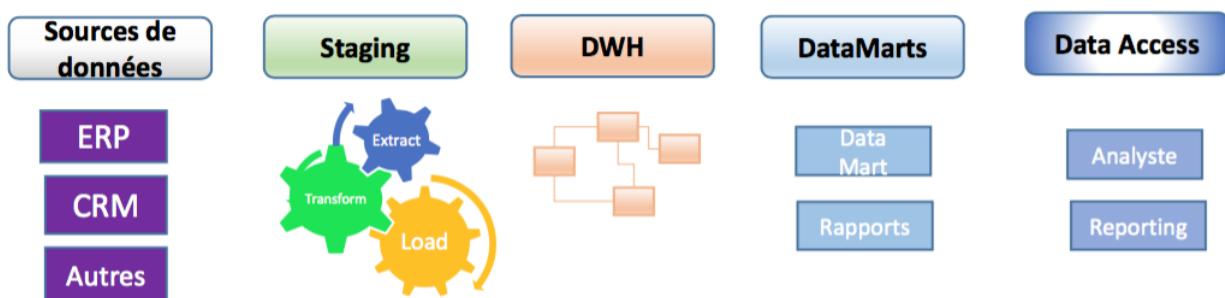


Figure 4.2: L'approche de Kimball [9]

### 4.2.2 Approche Bill Inmon

La méthode d'Inmon est généralement le contraire de Ralph Kimball se caractérise par un "top-down" Du point de vue d'Inmon, "Nous ne pouvons pas entrer dans la phase de compensation et d'analyse Seulement si et seulement si l'entrepôt de données est prêt et terminé. De plus, l'entrepôt sera ensuite subdivisé en plusieurs magasins de données.

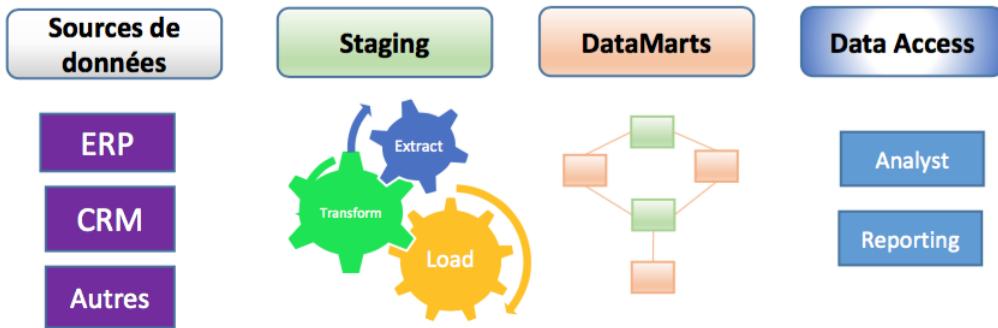


Figure 4.3: L'approche d'Inmon [10]

Le tableau suivant résume ces deux méthodes :

	Ralph Kimball	Bill Inmon
Processus	Bottom-Up	Top-Down
Organisation	Data marts	Data warehouse
Schématisation	Etoile	Flocon

Tableau 4.1: Comparaison des approches

Par conséquent, nous choisirons la méthode Bill Inmon car elle est plus adaptable pour l'architecture de notre système et répondre à nos besoins.

### 4.3 Entrepôt de données

L'entrepôt de données est une base de données relationnelle conçue pour les requêtes et pour l'analyse de données et de prise de décision et l'informatique décisionnelle pour le traitement des transactions ou à d'autres fins traditionnelles de base de données.

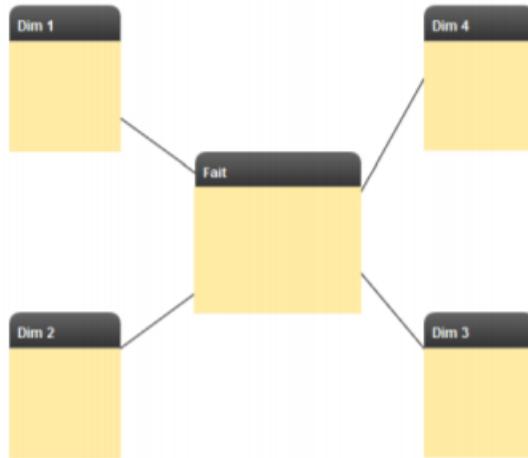
Les informations stockées dans l'entrepôt de données sont sauvegardées et fournissent un aperçu sur les différentes transactions qui se sont produites au cours du temps. Les données redondantes sont généralement incluses dans l'entrepôt de données pour fournir aux utilisateurs plusieurs vues des informations pertinentes. C'est Pourquoi les données stockées dans l'entrepôt sont généralement agrégées pour permettre aux utilisateurs d'y accéder plus facilement.

En plus de la base de données, l'environnement de l'entrepôt de données comprend également des outils d'extraction, Transfert, conversion et chargement de données (Extract, Transform and Load ETL). Il y a aussi un moteur Traitement analytique en ligne (**OnLine Analytical Processing OLAP**), outils d'analyse client et autres applications pour gérer le traitement des données collectées.

Dans un entrepôt de données, les données sont en générale redondantes et non standardisées Il n'y a absolument aucune raison de modéliser sous la troisième forme normale, il est donc possible de promouvoir à utiliser lors de l'analyse des données et améliorer les performances. Trois types de schémas , On rencontre souvent des schémas en étoiles, des schémas de flocon de neige et des schémas de constellation fait.

#### 4.3.1 Modèle en étoile

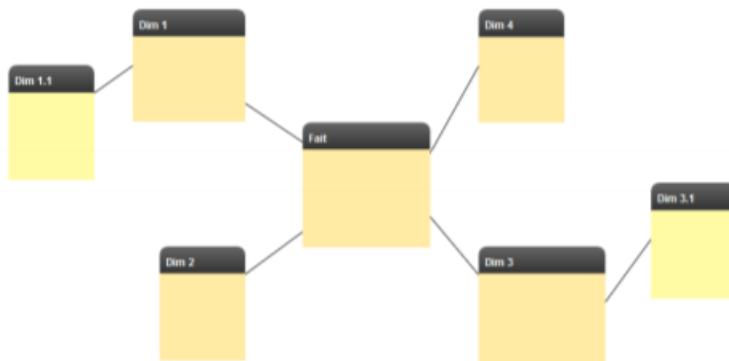
Le nom du modèle de données en étoile vient de sa forme. Il se compose d'une table de faits centrale et d'un ensemble de tables de dimensions, Comme le montre la figure 4.4



**Figure 4.4:** Modèle en étoile

#### 4.3.2 Modèle en flocon

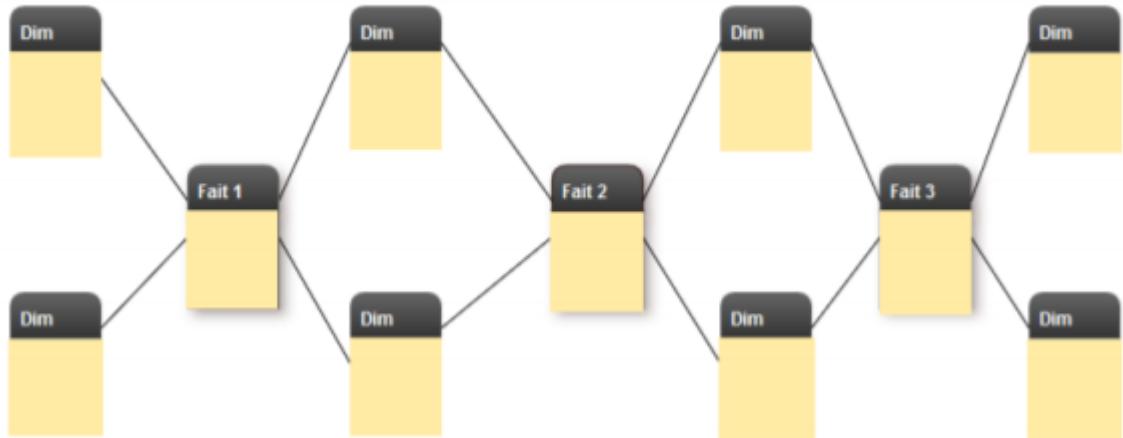
Le modèle de données en flocon est illustré à la figure 4.5, est une variante du modèle en étoile, où les dimensions sont normalisées, donc par conséquent, la hiérarchie des dimensions est affichée explicitement .



**Figure 4.5:** Modèle en flocon

#### 4.3.3 Modèle en constellation

Généralement connu sous le nom de flocon de faits, le modèle constellation se compose de plusieurs tables de faits qui peuvent partager ou non une dimension commune.



**Figure 4.6:** Modèle en constellation

Enfin, nous choisissons le modèle de constellation car il peut répondre aux besoins requis. Il nous permet de créer un modèle d'entrepôt de données contenant plusieurs tables de faits et leurs dimensions respectives, qui créent une fusion de modèles en flocon, dimensions communes entre plusieurs tables de faits. Cela nous permet de réduire l'espace besoin de stockage.

#### 4.3.4 Détermination des tables de dimensions

L'identification des dimensions se présente sur l'axe d'analyse nécessaire pour une capture correcte de dans notre projet sera l'objectif que nous proposerons dans la prochaine partie du rapport

où nous allons présenter chaque dimension ainsi ses attributs respectifs et une description, comme indique le tableau ci-dessous.

<b>Dimensions</b>	<b>Attributs</b>	<b>Descriptions</b>
Dim-Activite	ACTIVITE-ID, ACTIVITE-DESC	Cette dimension correspond aux différentes activités
Dim-Accord	Accord-id, Accord-code, Accord-description	Cette dimension correspond aux différents Accords
Dim-cdeType	CdeType-id, CdeType-code, CdeType-description	Cette dimension correspond aux différents types de commandes
Dim-Client	CLT-ID, CLT-DESC, TYPE-ID, SECTEUR-ID, RESEAU-ID, RESEAU-DESC, SRESEAU1-ID, SRESEAU1-DESC, SRESEAU2-ID, SRESEAU2-DESC, SRESEAU3-ID, SRESEAU3-DESC, SRESEAU4-ID, SRESEAU4-DESC	Cette dimension correspond aux différents informations des clients
Dim-date	DATE-ID, CHDATE-ID, CHDATE-DESC, CH-DATE-JOUR, CHMOIS-ID, CHMOIS-DESC, CHSEMAINE-ID, CHSEMAINE-DESC, CHSEMAINE-SHORT-DESC, CHSEMAINE-LONG-DESC, CHMOIS-SHORT-DESC, CHMOIS-LONG-DESC, CHMOIS-NUM, CHANNEE-ID, CHJO, CHJOB, CHJOC, CHJOE, CHJOS, CHIS-RAM, CHJOUR-RAM	Cette dimension correspond à l'axe temporel en terme de date
Dim-Doctype	DocType-id, DocType-code, DocType-description	Cette dimension correspond aux types des documents

Dim-Magasin	Magasin-id, Magasin-description	Magasin-code, Magasin-description	Cette dimension correspond aux différents magasins
Dim-Nature	Nature-id, Nature-description	Nature-code, Nature-description	Cette dimension correspond aux différents natures de ventes
Dim-produit	ID, CODE, DESC, UMSTD-ID, UMACH-ID, MARQUE-ID, FAMILLE-ID		Cette dimension correspond aux différents produits
Dim-Remise	Remise-id, Remise-description, REMISE-SEG1-ID, REMISE-SEG1-DESC, REMISE-SEG2-ID, REMISE-SEG2-DESC	Remise-code,	Cette dimension correspond aux différents remises
Dim-societe	Societe-id, Societe-description	Societe-code, Societe-description	Cette dimension correspond aux différents sociétés
Dim-Source	Source-id, Source-description	Source-code, Source-description	Cette dimension correspond aux différents sources
Dim-Taxes	Taxe-id, Taxe-Taux, Taxe-Type, Taxe-Taxable, Taxe-Valeur		Cette dimension correspond aux différents taxes
Famille		FAMILLE-ID, FAMILLE-DESC	Cette dimension correspond aux différents familles des produits
Marque	MARQUE-ID, MARQUE-DESC, ACTIVITE-ID		Cette dimension correspond aux différents marques des produits
Region		REGION-ID, REGION-DESC	Cette dimension correspond aux différents régions des clients
Sous-region		SOUS-REGION-ID, SOUS-REGION-DESC, REGION-ID	Cette dimension correspond aux différents sous régions des clients
Zone	ZONE-ID, SOUS-REGION-ID	ZONE-DESC,	Cette dimension correspond aux différents zones des clients

Secteur	SECTEUR-ID, SECTEUR-DESC, ZONE-ID	Cette dimension correspond aux différents secteurs des clients
Type-client	ID, DESC	Cette dimension correspond aux différents types des clients

**Tableau 4.2:** Dimensions et leur descriptions

#### 4.3.5 Identification des tables de faits

La table de faits est la table centrale du modèle dimensionnel et contient des clés étrangères liées aux dimensions et aux valeurs qu'on désire calculer. Après qu'on a identifié toutes les dimensions, nous venons de présenter les tables des faits en détail ci-dessus

- **Ventes-Commandes :** C'est une table de faits de transactionnelle qui représente l'historique des commandes et qui contient les mesures suivantes :
  - QUANTITE-COMM
  - TONNAGE-COMM
  - QUANTITE-LOG
  - TONNAGE-LOG
- **Ventes-Livraison :** C'est une table de faits de transactionnelle qui représente l'historique des Livraisons et qui contient les mesures suivantes :
  - QUANTITE
  - TONNAGE
  - CAB
  - CAF
  - COUT-TOTAL
- **Ventes-Remises :** C'est une table de faits de transactionnelle qui représente l'historique des Remises et qui contient les mesures suivantes :
  - REMISE-UNITAIRE
  - REMISE-VALEUR

Avant de passer à la partie suivante pour décrire les différents indicateurs liés aux tables des faits, nous allons d'abord expliquer quelques concepts techniques comme indiqué ci-dessous :

- **Unité de stockage** : Nombre / Montant / Taux
- **Règle de calcul** : Complément de définition.

#### **4.3.5.1 Indicateurs de la fait :Ventes-Commandes**

Les indicateurs suivants appartiennent à la table de fait "Ventes-Commandes " :

- Unités par commande= produits vendus / nombre des commandes
- Nombre des commandes

#### **4.3.5.2 Indicateurs de la fait :Ventes-Livraisons**

Les indicateurs suivants appartiennent à la table de fait "Ventes-Livraison " :

- Chiffre d'affaire : La somme de la mesure CAF .
- Coût total : La somme de la mesure COUT-TOTAL .
- Perte et casse : La somme de la mesure CAF ou la mesure QUANTITE < 0 .
- Taux de conversion : (nombre de clients / nombre d'achats) \* 100 .
- Marge bénéficiaire brute : ((La somme de la mesure CAB - la somme de la mesure COUT-TOTAL) / la somme de la mesure CAB) \* 100 .
- Croissance d'une année à l'autre = Total des ventes de l'année en cours - Total des ventes de l'année précédente .
- Taux de vente : La somme de la mesure CAF ou la mesure QUANTITE > 0 .
- Nombre moyen de vente= nombre de vente au cours de la période / période .

#### **4.3.5.3 Indicateurs de la fait :Ventes-Remises**

Les indicateurs suivants appartiennent à la table de fait "Ventes-Remises " :

- Taux de vente par promotion : La somme de la mesure CAF pour chaque type de promotion ou la mesure QUANTITE > 0
- Taux de retour et casse par promotion : La somme de la mesure CAF pour chaque type de promotion ou la mesure QUANTITE < 0

#### **4.3.6 Conception du modèle**

Après avoir présenté les dimensions et les tables de faits composées de clés étrangères et de mesures, nous avons élaboré un modèle de données logique, cohérent et qui possède une haute

efficacité de calcul. L'image ci-dessous représente notre modèle.

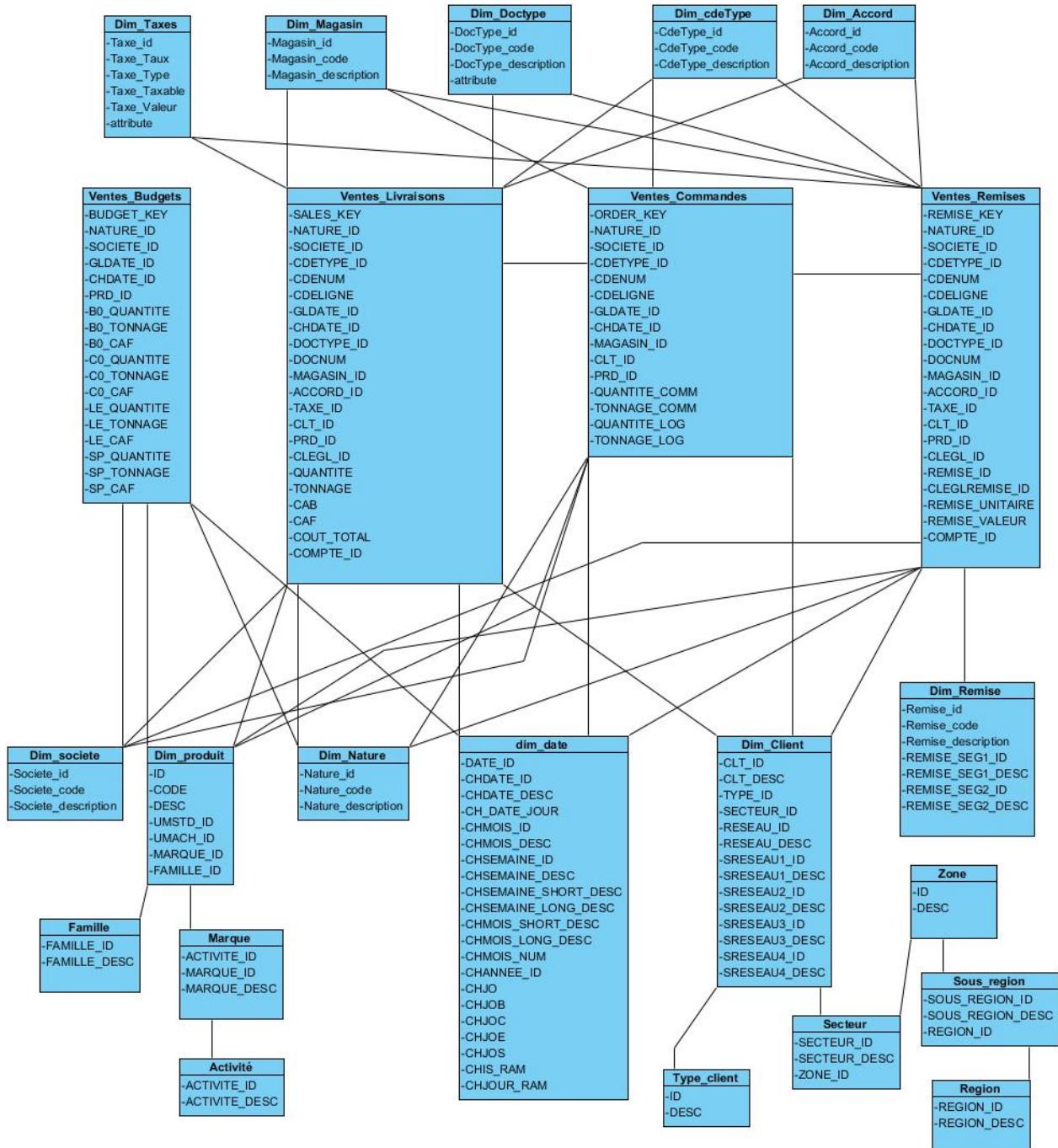


Figure 4.7: Conception de l'entrepôt des données

## 4.4 Développement et préparation des données

L'étape de préparation des données est une étape essentielle du processus de BI. En fait, cette étape passera par trois phases :

- Alimentation de la base de données intermédiaire.

- Extraction, transformation et chargement des données dans l'entrepôt de données .
- Développement des indicateurs et calcul des mesures .

#### 4.4.1 Etape d'alimentation de Data Staging Area

Avant de passer à la partie d'alimentation il faut tout d'abord extraire les données de la base de données de production dans une base donnée nommée **Staging Area**. Le Staging Area est une zone de stockage intermédiaire utilisée pour le traitement des données pendant le processus d'extraction, de transformation et de chargement (ETL). Les Staging Areas sont souvent de nature transitoire, leur contenu étant effacé avant d'exécuter un processus ETL ou immédiatement après l'achèvement réussi d'un processus ETL. Après que nous avons définis la Data Staging Area, nous allons ensuite démarrer la partie alimentation de la base de données Staging Area. Pour ce faire, nous suivrons les étapes suivantes :

- Premièrement, nous générerons une base de données Backup générées depuis la base de données de production, Et pour des raisons de confidentialité on ne peut pas présenter cette étape .
- Deuxièmement, nous restaurons la base de données Backup dans SQL SERVER qui est installé dans notre machine comme indique la figure 4.8 :

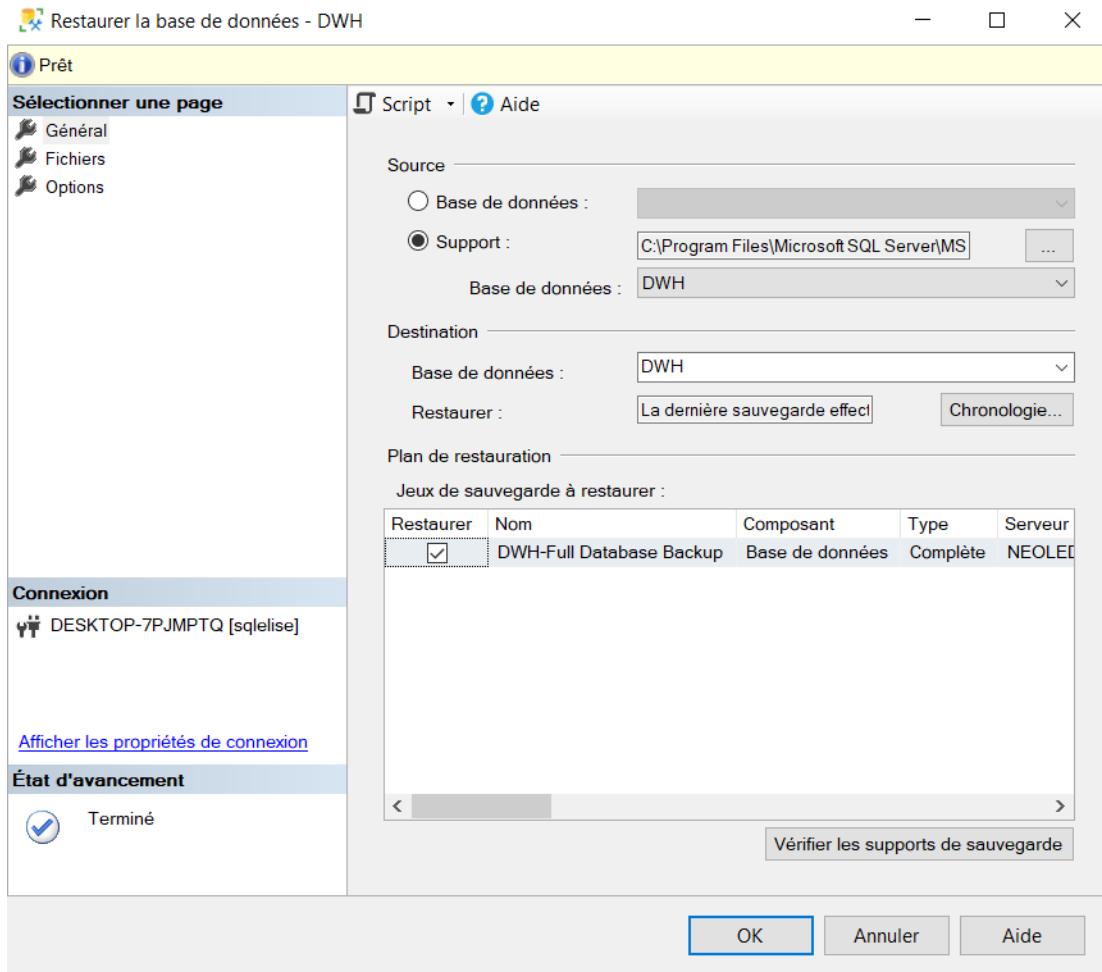


Figure 4.8: Restauration de la base de données Backup

Cette figure représente l'interface de restauration de la base de données Backup dans SSIS.

- Finalement, nous obtenons notre base de donnée que nous désirons la supprimer après la phase ETL (Extract, Transform and Load) et pour des raisons de confidentialité nous ne pouvons pas présenter sa structure .

#### 4.4.2 Etape de modélisation

L'étape la plus importante du processus de BI est l'ETL . Cette étape nous permet d'extraire les données de la base de données «Staging Area». Ensuite, nous apporterons les modifications nécessaires pour unifier et homogénéiser les données. Pour les rendre prêtes a exploiter dans l'étape suivante. De plus, les données doivent être intégrées pour obtenir de bons résultats. Le changement peut être : une transformation simple ou une transformation complexe, son traitement nécessite plusieurs étapes et plusieurs composants de traitement sous l'outil d'intégration de données **SSIS**.

Enfin, nous stockons les données traitées dans différentes tables de l'entrepôt de données.

#### 4.4.2.1 Extraction des données

Dans cette étape, nous allons extraire les tables de la base Staging area générée de l'étape précédente après avoir établi une connexion avec cette base en utilisant le composant "Source OLE DB" comme indique les figures ci-dessous.

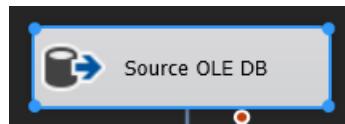


Figure 4.9: Composant Source OLE DB

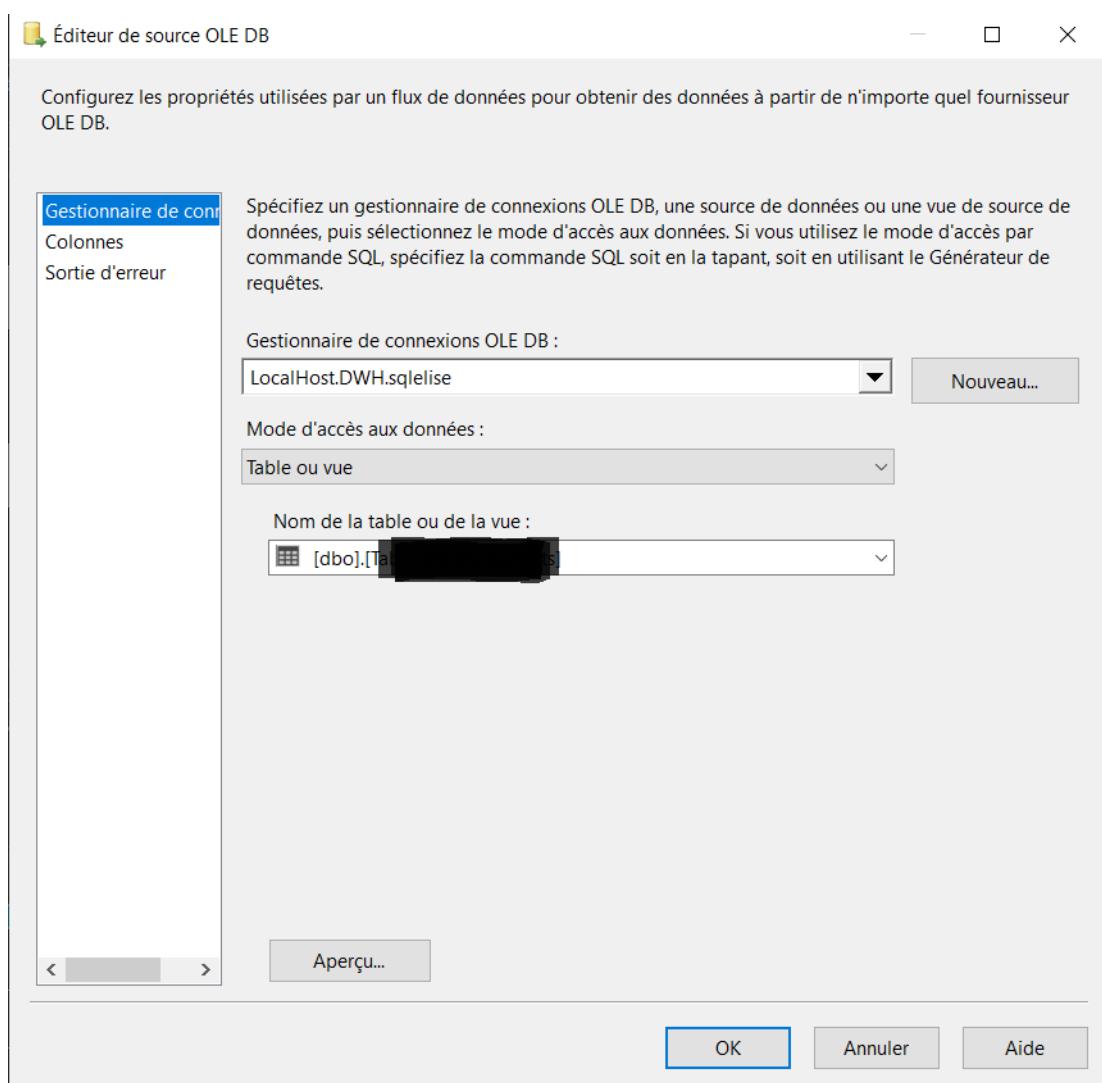


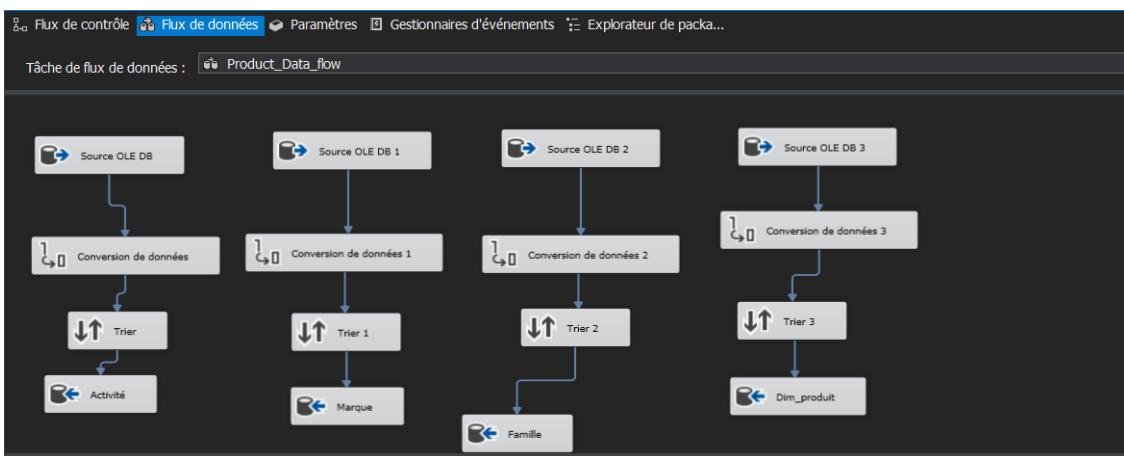
Figure 4.10: Connexion à une table

#### 4.4.2.2 Transformation des données

Après l'extraction des données, nous nous occuperons de l'étape de conversion des données. Cette étape nous permettra d'effectuer diverses opérations de traitement nécessaires pour fournir des données cohérentes après transformation pour répondre à divers besoins.

- Alimentation des tables de dimensions :

Après avoir transformé les données, nous commencerons par le chargement des dimensions mentionnées dans la section précédente pour concrétiser le concept théorique. L'image ci-dessous montre un exemple d'alimentation de "Dim-produits".



**Figure 4.11:** Flux de données d'alimentation de Dim-produit

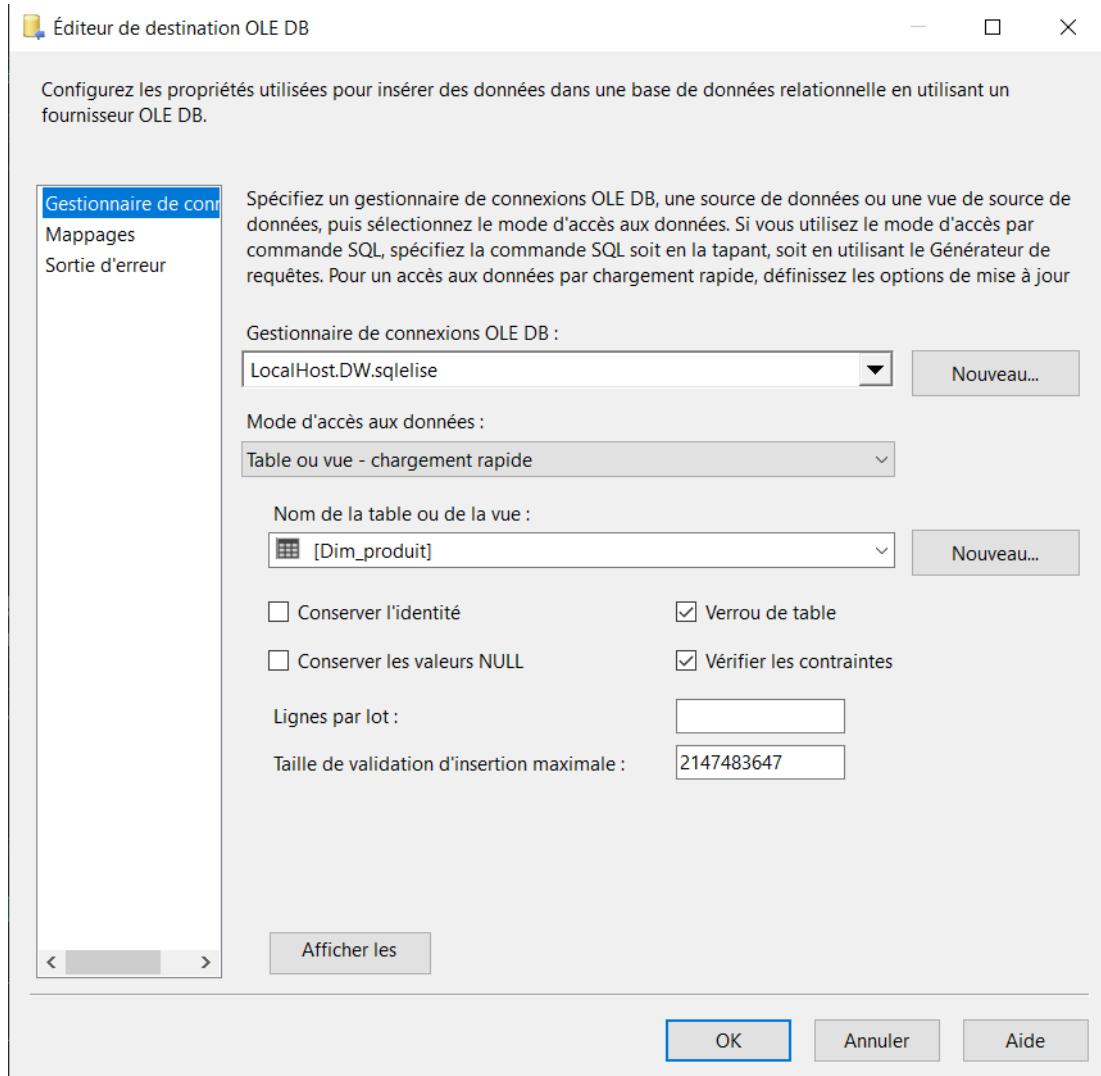
Le processus consiste à récupérer les données contenant les informations des produits de la base Satging Area . Tout d'abord, après avoir restauré les données via le composant "Source OLE DB", nous allons servir le composant "Conversion de données" qui sert à convertir le type de données . Ensuite, nous allons servir le composant "Trier" qui sert à trier les données. En fait, il s'agit du même processus effectué pour la transformation nécessaire des autres dimensions.

#### 4.4.2.3 Chargement des données

Après avoir appliqué l'extraction et la transformation des données nécessaires, nous passerons à l'étape finale d'ETL, qui est le stockage des données. Cette étape se déroule comme indiqué ci-dessous :

- Utilisez le composant "Destination OLE DB" pour créer une connexion à la base de données finale "DW" Comme indiqué dans la figure 4.12.

- Ecriture des données dans la table correspondante.



**Figure 4.12:** Destination OLE DB

## 4.5 Conclusion

Dans ce chapitre, nous avons présenté la partie modélisation, où les modifications nécessaires ont été apportées au traitement et homogénéiser les données afin de les fournir à notre entrepôt de données. Ce travail nous mènera à l'étape suivante qui est l'analyse de données et web scraping qui sera décrite lors du chapitre suivant.

# DATA MINING ET WEB SCRAPING

---

## Plan

<b>1</b>	<b>Introduction</b>	<b>46</b>
<b>2</b>	<b>Data Mining</b>	<b>46</b>
<b>3</b>	<b>Web scraping</b>	<b>61</b>
<b>4</b>	<b>Conclusion</b>	<b>64</b>

## 5.1 Introduction

La partie Business Intelligence est terminée. Cependant, à l'aide des tableaux de bord que nous avons produits, les décideurs ne pourront jamais prédire l'état de ventes et du chiffre d'affaire que ce soit à court ou à long terme, ne pourront **jamais** connaître comment leurs clients achètent les produits , le comportement de leurs consommateurs... Le but de ce chapitre est de prédire la vente des produits et le chiffre d'affaire, segmenter les clients, élaborer des règles d'association et avoir une analyse sentimentale sur leurs consommateurs . Dans notre cas les clients de l'entreprise sont les vendeurs de détails (hypermarchés, supermarchés...) et les vendeurs en gros. Pour cela, nous utiliserons Python pour développer des algorithmes pour appliquer des modèles de séries temporelles, segmentation, règles d'associations, web scraping et text Mining

## 5.2 Data Mining

La Data mining est un ensemble de méthodes et d'algorithmes utilisés pour prendre des décisions correctes et prédire différents changements dans l'environnement de l'entreprise.

### 5.2.1 Règles d'associations

#### 5.2.1.1 Apriori

Apriori est un algorithme utilisé pour identifier les ensembles d'éléments fréquents (dans notre cas, les paires d'éléments). Il le fait en utilisant une approche «Bottom-up», en identifiant d'abord les éléments individuels qui satisfont à un seuil d'occurrence minimum. Il étend ensuite l'ensemble d'éléments, en ajoutant un élément à la fois et en vérifiant si l'ensemble d'éléments résultant satisfait toujours au seuil spécifié. L'algorithme s'arrête lorsqu'il n'y a plus d'éléments à ajouter qui répondent à l'exigence d'occurrence minimale.

#### 5.2.1.2 Les règles d'association Mining

L'exploration de règles d'association trouve des associations et des relations intéressantes entre de grands ensembles d'éléments de données. Cette règle indique la fréquence à laquelle un ensemble d'éléments se produit dans une transaction. Un exemple typique est l'analyse basée sur le marché.

L'analyse basée sur le marché est l'une des techniques clés utilisées par les grandes relations

pour montrer les associations entre les éléments. Il permet aux détaillants d'identifier les relations entre les articles que les consommateurs achètent fréquemment ensemble.

Voici les trois indicateurs clés à prendre en compte lors de l'évaluation des règles d'association :

- **Support :**

Il s'agit du pourcentage de commandes contenant l'ensemble d'articles . Par exemple, il y a 5 commandes au total et A, B se produit dans 3 d'entre elles, donc : **support[A,B] = 3/5**

- **Confidence :**

Étant donné deux articles, A et B, la confiance mesure le pourcentage de fois où l'article B est acheté, étant donné que l'article A a été acheté. Ceci est exprimé comme : **confiance[A->B] = support[A,B] / support[A]**.

Les valeurs de confiance vont de 0 à 1, où 0 indique que B n'est jamais acheté lorsque A est acheté, et 1 indique que B est toujours acheté chaque fois que A est acheté. Notez que la mesure de confiance est directionnelle. Cela signifie que nous pouvons également calculer le pourcentage de fois où l'article A est acheté, étant donné que l'article B a été acheté : **confidence[B->A] = support[A,B] / support[B]**

- **Lift :**

Étant donné deux éléments, A et B, lift indique s'il existe une relation entre A et B, ou si les deux éléments surviennent ensemble dans les mêmes commandes simplement par hasard (c'est-à-dire au hasard). Contrairement à la métrique de confiance dont la valeur peut varier en fonction de la direction (par exemple : la confiance [A-> B] peut être différente de la confiance [B-> A]), le lift n'a pas de direction. Cela signifie que  $\text{lift}[A, B] = \text{lift}[B, A]$  : **lift[A,B] = lift[B,A] = support[A,B] / (support[A] \* support[B])**

En résumé, le lift peut prendre les valeurs suivantes :

- **lift = 1** n'implique aucune relation entre A et B. (c'est-à-dire : A et B se produisent ensemble uniquement par hasard)
- **lift > 1** implique qu'il existe une relation positive entre A et B. (c'est-à-dire : A et B se produisent plus souvent ensemble qu'au hasard)
- **lift < 1** implique qu'il existe une relation négative entre A et B. (c'est-à-dire : A et B se produisent moins souvent ensemble qu'au hasard)

Une fois que nous avons présenté l'algorithme à utiliser, nous passons maintenant à la préparation de données.

### 5.2.1.3 Préparation de données

Dans cette section, nous concentrerons sur la préparation des données disponibles et existantes dans l'entrepôt de données.

Pour effectuer cette partie, nous allons utiliser la bibliothèque de python "Pyodbc [11]" pour sélectionner un échantillon d'historique de ventes des produits à cause du grand volume de données...

Pour cela, nous connectons à l'entrepôt de données pour charger la table des fait "Ventes-Livraison" et appliquer les traitements nécessaires. Les variables extraites sont illustrés dans le tableau 5.1 .

Variable	Type	Description
CDENUM	Entier	Représente le numéro de commande
Clt	Entier	Représente l'id du client
PRD	Entier	Représente l'id du produit
Quantite	Nombre	Représente la quantité achetée
CAB	Nombre	Représente le montant brût
CAF	Nombre	Représente le montant facturé
cout-total	Nombre	Représente le coût total
CHDATE-ID	Date	Représente la date de l'achat

Tableau 5.1: Variables utilisées dans les règles d'associations

Ensuite, nous allons convertir les données pour respecter les normes de la fonction de règles d'associations .

```
sql_query = sql_query.set_index('CDENUM')[['PRD']].rename('item_id')
display(sql_query.head(10))
type(sql_query)|
```

Figure 5.1: Convertir les données de ventes

### 5.2.1.4 Fonctions de règles d'association

L'une des caractéristiques de Python est le grand nombre de fonctions et de processus qui peuvent être utilisés. Dans notre cas, nous allons développer un modèle qui nous permet d'élaborer une liste des produit fréquents et qui repekte les normes de règles d'associations. Après qu'on a développé ce modèle , nous passons à l'étape suivante qui est l'exécution de ce modèle et

l'interprétation des résultats.

#### 5.2.1.5 Exécution du modèle

Dans cette partie, nous allons exécuter le modèle qu'on a développé comme indique ci-dessous :

```
%time  
rules = association_rules(sql_query, 0.01)  
  
Starting order_item: 1000000  
Items with support >= 0.01: 197  
Remaining order_item: 999957  
Remaining orders with 2+ items: 65632  
Remaining order_item: 983880  
Item pairs: 24861  
Item pairs with support >= 0.01: 18442  
  
Wall time: 3.98 s
```

**Figure 5.2:** Fonction de règles d'associations

Après l'exécution du modèles précédent, nous avons constaté la liste produits associés d'après les règles d'associations comme indiqué dans la figure 5.6 .

rules											
	item_A	item_B	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
8089	437	438	37	0.056375	57	0.086848	62	0.094466	0.649123	0.596774	6.871488
18076	413	413	12	0.018284	35	0.053328	35	0.053328	0.342857	0.342857	6.429257
8673	280	281	18	0.027426	27	0.041138	83	0.126463	0.666667	0.216867	5.271647
8102	439	440	32	0.048757	76	0.115797	98	0.149317	0.421053	0.326531	2.819850
18194	281	281	24	0.036568	83	0.126463	83	0.126463	0.289157	0.289157	2.286497
...	...	...	...	...	...	...	...	...	...	...	...
7701	66	205	8	0.012189	29112	44.356412	7753	11.812835	0.000275	0.001032	0.000023
14990	354	144	9	0.013713	18729	28.536385	15282	23.284373	0.000481	0.000589	0.000021
12921	354	14	8	0.012189	18729	28.536385	15420	23.494637	0.000427	0.000519	0.000018
14526	354	156	7	0.010666	18729	28.536385	20551	31.312470	0.000374	0.000341	0.000012
13715	354	12	7	0.010666	18729	28.536385	22360	34.068747	0.000374	0.000313	0.000011

**Figure 5.3:** Résultats des règles d'associations

Après que nous avons analysé les associations des produits, nous passons dans ce qui suit à l'analyse des clients.

### 5.2.2 segmentation des clients avec RFM

L'analyse RFM (Recency, Frequency et Monetary) est une technique de segmentation des clients qui utilise le comportement d'achat passé pour diviser les clients en groupes. RFM permet de diviser les clients en différentes catégories ou clusters pour identifier les clients les plus susceptibles de répondre aux promotions et également pour les futurs services de personnalisation. Ci-dessous, nous donnons les significations des trois critères.

- **RECENCY (R)** : Nombre de jours depuis le dernier achat .
- **FREQUENCY (F)** : Nombre total d'achats .
- **MONETARY VALUE (M)** : Total de l'argent dépensé par ce client.

Dans ce qui suit, on vas créer ces 3 attributs client pour chaque client.

#### 5.2.2.1 Préparation de données

Après avoir définit les objectifs de cette section, nous nous concentrerons sur la préparation des données disponibles et existantes dans l'entrepôt de données.

Pour effectuer cette partie, nous sélectionnerons un échantillon d'historique de ventes des produits à cause du grand volume de données... Pour cela, nous nous connectons à l'entrepôt de données pour charger la table des fait "Ventes-Livraison" sous Python et appliquer les traitements nécessaires.

Variable	Type	Description
CDENUM	Entier	Représente le numéro de commande
Clt	Entier	Représente l'id du client
PRD	Entier	Représente l'id du produit
Quantite	Nombre	Représente la quantité achetée
CAB	Nombre	Représente le montant brût
CAF	Nombre	Représente le montant facturé
cout-total	Nombre	Représente le coût total
CHDATE-ID	Date	Représente la date de l'achat
type	Chaine de caractères	Représente le type des clients

**Tableau 5.2:** Variables utilisées dans le RFM

Ensuite, nous allons nettoyer les données en supprimant les données de retour, le données de casse et la redondance des données comme indiqué dans les figures 5.4 et 5.5.

```
retail_uk = sql_query[sql_query['Quantite'] > 0]
retail_uk.shape
```

**Figure 5.4:** Suppression des données de retour et casse

```
print("Summary..")
print("Number of transactions: ", retail_uk['CDENUM'].nunique())
print("Number of products bought: ", retail_uk['PRD'].nunique())
print("Number of customers: ", retail_uk['Clt'].nunique())
print("Percentage of customers NA: ", round(retail_uk['clt'].isnull().sum() * 100 / len(sql_query), 2), "%")
```

Summary..  
Number of transactions: 49621  
Number of products bought: 183  
Number of customers: 501  
Percentage of customers NA: 0.0 %

**Figure 5.5:** Explorer les valeurs uniques de chaque attribut

### 5.2.2.2 Recency

Pour calculer la recency, nous devons choisir un point de date à partir duquel nous évaluons combien de jours a eu lieu le dernier achat du client.

```
retail_uk['CHDATE_ID'].max()
Timestamp('2015-02-21 00:00:00')
```

**Figure 5.6:** Dernière date d'achat

Pour l'échantillon sur lequel nous travaillons, la dernière date que est le 21/02/2015, nous allons donc l'utiliser comme référence pour passer au calcul du recency.

### 5.2.2.3 Frequency

La fréquence nous aide à savoir combien de fois un client a acheté chez nous. Pour ce faire, nous devons vérifier le nombre de factures enregistrées par le même client comme indique ci-dessous :

```
retail_uk_copy = retail_uk
retail_uk_copy.drop_duplicates(subset=['CDENUM', 'clt'], keep="first", inplace=True)
frequency_df = retail_uk_copy.groupby(by=['clt'], as_index=False)[['CDENUM']].count()
frequency_df.columns = ['clt', 'Frequency']
frequency_df.head()
```

**Figure 5.7:** Calcul du fréquence

### 5.2.2.4 Monetory

L'attribut monétaire répond à la question : combien d'argent le client a-t-il dépensé au fil du temps ? Pour ce faire, nous allons d'abord créer une nouvelle colonne de coût total pour avoir le prix total par facture.

```
retail_uk['TotalCost'] = retail_uk['cout_total']

monetary_df = retail_uk.groupby(by='clt',as_index=False).agg({'TotalCost': 'sum'})
monetary_df.columns = ['clt','Monetary']
monetary_df.head()
```

**Figure 5.8:** Calcul du monétaire

### 5.2.2.5 Crédation d'une table RFM

Cette étape consiste à fusionner les 3 tables Recency, Frequency et Monetary dans une seule table .

```
rfm_df = temp_df.merge(monetary_df,on='clt')
rfm_df.set_index('clt',inplace=True)
rfm_df.head()
```

**Figure 5.9:** Table RFM

### 5.2.2.6 Segmentation des clients avec modèle RFM

Le moyen le plus simple de créer des segments de clients à partir du modèle RFM consiste à utiliser des quartiles. Nous attribuons un score de 1 à 4 à Récence, Fréquence et Monétaire. La valeur 4 la meilleure / la plus élevée, et la valeur 1 la plus basse / la pire. Un score RFM final est calculé simplement en combinant des numéros de score RFM individuels.

**Remarque :** les quantiles offrent une meilleure granularité, au cas où l'entreprise en aurait besoin, mais il sera plus difficile de créer des segments car nous aurons un nombre élevé de combinaisons possibles. Donc, nous utiliserons des quartiles.

```

def RScore(x,p,d):
    if x <= d[p][0.25]:
        return 4
    elif x <= d[p][0.50]:
        return 3
    elif x <= d[p][0.75]:
        return 2
    else:
        return 1
def FMScore(x,p,d):
    if x <= d[p][0.25]:
        return 1
    elif x <= d[p][0.50]:
        return 2
    elif x <= d[p][0.75]:
        return 3
    else:
        return 4

```

**Figure 5.10:** Fonction de calcul de score RFM

Meilleur score de recency = 4 : achat le plus récent. Meilleur score de fréquence = 4 : la plupart des achats en quantité. Meilleur score monétaire = 4 : le plus dépensé.

Voyons dans la figure ci-dessous combien de clients avons-nous dans chaque segment.

```

print("Meilleurs clients: ",len(rfm_segmentation[rfm_segmentation['RFMScore']=='444']))
print('Clients fidèles: ',len(rfm_segmentation[rfm_segmentation['F_Quartile']==4]))
print("Gros dépensiers: ",len(rfm_segmentation[rfm_segmentation['M_Quartile']==4]))
print('Presque perdu: ', len(rfm_segmentation[rfm_segmentation['RFMScore']=='244']))
print('Clients perdus: ',len(rfm_segmentation[rfm_segmentation['RFMScore']=='144']))
print('Mauvais clients: ',len(rfm_segmentation[rfm_segmentation['RFMScore']=='111']))

```

```

Meilleurs clients: 63
Clients fidèles: 123
Gros dépensiers: 125
Presque perdu: 0
Clients perdus: 0
Mauvais clients: 32

```

**Figure 5.11:** Total des clients dans chaque segment

### 5.2.3 Série Temporelle

Une série temporelle ou une série chronologique est une série de valeurs numériques qui représentent le changement d'une quantité spécifique au fil du temps. Les séries temporelles permettent de prédire la valeur future sur la base des valeurs précédentes. les séries temporelle peuvent être utilisées pour prédire les tendances de l'économie, de la météo et de la vente... Les propriétés spécifiques des données de séries temporelles nécessitent des méthodes statistiques spécialisées sont

souvent nécessaires. Dans notre projet, nous avons utilisé les séries temporelles pour prédire la vente d'un produit déterminé et le chiffre d'affaire.

### 5.2.3.1 Prédiction de vente des produits et du chiffre d'affaire

Dans ce qui suit, nous détaillons les différentes étapes pour prédire la vente d'un produit spécifique ainsi que du chiffre d'affaire. Pour ce faire, nous suivons trois étapes principales, à savoir :

- Préparation de données.
- Construction du modèle.
- Validation du modèle.

Nous commençons par présenter la première étape.

- **Préparation de données :**

Pour effectuer cette partie, nous sélectionnerons un historique de ventes d'un produit spécifique et l'historique de vente. Pour cela, nous exécutons une requête SQL pour charger les données de la table des fait "Ventes-Livraison" sous Python et appliquer les traitements nécessaires.

Après qu'on a chargé les données, nous allons les convertir en une série temporelle comme indique Ci-dessous :

```
from datetime import datetime
con=df[ 'Date' ]
df[ 'Date' ]=pd.to_datetime(df[ 'Date' ])
df.set_index('Date', inplace=True)
ts = df[ 'Quantite' ]
```

**Figure 5.12:** Convertir les données à une série temporelle

Les données quotidiennes peuvent être difficiles à utiliser, car elles sont plus volumineuses, alors utilisons plutôt des moyennes mensuelles. Nous allons effectuer la conversion avec la fonction de ré-échantillonnage.

La figure 5.19 illustre la représentation de la série temporelle vente repérées en moyennes mensuelles.

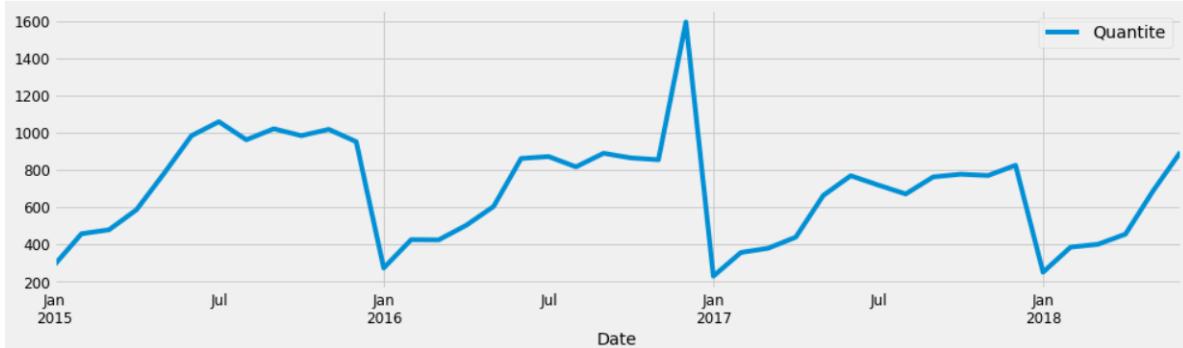


Figure 5.13: Traçage des données

- **Construction du modèle :**

L'une des méthodes les plus courantes utilisées dans la prévision des séries temporelles est connue sous le nom du modèle ARIMA [12], qui signifie AutoRegressive Integrated Moving Average. Le modèle ARIMA est un modèle qui peut être ajusté aux données de séries chronologiques afin de mieux comprendre ou prédire les points futurs de la série. Les modèles ARIMA sont désignés par la notation ARIMA ( $p, d, q$ ) :

- **p** est le paramètre de la partie auto-régressive du modèle. Cela nous permet d'incorporer l'effet des valeurs passées dans notre modèle.
- **d** est le paramètre de la partie intégrée du modèle. Cela inclut les termes du modèle qui intègrent la quantité de différenciation à appliquer à la série temporelle.
- **q** est le paramètre de la partie moyenne mobile du modèle. Cela nous permet de définir l'erreur de notre modèle comme une combinaison linéaire des valeurs d'erreur observées à des moments précédents dans le passé.

Comme nous avons remarqué une saisonnalité, nous utilisons le modèle ARIMA saisonnier, désigné par SARIMA ( $p, d, q$ ) ( $P, D, Q$ )-s. Ici, ( $p, d, q$ ) sont les paramètres non saisonniers décrits ci-dessus, tandis que ( $P, D, Q$ ) suivent la même définition mais sont appliqués à la composante saisonnière de la série chronologique. Le terme s est la périodicité de la série chronologique (4 pour les périodes trimestrielles, 12 pour les périodes annuelles, etc.).

Lorsque nous cherchons à ajuster des données de séries chronologiques avec un modèle ARIMA saisonnier, notre premier objectif est de trouver les valeurs d'ARIMA ( $p, d, q$ ) ( $P, D, Q$ ) qui optimisent une métrique d'intérêt. nous allons résoudre ce problème en écrivant du code Python pour sélectionner par algorithme les valeurs de paramètres optimales pour notre modèle de série chronologique ARIMA ( $p, d, q$ ) ( $P, D, Q$ ). Nous utiliserons une "recherche par

grille" pour explorer itérativement les différentes combinaisons de paramètres. Pour chaque combinaison de paramètres, nous ajustons un nouveau modèle saisonnier SARIMA avec la fonction SARIMAX() de la bibliothèque statsmodels [13] et évaluons sa qualité globale.

```
p = d = q = range(0, 2)
pdq = list(itertools.product(p, d, q))
seasonal_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, d, q))]
```

**Figure 5.14:** Génération des différentes combinaisons de paramètres

Nous pouvons maintenant utiliser les triplets de paramètres définis ci-dessus pour automatiser le processus de formation et d'évaluation des modèles SARIMA sur différentes combinaisons. En statistique et en apprentissage automatique, ce processus est connu sous le nom de recherche de grille (ou optimisation des hyperparamètres) pour la sélection des modèles. Lors de l'évaluation et de la comparaison de modèles statistiques adaptés à différents paramètres, chacun peut être classé par rapport à un autre en fonction de sa pertinence par rapport aux données ou de sa capacité à prévoir avec précision les points de données futurs. Nous utiliserons la valeur AIC [14] (Akaike Information Criterion), qui est renvoyée de manière pratique avec les modèles SARIMA ajustés à l'aide de statsmodels. L'AIC mesure l'adéquation d'un modèle aux données tout en tenant compte de la complexité globale du modèle. Un modèle qui s'ajuste très bien aux données tout en utilisant de nombreuses caractéristiques se verra attribuer un score AIC plus élevé qu'un modèle qui utilise moins de caractéristiques pour obtenir la même qualité d'ajustement. C'est pourquoi nous souhaitons trouver le modèle qui donne la valeur AIC la plus faible. Le morceau de code ci-dessous itère par des combinaisons de paramètres et utilise la fonction SARIMAX des statsmodels pour s'adapter au modèle SARIMA saisonnier correspondant. Ici, l'argument order spécifie les paramètres (p, d, q), tandis que l'argument seasonal\_order spécifie la composante saisonnière (P, D, Q, S) du modèle Seasonal ARIMA. Après ajustement de chaque modèle SARIMAX(), le code imprime son score AIC respectif.



```
warnings.filterwarnings("ignore")
ans=[]
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod = sm.tsa.statespace.SARIMAX(y,
                                              order=param,
                                              seasonal_order=param_seasonal,
                                              enforce_stationarity=False,
                                              enforce_invertibility=False)

            results = mod.fit()
            ans.append([param,param_seasonal,results.aic])
        except:
            continue
ans_df= pd.DataFrame(ans,columns=['pdq','pdqs','aic'])
ans_df.loc[ans_df['aic'].idxmin()]

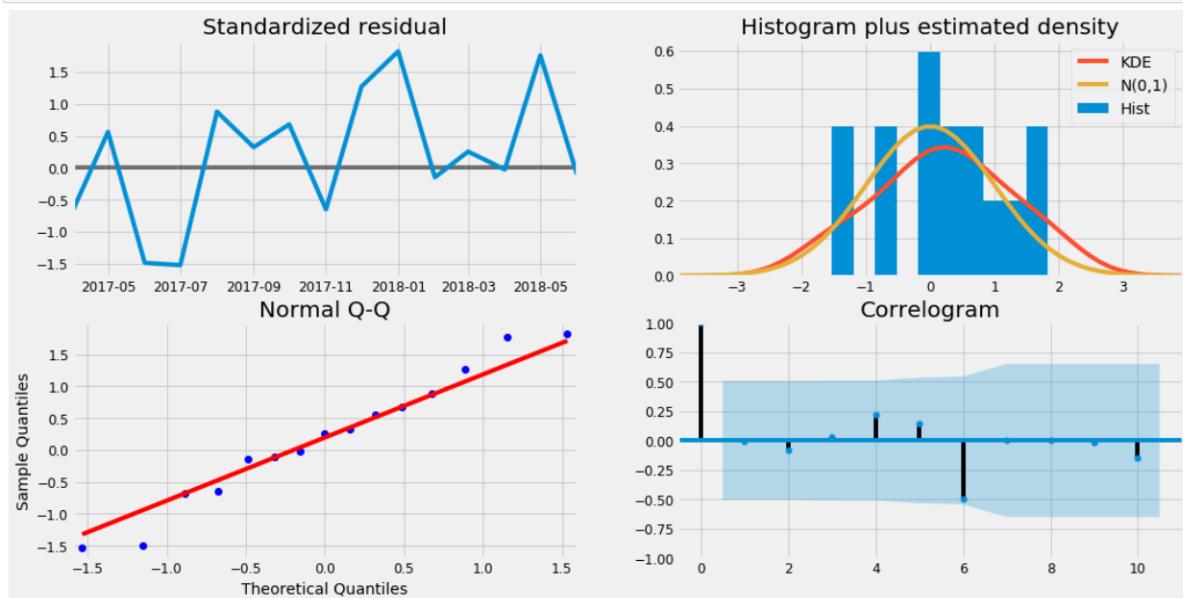
pdq      (0, 1, 1)
pdqs     (1, 1, 1, 12)
aic      177.194
Name: 31, dtype: object
```

**Figure 5.15:** Itérations des différentes combinaisons des paramètres

Pour les données de vente des produits, nous représentons ci-dessous un exemple relatif à un produit bien déterminé. Le résultat de notre code suggère que SARIMAX(0, 1, 1)x(1, 1, 1, 12) donne la valeur AIC la plus basse, soit 177,78. Nous devrions donc considérer cette option comme optimale parmi tous les modèles que nous avons examinés.

Lors de l'ajustement des modèles saisonniers ARIMA (et de tout autre modèle d'ailleurs), il est important d'effectuer des diagnostics de modèle pour s'assurer qu'aucune des hypothèses formulées par le modèle n'a été violée. La fonction `plot_diagnostics` nous permet de générer rapidement des diagnostics de modèle et de rechercher tout comportement inhabituel.

```
results.plot_diagnostics(figsize=(16, 8))
plt.show()
```



**Figure 5.16:** Diagnostic du modèle

Notre principale préoccupation est de veiller à ce que les résidus de notre modèle soient non corrélés et normalement distribués avec une moyenne nulle. Si le modèle saisonnier ARIMA ne satisfait pas ces propriétés, c'est une bonne indication qu'il peut être encore amélioré. Dans ce cas, nos diagnostics de modèle suggèrent que les résidus de modèle sont normalement distribués sur la base de ce qui suit :

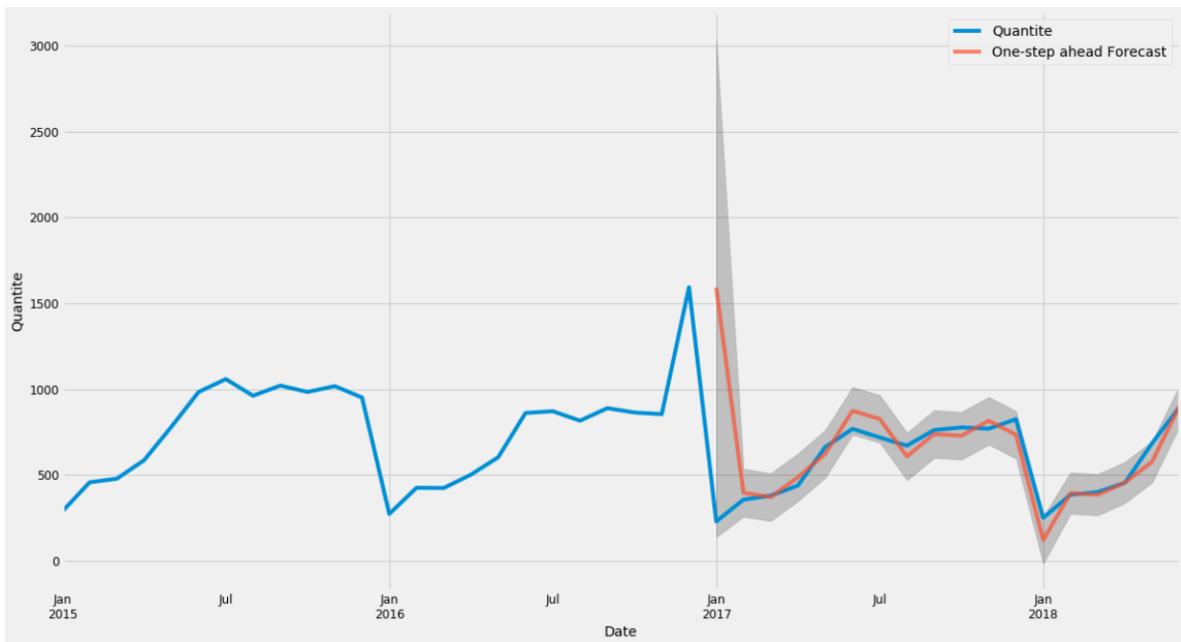
- Dans le graphique en haut à droite, nous voyons que la ligne rouge de KDE semble être normale avec la ligne  $N(0,1)$  (où  $N(0,1)$ ) est la notation standard pour une distribution normale avec une moyenne de 0 et un écart-type de 1).
- Le graphique qq en bas à gauche montre que la distribution ordonnée des résidus (points bleus) suit la tendance linéaire des échantillons prélevés à partir d'une distribution normale standard avec  $N(0, 1)$ . Là encore, cela indique clairement que les résidus sont normalement distribués.
- Les résidus au fil du temps (graphique en haut à gauche) ne présentent pas de saisonnalité évidente et semblent être un bruit blanc. Ceci est confirmé par le tracé d'autocorrélation (c'est-à-dire de corrélation) en bas à droite, qui montre que les résidus de la série temporelle ont une faible corrélation avec les versions décalées de celle-ci.

Ces observations nous amènent à conclure que notre modèle produit un ajustement satisfaisant qui pourrait nous aider à comprendre nos données de séries chronologiques et à prévoir les

valeurs futures.

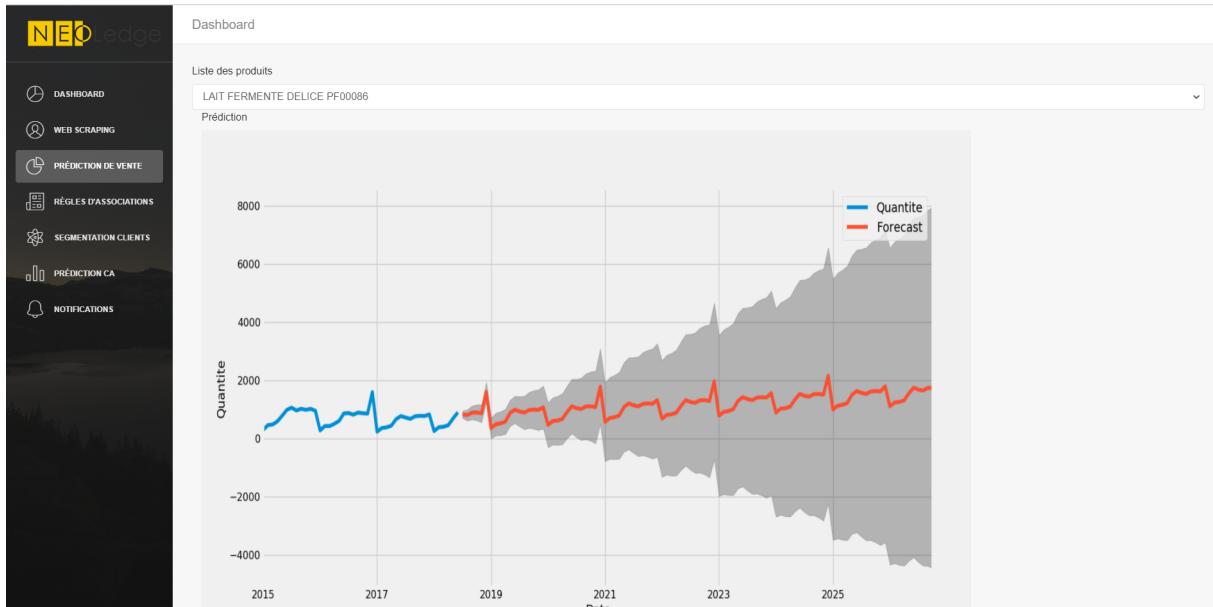
- **Validation du modèle :**

Nous avons obtenu un modèle pour nos séries chronologiques qui peut maintenant être utilisé pour faire des prévisions. Pour valider ensemble le modèle, nous comparerons les valeurs prévues aux valeurs réelles de la série temporelle étudiée de vente d'un produit déterminé, ce qui nous aidera à comprendre la précision de nos prévisions. Les attributs `get_prediction()` et `conf_int()` nous permettent d'obtenir les valeurs et les intervalles de confiance associés pour les prévisions de la série temporelle.



**Figure 5.17:** Comparaison des valeurs

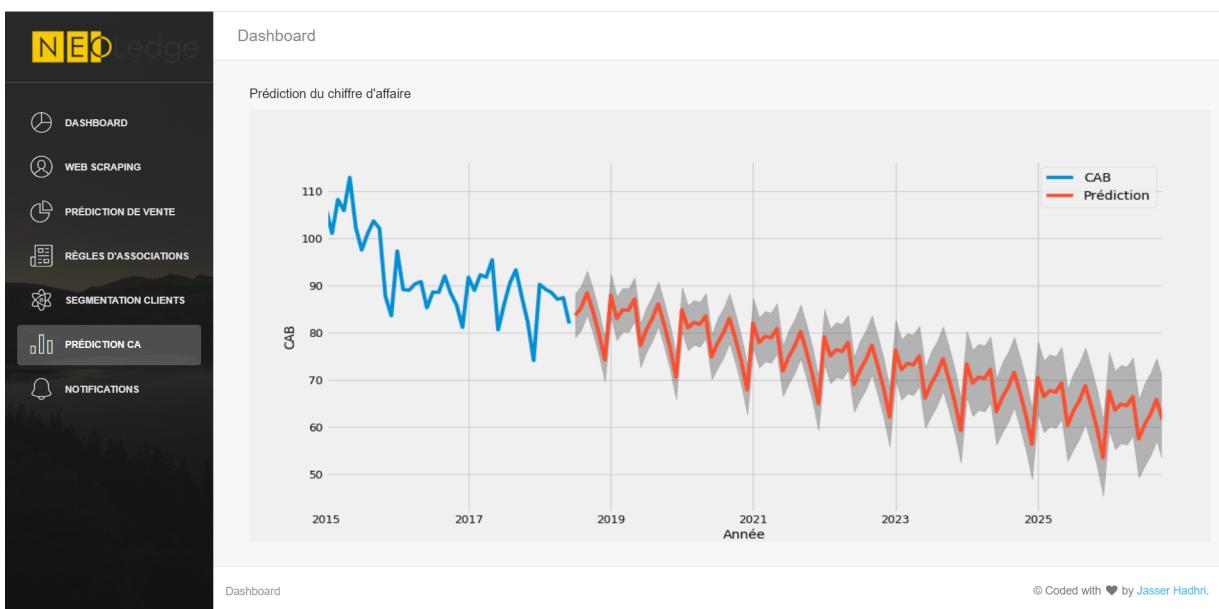
Dans l'ensemble, nos prévisions s'alignent **presque** sur les valeurs réelles, montrant une tendance générale à la hausse. Enfin, nous décrivons comment exploiter notre modèle saisonnier de séries chronologiques ARIMA pour prévoir les valeurs futures. L'attribut `get_forecast()` de notre objet séries temporelles permet de calculer les valeurs prévues pour un nombre déterminé de pas en avant.



**Figure 5.18:** Prédiction de vente d'un produit déterminé

Les prévisions et l'intervalle de confiance associé que nous avons générés peuvent maintenant être utilisés pour mieux comprendre les séries chronologiques et prévoir ce à quoi il faut s'attendre. Nos prévisions montrent que la série chronologique devrait continuer à augmenter à un rythme régulier.

À mesure que nous faisons des prévisions à plus long terme, il est naturel que nous devenions moins confiants dans nos valeurs. Cela se reflète dans les intervalles de confiance générés par notre modèle, qui s'élargissent au fur et à mesure que nous nous éloignons dans l'avenir.

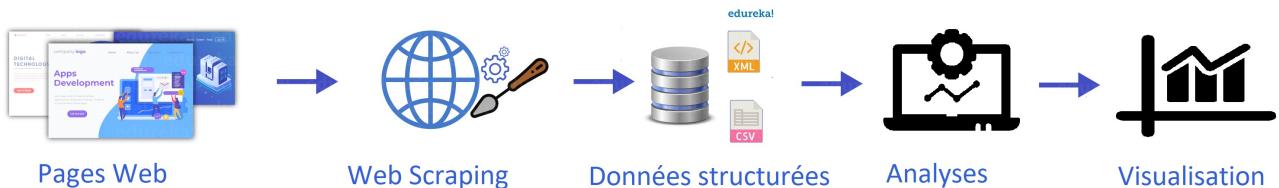


**Figure 5.19:** Prédiction du Chiffre d'affaire

Les prévisions et l'intervalle de confiance associé que nous avons générés peuvent maintenant être utilisés pour mieux comprendre les séries chronologiques et prévoir ce à quoi il faut s'attendre. Nos prévisions montrent que la série chronologique devrait continuer à diminuer à un rythme régulier.

### 5.3 Web scraping

Le Web scraping (également appelé Screen Scraping, Web Data Extraction, Web Harvesting, etc.) est une technique utilisée pour extraire de grandes quantités de données d'un site Web, qui seront extraites et enregistrées sur un fichier local sur ordinateur ou dans une base de données. Les données affichées sur la plupart des sites Web ne peuvent être consultées qu'à l'aide d'un navigateur Web. Ils ne fournissent pas la fonction de sauvegarde d'une copie de ces données pour un usage personnel. Ensuite, la seule option est de copier et coller manuellement les données - c'est une tâche très fastidieuse, qui peut prendre plusieurs heures, voire plusieurs jours. Web Scraping est une technologie qui automatise le processus, de sorte que le Web Scraping peut effectuer la même tâche en peu de temps, au lieu de copier manuellement les données du site Web.



**Figure 5.20:** Architecture technique du web scraping

Notre algorithme de scraping Web chargera et extraira automatiquement les données de la page Facebook de notre client . Il est soit conçue sur mesure pour une page facebook spécifique, soit il peut être configuré pour fonctionner avec plusieurs pages Facebook. En lançant l'algorithme, on peut facilement enregistrer les données disponibles sur la page facebook pour qu'on puisse ensuite les nettoyer et les analyser .

Dans ce qui suit, nous allons présenter les différents étapes de l'extraction , nettoyage et analyse de données.

#### 5.3.1 Extraction de données

Pour réussir cette partie, nous allons préparer un algorithme avec Python permettant d'enregistrer les données(Publications, images, interactions, commentaires ...) dans un fichier JSON .

Pour ce faire , nous allons utiliser quelques bibliothèques Python :

- **Facebook Scrapper [15]** : Cette bibliothèque nous permet d'enregistrer les publications, images, interactions, dates des publications, leurs liens et le nombre des commentaires. Mais malheureusement elle ne permet pas d'extraire le contenu des commentaires ~~d'où le but d'extraire les données des pages Facebook c'est d'analyser les commentaires des consommateurs~~. Pour résoudre ce problème , nous allons extraire les commentaires à travers d'autres bibliothèques et techniques.
- **Requests [16]** : Requests est une bibliothèque HTTP Python publiée sous la licence Apache 2.0. Le but de cette bibliothèque est de rendre les requêtes HTTP plus simples et plus conviviales.Avec cette bibliothèque nous allons envoyer une requête de connexion pour qu'on puisse se connecter à un compte Facebook.
- **BeautifulSoup [17]** : BeautifulSoup est une bibliothèque qui permet de récupérer facilement des informations à partir de pages Web. Il se trouve au sommet d'un analyseur HTML ou XML, fournissant des idiomes pythoniques pour itérer, rechercher et modifier l'arborescence d'analyse. Après qu'on a reçu les liens des publications, nous allons ouvrir un lien à la fois pour pouvoir extraire les commentaires de chaque publication .

### 5.3.2 Analyse de données

Après que nous avons enregistré les données dans un fichier JSON et pour réussir cette partie, nous allons préparer un algorithme pour analyser les sentiments des consommateurs. Pour ce faire, on va préparer un dictionnaire de sentiments. Parce que la plupart des algorithmes de Traitement automatique des langues(NLP) fonctionnent avec les dictionnaires de Français, Anglais et d'autres langues. Malheureusement dans notre cas, les commentaires sont écrits avec la langue commune Tunisienne ce qui nécessite un dictionnaire des mots personnalisé qui nous permet de distinguer les mots de sentiment positives et les mots de sentiment négative comme **ci-dessous**.

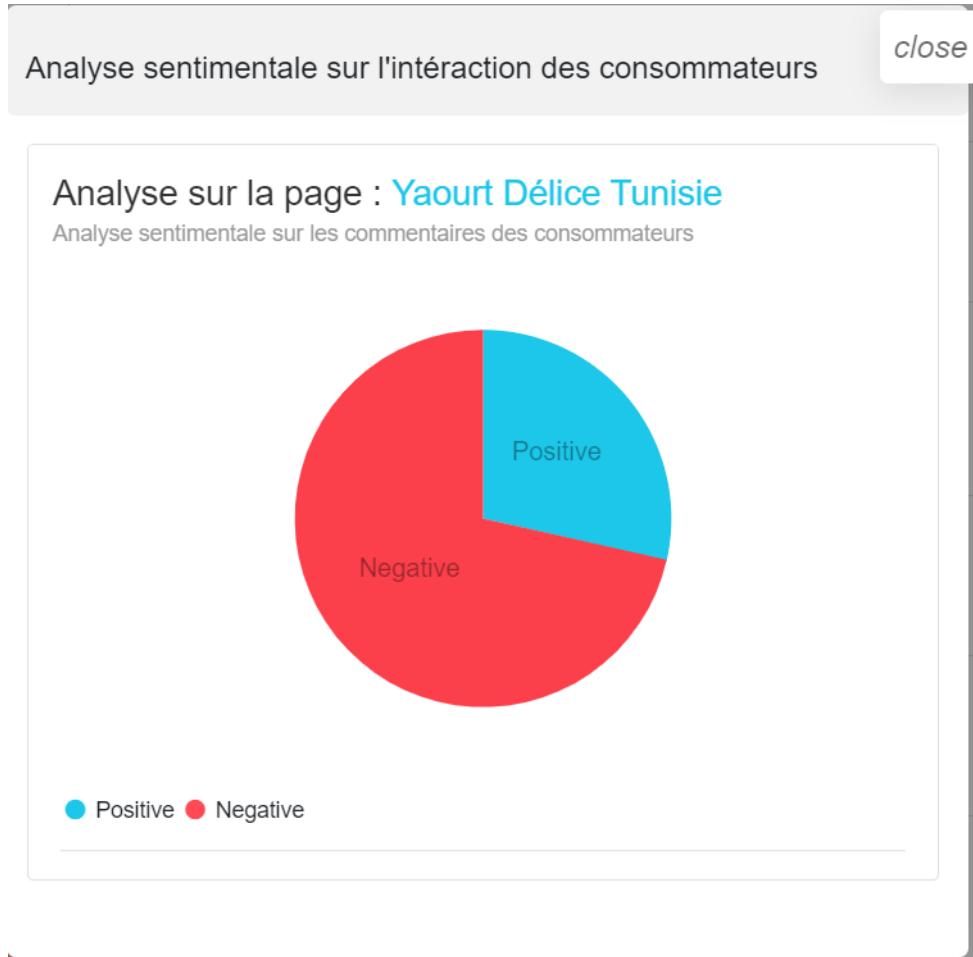
word	sentiment
<3	p
😢	n
❤️	p
جوده	p
عاليه	p
اعز	p
منتوج	p
😢	n
mafameche	n

Figure 5.21: Dictionnaire des sentiments

Ensuite, nous allons développer un algorithme d'analyse sentimentale qui va chercher les mots du dictionnaire dans les commentaires et les classifier sous forme des commentaires positives et négative et enfin attribuer les résultats à chaque publication pour qu'on puisse ensuite visualiser dans la plateforme web **qu'on** a développé et connaitre les interactions des consommateurs sur les produits de notre client .

SENTIMENT	IMAGE	PUBLICATION	DATE PUBLICATION	NOMBRE DES COMMENTAIRES	LIKES	GÉRER
😊		Healthy Nature au sucre brun 🍃 #Délice #YaourtDélice #DéliceNature #Nature #Yummy #Delicious #Tasty #Healthy	2020-07-22 11:32:05	3	67	<a href="#">voir les sentiments</a>
😊		Trouvez l'intrus ! 😊 Faites un screenshot et mettez le en commentaire 📸 #Délice #YaourtDélice #Banane #Game	2020-07-12 11:52:00	55	58	<a href="#">voir les sentiments</a>
😊		شكرون فيكم يختم الباهورنة بالصحيح !! 😊 #YaourtDélice #Délice #Ordre #Fraise #Strawberry	2020-07-04 13:26:00	144	294	<a href="#">voir les sentiments</a>
😊		مركي صاحك المفروم بل 💛 وفیت کعبه نلیس Vanille 🍃 #YaourtDélice #Délice #Vanille #Yummy	2020-06-30 11:14:00	9	364	<a href="#">voir les sentiments</a>
😊		الصحيح !! 😊 #YaourtDélice #Délice #Fruits #Yummy رحمة شنطة ال	2020-06-28 11:11:00	187	361	<a href="#">voir les sentiments</a>

Figure 5.22: Résultat du web scraping et de l'analyse sentimentale



**Figure 5.23:** Courbe des interactions sur une publication

## 5.4 Conclusion

Dans ce chapitre, nous avons présenté la partie analyse de données basée sur plusieurs algorithmes et techniques de data mining, Web scraping et analyse sentimentale pour répondre aux besoins de notre client. Le chapitre suivant inclura la section de récupération de données, où nous présenterons le tableau de bord résultant.

# RÉALISATION

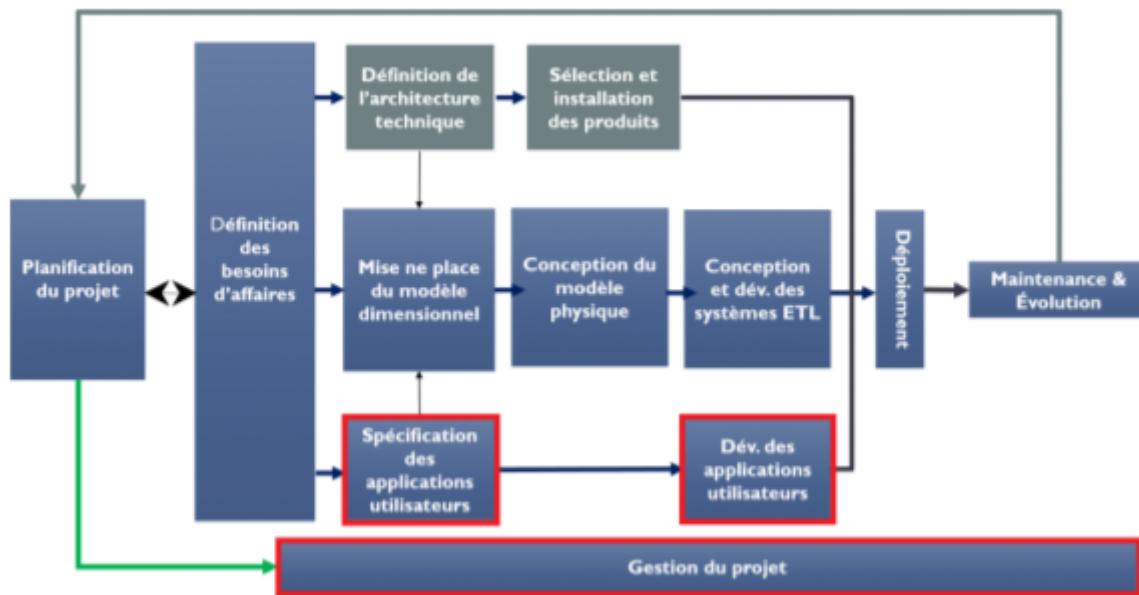
---

## Plan

<b>1</b>	<b>Introduction</b>	<b>66</b>
<b>2</b>	<b>Représentation des applications utilisateurs</b>	<b>66</b>
<b>3</b>	<b>Construction de l'application utilisateur</b>	<b>69</b>
<b>4</b>	<b>Performance Dashboard</b>	<b>72</b>
<b>5</b>	<b>Conclusion</b>	<b>73</b>

## 6.1 Introduction

Dans ce chapitre, nous présenterons les étapes de réalisation et d'implémentation de notre application. Correspondant à la branche de conception de notre cycle de vie dimensionnel souligné par Ralph KIMBALL, comme le montre la figure ci-dessous . Nous montrerons d'abord les maquettes qui nous permettra de vérifier que notre travail répond aux besoins des clients, puis le tableau de bord produit par PowerBI.



**Figure 6.1:** Crédit des maquettes et développement de la solution

## 6.2 Représentation des applications utilisateurs

La spécification des applications utilisateur est l'étape préalable à la création des tableaux de bords finaux de notre solution, elles sont la base pour atteindre facilement les objectifs fixés et satisfaire les besoins des utilisateurs. À ce stade, notre tâche est de produire les modèles nécessaires qui reflètent les exigences précédemment fixées, de plus, nous concevrons ces maquettes en fonction de l'axe analysé et des modules (tables de faits) définis lors du traitement des données et se conforment aux normes de conception de tableau de bord. Dans ce qui suit , nous allons présenter les modèles des tableaux de bords.

### 6.2.1 Maquette : Vente dashboard

La première maquette illustrée dans la figure 6.2 montre une vue d'ensemble sur le trafic de vente de notre client, elle se compose de différents types de graphiques et d'indicateurs de performance.

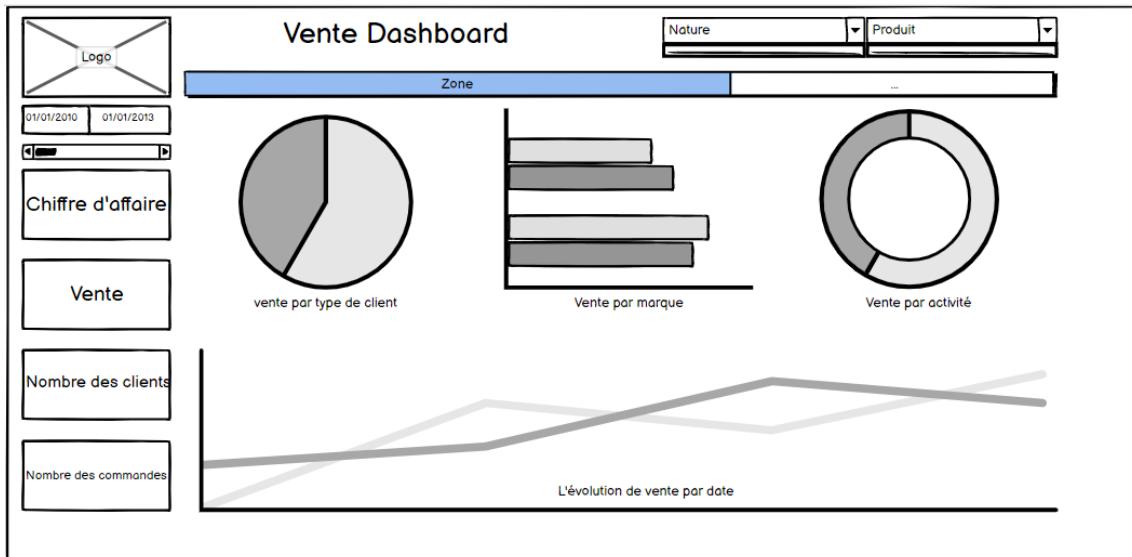


Figure 6.2: Maquette vente dashboard

### 6.2.2 Maquette : Perte et casse dashboard

La figure suivante est le modèle du rapport de suivi de la perter et casse. Il y aura des indicateurs de performance et des graphiques pour les suivre .

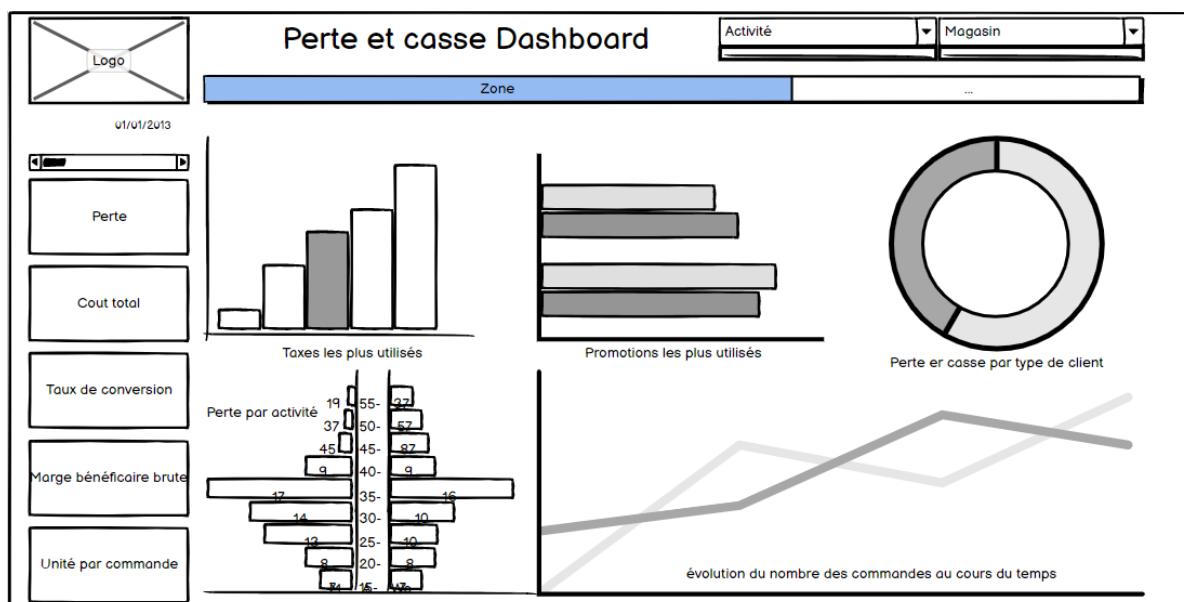


Figure 6.3: Maquette Perte et casse dashboard

### 6.2.3 Maquette : Détails de vente

La maquette illustrée dans la figure 6.4 plus de détails sur le trafic de vente de notre client.

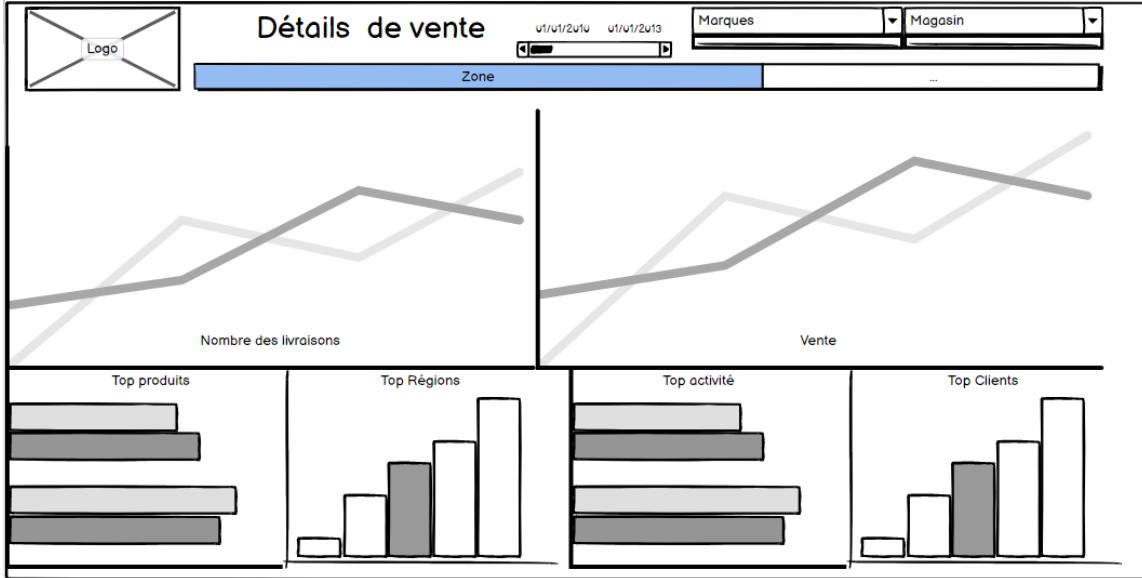


Figure 6.4: Maquette Détails de vente

### 6.2.4 Maquette : Suivi de vente

Cette maquette, et comme illustrée dans la figure 6.5 représente un suivi détaillé sur le trafic de vente de notre client.

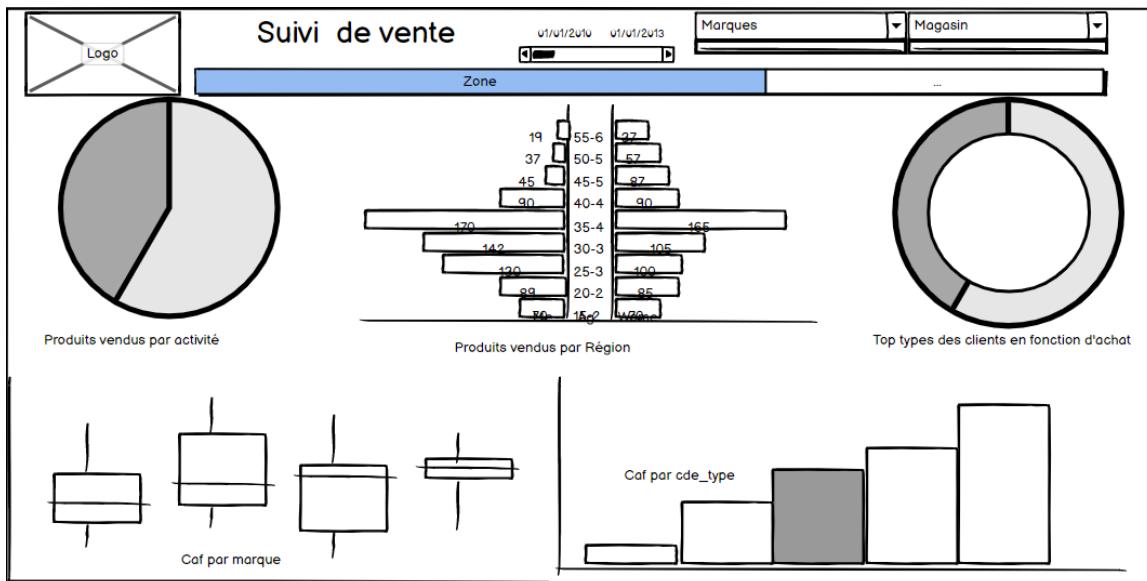


Figure 6.5: Maquette Suivi de vente

### 6.2.5 Maquette : Performances

La figure suivante est le modèle du rapport de suivi de la performance de vente.



**Figure 6.6:** Maquette de Performance

### 6.3 Construction de l'application utilisateur

Le but de cette section est de développer des applications utilisateur simples et faciles à utiliser. Après avoir correctement traité les données extraites de ELISE, nous les utiliserons pour créer un tableau de bord dédié aux décideurs de notre client afin de simplifier le processus décisionnel en respectant les acquis de la partie conception des modèles.

### 6.3.1 Vente Dashboard

Le premier tableau de bord illustré à la figure 6.7 nous fournit une vue globale de toutes les activités de vente en illustrant le développement des ventes, les types de clients, les marques et les activités avec la plus grande part de ventes. Et certains indicateurs, tels que le chiffre d'affaires, le taux de vente, le nombre de clients et le nombre de commandes.

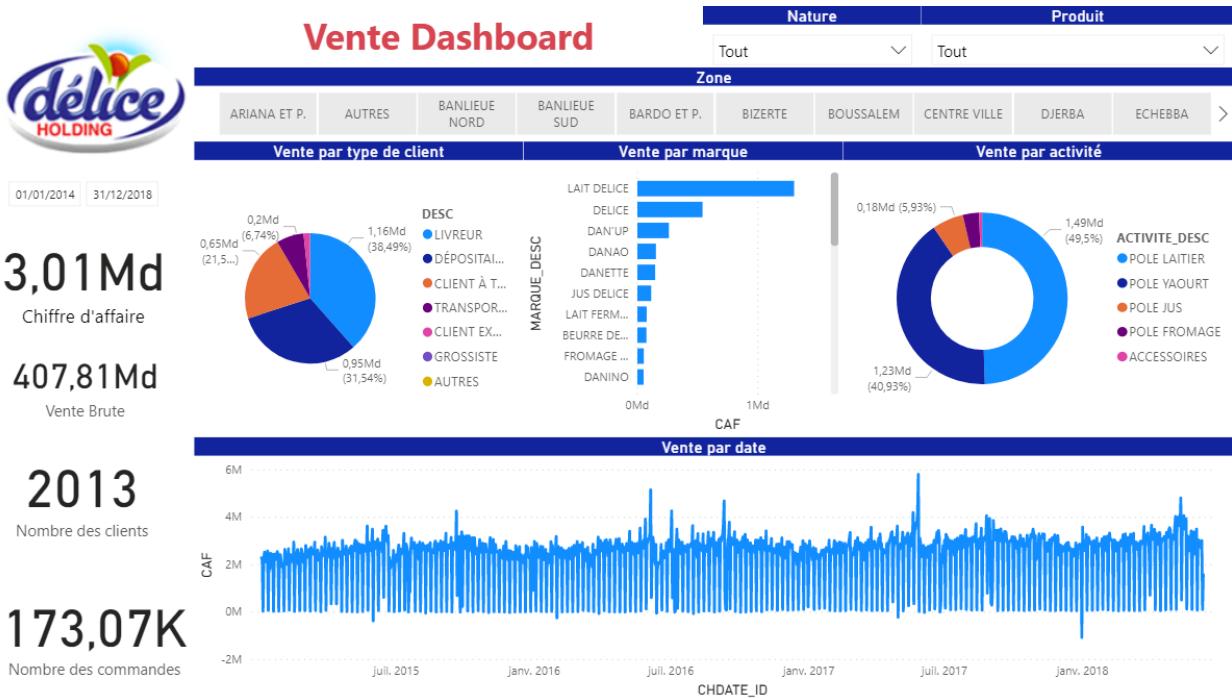


Figure 6.7: Vente Dashboard

### 6.3.2 Perte et casse Dashboard

Le deuxième tableau de bord représenté à la figure 6.8 nous fournit une vue globale sur le trafic de perte et casse de notre client et l'évolution des nombres de commandes, les taxes et les promotions les plus utilisés. Ainsi que les types des clients et activités ayant la grande part de perte et casse. Et certains indicateurs, tels que la perte, le Cout total, le taux de conversion, la marge bénéficiaire brute et l'unité par commande

## Chapitre 6. Réalisation

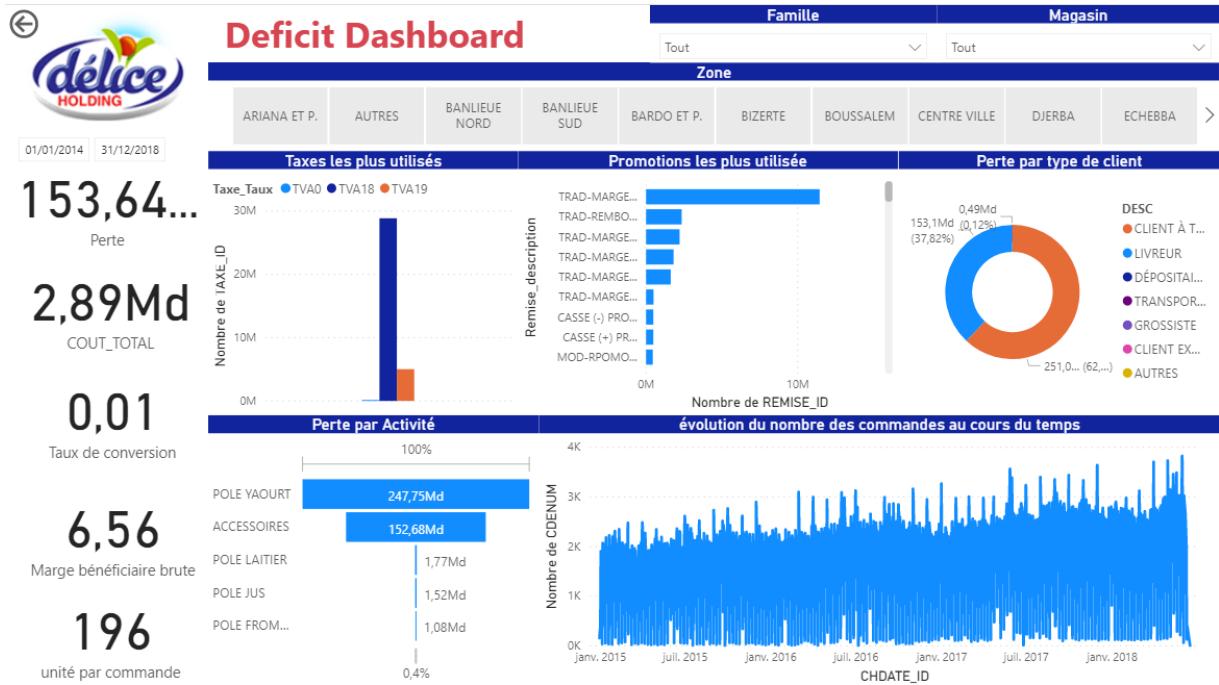


Figure 6.8: Perte et casse dashboard

### 6.3.3 Détails de vente dashboard

Ce tableau de bord représenté à la figure 6.9 nous fournit plus de détails sur la vente en fonction de l'évolution de vente et nombre des livraisons par année. Aussi que les top produits, Régions, Activités et Clients.



Figure 6.9: Détails de vente Dashboard

### 6.3.4 Suivi de vente Dashboard

Ce quatrième tableau de bord illustré à la figure 6.10 nous fournit une vue sur le suivi détaillé de la vente en fonction de nombre de produits vendus par activité, par région et le nombre des clients par chaque type. Aussi les marques et les cde\_types ayant la grande part du chiffre d'affaire.

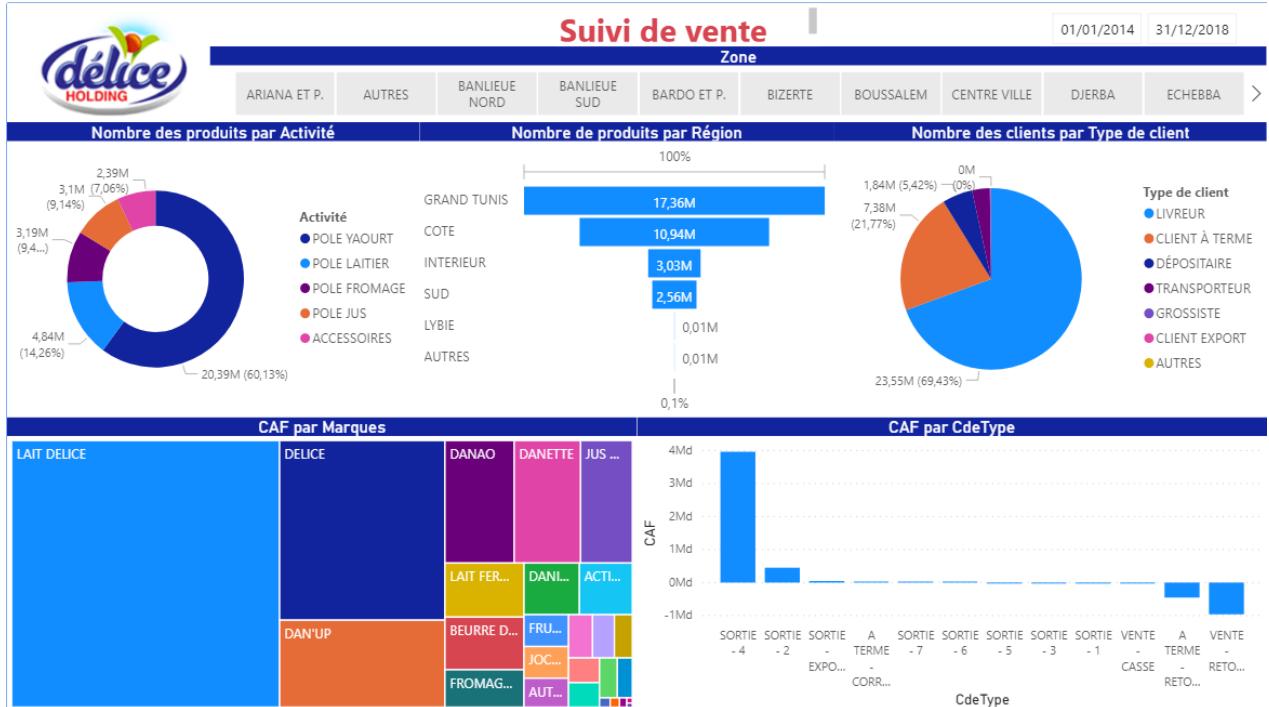


Figure 6.10: Suivi de vente Dashboard

## 6.4 Performance Dashboard

Ce tableau de bord illustré à la figure 6.11 nous fournit une vue sur la performance de vente de notre client, en représentant une comparaison entre la vente et la perte casse et une autre comparaison entre le nombre des livraisons et l'utilisations des promotions.

## Chapitre 6. Réalisation

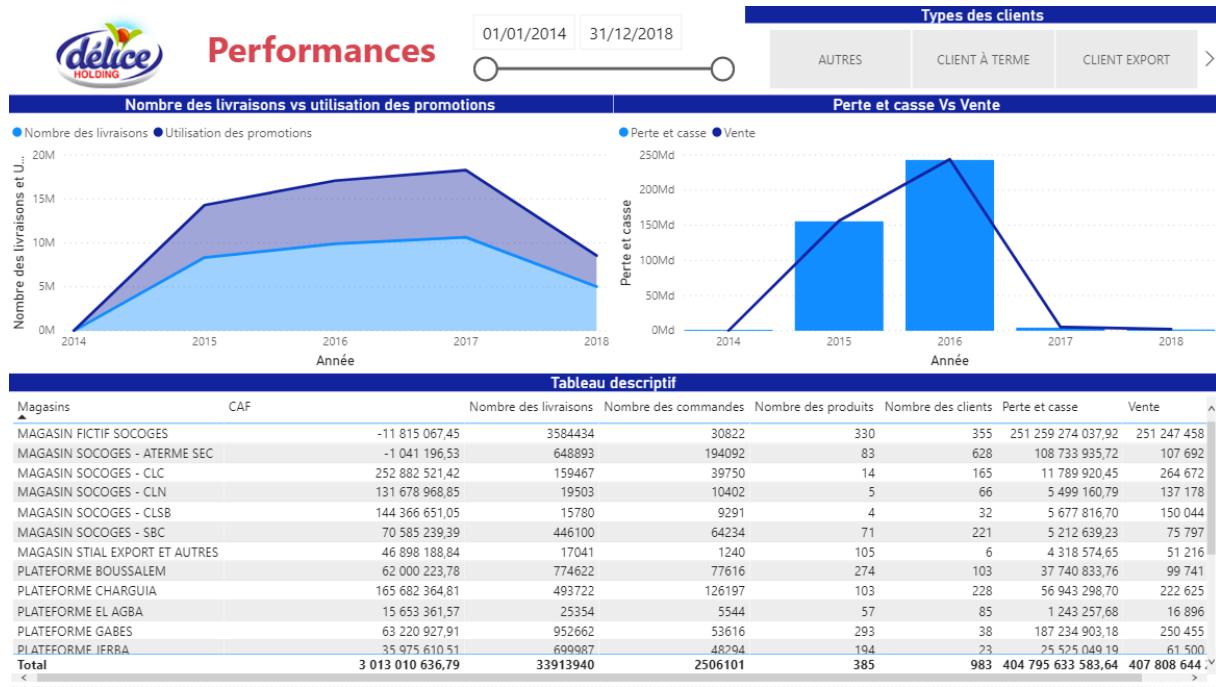


Figure 6.11: Performances Dashboard

## 6.5 Conclusion

Dans ce chapitre, nous avons illustré le travail dont nous avons besoin en représentant les modèles du tableau de bord demandés. Ensuite, nous montrons quelques captures d'écran de tableaux de bord réalisés avec power bi. La prochaine étape consiste à terminer notre rapport avec la conclusion générale et les perspectives.

# Conclusion générale

Afin de créer un système de prise de décision pour le suivi des actions de ventes et marketing, ce stage de fin d'études a été réalisé au sein de **NeoLedge**. En effet, l'application vise à visualiser un ensemble de tableaux de bord et analyses afin que les décideurs puissent avoir une vision globale et complète de la gestion et du suivi de vente afin qu'ils puissent enfin prendre la décision la plus appropriée.

L'objectif global est donc de développer un système décisionnel afin d'analyser les données provenantes de ELISE pour avoir une visibilité claire sur l'historique de vente ainsi qu'une analyse sentimentale sur le comportement des consommateurs de notre client.

Afin de mettre en œuvre cette solution, nous avons d'abord découvert ELISE et générée les données nécessaires. Cette étape, nous a permis de mieux comprendre notre base de données et donc d'avoir une compréhension plus claire des indicateurs à mesurer. Par la suite, nous nous sommes concentrés sur le traitement des données, tout en assurant trois étapes : Utilisez les outils SSIS pour extraire, transformer et charger des données dans l'entrepôt de données , SSAS pour développer les mesures et indicateurs de performances nécessaires et Python pour développer les algorithmes de data mining et pour extraire les données du web et les analyser ensuite

La troisième partie du projet consiste à utiliser les outils Power BI pour créer des tableaux de bord interactifs et détaillés.

Finalement, nous avons développé une plateforme web développée par VueJs pour afficher les différents tableaux de bord et les résultats d'analyses et web scraping qu'on a réalisé d'une façon plus claire et bien structurés.

Ce projet d'une importance capitale pour moi il m'a donné l'opportunité d'intégrer dans le domaine professionnel malgré les circonstances actuelles causées par le covid19. En effet, c'était un stage enrichissant parce que j'ai eu la possibilité de renforcer mes compétences techniques que j'ai acquis au cours de mes études universitaires à Esprit bénéficiant de la découverte de nouveaux concepts et technologies, et parce que j'ai eu la possibilité de travailler pour des utilisateurs finaux dans un environnement réel.

## Conclusion générale

---

En guise de perspectives, notre projet sera déployer sur les serveurs de notre client comme première étape . Ensuit nous allons intégrer la base de données de production dans notre plateforme. Et nous allons enrechir le dictionnaires des mots Tunisiennes et améliorer l'algorithme de web scraping pour qu'il soit plus rapide et intélligent.

# Bibliographie

- [1] *Archimed*, <https://www.archimed.fr/>, [Consultée en ; Février 2020].
- [2] *Neoledge*, <https://www.neoledge.com/>, [Consultée en ; Février 2020].
- [3] *SqlServer*, [https://fr.wikipedia.org/wiki/Microsoft\\_SQL\\_Server](https://fr.wikipedia.org/wiki/Microsoft_SQL_Server).
- [4] *MSBI*, <https://thwack.solarwinds.com/t5/THWACK-EMEA/What-is-meant-by-MSBI-Technology/gpm-p/445802>.
- [5] *POWER BI*, [https://en.wikipedia.org/wiki/Microsoft\\_Power\\_BI](https://en.wikipedia.org/wiki/Microsoft_Power_BI).
- [6] *Python*, [https://fr.wikipedia.org/wiki/Python\\_\(langage\)](https://fr.wikipedia.org/wiki/Python_(langage)).
- [7] *VueJs*, <https://fr.wikipedia.org/wiki/Vue.js>.
- [8] *Architecture MSBI*, <https://www.google.com/search?q=architecture+msbi&sxsrf=ALeKk03gLS3O5VsUCidCxDxPA:1602420395928&source=lnms&tbo=isch&sa=X&ved=2ahUKEwiBpuuTyazsAhVEzBoKHSkuDQAUoAXoECAQQAw&biw=1536&bih=754#imgrc=kh3IeAd-2RCFCM>.
- [9] *Approche de Kimball*, <https://www.aerow.group/a16u1509/>.
- [10] *Approche d'Immon*, <https://www.aerow.group/a16u1509/>.
- [11] *Pyodbc*, <https://pypi.org/project/pyodbc/>.
- [12] *ARIMA*, [https://en.wikipedia.org/wiki/Autoregressive\\_integrated\\_moving\\_average](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average).
- [13] *Statsmodels*, <https://www.statsmodels.org/stable/index.html>.
- [14] *AIC*, [https://en.wikipedia.org/wiki/Akaike\\_information\\_criterion](https://en.wikipedia.org/wiki/Akaike_information_criterion).
- [15] *Facebook Scrapper*, <https://pypi.org/project/facebook-scraping/>.
- [16] *Requests*, [https://en.wikipedia.org/wiki/Requests\\_\(software\)](https://en.wikipedia.org/wiki/Requests_(software)).
- [17] *BeautifulSoup*, [https://fr.wikipedia.org/wiki/Beautiful\\_Soup](https://fr.wikipedia.org/wiki/Beautiful_Soup).