**Entrepôt de données**

**Rapport Final**

***Sébastien Quiquerez***

***Johan Steiner***

Table des matières

[1. Introduction 4](#_Toc471845846)

[2. Analyse de l’existant 4](#_Toc471845847)

[3. Analyse des besoins 4](#_Toc471845848)

[4. Cahier des charges 4](#_Toc471845849)

[5. Modèle dimensionnel en étoile 5](#_Toc471845850)

[5.1. Explications des choix 6](#_Toc471845851)

[5.1.1. Tables de faits 6](#_Toc471845852)

[5.1.2. Mini dimension Commande 6](#_Toc471845853)

[5.1.3. Dimension Temps : Liaisons multiples 6](#_Toc471845854)

[5.1.4. Dimensions Client et Geographie : flocon 6](#_Toc471845855)

[5.2. Hiérarchies 7](#_Toc471845856)

[5.2.1. Hiérarchie du temps 7](#_Toc471845857)

[5.2.2. Hiérarchie des vendeurs 7](#_Toc471845858)

[5.2.3. Hiérarchie des articles 8](#_Toc471845859)

[5.2.4. Hiérarchie de la géographie 8](#_Toc471845860)

[6. Données : importation/transormation 8](#_Toc471845861)

[6.1. Staging 8](#_Toc471845862)

[6.1.1. Types de données 9](#_Toc471845863)

[6.2. Data Warehouse 9](#_Toc471845864)

[6.2.1. Master data 9](#_Toc471845865)

[6.2.2. Qualité des données 9](#_Toc471845866)

[6.2.3. Audit 9](#_Toc471845867)

[6.3. Cube 9](#_Toc471845868)

[7. Indicateurs 9](#_Toc471845869)

[8. Conclusion 9](#_Toc471845870)

[9. Bibliographie 9](#_Toc471845871)

# Introduction

Mediazon est spécialisée dans la vente de produits multimédias. L’entreprise nous a mandaté pour mettre en place la structure de son Data Warehouse qui stockera de grandes quantités de données issues des commandes de ses clients.

Les données source sont donc sous une forme « brute », provenant de différentes sources, avec un certain nombre d’erreurs et il est nécessaire de mettre en place un système permettant d’obtenir des données d’une certaine qualité afin de permettre le développement d’un système décisionnel bien conçu.

Le développement du système de test a été réalisé depuis une machine virtuelle Microsoft Server 2012. Le logiciel Microsoft Visual Studio ainsi que Microsoft SQL Server Management Studio nous ont permis de réaliser la plus grande partie de notre travail.

# Analyse de l’existant

Au départ, les données sont dispersées entre plusieurs sources de données de différents types. A notre disposition, nous avons les types de fichiers suivants :

* Fichier CSV délimité
* Fichier Microsoft Access Database
* Fichier texte non délimités
* Fichier texte délimité

Les informations contenues dans les différents fichiers sont décrites ci-dessous.

## Fichier CSV délimité

Dans ce fichier, les différentes villes de France sont présentes, avec leur NPA respectif et le département dans lequel elles se situent. Chaque ville est identifiée par un ID unique.

## Fichier Microsoft Access Database

Ce fichier contient 2 catégories de données distinctes.

### Devises

Mediazon livre ses produits dans plusieurs régions du monde, ce qui signifie qu’elle doit adapter la devise de paiement en fonction de ces régions. Elle possède donc la liste des devises présentes dans le monde avec, pour chacune d’elles, un code à trois chiffres uniques.

### Vendeurs

L’entreprise possède également une liste de ses vendeurs avec pour chacun d’eux son nom et prénom, la division et le groupe dans laquelle il travaille, ainsi que son manager respectif.

En plus de cela, des informations plus personnelles telles que le sexe de la personne, son quota, bonus ou commission sont renseignées. Les vendeurs sont aussi identifiés à l’aide d’un ID unique.

## Fichier texte non délimité

Le fichier texte non délimité ne contient pas de délimiteur (virgule, point-virgule, etc…) mais chaque « champ » a un nombre de caractères défini. Si le contenu de ce champ est plus petit que le nombre défini, les caractères disponibles restants sont comblés par des espaces. Cela crée des « pseudo-colonnes ».

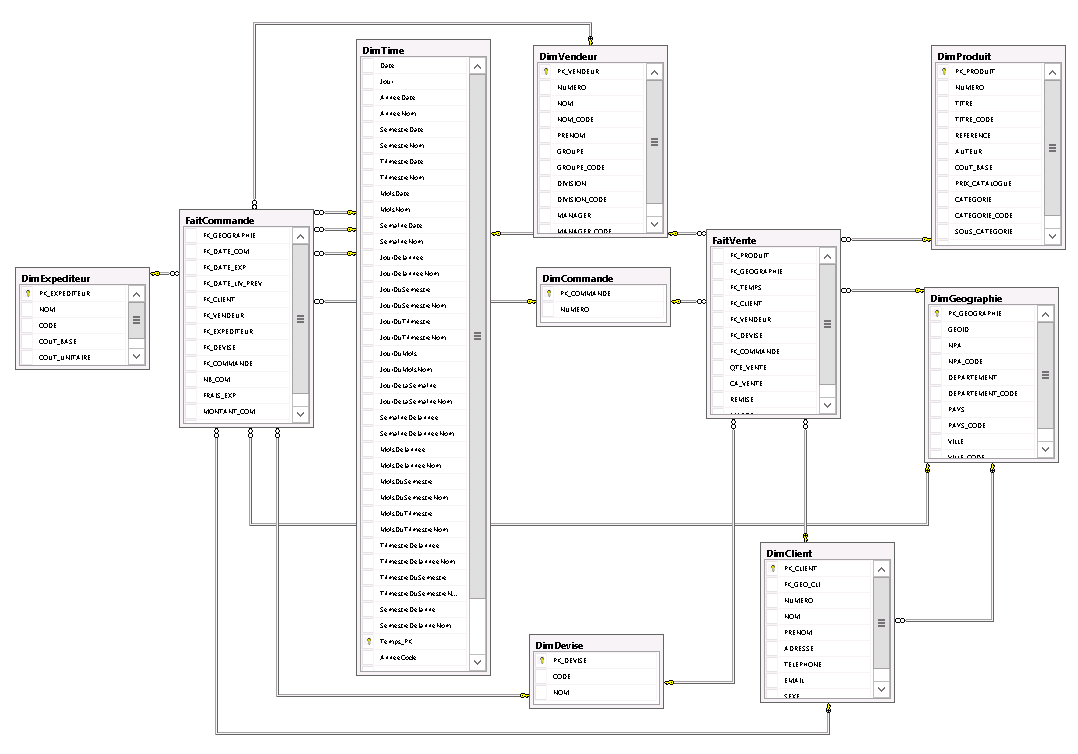
Les fichiers textes ne contiennent pas d’entêtes, néanmoins nous avons eu les informations sur la signification des valeurs/textes présents dans chaque « colonne ».

## Fichier texte délimité

# Analyse des besoins

# Cahier des charges

# Modèle dimensionnel en étoile



## Explications des choix

### Tables de faits

Notre schéma contient 2 tables de faits : la table **Commande** et la table **Vente**.

La table *Vente* contient les informations propres aux produits et plutôt à une ligne de la commande (En quelle quantité un produit a-t-il été acheté ? De quel pourcentage de remise a-t-on bénéficié sur chaque produit ? Quelle marge a été réalisée sur la vente ? Etc…).

La table *Commande* est plus orientée sur la commande en générale, par exemple pour connaître les dates de commande, livraison et expédition, l’expéditeur ainsi que ses frais ou encore le client qui a passé la commande.

### Mini dimension Commande

Avant la création de cette dimension, il était impossible de lier une commande aux ventes. On ne pouvait par exemple pas connaître le prix total de la commande.

La mini dimension Commande permet de créer cette liaison : elle contient le numéro de chaque commande et dans chaque table de faits, une clé secondaire référence cette table afin qu’il soit possible de faire correspondre les informations des deux tables de faits.

Par exemple, si l’on souhaite connaitre le montant total des produits d’une commande (laissons de côté les frais d’expédition pour cet exemple), il suffit de multiplier les prix d’un produit par sa quantié et de répéter cette opération pour chaque produit de la commande. Grâce à notre mini dimension Commande, nous pouvons lier le numéro de commande de la table de faits *Commande* à la table de faits *Vente* et donc savoir quel produit appartient à quelle commande.

### Dimension Temps : Liaisons multiples

Nous avons créé 3 liaisons entre la dimension Temps et la table de faits Commande, car dans le cas, il est mentionné 3 types de dates distinctes :

* La date de commande
* La date de livraison prévue
* La date d’expédition

### Dimensions Client et Geographie : flocon

Un flocon a été mis en place entre les tables Client et Geographie car nous avons jugé intéressant de connaitre la provenance de nos clients sans qu’on doive forcément se lier à une vente. Cela permettra la mise en place d’indicateurs supplémentaires.

## Hiérarchies

### Hiérarchie du temps

**Année**

**Année**

**Semaine**

**Mois**

**Semestre**

**Jour**

**Jour**

**Trimestre**

**Mois**

**Jour**

### Hiérarchie des vendeurs

**Groupe**

**Division**

**Manager**

**Vendeur**

### Hiérarchie des produits

**Catégorie**

**Sous-catégorie**

**Article**

### Hiérarchie de la géographie

**Pays**

**Département**

**Ville**

# Données : importation/transormation

L’importation des données dans notre Data Warehouse a été réalisée à l’aide de la méthode ETL (Extract, Transform, Load).

## Staging

La zone staging représente la zone contenant les données qui n’ont pas encore été transformées. Nous extrayons les données depuis nos différentes sources (fichiers Excel, fichiers Access, fichiers texte, etc…), les plaçons dans notre staging area, à la suite de quoi nous les transformons (conversion de type, modification de la longueur des champs, etc…) pour qu’elles correspondent à ce qu’attendent les tables présentes dans notre Data Warehouse et les chargeons dans celui-ci.

### Types de données

L’importation des données dans la staging a été simplifiée au niveau des types de données. En effet, nous n’avons à ce moment-là pas encore procédé à une transformation des différents types, mais avons simplement importé toutes les données sous forme de chaînes de caractères (sauf les identifiants).

## Data Warehouse

### Master data

### Qualité des données

### Audit

## Cube

# Indicateurs

# Conclusion

# Bibliographie