

Entrepôt de données

Par Maggie LEKPA

INTRODUCTION

Depuis les années 80, les entreprises ont vu la quantité de données évoluée à la vitesse grand V;

D'ici 2035 → 2142 zettaoctets

On parle de 3V

Besoin de tirer profit des données; les analyser pour en tirer profit notamment pour savoir ce qui marche ou ne marche pas bien; ce qui est à améliorer ou qui peut être anticipé.

INTRODUCTION

Une entreprise a généralement plusieurs systèmes opérationnels :

CRM (Client Relationship Management) pour la gestion des clients

ERP (Entreprise Ressource Planning) pour la gestion des ressources...

Système de gestion des stocks/commandes...

Ou encore plusieurs entités indépendants localisées dans différents pays.

INTRODUCTION

Pour une bonne prise de décision, il est nécessaire de mettre toutes les données en relation (dans un lieu unique) pour pouvoir les croiser.

Exemple: le calcul du chiffre d'affaire d'une entreprise internationale nécessite la prise en compte des données de ses différentes entités dans le monde.

Problématique : besoin d' un système centralisé de données pour l'aide à la décision / visualiser les KPI de l'entreprise...

INTRODUCTION

Caractéristique du système central:

Simple et Rapide : les personnes chargées de prendre les décisions ne sont pas forcément des informaticiens et ont un temps très précieux.

Performant : afin de traiter de gros volume de données

Indépendant du système de production : important de dissocier la production avec le système décisionnel

Gestion d'accès: tout le monde n'est pas habilité à prendre les décisions dans l'entreprise.

Hétérogène : toutes les données de l'entreprise doivent pouvoir y être agrégées pour une meilleure prise de décision.

INTRODUCTION

Les entrepôts de données ou data Warehouse sont nés

Un **entrepôt de données** peut être défini comme étant un système de stockage qui connecte et harmonise de gros volumes de données provenant de nombreuses sources.

Un entrepôt de données → base de données décisionnelle

Deux mondes appelés à coexister :

Le monde **opérationnel** et le monde **décisionnel**

Système OPERATIONNEL (OLTP)

OLTP : OnLine Transactionnel Processing

Destiné à l'ensemble des opérations (transactions) de l'entreprise (traitement d'une commande, gestion des stocks, édition d'une facture...)

Les données de l'entreprise qui y sont stockées ont les données de vie de l'entreprise (les données à l'instant présent).

Systeme DECISIONNEL (OLAP)

OLAP : OnLine Analytical Processing

Aide à piloter l'entreprise et à prendre des décisions.

La prise de décision se fait en se basant sur des KPI.

KPI : Key Performance Indicator – indicateur clé de performance

Exemple de KPI : Chiffre d'affaire, nombre de nouveaux clients

OLTP VS OLAP

OLTP	OLAP
Accessible à toutes personnes contribuant à la vie de l'entreprise	A destination d'un nombre restreint d'utilisateurs : les décideurs
Petite volumétrie de données et requêtes simples	Gros volume de données (agrégées) et requêtes complexes
Système très rapide	Idéalement rapide

ARCHITECTURE D'UN SYSTÈME DECISIONNEL

Intégration de données

Une entreprise a généralement plusieurs systèmes opérationnels :

CRM (Client Relationship Management) pour la gestion des clients

ERP (Entreprise Ressource Planning) pour la gestion des ressources...

Système de gestion des stocks/commandes...

Pour une bonne prise de décision, il est nécessaire de mettre toutes les données en relation (dans un lieu unique) pour pouvoir les croiser.

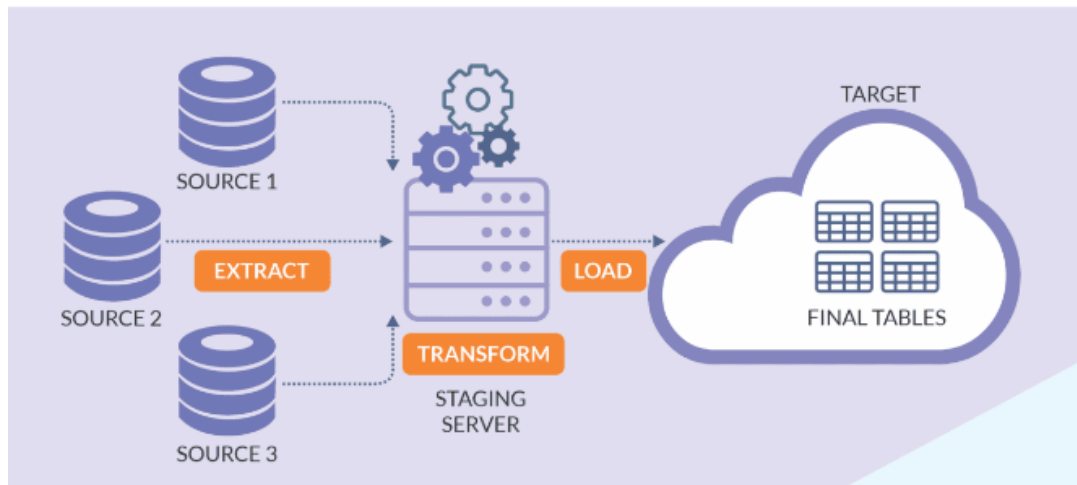
Intégration de données

Comment quitte t'on d'un ou plusieurs systèmes opérationnels à un système unique décisionnel ? → **ETL (Extract Transform Load)**

ETL est le processus par lequel les données vont transiter du système opérationnel au système décisionnel. Les données vont être nettoyées et ou transformées avant d'être stockées.

Intégration de données - ETL

ETL (Extract Transform Load)

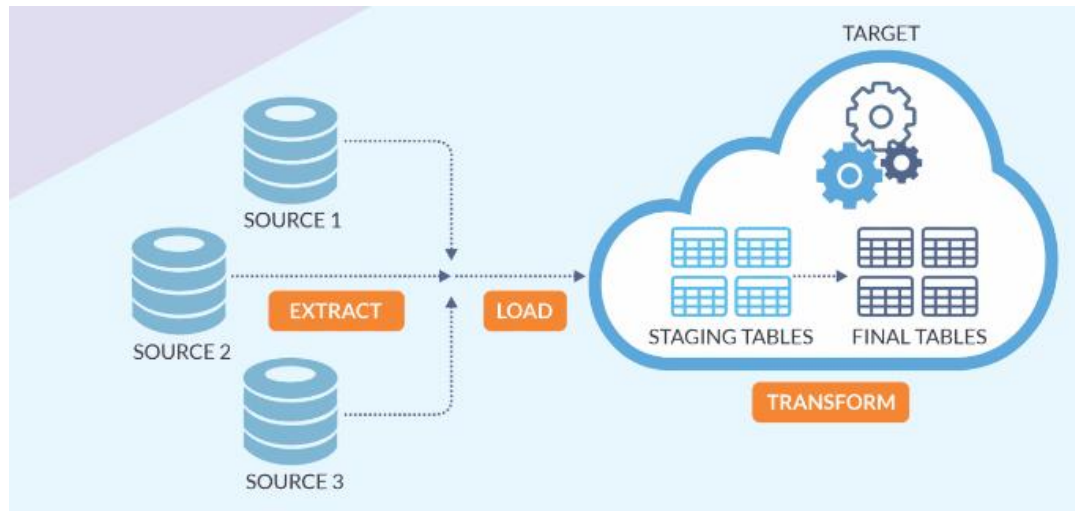


Système opérationnel

Système décisionnel

Intégration de données

Variante : ELT (Extract Load Transform)



Intégration de données

ETL	ELT
La transformation est faite avant le chargement dans les tables → données sensibles masquées ou supprimées	Transformation faite après le chargement dans les tables → demande un bon niveau de sécurité
Le chargement des données prend du temps	Données disponibles très rapidement.
Gestion de l'évolution des besoins de données difficile à gérer.	Plus flexible car toutes les données sont à disposition

DWH

Les données intégrées via ETL sont stockées dans un **entrepôt de données (DWH)**.

Un **entrepôt de données** est une base de données dans laquelle sont stockées les données historiques et structurées de l'entreprise dans le but de les analyser.

Les entrepôts de données étant très volumineux → besoin de les diviser en élément plus petits et facile à gérer → **DATAMART**

Un DATAMART est un sous ensemble d'un entrepôt; il contient les données liées à une fonction de l'entreprise (vente, client..) ou a une branche de l'entreprise.

Moins volumineux

Comment est modélisé l'entrepôt de données?

DWH – Modélisation

Modélisations les plus utilisées :

- **Modélisation en étoile**
- **Modélisation en Flocon**

On parle de Fait et Dimension dans les deux modélisations.

Dimension : ce qu'on utilise pour faire l'analyse (le temps, par société, par poste...)

Table de fait : contient les données à analyser; ce sont des valeurs, quantité (le chiffre d'affaires, la rémunération...)

DWH – Modélisation

Dans un DWH, on parle de Fait et Dimension.

Dimension : ce qu'on utilise pour faire l'analyse (le temps, par société, par poste...)

Table de fait : contient les données à analyser; ce sont des valeurs, quantité (le chiffre d'affaire, la rémunération...)

DWH – Modélisation

Fait ou Dimension?

Produit	
ID 	integer
Nom_Produit	varchar
Type_Produit	varchar

Date	
ID 	integer
Annee	varchar
Mois	varchar

Ventes	
ID 	integer
Id_Magasin	integer
Id_Produit	integer
Id_date	integer
vente	decimal
profits	decimal
created_at	timestamp

Magasin	
ID	integer
Description	varchar
Ville	varchar

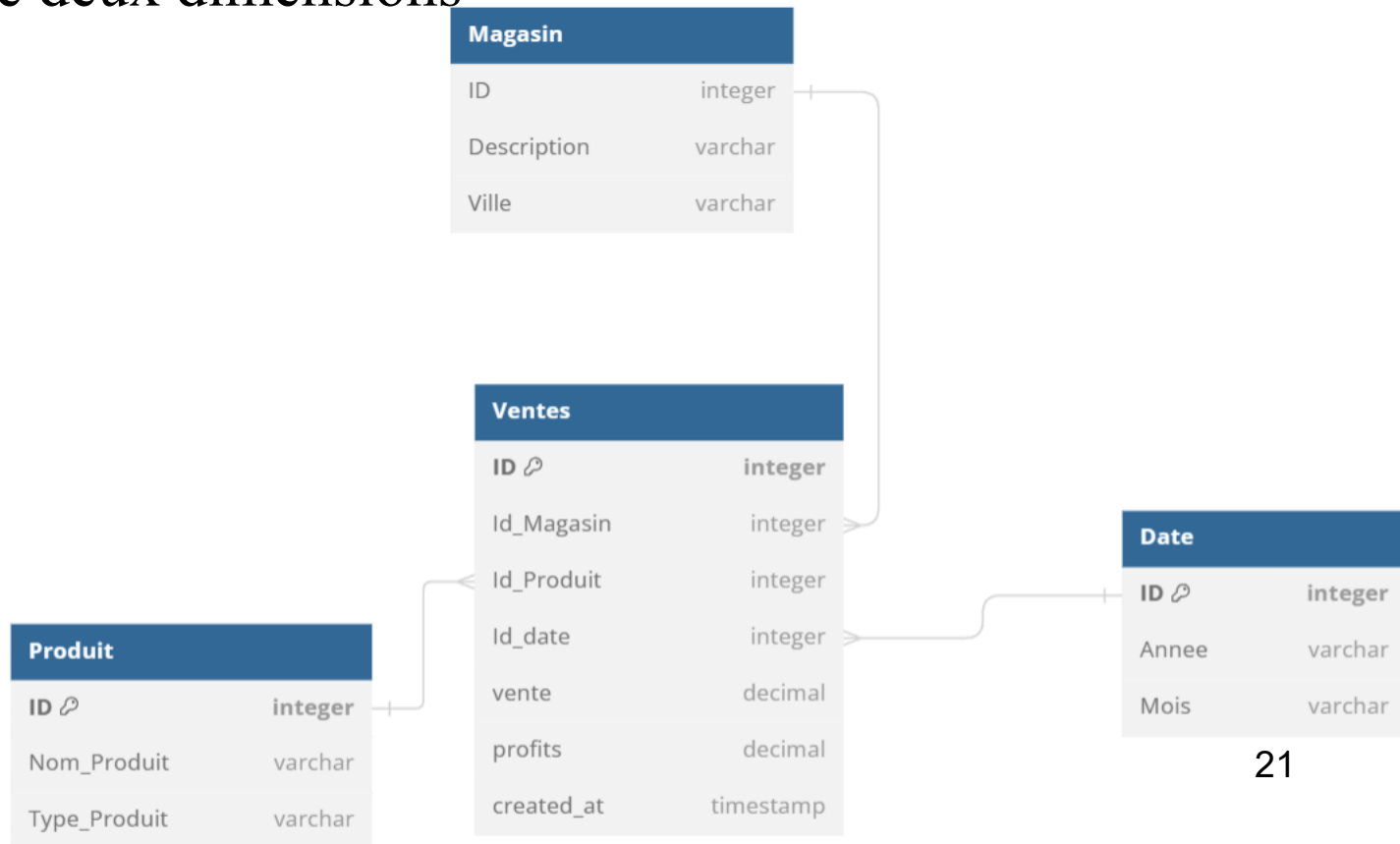
DWH – Modélisation en étoile

Modélisation en **Etoile** :

Une table de fait centrale

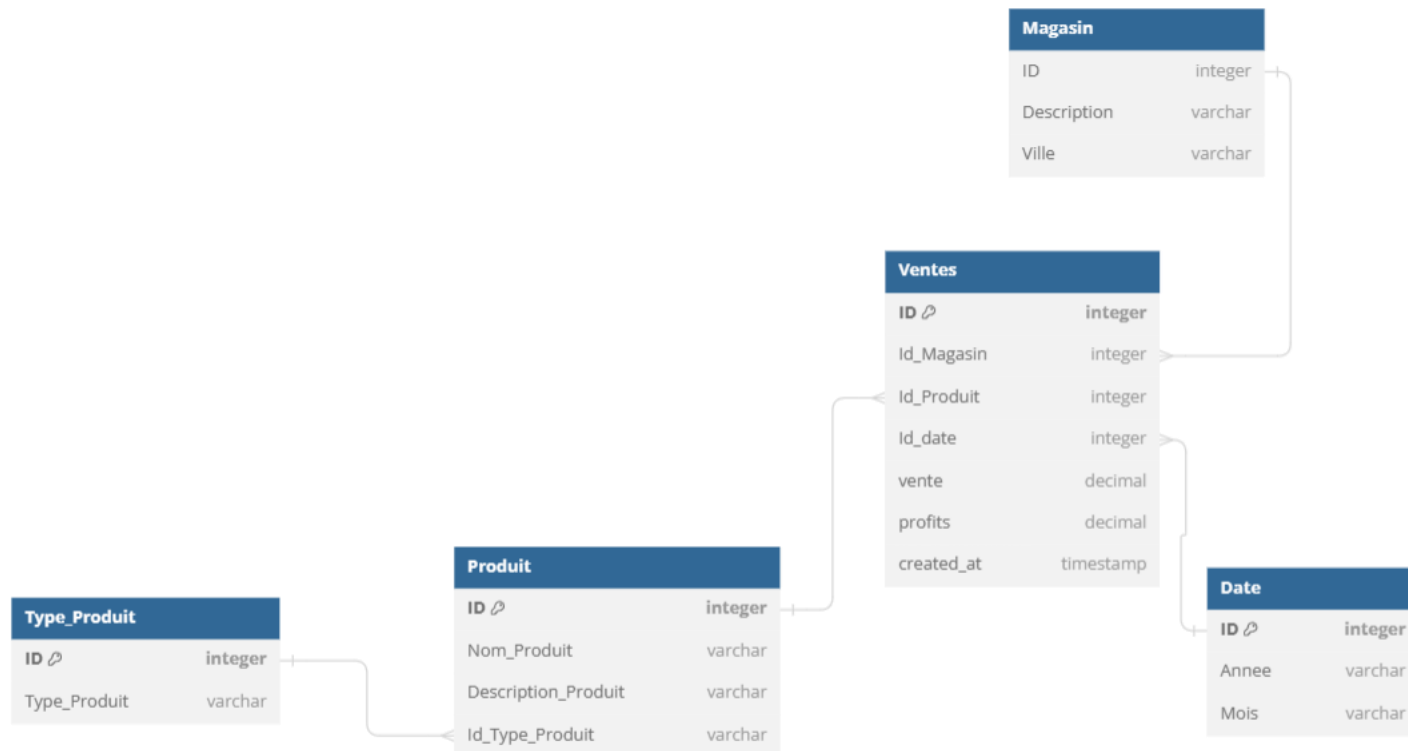
Chaque dimension est directement reliée à un fait

Pas de liaison entre deux dimensions



Architecture BI – Modélisation en flocon

Modélisation en **Flocon** : certaines dimensions peuvent être reliées à d'autres dimensions créant ainsi une hiérarchie.



Données multidimensionnelles

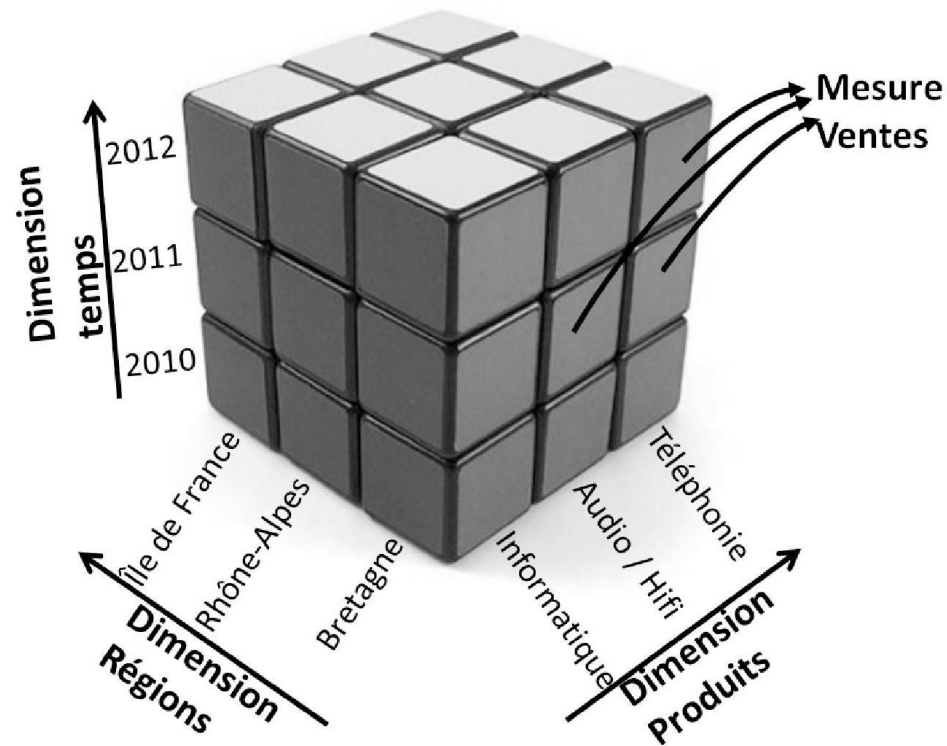
Les DWH sont basés sur les bases de données multidimensionnelle (cube ou hypercube).

Les données y sont vues comme des data cubes.

Les axes sont les dimensions et les mesures sont les données du cube.

Exemple : calcul du chiffre d'affaires par client, par produit et par zone géographique.

Données multidimensionnelles

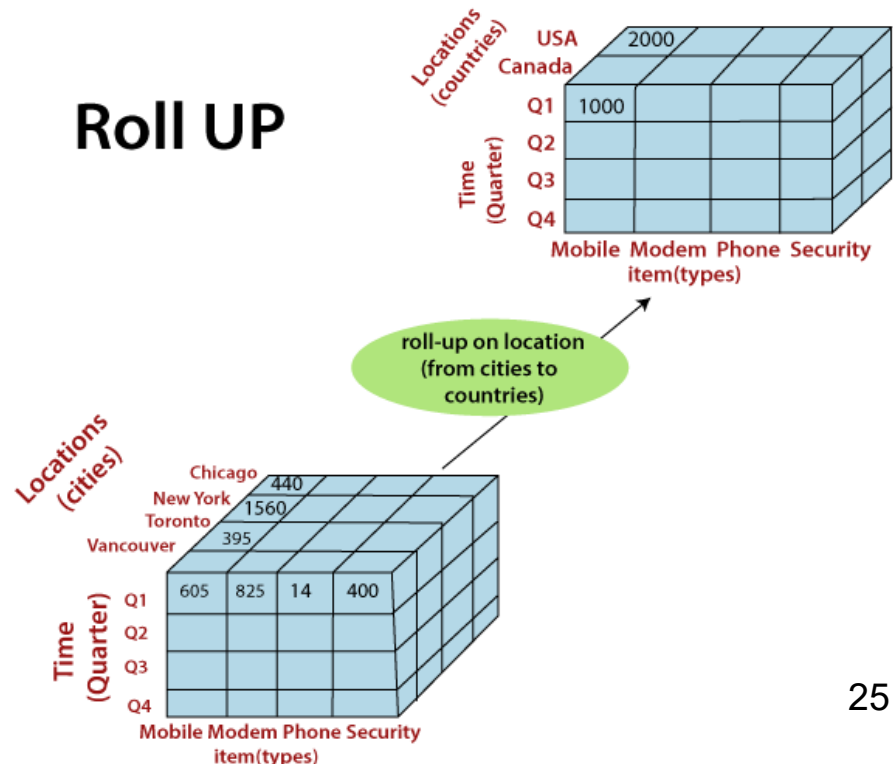


Données multidimensionnelles

Opération sur les cubes:

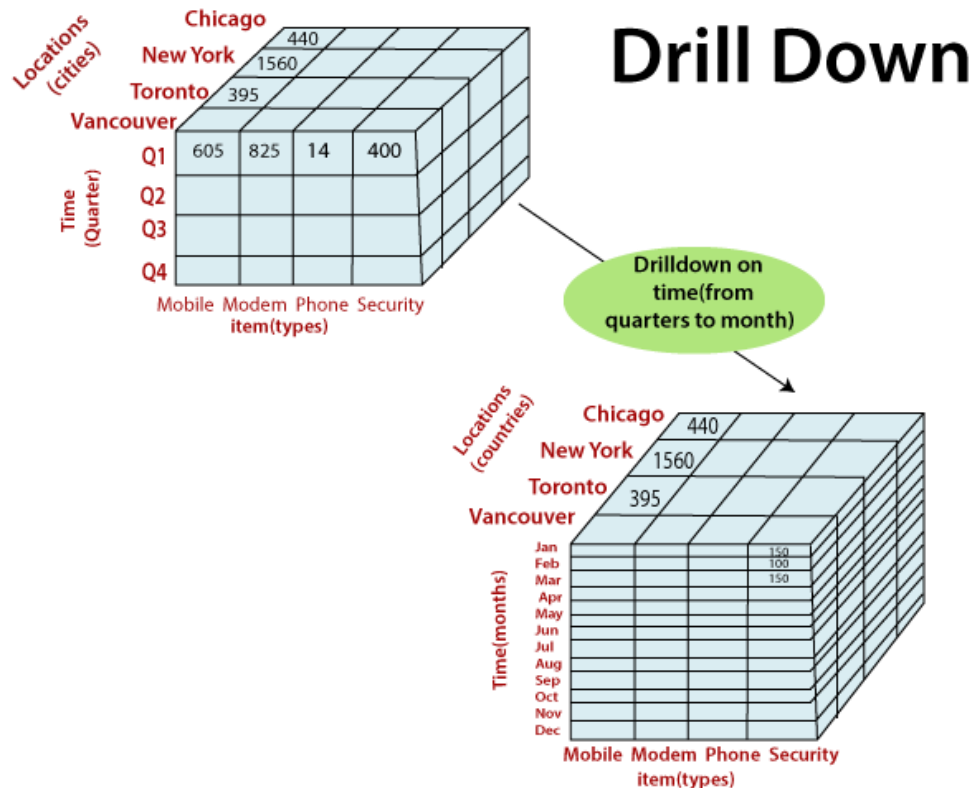
Roll-up (drill up) : opération qui permet de rassembler les données

Roll UP



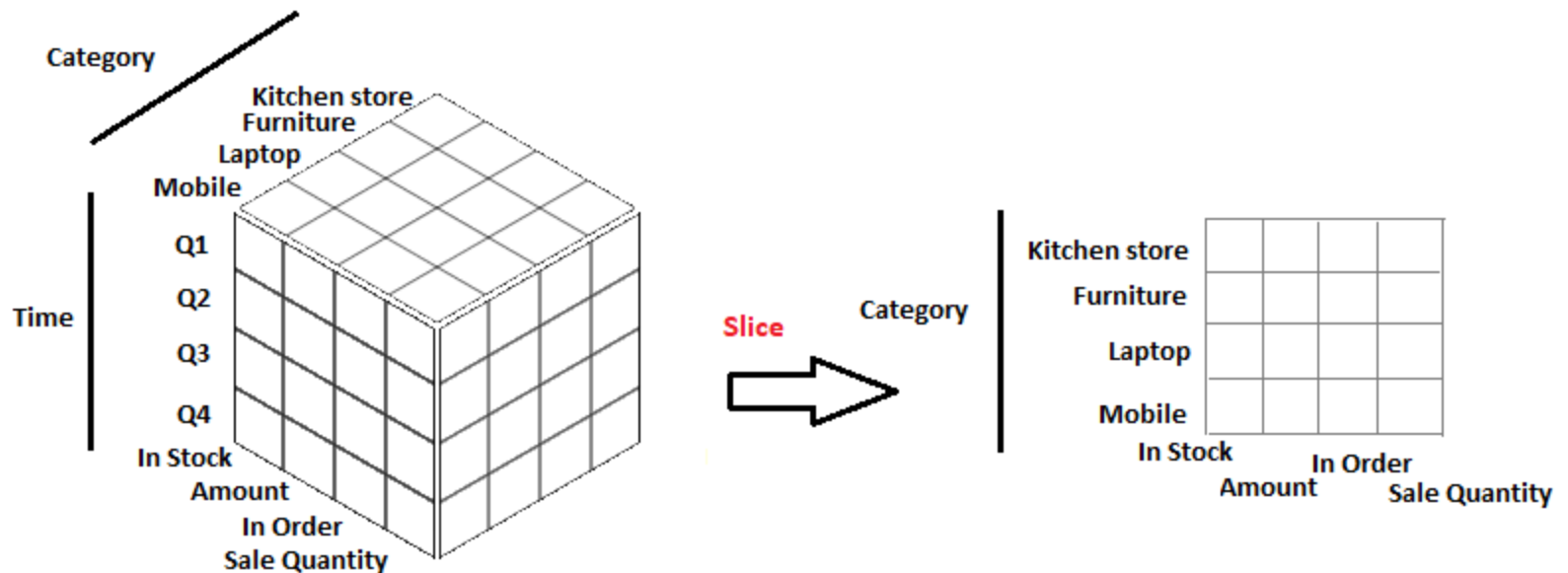
Données multidimensionnelles

Drill-down: opération qui permet d'affiner l'analyse.



Données multidimensionnelles

SLICE: opération qui consiste pour une valeur d'un axe, d'obtenir le sous cube (dimension 2)



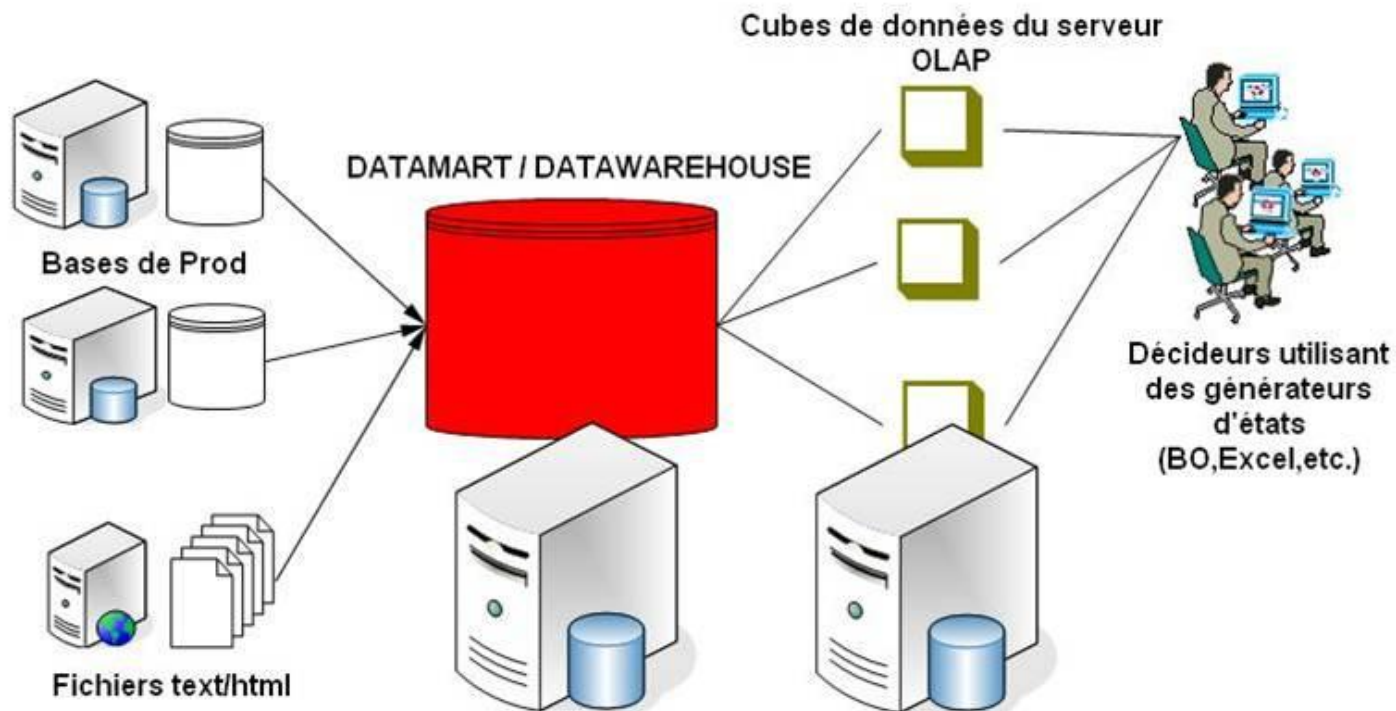
Outil d'analyse et de restitution

Un fois l'entrepôt de données mis en place → visualiser les données pour prise de décisions.

Outils de visualisation : PowerBI

Architecture BI

Architecture BI



Fin