Esprit School of Business
ESB

Modélisation Multidimensionnelle Entrepôt de Données

Préparé et Présenté par :
Dr. Mohamed Aymen Ben HajKacem

Année universitaire 2021-2022

Avant de commencer

- Contact : medaymen.hajkacem@gmail.com
- Page web du cours : moodle
  - Planning, dates d'Exam ;
  - Transparents de cours, sujets de TP et de TD ;
  - Bibliographie et liens utiles ;
- 14 Séances de cours 3h (une par semaine)
  - 2LBI1 + 2LBI2 ;
- Contrôle des connaissances :
  - Contrôle continu (Quiz, Ex. à rendre, orale, TD) ;
  - 1 Projet ;
  - 1 Examen ;

Avant de commencer

Objectifs

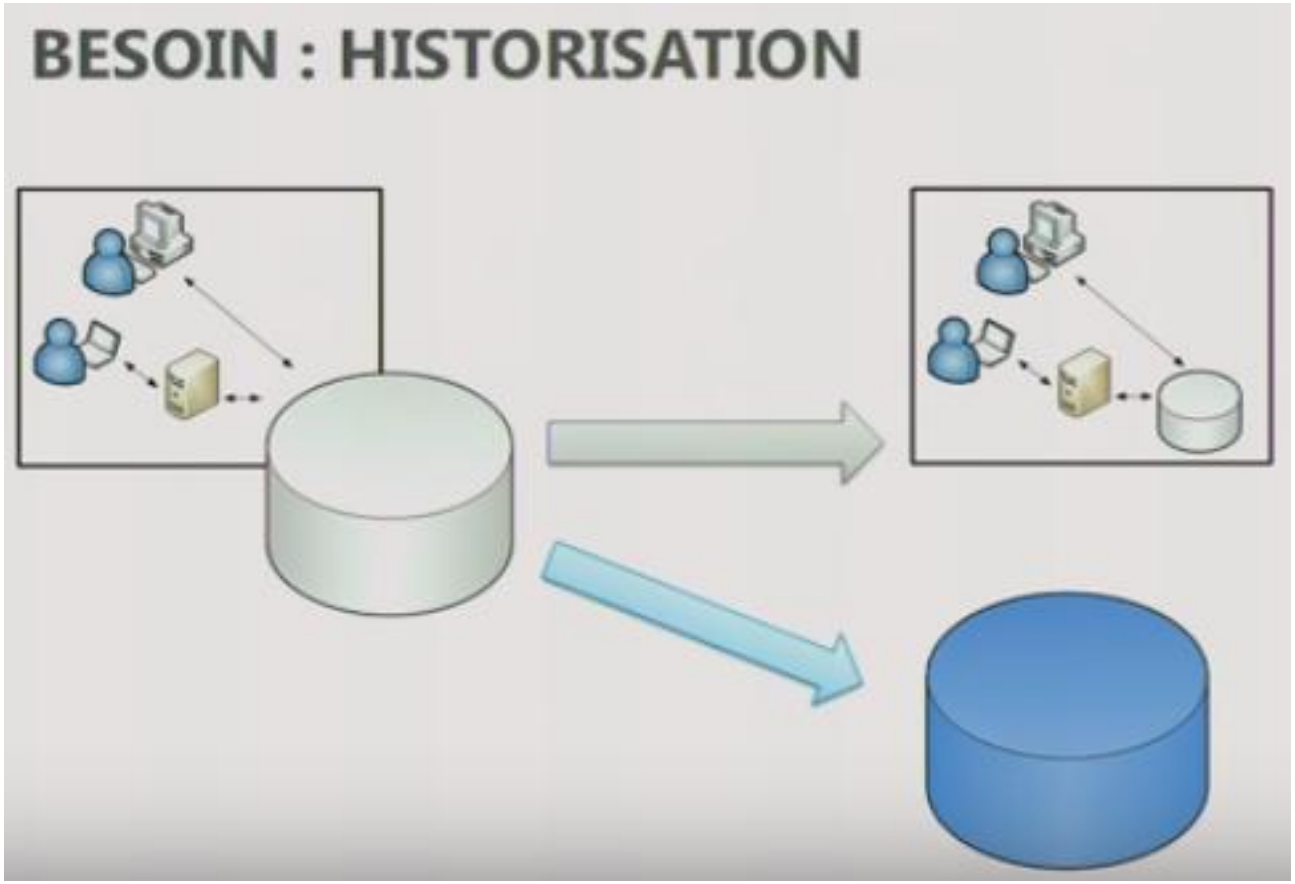
- Découvrir les principaux concepts du Business Intelligence.
- Apprendre la modélisation multidimensionnelle.
- Découvrir les principaux concepts de l'ETL.
- Apprendre à installer et utiliser Talend.
- Apprendre à installer et utiliser Power BI.

Prérequis

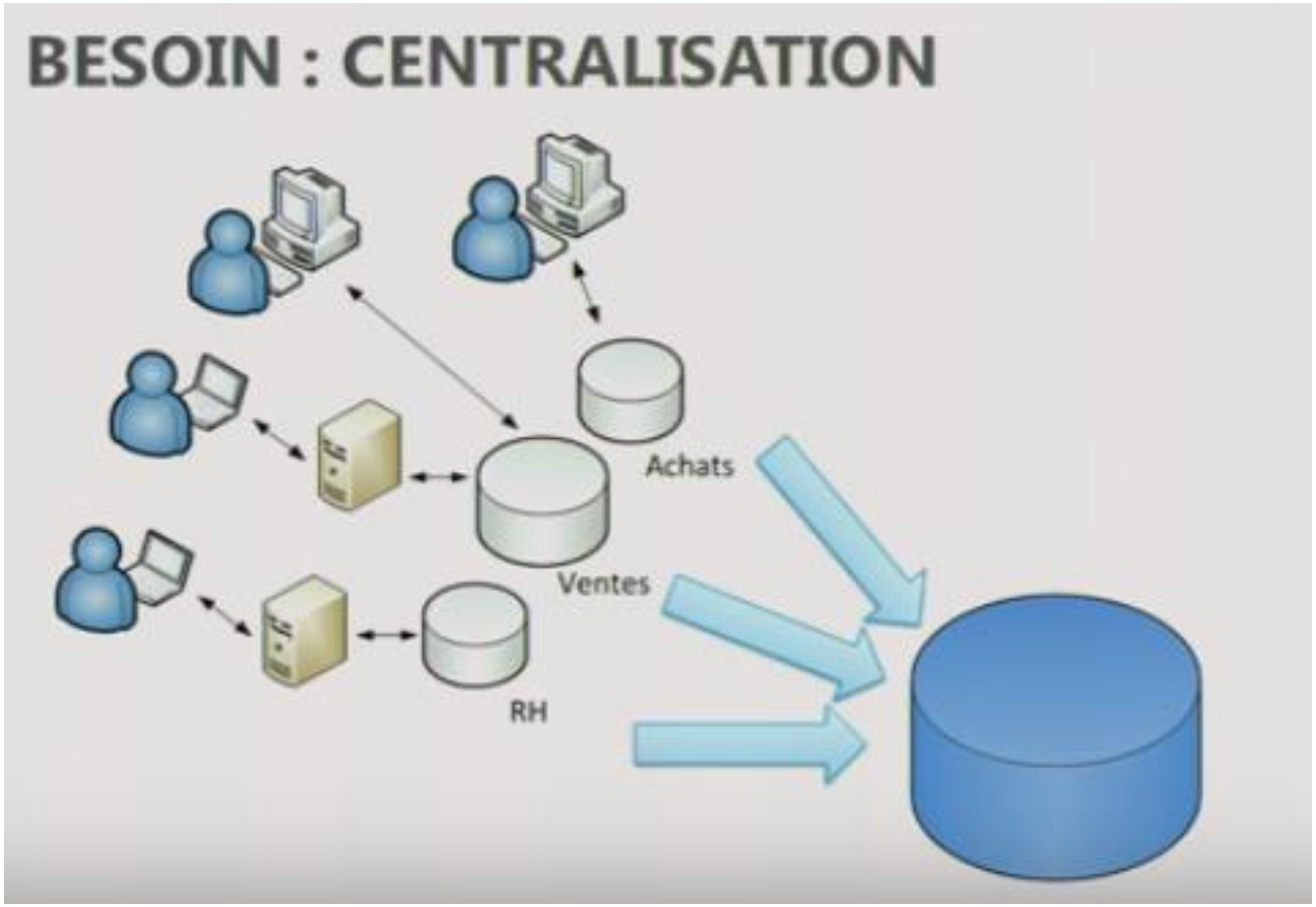
- Langage SQL
- Langage Java
- Conception des bases de données

Plan du cours

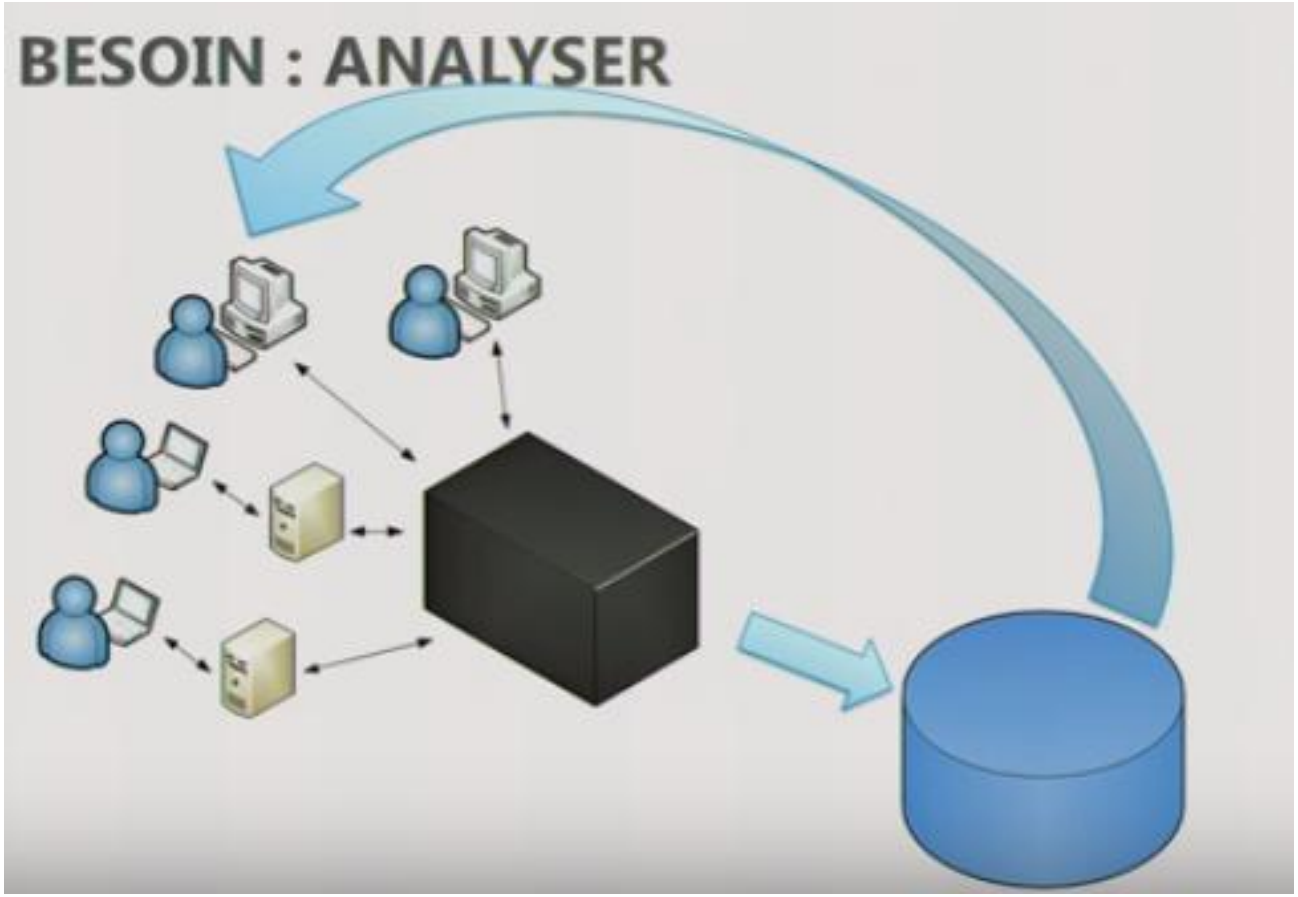
Context



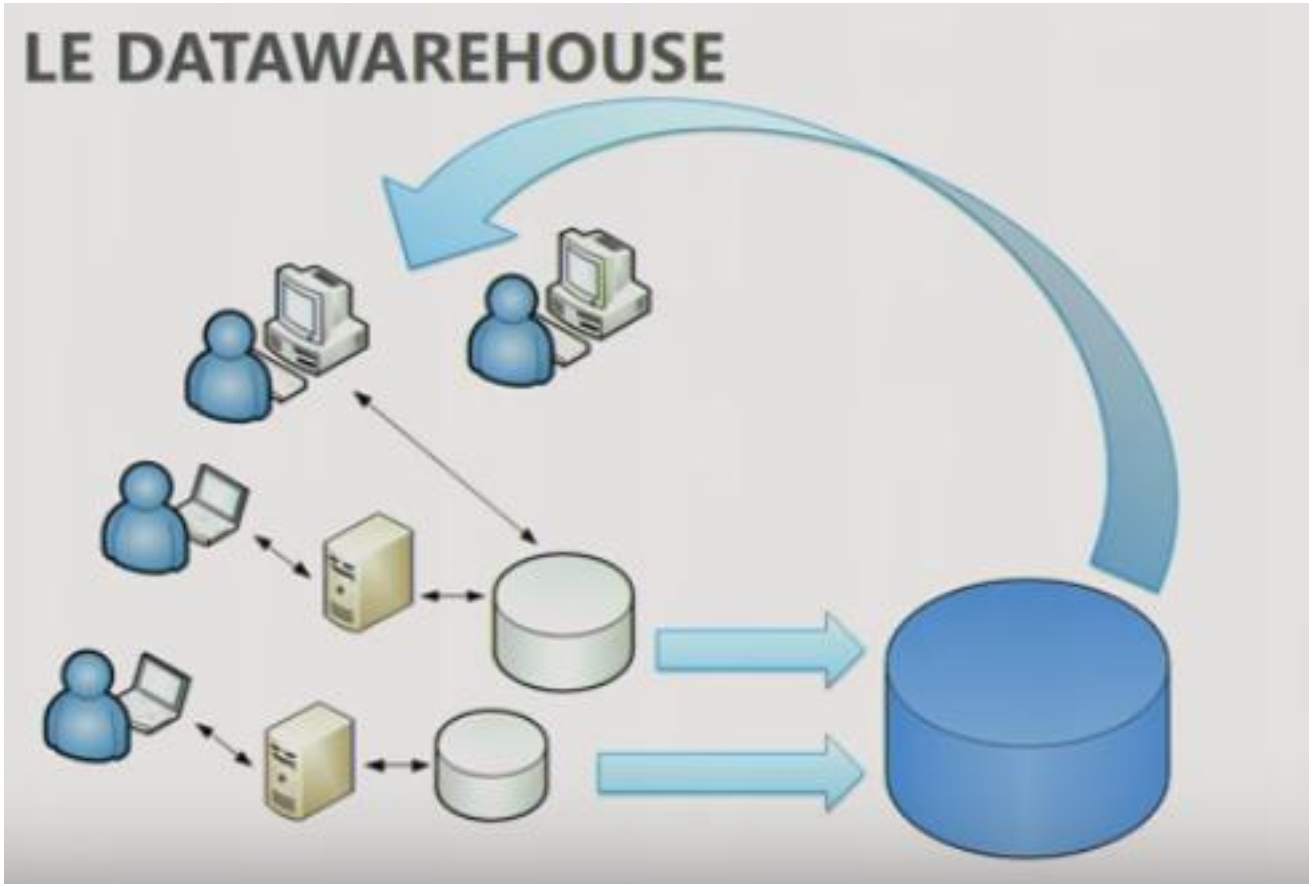
Context



Context



Context



- Besoin des entreprises :
  - Accéder facilement à toutes les données de l'entreprise
  - Analyser et organiser les données provenant des différentes sources.
  - Retrouver les informations pertinentes à la prise de décision.
  - Appliquer des prévisions pour mieux orienter les choix de l'entreprise.
  - Prendre rapidement des décisions stratégiques et tactiques.


Qui sont mes meilleurs clients?
Pourquoi et comment le chiffre d'affaire a baissé?
Quels français consomment beaucoup de poisson?
A combien s'élèvent mes ventes journalières?

Applications

- Commerce
  - Ciblage de clientèle,
  - Déterminer les promotions,
  - Aménagement des rayons (2 produits en corrélation).
- Banque
  - Déterminer les profils client,
  - Gestion des portefeuilles clients.
- Mailing
  - Amélioration du taux de réponse
  - Détection des e-mails spam

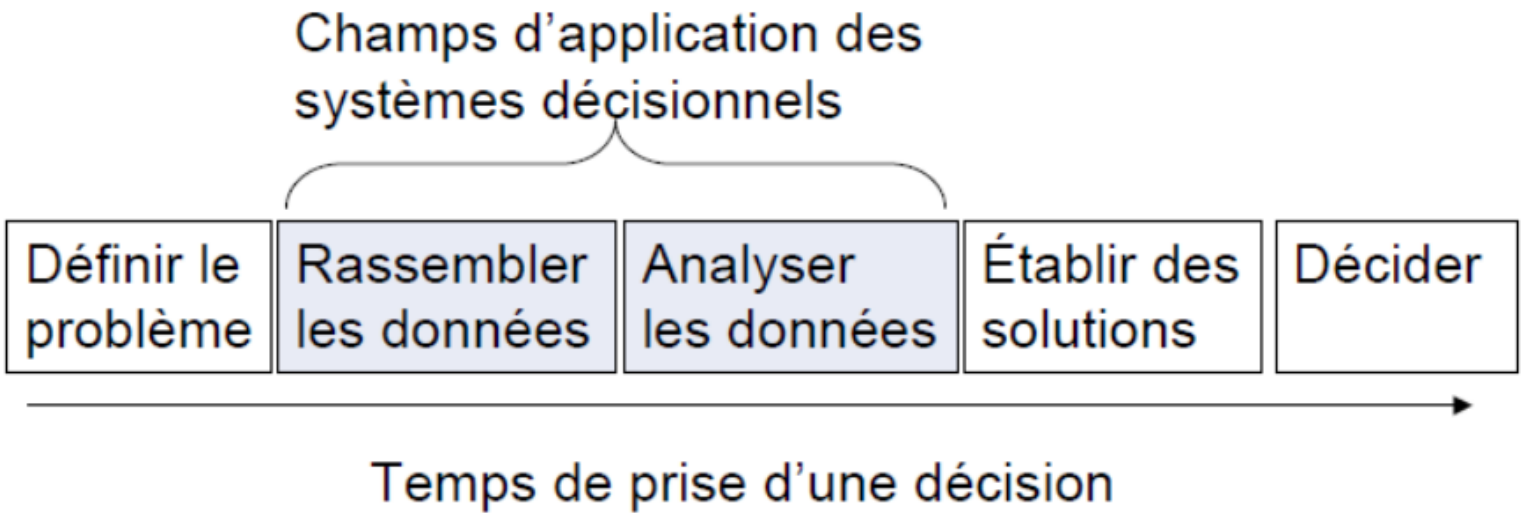
Prise de décision

- Comment faciliter la prise de décision ?
- Utiliser les données de l'entreprise et donner un accès rapide et simple à l'information.
- Produire une vision transversale sur les données sources.
- Intégrer des différentes bases de données.
- Disposer des outils performants à l'exploration et l'analyse des données.

Prise de décision

- Solution : Mettre en place un entrepôt de données (Data Warehouse)
- Transformer les données de production en informations stratégiques $\Longrightarrow$ Prise de décision

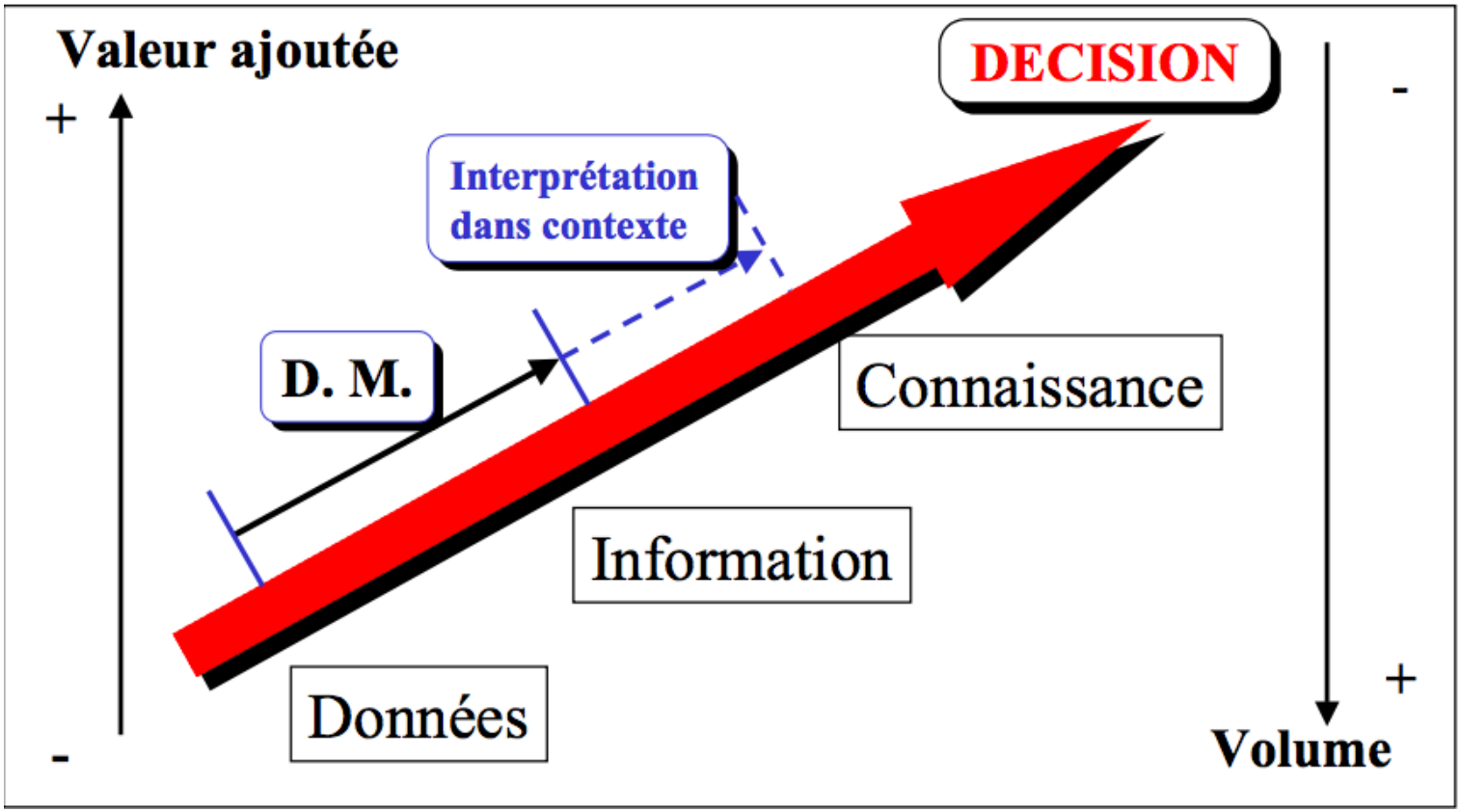
Processus de prise de décision



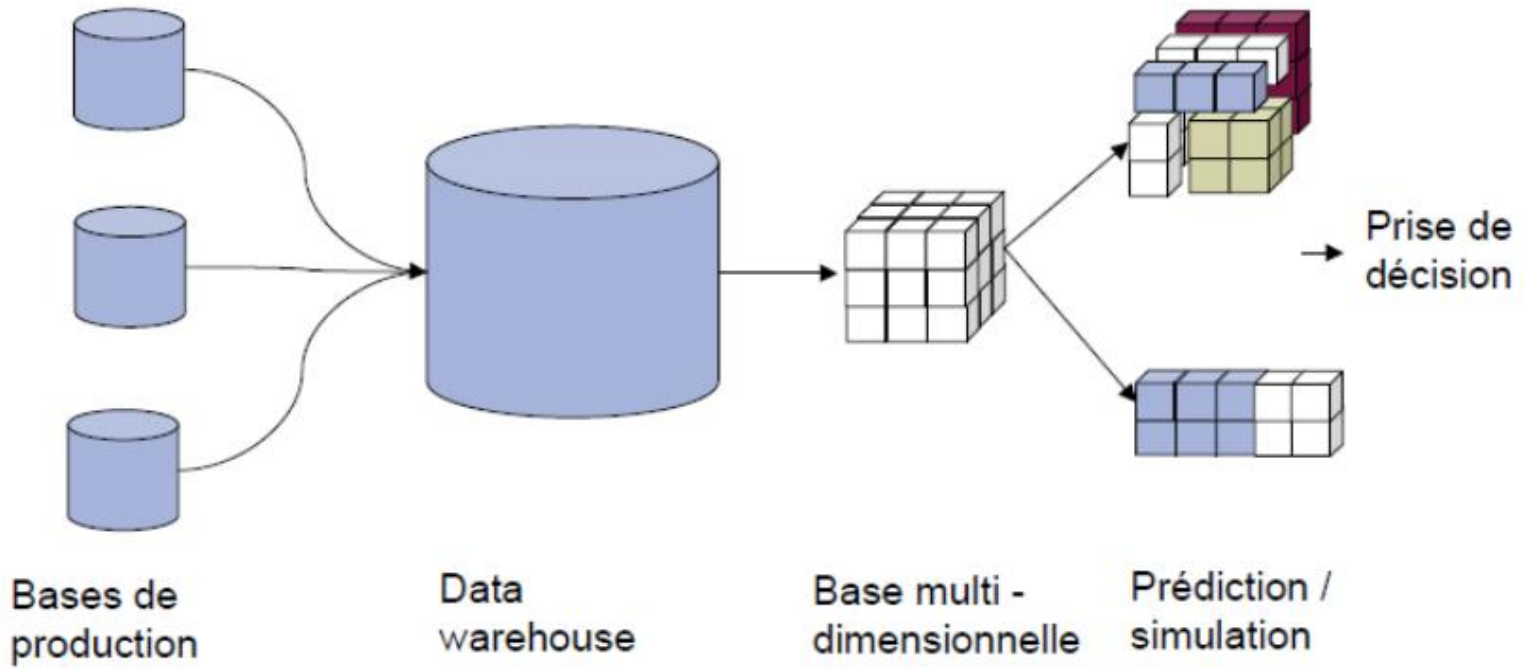
Données $\Longrightarrow$ Décisions

- Données : Points de ventes, géographiques, démographiques, ...
- Informations : I vit dans R, I est âgé de A, ...
- Connaissances : Dans X%, le produit Y est vendu en même temps que le produit Z, ...
- Décisions : Lancer la promotion de Y Z dans R auprès des clients plus âgés que A, ...

Données $\Longrightarrow$ Décisions



Processus de la prise de décision (DW)



Pour quoi ne pas SGBD ?

SGBD Vs DW

SGBD

- Fonctions : insérer, modifier, interroger les données.
- Accès à de nombreux utilisateurs simultanément.
- Processus transactionnels en ligne / **O**n-**L**ine **T**ransactional **P**rocessing (OLTP).
- Exemple : Le client a débité 1000dt de son compte.

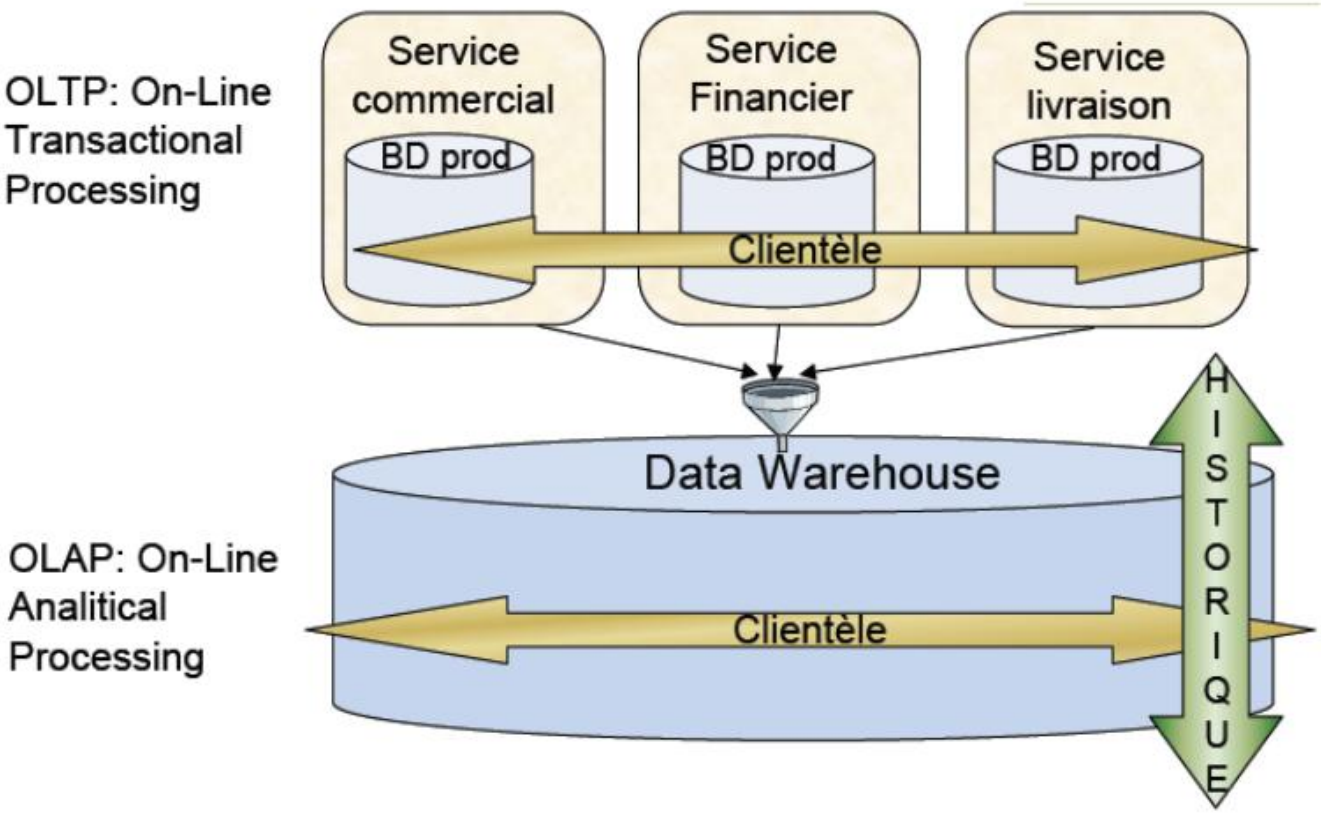
DW

- Fonctions : Regrouper, organiser des informations provenant de sources diverses.
- Accès seulement aux analystes (décideurs).
- Processus analytiques en ligne / **O**n-**L**ine **A**nalytical **P**rocessing) (OLAP)
- Exemple : Quel est le volume de ventes dans la période 2017-2018 ?

SGBD Vs DW

	SGBD	**DW**
Utilisateurs	Nombreux Employés	Peu Analystes
Données	Alphanumériques Détaillées Orientées applications Dynamiques	Numériques Agrégées/Resumées Orientées sujets Statiques
Requétes	Préfdéfines	One-use
Accés	Peu de données Courantes	Beaucoup d'informations Historisées
Temps	Court	Long
Mise à jour	Trés souvent	Périodique

SGBD VS DW


OLTP: On-Line Transactional Processing
Service commercial
BD prod
Service Financier
BD prod
Service livraison
BD prod
Clientèle
Data Warehouse
HISTORIQUE
OLAP: On-Line Analitical Processing
Clientèle

Définition d'un DW

Définition de Bill Inmon (1996)

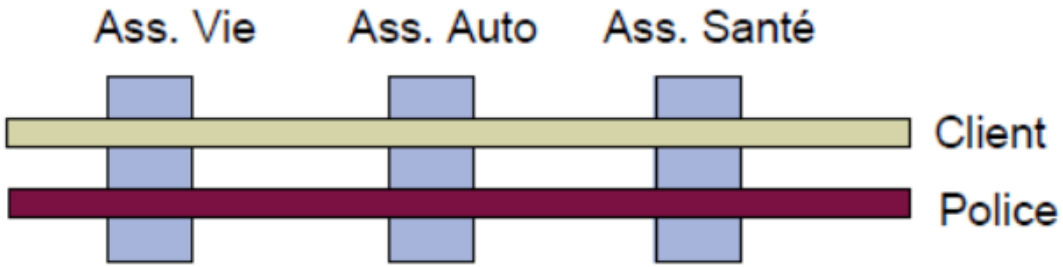
- Le Data Warehouse est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision.

Caractéristiques

- Une Base de données utilisée pour la prise de décision.
- Caractéristiques : orintée sujets, intégrée, non volatile et historisée.

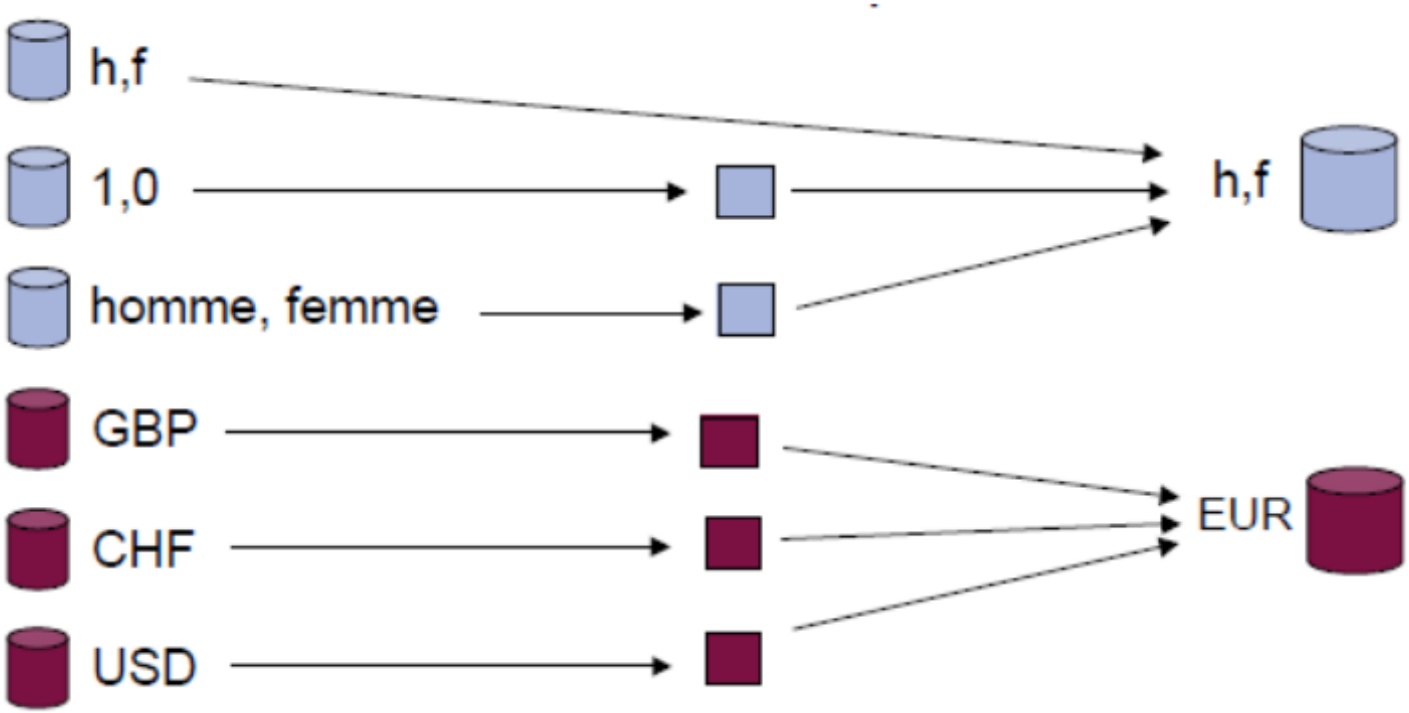
Caractéristiques d'un DW

- Données orientées sujet :
  - Fusionner les données issues des différents métiers de entreprise.
  - Données pertinentes pour un sujet d'analyse.



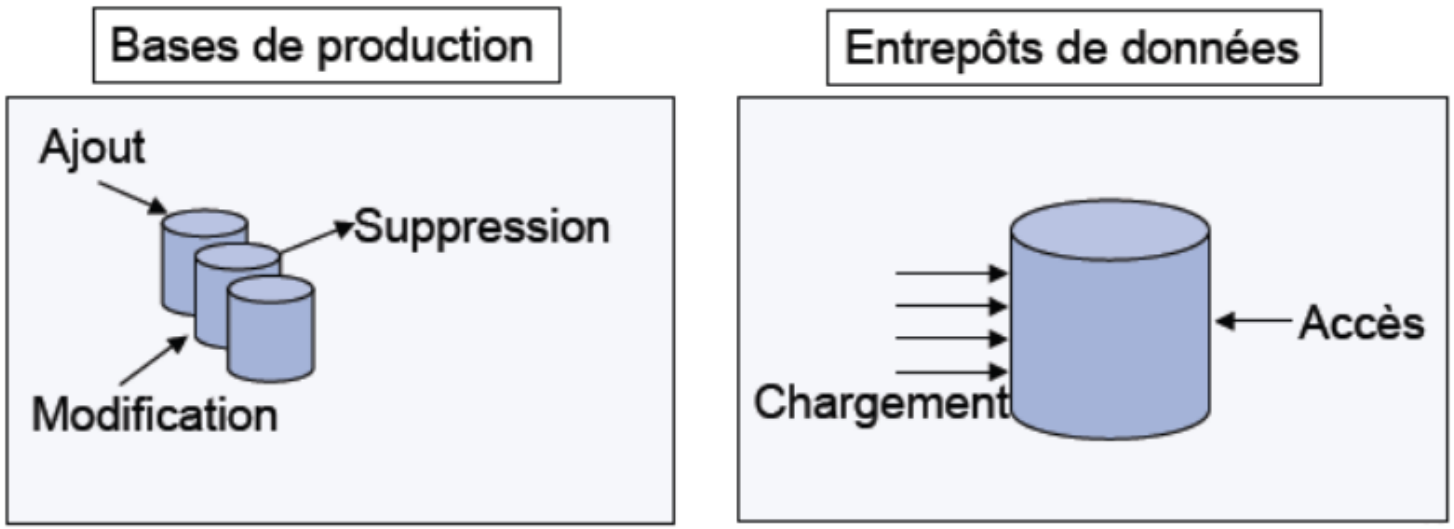
Caractéristiques d'un DW

- Données intégrées :
  - Représente les informations provenant de différentes sources qui sont hétérogènes.



Caractéristiques d'un DW

- Données non volatiles
  - Représente l'activité de l'entreprise durant une certaine période.
  - Traçabilité des informations et des décisions prises.



Caractéristiques d'un DW

- Données historisées :
  - Les données enregistrées dans le temps et sont utilisées en mode consultation (Pas de mise à jour).
  - Mise en place d'un référentiel temps.

Base de production

Image de la base en Mai 2005

Répertoire

Nom	Ville
Dupont	Paris
Durand	Lyon

Image de la base en Juillet 2006

Répertoire

Nom	Ville
Dupont	Marseille
Durand	Lyon

Entrepôt de données

Calendrier

Code	Année	Mois
1	2005	Mai
2	2006	Juillet

Répertoire

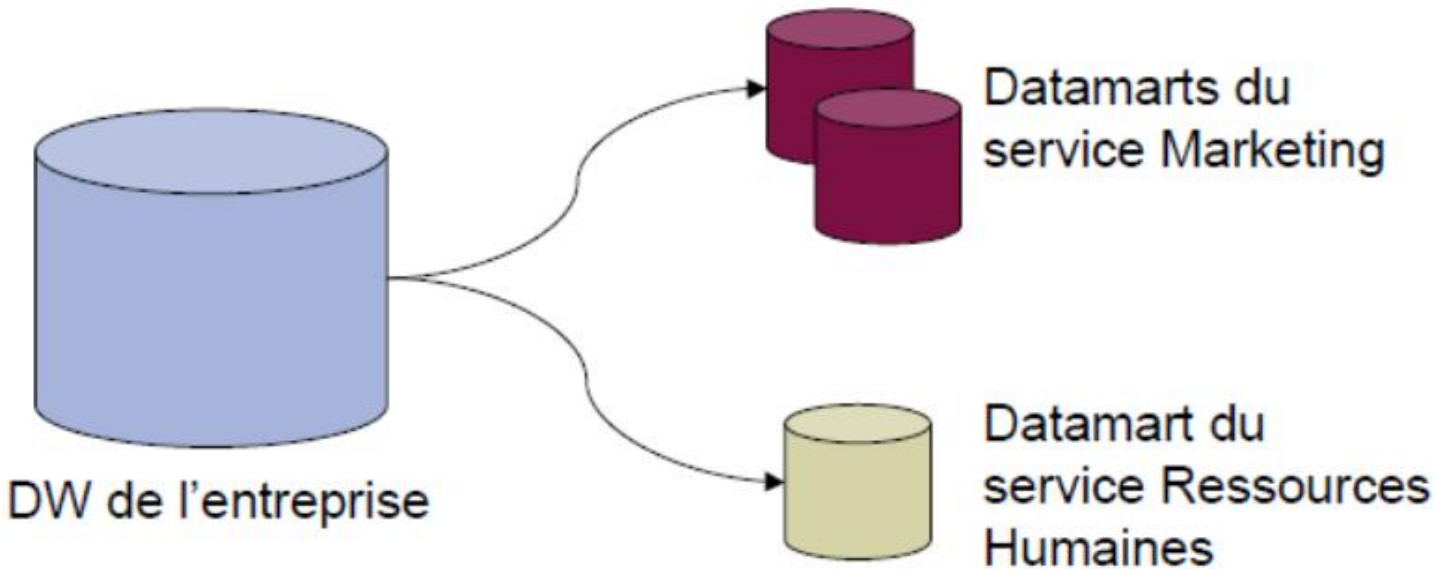
Code	Année	Mois
1	Dupont	Paris
1	Durand	Lyon
2	Dupont	Marseille

Exemple : un DW dans les télécoms

- Sujets :
  - Comportement du marché.
  - Comportement de la clientèle.
  - Comportement du réseau.
- Historique :
  - 5 ans pour le suivi du marché.
  - 1 an pour le comportement de la clientèle.
  - 1 mois pour le comportement du réseau.
- Source :
  - Fiches clients élaborés par les agences.
  - Fichiers de facturation.
  - Fichiers issus de CRM.
- Requêtes :
  - Nombre moyen d'heures par client, par mois et par région.
  - Durée moyenne d'une communication téléphone par client.
  - Durée moyenne d'une communication internationale par ville.

Data Mart (DM)

- Les magasins de données (sous-ensemble d'un entrepôt de données).
- Destiné à fournir des données pour un secteur ou une fonction particulière de l'entreprise.



Data Warehouse (DW) VS Data Mart (DM)

DW

- Collecte l'ensemble de l'information utile aux décideurs.
- Centralise l'information décisionnelle en assurant l'intégration des données extraites.

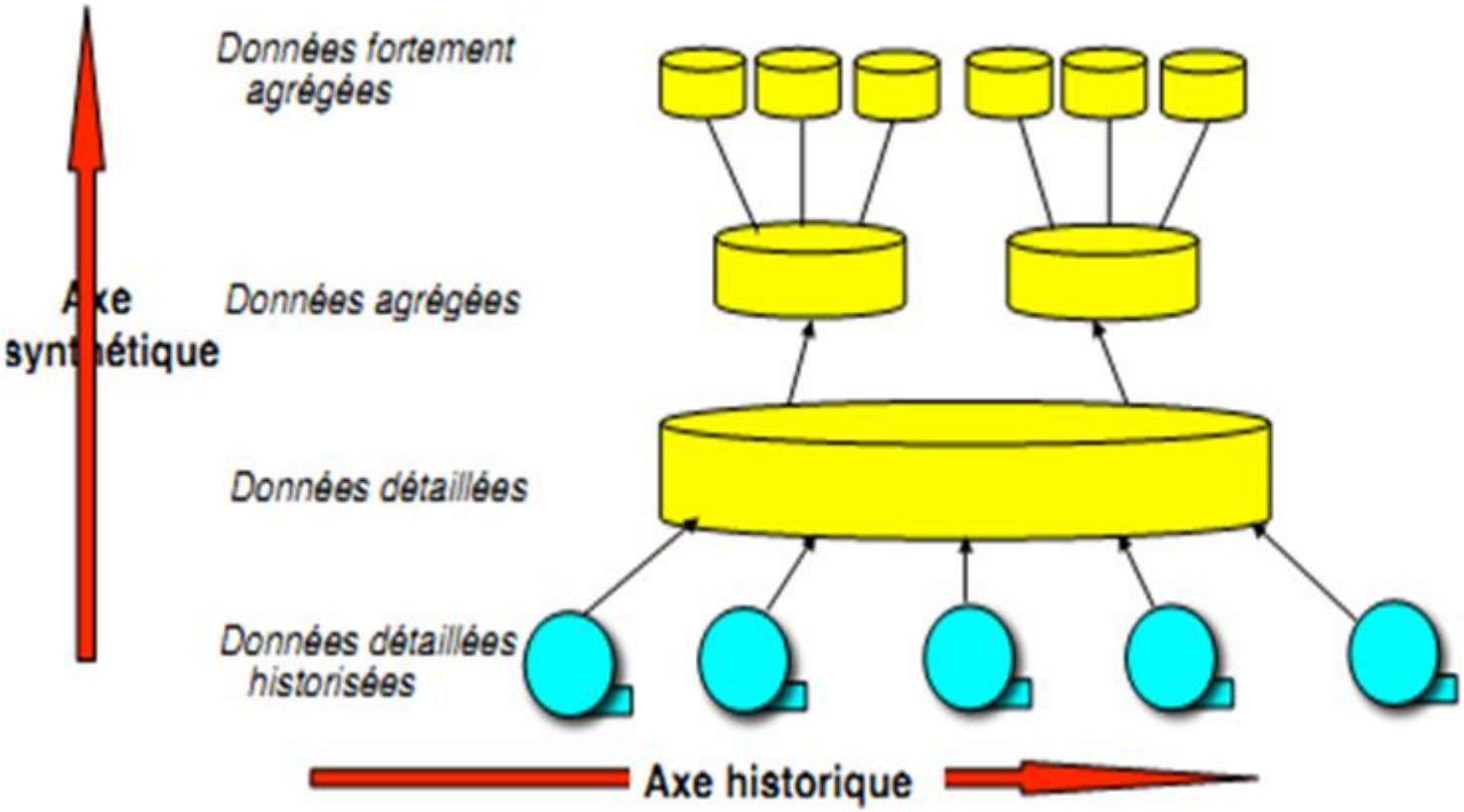
DM

- Extraire pour chacune classe d'utilisateurs une partie de l'information décisionnelle de l'entrepôt.
  - Un besoin d'analyse spécifique.

	DW	**DM**
Objectif	facilite la gestion des données	facilite les traitements décisionnels
Données	Centralisées	Adaptées
Taille	Grande	Moyenne
Mise à jour	Des machines puissantes	infrastructure plus légère

Axes historique et synthétique des données d'un DW

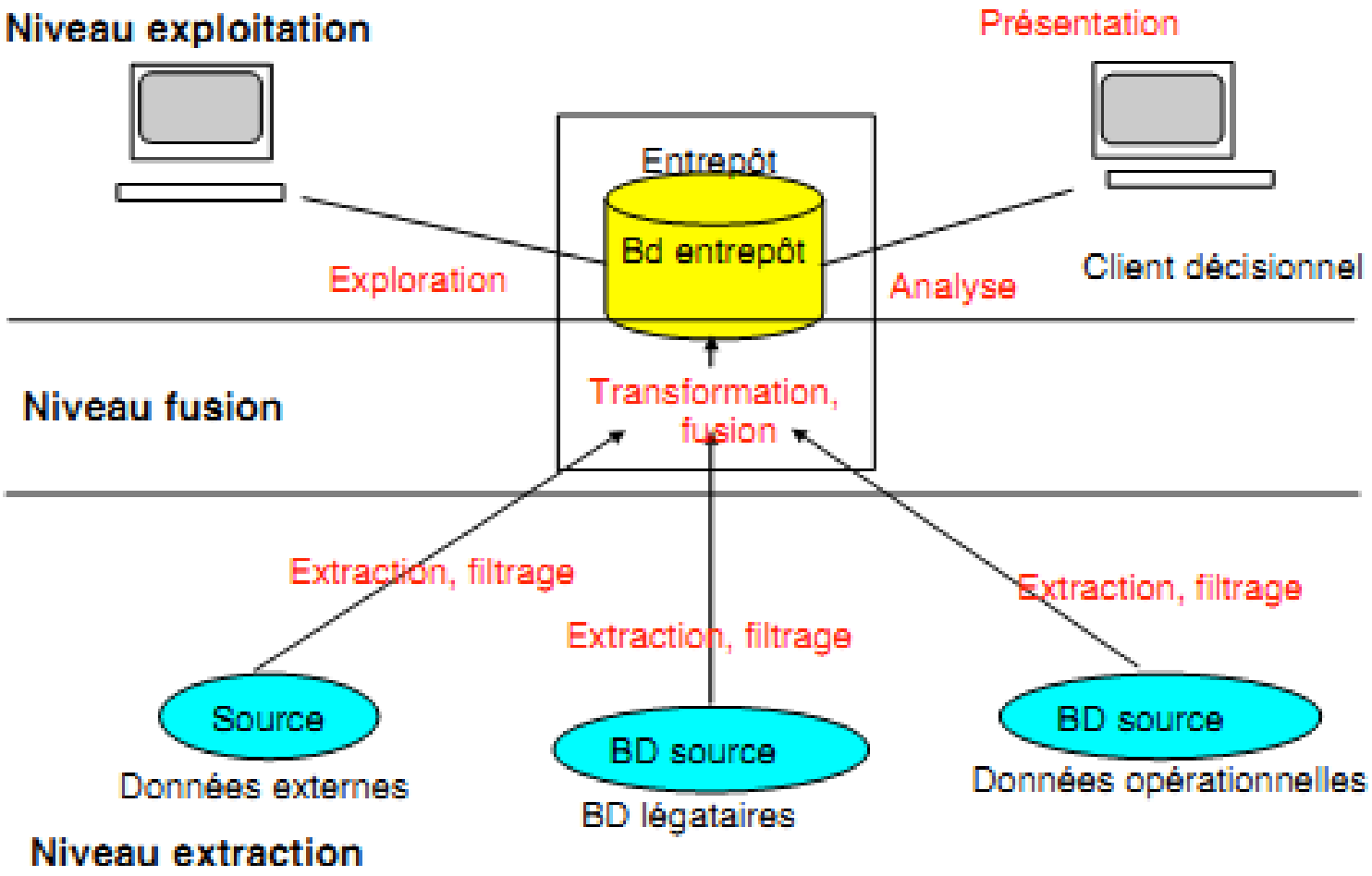
- Les données dun DW s'organisent autour de 2 axes : synthétique et historique.



Axes historique et synthétique des données d'un DW

- Axe synthétique : représente une hiérarchie dagrégation entre les données comprenant :
  - les données détaillées définies par les événements les plus récents.
  - les données agrégées définies par des résumés des données détaillées.
  - les données fortement agrégées qui synthétisent à un niveau supérieur les données agrégées.
- Axe historique : représente les données détaillées historisées dans un intervalle de temps.

Architecture fonctionnelle



Modélisation multidimensionnelle

Principe

- Modélisation des données pour supporter efficacement les opérations OLAP.
- Organisé autour des sujets majeurs, tels que vente, client,. . .
- Sujet = Faits + Dimensions
- Consiste à considérer un sujet (fait) comme un point dans un espace à plusieurs dimensions.
- Fournit une vision simple, concise et multidimensionnelle sur des sujets particuliers.

Modélisation multidimensionnelle

Démarche

- 3 Niveaux d'abstraction identiques à la modélisation Relationnelle.
  1. Niveau conceptuel : Analyse des besoins de décideurs (Quoi ?)
     - Description de la base multidimensionnelle indépendamment des choix d'implantation.
  2. Niveau logique : Mode de Stockage (Comment ?)
     - Description de la base multidimensionnelle suivant la technologie utilisée.
  3. Niveau physique : Choix du logiciel (Avec quel outil ?)
     - Implantation physique qui dépend du logiciel utilisé.

Niveau conceptuel

Principe

- Description du Data Warehouse indépendamment de la stratégie d'implémentation.
- Définition les éléments autour des :
  - Faits et mesures.
  - Dimensions et hiérarchies.

Fait / Mesure

Principe

- Un fait :
  - modélise le sujet d'analyse
  - est formé d'un ensemble d'attributs appelés mesures (indicateurs de performance).
- Mesure :
  - Est un attribut numérique sur lequel portent les analyses, en fonction des différents axes d'analyse.
  - Ces valeurs sont définies via des fonctions d'agrégation sur les données.
  - Exemples : Coût des travaux, Nombre d'accidents, Chiffre d'affaire, ...

Dimension / Hiérarchie

Principe

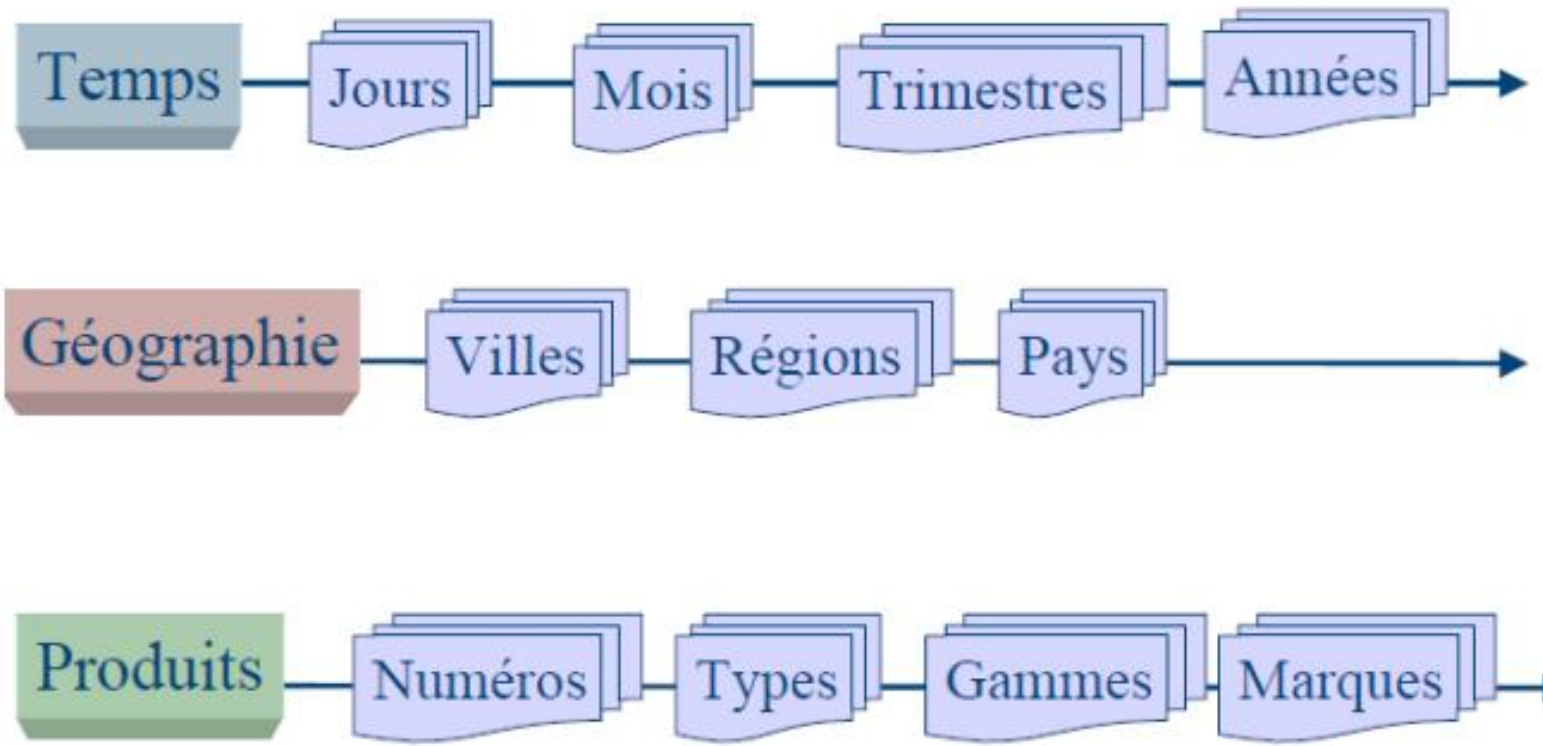
- Le sujet (fait) est analysé selon différents axes qui caractérisent les mesures $\Longrightarrow$ Dimension.
- Une dimension :
  - modélise un axe d'analyse et comporte un ou plusieurs attributs/membres.
  - Les attributs/membres d'une dimension sont définies selon une organisation hiérarchique.

Dimension / Hiérarchie

Principe

- Chaque attribut représente un niveau hiérarchique (ou niveau de granularité) bien déterminé .
- Exemples :
  - Dimension Temps : jour, semaine, mois, année.
  - Dimension Localisation : magasin, ville, région, pays.
  - Dimension Produit : produit, catégorie, sous-catégorie.

Dimension / Hiérarchie



Exemple

- 150 000 euros est le coût des travaux pour le membre 2020 de l'attribut Année de la dimension Temps et le membre Versailles de l'attribut Ville de la dimension Localisation.
- Fait : Travaux.
- Mesure : Coût des travaux.
- Dimensions : Temps, Localisation.

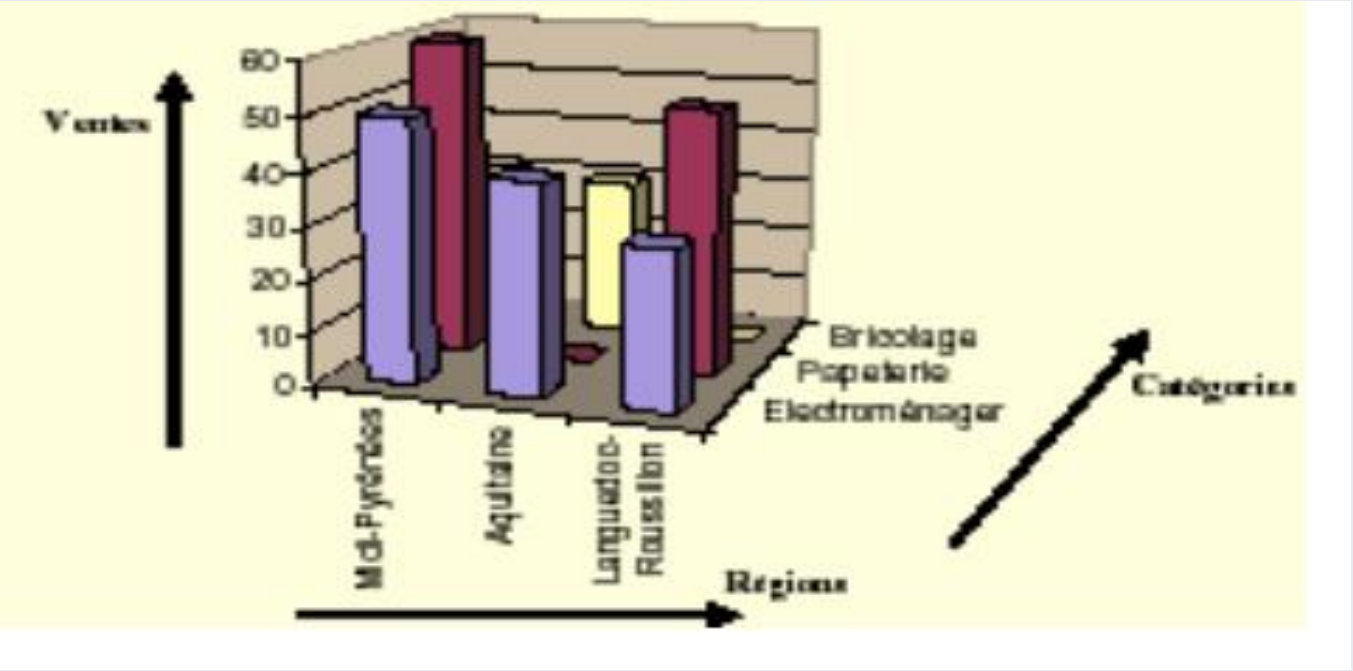
Exemple

- Les données suivantes comportent les ventes de 1999 d'une entreprise de distribution :

Catégories des produits	Régions	Montant des ventes
Electroménager	Midi-Pyrénées	50
Electroménager	Aquitaine	40
Electroménager	Languedoc-Roussillon	30
Papeterie	Midi-Pyrénées	60
Papeterie	Languedoc-Roussillon	50
Bricolage	Midi-Pyrénées	30
Bricolage	Aquitaine	30

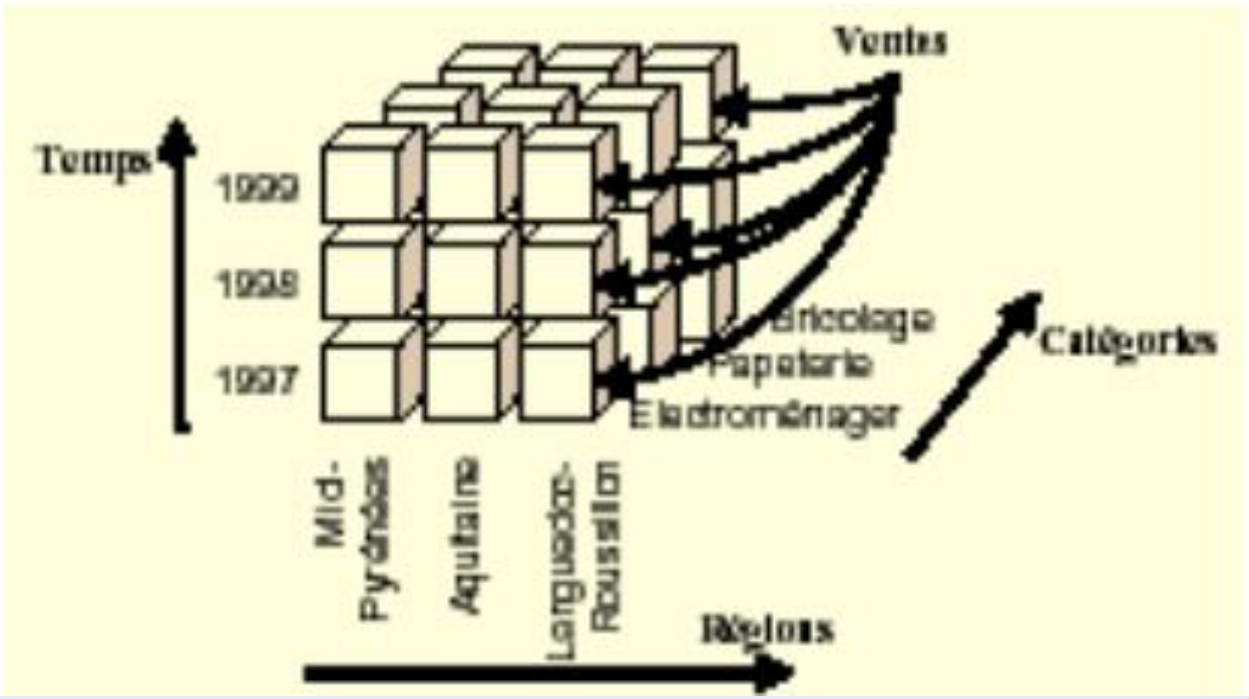
Exemple

- On peut définir différents axes pour analyser ces données :
  - une dimension catégorie des produits.
  - une dimension région.



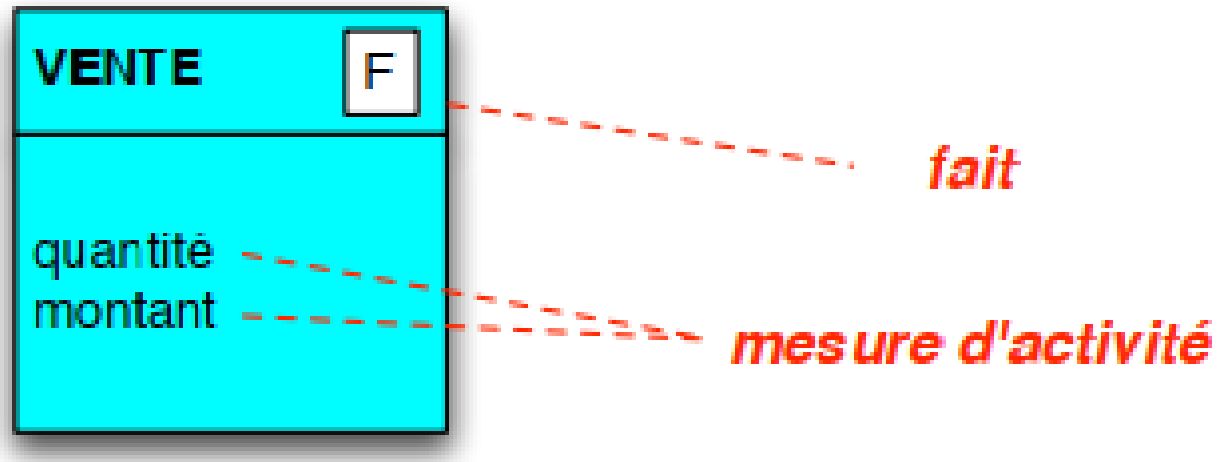
Exemple

- On peut alors analyser les données dans un espace à 3 dimensions :
  - la dimension produit
  - la dimension localisation
  - la dimension temps
- Chaque intersection de ces dimensions représente un élément qui définie le montant des ventes :



Exemple

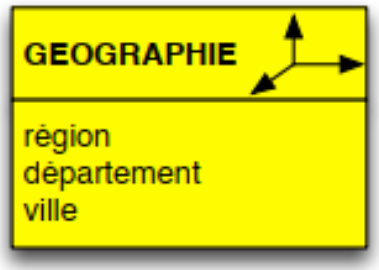
- Le fait Vente est défini par des mesures suivantes :
  - quantité de produits vendus et
  - montant total des ventes

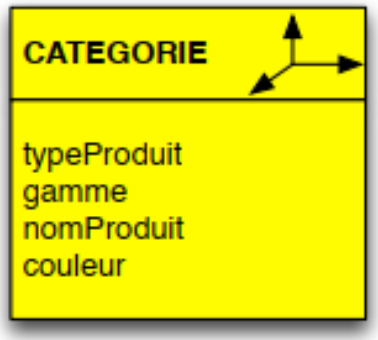


Exemple

Principe

- Le fait Vente peut être analysé suivant les différentes dimensions suivantes :
  - la dimension Temps,
  - la dimension Localisation et
  - la dimension Produit :





Exemple

- La hiérarchie d'une dimension :
  - Dimension « temps» : jour / mois / trimestre / année
  - Dimension « géographie » : ville / département / région
  - Dimension « catégorie» : couleur / nomProduit / gamme / typeProduit

Niveau logique

Principe

- Description de la base multidimensionnelle suivant la technologie utilisée.
- 3 technologie pour créer un DW.
  - Relational OLAP (ROLAP)
  - Multidimensional OLAP (MOLAP)
  - Hybrid OLAP (HOLAP)

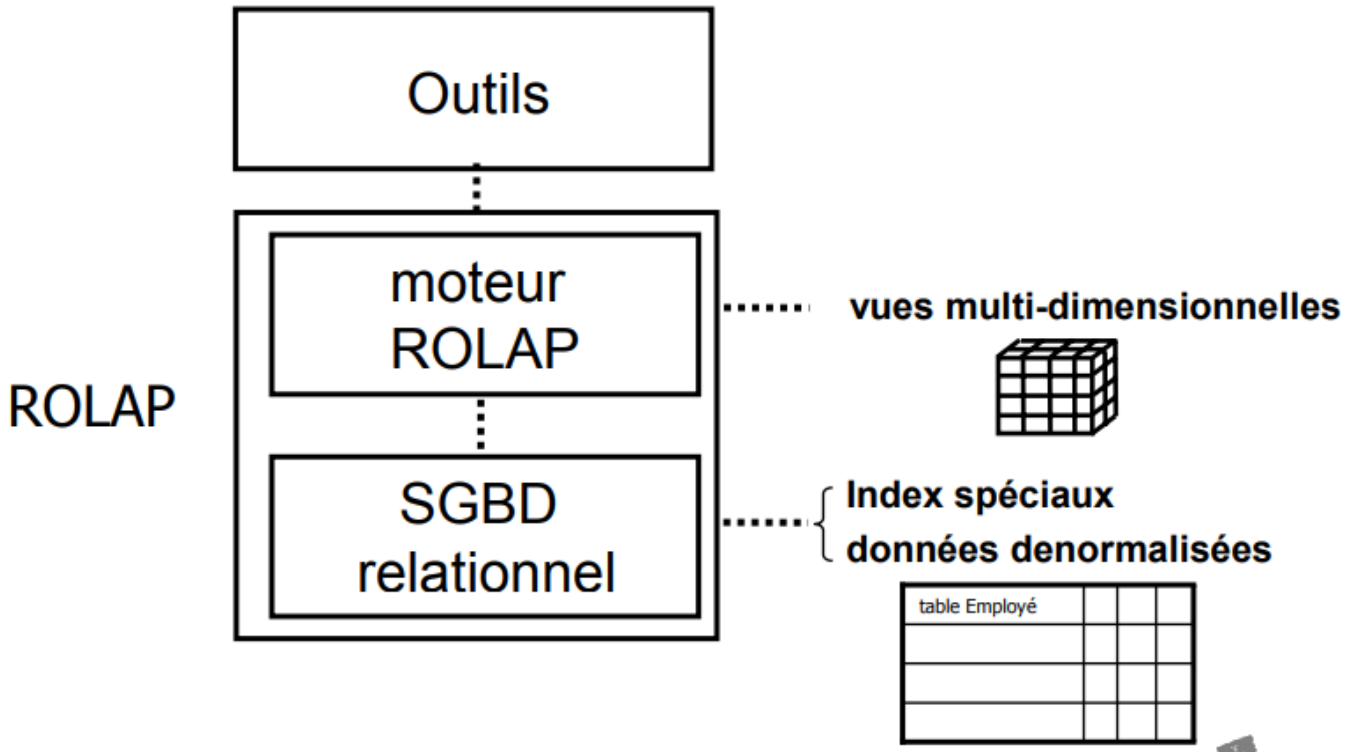
Approches pour créer un DW

- ROLAP :
  - Données sont stockées dans un SGBD relationnel.
  - Un moteur OLAP permet d'analyser la base. multidimensionnelle
- MOLAP :
  - Données sont stockées dans des cubes.
  - Accès direct aux données dans le cube.
- HOLAP :
  - Les données de base stockées dans un SGBD relationnel.
  - Les données agrégées stockées dans un cube.

ROLAP

- Modélisation en relationnel.
- La conception du schéma est particulière : schéma en étoile et schéma en flocon.
- Des vues sont utilisées pour la représentation multidimensionnelle.
- Les requêtes OLAP sont traduites en SQL.
- +++ Souple et facile à évoluer.
- +++ Capable de stocker de gros volume de données.
- — Peu efficace pour les calculs complexes.

ROLAP

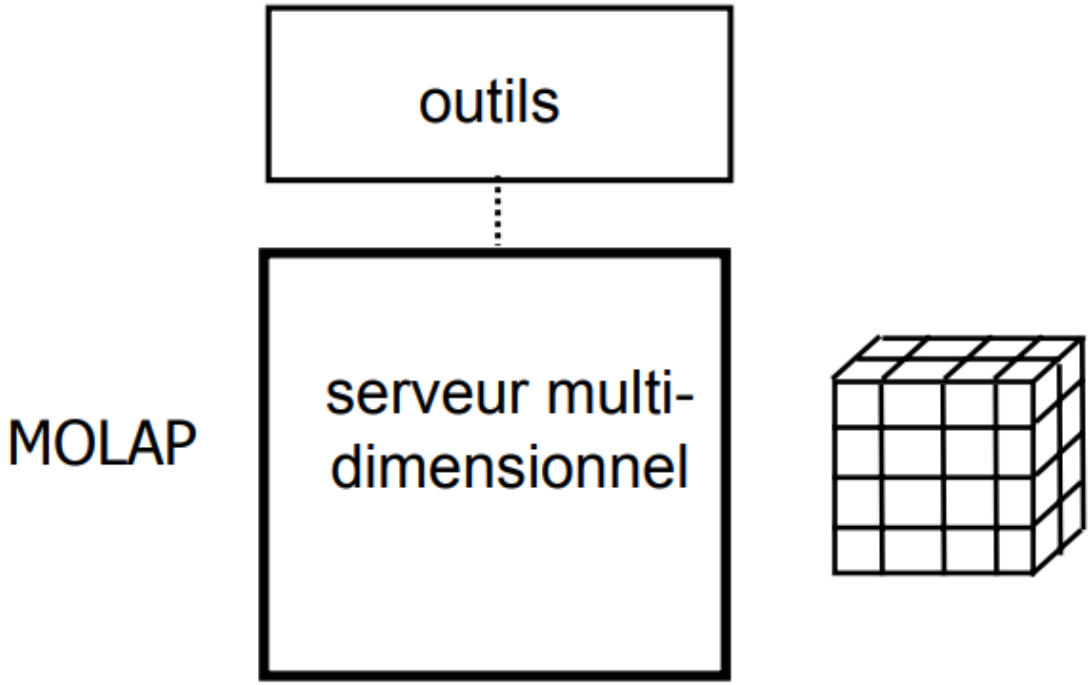

Outils
moteur ROLAP
SGBD relationnel
ROLAP
vues multi-dimensionnelles
Index spéciaux
données denormalisées
table Employé

MOLAP

Principe

- Modélisation par des cubes.
- Ces cubes sont définis sous forme de matrices multidimensionnelles.
- Le cube est indexé sur ses dimensions.
- +++ Rapide et support tous les formats des données.
- — Ne supporte pas une grande volumétrie des données.

MOLAP



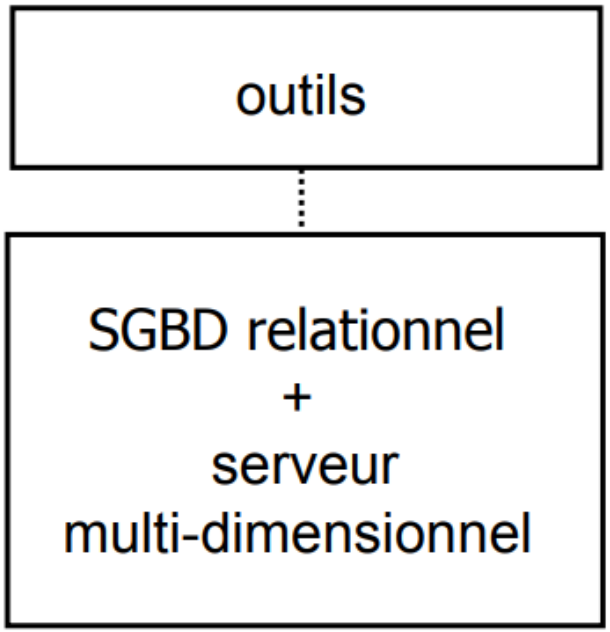
HOLAP

Principe

- MOLAP + ROLAP
- Données détaillerais stockées dans des tables relationnelles.
- Données agrégées stockées dans des cubes.
- Les traitements vont porter sur les données des tables et les cubes.

HOLAP

HOLAP


outils
SGBD relationnel
+
serveur
multi-dimensionnel

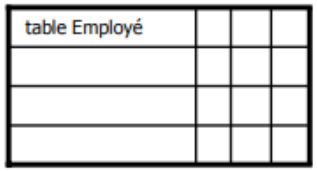
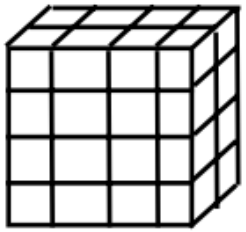


table Employé



Niveau logique ROLAP

- 3 types de schémas à concevoir :
  - schéma en étoile
  - schéma en flocon
  - schéma en constellation

Le schéma (modèle) en étoile

Principe

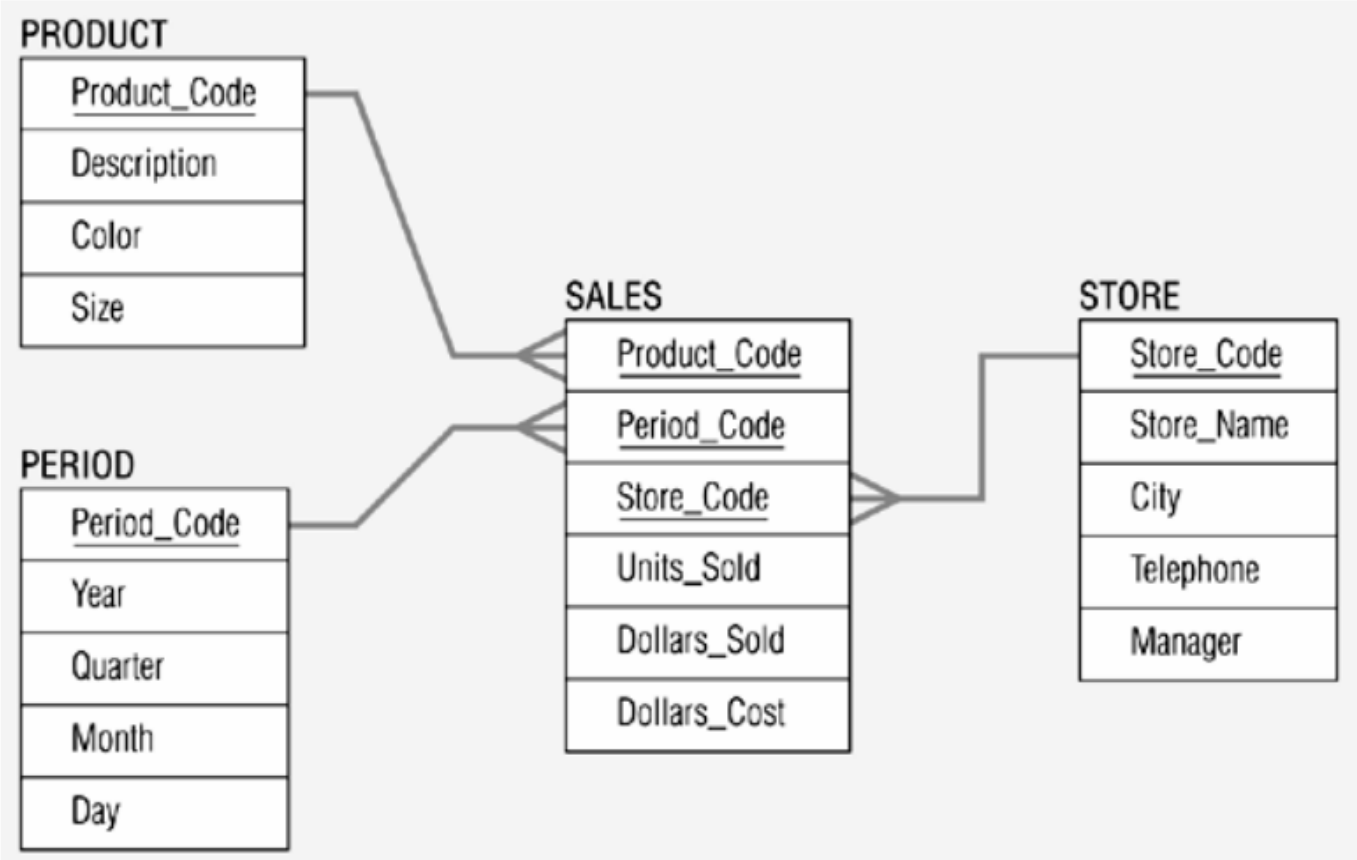
- La table de fait comporte une ou plusieurs mesures.
- Les tables de dimension comportent des descripteurs des dimensions (attributs).
- Remarques :
  - Les tables de dimension ne sont pas liées entre elles.
  - La table de dimension comporte une clé primaire.
  - La table de fait comporte une clé composée, définie par les clés étrangères des tables de dimension.
  - La table de fait comporte les valeurs des mesures et les clés étrangères vers les tables de dimensions.

Le schéma (modèle) en étoile

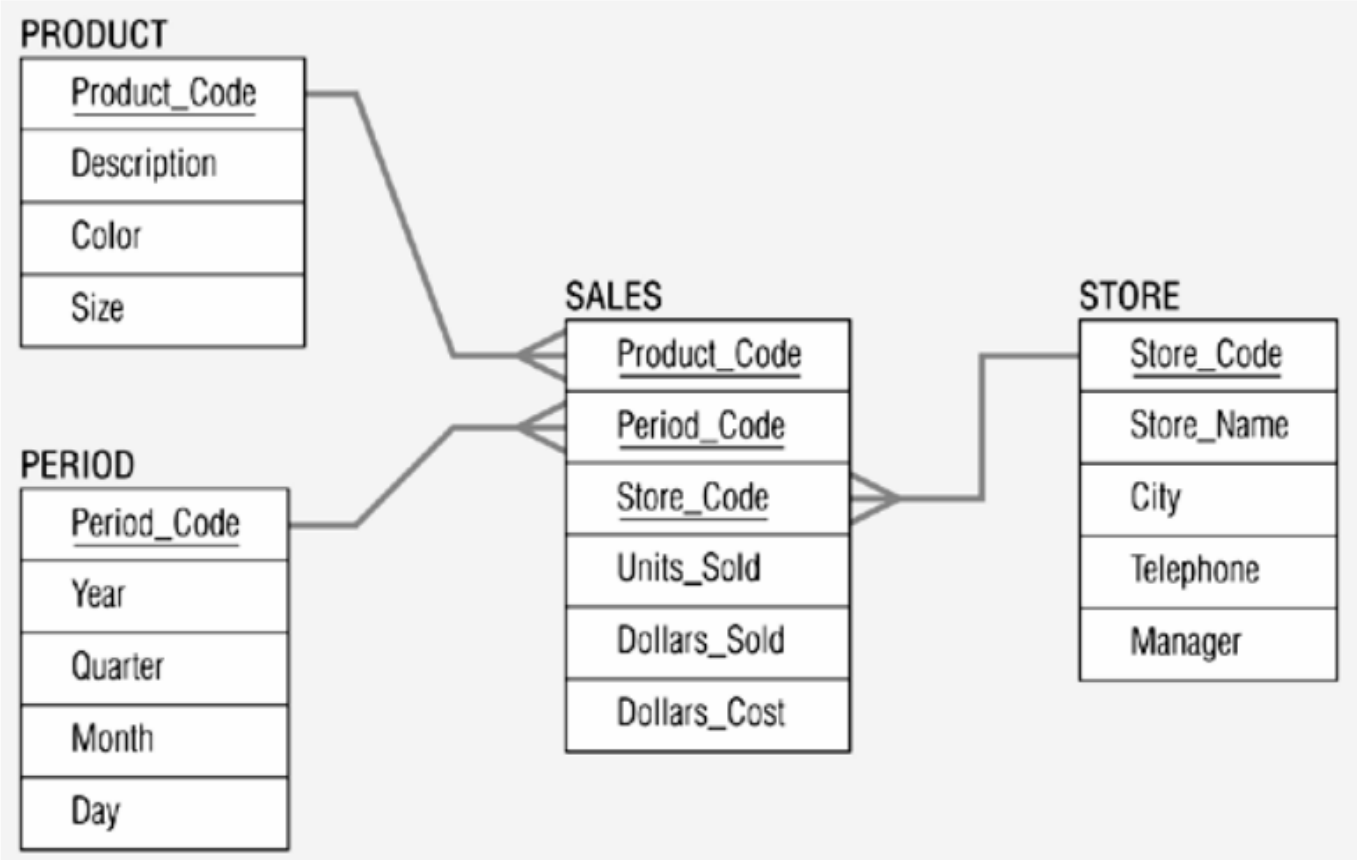
Règles de transformation

- R1 : Toute dimension est transformée en relation (table) où :
  - Tous les attributs de la dimension deviennent des attributs de la relation.
  - L'attribut le plus bas (granularité fine) devient la clé primaire
- R2 : Tout fait est transformé en une relation où :
  - La clé primaire est la concaténation des clés étrangères des dimensions.
  - Les attributs sont les mesures du fait.

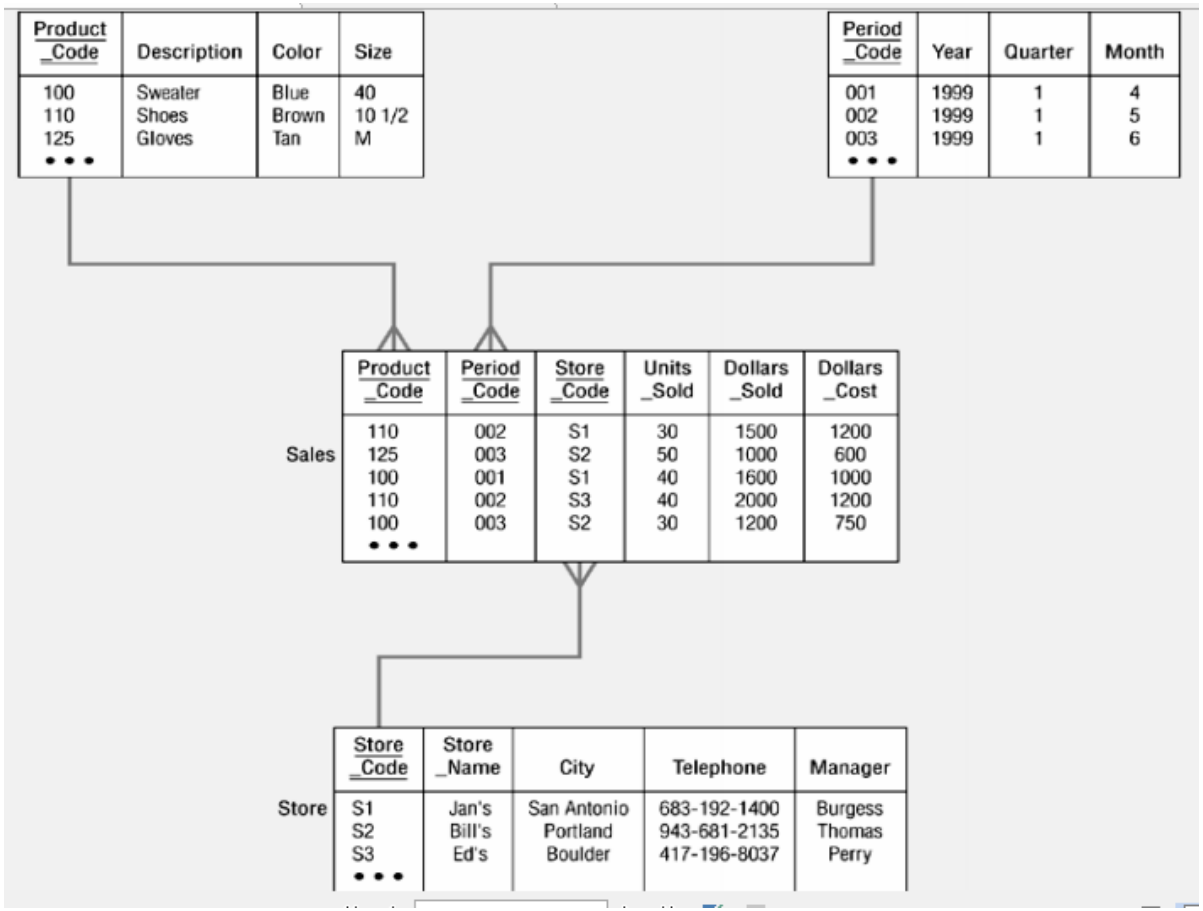
Schéma général



Exemple


PRODUCT
Product_Code
Description
Color
Size
PERIOD
Period_Code
Year
Quarter
Month
Day
SALES
Product_Code
Period_Code
Store_Code
Units_Sold
Dollars_Sold
Dollars_Cost
STORE
Store_Code
Store_Name
City
Telephone
Manager

Exemple



Product_Code	Description	Color	Size
100	Sweater	Blue	40
110	Shoes	Brown	10 1/2
125	Gloves	Tan	M
• • •			

Period_Code	Year	Quarter	Month
001	1999	1	4
002	1999	1	5
003	1999	1	6
• • •			

Sales

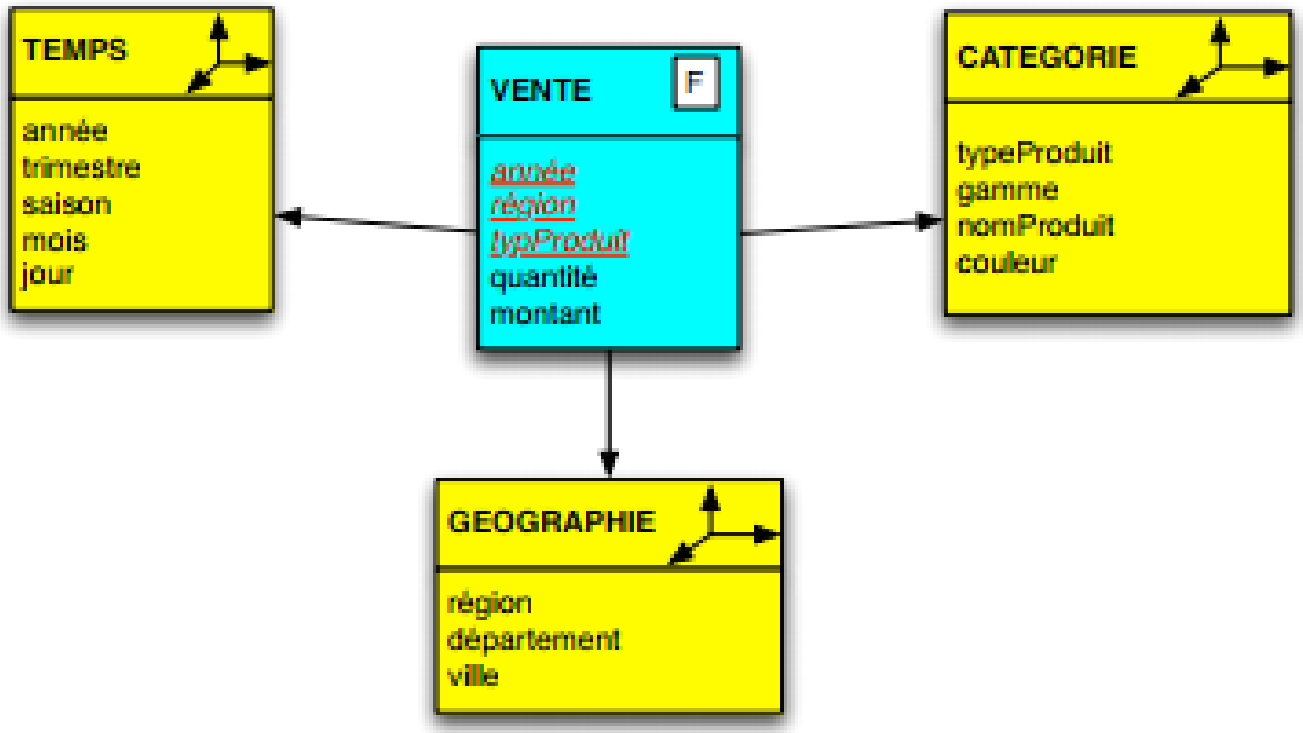
Product_Code	Period_Code	Store_Code	Units_Sold	Dollars_Sold	Dollars_Cost
110	002	S1	30	1500	1200
125	003	S2	50	1000	600
100	001	S1	40	1600	1000
110	002	S3	40	2000	1200
100	003	S2	30	1200	750
• • •					

Store

Store_Code	Store_Name	City	Telephone	Manager
S1	Jan's	San Antonio	683-192-1400	Burgess
S2	Bill's	Portland	943-681-2135	Thomas
S3	Ed's	Boulder	417-196-8037	Perry
• • •				

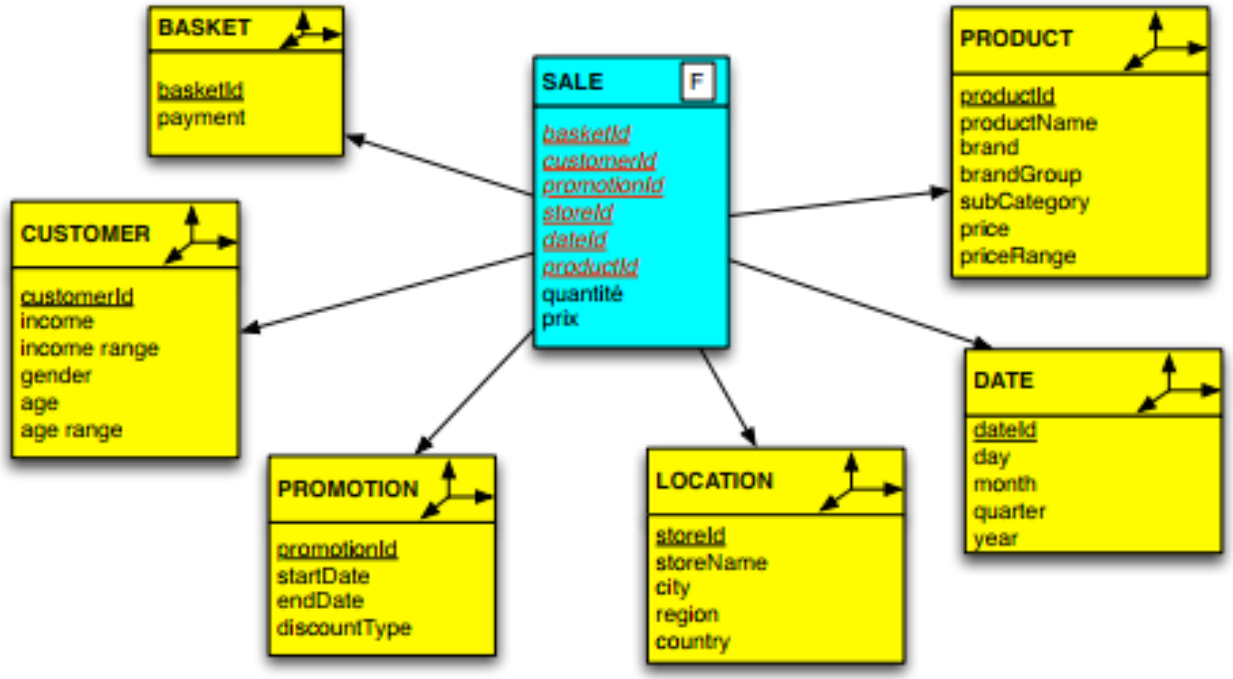
Exemple

- Vente de médicaments dans des pharmacies.
  - Table de faits : Vente
  - Tables de dimension : Temps, Catégorie, Géographie.



Exemple

- Ventes des articles dans un supermarché
  - Table de faits : Vente
  - Tables de dimension : client, produit, magasin, date, chariot et promotion.



+/-

- Avantages :
  - Facile à naviguer entre les attributs de dimension.
  - Nombre de jointures limité entre les tables de dimension.
- Inconvénients :
  - Redondance des données dans les tables de dimension.
  - Toutes les tables de dimension ne concernent pas les mesures.

Le schéma (modèle) en flocon

Principe

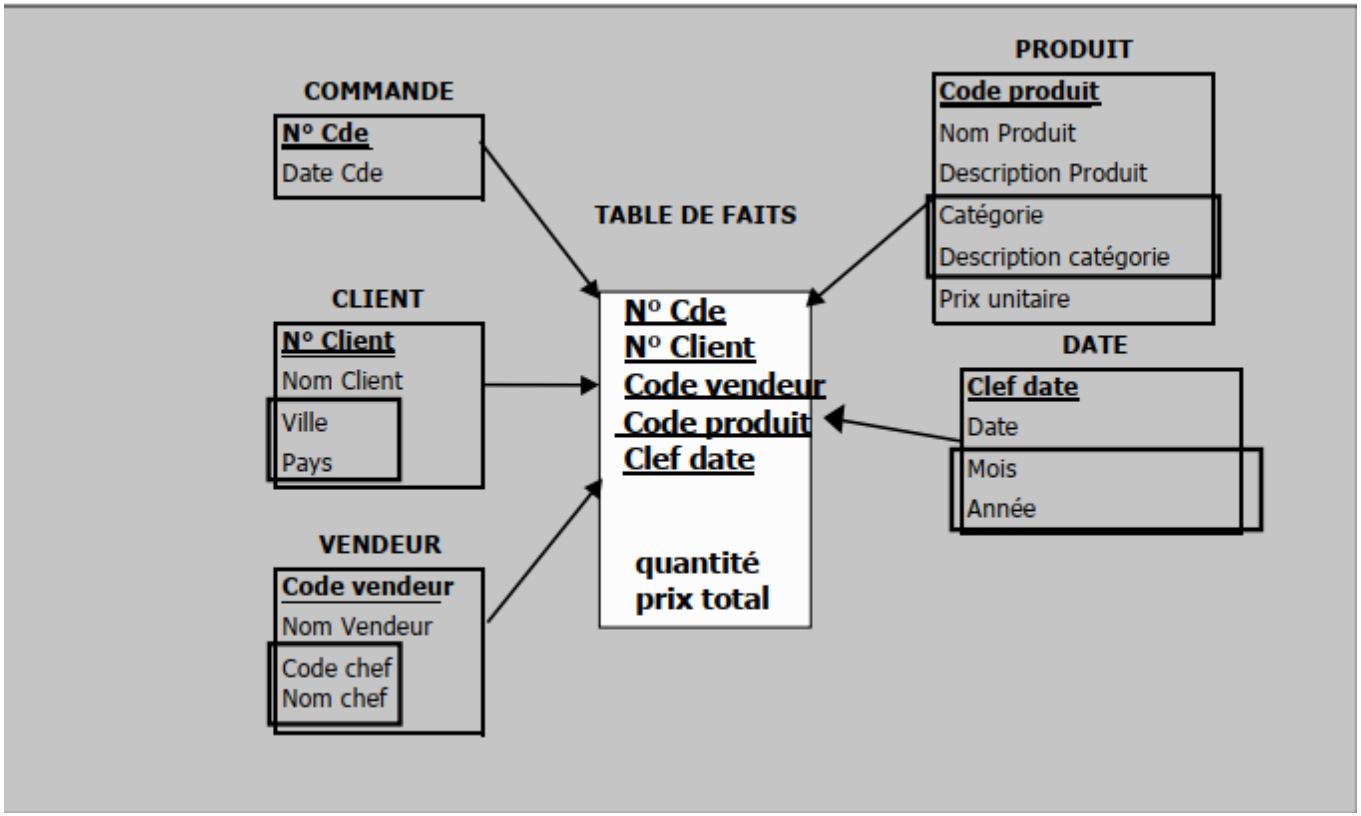
- Le schéma en flocon est une évolution du schéma en étoile avec :
  - Une décomposition de chaque dimensions selon des hiérarchies.
  - Une conservation de la table des faits.
- Normalisation des tables de dimensions :
  - Structure hiérarchique des dimensions.
  - Un niveau inférieur identifie un niveau supérieur.

Le schéma (modèle) en étoile

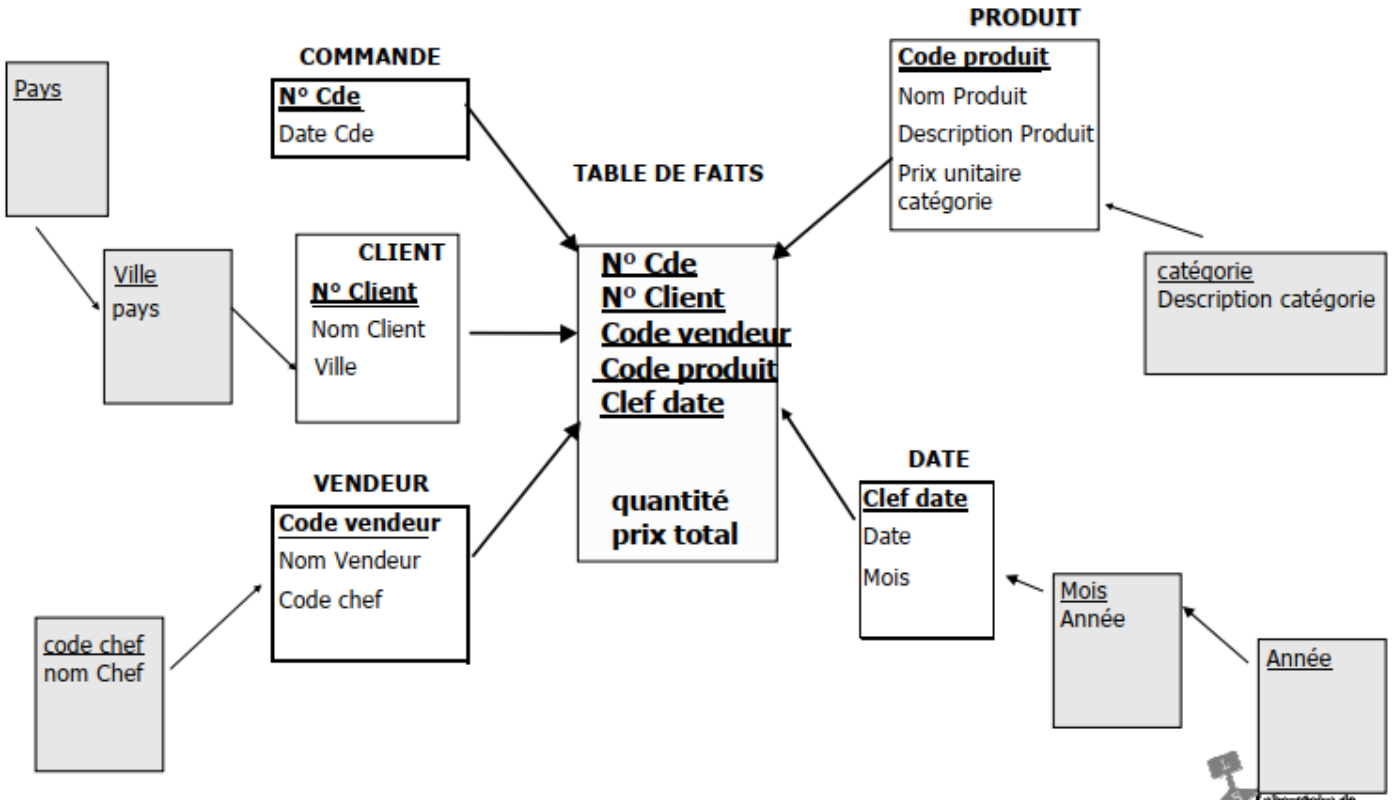
Règles de transformation

- R1 : Une dimension est représentée par plusieurs tables :
  - Chaque table représente un niveau d'agrégation.
  - Chaque table est composée de : une clé primaire (niveau d'agrégation), une clé étrangère qui permet de référencier le niveau d'agrégation supérieur et un ensemble d'attributs associés.
- R2 : Tout fait est transformé en une relation où :
  - La clé primaire est la concaténation des clés étrangères référençant les dimensions.
  - Les attributs sont les mesures du fait.

Exemple

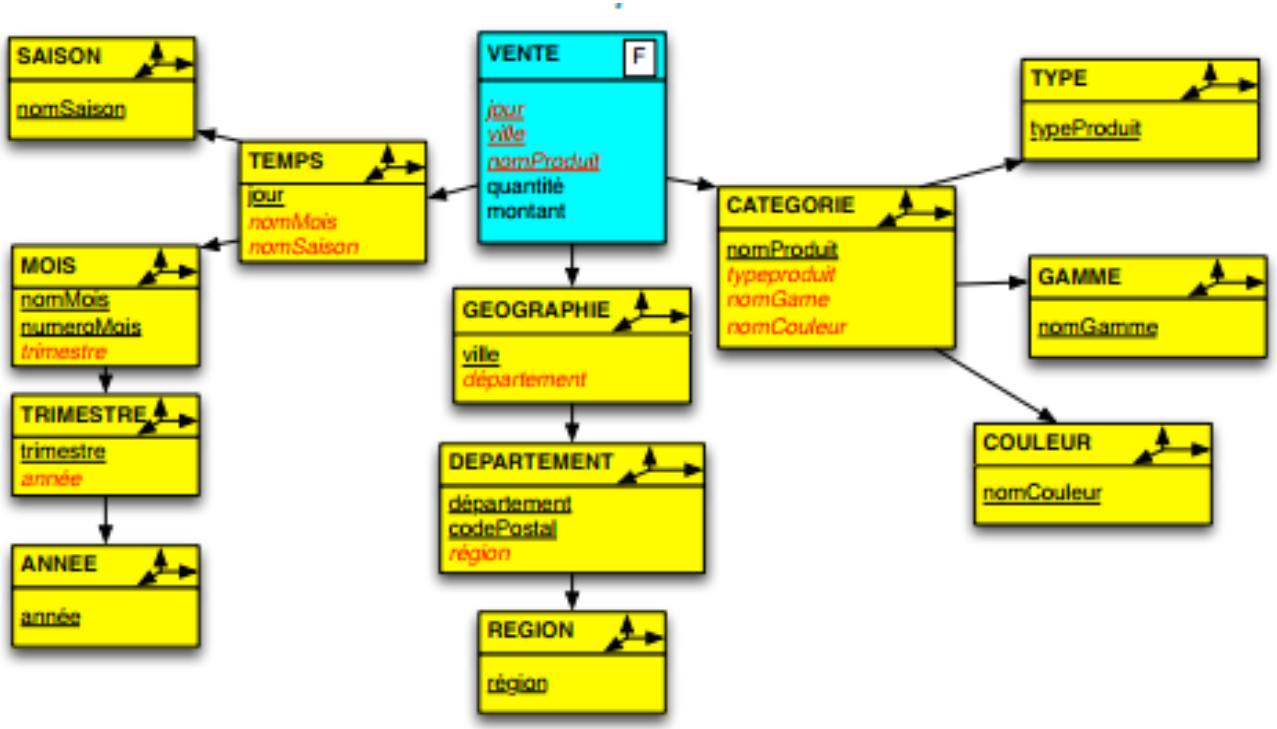

COMMANDE
N° Cde
Date Cde
CLIENT
N° Client
Nom Client
Ville
Pays
VENDEUR
Code vendeur
Nom Vendeur
Code chef
Nom chef
TABLE DE FAITS
N° Cde
N° Client
Code vendeur
Code produit
Clef date
quantité
prix total
PRODUIT
Code produit
Nom Produit
Description Produit
Catégorie
Description catégorie
Prix unitaire
DATE
Clef date
Date
Mois
Année

Exemple


PRODUIT
Code produit
Nom Produit
Description Produit
Prix unitaire
catégorie
COMMANDE
N° Cde
Date Cde
Pays
TABLE DE FAITS
CLIENT
N° Client
Nom Client
Ville
Ville
pays
N° Cde
N° Client
Code vendeur
Code produit
Clef date
quantité
prix total
catégorie
Description catégorie
DATE
Clef date
Date
Mois
VENDEUR
Code vendeur
Nom Vendeur
Code chef
Mois
Année
code chef
nom Chef
Année

Exemple

- Vente de médicaments dans des pharmacies.
  - Table de faits : Vente
  - Tables de dimension : Temps, Catégorie, Géographie.



+/-

- Avantages :
  - Hiérarchie explicite
  - réduction de la redondance
- Inconvénients :
  - Nombreuses jointures
  - navigation difficile

Schéma en étoile VS Schéma en flocon

	Schéma en étoile	**Schéma en flocon**
Structure	Tables et dimensions	Tables, dimensions et sous dimensions
Rquétes	Faible	Haute
Jointure	Moins	Grand nombre
Espace	Plus	Moins
Temps	Moins	Plus

Le modèle en constellation

Principe

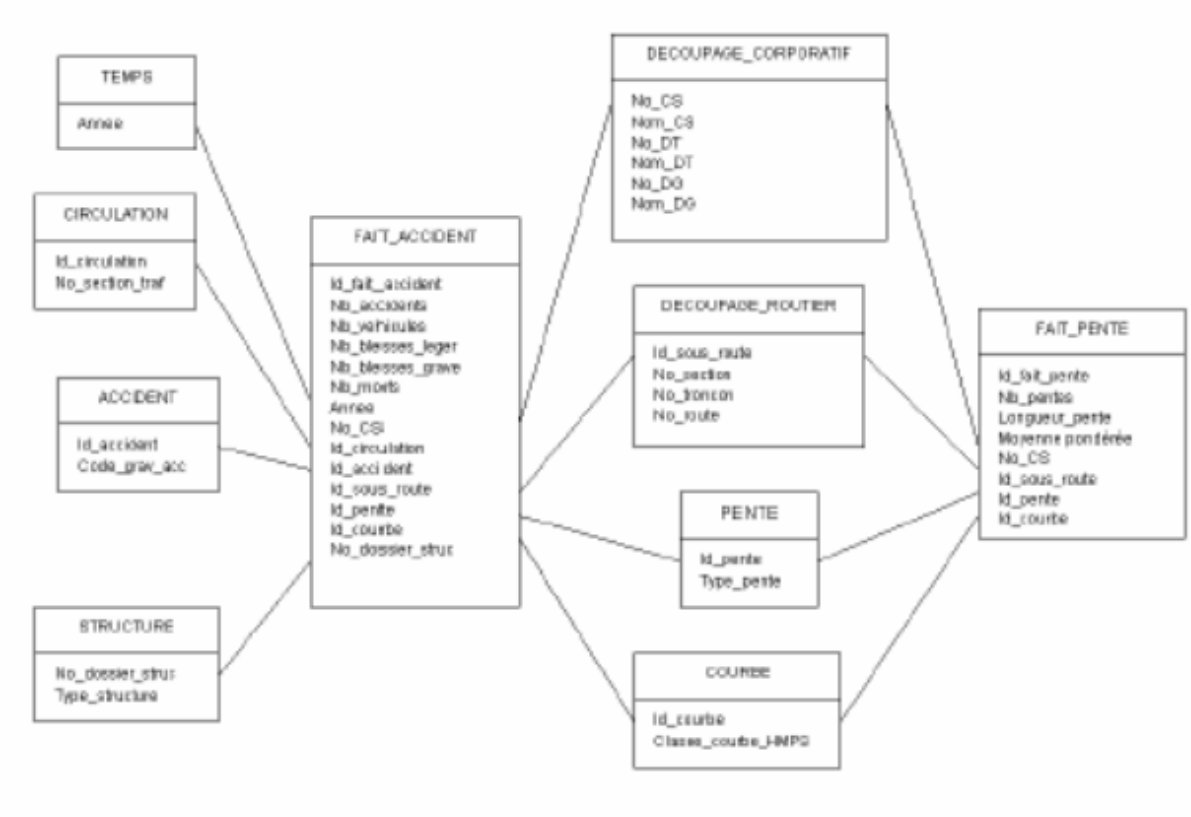
- Le modèle en constellation fusionne plusieurs modèles en étoile qui partagent des dimensions communes.
- Il comporte plusieurs tables de faits et des tables de dimensions communes ou non à ces tables de faits.

Le modèle en constellation

Règles de transformation

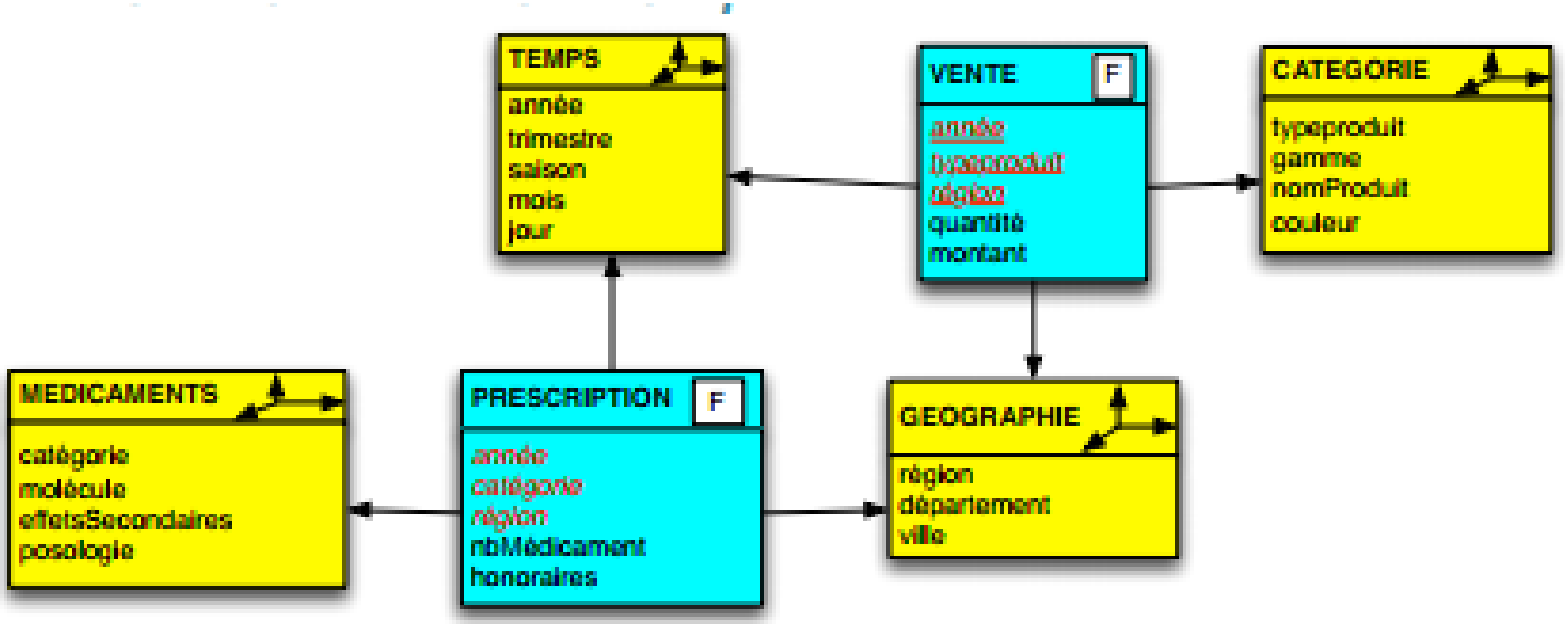
- R1 : Une dimension monofait est représentée par une table.
- R2 : Une dimension multifaits partagée par plusieurs faits est représenté par une table.
- R3 : Toute dimension partagée par plusieurs faits est représentée par plusieurs tables .
- R4 : Tout fait est représenté par une table

Exemple


TEMPS
Annee
CIRCULATION
Id_circulation
No_section_traf
ACCIDENT
Id_accident
Code_grav_acc
STRUCTURE
No_dossier_struc
Type_structure
FAIT_ACCIDENT
Id_fait_accident
Nb_accidents
Nb_vehicules
Nb_blesses_leger
Nb_blesses_grave
Nb_morts
Annee
No_CS
Id_circulation
Id_accident
Id_sous_route
Id_pente
Id_courbe
No_dossier_struc
DECOUPAGE_CORPORATIF
No_CS
Nom_CS
No_DT
Nom_DT
No_DG
Nom_DG
DECOUPAGE_ROUTIER
Id_sous_route
No_section
No_tronçon
No_route
PENTE
Id_pente
Type_pente
COURBE
Id_courbe
Classe_courbe_HMPS
FAIT_PENTE
Id_fait_pente
Nb_pentes
Longueur_pente
Moyenne pondérée
No_CS
Id_sous_route
Id_pente
Id_courbe

Exemple

- Vente de médicaments dans des pharmacies.
  - Table de faits : Vente
  - Tables de dimension : Temps, Catégorie, Géographie.



Niveau Physique

Principe

- L'implantation (implémentation) du DW et dépend du logiciel utilisé.
- CREATE TABLE NomTable AS SELECT ... : recopie physique d'une table. Si mise à jour de la table source, pas de mise à jour de la table «NomTable»
- CREATE VIEW NomView AS SELECT ... : recalculé à chaque requête.

Exemple

- Création de l'alias de la BD

```
CREATE DATABASE LINK myRelDB CONNECT TO user
IDENTIFIED BY user USING localDB;
```

- Création du fait

```
CREATE MATERIALIZED VIEW salesMv
BUILD IMMEDIATE|DEFERRED
REFRESH FAST|COMPLETE|FORCE ON COMMIT|ON DEMAND
AS SELECT t.calendarYear, p.prodId,
SUM(s.amountSold) AS sumSales
FROM times@myRelDB t, products@myRelDB p,
sales@myRelDB s
WHERE t.timeId = s.timeId AND p.prodId = s.prodId
GROUP BY t.calendarYear, p.prodId;
ALTER TABLE salesMv ADD CONSTRAINT pk_salesmv
PRIMARY KEY (prodId,calendarYear);
```

Exemple

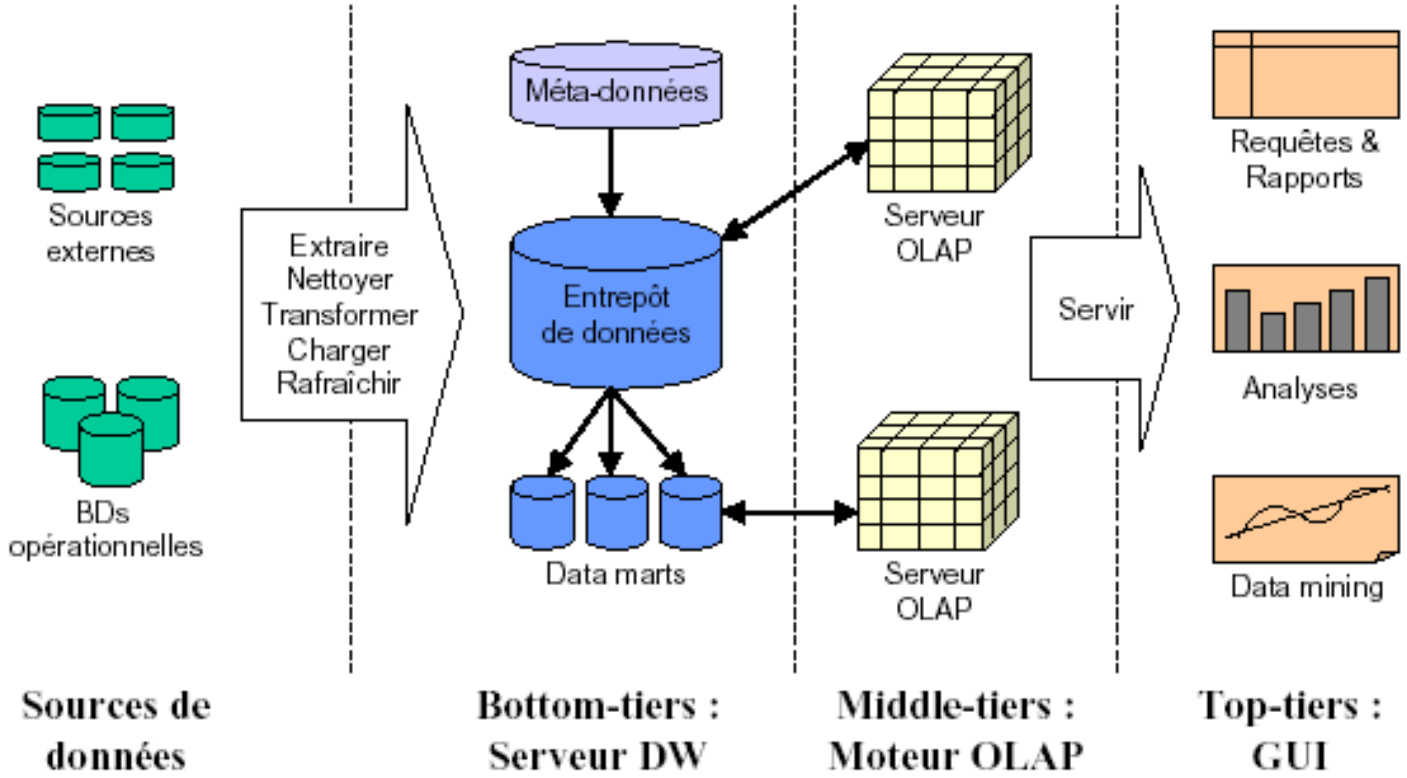
- Création des dimensions

```
CREATE MATERIALIZED VIEW location
BUILD IMMEDIATE|DEFERRED
REFRESH FAST|COMPLETE|FORCE ON COMMIT|ON DEMAND
AS SELECT g.ville, g.pays, g.continent
FROM geographie@myRelDB g;
ALTER TABLE location ADD CONSTRAINT pk_location
PRIMARY KEY (ville);
```

- Création des hiérarchies

```
CREATE DIMENSION geo LEVEL n1 is (ville) LEVEL n2
is (pays) LEVEL n3 is (continent) HIERARCHY H1 (n1
CHILD OF n2 CHILD OF n3 )
```

Architecture du DW



Alimentation des données

Principe

- Après avoir conçu le modèle des données, comment alimenter le DW ?
- Le processus dalimentation dun DW consiste à rassembler les données sources souvent hétérogènes.
- L'alimentation est faite selon des règles précises qui sont stockées sous forme de méta-données.
- Les règles permettent d'assurer la maintenance et l'administration des data warehouses.

Alimentation des données

ETL

- Solution : un outil qui automatise l'alimentation et le chargements de l'entrepôt.
- ETL (Extracting, Transforming and Loading)
- 3 étapes a suivre :
  - Extraction des données
  - Nettoyage et Transformation des données
  - Alimentation des données

Objects

- Extraction : Accés aux differentes données.
- Nettoyage : Recherche et résolution les inconsistances des données.
- Transformation : Traitement des differents formats des données.
- Chargement : Alimentation des données dans le DW.

Extraction

Principe

- Identification des sources.
- Quelles données de production qui il faut sélectionner pour créer un DW ?
- Remarque : Toutes les données sources ne sont pas forcément utiles à la prise de décision.
- Exemple : Est ce qu'on doit alimenter ladresse complète ou séparer le code postal ?

Type

- Un extracteur (wrapper) est implémenté pour chaque source de données :
  - sélectionne et extrait les données
  - transforme les dans un format commun
  - Exploite les interfaces de connexion comme ODB, JDBC.
- Deux types d'extractions :
  - Extraction complète
  - Extraction incrémentale

Extraction

Extraction complète

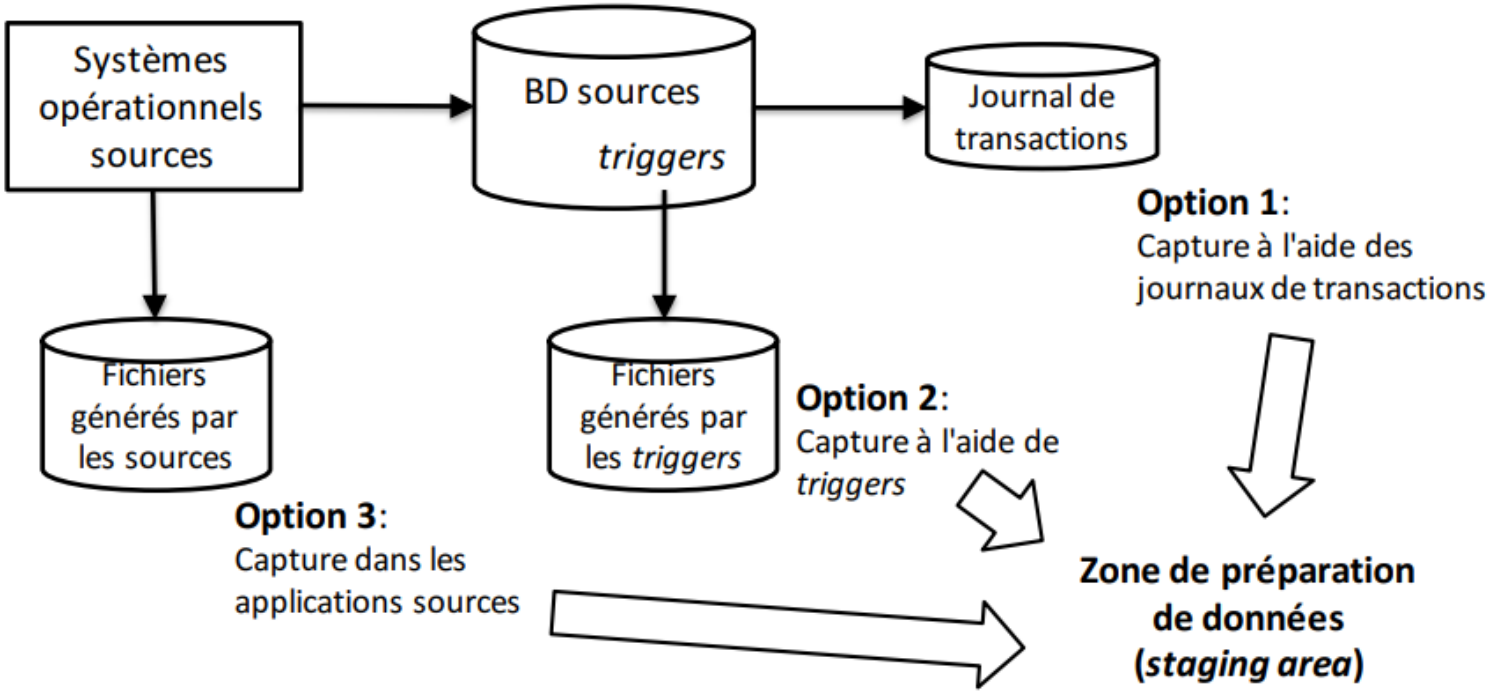
- Capture l'ensemble des données à une période définie et elle est employée dans deux cas :
  1. Chargement initial des données ;
  2. Rafraîchissement complet des données (ex : modification d'une source).

Extraction incrémentale

- Capture uniquement les données qui ont changées ou ont été ajoutées depuis la dernière extraction ;
- Elle est faite de deux façons :
  1. Extraction temps-réel ;
  2. Extraction différée .

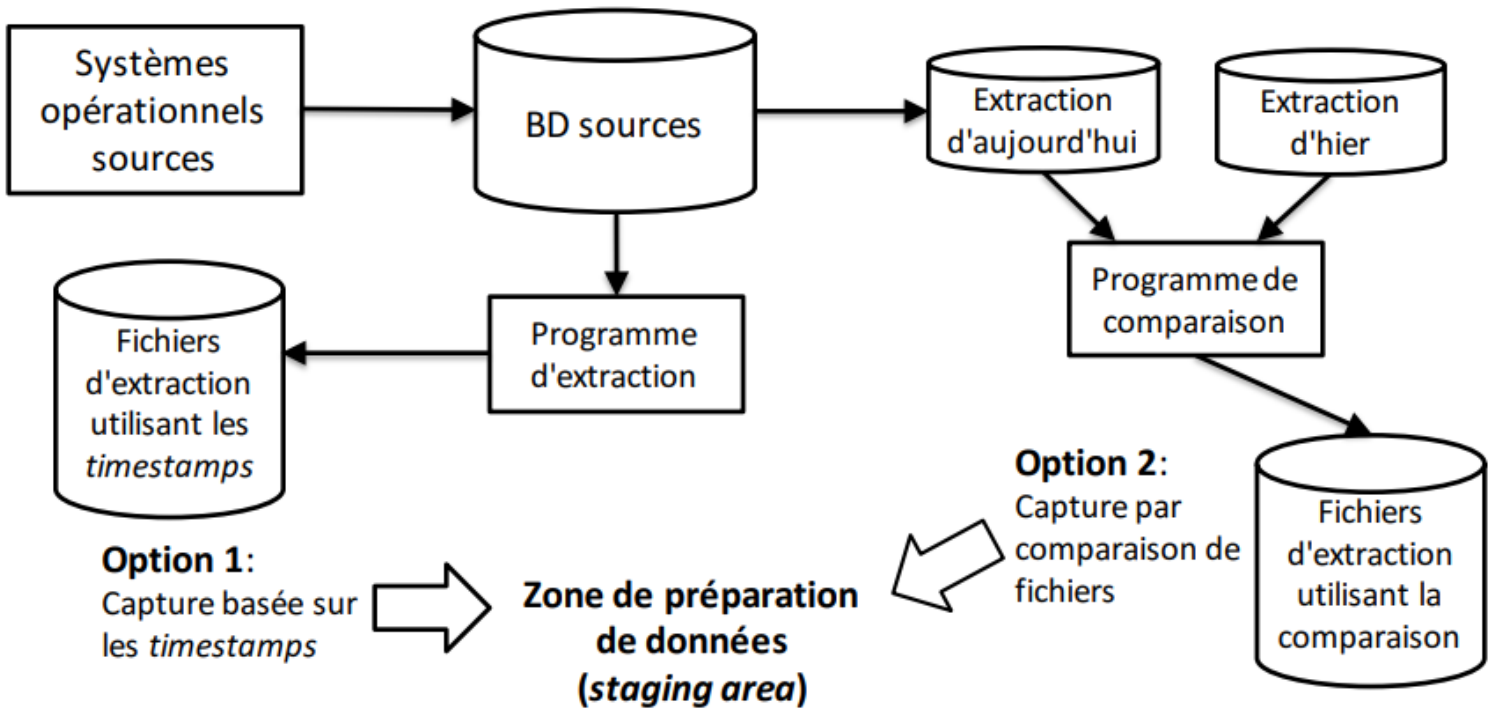
Extraction en temps réel

- S'effectue au moment où les transactions surviennent dans les base de données sources.



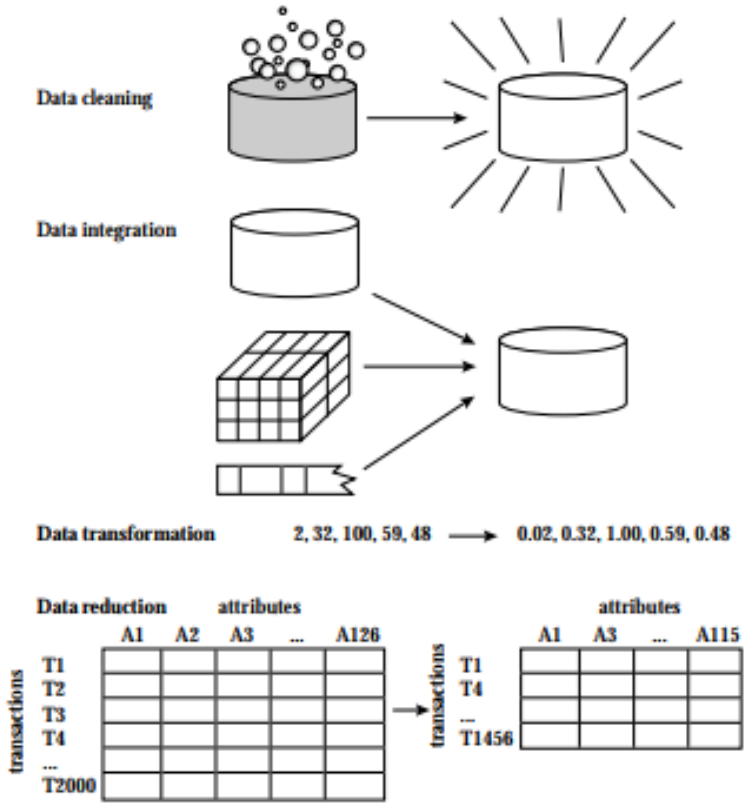
Extraction en temps différé

- Extrait tous les changements effectués durant une période donnée (ex : heure, jour, semaine, mois).



Transformation

- Etape très importante qui assure la cohérence et l'integrité des données DW.
- Résultats de l'analyse dépendent de la qualité des données.
- Données erronées $\Longrightarrow$ Analyse erroné.



Nettoyage des données

- Résoudre le problème d'inconsistance et d'ambigüité qui se trouvent dans les données sources.
- Plusieurs types dinconsistances ont été détectés.
- 5 à 30 % des données sources sont erronées.
- Données réelles sont souvent :
  - Incomplètes : valeurs manquantes, valeur nulle ...
  - Bruitées : erreurs et exceptions, présence de données fausses dés leur saisie, faute de frappe ...
  - Incohérentes : incompatibilité entre la valeur et la description de la colonne, différents formats dans une même colonne ...
  - Dupliquées : duplication dinformation ...

Nettoyage des données

Traitement des données manquantes

- Données manquantes : certains attributs n'ont pas de valeur.
  - Ignorer les données manquantes.
  - Compléter manuellement les données manquantes.
  - Utiliser une constante globale exemple : « inconnue »
  - Utiliser la moyenne de l'attribut
  - Utiliser la valeur la plus probable.

Traitement des données bruitées

- Données bruitées : erreur ou variance aléatoire d'une variable mesurée.
  - Lissage des données (Smoothing) réduire le bruit dans les données.
    - Par partitionnement (Binning) : Partitionner les données et lisser les partitions par la moyenne, la médiane, les bornes.
    - Par clustering : Détecter et supprimer les exceptions.
    - Par régression : Lisser les données par des fonctions de régression.

Nettoyage des données

Traitement de conflits

- Résoudre les conflits des valeurs et réduire les incohérences entre les différentes sources des données.
  - conflit sémantique : choix de différents niveaux dabstraction pour un même objet.
  - conflits de structures : choix de différentes propriétés pour un même objet.
  - conflits de représentation : choix de différentes représentations pour un même objet.

Traitement de conversion

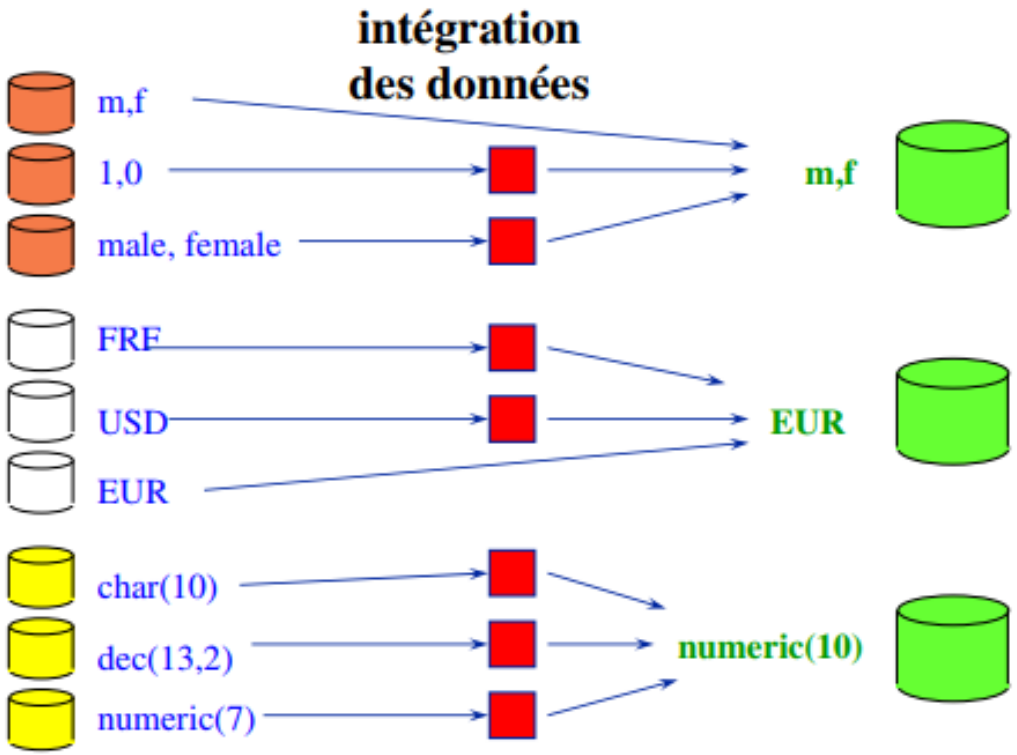
- Conversion de jeu de caractères :
  - Exemple : EBCDIC (IBM) vers ASCII.
- Conversion des unités de mesure :
  - Exemple : impérial à métrique.
- Conversion de dates :
  - Exemple : "24 FEB 2011" vs "24/02/2011" vs "02/24/2011".

Nettoyage des données

Décodage des champs

- Identifier les différents noms des mêmes données.
  - Exemple : num-client Ou client-id
- Consolider les données de sources multiples.
  - Exemple : ["homme", "femme"] vs ["M", "F"] vs [1,2].
- Traduire les valeurs cryptiques
  - Exemple : "AC", "IN", "SU" pour les statuts actif, inactif et suspendu.

Nettoyage des données



Nettoyage des données

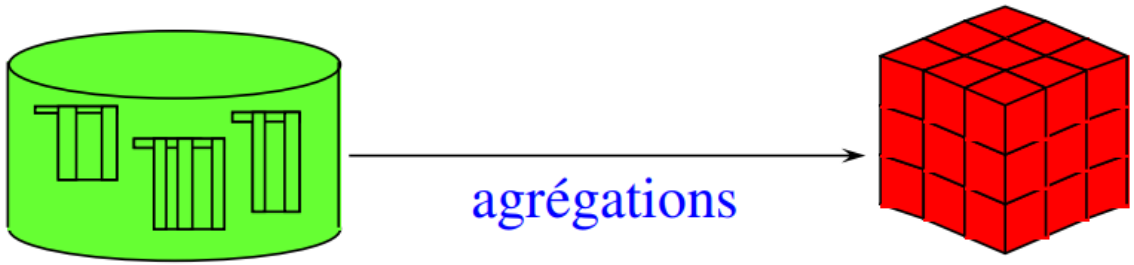
Traitement des données dupliquées

- Découpage de champs complexes :
  - Exemple : extraire les valeurs prénom, secondPrénom et nomFamille à partir du nomComplet.
- Fusion de plusieurs champs :
  - Exemple : information d'un produit :
    - Source 1 : code et description ;
    - Source 2 : types de forfaits ;
    - Source 3 : coût.

Chargement

- Charger les données nettoyées et préparées dans le DW.
- Il est nécessaire de mettre en place des stratégies de chargement et politiques de rafraîchissement.
- 3 Types de chargement :
  - Chargement initial : Fait une seule fois lors de l'activation de l'entrepôt de données ;
  - Chargement incrémental : Fait une fois le chargement initial est complété ;
  - Rafraîchissement complet : Employé lorsque le nombre de changements rend le chargement incrémental trop complexe.

Analyse Multidimensionnelle



Analyse multidimensionnelle = capacité à manipuler des données qui ont été aggrégées selon différentes dimensions

Analyse Multidimensionnelle

- Objectif : Produire des informations déjà agrégées dans le DW selon les requêtes utilisateurs.
- Cube OLAP : représentation de l'information décisionnelles dans un cube à N dimensions.
- Opérations OLAP : fonctions qui facilitent l'analyse multidimensionnelle.

Exemple

- Soit le schéma de la base de donnée suivant :
  - Produit(GENCOD, Designation, Marque, Nature, PrixAchat, PrixReventeConseille)
  - Vente (GENCOD, NMAG, Date, Qte, PrixVente)
  - Magasin(NMAG, Enseigne, Adresse, Ville, Dept
  - Nat2Cat(Nature, Categorie)
  - Cat2Ray(Categorie, Rayonnage)
  - Dep2Reg(Dept, Region)

Analyse Multidimensionnelle

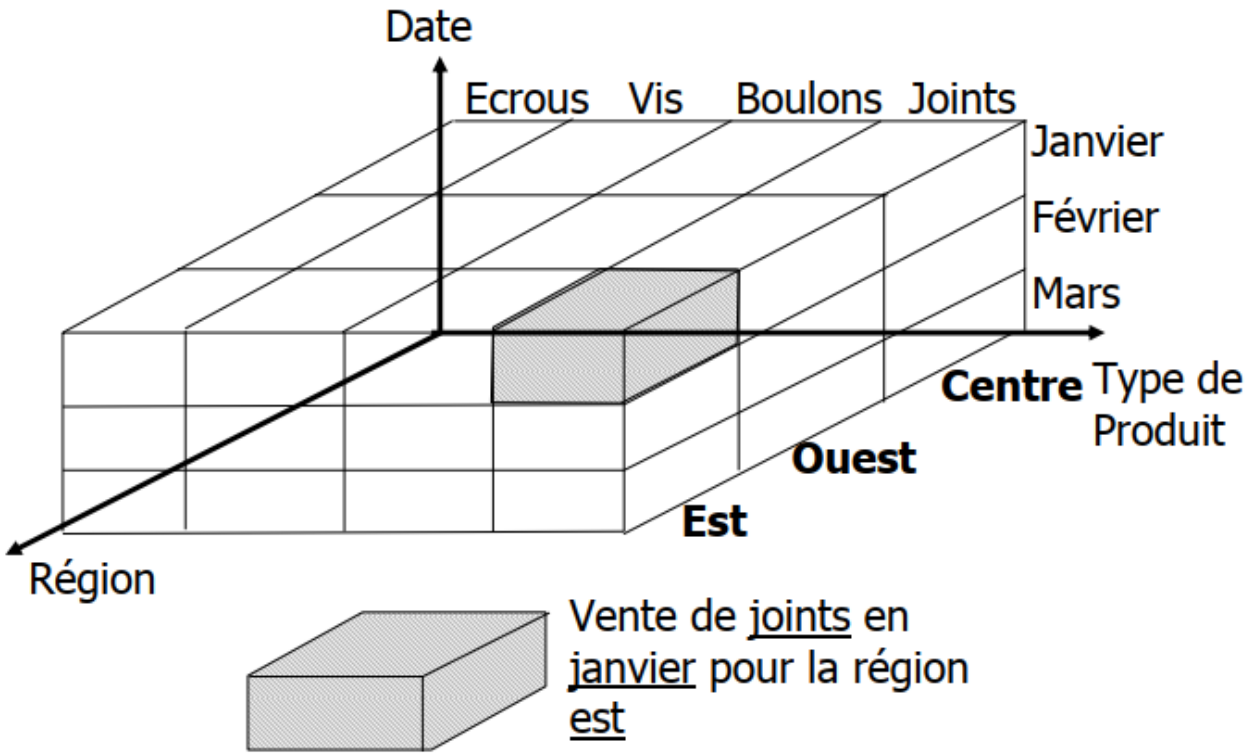
- Un data warehouse est basé sur un modèle de données multidimensionnel qui traite les données sous forme des data cubes.
- Cube = BD multidimensionnel.
- Axes = dimensions.
- Chaque cellule de l'hypercube contient une mesure calculée.
- Exemple : Un data cube pour les ventes, permet de modéliser et de voir les données relatives aux ventes en de multiples dimensions (date, type de produits, région).

Exemple Data cube

produit	région	vente
écrou	Est	50
écrou	Ouest	60
écrou	Centre	110
vis	Est	70
vis	Ouest	80
vis	Centre	90
boulon	Est	120
boulon	Ouest	10
boulon	Centre	20
joint	Est	50
joint	Ouest	40
joint	Centre	70

	Est	Ouest	Centre
écrous	50	60	110
vis	70	80	90
boulons	120	10	20
joints	50	40	70

Exemple Data cube

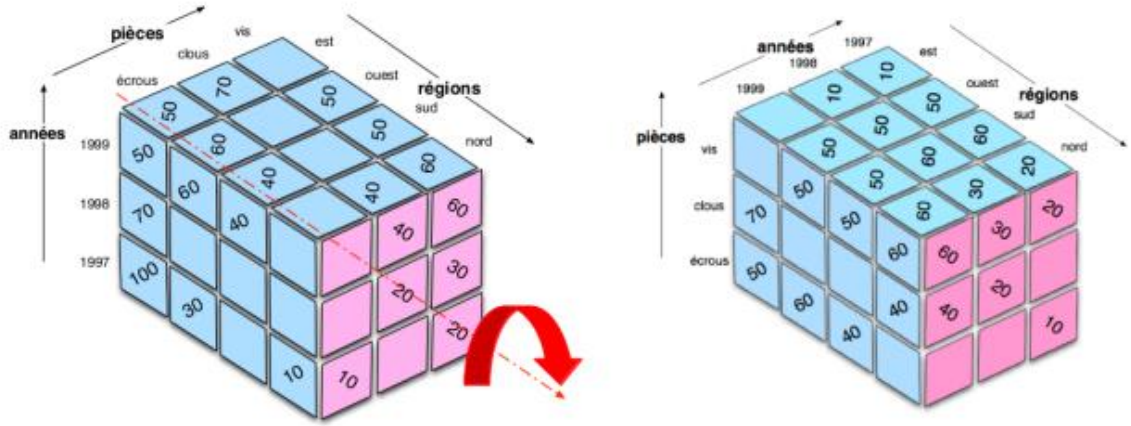


Opérateurs OLAP

- Opérateurs de visualisation du cube (Cube -> Cube) :
  - Restructuration ou réorientation du cube ;
  - Transformation de la granularité des données (Forage) ;
  - Sélection ou projection du cube.

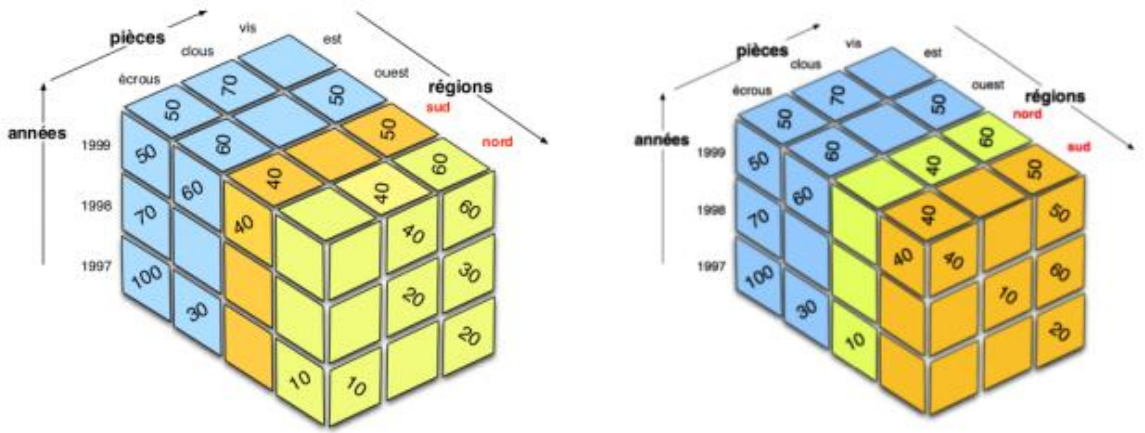
Opérations de restructuration

- Rotate/pivot : sélection de faces.
- Effectuer une rotation du cube autour d'un de ses axes.
- Présentation d'un ensemble de faces différentes.



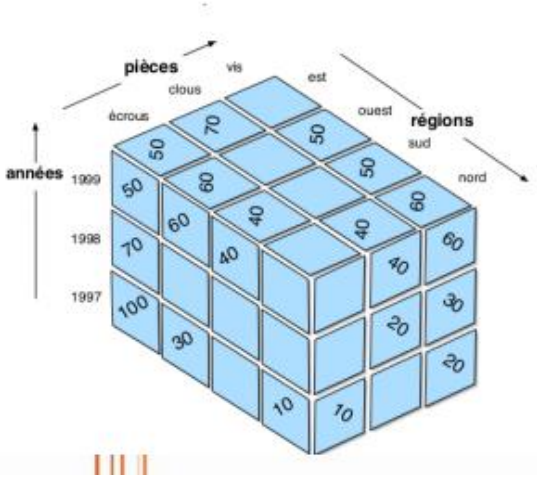
Opérations de restructuration

- Switch ou permutation : interchanger la position des membres d'une dimension



Opérations de restructuration

- Split ou division : Présentation de chaque tranche du cube
- Passer d'une présentation tridimensionnelle à une présentation d'un ensemble de tables
- Généralisation : découpage d'un hypercube de dimension 4 en cubes (3D)



ventes est	1999	1998	1997
écrous	50	70	100
vis		10	10
clous	70	70	100

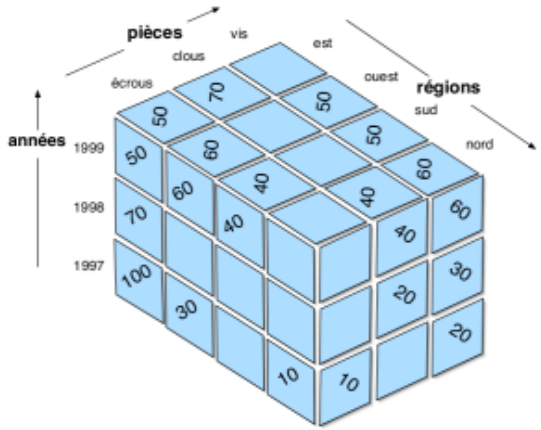
ventes sud	1999	1998	1997
écrous	40	20	
vis	50	60	60
clous		10	

ventes ouest	1999	1998	1997
écrous		10	30
vis	50	50	50
clous		10	40

ventes nord	1999	1998	1997
écrous			10
vis	60	30	20
clous	40	20	

Opérations de restructuration

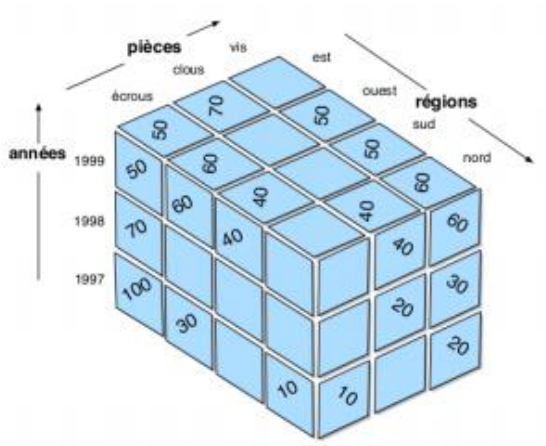
- Nest ou l'emboitement : Imbrication des membres à partir du cube.
- Produire sur une même représentation à 2 dimensions toutes les informations (mesures et membres) d'un cube peu importe le nombre de dimensions.



ventes nest		1999	1998	1997
écrous	est	50	70	100
	ouest		10	30
	nord			10
	sud	40	20	
vis	est		10	10
	ouest	50	50	50
	nord	60	30	20
	sud	50	60	60
clous	est	70	70	100
	ouest		10	40
	nord	40	20	
	sud		10	

Opérations de restructuration

- Push ou l'enfoncement : Combiner les membres d'une dimension avec les mesures.
- Mettre les membres et les mesures dans les mêmes cellules.



ventes push	est	ouest	nord	sud
écrous	1999 50 1998 70 1997 100	1998 10 1997 30	1997 10	1999 40 1998 20
vis	1998 10 1997 10	1999 50 1998 50 1997 50	1999 60 1998 30 1997 20	1999 50 1998 60 1997 60
clous	1999 70 1998 70 1997 100	1998 10 1997 40	1999 40 1998 20	1998 10

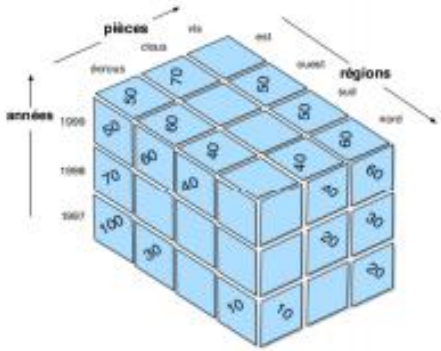
Opérations de restructuration

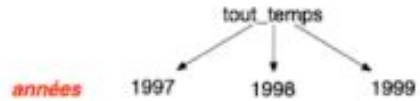
- Navigation entre les différents niveaux de granularité.
- Roll-up (forage vers le haut) :
  - Représente les données à un niveau de granularité supérieur selon l'organisation hiérarchique d'une dimension.
- Drill-down (forage vers le bas) :
  - Inverse du roll-up
  - Représente les données à un niveau de granularité inférieur.

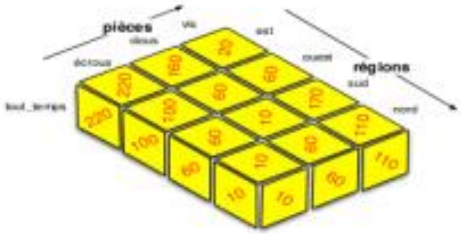
Opérations de forage

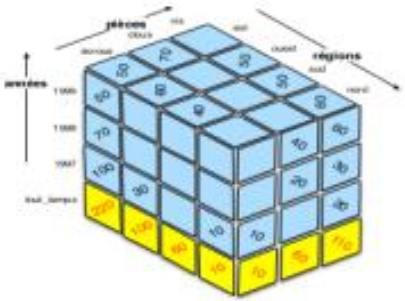
- Roll-up/forage vers le haut :
  - Représente les données à un niveau de granularité supérieur selon l'organisation hiérarchique d'une dimension.
  - Il faut définir une fonction d'agrégation pour savoir comment produire les valeurs du niveau supérieur à partir de celles du niveau inférieur.

Opérations de forage





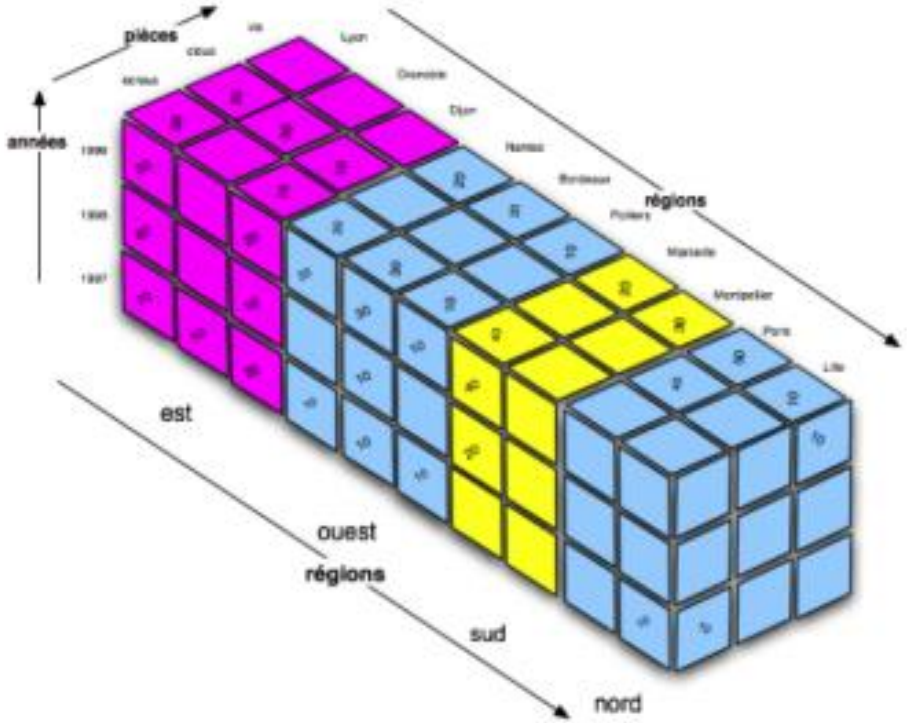




Opérations de forage

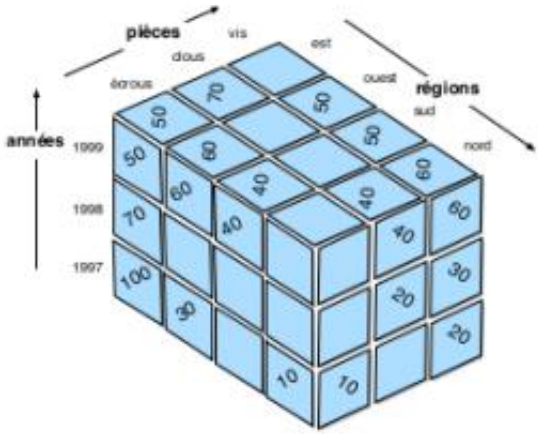
- Drill-down/forage vers le bas :
  - Représente les données à un niveau de granularité inférieur selon l'organisation hiérarchique d'une dimension.
  - Données produites sous forme plus détaillée en fonction de la hiérarchie de la dimension.

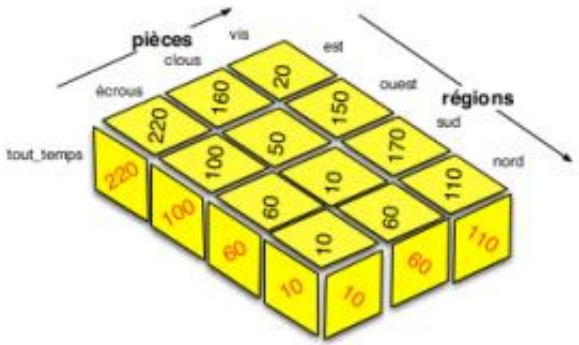
Opérations de forage



Opérations de sélection / projection

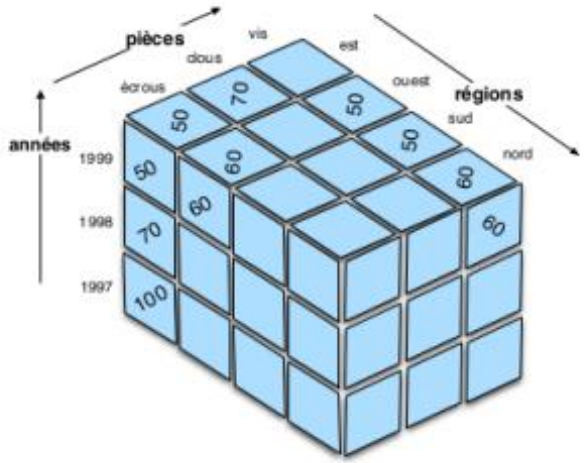
- Dice : projection sur une dimension du cube.





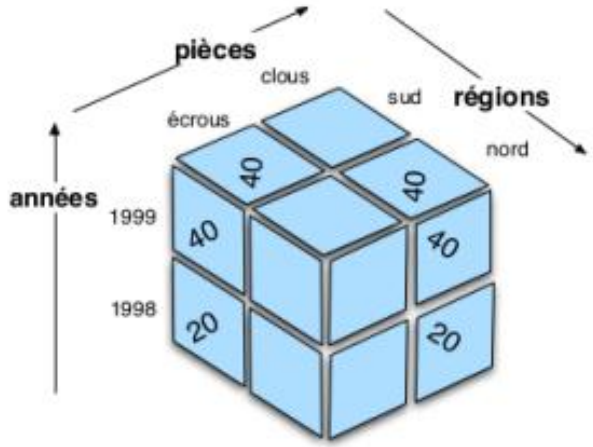
Opérations de sélection / projection

- Slice : Produire une partie du cube en utilisant des conditions selon les dimensions.
- Exemple : Sélection du cube dont les Ventes 50.



Opérations de sélection / projection

- Sélection des ventes des écrous ou des clous, durant les années 1998 ou 1999, dans les régions nord ou sud.



Bibliographie

- Transparents :
  - C. Vangenot : "Datawarehouse"
  - E. GRISLIN-LE STRUGEON : "Systèmes d'information décisionnels"
  - N. Essoussi : "Data Warehouse"
  - L. Soulier : "Business Intelligence"
- Livres :
  - Franco, J. M. (1997). Le data warehouse : le data mining. Eyrolles : Informatiques magazine.
  - Goglin, J. F. (1998). La construction du datawarehouse : du datamart au dataweb. Hermès.
  - Inmon, W. H. (2005). Building the data warehouse. John wiley sons.