

Cours 5

Entrepôt de Données Datawarehouse

BDA
Master GL 2019- 2020

OUARED Abdelkader

Plan

1. Motivation
2. Solution 1: Architecture non entrepôt
3. Pourquoi un ED ?
4. Définition d'un ED
5. Architecture d'un ED
6. OLAP vs OLTP
7. Modélisation Multidimensionnelle
8. Solutions Open source

Motivation

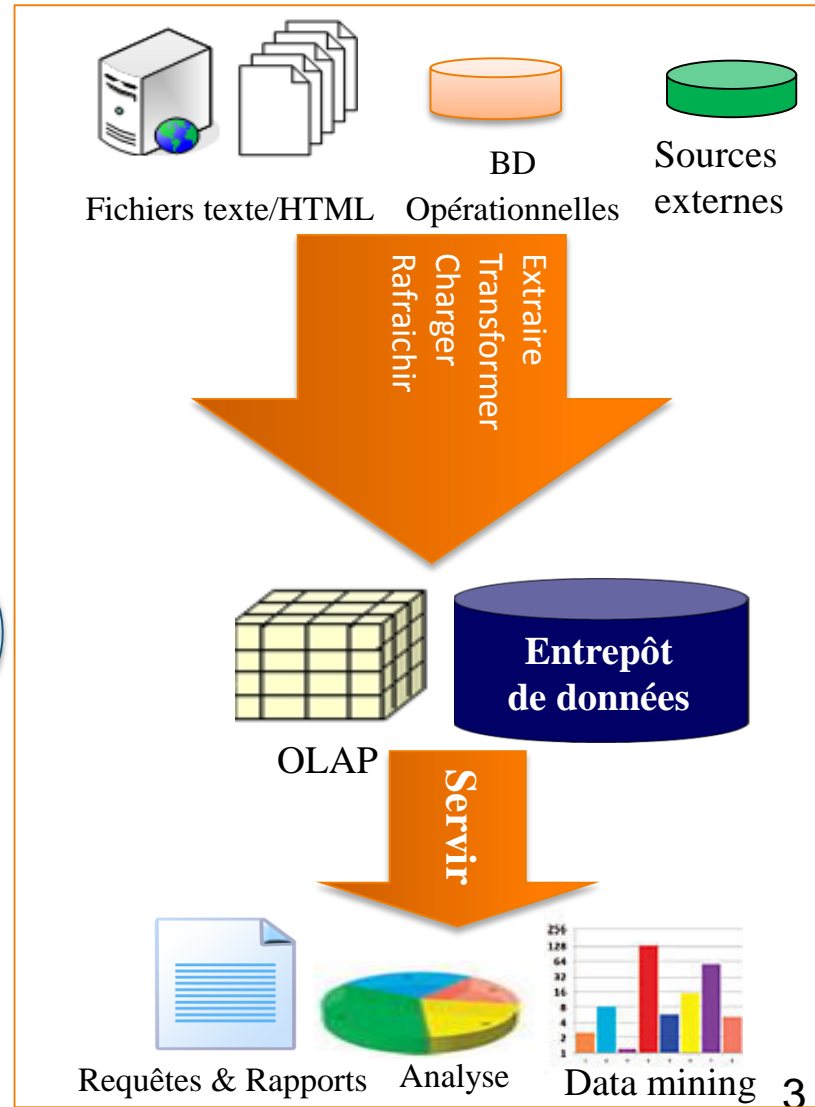
Motivation

- ❑ L'informatique opérationnelle est maîtrisée
- ❑ Autre type d'informatique => **informatique décisionnelle (BI)**

Besoin: prise de décisions stratégiques et tactiques



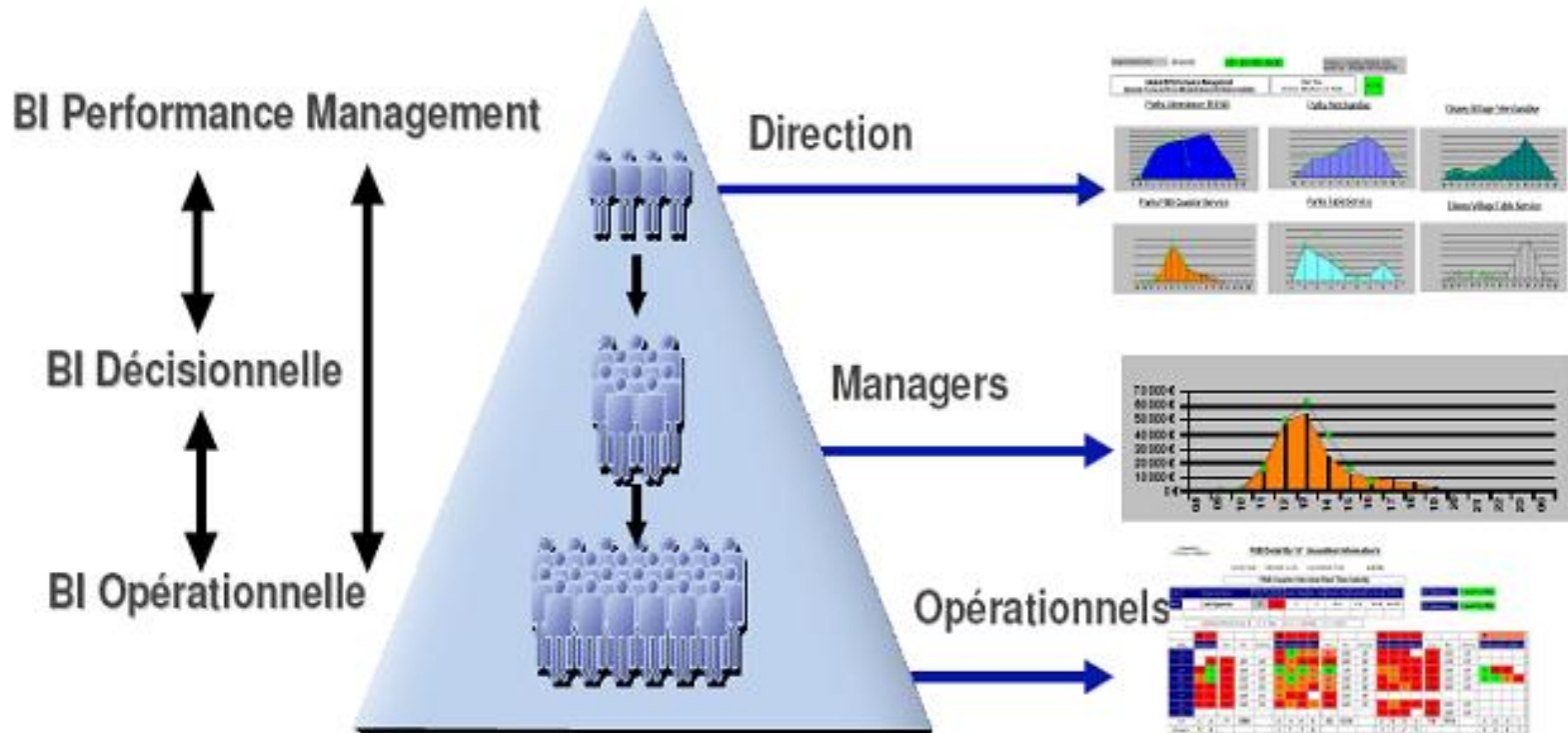
-Qui sont mes meilleurs clients?
-Pourquoi et comment le chiffre d'affaire a baissé?
-A combien s'élèvent mes ventes journalières?
.....



La place de la BI dans l'entreprise

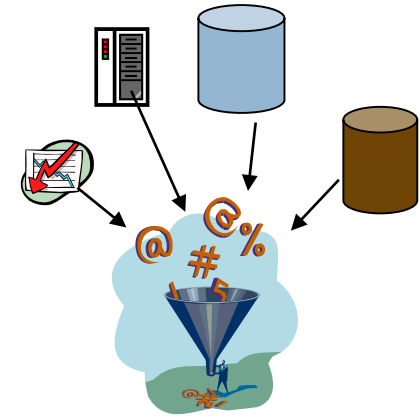
- La Business Intelligence aligne & connecte tous les niveaux de l'entreprise

Les besoins d'information sont différents à chaque échelon de l'entreprise



Les données utilisables par les décideurs

- Données opérationnelles (de production)
 - ▣ Bases de données (Oracle, SQL Server)
 - ▣ Fichiers, ...
 - ▣ Paye, gestion des RH, gestion des commandes...
- Caractéristiques de ces données:
 - ▣ Distribuées: systèmes éparpillés
 - ▣ Hétérogènes: systèmes et structures de données différents
 - ▣ Détaillées: organisation des données selon les processus fonctionnels, données surabondantes pour l'analyse
 - ▣ Peu/pas adaptées à l'analyse : les requêtes lourdes peuvent bloquer le système transactionnel
 - ▣ Volatiles: pas d'historisation systématique



Problématique

- Comment répondre aux demandes des décideurs?
 - ▣ En donnant un accès rapide et simple à l'information stratégique

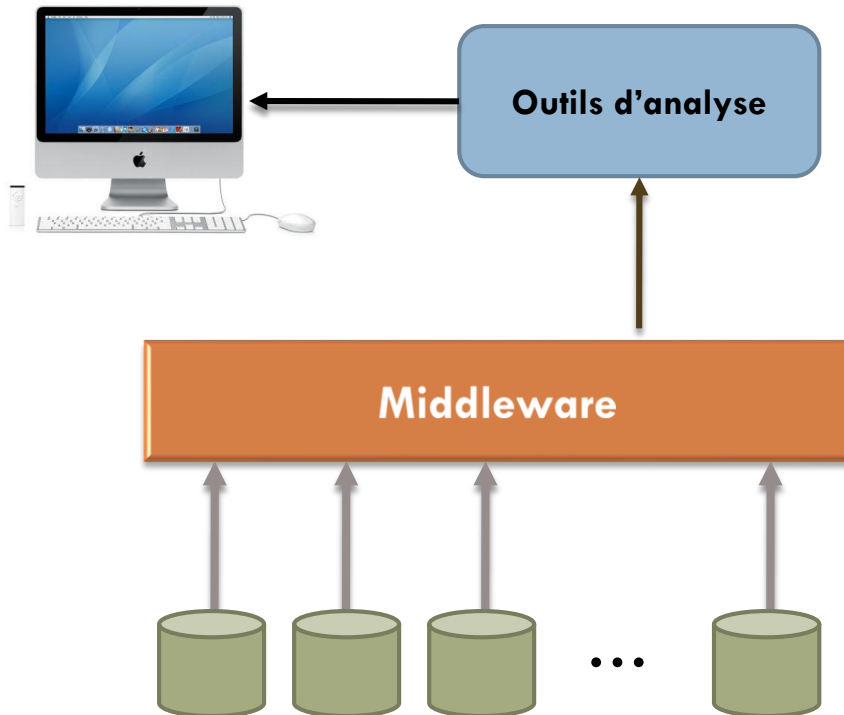
➔ Mettre en place un système d'information dédié aux applications décisionnelles:

datawarehouse

Solution 1 ☹️

Architecture d'un entrepôt de données (1)

□ Approche virtuelle (ou le non-entrepôt)

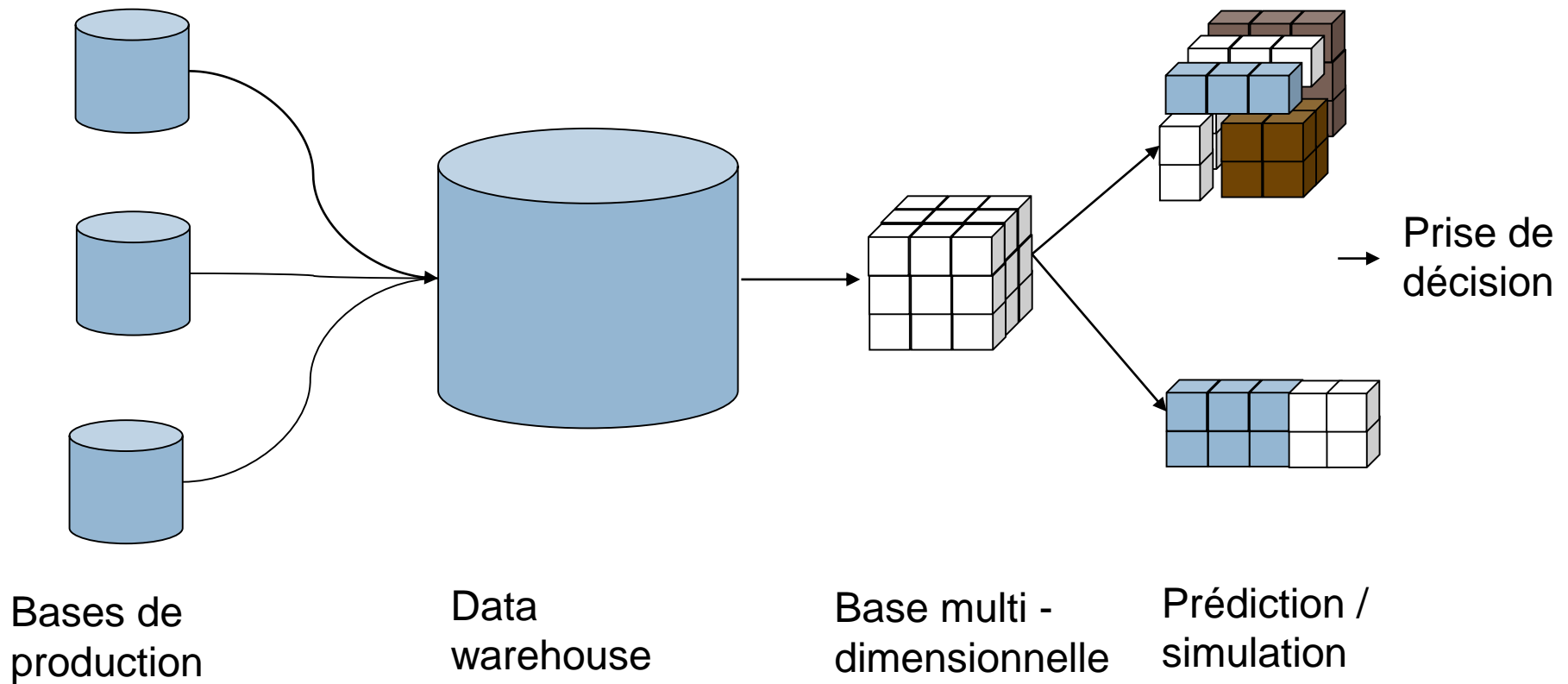


Inconvénients

- Pas de réelle intégration des données
- Pas de vues dans le temps
- Les requêtes peuvent facilement bloquer les transactions en cours

Solution 2 ☺

Le processus de prise de décision



Pourquoi un entrepôt de données?

□ Raisons d'être d'un entrepôt de données

- ▣ Rassembler les données de l'entreprise dans un **même lieu sans surcharger** les BD (systèmes opérationnels)
- ▣ Permettre **un accès universel** à diverses sources de données et assurer la **qualité** des données
- ▣ **Extraire, filtrer, et intégrer** les informations pertinentes, à l'avance, pour des requêtes ultérieures
- ▣ Dégager des **connaissances** et faire un apprentissage sur l'entreprise, le marché et l'environnement

Domaines d'applications

□ Banque, Assurance

- Détermination des profils client (prêt, ...)
- Risques d'un prêt, prime plus précise

□ Commerce

- Ciblage de clientèle
- Déterminer des promotions
- Aménagement des rayons (2 produits en corrélation)

□ Logistique

- Adéquation demande/production

□ Compagnies téléphoniques

□ Santé

- Risque alimentaire

C'est quoi un entrepôt de données?

C'est quoi un entrepôt de données?

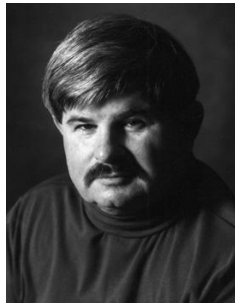
Définition ED

Entrepôt de Données : système central pour la prise de décision , « Un entrepôt est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision »

[Bill Inmon ,1992]



Entrepôt de données



C'est quoi un entrepôt de données?

Comme le Père du DataWarehouse le définit :



*« L'entrepôt de données est la ressource de présentation interrogeable des données d'une entreprise et elle ne doit pas être organisée autour d'un modèle entité relation, qui lui ferait perdre sa clarté et ses performances, Ces Source de données interrogeable de l'entreprise. C'est l'union des **DataMarts** qui le composent »*

[Kimball ,2010].

C'est quoi un entrepôt de données?

□ **Motés Clés ??**

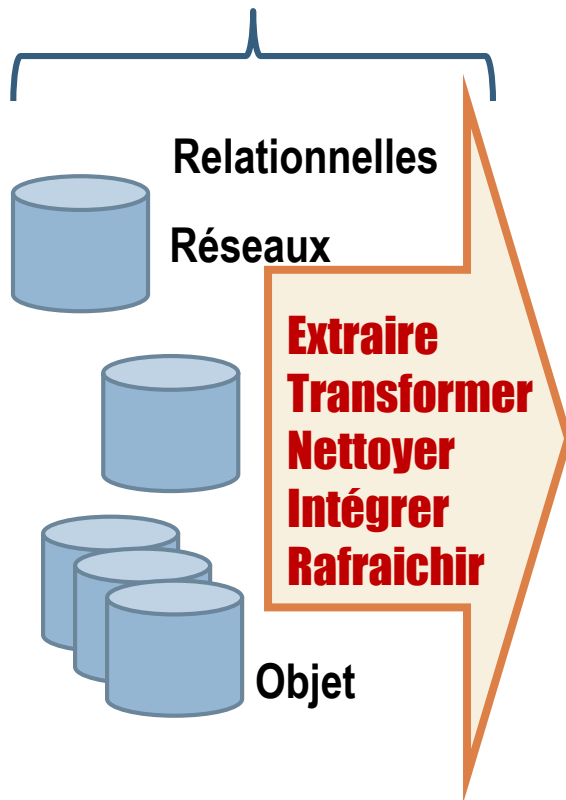
- Collection de données **orientées sujets**
- Consolidées dans une **base de données unique**
- **Non volatiles** et **historisées variante** dans le **temps**
- organisées pour le support d'un **processus d'aide à la décision**

- Dispositif de stockage d'informations **intégrées** de sources **distribuées, autonomes, hétérogènes**

Architecture de l'entrepôt de données?

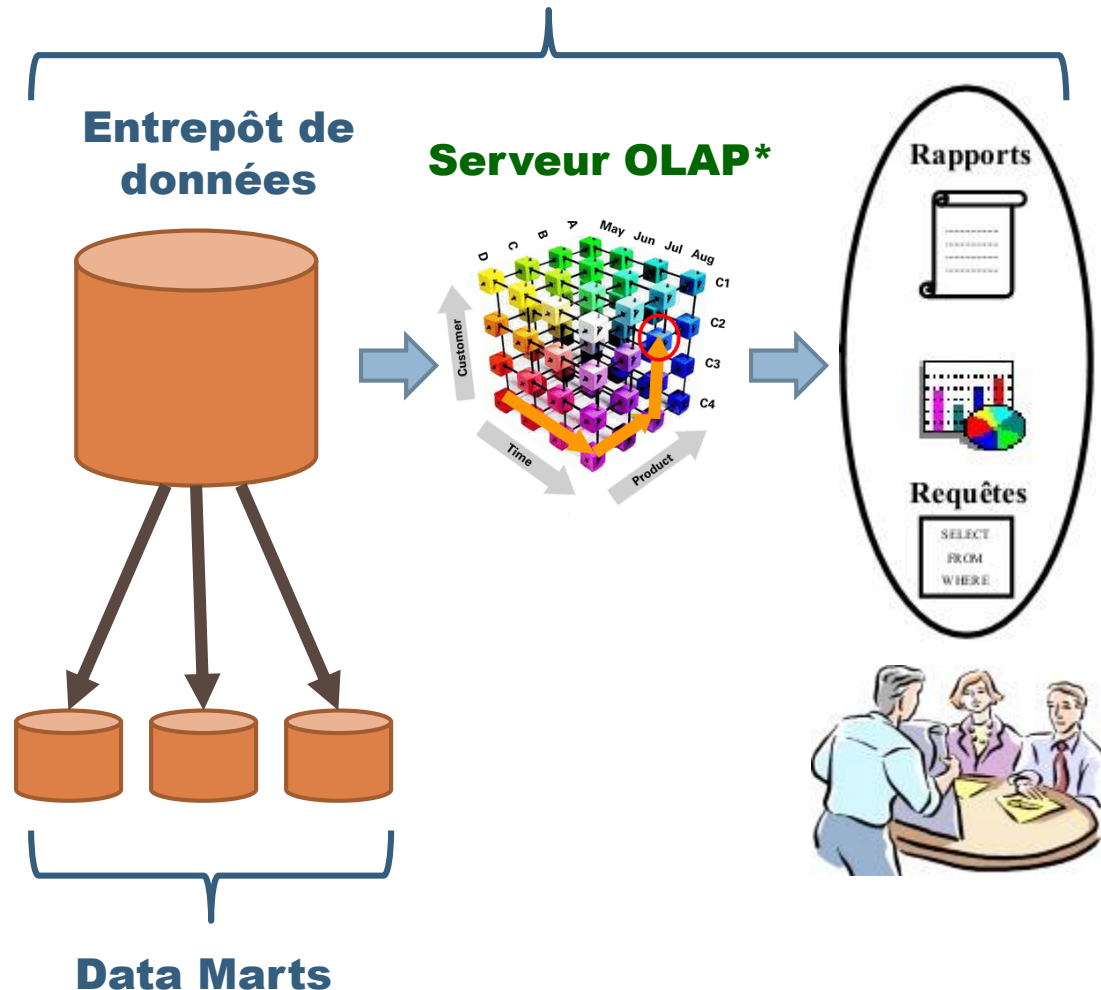
Architecture d'un entrepôt de données (2)

Phase Intégration



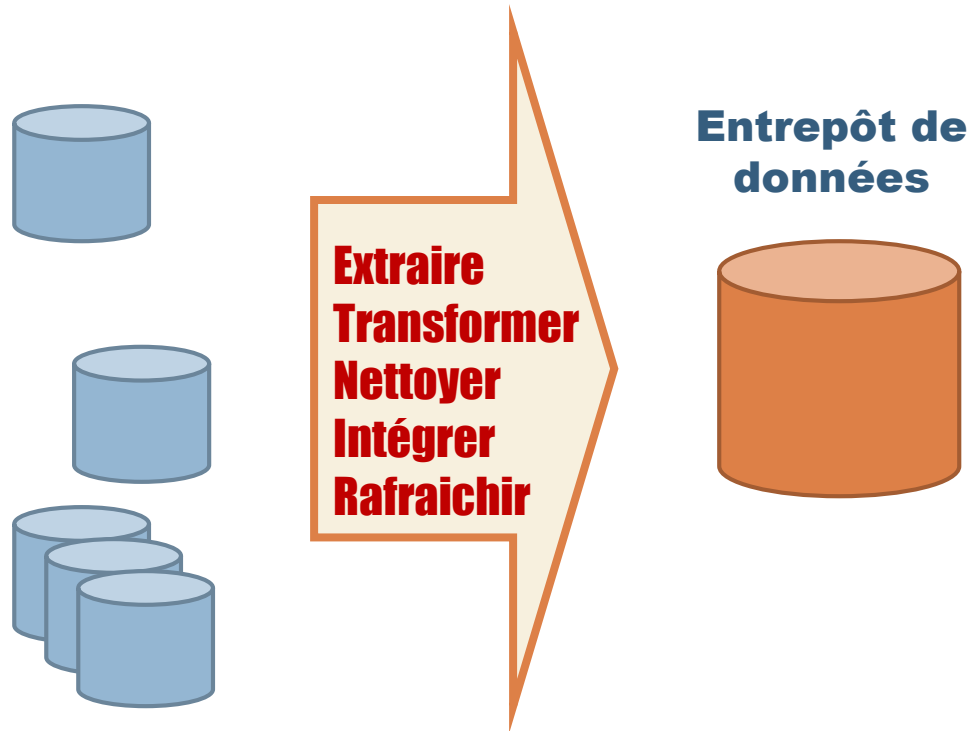
METADATA

Phase Traitement & Analyse



* On-Line Analytical Processing

Alimentation (ETL) d'un entrepôt de données



□ **Extraction**

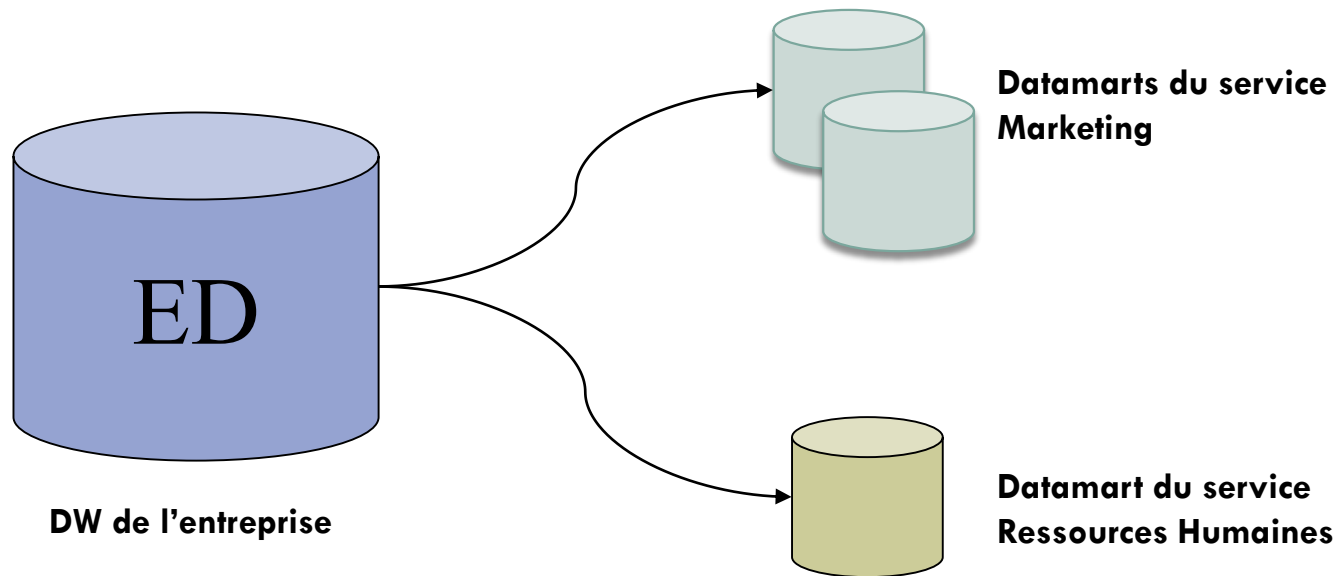
□ **Transformation**

- filtrer
- trier
- homogénéiser
- nettoyer
- ...

□ **Chargement(Loading)**

DATAMART

- Sous-ensemble d'un entrepôt de données
- Destiné à répondre aux besoins d'un secteur ou d'une fonction particulière de l'entreprise
- Point de vue spécifique selon des critères métiers



Intérêts des DATAMART

- ❑ Nouvel environnement structuré et formaté en fonction des besoins d'un métier ou d'un usage particulier
- ❑ Moins de données que DW
 - ▣ *Plus facile à comprendre, à manipuler*
 - ▣ *Amélioration des temps de réponse*
- ❑ Utilisateurs plus ciblés: DM plus facile à définir

Exploitation de l'entrepôt

□ Business Intelligence

- ▣ Possibilité de **visualiser** et **d'exploiter** une masse **importante** de données **complexes**

□ Trois principaux outils

- ① **OLAP** : **O**n-**L**ine **A**nalytical **P**rocessing
- ② **Data mining**: fouille de données
- ③ Formulation de **requêtes** et **visualisation** des résultats

Cube OLAP

Base de données vs. Entrepôt de données

□ Pourquoi dissocier une BD d'un ED?

- ▣ Les objectifs de **performances** dans les BD ne sont pas les mêmes que ceux dans les Eds
 - **BD** : requêtes **simples**, méthodes d'accès et d'indexation
 - **ED** : requêtes OLAP souvent **complexes** !!!
- ▣ La nécessité de **combiner** des données provenant de diverses sources, d'effectuer des **agrégations** dans un ED et d'offrir des **vues multidimensionnelles**
- ▣ Les données d'un ED sont souvent non **volatiles** et ont donc une plus longue durée de vie que celles d'une BD

Modélisation classique - OLTP

□ Le modèle relationnel

- ▣ Table, attributs, tuples, vues, ...
- ▣ Normalisation (*redondance*)
- ▣ Requêtes simples (*sélection, projection, jointure, ...*)

⇒ **Analyse difficile de l'activité**

□ Le critère temps

- ▣ Représentation du passé
 - *Un fardeau pour les systèmes OLTP*

Exemple


- Table historique

- ▣ Compte(NC, DateOp, Solde)

- Questions (ou Requêtes)

- ▣ Quel est le solde *courant* du client **525**? *% critère temps*

```
SELECT Solde
FROM Compte
WHERE NC = 525
AND DateOp = (SELECT MAX (DateOp)
               FROM Compte
               WHERE NC = 525)
```



- Quels sont les soldes courants de mes clients?

Requêtes décisionnelles plus complexes !!

□ **Exemples**

- Combien de clients âgés entre 20 et 30 ans et résidant à Alger ont-ils acheté une caméra vidéo au cours des **5 dernières années** ?
- Quelle est la répartition des ventes par produit, ville et par mois au cours de la présente **année**?
- Quelles sont les composantes des machines de production ayant eu le plus grand nombre d'incidents imprévisibles au cours de la période **1992-97** ?

⇒ Critère **temps** est la base de l'analyse décisionnelle

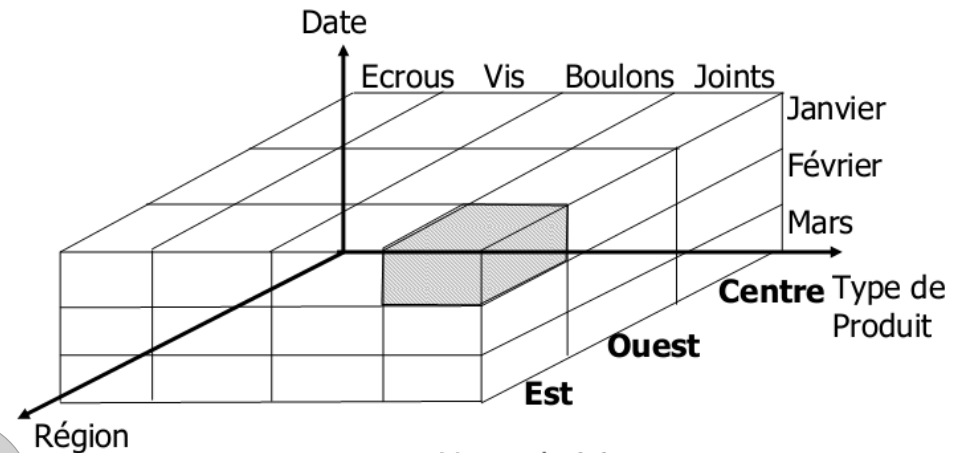
OLAP (On-Line Analytical Processing)

□ **Exemples**

- Traitement analytique interactif (**Codd**) typique dans les systèmes informationnels
- Catégorie de traitements dédiés à l'aide à la décision
- Analyses diverses (multidimensionnelles)
- Information : *surtout dérivée et sommaire*
- Aide à la prise de décision

Modélisation de l'entrepôt de données? MCD

Modélisation Multidimensionnelle



Vente de joints en janvier pour la région est

-Qui sont mes meilleurs clients?
-Pourquoi et comment le chiffre d'affaire a baissé?
-A combien s'élèvent mes ventes journalières?



Modélisation Multidimensionnelle

□ **Dimension**

- ▣ Présente le point de vue selon lequel on veut voir les données décrites par un ensemble d'attributs \Rightarrow **Axe de l'analyse.**
- ▣ Exemple: *Commandes, achats, réclamations, produits, clients,...*

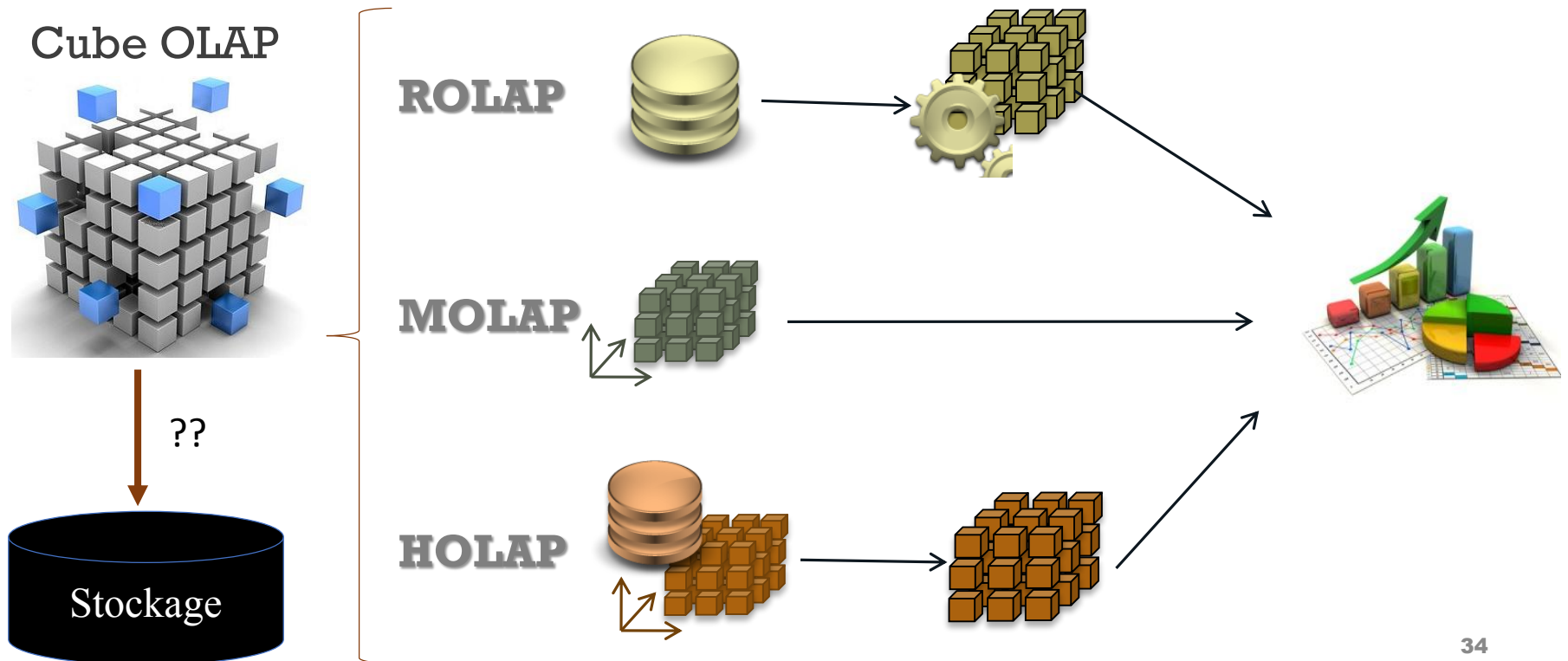
□ **Mesures / faits**

- ▣ Fonction numérique qui peut être évaluée en tout point du data cube en agrégeant les données correspondant à ce point \Rightarrow **mesure d'activité (critère d'analyse)**

Exemples des Faits

- ***Chiffre d'affaire, nombre de ventes, gain, ...***
- ***le fait de vente***
 - ▣ Chaque enregistrement de fait représente le total des ventes d'un produit dans un magasin dans une journée

Modélisation de l'entrepôt de données? MLD (ROLAP)



Comment stocker le cube de données ?

- **ROLAP: Relational On-Line Analytical Processing**

- ▣ The data cube is stored as relational table(s): a fact table with dimension tables.

- **MOLAP: Multidimensional On-Line Analytical Processing**

- ▣ The data cube is stored as multi-dimensional array(s).

- **HOLAP: Hybrid On-Line Analytical Processing**

- ▣ Is a **combination** of **ROLAP** and **MOLAP**

ROLAP

Requêtes de jointure en étoile

```
SELECT  P.Marque, sum(montant)
FROM    Ventes V, Produit P, Temps T, Client C
WHERE   V.PID = P.PID                (jointure)
AND     V.TID = T.TID                (jointure)
AND     V.CID = C.CID                (jointure)
AND     T.année = 2006                (sélection)
AND     P.Catégorie = 'Beauté'       (sélection)
AND     C.Ville = 'Poitiers'         (sélection)
GROUP BY P.Marque
```

-Qui sont mes meilleurs clients?
-Pourquoi et comment le chiffre d'affaire a baissé?
-A combien s'élèvent mes ventes journalières?

Schéma en étoile

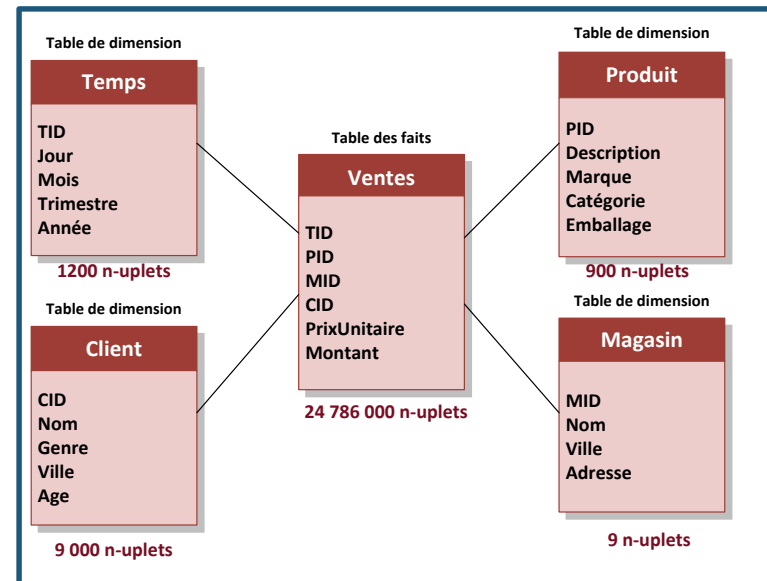


Table des faits (1)

- Table **principale** du modèle dimensionnel
- Contient les données observables (**les faits**) sur le sujet étudié selon divers axes d'analyse (**les dimensions**)

Table de faits des ventes	
Clés étrangères vers les dimensions	Clé date (CE)
	Clé produit (CE)
	Clé magasin (CE)
Faits	Quantité vendue
	Coût
	Montant des ventes

Table des faits (2)

□ Faits:

- ▣ *Ce que l'on souhaite mesurer*

 - Quantités vendues, montant des ventes...

- ▣ *Contient les clés étrangères des axes d'analyse (dimension)*

 - Date, produit, magasin

- ▣ *Trois types de faits:*

 - **Additif**

 - **Semi additif**

 - **Non additif**

Table des faits (3) - Typologie des faits

- **Additif:** additionnable suivant toutes les dimensions
 - ▣ *Quantités vendues, chiffre d'affaire*
 - ▣ *Peut être le résultat d'un calcul:*
 - **Bénéfice = montant vente - coût**
- **Semi additif:** additionnable suivant certaines dimensions
 - ▣ *Solde d'un compte bancaire:*
 - **Pas de sens d'additionner sur les dates car cela représente des instantanés d'un niveau**
 - **Σ sur les comptes: on connaît ce que nous possédons en banque**
- **Non additif:** fait non additionnable quelque soit la dimension
 - ▣ *Prix unitaire: l'addition sur n'importe quelle dimension donne un nombre dépourvu de sens*

Table des faits (4) - Granularité de la table des faits

□ Répondre à la question :

▣ Que représente un enregistrement de la table de faits?

□ ***La granularité définit le niveau de détails de la table de faits:***

▣ **Exemple:** une ligne de commande par produit, par client et par jour

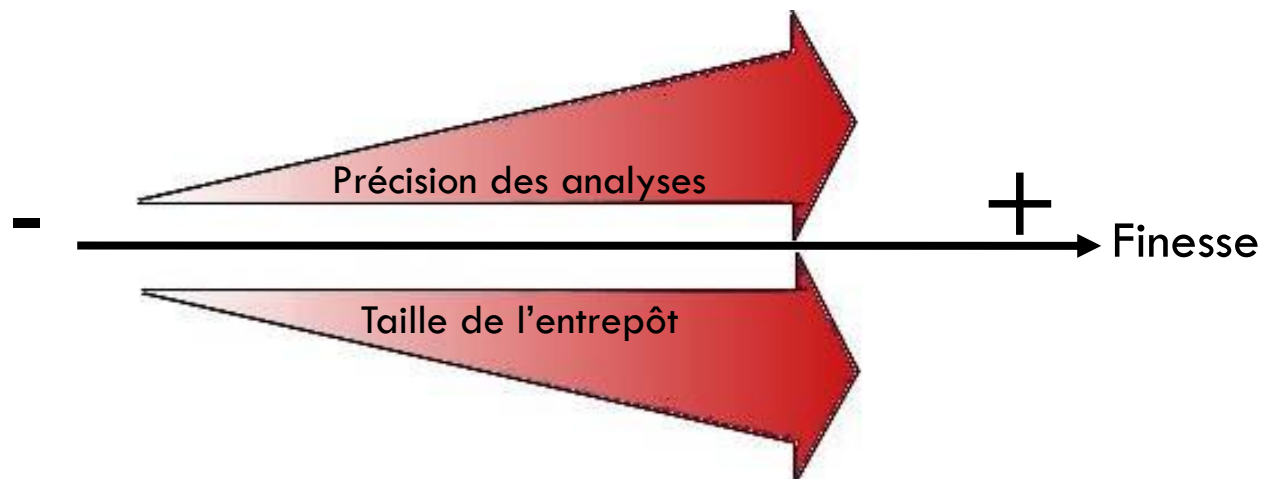


Table de dimension (1)

- **Axe d'analyse** selon lequel vont être étudiées les données observables (faits)
- Contient le **détail** sur les faits

Dimension produit	
Clé de substitution	Clé produit (CP)
	Code produit
Attributs de la dimension	Description du produit
	Famille du produits (Marque)
	Emballage
	Poids

Table de dimension (2)



- **Dimension = axe d'analyse**
 - ▣ Client, produit, période de temps...
- **Contient souvent un grand nombre de colonnes**
 - ▣ L'ensemble des informations descriptives des faits
- **Contient en général beaucoup moins d'enregistrements qu'une table de faits**

La dimension Temps

- **Commune** à l'ensemble des **DW**
- **Reliée** à toute **table de faits**

Dimension Temps	
Clé de substitution	Clé temps (CP)
	Jour
Attributs de la dimension	Mois
	Trimestre
	Semestre
	Année
	Num_jour_dans_année
	Num_semaine_ds_année

Modèle ROLAP

- **ROLAP: Relational On-Line Analytical Processing**
- Exploiter l'expérience des modèles relationnels (**un grand succès!!**)
- Il faut des modèles bien adaptés aux ED!
 - ▣ Schéma en **étoile** (star schema) 
 - ▣ Schéma en **flocon de neige** (snowflake schema) 

Modèle en étoile

- **Autant de tables de dimension qu'il existe de dimensions.**
 - **Exemple**
 - Temps, Produit, Client...
- **Une table de faits contenant la clé de chaque dimension et des mesures**
 - **Exemple**
 - montant en dollars, nombre d'unités vendues

Schéma en étoile

Table de dimensions

TEMPS
Code temps
Date
Année
Mois
Jour

1094 n-uplets

Table de dimensions

CLIENT
Code client
Sexe
Etat
Ville
Age

3 000 000 n-uplets

Table des faits

VENTES
Code temps
Code produit
Code client
Quantité vendue
Coût_dollars
Coût_unitaire

100 000 000 n-uplets

Table de dimensions

PRODUIT
Code produit
Nom produit
Prix unitaire
Taille
Poids
gamme
Type_paquet

300 000 n-uplets

Une Requête type

SELECT P.brand, sum(dollars_sold), sum(units_sold)
FROM SALES S, PRODUCT P, TIME T
WHERE S.PID = P.PID (**Jointure**)
AND S.TID = T.TID (**Jointure**)
AND T.Quarter = "1 Q 97" (**Sélection**)
GROUP BY P.brand
ORDER BY P.brand

Requêtes de jointure en étoile

- Plusieurs jointures
- Suivies par des sélections

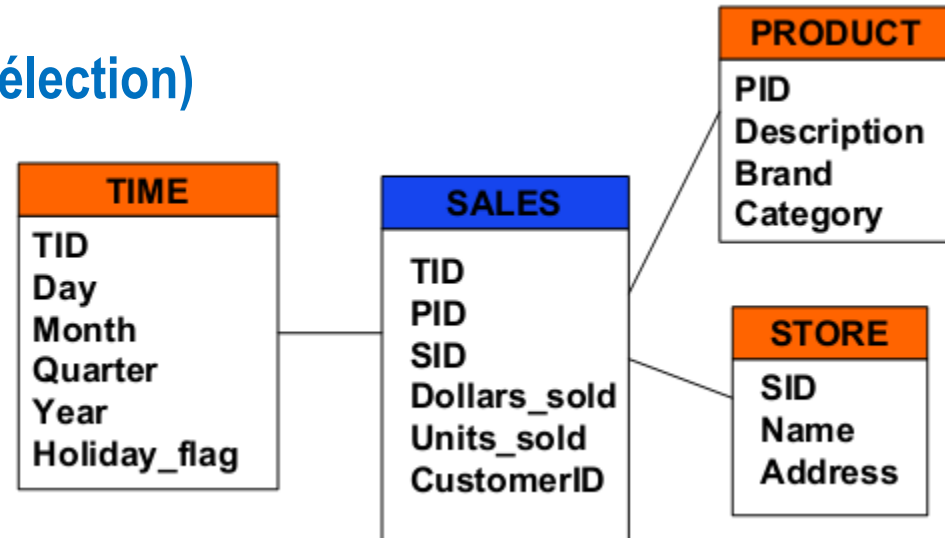


Schéma en étoile

Avantages & Inconvénients

□ Avantages

- ▣ Simple
- ▣ Le plus utilisé !!!

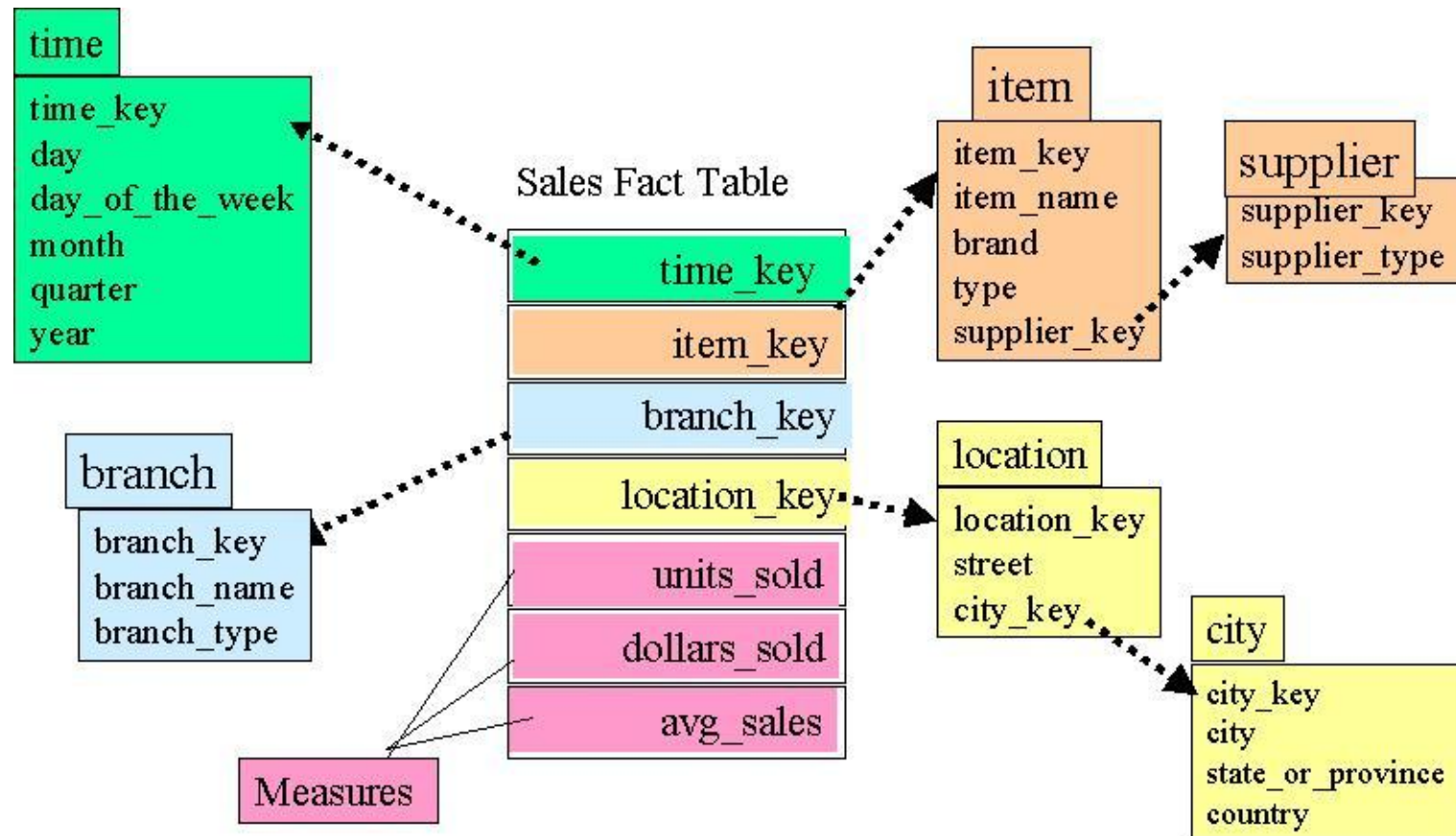
□ Inconvénients

- ▣ Possibilité de redondance car les tables de dimension ne sont pas nécessairement **normalisées**.
- ▣ Taille de dimensions plus grosse

Modèle en flocon de neige

- ❑ Variante du modèle en étoile.
- ❑ Les tables de dimensions sont **normalisées**
- ❑ Réduction de la redondance mais exécution parfois **plus lente** des requêtes (jointure de tables).
- ❑ Modèle adopté par **Oracle!!**
- ❑ Modèle mixte
 - ▣ Seules certaines tables sont normalisées

Exemple d'un modèle en flocon de neige



Granularité d'une dimension

- Une dimension contient des membres organisés en hiérarchie :
 - ▣ Chacun des membres appartient à un niveau hiérarchique (ou niveau de granularité) particulier
 - ▣ Granularité d'une dimension ⇔ nombre de niveaux hiérarchiques
 - ▣ Temps :
 - *année – semestre – trimestre - mois*

Modélisation de l'entrepôt de données? MLD (MOLAP)

Modèle MOLAP

- **MOLAP: Multidimensional On-Line Analytical Processing**
 - ▣ Utiliser un **système multidimensionnel « pur »** qui gère les structures multidimensionnelles natives (les cubes).
 - ▣ Accès direct aux données dans le cube.
- Plus difficile à mettre en place
- Formats souvent propriétaires
- Conçu exclusivement pour l'analyse multidimensionnelle.
- Exemples de moteurs MOLAP:
 - ▣ Microsoft Analysis Services
 - ▣ Hyperion

Manipulation des données multidimensionnelles - [1]

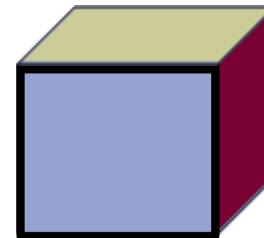
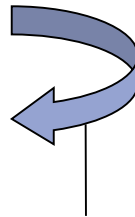
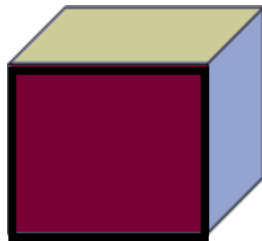
■ Opération agissant sur la structure

- ▣ Rotation (**rotate**): présenter une autre face du cube

	05	06	07
Œuf	221	263	139
Viande	275	257	116



	05	06	07
Idf	101	120	52
Ain	395	400	203



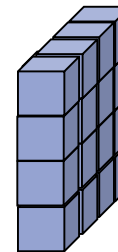
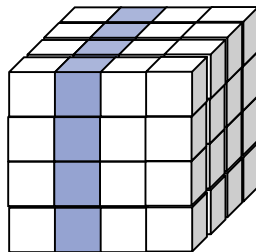
Manipulation des données multidimensionnelles - [2]

■ Opération agissant sur la structure

- ▣ Tranchage (**slicing**): consiste à ne travailler que sur une tranche du cube. Une des dimensions est alors réduite à une seule valeur

		05	06	07
Œuf	Idf	220	265	284
	Ain	225	245	240
Viande	Idf	163	152	145
	Ain	187	174	184

		06
Œuf	Idf	265
	Ain	245
Viande	Idf	152
	Ain	174



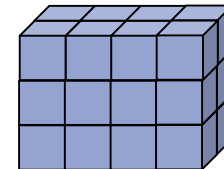
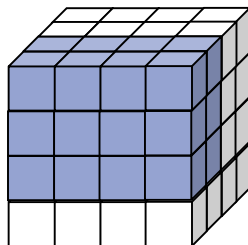
Manipulation des données multidimensionnelles - [3]

■ Opération agissant sur la structure

- ▣ Extraction d'un bloc de données (**dicing**): ne travailler que sous un sous-cube

		05	06	07
Œuf	Idf	220	265	284
	Ain	225	245	240
Viande	Idf	163	152	145
	Ain	187	174	184

		05	06	07
Œuf	Idf	220	265	284
	Ain	225	245	240



Manipulation des données multidimensionnelles - [4]

■ Opérations agissantes sur la granularité

▣ Forage vers le haut (**roll-up**): « *dézoomer* »

- Obtenir un niveau de granularité supérieur
- Utilisation de fonctions d'agrégation

▣ Forage vers le bas (**drill-down**): « *zoomer* »

- Obtenir un niveau de granularité inférieur
- Données plus détaillées

Manipulation des données multidimensionnelles - [4]

Roll-up, Drill-down

Roll up

Roll up

	05	06	07
Alim.	496	520	255

*Dimension
Temps*

	05-07
Fruits	623
Viande	648

	05	06	07
Fruits	221	263	139
Viande	275	257	116

	1S05	2S05	1S06	2S06	1S07
Fruits	100	121	111	152	139
Viande	134	141	120	137	116

Drill down

Drill down

*Dimension
Produit*

Approche suivie pour la modélisation

Middle-out

KIMBALL

Utilisateurs



INMON

Sources de données



Approche suivie pour la modélisation

Méthodologie de Kimball

« Se base sur l'architecture des DM indépendants »

1. Choisir le sujet
2. Choisir la granularité des faits
3. Identifier et adapter les dimensions
4. dimensions
5. Choisir la durée de la base
6. Suivre les dimensions lentement évolutives
7. Décider des requêtes prioritaires, des modes de requêtes
8. Choisir les faits
9. Stocker les pré-calculs

Quelques solutions open source

ETL	Entrepôt de données	OLAP	Reporting	Data Mining
<ul style="list-style-type: none">▪ Octopus▪ Kettle▪ CloverETL▪ Talend	<ul style="list-style-type: none">▪ MySql▪ Postgresql▪ Greenplum/Bizgres	<ul style="list-style-type: none">▪ Mondrian▪ Palo	<ul style="list-style-type: none">▪ Birt▪ Open Report▪ Jasper Report▪ JFreeReport	<ul style="list-style-type: none">▪ Weka▪ R-Project▪ Orange▪ Xelopes

Intégré

- Pentaho (Kettle, Mondrian, JFreeReport, Weka)
- SpagoBI

Les erreurs à éviter

[Barquin97]

- | Démarrer le projet sans la **bénédiction** des personnes clés dans l'entreprise
- | Placer la barre trop haut
 - | Objectifs trop ambitieux lors de la phase de l'élaboration de l'ED
 - | Frustration des membres de l'exécutif lorsque les objectifs ne sont pas atteints
- | Adopter un comportement "**politiquement naïf**"
- | L'ED va aider les gestionnaires à prendre de **meilleures décisions** 😊

Les erreurs à éviter

[Barquin97]

- | Charger l'ED avec trop d'information
 - | Simplement parce qu'elle est disponible dans les BD
- | La conception d'un ED est identique à celle d'une BD!!
- | Sous-estimer les besoins en performance et en capacité d'expansion
- | Choisir comme gestionnaire d'ED une personne davantage orientée vers la technologie que vers les usagers /