Fonctions d'un Data Warehouse

- Récupérer des données existants dans différentes BD sources
- Stocker les données (historisées)
- Mettre à disposition les données pour :
  - Interrogation
  - Visualisation
  - Analyse

3- Pourquoi ne pas utiliser un SGBD?

- SGBD et DW :
  - ont des objectifs différents et font des traitements différents
  - stockent des données différentes
  - font l'objet de requêtes différentes

  -> SGBD et DW ont besoin d'une organisation différente des données

  -> SGBD et DW doivent être physiquement séparés.

SGBD: Objectifs et traitements

- Les SGBD sont des systèmes dont le mode de travail est transactionnel (OLTP On-Line Transaction Processing).
- Permet d'insérer, modifier, interroger des informations rapidement, efficacement, en sécurité.
- Deux objectifs principaux :
  - Sélectionner, ajouter, mettre à jour et supprimer des tuples
  - Ces opérations doivent pouvoir être effectuées très rapidement, et par de nombreux utilisateurs simultanément.

DW: Objectifs et traitements

- Les datawarehouse sont des systèmes conçus pour l'aide à la prise de décision. (Mode de travail: OLAP On-Line Analytical Processing)
- La plupart du temps sont utilisés en lecture (utilisateurs)
- Les objectifs principaux sont
  - regrouper, organiser des informations provenant de sources diverses,
  - les intégrer et les stocker pour donner à l'utilisateur une vue orientée métier,
  - retrouver et analyser l'information facilement et rapidement.

Données différentes

- D'après BILL Inmon :

  "Un DW est une collection de données orientées sujet, intégrées, non volatiles, historisées, organisées pour la prise de décision."

  - Orientées sujet: thèmes par activités majeures ;
  - Intégrées: divers sources de données ;
  - Non volatiles: ne pas supprimer les données du DW ;
  - Historisées: trace des données, suivre l'évolution des indicateurs.

Requêtes (1)

- BD-OLTP représentent les données sous forme aplatie: relation, données normalisées

produit	région	vente	date	vendeur
écrou	Est	50	01012004	X
écrou	Ouest	60	12122003	X
écrou	Centre	110	01112003	Y
vis	Est	70	01042004	Y
vis	Ouest	80	10022004	Z
vis	Centre	90	29032004	Y
boulon	Est	120	05052004	X
boulon	Ouest	10	24042004	Z
boulon	Centre	20	11022004	Y
joint	Est	50	01032004	X
joint	Ouest	40	01102003	Y
joint	Centre	70	01012003	Z

produit	prix	fournisseur
écrou	44	CC
vis	2	DD
boulon	3	VV
joint	1	BB

fournisseur	ville
...	...

Requêtes (2)

- OLTP: Requêtes simples "qui, quoi"
  - par ex. les ventes de X.
  - jointures: les ventes de X à quel prix de quel fournisseur,
- OLAP: besoin de données agrégées, synthétisées
  - nombre de ventes par vendeur, par région, par mois,
  - nombre de ventes par vendeur, par fournisseur, par mois,
  - ...
- SQL: Possibilité d'agréger les données (group by) mais très coûteux (parcourir toutes les tables) et il faut recalculer à chaque utilisation
- Sur plusieurs tables (ex : somme des ventes par fournisseur), nécessité de faire des jointures souvent coûteuses

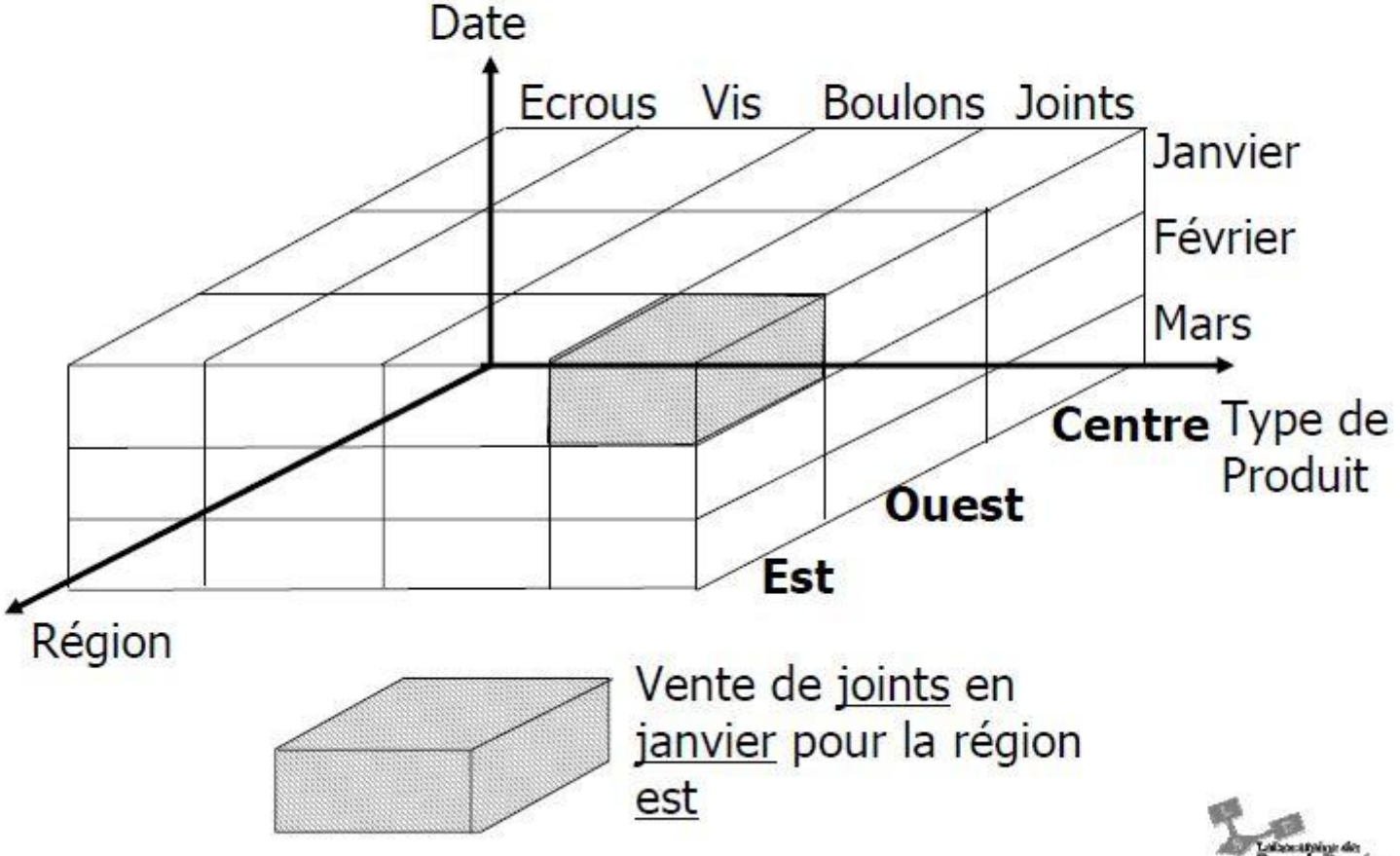
Différences BD – DW

Caractéristiques	OLTP	OLAP
Applications	production	aide à la décision
Utilisateurs	un département professionnel IT	transversal (entreprise) décideur non IT
Données	normalisées, non agrégées	dénormalisées, agrégées
Requêtes	simples, nombreuses, régulières, prévisibles, répétitives	complexes, peu nombreuses, irrégulières, non prévisibles
Nb tuples invoqués par requête (moyenne)	dizaines	millions
Taille données	100 MB à 1 GB	1 GB à 1 TB
Ancienneté des données	récente, mises à jour	historique

Nécessité d'une structure muti-dimensionnelle

- Les BD relationnelles ne sont pas adaptées à l'OLAP car :
  - Pas les mêmes objectifs
  - Pas les mêmes données:
    - Les données nécessaires à l'OLAP sont multi-dimensionnelles (i.e. ventes par vendeur, par date, par ville…). Les tables en représentent une vue aplatie.
  - Pas les mêmes traitements et requêtes:
    - Non seulement perte de performances mais aussi nécessité pour les utilisateurs de savoir comment trouver les liens entre les tables pour recréer la vue multi-dimensionnelle.
- Il est donc nécessaire de disposer d'une structure de stockage adaptée à l'OLAP, i.e. permettant de
  - représenter les données dans plusieurs dimensions,
  - manipuler les données facilement et efficacement.

Cube: représentation des données sous forme multidimensionnelle

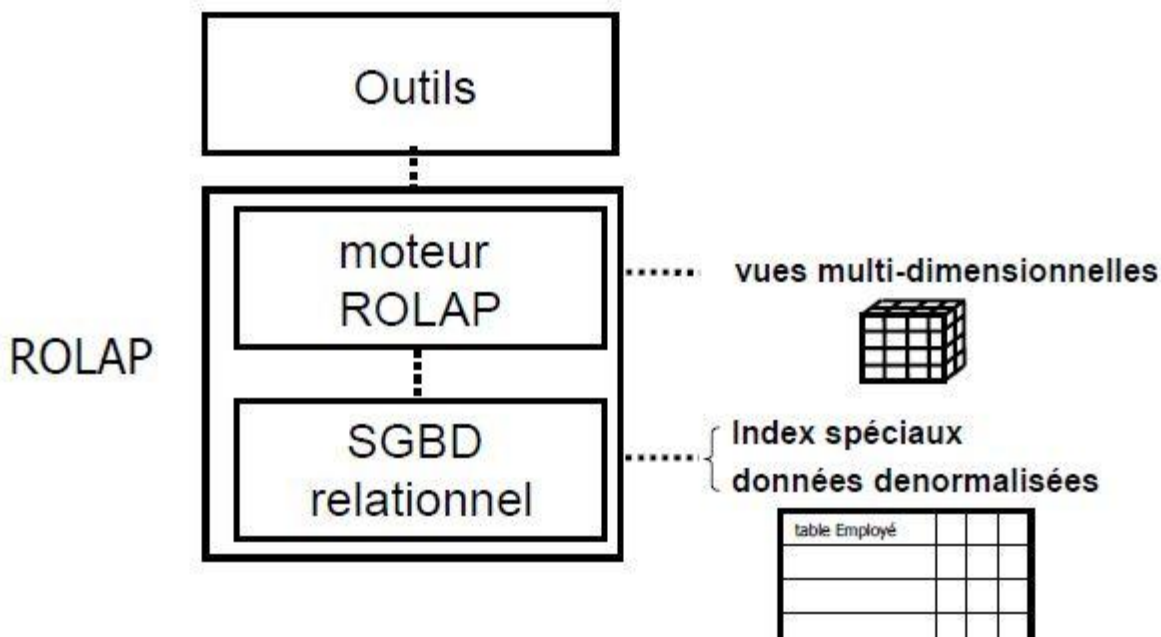


Approches pour créer un DW

- 3 possibilités:
- (1) Relational OLAP (ROLAP)
  - Données sont stockées dans un SGBD relationnel
  - Un moteur OLAP permet de simuler le comportement d'un SGBD multi-dimensionnel
- (2) Multidimensional OLAP (MOLAP)
  - Structure de stockage en cube
  - Accès direct aux données dans le cube
- (3) Hybrid OLAP (HOLAP)
  - Données stockées dans SGBD relationnel (données de base)
  - + structure de stockage en cube (données agrégées)

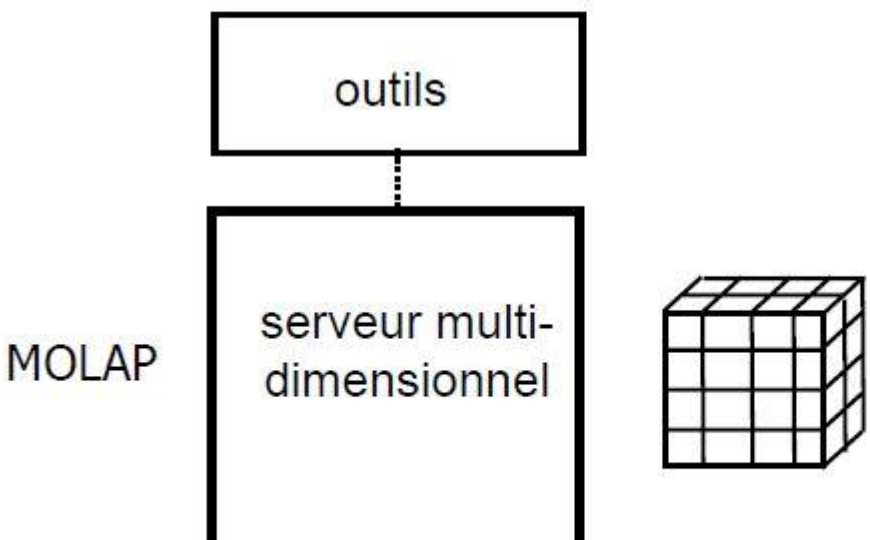
ROLAP

- Relational OLAP



MOLAP

- Multi-Dimensional OLAP

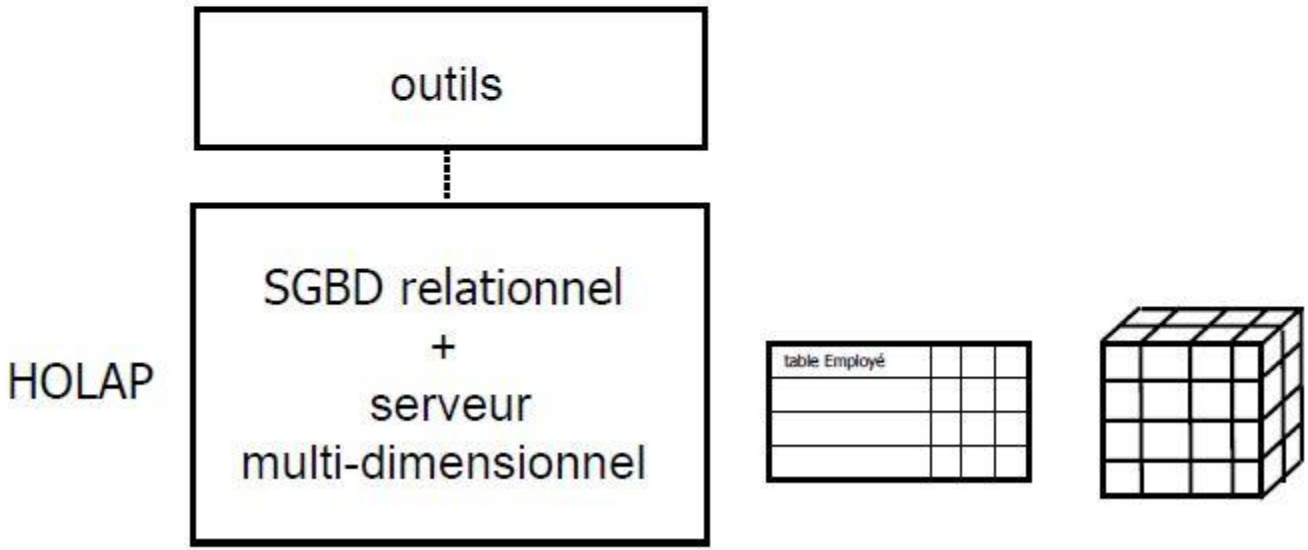


HOLAP

- Idée:
  - MOLAP + ROLAP
  - Données stockées dans des tables relationnelles
  - Données agrégées stockées dans des cubes.
  - Les requêtes vont chercher les données dans les tables et les cubes

HOLAP

- Hybrid OLAP



Définitions

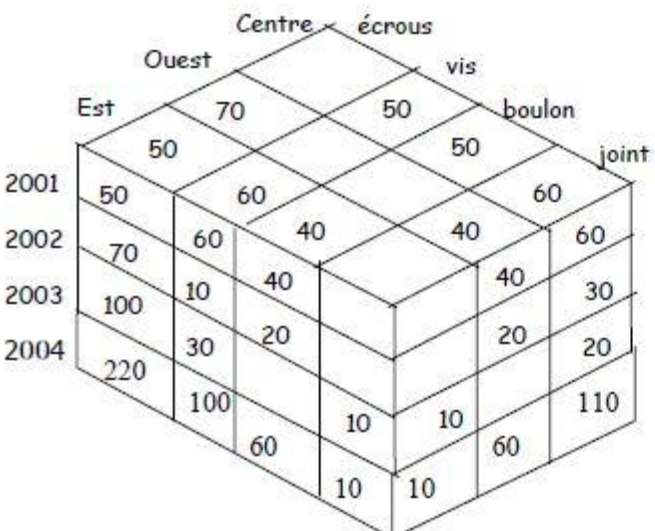
- Principe de base : ce sont les analyses des indicateurs qui intéressent l'utilisateur
- Le modèle multidimensionnel contient 2 types d'attributs : les *dimensions* et les *mesures*
- Les mesures sont les valeurs numériques que l'on compare (ex : montant_ventes, qte_vendue)
  - Ces valeurs sont le résultat d'une opération d'agrégation des données
- Les dimensions sont les points de vues depuis lesquels les mesures peuvent être observées :
  - Ex : date, région, type de produit, etc.

Slicing et dicing

- Slicing: Sélection de tranches du cube par des prédicats selon une dimension
  - filtrer une dimension selon une valeur
  - Exemple: Slice (2004) : on ne retient que la partie du cube qui correspond à cette date
- Dicing: extraction d'un sous-cube

Exemple: slicing

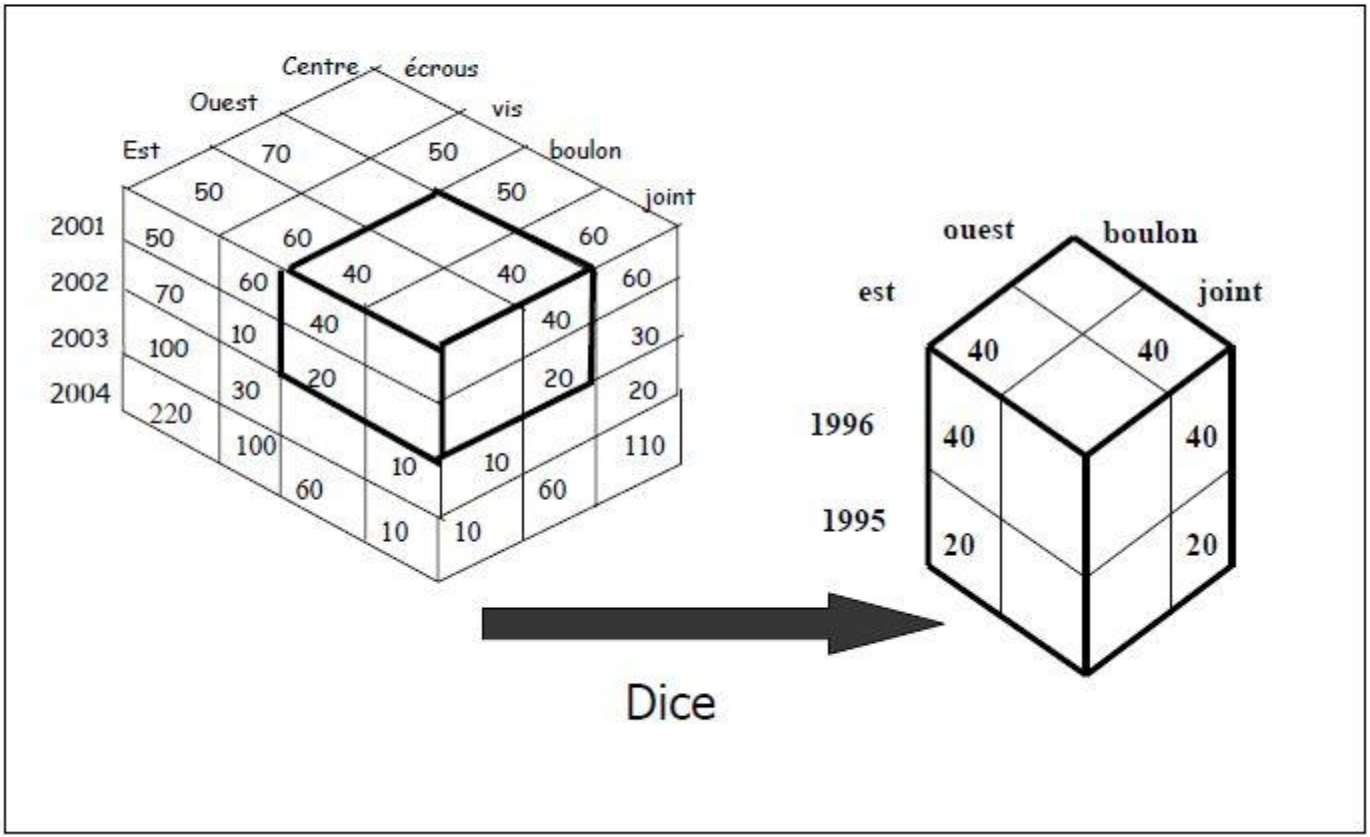
Slice



Slice (2004)

Ventes 2004	écrou	vis	boulon	joint
est	220	100	60	10
ouest	160	50	10	60
centre	20	150	170	110

Exemple: Dicing



Modélisation en étoile ou en flocons

- Modélisation conceptuelle BD : entité et relation
- Modélisation de DW : dimension et mesure
- Les mesures sont les valeurs numériques que l'on compare (ex : montant_ventes, qte_vendue)
  - Ces valeurs sont le résultat d'une opération d'agrégation des données
- Les dimensions sont les points de vues depuis lesquels les mesures peuvent être observées :
  - Ex : date, localisation, produit, etc.
  - Elles sont stockées dans les tables de dimensions

Les dimensions

- Une dimension peut être définie comme :
  - un thème, ou un axe (attributs), selon lequel les données seront analysées
  - Ex : Temps, Découpage administratif, Produits
- Une dimension contient des membres organisés en hiérarchie :
  - Chacun des membres appartient à un niveau hiérarchique (ou niveau de granularité) particulier
  - Ex : pour la dimension Temps: année – semestre – mois – jour

Les mesures

- Une mesure est un élément de donnée sur lequel portent les analyses, en fonction des différentes dimensions
  - Ex : coût des travaux, nombre d'accidents, ventes

Les faits

- Un fait représente la valeur d'une mesure, mesurée ou calculée, selon un membre de chacune des dimensions
  - Exemple : «250 000 euros » est un fait qui exprime la valeur de la mesure « coût des travaux » pour le membre « 2002 » du niveau année de la dimension « temps » et le membre « Versailles » du niveau « ville » de la dimension « découpage administratif »

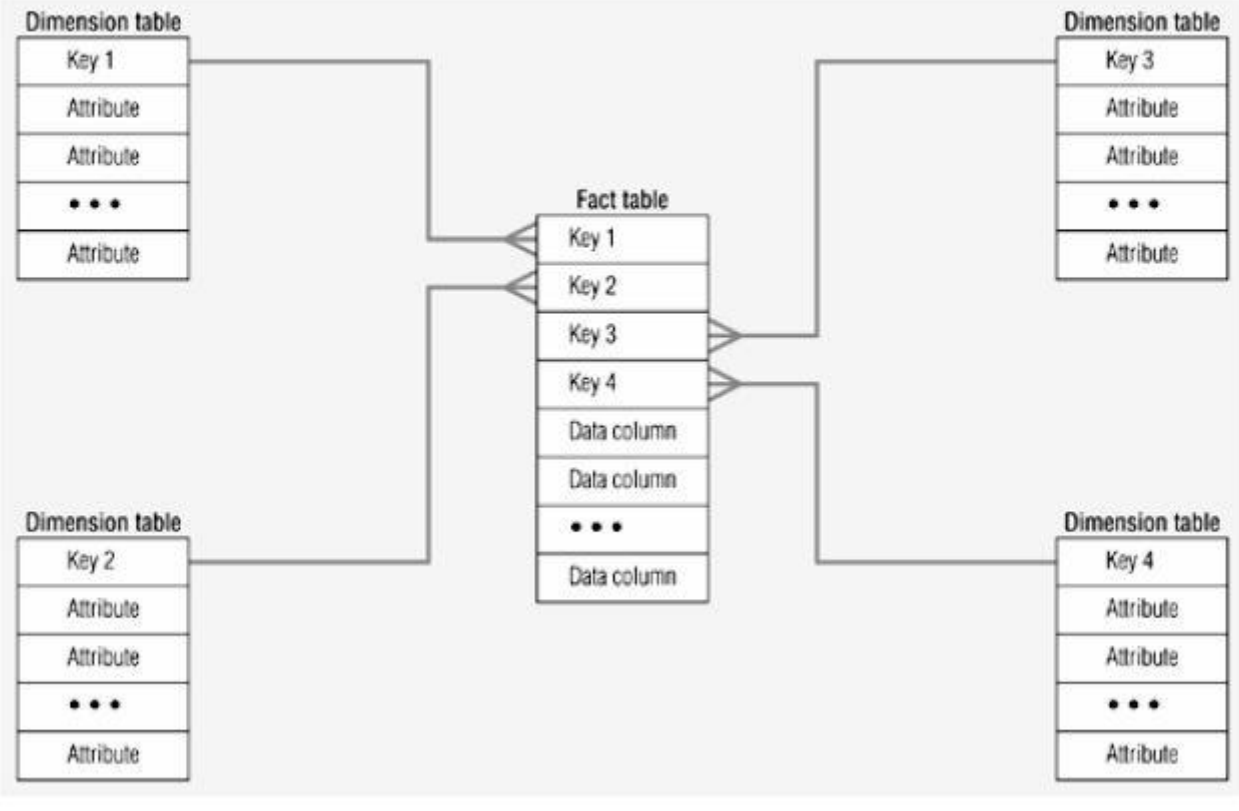
La table de faits

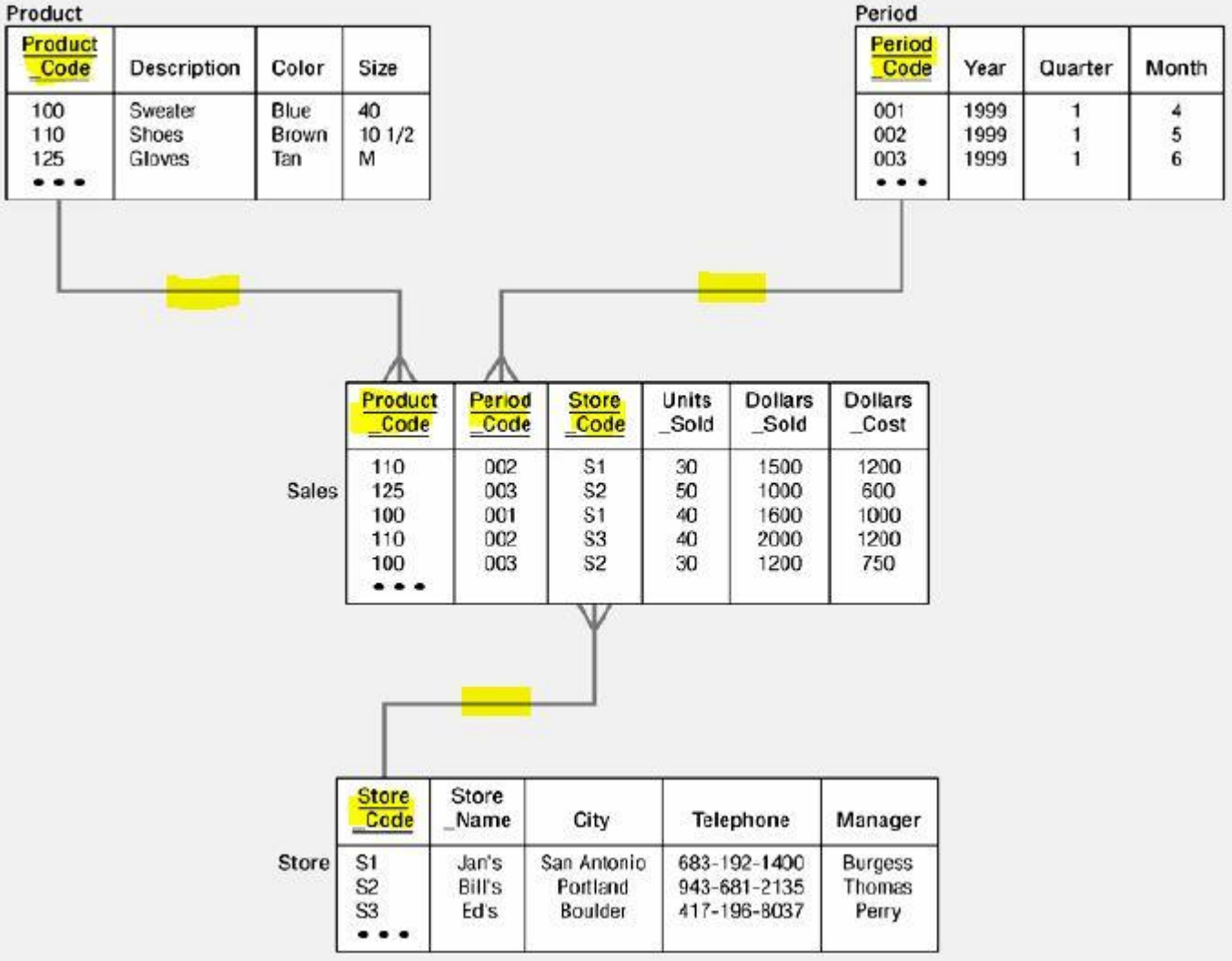
- Les mesures sont stockées dans les tables de faits
  - Table de fait contient les valeurs des mesures et les clés vers les tables de dimensions

Le modèle en étoile

- Une (ou plusieurs) table(s) de faits comprenant une ou plusieurs mesures.
- Plusieurs tables de dimension dénormalisées: descripteurs des dimensions.
- Les tables de dimension n'ont pas de lien entre elles
- Avantages :
  - Facilité de navigation
  - Performances : nombre de jointures limité ; gestion des données creuses.
  - Gestion des agrégats
- Inconvénients :
  - Toutes les dimensions ne concernent pas les mesures
  - Redondances dans les dimensions
  - Alimentation complexe.

Modèle en étoile




Product
Product _Code
Description
Color
Size
100
Sweater
Blue
40
110
Shoes
Brown
10 1/2
125
Gloves
Tan
M
Period
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

Le modèle en flocons

- Le schéma en flocon est dérivé du schéma en étoile où les tables de dimensions sont normalisées (la table des faits reste inchangée).
- Avec ce schéma, chacune des dimensions est décomposée selon sa (ou ses) hiérarchie(s).
- Exemple : Commune, Département, Région, Pays, Continent

Client	Continent	Pays	Region	Département	Commune
Pepone	Europe	France	RhôneAlpes	Rhône	Lyon1
Testut	Europe	France	RhôneAlpes	Rhône	Lyon2
Soinin	Europe	France	RhôneAlpes	Rhône	Lyon3
Vepont	Europe	France	Ile de France	Paris	Paris1
Martin	Europe	France	Ile de France	Paris	Paris2
Elvert	Europe	France	Ile de France	Yvelines	Versailles