

DATA WAREHOUSE

Départements : Logiciel et systèmes d'information

Disciplines: Sciences des données et intelligence artificielle

Enseignant : EL GHEBOULI Ayoub

2022/2023



SEANCE 3

Les concepts de base de Data Warehouse

Données multidimensionnelles

- Notion de dimension : C'est une catégorie linguistique selon laquelle les données sont organisées:
 - Nom d'un attribut
 - Valeur d'un attribut

Représentation :

| DuréeMoy | Départ. | Mois | Année |
|----------|---------|------|-------|
| 5 | Info | Janv | 1998 |
| 5 | Phys | Janv | 1998 |
| 18 | Philo | Janv | 1998 |
| 7 | Droit | Janv | 1998 |
| 12 | Info | Févr | 1998 |
| 8 | Phys | Févr | 1998 |
| 9 | Philo | Févr | 1998 |
| 15 | Droit | Févr | 1998 |
| 18 | Info | Mars | 1998 |
| 12 | Phys | Mars | 1998 |
| 22 | Philo | Mars | 1998 |
| 25 | Droit | Mars | 1998 |

Tableau simple

| 1998 | Janv | Févr | Mars |
|-------|------|------|------|
| Info | 5 | 12 | 18 |
| Phys | 5 | 8 | 12 |
| Philo | 18 | 9 | 22 |
| Droit | 7 | 15 | 25 |

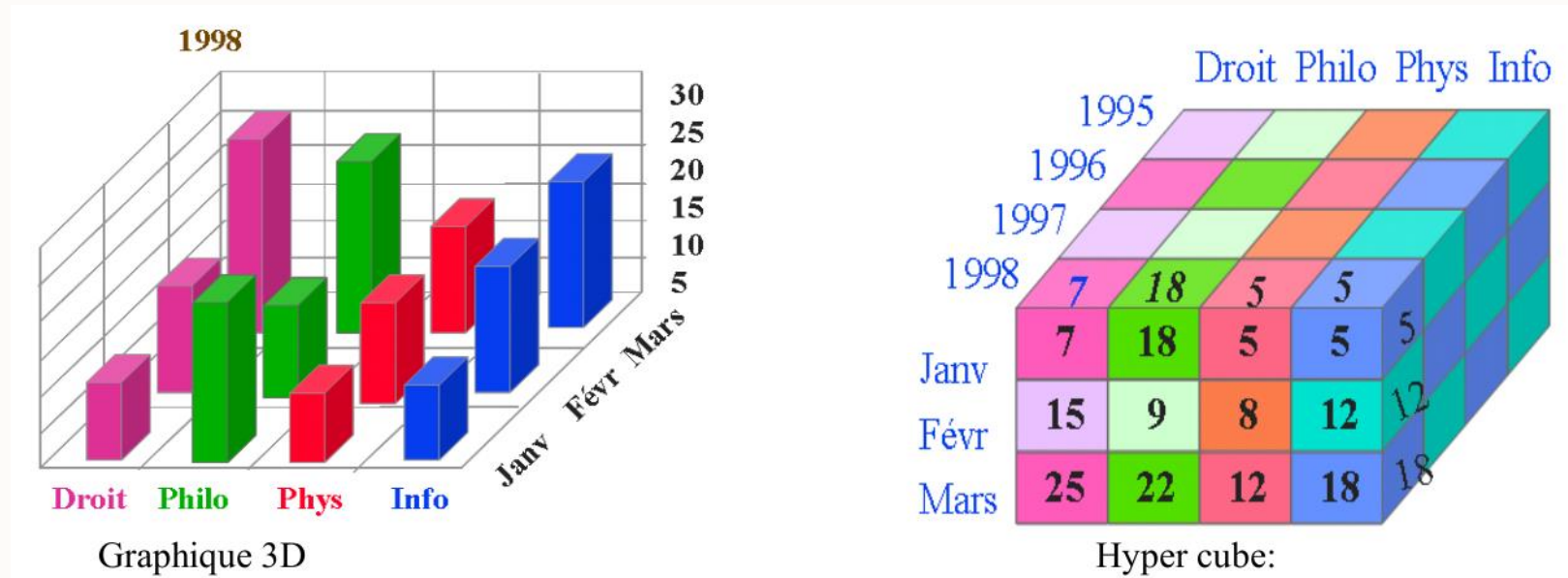
Tableau croisé

Les concepts de base de Data Warehouse

Données multidimensionnelles

- Notion de dimension : C'est une catégorie linguistique selon laquelle les données sont organisées:
 - Nom d'un attribut
 - Valeur d'un attribut

Représentation:



Les concepts de base de Data Warehouse

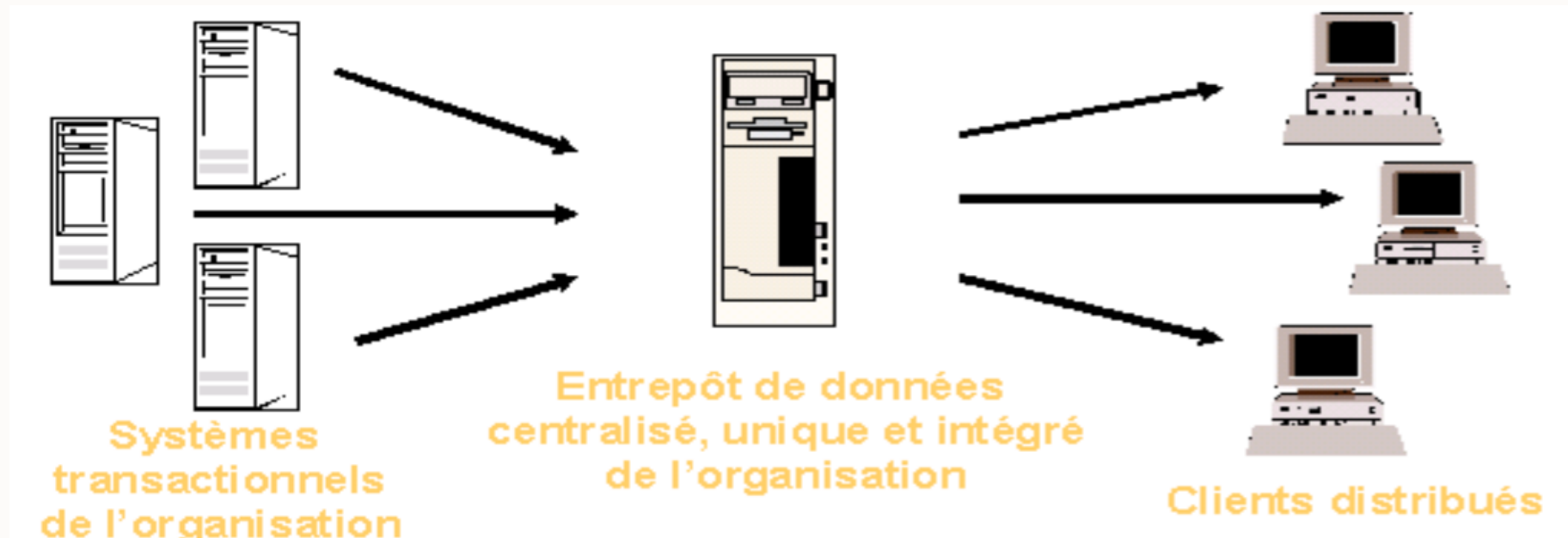
Exemple: un DW dans les télécoms

- **Sujets :**
 - Suivi du marché: lignes installées/ désinstallées, services et options choisis, répartition géographique, répartition entre public et différents secteurs d'organisations
 - 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
- **Sources**
 - Fichiers clients élaborés par les agences
 - Fichiers de facturation
- **Requêtes**
 - Comportement clientèle
 - Nombre moyen d'heures par client, par mois et par région
 - Durée moyenne d'une communication urbaine par ville
 - Durée moyenne d'une communication internationale

Les concepts de base de Data Warehouse

Architecture d'un Datawarehouse

Architecture centralisée

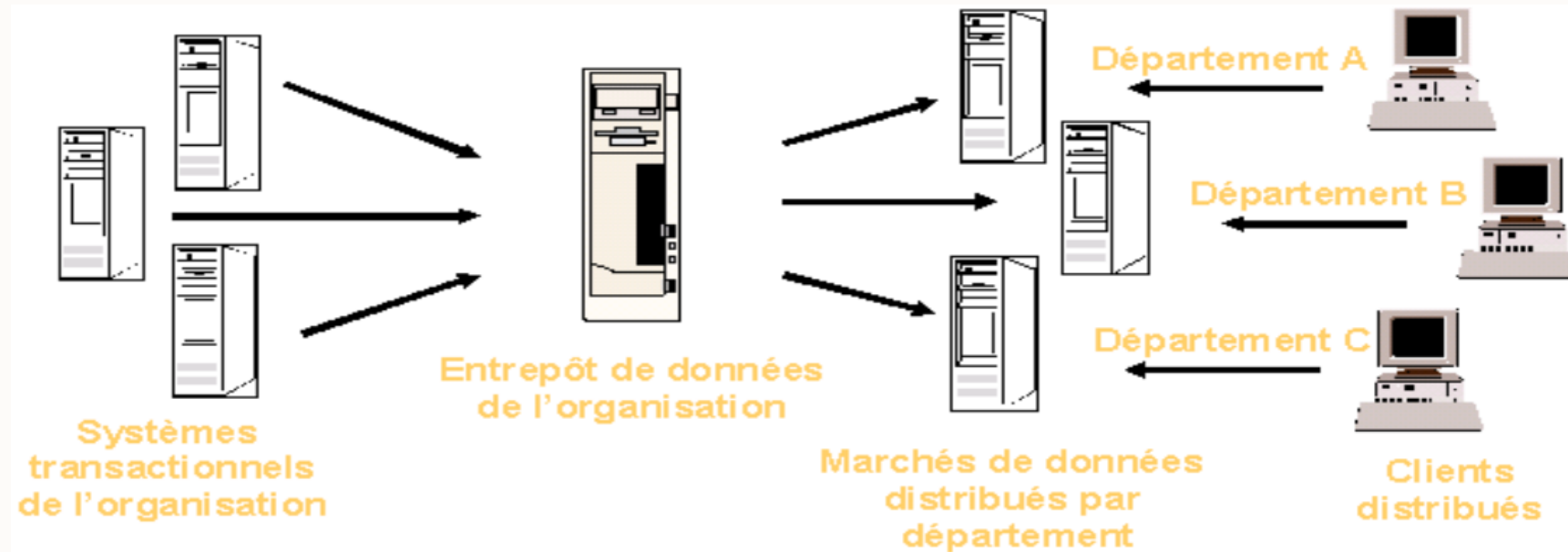


-> Il s'agit de la version centralisée et intégrée d'un entrepôt regroupant l'ensemble des données de l'entreprise. Les différentes bases de données sources sont intégrées et sont distribuées à partir de la même plate-forme physique

Les concepts de base de Data Warehouse

Architecture d'un Datawarehouse

Architecture fédérée



-> Il s'agit de la version intégrée d'un entrepôt où les données sont introduites dans les marchés de données orientés selon les différentes fonctions de l'entreprise

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

OLAP :

- « Il s'agit d'une catégorie de logiciels axés sur l'exploration et l'analyse rapide des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation ».
- OLAP vise à assister l'utilisateur dans son analyse en lui facilitant l'exploration de ses données et en lui donnant la possibilité de le faire rapidement.
 - L'utilisateur n'a pas à maîtriser des langages d'interrogation et des interfaces complexes
 - L'utilisateur interroge directement les données, en interagissant avec celles-ci

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

OLAP :

- OLAP (Online Analytical Processing) permet aux utilisateurs d'analyser des données présentes de plusieurs systèmes de bases de données en même temps. Les données OLAP sont multidimensionnelles, ce qui signifie que l'information peut être comparée de nombreuses façons différentes. Par exemple, une entreprise peut comparer ses ventes d'ordinateurs en juin avec ses ventes en juillet, puis comparer ces résultats avec les ventes d'un autre endroit, qui pourraient être stockées dans une base de données différente.
- Un serveur OLAP est nécessaire pour organiser et comparer les informations. Les clients peuvent analyser différents ensembles de données à l'aide des fonctions intégrées au serveur OLAP

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Différence entre OLTP et OLAP ?

| | OLTP | OLAP |
|---------------|---|---|
| Définition | C'est un système transactionnel en ligne qui sert à effectuer des modifications dans une base de données. | C'est un système de récupération de données et d'analyse de données en ligne. |
| Transaction | OLTP a des transactions courtes. | OLAP a des transactions longues. |
| Les données | OLTP et ses transactions constituent la source originale de données. | Différentes bases de données OLTP deviennent la source de données pour OLAP. |
| Intégrité | La base de données OLTP doit maintenir la contrainte d'intégrité des données. | La base de données OLAP n'est pas fréquemment modifiée. Par conséquent, l'intégrité des données n'est pas affectée. |
| Normalisation | Les tables dans la base de données OLTP sont normalisées (3NF). | Les tables dans la base de données OLAP ne sont pas normalisées. |
| Requêtes | Des requêtes plus simples. | Des Requêtes plus complexes |

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Modèle conceptuel

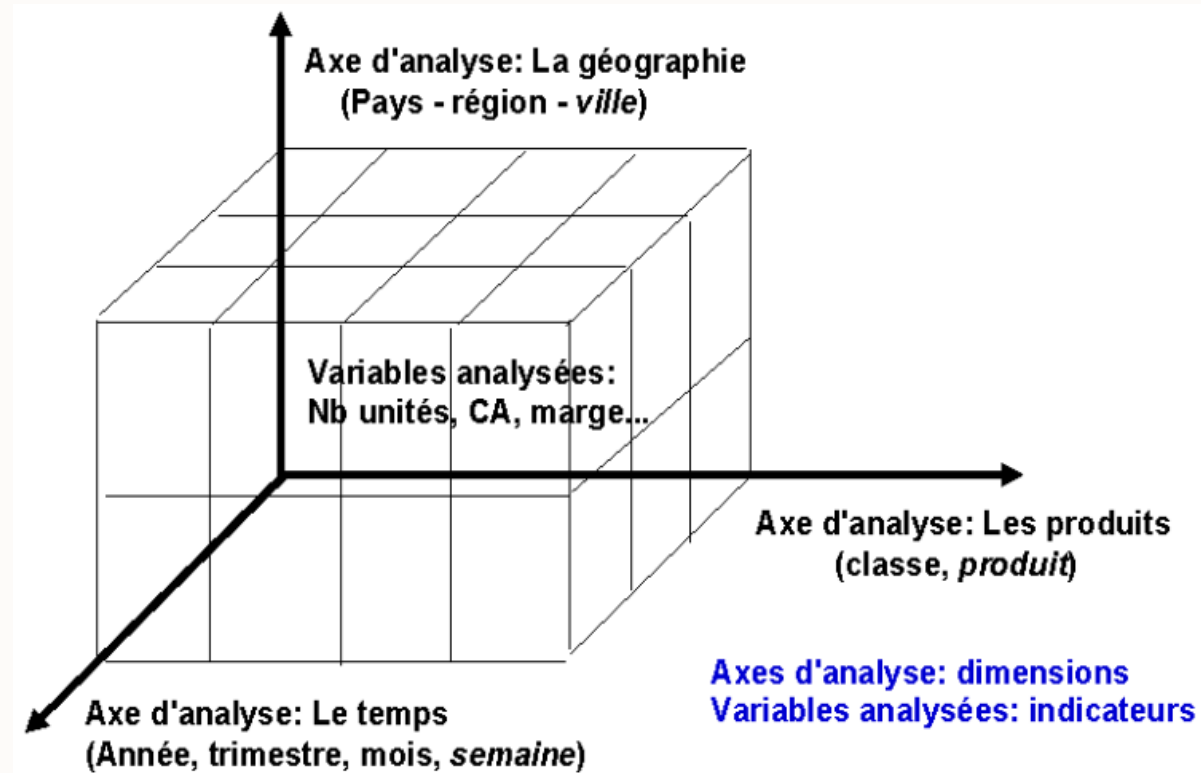
--> Approche multidimensionnelle

- Souv ent représentés par une structure à plusieurs dimensions
- Une dimension est un attribut ou un ensemble d'attributs:
 - Temps
 - Géographie
 - Produits
 - Clients
- Les cellules contiennent des données agrégées appelées Faits ou Indicateurs:
 - Nombre d'unités vendues
 - Chiffre d'Affaire
 - Coût
- Représentations:
 - Relations,
 - Cube de données,
 - hyper cube de données

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Vue multidimensionnelle



Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

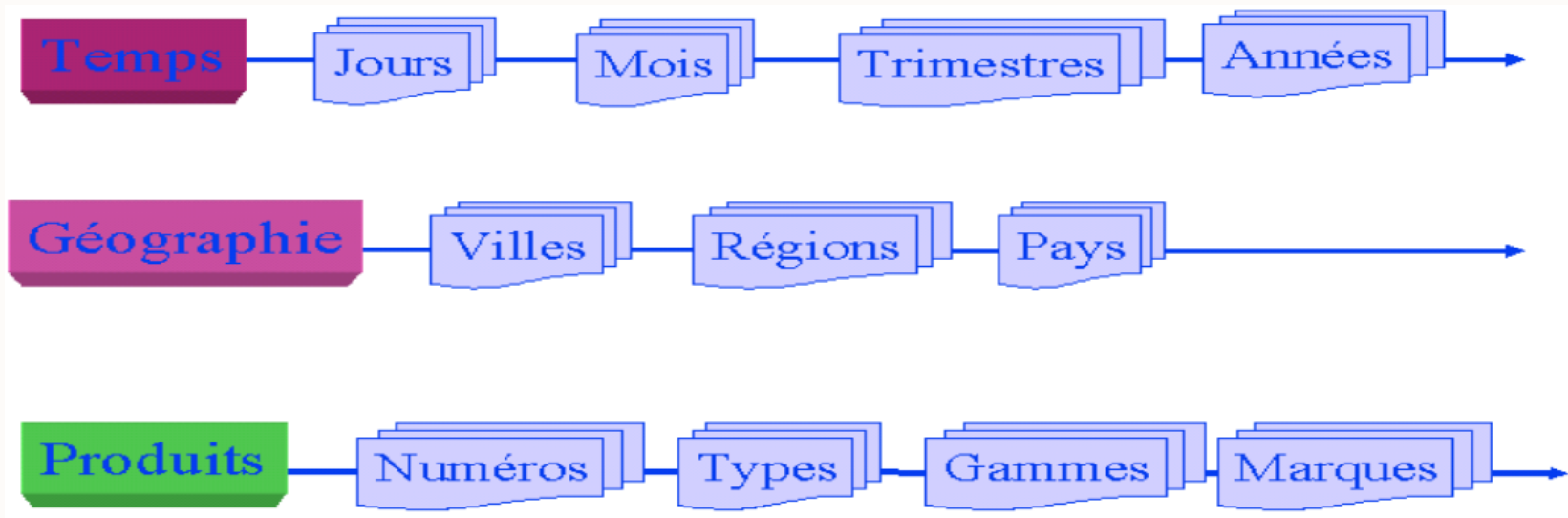
Agrégation des données

- Plusieurs niveaux d'agrégation
 - Les données peuvent être groupées à différents niveaux de granularité
 - Les regroupements sont pré-calculés,
--> Par exemple, le total des ventes pour le mois dernier calculé à partir de la somme de toutes les ventes du mois.
- Granularité : niveau de détail des données emmagasinées dans un Datawarehouse

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Granularité des dimensions



Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Les 12 règles OLAP (Edgar Frank Codd, 1993)

1. Vue multidimensionnelle:

- Comme par exemple lorsqu'on souhaite analyser les ventes selon plusieurs dimension: par produit par région ou par période.

2. Transparence du serveur OLAP à différents types de logiciels

- Elle s'appuie sur une architecture ouverte permettant à l'utilisateur d'implanter le système OLAP sans affecter les fonctionnalités du système central.

3. Accessibilité à de nombreuses sources de données

- Le système OLAP doit donner accès aux données nécessaires aux analyses demandées.
- Les outils OLAP doivent avoir leur propre schéma logique de stockage des données physiques

4. Performance du système de Reporting

- L'augmentation du nombre de dimensions ou du volume de la base de données ne doit pas entraîner de dégradation visible par l'utilisateur.

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Les 12 règles OLAP (Edgar Frank Codd, 1993)

5. Architecture Client/Serveur

- La plus part des données pour OLAP sont stockées sur des gros systèmes. Il est nécessaire que les outils OLAP soient capables de travailler dans un environnement Client/Serveur.

6. Dimensions Génériques

- Toutes les dimensions doivent être équivalentes en structure et en calcul.
- Toute fonction qui s'applique à une dimension doit être aussi applicable à une autre dimension.

7. Gestion dynamique des matrices creuses

- Le schéma physique des outils OLAP doit s'adapter entièrement au modèle d'analyse spécifique créé pour optimiser la gestion des matrices creuses

8. Support Multi-Utilisateurs

- Les outils OLAP doivent supporter les accès concurrents,
- Garantir l'intégrité et la sécurité afin que plusieurs utilisateurs accèdent au même modèle d'analyse.

Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Les 12 règles OLAP (Edgar Frank Codd, 1993)

9. Opération sur les dimensions

- Les opérations doivent pouvoir s'effectuer sur toutes les dimensions.

10. Manipulation intuitive des données

- Toute manipulation doit être accomplie via une action directe sur les cellules du modèle sans utiliser de menus ou des chemins multiples à travers l'interface utilisateur.

11. Souplesse et facilité de constitution des rapports

- La création des rapports dans les outils OLAP doit permettre aux utilisateurs de présenter comme ils le désirent des données synthétiques ou des résultats en fonction de l'orientation du modèle.

12. Nombre illimité de niveaux d'agrégation et de dimensions

- Tout outil OLAP doit gérer au moins 15 à 20 dimensions.

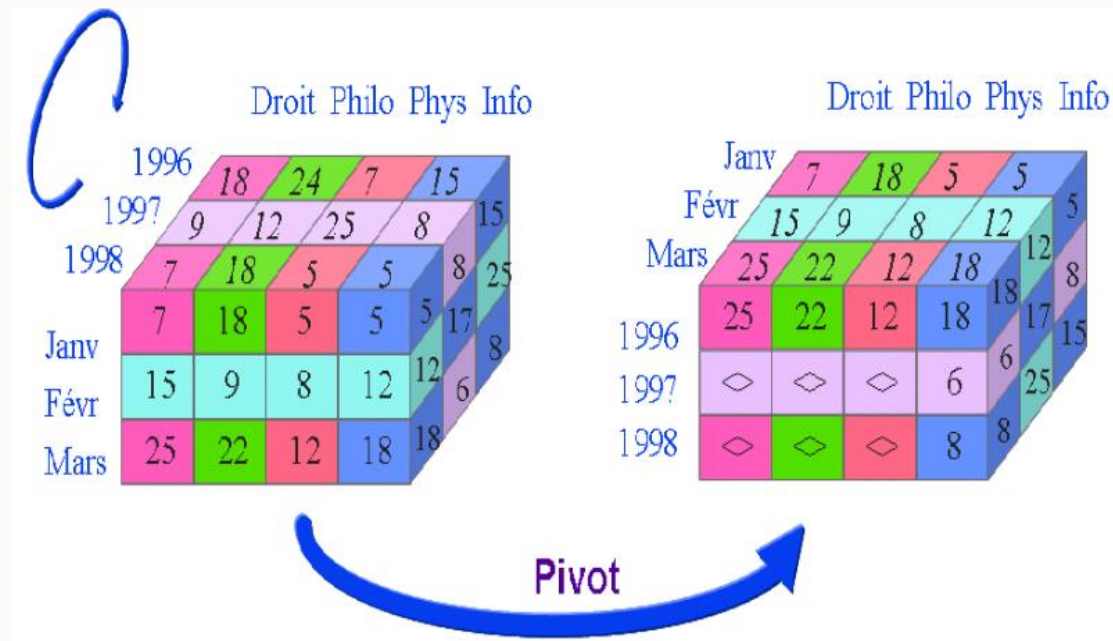
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations sur la structure des cubes

- Pivot (Rotation)



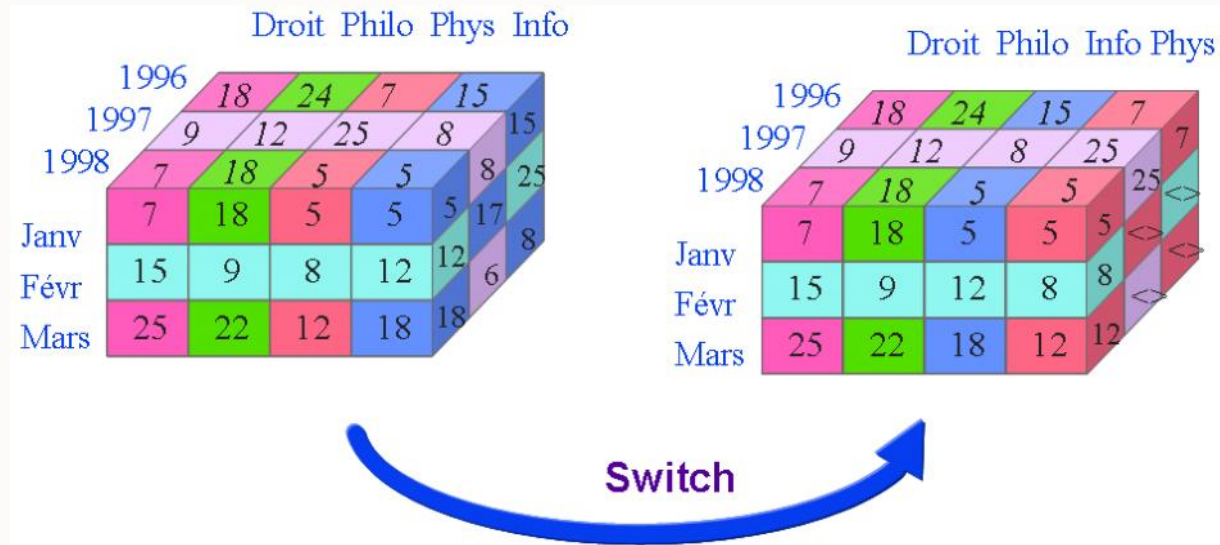
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations sur la structure des cubes

- Switch (Permutation)



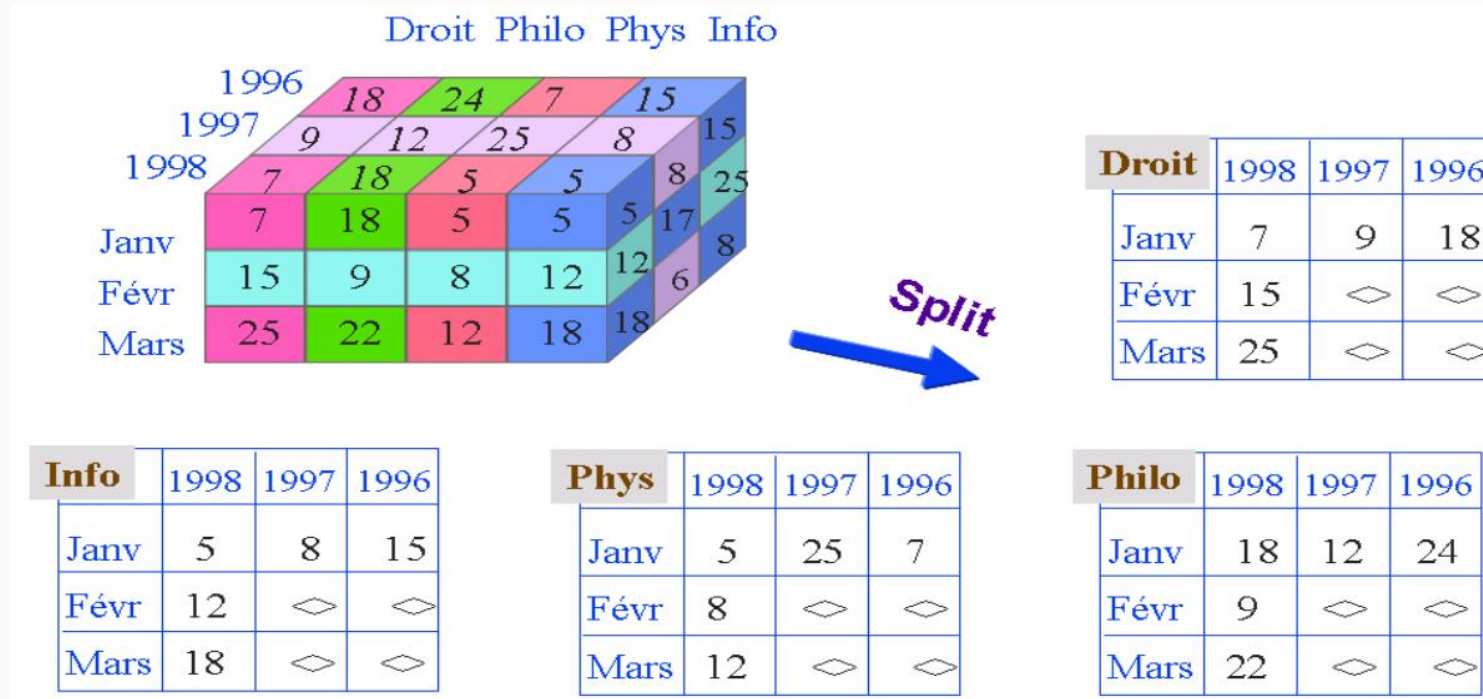
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations sur la structure des cubes

- Split (Décomposition)



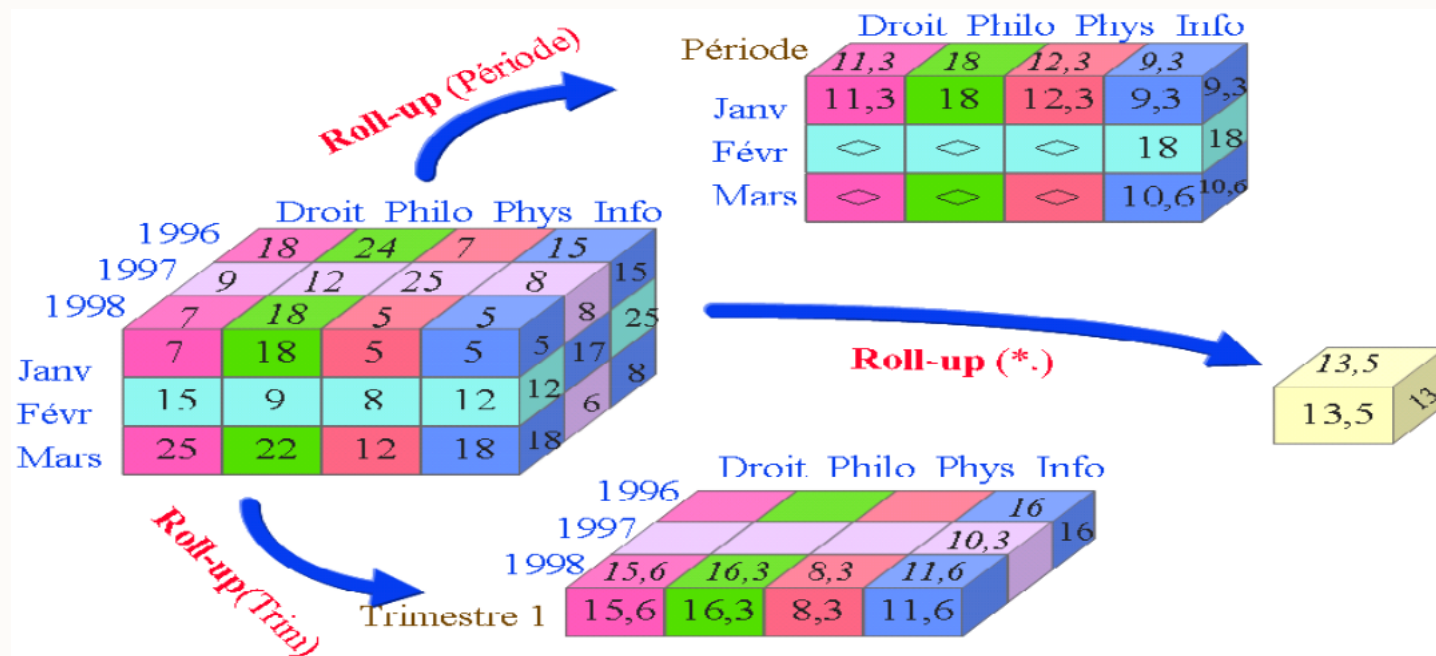
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations sur le contenu des cubes

- Roll-up (passage au grain supérieur) / Drill-down (passage au grain inférieur)



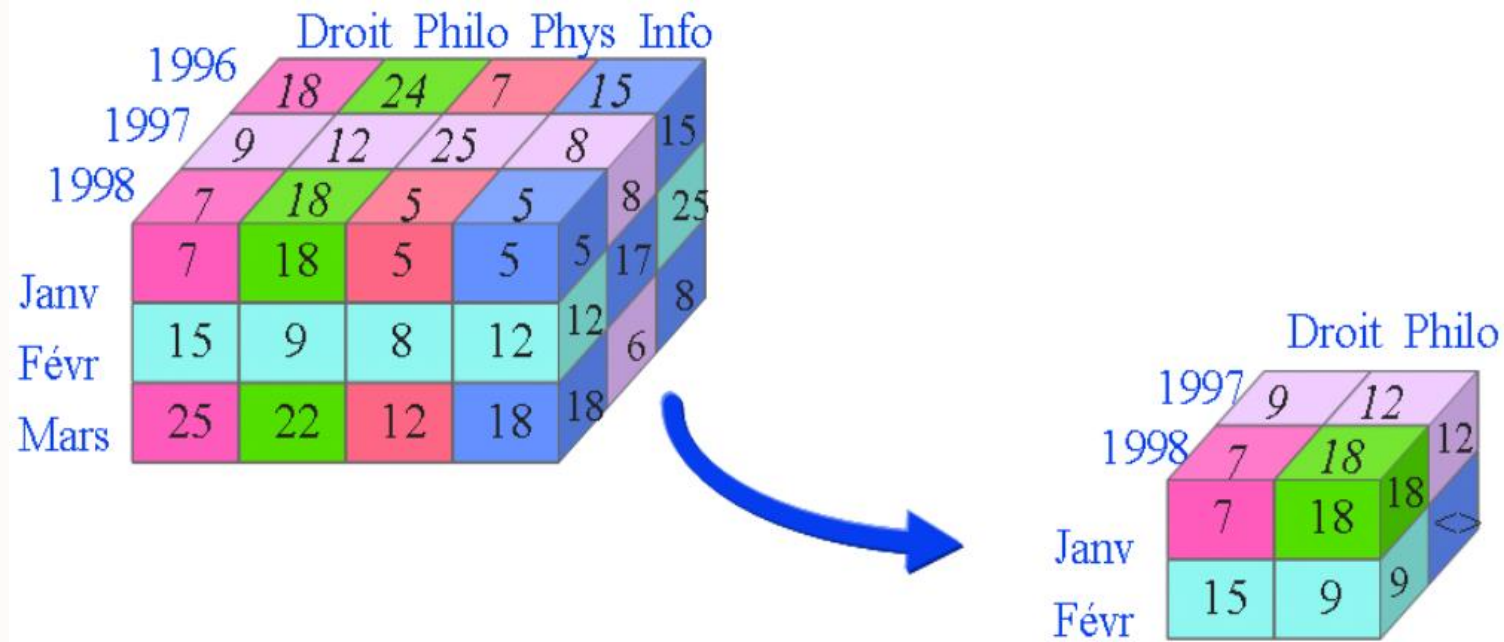
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations sur la structure des cubes

- Slice (Restriction)



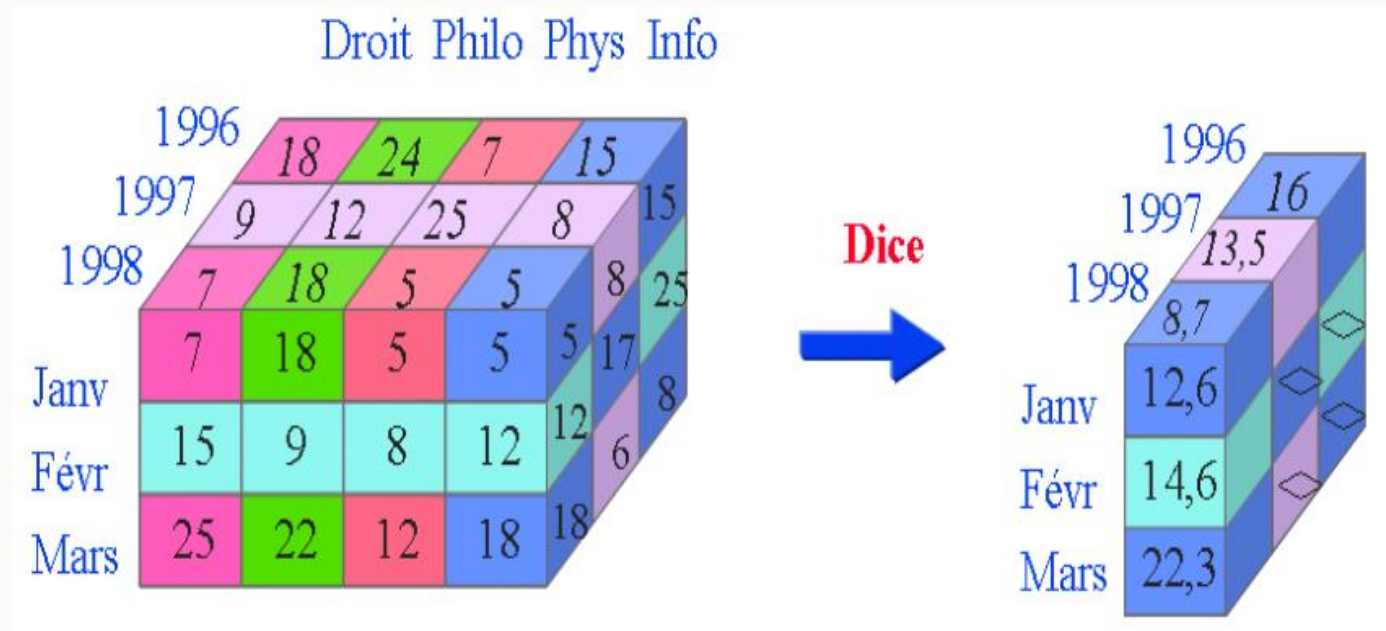
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations sur la structure des cubes

- Dice (Projection)



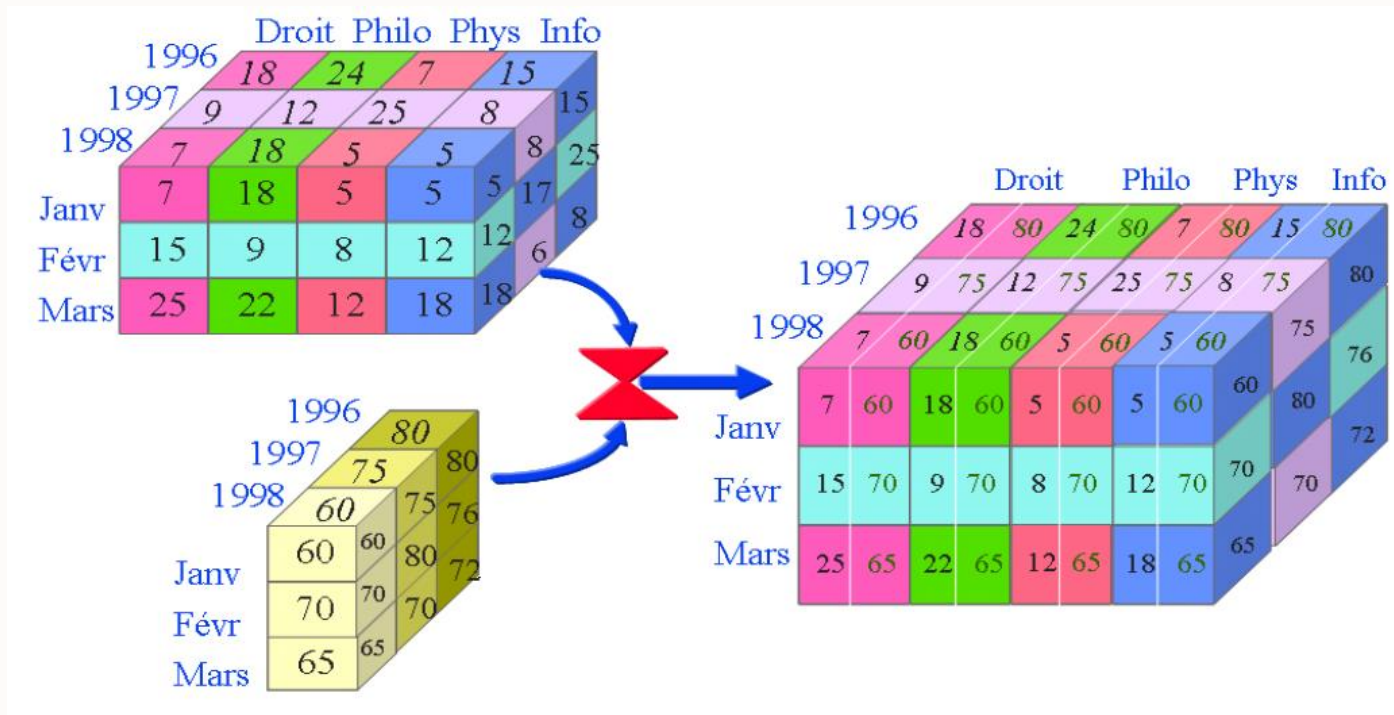
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations entre cubes

- Jointure



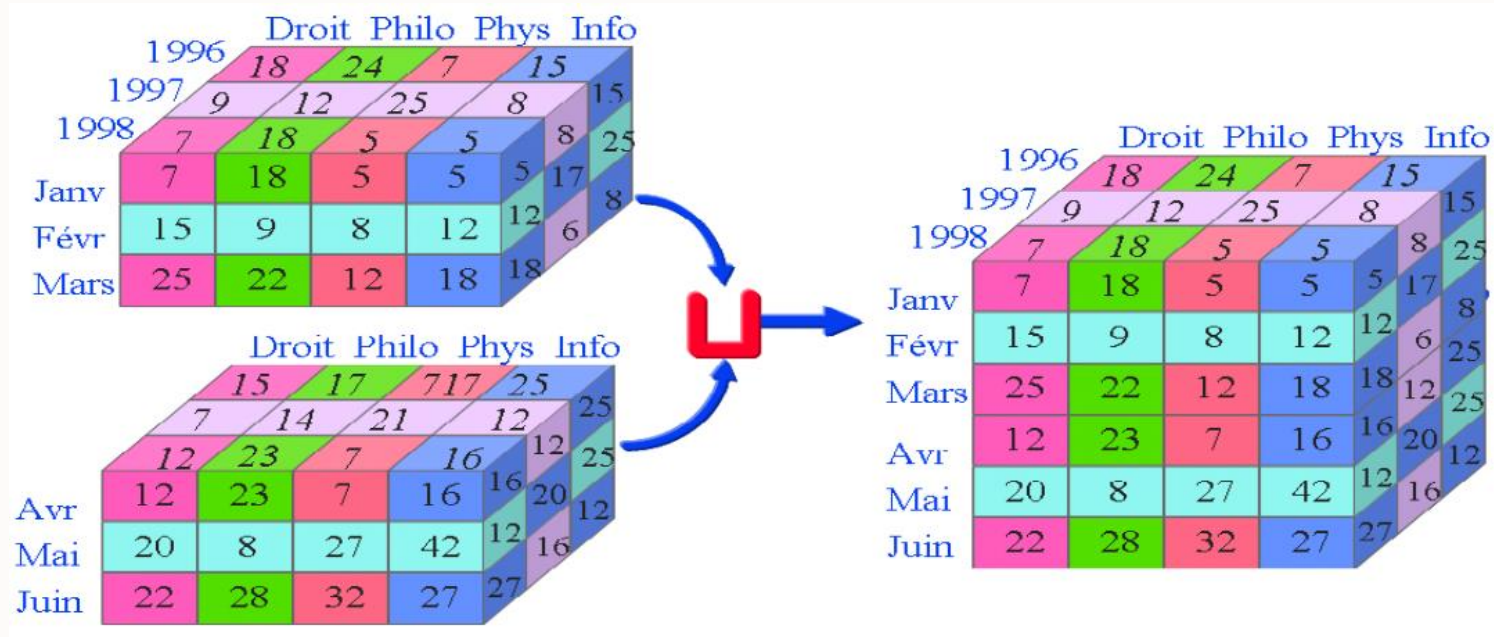
Les concepts de base de Data Warehouse

OLAP et Analyse multidimensionnelles

Opérations OLAP

-> Opérations sur la structure des cubes

- Union



Modélisation et Conception d'un DW

Les concepts de base de Data Warehouse

Modélisation et Conception d'un DW

Construction d'un Datawarehouse

Caractéristiques:

- Le Datawarehouse est différent des bases de données de production:
 - Les besoins pour lesquels on veut le construire sont différents
 - Il contient des informations historisées, organisées selon les métiers de l'entreprise pour le processus d'aide à décision
- Le Datawarehouse n'est pas un produit ou un logiciel mais un environnement, qui se bâtit et ne s'achète pas.

Phases de construction d'un DW:

- Il y'a trois parties interdépendante qui relève la construction d'un Datawarehouse:
 - L'étude préalable qui va définir les objectifs, la démarche à suivre, le retour sur investissement,...
 - L'étude du modèle de données qui représente le DW conceptuellement et logiquement
 - L'étude de l'alimentation du Datawarehouse

Les concepts de base de Data Warehouse

Modélisation et Conception d'un DW

Etude préalable

- Etude des besoins:
 - Définir les objectifs du DW
 - Déterminer le contenu du DW et son organisation, d'après:
 - Les résultats attendus par les utilisateurs,
 - Les requêtes qu'ils formuleront,
 - Les projets qui ont été définie
 - Recenser les données nécessaires à un bon fonctionnement du DW:
 - Recenser les données disponibles dans les bases de production
 - Identifier les données supplémentaires requises
 - Choisir les dimensions
 - Typiquement: le temps, le client, le produit, le magasin...
 - Choisir les mesures de fait
 - De préférences de quantités numériques additives
 - Choisir la granularité des faits
 - Niveau de détails des dimensions

Les concepts de base de Data Warehouse

Modélisation et Conception d'un DW

Etude préalable

- Coûts de déploiement:
 - Nécessite des machines puissantes, souvent une machine parallèle
 - Capacité de stockage très importante (historisation des données)
 - Evaluer la capacité de stockage
 - Equipes de maintenance et d'administration
 - Les coûts des logiciels
 - Les logiciels d'administration du DW
 - Les outils ETL (Extract-Transform-Loading)
 - Les outils d'interrogation et de visualisation
 - Les outils de Datamining

Les concepts de base de Data Warehouse

Modélisation et Conception d'un DW

Modélisation

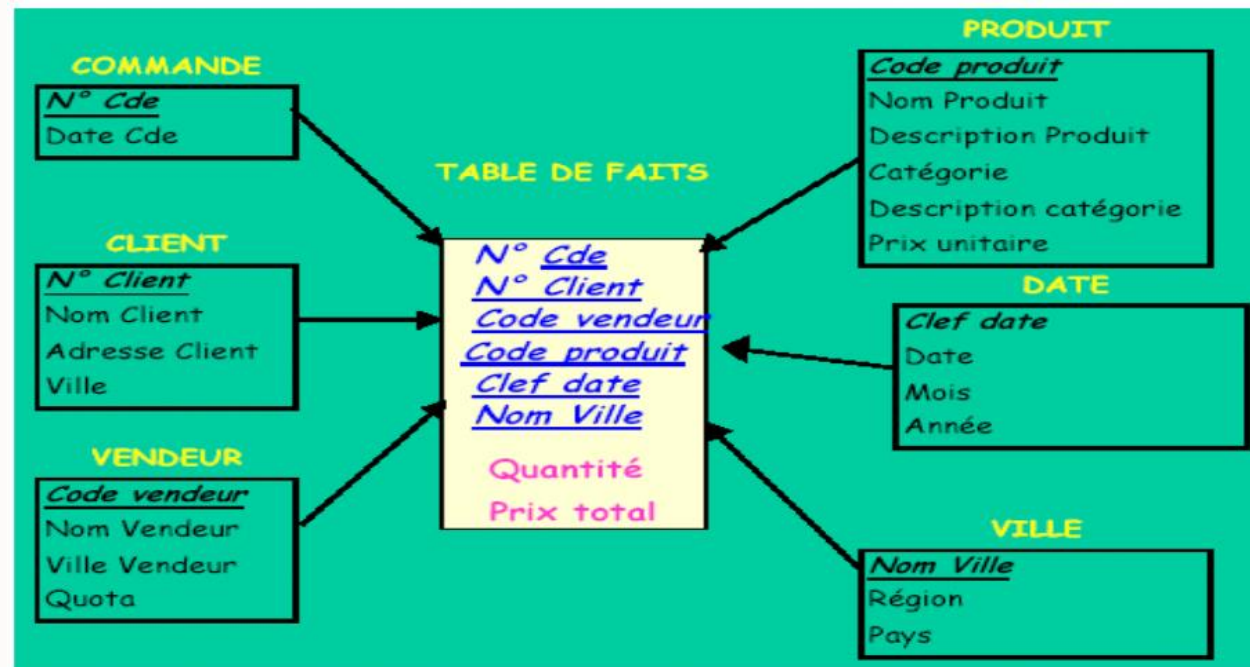
- Niveau conceptuel:
 - Un DW est basé sur une modélisation multidimensionnelle qui représente les données dans un cube
 - Un cube permet de voir les données suivant plusieurs dimensions:
 - Tables de dimensions
 - La table des faits contient les mesures et les clés des dimensions
- Niveau Logique:
 - Plusieurs schémas types sont proposés pour représenter un DW:
 - Schéma en étoile;
 - Schéma en flocon;

Les concepts de base de Data Warehouse

Modélisation et Conception d'un DW

Modélisation

- Schéma en étoile
 - Une (ou plusieurs) table(s) de faits : identifiants des tables de dimension ; une ou plusieurs mesures.
 - Plusieurs tables de dimension : descripteurs des dimensions.

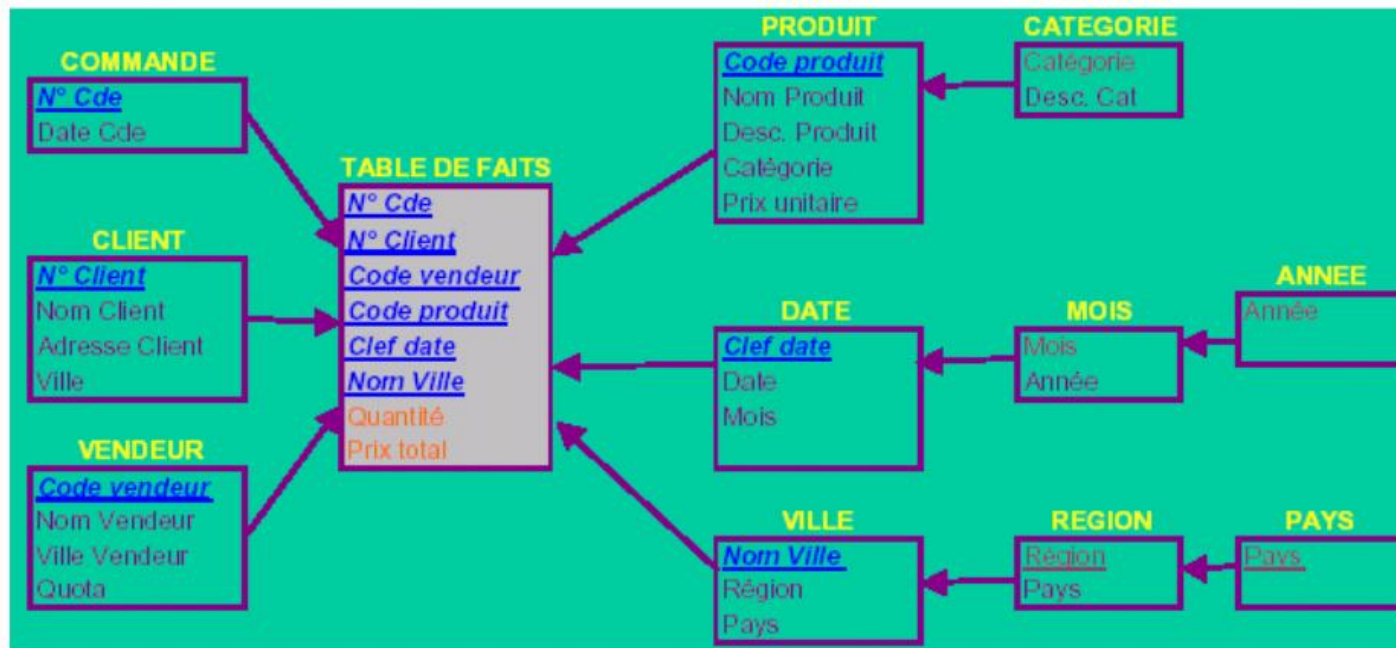


Les concepts de base de Data Warehouse

Modélisation et Conception d'un DW

Modélisation

- Schéma en flocons
 - Raffinement du schéma étoile avec des tables normalisées par dimensions.



Les concepts de base de Data Warehouse

Modélisation et Conception d'un DW

La **table de faits** et la **table de dimensions** sont utilisées pour créer des schémas. L'enregistrement d'une **table de faits** est une combinaison d'attributs de différentes **tables de dimension**. La **table des faits** aide l'utilisateur à analyser **les dimensions** de l'entreprise.

