

# DATA WAREHOUSE

Départements : Logiciel et systèmes d'information

Disciplines: Sciences des données et intelligence artificielle

Enseignant : EL GHEBOULI Ayoub

2022/2023



## **Modalités d'évaluation :**

Le module sera validé par :

- Un **examen final** (devoir écrit)
  - > Durée : 2h
  - > Pondération : **55.00%**
- **Exercices, Resumé et Présence 15 %**
- Un **mini-projet** réalisé (TP) en binôme ou en trinôme :
  - > Pondération : **30.00%**

# PLAN DU COURS

- I. Introduction à l’Informatique Décisionnelle.
- II. Les concepts de base de data warehousing.
- III. La différence entre un data warehouse et une base de données transactionnelle.
- IV. Le système de data warehouse et ses composants.
- V. Les étapes du processus de transfert des données.
- VI. Le traitement analytique en ligne (OLAP) et les outils OLAP.
- VII. Les applications de data warehousing.

*Business Intelligence*

*base de données*

*Big data*

*ETL*

*Informatique Décisionnelle*

*Data Virtualization*

**Data Warehouse**

*SQL*

*OLTP et OLAP*

*Données transactionnelles*

*Data marts*

# **Chapitre 1 : Introduction à l'Informatique Décisionnelle**

## Aide à la Décision : Mise en Situation

Pour prendre la bonne décision, il faut savoir:

- Comment a-t-il baissé ?
- Dans quelle gamme de produits ?
- Dans quels pays, quelles régions ?
- Dans le portefeuille de clientèle de quels commerciaux ?
- Dans quel segment de distribution ?
- N'avait-on pas une baisse semblable en octobre chaque année ?

## Business Intelligence

La Business Intelligence (BI) est un processus technologique d'analyse des données et de présentation d'informations pour aider les dirigeants, managers et autres utilisateurs finaux de l'entreprise à prendre des décisions business éclairées. **C'est un terme générique qui englobe une grande variété d'outils, d'applications et de méthodologies qui permettent l'accès et l'analyse de l'information afin d'améliorer et d'optimiser les décisions et les performances.** Ces données sont préparées pour l'analyse afin de créer des rapports, tableaux de bord et autres outils de Data visualisation (représentation graphique des données) pour rendre les résultats analytiques disponibles aux décideurs et aux opérations.

## Business Intelligence

Aujourd'hui, les entreprises s'appuient sur les logiciels de Business Intelligence pour **identifier et extraire des informations précieuses des grands volumes de données qu'elles stockent**. Ces outils permettent d'en tirer des informations tels que des veilles concurrentielles et les tendances du marché, ainsi que des informations internes tel que trouver les raisons des opportunités perdues.

## Business Intelligence

Les applications de Business Intelligence aident les entreprises à **regrouper les différentes sources** (telles que les systèmes de gestion de la relation client (CRM), les informations sur la chaîne logistique, les tableaux de bord des performances commerciales, les analyses marketing, les données d'appel des call center) **disparates en une seule vue unifiée fournissant des rapports, des tableaux de bord et des analyses en temps réel**. Elles peuvent apporter de nombreux bénéfices à une entreprise. Ils permettent d'accélérer et d'améliorer la prise de décision, d'optimiser des processus d'affaires internes, d'augmenter l'efficacité opérationnelle, la génération de nouveaux revenus et d'obtenir un avantage concurrentiel face à la concurrence.

CRM (gestion de la relation client) est un système logiciel complet qui gère les relations avec la clientèle

Chaîne logistique est le processus qui est généré lorsqu'un client passe une commande jusqu'à ce que le produit ou le service soit livré et payé.

**Pourquoi construire  
un système décisionnel ?**

# Introduction à l'Informatique Décisionnelle

## Pourquoi construire un système décisionnel ?

1. Servir une information considérée comme stratégique
2. Quelques constats
3. Les besoins justifiant un système décisionnel
4. Les principaux défis des projets décisionnels

# Introduction à l'Informatique Décisionnelle

## Pourquoi construire un système décisionnel ?

### 1. Servir une information considérée comme stratégique

- Un des actifs les plus importants des sociétés, c'est leur **capital d'informations** qu'elles collectent au jour le jour.
- Généralement, la plupart de ces informations sont **inaccessibles**, ou réparties dans une multitude de systèmes.
- Le système d'Information Décisionnel résulte d'un processus qui consiste à extraire les données à partir des systèmes opérationnels et d'autres sources externes à l'entreprise, de les transformer en information de pilotage et de les rendre accessibles aux utilisateurs.
- La base Décisionnelle est aujourd'hui reconnue comme **un actif stratégique** par beaucoup d'entreprises.

# Introduction à l'Informatique Décisionnelle

## Pourquoi construire un système décisionnel ?

### 2. Quelques constats

- L'information existante est souvent très riche mais il est difficile d'avoir une vision globale homogène et cohérente des informations manipulées par l'ensemble des départements
- Il n'est pas facile accéder directement à l'information nécessaire : il existe plusieurs sources utilisant des supports différents (papier, base de données, fichiers Excel).
- Les données de gestion peuvent avoir des significations différentes selon l'utilisation qui en est faite, exemples : la marge..., Mais le reporting de Direction Générale n'accepte qu'un seul sens à une valeur restituée.

# Introduction à l'Informatique Décisionnelle

## Pourquoi construire un système décisionnel ?

### 3. Les besoins justifiant un système décisionnel

- Meilleur accès aux données
- Amélioration de la qualité des informations
- Intégration des données provenant de systèmes différents
- Définition commune des informations
- Meilleur accès aux données historiques

# Introduction à l'Informatique Décisionnelle

## Pourquoi construire un système décisionnel ?

### 4. Les principaux défis des projets décisionnels

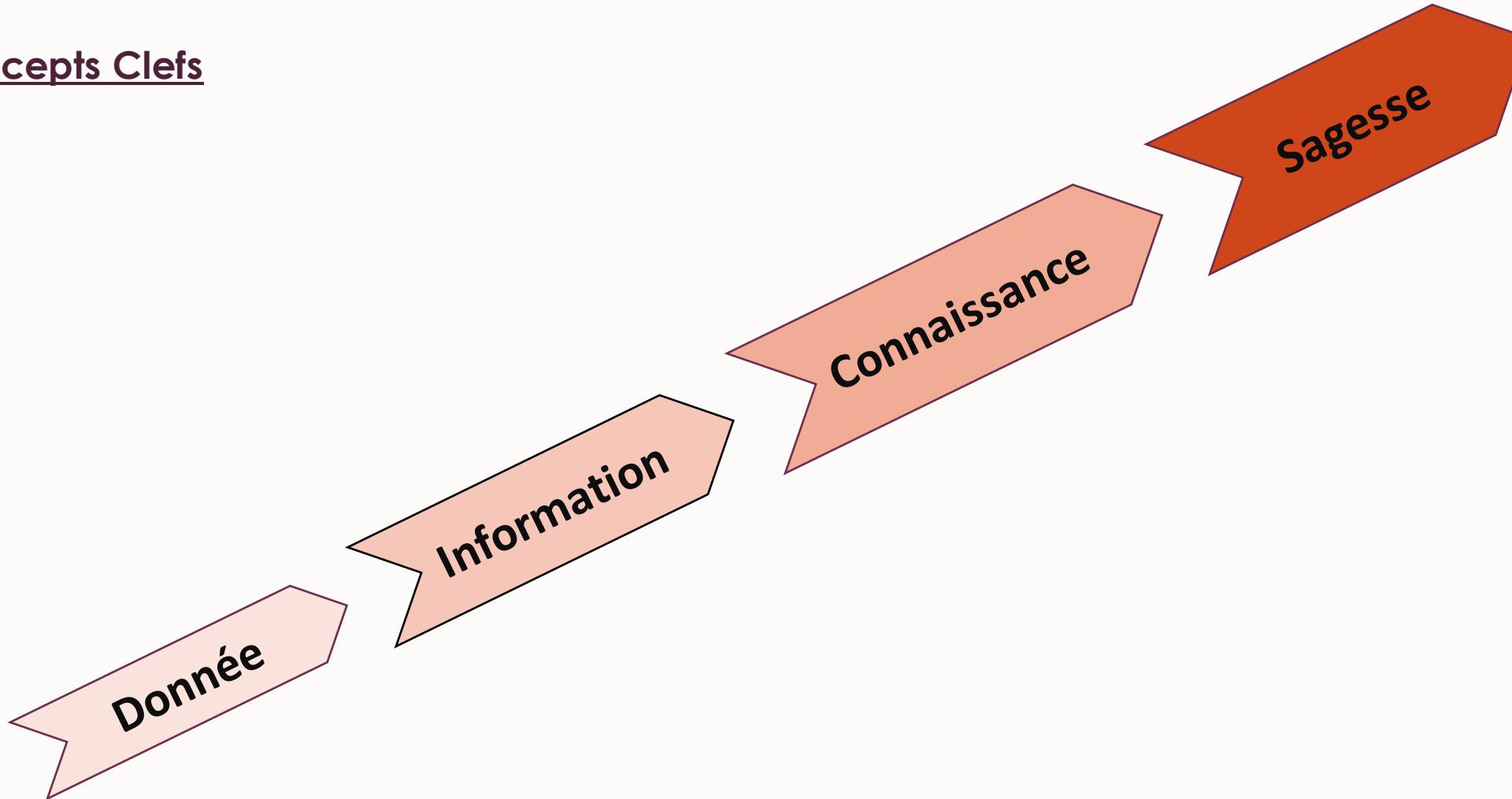
- Compréhension des besoins utilisateurs
- Intégrité des données
- Coût des alimentations en données
- Définition du périmètre du projet
- Performances du système
- Règles de gestion commune

## Métriques d'aide à la décision

# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

### Concepts Clefs



# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

### Concepts Clefs : Donnée

- **Donnée :**
  - Résultat direct d'une mesure
  - Peut-être collectée par un outil de mesure, ou être présente dans une base de données
  - Ne permet pas de prendre de décision sur une action à lancer
  
- **Exemple**
  - Le mois dernier, on a enregistré 1217 incidents au centre de services
  - 10 nouveaux prestataires ont été employés à la direction informatique

# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

### Concepts Clefs : Information

- **Information:**
  - Donnée à laquelle un sens et une interprétation ont été donnés
  - Permet au responsable de prendre une décision sur une action
- **Exemple**
  - Le mois dernier, on a enregistré une augmentation de 240% du nombre d'incidents par rapport au mois précédent
  - L'emploi des 10 prestataires est lié à une augmentation temporaire de la charge de travail

# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

### Concepts Clefs : Connaissance

- **Connaissance :**
  - Résultat d'une réflexion sur les informations analysées
  - Se base sur les expériences, les idées, valeurs, avis des personnes consultées
- **Exemple**
  - Le gestionnaire de chargement peut établir une corrélation entre l'arrivée des nouveaux prestataires et l'augmentation de nombre d'incidents en ayant connaissance de certains éléments

# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

### Concepts Clefs : Sagesse

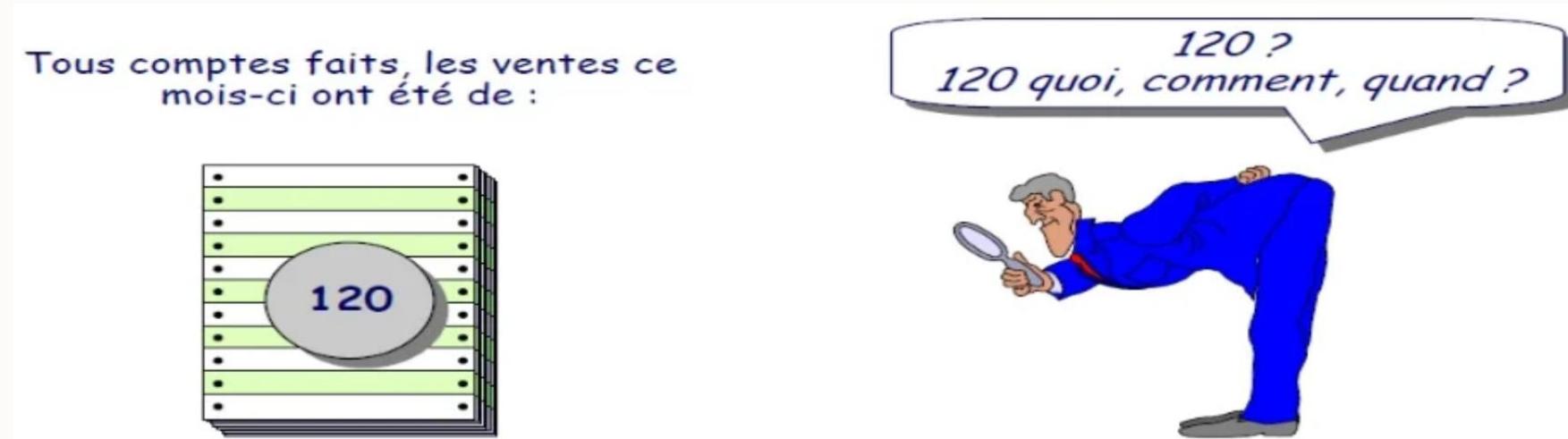
- **Sagesse :**
  - État d'esprit général de discernement final sur le contenu et de jugement de bon sens
  - Permet de lancer des actions d'adaptation, des personnes, des processus et outils
- **Exemple**
  - Le responsable senior de l'organisation prend des décisions à long terme et des décisions stratégiques pour l'organisation informatiques

# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

## Illustration d'un reporting imprécis

Reporting classique présentant une information brute, statique et peu précise



# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

### Illustration d'un reporting précis

Reporting amélioré présentant une information préparée, dynamique et plus précise



L'important  
est dans  
la variation



*Donc, tout va bien ...*



# Introduction à l'Informatique Décisionnelle

## Métriques d'aide à la décision

Une information peut en cacher une autre ... !!

Reporting décisionnel présentant une information enrichie, analytique et pertinente



*Rien ne va plus !  
Bazar à la baisse...  
Si je savais pourquoi,  
je saurais quoi faire !*



PGC : Produits de Grande Consommation

## **Chaîne Décisionnelle**

# Introduction à l'Informatique Décisionnelle

## Chaîne Décisionnelle

Les 5 grandes étapes :



# Introduction à l'Informatique Décisionnelle

## Chaîne Décisionnelle

### Planification

- Pour mettre en place une plate-forme décisionnelle d'entreprise intégrée, la première étape est donc la planification de ce projet
- Un tel projet nécessite une administration solide
- Exemple : les ressources humaines
  - Un responsable peut voir le salaire des personnes de son équipe
  - Mais ne peut pas voir celui de son chef
  - Nécessité d'une stratégie de sécurité rigoureuse

# Introduction à l'Informatique Décisionnelle

## Chaîne Décisionnelle

### ETL : Extract, Transform, Load

- Extraction des données à partir d'une ou plusieurs sources de données : fichier texte, Excel, base de données ...
- Transformation des données agrégées
- Chargement des données dans la banque de données de destination (datawarehouse)

# Introduction à l'Informatique Décisionnelle

## Chaîne Décisionnelle

### Stockage

- Plusieurs manières de stocker la donnée dans un data warehouse
- Chacune ayant ses avantages et ses inconvénients
- L'administrateur des bases de données décisionnelles pourra notamment choisir entre : DDS (Detail Data Store), les schémas en étoile, schéma en flocon ...

**DDS** : Un Data Store (littéralement « dépôt de données ») est un référentiel servant au stockage permanent d'ensembles de données.

# Introduction à l'Informatique Décisionnelle

## Chaîne Décisionnelle

### Analyse

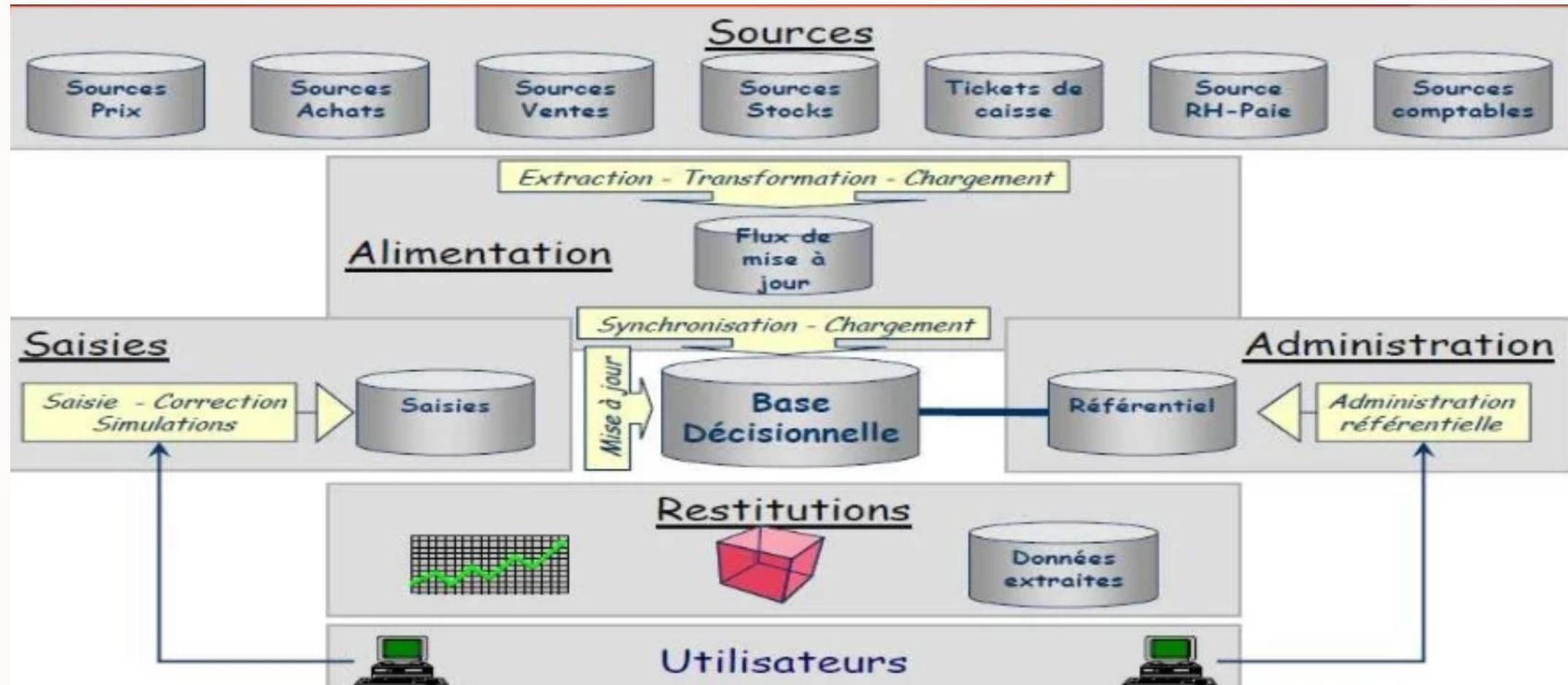
- Regroupement de l'ensemble des techniques de statistique, d'économétrie, de Data Mining, et de recherche opérationnelle
- Demande souvent des compétences statistiques avancées
- Néanmoins certaines solutions embarquent ces fonctionnalités pré-paramétrées à des cas de figures bien définies, afin d'offrir leur valeur ajoutée à des personnes fonctionnelles

Data mining est le processus d'analyse d'un grand nombre d'informations pour discerner les tendances et les modèles

# Introduction à l'Informatique Décisionnelle

## Chaîne Décisionnelle

### Architecture d'un système décisionnel



## **Chapitre 2 :**

# **Les concepts de base de Data Warehouse**

# Les concepts de base de Data Warehouse

## Définition

Le concept de data warehousing remonte à la **fin des années 1980** lorsque les chercheurs d'IBM Barry Devlin et Paul Murphy ont développé le « business data warehouse ». Essentiellement, le concept d'entreposage de données visait à fournir un modèle architectural pour le flux de données des systèmes opérationnels aux environnements d'aide à la décision.

Un Data Warehouse (entrepôt de données) est une technologie qui regroupe des données structurées provenant d'une ou de plusieurs sources afin qu'elles puissent être comparées et analysées pour une meilleure business intelligence.



# Les concepts de base de Data Warehouse

## Définition

Le Data warehouse (entrepôt de données) est une collection de données orientées sujet, intégrées, non-volatiles et historisées, organisées pour le support d'un processus d'aide à la décision (Inmon, 94).



# Les concepts de base de Data Warehouse

## Définition

### 1. **Données orientées sujet** : réorganisation des données par sujets.

- Il n'y a pas de duplication des informations communes à plusieurs sujets.
- La base de données est construite selon les thèmes qui touchent aux métiers de l'entreprise (clients, produits, risques, rentabilité, ...).
- Les données de base sont toutefois issues des Systèmes d'Information Opérationnels (SIO)

# Les concepts de base de Data Warehouse

## Définition

**1. Données intégrées :** l'intégration dans un data warehouse des systèmes sources différents.

- Les données, issues de différentes applications de production, peuvent exister sous toutes formes différentes.
- Il faut les intégrer afin de les homogénéiser et de leur donner un sens unique, compréhensible par tous les utilisateurs.
- Ils doivent posséder un codage et une description unique.

# Les concepts de base de Data Warehouse

## Définition

**1. Données intégrées :** l'intégration dans un data warehouse des systèmes sources différents.

- La phase d'intégration est longue et pose souvent des problèmes de qualification sémantique des données à intégrer (synonymie, etc...).
- Ce problème est amplifié lorsque des données externes sont à intégrer avec les données du SI O.

# Les concepts de base de Data Warehouse

## Définition

### 1. **Données non-volatiles** : Traçabilité des informations et des décisions prises

- Une information est considérée volatile quand les données sont régulièrement mises à jour comme dans les Systèmes d'Information Opérationnels.
- Dans un SIO, les requêtes portent sur les données actuelles. Il est difficile de retrouver un ancien résultat.
- Dans un DW, il est nécessaire de conserver l'historique de la donnée. Ainsi, une même requête effectuée à deux mois d'intervalle en spécifiant la date de référence de la donnée, donnera le même résultat.

# Les concepts de base de Data Warehouse

## Définition

**1. Données historisées :** les DW contiennent des données historiques, pas seulement des données actuelles.

- Dans un SIO, les transactions se font en temps réel, et les données sont mises à jour constamment. L'historique des valeurs de ces données n'est généralement pas conservé car il est inutile.
- Dans un DW, la donnée n'est jamais mise à jour.
- Les données du DW s'ajoutent aux données déjà stockées => ajout de couches de données successives, à la manière des strates géologiques

# Les concepts de base de Data Warehouse

## Définition

**1. Données historisées :** les DW contiennent des données historiques, pas seulement des données actuelles.

- Le DW stocke donc l'historique des valeurs que la donnée aura prises au cours du temps.
- Un référentiel de temps est alors associé à la donnée afin d'être capable d'identifier une valeur particulière dans le temps.
- Les utilisateurs possèdent un accès aux données courantes ainsi qu'à des données historisées.

# **Les concepts de base de Data Warehouse**

## **Les raisons de créer un data warehouse**

# Les concepts de base de Data Warehouse

## Les raisons de créer un data warehouse

Deux raisons principales :

- Le premier est de soutenir **la prise de décisions en fonction des données** plutôt que de devoir se fier uniquement à l'expérience et à l'intuition.
- La seconde est l'idée d'un **guichet unique**. En d'autres termes, les données dont nous avons besoin se trouvent toutes au même endroit plutôt que d'être dispersées entre les applications transactionnelles et opérationnelles d'où nous obtenons ces données lorsqu'il s'agit de prendre des décisions fondées sur les données.



# Les concepts de base de Data Warehouse

## Les raisons de créer un data warehouse

Avant l'idée d'un guichet unique, toute tentative de prise de décision basée sur les données nécessite de rechercher des données, soit dans les applications d'origine elles-mêmes, soit dans ce que nous appelons des fichiers d'extraction qui extrait les données d'une ou plusieurs applications. Mais pour la plupart, la prise de décision basée sur les données dans le passé était une grande problématique en raison des données qui sont dispersées.

Si nous considérons toutes ces vues de nos entreprises représentées par nos données ensemble comme une seule, nous avons en fait une discipline connue sous le nom de business intelligence, et par le data warehouse (Les deux sont entrés en scène à peu près au même moment vers 1990).



# Les concepts de base de Data Warehouse

## Les raisons de créer un data warehouse



Avec le Data Warehouse, nous intégrons toutes les données en un seul endroit et fournissons un guichet unique pour les données. Donc, nous pouvons nous concentrer sur l'analyse des données plutôt que sur la collecte et l'intégration répétées des données. Et nous avons Business Intelligence et le data warehouse comme une sorte de disciplines qui offrent une valeur énorme aux entreprises.

### **Différence entre Data Warehouse et une base de données**

# Les concepts de base de Data Warehouse

## Différence entre Data Warehouse et une base de données

- Une **base de données** est une collection de données organisées. Par exemple, une base de données peut regrouper toutes les informations sur les clients ou sur les transactions.
- Un **data warehouse** est un système de reporting et d'analyse de données. Il fournit des performances élevées pour les requêtes analytiques.

# Les concepts de base de Data Warehouse

## Différence entre Data Warehouse et une base de données

### 4 différences phare entre database et data warehouse

#### **1. Stockage vs analyse :**

Une base de données est conçue principalement pour enregistrer des données. Un entrepôt de données, d'autre part, est conçu principalement pour analyser les données. Une base de données est normalement optimisée pour effectuer des opérations de lecture-écriture de transactions ponctuelles. Il n'est pas conçu pour effectuer de grandes requêtes analytiques de la même manière qu'un entrepôt de données.

# Les concepts de base de Data Warehouse

## Différence entre Data Warehouse et une base de données

### 4 différences phare entre database et data warehouse

#### 2. Collecte vs catégorie :

Alors qu'une base de données est une collecte de données axée sur les applications, un entrepôt de données est plutôt axé sur une catégorie de données. Une base de données est normalement limitée à une seule application, ce qui signifie qu'une base de données équivaut habituellement à une application ; elle cible habituellement un processus à la fois. Un entrepôt de données, d'autre part, stocke les données d'un nombre quelconque d'applications. Un entrepôt de données comprend un nombre infini d'applications et cible autant de processus que nécessaire.

# Les concepts de base de Data Warehouse

## Différence entre Data Warehouse et une base de données

### 4 différences phare entre database et data warehouse

#### **3. Fournisseur de données vs source d'analyse de données :**

L'une des différences pratiques entre une base de données et un entrepôt de données est que le premier est un fournisseur de données en temps réel, tandis que le second est davantage une source d'analyse des données à mesure qu'elles sont enregistrées. Toutes les données peuvent être extraites d'un entrepôt de données pour être analysées chaque fois que cela est nécessaire.

# Les concepts de base de Data Warehouse

## Différence entre Data Warehouse et une base de données

### 4 différences phare entre database et data warehouse

#### **1. Rapidité de stockage vs temps d'analyse :**

Une base de données comporte généralement des tables complexes parce que les données sont organisées de telle sorte qu'aucun élément n'est dupliqué. Cette structure organisationnelle permet un traitement et un stockage très efficaces des données ; une réponse est très rapide. Un entrepôt de données, par contre, n'est pas conçu pour des transactions rapides, mais plutôt pour améliorer les requêtes analytiques, ce qui est obtenu en utilisant moins de tables et une structure plus simple.

# Les concepts de base de Data Warehouse

## Un système de gestion de base de données (SGBD)

Un système de gestion de base de données (**SGBD**) est le **logiciel qui facilite la gestion des bases de données**. Certains SGBD populaires incluent MySQL, MSSQL, Oracle et PostgreSQL. L'utilisateur peut écrire des requêtes en langage SQL (Structured Query Language) pour manipuler des données dans la base de données. Le processus d'exécution des requêtes dans la base de données s'appelle OLTP ou traitement transactionnel en ligne. Par conséquent, une base de données utilise OLTP. Globalement, une base de données aide à organiser un ensemble de données.



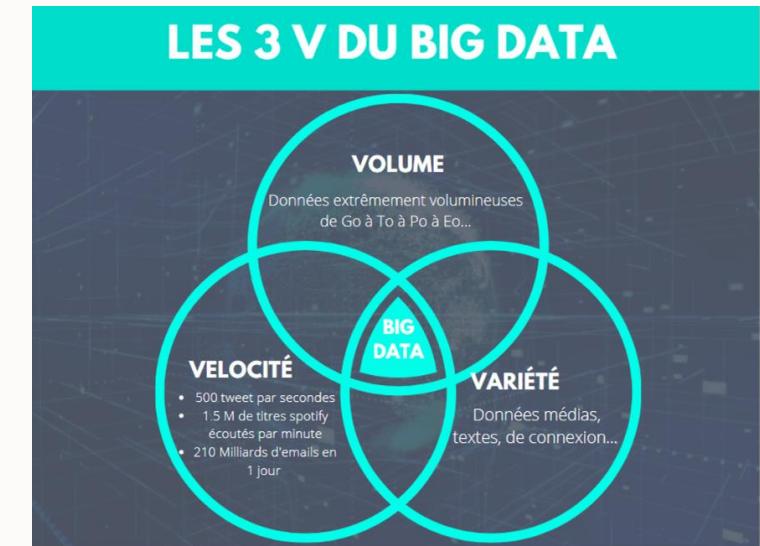
## Big Data

# Les concepts de base de Data Warehouse

## Big data

Les Big data (données massives ou méga données) sont des ressources d'information très volumineuses, à vitesse très élevée et/ou de très grande variété qui nécessitent de nouvelles formes de traitement pour permettre une meilleure prise de décision, la découverte d'informations et l'optimisation des processus.

- **Volume:** désigne une très grande masse de données collectées, allant du téraoctet (1 To =  $10^{12}$  octets) au zettaoctet (1 Zo =  $10^{21}$  octets).
- **Vélocité:** désigne une très haute fréquence à laquelle les données sont générées, traitées et mises en réseau.
- **Variété:** désigne une très grande variété de données qui sont soit structurées ou non structurées (textuelles, visuelles ou sonores, scientifiques ou provenant de la vie courante...).



# Les concepts de base de Data Warehouse

## **Différences entre Data Store, Database, Data Warehouse, Datamart et Data Lake**

Une base de donnée (Database) est un type particulier de Data Store. Et un entrepot de données (Data Warehouse) est un type particulier de base de données. Un Datamart est un sous-ensemble d'un entrepôt de données mis en place pour répondre aux besoins précis d'un groupe particulier d'utilisateurs ; par exemple, les ressources humaines, et leur fournir un accès aux informations dont ils ont besoin.

Un Data Lake décrit pour sa part tout réservoir de données de grande envergure dans lequel aucune exigence de schéma et de données n'est définie avant interrogation des données (contrairement aux Data Warehouses).

Un Datamart, un Data Lake et un Data Warehouse sont donc des formes très particulières de Data Store.

## Qu'est-ce que la Data Virtualization ?

# Les concepts de base de Data Warehouse

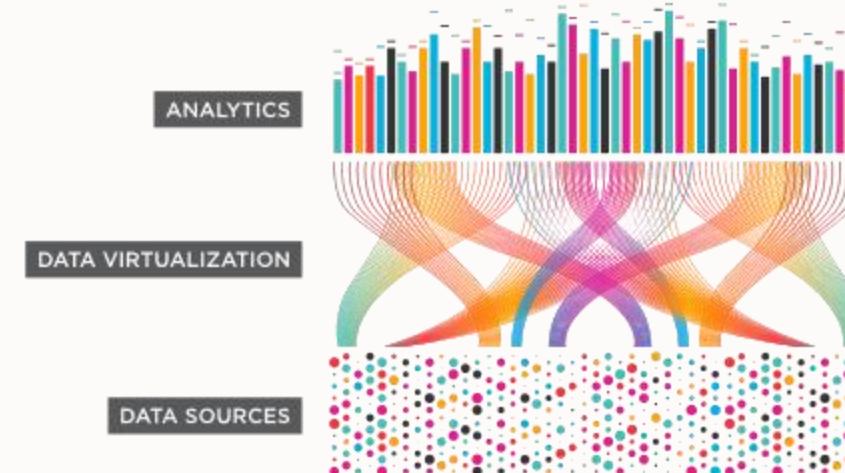
## Data Virtualization

- Le logiciel de Data Virtualization agit comme un pont entre des sources de données multiples et diverses, rassemblant les données essentielles à la prise de décision en un seul endroit virtuel, pour alimenter les analyses.
- La Data Virtualization fournit une couche de données moderne qui permet aux utilisateurs d'accéder à des ensembles de données, de les combiner, de les transformer et de les livrer à une vitesse et une rentabilité révolutionnaire.
- La technologie de Data Virtualization permet aux utilisateurs d'accéder rapidement aux données hébergées dans l'ensemble de l'entreprise, y compris dans les bases de données traditionnelles, les sources de big data et les systèmes cloud et IoT, pour une fraction du temps et du coût de l'entreposage physique et de l'extraction/transformation/chargement (ETL)

# Les concepts de base de Data Warehouse

## Data Virtualization

- Grace à la Data Virtualization, les utilisateurs peuvent appliquer toute une gamme d'analyses, y compris des analyses visualisées, prédictives et en continu, sur des mises à jour de données fraîches et actualisées à la minute près.
- Grace à une gouvernance et une sécurité intégrées, les utilisateurs de la Data Virtualization sont assurés de la cohérence, de l'extrême qualité et de la protection de leurs données.
- En outre, la Data Virtualization permet d'obtenir des données plus conviviales pour l'entreprise, en transformant les structures et la syntaxe informatiques en services de données faciles à comprendre, élaborés par l'informatique et faciles à trouver et à utiliser via un répertoire professionnel en libre-service.



# Les concepts de base de Data Warehouse

## Data Virtualization / Cloud Computing

- Le Cloud Computing (l'informatique en nuage) est une révolution technologique de cette décennie avec le Big Data. Le Big Data propose des solutions de traitement des données massives alors que le Cloud Computing offre des services de dématérialisation des ressources informatiques.



# Les concepts de base de Data Warehouse

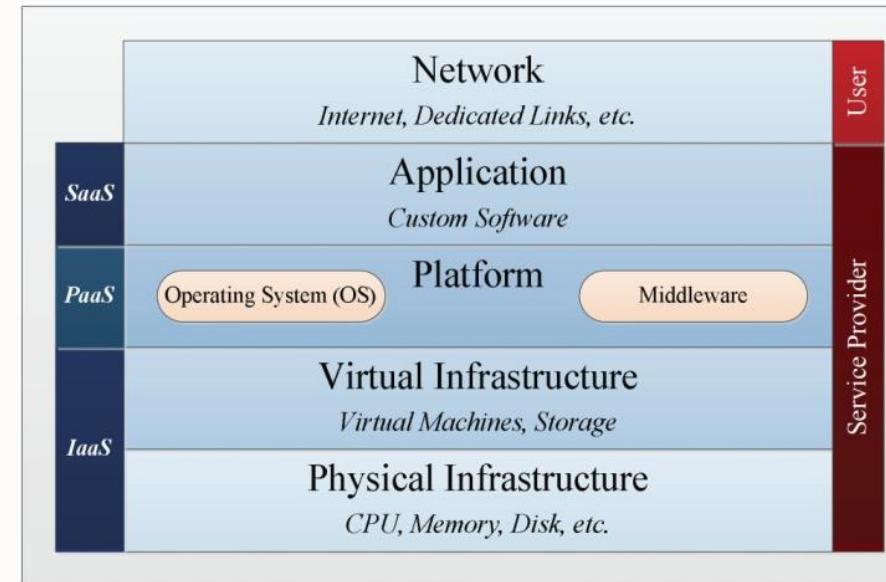
## Data Virtualization / Cloud Computing

- La définition la plus acceptable du cloud computing a été introduite par l'organisme NIST (National Institute of Standards and Technology): "**Le cloud Computing est un modèle fournissant, à la demande et au travers d'un réseau, un ensemble partagé de ressources informatiques incluant des serveurs, des espaces de stockage, des applications, des traitements et des plates-formes de déploiement qui peuvent être rapidement mises en service avec un effort minimum de gestion et d'interaction avec le fournisseur de ce service**".
- Une autre définition: "**Le Cloud Computing est un modèle dans lequel les ressources telles que la puissance de traitement, le stockage, la connectivité et le partage, etc. sont proposées sous forme de services par un mécanisme d'accès à distance. Il présente plusieurs caractéristiques souhaitables telles que la distribution et l'élasticité rapide, la sécurité, les self-services à la demande, l'accès omniprésent au réseau et la mise en commun des ressources**".

# Les concepts de base de Data Warehouse

## Data Virtualization / Cloud Computing

- L'**architecture des environnements de Cloud Computing** est composée de cinq grandes couches : Infrastructure physique, infrastructure virtuelle, plateforme, application et réseau. En fonction de ces couches, trois modèles de services de cloud computing ont été définis pour être fournis aux utilisateurs : Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) et Software-as-a-Service (SaaS).



## Environnement simple d'un Data Warehouse

# Les concepts de base de Data Warehouse

## Environnement simple d'un Data Warehouse

- Pour comprendre comment les différentes pièces d'un data warehouse s'emboîtent, nous allons jeter un coup d'œil à un environnement simple end to end d'un data warehouse.
- Un **data warehouse** est construit en extrayant des **données** d'autres applications et systèmes. Nous identifions nos sources de données ainsi que notre data warehouse. Entre les deux, nous avons un aspect essentiel appelé **ETL**. Ce dernier signifie extraction, transformation et chargement.



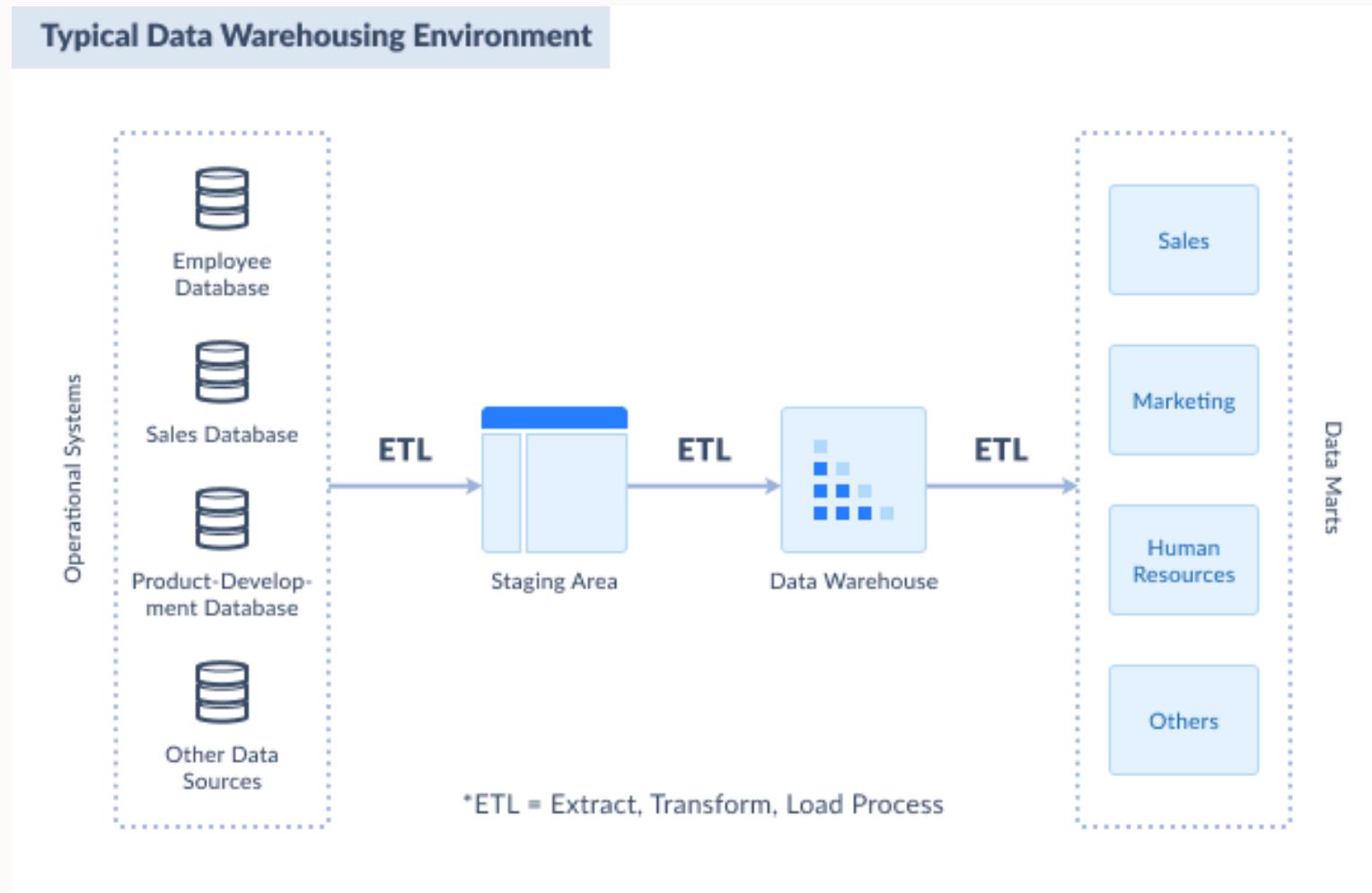
# Les concepts de base de Data Warehouse

## Environnement simple d'un Data Warehouse

- Parfois, nous ne nous arrêtons pas à extraire et à copier des données de nos sources de données dans un data warehouse via ETL. Parfois, nous continuerons ensuite à regrouper et à copier à nouveau les données, en les envoyant en aval dans des environnements plus petits, généralement appelés Data marts .
- Un data mart est un sous-ensemble d'un entrepôt de données généralement utilisé pour accéder aux informations destinées aux clients. Il s'agit d'une structure spécifique aux paramètres d'entreposage de données. Ainsi, il est généralement axé sur un secteur d'activité ou une équipe et tire des informations d'une seule source particulière.
- Contrairement à la mise en œuvre d'un entrepôt de données d'entreprise qui peut s'étendre sur plusieurs mois, voire plusieurs années, un magasin de données est généralement mis en œuvre en quelques mois, offrant une assistance rapide.



# Les concepts de base de Data Warehouse



**FIN DE SEANCE**

# 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   |

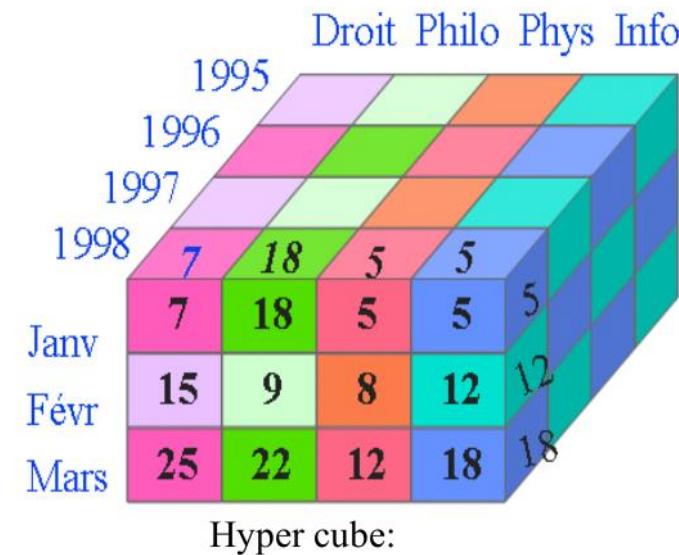
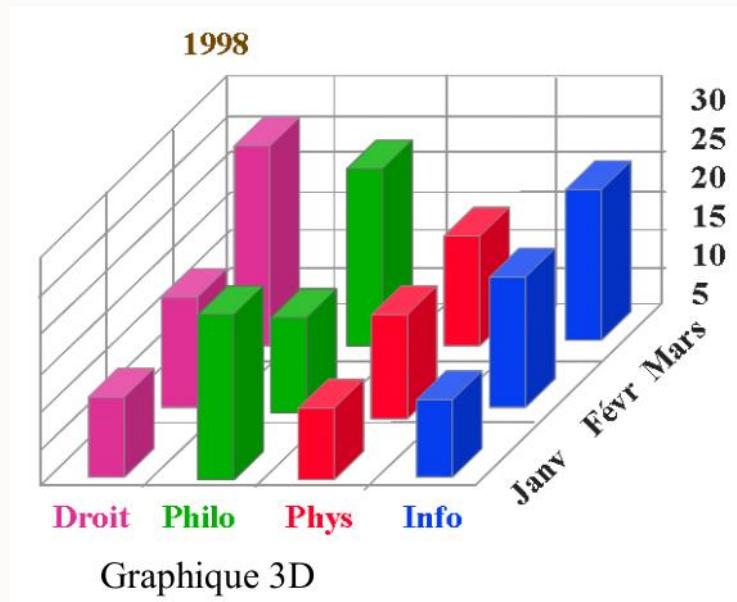
Tableau croisé

# 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

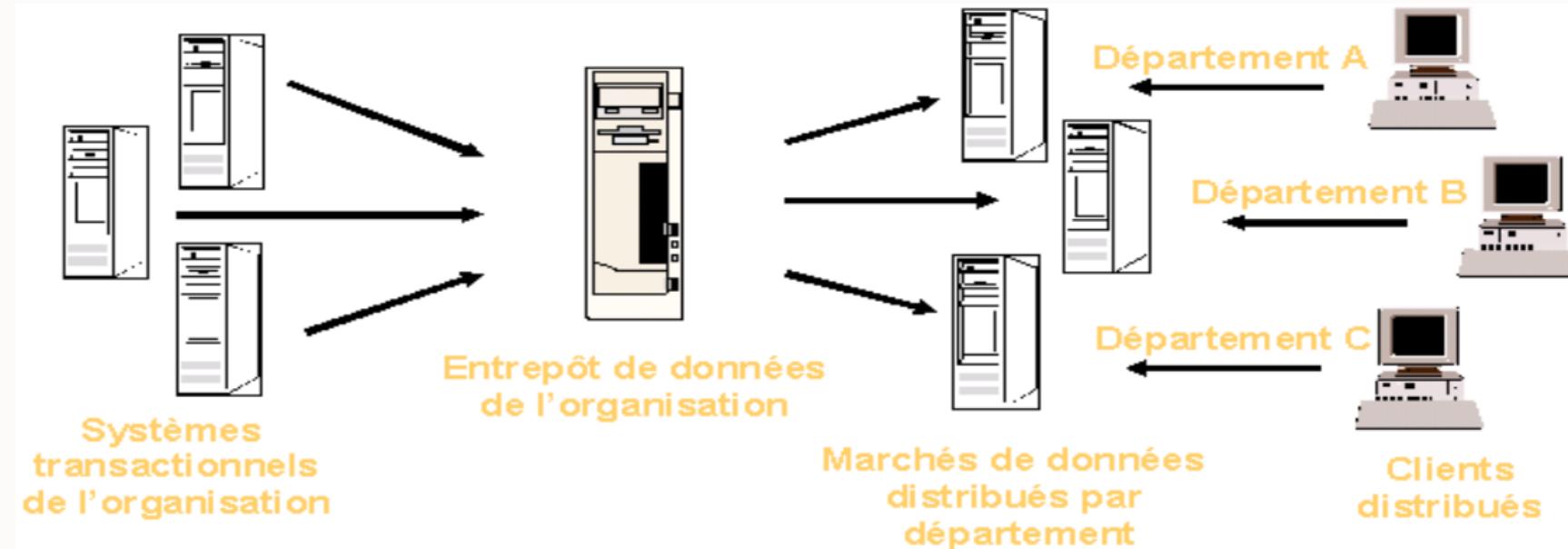


-> Il s'agit de la version centralisée et intégrée d'un entrepôt regroupant l'ensemble des données de l'entreprise. Les différentes bases de données sources sont intégrées et sont distribuées à partir de la même plate-forme physique

# 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'usager dans son analyse en lui facilitant l'exploration de ses données et en lui donnant la possibilité de le faire rapidement.
  - L'usager n'a pas à maîtriser des langages d'interrogation et des interfaces complexes
  - L'usager 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

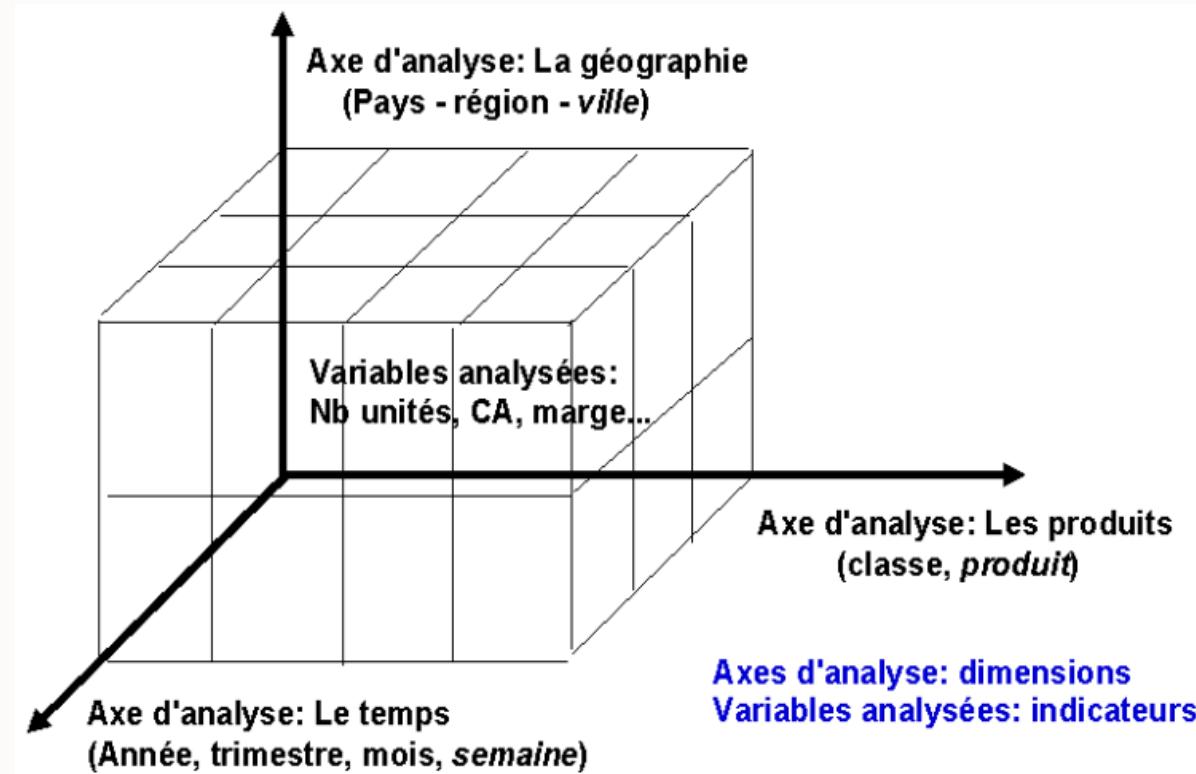
--> Approche multidimensionnelle

- Souvent 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,
  - hypercube 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

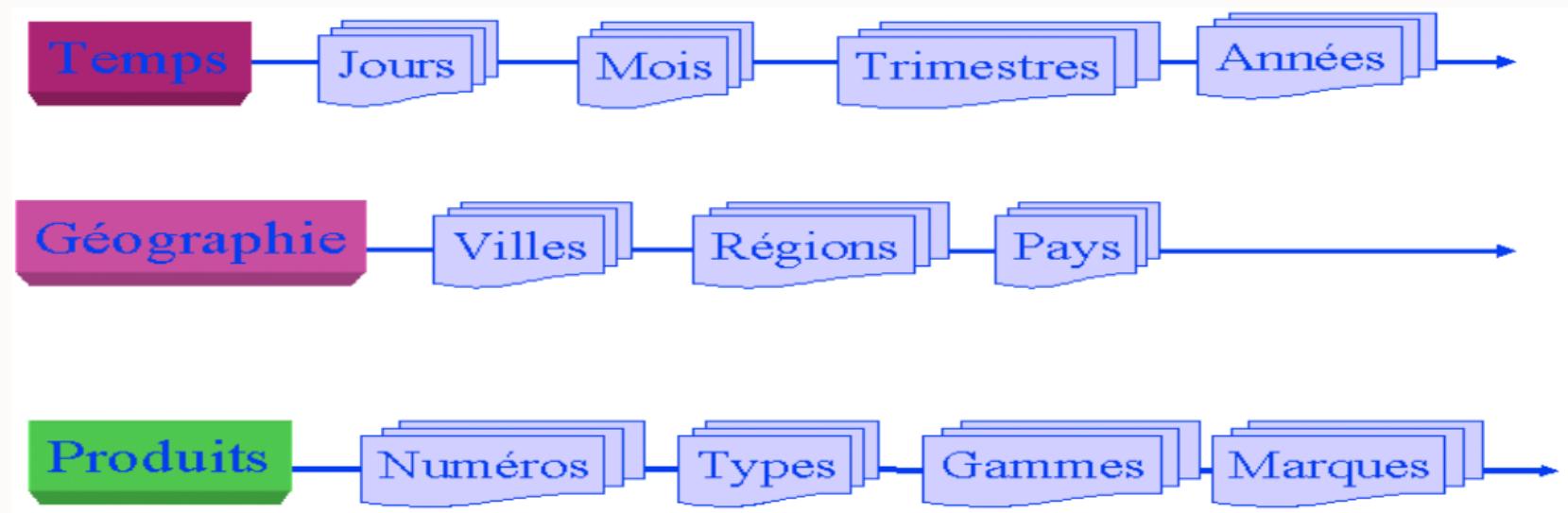
### Agrégation des données

- 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.

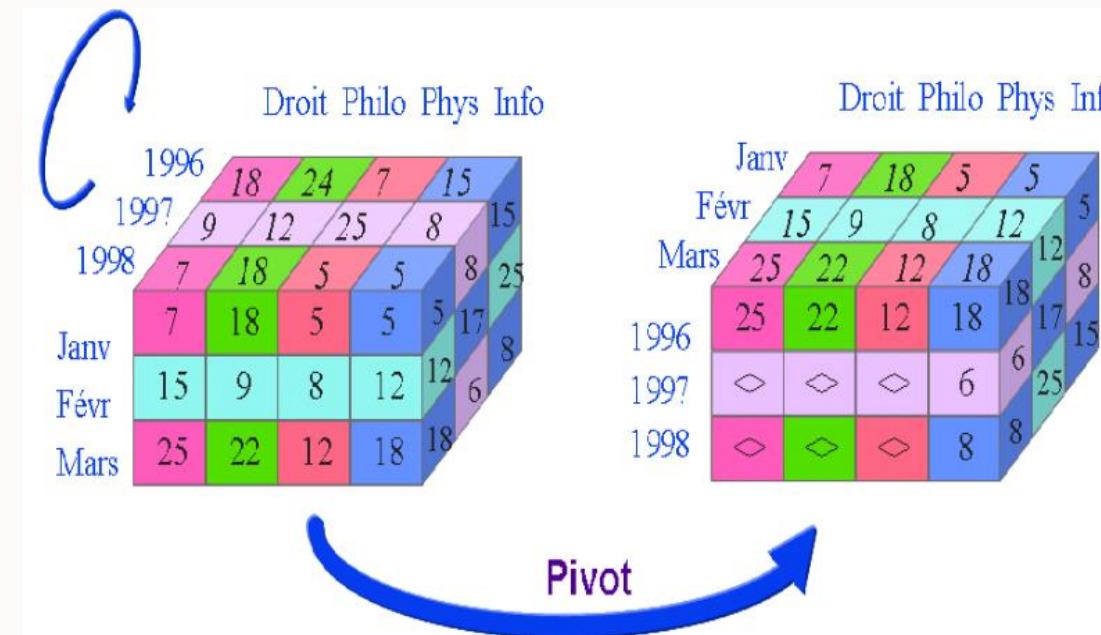
# Les concepts de base de Data Warehouse

## OLAP et Analyse multidimensionnelles

### Opérations OLAP

-> Opérations sur la structure des cubes

- Pivot (Rotation)



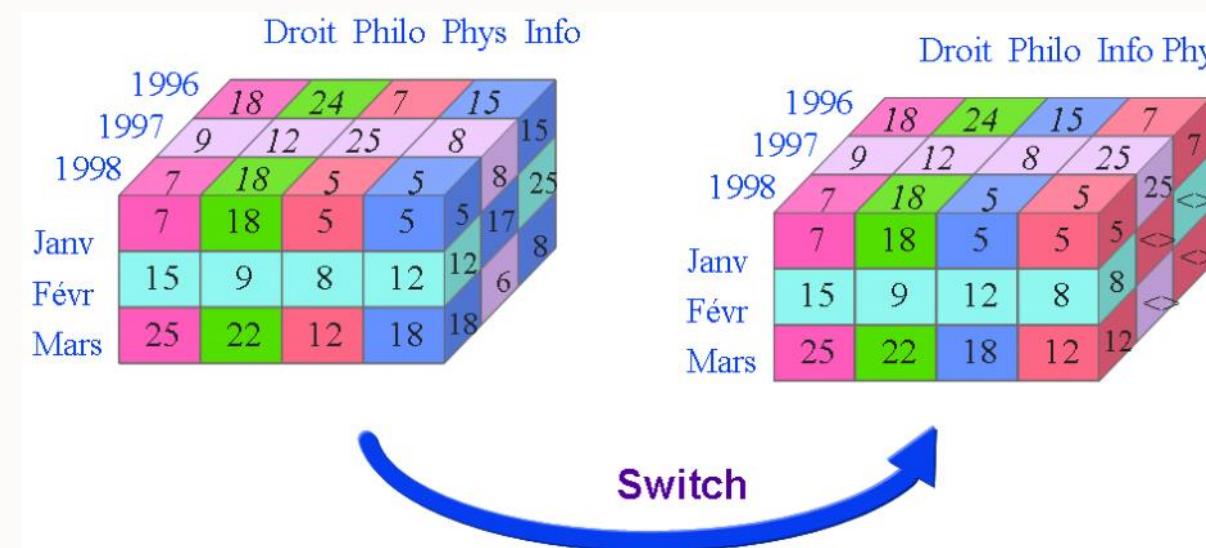
# Les concepts de base de Data Warehouse

## OLAP et Analyse multidimensionnelles

### Opérations OLAP

-> Opérations sur la structure des cubes

- Switch (Permutation)



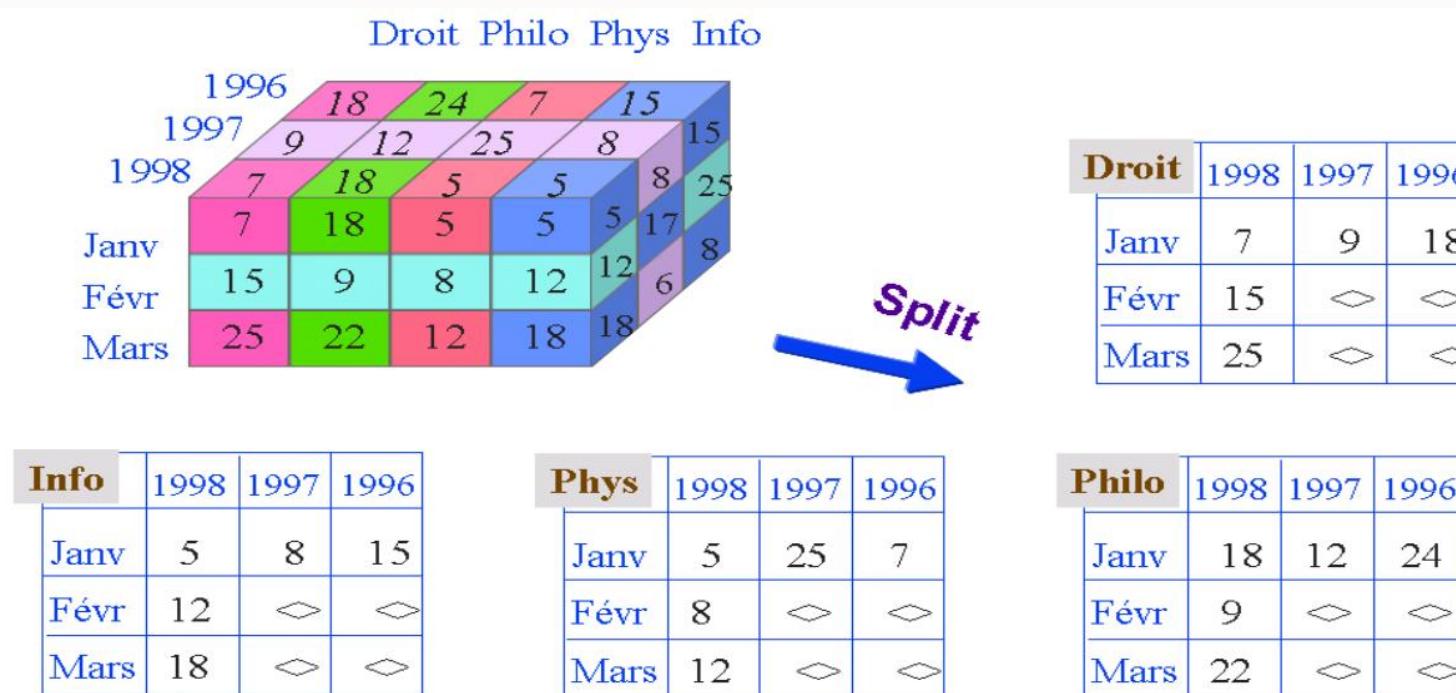
# Les concepts de base de Data Warehouse

## OLAP et Analyse multidimensionnelles

### Opérations OLAP

-> Opérations sur la structure des cubes

- Split (Décomposition)



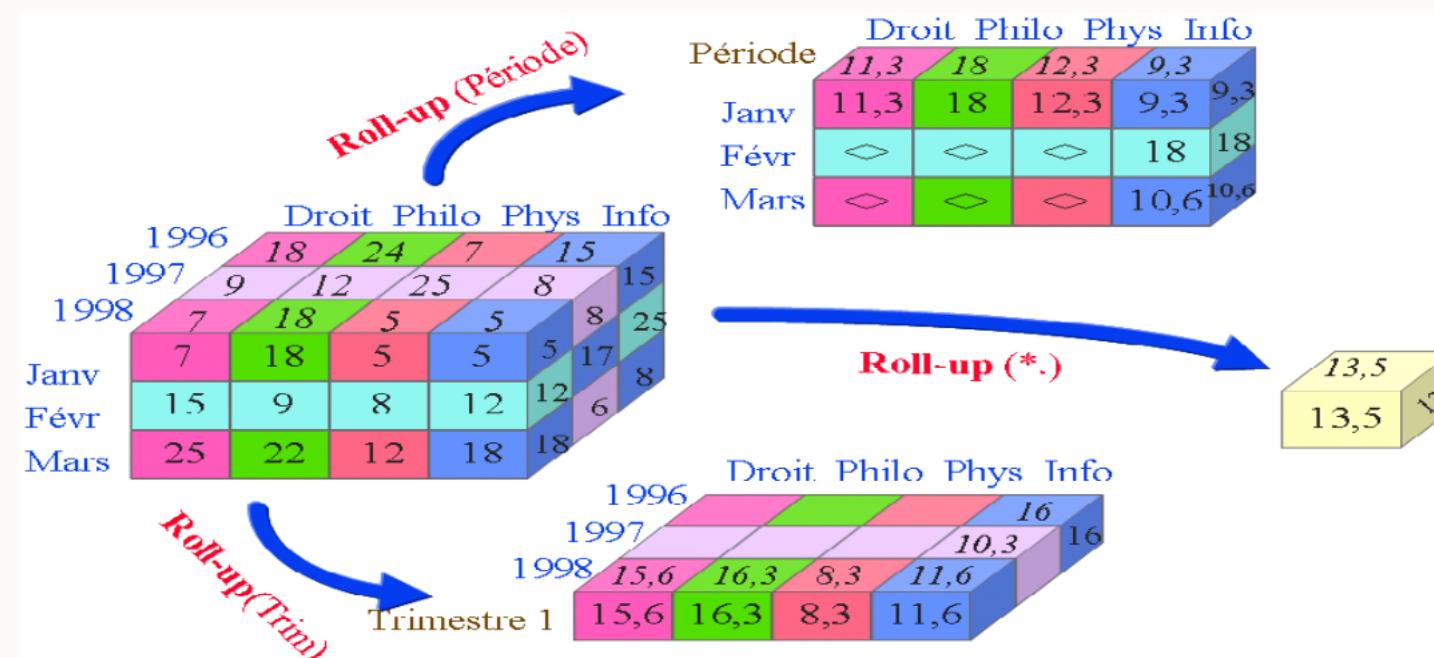
# **Les concepts de base de Data Warehouse**

# OLAP et Analyse multidimensionnelles

# Opérations OLAP

## -> Opérations sur le contenu des cubes

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



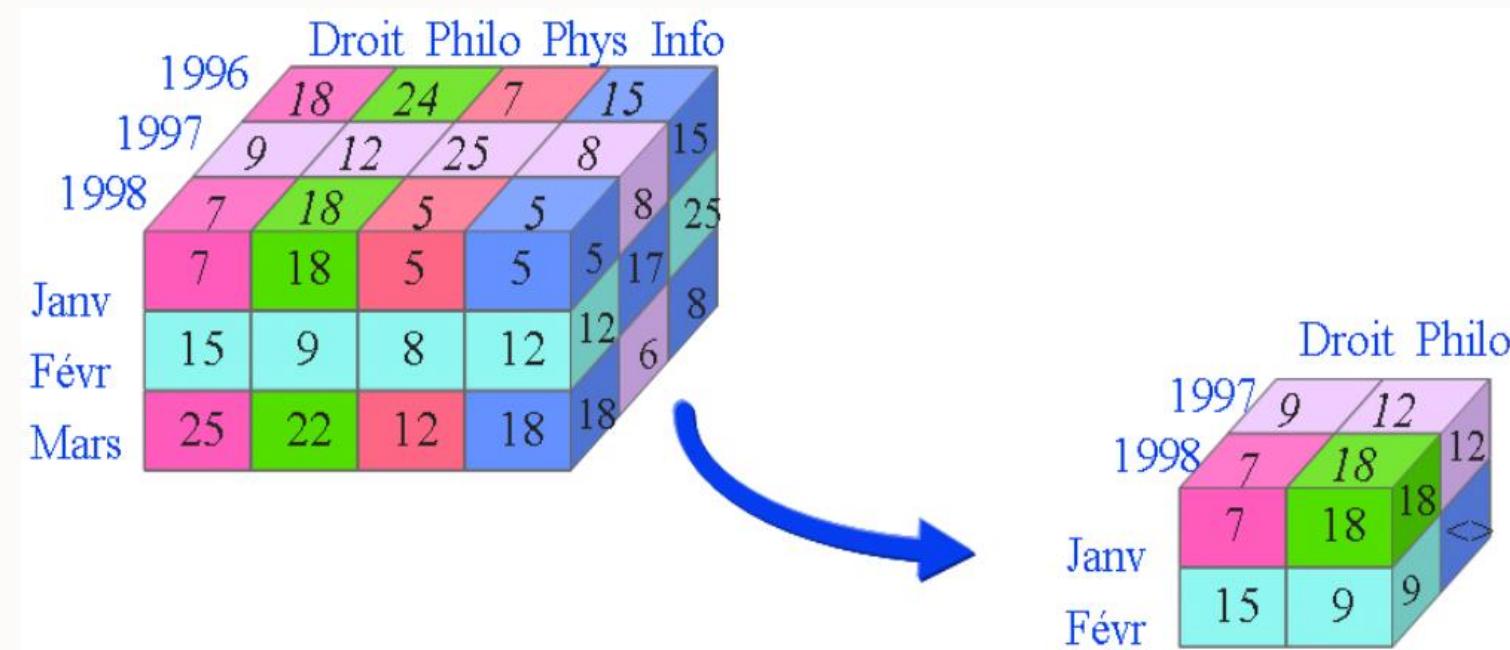
# Les concepts de base de Data Warehouse

## OLAP et Analyse multidimensionnelles

### Opérations OLAP

-> Opérations sur la structure des cubes

- Slice (Restriction)



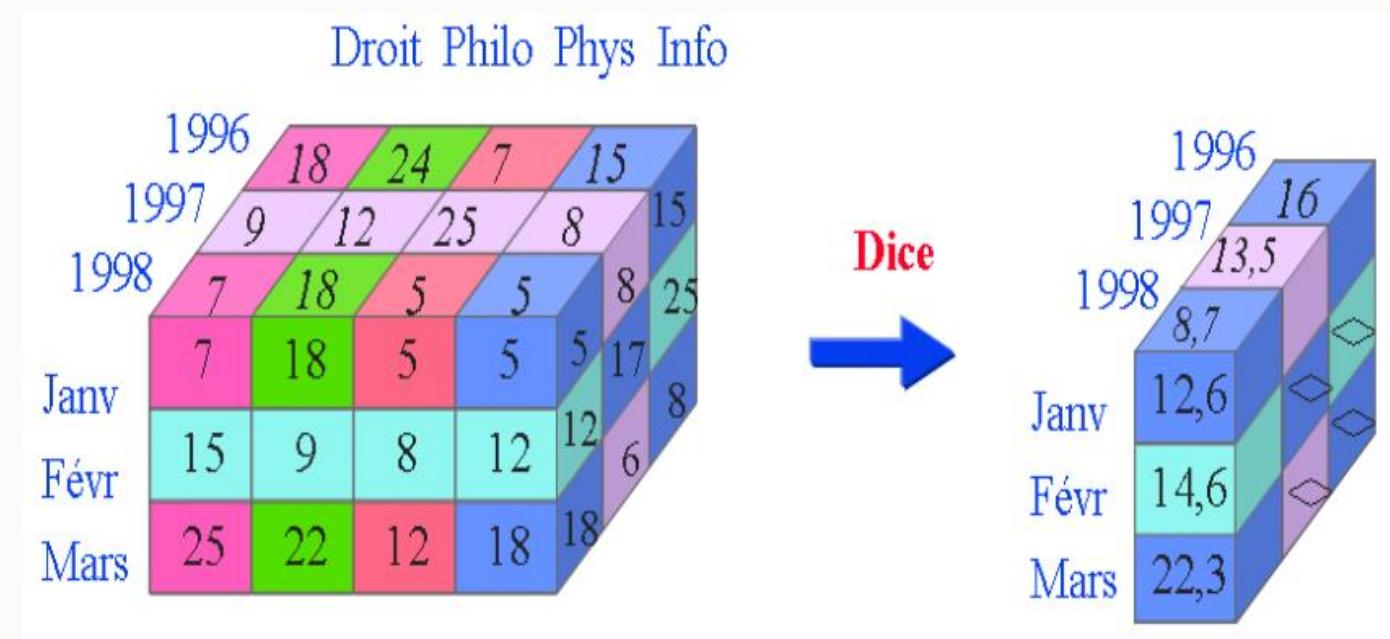
# Les concepts de base de Data Warehouse

## OLAP et Analyse multidimensionnelles

### Opérations OLAP

-> Opérations sur la structure des cubes

- Dice (Projection)



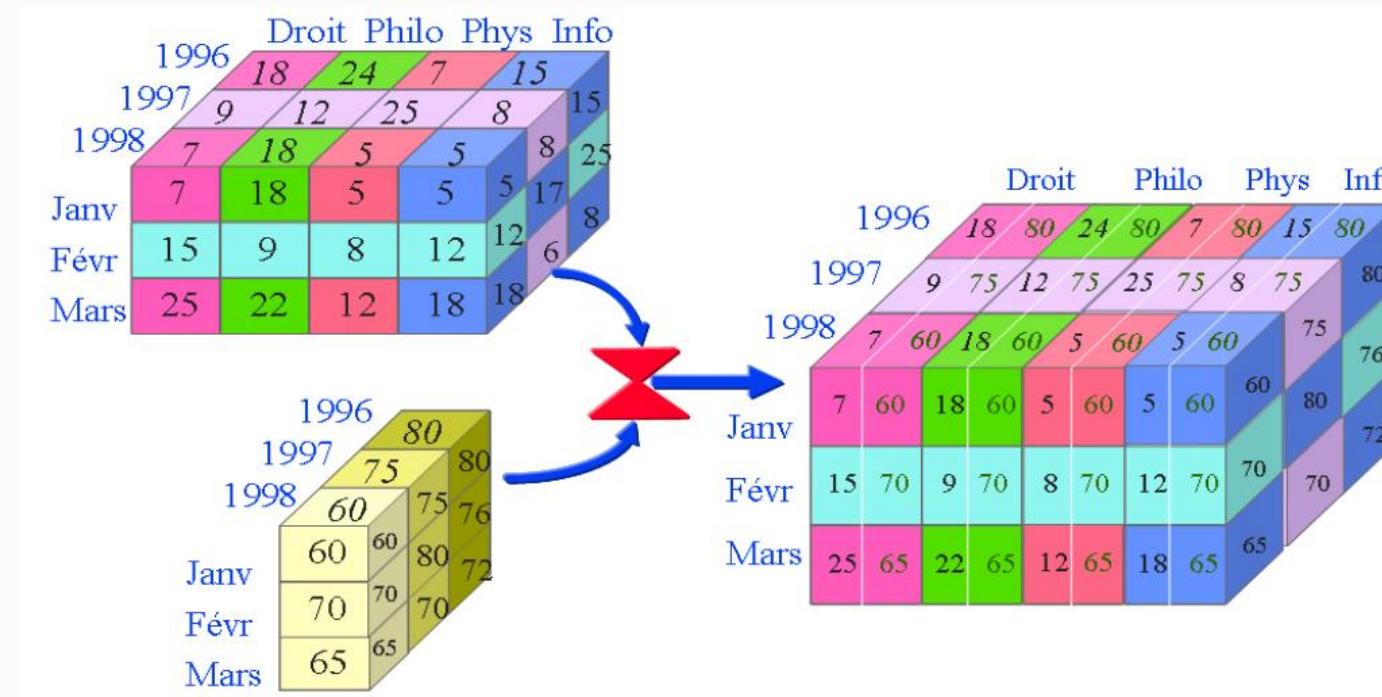
# Les concepts de base de Data Warehouse

## OLAP et Analyse multidimensionnelles

### Opérations OLAP

-> Opérations entre cubes

- Jointure



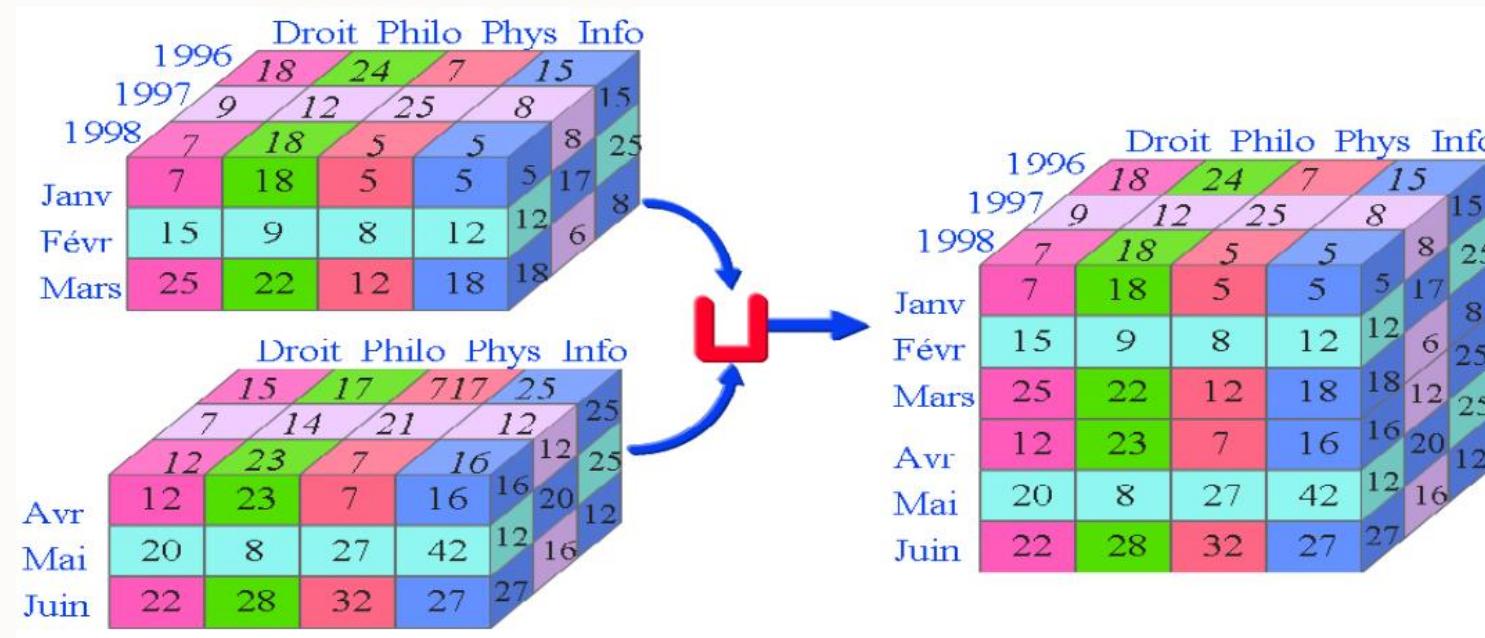
# Les concepts de base de Data Warehouse

## 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 définis
  - 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

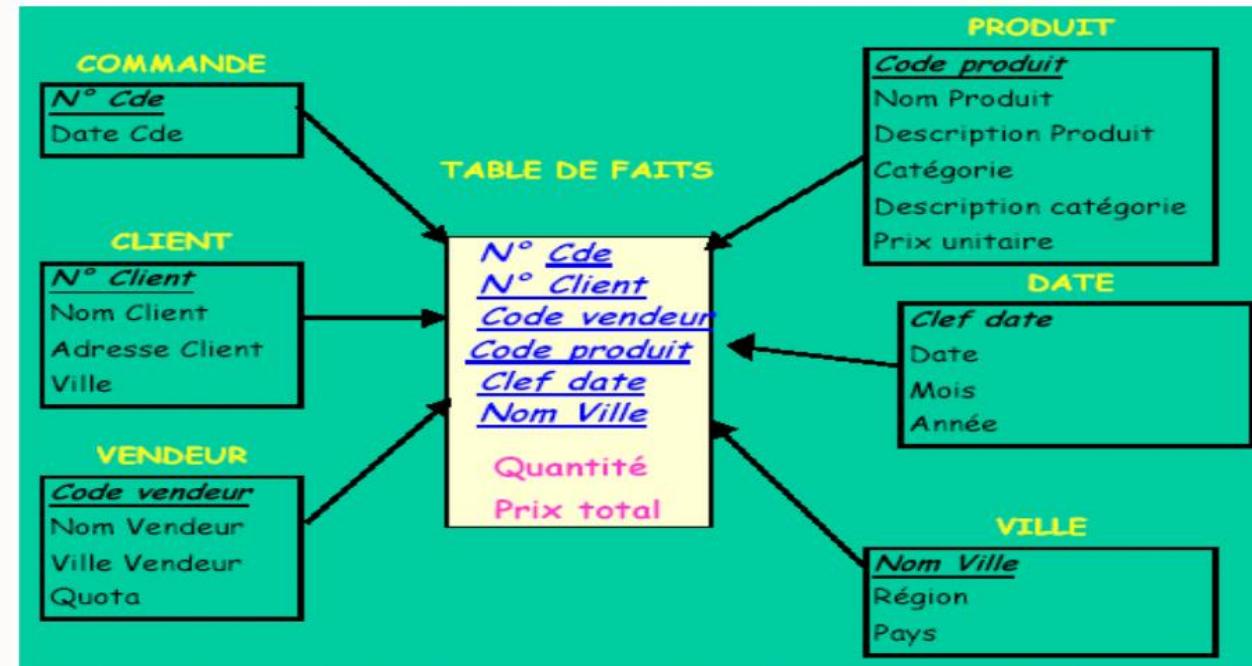
- Niveau conceptuel:
  - Un DW est basé sur une modélisation multidimensionnelle qui représente les données dans un cube
  - Un cube permet de voir les données suivant plusieurs dimensions:
    - Tables de dimensions
    - La table des faits contient les mesures et les clés des dimensions
- Niveau Logique:
  - 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.

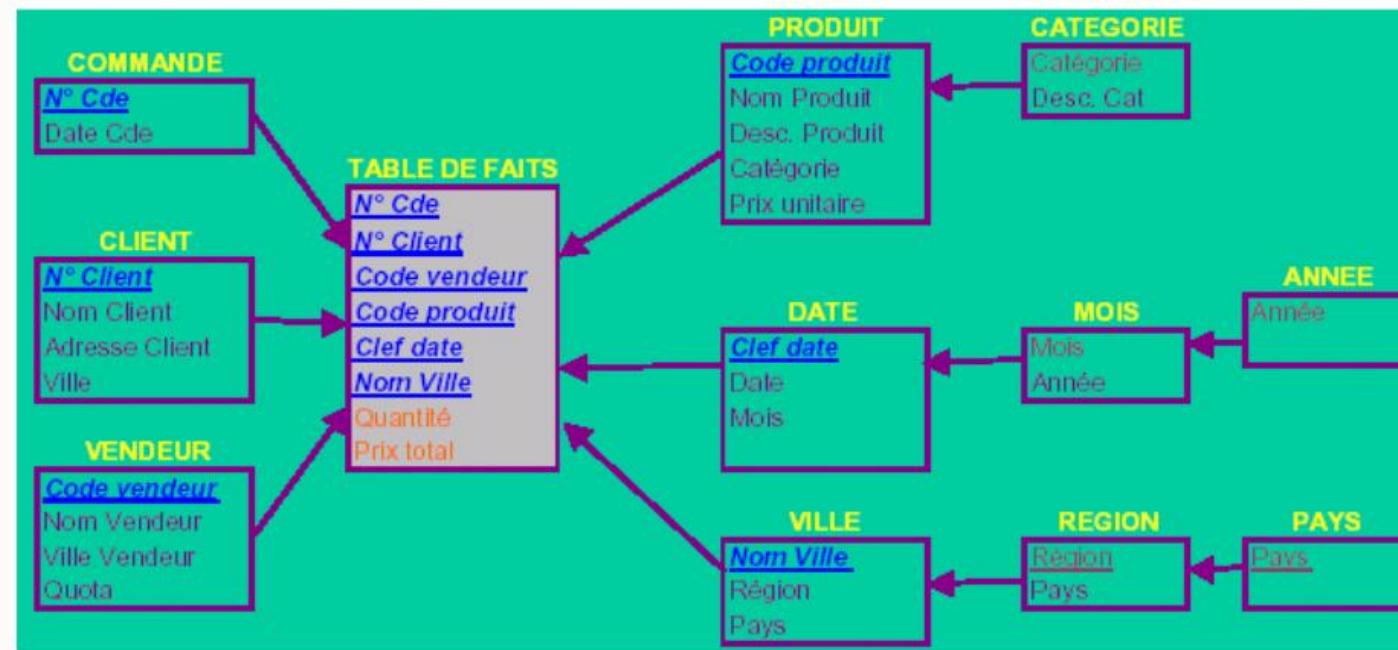


# Les concepts de base de Data Warehouse

## Modélisation et Conception d'un DW

### Modélisation

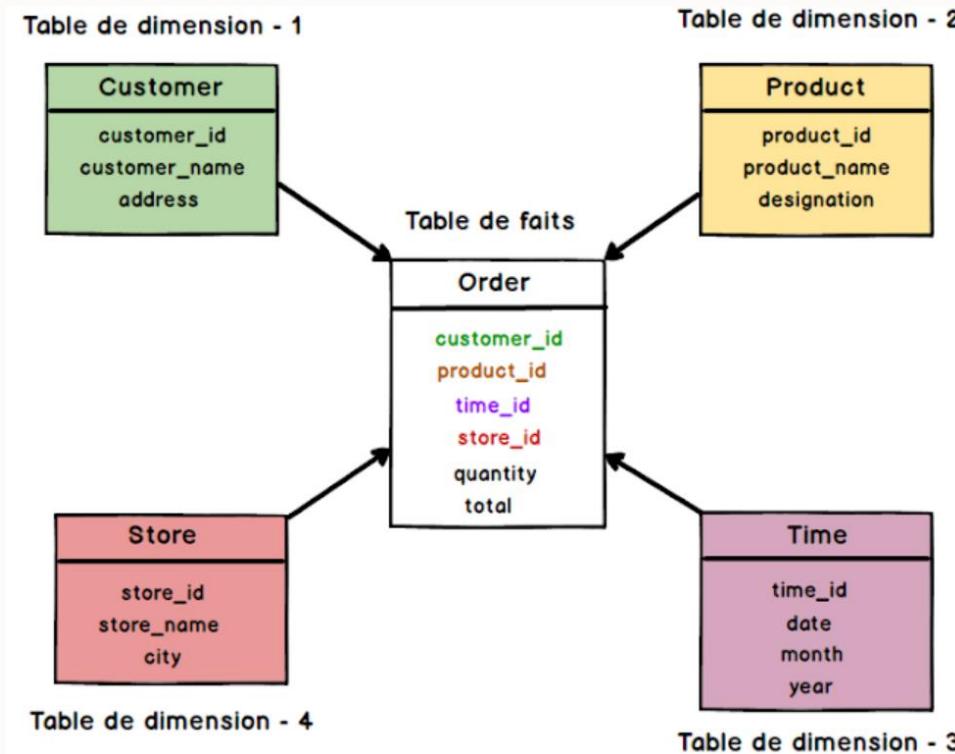
- Schéma en flocons
  - Raffinement du schéma étoile avec des tables normalisées par dimensions.



# 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.



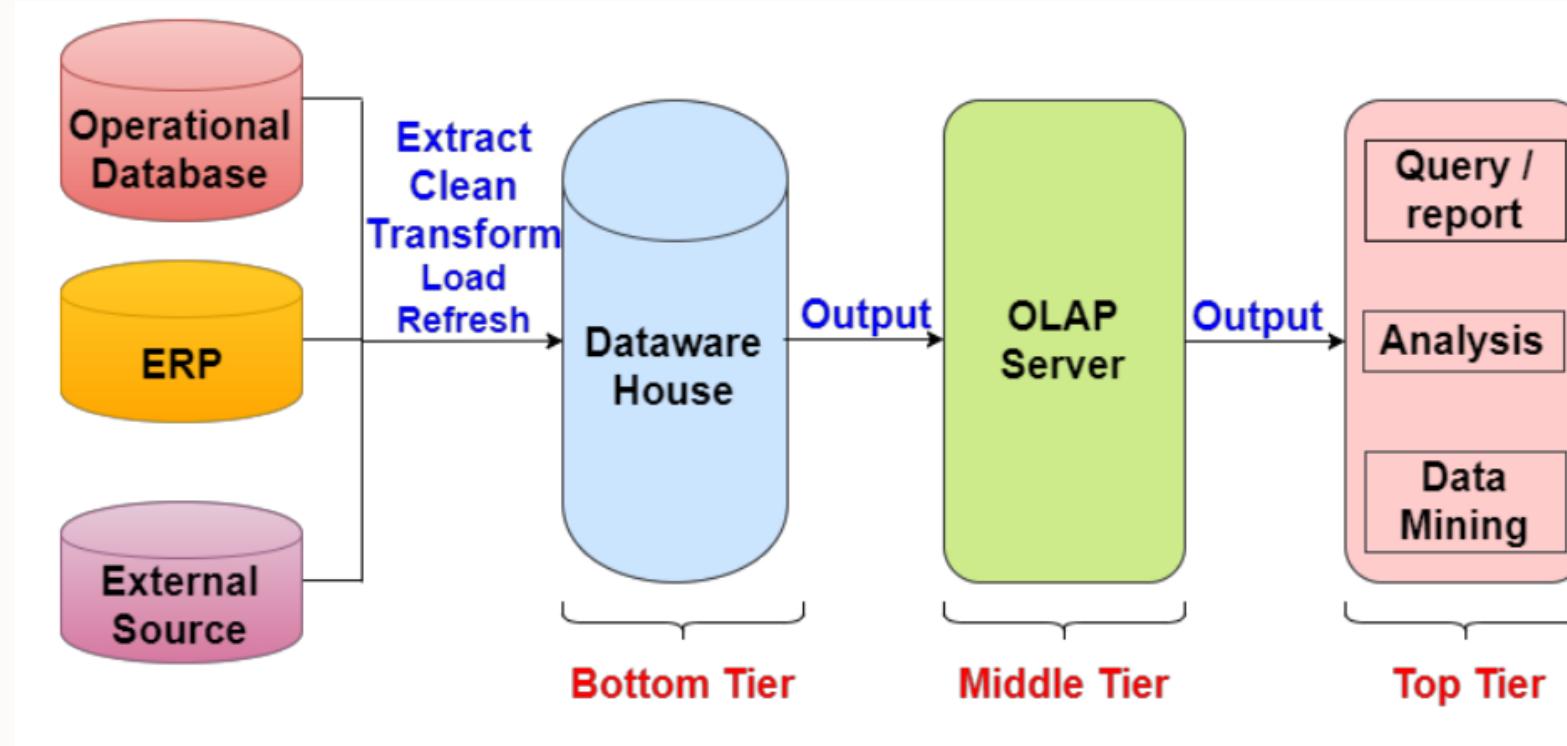
# **Chapitre 3 :**

## **Le système de Data Warehouse et ses composants**

### **(Architecture d'un Data Warehouse)**

# Le système de Data Warehouse et ses composants

## Architecture d'un Data Warehouse



Architecture générale Data Warehouse

# Le système de Data Warehouse et ses composants

## Architecture d'un Data Warehouse

Généralement, un entrepôt de données adopte une architecture à trois niveaux :

- **Niveau inférieur ou Bottom Tier** : composé généralement du système de base de données relationnel de l'entrepôt. Les programmes d'applications et les utilitaires ETL sont utilisés pour fournir les données au niveau inférieur.
- **Niveau intermédiaire ou Middle Tier** : le niveau où se trouve le serveur OLAP implémenté par deux modèles OLAP relationnel (ROLAP) et OLAP multidimensionnel (MOLAP).
- **Niveau supérieur ou Top Tier** : c'est la couche client. Elle contient les outils de requête et les outils de génération de rapports, les outils d'analyse et les outils d'exploration des données

la **différence** la plus importante entre les deux est que **ROLAP** fournit des données, directement à partir de l'entrepôt de données(data warehouse) principal, alors que **MOLAP** fournit des données à partir des bases de données propriétaires MDDB(Multi Dimensional Data Base)

# **Le système de Data Warehouse et ses composants**

## **Processus Extract, Tranform, Load (ETL)**

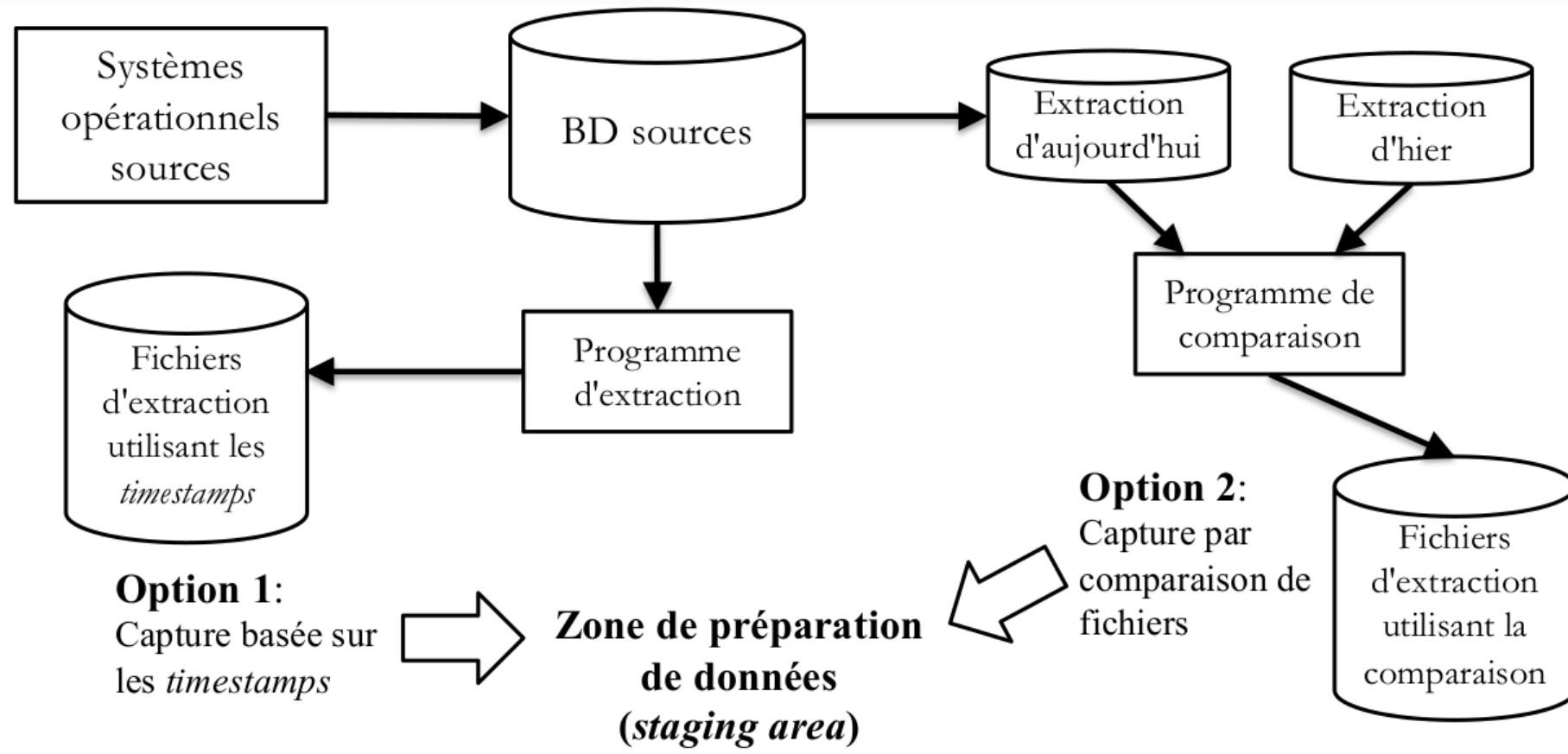
- **Extraire les données des sources hétérogènes (extract)**
  - Identifier les données sources utiles
  - Déterminer les données qui ont changé
- **Consolider les données (transform)**
  - Données redondantes, manquantes, incohérentes, etc.
  - Découpage, fusion, conversion, aggrégation, etc.
- **Charger les données intégrées dans l'entrepôt (load)**
  - Mode différé (batch) ou quasi temps-réel.
- Partie la plus longue du développement (jusqu'à 70% du temps total).

# Le système de Data Warehouse et ses composants

## Processus Extract, Tranform, Load (ETL)

### Extraction des données (différée)

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

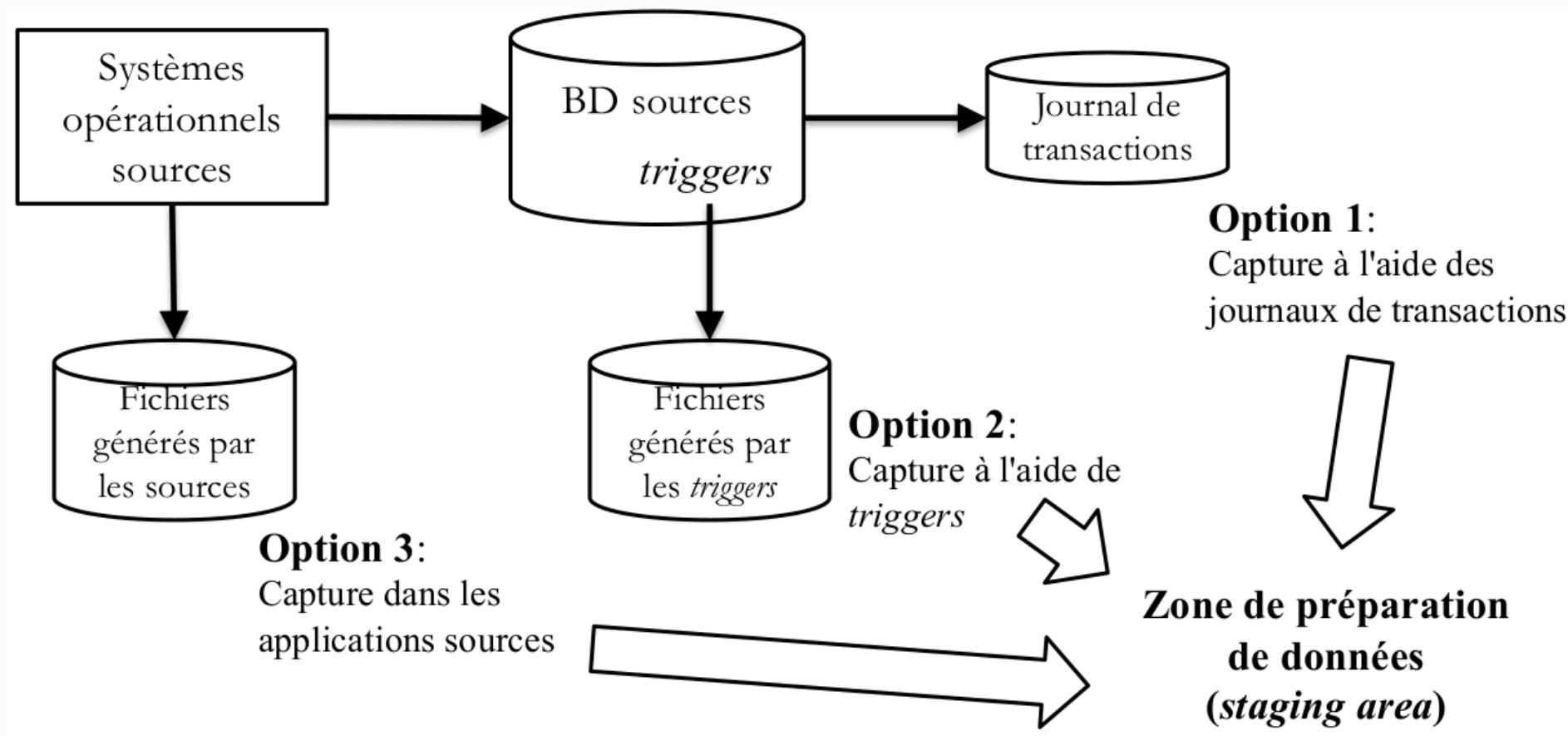


# Le système de Data Warehouse et ses composants

## Processus Extract, Tranform, Load (ETL)

### Extraction des données (temps-réel)

- S'effectue au moment où les transactions surviennent dans les systèmes sources.



# Le système de Data Warehouse et ses composants

## Processus Extract, Tranform, Load (ETL)

### Transformation des données

- Révision de format:
  - Ex: Changer le type ou la longueur de champs individuels.
- Décodage de champs:
  - Ex: ['homme', 'femme'] vs ['M', 'F'] vs [1,2].
- Pré-calcul des valeurs dérivées:
  - Ex: profit calculé à partir de ventes et coûts.
- Découpage de champs complexes:
  - Ex: extraire les valeurs prénom, secondPrénom et nomFamille à partir d'une seule chaîne de caractères nomComplet.
- Pré-calcul des agrégations:
  - Ex: ventes par produit par semaine par région.
- Déduplication
  - Ex: Plusieurs enregistrements pour un même client

# **Le système de Data Warehouse et ses composants**

## **Processus Extract, Tranform, Load (ETL)**

### **Chargement des données**

- Faire les chargements en lot dans une période creuse (entrepôt de données non utilisé);
- Considérer la bande passante requise pour le chargement;
- Avoir un plan pour évaluer la qualité des données chargées dans l'entrepôt;
- Commencer par charger les données des tables de dimension;
- Désactiver les indexées et clés étrangères lors du chargement.

# **Le système de Data Warehouse et ses composants**

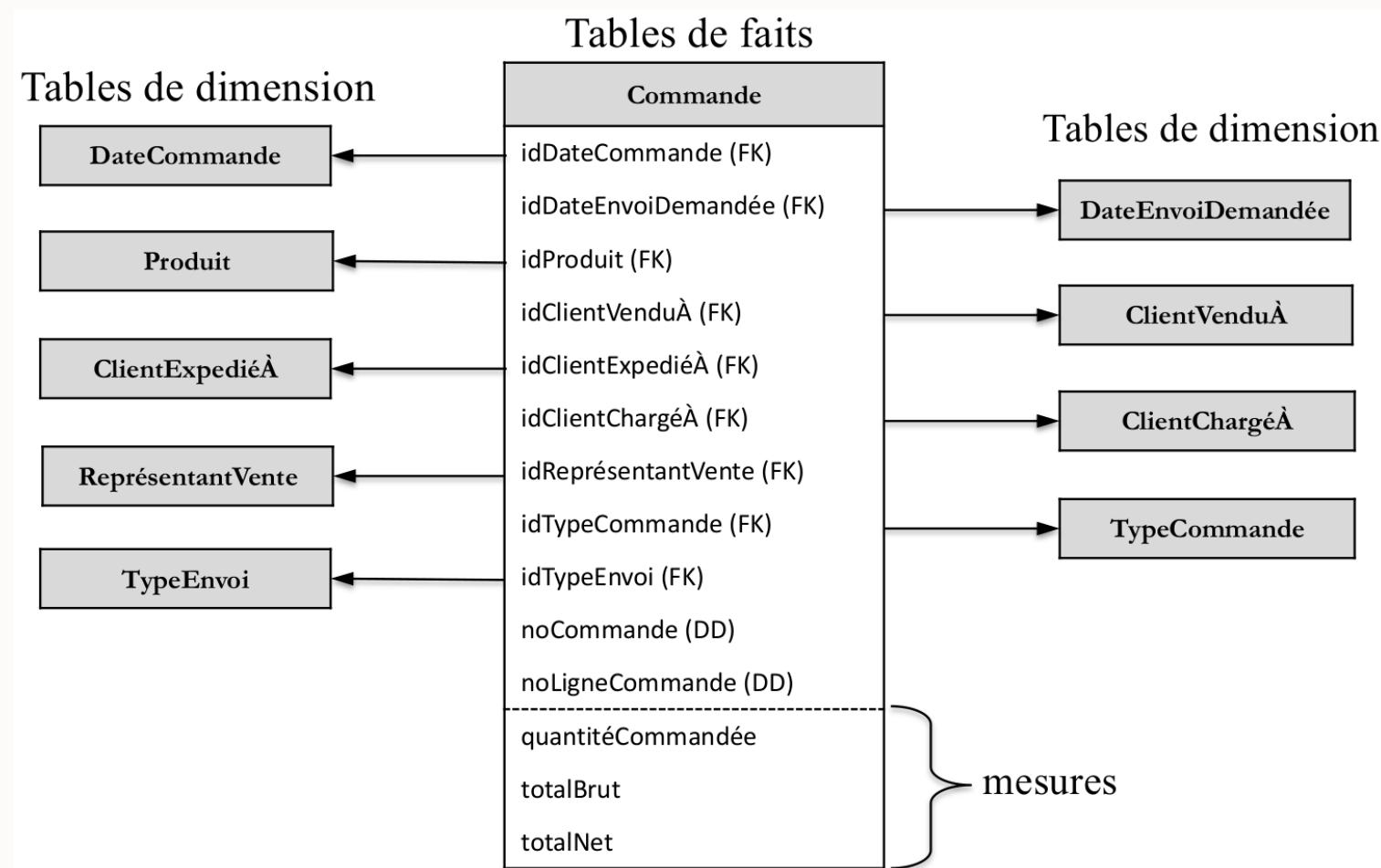
## **Processus Extract, Tranform, Load (ETL)**

### **Modélisation dimensionnelle**

- Représente les données sous la forme d'un schéma en étoile:
  - Table de faits entourée de plusieurs tables de dimension (normalement entre 8 et 15)
- Les faits (mesures) sont généralement des valeurs numériques provenant des processus d'affaires;
- Les dimensions fournissent le contexte (qui, quoi, quand, où, pourquoi et comment) des faits;
- Les tables ne sont pas normalisées

# Le système de Data Warehouse et ses composants

## Exemple de schéma en étoile (Commande de produits)



# Le système de Data Warehouse et ses composants

## Exemple de schéma en étoile (Commande de produits)

Tables de dimension:

| DateCommande         |
|----------------------|
| idDate (PK)          |
| date                 |
| jourDeSemaine        |
| jourDuMois           |
| jourDeAnnée          |
| jourDansMoisFiscal   |
| jourDansAnnéeFiscale |
| congéFérié           |
| jourDeTravail        |
| semaineDuMois        |
| ...                  |

| Produit        |
|----------------|
| idProduit (PK) |
| description    |
| SKU            |
| marque         |
| sousCatégorie  |
| catégorie      |
| département    |
| poids          |
| taille         |
| couleur        |
| ...            |

| ClientExpédiéÀ    |
|-------------------|
| idClient (PK)     |
| nomFamille        |
| prénom            |
| sexe              |
| dateNaissance     |
| dateAbonnement    |
| forfaitAbonnement |
| adresseRue        |
| adresseVille      |
| adresseProvince   |
| ...               |

## Tables de faits

- Correspondent à un événement d'affaires
  - Ex: achat d'un produit par un client, envoi du produit au client, commande de matériaux auprès d'un fournisseur, etc.
- Contiennent deux types de colonnes:
  - Des métriques associées à l'événement d'affaire:
    - Ex: total des ventes, nombre d'items commandés, etc.
  - Des clés étrangères vers les tables de dimension:
    - Ex: ID du client qui fait la commande, ID du produit commandé, etc.
- Contiennent typiquement un très grand nombre de lignes:
  - Jusqu'à plusieurs milliards de lignes;
  - Souvent plus de 90% des données du modèle.

## Tables de dimension

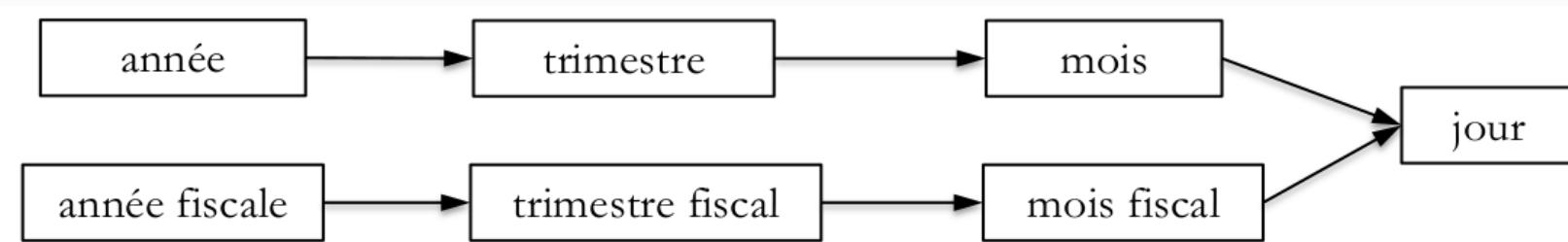
- Ensemble hautement corrélé d'attributs (jusqu'à plusieurs dizaines) regroupés selon les objets clés d'une entreprise:
  - Ex: produits, clients, employés, installations, etc.
- Propriétés des attributs:
  - Descriptif (ex: chaînes de caractères);
  - De qualité (ex: aucune valeur manquante, obsolète, erronée, etc.);
  - Valeurs discrètes (ex: jour, âge d'un client);
- Rôles des attributs:
  - Filtrer/agréger les données (ex: ville, catégorie produit, etc.);
  - Étiqueter les résultats (ex: champs descripteurs).

# Le système de Data Warehouse et ses composants

## Hiérarchies dimensionnelles

- Ensemble d'attributs d'une table de dimension ayant une relation hiérarchique (x est inclus dans y);
- Correspondent à des relations de type 1 à plusieurs;
- Définissent les chemins d'accès dans les données (drill-down paths);
- Peuvent être simples:
  - Produit : tous → catégorie → marque → produit;
  - Lieu : tous → pays → province → ville → code postal.

Ou multiples:



# Le système de Data Warehouse et ses composants

## Dimension temporelle

| Dimension: Temps     |
|----------------------|
| idDate (PK)          |
| date                 |
| jourDeSemaine        |
| jourDuMois           |
| jourDeAnnée          |
| jourDansMoisFiscal   |
| jourDansAnnéeFiscale |
| congéFérié           |
| jourDeTravail        |
| semaineDuMois        |
| ...                  |

- Mettre toutes ces valeurs, même si la plupart peuvent être déduites d'une seule colonne;
- Pré-générer les lignes de la table (ex: 10 prochaines années) pour faciliter la référence et éviter les mises à jour

# Le système de Data Warehouse et ses composants

## Dimension temporelle

- Problème: avoir un grain trop fin dans la dimension temporelle (ex: temps du jour) peut causer l'explosion du nombre de rangées:
  - Ex: 31,000,000 secondes différentes dans une année.
- Solution: mettre le temps du jour dans une dimension séparée:
  - Dimension Date: année → mois → jour;
  - Dimension TimeOfDay : heure → minute → secondes;
  - 86,400 + 365 lignes au lieu de 31,000,000 lignes.
- Note: la dimension TimeOfDay est souvent modélisée comme un simple champs dans la table de faits.

## Dimensions à évolution lente (SCD)

- Slowly Changing Dimensions (SCD);
- Même si elles sont plus statiques que les tables faits, les dimensions peuvent également changer:
  - Ex: adresse d'un client, catégorie d'un produit, etc.
- Stratégies d'historisation:
  - SCD Type 1: Écraser l'ancienne valeur avec la nouvelle
  - SCD Type 2: Ajouter une ligne dans la table de dimension pour la nouvelle valeur
  - SCD Type 3: Avoir deux colonnes dans la table de dimension correspondant à l'ancienne et la nouvelle valeur

# Le système de Data Warehouse et ses composants

## Dimensions à évolution lente (SCD)

Stratégie SCD Type 1

| Product Key | Product Description | Department | SKU Number (Natural Key) |
|-------------|---------------------|------------|--------------------------|
| 12345       | IntelliKidz 1.0     | Education  | ABC922-Z                 |

| Product Key | Product Description | Department | SKU Number (Natural Key) |
|-------------|---------------------|------------|--------------------------|
| 12345       | IntelliKidz 1.0     | Strategy   | ABC922-Z                 |

- Impossible de faire des analyses sur l'ancienne valeur;
- À utiliser seulement lorsque l'ancienne valeur n'est pas significative pour les besoins d'affaires;
- Exige de mettre à jour les données agrégées avec l'ancienne valeur.

## Dimensions à évolution lente (SCD)

Stratégie SCD Type 2

| Product Key | Product Description | Department | SKU Number (Natural Key) |
|-------------|---------------------|------------|--------------------------|
| 12345       | IntelliKidz 1.0     | Education  | ABC922-Z                 |
| 25984       | IntelliKidz 1.0     | Strategy   | ABC922-Z                 |

- Permet de faire des analyses historiques;
- Demande l'ajout d'une nouvelle ligne par changement;
- À utiliser lorsque l'ancienne valeur a une signification analytique ou si le changement est une information en soi.

# Le système de Data Warehouse et ses composants

## Dimensions à évolution lente (SCD)

Stratégie SCD Type 3

| Product Key | Product Description      | Prior Department | SKU Number (Natural Key) |
|-------------|--------------------------|------------------|--------------------------|
| 12345       | IntelliKidz 1.0 Strategy | Education        | ABC922-Z                 |

- Rarement employée;
- Profondeur de l'historique est de un seul changement;
- Utilisé lorsqu'on veut vouloir comparer les faits avec l'ancienne ou la nouvelle valeur;

# **Le système de Data Warehouse et ses composants**

## **Types d'entrepôts de données**

### **Types d'entrepôts de données**

1. Magasins de données
2. Entrepôts de données d'entreprise (EDW)
  - Bus de magasins de données (datamart bus)
  - Hub-and-spokes
  - Entrepôts de données fédérés

# **Le système de Data Warehouse et ses composants**

## **Types d'entrepôts de données**

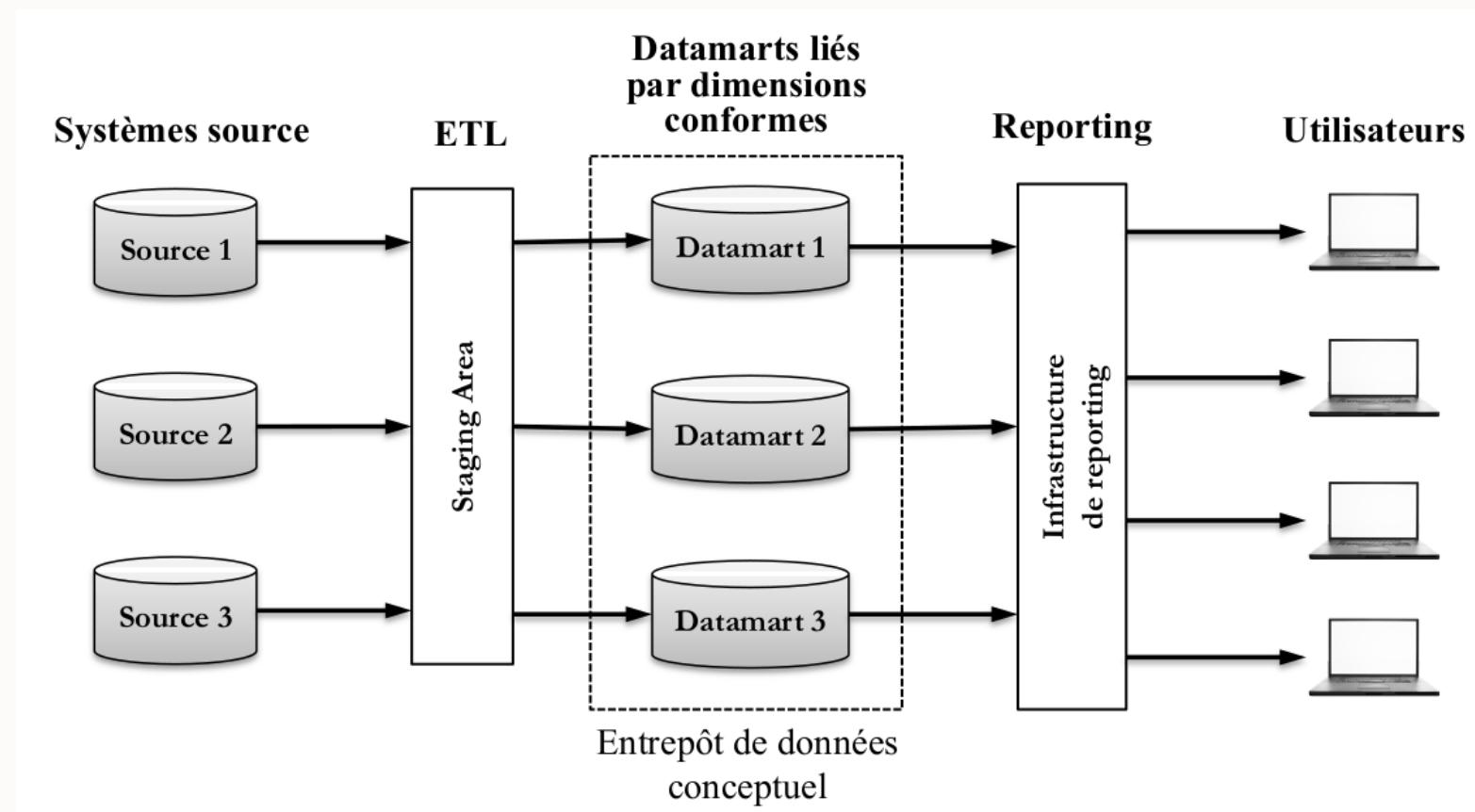
### **Magasins de données (datamart)**

- Contiennent une portion du contenu de l'entrepôt de données;
- Se concentre sur un seul sujet d'analyse (ex: les ventes OU l'inventaire, mais pas les deux);
- Servent à faire des analyses simples et spécialisées (ex: fluctuations des ventes par catégorie de produits);
- Nombre de sources limitées, provenant la plupart du temps d'un même département;
- Modélisés sous la forme d'un schéma en étoile.

# Le système de Data Warehouse et ses composants

## Types d'entrepôts de données

### Architecture Datamart bus



# Le système de Data Warehouse et ses composants

## Types d'entrepôts de données

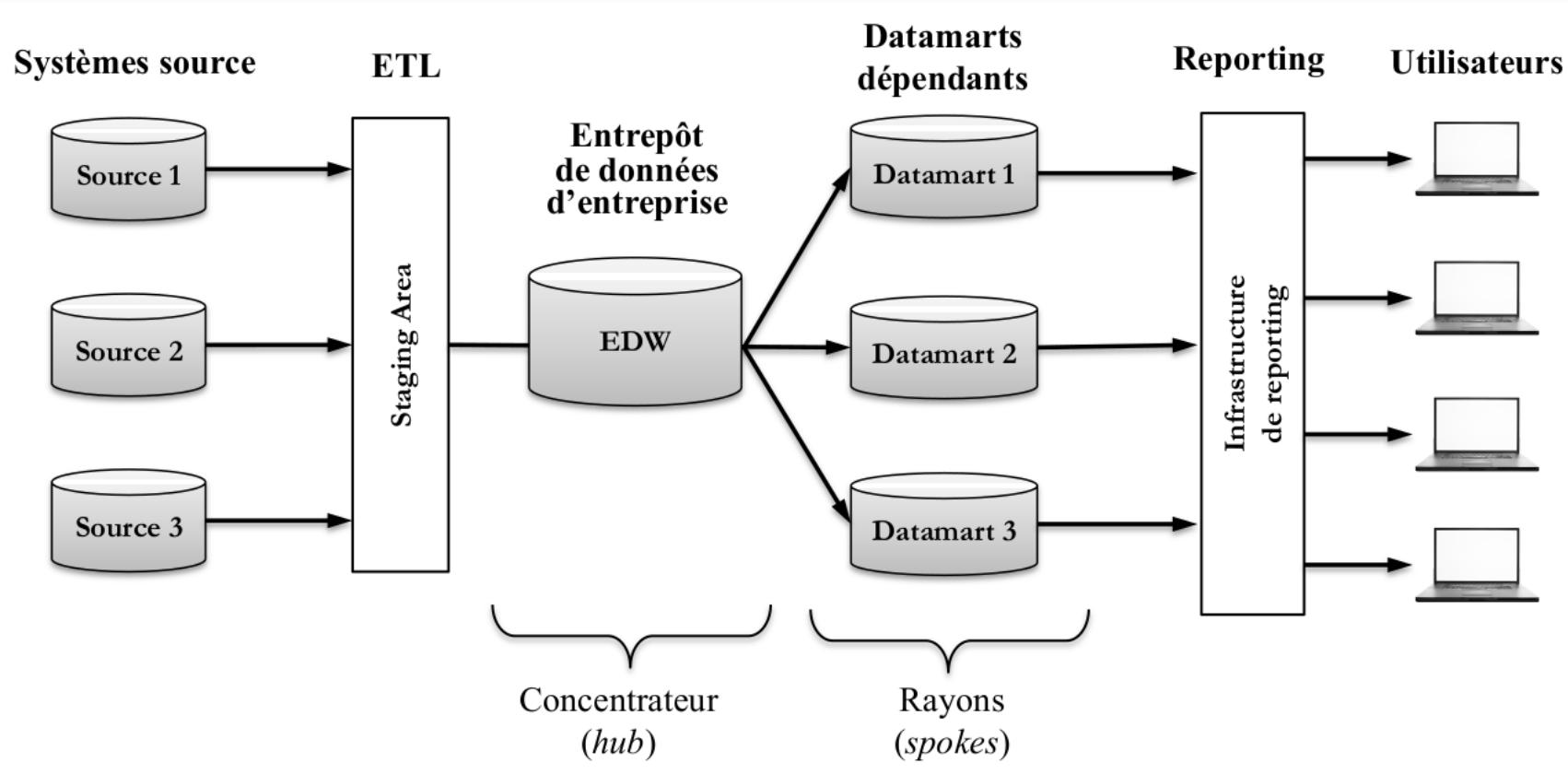
### Architecture Datamart bus

- Approche bottom-up, où on construit l'entrepôt un datamart à la fois;
- Modélisation dimensionnelle (schéma en étoile) des datamarts, au lieu du diagramme entité-relation;
- Entrepôt de données conceptuel, formé de magasins de données inter-reliés à l'aide d'une couche d'intergiciels (middleware);
- Intégration des données assurée par les dimensions partagées entre les datamarts (i.e., dimensions conformes);
- Approche incrémentale qui donne des résultats rapidement (développement agile);

# Le système de Data Warehouse et ses composants

## Types d'entrepôts de données

### Architecture Hub-and-spoke



# Le système de Data Warehouse et ses composants

## Types d'entrepôts de données

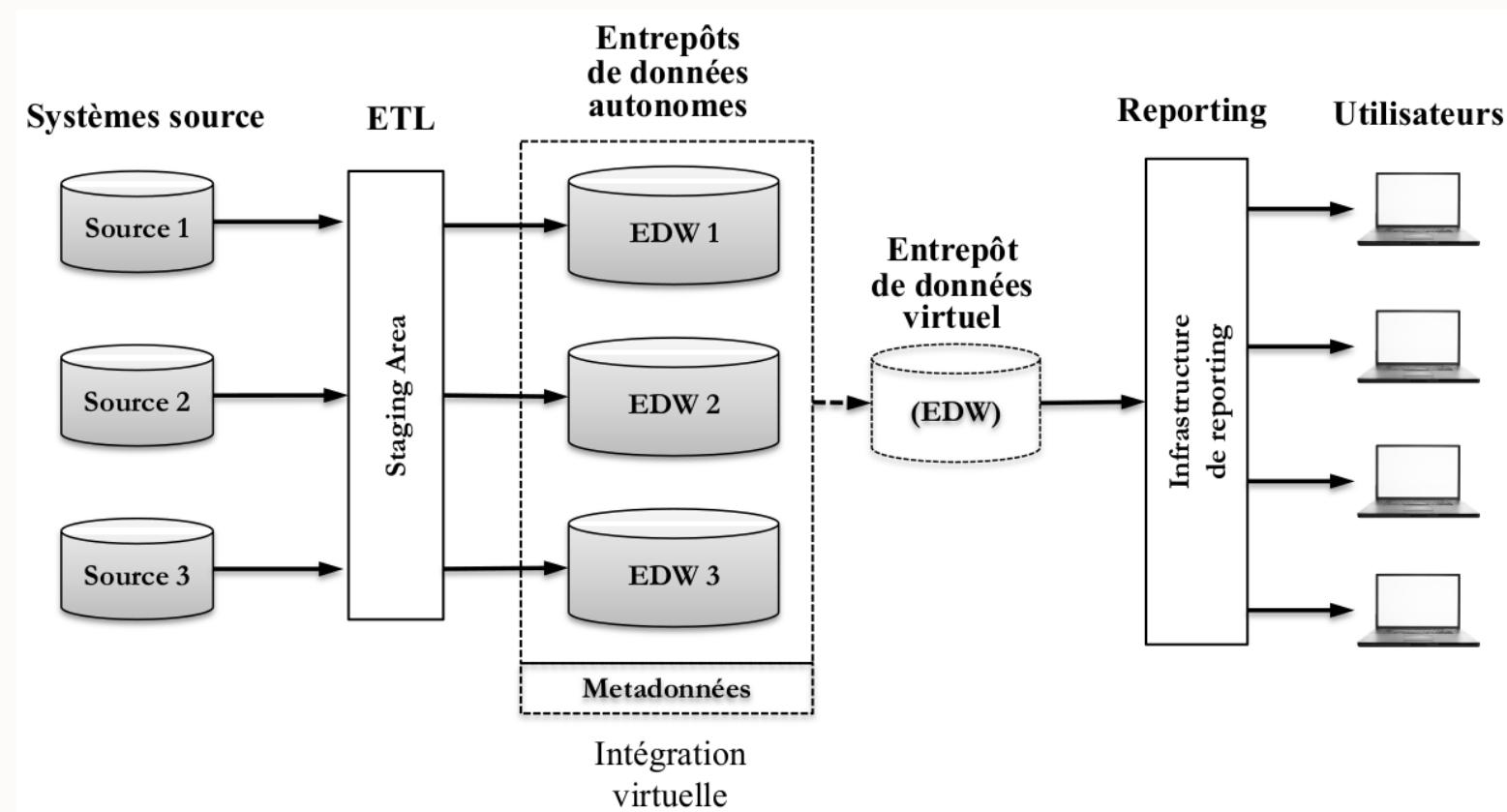
### Architecture Hub-and-spoke

- Approche top-down, favorisant l'intégration et consolidation complète des données de l'entreprise
- Entrepôt (hub) contient les données atomiques (niveau de détail le plus fin) et normalisées (3FN);
- Les datamarts (spokes) contiennent principalement des données agrégées (pas atomique) et suivant le modèle dimensionnel;
- La plupart des requêtes analytiques sont faites sur les datamarts;
- Développement plus long, dû à la complexité du processus ETL et de la modélisation;
- Meilleure qualité de données que l'architecture par bus de datamarts.

# Le système de Data Warehouse et ses composants

## Types d'entrepôts de données

### Architecture fédérée



# **Le système de Data Warehouse et ses composants**

## **Types d'entrepôts de données**

### **Architecture fédérée**

- Entrepôt de données distribué sur plusieurs systèmes hétérogènes;
- Données intégrées logiquement ou physiquement à l'aide de méta-données (ex: XML);
- Opère de manière transparente (l'utilisateur ne voit pas que les données sont réparties);
- Utile lorsqu'il y a déjà un entrepôt en place (ex: acquisitions ou fusions de compagnies);
- Très complexe (synchronisation, parallélisme, concurrence, etc.) et faible performance.

**FIN DE SEANCE**

## Résumé

1. Faites resumé détaillé sur l'architecture de Data Warehouse
2. Que-ce que vous avez compris des tables de dimensions et de faits ?
3. Faites un résumé sur les types d'un Data Warehouse