École Nationale des Sciences de l'Informatique

DATA WAREHOUSE

Cours ILSI II3 (30 h)

Enseignante: Rahma AMRI

2014 - 2015

Evaluation

Formules:

- M = 65% examen + 35% CC
- CC = f(assiduité, participation, projet, test, …)

Objectifs de module

- **Comprendre et assimiler les nouveaux besoins liés à l'apparition du data warehouse (DW)**
- **Maîtriser les notions de bases liés au DW**
- **Être capable de modéliser et concevoir un DW**
- **Manipuler des outils utilisés pour le data warehousing**

Introduction

Systèmes de gestion de données:

1. Systèmes transactionnels
2. Systèmes décisionnels

Introduction:
Systèmes transactionnels: les transactions

- Assurer l'exécution « correcte » d'un ensemble d'opérations sur des bases de données.
- Exemple :

```
begin-transaction
    read(account1, v1) -- opération 1
    read(account2, v2) -- opération 2
    v1 ← v1 – 1
    v2 ← v2 + 1
    write(account1, v1) -- opération 3
    write(account2, v2) -- opération 4
end-transaction
```

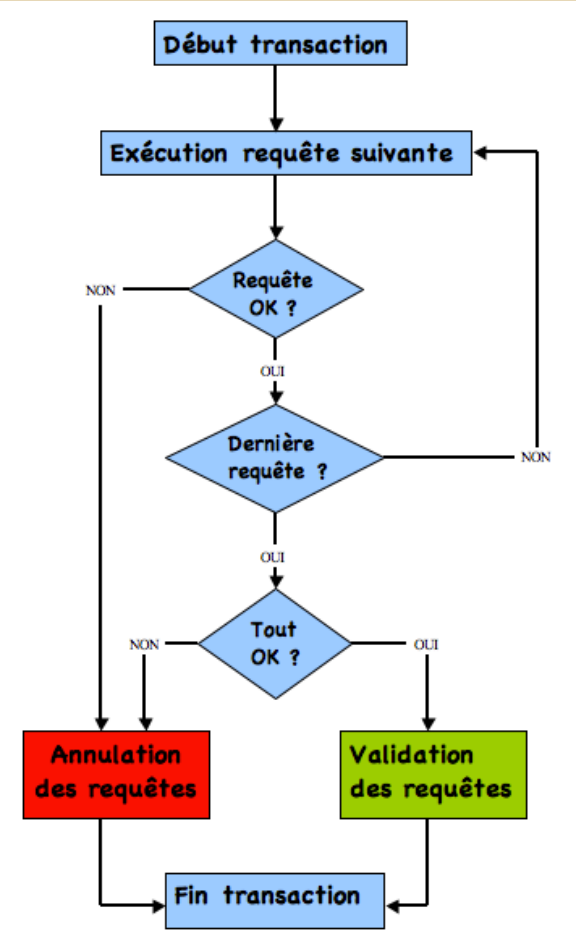
- Deux résultats sont possibles pour une transaction:

  Commit (valider), si la transaction s'exécute avec succès

  Abort (avorter), sinon

Introduction:

Systèmes transactionnels: les transactions



Introduction:
Systèmes transactionnels: les transactions

Une transaction est formellement définie par les propriétés ACID:

- **A: Atomicité**

soit toutes les opérations de la transaction sont exécutées, soit aucune.

- **C: Cohérence**

la transaction transforme un état cohérent de la base de données en un nouvel état cohérent.

- **I: Isolation**

La transaction va travailler dans un mode isolé où elle seule peut voir les données qu'elle est en train de modifier, cela en attente d'un nouveau point de synchronisation.

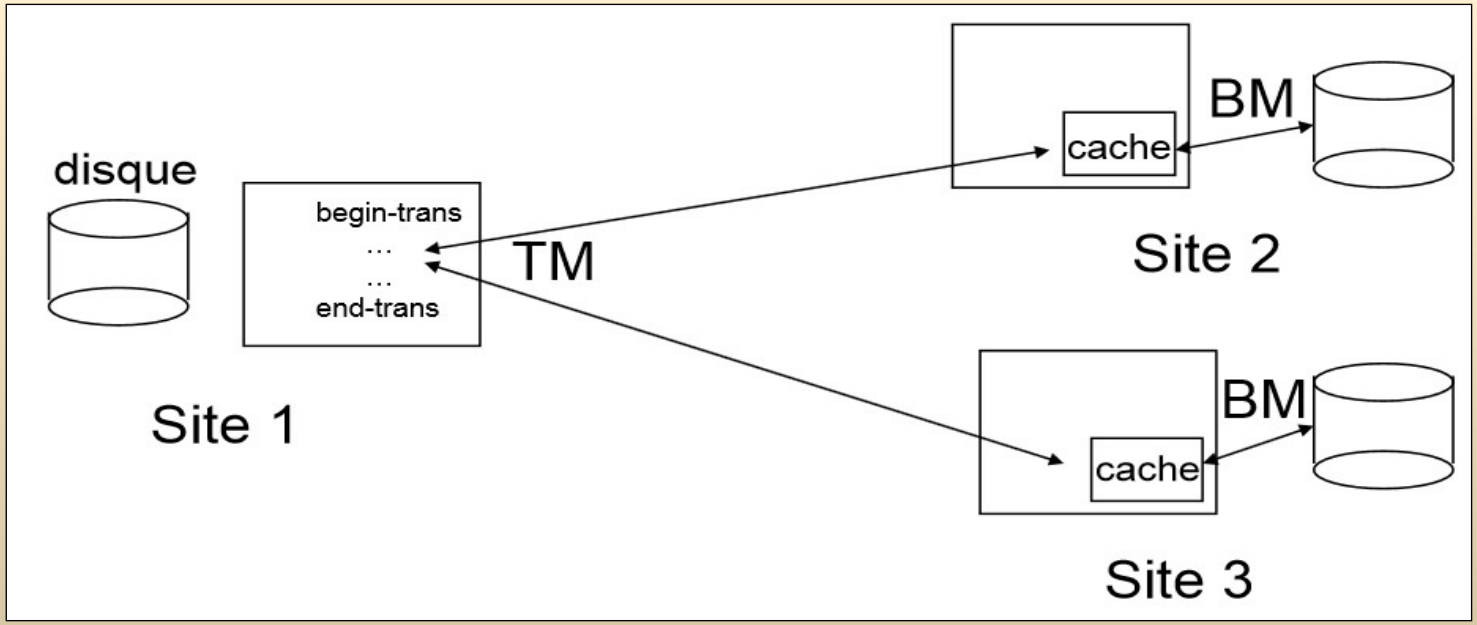
- **D: Durabilité**

Lorsque la transaction est achevée, le système est dans un état stable durable.

Introduction:
Systèmes transactionnels

Les composants d'un système transactionnel :

- Transaction Manager (TM)
- Buffer manager (BM) ou Data Manager (DM)
- Scheduler
- Local Recovery Manager (LRM)



Introduction:
Systèmes transactionnels

- Système transactionnel: un environnement informatique à travers lequel des transactions sont réalisées quasi-instantanément, impliquant une ou plusieurs applications ou acteurs internes et/ou externes à l'entreprise (départements, clients, partenaires, etc.), chaque transaction engendrant un certain nombre d'invocations et de mises à jour de base de données.

- On parle de système transactionnel dans les contextes où des transactions (financières, administratives, commerciales, etc) entre deux ou plusieurs parties doivent être traitées en temps réel, et où il est impératif que chaque transaction entraîne toutes les mises à jour de base de données nécessaires.

- Exemple: dans le cas d'un service automatisé de réservation de billets d'avion, il est indispensable que la réservation d'un client effectuée sur site Internet provoque tous les enregistrements utiles.

Introduction:
Systèmes transactionnels

- Dans ces applications transactionnelles, le système doit être capable de traiter correctement de nombreuses demandes simultanées.

- Un système transactionnel doit mettre en œuvre des stratégies particulièrement poussées de traitement des pointes de charge, des incidents possibles, des demandes simultanées d'écriture, tout en prenant toutes les mesures possibles pour préserver à tout moment la cohérence de la base de données.

- Le système transactionnel doit également mettre en œuvre les meilleures solutions possibles en cas de désastre informatique, avec en tout premier lieu des procédures prédéfinies de rollback, c'est-à-dire de retour en arrière, à un état connu et cohérent de la base de données, avec annulation des dernières transactions.

Introduction:
Systèmes transactionnels

- Systèmes transactionnels = OLTP (On-Line Transaction Processing)
- Opérations dans les systèmes transactionnels :
  - Ajout
  - Suppression
  - Mise à jour
  - Requêtes simples
  - Interrogations et modifications fréquentes des données par de nombreux utilisateurs
- Le système transactionnel est développé pour gérer les transactions quotidiennes
- Ces bases de données supportent habituellement des applications particulières telles que les inventaires de magasins, les réservations d'hôtel, etc.
- Le contenu est fait de données actuelles, pas d'archives
- Les données sont très détaillées
- Très souvent plusieurs de ces systèmes existent indépendamment les uns des autres

Introduction:
Naissance de nouveaux besoins

- Transformer un système d'information (SI) qui avait une vocation de production en un SI décisionnel
- Transformation des données de production en informations stratégiques

- Exemple de requêtes décisionnelles :
  - Catégorie socioprofessionnelle des meilleurs clients de chaque région?
  - Evolution de la part de marché d'un produit particulier?
  - Quel est le profil des employés les plus performants?

- Gestion et visualisation des données doit être rapide et intuitive
- Visualisation multidimensionnelle des données
- Nécessaire de retrouver et d'analyser rapidement les données provenant de diverses sources

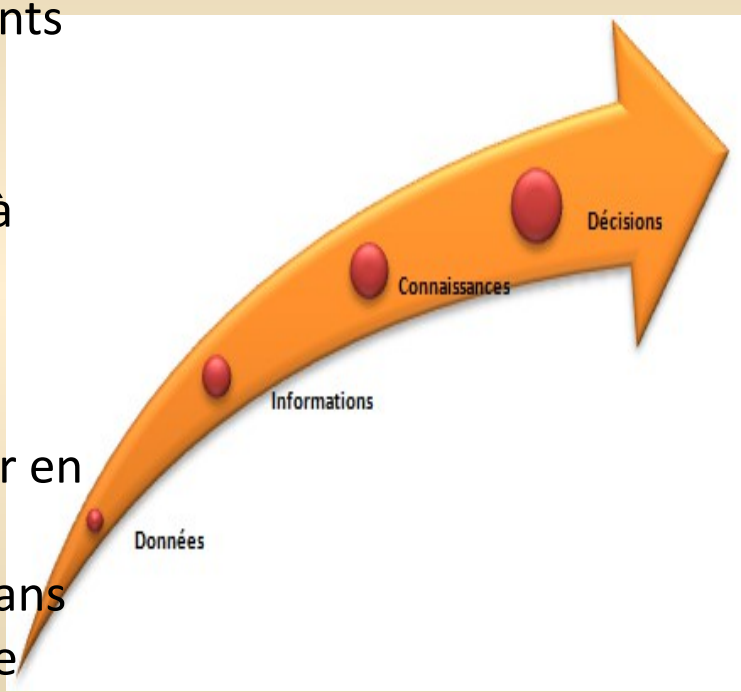
Introduction:
Naissance de nouveaux besoins

Données : Informations brutes présentées sous forme conventionnelle, en vue d'être traitée (Points de ventes, géographiques, démographiques, …)

Informations : Une information est une donnée à laquelle un sens et une interprétation ont été donnés. (I vit dans R, I est âgé de A, …)

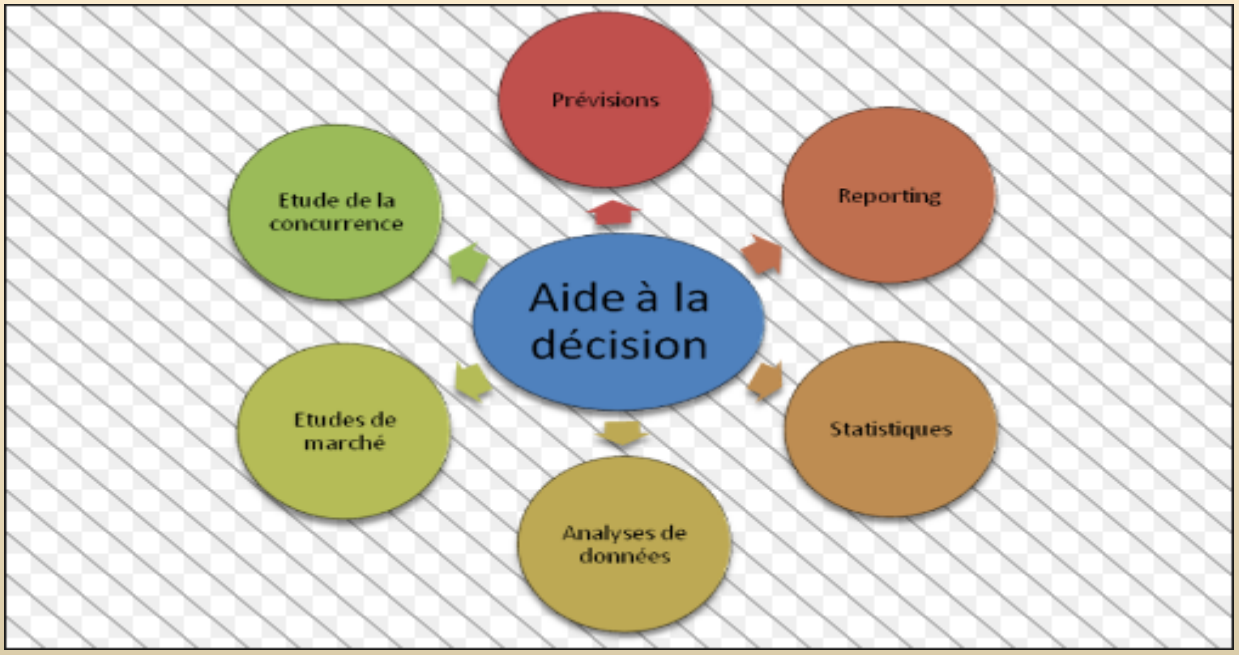
Connaissances : Règles utilisant les données pour en déduire d'autres, La connaissance est le résultat d'une réflexion sur les informations analysées (Dans X%, le produit Y est vendu en même temps que le produit Z, …)

Décisions : Lancer la promotion de Y & Z dans R auprès des clients plus âgés que A, ...



Introduction:
Les systèmes décisionnels

Le terme décisionnel « Business Intelligence » couvre l'ensemble des technologies permettant en bout de chaîne d'apporter une aide à la décision.



Introduction:
Les systèmes décisionnels

Les systèmes décisionnels (On-Line Analytical Processing= OLAP)sont:

- SI capable d'agréger les données internes ou externes et de les transformer en information servant à une prise de décision rapide.

- SI capable de répondre à certains types de questions:
  - Quelles sont les ventes du produit X pendant le trimestre A de l'année B dans la région C ?
  - Comment se comporte le produit X par rapport au produit Y?
  - Quel type de client peut acheter le produit X?
  - Est-ce qu'une baisse de prix de 10% par rapport à la concurrence ferait redémarrer les ventes du produit X ?

Introduction:
Les systèmes décisionnels

Ces exemples mettent en évidence les faits suivants:

- Les questions doivent pouvoir être formulées dans le langage de l'utilisateur en fonction de son secteur d'activité: Service marketing, Service économique, service relation clients…
- La prévision des interrogations est difficile car elles sont du ressort de l'utilisateur.
- Les questions vont varier selon les réponses obtenues: Si le produit X s'est vendu moins bien que l'année précédente, il va être utile de comprendre les raisons: Détailler les ventes par région, par type de magasin,…
- Des questions ouvertes vont nécessiter la mise en place de méthodes d'extraction d'informations

Introduction:
Les systèmes décisionnels

Les domaines d'application du décisionnel:

- La gestion de la relation client (CRM) est l'un des premiers champs d'application de la Business Intelligence.
- Le contrôle de gestion pour l'analyse des coûts, l'analyse de la rentabilité, l'élaboration budgétaire, les indicateurs de performance…
- La direction marketing pour le ciblage, le pilotage de gamme, les applications de géomarketing, de fidélisation clients…
- La direction commerciale pour le pilotage des réseaux, les prévisions des ventes, l'optimisation des territoires…
- Les ressources humaines pour la gestion des carrières,
- La direction de la production pour l'analyse qualité, la prévision des stocks, la gestion des flux, la fiabilité industrielle…
- La direction générale pour les tableaux de bord, indicateurs de pilotage, gestion d'alertes…

Introduction:
Les systèmes décisionnels

Applications transactionnelles v.s Applications décisionnelles

Décisionnel	Transactionnel
Gros volumes de données à gérer.	Petits volumes de données à gérer.
Nombre d'utilisateur restreint (décideurs, analystes).	Utilisé par toute l'entreprise.
Processus ouverts pour permettre la génération de connaissance.	Processus fermés, transactionnels, le but est de donner le moins de marge de manœuvre possible.
Données en lecture seule.	Données en lecture - Écriture.
Rapidité moyenne comparée aux systèmes opérationnels.	Réponses très rapides.
Niveau de granularité très grand (on peut avoir des résumés sur ce qui c'est passé durant les 10 dernières années par exemple).	Niveau de granularité fin.
Centralisés (on veut avoir toutes les données de l'entreprise dans une seule structure).	Décentralisés.

Introduction:
Les systèmes décisionnels

- Le support efficace d'une activité OLAP nécessite la constitution d'un système d'information propre: Le Data warehouse

- Data warehouse :
  - Collection de données orientées sujets, intégrées, non volatiles et historisées, organisées pour le support du processus d'aide à la décision.
  - Base de données dans laquelle sont déposées après nettoyage et homogénéisation les informations en provenance des différents systèmes de production de l'entreprise OLTP.

Concepts fondamentaux:
Data warehouse (Entrepôt de données)

Qu'est ce qu'un data warehouse?

- Collection de données orientées sujets, intégrées, non volatiles et historisées, organisées pour le support du processus d'aide à la décision.
- Une base de données dans laquelle sont déposées après nettoyage et homogénéisation les informations en provenance des différents systèmes de production de l'entreprise (OLTP).
- Ensemble de données :
  - destinées aux décideurs
  - souvent une « copie » des données de production
  - avec une valeur ajoutée (agrégation, historique)
  - intégrées
  - historisées
- Ensemble d'outils permettant :
  - de regrouper les données
  - de nettoyer, d'intégrer les données, ...
  - de faire des requêtes, rapports, analyses
  - de faire du data mining
  - faire l'administration du data warehouse

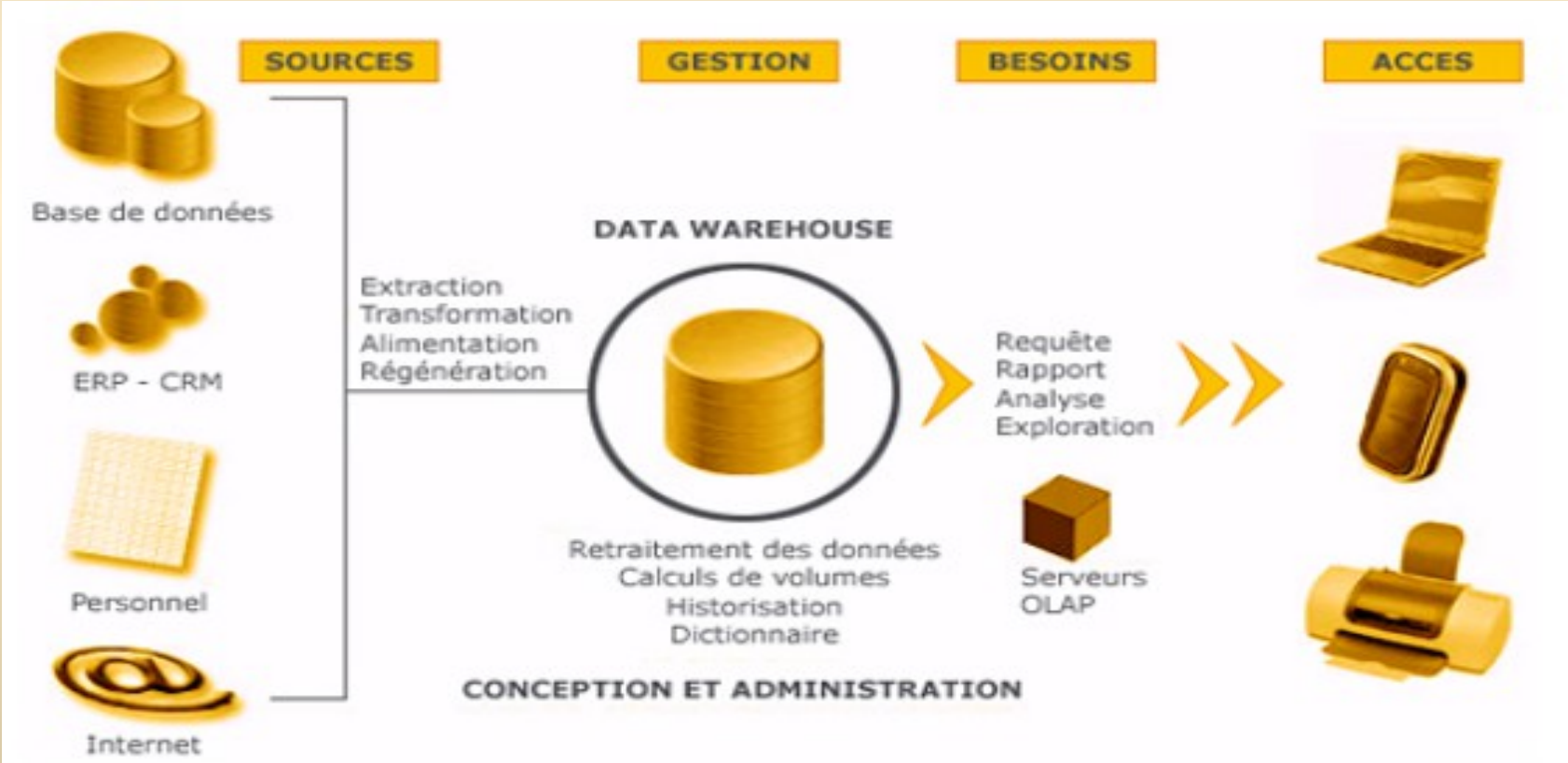
Concepts fondamentaux:
Data warehouse (Entrepôt de données)

- Fonctions d'un Data Warehouse :
  - Récupérer des données existantes dans différentes sources
  - Stocker les données (historisées)
  - Mettre à disposition les données pour: Interrogation, Visualisation et Analyse
- Un entrepôt de données, ou data Warehouse:
  - Est une vision centralisée et universelle de toutes les informations de l'entreprise.
  - Il a pour but de regrouper les données de l'entreprise pour des fins analytiques et pour aider à la décision stratégique.
  - La décision stratégique étant une action entreprise par les décideurs de l'entreprise et qui vise à améliorer, quantitativement ou qualitativement, la performance de l'entreprise.

  => En gros, c'est un gigantesque tas d'informations épurées, organisées, historisées et provenant de plusieurs sources de données, servant aux analyses et à l'aide à la décision.

Concepts fondamentaux:
Data warehouse (Entrepôt de données)

L'entrepôt de données est l'élément central de l'informatique décisionnelle.



architecture d'un environnement décisionnel

Concepts fondamentaux:
Caractéristiques des données d'un DW

- **Orientées sujet :**
  - Organisées autour de sujets majeurs de l'entreprise
  - Données pour l'analyse et la modélisation en vue de l'aide à la décision, et non pas pour
  - les opérations et transactions journalières
  - Vue synthétique des données selon les sujets intéressant les décideurs

- **Intégrées :**
  - Construit en intégrant des sources de données multiples et hétérogènes: BD
  - relationnelles, fichiers, enregistrements de transactions
  - Les données doivent être mises en forme et unifiées afin d'avoir un état cohérent
  - Phase la plus complexe (60 à 90 % de la charge totale d'un projet DW)

Concepts fondamentaux:
Caractéristiques des données d'un DW

- **Orientées sujet**
- **Intégrées**
- **Historisées :**
  - Fournies par les sources opérationnelles
  - Matière première pour l'analyse
  - Stockage de l'historique des données, pas de mise à jour
  - Un référentiel temps doit être associé aux données

- **Non volatiles :**
  - Conséquence de l'historisation
  - Une même requête effectuée à intervalle de temps, en précisant la date
  - Référence de l'information donnera le même résultat
  - Stockage indépendant des BD opérationnelles
  - Pas de mises à jour des données dans le DW

Concepts fondamentaux:
Exemple de DW: 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

Concepts fondamentaux:
Data mart (magasin de données)

- Les Data Warehouses étant, en général, très volumineux et très complexes à concevoir, on a décidé de les diviser en bouchées plus faciles à créer et entretenir. Ce sont les Data Marts.
- On peut faire des divisions:
  - Par fonction (un data mart pour les ventes, pour les commandes, pour les ressources humaines)
  - Par sous-ensemble organisationnel (un data mart par branche: pour le service marketing, pour le serv).
- Définition: « C'est un sous-ensemble de données dérivées du DW ciblé sur un sujet unique».

Concepts fondamentaux:
Data mart (magasin de données)

Caractéristiques des data marts:

- Orienté vers un sujet unique, exemple: comportement de la clientèle
- Données fortement agrégées: Le DW joue le rôle de source et d'historique pour le Data mart
- Organisation multidimensionnelle (cubique), dont l'une des dimensions indique souvent le temps
- Lien dynamique avec le DW: Association entre valeur agrégée et valeur détaillée
- Interfaces simples et conviviales

Concepts fondamentaux:
Les dimensions

- En BDD on parle en termes de tables et de relations, une table étant une représentation d'une entité et une relation une technique pour lier ces entités.
- En BI, on parle en termes de Dimension et de Faits. C'est une autre approche des données.
- Les dimensions sont les axes avec lesquels on veut faire l'analyse.
- Exemple: dimension client, dimension produit, dimension géographie (pour faire des analyses par secteur géographique), etc.

=> Une dimension est tout ce qu'on utilisera pour faire nos analyses.

=> Il existe toujours une dimension temps

Concepts fondamentaux:
Les faits

- Les faits, en complément aux dimensions, sont ce sur quoi va porter l'analyse.
- Ce sont des tables qui contiennent des informations opérationnelles et qui relatent la vie de l'entreprise.
- Exemple:
  - des faits pour les ventes (chiffre d'affaire net, quantités et montants commandés, quantités facturées, quantités retournées, volumes des ventes, etc.)
  - Des faits pour les stocks (nombre d'exemplaires d'un produit en stock, niveau de remplissage du stock, taux de roulement d'une zone, etc.), ou peut être sur les ressources humaines (performances des employés, nombre de demandes de congés, nombre de démissions, taux de roulement des employés, etc.).

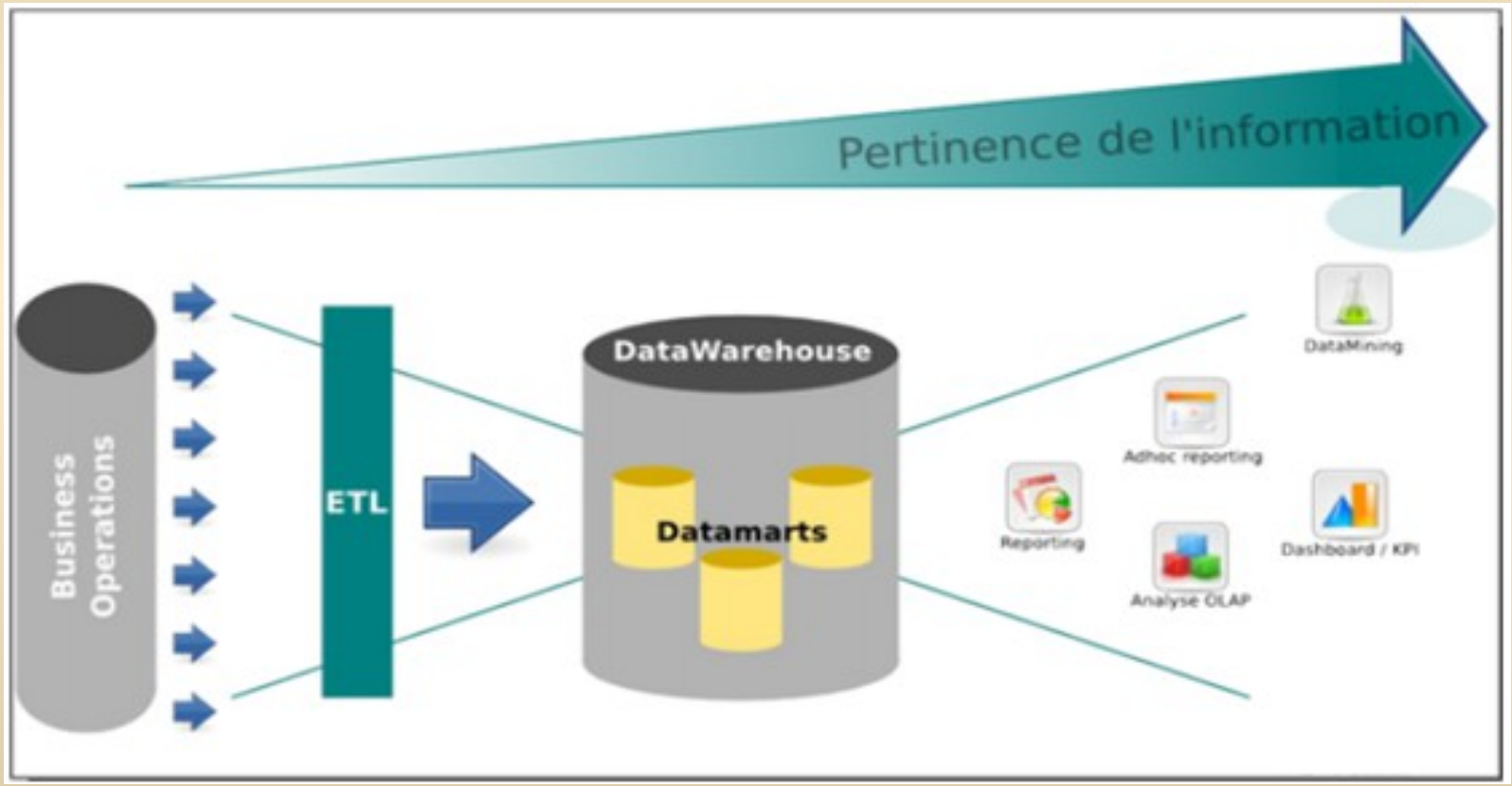
=> Un fait est tout ce qu'on voudra analyser.

Concepts fondamentaux:
ETL (ou ETC)

- L’ ETL : *Extraction, Transformation, Loading* est un système par lequel vont passer toutes les données des systèmes opérationnels avant d'arriver dans la forme souhaitée dans l'entrepôt.
- Une sorte de moulinette par laquelle vous ferez passer toutes les données de votre entreprise. Les données en sortie (passées à la moulinette) seront nettoyées, purifiées (les gros morceaux seront mis de coté), contextualiseés (les données des différents systèmes s'homogénéiseront) et prêts à être reçus dans l'entrepôt.
- Le système d'ETL est la partie la plus importante d'un projet décisionnel. Car c'est avec l'ETL que les systèmes seront mis en relation, les erreurs détectées, les calculs complexes effectués, etc. On peut dire que la solidité d'un ETL détermine la viabilité du projet.
- L'ETL sert à transposer le modèle entité-relation des bases de données de production ainsi que les autres modèles utilisés dans les opérations de l'entreprise, en modèle à base de dimensions et de faits.

Concepts fondamentaux

Schéma récapitulatif



Modélisation et conception de DW

Modélisation en étoile / flocon

- Une étoile est une façon de mettre en relation les dimensions et les faits dans un entrepôt de données.
- Le principe est que les dimensions sont directement reliées à un fait (schématiquement, ça fait comme une étoile).
- Un flocon est un autre modèle de mise en relation des dimensions et des faits dans un entrepôt de données. Le principe étant qu'il peut exister des hiérarchies de dimensions et qu'elles sont reliées au faits, ça fait comme un flocon.
- Les flocons et les étoiles peuvent être vus comme une manière de diviser les entrepôts de données et les magasins de données.
- On peut les voir comme l'atome de l'informatique décisionnelle : le plus petit élément avec lequel ont peut faire des analyses et avec lequel ont peut faire des magasins de données qui, mis ensemble, forment un entrepôt de données.

Modélisation et conception de DW
Un cas d'étude

- On vous demande de créer un data mart pour l'analyse de l'activité des représentants d'une entreprise de vente d'imprimantes.
- Le chef d'entreprise veut savoir:
  - Ce qui se passe pour ses vendeurs,
  - Les employés font ils leur travail,
  - Quelle est la zone de couverture des vendeurs,
  - Où sont les endroits où les vendeurs sont le moins efficaces,
  - Quelle est la moyenne de ventes des représentants,
  - Etc.
- L'entreprise possède:
  - Un système de gestion de ressources humaines,
  - Un système de gestion des ventes et des feuilles de routes avec des informations concernant les vendeurs : kilomètres parcourus, litres d'essence utilisée, frais de voyage, ventes, promesses de ventes, etc.