Partie 1

Data Warehouse

Entrepôts de données

Chapitre 1 : Introduction et architectures des entrepôts de données

Dr H. EL BOUHISSI Epse BRAHAMI

Plan

1. Introduction et contexte : Besoins décisionnels.
2. Challenges.
3. Définition d’un Datawarehouse.
4. Caractéristiques d’un DWH.
5. Domaines d’application d’un DWH.
6. Processus de conception d’un DWH
7. Types de DWH.

Introduction

La majeure partie des applications Bases de Données reposent aujourd'hui sur trois couches :

- La couche la plus externe : permet de présenter les données aux utilisateurs. (appelée *Graphical User Interfaces GUI)*.
- La couche application (programme de l'application) : ne stocke pas les données.
- La couche interne : Base de Données.

Les applications interrogent les données avec, par exemple le langage SQL (Select) et les mettent à jour par l'intermédiaire des opérations Insert, Update et Delete qui constituent des transactions.

Ce type d'application est appelé On-Line Transaction Processing (OLTP)

Contexte : Besoins décisionnels

- **Besoin** : Prise de décisions stratégiques et tactiques, Réactivité.
- **Qui** : Les décideurs (non informaticiens, non statisticiens)
- **Comment** : Répondre aux demandes d'analyse de données et dégager des informations qualitatives nouvelles

Challenges

Comment répondre aux besoins des décideurs afin d'améliorer les performances décisionnelles de l'entreprise?

- ⇨ En donnant un accès rapide et simple à l'information stratégique.
- ⇨ En donnant du sens aux données.
- ⇨ En donnant une vision transversale des données de l'entreprise (intégration de différentes bases de données).
- ⇨ En extrayant, groupant, organisant, corrélant et transformant (résumé, agrégation) les données.

***Solution** : Mettre en place un SI dédié aux applications décisionnelles : un entrepôt de données (datawarehouse)*

Transformer des données de production en informations stratégiques

Datawarehouse : Définition

Dépôt centralisé d'intégration de données d'une ou plusieurs sources disparates

Base de données relationnelle hébergée sur un serveur dans un Data Center ou dans le Cloud dans laquelle sont déposées après nettoyage et homogénéisation les informations en provenance des différents systèmes de production de l'entreprise OLTP. Il recueille des données de sources variées et hétérogènes dans le but principal de soutenir l'analyse et faciliter le processus de prise de décision.

L'objectif du DWH est de permettre des requêtes sur de grands ensembles de données, la plupart du temps sous forme d'agrégats (GROUP BY) afin d'en obtenir une vision synthétique (propre à la prise de décision).

Datawarehouse : Définition

« Collection de données orientées sujets, intégrées, non volatiles et historisées, organisées pour le support du processus d'aide à la décision » **(Inmon, 1991)**

Base de Données utilisée à des fins d'analyse :

- ⇨ Regrouper les données sources.
- ⇨ Concevoir le schéma de l'entrepôt.
- ⇨ Remplir l'entrepôt.
- ⇨ Maintenir l'entrepôt.

Datawarehouse : Types

Entreprise warehouse : EW

Collecte de toutes les informations concernant les sujets traités au niveau de l'organisation.

Data Mart

Un sous ensemble d'un entreprise warehouse. Il est spécifique à un groupe d'utilisateurs (ex: data mart du marketing)

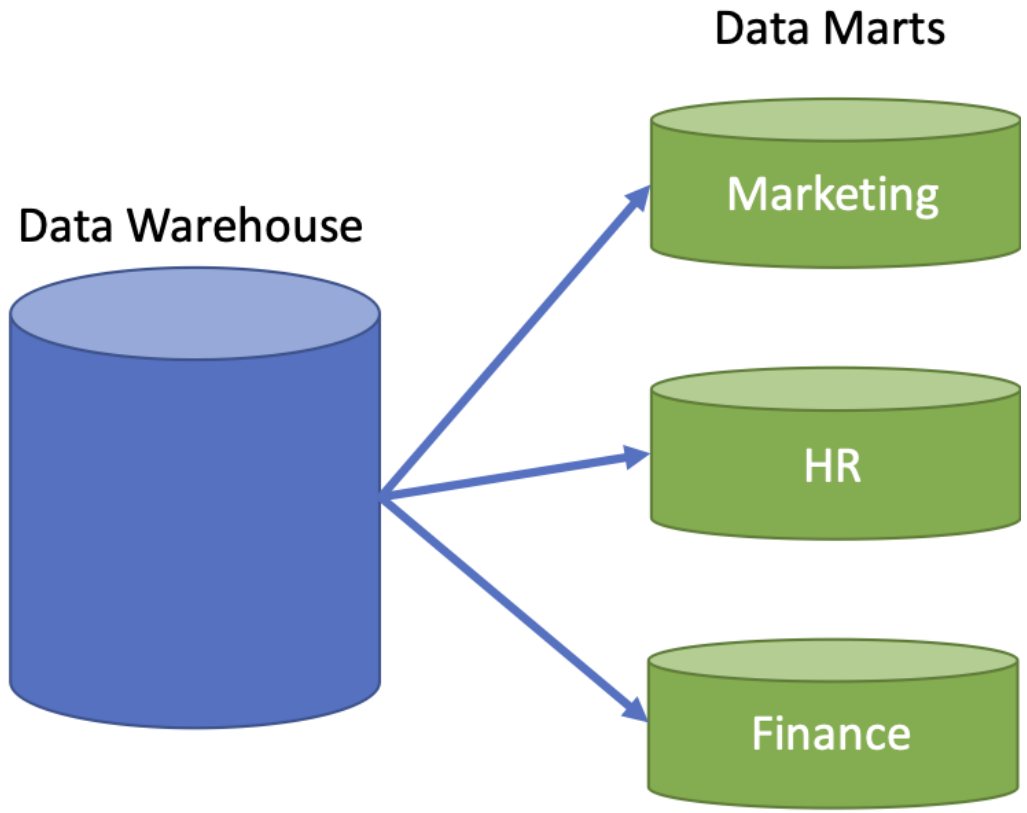
Data warehouse virtuel

Un ensemble de vues définies à partir de la base opérationnelle.
Seulement un sous ensemble des vues sont matérialisées.

Datawarehouse : Data mart

Sous-ensemble d'un entrepôt de données

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



Pourquoi pas un SGBD ?

Fonctions d'un SGBD :

- Systèmes transactionnels (OLTP : Online Transactional Processing).
- Permettre d'insérer, modifier, interroger rapidement, efficacement et en sécurité les données de la base.
- Sélectionner, ajouter, mettre à jour, supprimer des tuples.
- Répondre à de nombreux utilisateurs simultanément.

Pourquoi pas un SGBD ?

Fonctions d'un DW :

- Systèmes pour l'aide à la prise de décision (OLAP).
- Regrouper, organiser des informations provenant de sources diverses.
- Intégrer et stocker les données pour une vue orientée métier.
- Retrouver et analyser l'information rapidement et facilement

Pourquoi pas un SGBD ?

	OLTP	DW
Utilisateurs	Nombreux Employés	Peu Analystes
Données	Alphanumériques Détaillées / atomiques Orientées application Dynamiques	Numériques Résumées / agrégées Orientées sujet Statiques
Requêtes	Prédéfinies	« one-use »
Accès	Peu de données (courantes)	Beaucoup d'informations (historisées)
But	Dépend de l'application	Prise de décision
Temps d'exécution	Court	Long
Mises à jour	Très souvent	Périodiquement

Caractéristiques des données d'un DWH

❶ 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.

❷ 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).

Caractéristiques des données d'un DWH

❸ 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.

❹ 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.

Datawarehouse : Domaines d'application

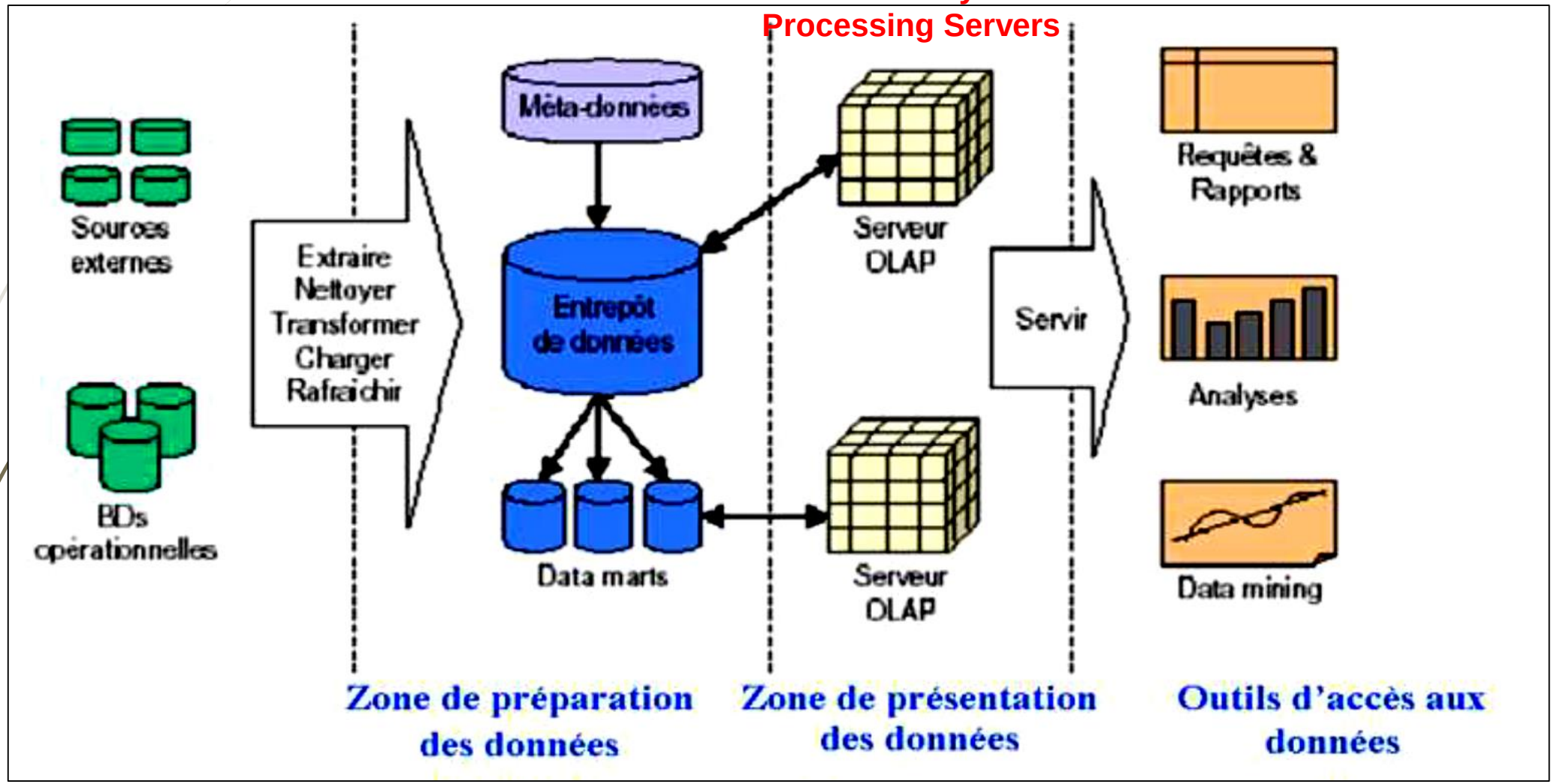
- L'informatique décisionnelle (*Business Intelligence*) pour aider atteindre les objectifs stratégiques d'une entreprise et faciliter son pilotage.
- Avoir une connaissance plus approfondie de l'entreprise.
- Anticiper les besoins clients.
- Prendre en compte les nouveaux canaux de distribution (vente en ligne, etc.)

Datawarehouse : Avantages

- Permettent de mener des analyses poussées sur différents sujets d'affaires.
- Fournissent une vue consolidée des données de l'entreprise (une seule vérité).
- Procurent de l'information de qualité, plus rapidement.
- Libèrent les ressources (ex: serveurs) dédiées au traitement des transactions des tâches d'analyse.
- Simplifient l'accès aux données.
- N'est pas un produit mais une solution : ne s'achète pas mais se construit.
- Offre une vue agrégée et permet de descendre aux données détaillées.

Inconvénient : DWH Pose le problème de performance à cause du grand volume de données et contient plus que nécessaire d'informations pour une classe de décideurs

Processus d'entreposage



Analyse OLAP (On-Line Analytical processing)

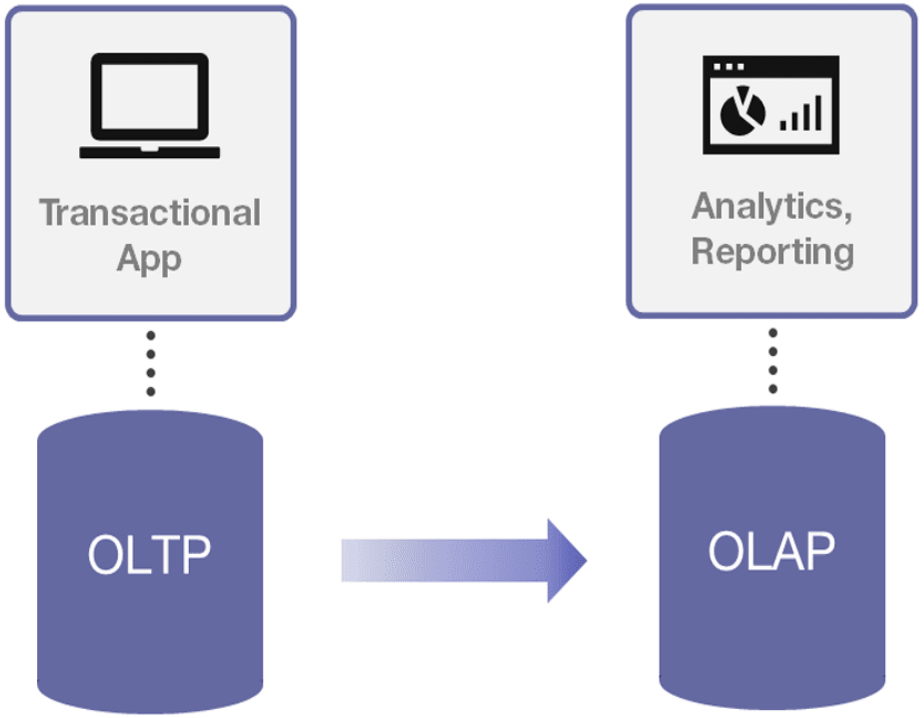
Une catégorie de logiciels axés sur l'exploration et l'analyse rapide des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation».

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.

OLAP VS OLTP

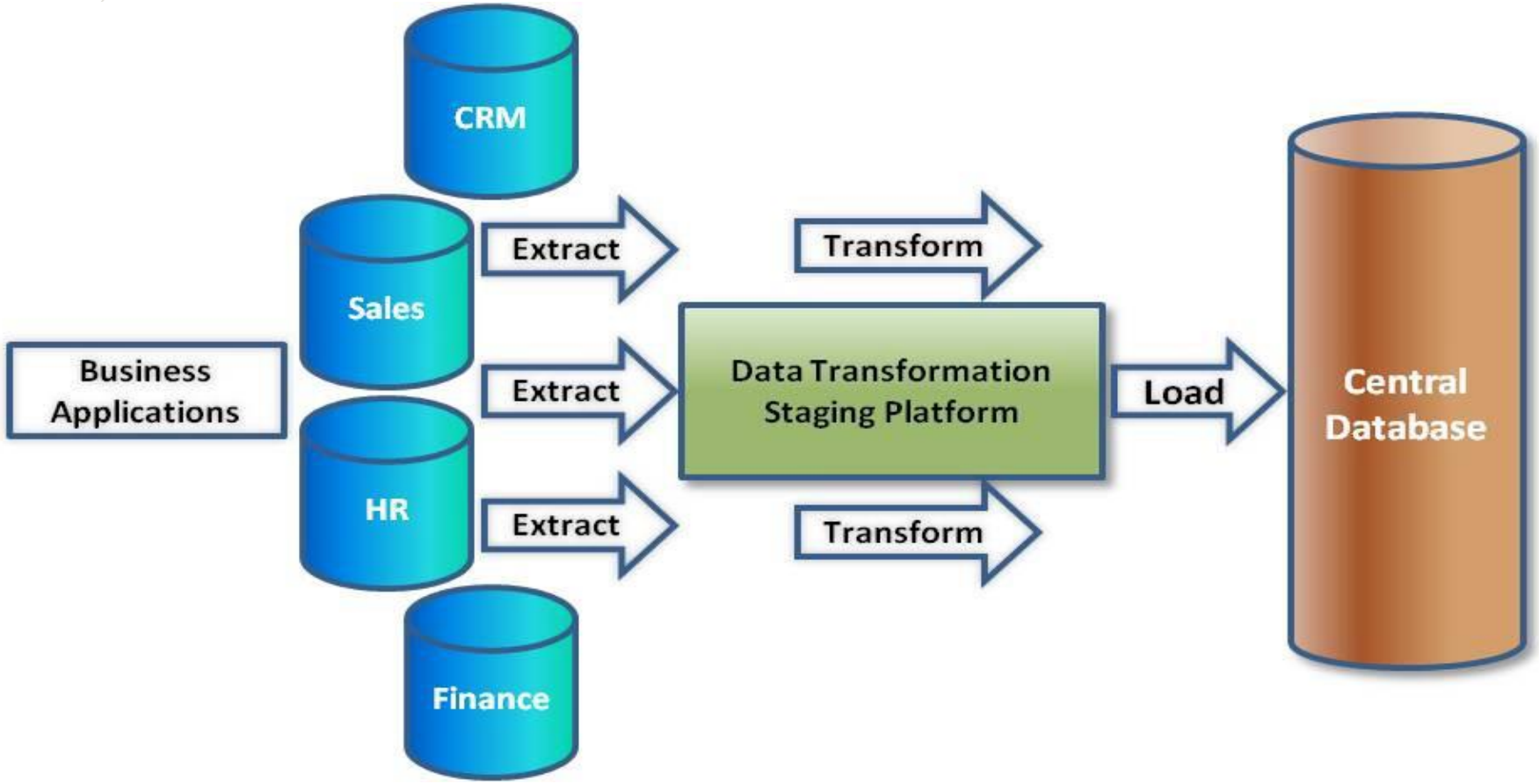
- High volume of transactions
- Fast processing
- Normalized data
- Many tables
- *"Who bought X?"*



- High volume of data
- Slow queries
- Denormalized data
- Fewer tables
- *"How many people bought X?"*

Alimentation d'un Datawarehouse

Extract, Transform, Load



Alimentation d'un Datawarehouse : Extraction

Sources de données variées

- Bases de données opérationnelles.
- Fichiers.
- Logs.
- Web…

Alimentation d'un Datawarehouse : Transformation

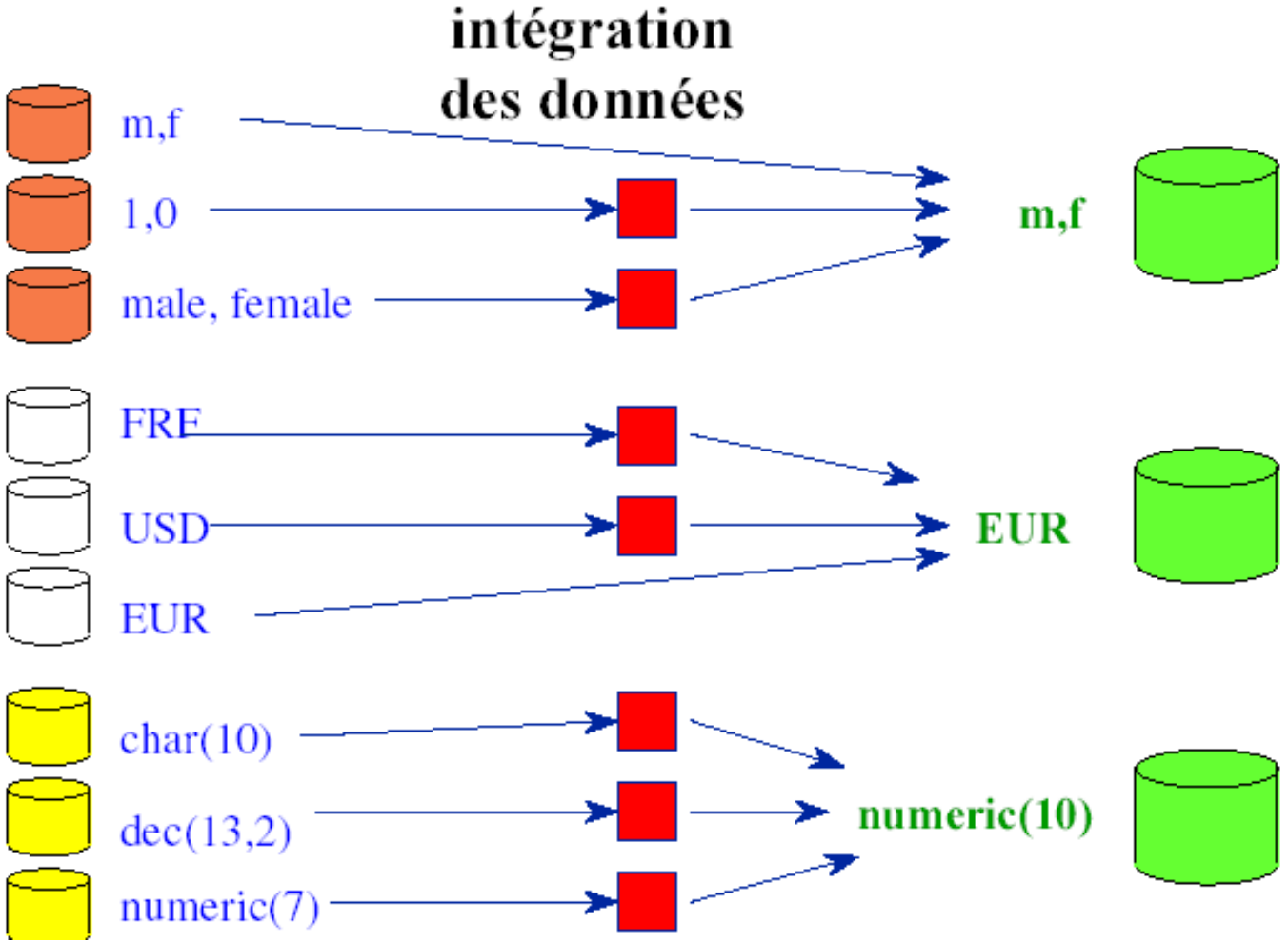
Unification des données

- Noms des attributs.
- Types (ex. précision numérique).
- Formats (ex. dates).
- Unités de mesure

Nettoyage des données

- Vérification des contraintes d'intégrité.
- Suppression des doublons.
- Traitement des valeurs manquantes.
- Détection des valeurs erronées ou incohérentes

Alimentation d'un Datawarehouse : Transformation



Alimentation d'un Datawarehouse : Load

Politiques de chargement

- Complet / incrémental.
- En ligne / hors ligne

Mises à jour des dimensions

- Écrasement de l'ancienne valeur.
- Versionnement.
- Traitement particulier des dimensions à évolution rapide

Rafraîchissement des index et vues matérialisées

Oubli des données anciennes

- Suppression.
- Agrégation

Datawarehouse : Deux grandes philosophies

Bill Inmon

EDW

Building the data warehouse that follows the top-down approach. In Inmon's philosophy, it is starting with building a big centralized enterprise data warehouse where all available data from transaction systems

Ralph Kimball

Data Mart

Building the data warehouse that follows the bottom-up approach. In Kimball's philosophy, it first starts with mission-critical data marts that serve analytic needs of departments

	Kimball	Inmon
Processus	Bottom-Up	Top-Down
Organisation	Datamarts	Datawarehouse
Schématisation	Etoile	Flocon

Datawarehouse : Processus

Que chacun construise ce qu'il veut, on intégrera ce qu'il faudra quand il faudra!

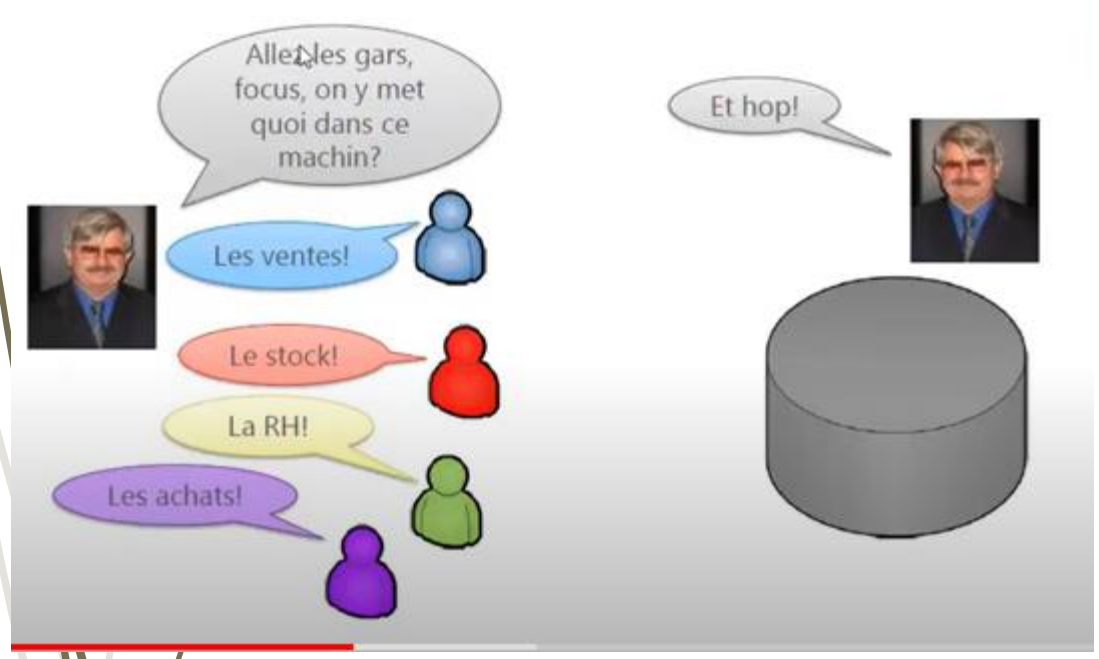
Ralph Kimball
Kimball Group
www.kimballgroup.com

On ne fait rien tant que tout n'est pas désigné, le datawarehouse doit être exhaustif!

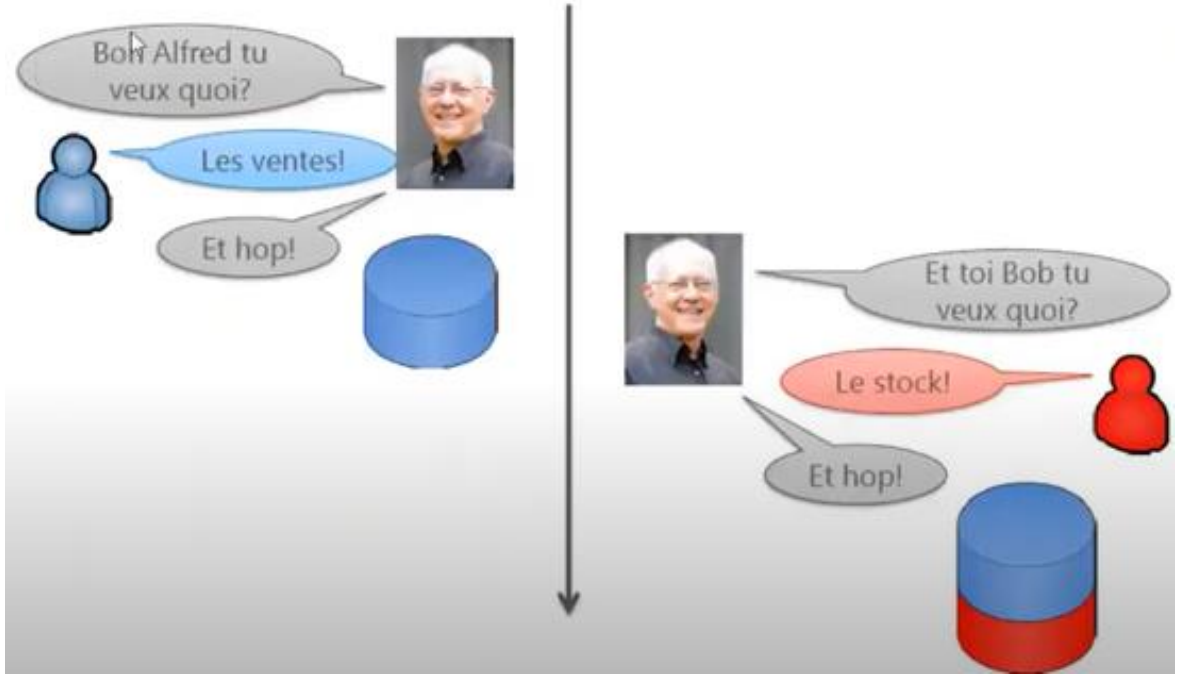
Bill Inmon
Corporate Information Factory
www.inmoncif.com

Datawarehouse : Processus

INMON : Top Down

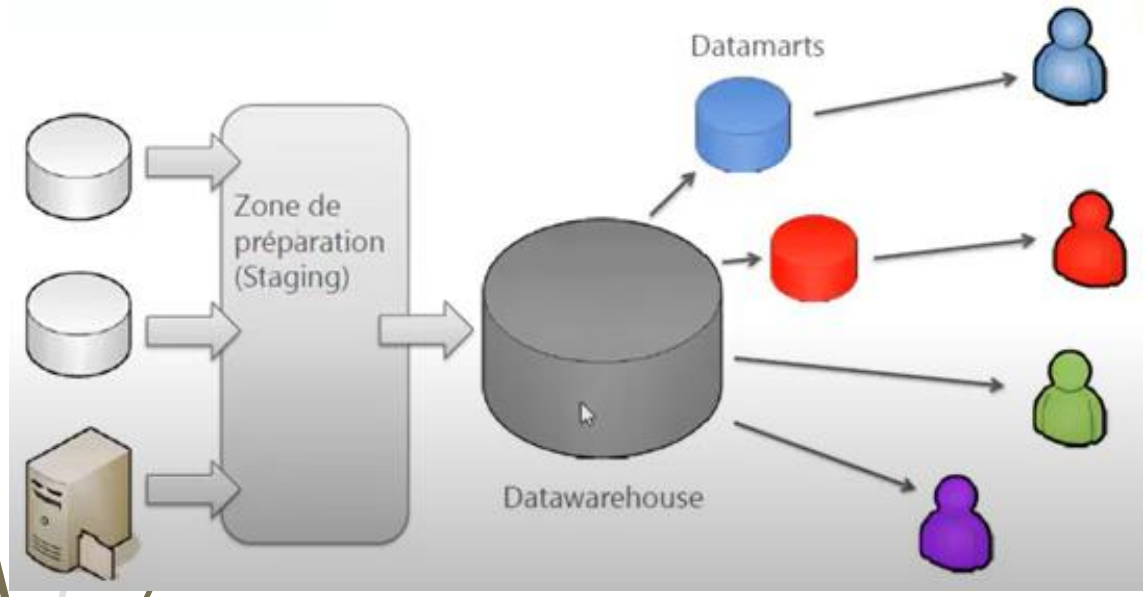


KIMBALL : Buttom up

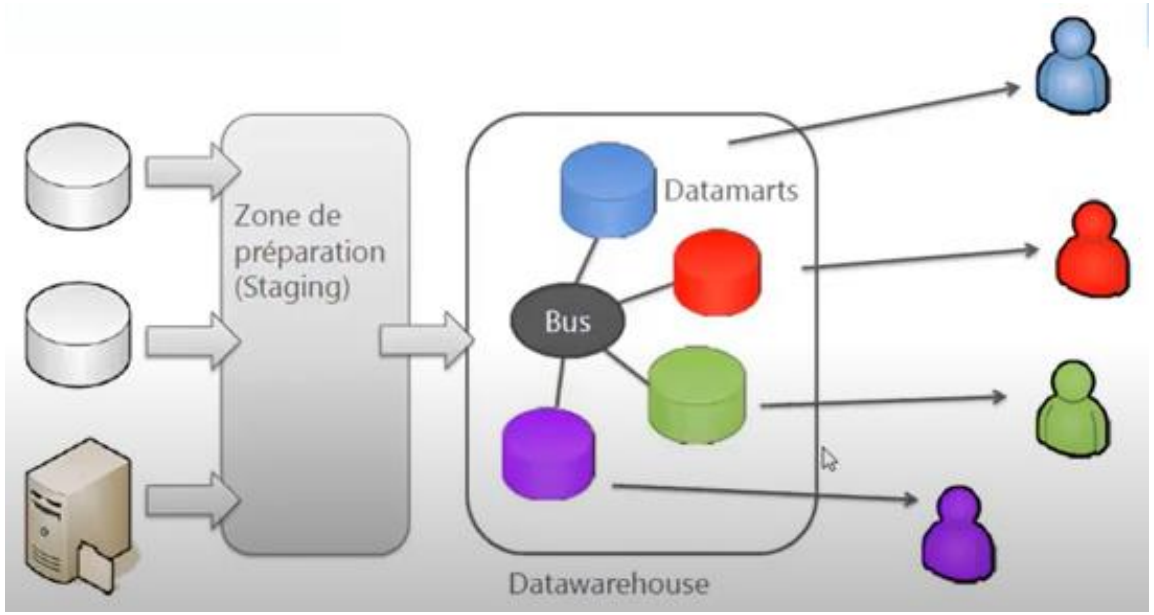


Datawarehouse : Processus

INMON : Organisation

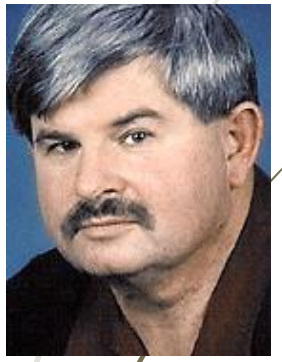


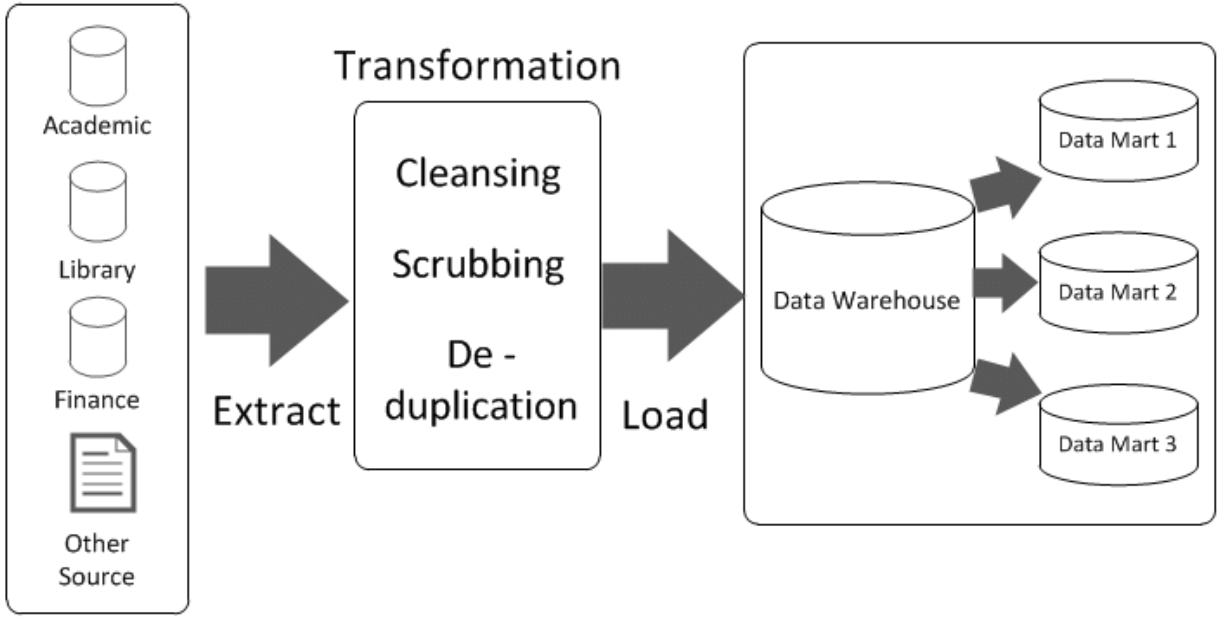
KIMBALL : Organisation



Datawarehouse Processus : Approche top-down

Conception intégrale de l'entrepôt à priori (Magasins de données (*datamarts*) extraits de l'entrepôt.





Avantages

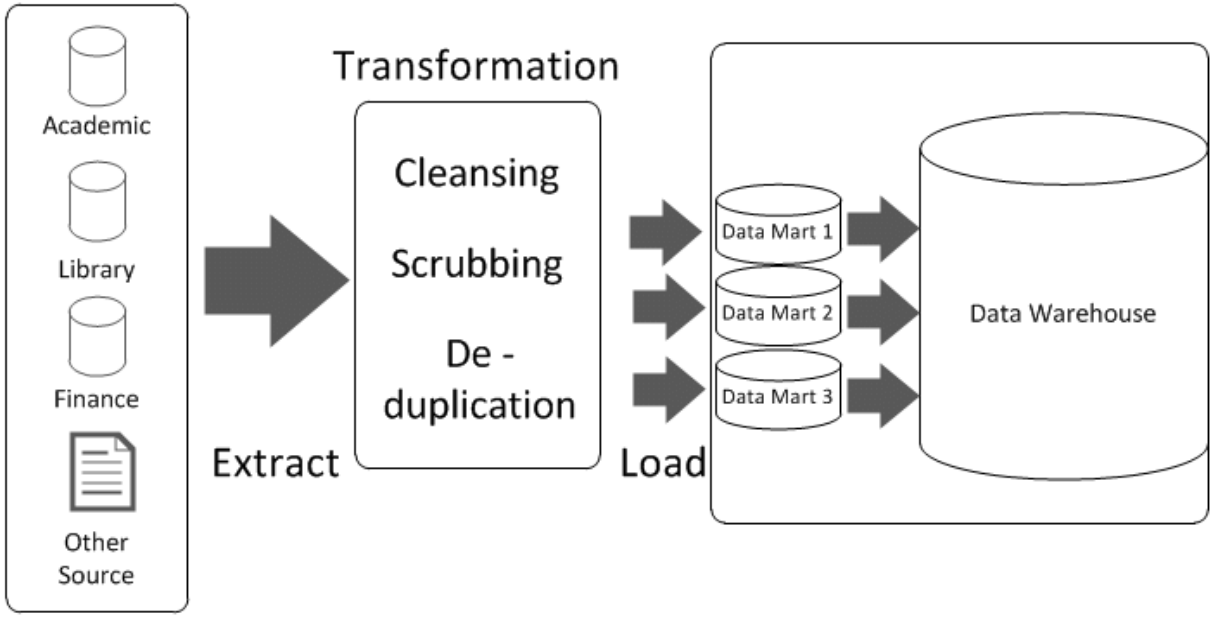
- Vision conceptuelle globale de l'entrepôt.
- Architecture intégrée.
- Normalisation des données, absence de redondance

Inconvénients

- Difficulté de mise en œuvre.
- Manque d'évolutivité

Datawarehouse Processus : Approche bottom-up

Construction incrémentale de l'entrepôt, l'entrepôt de données est une union de magasins de données.



Avantages

- Simplicité de mise en œuvre.
- Résultats rapides

Inconvénient

- Problèmes d'intégration des magasins de données.

Datawarehouse : Tools

ETL	Reporting
ODI	OBIEE/OAC
Data Stage	BI Publisher
Informatica	Cognos

Les programmes open source **Pentaho DI**, **Talend OS** et **Jasper ETL** sont idéaux pour l'acquisition de données et pour une intégration dans un processus ETL (extraction, transformation, chargement).

Les outils OLAP réputés sous licence open source sont Pentaho Mondrian et Jedox.

Des produits open source sont également disponibles pour les utilisateurs dans le domaine du data mining sous une licence open source : KNIME, RapidMiner et Weka.

Fin

Dr. H. EL BOUHISSI BRAHAMI