BIUM = Business Intelligence and User Modeling Master Data-Science

Laure Soulier - laure.soulier@lip6.fr
Benjamin Piwowarski - benjamin.piwowarski@lip6.fr

Sorbonne Université

25 janvier 2021

Contenu du cours

1. Business Intelligence (Laure Soulier et Amine Aouini)
   - Rôle de la gestion des données en entreprise
   - Agrégation, stockage des données à but décisionnel
2. Recommandation (Vincent Guigue)
3. Visualisation des données
   - Analyses factorielles, T-SNE, .. (Benjamin Piwowarski)
   - Dataiku (Malick)
4. User Modelling (Benjamin Piwowarski Laure Soulier)
   - Analyse des données utilisateur
   - Evaluation à partir des utilisateurs
   - Ouverture vers leur utilisation dans les entreprises

→ Intervenants industriels (en cours de discussion)

→ TME assurés par Agnès Mustar

Organisation

Horaires :

- Cours : Lundi de 13h30 à 15h30
- TME : Mardi de 8h30 à 12h45

Evaluation

- Contrôle continu (40%)
  - BI : Evaluation sur projet (TP + travail maison)
  - UM : Compte-rendus de TME
  - Evaluation continue (présentielle)
- Examen final (60%)

Evaluation contrôle continu

Projet - BI

- Mise en place d'une architecture complète de BI.
- Travail en équipe selon les méthodologies Agile
- Réalisation d'un démonstrateur sur une thématique donnée

Compte-rendu de TME

- Recommandation
- Visualisation

Objectifs

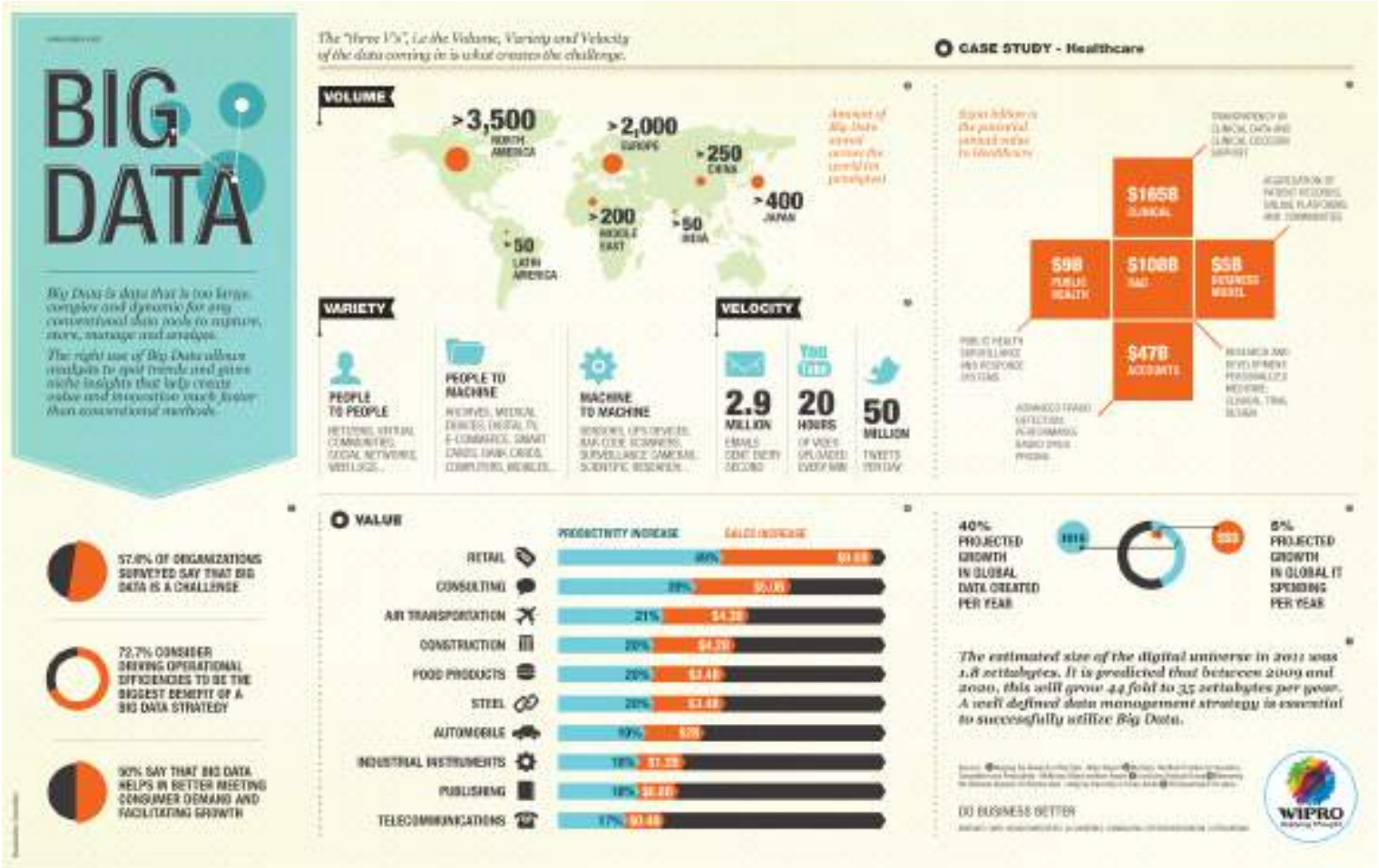
Etudiants ⇒ Data Integrator ⇒ Data Analyst ⇒ Data scientists

- Donner des clefs de compréhension autour du rôle et de la gestion des données en entreprise
- Aborder des problématiques de traitement/intégration de données sur des exemples concrets
- Présenter des outils du domaine pro

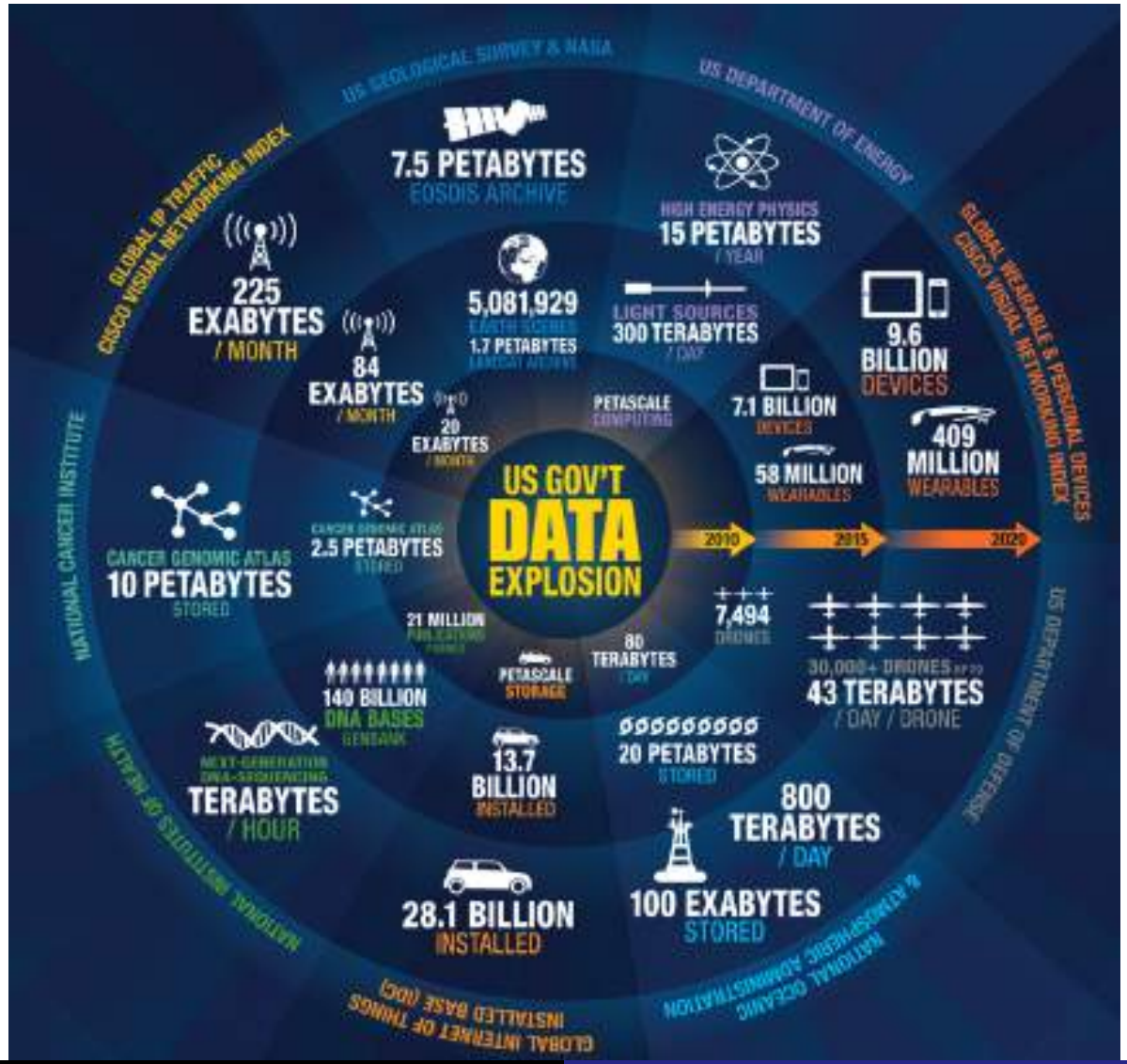
Et puis...

- Développer la créativité autour du traitement de données et de ses applications
- Travailler en équipe

Contexte


BIG DATA
Big Data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze.
The right use of Big Data allows analysts to spot trends and gives niche insights that help create value and innovation much faster than conventional methods.
The "three Vs", i.e the Volume, Variety and Velocity of the data coming in is what creates the challenge.
VOLUME
>3,500 NORTH AMERICA
>2,000 EUROPE
>250 CHINA
>400 JAPAN
>200 MIDDLE EAST
>50 INDIA
>50 LATIN AMERICA
Amount of Big Data stored across the world (in petabytes)
VARIETY
PEOPLE TO PEOPLE
PEOPLE TO MACHINE
MACHINE TO MACHINE
VELOCITY
2.9 MILLION EMAILS SENT EVERY SECOND
20 HOURS OF VIDEO UPLOADED EVERY MIN
50 MILLION TWEETS PER DAY
CASE STUDY - Healthcare
Some Advisors to the potential annual value in Healthcare
$165B CLINICAL
$90 PUBLIC HEALTH
$108B R&D
$8B BUSINESS MODEL
$47B ACCOUNTS
57.8% OF ORGANIZATIONS SURVEYED SAY THAT BIG DATA IS A CHALLENGE
72.7% CONSIDER DRIVING OPERATIONAL EFFICIENCIES TO BE THE BIGGEST BENEFIT OF A BIG DATA STRATEGY
90% SAY THAT BIG DATA HELPS IN BETTER MEETING CONSUMER DEMAND AND FACILITATING GROWTH
VALUE
PRODUCTIVITY INCREASE
SALES INCREASE
RETAIL
CONSULTING
AIR TRANSPORTATION
CONSTRUCTION
FOOD PRODUCTS
STEEL
AUTOMOBILE
INDUSTRIAL INSTRUMENTS
PUBLISHING
TELECOMMUNICATIONS
40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR
5% PROJECTED GROWTH IN GLOBAL IT SPENDING PER YEAR
The estimated size of the digital universe in 2011 was 1.8 zettabytes. It is predicted that between 2009 and 2020, this will grow 44 fold to 35 zettabytes per year. A well defined data management strategy is essential to successfully utilize Big Data.
DO BUSINESS BETTER
WIPRO

Contexte


US GOV'T
DATA
EXPLOSION
2010
2015
2020
7.5 PETABYTES
EOSDIS ARCHIVE
HIGH ENERGY PHYSICS
15 PETABYTES
/ YEAR
225
EXABYTES
/ MONTH
84
EXABYTES
/ MONTH
5,081,929
1.7 PETABYTES
LIGHT SOURCES
300 TERABYTES
/ DAY
9.6
BILLION
DEVICES
7.1 BILLION
58 MILLION
WEARABLES
409
MILLION
WEARABLES
20
EXABYTES
/ MONTH
PETASCALE
COMPUTING
CANCER GENOMIC ATLAS
10 PETABYTES
STORED
2.5 PETABYTES
STORED
21 MILLION
80
TERABYTES
/ DAY
PETASCALE
STORAGE
7,494
DRONES
30,000+ DRONES
43 TERABYTES
/ DAY / DRONE
140 BILLION
DNA BASES
TERABYTES
/ HOUR
13.7
BILLION
INSTALLED
20 PETABYTES
STORED
800
TERABYTES
/ DAY
28.1 BILLION
INSTALLED
100 EXABYTES
STORED
US GEOLOGICAL SURVEY & NASA
US DEPARTMENT OF ENERGY
GLOBAL IP TRAFFIC
CISCO VISUAL NETWORKING INDEX
GLOBAL WEARABLE & PERSONAL DEVICES
CISCO VISUAL NETWORKING INDEX
NATIONAL CANCER INSTITUTE
US DEPARTMENT OF DEFENSE
NATIONAL INSTITUTES OF HEALTH
NATIONAL OCEANIC & ATMOSPHERIC ADMINISTRATION
GLOBAL INTERNET OF THINGS
INSTALLED BASE (IDC)

Data driven science : le 4e paradigme (Jim Gray - Prix Turing)

SNR 2013

Extrait : "A l'heure actuelle, la science vit une révolution qui conduit à nouveau paradigme selon lequel 'la science est dans les données', autrement dit la connaissance émerge du traitement des données [...] **Le traitement de données et la gestion de connaissances représentent ainsi le quatrième pilier de la science après la théorie, l'expérimentation et la simulation.** L'extraction de connaissances à partir de grands volumes de données (en particulier quand le nombre de données est bien plus grand que la taille de l'échantillon) , l'apprentissage statistique, l'agrégation de données hétérogènes, la visualisation et la navigation dans de grands espaces de données et de connaissances sont autant d'instruments qui permettent d'observer des phénomènes, de valider des hypothèses, d'élaborer de nouveaux modèles ou de prendre des décisions en situation critique"

Traitement de données, quelques stats

....

- Meanwhile, the global mobile business intelligence market was valued at $6.18 billion in 2018 and is expected to reach $20.81 billion by 2024, at a 22.43% CAGR. (360 Market Updates, 2020)
- More than 46% of small businesses use business intelligence tools' virtual networking features as a core element of their business strategy. (Grand View Research, 2019)
- Reporting, dashboards, data integration, data warehousing, and data preparation are the top five most important technologies and initiatives strategic to BI. (Dresner, 2020)
- The top three business intelligence trends are data visualization, data quality management, and self-service business intelligence (BI). https://techjury.net
- The total enterprise data volume worldwide is estimated to

Traitement de données en entreprise

...

La donnée est donc l'un des principaux actifs immatériels de nos organisations, et pour autant, n'est pas encore gérée avec la même rigueur ni les mêmes moyens que les autres ressources, capital et ressources humaines notamment. Dans un contexte où elle est devenue critique pour l'activité de l'entreprise, la mise en place d'une gestion structurée et industrielle de la donnée est impérative.
* *Enjeux Business des données - CIGREF 2014*

La BI

L'Informatique Décisionnelle (ID), en anglais Business Intelligence (BI), est l'informatique à l'usage des décideurs et des dirigeants des entreprises. Les systèmes de ID/BI sont utilisés par les décideurs pour obtenir une connaissance approfondie de l'entreprise et de définir et de soutenir leurs stratégies d'affaires, par exemple :

- d'acquérir un avantage concurrentiel,
- d'améliorer la performance de l'entreprise,
- de répondre plus rapidement aux changements,
- d'augmenter la rentabilité, et
- d'une façon générale la création de valeur ajoutée de l'entreprise.

BI

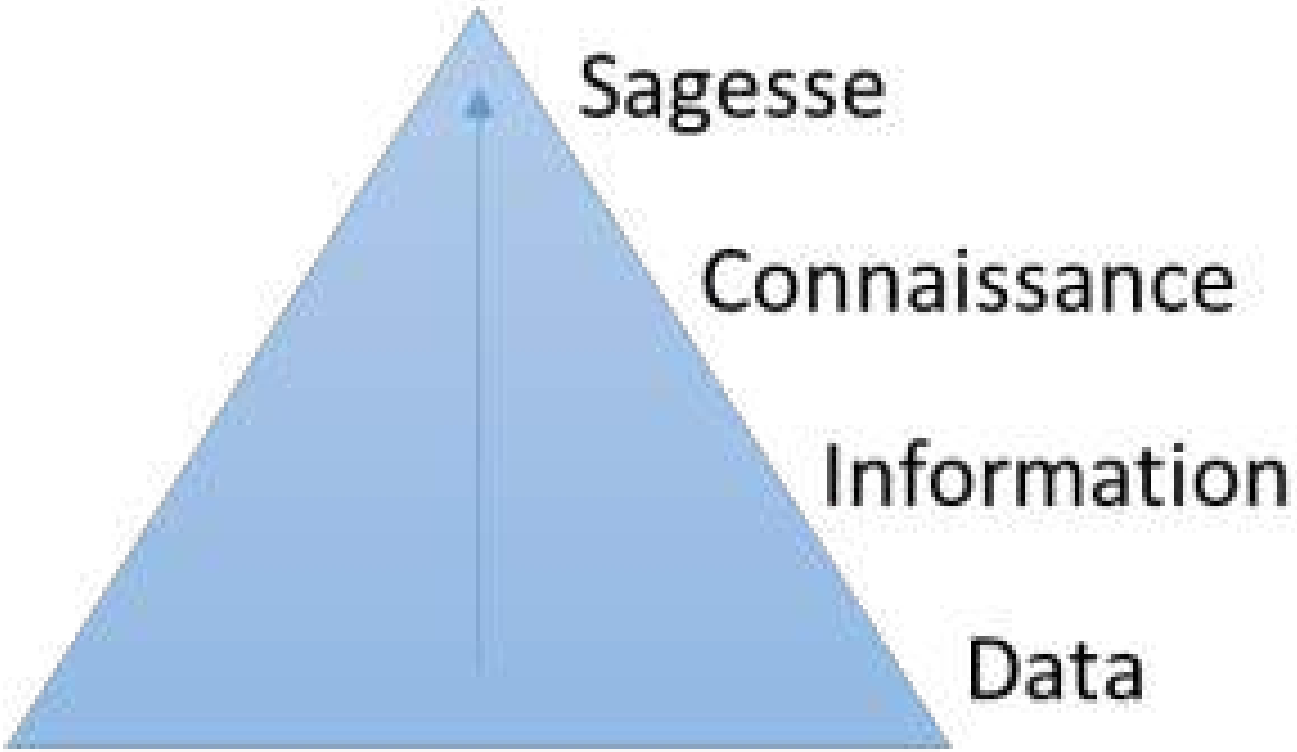
source : Rapport CIGREF 2009

Les domaines d'utilisation de la BI touchent la plupart des Métiers de l'entreprise :

- Finance, avec les reportings financiers et budgétaires par exemple ;
- Vente et commercial, avec l'analyse des points de ventes, l'analyse de la profitabilité et de l'impact des promotions par exemple ;
- Marketing, avec la segmentation clients, les analyses comportementales par exemple ;
- Logistique, avec l'optimisation de la gestion des stocks, le suivi des livraisons par exemple ;
- Ressources humaines, avec l'optimisation de l'allocation des ressources par exemple ;
- ...

La pyramide de la BI

But :

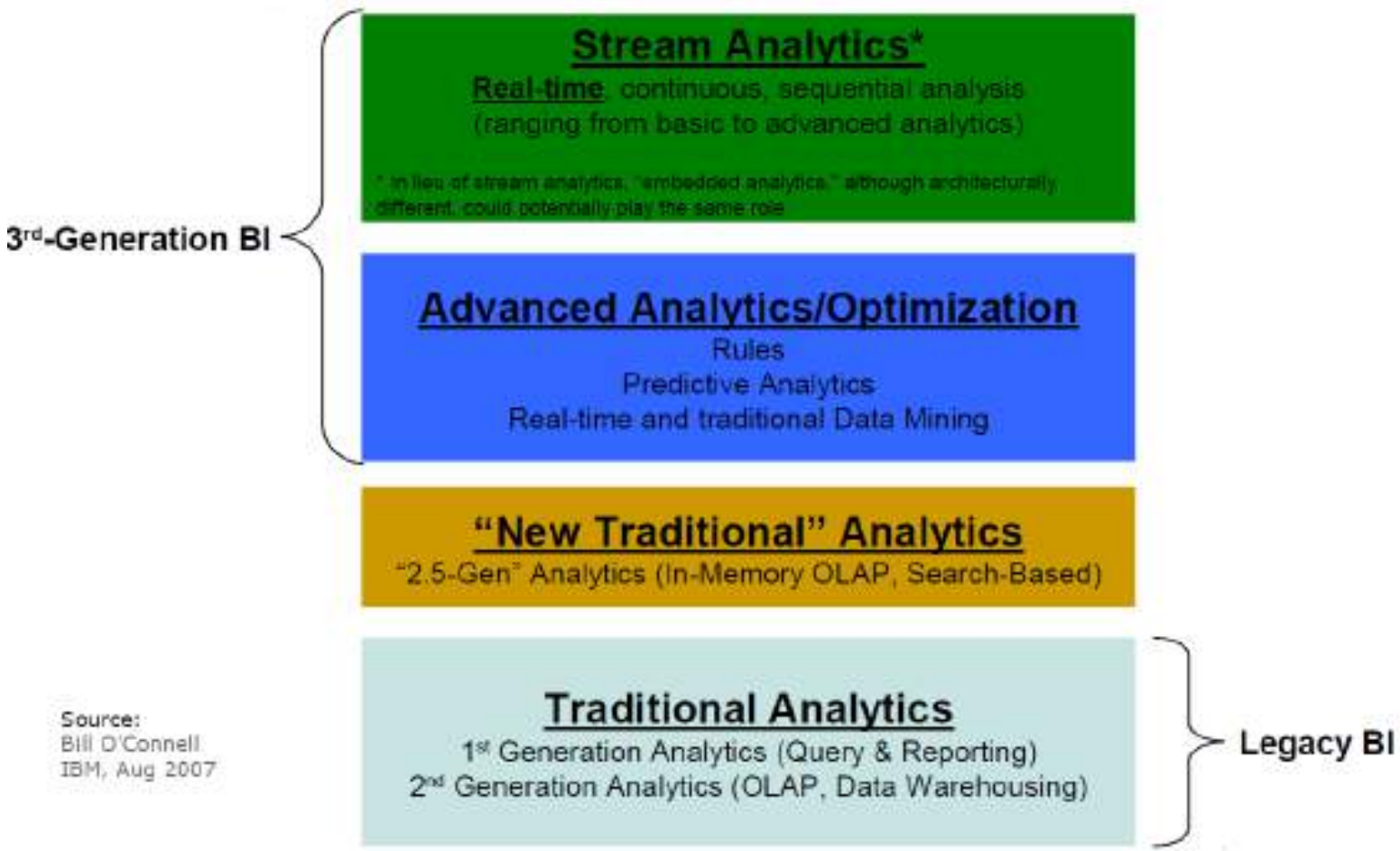


source : Blog : La BI ca vous gagne, Ah non, c'est pas la BI, c'est la montagne

Historique

- 1st Generation - Traditional analytics (query and reporting)
- 2nd Generation - Traditional generation (OLAP, data warehousing)
- 2.5nd Generation - New traditional generation
- 3rd Generation - Advanced analytics Rules, predictive analytics and realtime data mining Stream analytics

Historique


3rd-Generation BI
Stream Analytics*
Real-time, continuous, sequential analysis
(ranging from basic to advanced analytics)
* In lieu of stream analytics, "embedded analytics," although architecturally different, could potentially play the same role
Advanced Analytics/Optimization
Rules
Predictive Analytics
Real-time and traditional Data Mining
"New Traditional" Analytics
"2.5-Gen" Analytics (In-Memory OLAP, Search-Based)
Traditional Analytics
1st Generation Analytics (Query & Reporting)
2nd Generation Analytics (OLAP, Data Warehousing)
Legacy BI
Source:
Bill O'Connell
IBM, Aug 2007
ISQS 6339, Data Mgmt & BI

Les fonctions

Source : Groupe de travail CIGREF, 2014

3 fonctions :

- Data Integrator
- Data Analyst
- Data Scientist

+ **Data Steward (Responsable des données)**

Différentes Fonctions

Data Integration

- Combiner des informations hétérogènes
- venants de sources différentes

Data Analyst

Inspection, nettoyage, transformation et modélisation des données.
Le **Data Mining** est une forme particulière de *Data Analysis* centrée sur la modélisation et l'extraction de connaissances à partir de données

- En lien étroit avec la *Data Vizualisation* qui s'intéresse à la visualisation de données
  - Rendre la données compréhensible
  - Communiquer à partir de la donnée

Différentes Fonctions

Data Scientist

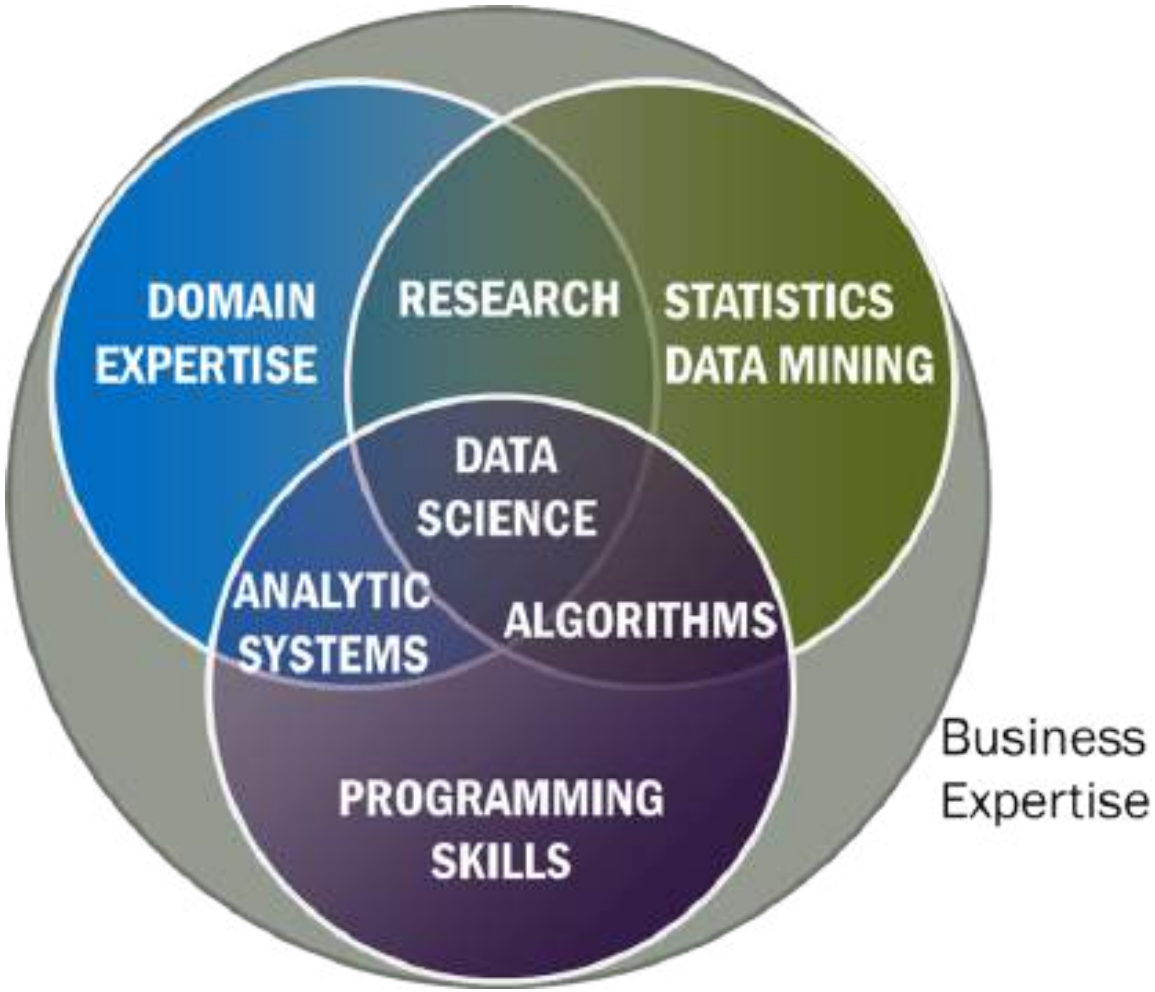
Il s'agit de disposer de compétences de haut niveau en matière d'analyse de données, en combinant à la fois les méthodes statistiques, mais aussi d'autres connaissances telles que la linguistique, la sémantique, utiles notamment pour travailler sur des données non structurées, sans oublier la bonne compréhension du métier sur lequel on travaille, et de mettre en oeuvre une démarche d'analyse itérative, en acceptant de tester des hypothèses sans a priori sur le résultat recherché.

Data Steward - Responsable des Données

[...] susceptibles sur un périmètre métier sur lequel ils détiennent une expertise reconnue, de spécifier les exigences sur les données et d'en contrôler la qualité. Ces responsables de données peuvent être positionnés à différents niveaux dans l'organisation, et peuvent être pilotés par des coordinateurs au niveau d'un métier, d'une fonction support ou d'une géographie.

Nouveau Métier

Contexte


DOMAIN EXPERTISE
RESEARCH
STATISTICS DATA MINING
DATA SCIENCE
ANALYTIC SYSTEMS
ALGORITHMS
PROGRAMMING SKILLS
Business Expertise

Architecture

Source : Groupe de travail CIGREF, 2014

Plusieurs éléments

- Data Sources
- Data Wharehouses et Data Marts
  - Extract, transform and load data
  - Multidimensional Exploratory Analysis
- Data Mining et Data Analytics
  - Extraction of Information and Knowledge from Data
  - Build Models of Prediction

Architecture

- Les données opérationnelles sont extraites périodiquement de sources hétérogènes : fichiers plats, fichiers Excel, base de données (DB2, Oracle, SQL Server, etc.), service web, données massives et stockées dans un entrepôt de données.
- Les données sont restructurées, enrichies, agrégées, reformatées, nomenclaturées pour être présentées à l'utilisateur sous une forme sémantique (vues métiers ayant du sens) qui permettent aux décideurs d'interagir avec les données sans avoir à connaître leur structure de stockage physique, de schémas en étoile qui permettent de répartir les faits et mesures selon des dimensions hiérarchisées, de rapports pré-préparés paramétrables, de tableaux de bords plus synthétiques et interactifs.
- Ces données sont livrées aux divers domaines fonctionnels (direction stratégique, finance, production, ressources humaines, etc.) à travers un système de sécurité ou de datamart spécialisés à des fins de consultations, d'analyse, d'alertes prédéfinies, d'exploration, etc.

Les fonctions de la BI

- Fonction de collecte de données
- Fonction d'intégration
- Fonction de diffusion (ou distribution)
- Fonction présentation

Collecte de données - *Data Pumping*

Définition

La fonction collecte (parfois appelée datapumping) recouvre l'ensemble des tâches consistant à détecter, sélectionner, extraire et filtrer les données brutes issues des environnements pertinents compte tenu du périmètre couvert par le SID.

Hétérogénité des données

Plusieurs types de sources

- fichiers plats
- fichiers Excel
- base de données (DB2, Oracle, SQL Server, etc.)
- services web
- données massives

Plusieurs natures d'informations

- Données quantitatives, texte, image, flux, ...
- Flux de données
- Données bruitées, données fausses

Difficultés

- Sources diverses et disparates ;
- Sources sur différentes plateformes et OS ;
- Applications utilisant des BDs et autres technologies obsolètes ;
- Historique de changement non-préservé dans les sources ;
- Qualité de données douteuse et changeante dans le temps ;
- Structure des systèmes sources changeante dans le temps ;
- Incohérence entre les différentes sources ;
- Données dans un format difficilement interprétable ou ambigu.

Fonction d'intégration

Définition

La fonction d'intégration consiste à concentrer les données collectées dans un espace unifié, dont le socle informatique essentiel est l'entrepôt de données (Datawarehouse)
Wikipedia

Cela inclut :

- Le nettoyage et filtrage des données
- La validation des données
- La synchronization
- La certification

ETL

La fonction de collecte s'appuie habituellement sur des outils d'ELT

ETL = Extract, Tranform, Load

Processus :

- **Extract :** Extraire les données de sources hétérogènes
- **Transform :** Transformation des données pour les stocker dans un datawharehouse
- **Load :** Chargement des données dans le datawharehouse

Les logiciels d'ETL sont des intergiciels = des logiciels dont le but est de faire passer des données entre plusieurs logiciels.

ETL

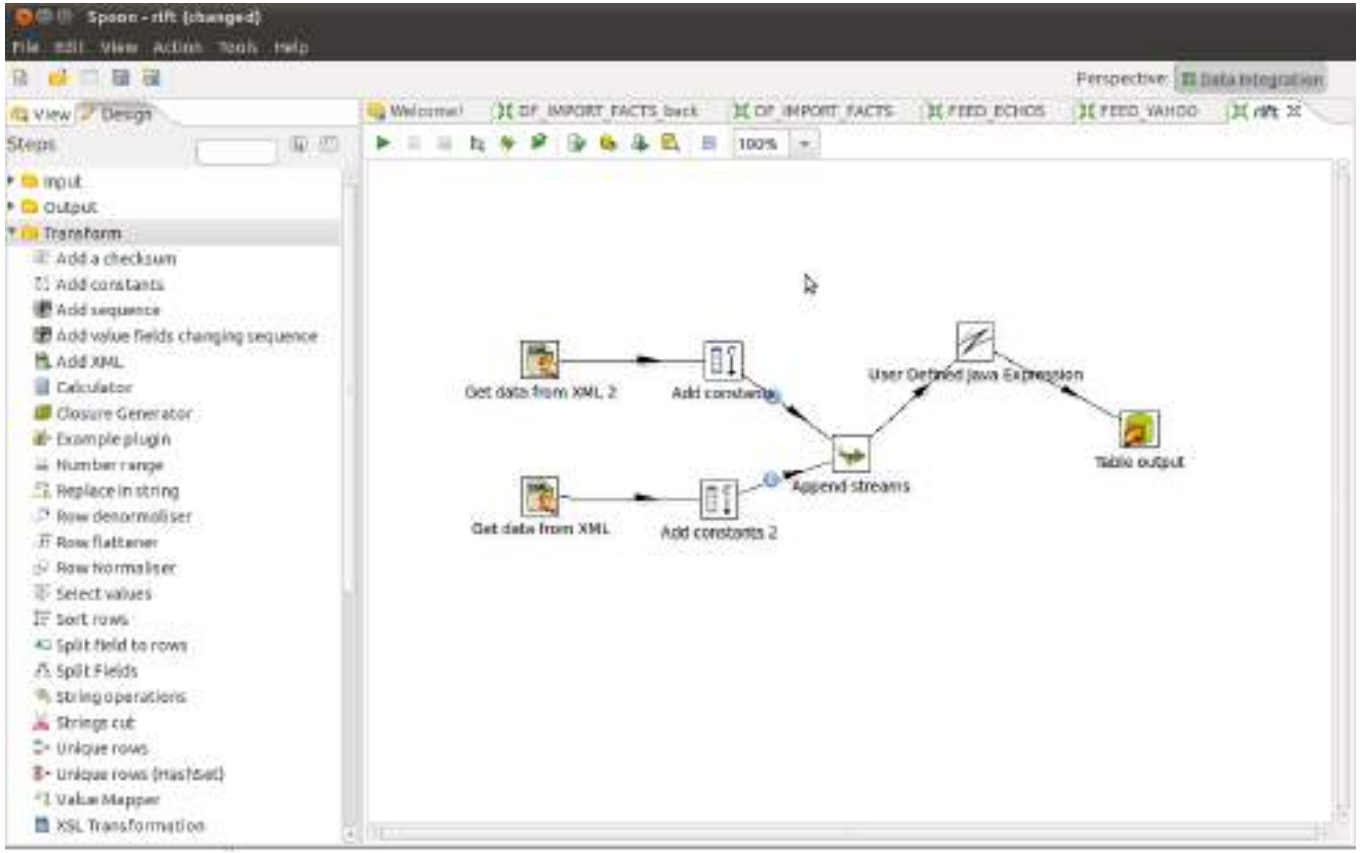
Les ETL sont basés sur un ensemble de connecteurs permettant la gestion de différentes sources de données

Logiciels ETL

Plusieurs logiciels sont disponibles sur le marché. Ils permettent d'effectuer de l'ETL sous forme de programmes "graphiques". Ils sont intégrés dans des suites de BI.

- Anatella2
- DataStudio (Data)
- Feature Manipulation Engine (FME)
- Hurence avec un ETL natif Hadoop
- IBM InfoSphere DataStage
- Informatica PowerCenter
- MapReport
- Microsoft SQL Server Integration Services (SSIS)
- OpenText Genio
- Oracle Data Integrator (Sunopsis)
- Oxio Data Intelligence solution ETL
- SAP Data Services
- SAS Data Integration Studio
- Stambia
- STATISTICA ETL (StatSoft)
- SynchroDB https ://synchrodb.com
- Talend

Kettle - Pentaho



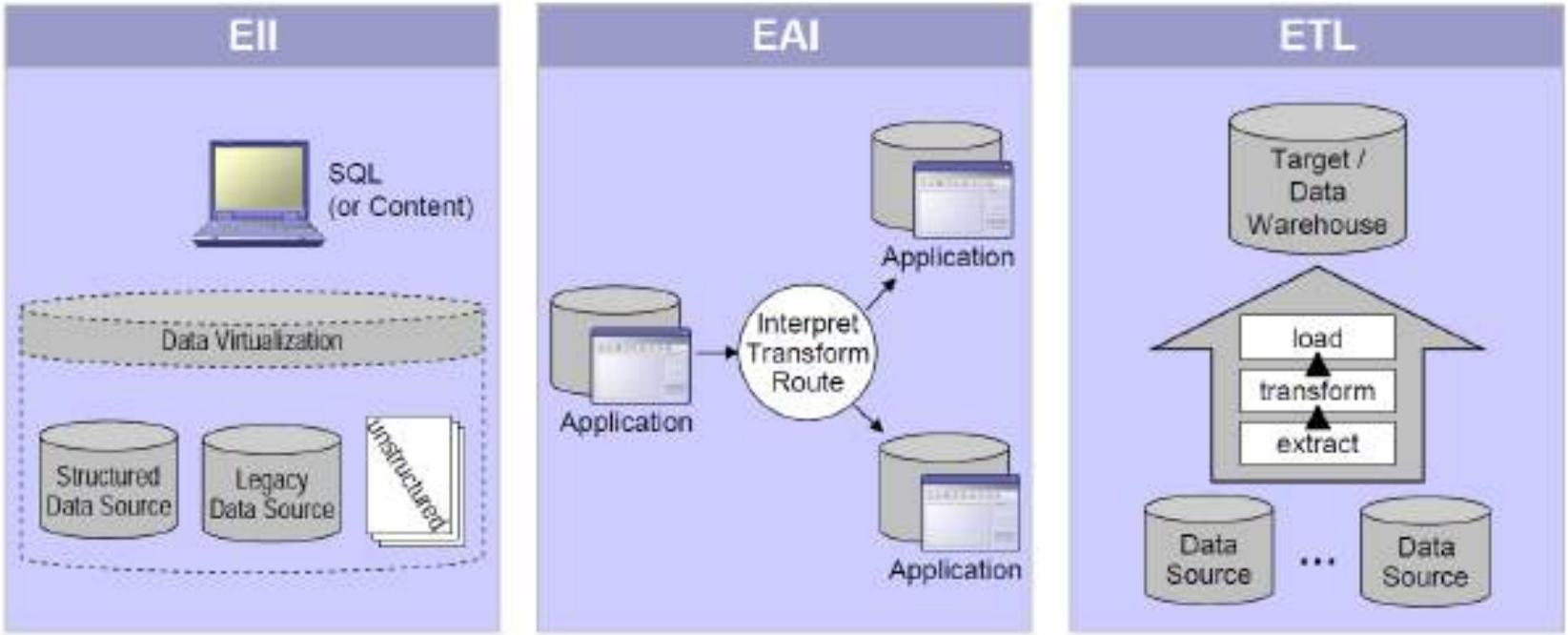
Définition

L'intégration de données appelé ETL (Extraction Transfer Loading) regroupe les processus par lesquels les données provenant de différentes parties du système d'information sont déplacées, combinées et consolidées. Ces processus consistent habituellement à extraire des données de différentes sources (bases de données, fichiers, applications, Services Web, emails, etc.), à leur appliquer des transformations (jointures, lookups, déduplication, calculs, etc.), et à envoyer les données résultantes vers les systèmes cibles.
Source : wikiversity.org

Il existe plusieurs système d'intégration de données :

- La médiation au service de l'intégration de données d'entreprise (EII).
- L'intégration de données via les applications (EAI).
- L'intégration de données via les services Web (ESB, SOA).
- L'intégration de données en nuage (Data Cloud).
- L'ETL (Extract - Transform - Load)

EII - EAI - ETL



- Real-time information access
- Federation of data from multiple sources
- Dynamic drill down
- Semi-structured & unstructured data

- Process based integration of application data
- Message-based, transaction-oriented processing
- Workflow and data orchestration, content-based routing

- Bulk data integration
- Set-based & hierarchical transformations
- High scale, batch-oriented data delivery

Source : IBM Software group

Conception

Le rapatriement des données peut se faire de trois façons différentes :

- **Push :** la logique de chargement est dans le système de production, il pousse les données vers le Staging quand il en a l'occasion.
- **Pull :** le Pull tire les données de la source vers le Staging.
- **Push-Pull :** La source prépare les données à envoyer et prévient le Staging qu'elle est prête. Le Staging va récupérer les données. Si la source est occupée, le Staging fera une autre demande plus tard.

Data warehouse

Définition

Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles et fournir ainsi un socle à l'aide à la décision en entreprise.
Wikipedia

Attention : Le data warehouse est différent des bases de donnees opérationnelles de l'entreprise qui l'alimentent.

Datawarehouse

Un entrepôt de données conserve une copie des informations des systèmes de transaction opérationnels. Il offre la possibilité de :

- Rassembler des données provenant de sources multiples en une seule base de données afin qu'un moteur de requête unique puisse être utilisé pour présenter des données.
- Permettre l'éxécution de requête longues, bloquantes, sur des données opérationnelles
- Maintenir l'historique des données, même si les systèmes de transaction source ne le font pas
- Intégrer des données provenant de multiples systèmes sources, permettant une vue centrale dans l'entreprise. Cet avantage est particulièrement valable lorsque l'organisation est issue de fusions successives
- Améliorer la qualité des données

Datawarehouse

Un entrepôt de données conserve une copie des informations des systèmes de transaction opérationnels. Il offre la possibilité de :

- Présenter l'information de l'organisation
- Fournir un seul modèle de données commun pour toutes les données d'intérêt, indépendamment de la source de données
- Restructurer les données de sorte qu'elles prennent sens (décisionnel)
- Ajouter de la valeur aux applications métiers opérationnels, notamment la gestion de la relation client (CRM).
- Faire des requêtes d'aide à la décision plus faciles à écrire.

Datawarehouse vs BD opérationnelle

Table 11-1 Comparison of Operational and Informational Systems

Characteristic	*Operational Systems*	*Informational Systems*
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

Datamart

Définition

Un DataMart (littéralement en anglais magasin de données) est un sous-ensemble d'un DataWarehouse destiné à fournir des données aux utilisateurs, et souvent spécialisé vers un groupe ou un type d'affaire. Techniquement, c'est une base de données relationnelle utilisée en informatique décisionnelle et exploitée en entreprise pour restituer des informations ciblées sur un métier spécifique, constituant pour ce dernier un ensemble d'indicateurs utilisés pour le pilotage de l'activité et l'aide à la décision.

Source : wikipedia

Le datawarehouse est **Général**, le datamart est **spécifique** à un métier.

Datamart vs datawarehouse

Table 11-2 Data Warehouse Versus Data Mart

Data Warehouse	*Data Mart*
Scope	*Scope*
• Application independent	• Specific DSS application
• Centralized, possibly enterprise-wide	• Decentralized by user area
• Planned	• Organic, possibly not planned
Data	*Data*
• Historical, detailed, and summarized	• Some history, detailed, and summarized
• Lightly denormalized	• Highly denormalized
Subjects	*Subjects*
• Multiple subjects	• One central subject of concern to users
Sources	*Sources*
• Many internal and external sources	• Few internal and external sources
Other Characteristics	*Other Characteristics*
• Flexible	• Restrictive
• Data-oriented	• Project-oriented
• Long life	• Short life
• Large	• Start small, becomes large
• Single complex structure	• Multi, semi-complex structures, together complex

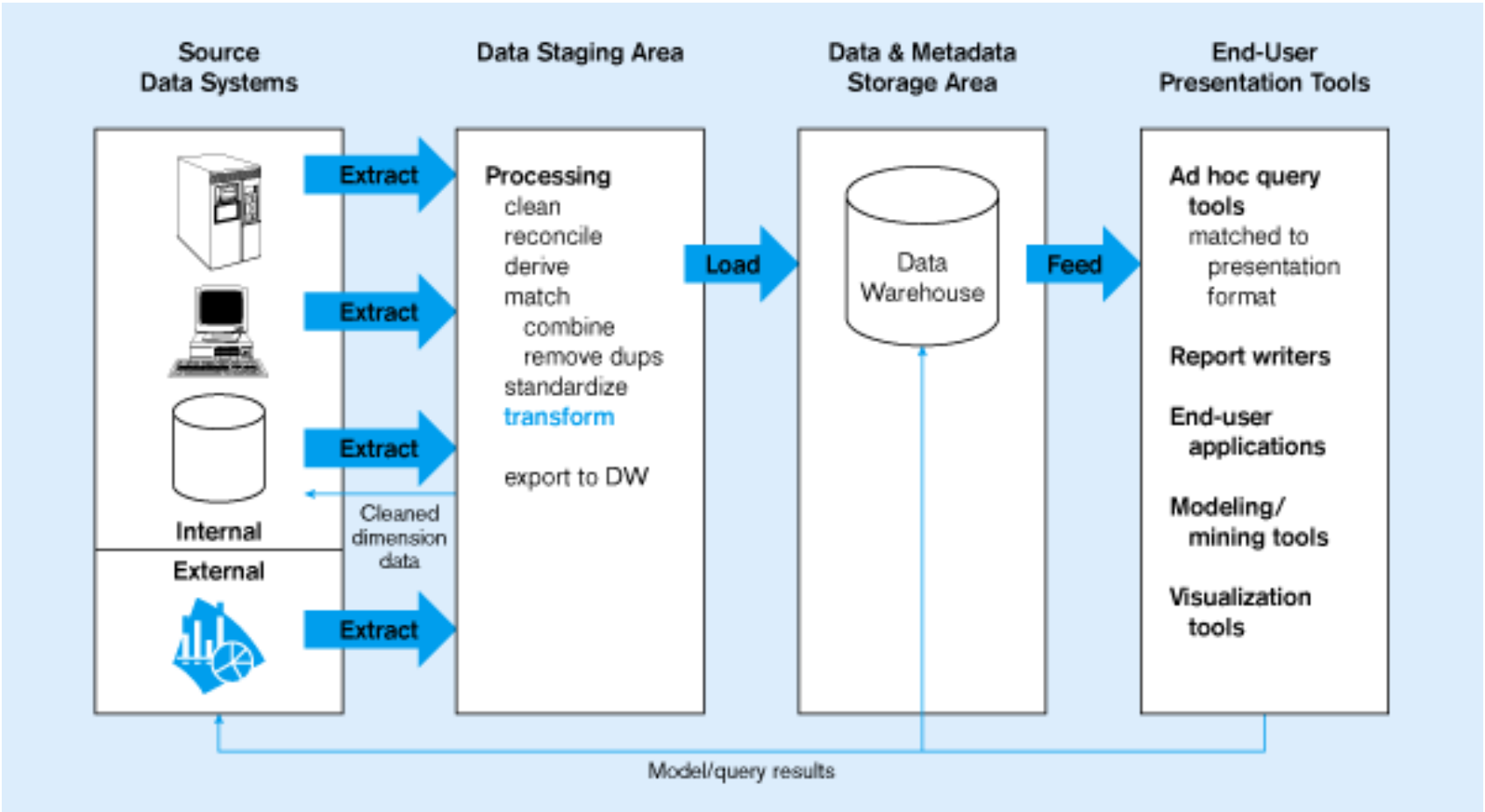
Adapted from Strange (1997)

Datamart vs datawarehouse

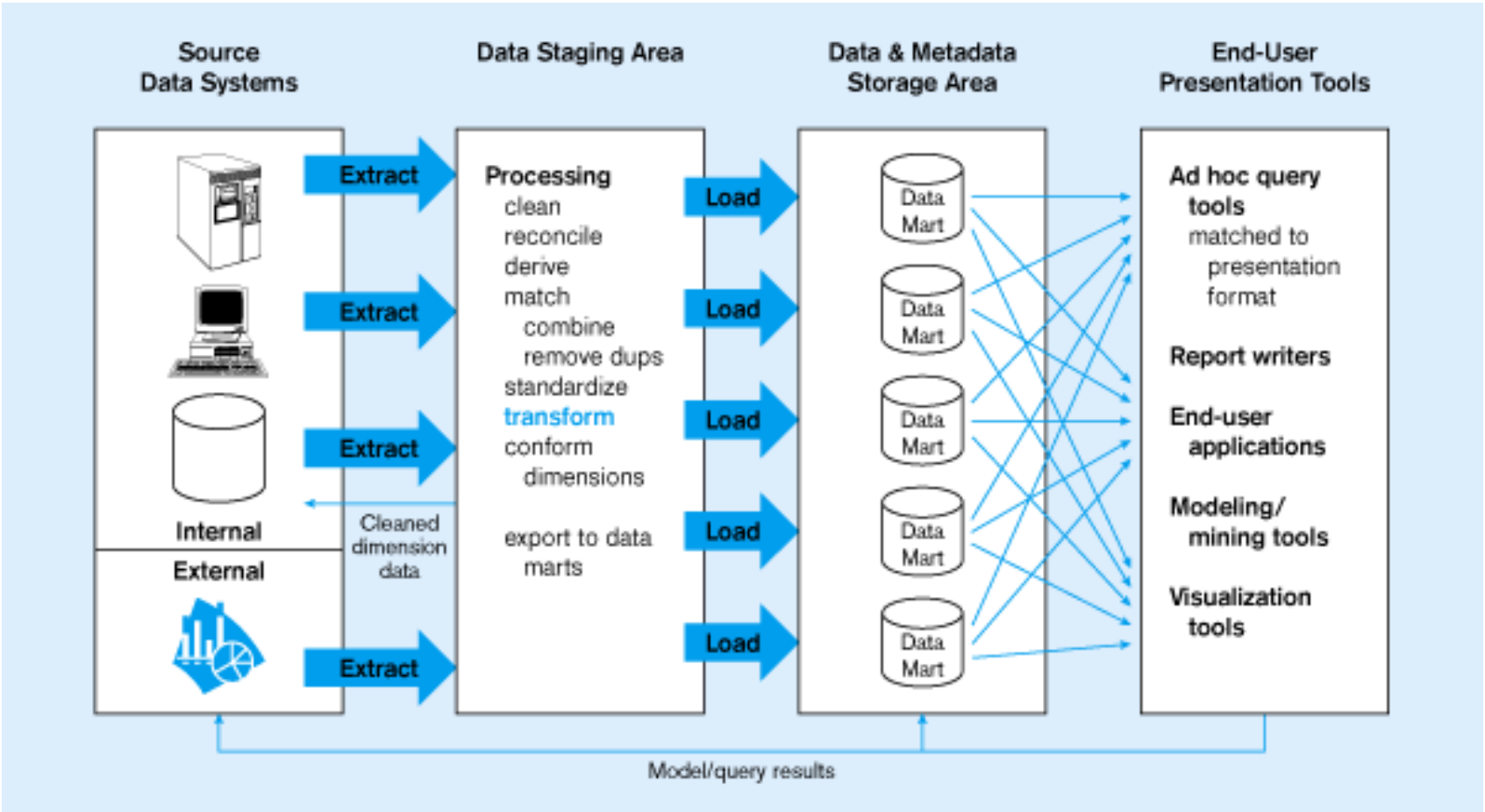
Deux conceptions existantes :

- Définition d'Inmon : Le DataMart est issu d'un flux de données provenant du DataWarehouse. Contrairement à ce dernier qui présente le détail des données pour toute l'entreprise, il a vocation à présenter la donnée de manière spécialisée, agrégée et regroupée fonctionnellement.
- Définition de Kimball : Le DataMart est un sous-ensemble du DataWarehouse, constitué de tables au niveau détail et à des niveaux plus agrégés, permettant de restituer tout le spectre d'une activité métier. L'ensemble des DataMarts de l'entreprise constitue le DataWarehouse.

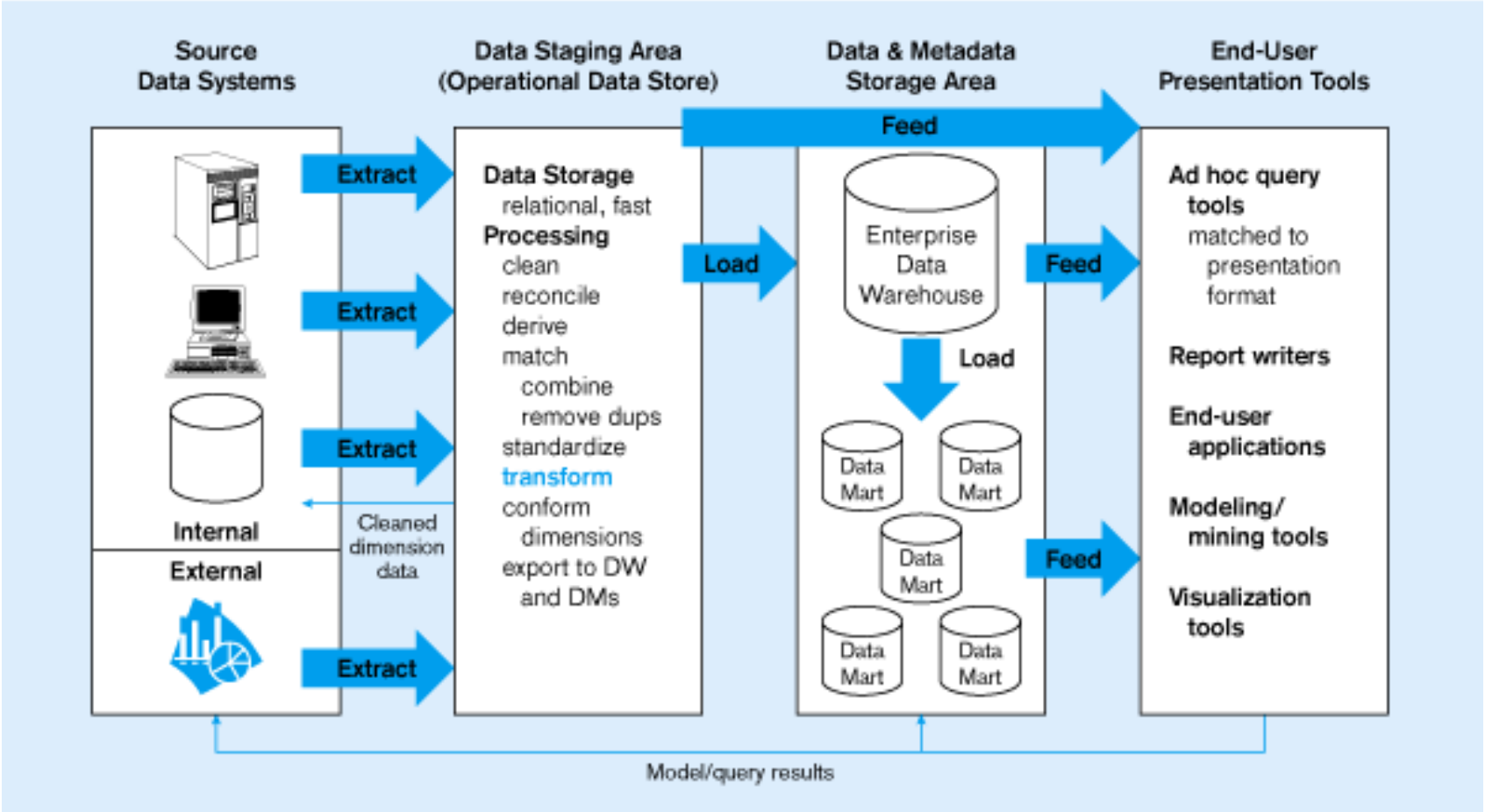
Différentes Architectures



Différentes Architectures

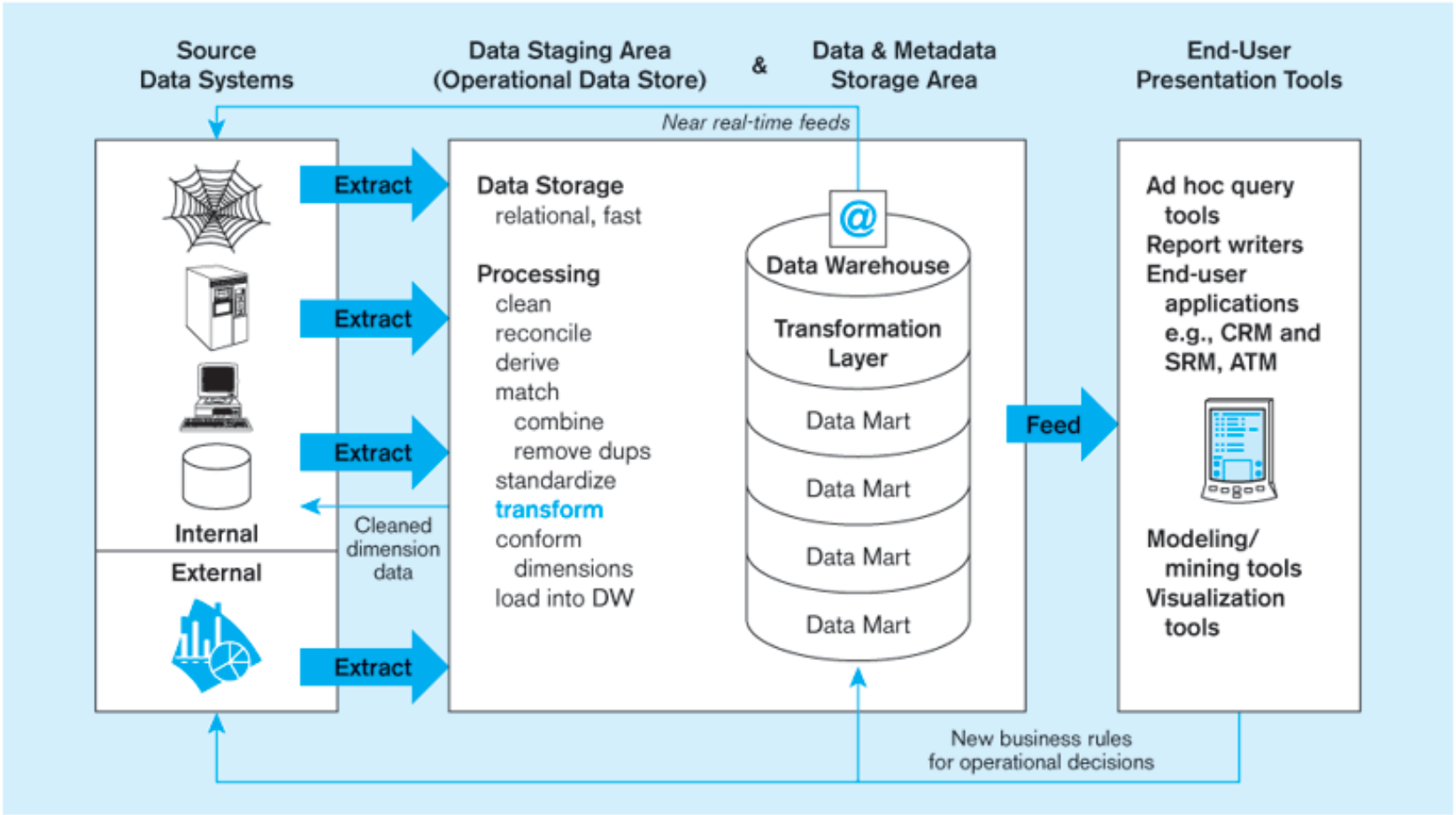


Différentes Architectures



Différentes Architectures

Figure 11-5 Logical data mart and @ctive warehouse architecture



Données orientées sujets

- En production : données organisées par processus fonctionnels
- Datawarehouse : données organisées autour de sujets majeurs
- Données structurées par thème, potentiellement transverses par rapport aux domains fonctionnels et organisationnelles

Examples (médecine) : Actes, Séjours vs Bases par services

Architecture

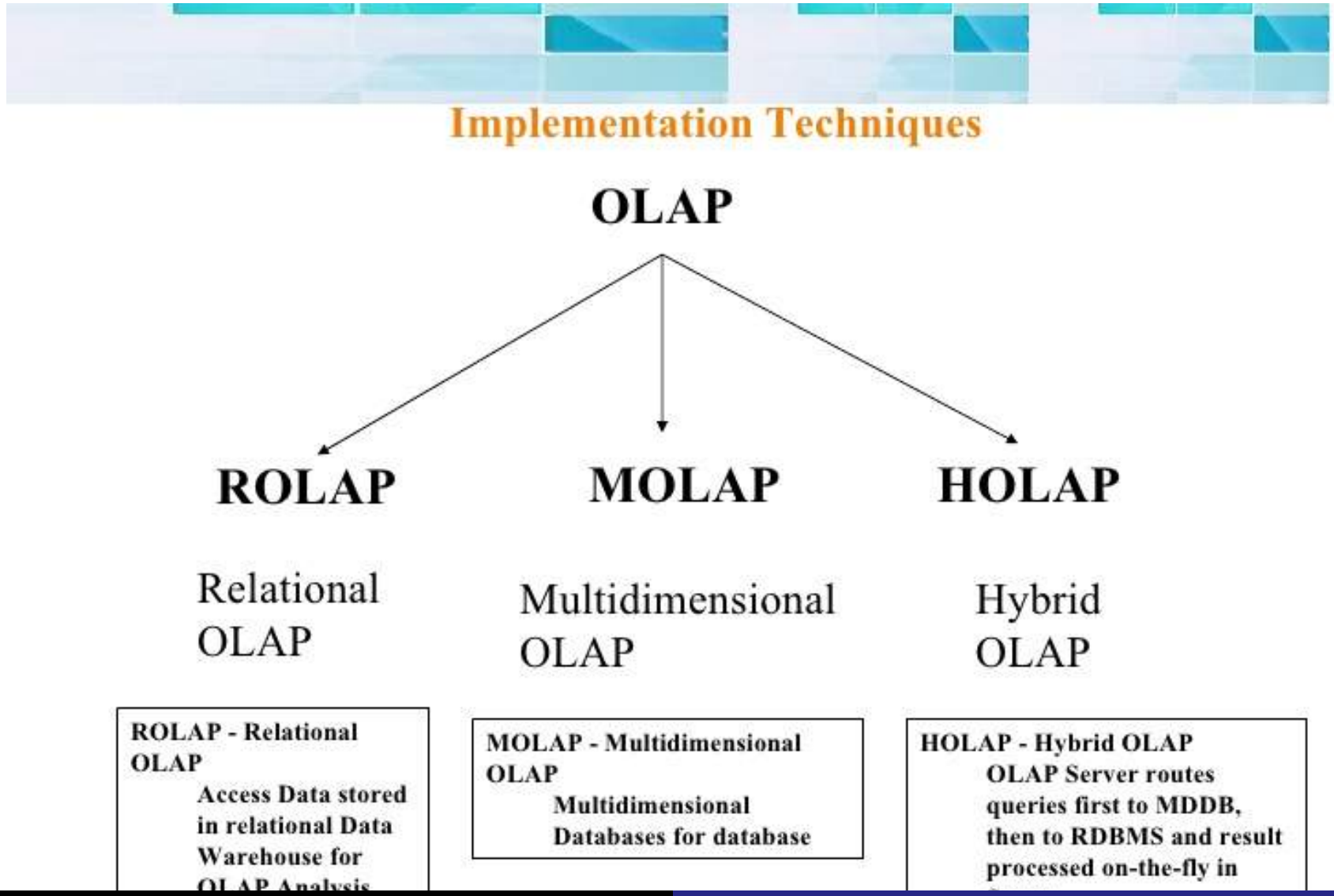
Good DW architecture

"It's not easy to describe a good design, but I'll know it when I see it"

Modèle relationnel

- Normalisation (3NF)
- Répond aux besoins transactionnels (OLTP)
- Avantages :
  - Réduction de l'entrée de données
  - Réduction du nombre d'index
  - Ajouts/destructions/modifications plus rapides
- Désavantages :
  - Peu efficace pour l'extraction de données analytiques
  - Beaucoup de relations
  - Trop complexe pour l'utilisateur BI

Modèle relationnel



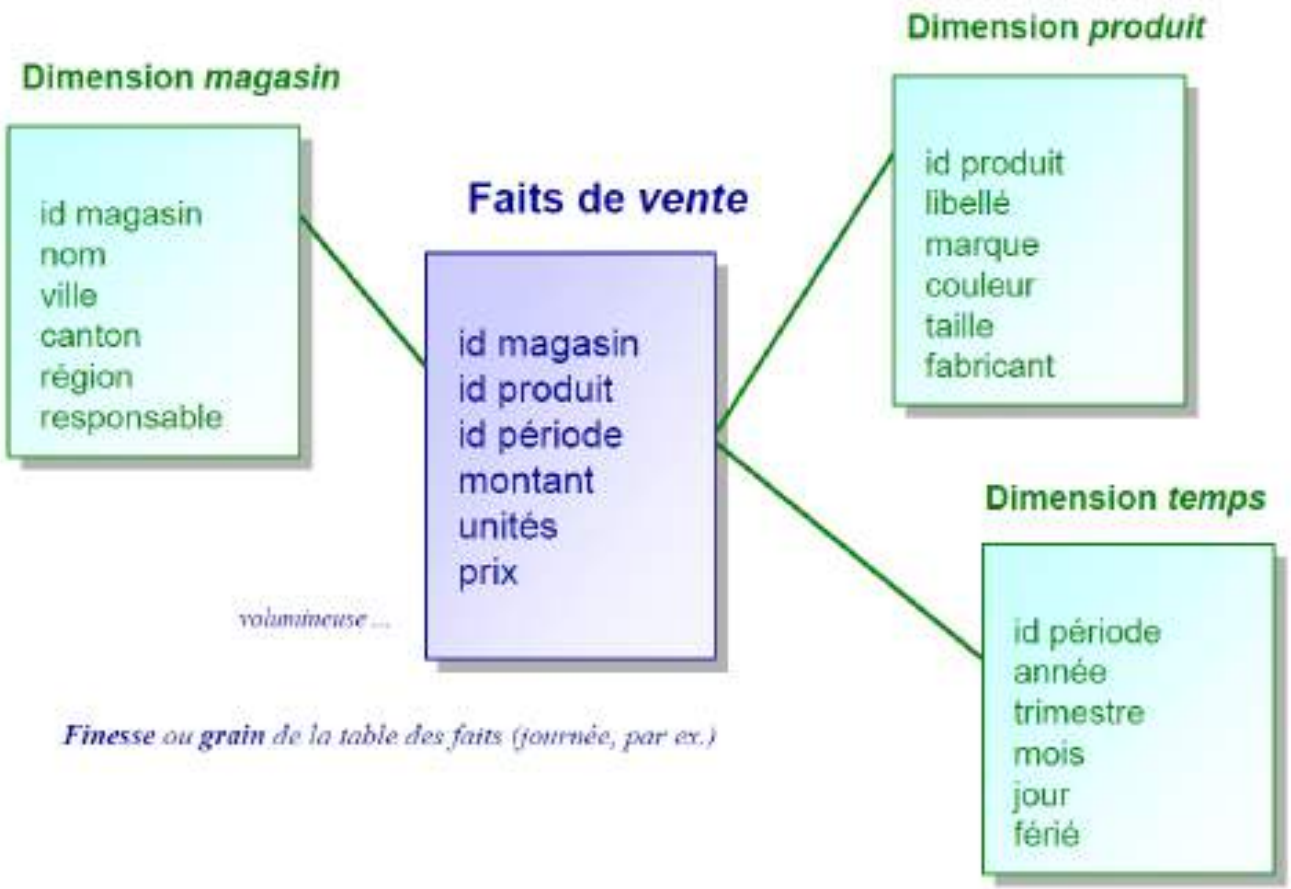
Modèle dimensionnel

Principes

On va partir du besoin "client" (quel analyse ?). On va définir des **faits** et des **dimensions**.

- **Faits :** les faits représentent un sujet d'analyse. Les faits sont caractérisées par plusieurs informations
- **Dimensions :** les dimensions sont les critères selon lesquels on souhaite faire de l'analyse.

Modèle dimensionnel



Aussi connu sous le nom de modèle en étoile