

# BI = Business Intelligence

## Master Data-Science

Ludovic DENOYER - ludovic.denoyer@lip6.fr

Laure SOULIER - laure.soulier@lip6.fr

UPMC

January 16, 2017

# Organisation

## Horaire

- Cours : Lundi de 13h30 à 15h30
- TP : Vendredi de 13h30 à 17h45

## Intervenants:

- Ludovic DENOYER - ludovic.denoyer@lip6.fr
- Divers industriels (en cours de construction)

## Evaluation

- Evaluation sur projet (TP + travail maison)
- Evaluation continue (présentielle)
- Examen final

# UE Orientée projet

## Projet

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

# Objectifs

Etudiants  $\Rightarrow$  Data Integrator  $\Rightarrow$  Data Analyst  $\Rightarrow$  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 (expérimental)

# Plan

- Partie 1 : BI de base
  - Cours 1 : Introduction
  - Cours 2 : ETL
  - Cours 3 : Datawarehouse : modélisation multidimensionnelle
  - Cours 4 : Intervention "Qualité des données" ?
  - Cours 5 : OLAP
- Partie 2 : Data Mining
  - Cours 6 : Introduction + règles d'association
  - Cours 7 : Classification
  - Cours 8 : visualisation
  - Cours 9-10 : Intervention "Méthodologie agile"

# Contexte

www.wipro.com

## 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

- NORTH AMERICA: >3,500
- EUROPE: >2,000
- CHINA: >250
- MIDDLE EAST: >200
- INDIA: >50
- JAPAN: >400
- LATIN AMERICA: >50

*Amount of Big Data stored across the world (in petabytes)*

### VARIETY

- PEOPLE TO PEOPLE:** NETWORKS, VIRTUAL COMMUNITIES, SOCIAL NETWORKS, WEB LOGS...
- PEOPLE TO MACHINE:** ARCHIVES, MEDICAL DEVICES, DIGITAL TV, E-COMMERCE, SMART CARDS, BANK CARDS, COMPUTERS, MOBILES...
- MACHINE TO MACHINE:** SENSORS, GPS DEVICES, RAW CODE SOURCES, SURVEILLANCE CAMERAS, SCIENTIFIC RESEARCH...

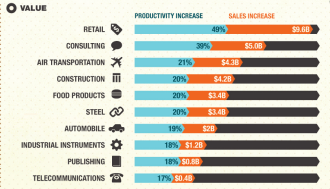
### VELOCITY

- 2.9 MILLION:** EMAILS SENT EVERY SECOND
- 20 HOURS:** OF VIDEO UPLOADED EVERY MIN
- 50 MILLION:** TWEETS PER DAY

### CASE STUDY - Healthcare

\$300 billion is the potential annual value to Healthcare

- \$165B CLINICAL:** TRANSPARENCY IN CLINICAL DATA AND CLINICAL DECISION SUPPORT
- \$9B PUBLIC HEALTH:** PUBLIC HEALTH SURVEILLANCE AND RESPONSE SYSTEMS
- \$108B RAD:** AGGREGATION OF PATIENT RECORDS, ONLINE PLATFORMS AND COMMUNITIES
- \$47B ACCOUNTS:** ADVANCED FRAUD DETECTION, PERFORMANCE BASED DRUG PRICING
- \$5B BUSINESS MODEL:** RESEARCH AND DEVELOPMENT, PERSONALIZED MEDICINE, CLINICAL TRIAL DESIGN



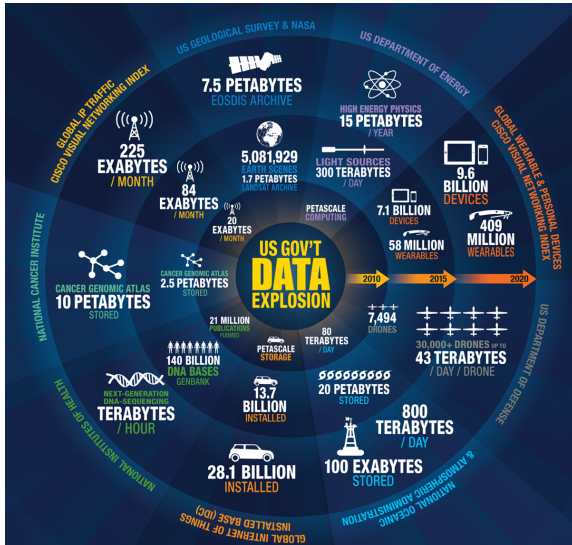
40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR

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

# Contexte



# 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

.....

68% des entreprises qui ont systématiquement recours à une analyse de données dans leurs prises de décision voient leurs bénéfices augmenter

\* *selon une étude menée par the Economist Intelligence Unit (2014)*

.....

Pour qui réussit à optimiser son usage, la donnée devient **information**, puis, bien partagée au sein de l'entreprise, elle se transforme en **connaissance** et constitue son **savoir**. Elle peut être une source de services et d'innovations, notamment lorsqu'on la croise avec d'autres données et qu'elle provient de sources diverses.

\* *Enjeux Business des données - CIGREF 2014*

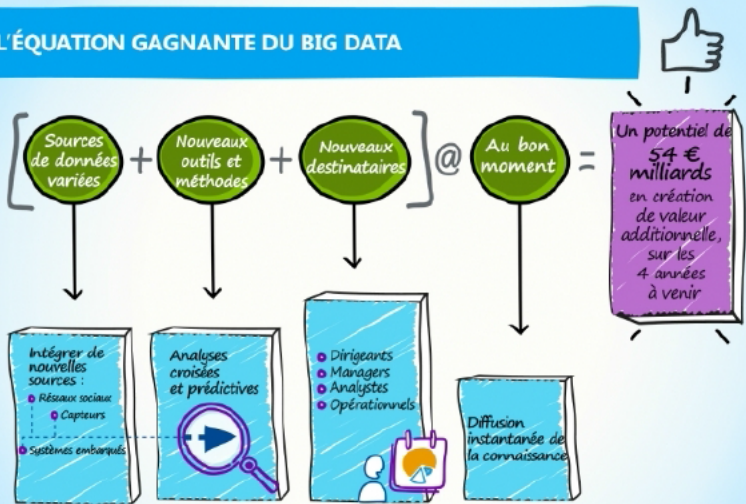
# Traitement de données en entreprise

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*

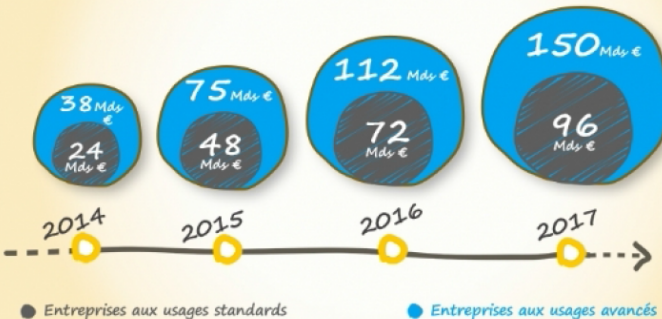
# Etude IDC - Microsoft 2014

## L'ÉQUATION GAGNANTE DU BIG DATA



# Etude IDC - Microsoft 2014

## ÉVOLUTION DU CUMUL DE VALEURS DE LA DONNÉE (en milliards d'euros)

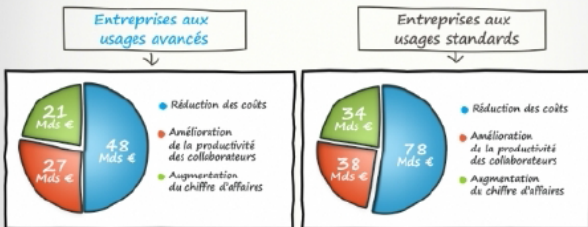


# Etude IDC - Microsoft 2014

## COMPÉTITIVITÉ DES ENTREPRISES ET DONNÉES



Une meilleure exploitation des données pourrait améliorer la compétitivité des entreprises à plusieurs niveaux



# 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.

# Je me repose

## BI

source: Rapport CIGREF 2009

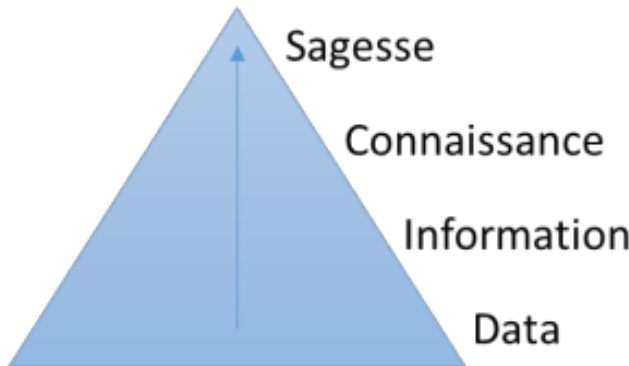
Les domaines d'utilisation de la BI touchent la plupart des Métiers de l'entreprise:

- Finance, avec les reportings financiers et budgétaires par exemple;
- Vente et commercial, avec l'analyse des points de ventes, l'analyse de la rentabilité 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

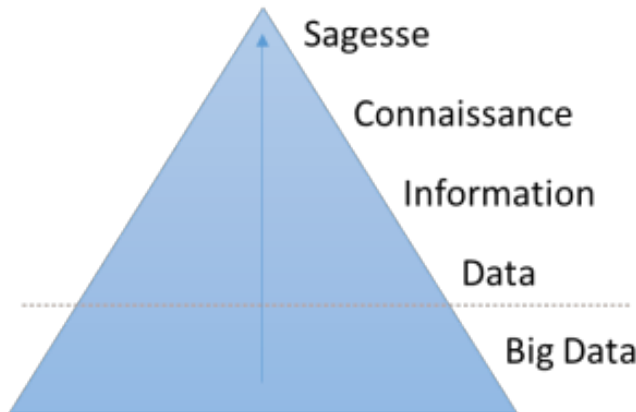
But:



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

# La pyramide de la BI

But:

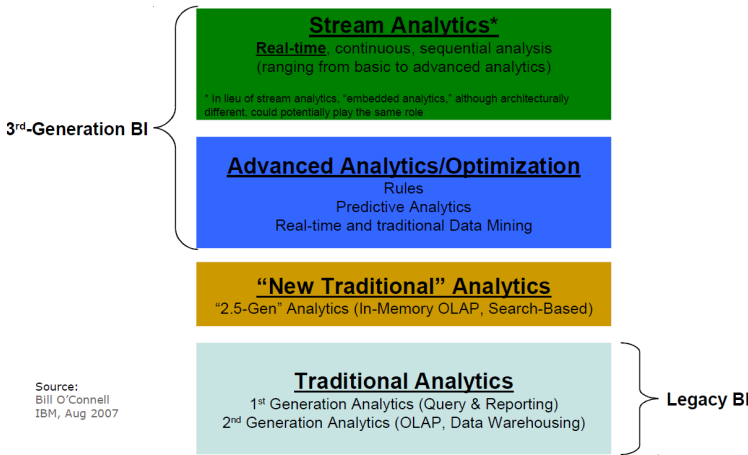


source: Blog: La BI ca vous gagne, Ah non, c'est pas la BI, c'est la montagne

# 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

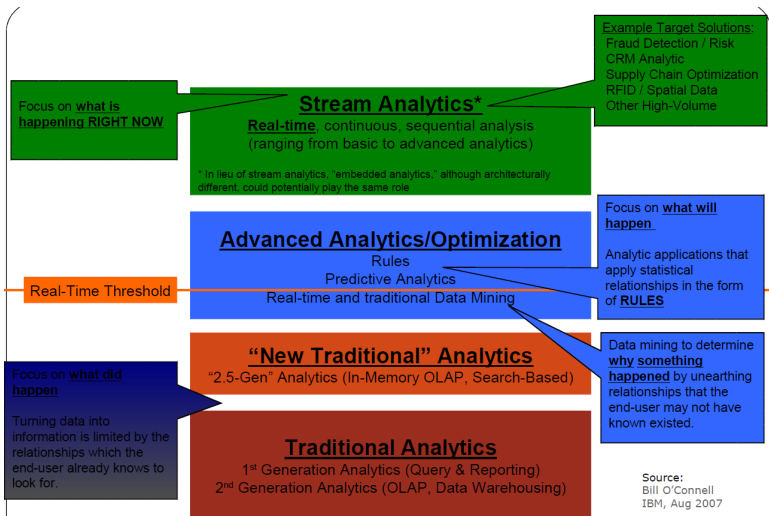
# Historique



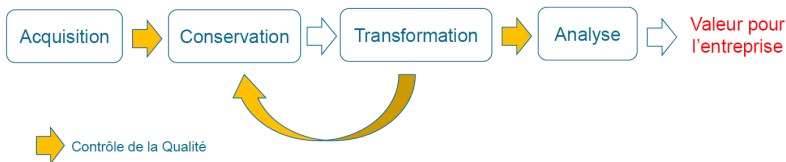
Source:  
 Bill O'Connell  
 IBM, Aug 2007

ISQS 6339, Data Mgmt & BI

# Historique



# 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 Analysis

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

## Characteristics of data scientists



**BIG  
DATA  
SCIENCE**

I feel comfortable operating with incomplete data

My data files are often messy

I explore data to see what it tells me

My dataset is so big, managing it is part of the challenge

My findings drive product and operational decisions

I want to have a complete set of data

My data files are usually clean

I report on what the data says

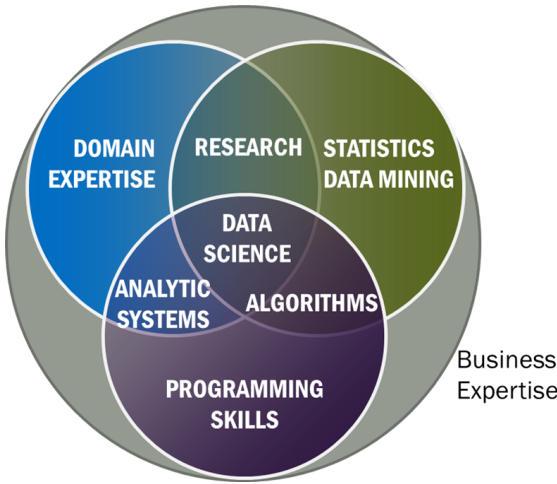
While my dataset is big, it's currently manageable

My findings measure past performance

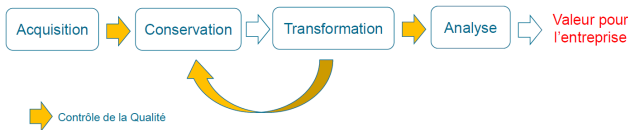


**NORMAL  
DATA  
SCIENCE**

# Contexte



# Architecture



Source : Groupe de travail CIGREF, 2014

## Plusieurs éléments

- Data Sources
- Data Warehouses 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



# 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.

# Hétérogénéité 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

# ETL

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

ETL = Extract, Transform, Load

Processus:

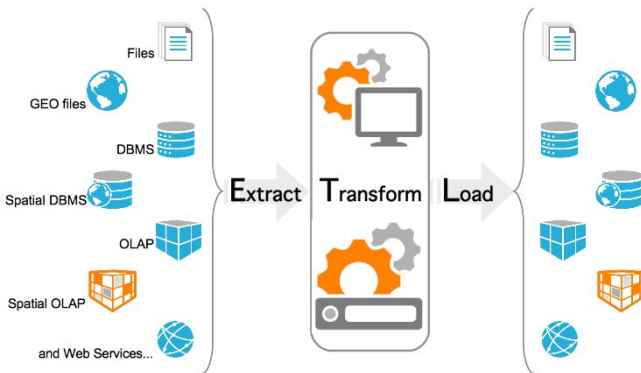
- **Extract:** Extraire les données de sources hétérogènes
- **Transform:** Transformation des données pour les stocker dans un datawarehouse
- **Load:** Chargement des données dans le datawarehouse

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



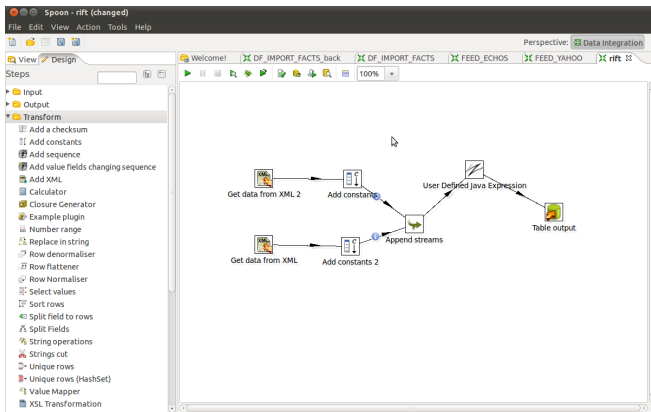
# Logiciels ETL

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

Fonction de collecte de données

# Kettle - Pentaho - TME



# 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 synchronisation
- La certification

# Data warehouse

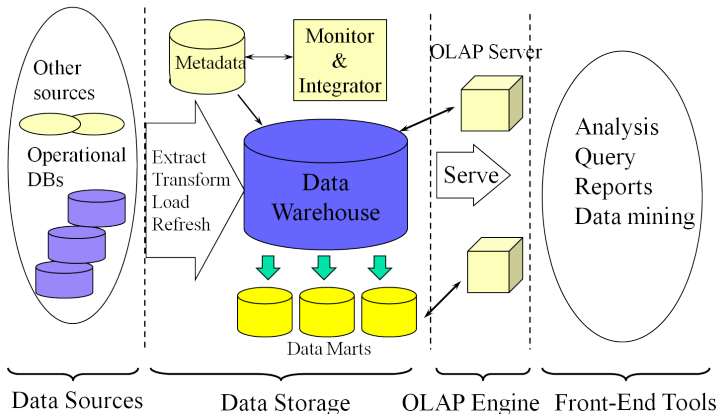
## Définition

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 données opérationnelles de l'entreprise qui l'alimentent.

# Data warehouse



# Data mart

## 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.

Wikipedia

# Data warehouse

Caractéristique	Base de données de production	Data warehouses	Datamarts
Opération	gestion courante, production	référentiel, analyse ponctuelle	analyse récurrente, outil de pilotage, support à la décision
Modèle de données	entité/relation	3NF, étoile, flocon de neige	étoile, flocon de neige
Normalisation	fréquente	maximum	rare (redondance d'information)
Données	actuelles, brutes, détaillées	historisées, détaillées	historisées, agrégées
Mise à jour	immédiate, temps réel	souvent différée, périodique	souvent différée, périodique
Niveau de consolidation	faible	faible	élevé
Perception	verticale	transverse	horizontale
Opérations	lectures, insertions, mises à jour, suppressions	lectures, insertions, mises à jour	lectures, insertions, mises à jour, suppressions
Taille	en gigaoctets	en téraoctets	en gigaoctets





# fonction de présentation

## Définition

La fonction de diffusion met les données à la disposition des utilisateurs, selon des schémas correspondant aux profils ou aux métiers de chacun, sachant que l'accès direct à l'entrepôt de données ne correspond généralement pas aux besoins spécifiques d'un décideur ou d'un analyste.

Wikipédia

La diffusion est souvent multi-dimensionnelle  $\Rightarrow$  Hypercube  $\Rightarrow$  OLAP

# OLAP - online analytical processing

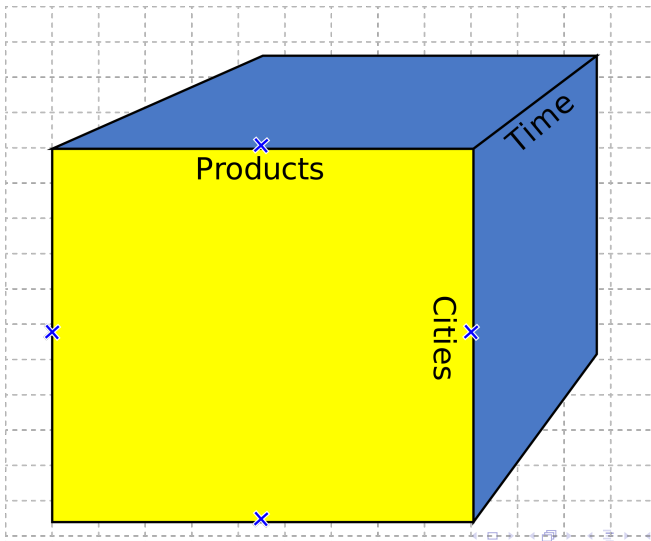
## Définition

OLAP est une approche d'analyse multi-dimensionnelle des données. Elle est basée sur le concept d'hypercube

## Hypercube

Une Hypercube (ou cube OLAP) est un tableau multi-dimensionnel de données - une feuille excel mais avec plus de dimensions....

# OLAP



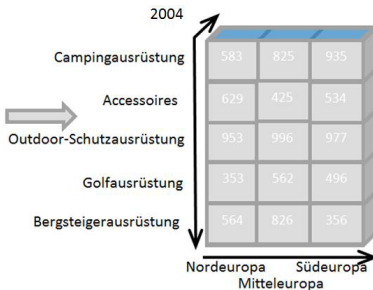
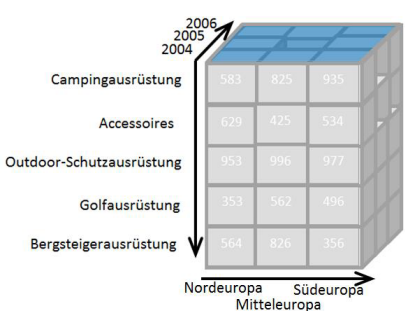
# OLAP

## Opérations usuelles

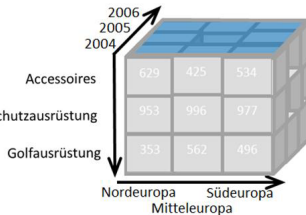
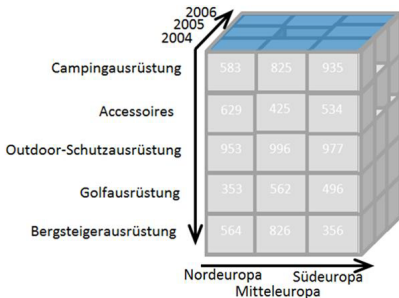
- Rotate : sélection du couple de dimensions à cibler,
- Slicing : extraction d'une tranche d'information,
- Scoping : extraction d'un bloc de données (opération plus générale que le slicing),
- Drill-up : synthèse des informations en fonction d'une dimension (exemple de drill-up sur l'axe temps : passer de la présentation de l'information jour par jour sur une année, à une valeur synthétique pour l'année),
- Drill-down : c'est l'équivalent d'un « zoom », opération inverse du drill-up,
- Drill-through : lorsqu'on ne dispose que de données agrégées (indicateurs totalisés), le drill through permet d'accéder au détail élémentaire des informations (voir notamment les outils H-OLAP).



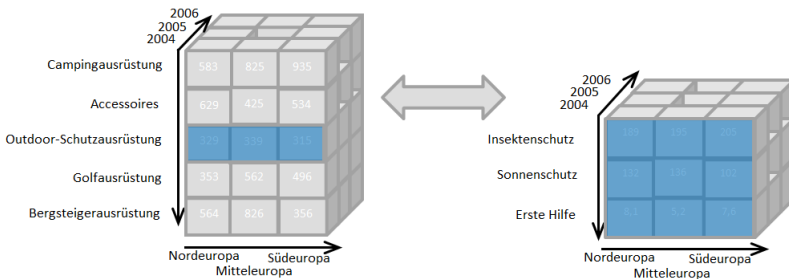
# OLAP - Slicing



# OLAP - Dicing

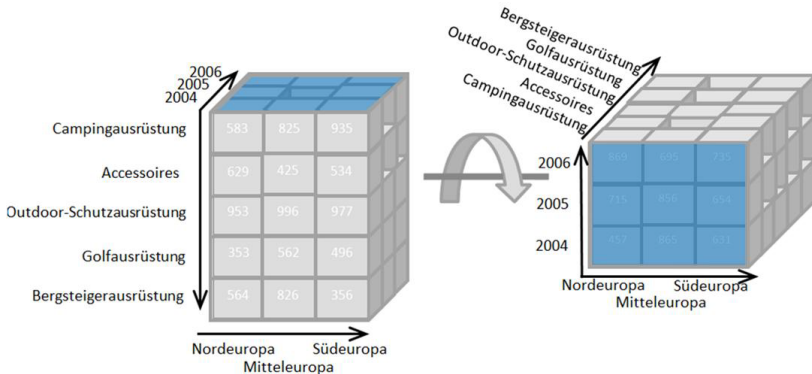


# OLAP - Drill-up and drill-down





# OLAP - Pivoting



# Fonction présentation

- Visualisation
- Reporting

## BI moderne

D'autres fonctions:

- Predictive Analysis : Machine Learning / Data Mining / ...  
(voir autre UEs)
- Real-time business intelligence (RTBI)
- Social BI
- ....

# Challenge Epidemium

Le challenge Epidemium va servir de base de travail en TP:  
<http://www.epidemium.cc>

## Epidemium

Epidemium est un programme de recherche scientifique participatif et ouvert dédié à la compréhension du cancer grâce aux Big Data, qui se concrétisera sous la forme d'un data challenge.