BIUM = Business Intelligence and User Modeling Master Data-Science

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

Sorbonne Université

28 janvier 2019

Contenu du cours

1. Business Intelligence (Laure Soulier)
   - Rôle de la gestion des données en entreprise
   - Agrégation, stockage des données à but décisionnel
2. User Modelling (Benjamin Piwowarski)
   - Analyse des données utilisateur et recommandation
   - Ouverture vers leur utilisation dans les entreprises

→ Intervenants industriels (en cours de discussion)
→ TME assurés par Clara Gainon de Forsan de Gabriac

Organisation

Horaires :

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

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 - UM

- ?
- ?
- ?

Objectifs

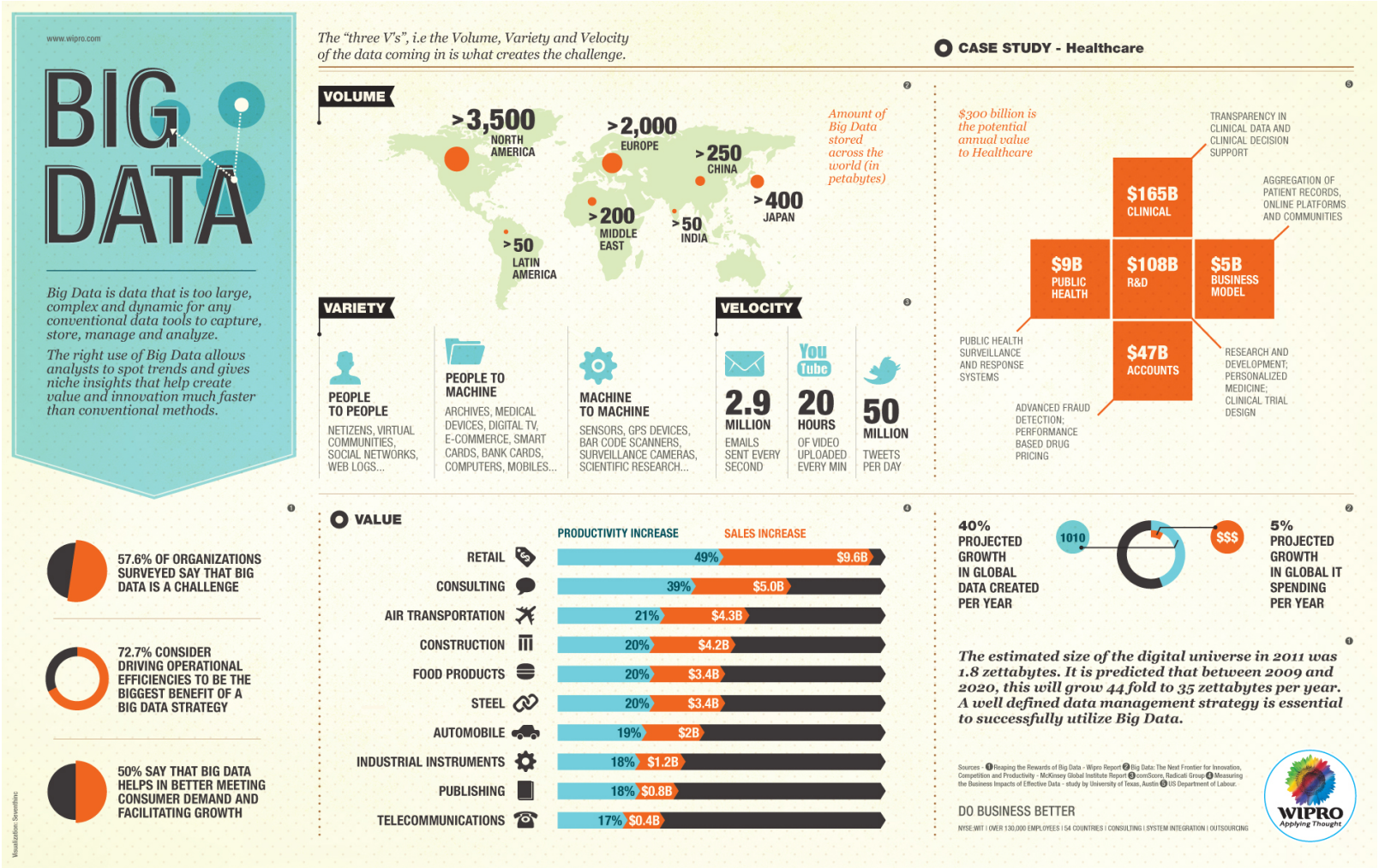
Etudiants ⇒ Data Integrator ⇒ Data Analyst ⇒ Data scientists

- 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

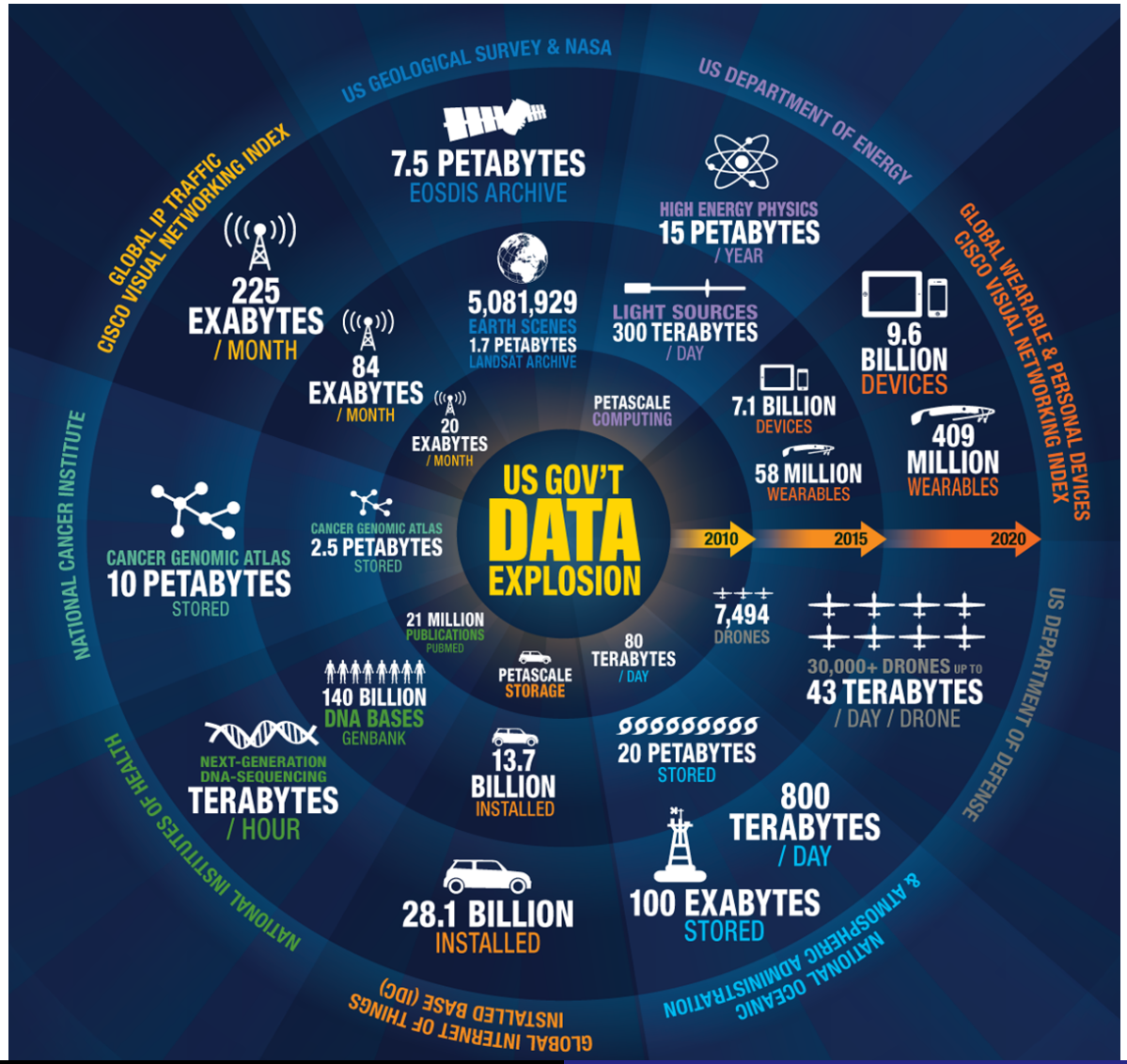
Et puis...

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

Contexte


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 V's", 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
NETIZENS, 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, BAR CODE SCANNERS, 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
AGGREGATION OF PATIENT RECORDS, ONLINE PLATFORMS AND COMMUNITIES
$9B PUBLIC HEALTH
$108B R&D
$5B BUSINESS MODEL
$47B ACCOUNTS
PUBLIC HEALTH SURVEILLANCE AND RESPONSE SYSTEMS
RESEARCH AND DEVELOPMENT; PERSONALIZED MEDICINE; CLINICAL TRIAL DESIGN
ADVANCED FRAUD DETECTION; PERFORMANCE BASED DRUG PRICING
57.6% 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
50% SAY THAT BIG DATA HELPS IN BETTER MEETING CONSUMER DEMAND AND FACILITATING GROWTH
VALUE
PRODUCTIVITY INCREASE
SALES INCREASE
RETAIL 49% $9.6B
CONSULTING 39% $5.0B
AIR TRANSPORTATION 21% $4.3B
CONSTRUCTION 20% $4.2B
FOOD PRODUCTS 20% $3.4B
STEEL 20% $3.4B
AUTOMOBILE 19% $2B
INDUSTRIAL INSTRUMENTS 18% $1.2B
PUBLISHING 18% $0.8B
TELECOMMUNICATIONS 17% $0.4B
40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR
1010
$$$
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 Applying Thought

Contexte


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)
7.5 PETABYTES
EOSDIS ARCHIVE
HIGH ENERGY PHYSICS
15 PETABYTES
/ YEAR
225
EXABYTES
/ MONTH
84
EXABYTES
/ MONTH
20
EXABYTES
/ MONTH
5,081,929
EARTH SCENES
1.7 PETABYTES
LANDSAT ARCHIVE
LIGHT SOURCES
300 TERABYTES
/ DAY
9.6
BILLION
DEVICES
7.1 BILLION
DEVICES
409
MILLION
WEARABLES
58 MILLION
WEARABLES
PETASCALE
COMPUTING
US GOV'T
DATA
EXPLOSION
2010
2015
2020
CANCER GENOMIC ATLAS
2.5 PETABYTES
STORED
CANCER GENOMIC ATLAS
10 PETABYTES
STORED
21 MILLION
PUBLICATIONS
PUBMED
80
TERABYTES
/ DAY
7,494
DRONES
30,000+ DRONES UP TO
43 TERABYTES
/ DAY / DRONE
PETASCALE
STORAGE
140 BILLION
DNA BASES
GENBANK
NEXT-GENERATION
DNA-SEQUENCING
TERABYTES
/ HOUR
13.7
BILLION
INSTALLED
20 PETABYTES
STORED
800
TERABYTES
/ DAY
28.1 BILLION
INSTALLED
100 EXABYTES
STORED

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.

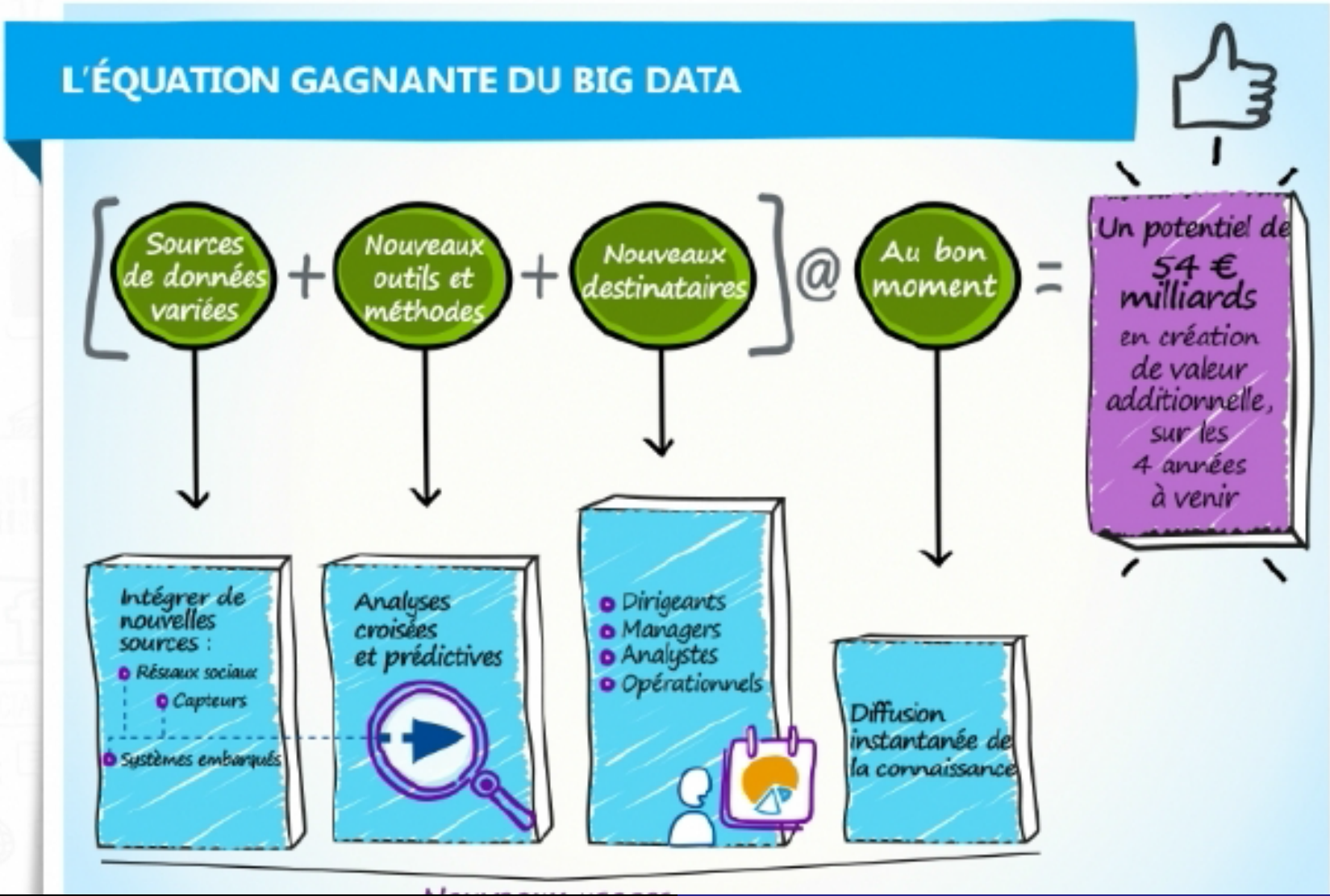
* *Enjeux Business des données - CIGREF 2014*

Traitement de données en entreprise

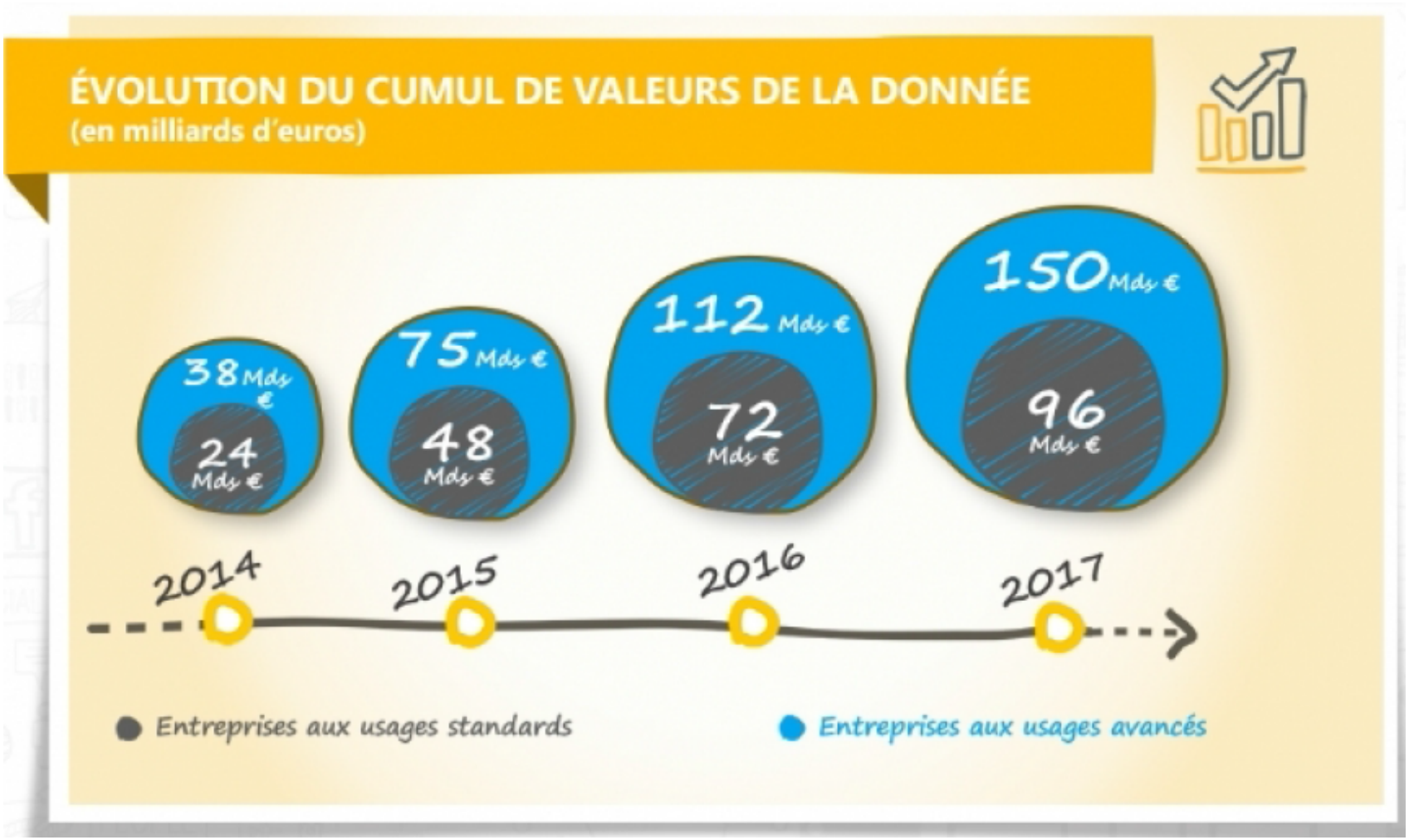
...

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*

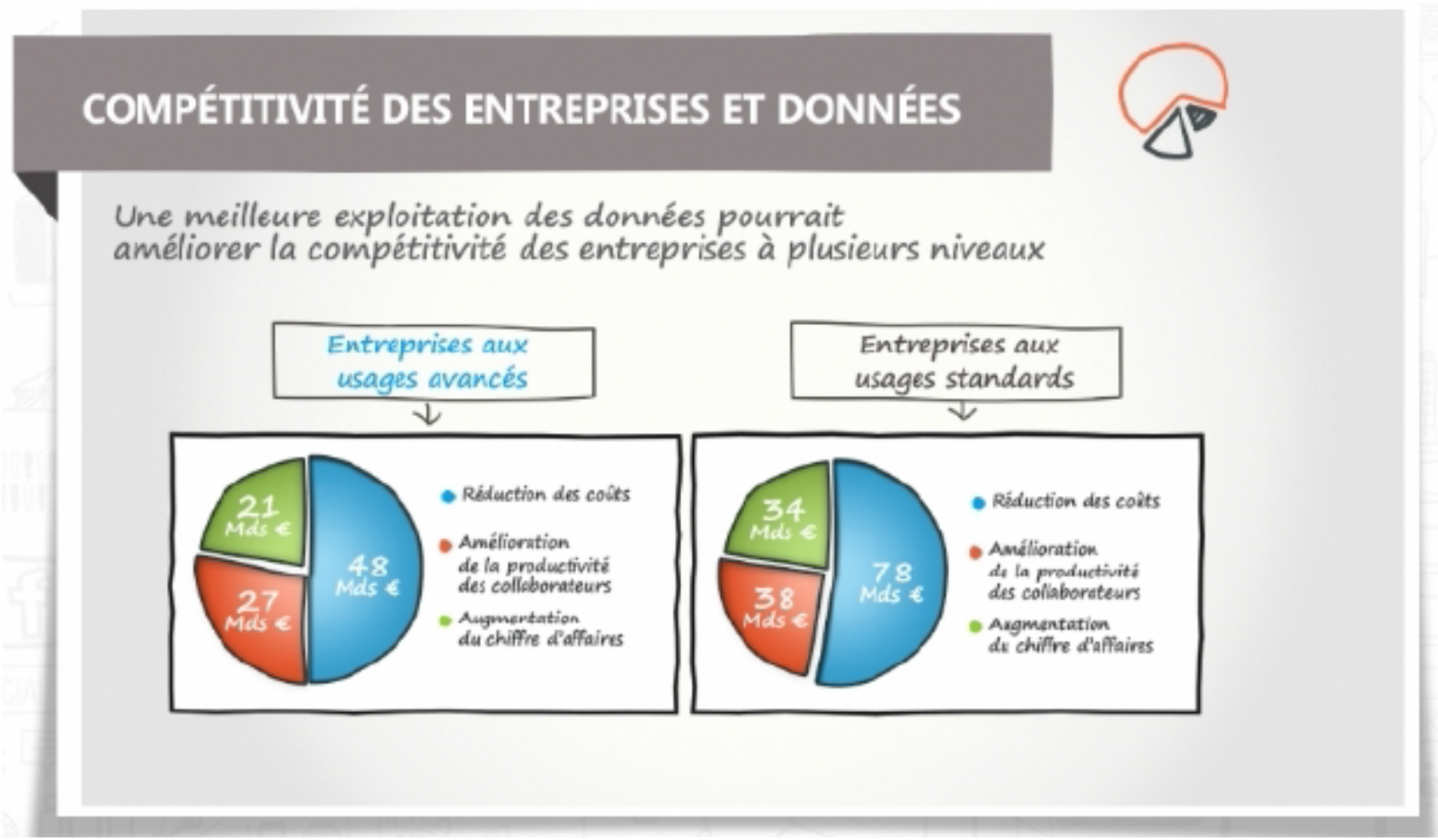
Etude IDC - Microsoft 2014



Etude IDC - Microsoft 2014



Etude IDC - Microsoft 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

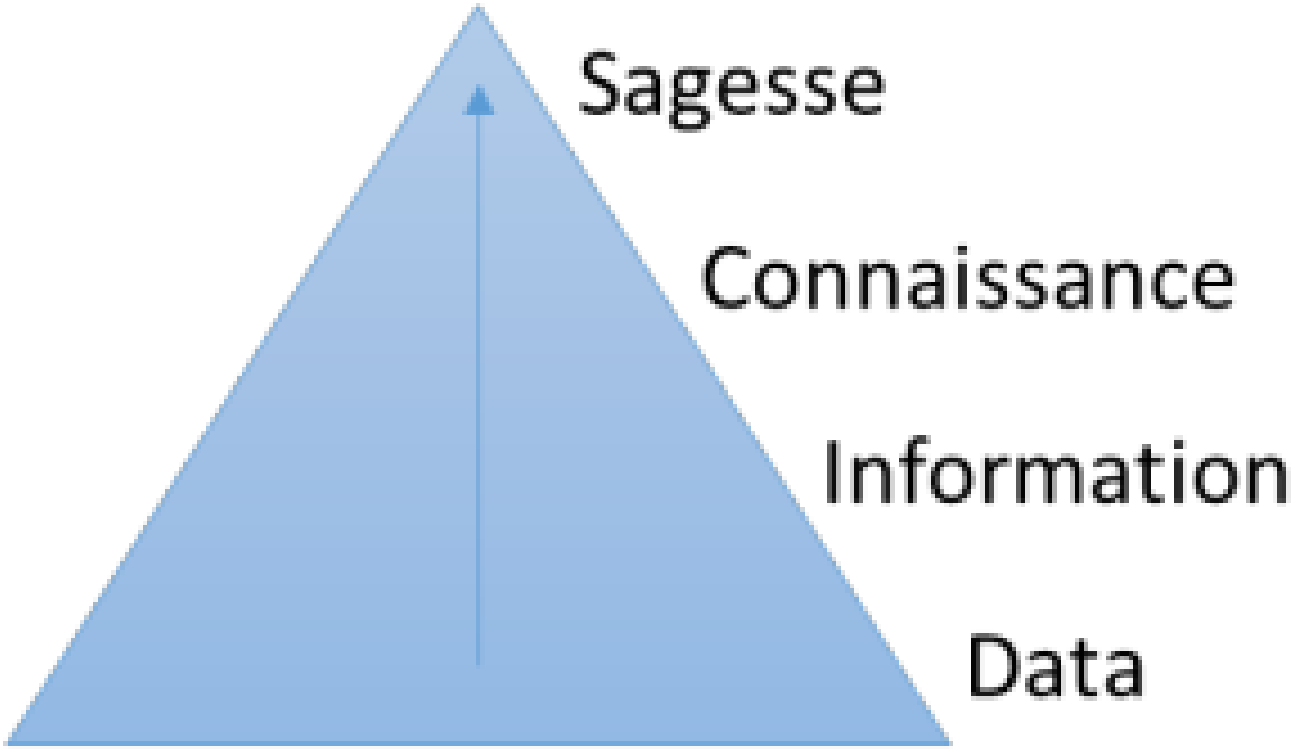
source : Rapport CIGREF 2009

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

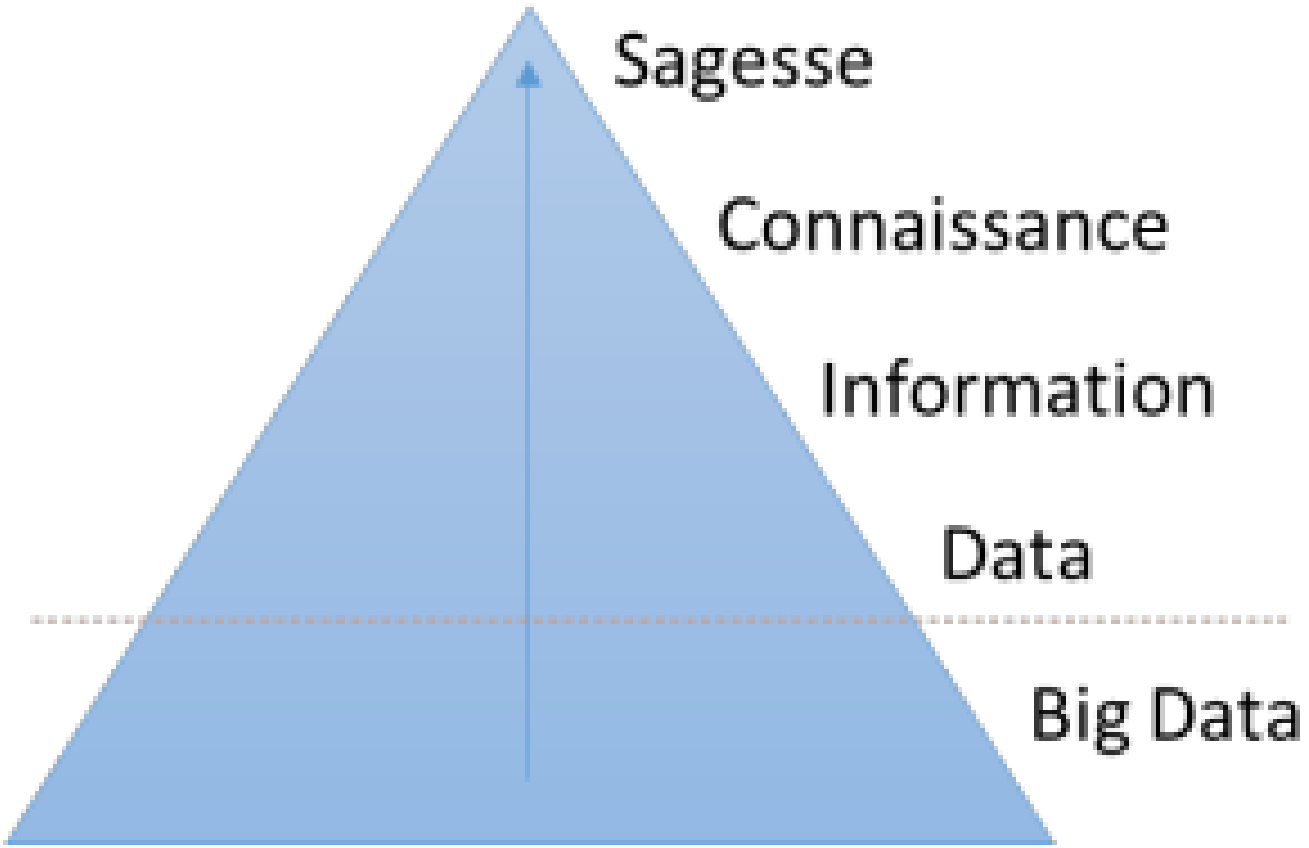
But :



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

La pyramide de la BI

But :

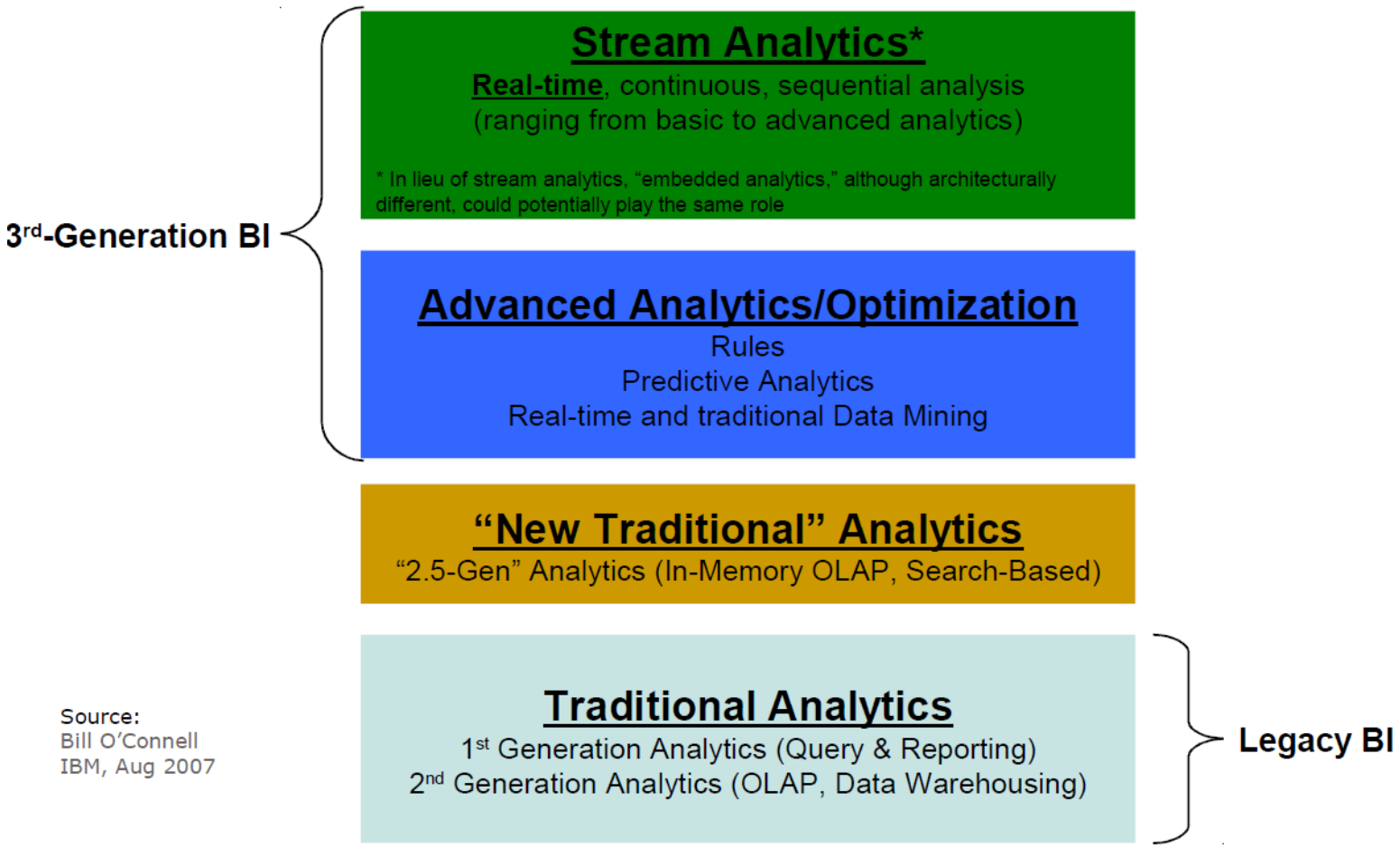


source : Blog : La BI ca vous gagne, Ah non, c'est pas la BI, c'est la montagne

Historique

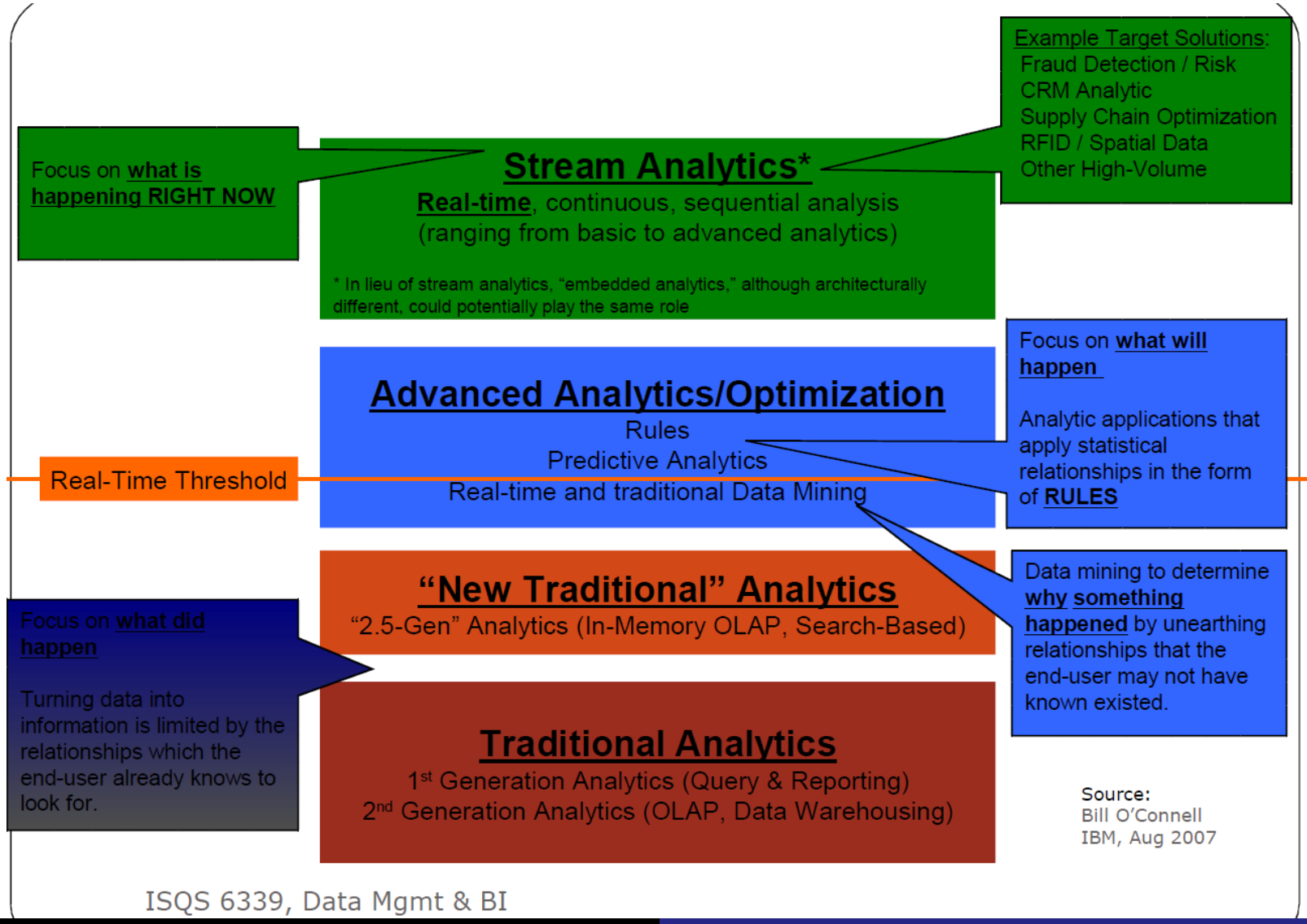
- 1st Generation - Traditional analytics (query and reporting)
- 2nd Generation - Traditional generation (OLAP, data warehousing)
- 2.5nd Generation - New traditional generation
- 3rd Generation - Advanced analytics Rules, predictive analytics and realtime data mining Stream analytics

Historique

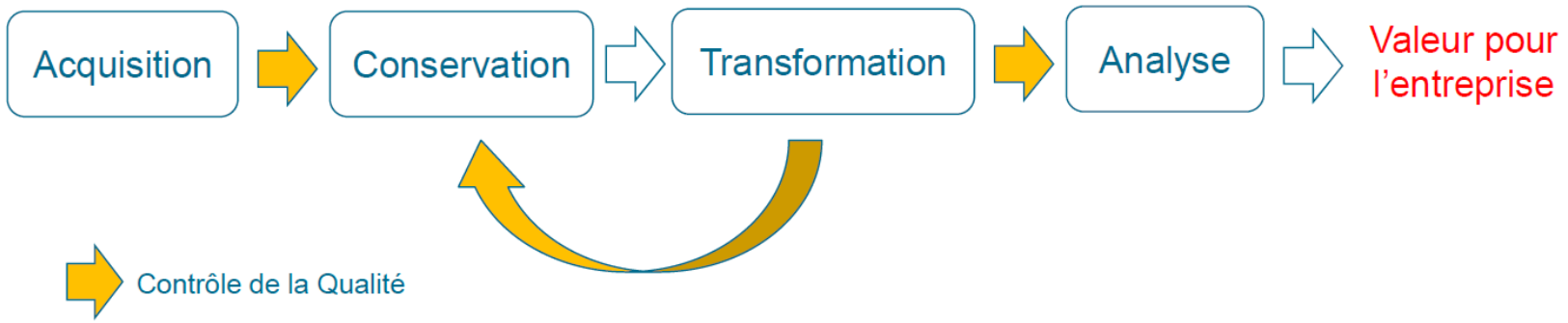


ISQS 6339, Data Mgmt & BI

Historique


Example Target Solutions:
Fraud Detection / Risk
CRM Analytic
Supply Chain Optimization
RFID / Spatial Data
Other High-Volume
Focus on what is happening RIGHT NOW
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
Focus on what will happen
Analytic applications that apply statistical relationships in the form of RULES
Advanced Analytics/Optimization
Rules
Predictive Analytics
Real-time and traditional Data Mining
Real-Time Threshold
Data mining to determine why something happened by unearthing relationships that the end-user may not have known existed.
"New Traditional" Analytics
"2.5-Gen" Analytics (In-Memory OLAP, Search-Based)
Focus on what did happen
Turning data into information is limited by the relationships which the end-user already knows to look for.
Traditional Analytics
1st Generation Analytics (Query & Reporting)
2nd Generation Analytics (OLAP, Data Warehousing)
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

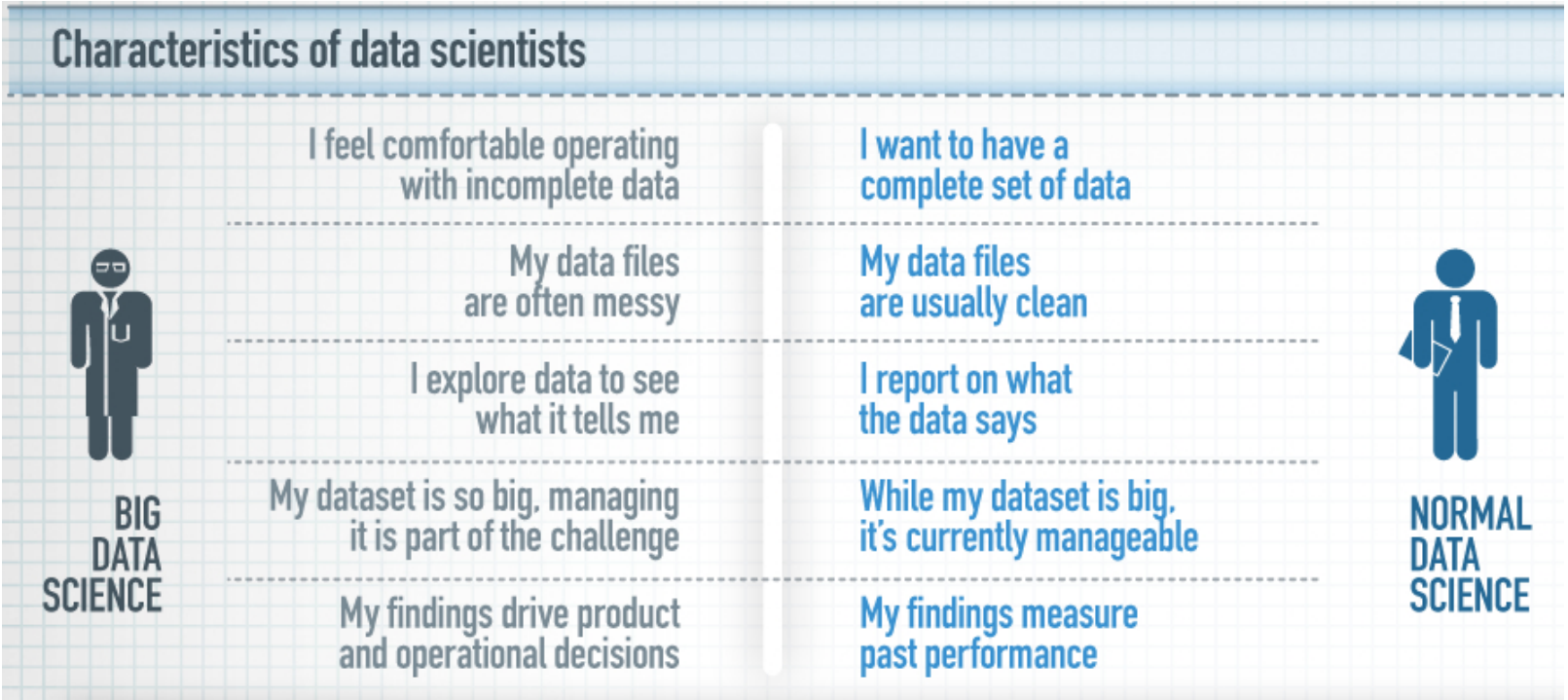
Data Scientist

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.

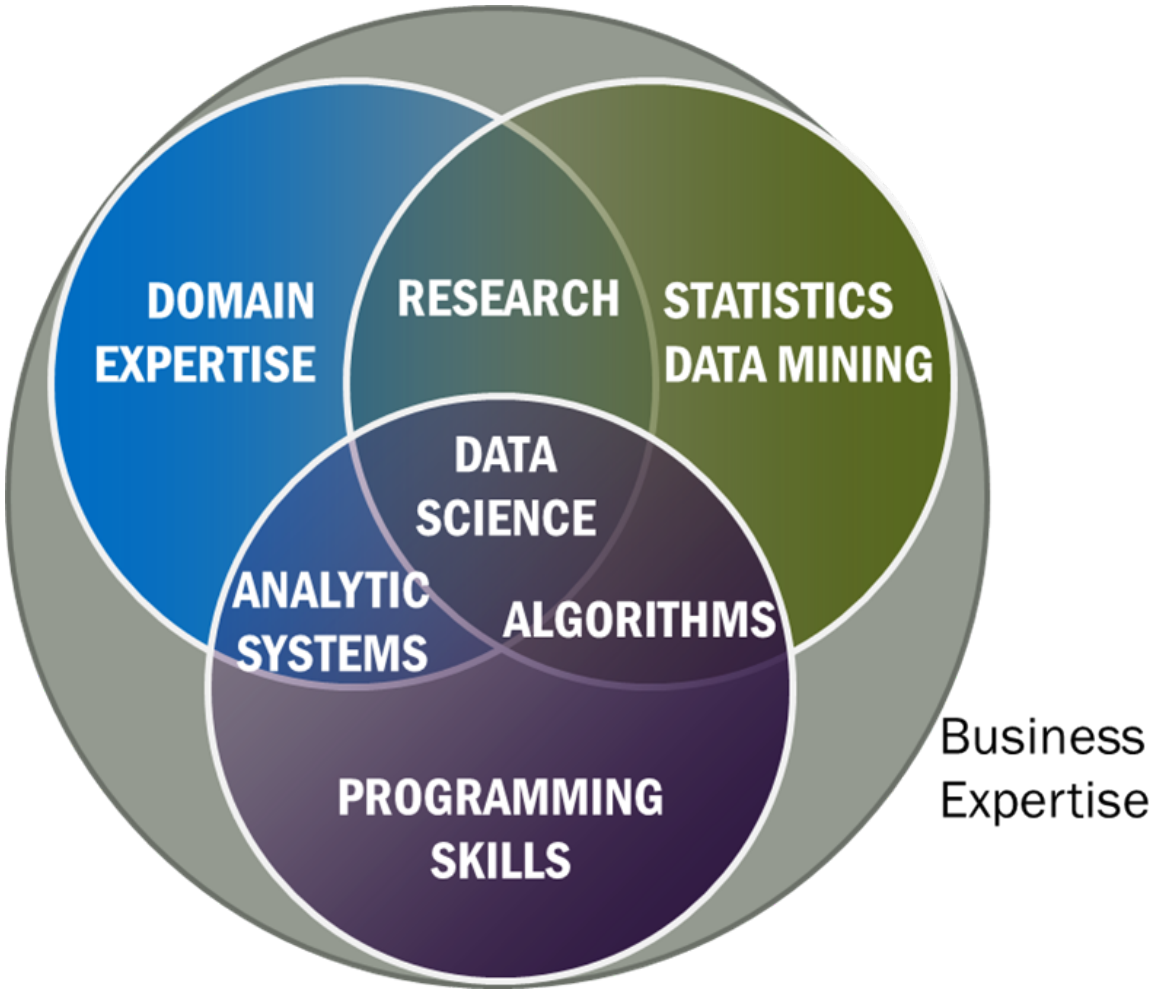
Data Steward - Responsable des Données

[...] 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.

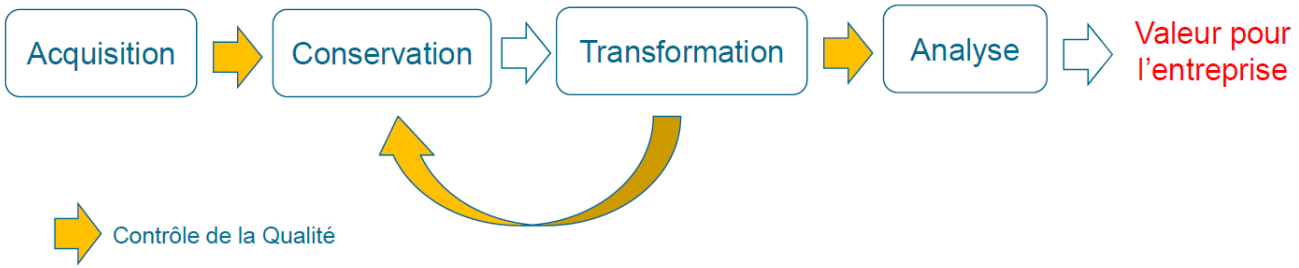
Nouveau Métier



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, comptabilité, ressources humaines, etc.) à travers un système de sécurité ou de datamart spécialisés à des fins de consultations, d'analyse, d'alertes prédéfinies,

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

ETL

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

ETL = Extract, Tranform, Load

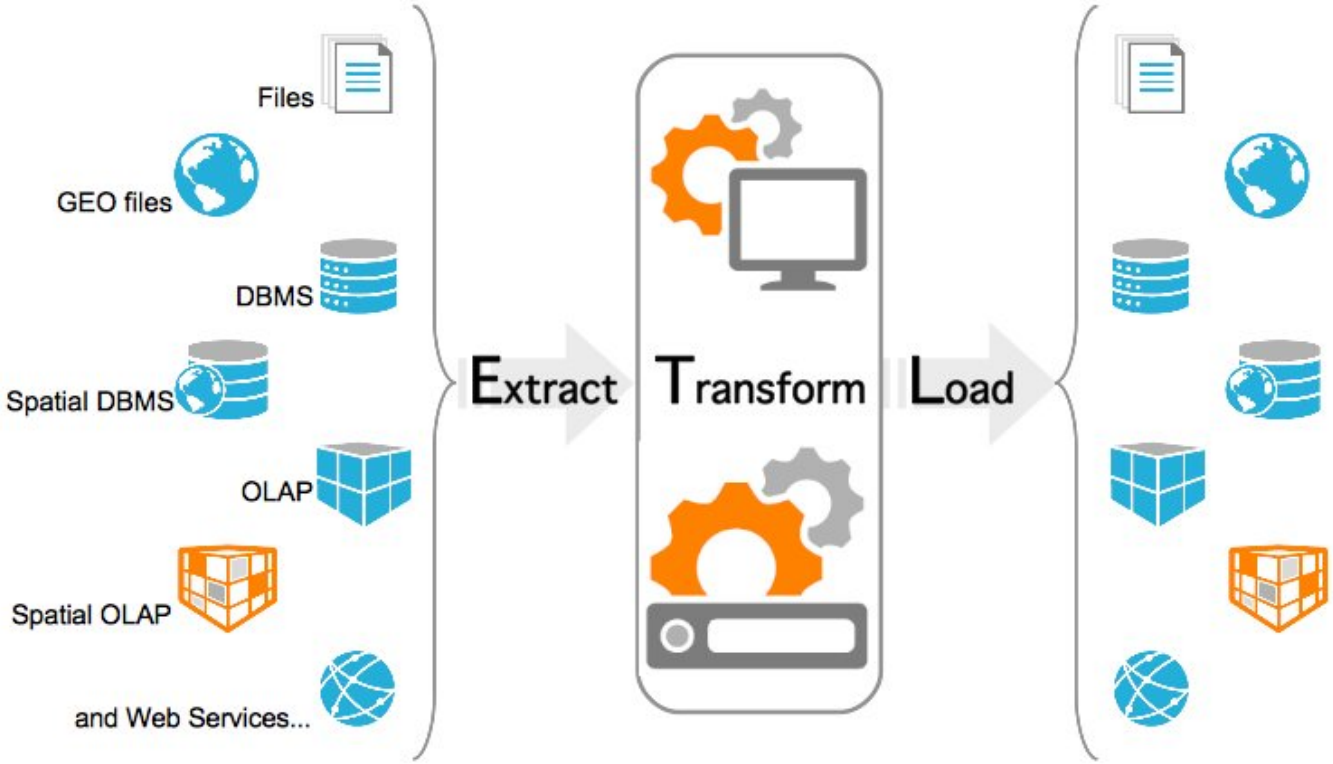
Processus :

- **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

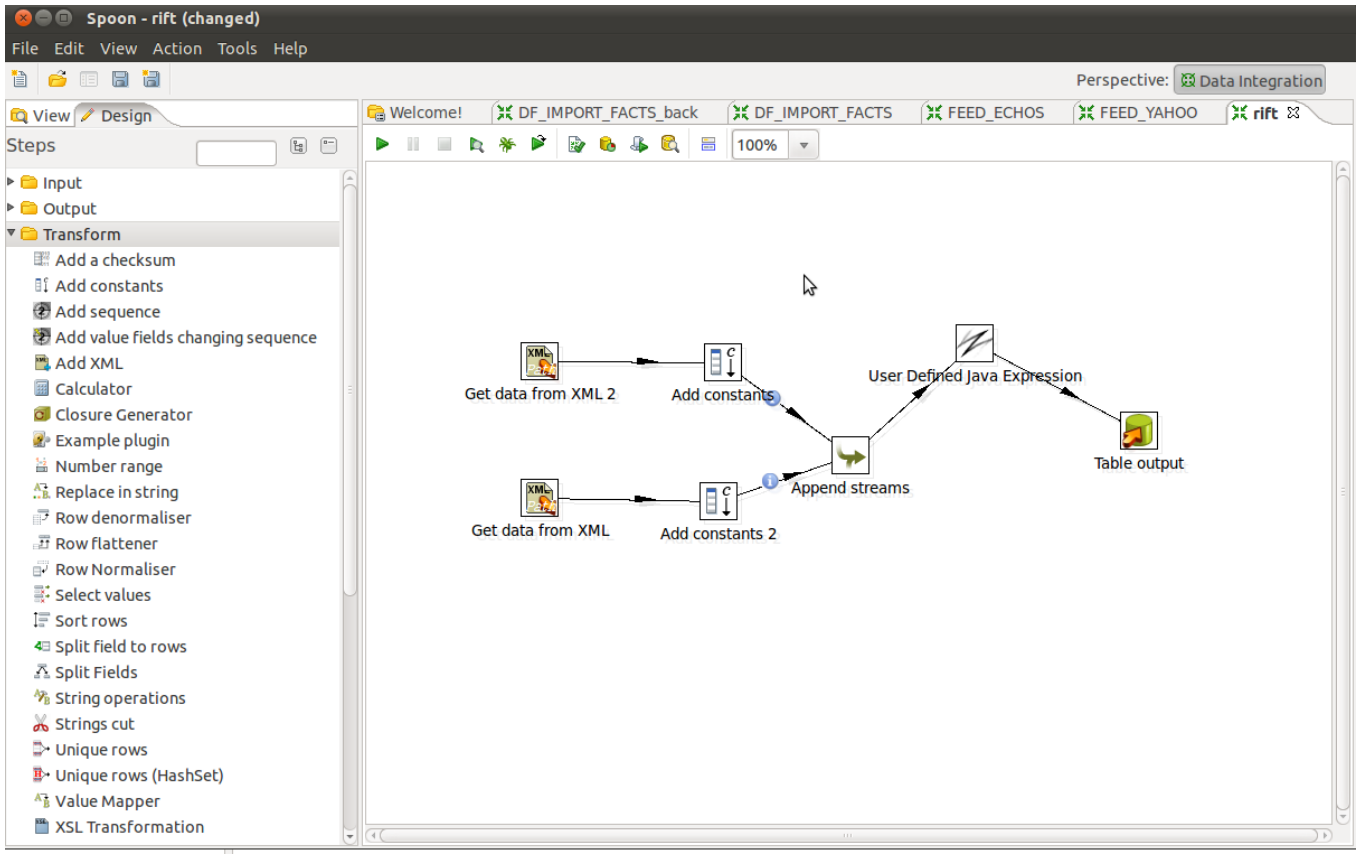


Logiciels ETL

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 - 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 synchronization
- La certification

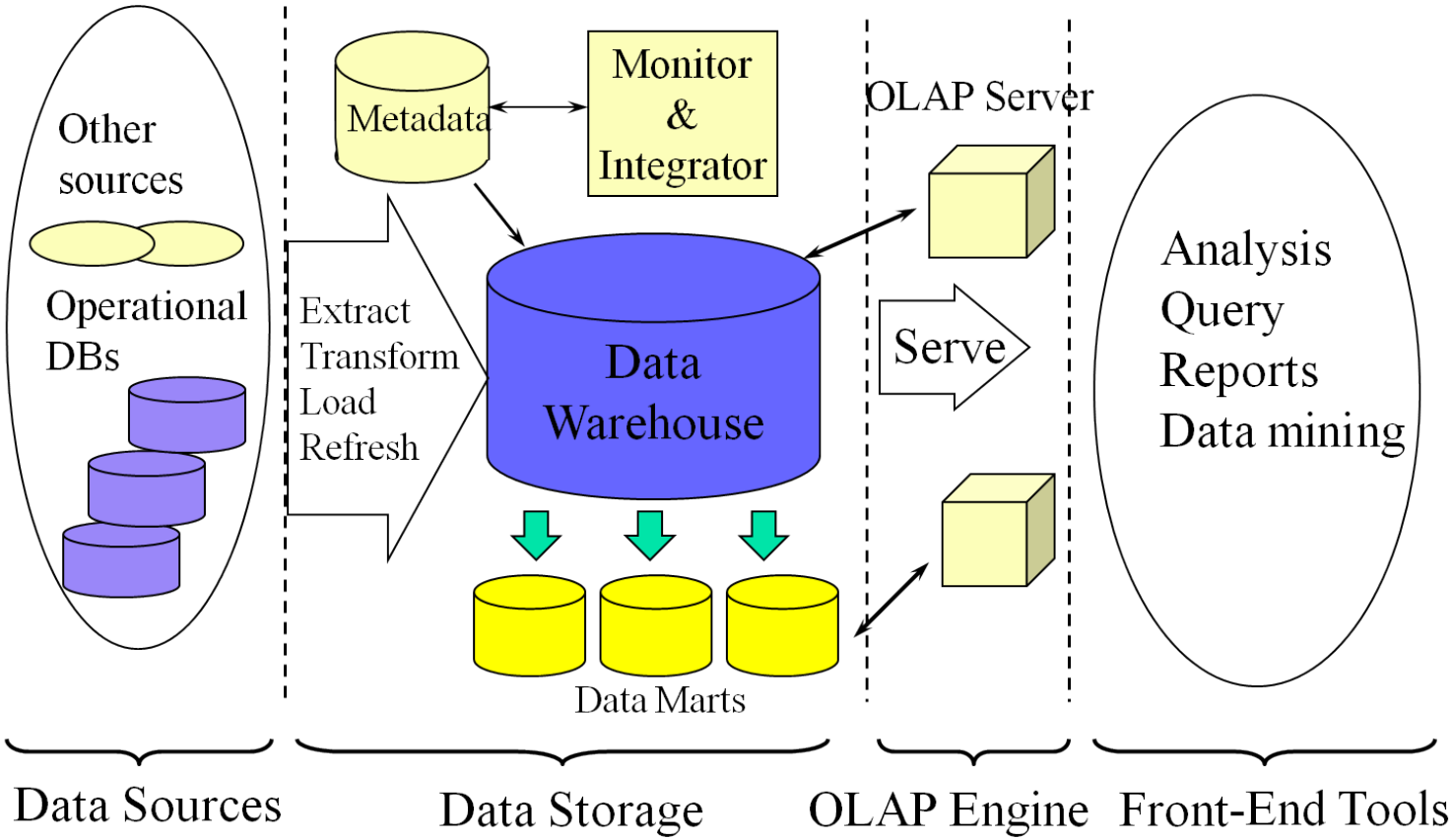
Data warehouse

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 donnees 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

Data warehouse - outils

Grands acteurs

IBM, Oracle, Teradata, Microsoft et Sybase IQ.

+ Tous les SGBDs.....

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.
Wikpiedia

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

OLAP - online analytical processing

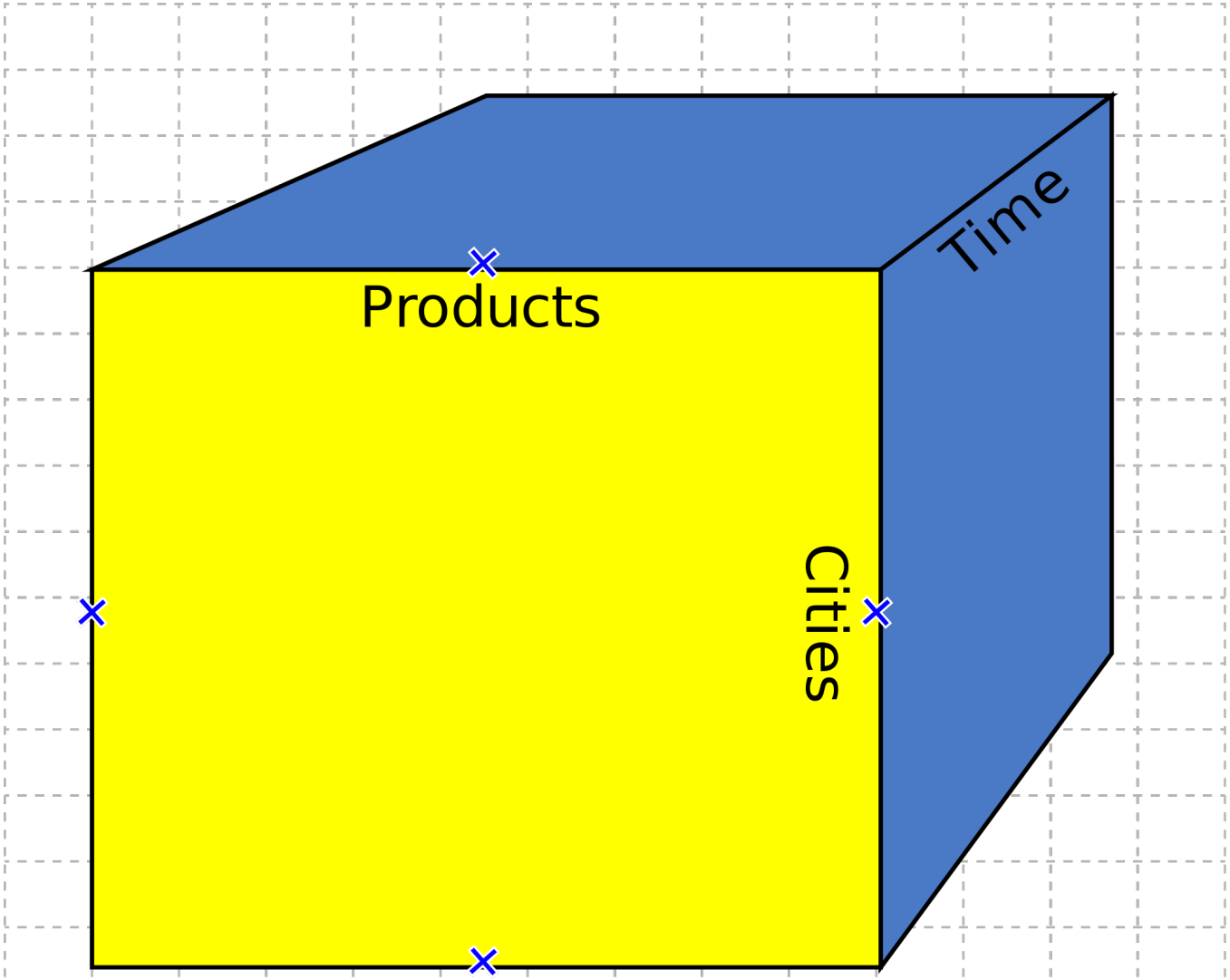
Définition

OLAP est une approche d'analyse multi-dimensionelle 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

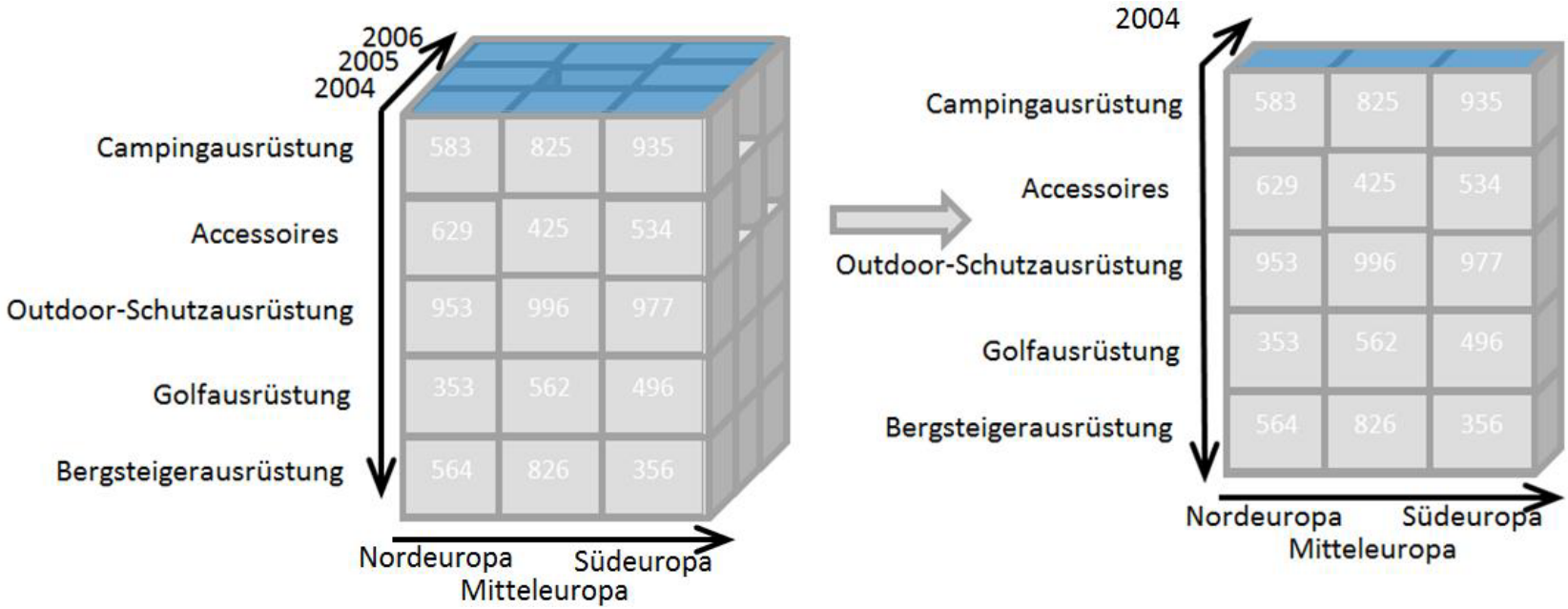


OLAP

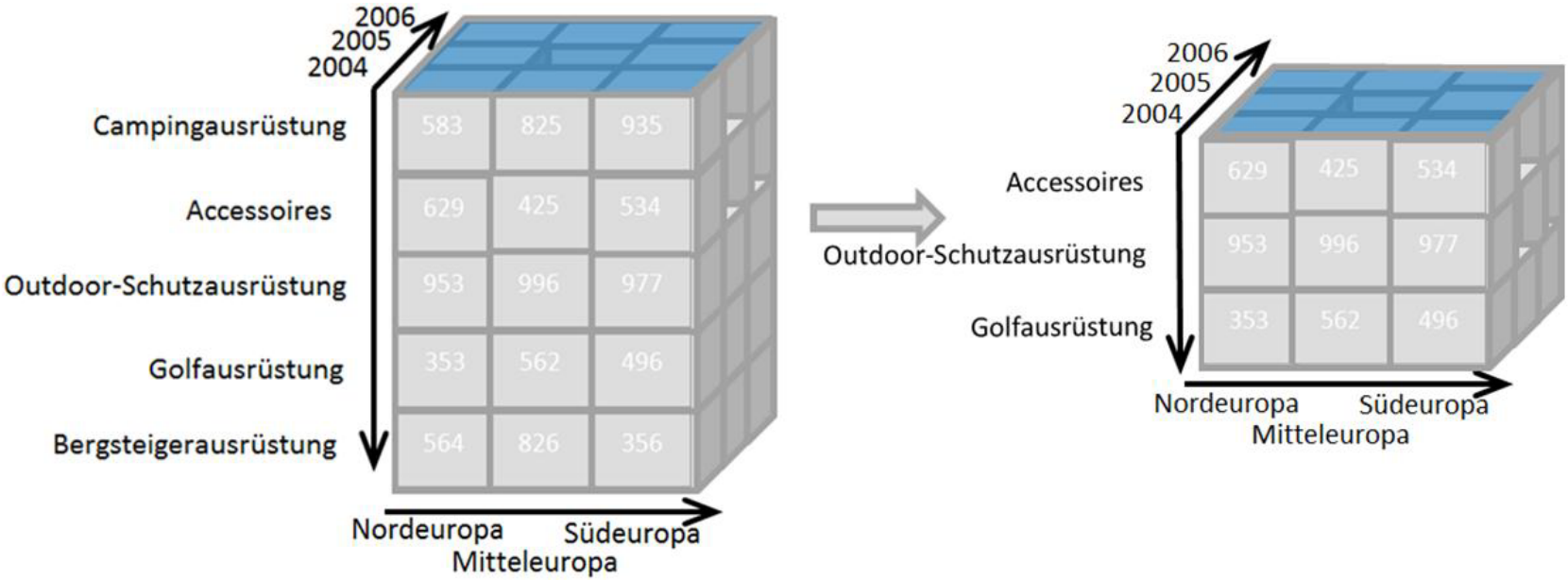
Opérations usuelles

- 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).

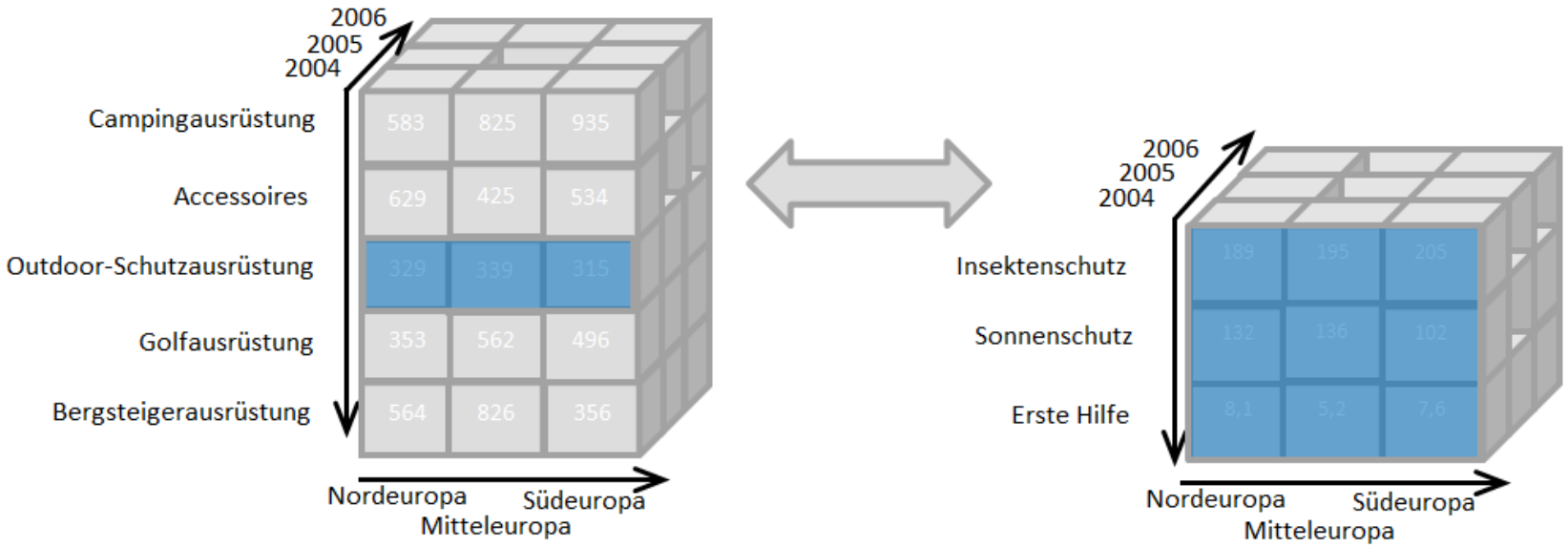
OLAP - Slicing



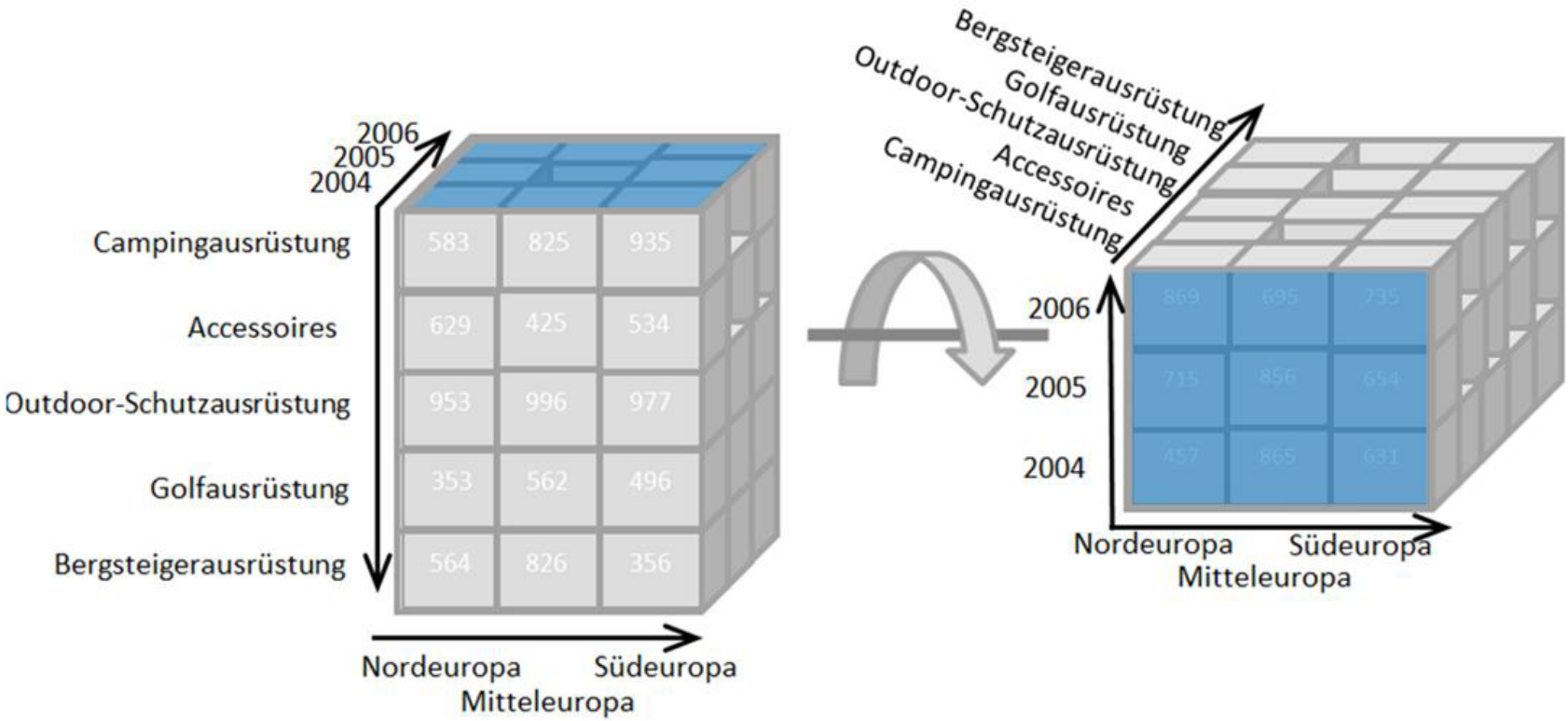
OLAP - Dicing


2006
2005
2004
Campingausrüstung
Accessoires
Outdoor-Schutzausrüstung
Golfausrüstung
Bergsteigerausrüstung
583 825 935
629 425 534
953 996 977
353 562 496
564 826 356
Nordeuropa
Südeuropa
Mitteleuropa
Outdoor-Schutzausrüstung
2006
2005
2004
Accessoires
Golfausrüstung
629 425 534
953 996 977
353 562 496
Nordeuropa
Südeuropa
Mitteleuropa

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
- ....