Business Intelligence - Introduction

Olivier Schwander <olivier.schwander@lip6.fr>

UPMC

Organisation du cours

`http://www-connex.lip6.fr/~schwander/enseignement/2015-2016/m2stat_bi/`

Horaires et salles

- Mardi de 14h à 17h, salle 1525-101 ou 1525-102
- Cours puis TD/TP (mais pas toujours)

Contenu

- Business intelligence
- Bases de données, extraction de données
- Interventions d'industriels

Évaluation

- Note de TP et travail à la maison
- Examen final

Inspiration

Cours de Ludovic Denoyer

- Master 1 *Données Apprentissage Connaissances*
- Beaucoup plus d'heures
- Pas le même public (informaticiens)

Cours de Bernard Espinasse

- Ecole Polytechnique Universitaire de Marseille
- Public encore plus spécialisé

Article Wikipedia *Informatique décisionnelle*

Autres sources

- Indiquées au fur et à mesure

Objectifs

Analyse de données pour l'entreprise

- 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

Analyse de données en pratique

- Donner des éléments de bases de données
- Présenter des cas concrets d'extaction de données

Contexte

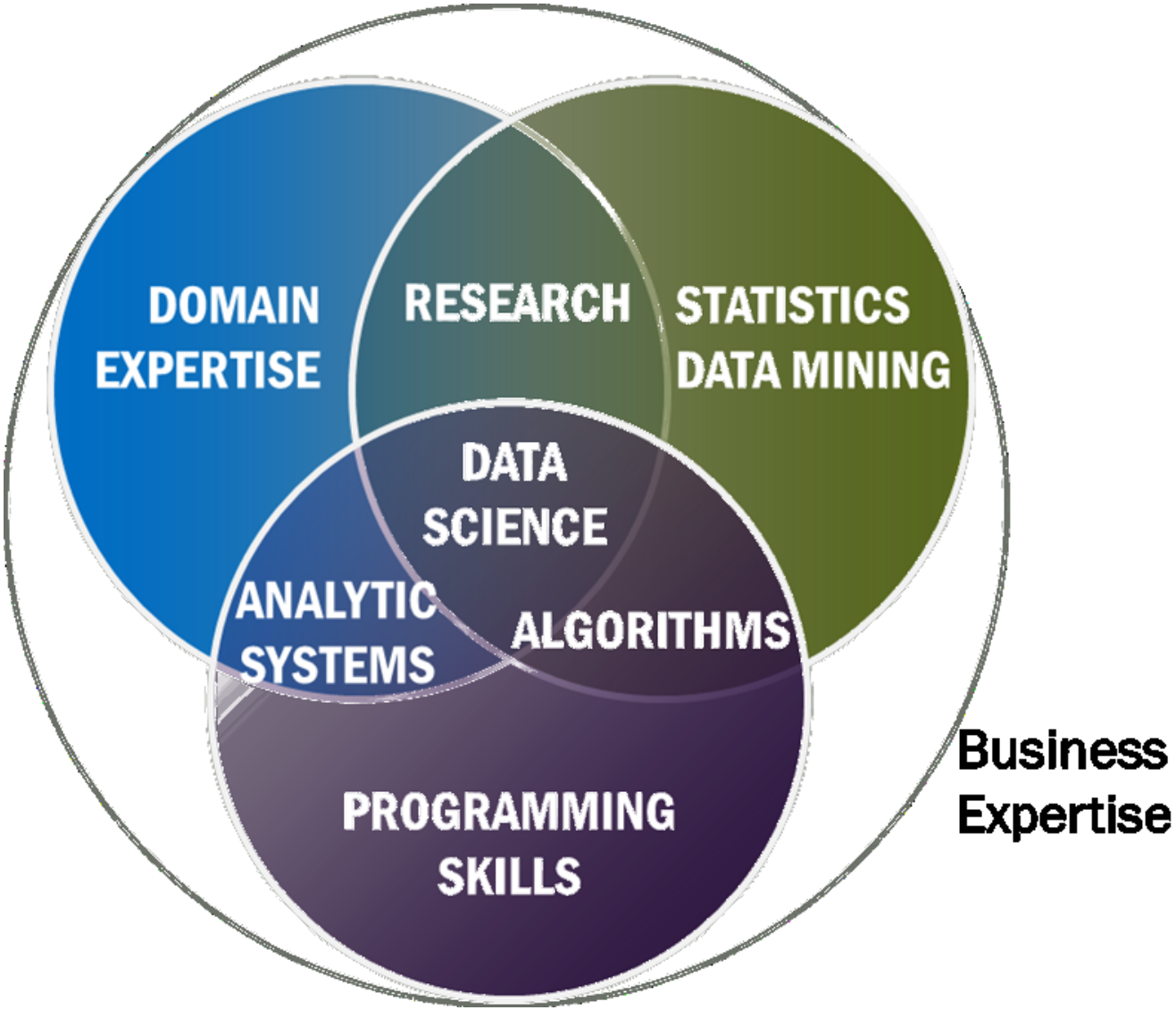
Entreprise

- On veut gagner de l'argent
- On cherche à faire des choix intelligents
- On peut collecter *beaucoup* de données
- On a les ressources pour les traiter
- On cherche les compétences pour les traiter

Vous

- Des mathématiciens, des statisticiens
- Des étudiants à intégrer dans le monde du travail
- Lien à faire entre vos compétences et le vocabulaire et les besoins de l'entreprise

Contexte


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

Définition

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.

Motivation

Enjeux Business des données - CIGREF 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.

Mots-clés

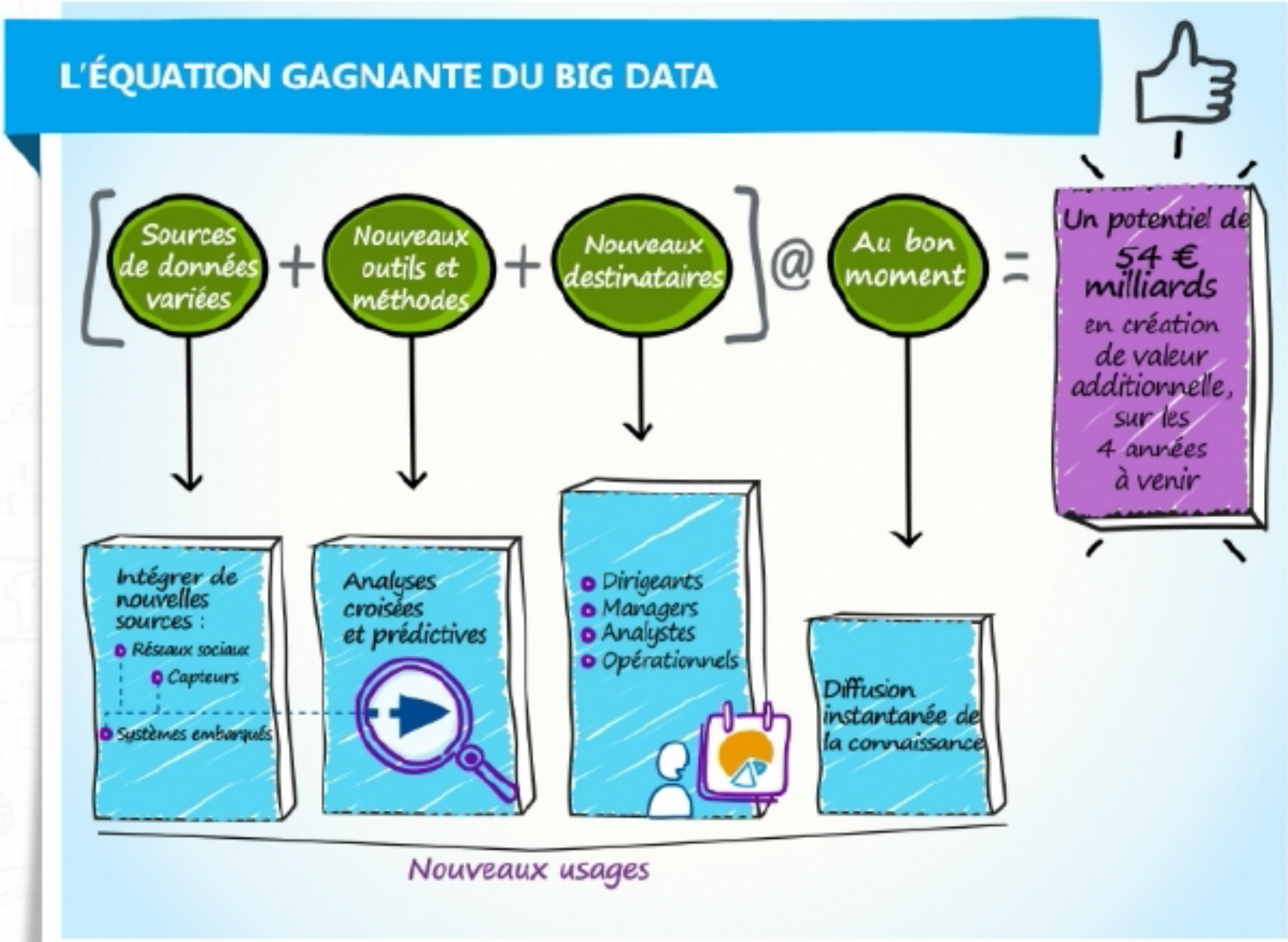
- données, information, connaissance, savoir
- optimiser, partager, sources diverses

Applications

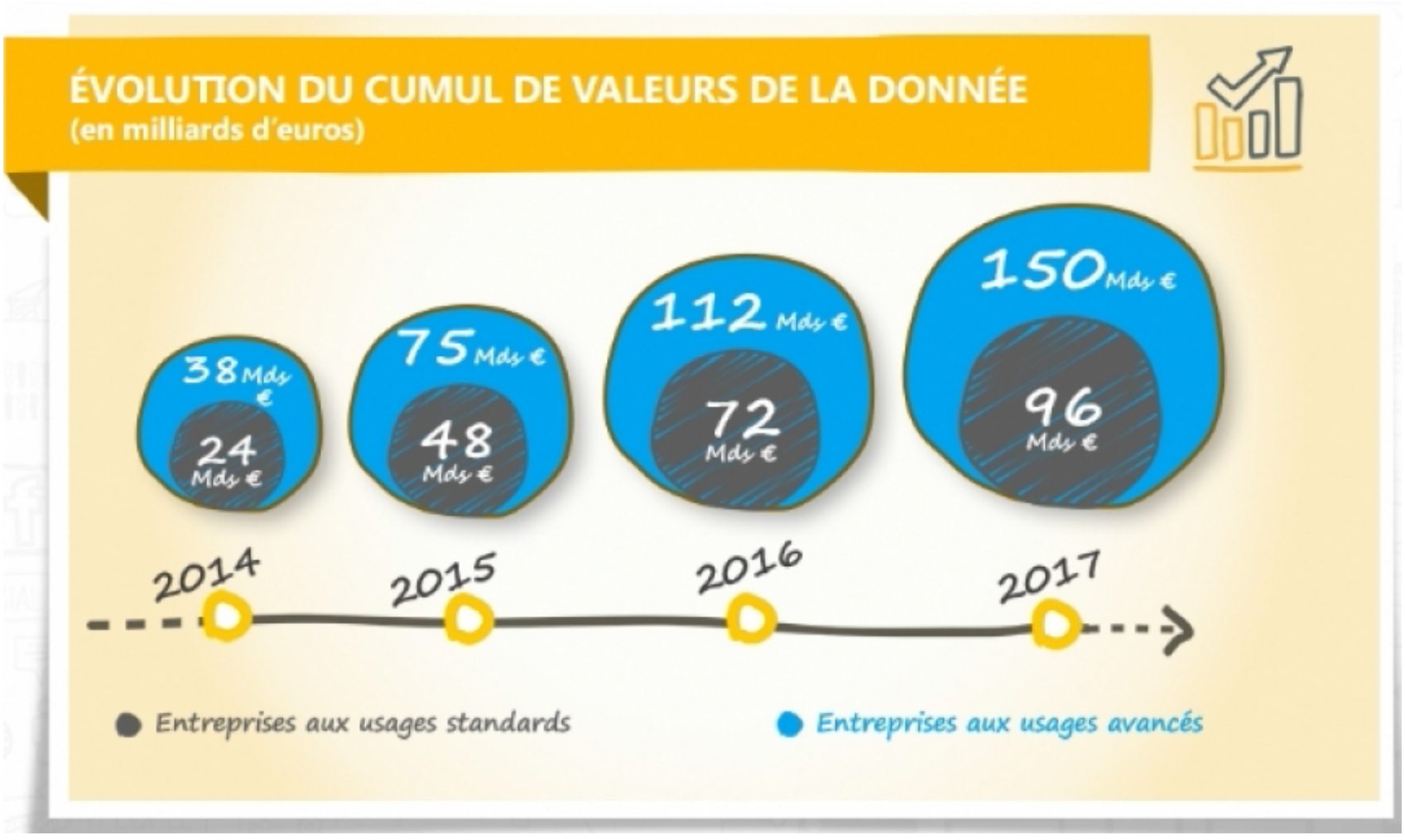
source : Rapport CIGREF 2009

- *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 ;

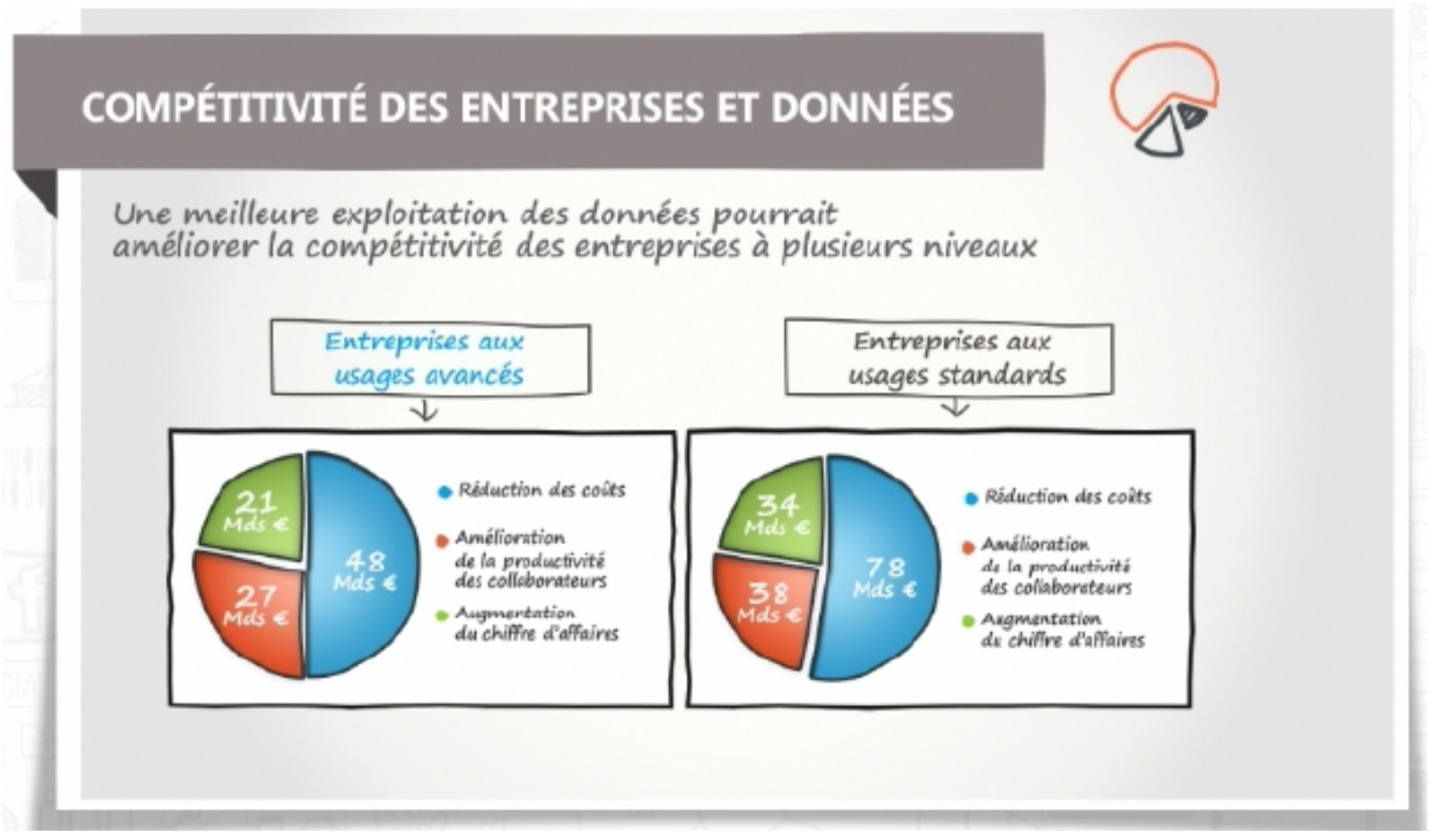
Etude IDC - Microsoft 2014



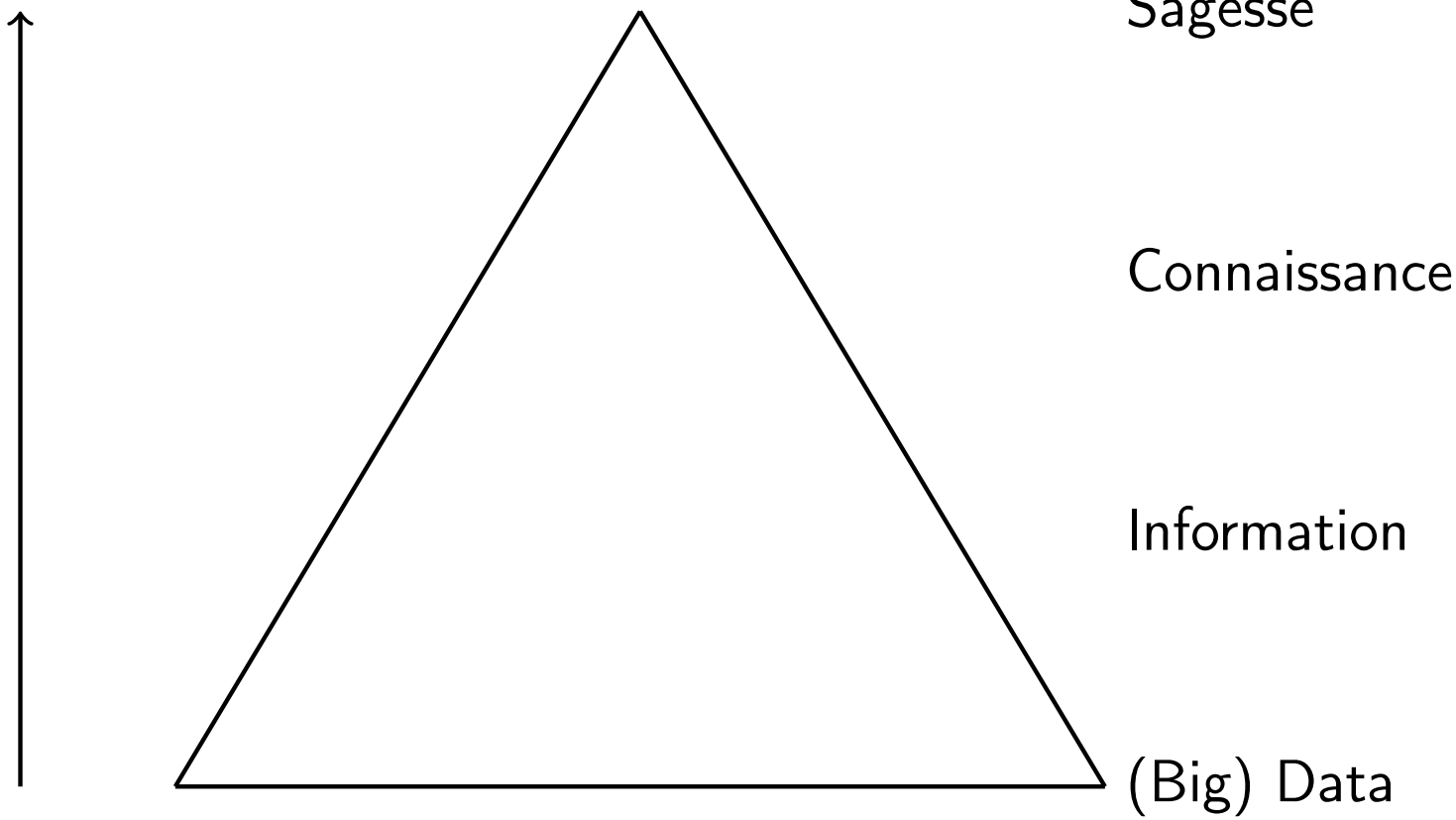
Etude IDC - Microsoft 2014



Etude IDC - Microsoft 2014



Pyramide du BI



Les métiers du BI

4 métiers

- Data Integrator
- Data Analyst
- Data Scientist
- Data Steward (Responsable des données)

Bas niveau

Data Integration

- Combiner des informations hétérogènes venants de sources différentes

Data Analysis

- Inspection, nettoyage, transformation et modélisation des données.
- Data Mining, Data Vizualisation
- Rendre la donnée compréhensible
- Communiquer à partir de la donnée

Haut niveau

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.

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,

Architecture

Extraction des données

- Bases de données
- Autres sources

Structuration des données

- Prétraitements
- Aggrégation
- Interface

Présentation des données

- Visualisation, alertes automatiques
- Pour une tâche donnée
- À destination d'un décideur

Stockage

Base de données opérationnelle

- Fonctionnement normal de l'entreprise
- Pas forcément un historique très grand
- Peut changer dans le temps

Datawarehouse

- Stockage pour le BI
- Archivage sur toute l'histoire de l'entreprise
- Format stable dans le temps

Datamart

- Vue métier
- À destination du décideur

Définitions

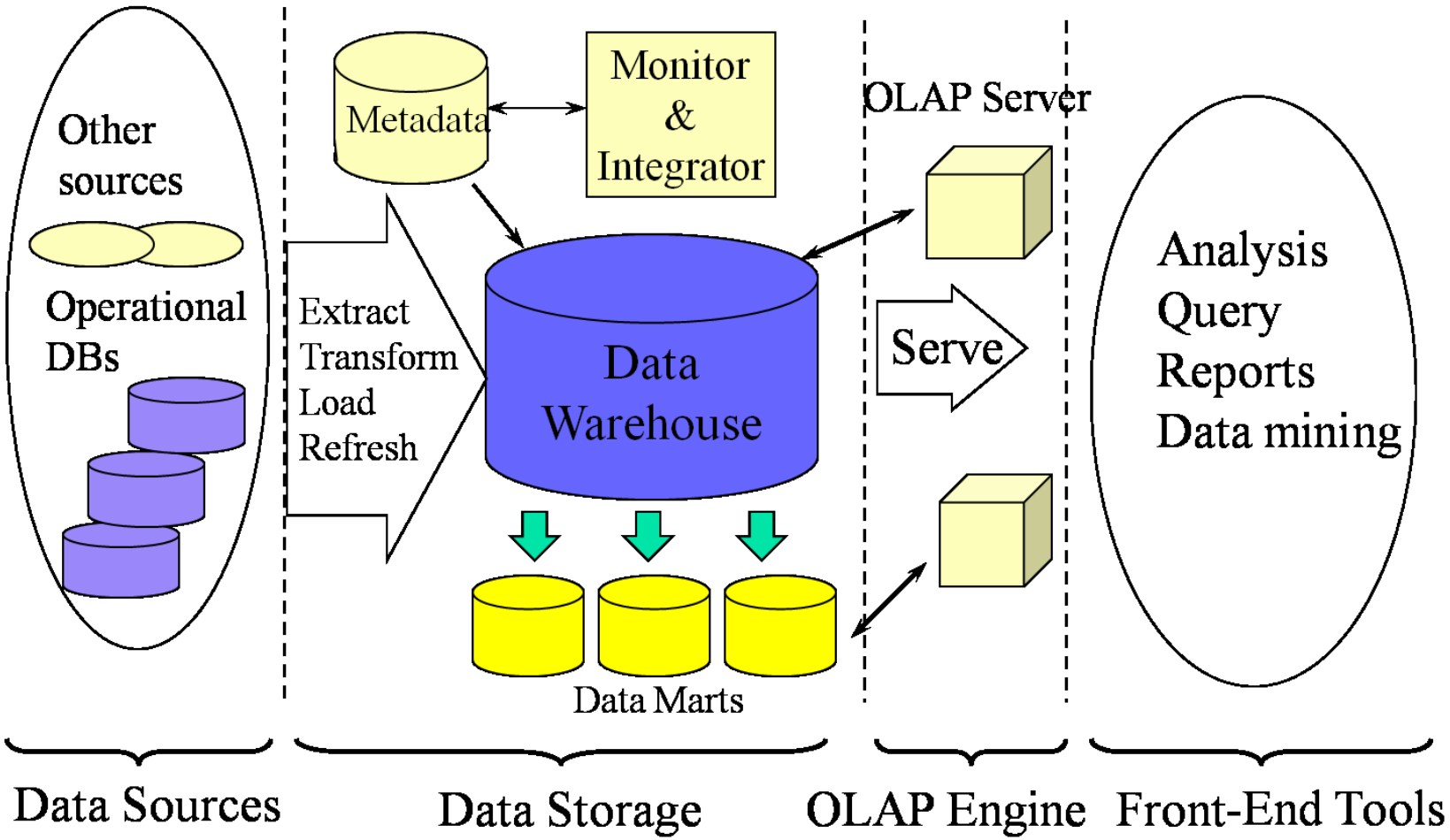
Datawarehouse

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.

Datamart

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.

Datawarehouse



Les fonctions

- Collecte de données
- Intégration
- Diffusion (ou distribution)
- Présentation

Fonction de collecte

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

Tâche

- Récupérer les données
- Méthodologie ETL

Données hétérogènes

Plusieurs types de sources

- Fichiers plats
- Fichers Excel
- Bases de données (SQL)
- Services web
- Systèmes de stockages pour données massives
- Interfaces exotiques

Plusieurs types de données

- Chiffres, texte, image
- Données statiques, flux
- Données bruitées, manquantes, erronées

Flux de données et données statiques

Données statiques

- Image à un instant donné de l'état de l'entreprise
- Rapports d'activité, bilans, inventaire

Flux de données

- Mise à jour en temps réel
- Compte rendus quotidiens, commandes, livraisons

Recodage

Mise sous forme canonique

- Choix d'une représentation unique
- Indépendante de la représentation en entrée

Stabilité dans le temps

Un changement dans les formats d'entrées en doit pas perturber l'analyse.

ETL

Méthodologie et outils

Extract

- Extraire les données de sources hétérogènes

Transform

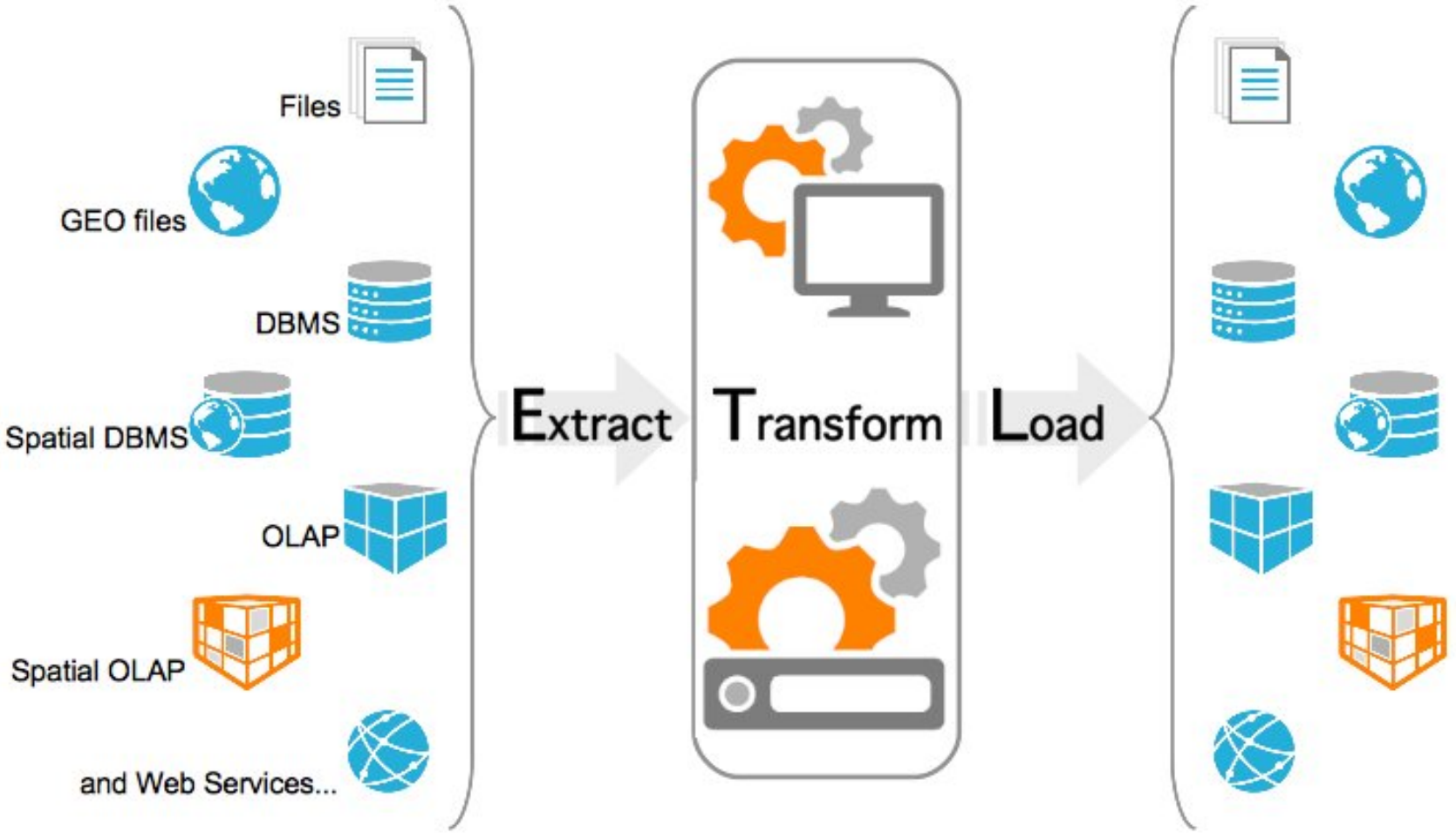
- Transformation des données pour les mettre dans un format acceptable

Load

- Charger les données dans le datawarehouse

ETL

Ensemble de connecteurs

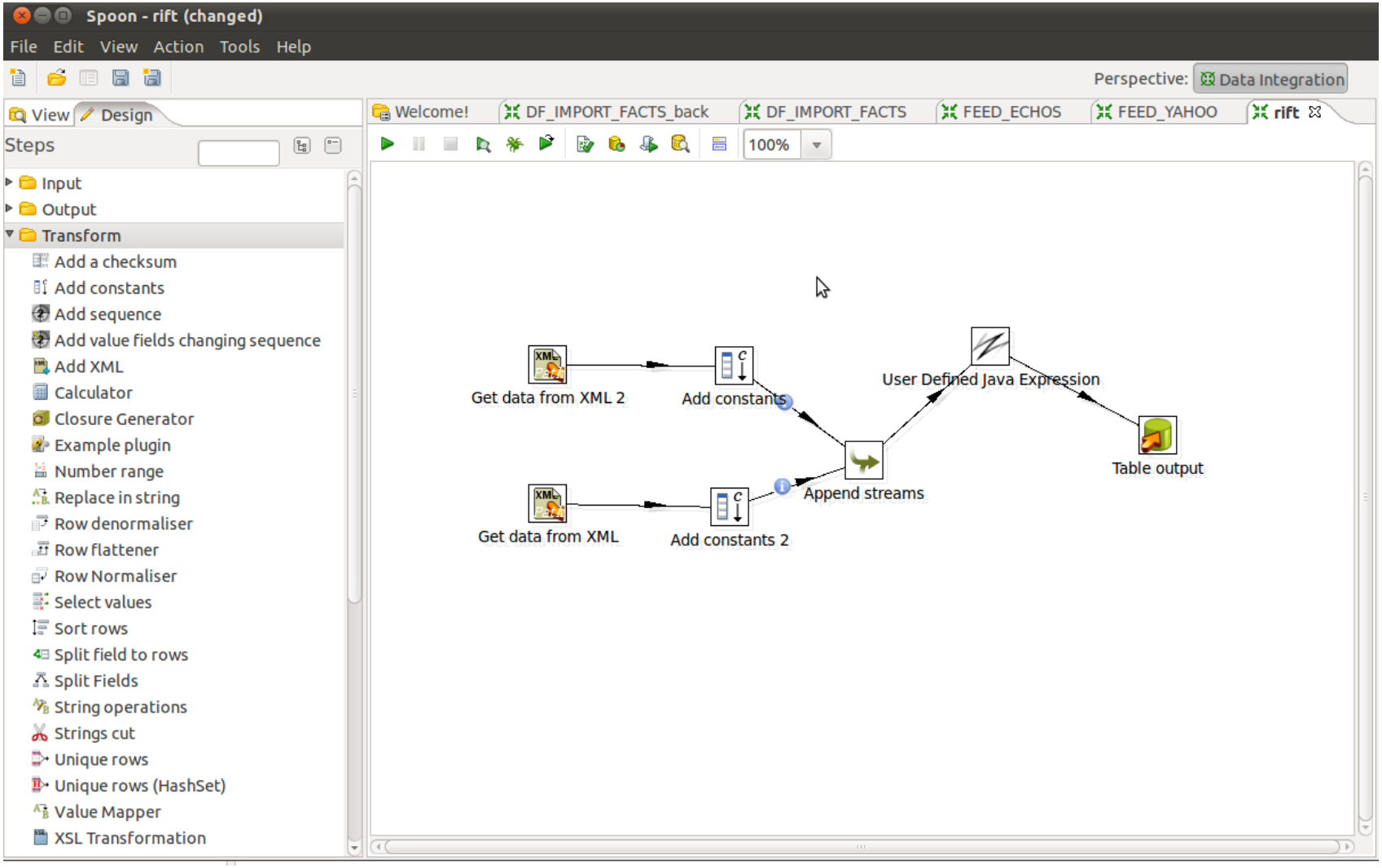


Logiciels d'ETL

Interfaces graphiques pour le non-spécialiste

- 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)

Pentaho Data Integration



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. Élément central du dispositif, il permet aux applications décisionnelles de masquer la diversité de l'origine des données et de bénéficier d'une source d'information commune, homogène, normalisée et fiable, au sein d'un système unique et si possible normalisé.

Tâches

- Deuxième passe de filtrage et validation
- Synchronisation
- Certification (liens avec des documents légaux)

Fonction de diffusion

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.

Tâche

- Choisir les données en fonction des besoins des utilisateurs
- Méthodoligie OLAP

OLAP - *Online Analytical Processing*

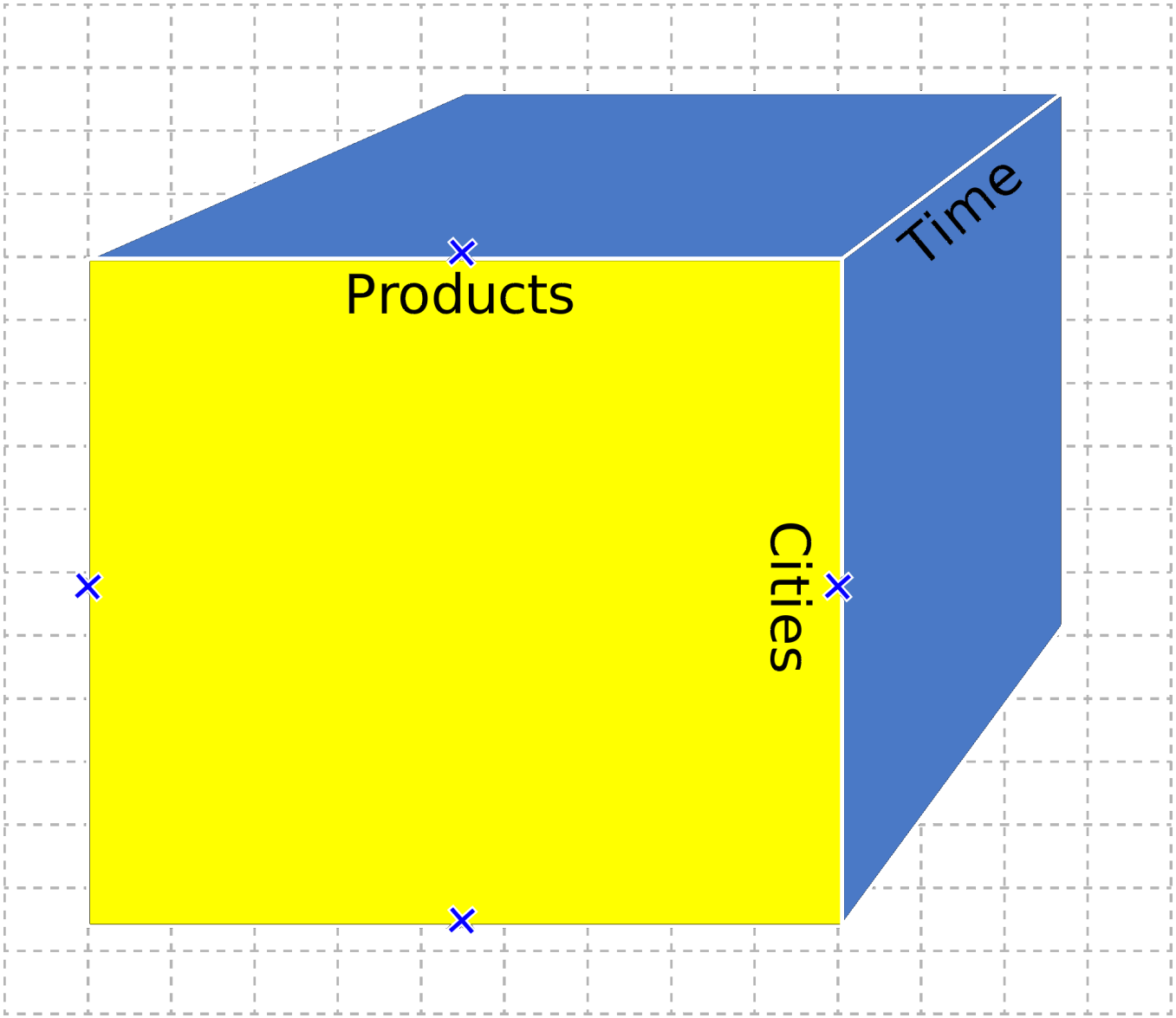
Définition

Analyse sur-le-champ d'informations selon plusieurs axes, dans le but d'obtenir des rapports de synthèse

But

- Les données sont dans un espace de grande dimension
- Beaucoup de données
- Comment gérer ça ?

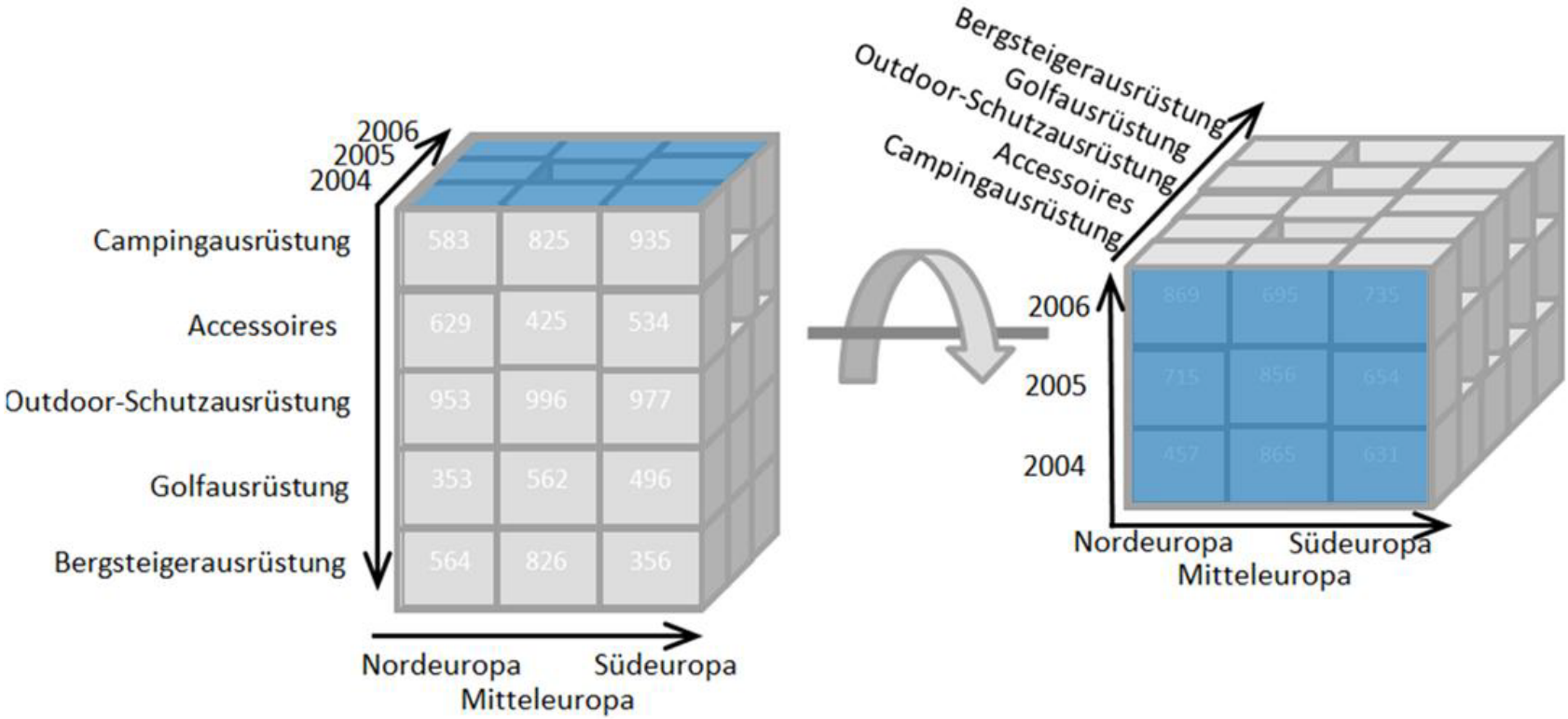
Hypercube



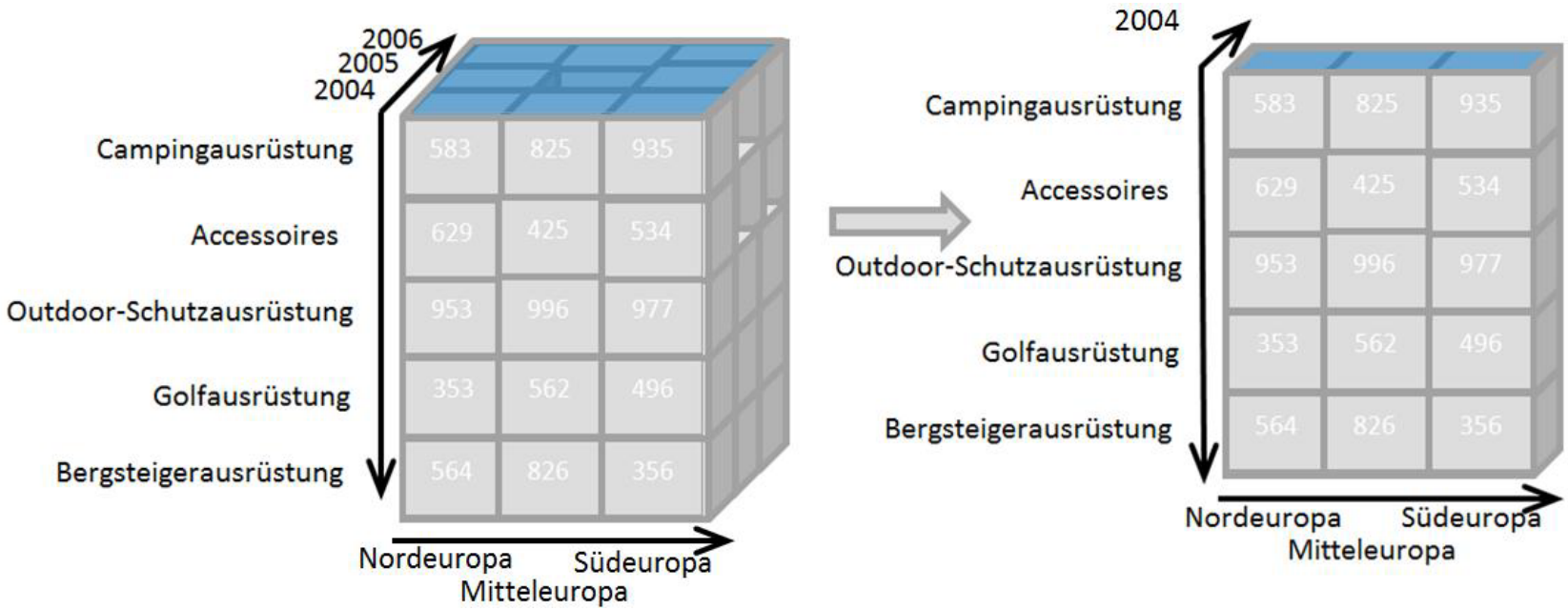
Opérations sur l'hypercube

- 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).

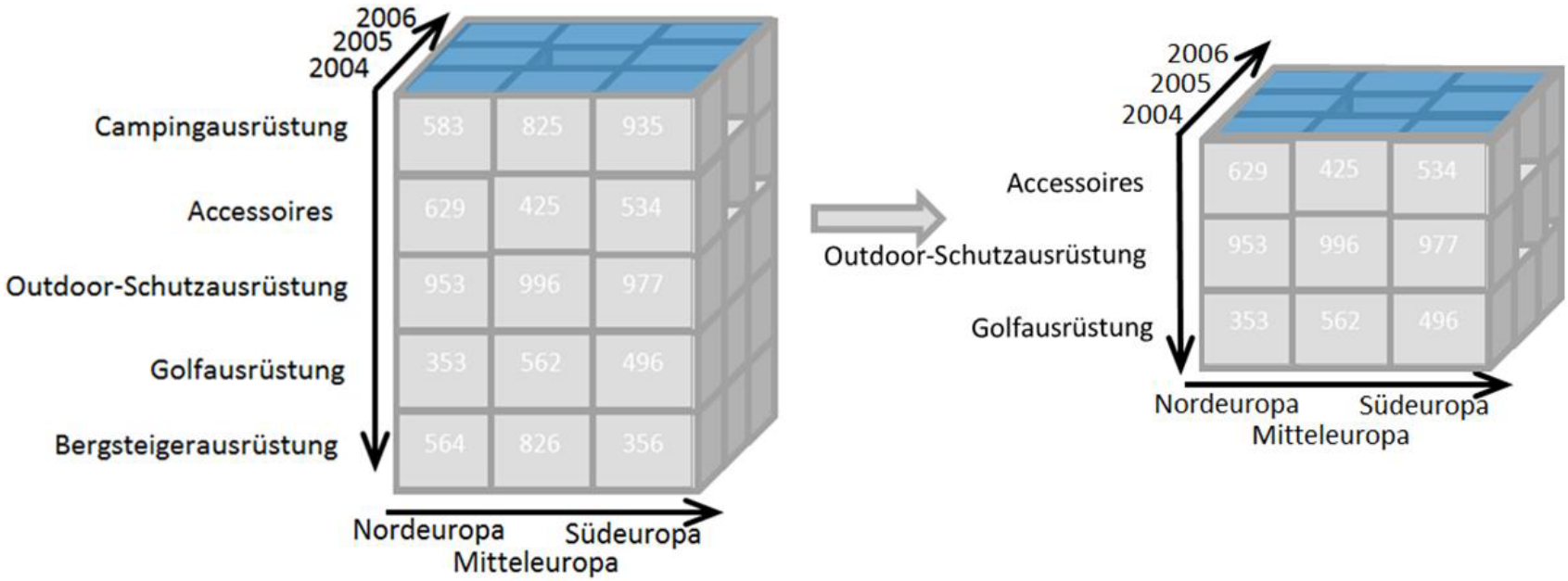
Rotate


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
Bergsteigerausrüstung
Golfausrüstung
Outdoor-Schutzausrüstung
Accessoires
Campingausrüstung
2006
2005
2004
869
695
735
715
856
654
457
865
631
Nordeuropa
Südeuropa
Mitteleuropa

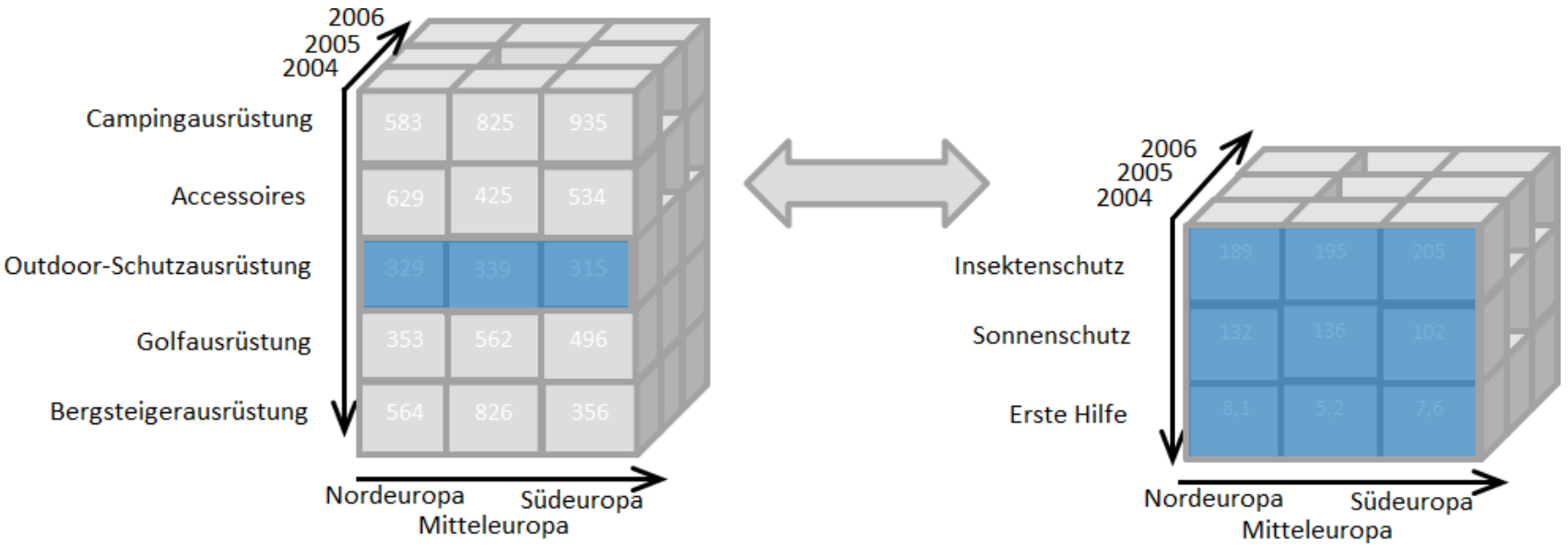
Slicing


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
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

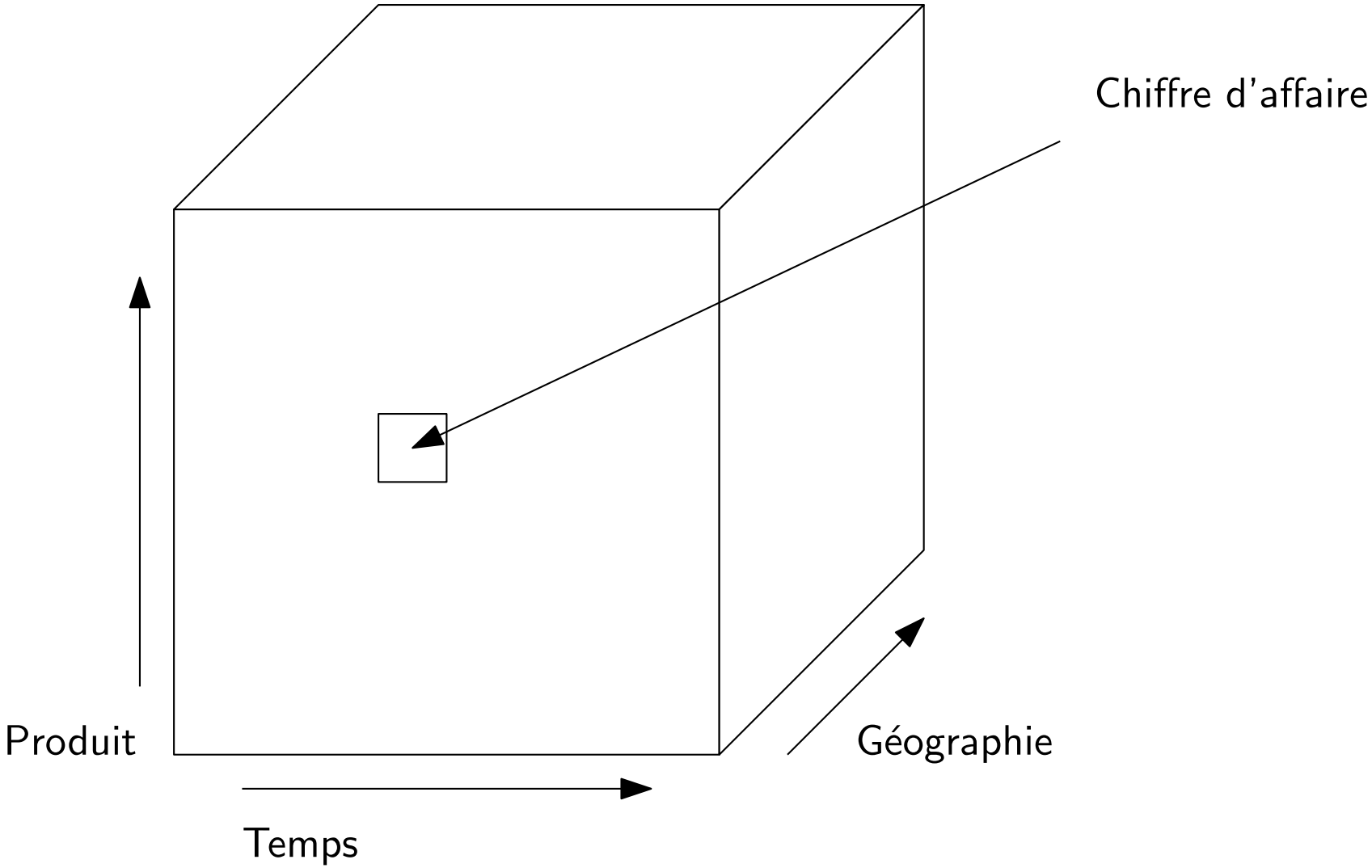
Scoping


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
Accessoires
Golfausrüstung
629
425
534
953
996
977
353
562
496
Nordeuropa
Südeuropa
Mitteleuropa

Drill-up, drill-down



Concrètement



Fonction présentation

Définition

Cette quatrième fonction, la plus visible pour l'utilisateur, régit les conditions d'accès de l'utilisateur aux informations, dans le cadre d'une interface Homme-machine déterminé (IHM).

Tâche

- Visualisation
- Rapports
- En lien direct avec l'utilisateur final

Bases de données relationnelles

Stockage organisé de données

- Base opérationnelle (les données de l'activié de l'entreprise)
- Datawarehouse et datamart

Un langage de requêtes standardisé : SQL

- `SELECT ... FROM ... WHERE ...`
- `INSERT INTO ... VALUES ...`

Extrêmement répandu

- pour toutes sortes d'utilisations

Sites web

Service web

- Sites coopératifs, publics ou privés
- Une interface documentée pour extraire des données
- Formats standardisés, gérés par les suites BI

Web scraping

- Sites non-coopératifs
- Analyse des pages webs fournies aux navigateurs webs
- Nécessite de programmer et de formater les données

Pentaho

Une suite complète

- ETL
- OLAP
- Visualisation et rapports
- Datamining

Datamining

Plus loin que la visualisation et les rapports

- Prédire à partir des données
- Aide à la décision : pas une boîte noire

Techniques de machine learning

- Classification (en particulier les méthodes interprétables, telles que les arbres de décision)
- Clustering

Logiciels

- Langages de programmation : R
- Interfaces graphiques : Weka, Orange, Tanagra
- Mixtes : SPSS, Matlab, Excel