Data Warehousing & Data Mining

Business Intelligence

(sélection des slides)

Jef Wijsen

Université de Mons (UMONS)

Outline

Le Business Intelligence en quelques mots clés

- Informatique opérationnelle (OLTP)

  E,g., gestion des commandes, livraison, facturation, paiement. . .

  $\Longrightarrow$ stockage de volumes de données énormes
- Informatique décisionnelle (DSS, BI, pilotage, stratégie)
  - Indicateurs préconçus (OLAP, tableau de bord, querying, reporting)

    E.g., un graphique montrant l'évolution du délai moyen entre la commande et la livraison. → "faisable en SQL + un peu de graphisme"
  - Découvertes de connaissances dans les données (KDD, data mining, machine learning, analytics)

    E.g., quels sont les facteurs qui impactent sur le délai entre la commande et la livraison ? → dépasse l'SQL

Data Mining vs OLAP

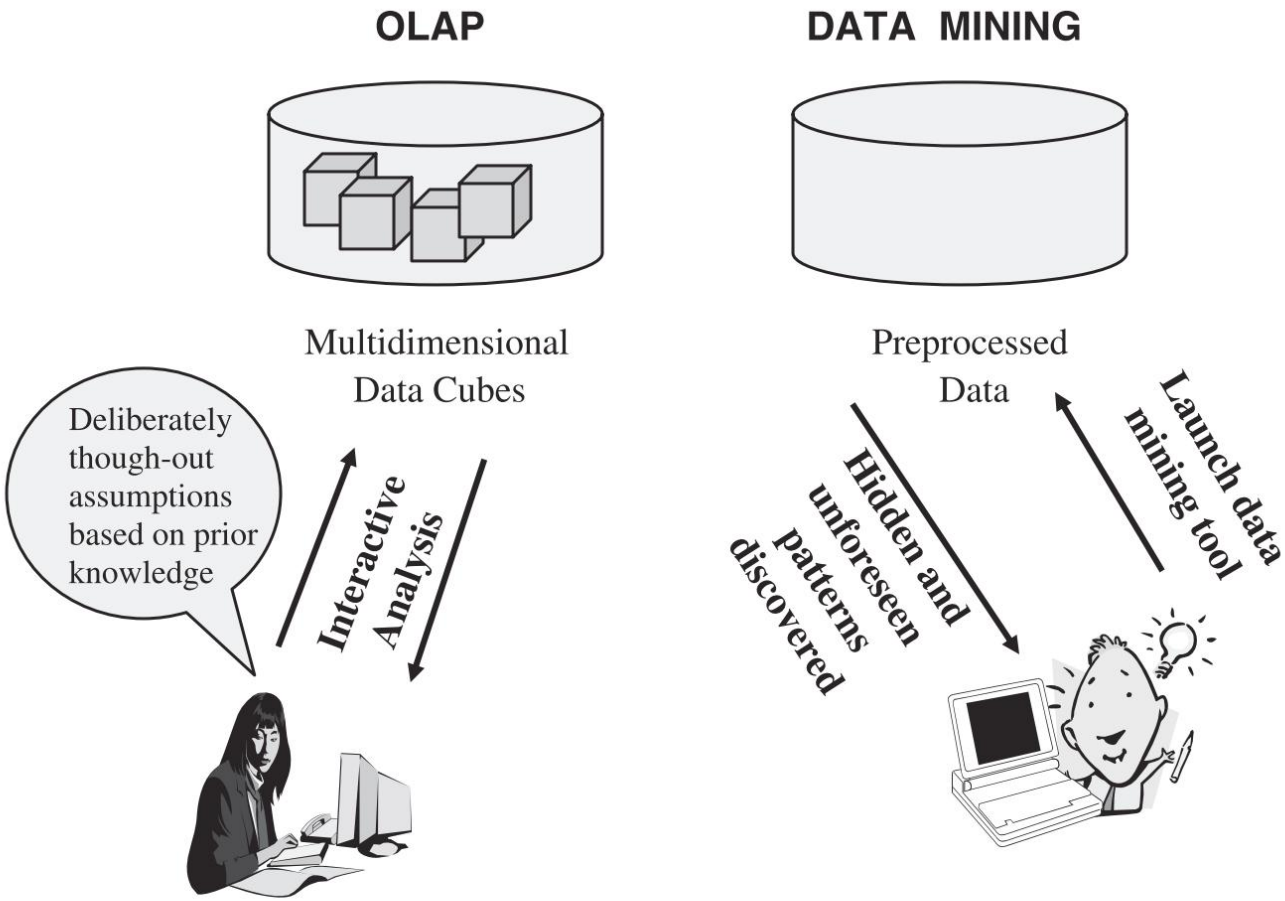


Figure 17-5 OLAP and data mining.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Étude de cas

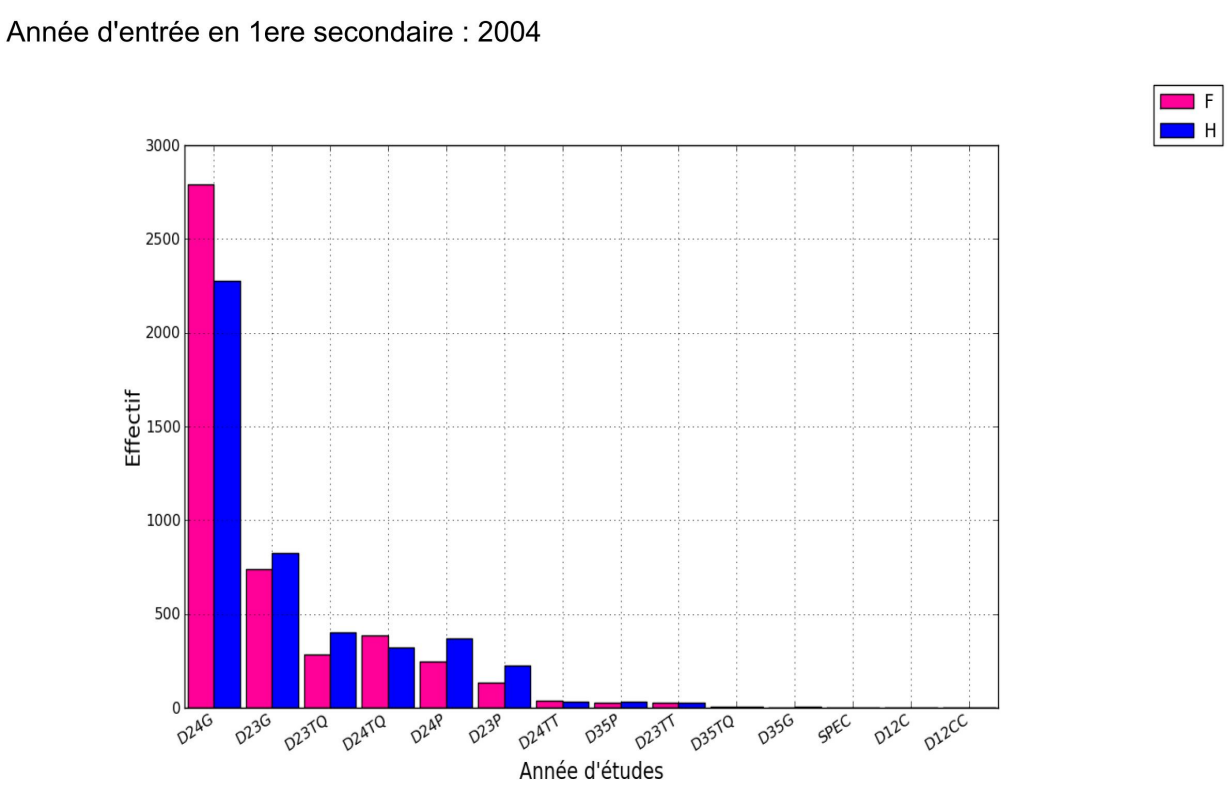
- Développement d'outils de pilotage effectif du réseau de la Communauté française
- Données opérationnelles :

COMPTAGE	Le comptage des élèves.
EDIFCf	Les infrastructures.
PERSONNEL	Le personnel de l'enseignement.
GESTELEV	Les grilles horaires et les attestations.
TEC	Les transports en commun, provenant de la Société Régionale Wallonne du Transport (SRWT) et de la Société de Transport Intercommunal de Bruxelles (STIB).
RESULTATS	Les résultats des évaluations externes certificatives (CEB et CE1D).

- Objectif décisionnel : améliorer le pilotage et faciliter la définition des actions pour améliorer la qualité de l'enseignement

Indicateur

Situation après 3 ans des élèves inscrits en 1ère secondaire commune une année donnée, selon le genre



D24G = 2e degré de transition quatrième général transition, D23TQ = 2e degré de transition troisième technique qualification,

D24P = 2e degré de qualification quatrième professionnel qualification...

Questions de type KDD ($\simeq$ data mining)

- Quels sont les facteurs (tels que le genre, le statut socio-économique...) expliquant le retard scolaire ?
- A quels endroits faut-il prévoir de nouvelles implantations ?
- ...

Difficultés à surmonter

Mismatch entre les données opérationnelles et les besoins au plan décisionnel. Les données opérationnelles sont typiquement

- dispersées,
- brutes (i.e., non moyennées, non agrégées. . .),
- bruitées (erronnées, non filtrées. . .),
- privées,
- . . .

Contenu du cours

1. Introduction générale
2. Data warehousing, ETL, data quality, OLAP
3. Data mining
   - Classification
   - Association rules
   - Clustering (avec lecture d'un article scientifique, si le temps le permet)

Outline

Operational to Decisional

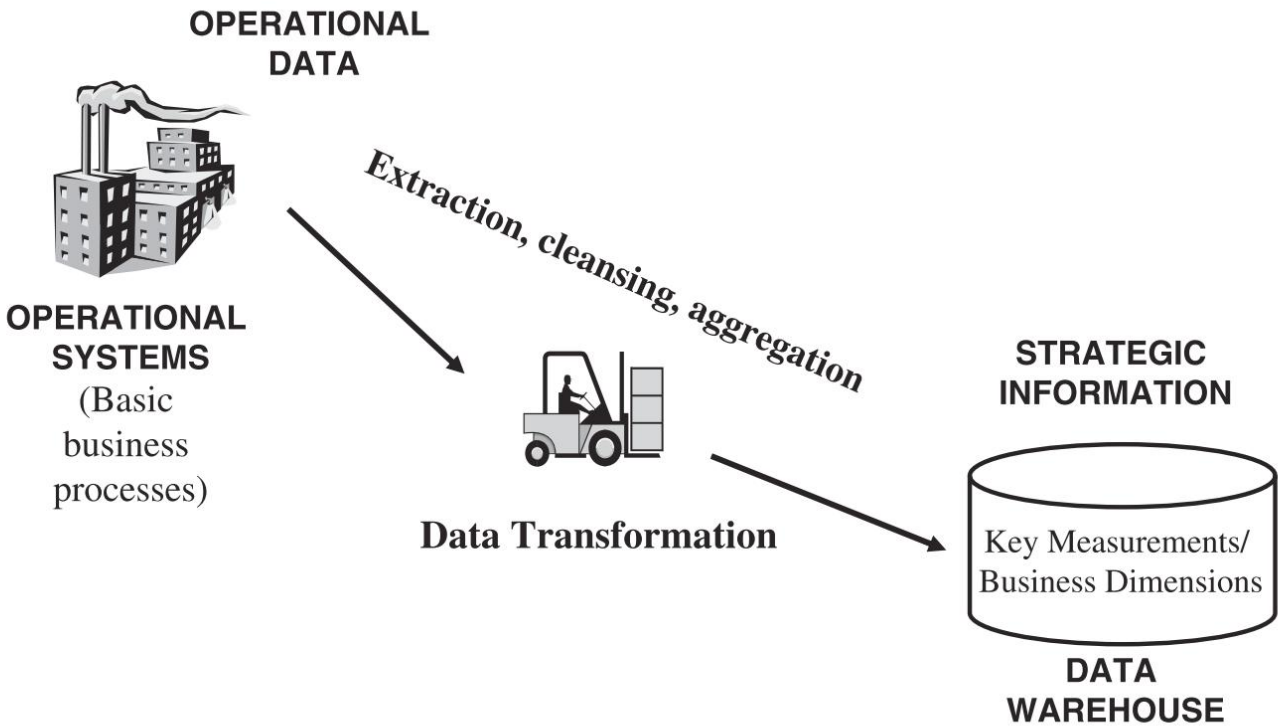


Figure 1-8 General overview of the data warehouse.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Running $\rightsquigarrow$ Monitoring the Business

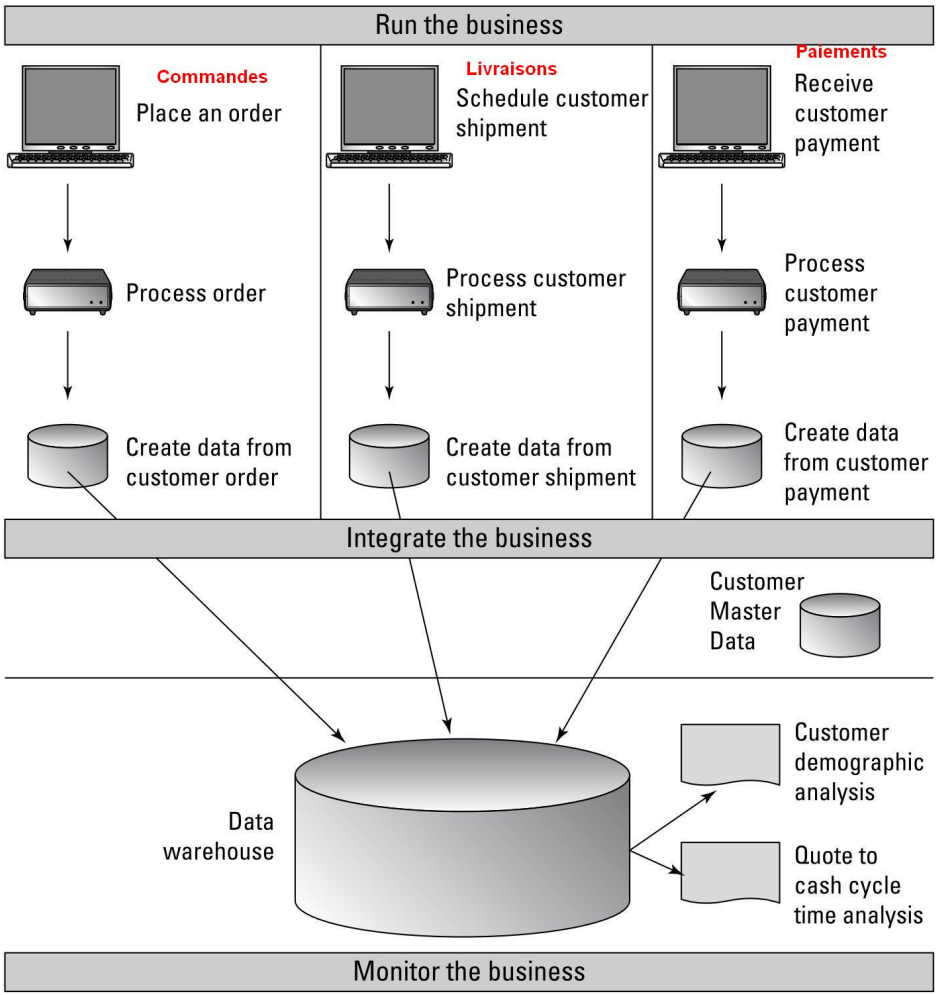


Figure 1-1: Most computer applications create data as a result of an activity or transaction; a data warehouse instead swipes data created elsewhere and converts it into information.

Source: Thomas C. Hammergren and Alan R. Simon: *Data Warehousing for Dummies* (2nd Edition). Wiley Publishing, 2009

Business Intelligence (BI): Two Environments

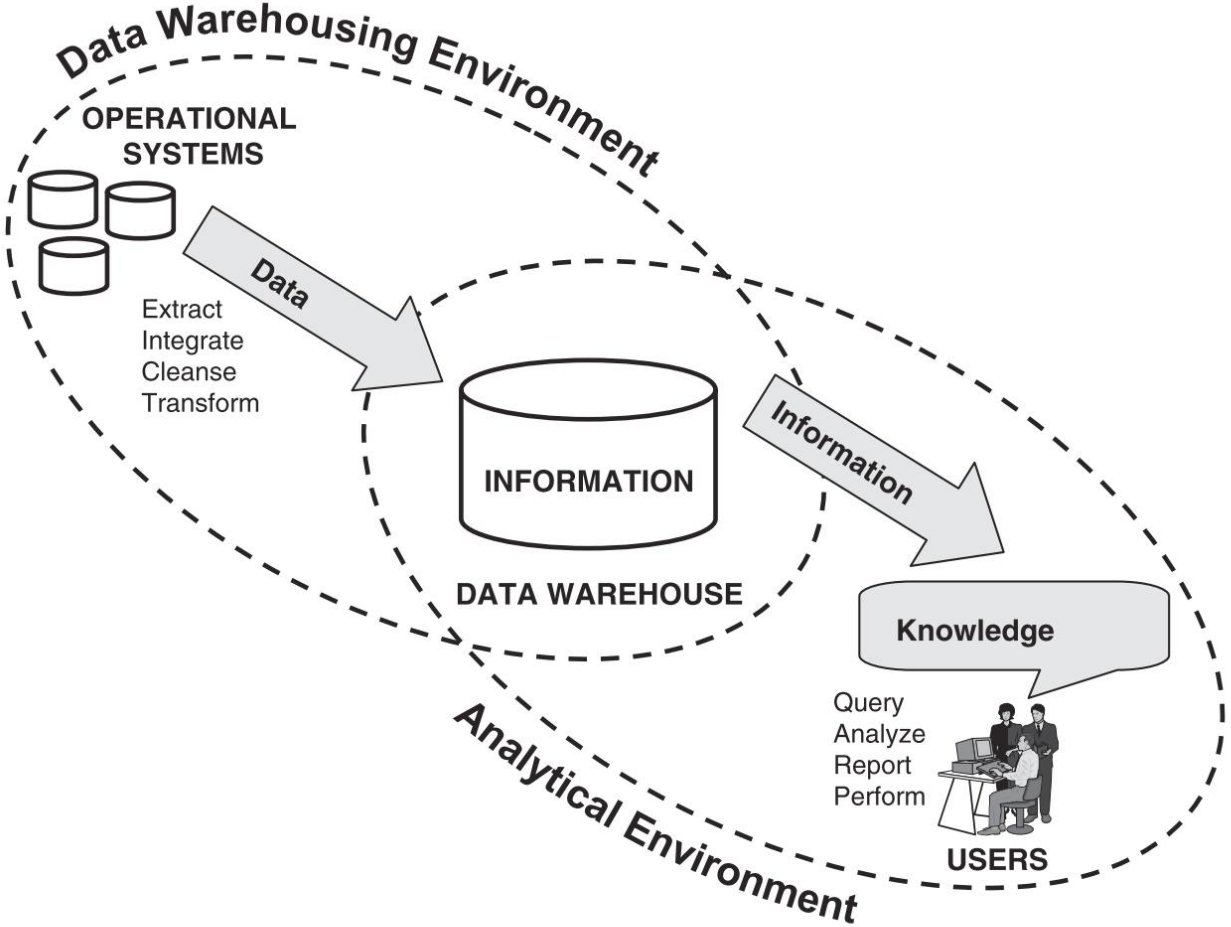


Figure 1-10 BI: data warehousing and analytical environments.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Data Warehousing, OLAP and Data Mining

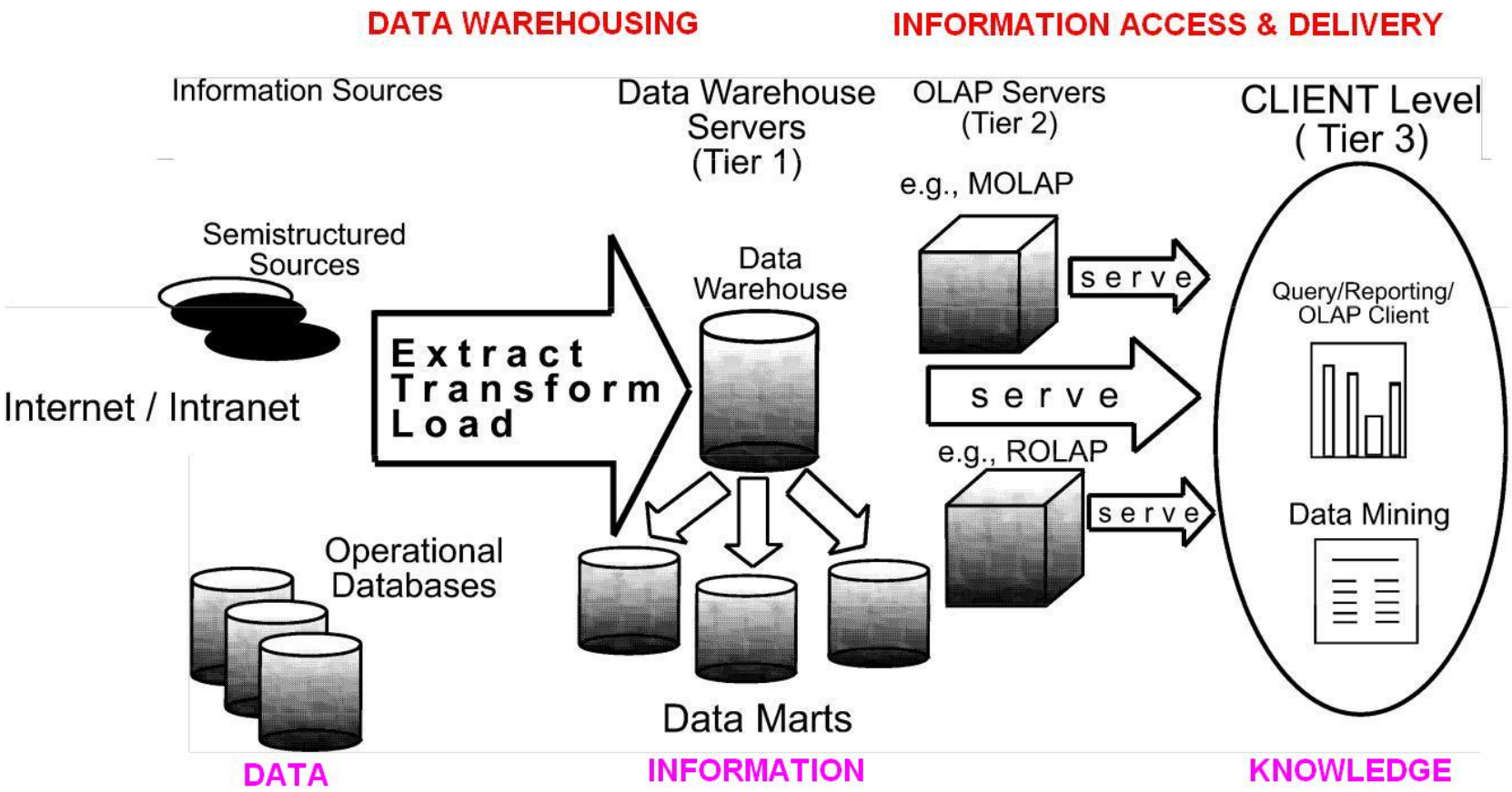


Fig. 1.3. The IT Decision Support Tiers.

Source: Oded Maimon, Lior Rokach (Eds.): *The Data Mining and Knowledge Discovery Handbook* (2nd Edition). Springer, 2010

Data Warehouse

Une collection de données organisée pour la prise de décisions, possédant les caractéristiques suivantes :

Thématique et intégrée. Les données OLTP sont souvent réparties sur plusieurs *applications* (facturation, livraison, production,...).
Les données seront intégrées dans un data warehouse autour d'un certain nombre de *thèmes* (client, produit, fournisseur,...).

Non volatile et historisée. Les données, une fois dans l'entrepôt, ne sont plus supposées être modifiées.
Les données couvrent une certaine période de temps (par ex. dix ans) pour en extraire des tendances.

Integrated and subject-oriented

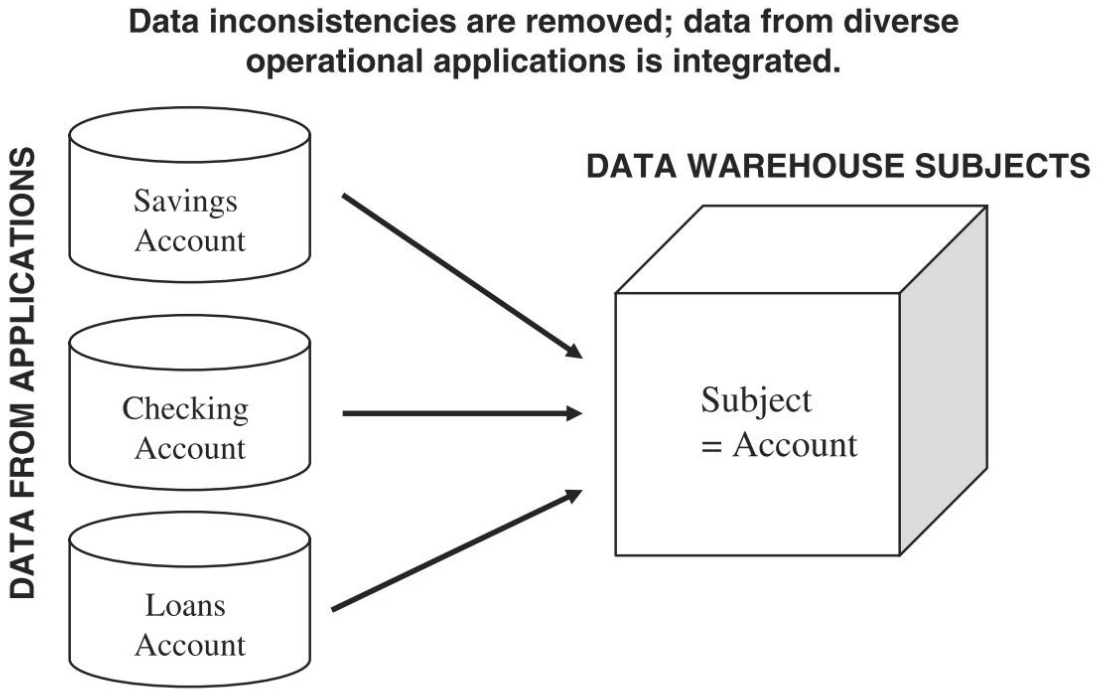


Figure 2-2 The data warehouse is integrated.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Nonvolatile

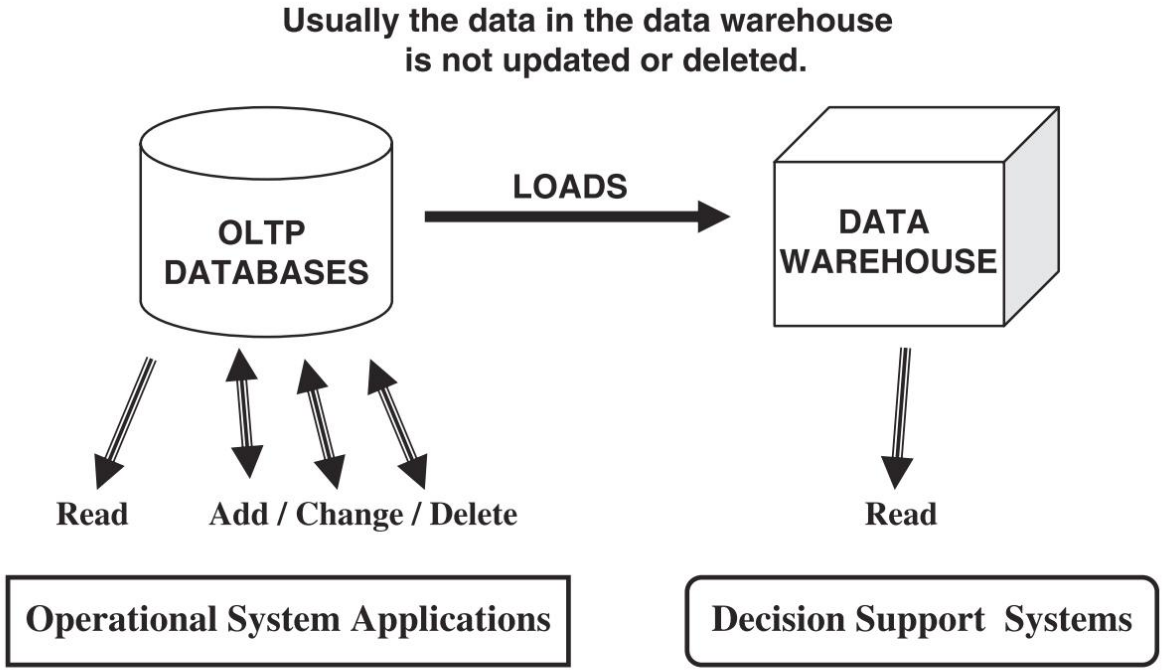


Figure 2-3 The data warehouse is nonvolatile.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Operational vs Decisional

Operational databases	Data warehouse
dispersed	integrated
detailed	summarized, aggregated
quality problems (errors, missing values,. . .)	cleaned

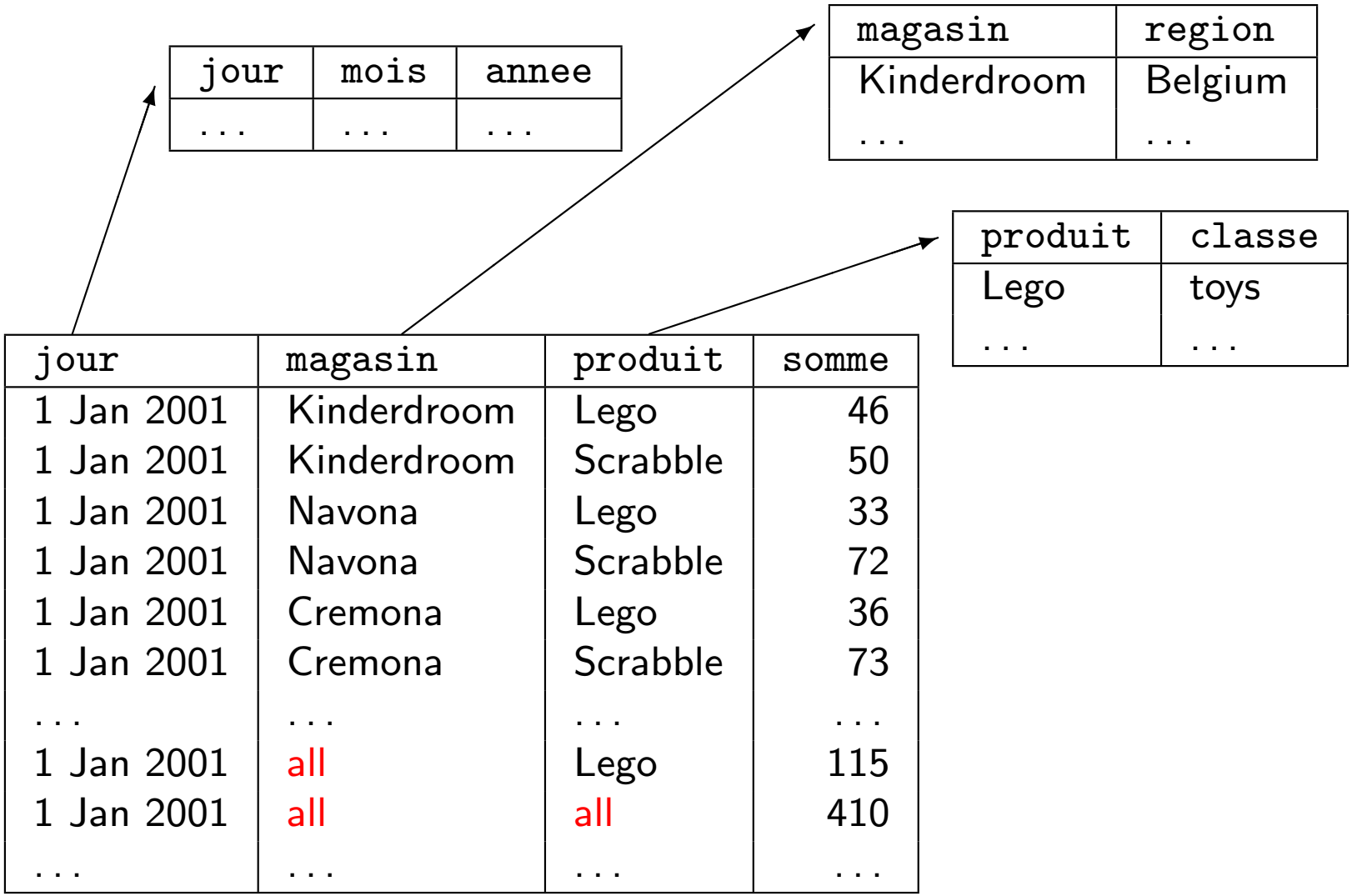
Outline

Outline

Recall from Relational Databases

- Relational table.
- Primary key.
- Foreign key.
- Boyce-Codd normal form (BCNF).

Schéma en étoile (STAR Schema)



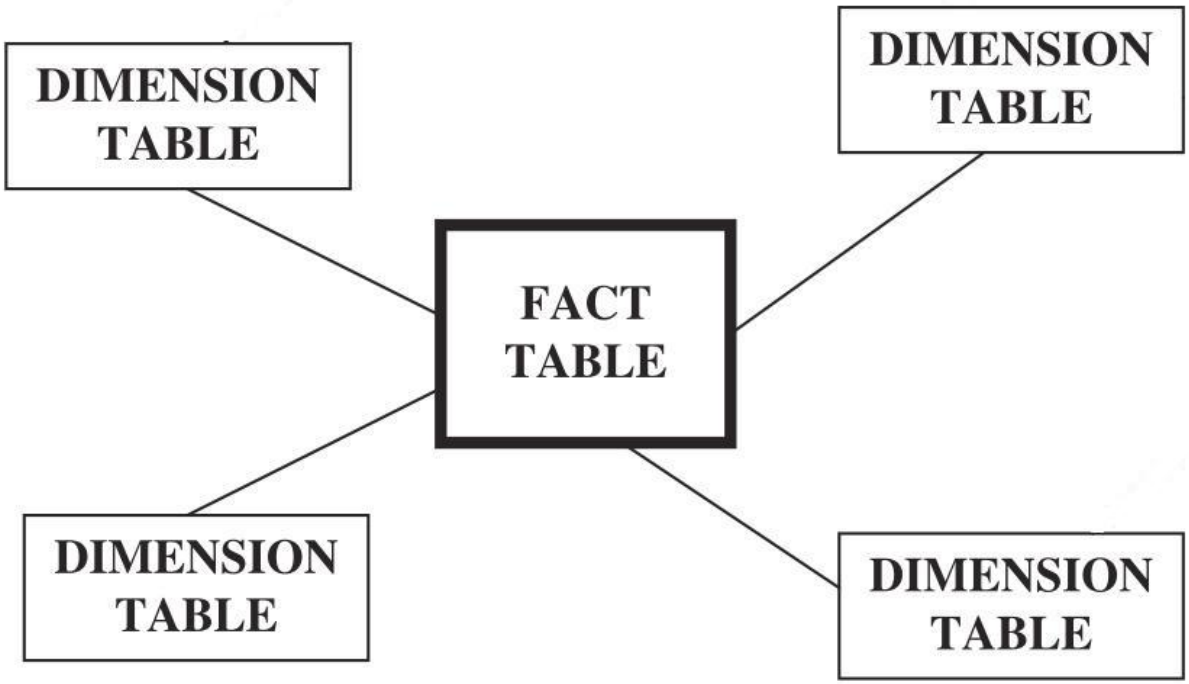
jour	mois	annee
...	...	...

magasin	region
Kinderdroom	Belgium
...	...

produit	classe
Lego	toys
...	...

jour	magasin	produit	somme
1 Jan 2001	Kinderdroom	Lego	46
1 Jan 2001	Kinderdroom	Scrabble	50
1 Jan 2001	Navona	Lego	33
1 Jan 2001	Navona	Scrabble	72
1 Jan 2001	Cremona	Lego	36
1 Jan 2001	Cremona	Scrabble	73
...	...	...	...
1 Jan 2001	all	Lego	115
1 Jan 2001	all	all	410
...	...	...	...

Fact and Dimension Tables



STAR Schema

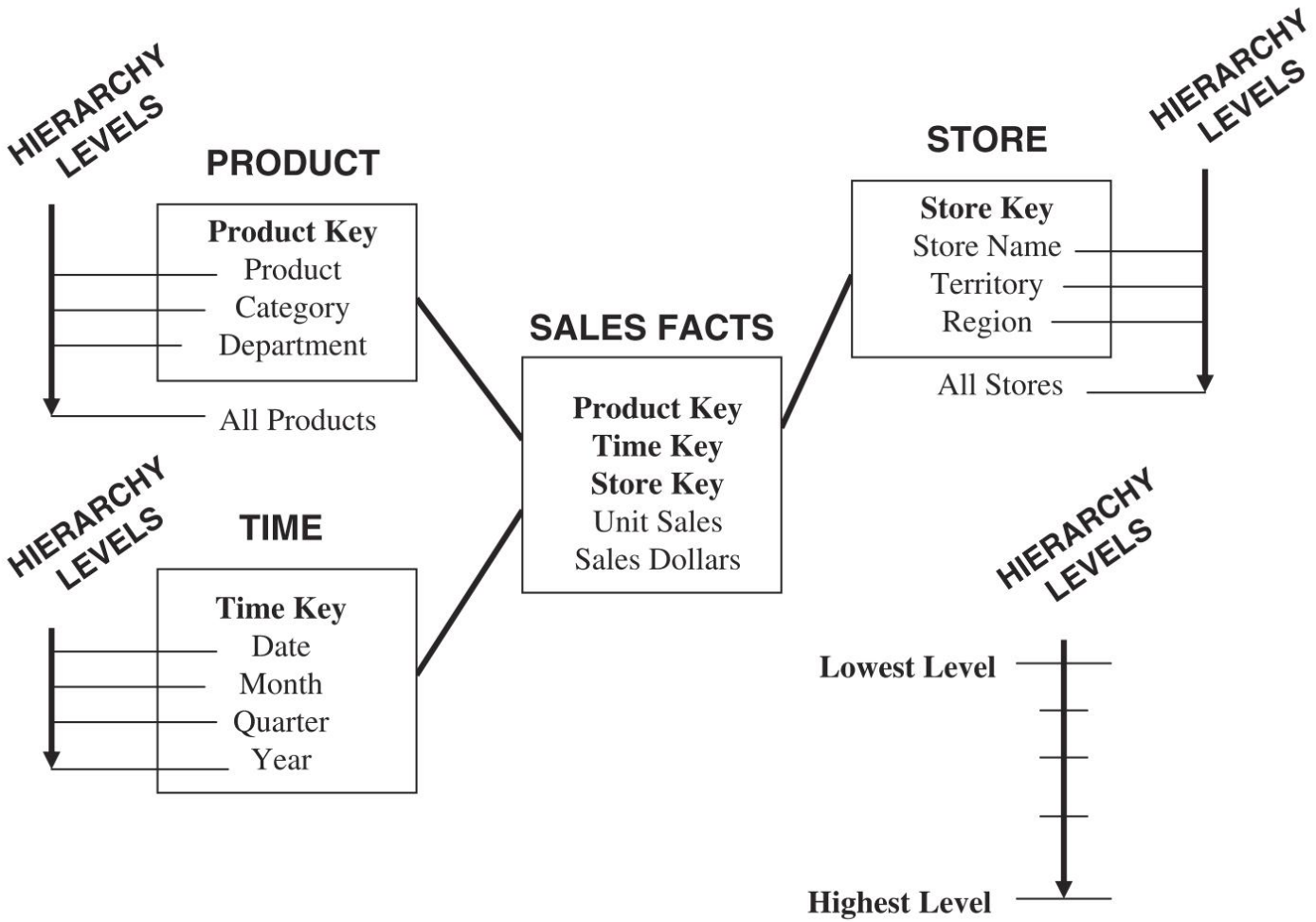


Figure 11-13 Dimension hierarchies.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

SNOWFLAKE Schema

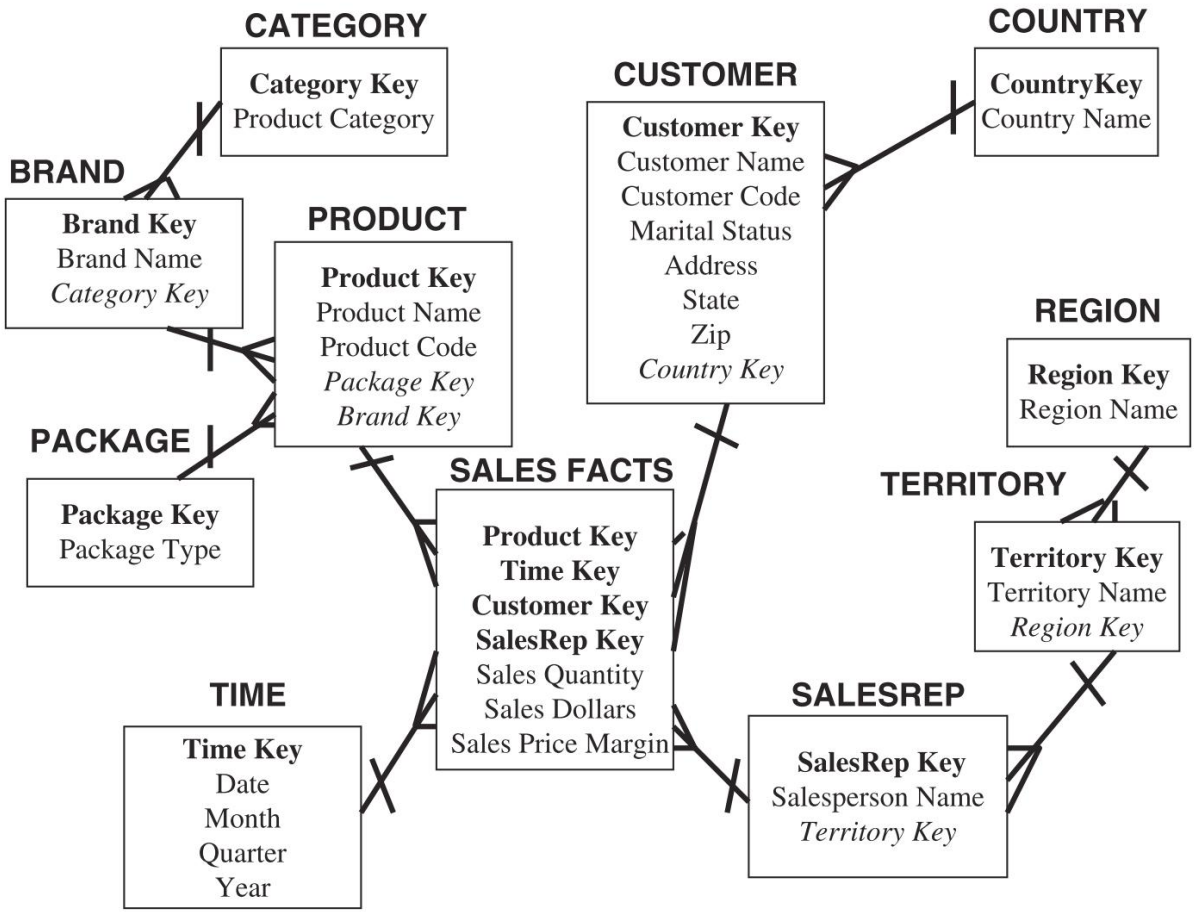


Figure 11-9 Sales: the "snowflake" schema.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Outline

Data Extraction, Transformation, and Loading

- Reshape relevant data from operational systems into useful information to be stored in the data warehouse.
- Data warehouse $\neq$ data junkhouse

Operational data	Strategic information
dispersed	integrated
detailed	summarized, aggregated
quality problems	clean[s]ed

- Not uncommon to spend 50% to 70% of project effort on ETL functions.

ETL Challenges I

- Source systems are very diverse and disparate.
- There is usually a need to deal with source systems on multiple platforms and different operating systems.
- Many source systems are older legacy applications running on obsolete database technologies.
- Generally, historical data on changes in values are not preserved in source operational systems. Historical information is critical in a data warehouse.
- Quality of data is dubious in many old source systems that have evolved over time.
- Source system structures keep changing over time because of new business conditions. ETL functions must also be modified accordingly.

ETL Challenges II

- Gross lack of consistency among source systems is prevalent. Same data is likely to be represented differently in the various source systems (example: prices in EUR, USD, BEF).
- Even when inconsistent data is detected among disparate source systems, lack of a means for resolving mismatches escalates the problem of inconsistency.
- Most source systems do not represent data in types or formats that are meaningful to the users. Many representations are cryptic and ambiguous (example: 1=male, 2=female).

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Data Extraction (ETL)

- Source identification
- Extract data for one-time initial full load of the data warehouse
- Extract data for ongoing incremental loads
  - Immediate data extraction in real-time
  - Deferred data extraction, e.g., at midnight every day

Note: a data staging area is an intermediate storage area between the sources of information and the data warehouse.

Data Transformation (E**T**L)

- Extracted raw data need to be transformed in usable information.
- Basic tasks:
  - Select, project, join records from many source systems
  - Conversion (e.g., standardize fields from disparate source systems)
  - Summarization
  - Enrichment

Major Transformation Types (with Examples)

Format Revisions. Changes to the data types and lengths of individual fields.

Decoding of Fields. $1 \rightsquigarrow$ M, $2 \rightsquigarrow$ F

Calculated and Derived Values. Net profit margin $= \frac{\text{Net profit (after taxes)}}{\text{Revenue}} \times 100\%$

Splitting of Single Fields.

Address	City	$\rightsquigarrow$	Street	Nr	ZIP	City
22 Rue de Ath	7000 Mons		Rue de Ath	22	7000	Mons

Merging of Information. Joining records coming from different sources.

Character set conversion. EBCDIC $\rightsquigarrow$ ASCII

Conversion of Units of Measurements. USD, GBP $\rightsquigarrow$ EUR

Date/Time Conversion. "October 11, 2008", "11/10/2008" $\rightsquigarrow$ 11 OCT 2008

Summarization. Daily sales amount.

Deduplication.

CName	Address	...
RAYTEC	Rue de Commerce 2	...
S.A. RAYTEC	2 Rue de Commerce	...

Data Integration and Consolidation

Data integration problems:

Entity Identification Problem The same customer may be stored with distinct identification numbers in different data sets.

Inconsistency Among Data Sources The same customer may be stored with distinct domicile addresses in different data sets.

Data Loading (ETL)

- Initial load. This may take several days to complete...
- Incremental update
- Refresh of some tables (refresh of all tables = initial load)

Outline

Étude de cas

- Développement d'outils de pilotage effectif du réseau de la Communauté française
- Données opérationnelles :

COMPTAGE	Le comptage des élèves.
EDIFCf	Les infrastructures.
PERSONNEL	Le personnel de l'enseignement.
GESTELEV	Les grilles horaires et les attestations.
TEC	Les transports en commun, provenant de la Société Régionale Wallonne du Transport (SRWT) et de la Société de Transport Intercommunal de Bruxelles (STIB).
RESULTATS	Les résultats des évaluations externes certificatives (CEB et CE1D).

- Objectif décisionnel : améliorer le pilotage et faciliter la définition des actions pour améliorer la qualité de l'enseignement

Problèmes Intra-Sources (1)

Année scolaire	Identifiant	Genre	Date de naissance	Année d'études
2009	_____831	M	2000-03-21	4P
2010	_____831	F	2000-03-21	5P
2009	_____121	F	1968-08-12	5P
2010	_____332	F	0000-00-00	1P
2009	_____534	M	2004-05-13	1P
2010	_____534	M	2003-03-21	2P
2010	_____726	I	2003-06-01	2P

Problèmes Intra-Sources (1)

Année scolaire	Identifiant	Genre	Date de naissance	Année d'études
2009	_____831	M	2000-03-21	4P
2010	_____831	F	2000-03-21	5P
2009	_____121	F	1968-08-12	5P
2010	_____332	F	0000-00-00	1P
2009	_____534	M	2004-05-13	1P
2010	_____534	M	2003-03-21	2P
2010	_____726	I	2003-06-01	2P

Problèmes Intra-Sources (1)

Année scolaire	Identifiant	Genre	Date de naissance	Année d'études
2009	_____831	M	2000-03-21	4P
2010	_____831	F	2000-03-21	5P
2009	_____121	F	1968-08-12	5P
2010	_____332	F	0000-00-00	1P
2009	_____534	M	2004-05-13	1P
2010	_____534	M	2003-03-21	2P
2010	_____726	I	2003-06-01	2P

Problèmes Intra-Sources (1)

Année scolaire	Identifiant	Genre	Date de naissance	Année d'études
2009	_____831	M	2000-03-21	4P
2010	_____831	F	2000-03-21	5P
2009	_____121	F	1968-08-12	5P
2010	_____332	F	0000-00-00	1P
2009	_____534	M	2004-05-13	1P
2010	_____534	M	2003-03-21	2P
2010	_____726	I	2003-06-01	2P

Problèmes Intra-Sources (1)

Année scolaire	Identifiant	Genre	Date de naissance	Année d'études
2009	_____831	M	2000-03-21	4P
2010	_____831	F	2000-03-21	5P
2009	_____121	F	1968-08-12	5P
2010	_____332	F	0000-00-00	1P
2009	_____534	M	2004-05-13	1P
2010	_____534	M	2003-03-21	2P
2010	_____726	I	2003-06-01	2P

Problèmes Intra-Sources (1)

Année scolaire	Identifiant	Genre	Date de naissance	Année d'études
2009	_____831	M	2000-03-21	4P
2010	_____831	F	2000-03-21	5P
2009	_____121	F	1968-08-12	5P
2010	_____332	F	0000-00-00	1P
2009	_____534	M	2004-05-13	1P
2010	_____534	M	2003-03-21	2P
2010	_____726	I	2003-06-01	2P

Problèmes Intra-Sources (2)

Relations entre tables non respectées

Cours		
Type option	Code	Libellé
C	0033	Education Artistique : Arts plastique
C	1589	Français : complément
S	0115	Génie chimique
G	7303	Chimie appliquée
G	8165	Agriculture
G	8165	Agriculture-Horticulture

Grille horaire				
Année scolaire	Code grilles	...	Type option	Code Cours
2009	1	...	C	0033
2009	1	...	G	0115
2009	2	...	C	NULL
2009	3	...	G	8165

Problèmes Intra-Sources (2)

Relations entre tables non respectées

Cours		
Type option	Code	Libellé
C	0033	Education Artistique : Arts plastique
C	1589	Français : complément
S	0115	Génie chimique
G	7303	Chimie appliquée
G	8165	Agriculture
G	8165	Agriculture-Horticulture

Grille horaire				
Année scolaire	Code grilles	...	Type option	Code Cours
2009	1	...	C	0033
2009	1	...	G	0115
2009	2	...	C	NULL
2009	3	...	G	8165

Problèmes Intra-Sources (2)

Relations entre tables non respectées

Cours		
Type option	Code	Libellé
C	0033	Education Artistique : Arts plastique
C	1589	Français : complément
S	0115	Génie chimique
G	7303	Chimie appliquée
G	8165	Agriculture
G	8165	Agriculture-Horticulture

Grille horaire				
Année scolaire	Code grilles	...	Type option	Code Cours
2009	1	...	C	0033
2009	1	...	G	0115
2009	2	...	C	NULL
2009	3	...	G	8165

Problèmes Intra-Sources (2)

Relations entre tables non respectées

Cours		
Type option	Code	Libellé
C	0033	Education Artistique : Arts plastique
C	1589	Français : complément
S	0115	Génie chimique
G	7303	Chimie appliquée
G	8165	Agriculture
G	8165	Agriculture-Horticulture

Grille horaire				
Année scolaire	Code grilles	...	Type option	Code Cours
2009	1	...	C	0033
2009	1	...	G	0115
2009	2	...	C	NULL
2009	3	...	G	8165

Problèmes Inter-Sources

Incohérence de syntaxe de champs théoriquement identiques

- Année scolaire : 2008 [OU] 2008-2009 [OU] 2008 - 2009
- Genre : M / F [OU] H / F
- Date de naissance : 21 mai 2003 [OU] 21/03/2003 [OU] 2003-03-21
- Adresse : 24 rue du blé [OU] RUE DU BLE 24 [OU] Rue du Blé | 24

Estimation du temps pour l'ETL

ETL = Extract-Transform-Load

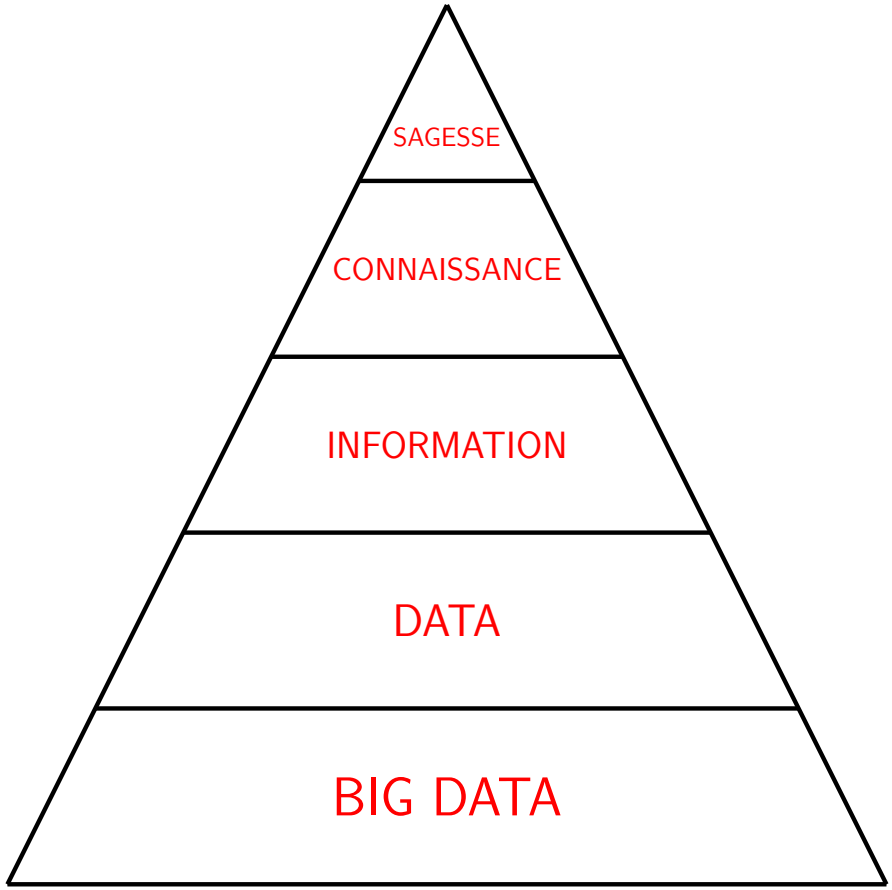
Pour ce projet :

- Extraction : 10%
- Transformation : 85%
- Chargement: 5%

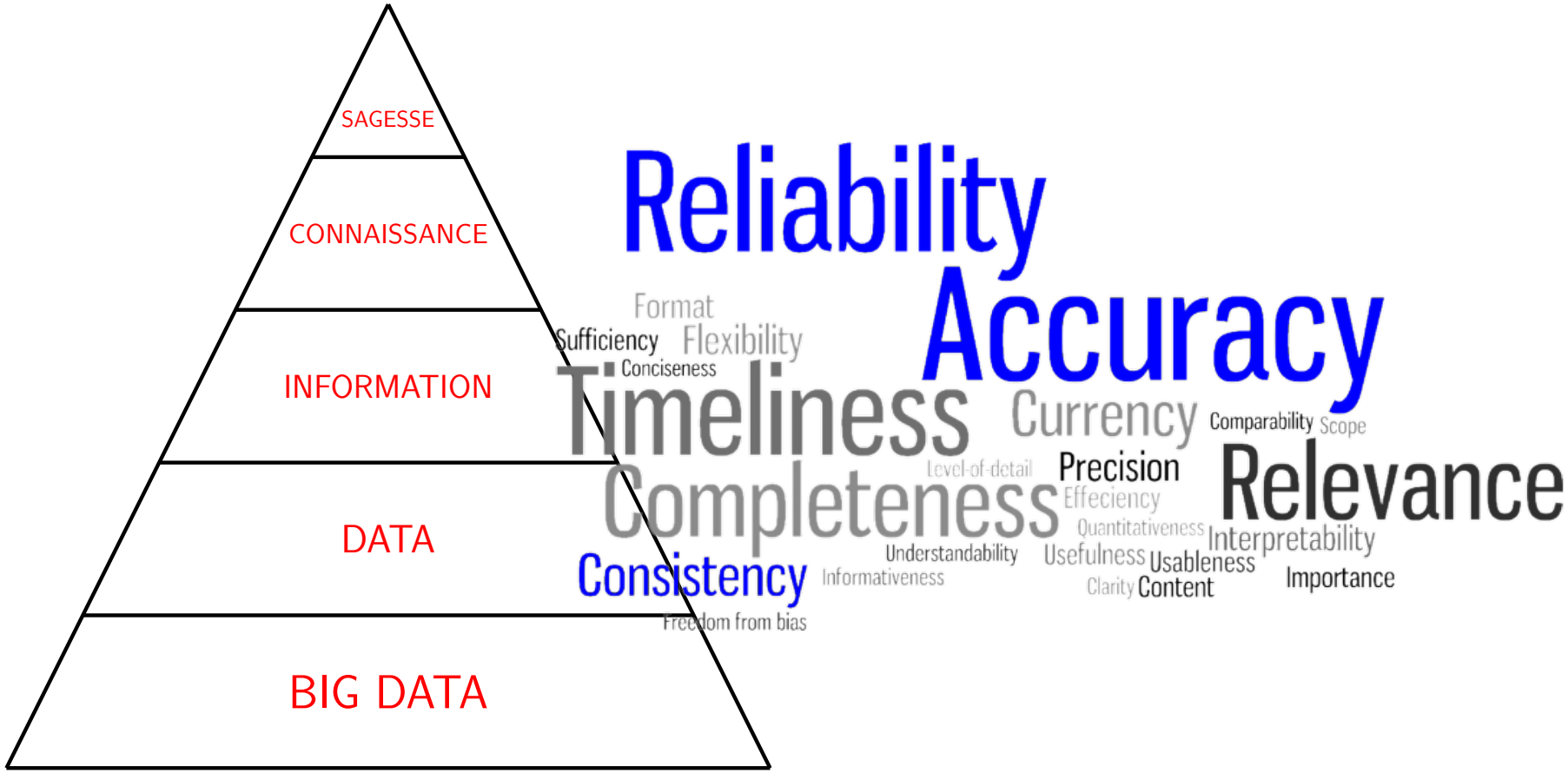
→ Impossible d'intégrer directement les données au Data Warehouse à partir des sources

Outline

L'idéal



L'idéal


SAGESSE
CONNAISSANCE
INFORMATION
DATA
BIG DATA
Reliability
Accuracy
Format
Sufficiency
Flexibility
Conciseness
Timeliness
Currency
Comparability
Scope
Level-of-detail
Precision
Completeness
Effeciency
Relevance
Quantitativeness
Interpretability
Understandability
Usefulness
Usableness
Consistency
Informativeness
Clarity
Content
Importance
Freedom from bias

La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

La réalité

Souvent les données sont incohérentes, incomplètes, manquantes...

Que peut-on faire avec ces données?

La réalité

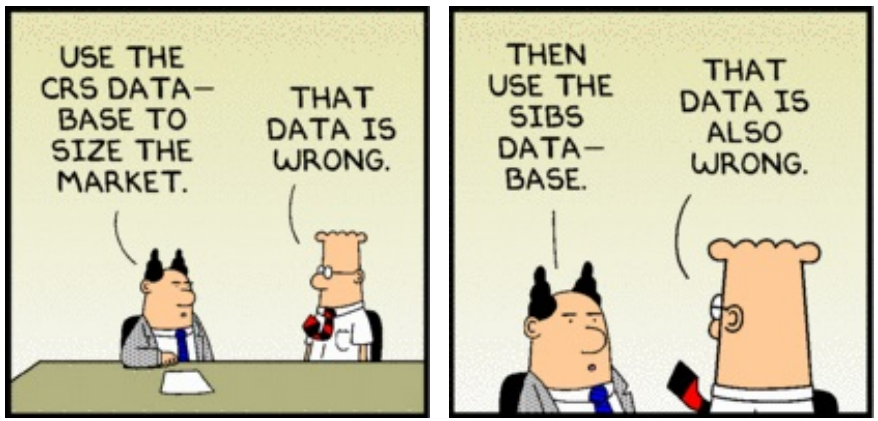
Souvent les données sont incohérentes, incomplètes, manquantes. . .

Que peut-on faire avec ces données?

La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

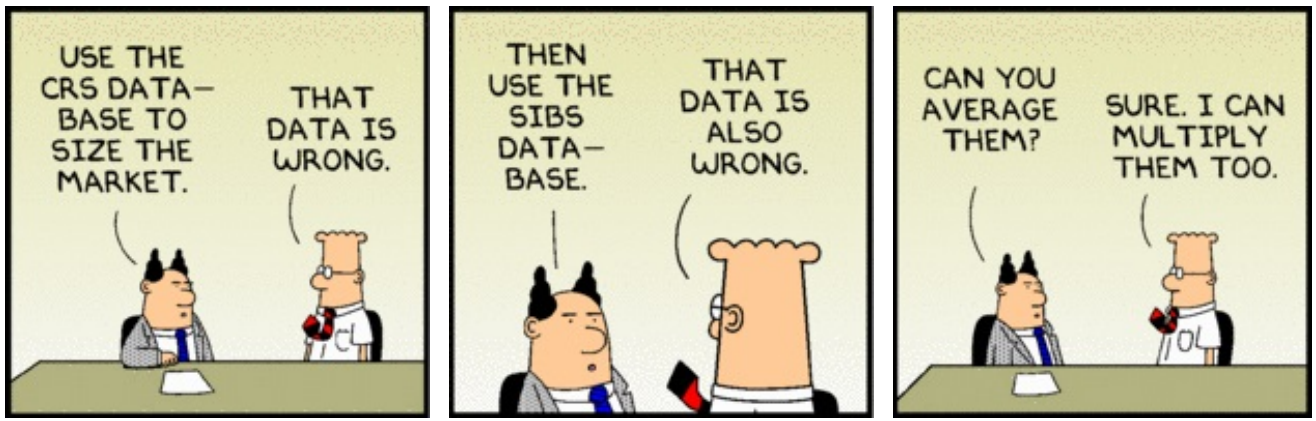
Que peut-on faire avec ces données?



La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

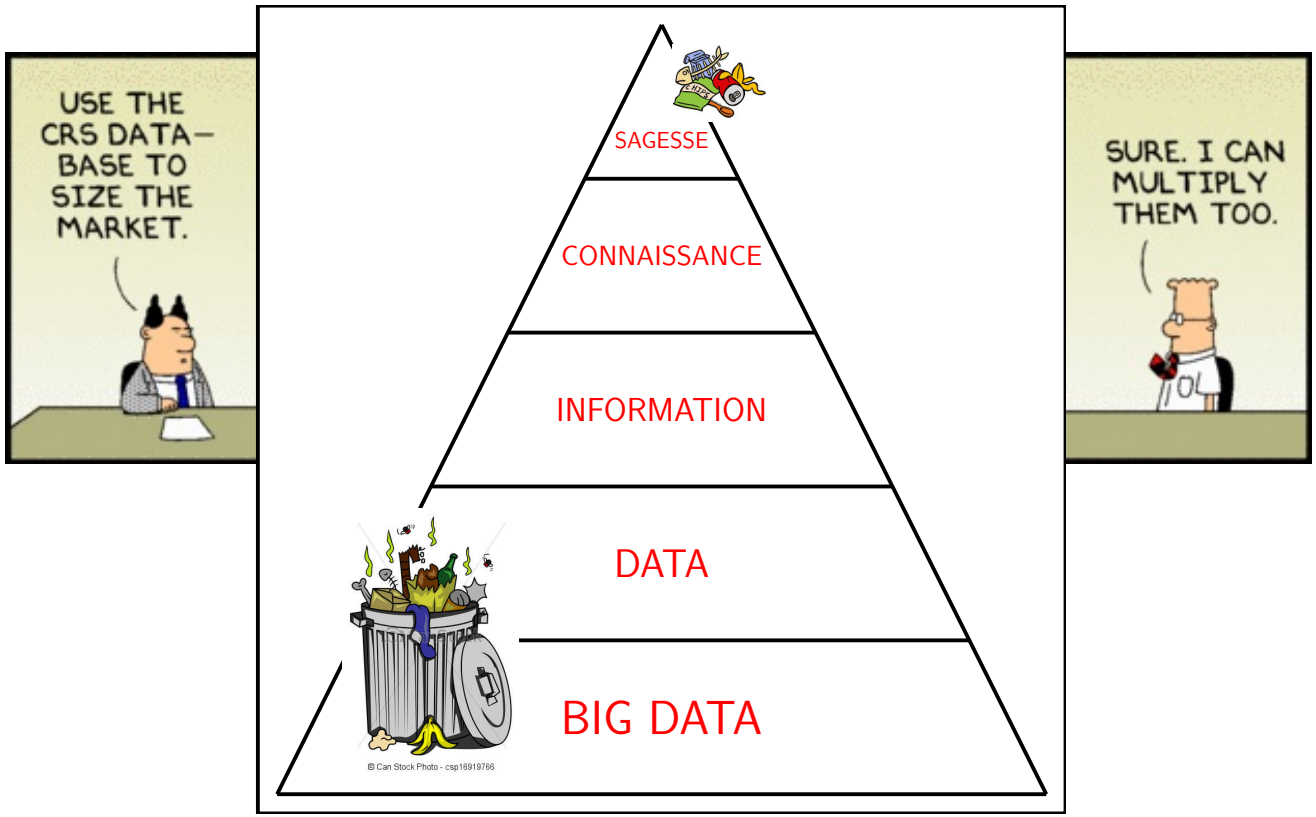
Que peut-on faire avec ces données?



La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

Que peut-on faire avec ces données?

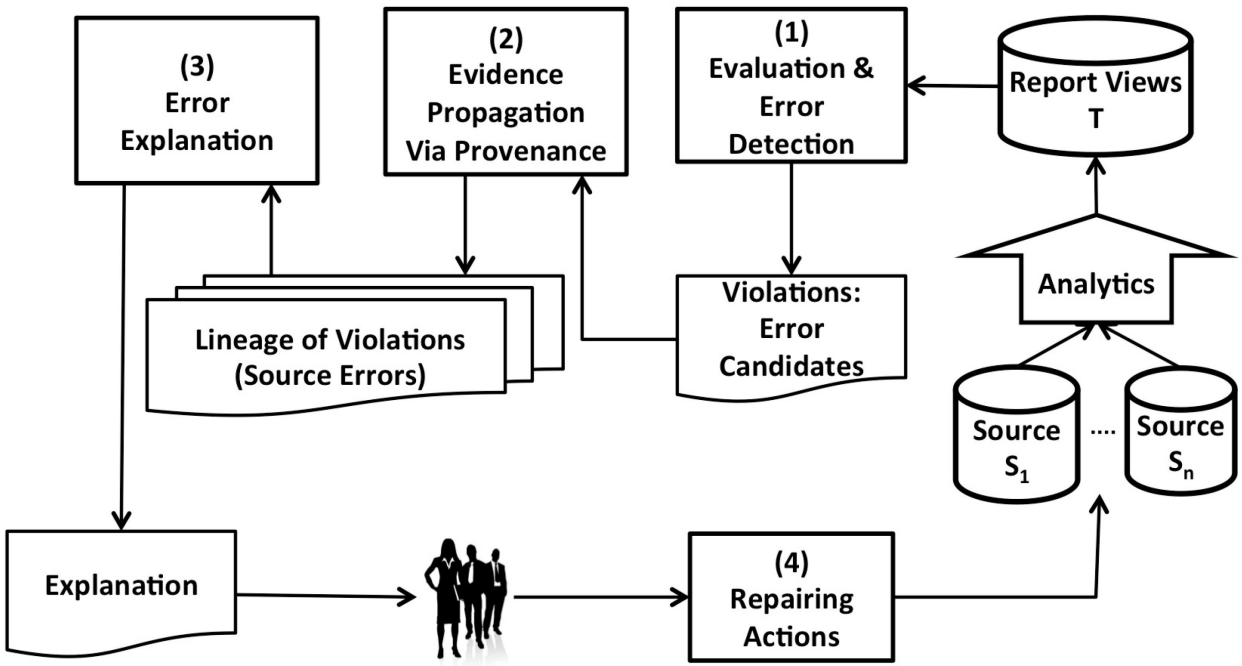


Clean Up

- Détection et suppression des doublons
- Détection et correction des erreurs
- Compléter des valeurs manquantes
- ...

Clean Up

- Détection et suppression des doublons
- Détection et correction des erreurs
- Compléter des valeurs manquantes
- ...



Source: Ihab F. Ilyas, *Effective Data Cleaning with Continuous Evaluation*

Sources of Data Pollution

System Conversions batch file systems → online processing monitor → hierarchical database systems → relational database systems

Heterogeneous System Integration

Poor Database Design

Input Errors

Incomplete Information at Data Entry For example, entry of NULL if birth date is unknown; or 9/9/99 if the birth date is unknown but mandatory.

Data Aging The older values lose their meaning and significance.

Internationalization/Localization As a company is internationalized, the existing data elements must adapt to newer and different values.

Fraud Incorrect data entries may be falsifications to commit fraud.

Lack of Data Quality Policies

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Data Purification

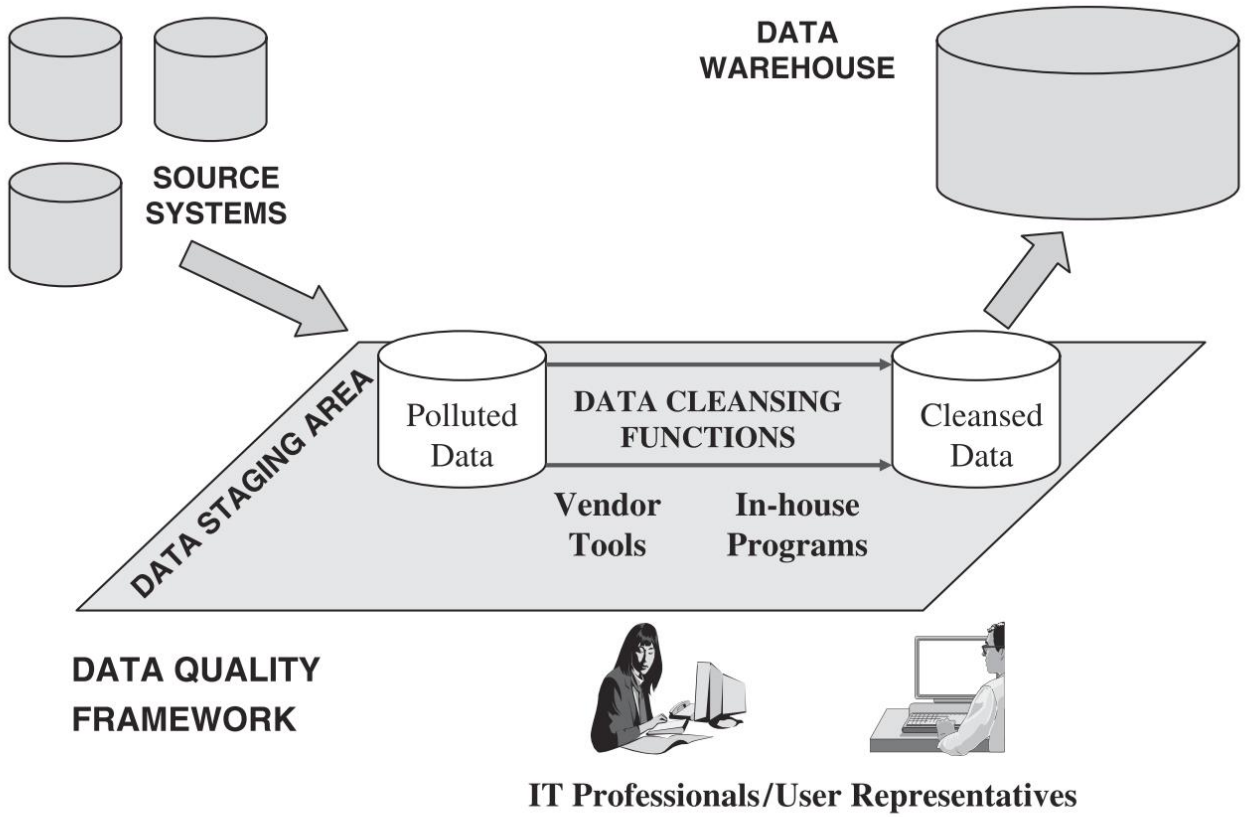


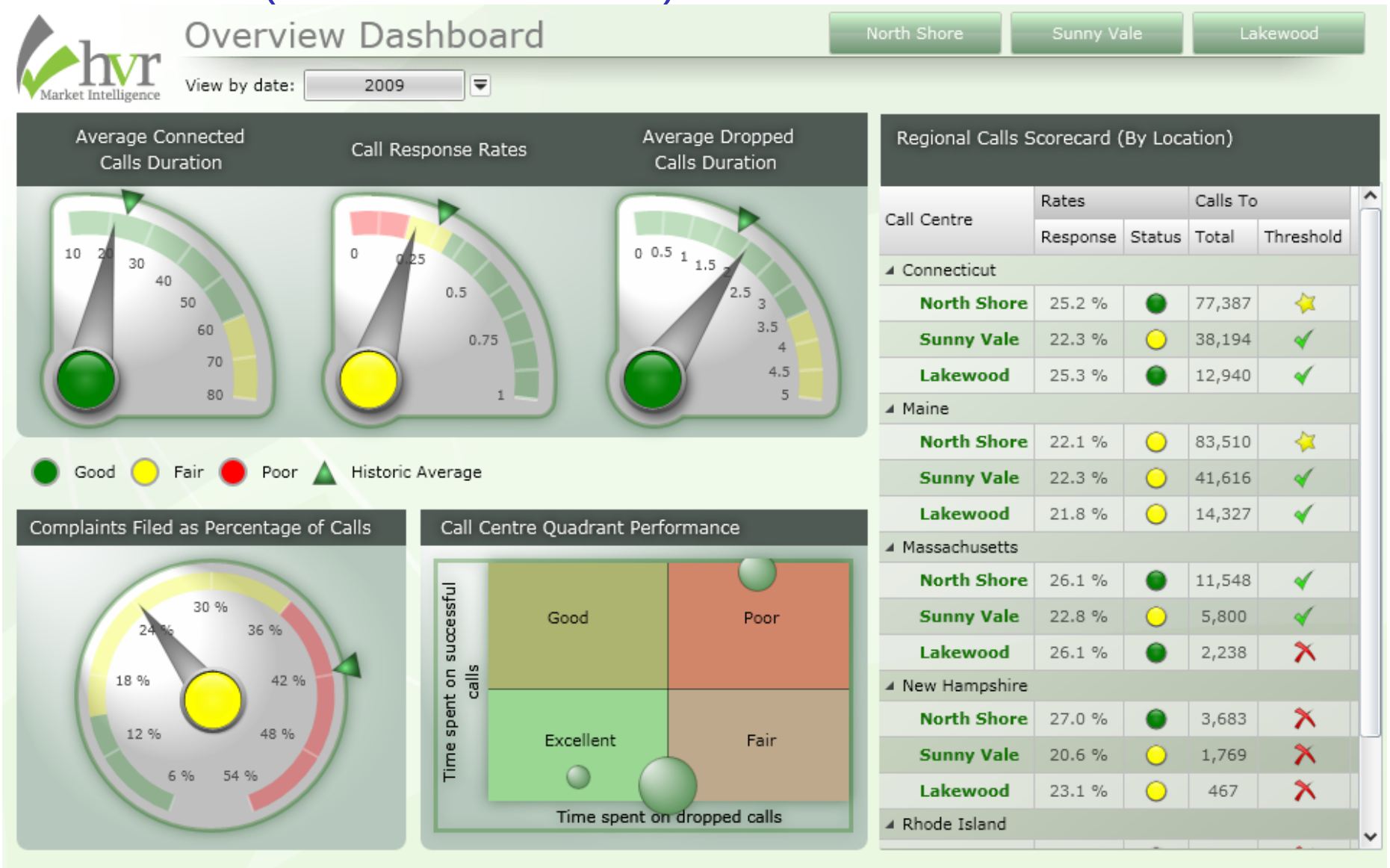
Figure 13-7 Overall data purification.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

Outline

Outline

Dashboard (tableau de bord)

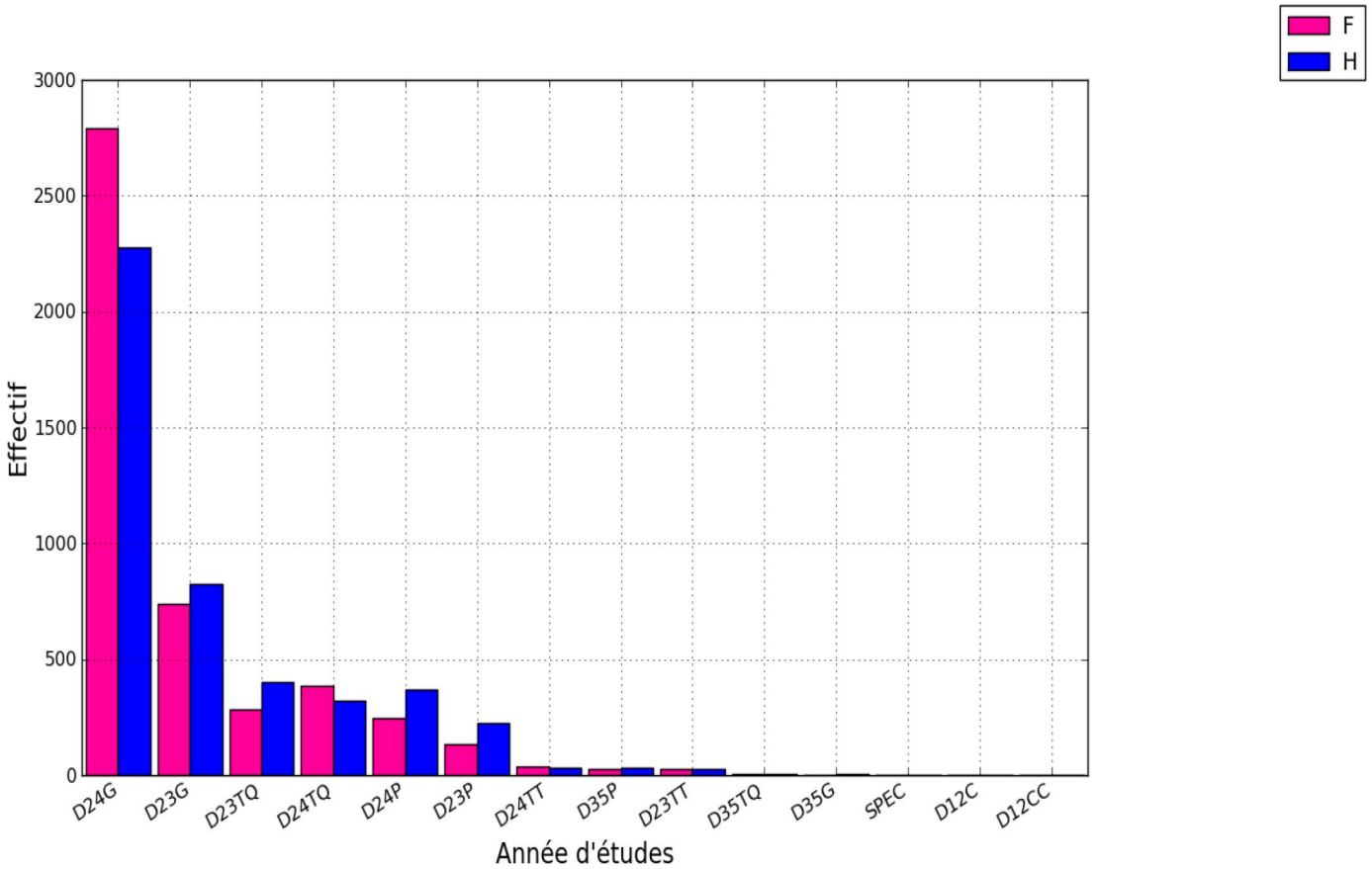


Source: www.dundas.com

Indicateur

Situation après 3 ans des élèves inscrits en 1ère secondaire commune une année donnée, selon le genre

Année d'entrée en 1ere secondaire : 2004



Outline

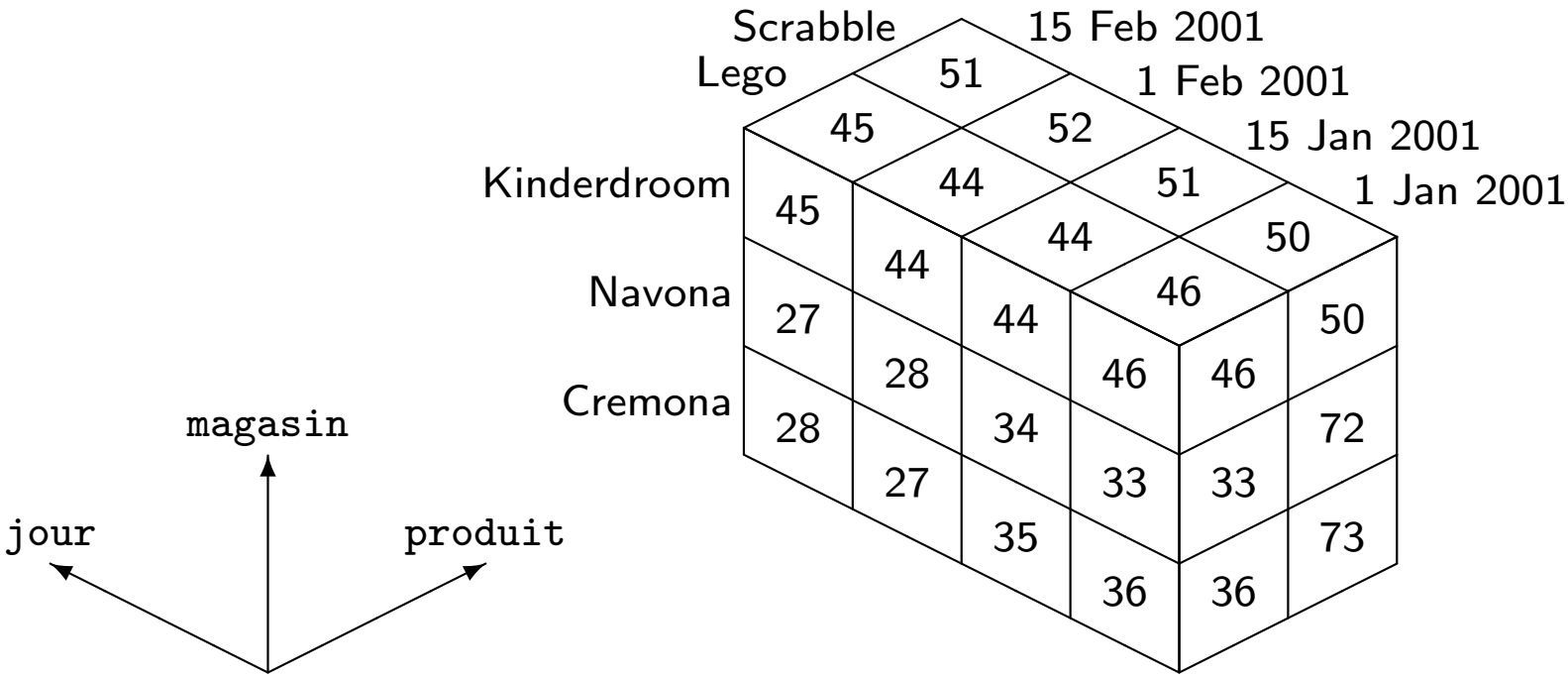
Dimensions et mesures

Typiquement, les analyses OLAP sont basées sur des rapports de résumé, par ex. les ventes quotidiennes par magasin et produit.
Les données peuvent être représentées de manière naturelle dans un data cube ("cube de données"):

- Les dimensions du cube correspondent aux variables indépendantes, par ex. `jour`, `magasin` et `produit`.
- Les cellules du cube contiennent les valeurs des variables dépendantes, par ex. le nombre de pièces vendues.

Les logiciels OLAP offrent différents types de *visualisation conviviales* des data cubes.

Cube de données



Un cube en 3D.

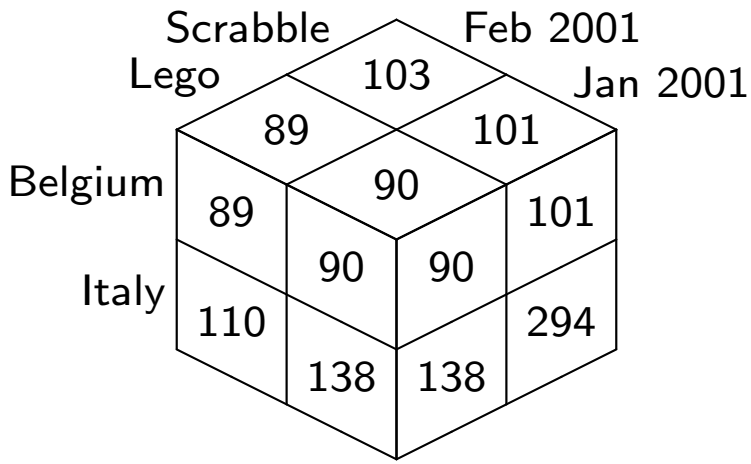
Hiérarchies de concepts

Les dimensions sont organisées en des hiérarchies conceptuelles qui déterminent les façons de regrouper les données.

Rollup

Les requêtes *rollup* donnent, pour chaque dimension, le niveau auquel l'information doit être présentée.

"Donne le total des ventes par produit, région et mois".

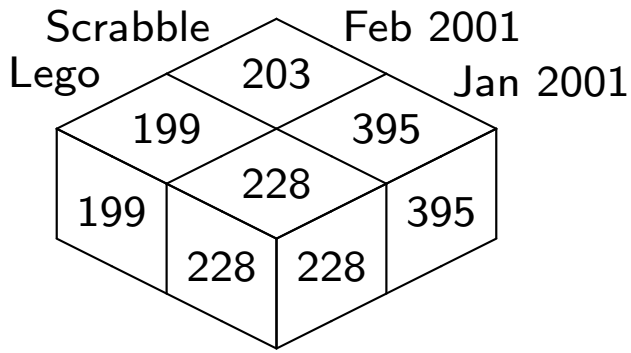


Le cube `mois region produit`.

Réduire la dimensionnalité

Les requêtes OLAP peuvent réduire le nombre de dimensions.

"Donne le graphique des ventes par produit, mois, pour tous les magasins."



Le cube `mois produit`.

Drilling

What's so great about this drilling stuff?

"Big deal," you might be thinking. "I can run reports with varying levels of detail in the query tool I've been using for years. What's so wonderful about this drill-down and drill-across business?"

The major advantage of business analysis (OLAP) drilling capability, as compared to traditional methods of getting this information, is that basic querying and reporting tools usually have had to run separate database access queries for each level of detail (often by using the SQL GROUP BY clause and along with an associated SQL WHERE clause). Each run is a separate SQL statement issued to the database, a separate pass through the database, a separate return of all the requested data, and a separate formatting of the results.

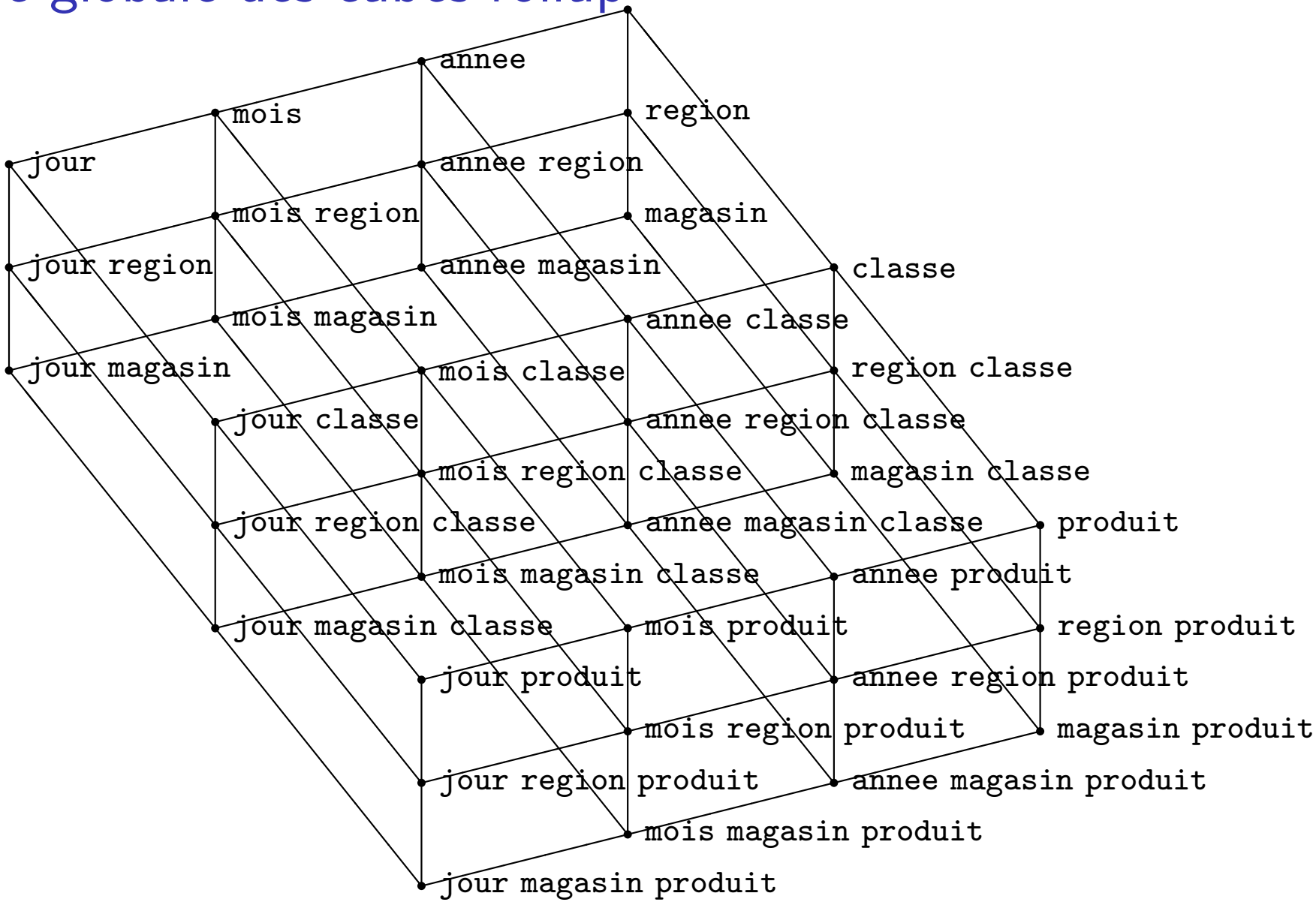
Multidimensional analysis and its drilling capability, on the other hand, are instantaneous because the information you need is staged for you. By clicking the mouse or selecting a command, you see less detail, more detail, or whatever you want. The tool and the database don't have to collaborate for successive data access requests — it's all there for you.

Hint: If you haven't used a drill-down feature and want to get a feel for it, try using the HIDE/UNHIDE features for rows and columns in your spreadsheet program. Set up a set of detailed rows of data, total them into another row, and then do the same thing again. When you HIDE the detail rows, you're performing a drill-up function; when you UNHIDE them, you're drilling down.

As mentioned in Chapter 9, some reporting tools now have business analysis (OLAP) drill-down capabilities, which blurs the distinction between members of these two classes of business intelligence tools.

Source: Thomas C. Hammergren and Alan R. Simon: *Data Warehousing for Dummies* (2nd Edition). Wiley Publishing, 2009

Vue globale des cubes rollup



Choix technologique : ROLAP ou MOLAP

Défis technologiques en OLAP :

> *Supporter de manière efficace les opérations arithmétiques sur les data cubes de plusieurs gigabytes.*

Dépendant de la technologie utilisée, on peut classer les logiciels OLAP en deux catégories :

- ROLAP (*Relational OLAP*), ou
- MOLAP (*Multidimensional OLAP*).

ROLAP

- Le data cube est stocké dans une base de données standard (i.e. SQL), dans un schéma "en étoile".
- Les serveurs de base de données sont munis d'extensions *middleware* pour le support de l'OLAP. Par ex. Microsoft SQL Server OLAP Services.
- Le langage de requête SQL est étendu avec des primitives OLAP.

ROLAP

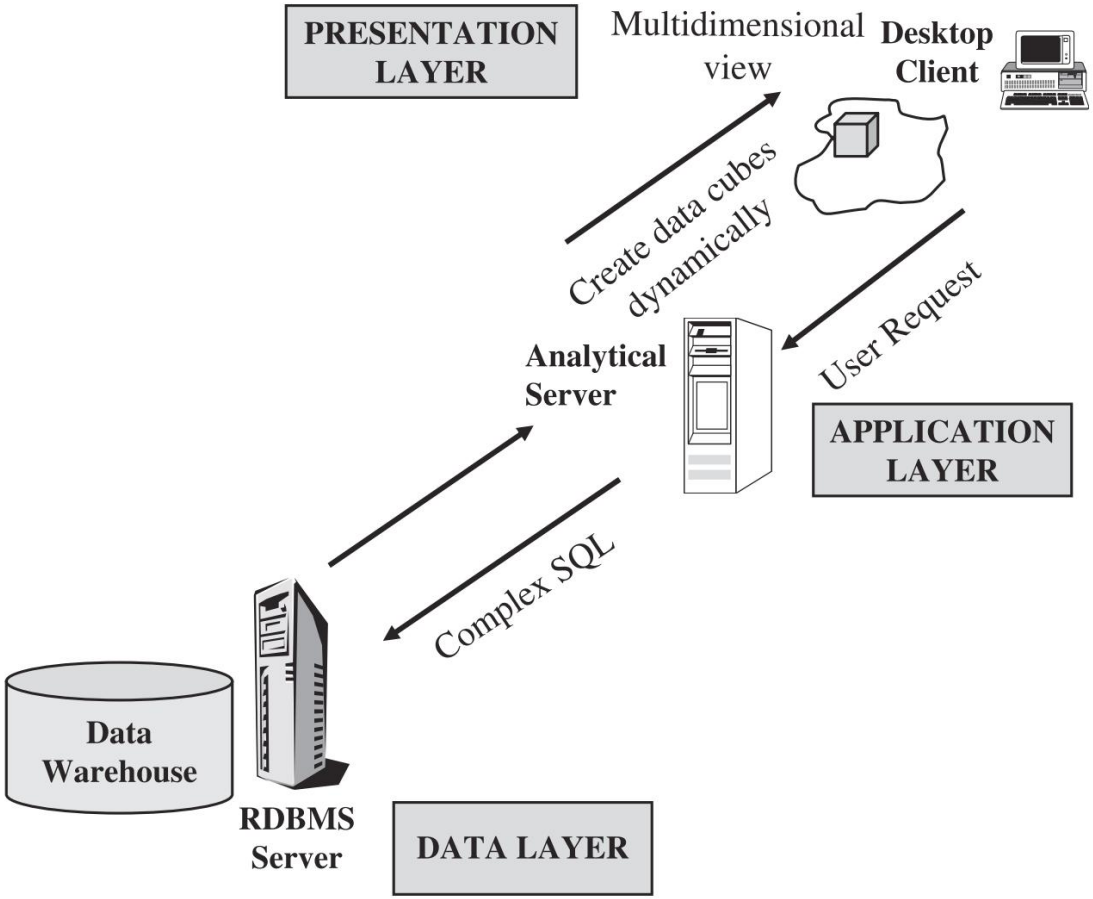


Figure 15-17 The ROLAP model.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

MOLAP

- Au lieu de s'appuyer sur des tables SQL, MOLAP stocke les data cubes dans des matrices multidimensionelles.
- Cette manière de stocker les données peut être plus efficace que ROLAP.
- Un inconvénient est que l'intégration avec les bases de données SQL existantes est plus difficile.

MOLAP

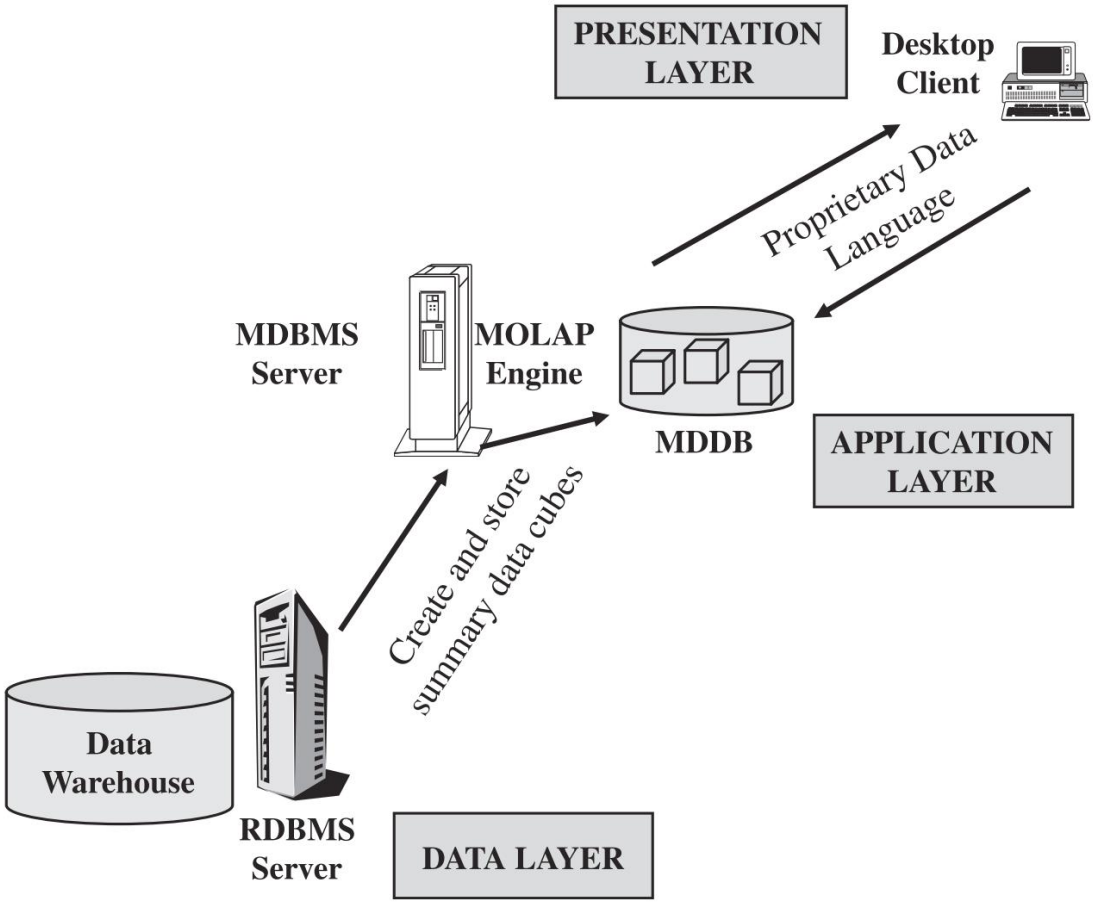


Figure 15-16 The MOLAP model.

Source: Paulraj Ponniah: *Data Warehousing. Fundamentals for IT professionals* (2nd Edition). John Wiley & Sons, 2010

OLAP $\rightsquigarrow$ Data Mining

- En OLAP, l'utilisateur final oriente l'analyse :
  1. le choix des dimensions et des variables, et
  2. la spécification des requêtes.
- Problème : le contenu du data warehouse n'est souvent pas bien compris et il est donc quasi impossible de choisir le bon data cube et de poser les bonnes questions.
- Point de départ du data mining : utiliser la puissance de l'ordinateur pour découvrir des modèles intéressants dans les bases de données – plutôt que de vérifier des hypothèses (idées préconçues).

Outline

What is Data Mining?

Knowledge Discovery in Databases (KDD) is the process of identifying valid, novel, useful, and understandable patterns from large datasets.

Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover previously unknown patterns.

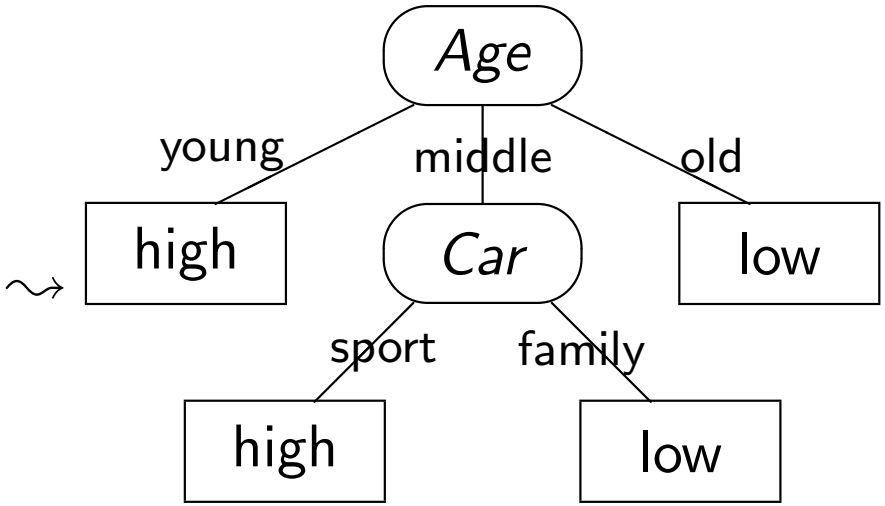
Source: Oded Maimon, Lior Rokach (Eds.): *The Data Mining and Knowledge Discovery Handbook* (2nd Edition). Springer, 2010

Applications

- *Credit scoring*.
- Classement automatique d'objets stellaires.
- Campagne de courrier ciblé.
- Détection de fraude.
- ...

DM Example: Model

Age	*Sex*	...	*Car*	*Risk*
young	M	...	sport	high
middle	M	...	sport	high
middle	F	...	family	low
⋮	⋮	⋮	⋮	
old	F	...	sport	low



Data Mining a Sky Survey

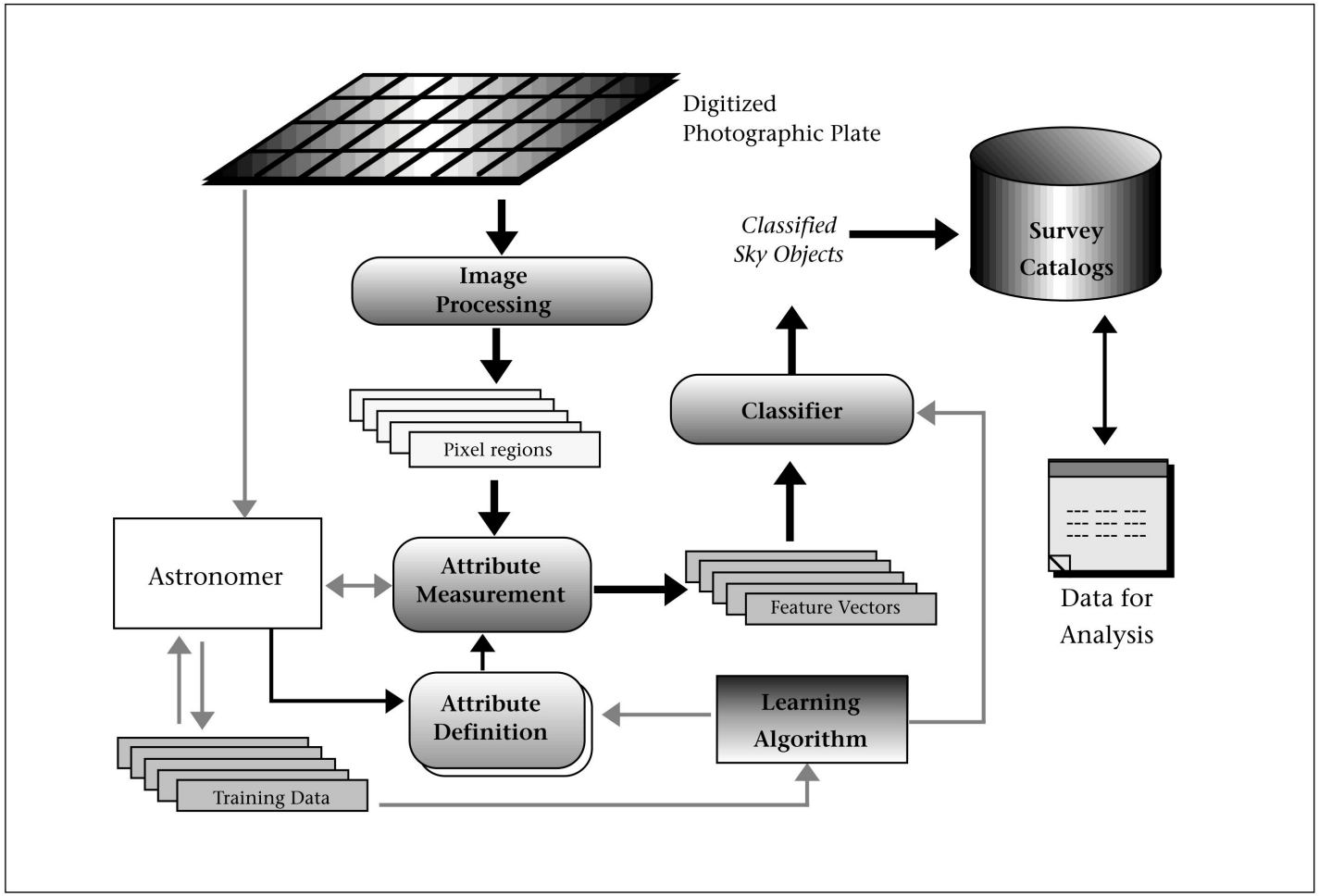


Figure 3. An Overview of the SKICAT Plate-Cataloging Process.

Source: Usama M. Fayyad et al.: *From Digitized Images to Online Catalogs. Data Mining a Sky Survey* AI Magazine. 17(2), 1996

Data Mining a Sky Survey

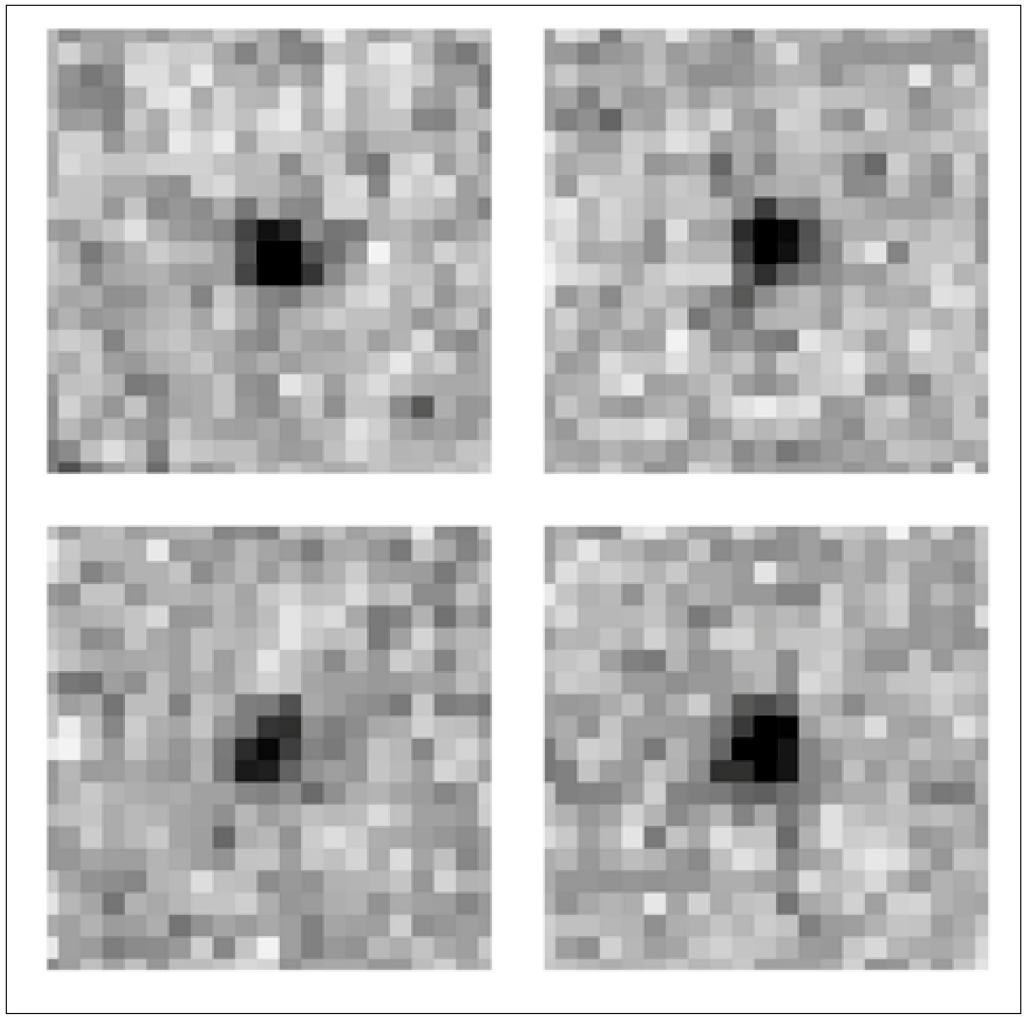


Figure 6. An Illustrative Example: Four Faint Sky Objects.

Source: Usama M. Fayyad et al.: *From Digitized Images to Online Catalogs. Data Mining a Sky Survey* AI Magazine. 17(2), 1996

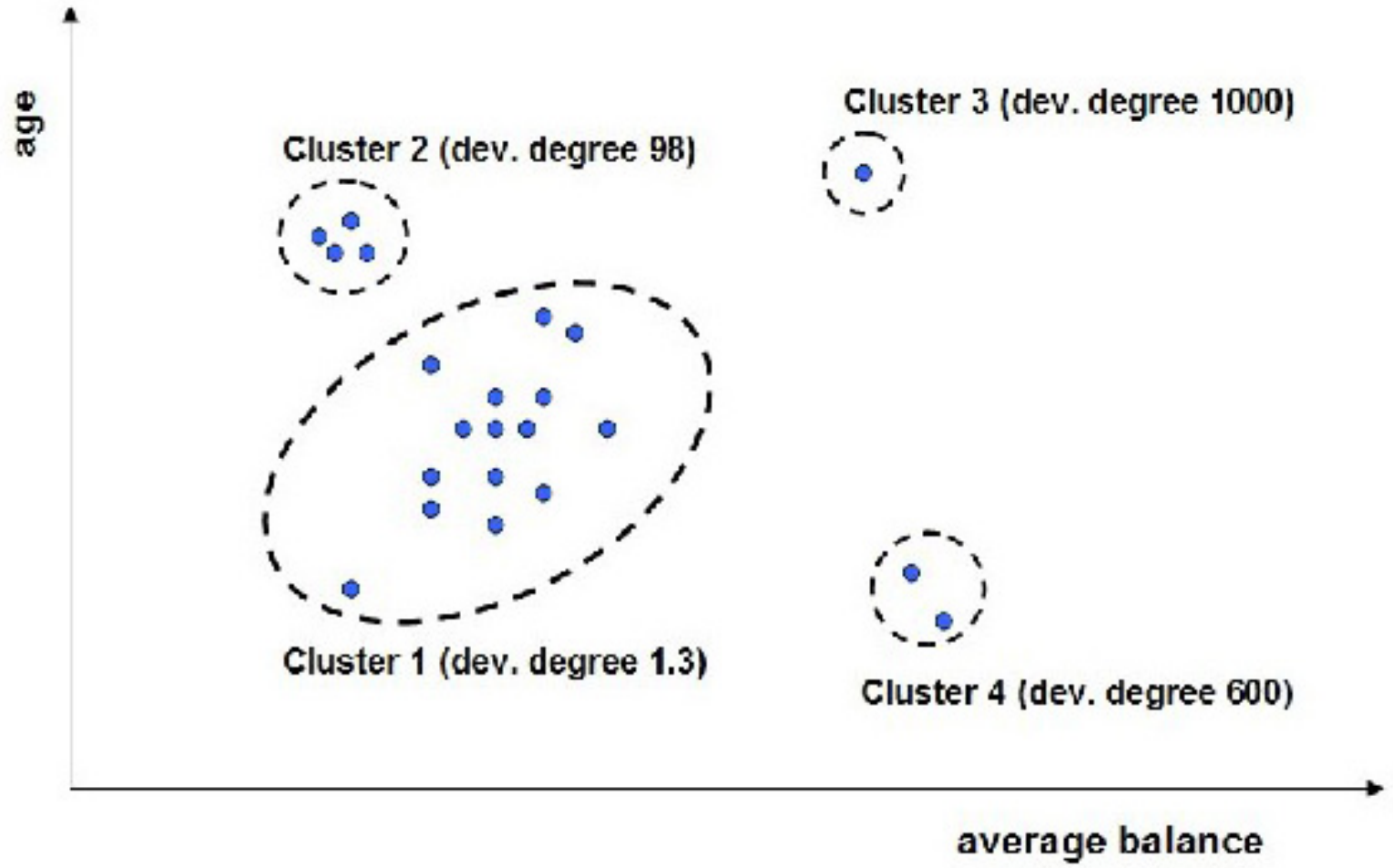
DM Example: Unknown Pattern

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

$\rightsquigarrow$ {Diapers}$\rightarrow${Beer}

"many customers who buy diapers also buy beer"

DM Example: Deviation Detection by Clustering



Source: www.ibm.com

The KDD Process

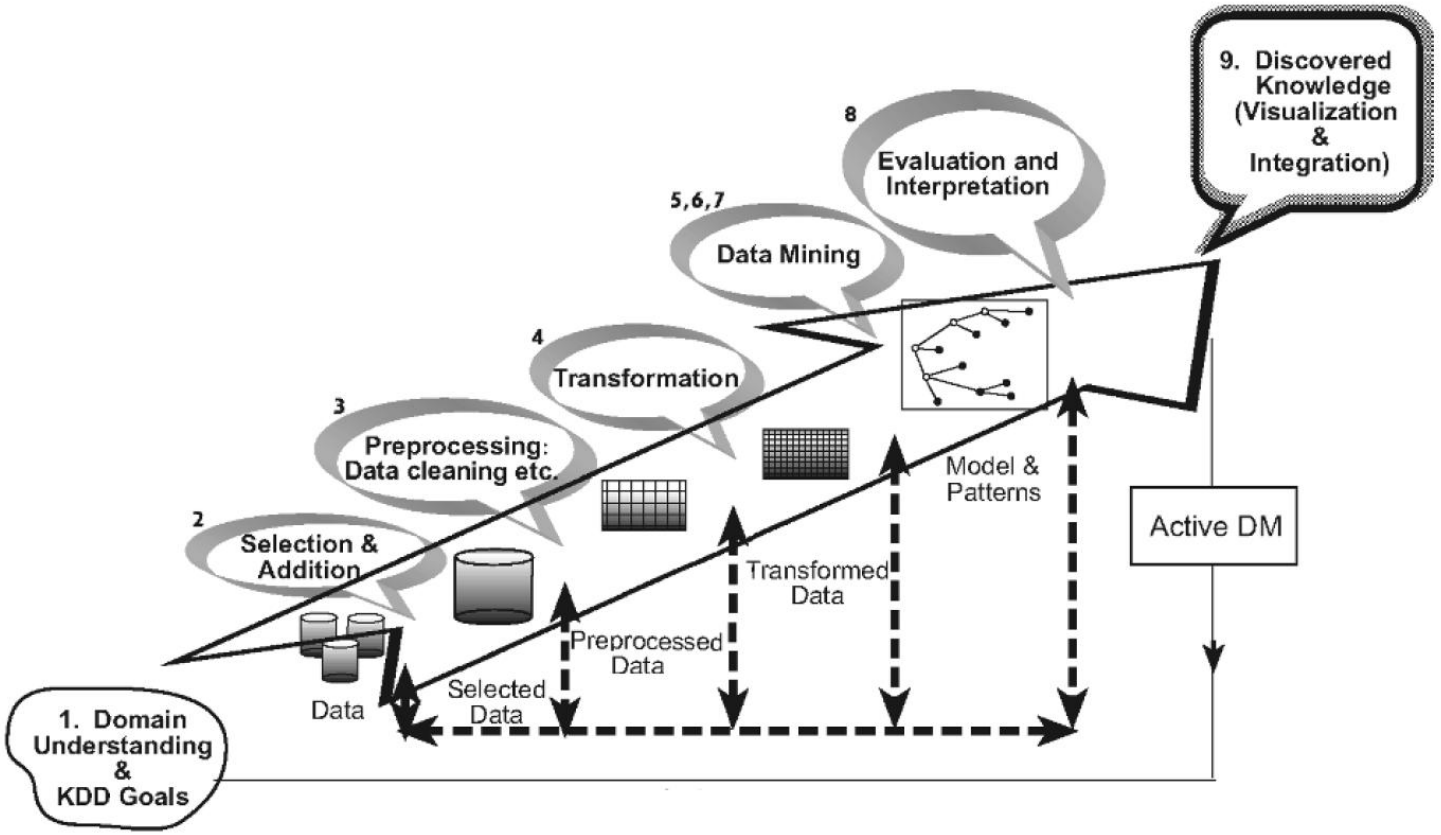


Fig. 1.1. The Process of Knowledge Discovery in Databases.

Source: Oded Maimon, Lior Rokach (Eds.): *The Data Mining and Knowledge Discovery Handbook* (2nd Edition). Springer, 2010

Data Mining Taxonomy

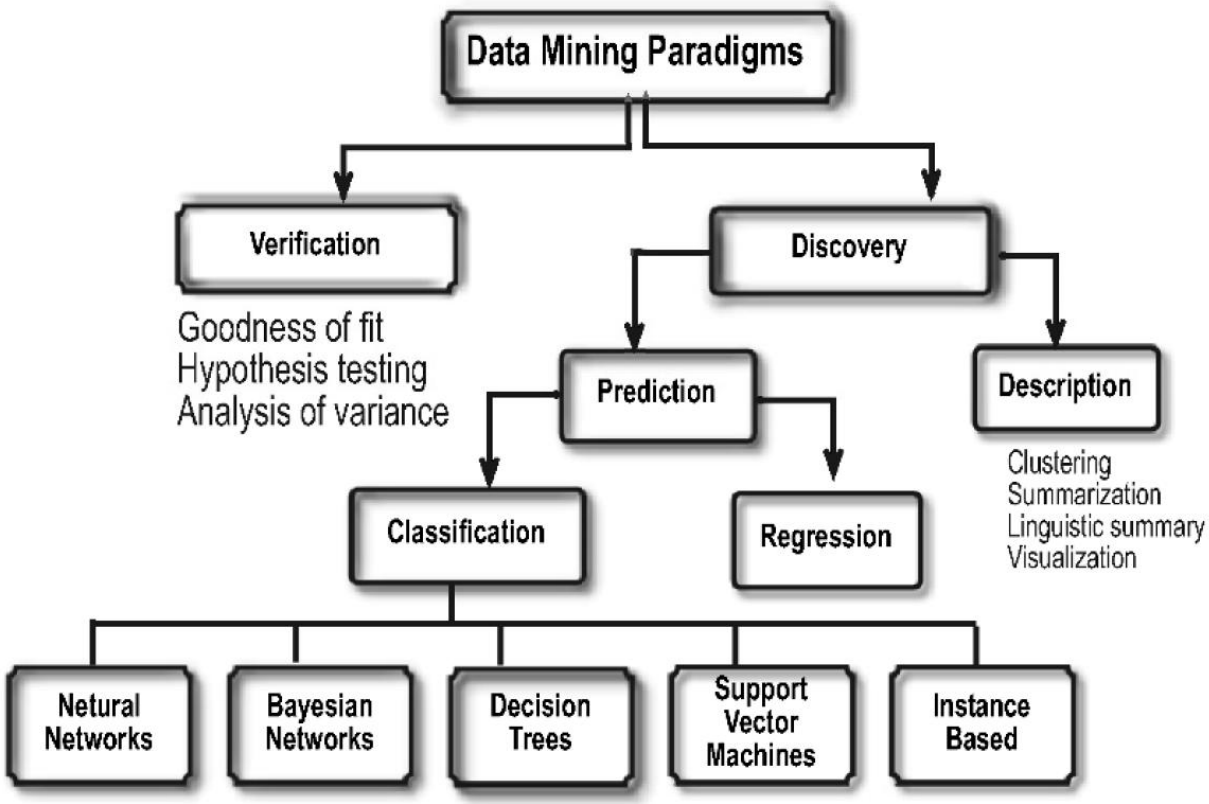


Fig. 1.2. Data Mining Taxonomy.

Source: Oded Maimon, Lior Rokach (Eds.): *The Data Mining and Knowledge Discovery Handbook* (2nd Edition). Springer, 2010