Data Warehousing & Data Mining

Business Intelligence

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. . .

  ☛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.
  - 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 ?

Étude de cas

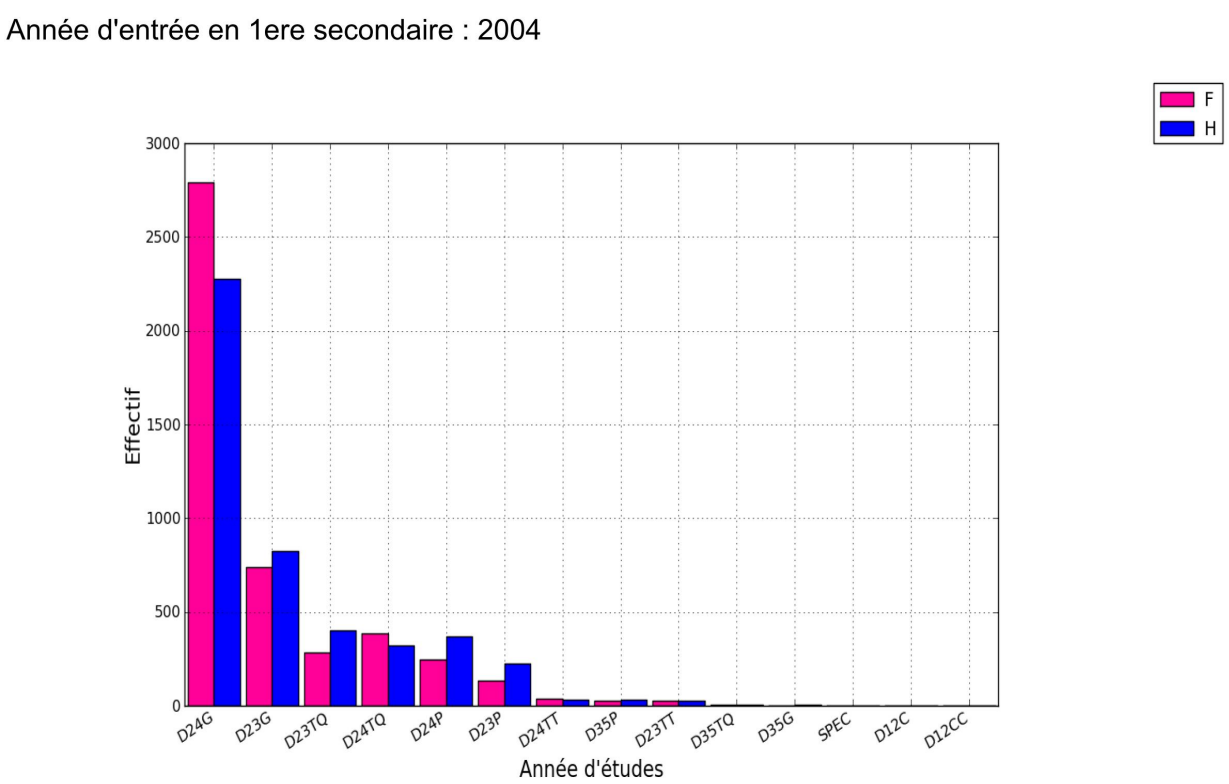
- Développement d'outils de pilotage effectif du réseau de la Communauté française
- Données opérationnelles :

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

- 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)
4. TP

Outline

Informatique décisionnelle

L'informatique décisionnelle (en anglais : DSS pour Decision Support System ou encore BI pour Business Intelligence) désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données [...] d'une entreprise en vue d'offrir une aide à la décision et de permettre aux responsables de la stratégie d'entreprise d'avoir une vue d'ensemble de l'activité traitée.

Source: Wikipedia

Operational to Decisional

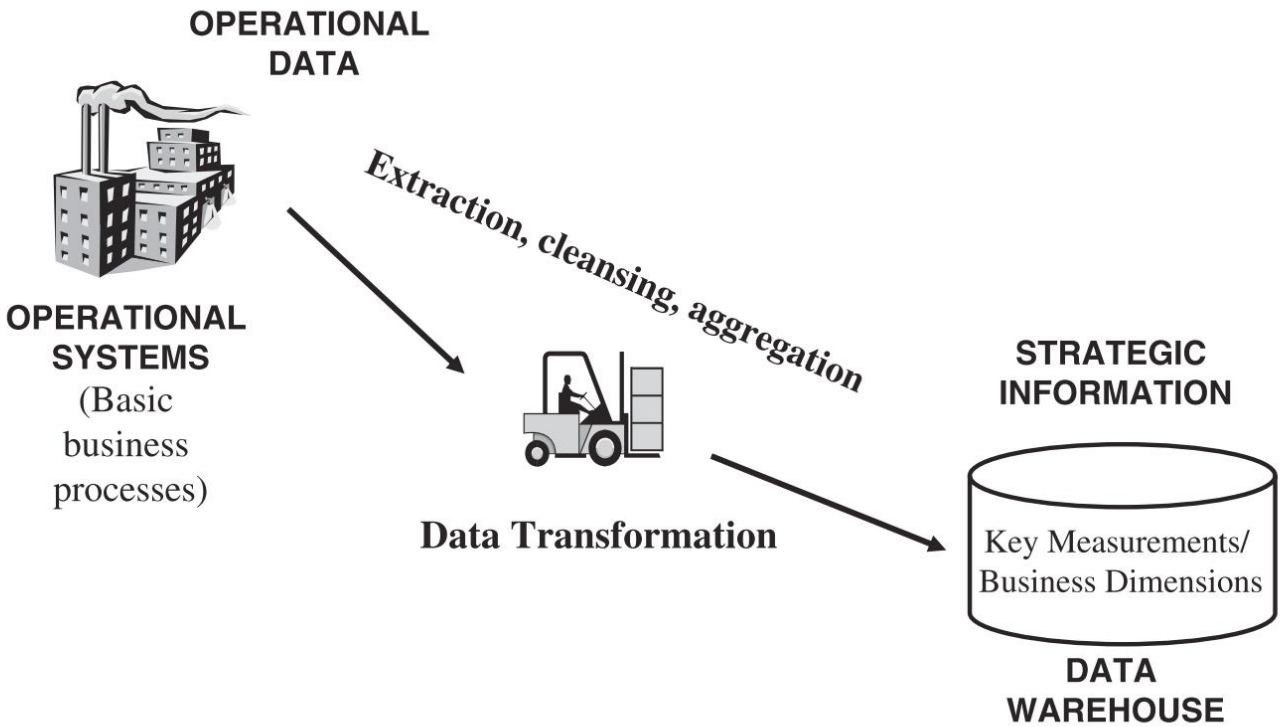


Figure 1-8 General overview of the data warehouse.

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

Étude de cas

Y. Baeyens. *Le data-mining chez Proximus : Prédiction des créances douteuses.* Travail de fin d'étude, Université de Mons-Hainaut.

- Peut-on assigner une classe de risque (risque de devenir un jour mauvais payeur) à chacun des nouveaux clients de Proximus ?
- Peut-on assigner une classe de risque à tous les clients existants ?
- Peut-on identifier, parmi les poursuites actuelles, celles qui sont susceptibles d'évoluer vers un état *bad-dept* ?
- Est-ce que les clients se mettent en ordre juste après une action de recouvrement ou est-ce qu'ils attendent quelques jours encore ?
- Est-ce que la qualité de la promesse de paiement est indépendante du moment où la promesse est faite ?
- Peut-on obtenir une carte de Belgique permettant d'identifier les communes ou les groupes de communes à forte concentration de mauvais payeurs ?

Running $\rightsquigarrow$ Monitoring the Business

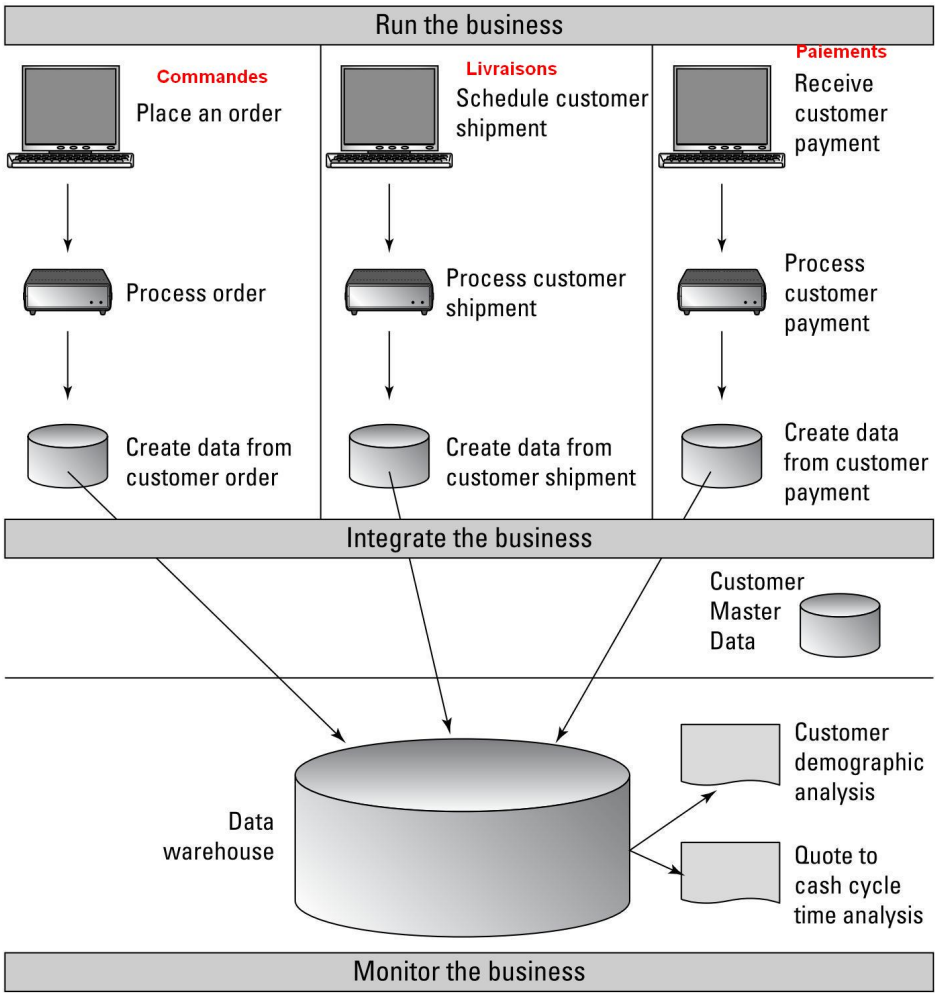


Figure 1-1: Most computer applications create data as a result of an activity or transaction; a data warehouse instead swipes data created elsewhere and converts it into information.

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

Business Intelligence (BI): Two Environments

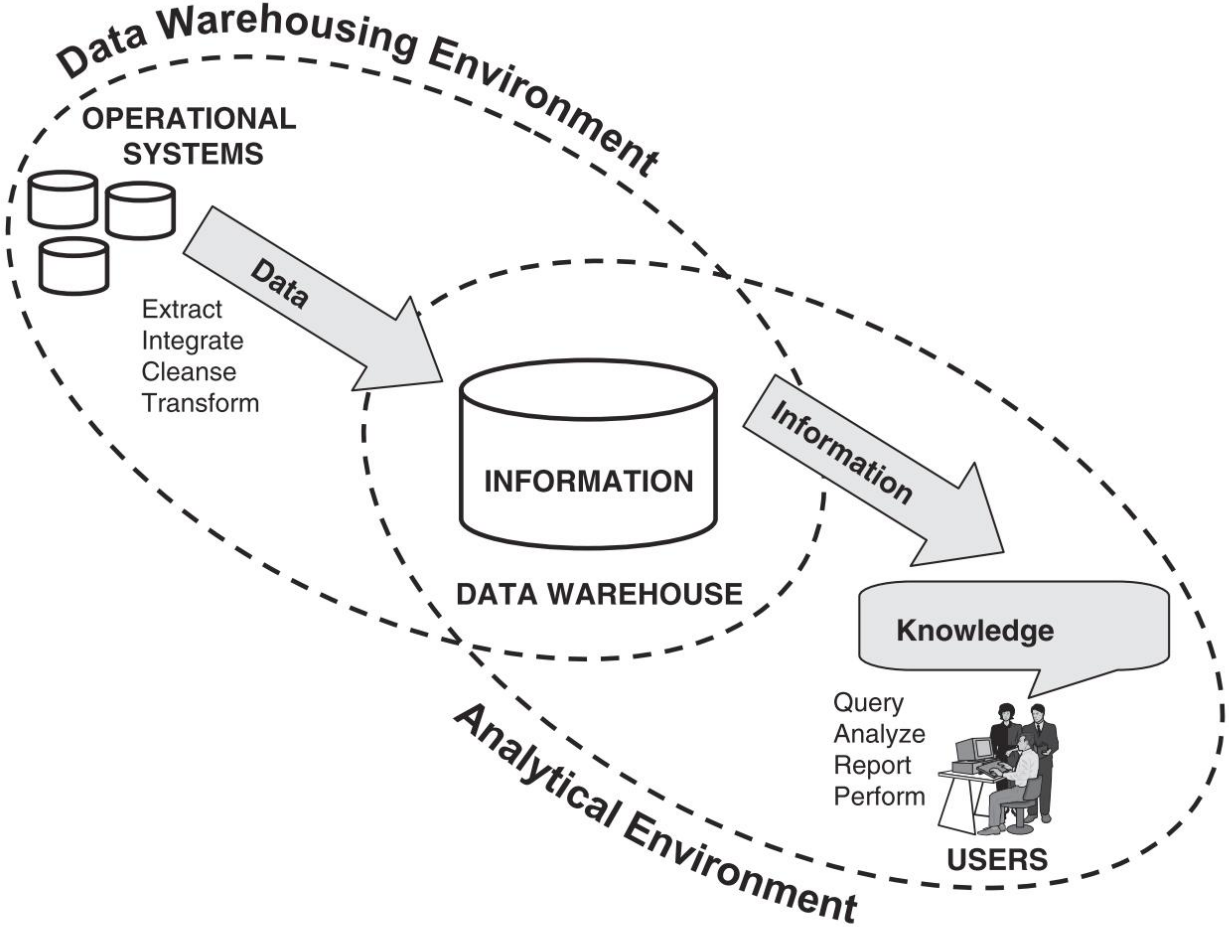


Figure 1-10 BI: data warehousing and analytical environments.

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

Data Warehousing, OLAP and Data Mining

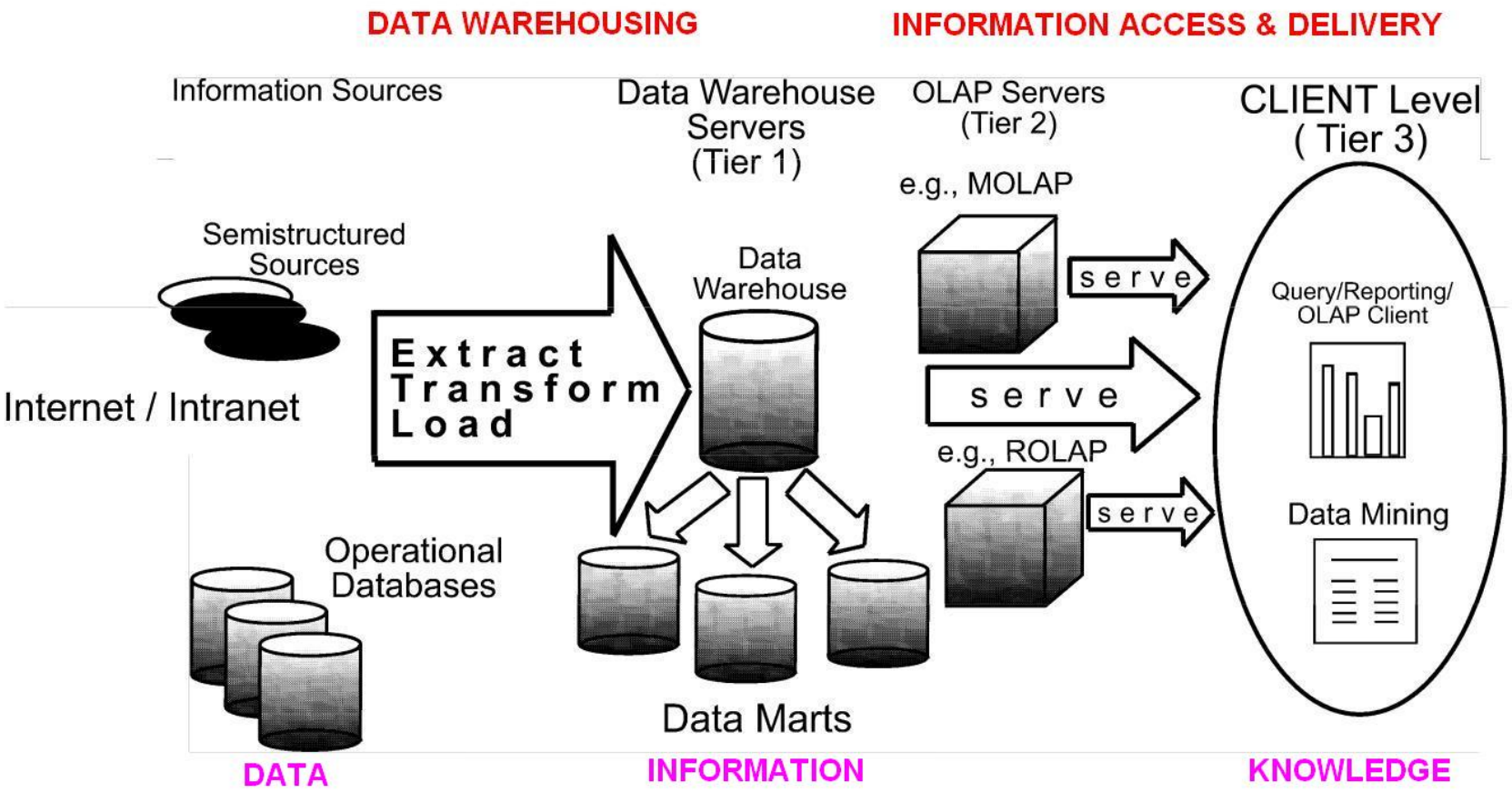


Fig. 1.3. The IT Decision Support Tiers.

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.

Decision Making

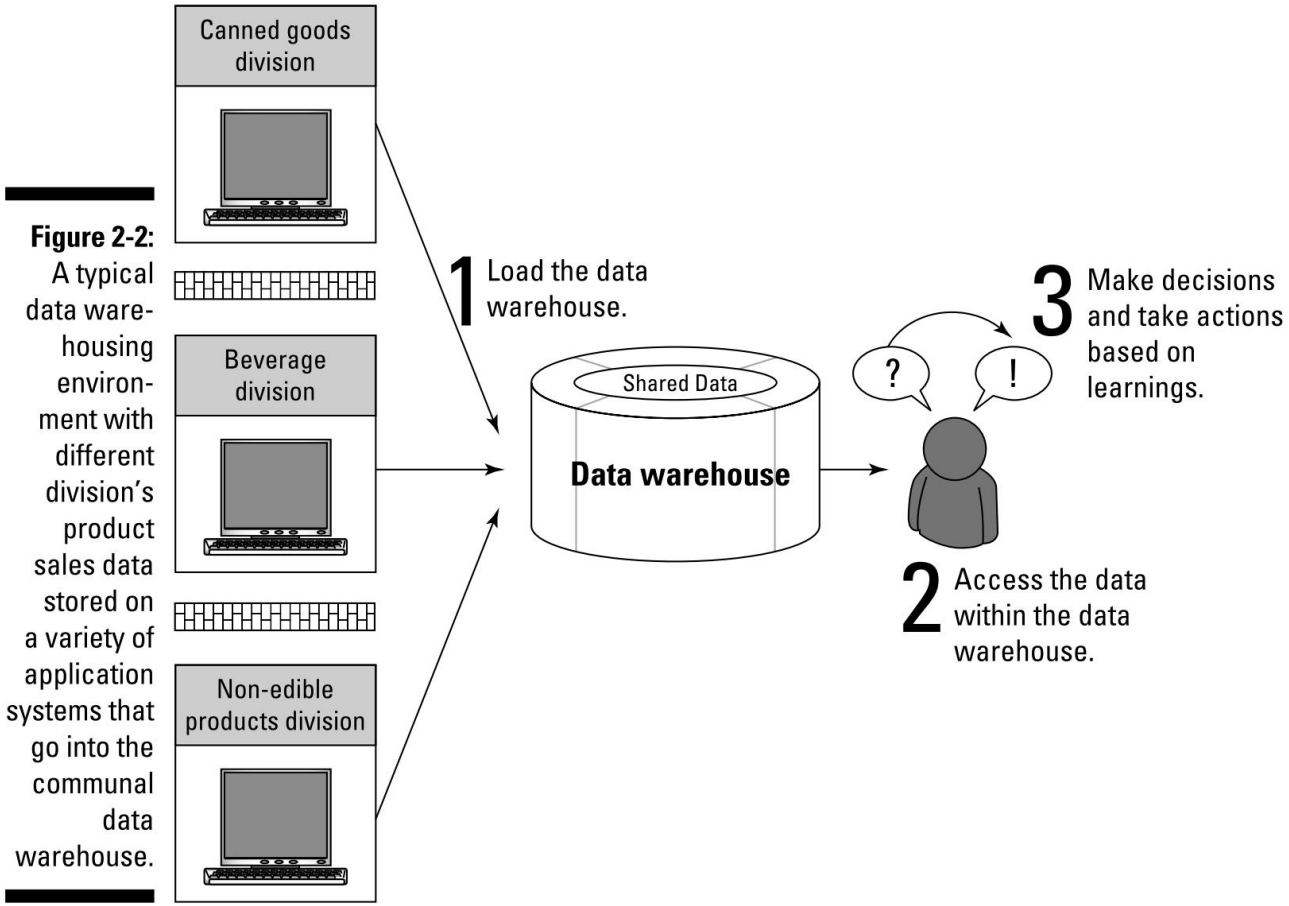


Figure 2-2: A typical data warehousing environment with different division's product sales data stored on a variety of application systems that go into the communal data warehouse.

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

Subject Oriented (Thématique)

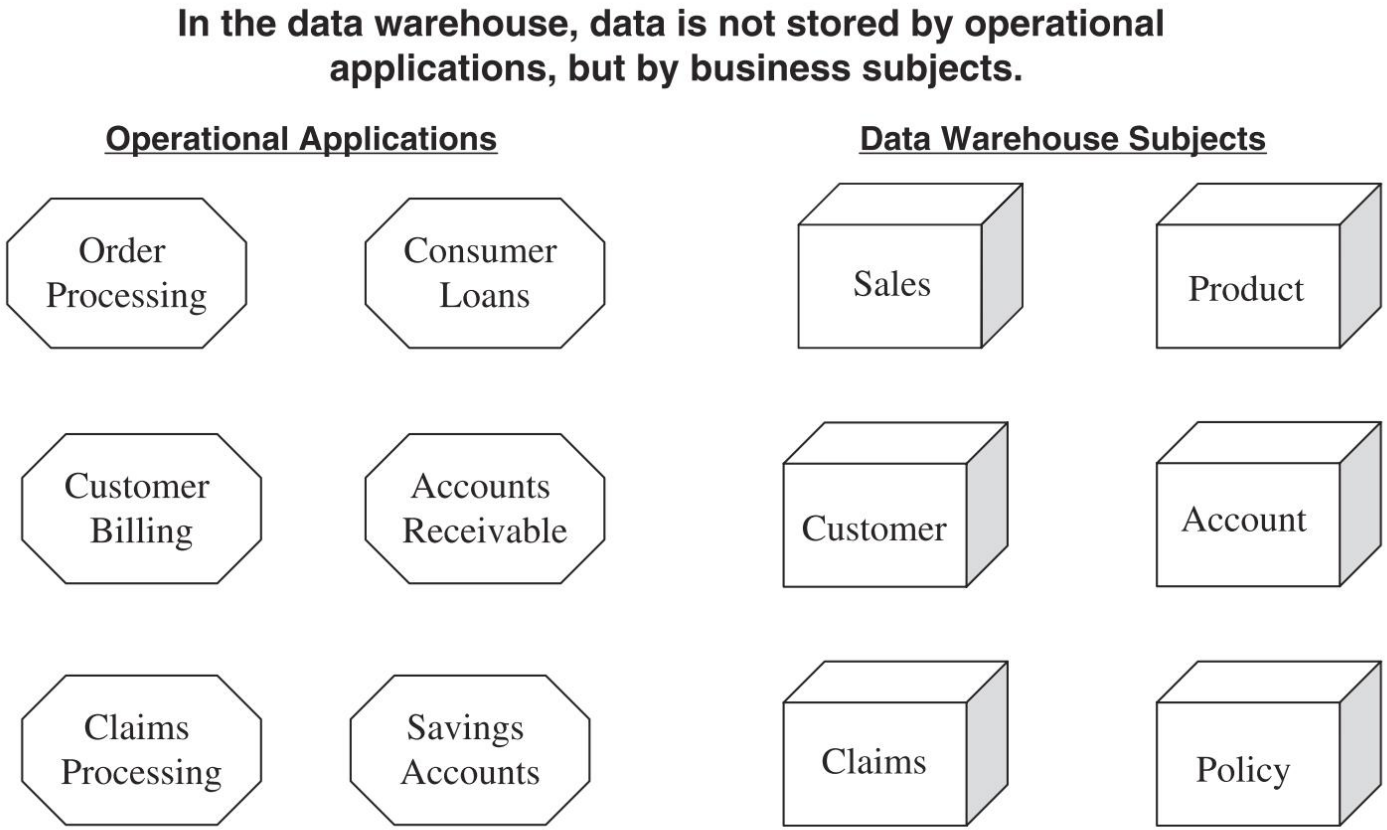


Figure 2-1 The data warehouse is subject oriented.

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

Integrated

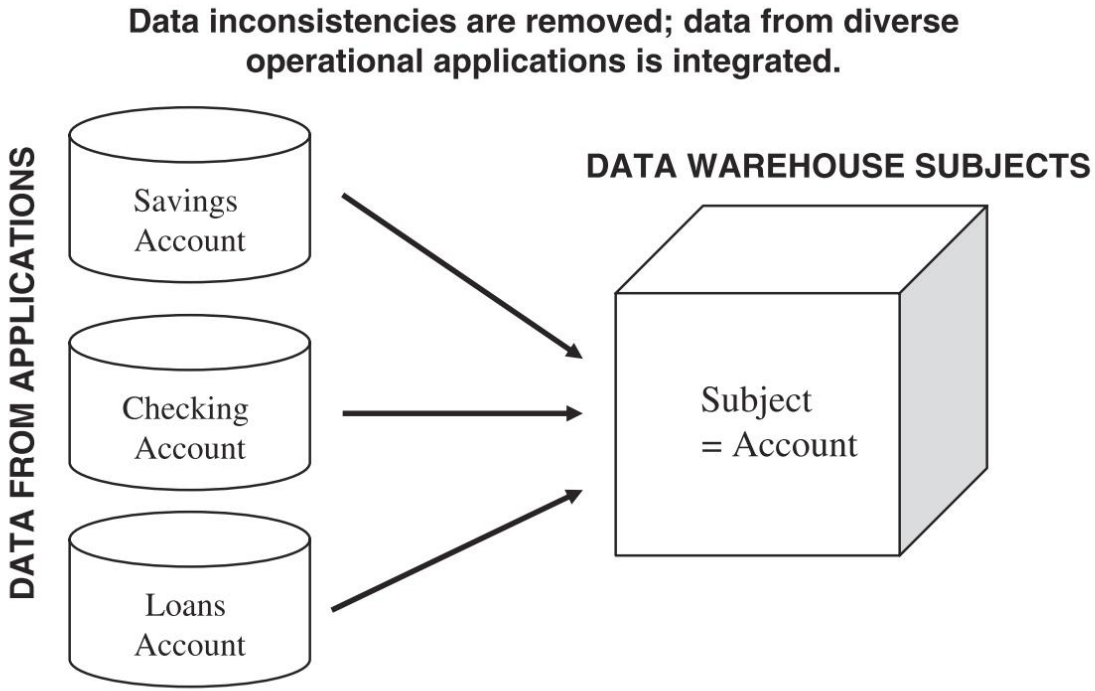


Figure 2-2 The data warehouse is integrated.

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

Nonvolatile

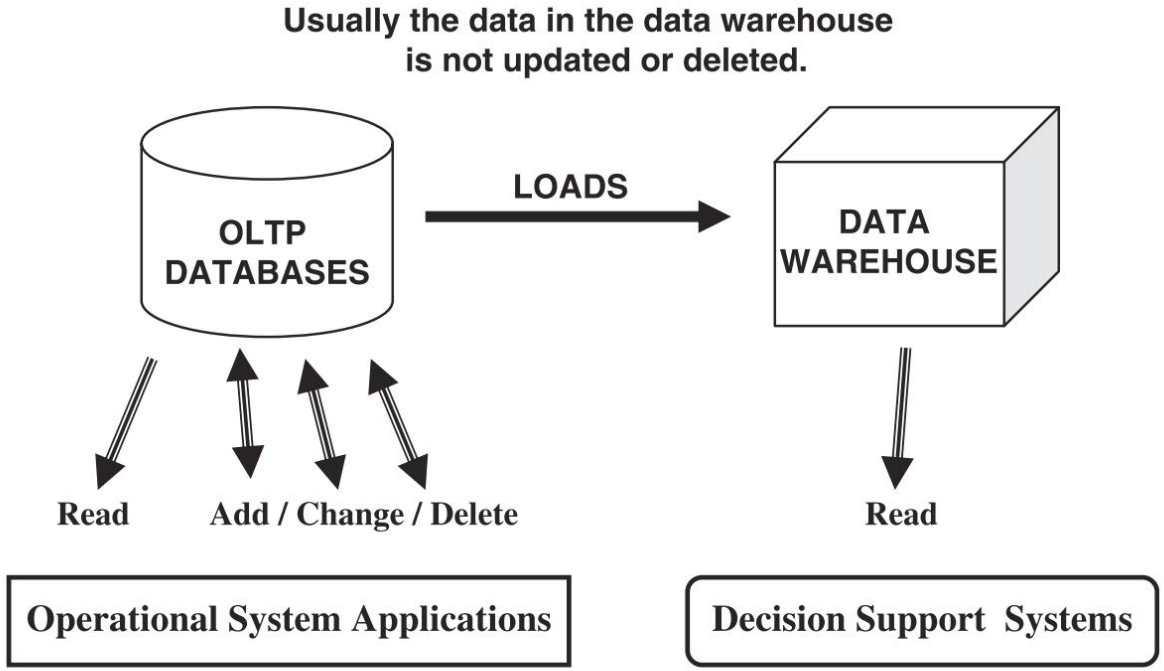


Figure 2-3 The data warehouse is nonvolatile.

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

Operational vs Decisional

	Transaction Processing	Analytical Processing
Users	numerous employees	relatively few managers
Workload	very frequent transactions:	less frequent analyses:
Access Type	read and write, individual records	read, database scans
Data Content	current values	current and historical values, derived and summarized
Database size	100 MB to GB	100 GB to TB ($= 10^6$ MB)

Data Mart

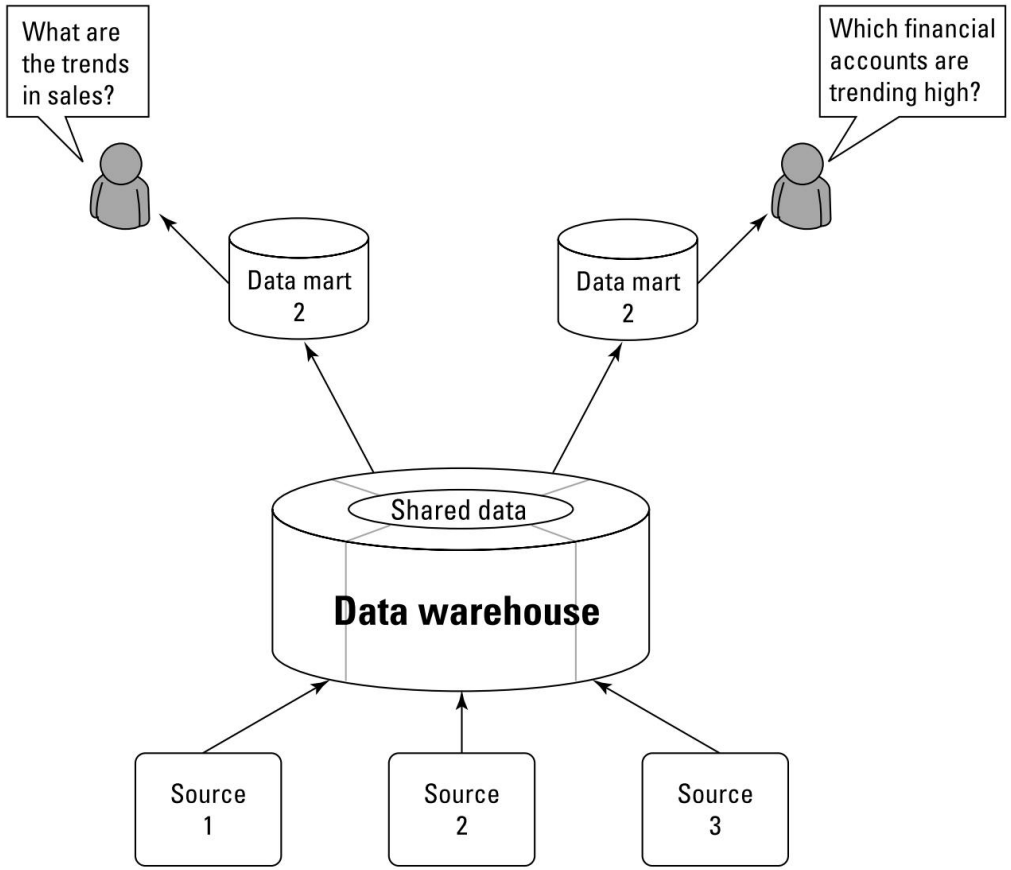


Figure 4-1: The retail-outlet approach to data marts: All the data comes from a data warehouse.

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

Data Mart

Un data warehouse au niveau d'un département sur une partie spécifique du business.
Par ex. data mart pour le marketing se focalisant sur les clients, produits et ventes.

Deux types de data marts :

Data mart sans data warehouse. Ces data marts peuvent être construits facilement car ils ne nécessitent pas de modèle de données conceptuel au niveau de l'entreprise. Néanmoins, ils peuvent à long terme faire apparaître des problèmes d'intégration complexes.

Data mart extrait du data warehouse. Pour des raisons de flexibilité et de performance.

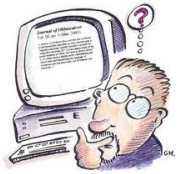
Operational vs Decisional

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

BI Categories

Business Intelligence Categories	
Type	***Information You Want***
Basic querying and reporting	"Tell me what happened."
Business analysis (OLAP)	"Tell me what happened and why."
Data mining	"Tell me what might happen" or "Tell me something interesting." prediction
Dashboards and scorecards	"Tell me a lot of things, but don't make me work too hard."

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



Travail Pratique

- Concrétiser le schema suivant pour une chaîne de supermarchés.

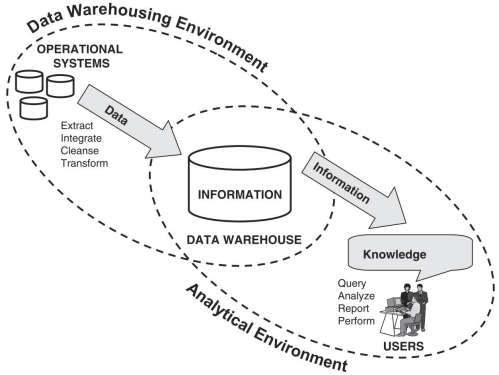
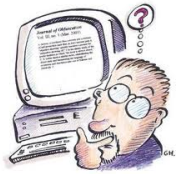


Figure 1-10 BI: data warehousing and analytical environments.

- Préciser la nature et l'origine des données opérationnelles.
- Quelle aide l'informatique décisionnelle peut-elle apporter aux décideurs ?
- Davantage qu'un outil de fidélisation, les cartes à points s'avèrent être de formidables outils de marketing direct pour les grandes surfaces. Quelles opportunités stratégiques offre la carte à points ?



La carte-plus chez Delhaize

☐ Monsieur ☐ Madame ☐ Mademoiselle ☐ Entreprise

Date de naissance: __/__/____

Nom: _____

Prénom: _____

Rue: _____

N°: ____ Boîte: ____ Code Postal: ____

Localité: _____

Pays: ☐ L ☐ B ☐ D ☐ F ☐ NL

Tél.*: _____ GSM*: _____

* utile: En cas de perte de votre carte ou de vos clés, nous vous contactons

E-mail**: _____

☐ ** J'accepte de recevoir des courriers électroniques publicitaires de Points Plus Punten S.A. et de ses partenaires commerciaux en rapport avec leurs produits et services, notamment la carte-*plus*.

Nombre de personnes faisant partie de votre ménage: __ Nombre d'enfants de -18 ans: __

Date de naissance des enfants: __/__/__ __/__/__ __/__/__
__/__/__ __/__/__ __/__/__

Indiquez dans quelle langue vous souhaitez recevoir nos informations:

☐ Français ☐ Nederlands ☐ Deutsch*** *** Les documents n'étant pas disponibles en allemand vous seront envoyés en version française.

Date: Signature:

Chaque carte-plus porte un code-barres.

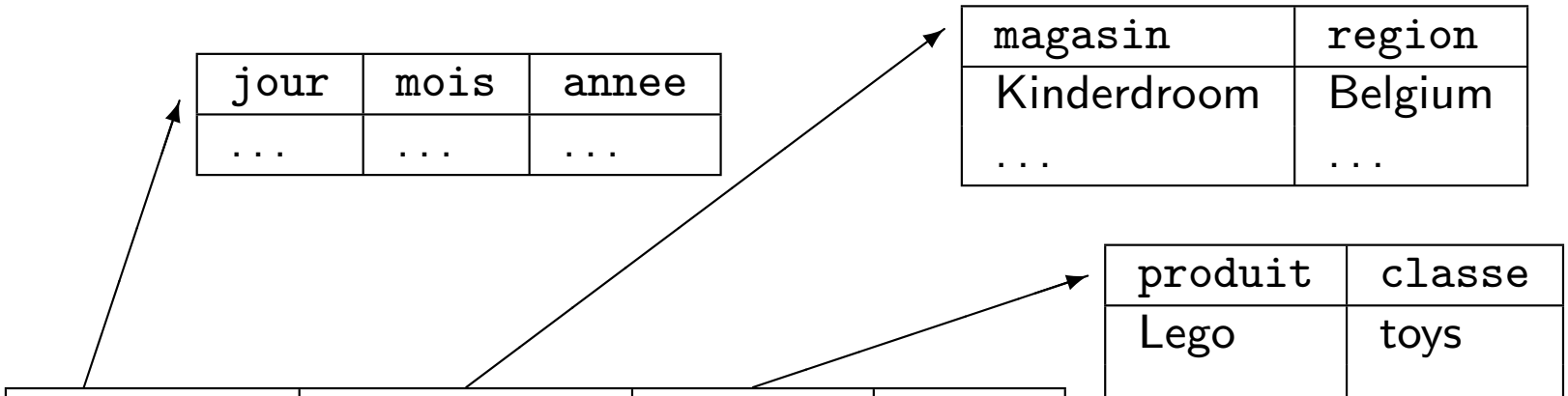
Outline

Outline

Recall from Relational Databases

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

Schéma en étoile (STAR Schema)



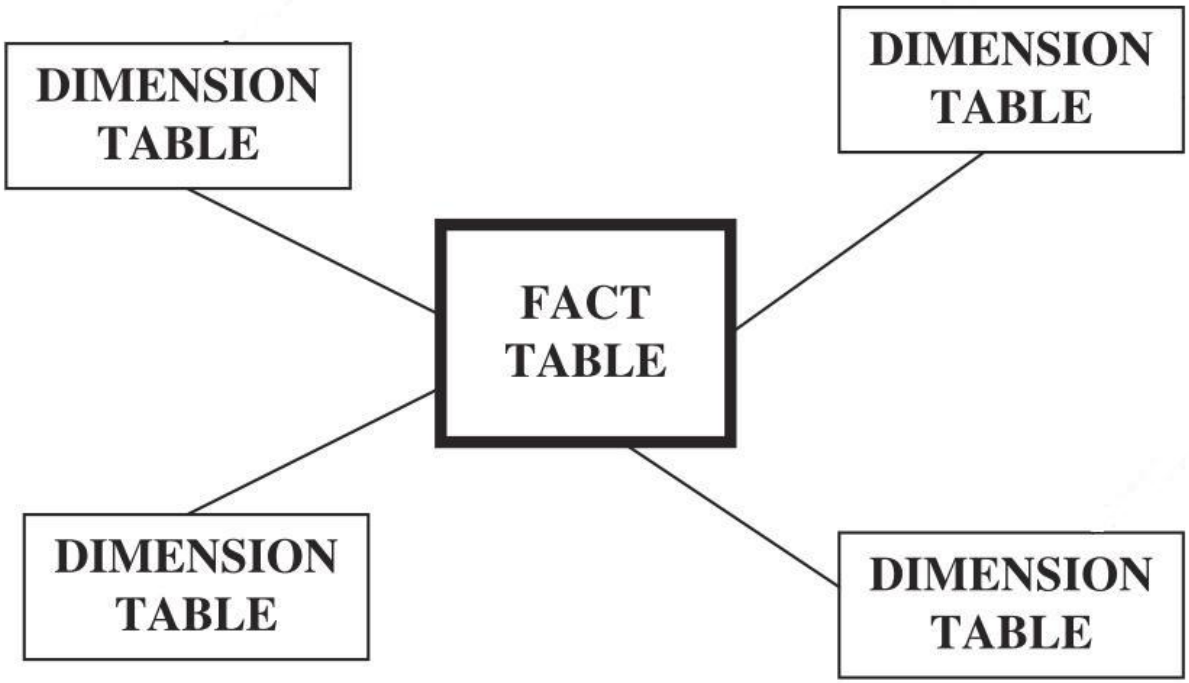
jour	mois	annee
...	...	...

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

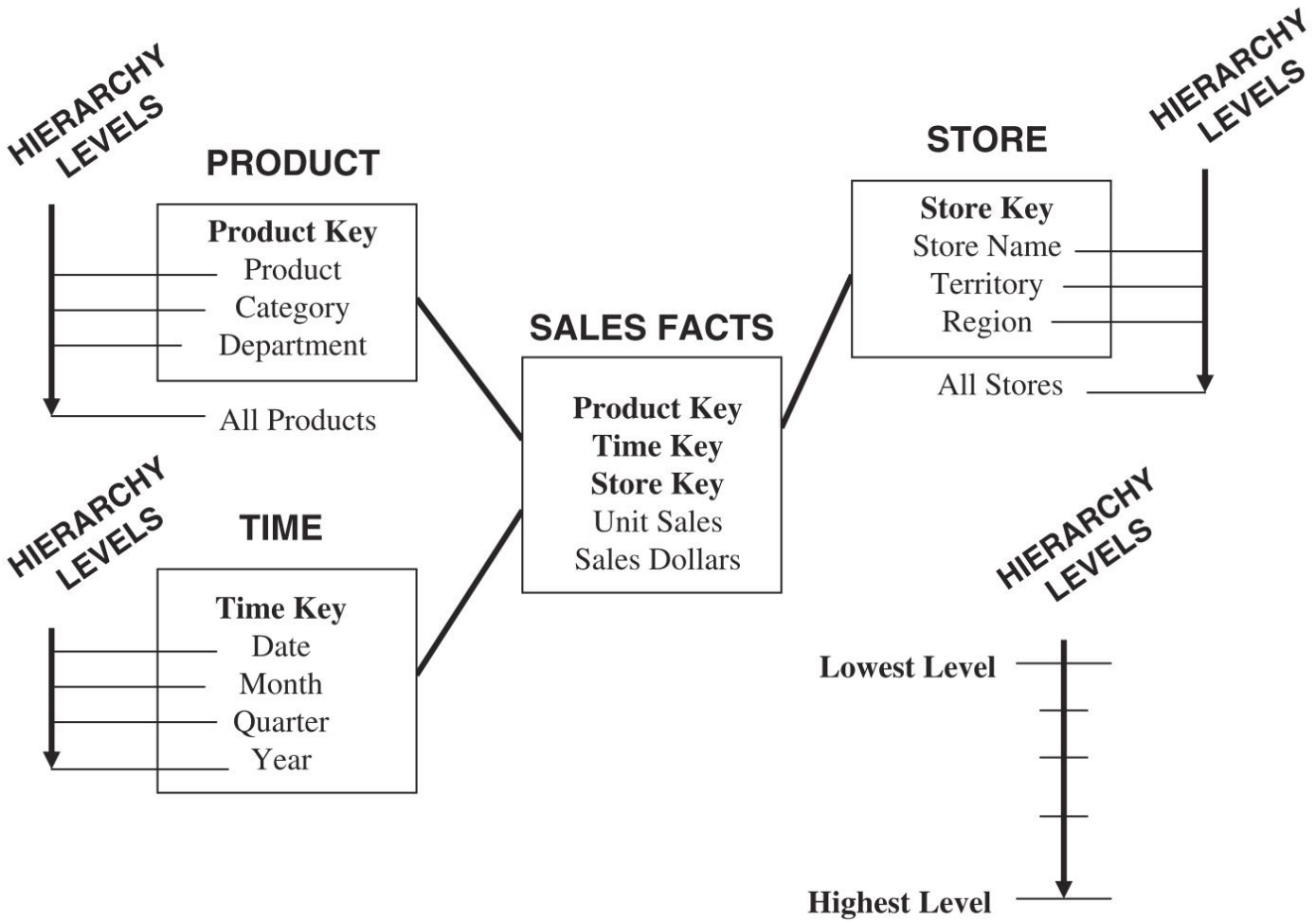


Figure 11-13 Dimension hierarchies.

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

SNOWFLAKE Schema

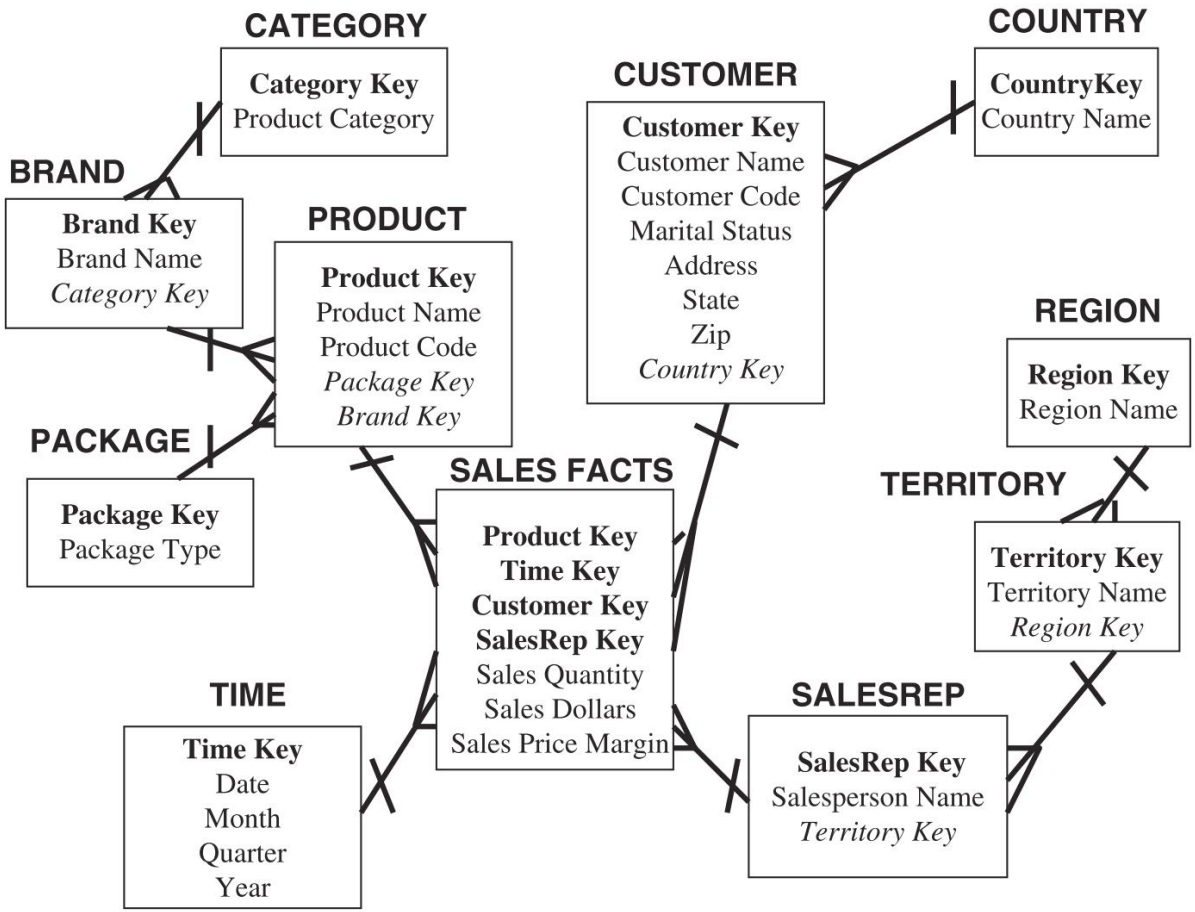
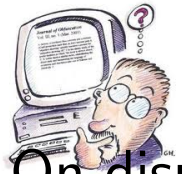


Figure 11-9 Sales: the "snowflake" schema.

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



Travail Pratique

On dispose d'un outil OLAP pour analyser les salaires selon l'âge, le niveau d'étude et la situation géographique des personnes.

- L'analyse selon l'âge peut se faire par année ou par décade (tranches de 10 années à partir de 14 ans et jusqu'à 73 ans).
- L'analyse du niveau d'étude peut se faire par le niveau d'enseignement atteint en fin d'études (primaire, secondaire, supérieur) ou par le dernier diplôme obtenu (Certificat d'études de base (CEB), CESS général, CESS technique, CESS artistique, CESS professionnel, Bac, Licence, Master, Doctorat...).
- L'analyse de la situation géographique peut se faire par ville, province ou communauté.

1. Développez un schéma relationnel en étoile pour cette analyse.
2. Écrivez une requête SQL qui rend le salaire moyen par niveau d'enseignement et province. Par exemple,

Niveau	Province	Salaire
primaire	Hainaut	3000
supérieur	Hainaut	5000
⋮	⋮	⋮

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.

Source Identification

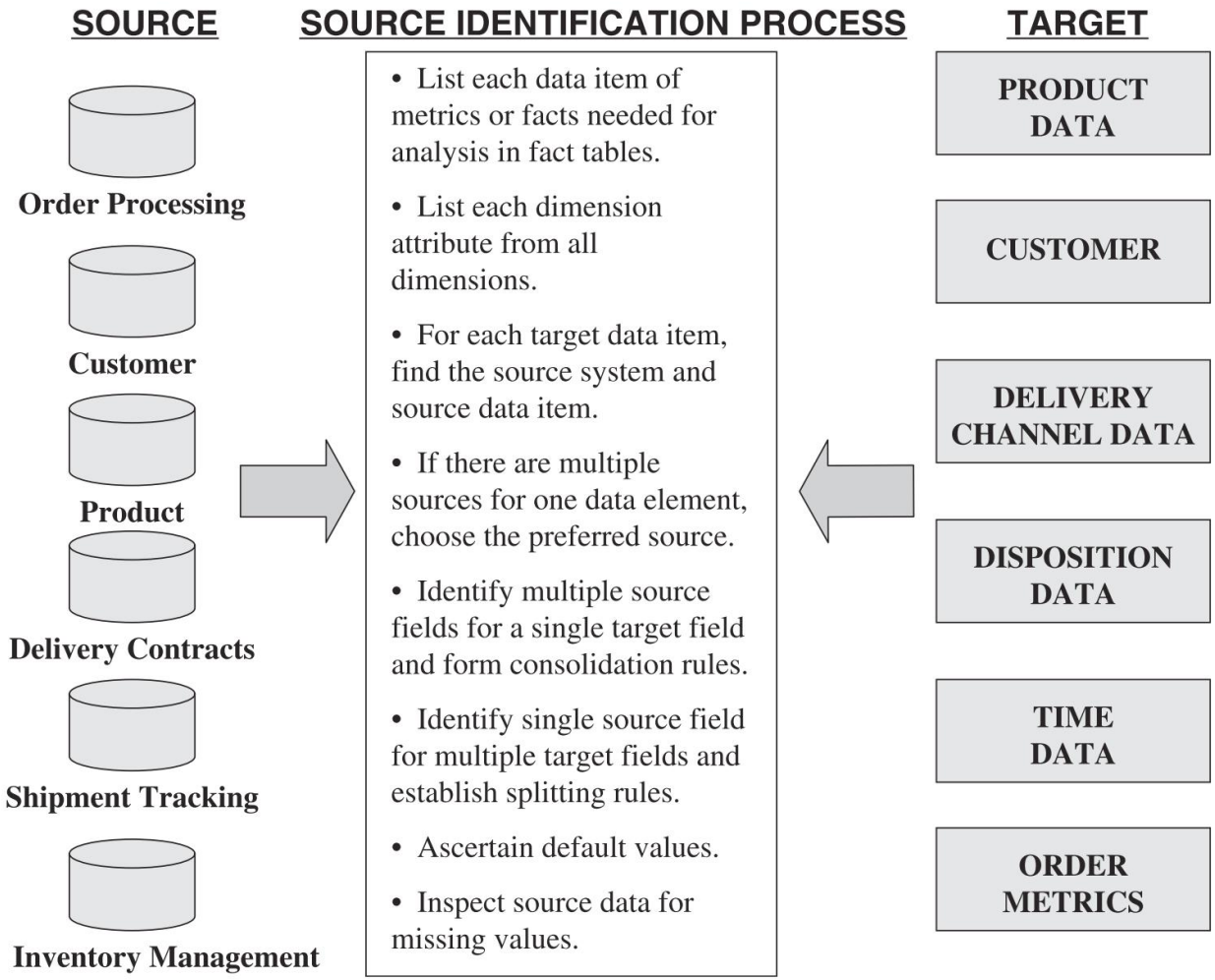


Figure 12-2 Source identification: a stepwise approach.

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

Immediate Data Extraction

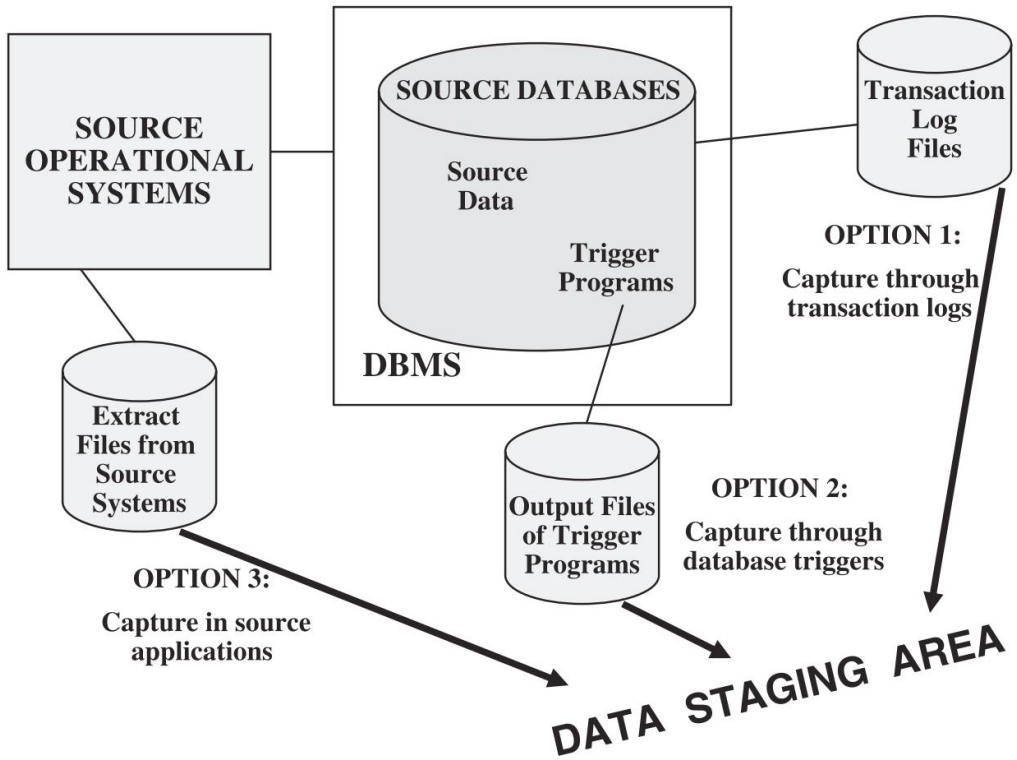


Figure 12-4 Options for immediate data extraction.

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

Deferred Data Extraction

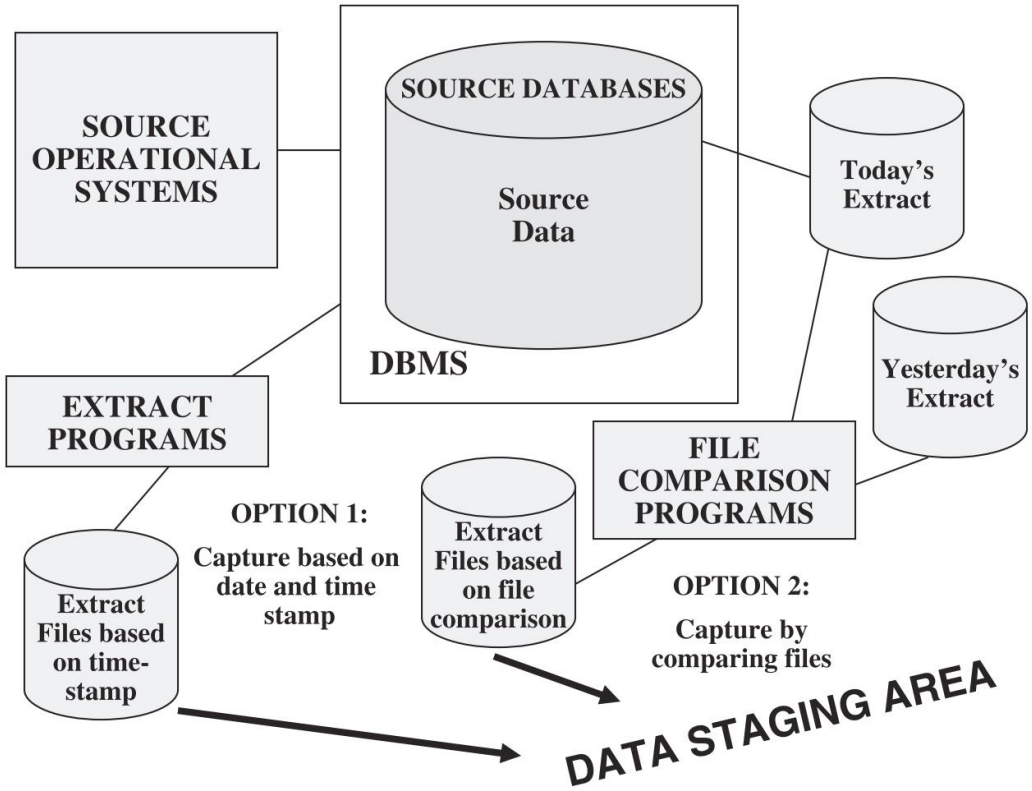


Figure 12-6 Options for deferred data extraction.

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

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.

Caveat

Summarized data can affect information so much that it becomes misleading.

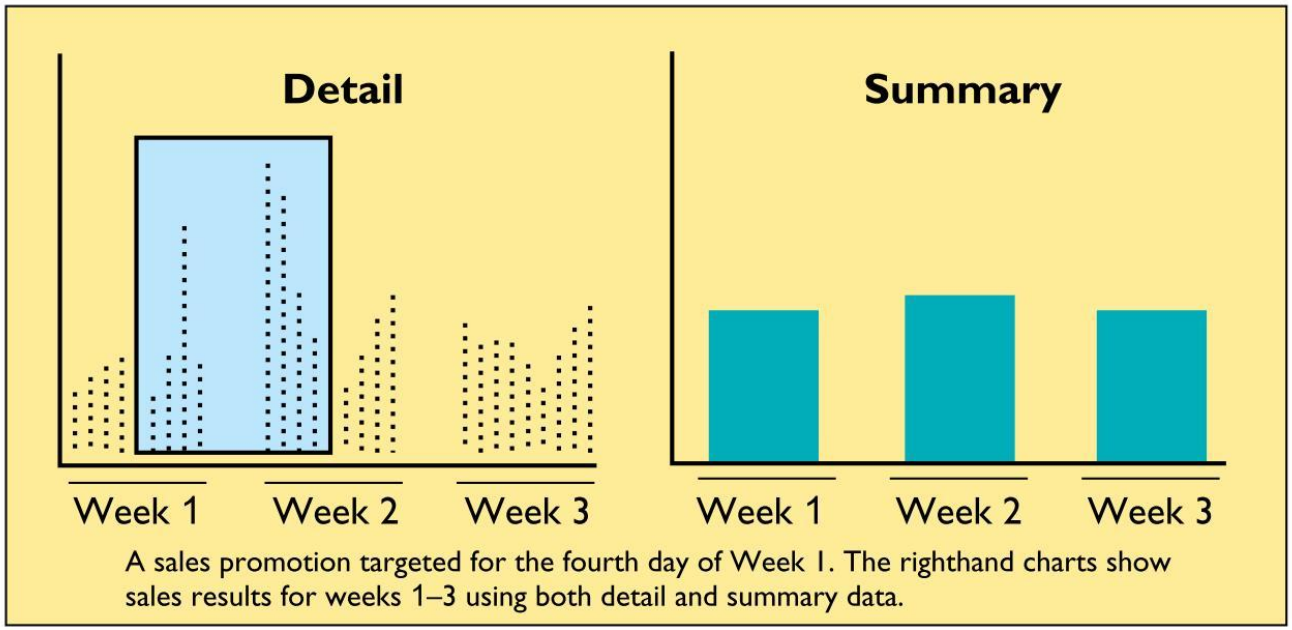
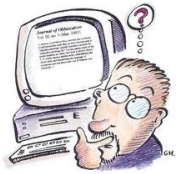


Figure 2. Why do I need detail data?

Source: Stephen R. Gardner: *Building the Data Warehouse*. Communications of the ACM 41(9): 52-60, 1998



Travail Pratique

Voir `pre.pdf`.

Data Loading (ETL)

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

ETL Metadata

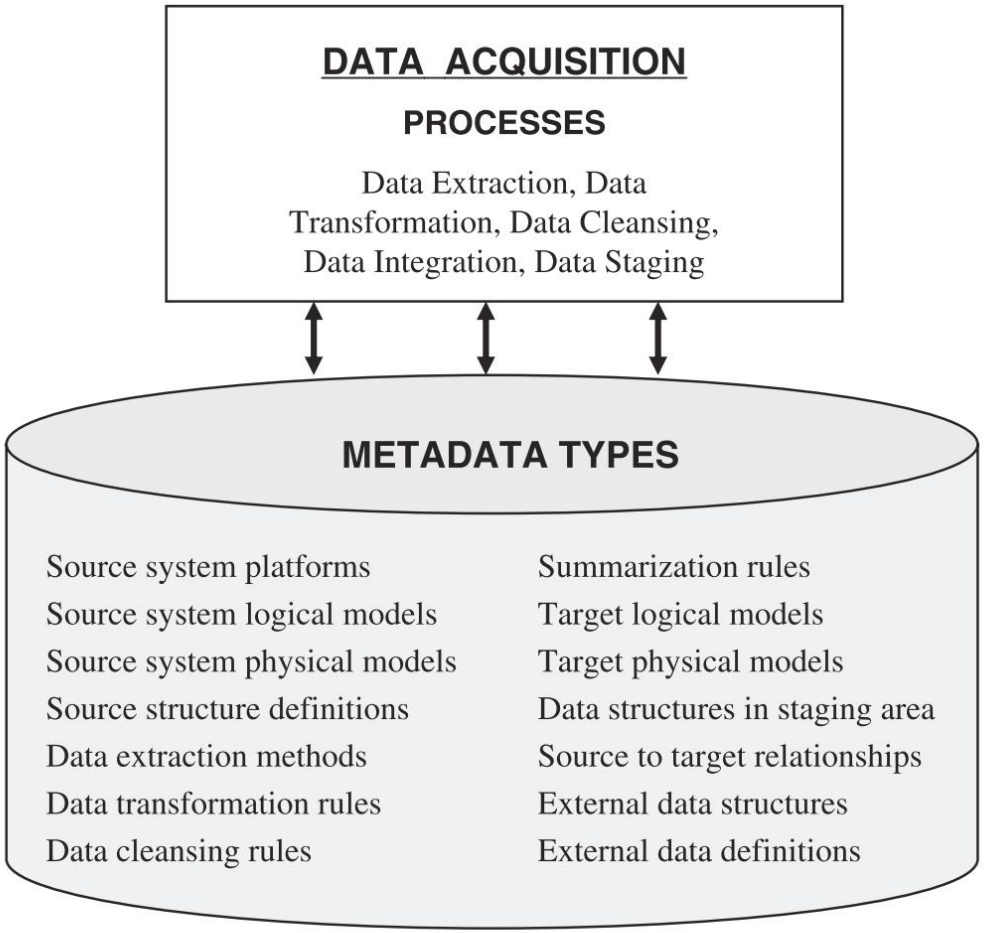


Figure 9-8 Data acquisition: metadata types.

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

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 (1)

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

Problèmes Inter-Sources (2)

Intégration et normalisation (dans une nouvelle table) des données sur les implantations scolaires venant de trois sources :

- Grilles horaires : utilise un compteur "auto-incrément" pour identifier les implantations
- Infrastructures : renseigne l'identification et le niveau d'enseignement de l'implantation
- Comptage : répète l'identification et le nom de l'implantation pour chaque élève inscrit

Problèmes liés aux fichiers reçus

Différents formats :

- Fichier Access
- Fichier Excel, CSV
- Dump
- ...

Différents encodages :

- UTF-8
- ISO-8859-1
- ...

Estimation du temps pour l'ETL

ETL = Extract-Transform-Load

Pour ce projet :

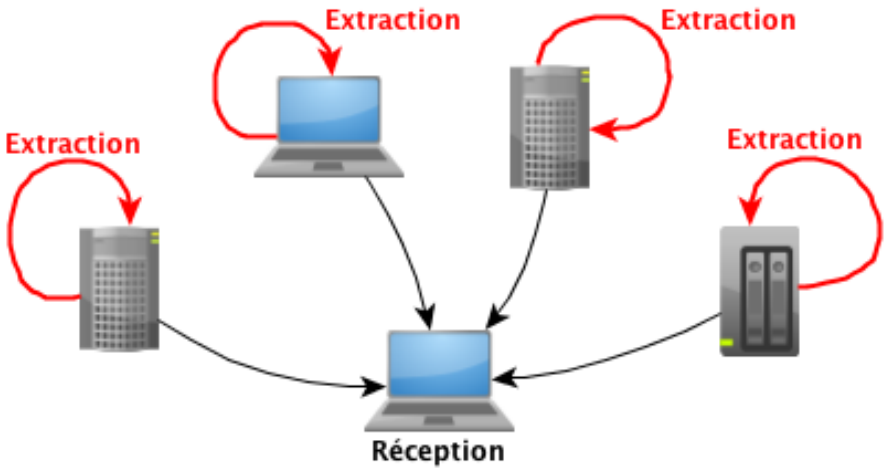
- Extraction : 10%
- Transformation : 85%
- Chargement: 5%

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

Extraction

L'extraction consiste en la collecte des données identifiées et sélectionnées. Pour cela, il faut accéder aux systèmes de stockage correspondants.

Dans le cas de ce projet, les systèmes de stockage ne sont pas directement accessibles. Les demandes sont donc effectuées auprès de personnes de contacts (administrateurs BDD, responsables de services, etc.). Les données sont donc directement extraites par les services en question avant de nous être envoyées. Il faut cependant constamment leur rappeler.



Transformation

Les données ne sont pas utilisables directement. Les données doivent être vérifiées, reformatées, standardisées, modifiées ou encore nettoyées afin de supprimer les doublons ou les valeurs non conformes, assurer les relations entre tables (de mêmes ou de différentes sources) et rendre les données conformes au modèle de destination.

Dans le cas de ce projet, les principales phases de transformation sont :

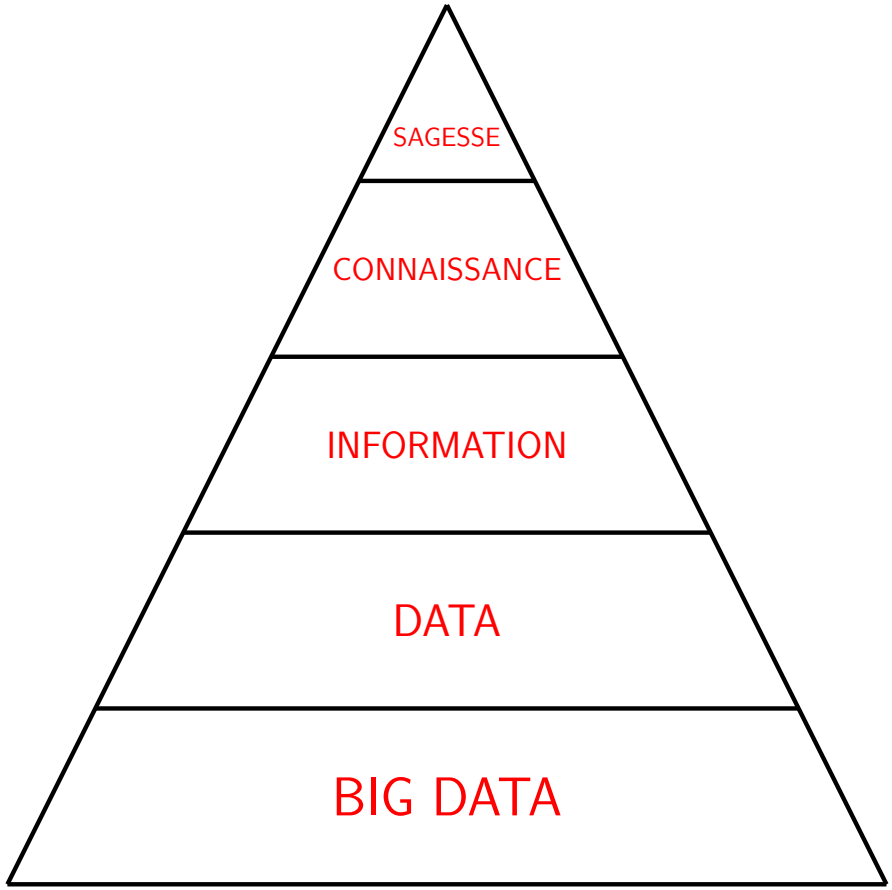
- La préparation des données (10%) : formater les données dans un fichier de type souhaité et avec l'encodage souhaité.
- La vérification des données (40%) : déterminer s'il y a des doublons, des valeurs manquantes, des problèmes de relation, etc.
- La modification des données (50%) : modifier les problèmes rencontrés dans le point précédent.

Chargement

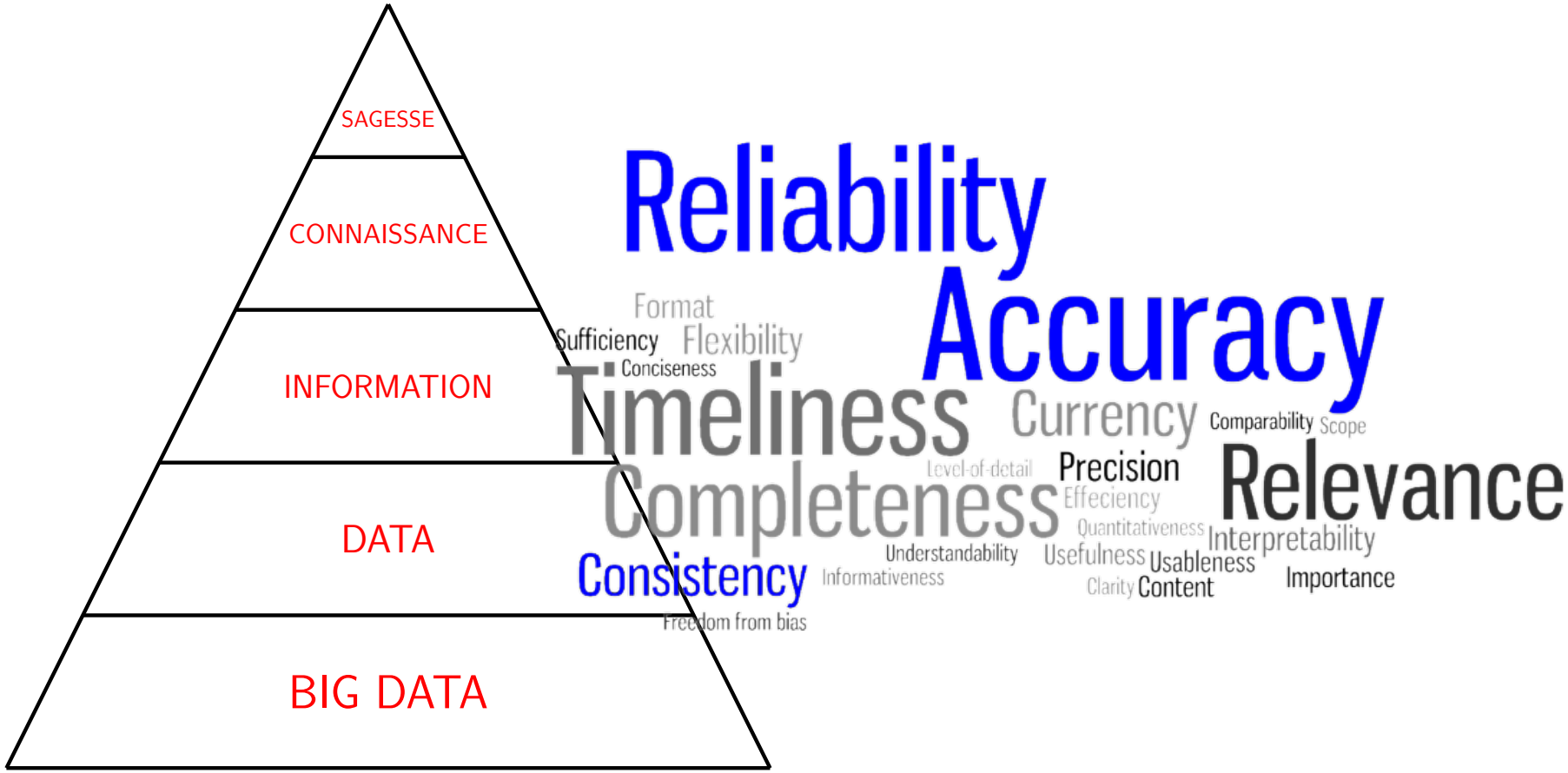
Le chargement consiste à insérer les données dans le Data Warehouse. Si le travail précédent a été effectué correctement, le chargement ne demande pas beaucoup de travail : une simple requête peut suffire.

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

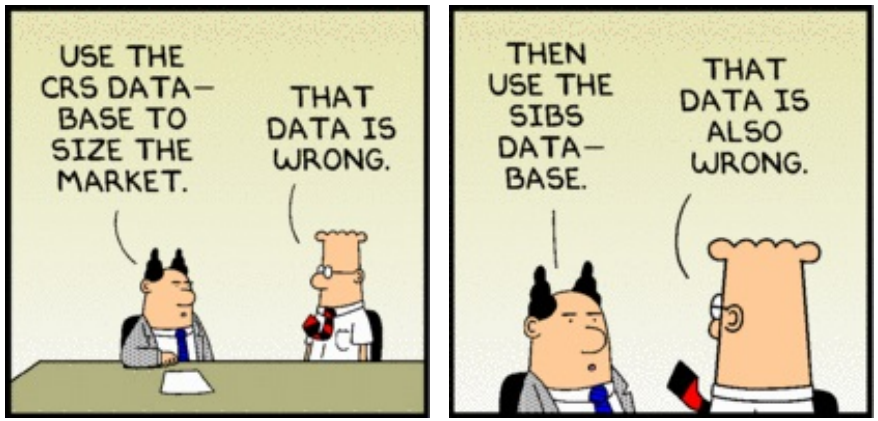
Souvent les données sont incohérentes, incomplètes, manquantes. . .

Que peut-on faire avec ces données?

La réalité

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

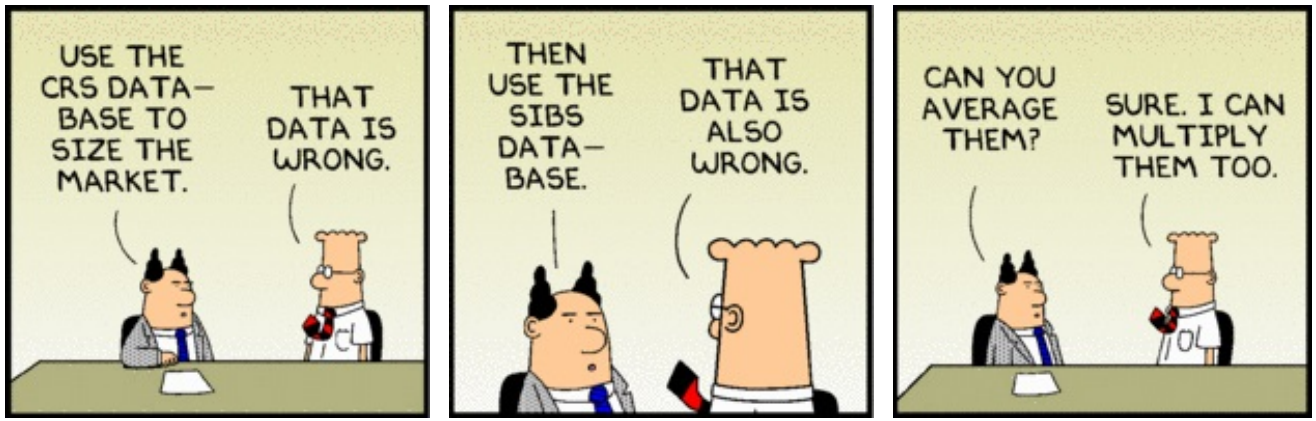
Que peut-on faire avec ces données?



La réalité

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

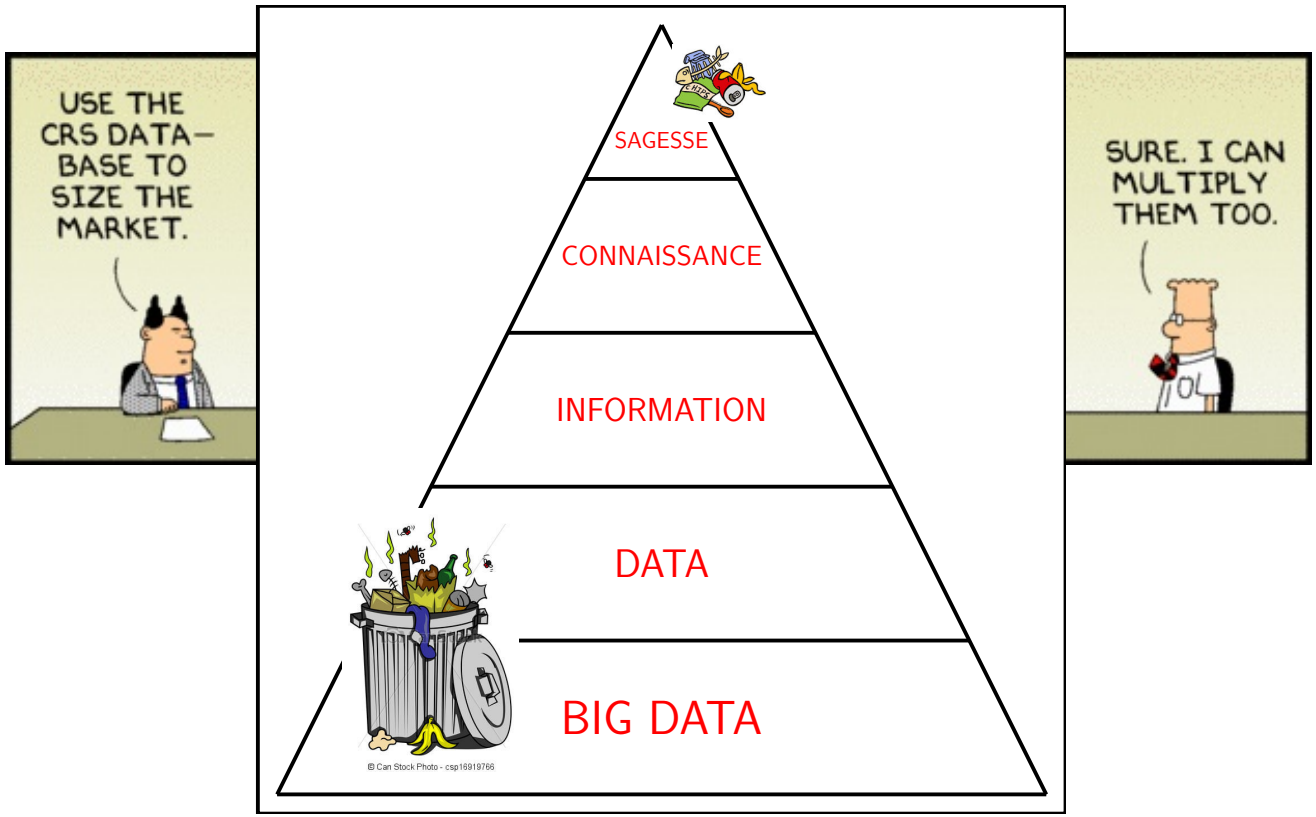
Que peut-on faire avec ces données?



La réalité

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

Que peut-on faire avec ces données?



Données imparfaites

Example

P

PID	Prénom	Nom	GroupeSanguin	Genre	···
1	John	Adams	NULL	♂	···
2	Jan	Peeters	A+	M	···
3	Jean	Lemaître	A+	M	···
3	Jean	Lemaitre	AB+	M	···

Données imparfaites

Example

P	PID	Prénom	Nom	GroupeSanguin	Genre	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Problèmes d'encodage

Données imparfaites

Example

P

PID	Prénom	Nom	GroupeSanguin	Genre	···
1	John	Adams	NULL	♂	···
2	Jan	Peeters	A+	M	···
3	Jean	Lemaître	A+	M	···
3	Jean	Lemaitre	AB+	M	···

Doublons

Données imparfaites

Example

P	PID	Prénom	Nom	GroupeSanguin	Genre	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Valeurs manquantes

Données imparfaites

Example

P	PID	Prénom	Nom	GroupeSanguin	Genre	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Valeurs impossibles

Gestion de données imparfaites

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

Que peut-on faire avec ces données?

Gestion de données imparfaites

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

Que peut-on faire avec ces données?

Gestion de données imparfaites

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

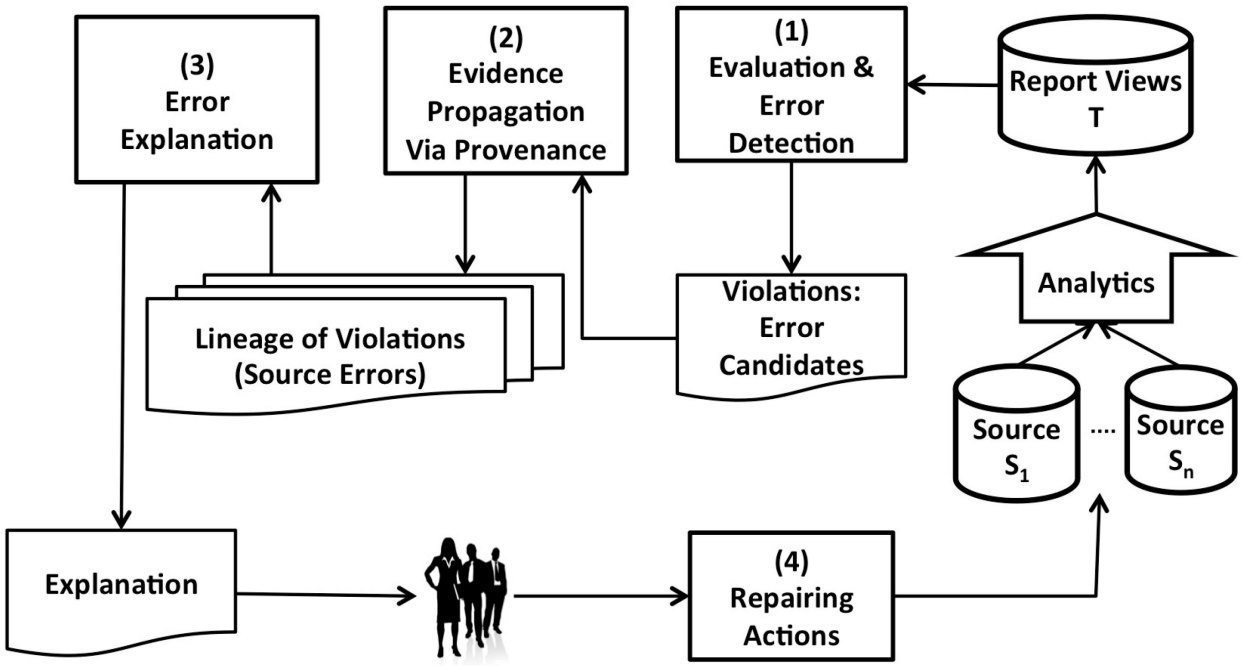
Que peut-on faire avec ces données?

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*

Embrace Imperfection

Example

P

PID	Prénom	Nom	GroupeSanguin	Genre	···
1	John	Adams	NULL	♂	···
2	Jan	Peeters	A+	M	···
3	Jean	Lemaître	A+	M	···
3	Jean	Lemaitre	AB+	M	···

Embrace Imperfection

Example

P

PID	Prénom	Nom	GroupeSanguin	Genre	···
1	John	Adams	NULL	♂	···
2	Jan	Peeters	A+	M	···
3	Jean	Lemaître	A+	M	···
3	Jean	Lemaitre	AB+	M	···

Embrace Imperfection

Example

P

PID	Prénom	Nom	GroupeSanguin	Genre	···
1	John	Adams	NULL	M	···
2	Jan	Peeters	A+	M	···
3	Jean	Lemaître	A+	M	···
3	Jean	Lemaître	AB+	M	···

Embrace Imperfection

Example

P

PID	Prénom	Nom	GroupeSanguin	Genre	···
1	John	Adams	NULL	M	···
2	Jan	Peeters	A+	M	···
3	Jean	Lemaître	A+	M	···
3	Jean	Lemaître	AB+	M	···

Embrace Imperfection

Example

P

PID	Prénom	Nom	GroupeSanguin	Genre	···
1	John	Adams	NULL	M	···
2	Jan	Peeters	A+	M	···
3	Jean	Lemaître	A+	M	···
3	Jean	Lemaître	AB+	M	···

Embrace Imperfection

Example

P	PID	Prénom	Nom	GroupeSanguin	Genre	···
	1	John	Adams	NULL	M	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaître	AB+	M	···

Attention

Si les données sont imparfaites et non nettoyables, nous devons repenser la façon de répondre aux requêtes.
Par exemple, comment répondre aux questions suivantes?

1. *Combien de patients ont un groupe sanguin de A+?*
2. *Combien de patients ont un groupe sanguin autre que A+?*

Qualité des données au sein d'un service public

H. Van Puyvelde. *De l'information opérationnelle à l'intelligence décisionnelle par le data mining–Etude de faisabilité appliquée au cas d'un service public.* Travail de fin d'étude, Université de Mons-Hainaut.

- La compagnie utilise une dizaine d'applications différentes sur quatre plates-formes SGBD différents.
- But initial : Appliquer le data mining pour répondre aux questions du type
  - *"Qui sont nos clients ?"*
  - *"Quels services sont les plus bénéfiques pour nos clients ?"*

Cependant, une préparation considérable des données était nécessaire...

Problèmes de qualité

Quelques problèmes difficiles :

- Doublons. Par ex.
  $\langle$*RAYTEC, Rue de Commerce 2,. . .*$\rangle$ et
  $\langle$*S.A. RAYTEC, 2 Rue de Commerce,. . .*$\rangle$.
- Usage fréquent du code fourre-tout *"autres"* pour des attributs tels que `formation` ou `profession`.
- Valeurs manquantes, impossibles ou périmées.

Ceci confirme l'expérience d'autres auteurs :

> *Preparation of the data [. . .] can easily take up to 80% of the time needed for the whole KDD [Knowledge Discovery in Databases]; this is not surprising, since the difficulties in data integration are well known.* [Mannila 96]

Data Quality Includes (but is not Limited to) Data Integrity

DATA INTEGRITY	DATA QUALITY
Specific instance of an entity accurately represents that occurrence of the entity.	The data item is exactly fit for the purpose for which the business users have defined it.
Data element defined in terms of database technology.	Wider concept grounded in the specific business of the company.
Data element conforms to validation constraints.	Relates not just to single data elements but to the system as a whole.
Individual data items have the correct data types.	Form and content of data elements consistent across the whole system.
Traditionally relates to operational systems.	Essentially needed in a corporate-wide data warehouse for business users.

Figure 13-1 Data accuracy versus data quality.

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

Data Warehouse Challenges

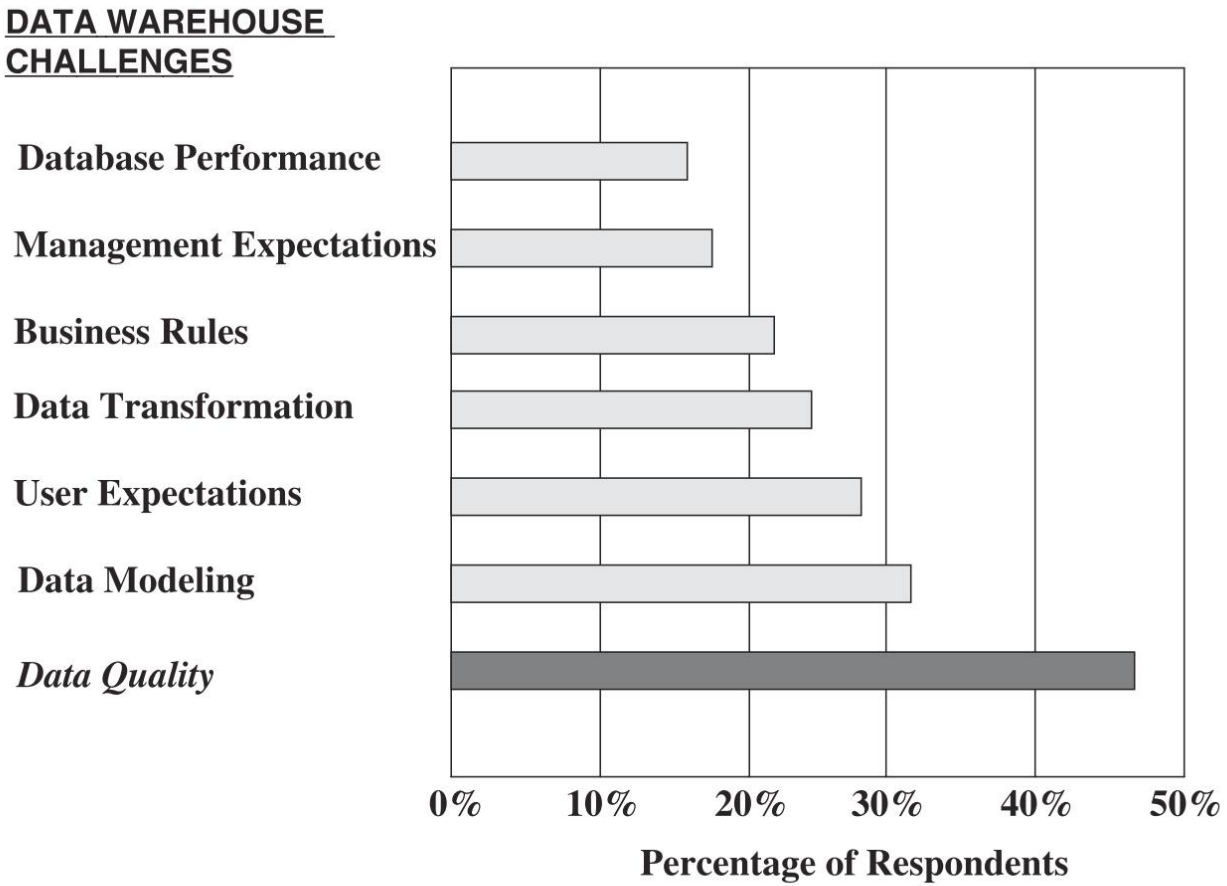


Figure 13-2 Data quality: the top challenge.

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

Data Quality Dimensions I

Adherence to Data Integrity Rules For example, primary and foreign keys must be satisfied.

Domain Integrity The data value of an attribute falls in the range of allowable values.

Data Type The data value for an attribute is of the right type.

Completeness There are no missing values for a given attribute.

Conformance to Business Rules The values of each data item adhere to prescribed business rules.

Accuracy The value stored is the right value.

Timely For example, if the users expect customer dimension data not to be older than one day, the changes to customer data in the source systems must be applied to the data warehouse daily.

Data Quality Dimensions II

Usefulness If a data element is of no value to the users, then it is totally unnecessary for that data element to be in the data warehouse.

Consistency For example, if the product code for product ABC in one system is 1234, then the code for this product is 1234 in every source system.

Structural Definiteness. For example, values for names of individuals must be stored as first name, middle initial, and last name.

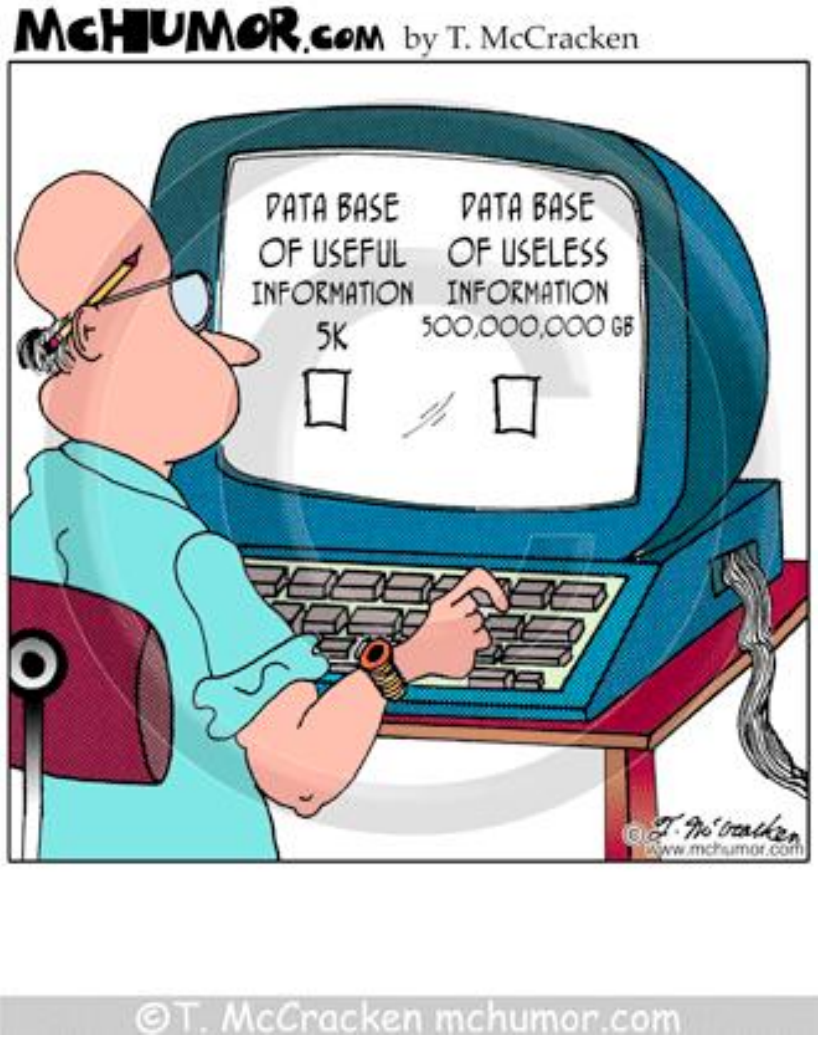
Data Anomaly A field must be used only for the purpose for which it is defined.

Redundancy The same data must not be stored in more than one place in a system.

Duplication Duplication of records in a system is completely resolved.

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

Usefulness



Source: www.enterpriseirregulars.com

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

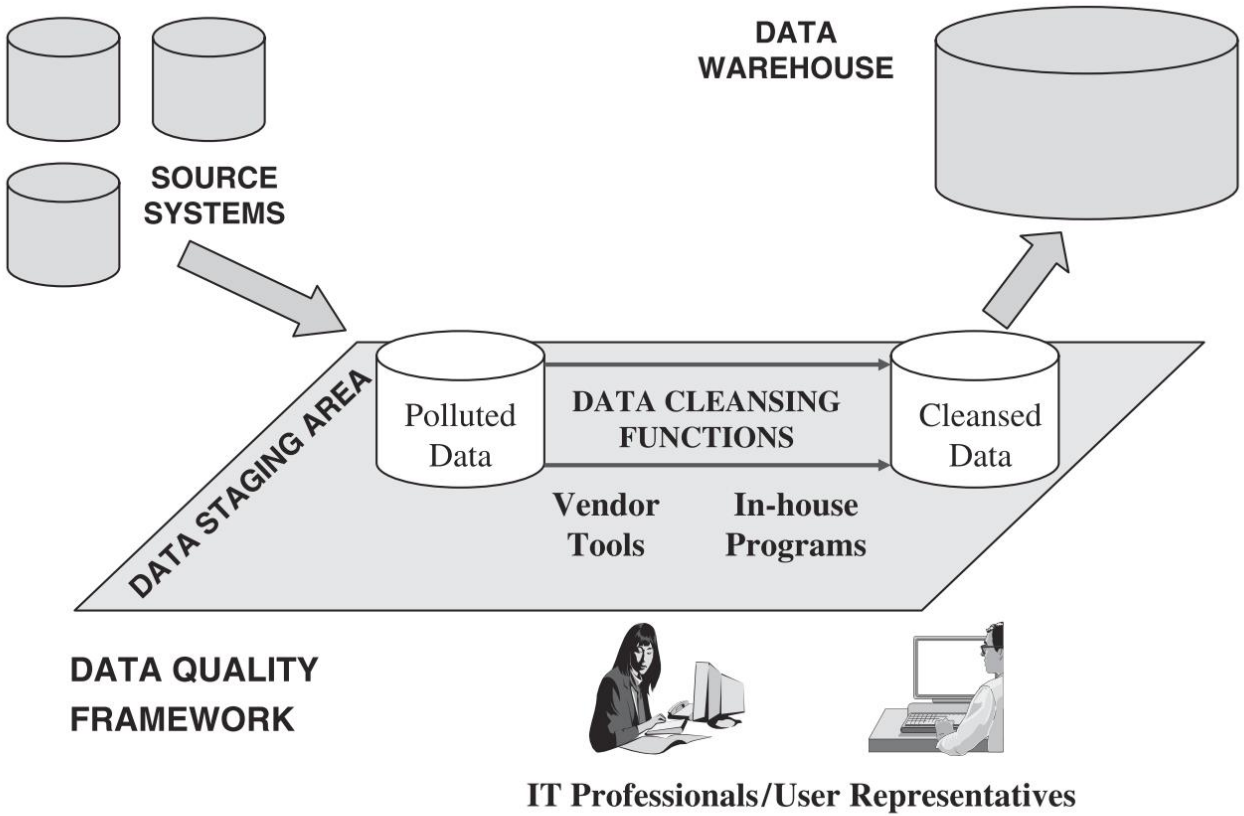
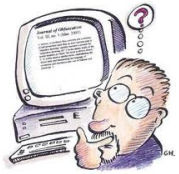


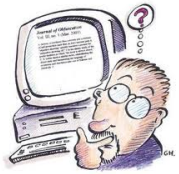
Figure 13-7 Overall data purification.

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



Travail Pratique

Regarder les vidéos de Google Refine.



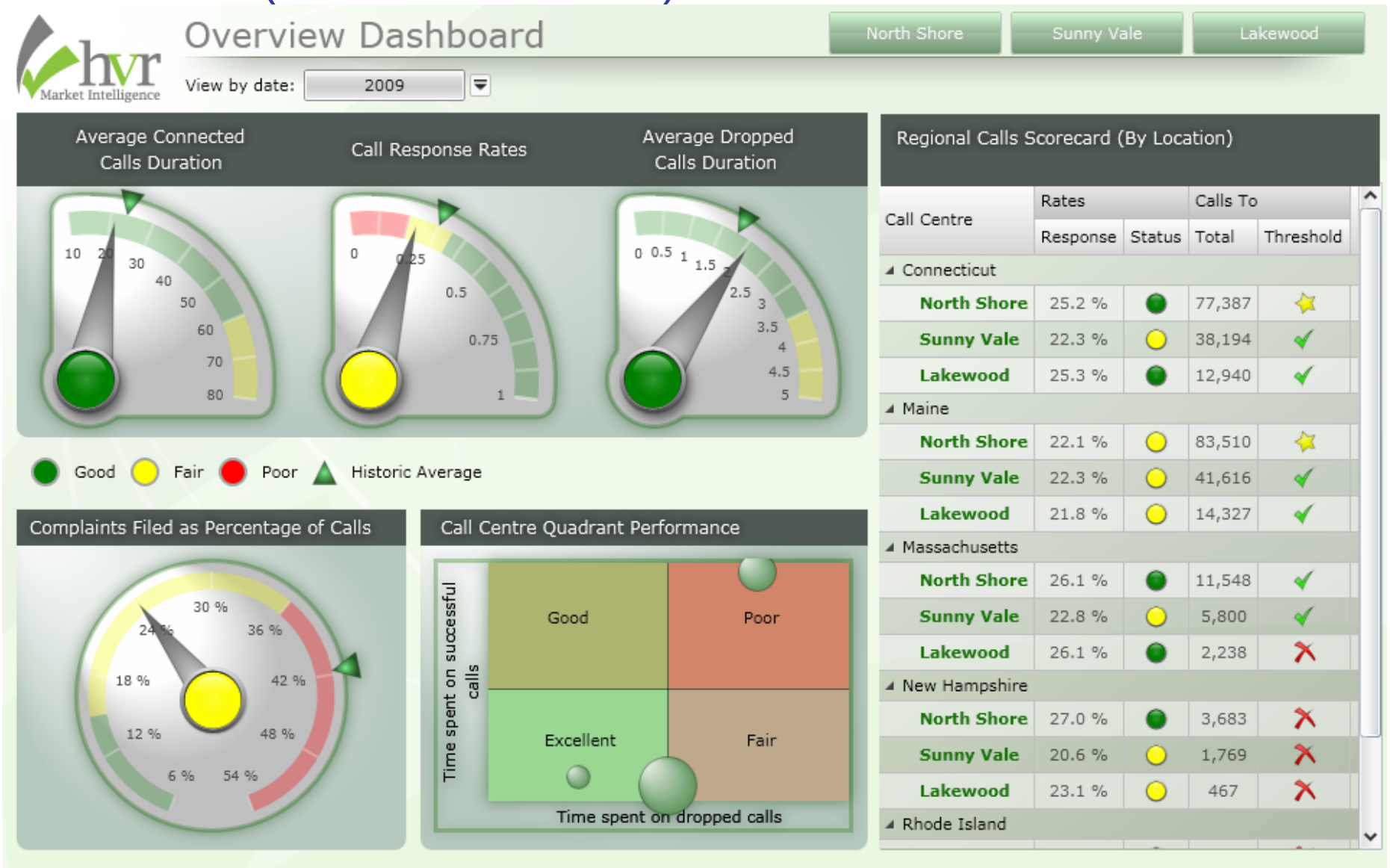
Travail Pratique

- Discuter les risques de pollution des données sur les teneurs d'une carte à points.
- Quel peut être l'impact de cette pollution sur la prise de décision ?
- Comment réduire ces risques de pollution ?
- Comment nettoyer les données "sales" ?

Outline

Outline

Dashboard (tableau de bord)

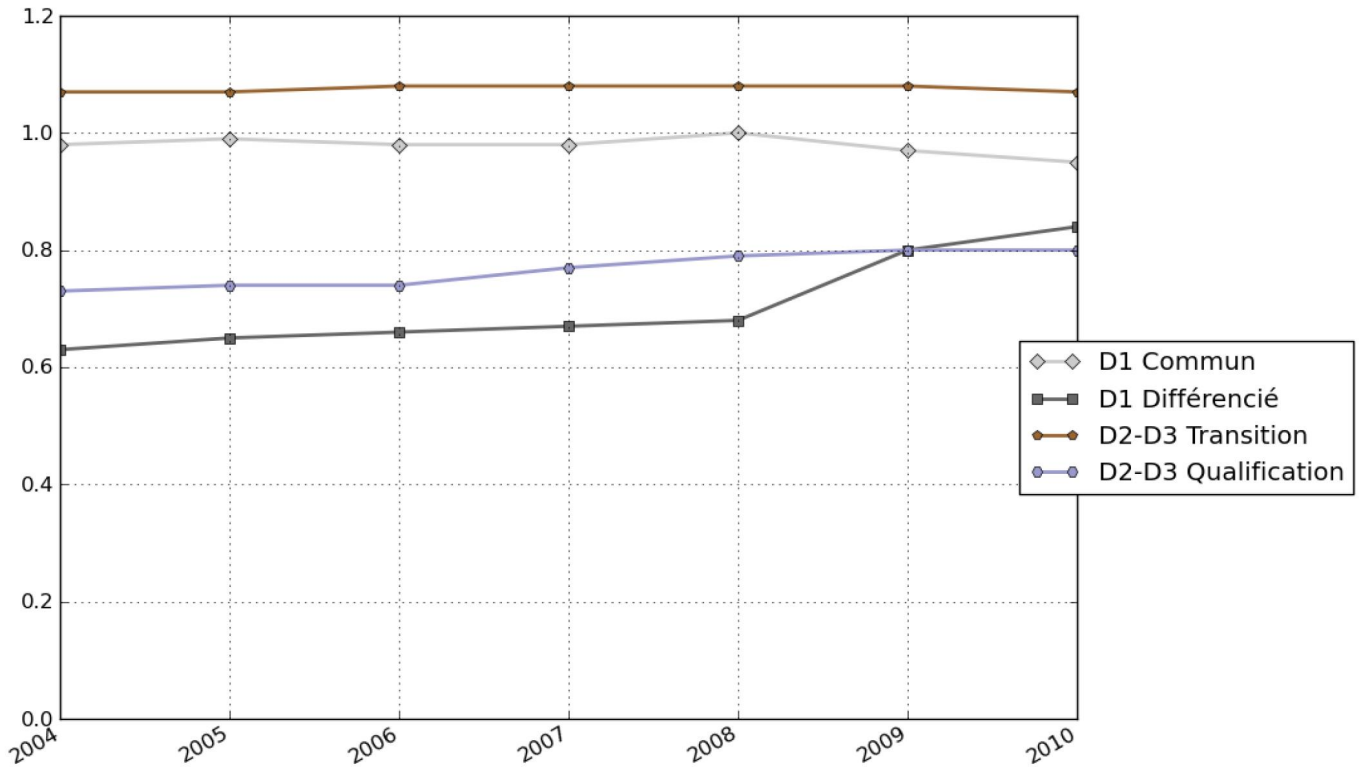


Source: www.dundas.com

Indicateur

Evolution de la proportion de filles, au 1er degré et aux 2e - 3e degrés , dans l'enseignement secondaire ordinaire, selon la filière d'enseignement

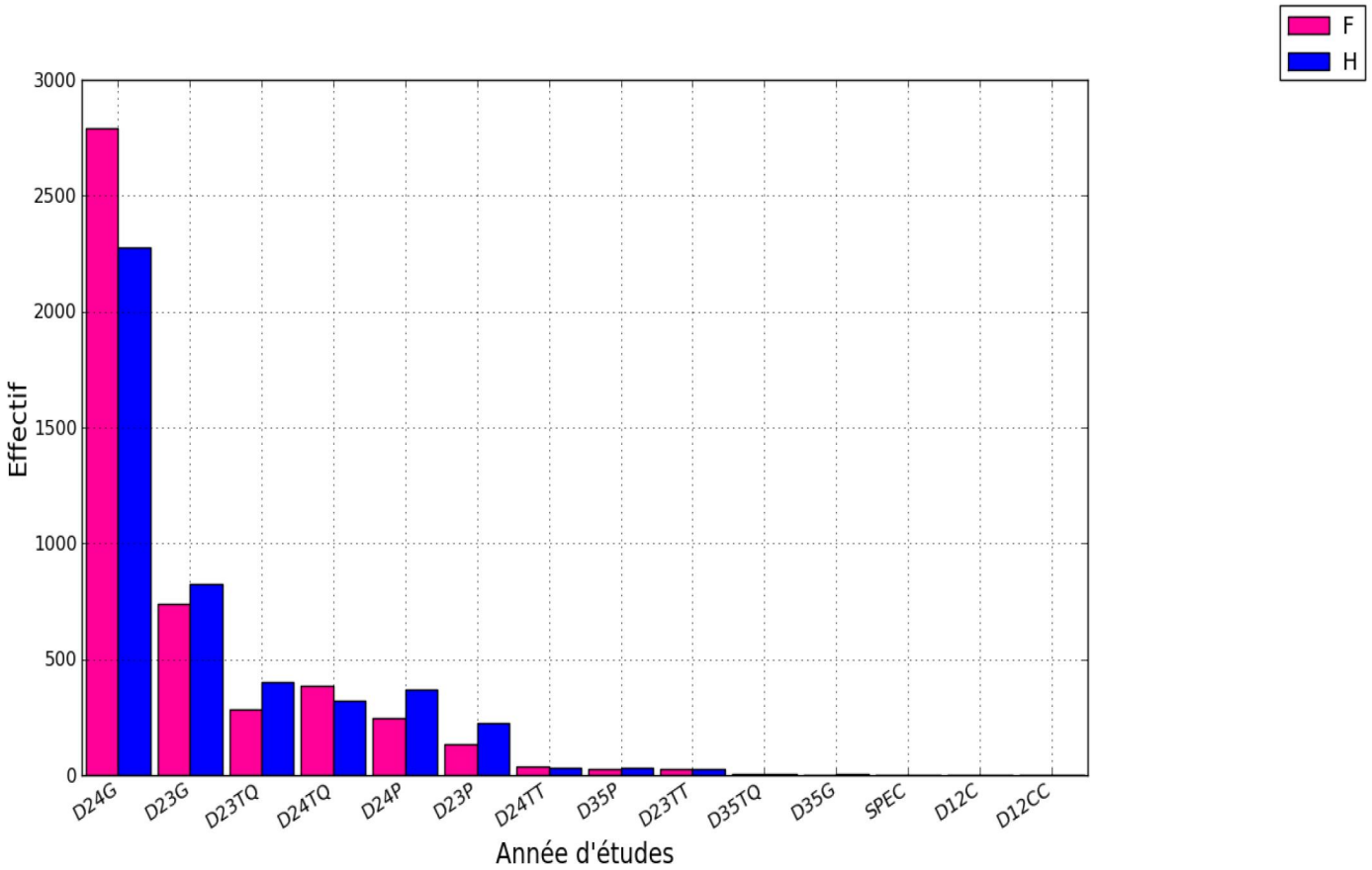
Ensemble de la FWB



Indicateur

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

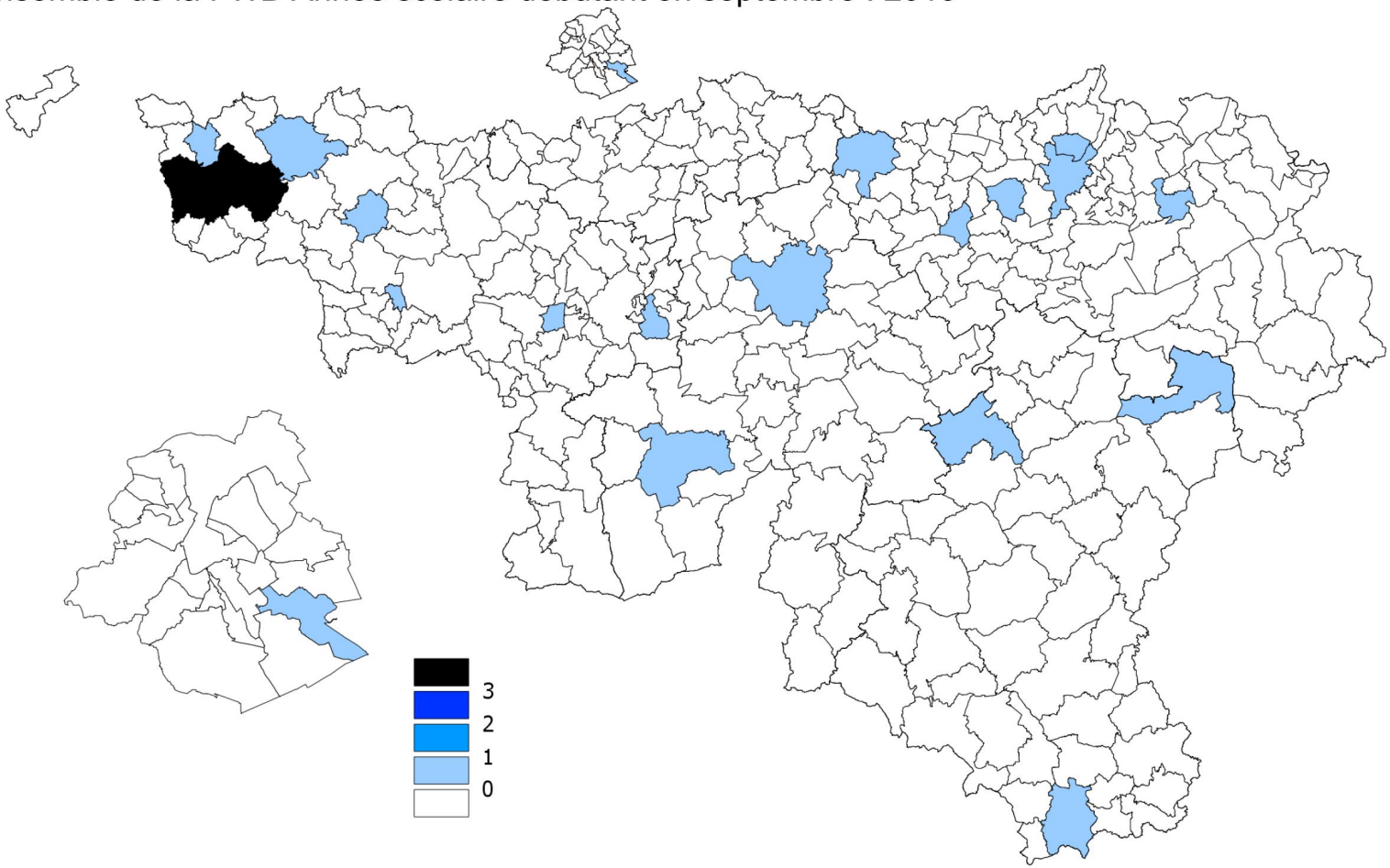
Année d'entrée en 1ere secondaire : 2004



Indicateur

Répartition géographique des établissements d''enseignement spécialisé organisant le niveau secondaire

Ensemble de la FWB Année scolaire débutant en septembre : 2010



Outline

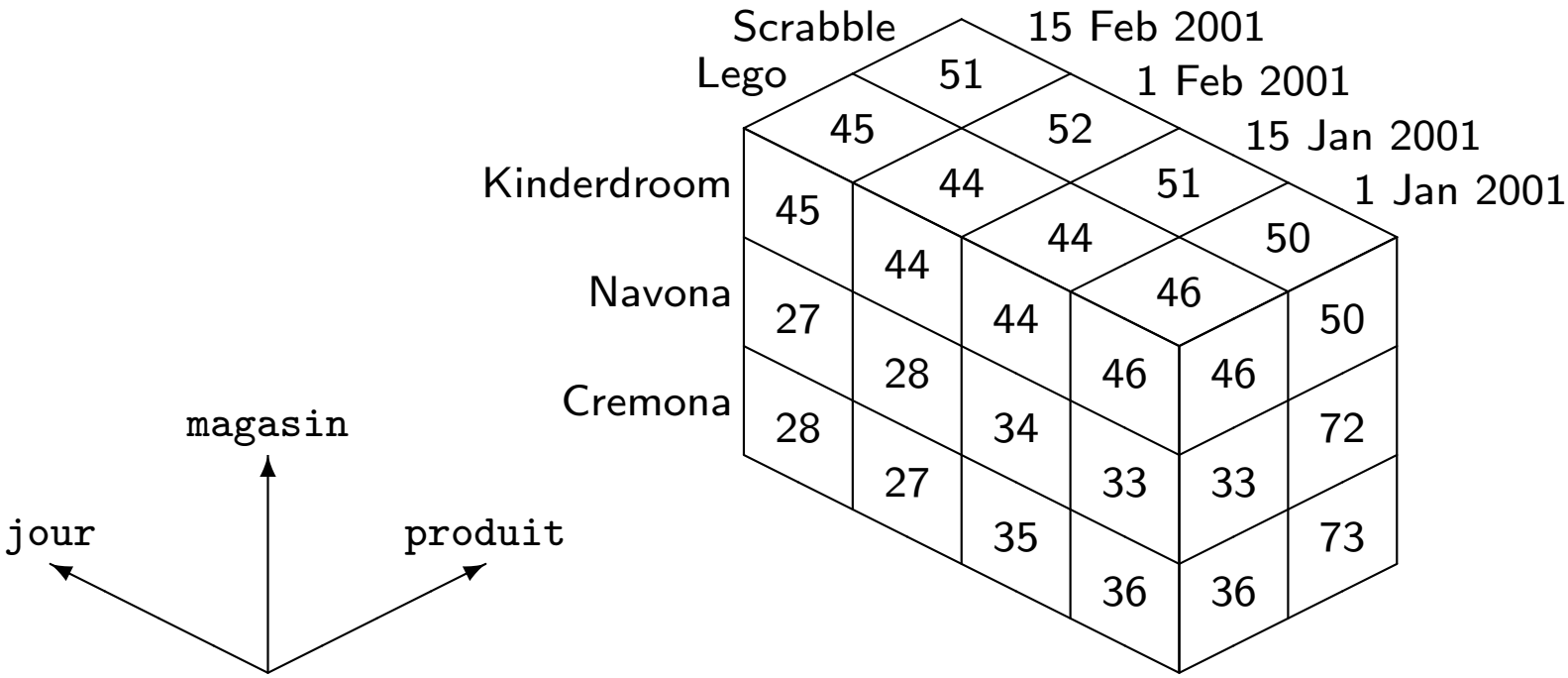
Dimensions et mesures

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.

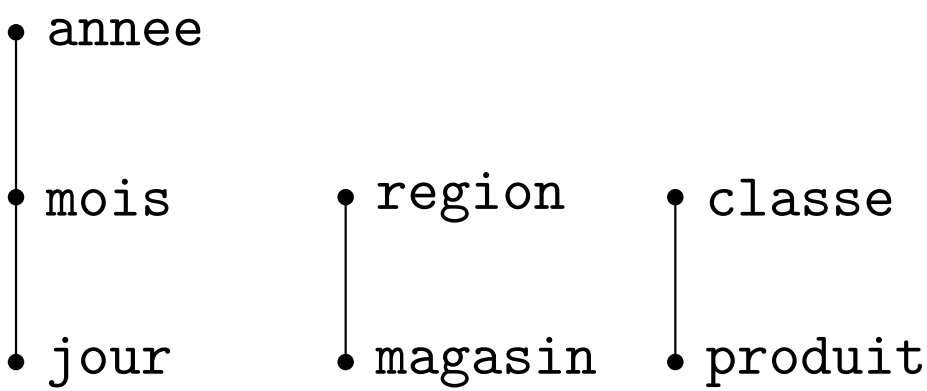
Cube de données



Un cube en 3D.

Hiérarchies de concepts

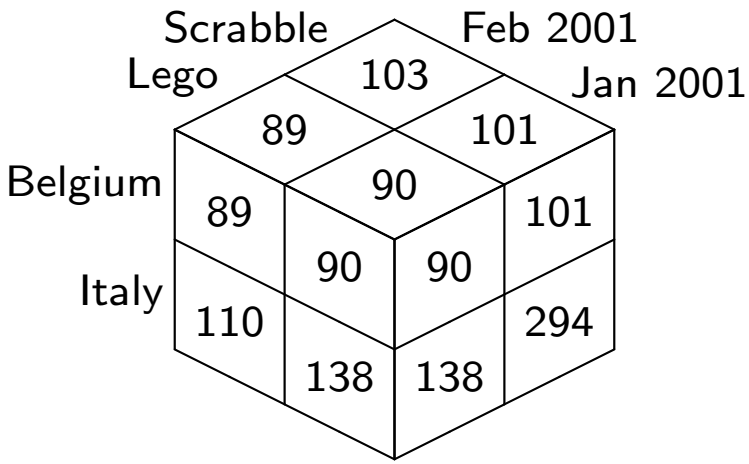
Les dimensions sont organisées en des hiérarchies conceptuelles qui déterminent les façons de regrouper les données.



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".

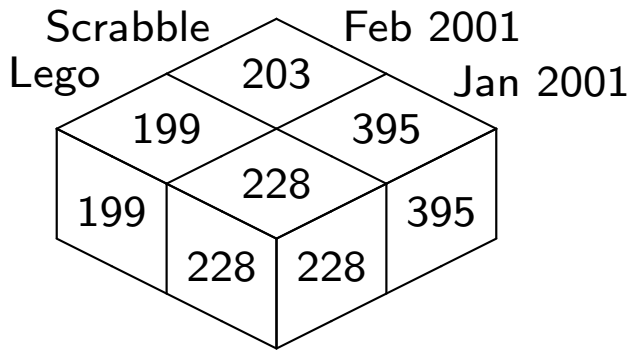


Le cube `mois region produit`.

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.

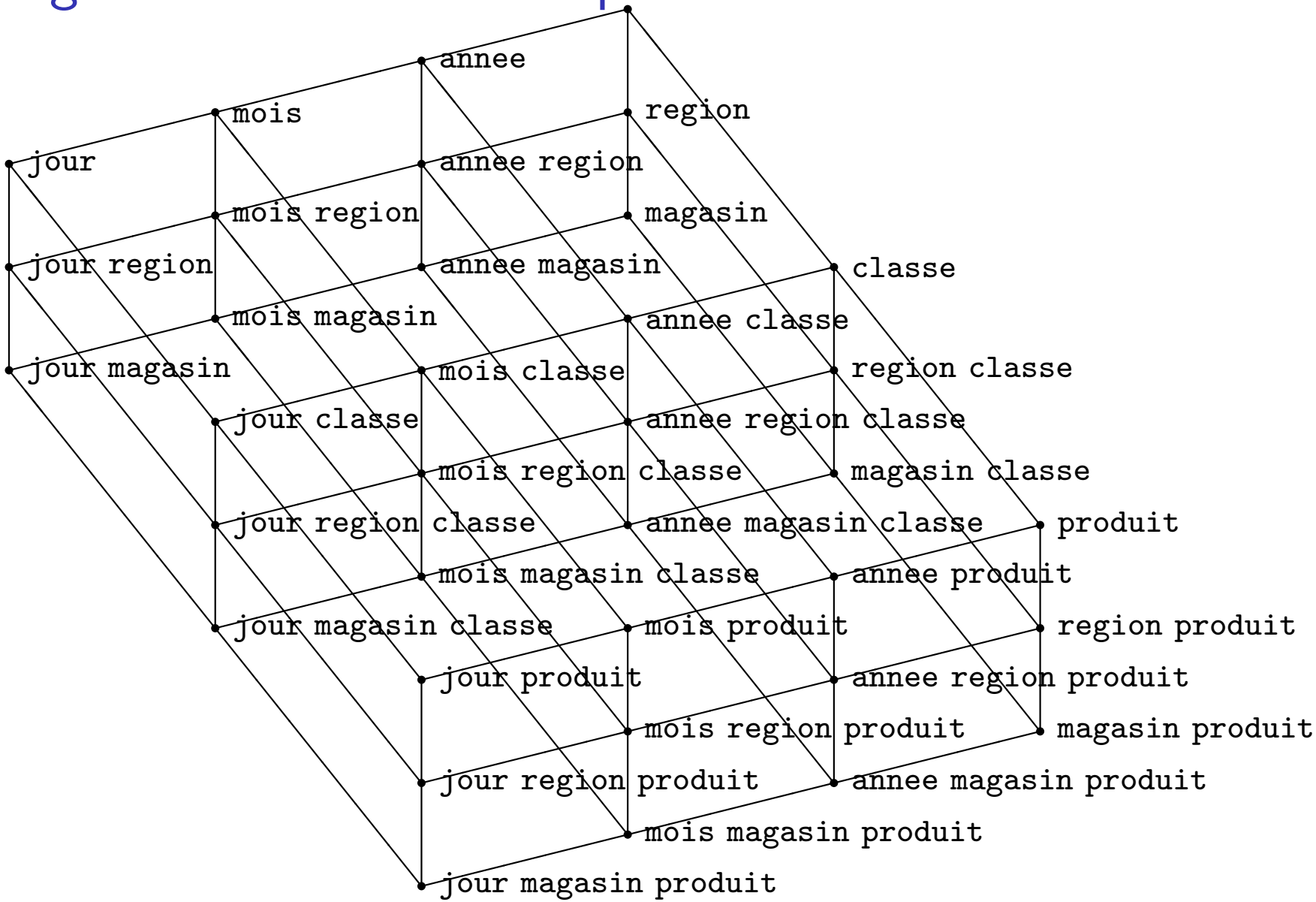
Multidimensional analysis and its drilling capability, on the other hand, are instantaneous because the information you need is staged for you. By clicking the mouse or selecting a command, you see less detail, more detail, or whatever you want. The tool and the database don't have to collaborate for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

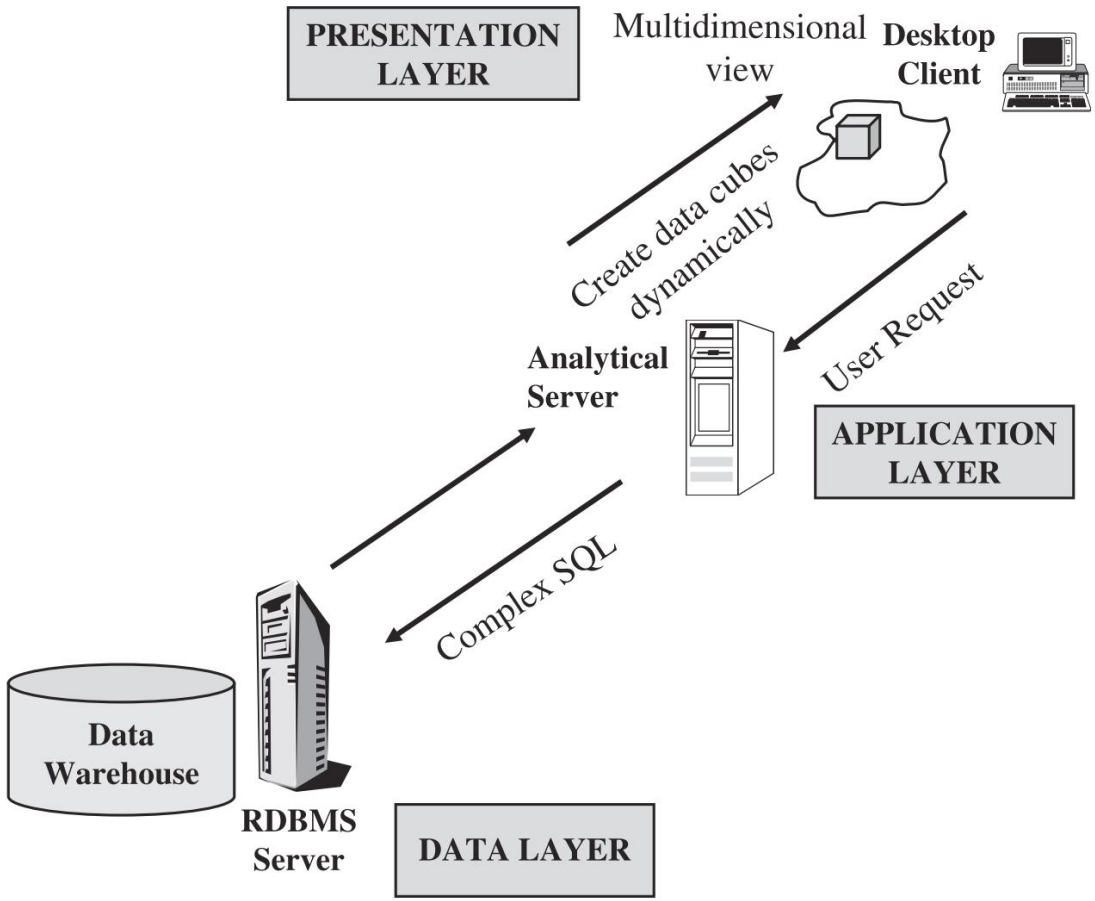


Figure 15-17 The ROLAP model.

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

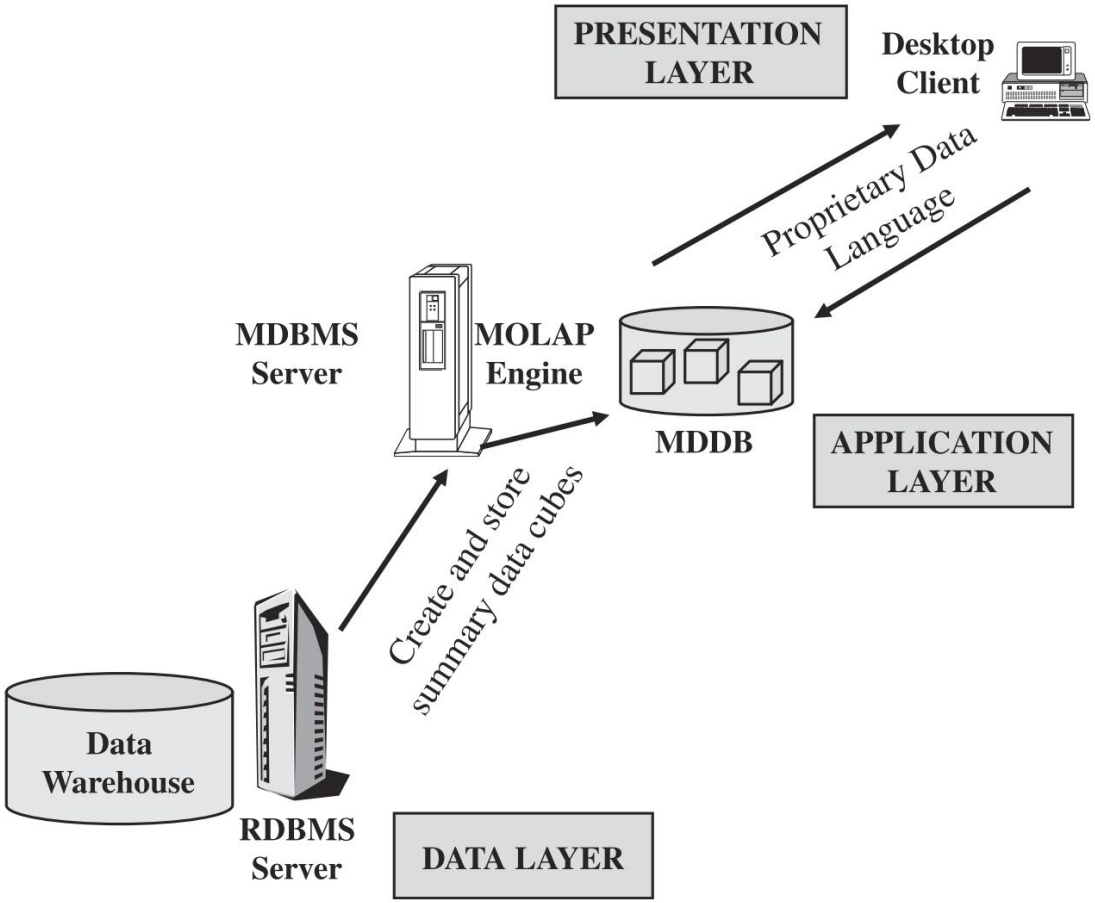


Figure 15-16 The MOLAP model.

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

ROLAP vs MOLAP

	Data Storage	Underlying Technologies	Functions and Features
ROLAP	Data stored as relational tables in the warehouse. Detailed and light summary data available. Very large data volumes. All data access from the warehouse storage.	Use of complex SQL to fetch data from warehouse. ROLAP engine in analytical server creates data cubes on the fly. Multidimensional views by presentation layer.	Known environment and availability of many tools. Limitations on complex analysis functions. Drill-through to lowest level easier. Drill-across not always easy.
MOLAP	Data stored as relational tables in the warehouse. Various summary data kept in proprietary databases (MDDBs) Moderate data volumes. Summary data access from MDDB, detailed data access from warehouse.	Creation of pre-fabricated data cubes by MOLAP engine. Propriety technology to store multidimensional views in arrays, not tables. High speed matrix data retrieval. Sparse matrix technology to manage data sparsity in summaries.	Faster access. Large library of functions for complex calculations. Easy analysis irrespective of the number of dimensions. Extensive drill-down and slice-and-dice capabilities.

Figure 15-19 ROLAP versus MOLAP.

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.

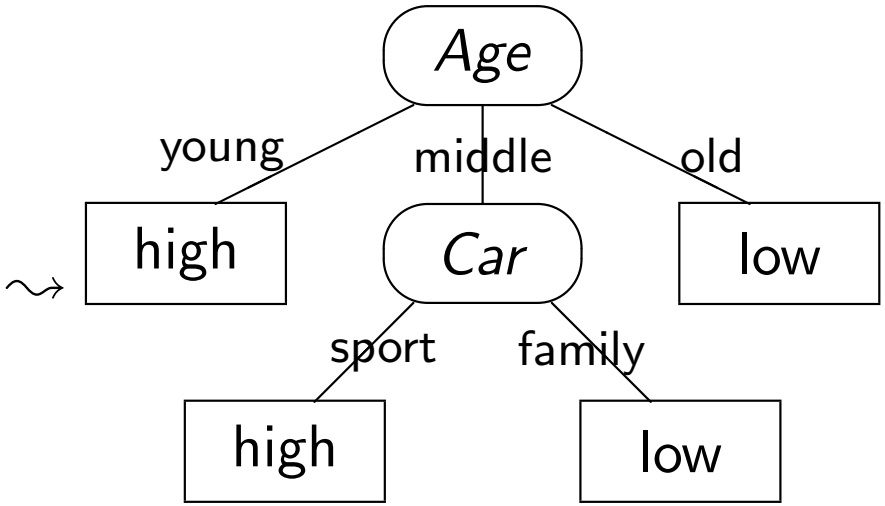
Source: Oded Maimon, Lior Rokach (Eds.): *The Data Mining and Knowledge Discovery Handbook* (2nd Edition). Springer, 2010

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

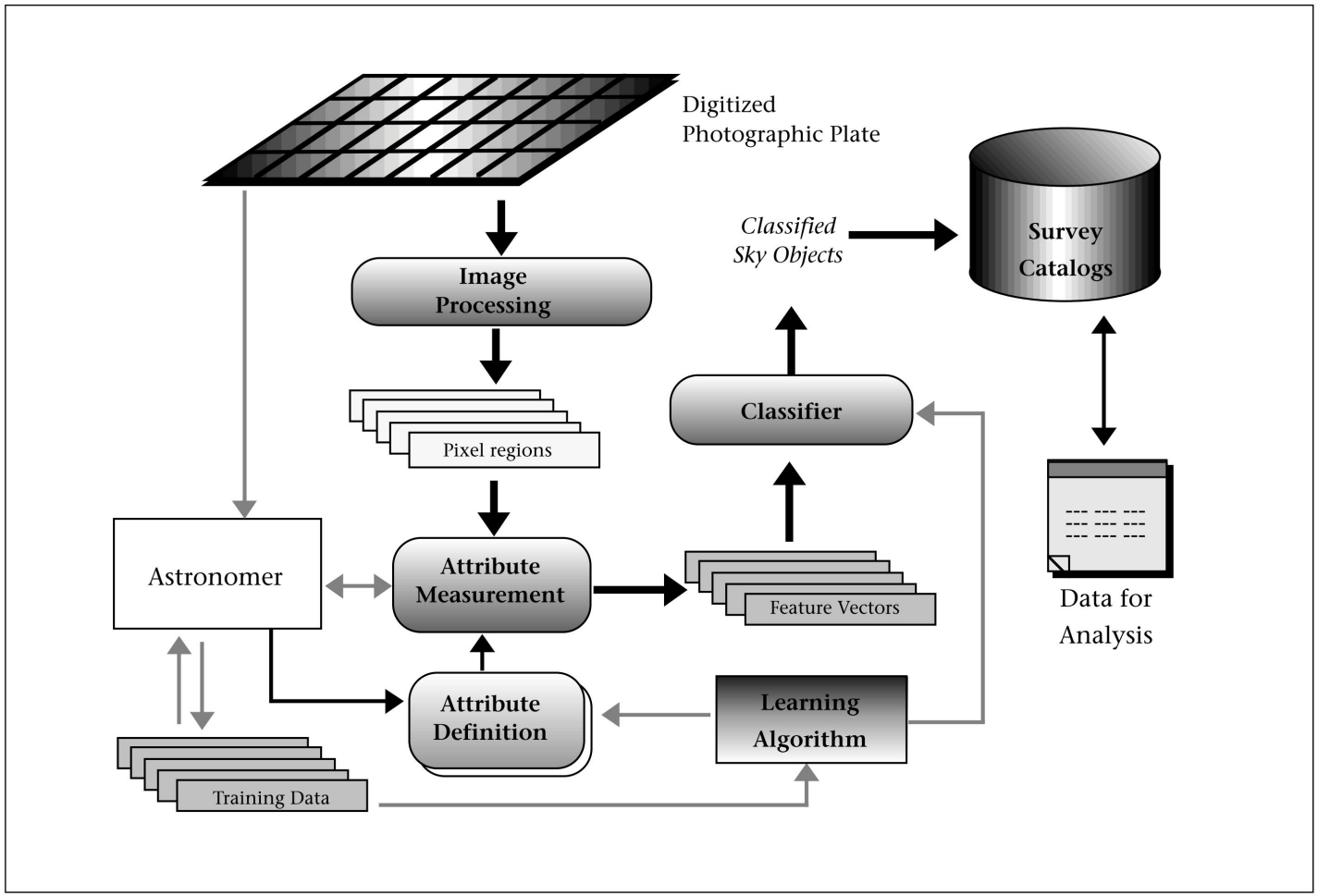


Figure 3. An Overview of the SKICAT Plate-Cataloging Process.

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

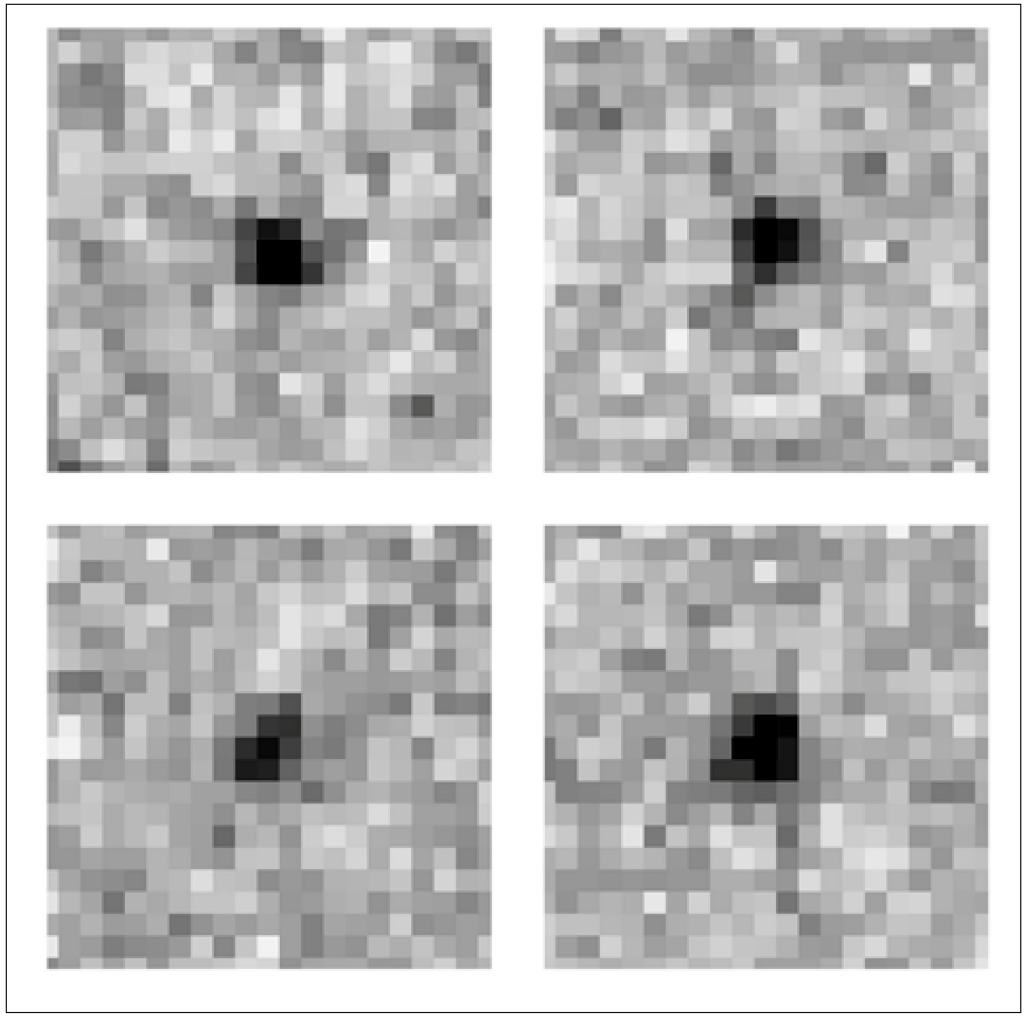
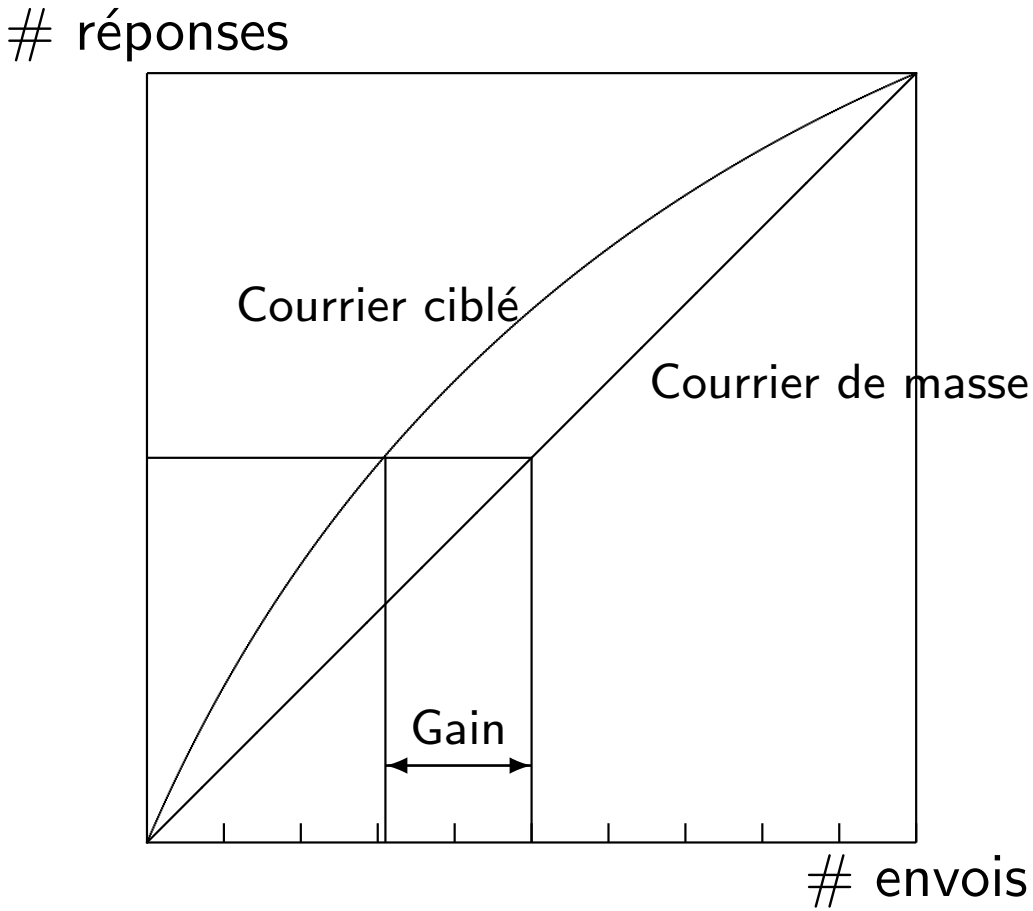


Figure 6. An Illustrative Example: Four Faint Sky Objects.

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

Campagne publicitaire ciblée



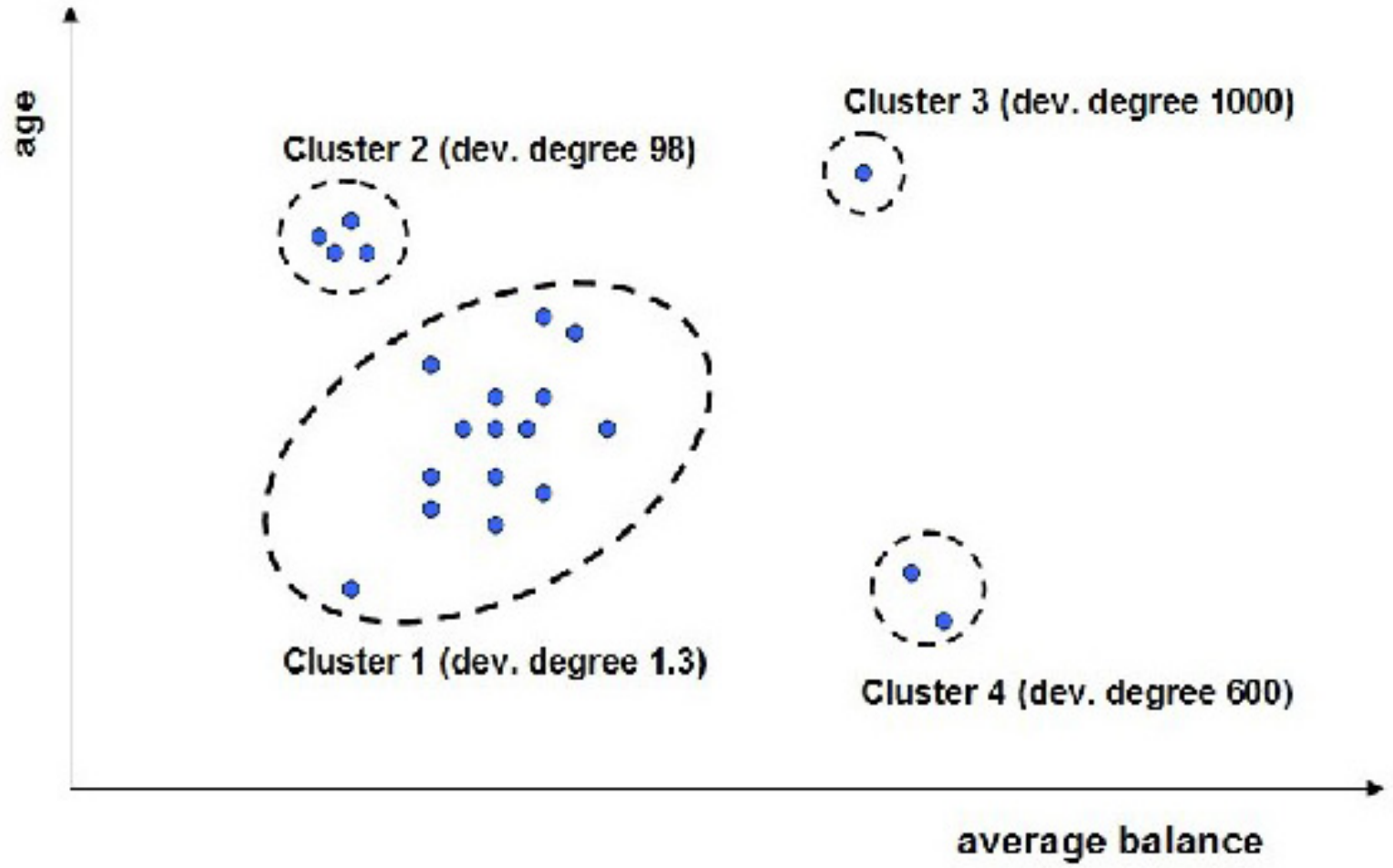
DM Example: Unknown Pattern

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

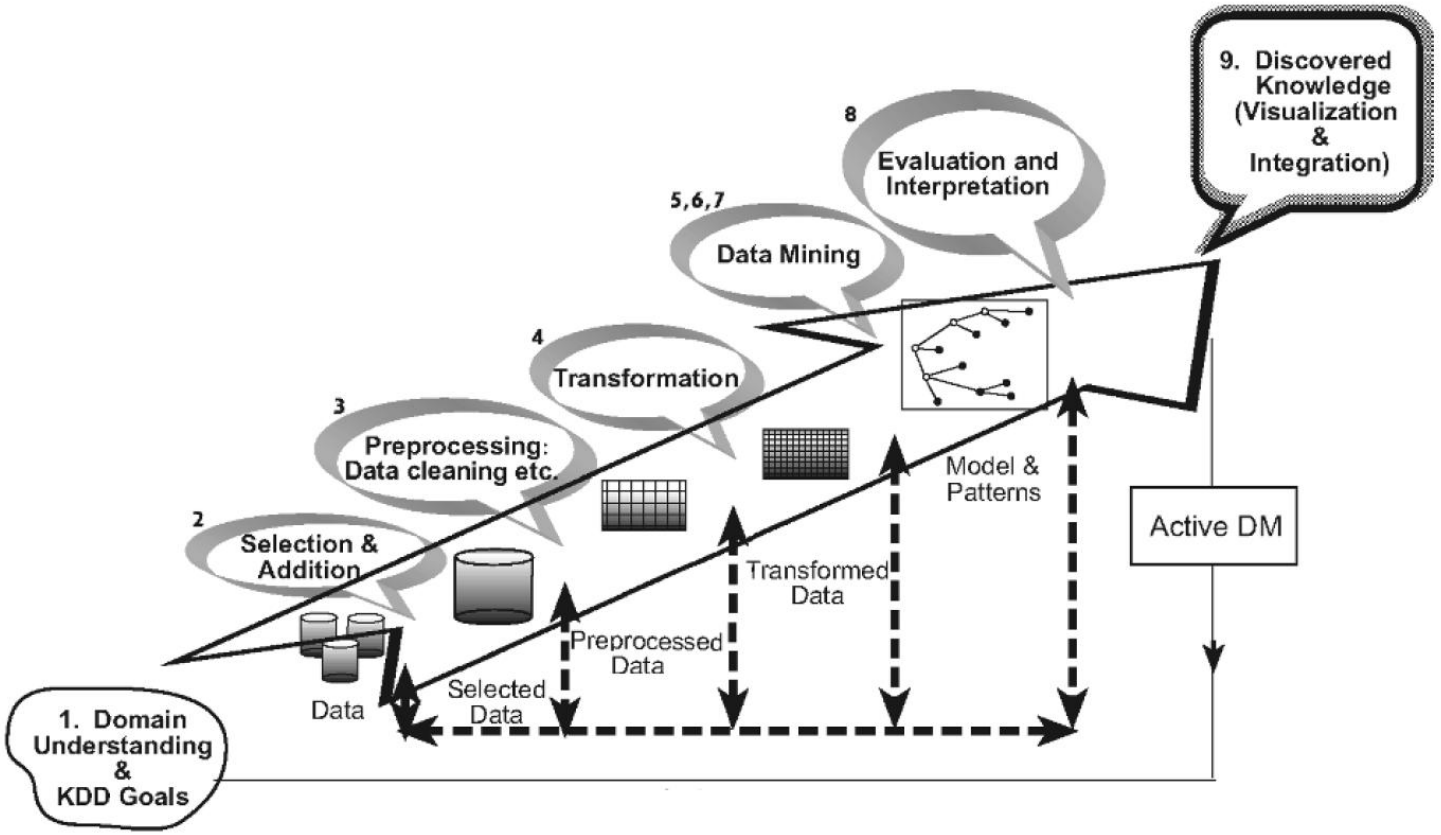


Fig. 1.1. The Process of Knowledge Discovery in Databases.

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

Data Mining vs OLAP

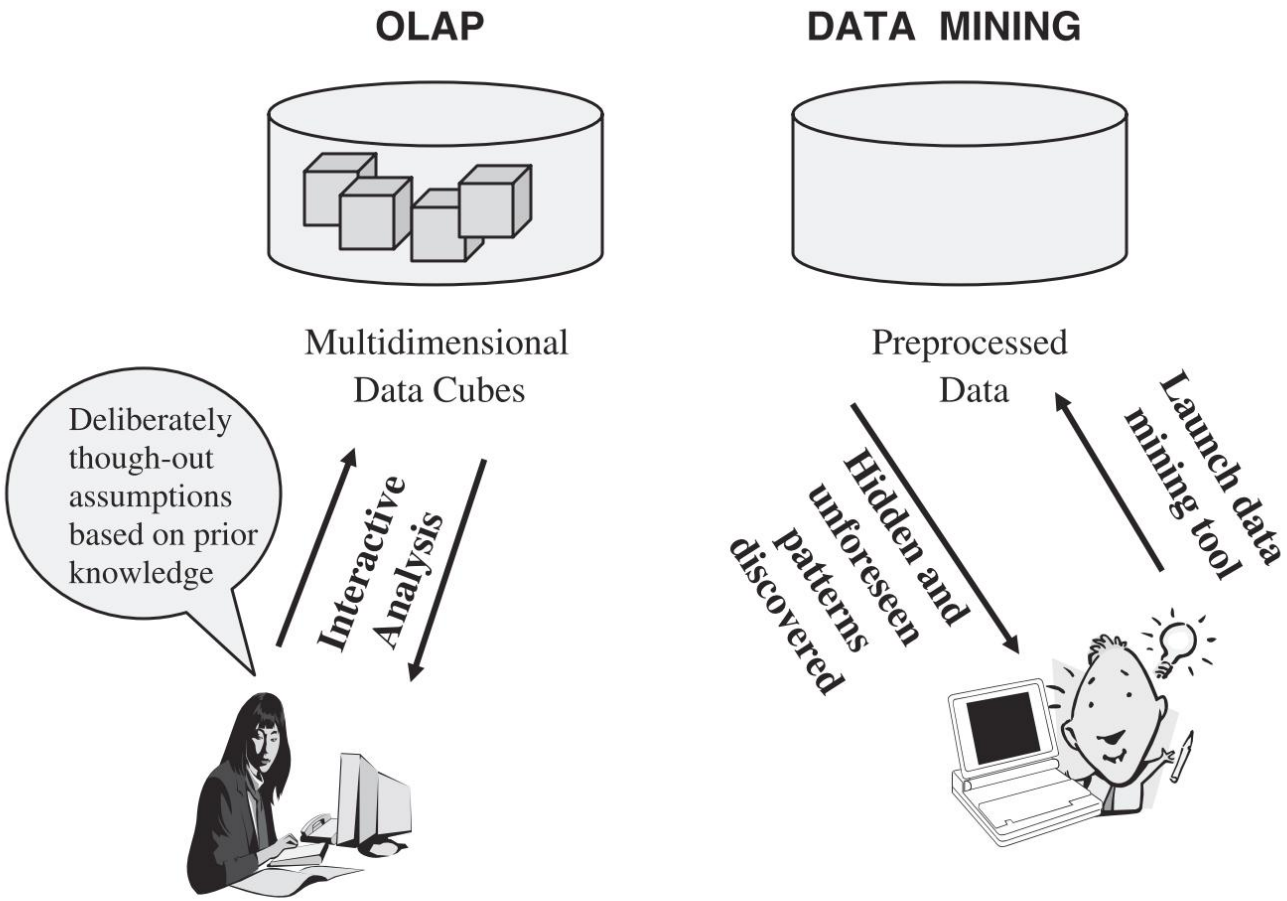


Figure 17-5 OLAP and data mining.

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

Decision Support Systems

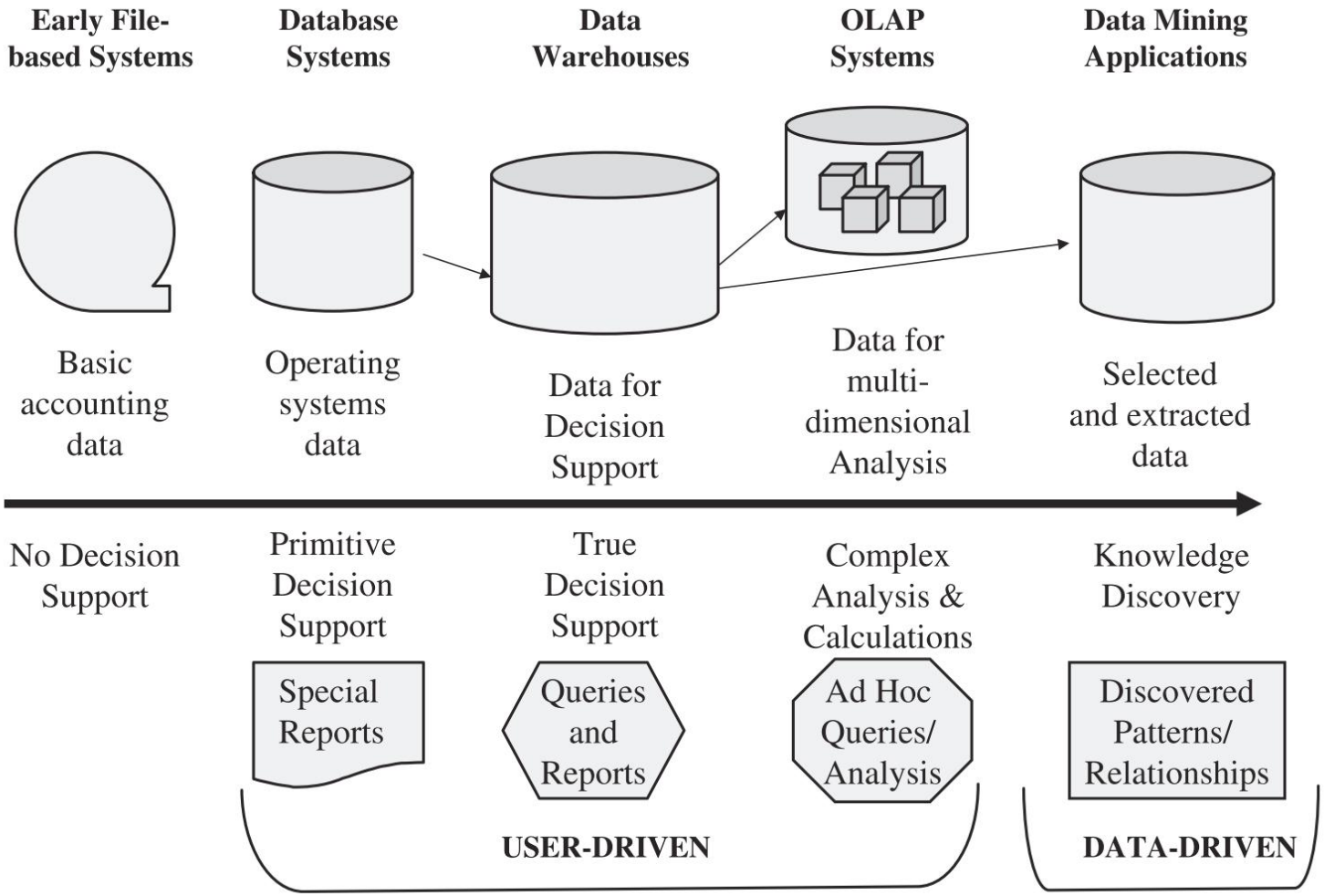


Figure 17-1 Decision support progresses to data mining.

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

Data Mining Taxonomy

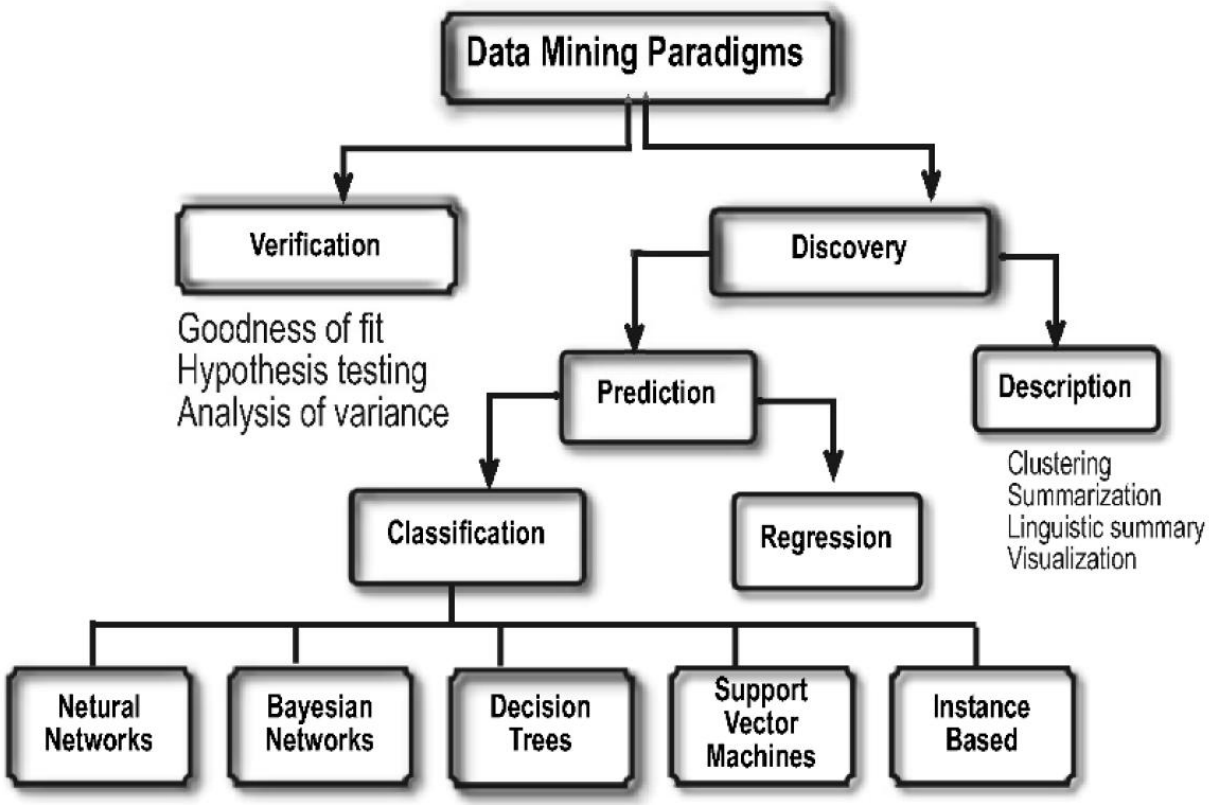
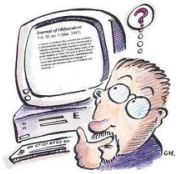


Fig. 1.2. Data Mining Taxonomy.

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



Travail Pratique

Qu'est-ce que le data mining peut apporter à une chaîne de supermarchés ?