Data Warehouses - Introduction & Overview

Databases and Information Systems

Fabian Panse

panse@informatik.uni-hamburg.de

University of Hamburg

Acknowledgements

These slides are based on slides provided by

- Prof. Dr. Erhard Rahm
  University of Leipzig
  `http://dbs.uni-leipzig.de/`

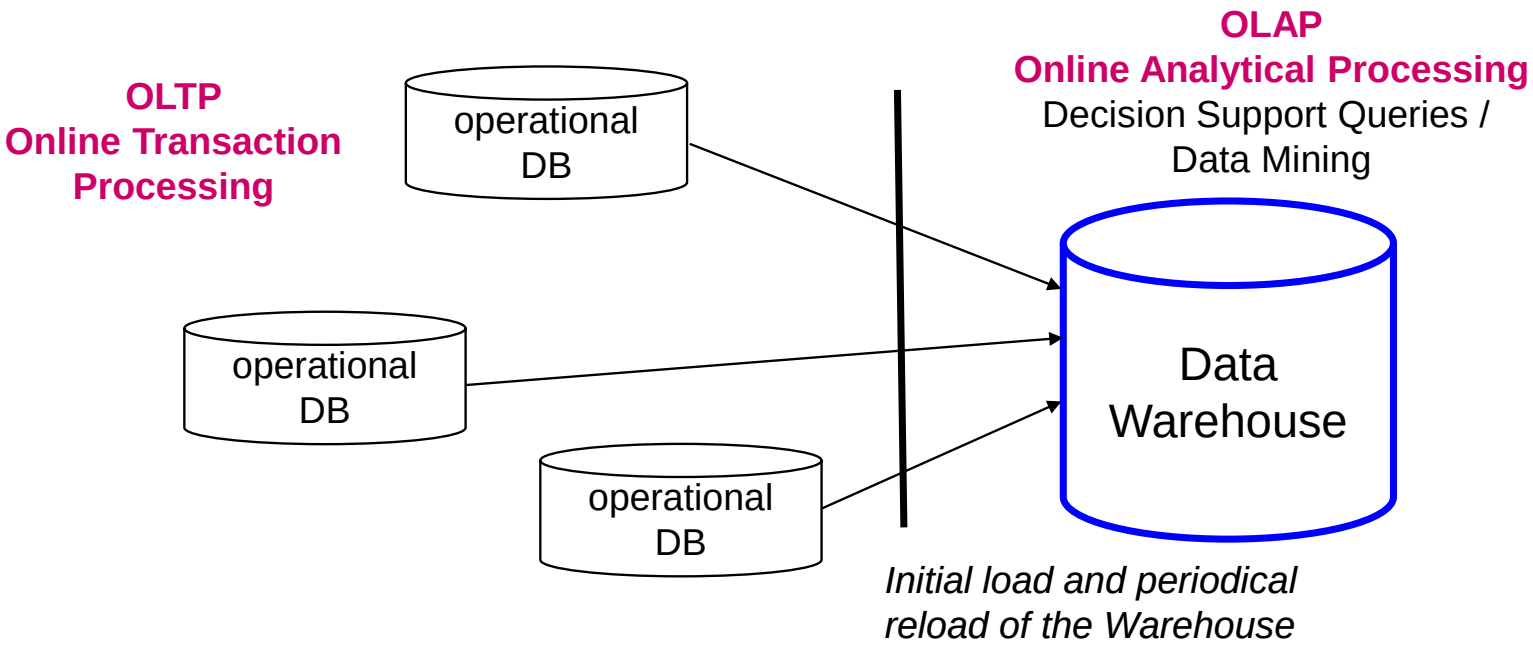
Overview

- Definition Data Warehouse
- Use cases
- OLTP vs. OLAP
- Architecture
- Virtual vs. Physical Data Integration
- Multi-dimensional perspective
- Star-schema, -queries
- Data Mining

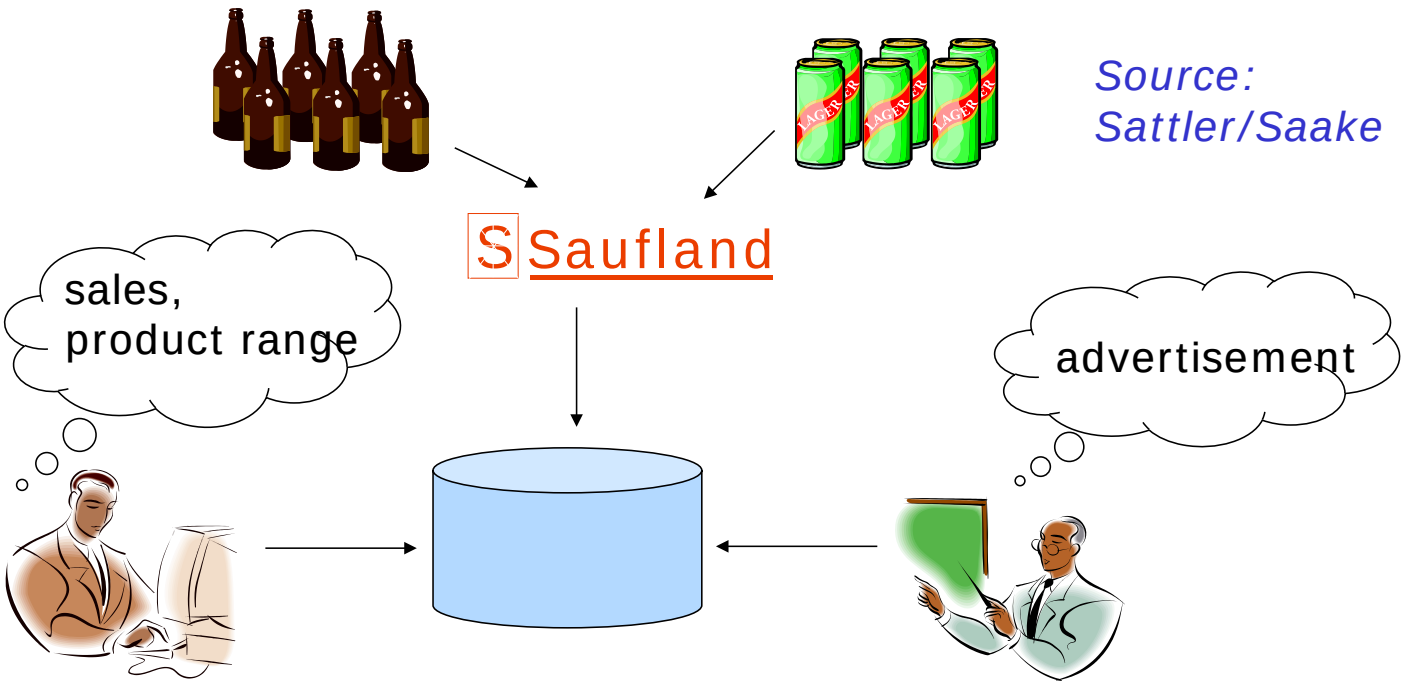
Data Warehouses - Definition

Problem: Many companies have vast quantities of data, but cannot derive much information or knowledge from their data that can be used in critical decision-making tasks

Data Warehouse (Def.): central database that is optimized for analyses and which combines and consolidates data from several heterogeneous sources (integration and transformation)



Scenario: Beverage store



Queries:

- How many bottles cola have been sold last month?
- How has the sale of red wine developed over the past year?
- Who are our premium customers?
- From which supplier do we get the most beverage crates?

Scenario: Beverage store (2)

Queries:

- Did we sell more beer in Hamburg than in Berlin?
- How much cola has been sold during the last summer in north Germany?
- More than water?

Use Cases

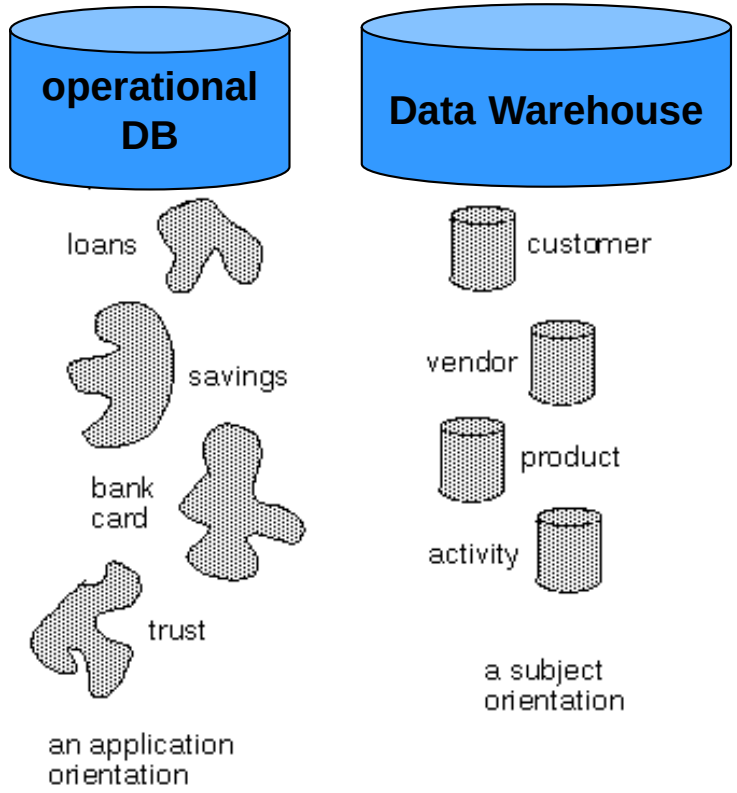
- **Department store chains**
  - Sales figures and inventories of department stores
  - Multi-dimensional analysis: Sales figures by products, regions, branches
  - Detection of bestsellers and non-sellers
  - Analysis on the buying behavior of customers (market basket analysis)
  - Success monitoring of marketing activities
  - Minimization of inventories and sold-out times
  - Optimization of the product range
  - Optimization of pricing
- **Insurance companies**
  - Rating of branches, sectors, ...
  - Automatic risk analyses
  - Faster credit ratings, Life insurance, Health insurance ...
- **Banks, mail-order companies, restaurant chains**
- **Scientific applications (e.g. bioinformatics)**

DW-Properties according to Inmon

A Data Warehouse is a **subject-oriented**, **integrated**, **non-volatile**, and **time variant** collection of data in support of managements decisions (*W. H. Inmon, Building the Data Warehouse, 1996*)

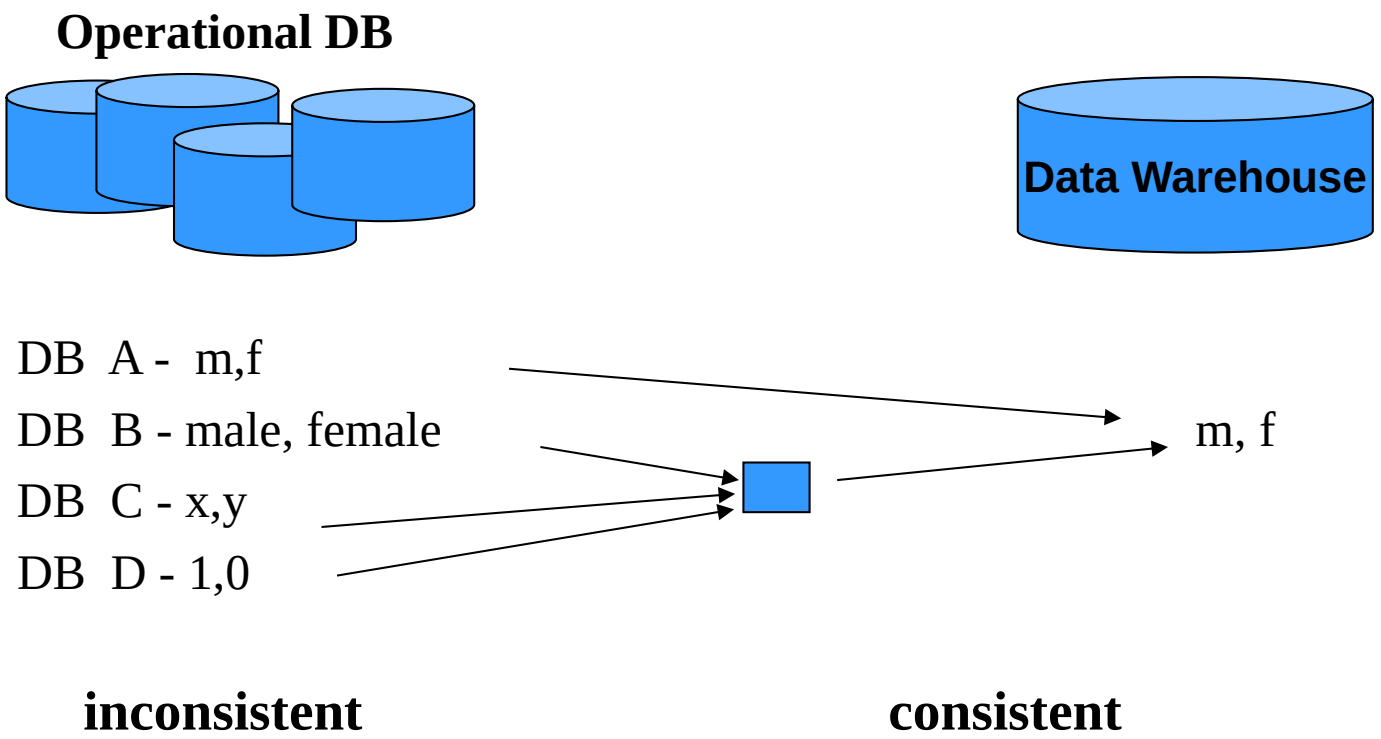
Subject-oriented:

- Purpose of the system is not the fulfillment of a dedicated task (e.g. personnel data management), but the support of methods to evaluate data across individual tasks from different perspectives
- All data – company-wide – about one subject (customer, product, region, ...) within a single system and not "hidden" in different applications



DW-Properties according to Inmon (2)

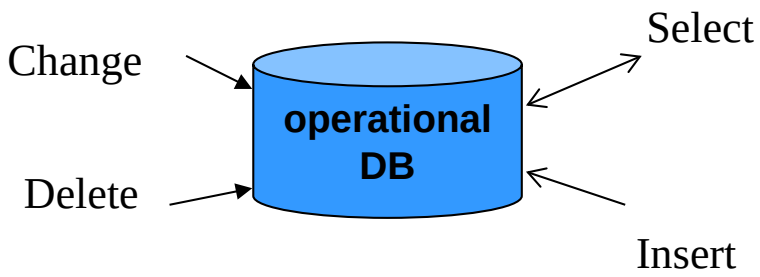
Integrated databases: Data from several distinct data sources

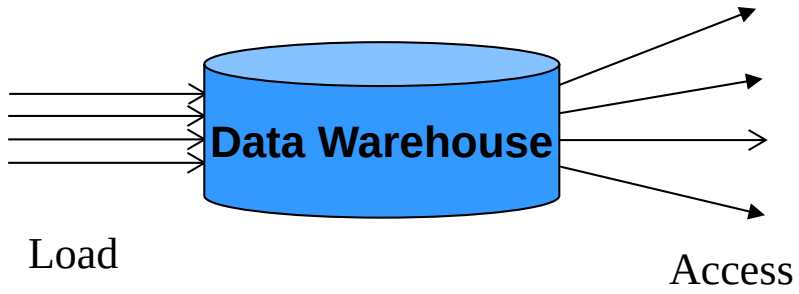


DW-Properties according to Inmon (3)

Non-volatile Databases:

- data values in DW are usually not changed anymore
- stable, persistent database



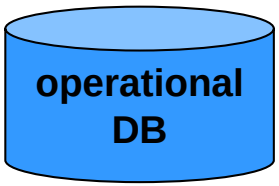


continual changes of data records

DW-Properties according to Inmon (4)

Historical data (time-variant):

- Comparison of data across different periods of time (time series analysis)
- Storage of data for a longer period of time



Time Variancy

Current data values:

- Reference to time only optional
- Time frame: 60-90 days
- Data changeable

Snapshot data

- Reference to a particular time for every object
- Time frame: 2-10 years
- No changes after the snapshot has been made

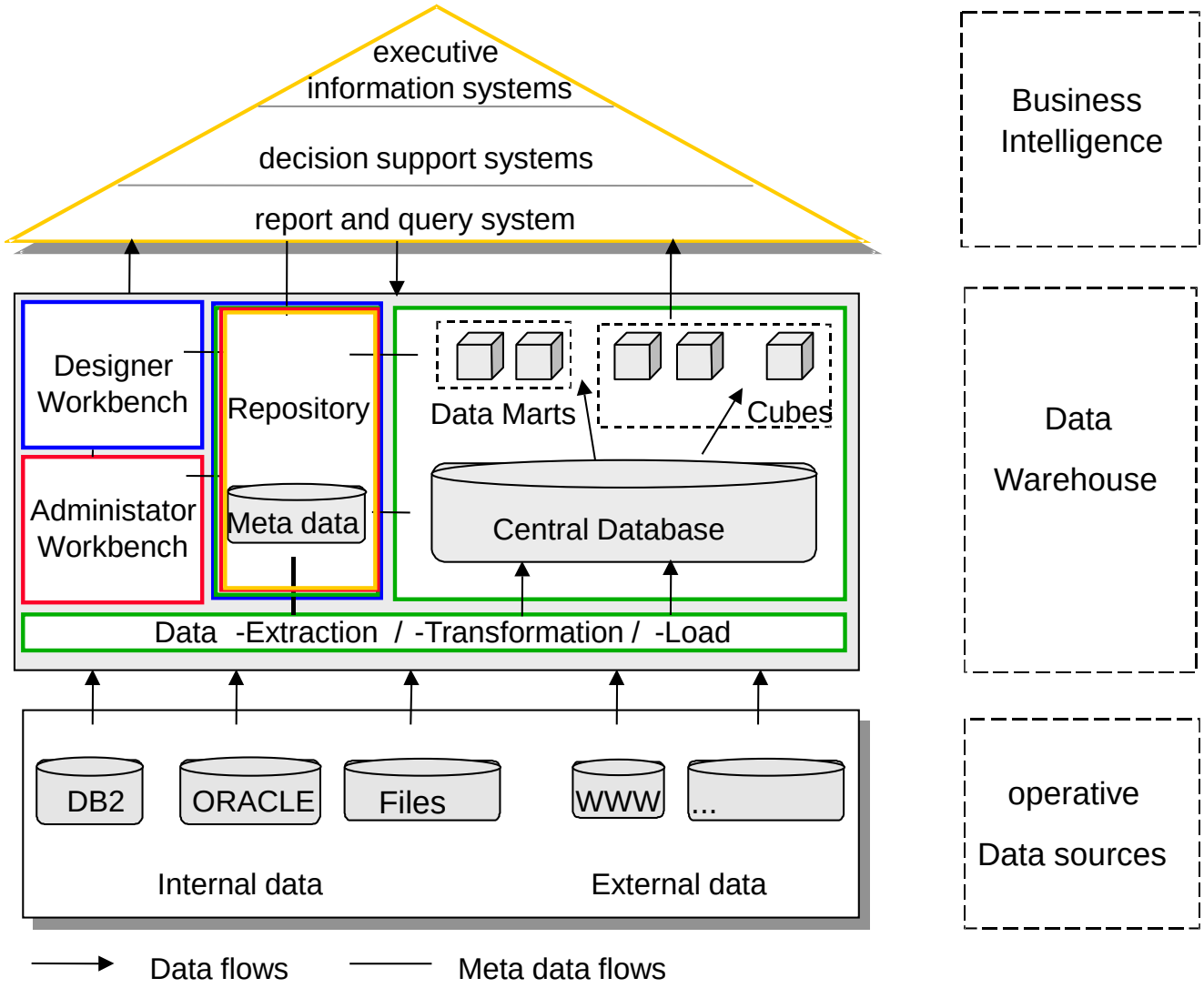
Operational DBs (OLTP) vs. Data Warehouses (OLAP)

	Operational Databases /OLTP	**Data Warehouses/OLAP**
Development	for one application or based on a particular perspective	several perspectives
Relevance	daily business	decision-making, planning tasks
User	case worker, online user	analyst, manager
Data Access	high access frequency; small amount of data per operation; read, write, update, delete	moderate access frequency; large amount of data; primarily read only
Changes/ Up-to-dateness	very often / always up-to-date	periodically / usually outdated
#Data sources	most often only one	several
Data characteristics	not derived, up-to-date, autonomous, dynamic	derived, not up-to-date, integrated, stable
Queries	fixed set of queries	not known in advance
Optimization goals	high throughput, short response time (ms .. s), high availability	acceptable response time for complex analysis, high flexibility

Why do we need a separate Data Warehouse?

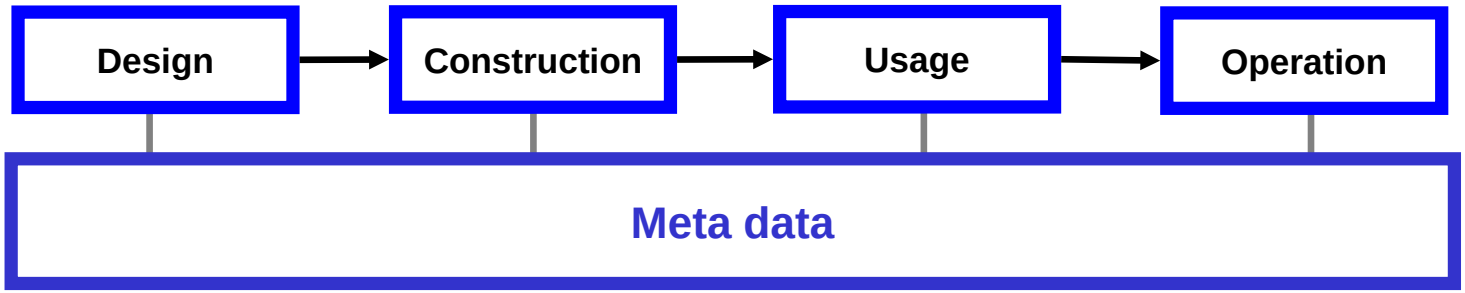
- **Different use cases and different data structures**
- **Performance**
  - OLTP is optimized for short transactions and known load profiles
  - Processing of complex OLAP-queries would decrease the throughput of simultaneously executed OLTP-transactions significantly
  - Multi-dimensional views/queries require a specific logical and physical database design
  - Properties of transactions (ACID) not important
- **Functionality**
  - Historical data
  - Consolidation (integration, cleaning and aggregation) of data from heterogeneous data sources
- **Drawbacks of a separate solution**
  - Data redundancy
  - Data is not always up-to-date
  - High administration effort, high costs (e.g. hardware)

Architecture of a DW-Enviroment



DW-Processes

- **Data Warehousing includes several sub-processes**
  - Design ("design it"),
  - Construction ("build it", "populate"),
  - Usage ("use it", "analyze") as well as
  - Operation and Administration ("maintain it"/"administer")



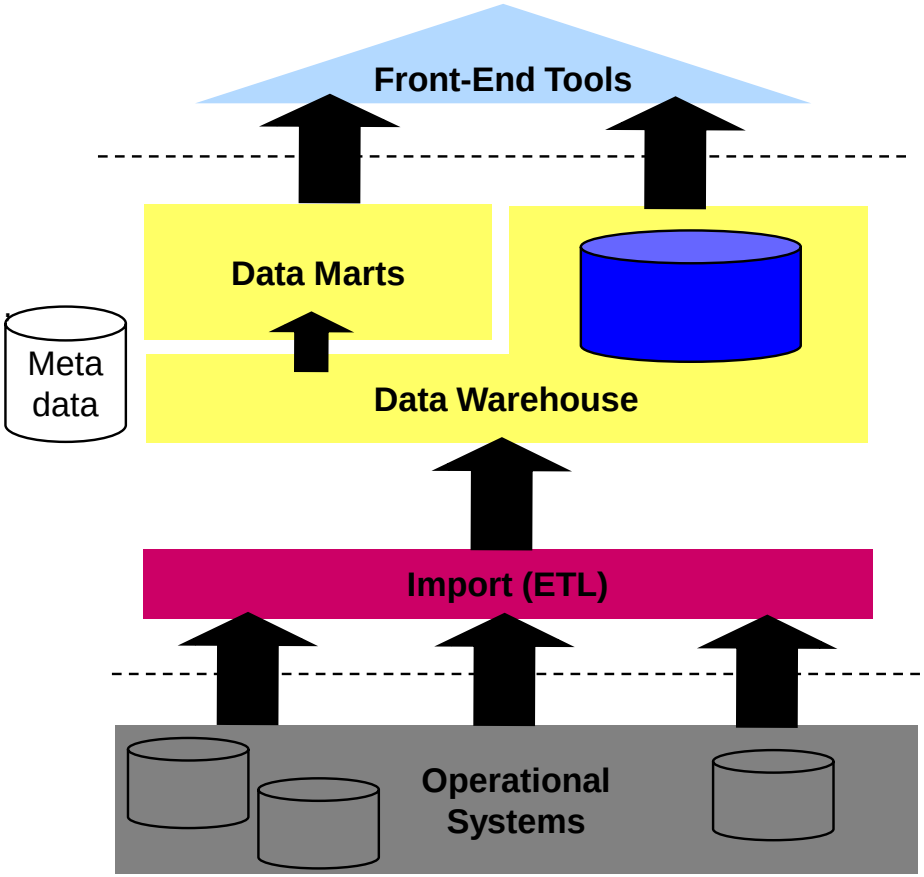
- **DW is usually not a monolithic system**
  - Most often use of tools/components from different producers as well as self-programmed components
- **Central importance of meta data, but often not sufficiently supported**

Problems in Setting Up a Data Warehouse

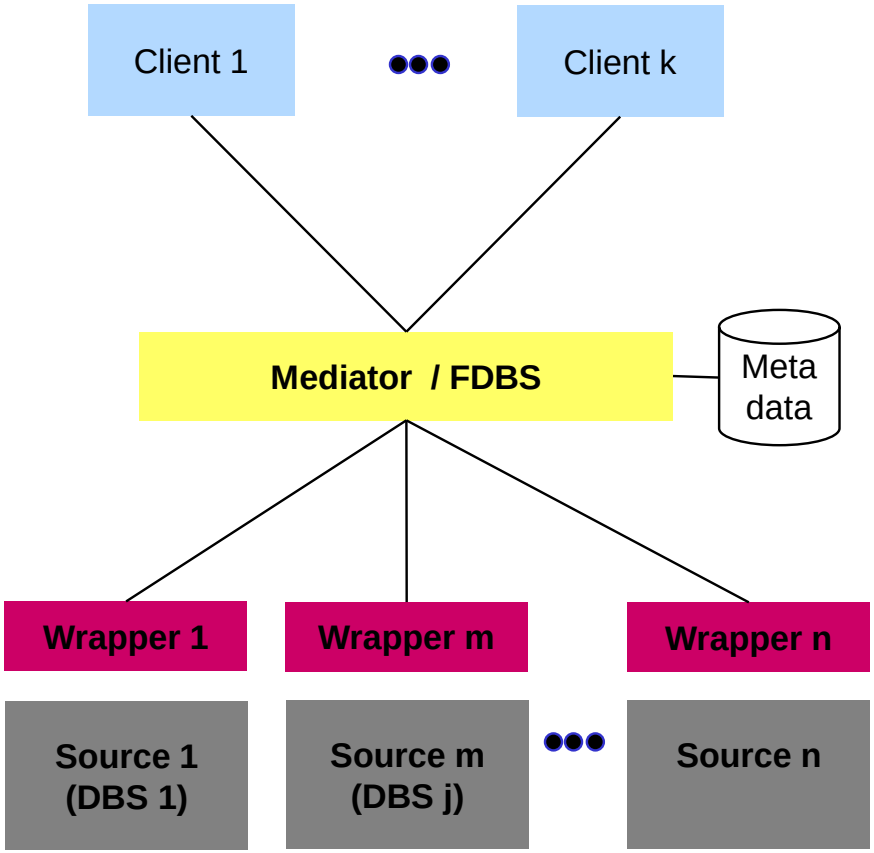
- Underestimation of resources for data loading
- Hidden problems with the source systems (e.g. missing data)
- Required data not captured
- Increased end-user demands
- Demanding resource requirements
- Conflicts between owners of data
- High maintenance requirements
- Long-duration project
- Complexity of integration

Data Integration: physical vs. virtual

Physical Integration
(Data Warehousing)



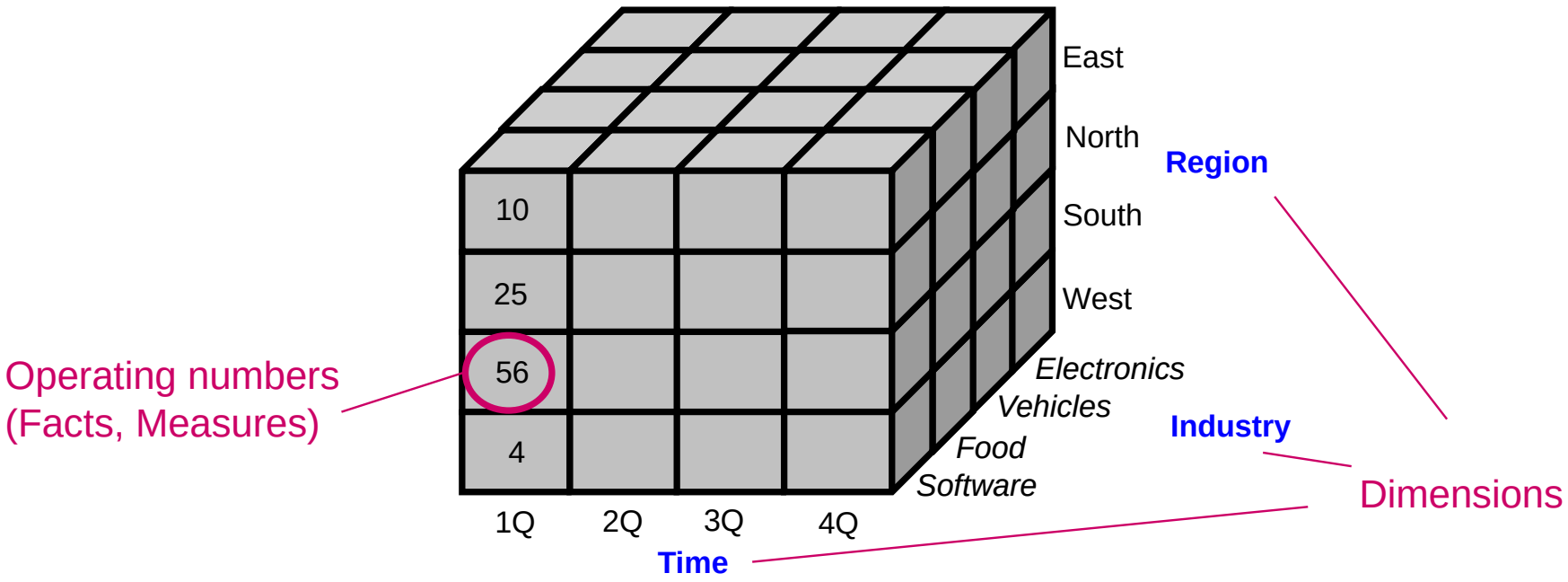
Virtual Integration
(Mediator/Wrapper-Architectures, federated DBS)



Data Integration: physical vs. virtual (2)

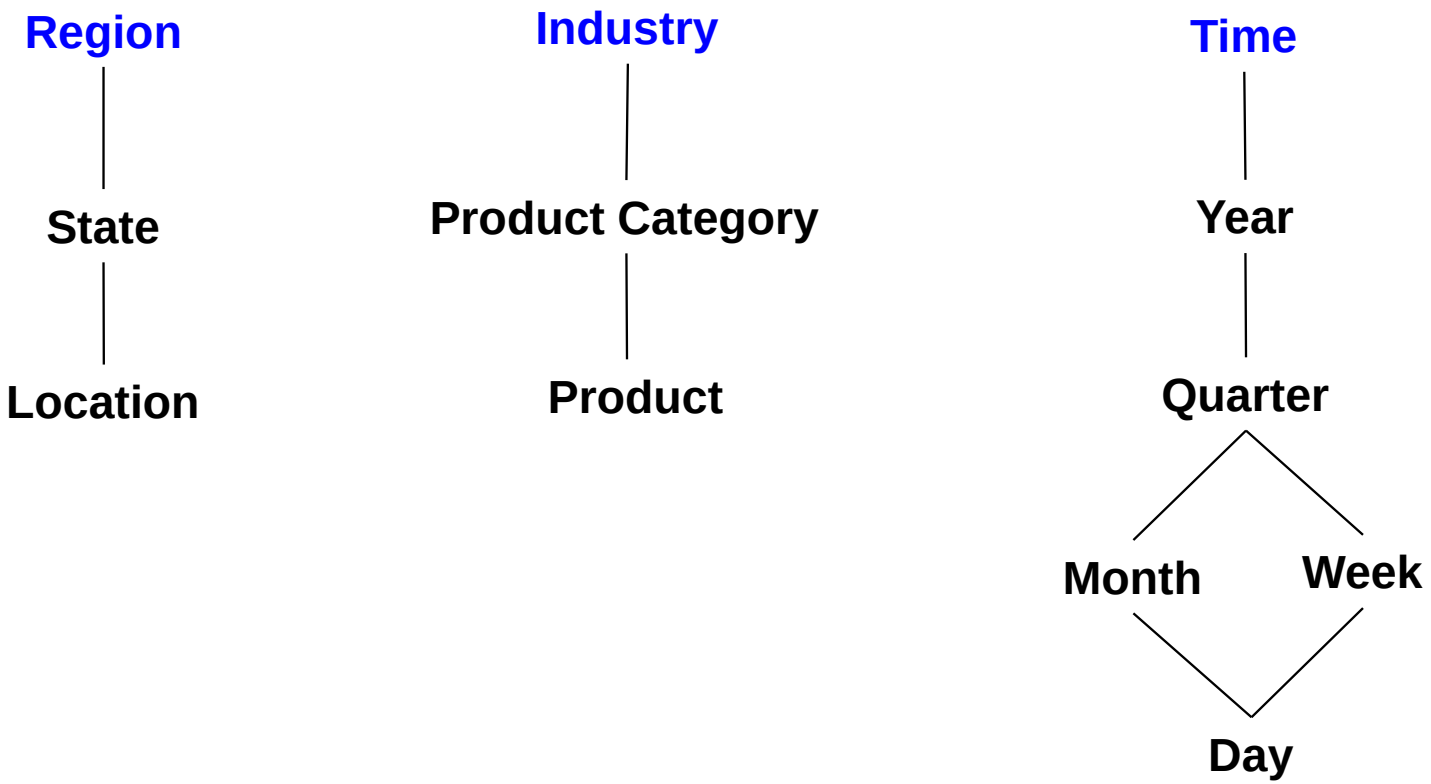
	Physical (Data Warehouse)	Virtual
Time of integration: Meta data	beforehand (DW-Schema)	beforehand (global schema)
Time of integration: Data	beforehand	dynamic (at query time)
Up-to-dateness	o	+
‚Autonomy of the data sources	o	+
Achievable data quality	+	o
Time requirements for analysis on large data sets	+	-
Hardware costs	-	o
Scalability with respect to number of data sources	-	-

Multi-dimensional view of data



- **Operating numbers:** numerical values as basis of aggregation / computation (e.g. sales figures, revenue)
- **Dimensions:** descriptive properties
- **Operations:**
  - Aggregation of the operating numbers over one or more dimension(s)
  - Slicing and Dicing: Restriction on particular (parts of) dimensions

Hierarchical Dimensioning



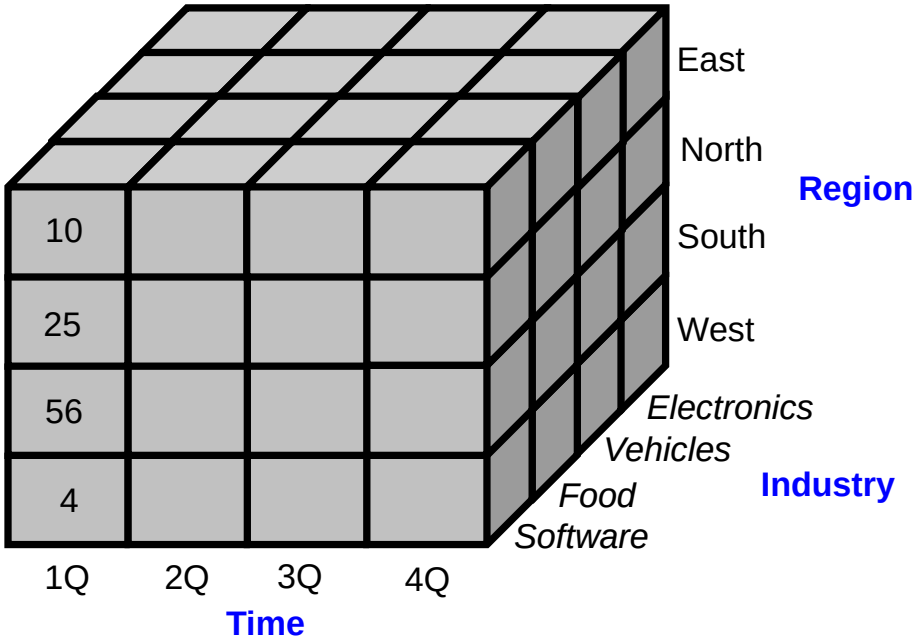
Operations to change the granularity of the individual dimensions

- Drill-Down
- Roll-Up

OLAP (Online Analytical Processing)

- **Interactive and multi-dimensional analyses on consolidated data of a company**
- **Characteristics / Requirements:**
  - Multi-dimensional, conceptual view of the data
  - Unlimited number of dimensions and aggregation levels
  - Operations across dimensions
  - Intuitive and interactive data manipulation/visualization
  - Transparent (integrated) access to heterogeneous databases with a logical overall view
  - Scalability with respect to large data sets
  - Stable and volume depending response time
  - Multi-client support
  - Client/Server-Architecture

Multi-dimensional vs. relational

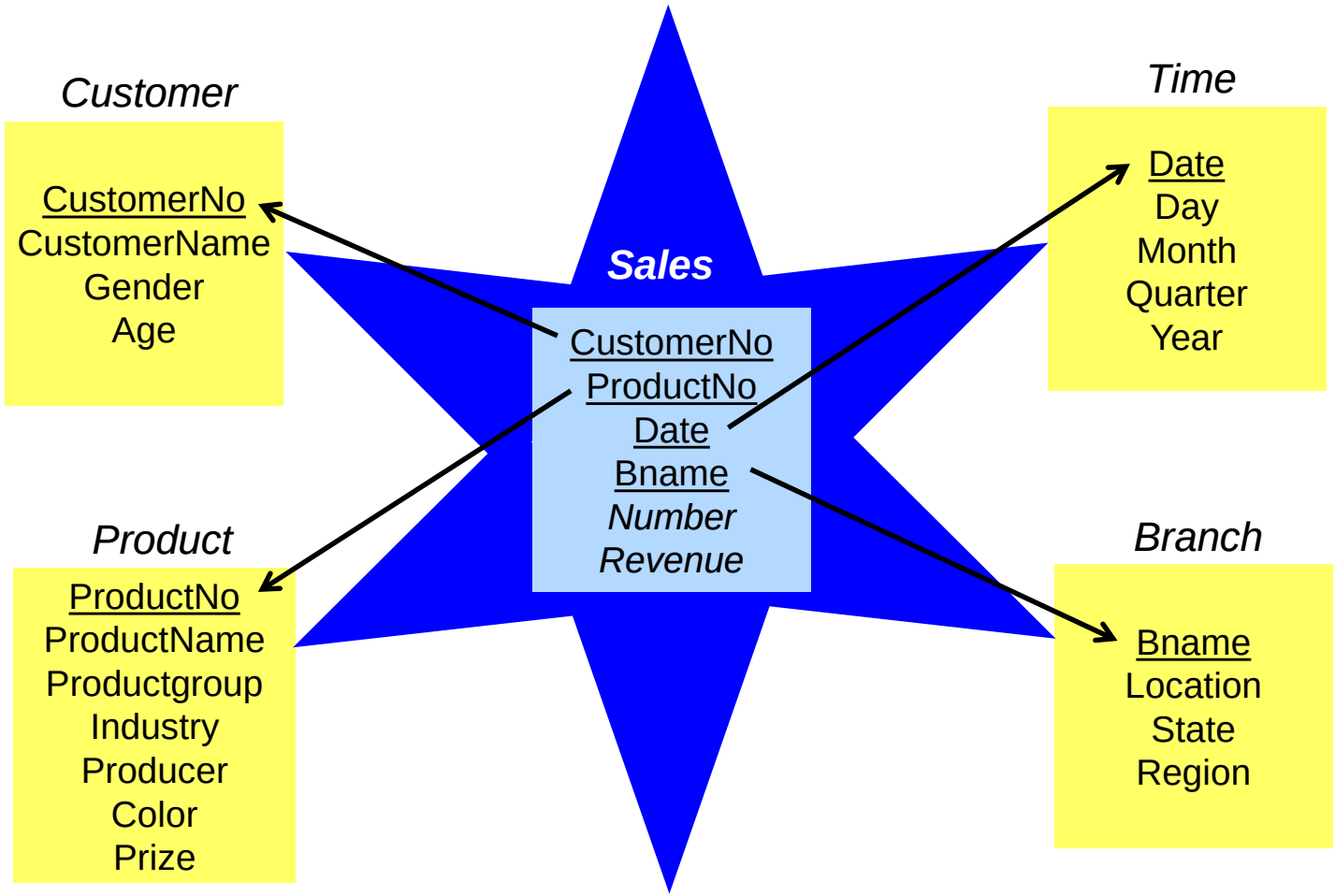


Order no.	Region	Industry	Time	Amount
1406	East	Vehicles	2Q	5
4123	West	Electronics	1Q	58
7829	South	Vehicles	2Q	30
5327	East	Food	4Q	3000
9306	North	Software	1Q	25
2574	East	Electronics	4Q	2

- **Multi-dimensional Representation (MOLAP):**
  - Cross product of all domains with aggregated value per combination
  - Assumption: almost all combinations occur
- **Relational Representation (ROLAP):**
  - Relation: Subset of the cross product of all domains
  - Only occurring combinations are stored
- **Hybrid OLAP (HOLAP): ROLAP + MOLAP**

Star Schema

Central fact table and one table per dimension



Queries

Sample query: *Which car producer was preferred by female customers in Hamburg in the first quarter of 2008?*

```
SELECT     p.Producer, SUM(s.Number)
FROM       Sales s, Branch b, Product p, Time t, Customer c
WHERE      t.Year = 2008              AND   t.Quarter = 1
  AND      c.Gender = 'W'             AND   p.Productgroup = 'Car'
  AND      b.State = 'Hamburg'        AND   s.Date = t.Date
  AND      s.BName = b.BName          AND   s.ProductNo = p.ProductNo
  AND      s.CustomerNo = c.CustomerNo
GROUP BY   p.Producer;
```

Star Join:

- Starlike Join of the (relevant) dimension tables with the fact table
- Restriction of the individual dimensions
- Consolidation of the operating numbers by grouping and aggregation

Analysis Tools

- **(Ad Hoc) Query tools**
- **Reporting tools, reports with flexible formatting options**
- **OLAP tools**
  - Interactive and multi-dimensional analyses and navigation (Drill Down, Roll Up, ...)
  - Grouping, statistical computations, ...
- **Data mining tools**
- **Representation**
  - Tables, particularly pivot-tables (cross tables)
  - Analyses by interchanging rows and columns, changing of table dimensions
  - Graphs as well as text and multimedia elements
- **Usage per Web Browser, Spreadsheet integration**

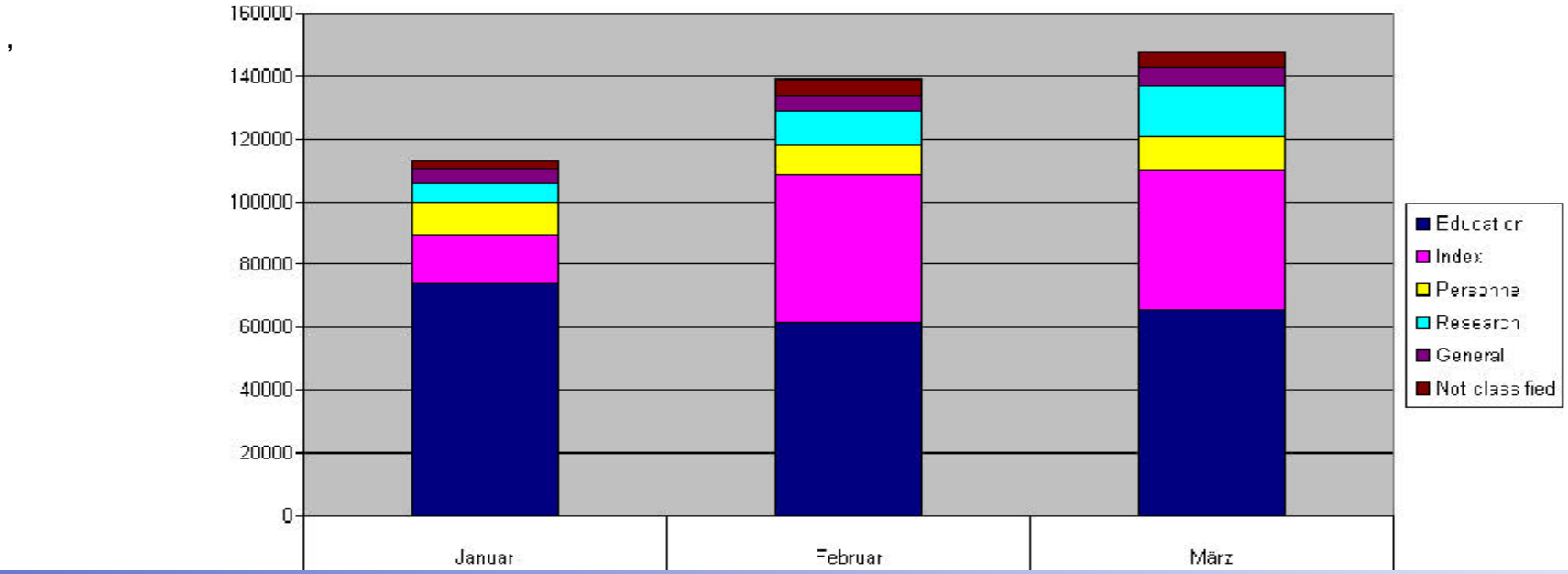
Example: OLAP-Output (Excel)

Zugriffe

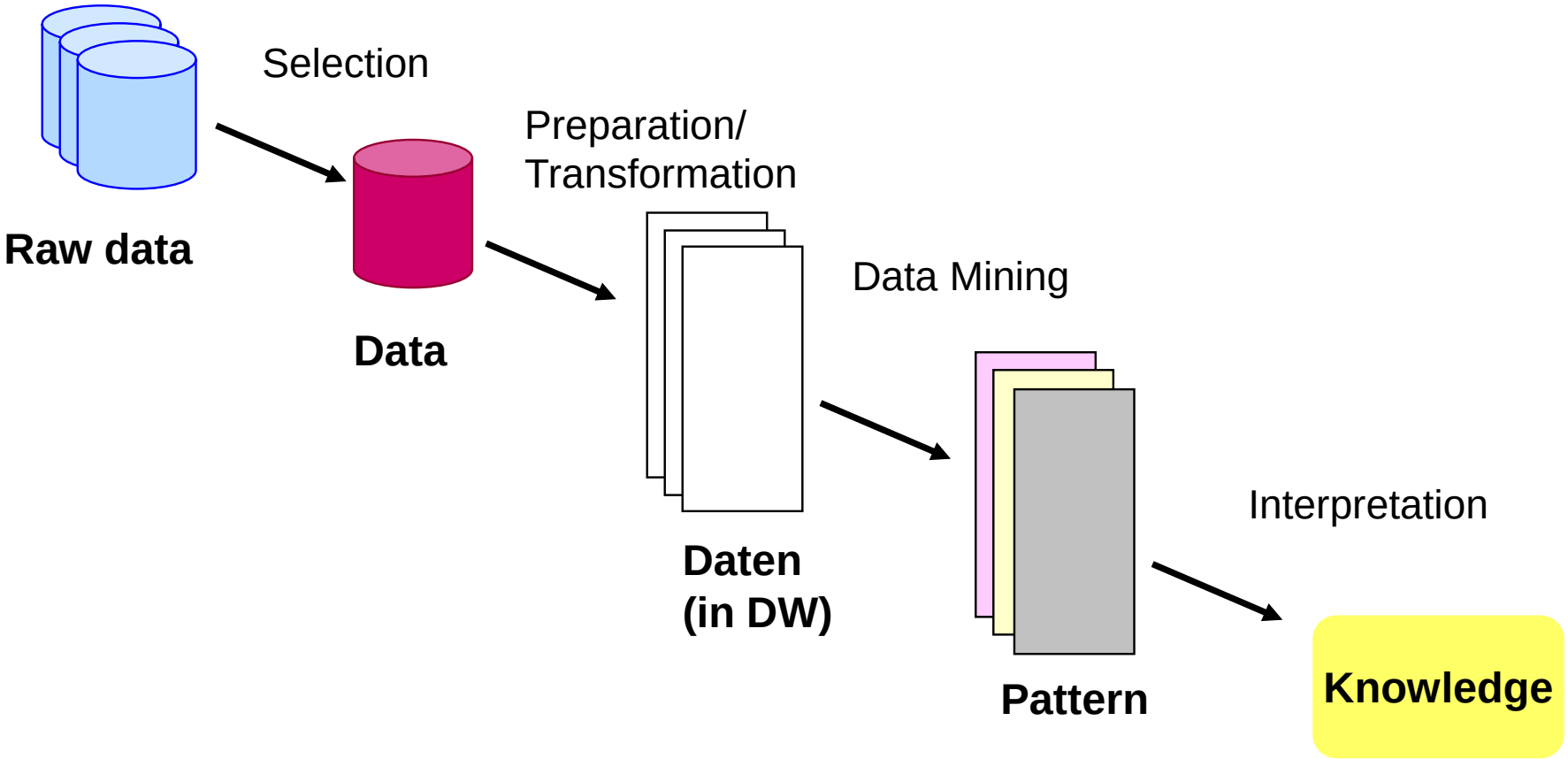
Source

Alle Source

		Category					
		Education	Index	Personnel	Research	General	Not classified
Jahr	Monat Tag	Hits	Hits	Hits	Hits	Hits	Hits
⊟ 2001	⊞ Januar	73961	15494	10559	5079	4915	2060
	⊞ Februar	61697	46666	9880	10686	4558	5504
	⊞ März	65642	44708	10439	15837	5871	5334
	Total *	213494	115430	32867	33501	16054	14275
Gesamtergebnis *		1106493	189912	111213	84560	46708	39735



Knowledge Discovery



Data Mining: Techniques

- **Data Mining:**
  - Usage of statistic- and knowledge-based methods
  - Detection of correlations, patterns or trends in the given data
  - "Knowledge Discovery": In contrast to OLAP ("knowledge verification"), KD does not require a formal model
- **Cluster analysis:**
  - Grouping of objects based on their similarities
  - Example: similar customers, similar webpage-user, ...
- **Association rules:**
  - Market basket analysis
    (e.g. customer buys A and B $\Rightarrow$ customer buys C)
- **Classification:**
  - Classification of objects
  - Construction of classification rules/predictions based on attribute values (e.g. "good customer" if age $>$ 25 and ...)
  - Possible realization: decision tree, Support Vector Machines

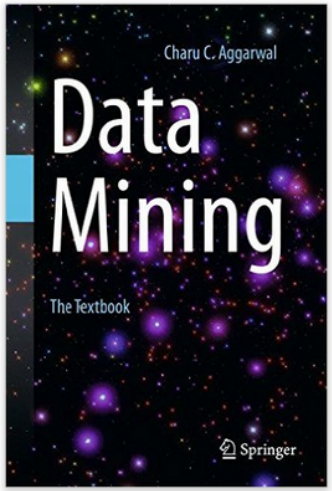
Example: Market basket analysis (Amazon)

Data Mining: The Textbook 2015th Edition

by Charu C. Aggarwal (Author)

8 customer reviews

Look inside



Hardcover $67.52 - $78.04 | Other Sellers from $53.34

Buy used $67.52

Buy new **$78.04**

Only 17 left in stock (more on the way).
Ships from and sold by Amazon.com. Gift-wrap available.

Want it tomorrow, May 19? Order within **4 hrs 3 mins** and choose **One-Day Shipping** at checkout.
Details

List Price: $89.99 Save: $11.95 (13%)
41 New from $53.34
FREE Shipping.
Qty: 1
Add to Cart
Turn on 1-Click ordering
Ship to:
Select a shipping address:

Frequently Bought Together

Total price: $181.29
Add all three to Cart
Add all three to List

This item: Data Mining: The Textbook by Charu C. Aggarwal Hardcover $78.04
Data Mining and Analysis: Fundamental Concepts and Algorithms by Mohammed J. Zaki Hardcover $65.26
Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking by Foster Provost Paperback $37.99

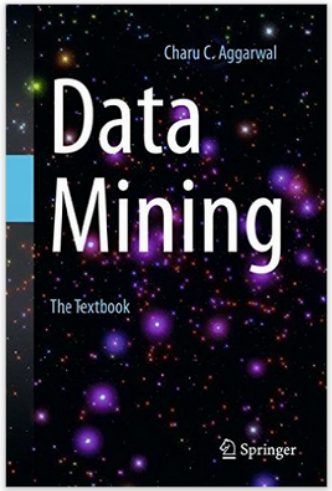
Example: Market basket analysis (Amazon)

Data Mining: The Textbook 2015th Edition

by Charu C. Aggarwal (Author)

8 customer reviews

Look inside



Hardcover $67.52 - $78.04 | Other Sellers from $53.34

Buy used $67.52

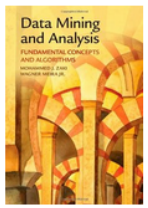
Buy new **$78.04**

Only 17 left in stock (more on the way).
Ships from and sold by Amazon.com. Gift-wrap available.

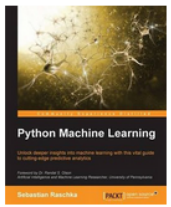
Want it tomorrow, May 19? Order within **4 hrs 3 mins** and choose **One-Day Shipping** at checkout.
Details

List Price: $89.99 Save: $11.95 (13%)
41 New from $53.34
FREE Shipping.
Qty: 1
Add to Cart
Turn on 1-Click ordering
Ship to:
Select a shipping address:

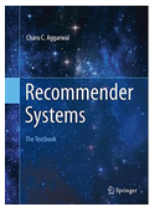
Customers Who Bought This Item Also Bought



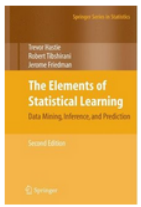
Data Mining and Analysis: Fundamental Concepts and Algorithms
› Mohammed J. Zaki
8
Hardcover
$65.26 Prime



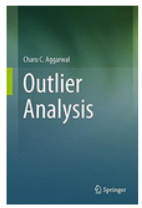
Python Machine Learning
› Sebastian Raschka
46
#1 Best Seller in Computer Neural Networks
Paperback
$40.49 Prime



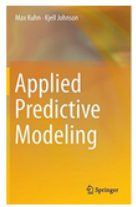
Recommender Systems: The Textbook
› Charu C. Aggarwal
Hardcover
$66.49 Prime



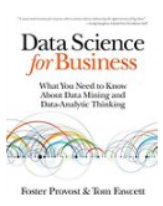
The Elements of Statistical Learning: Data Mining, Inference, and Prediction,...
Trevor Hastie
70
#1 Best Seller in Bioinformatics



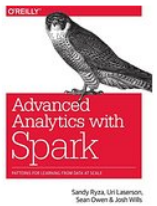
Outlier Analysis
› Charu C. Aggarwal
3
Hardcover
$101.81 Prime



Applied Predictive Modeling
› Max Kuhn
49
Hardcover
$83.79 Prime



Data Science for Business: What You Need to Know about Data Mining and...
› Foster Provost
136
Paperback
$37.99 Prime



Advanced Analytics with Spark: Patterns for Learning from Data at...
Sandy Ryza
20
#1 Best Seller in Website Analytics

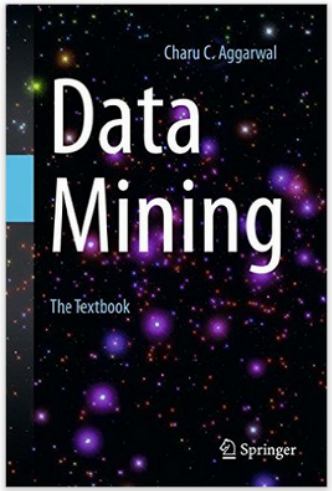
Example: Market basket analysis (Amazon)

Data Mining: The Textbook 2015th Edition

by Charu C. Aggarwal (Author)

8 customer reviews

Look inside



Hardcover $67.52 - $78.04 | Other Sellers from $53.34

Buy used $67.52

Buy new **$78.04**

Only 17 left in stock (more on the way).

Ships from and sold by Amazon.com. Gift-wrap available.

Want it tomorrow, May 19? Order within **4 hrs 3 mins** and choose **One-Day Shipping** at checkout.

Details

List Price: $89.99 Save: $11.95 (13%)

41 New from $53.34

FREE Shipping.

Qty: 1

Add to Cart

Turn on 1-Click ordering

Ship to:

Select a shipping address:

What Other Items Do Customers Buy After Viewing This Item?

Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking Paperback

› Foster Provost

136

$37.99 Prime

Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (Morgan Kaufmann Series in Data Management… Paperback

› Ian H. Witten

70

$46.83 Prime

Data Science from Scratch: First Principles with Python Paperback

› Joel Grus

55

$32.43 Prime

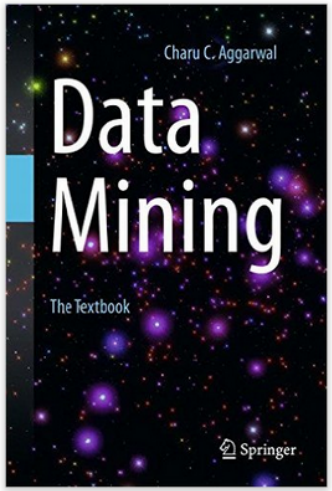
Example: Market basket analysis (Amazon)

Data Mining: The Textbook 2015th Edition

by Charu C. Aggarwal (Author)

8 customer reviews

Look inside



Hardcover $67.52 - $78.04

Other Sellers from $53.34

Buy used $67.52

Buy new **$78.04**

Only 17 left in stock (more on the way).

Ships from and sold by Amazon.com. Gift-wrap available.

List Price: $89.99 Save: $11.95 (13%)

41 New from $53.34

Want it tomorrow, May 19? Order within **4 hrs 3 mins** and choose **One-Day Shipping** at checkout.

Details

FREE Shipping.

Qty: 1

Add to Cart

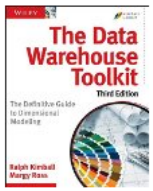
Turn on 1-Click ordering

Ship to:

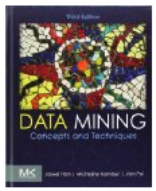
Select a shipping address:

Your Recently Viewed Items and Featured Recommendations

Inspired by your browsing history



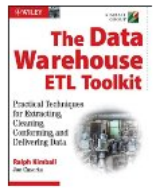
The Data Warehouse Toolkit: The... › Ralph Kimball 62 Paperback $48.87 Prime



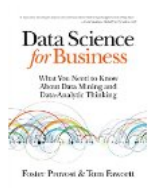
Data Mining: Concepts and Techniques, Third... › Jiawei Han 34 Hardcover $58.74 Prime

Agile Data Warehouse Design: Collaborative... › Lawrence Corr 29 Paperback $31.28 Prime

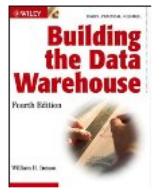
Storytelling with Data: A Data Visualization... › Cole Nussbaumer... 63 Paperback $22.21 Prime



The Data WarehouseETL Toolkit: Practical... › Ralph Kimball 34 Paperback $36.96 Prime



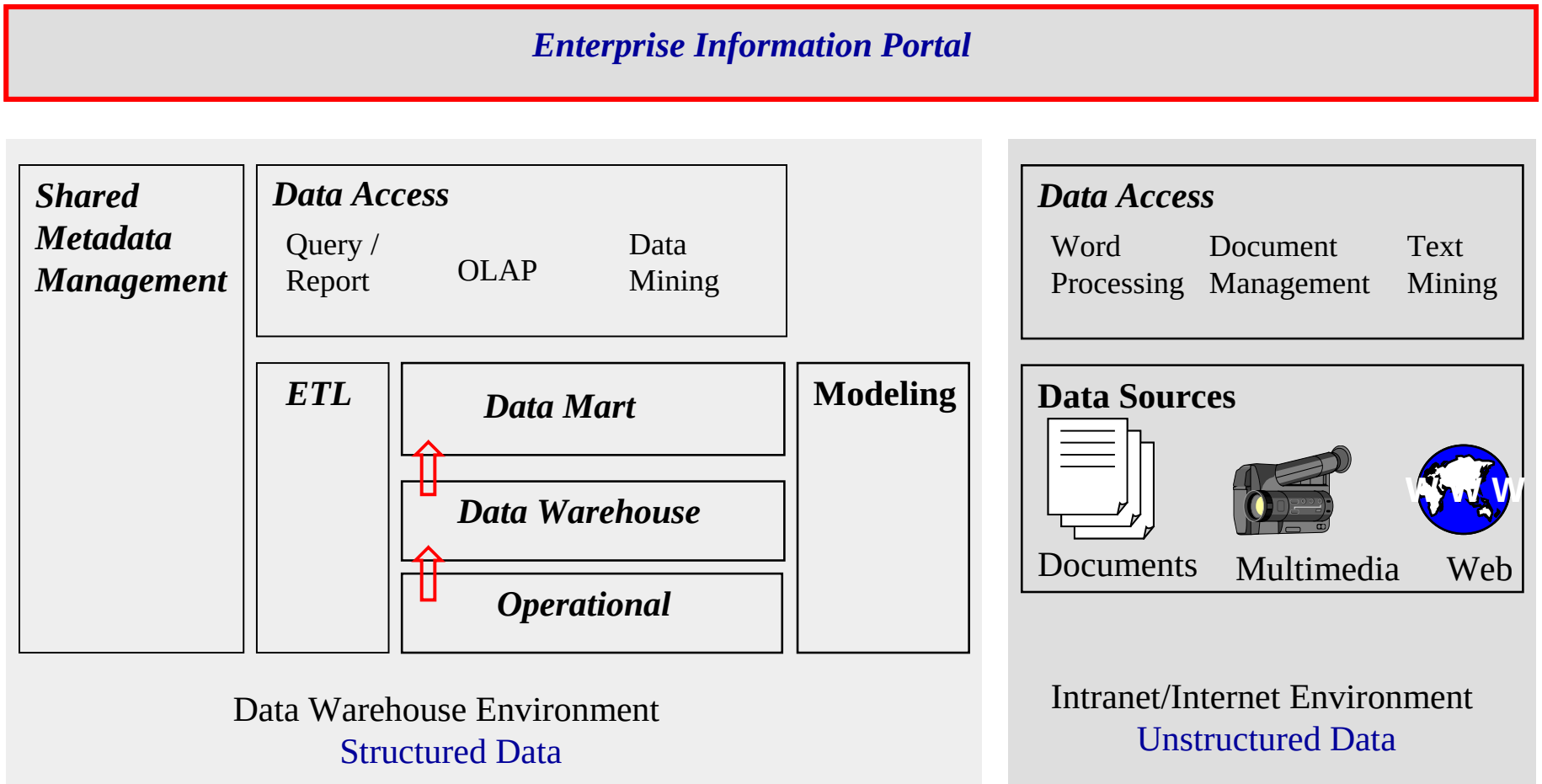
Data Science for Business: What You... › Foster Provost 136 Paperback $37.99 Prime



Building the Data Warehouse W. H. Inmon 11 Paperback $48.05 Prime

Enterprise Information Portals

Uniform and company-wide access to structured and unstructured data



Data Warehouse Hype & Reality

- "Turning data into knowledge"
- "360° view of customer"
- "A single version of the truth"
- "Getting you closer to the customer"
- "Better decision making"

- **Questions:**
  - In which way is the customer data used?
  - How can we guarantee a high degree of data quality?
  - How can we preserve the individual rights of the customers?

Summary

- Data Warehousing: DB query evaluation and analyses on an integrated database for Decision Support (OLAP)
- Huge volume of data
- Main difficulty: Integration of heterogeneous data sets as well as cleaning of primary data (raw data)
- Physical data integration enables complex data preparation activities and efficient analyses on large data sets
- Multi-dimensional modeling and organization of data
- Wide range of methods to evaluate and analyze the given data
- Data Mining: detection of relevant pattern in data
- Data Warehouse is much more than a database