OLAP and Data Warehousing: Concepts and cases

Juan D. Velásquez Silva
PhD. Information Engineering, U. of Tokyo
Post Doctoral Fellow, U. of Oxford
jvelasqu@dii.uchile.cl
http://wi.dii.uchile.cl
Web Intelligence Research Group
Department of Industrial Engineering
University of Chile

Outline

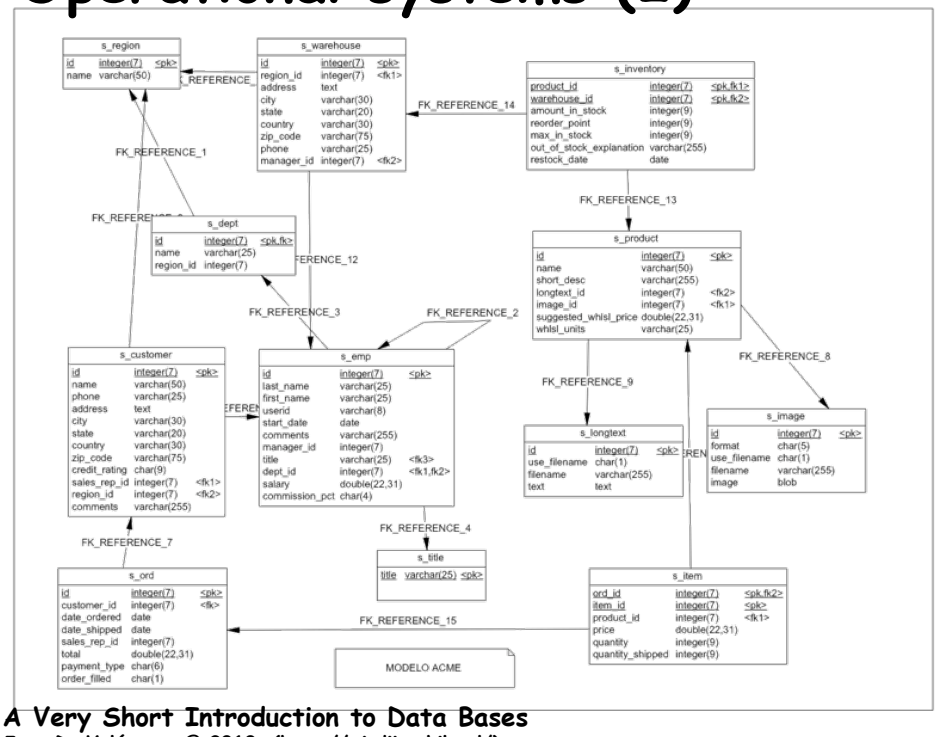
1. Operational Systems.
2. OLAP.
3. Data Warehousing.
4. Data Webhousing.

1.-Operational Systems

Operational systems

- These are the "day-to-day" systems, for instance, the sales system.
- Mainly transaction oriented.
- OLTP (On Line Transaction Process) systems.
- Normally, the data are stored in files or in relational data bases.
- In the last case, the Entity Relational Model is the base for storing the data.

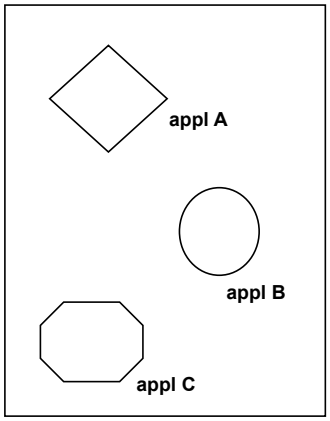
Operational systems (2)



Operational systems (3)

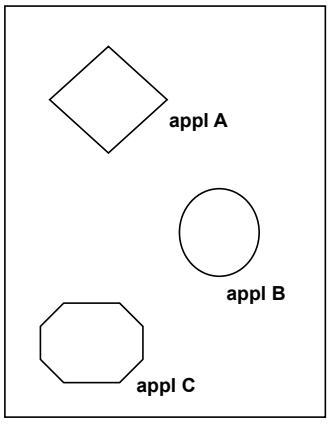
- By using an ad hoc interface, the data base in the OpS. is storage.
- Another possibility is to save the data through an application, such as a batch program.
- The OpS. are the base for extracting information about the business.
- The big problem is "how to get the information without disturb the operational users?"

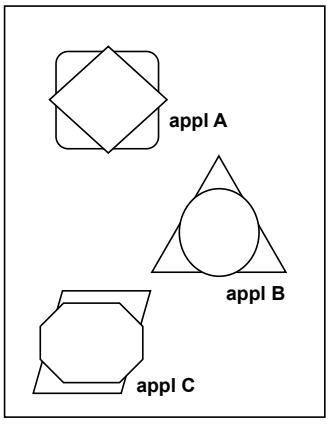
"Progression" of Systems



Applications are designed and developed to meet *specific* business requirements.

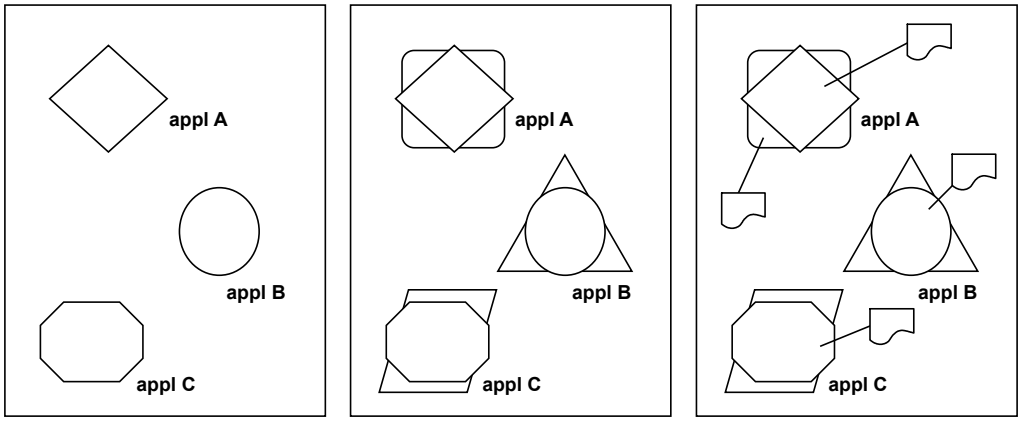
"Progression" of Systems





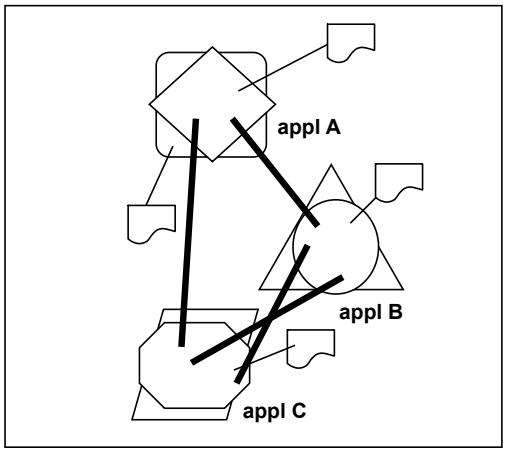
These applications evolve over time to meet new *operational* business requirements.

"Progression" of Systems



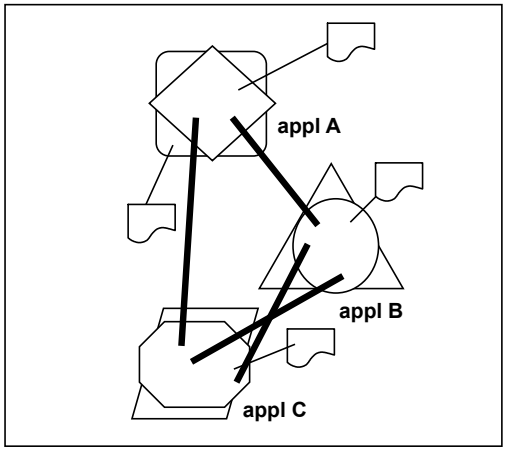
The organization's need for *information* begins to tax the capabilities of the operational systems.

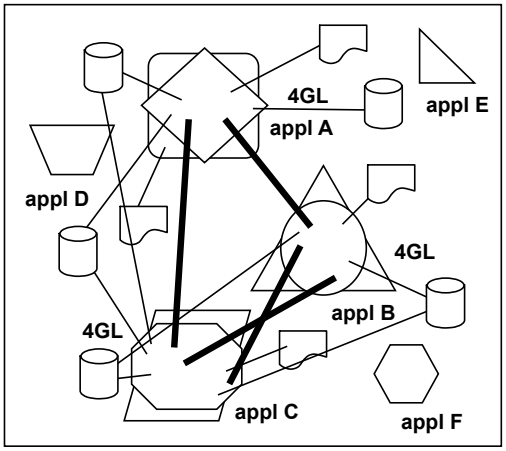
"Progression" of Systems



The need for shared data and the existence of redundant data begins to *distort* the true source.

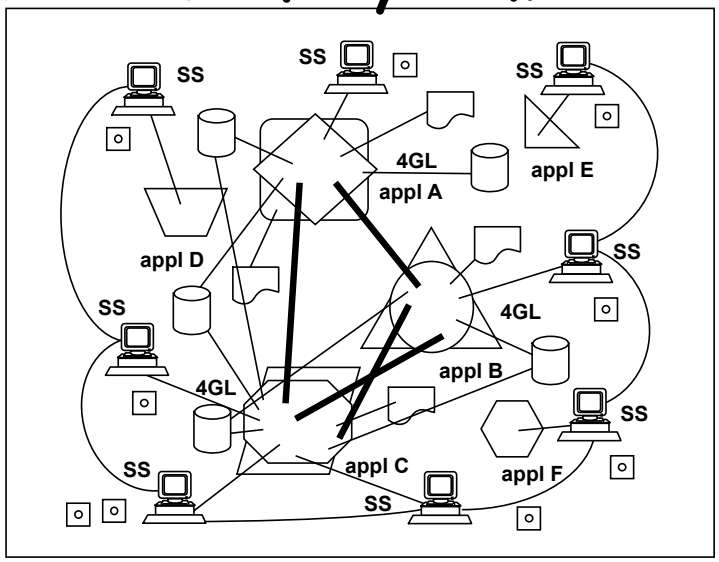
"Progression" of Systems





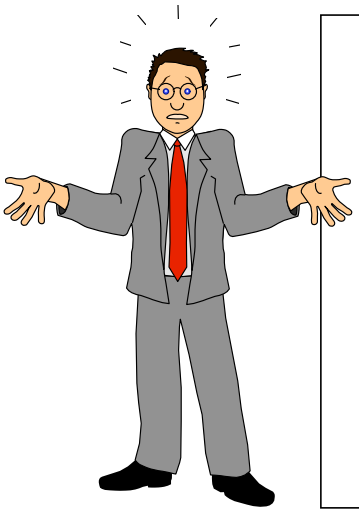
Extracts and stand-alone applications increase the proliferation of *unaudited* data.

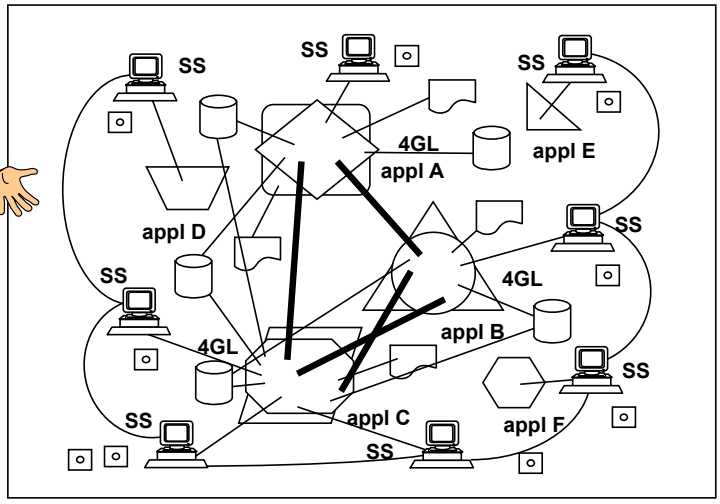
"Progression" of Systems



Powerful end user tools and networks increase the speed and distance that distorted data can cover.

"O.K., Where do I go?"





The typical spider wed of stand-alone, piecemeal applications containing mostly current data was never *architected* for informational processing.

Information systems

- Business users as end users need information to take strategic decisions.
- As most business information requests are recurrent, the logical solution is to package queries into a system that could be operated for the business user.
- Hence corporate business information systems were born.
- With information systems the companies can create client loyalty and solidifying customer relationships.

Information systems (2)

- Information systems need be updated and often changed to newer versions.
- The information needs are a "no ending story".
- Much will depend on the changing needs for information within the company and which in turn depends on the dynamism of the business.
- Then the flexibility is an important characteristic in a information system. However, it is quite difficult to get it.

2.- OLAP

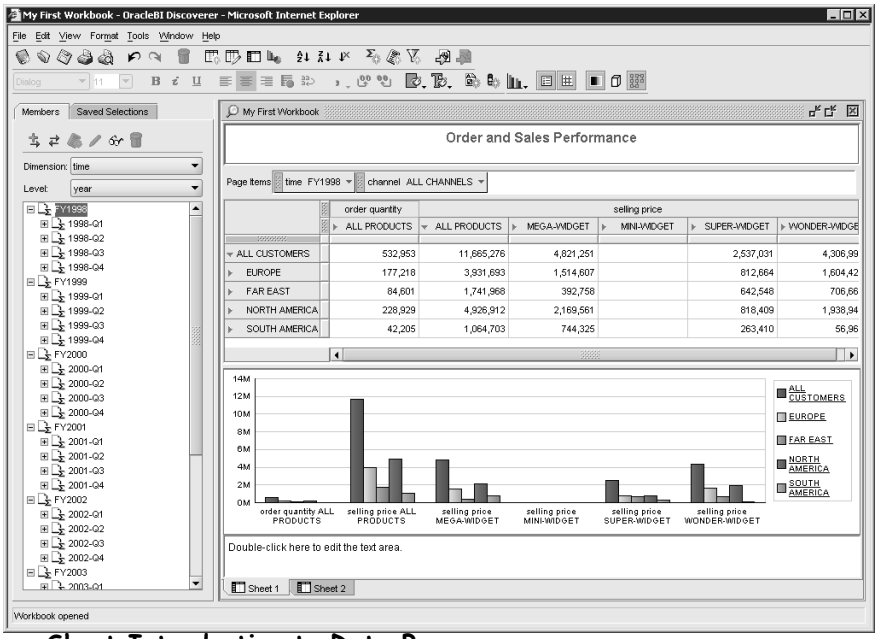
Online Analytical Processing (2)

- In 1993, Dr. E. Codd proposed the concept of for capturing enterprise data.
- Although the definition is sufficiently broad to apply to any data format, it is more convenient store data in files and specify how they are related.
- The concept is related with a multidimentional query, i.e, a query that use multiple data source for structuring the answer.

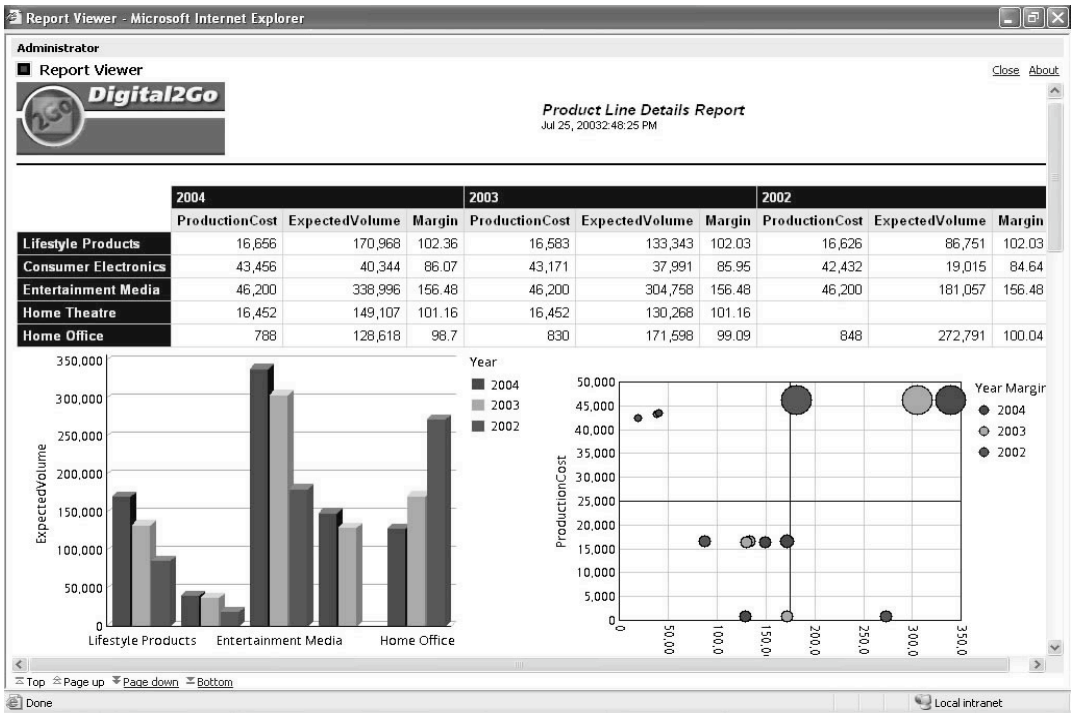
Online Analytical Processing (3)

- In order to facilitate the user's work, the OLAP tools need to support the following functionalities:
  - Querying. Ability to pose powerful ad-hoc queries through a simple and declarative interface.
  - Restructuring. Ability to restructure information in a multidimensional database, exploiting the dimensionality of data and bringing out different data perspectives.
  - Classification. Ability to classify or group data sets in a manner appropriate for subsequent summarization.
  - Summarization/Consolidation. This is a generation of the aggregate operators in standard SQL. In general, summarization maps multisets of values of a numeric type to a single consolidated value.

Tools: Oracle Discoverer



Tools: Cognos ReportNet



3.- Data warehouse architecture

Definition

"Subject oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process" [Inmon96]

Definition (2)

The Data Warehouse is an *architecture*, not a technology.

Multidimensional analysis

- The end users think the world in several dimensions, i.e., they associate to a fact a set of dependent dimensions.
- For instance, a sale indicator depend of the market place, the price, time and others elements.
- Then the business users make questions like "what were the sales of pencils in the region V during the first semester, 2001".
- In others words, they think the world multi-dimensionally [Teste01].

A simple sale business report

SALES			TIME													
			Year	2000				2001				2002				
			Semester	1		2		1		2		1		2		
			Month	Jan	…	Jul	…	Jan	…	Jul	…	Jan	…	Jul	…	
PRODUCT	Type	City	(Cost,Sale)													
	Rice	Tokyo		(4,5)	…	(3,2)	…	(3,4)	…	(5,6)	…	(6,7)	…	(2,3)	…	
		Osaka		(4,2)	…	(3,1)	…	(3,6)	…	(5,2)	…	(6,3)	…	(2,4)	…	
		⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	Fish	Tokyo														
		Osaka														
		⋮														
	Egg	Tokyo														
		Osaka														
		⋮														
	Meat	Tokyo														
		Osaka														
		⋮														

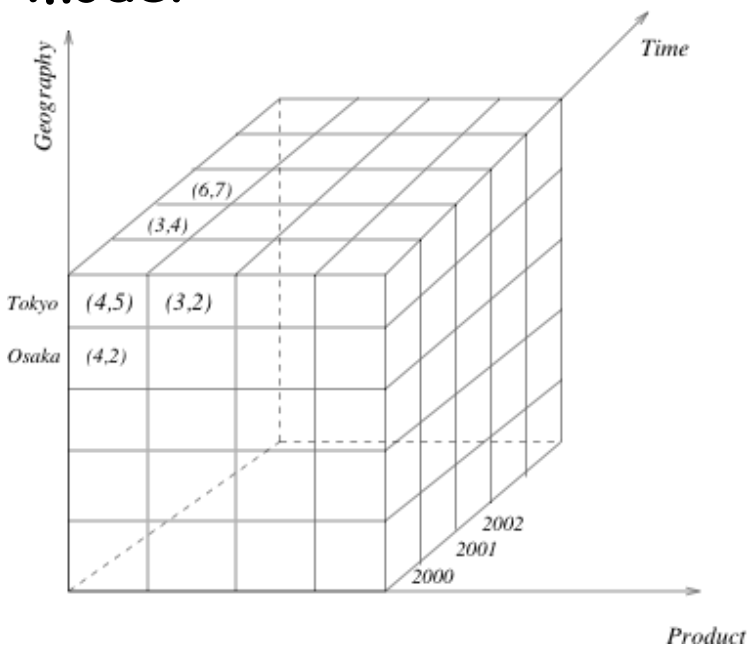
The need

- Applying tools, the end user can create and manipulate a report.
- In essence, it contains the information from several sources.
- The problem now is how to represent the data sources to facilitate the user manipulation to create the report.

Multidimensional Data Model

- Defines a way to store the information for the next step, i.e., the report creation or the application of a pattern extraction tool [Agrawal97].
- The MDM model must provide the basic information to satisfy the end user queries in different aggregation levels.
- The model's granularity is the minimum information to store in order to answer any information query that the end user realize on the data.
- Then the grain is the core of the MDM model and from this definition, different aggregations levels or hierarchies can be created.
- Using the above report, a grain expression would be "*the price and cost of a product sold*".

Cube model



Cube model (2)

- It is implemented in a Multidimensional Data Base Manager System (MDBMS).
- Some cube operations:
  - Pivoting. Rotate the cube and show a particular face.
  - Slicing. Select one dimension of the cube.
  - Dicing. Select one or more dimension of the cube.
  - Drill-down. Show the details of aggregation point.
  - Roll-up. The inverse operation to the previous point.

Cube model (3)

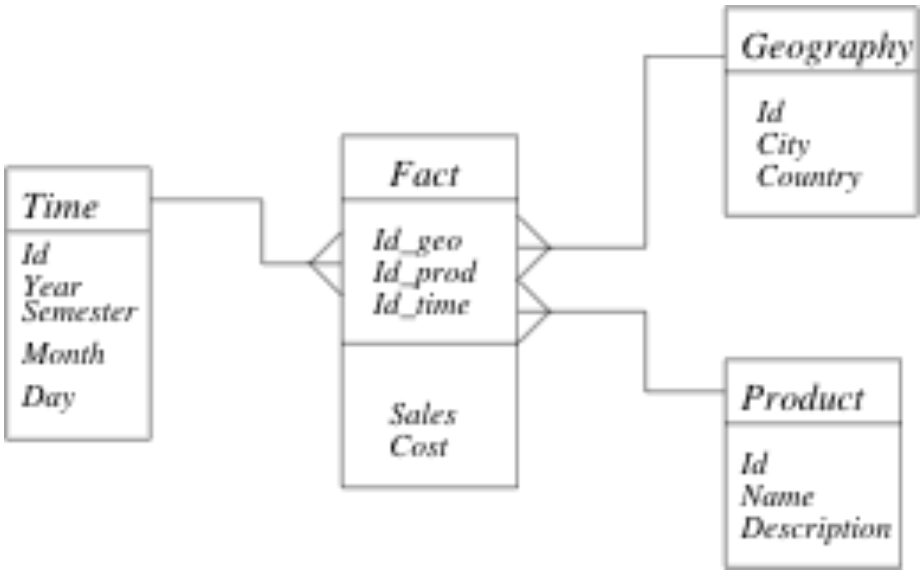
- The computational cube representation uses multidimensional arrays.

$$a : array \quad 1...\alpha_1,....1...\alpha_n \quad of \quad < TYPE >$$

- *Which was the total cost of the product code 200 acquired in Tokyo branch for the year 2004?*

Cube[Geography.city=Tokyo,Time.Year=2004,Product.code=200].cost➔

Star Model



Star model (2)

- It is implemented in a Relational Data Base Manager System (RDBMS).
- However, while ER imposes a strict normalization, it is inappropriate for the star model.
- Because the RDBMS is more popular, the star model has prevailed becomes almost standard in the data warehouse world.
- The star model consists in a single fact table and a set of dimensional tables. From the ER view point, the model shows a master-detail relation.

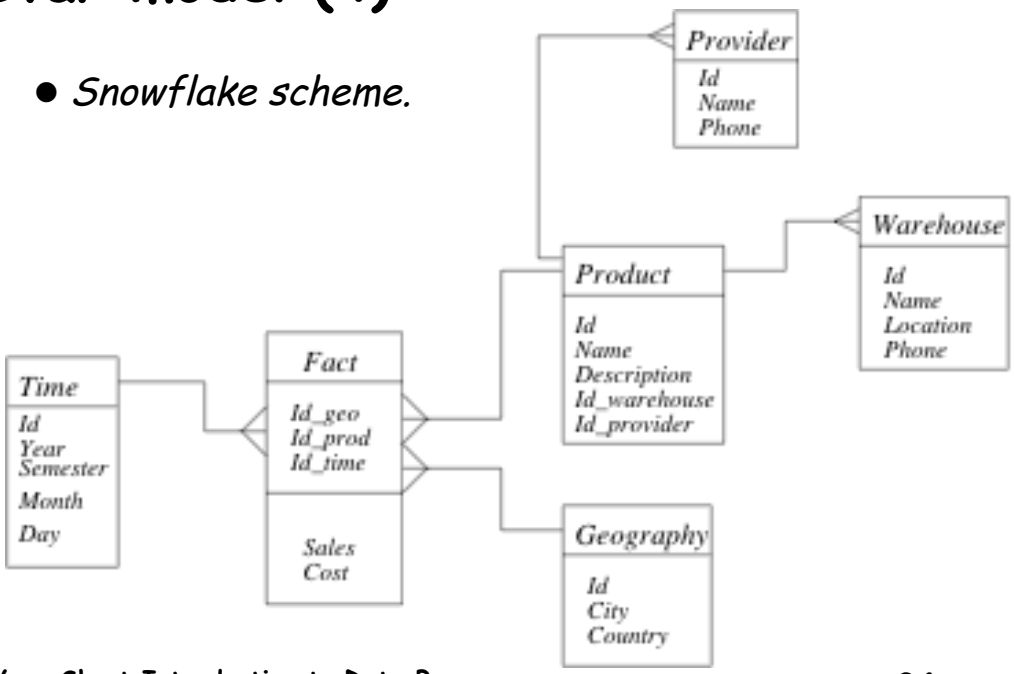
Star model (3)

- *Which was the total cost of the product code 200 acquired in Tokyo branch for the year 2004?*

```
select cost
from fact, geography, time, product
where time.year=2004 and geography.city=Tokyo and
product.code=200 /* start join */
and  fact.id_geo=geography.id and fact.id_time=time.id
and fact.id_prod=product.id
```


Star model (4)

- *Snowflake scheme.*



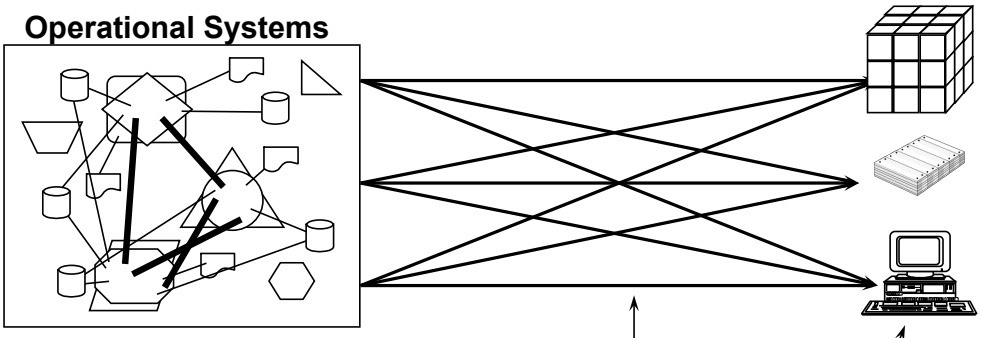
The Extraction, Transformation and Loading (ETL) process

- This process groups the techniques and tools used in the data extraction from several sources, the data transformation invaluable information, and its loading in a repository. [Golfarelli99].
- The ETL, concentrate a huge effort in the construction of the data warehouse. It is a non-trivial process because the data has different origin, type, format, etc

ETL (2)

- Extraction. It define the data source for information extraction.
- Transformation. The data are transformed into information in a Data Staging Area.
- Loading. The information is storage in an information repository.

Impact of the DW on Acquisition & Access



- Multiple extracts per source system
- Data acquisition & access limiting productivity
- No assurance of data accuracy

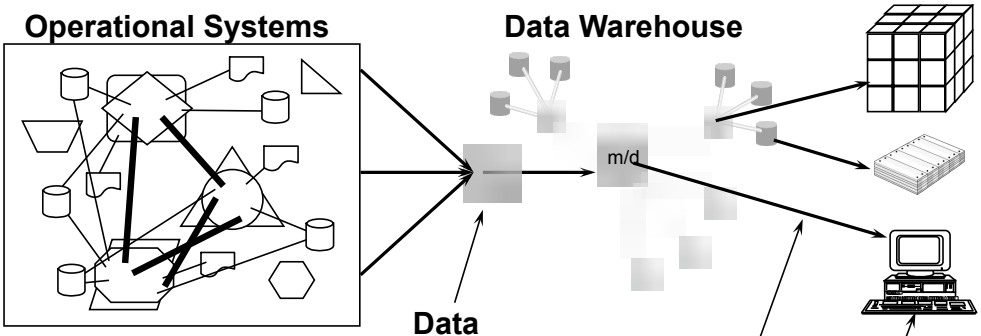
Data Acquisition & Access

80%

Data Analysis

20%

Impact of the DW on Acquisition & Access



- Single extract per operational source
- Metadata ensures accuracy
- Overall productivity triples!

Data Acquisition

20%

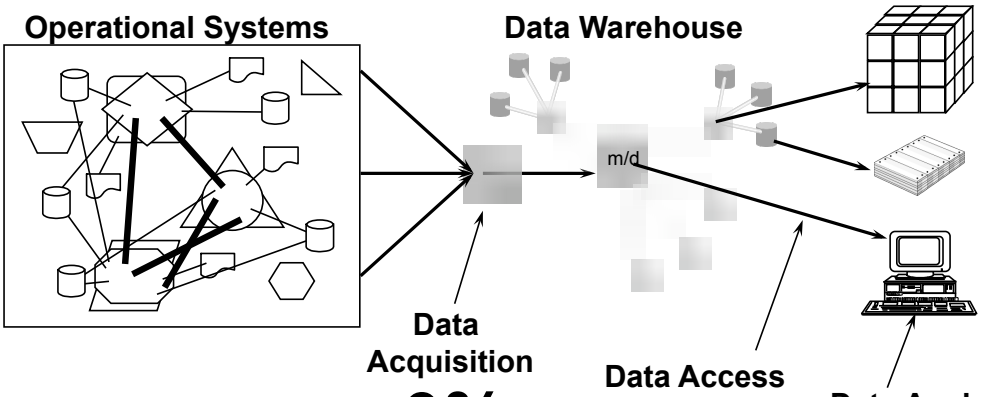
Data Access

20%

Data Analysis

60%

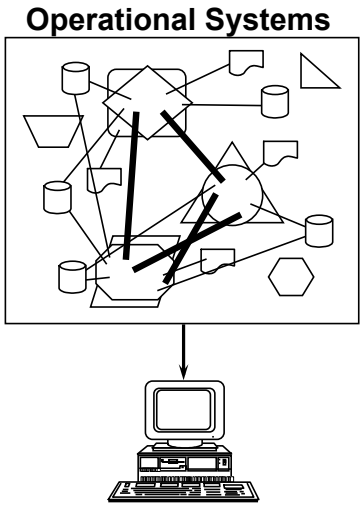
Impact of the DW on Acquisition & Access



0% **20%** **80%**

- Single extract per operational source
- Metadata ensures accuracy
- Analyst productivity quadruples!

Impact of the DW on Acquisition & Access



Compare these costs to the data warehouse environment

Total the costs for the:

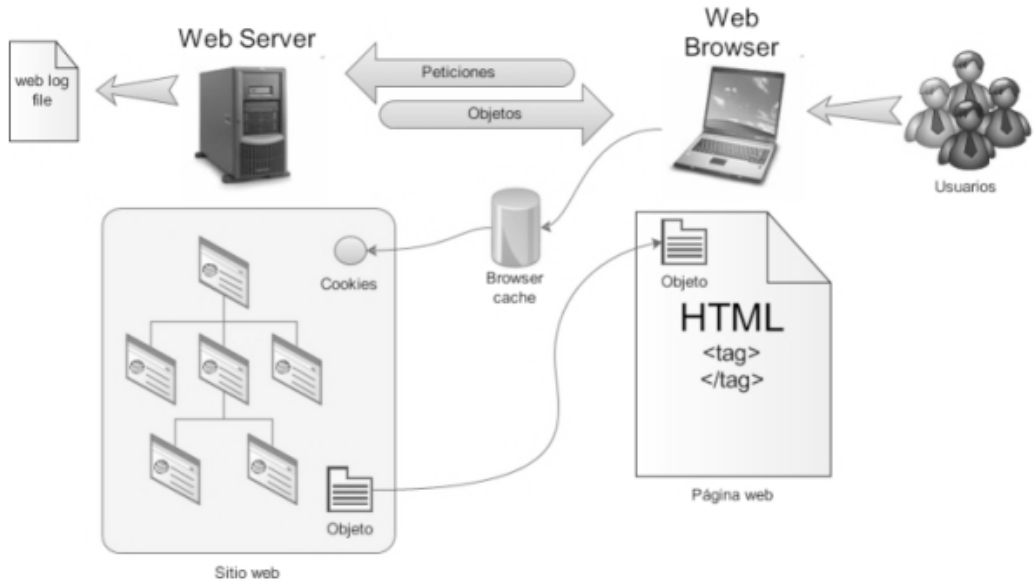
- ▲CPU time required for multiple extract processing
- ▲Disk storage required for multiple extract processing and storage
- ▲Number of extract programs currently maintained
- ▲Number of extract programs required/ planned
- ▲Cost of unmanaged redundant data
- ▲DBMS/system software licenses
- ▲Time spent by analysts satisfying informational requests

4.- Webhouse

The web data

- The Web is a primary source of data to analyze the user behavior in a web site [Thor04].
- The data are transformed into information, like indicators about how many users have visited the web site per month.
- This information is useful for the web user maintainers, i.e., the persons responsible for the web site structure and content modifications.
- Here the web manager, commercial areas, marketing department, etc., can be potential information's users, over all if the web site is the core business of the institution, situation so frequent today.

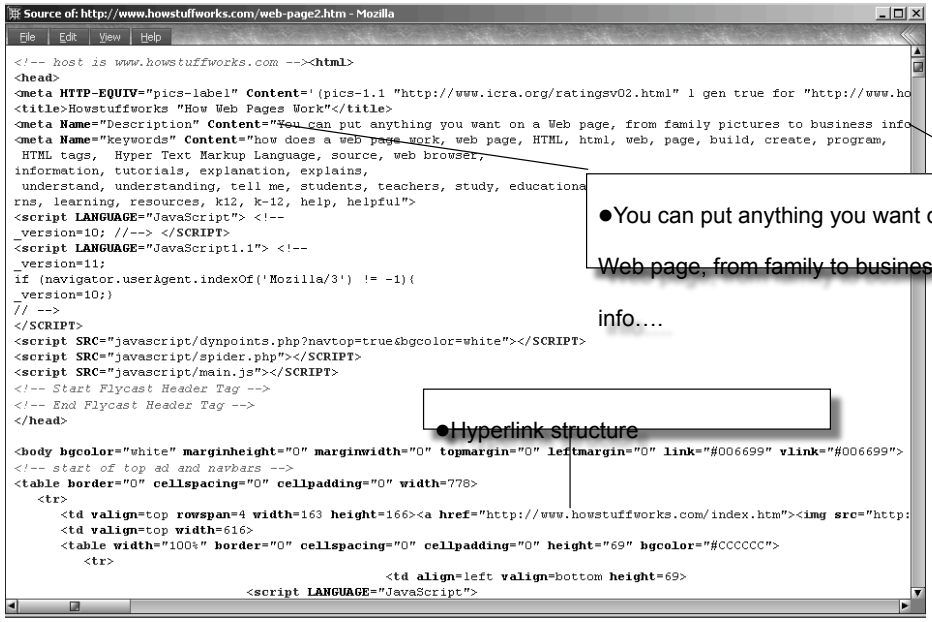
Web server and web browser



Web logs

#	IP	Id	Acces	Time	Method/URL/Protocol	Status	Bytes	Referer	Agent
1	165.182.168.101	-	-	16/06/2002:16:24:06	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
2	165.182.168.101	-	-	16/06/2002:16:24:10	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
3	165.182.168.101	-	-	16/06/2002:16:24:57	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
4	204.231.180.195	-	-	16/06/2002:16:32:06	GET p3.htm HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
5	204.231.180.195	-	-	16/06/2002:16:32:20	GET C.gif HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
6	204.231.180.195	-	-	16/06/2002:16:34:10	GET p1.htm HTTP/1.1	200	3821	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
7	204.231.180.195	-	-	16/06/2002:16:34:31	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
8	204.231.180.195	-	-	16/06/2002:16:34:53	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
9	204.231.180.195	-	-	16/06/2002:16:38:40	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
10	165.182.168.101	-	-	16/06/2002:16:39:02	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
11	165.182.168.101	-	-	16/06/2002:16:39:15	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
12	165.182.168.101	-	-	16/06/2002:16:39:45	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
13	165.182.168.101	-	-	16/06/2002:16:39:58	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
14	165.182.168.101	-	-	16/06/2002:16:42:03	GET p3.htm HTTP/1.1	200	4036	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
15	165.182.168.101	-	-	16/06/2002:16:42:07	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
16	165.182.168.101	-	-	16/06/2002:16:42:08	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
17	204.231.180.195	-	-	16/06/2002:17:34:20	GET p3.htm HTTP/1.1	200	2342	out.htm	Mozilla/4.0 (MSIE 6.0; Win98)
18	204.231.180.195	-	-	16/06/2002:17:34:48	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 6.0; Win98)
19	204.231.180.195	-	-	16/06/2002:17:35:45	GET p4.htm HTTP/1.1	200	3523	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
20	204.231.180.195	-	-	16/06/2002:17:35:56	GET D.gif HTTP/1.1	200	3231	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)
21	204.231.180.195	-	-	16/06/2002:17:36:06	GET E.gif HTTP/1.1	404	0	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)

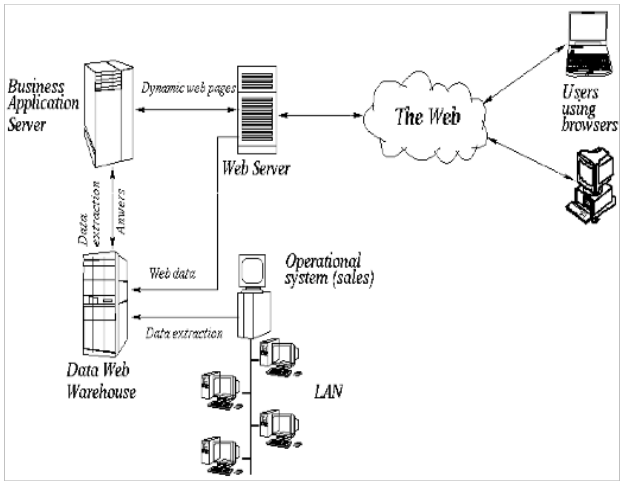
Web site contents



Web warehousing

- This is the name that receive the application of the data warehouse architecture on data originated in the web [Bhowmick98].
- The web data warehouse, or simply **webhouse**, was introduced by Kimball [Kimball00] as the solution to storage everything concerning with the cleackstreamin the web site.

Web warehousing (2)



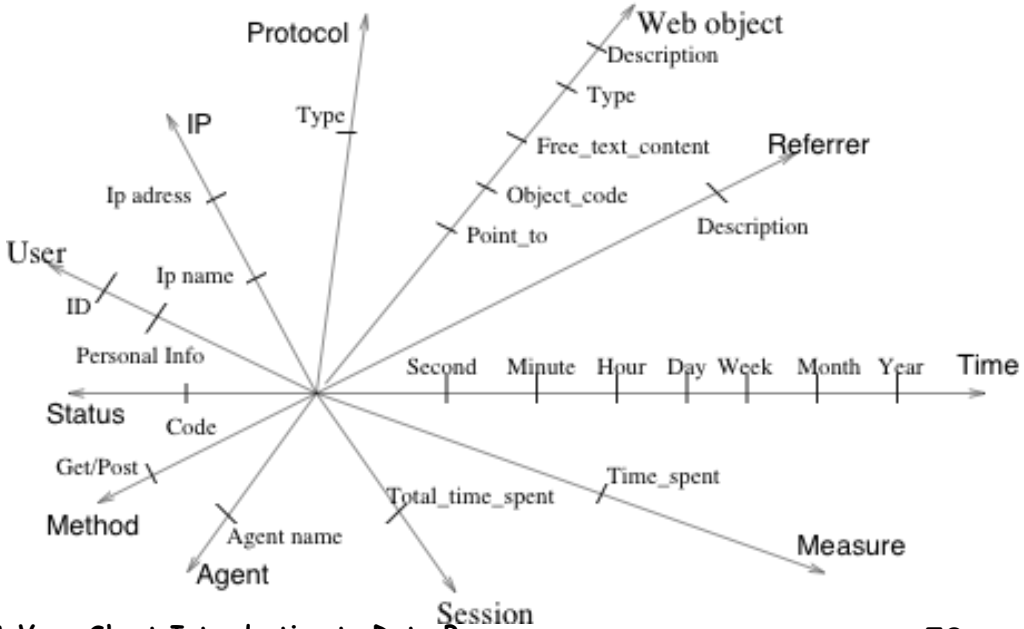
Information repository for web data

- Two objectives: to be the primary information source for end users and the input source for web mining algorithms.
- Due to the complexity of web data, these need special preprocessing tasks in order to clean and transform them in feature vectors to be storage in a information repository.
- The webhouse architecture shows a way to create a web information repository (WIR) to analyze the user behavior in a web site.

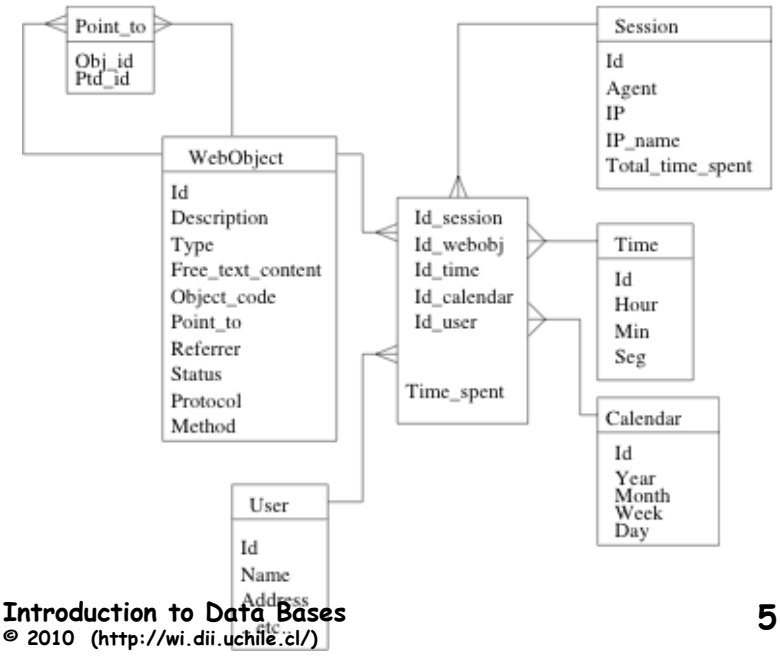
Thinking the web data in several dimensions

- Which end user questions are necessary to answer?
- Some examples:
  - Which is the most visited page in a period?
  - Time spent by page per session?
  - Bytes transmitted by session?
  - Session average (in pages visited and time spent)?
  - Amount of visit per page in a period?
- The grain "*the spent time per web object visited*"

A generic MDM for web data [Velasquez03]



A generic star model for web data [Velasquez03]



ETL process applied on web data

- Data sources: web objects with text content, web hyperlinks inner structure and web logs of a particular web site.
- Extraction.
  - In the case of the web pages, the text is extracted directly from the page content.
  - The hyperlink structure also is extracted from the web page content (href tag)
  - The web logs are extracted from the respective files.

ETL process applied on web data (2)

- Transformation:
  - The page transformation is by applying the vector space model. This task is performed by a code in a language in the S.O.
  - The hyperlink structure don't need transformation in this case.
  - The web logs need a session reconstruction task (sessionization), which is performed using the RDBMS capabilities. Then the logs are storage in tables before to be transformed.

ETL (3) : DSA example for web logs

Weblogs	
Ip	varchar2(18)
TimeStamp	date
Method	varchar2(20)
Status	numer(4)
Bytes	number(8)
Url	varchar2(20)
Agent	varchar2(20)

Logclean	
Ip	varchar2(18)
TimeStamp	date
Bytes	number(8)
Url	number(4)
Agent	number(2)
Session	number(4)
Timespent	number(4)

ETL (4)

- Loading. Finally by using a code the data are storage in the WIR.
- In the case of web logs sessionizated, it is more simple, because these are in tables of the RDBMS

5.- Summary

Summary

- The KDD process shows the needs to consolidate and maintain the information extracted from the data sources.
- In that sense, the information repository construction seems be the logical solution, allowing further retrospective analysis and also the application of advanced data processing techniques as datamining algorithms.
- The problem here is how understand the end user information requirements?.

Summary (2)

- In the multidimensional data analysis, we find a methodology for structuring the end user information requirements, by using the cube or the star models.
- An evolution of the data warehouse is the web data warehouse architecture, in short webhouse.
- By using the webhouse architecture and the star model, a Web Information Repository (WIR) was defined.
- The WIR is destined to be the main information platform for web mining algorithms and end user queries.

References

- [Agrawal97] R. Agrawal, A. Gupta and S. Sarawagi, *Modeling Multidimensional Databases*, Procs. 13th Int. Conf. Data Engineering ICDE, pages 232-243, 1997.
- [Bhowmick98] S.S. Bhowmick, S.K. Madria, W.K. Ng and E.P. Lim, *Web Warehousing: Design and Issues*, Procs. Int. of ER Workshops, pages 93-104, 1998.
- [Frawley92], W.J. Frawley, G. Piatetsky-Shapiro and C.J. *Matheus, Knowledge Discovery in Databases: An Overview*, AI Magazine, pages 57-70, 1992.
- [Golfarelli99] M. Golfarelli and S. Rizzi, Designing the data warehouse: key steps and crucial issues, *Journal of Computer Science and Information Management, 2(1)*:1-14, 1999.
- [Kimball00] R. Kimball and R. Merx, The Data Webhouse Toolkit, Wiley Computer Publisher, New York, 2000.

References (2)

- [Inmon96] W. H. Inmon, *Building the data warehouse (2nd ed.)*, John Wiley and Sons, New York, 1996.
- [Thor04] A. Thor, N. Golovin and E. Rahm, *AWESOME - A Data Warehouse-based System for Adaptive Website Recommendations*, In Proc. of Int. Conf VLDB, pages 384-395, 2004.
- [Teste01] O. Teste, Towards Conceptual Multidimensional Design in Decision Support Systems, Fifth East-European Conf. on Advances in Databases and Information Systems, pages 25-28, 2001.
- [Velasquez03] J. D. Velásquez et al, *A generic Data Mart architecture to support Web mining*, Procs. 4th Int. Conf. on Data Mining, pages 389-399, 2003.