

Data Warehouse Designing: Dimensional Modelling and E-R Modelling

Geetika Saxena¹, Bharat Bhushan Agarwal²
COMPUTER SCIENCE DEPARTMENT, IFTM UNIVERSITY, MORADABAD

Abstract: *The Data Warehouse (DW) is considered as a collection of integrated, detailed, historical data, collected from different sources. DW is used to collect data designed to support management decision making. There are so many approaches in designing a data warehouse both in conceptual and logical design phases. The conceptual design approaches are dimensional fact model, multidimensional E/R model, starER model and object-oriented multidimensional model. And the logical design approaches are flat schema, star schema, fact constellation schema, galaxy schema and snowflake schema. In this paper we have focused on comparison of Dimensional Modelling AND E-R modelling in the Data Warehouse. Dimensional Modelling (DM) is most popular technique in data warehousing. In DM a model of tables and relations is used to optimize decision support query performance in relational databases. And conventional E-R models are used to remove redundancy in the data model, facilitate retrieval of individual records having certain critical identifiers, and optimize On-line Transaction Processing (OLTP) performance.*

Keywords: *Data Warehouse, DM Models, E-R Models, flat schema, star schema, fact constellation schema, galaxy schema, snowflake schema.*

I. Introduction

Information is an asset for the organization or enterprise (resource like capital, first matters, plants and people) which is used to provide benefit and competitive advantage to any organization. Hence, understanding the value of information is important. Today, every organization have a relational database management system that is used for organization's daily operations. The organization wants to increase the value of their organizational data by making it accessible information.

Organizations usually complaints,

“We have tons of data but we cannot access them!”

“Show me only what is important!”

As the amount of the organizational data increases, it becomes harder to access and get the most information out of it, because it is in different formats, exists on different platforms and resides on different structures. Data warehousing provides an excellent approach in converting operational data into useful, accessible and reliable information to support the decision making process and also provides the basis for data analysis techniques like data mining and multidimensional analysis. Data mining is the process of identifying valid, novel, useful and understandable patterns in data. Data Mining is also known as KDD (Knowledge Discovery in Databases).

II. Data Warehouse Concepts

2.1. DEFINITION OF DATA WAREHOUSE

The data warehouse always contains data and information, on which management decisions can be reliably tested, analyzed, assessed and monitored using the data and information integration.

“A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.”—W. H. Inmon [1, 2, 3, 6, 10, 11].

- **Subject Oriented:** Data that gives information about a particular subject instead of about a company's ongoing operations.
- **Integrated:** Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- **Time-variant:** All data in the data warehouse is identified with a particular time period.
- **Non-volatile:** Data are not updated or changed in any way once they enter the data warehouse, but are only loaded and accessed.

III. Designing A Data Warehouse

3.1 CONCEPTUAL DESIGN MODELS

The conceptual design allows having closer ideas about the ways that a user can perceive an application domain. In fact, it is considered as a key step that ensures the successful of the DW projects since it defines the expressivity of the multidimensional schemata, and the result of this step is a graphical notation which facilitates to the designer and the user different tasks such as writing, understanding and managing the conceptual schemata.

The main goal of conceptual design modelling is developing a formal, complete, abstract design based on the user requirements [34].

At this phase of a DW there is the need to:

Represent facts and their properties

Represent objects and capture their properties with the associations among them:

Record the associations between objects and facts:

Distinguish dimensions and categorize them into hierarchies

3.1.1. THE DIMENSIONAL FACT MODEL

This model is built from ER schemas [9, 15, 16, 17, 33]. The Dimensional Fact (DF) Model is a collection of tree structured fact schemas whose elements are facts, attributes, dimensions and hierarchies. A fact schema is structured as a tree whose root is a fact. The fact is represented by a box. Sub-trees rooted in dimensions are hierarchies. The circles represent the attributes and the arcs represent relationship between attribute pairs.

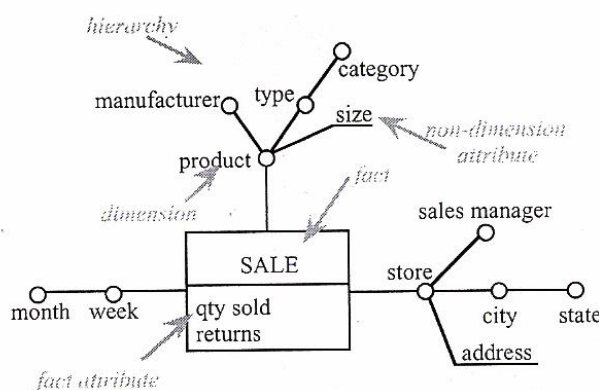


Fig.1 A dimensional fact schema

3.1.2. MULTIDIMENSIONAL E/R MODEL

Multidimensional E/R (ME/R) model includes some key points [14]:

- Specialization of the ER Model
- Minimal extension of the ER Model; this model should be easy to learn and use for an experienced ER Modeler.

This model allows the generalization concepts. There are some specializations:

- A special entity set: dimension level
- Two special relationship sets connecting dimension levels:
 - a special n-ary relationship set: the 'fact' relationship set
 - a special binary relationship set: the 'roll-up to' relationship set

The "roll-up to" relationship set; it relates a dimension level A to a dimension level B representing concepts of a higher level of abstraction (city roll-up to country).

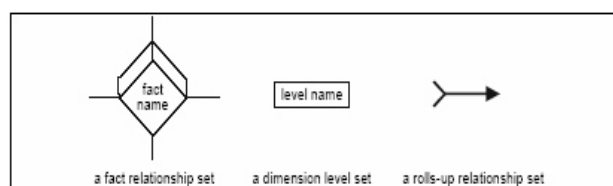


Fig.2. notation of Multidimensional E/R elements

3.1.3. STARER

This model combines star structure with constructs of ER model [13]. The starER contains facts, entities, relationships and attributes. This model has the following constructs:

- Fact set
- Entity set
- Relationship set

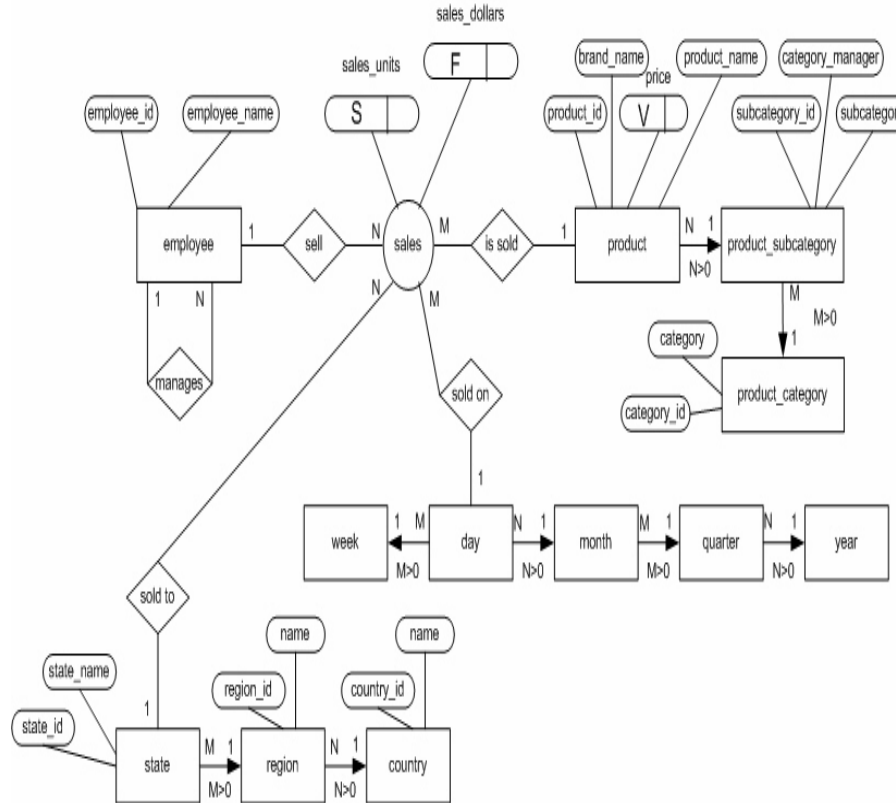


Fig.3. StarER Model

3.2 LOGICAL DESIGN MODELS

Dimensionality modelling uses the ER Modelling with some important restrictions. Dimensional model composed of one table with a composite primary key, called fact table, and a set of smaller tables called dimension tables. Each dimension table has a simple (non-composite) primary key that corresponds exactly to one of the components of the composite key in the fact table. This characteristic structure is called star schema or star join.

There are following models used in dimensional modelling.

3.2.1 DIMENSIONAL MODEL DESIGN

Dimensional Model is used by OLTP systems. It contains no redundancy, but high efficiency of updates, shows all data and relationships between them. Simple queries require multiple table joins and complex subqueries. It is suitable for technical specialist.

3.2.1.1 FLAT SCHEMA

This schema is the very simple schema. This is formed by collapsing all entities in the data model down into the minimal entities. This minimizes the number of tables in the database and joins in the queries. We end up with one table for each minimal entity in the original data model [12]. This structure does not lose information from the original data model. It contains redundancy, in the form of transitive and partial dependencies, but does not involve any aggregation. It contains some problems; first it may lead to aggregation errors when there are hierarchical relationships between transaction entities. When we collapse numerical amounts from higher level transaction entities in to other they will be repeated. Second this schema contains large number of attributes. Therefore while the number of tables (system complexity) is minimized, the complexity of each table (element complexity) is increased.

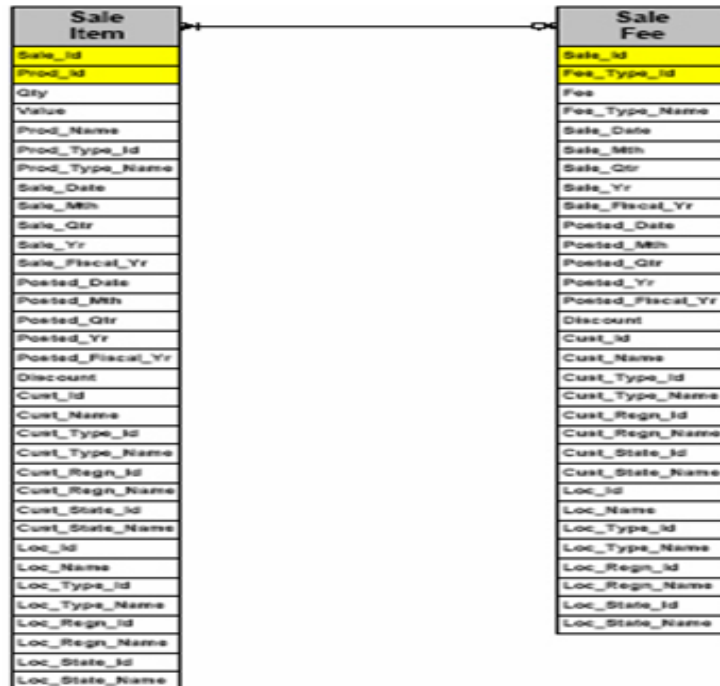


Fig. 4. Flat Schema

3.2.1.2 STAR SCHEMA

It is the basic structure for a dimensional model. A fact table in the middle connected to a set of dimension tables

It contains:

- A large central table (fact table)
- A set of smaller attendant tables (dimension table), one for each dimension

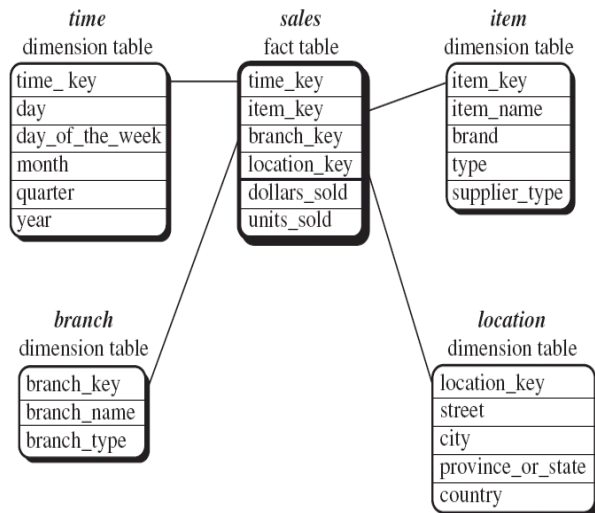


Fig. 5. Star Schema

3.2.1.3 SNOWFLAKE SCHEMA

A refinement of star schema where some dimensional hierarchy is further splitting (normalized) into a set of smaller dimension tables, forming a shape similar to snowflake

- However, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed.
- Snowflake schema is an extension of star schema in a way; it separates itself from Star when it comes to handling large dimension tables. A star schema focuses on a centralized design with a fact table in it connecting to different dimension tables end to end.

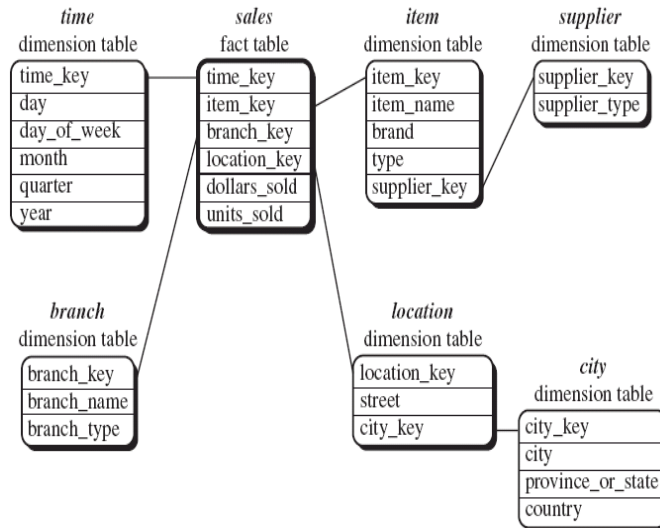


Fig. 6. Snowflake Schema

3.2.1.4 FACT CONSTELLATION SCHEMA

Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation.

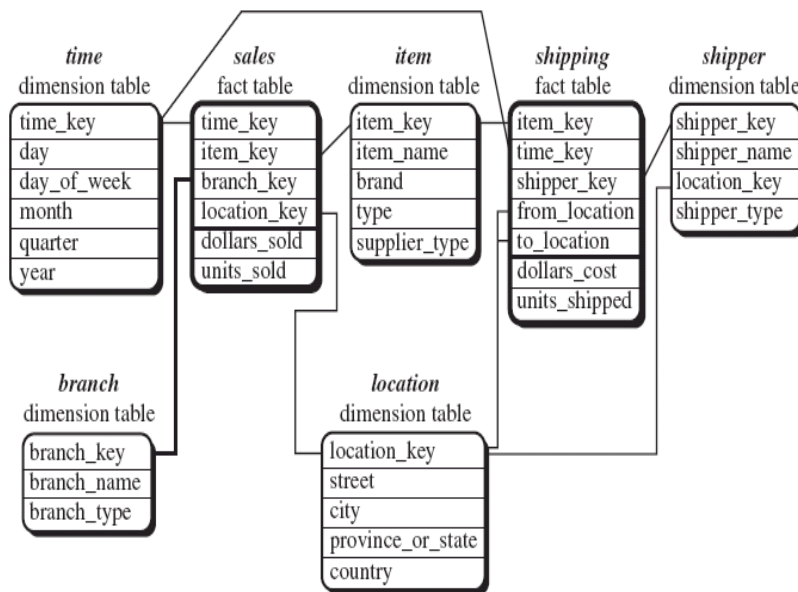


Fig. 7. Fact constellations

3.2.1.5 GALAXY SCHEMA

Galaxy schema is a schema where multiple fact tables share dimension tables. Unlike a fact constellation schema, the fact tables in a galaxy do not need to be directly related [12]. The following figure, Fig 5, illustrates a sample of a galaxy schema.

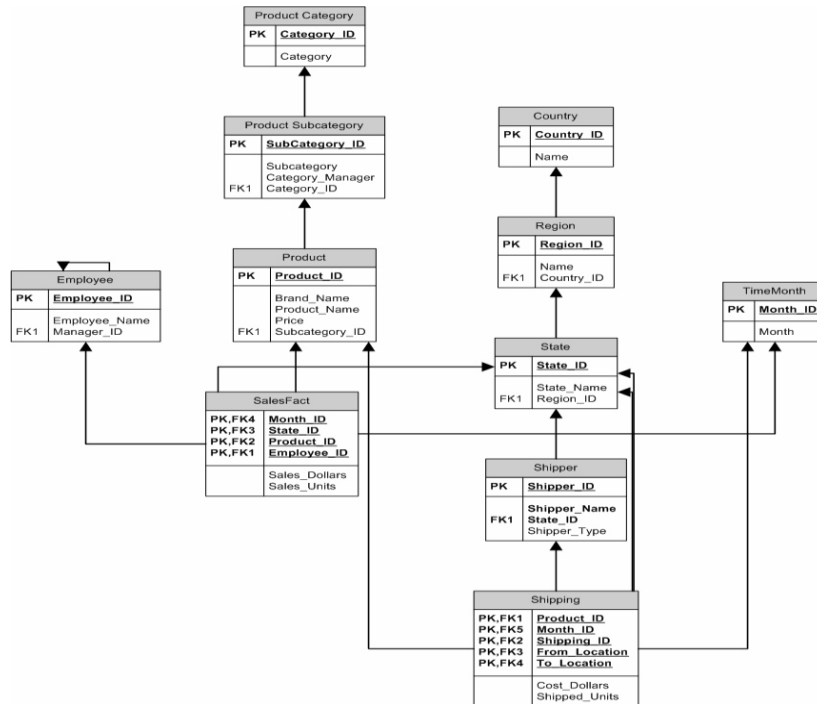


Fig.8. Galaxy Schema

IV. Comparison Of Dimensional Models And Er Model

ER modelling

The main goal of ER modelling is to remove redundancy from data. To remove redundancy, designers must use hundreds of entities and relations between entities, which makes ER model complex. There is no easy way to enable end users navigate through the data in ER models.

ER modelling aims to optimize performance for transaction processing. It is also hard to query ER models because of the complexity; many tables should be joined to obtain a result set. Therefore ER models are not suitable for high performance retrieval of data.

Dimensional model

The dimensional model is a standard framework. End user tools can make strong assumptions about the dimensional model. It helps, to make user interfaces more user friendly and processing more efficient [20].

Dimensional model is more useful to random changes in user behaviour and requirements. The logical design can be made independent of expected query patterns. All dimensions can be thought as symmetrically equal entry points into the fact table. Dimensional model is extensible to new design decisions and data elements. All existing fact and dimension tables can be changed in place without having to reload data. End user query and reporting tools are not affected by the change.

Dimensional model involves business rules but ER modelling does not involve business rules, it involves data rules.

V. Conclusion

In this paper we focused on comparison of Dimensional Modelling AND E-R modelling in the Data Warehouse. We have discussed Conceptual Design Models and all types of Logical Design Models. The conceptual design approaches are dimensional fact model, multidimensional E/R model, starER model and object-oriented multidimensional model. And the logical design approaches are flat schema, star schema, fact constellation schema, galaxy schema and snowflake schema. We got that Dimensionality modelling uses the ER Modelling with some important restrictions like Dimensional model composed of one table with a composite primary key, called fact table, and a set of smaller tables called dimension tables.

According to my research, there is no complete study in literature on DW models providing a mapping of models to development phases and giving a comparison of the models according to these phases. Also, there are very few articles for covering all phases of data warehousing.

VI. Future Work

One future work may be implementing a more complex case study using real world application data, perform performance tests using the three logical models compared to support the comparison on logical design models.

Another future work may be by adding more quality factor in the comparison we can improve comparison of logical design models.

REFERENCES

- [1] Romm M., Introduction to Data Warehousing, San Diego SQL User Group
- [2] Goyal N., Introduction to Data Warehousing, BITS, Pilani Lecture Notes
- [3] Franconi E., Introduction to Data Warehousing, Lecture Notes, <http://www.inf.unibz.it/~franconi/teaching/2002/cs636/2>, 2002.
- [4] Pang L., Data Warehousing and Data Mining, Leslie Pang Web Site and Lecturer Notes
- [5] Gatzui S. and Vavouras A., Data Warehousing: Concepts and Mechanisms, 1999
- [6] Thomas Connolly & Carolyn Begg., "Database Systems, 3th Edition", Addison- Wesley, 2002
- [7] Gatierez A. and Marotta A., An Overview of Data Warehouse Design Approaches and Techniques, Uruguay, 2000
- [8] Reed Jacobson., "Microsoft® SQL Server 2000 Analysis Services", ISBN 0-7356- 0904-7, 2000
- [9] Rizzi S., Open Problems in Data Warehousing., <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-77/DMDW 2003>, Berlin, Germany
- [10] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Chapter2: Data Warehouse and OLAP Technology for Data Mining, Barnes & Nobles, 2000
- [11] W. H. Inmon, "Building the Data Warehouse, 3th Edition", John Wiley, 2002 88
- [12] Moody D. L. and Kortink M. A. R., From Enterprise Models to Dimensional Models: Methodology for Data Warehouse and Data Mart Design, <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-28/DMDW 2000> , Stockholm, Sweden
- [13] Tryfona N., Busborg F., Christiansen J. G., starER: A Conceptual Model for Data Warehouse Design, Proceeding of the ACM 2nd International Workshop Data Warehousing and OLAP (DOLAP99), 1999
- [14] Sapia C., Blaschka M., Höfling G., Dinter B., Extending the E/R Model for the Multidimensional Paradigm, Proceeding 1st International Workshop on Data Warehousing and Data Mining (DWDW98), 1998
- [15] Golfarelli M., Maio D., Rizzi S., Conceptual Design of Data Warehouses from E/R Schemas, Proceeding of the 31st Hawaii International Conference on System Sciences (HICSS-31), Vol. VII, 1998
- [16] Golfarelli M., Maio D., Rizzi S., The Dimensional Fact Model: A Conceptual Model For Data Warehouses, International Journal of Cooperative Information Systems (IJCIS), Vol. 7, 1998
- [17] Golfarelli M, Rizzi S., A Methodological Framework for Data Warehouse Design, Proceeding of the ACM DOLAP98 Workshop, 1998
- [18] Lujan-Mora S., Trujillo J., Song I., Multidimensional Modeling with UML Package Diagrams, 21st International Conference on Conceptual Modeling (ER2002), 2002
- [19] Trujillo J., Palomar M., An Object Oriented Approach to Multidimensional Database Conceptual Modeling (OOMD) Proceeding 1st International Workshop on Data Warehousing and OLAP (DOLAP98), 1998
- [20] Kimball R., <http://www.dbmsmag.com/9708d15.html> "A Dimensional Modeling Manifesto", DBMS Magazine, Aug 1997
- [21] Kimball R., "The Data Warehouse Toolkit", John Wiley, 1996
- [22] Martyn T., Reconsidering Multi-Dimensional Schemas, SIGMOD Record, Vol. 33, No. 1, 2004
- [23] Elmasri R., Navathe S., "Fundamentals of Database Systems", 3rd Edition, Addison-Wesley, 2000
- [24] Ballard C., Herreman D., Schau D., Bell R., Kim E., and Valencic A., "Data Modeling Techniques for Data Warehousing", IBM Redbook, IBM International Technical Support Organization, 1998
- [25] Firestone J., Object-Oriented Data Warehousing, 1997
- [26] Kimball R., Enforcing the Rules, 2000, http://www.intelligententerprise.com/000818/webhouse.jhtml?_requestid=380244
- [27] Kimball R., The Software Developer in Us, 2000, <http://www.intelligententerprise.com/000908/webhouse.jhtml>
- [28] Microsoft Developer Network (MSDN) Library, XML Web Services Overview, October 2004
- [29] Hahn K., Sapia C., and Blaschka M., Automatically Generating OLAP Schemata from Conceptual Graphical Models, Proceedings ACM 3rd International Workshop Data Warehousing and OLAP (DOLAP 2000), 2000
- [30] Mora-Lujan S., Multidimensional Modeling Using UML and XML, Proceedings 16th European Conference on Object-Oriented Programming (ECOOP 2002), 2002
- [31] Golfarelli M., Rizzi S., WAND: A Case Tool for Data Warehouse Design, Demo Proceedings of The 17th International Conference on Data Engineering (ICDE 2001), 2001
- [32] Chaudhuri S., Dayal U., An Overview of Data Warehousing and OLAP Technology, ACM Sigmod Record, vol.26, 1997
- [33] Golfarelli M., Rizzi S., Designing the Data Warehouse: Key Steps and Crucial Issues, Journal of Computer Science and Information Management, 1999
- [34] Phipps C., Davis K., Automating Data Warehouse Conceptual Schema Design and Evaluation, DMDW'02, 2002
- [35] Peralta V., Marotta A., Ruggia R., Towards the Automation of Data Warehouse Design, 2003
- [36] Batini C., Ceri S., Navathe S., "Conceptual Database Design-An Entity Relationship Approach", Addison-Wesley, 1992
- [37] Abello A., Samos J., Saltor F., A Data Warehouse Multidimensional Data Models Classification, Technical Report, 2000
- [38] Abello A., Samos J., Saltor F., A Framework for the Classification and Description of Multidimensional Data Models, Database and Expert Systems Applications, 12th International Conference, 2001
- [39] Teklitz F., The Simplification of Data Warehouse Design, Sybase, 2000
- [40] Prosser A., Ossimitz M., Data Warehouse Management, University of Economics and Business Admin., Vienna, 2000
- [41] Ahmad I., Azhar S., Data Warehousing in Construction: From Conception to Application, First International Conference on Construction in the 21st Century (CITC2002) "Challenges and Opportunities in Management and Technology", 2002
- [42] Kimball R., Letting the Users Sleep, Part 1, DBMS, 1996, <http://www.dbmsmag.com/9612d05.html>
- [43] Kimball R., Letting the Users Sleep, Part 2, DBMS, 1997, <http://www.dbmsmag.com/9701d05.html>