ISSN 2229-5518

Data Warehouse Technological Infrastructure and Methodology

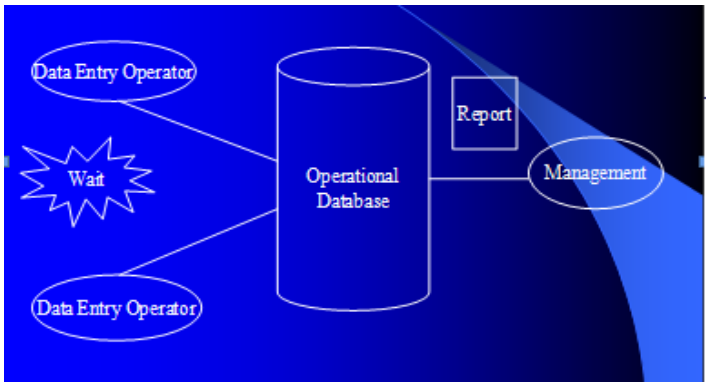
S.Nasira Tabassum

Abstract— Data warehouse Infrastructure basically supports a data warehousing environment with the help of a combination of technologies. In its most general definition, a data warehouse is large repository of all sorts of data the implementing organization would need in the present and in the future. But the real data warehouse and its functions and features may vary depending upon the need of the organization and what it can afford. So, the overall design and methodology of data warehouse will be depending on the data life cycle management policy of the organization. A data warehouse is a subjectoriented, integrated, time-variant, and nonvolatile collection of data that supports managerial decision making. Data warehousing has been cited as the highest-priority post-millennium project of more than half of IT executives. A large number of data warehousing methodologies and tools are available to support the growing market. The general life cycle starts with pre-data warehouse, data cleansing, data repository, and front-end analytics. The pre-data warehouse is like stage or area where the designers need to determine which data contains business value for insertion. Some of the Infrastructure found in this area includes online transaction processing (OLTP) database which store operational data. These OLTP databases may be residing in some transactional business software solutions such as Supply Chain Management (SCM), Point of Sale, Customer Serving Software and Enterprise Resource Planning (ERP) and management software. OLTP databases need to have very fast transactional speeds and up to the point accuracy. During the data cleansing, data undergoes a collective process referred to as ETL which stands for extract, transform, and load. Data are extracted from outside sources like those mentioned in the pre-warehouse.

Index Terms— Data Warehouse, ETL, Data Cleansing, Data Warehouse Methodology, Data Mining.

1 Introduction

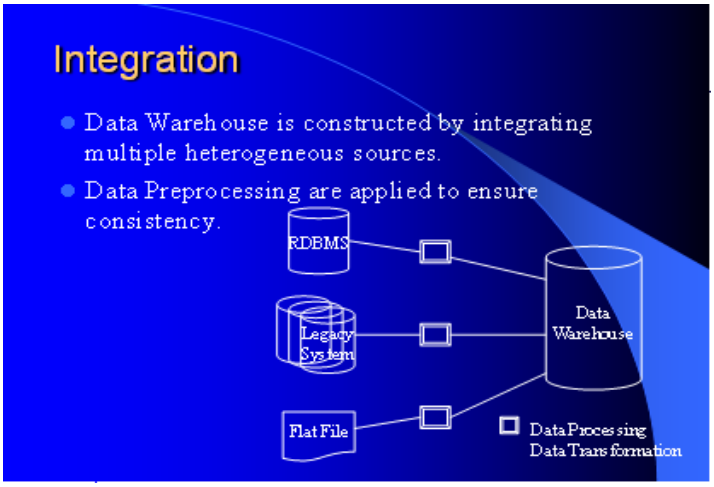
Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. This is generic information on what is needed to consider to set-up the infrastructure for a Data Warehouse. These are the considerations for Data Warehouse Platform alone. If you are placing OLAP Server the end-user tools (like data mining, enterprise reporting, analytics), they will be having their own considerations. One of the requirements of a good ETL tool is that it could efficiently communicate with the many different relational databases. It should also be able to read various file formats from different computer platforms. At the data repository phase, data are stored in corresponding databases. This is also the phase where active data of high business value to an organization are given priority and special treatment. Metadata computer application servers also can be found within this area. Metadata, which means data about data, make sure that data are accurate and clean. It also makes sure that they are well defined because metadata can help speed up searches in the future. It should also be able to read various file formats from different computer platforms. At the data repository phase, data are stored in corresponding databases. This is also the phase where active data of high business value to an organization are given priority and special treatment. Data repositories may be implemented as data mart of operational data store (ODS).A data mart is smaller than a data warehouse and is more specific as it is built on a departmental level instead of company level. An ODS (operational data store) are sort of resting place for data and they hold recent data before they are migrated to the data warehouses. Whether a data warehouse implements both or not, the tools in this stage are all related to databases and database computer servers. The front-end analysis may be considered the last and most critical stage of the data warehouse cycle. This is the stage where data consumers will interact with the data warehouse to get the information they need. Some of the tools used in this area are data mining applications which are used to discover meaningful patterns from a chaotic system or repository.

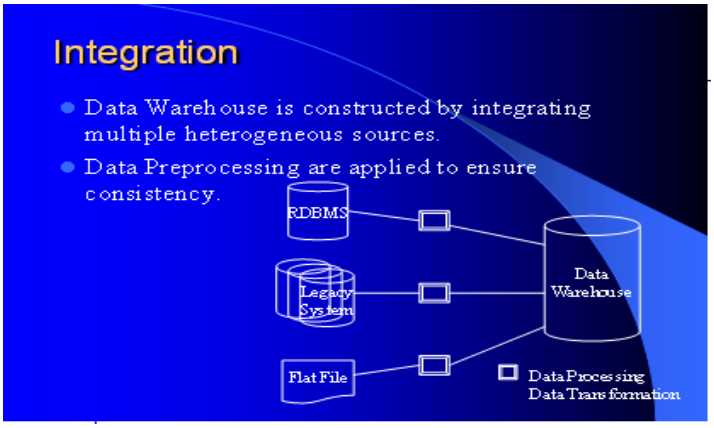


A Data Warehouse is a 'business infrastructure'. In a prac-

• *S. Nasira Tabassum is currently pursuing masters degree program in software engineering from Nizam Institute of Engineering and Technology, Deshmukhi, Nalgonda Dist,Affiliated to Jawaharlal Nehru Technology University, Hyderabad E-mail:nasira.tabassum@gmail.com*

tical world, it does not do anything on its own, but provides sanitized, consistent and integrated information for host of applications and end-user tools. Therefore, the stability, availability and response time of this platform is critical. Just like a foundation pillar, its strength is core to your information management success.





It is refreshed at regular interval so that it contains up to date information for analysis. Historical data is also stored in warehouse which helps in comparing data across different time period. The quality of data is improved before loading it into the warehouse by performing data cleaning and transformation before loading the data.A data warehouse is defined as a subject-oriented, integrated, time variant, nonvolatile collection of data in support of management's decision making process. Time-variant means information from historical perspective. Every key structure contains either implicitly or explicitly an element of time. A datawarehouse is mostly divided into dimensions and fact tables. Dimensions are tables which contain only attributes of elements. Dimension tables are normalized. Facts are mostly tables containing pointers to dimension tables and additional attributes

2 Online Analysis Processing Tool

It is a platform for consolidated historical data for analysis. It stores data of good quality so that knowledge worker can make correct decisions. Online Analytical Processing (OLAP) tool is used in analyzing historical data of the organizations and slice the required business information. Some of the other tools are generic reporting or data visualization tools so that end users can see the information in visually appealing layouts. Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses.Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic ad-vances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

3 Data Model

There are three basic styles of data models; conceptual data model, logical data model and physical data model. The conceptual data model is sometimes called the domain model and it is typically used for exploring domain concepts in an enterprise with stakeholders of the project. The logical model is used for exploring the domain concepts as well...Reverse Data Modeling: Reverse data modeling is basically a form of reverse IT code engineering and it is a process wherein an IT expert tries to extract information from an existing system in order to work backward and derive a physical model and work further back to a logical model in the case of data modeling.

Star Schema: The star schema which is some-times called a star join schema is one of the sim-plest styles of a data warehouse schema. It consists of a few fact tables that reference any number of dimension tables. The facts tables hold the main data with the typically smaller dimension tables describing each individual value of a dimension.

Enterprise Data Model: An Enterprise Data Model is a representation of single definition of data of an enterprise is and the representation is not based on any system application. It indepen-dently defines how the data is sourced, stored, processed or accessed physically. Enterprise Data Model gives overall picture of an industry perspective by offering an integrated blueprint. While there is little doubt that Software as a Service is convenient, flexible and very robust because it is being hosted over the web there are a number of security issues that must be considered. If security issues are not carefully dealt with problems could occur and some organizations may become reluctant to use the technology.

4 TECHNOLOGICAL INFRASTRUCTURE

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. NCR has the capacity to deliver applications exceeding 100 terabytes.

There are two critical technological drivers:

- **Size of the database:** the more data being processed and maintained, the more powerful the system required.
- **Query complexity:** the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers. and deployment The goal is to create individual data marts in a bottom-up fashion, but in conformance with a skeleton schema known as the "data warehouse bus." The data warehouse for the entire organiza-tion is the union of those conformed data marts.

ACKNOWLEDGMENT

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose encouragement and guidance have crowned our efforts with success. We extend our deep sense of gratitude to Principal Mr. M. S Qaseem, HOD, Information Technology, Nizam Institute of Engineering and Technology, Deshmukhi, for his support and encouragement. I am indebted to Ms. Asma, Associate professor, Nizam Institute of En-gineering and Technology, Deshmukhi, Nalgonda (A.P), India for giving her helpful comments and sharing ideas in carrying out this research work.

REFERENCES

[1] Kimball, R., Reeves, L., Ross, M., and Thronth-waite, W. The Data Warehouse Lifecycle Toolkit. Wiley, New York, 1998.

[2] DCI Seminar Workbook—Strategies and Tools for Successful Data Warehouses. DCI, Andover, MA, 1999; www.dciexpo.com.

[3] Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. 1st Edn., John Wiley and Sons, New York, ISBN-13:9780471412540, pp: 516.

[4] A model of data warehousing process maturity. IEEE Trans. Software Eng., 1-1. DOI: 10.1109/TSE.2011.2 Sharma, R.D. and R. Rishi, 2011. Information Technology Infrastructure Li-brary.

[5] Batini, C., Ceri, S., and Navathe, S.K. Conceptual Database Design: An Enity-Relationship Approach. Benja-min/Cummings, Redwood City, CA, 1992.

[6] Inmon, W. Metadata in the data warehouse, White Paper, 2000;

[7] www.inmoncif.com/library/whiteprs/earlywp/ttmeta.pdf.

[8] Kimball, R. and Ross, M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd edition, Wiley, New York, 2002

AUTHOR PROFILE

S. Nasira Tabassum has received her Master of Computer Application from Muffakham Jah Col-lege of Engineering and Technology, Affiliated to Osmania University, Hyderabad, AP India. She is currently pursuing M.Tech in Software Engineering from Nizam Institute of Engineering and Technology, Deshmukhi, Nalgonda Dist, Affiliated to JNTU Hyderabad, AP India. (Email: Nasira.tabassum@gmail.com)