International Journal of Computer Science and Informatics

Volume 3 | Issue 4 Article 15

April 2014

DATA WAREHOUSING APPLICATIONS: AN ANALYTICAL TOOL FOR DECISION SUPPORT SYSTEM

MOHAMMED SHAFEEQ AHMED
Department of Computer Science, Gulbarga University, Gulbarga, Karnataka, India, md_shafeeqa@rediffmail.com

DATA WAREHOUSING APPLICATIONS: AN ANALYTICAL TOOL FOR DECISION SUPPORT SYSTEM

MOHAMMED SHAFEEQ AHMED

Guest Lecturer, Department of Computer Science, Gulbarga University, Gulbarga, Karnataka, India (e-mail: Email:md_shafeeqa@rediffmail.com)

Abstract- Data-driven decision support systems, such as data warehouses can serve the requirement of extraction of information from more than one subject area. Data warehouses standardize the data across the organization so as to have a single view of information. Data warehouses (DW) can provide the information required by the decision makers. The data warehouse supports an on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Data warehousing and OLAP have emerged as leading technologies that facilitate data storage, organization and then, significant retrieval. Both are essential elements of decision support, which has increasingly become a focus of the database industry.

This paper provides a detailed picture of Data warehousing (DW), exploring the features of it, applications and the architecture of DW over Data Mining, Online Analytical Processing (OLAP), On-line Transaction Processing (OLTP) technologies.

***Keywords**: Data Warehousing, Data Mining, OLAP, OLTP, Decision Making and Decision Support & Managerial decision making.*

I. INTRODUCTION

Data Warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions for smooth functioning of the enterprise. It serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. It provides architecture and tools for business executives to systematically organize, understand and use their data to make strategic decisions. Data Warehouse is a database of historical, summarized and consolidated data is more important than detailed, individual records used for reporting and analysis. It refers to the database that is maintained separately from an organization's operational databases which contains consolidated data. The data stored in the data warehouse is uploaded from the operational systems, over potentially long time they tend to be much larger than operational databases. DW systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historical data for analysis. DW technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for inventory management), financial services (for credit card analysis, risk analysis, and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). This paper presents an overview of data warehousing technologies, focusing on the special requirements that data warehouses place on Database Management Systems (DBMSs).

Bill Inmon (father), the world-renowned expert, said that the definition for a Data Warehousing was and still is today. "A source of data that is subject-oriented, integrated, nonvolatile, and time-variant for the purpose of management's decision processes".

Sean Kelly, Defines the data warehouse in the following way. "The data in the data warehouse is: Separate, Available, Integrated, Time stamped, Subject oriented, Nonvolatile, Accessible".

Most queries on data warehouses are ad hoc and are complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Due to the complexity query throughput and response times are more important than transaction throughput.

A DW (or smaller-scale data mart) is a specially prepared repository of data designed to support decision making. The data comes from operational systems and external sources. To create the data warehouse, data are extracted from source systems, cleaned (e.g., to detect and correct errors), transformed (e.g., put into subject groups or summarized), and loaded into a data store (i.e., placed into a DW).

Benefits of using data warehouse in decision making

1. Improvements in turnaround time for data access and reporting generation.

2. Improved productivity of analytical staff due to availability of data
3. Standardizing data across the organization so that there will be one view of information.
4. Merging data from various source systems to create a single information source system.
5. Business improvement resulting from analysis of warehouse data.
6. Encouraging and improving fact-based decision making.

The data warehouse supports OLAP, the functional and performance requirements of which are quite different from those of the OLTP applications traditionally supported by the operational databases.

OLTP covers most of the day to day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration and accounting. An OLTP system is customer oriented and is used for transaction and query processing by clerks, clients and information technology professionals. It manages the current data that, typically, are too detailed to be easily used for decision making. An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. The access patterns of OLTP systems consist mainly of short, atomic transactions. Such system requires concurrency control and recovery mechanisms.

DW systems, on the other hand, are targeted for decision support. It serves users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users.

An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts. These systems manages large amount of historical data, provides facilities for summarization and aggregation, and store and manages information at different levels of granularity. These features make the data easier to use in informed decision making. An OLAP system typically adopts either a star or a snowflake model and a subject-oriented database design. This system often spans multiple versions of a database schema, due to the evolutionary process of an organization. These systems also deal with information that originates from different organization, integrating information from multiple data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

To facilitate complex analysis and visualization, the data in a warehouse is typically modeled multi dimensionally. For example, in a sales data warehouse, time of sales, sales district, salesperson, and product might be some of the dimensions of interest. Often these dimensions are hierarchical; time of sale may be organized as: day – month – quarter – year hierarchy, product as a product-category-industry hierarchy.

An operational database is designed and tuned from known tasks and workloads such as indexing and hashing using primary keys, searching for a particular record and optimizing ‘canned’ queries. On the other hand, data warehouse queries are often complex. They involve the computation of large amount of data at summarized levels and may require the use of special data organization, access and implementation methods based on multidimensional views.

DW might be implemented on Relational OLAP (ROLAP) servers. These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces. They support extensions to SQL and special access and implementation methods to efficiently implement the multidimensional data model and operations. In contrast, Multidimensional OLAP (MOLAP) servers support multidimensional views of data through array-based multidimensional storage engines. They map multidimensional views directly to data cube array structures.

There is more to building and maintaining the data warehouse than selecting an OLAP server, defining a schema and some complex queries for the warehouse. Different architectural alternatives exist. Many organizations want to implement an integrated enterprise warehouse that collects information about all subjects spanning the whole organization. However, building an enterprise warehouse is long and complex process. Some organizations are settling for data marts instead, which are departmental subsets focused on selected subjects. These data marts enable faster roll out since they do not require enterprise-wide consensus.

II. ARCHITECTURE AND PROCESS DESIGN

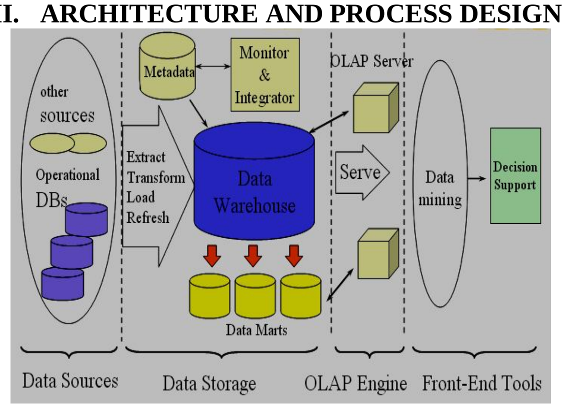

Figure: OLAP with data warehousing architecture

A DW can be built using a top-down approach, a bottom-up approach, or a combination of both. The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood. The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate

the benefits of the technology before making significant commitments. In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

The design and construction of the DW consists of the following steps: planning, requirements study, problem analysis, warehouse design, data integration and testing, and finally deployment of the DW.

1. Choose a business process to model, for example, shipments, inventory, sales etc
2. Choose the grain of the business process. The grain is the Fundamental, atomic level of data to be represented in the fact table for this process.
3. Choose the dimensions that will apply to each fact table record.
4. Choose the measures that will populate each fact table record.

Once a data warehouse is designed and constructed, the initial deployment of the warehouse includes initial installation, roll out planning, training and orientation.

Designing and rolling out a data warehouse is a complex process, consisting of the following:

- Define the architecture, do capacity planning, and select the storage servers, database and OLAP servers, and tools.
- Integrate the servers, storage, and client tools.
- Design the warehouse schema and views.
- Define the physical warehouse organization, data placement, partitioning, and access methods.
- Connect the sources using gateways, ODBC drivers, or other wrappers.
- Design and implement scripts for data extraction, cleaning, transformation, load, and refresh.
- Populate the repository with the schema and view definitions, scripts, and other metadata.
- Design and implement end-user applications.
- Roll out the warehouse and applications.

III. BACK-END TOOLS AND UTILITIES

Data warehouse systems use back-end tools and utilities to populate and refresh their data.

Data Cleaning: Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Since a data warehouse is used for decision making, it is important that the data in the warehouse must be correct. Some examples where data cleaning becomes necessary are: inconsistent field length, inconsistent descriptions, inconsistent value assignments, missing entries and violation of integrity constraints.

Load: After extracting, cleaning and transforming, data must be loaded into the warehouse. Additional preprocessing may still be required: checking integrity constraints; sorting; summarization; aggregation; and other computations to build the derived tables stored in the warehouse. In addition, load utility also allows the system administrator to monitor status, to cancel, to suspend and resume a load, and to restart after failure with no loss of data integrity. The load utilities for data warehouses have to deal with much larger data volumes than for operational databases

Refresh: Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse. There are two sets of issues to consider: when to refresh and how to refresh. Usually, the warehouse is refreshed periodically. The refresh policy is set by the warehouse administrator, depending on user needs and traffic, and may be different for different sources. Refresh techniques also depends on the characteristics of the source and capabilities of the database servers. Replication servers can be used to refresh a warehouse when the sources change.

IV. MULTIDIMENSIONAL DATA MODEL

DW and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions. Dimensions are perspectives or entities with respect to which an organization wants to keep records. Each dimension has a table associated with it, called a dimension table, which further describes the dimension.

A multidimensional data model is typically organized around a central theme. This theme is represented by a fact table. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

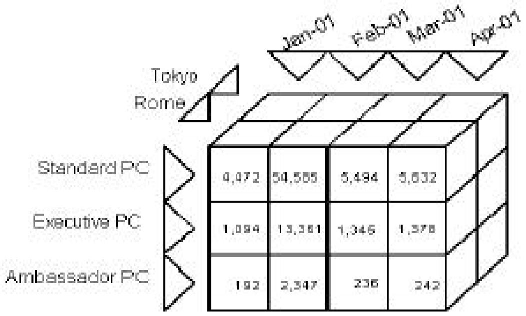

Figure: A logical view of representation a multidimensional data model in an OLAP cube

V. FRONT-END TOOLS

The multidimensional data model grew out of the view of business data popularized by spreadsheet programs that are extensively used by business analysts. One of the popular operations that are supported by the multidimensional spreadsheet is pivoting. Pivot also called rotate, is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data. Other operations are roll-up, drill-down, slice and dice. The roll-up operation performs the aggregation on a data cube, either by climbing up the concept hierarchy for a dimension or by dimension reduction. Drill-down is the reverse of the roll-up. It navigates from less detailed data to more detailed data. The slice operation performs a selection on one dimension of the cube. The dice operation performs a selection on two or more dimensions.

VI. DATA MINING

Data Mining is the extraction or "Mining" of knowledge from a large amount of data or data warehouse data mining should have been more appropriately named as "Knowledge Discovery from Data".

Joseph P. Bigus, Defines the Data Mining in the following way. "The efficient discovery of valuable, non-obvious information from a large collection of data".

Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the web, or other information repositories, or data that are streamed into the system dynamically.

To do this extraction data mining combines artificial intelligence, statistical analysis and database management systems to attempt to pull knowledge form stored data. Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools. The spreadsheet is still the most compiling front-end application for OLAP. The challenges in supporting a query environment for OLAP can be crudely summarized as that of supporting spreadsheet operation effectively over large multi-gigabytes databases. To distinguish information extraction through data mining from that of a traditional database querying, the following main observation can be made. In a database application the queries issued are well defined to the level of what we want and the output is precise and is a subset of operational data. In data mining there is no standard query language and the queries are poorly defined. Thus the output is not precise and do not represent a subset of the database. Beside the data used not the operational data that represents the to day transactions. For instance during the process of building a data warehouse the operational data are summarized over different characteristics, such as borrowings during 3 months period. Queries can be of the type of "identify all borrowers who have similar interest" or "items a member would frequently borrow along with movies", which is not a precise as the list of books borrowed by a member. The nature of the database and the query result in extracting non-subset of data. In supermarkets such relationships have already been identified using data mining. Thus related items such as "bread and milk' or "beer and potato chips" would be kept together. Mobile companies decide on peak hours, rates and special packages based similar market research. Users can use data mining techniques on the data warehouse to extract different kinds of information which would eventually assist the decision making process of an organization Such knowledge could only be discovered through sharing experiences of librarians or by capturing the knowledge through database and integrating them as done when building DW. DSS tools assist users in discovering knowledge.

VII. OLTP and OLAP

Previously, the job of on-line operational systems was to perform transaction and process the query. So, they are also termed as OLTP. Data warehouse systems serve users or knowledge workers in the role of data analysis and decision-making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are called OLAP systems.

Pre-requisite of data warehousing and OLAP

Data warehousing developed, despite the presence of operational databases due to following reasons:

- An operational database is designed and tuned from known tasks and workloads, such as indexing using primary keys, searching for particular records and optimizing 'canned queries'. As data warehouse queries are often complex, they involve the computation of large groups of data at summarized levels and may require the use of special data organization, access and implementation methods based on multidimensional views.

- An operational database supports the concurrent processing of multiple transactions. Concurrency control and recovery mechanisms, such as locking and logging are required to ensure the consistency and robustness of transactions. While and OLAP query often needs read-only access of data records for summarization and aggregation. Concurrency control and recovery mechanisms, if applied for such OLAP operations, may jeopardize the execution of concurrent transactions.
- Decision support requires historical data, whereas operational databases do not typically maintain historical data. So, the data in operational databases, though abundant, is always far from complete for decision-making.
- Decision support needs consolidation (such as aggregation and summarization) of data from heterogeneous sources; and operational databases contain only detailed raw data, which serves as base for decisions which are outputs of the decision process one has to identify the problem first to arrive at proper decision.

VIII. DATABASE DESIGN

Most data warehouse use a star schema to represent the multidimensional data model. The database consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a pointer to each of the dimension that provides its multidimensional coordinates and stores the numeric measures for that coordinates. Each dimension table consists of columns that correspond to attributes of the dimension.

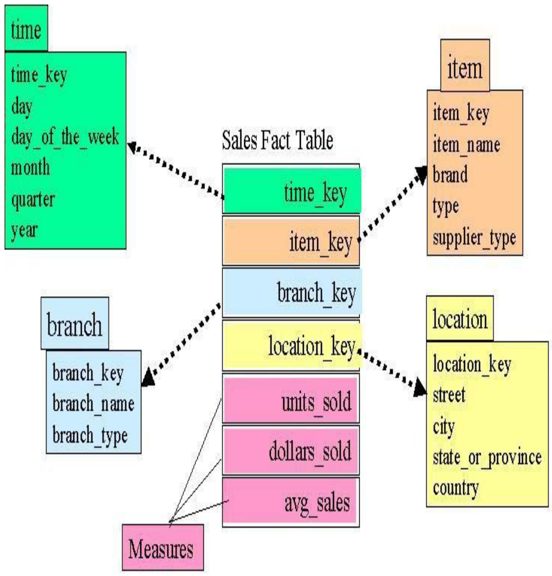

Figure: A Star Schema

The Snowflake schema is the variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms the shape similar to a snowflake.

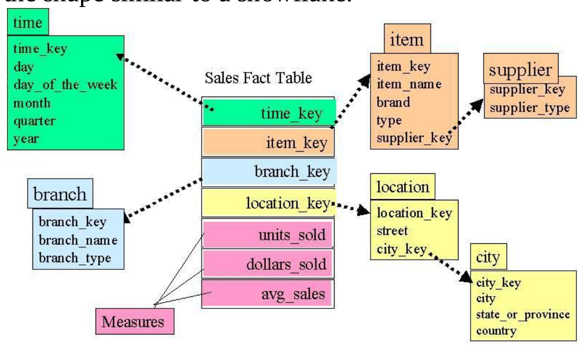

Figure: A Snowflake Schema

The major difference between the snowflake and star schema models is that the dimension table of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

IX. METADATA REPOSITORY

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Administrative metadata includes all of the information necessary for setting up and using a warehouse; description of the source databases; back-end and front-end tools. Business metadata includes business terms and definitions; ownership of the data. Operational metadata includes information that is created during the operation of the warehouse; monitoring information such as usage statistics, error reports, and audit trails. Metadata repository is used to store and manage all the metadata associated with the warehouse. The repository enables the sharing of the metadata among tools and processing for designing, setting up, using, operating and administering a warehouse.

Metadata play a very different role than other data warehouse, and are important for many reasons. For example, metadata are used as a directory to help the decision support system analyst to locate the contents of the data warehouse, as a guide to the mapping of the data when the data are transformed from the operational environment to the data warehouse environment. So, metadata should be stored and managed persistently.

X. CONCLUSION

Data warehouses have become base for effective tool for taking Managerial decisions so that tome lag is reduces and effective implementation takes place. how this help in strategic decision making, which are required to be taken by top management of organization to run it effectively and successfully reaching to achieve the objective of business. Next, is construction the architecture of the data warehouse and the process of a data warehouse design by integrating data from multiple heterogeneous sources to support and /or adhoc queries, analytical reporting and decision making. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Data warehousing and online analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. OLTP is customer-oriented and is used for transaction and query processing by clerks, clients and information technology professionals. The job of earlier on-line operational systems was to perform transaction and query processing. Data warehouse systems serve users or knowledge workers in the role of data analysis and decision making. Next is designing of data warehouses, data mining, distinguished between data warehouse and other techniques (OLAP, OLTP etc).

Data warehouse do not contain the current information. However, data warehouse brings high performance to the integrated heterogeneous database system. It can store and integrate historical information and support complex multidimensional queries. As a result, data warehousing has become very popular in industry.

RERERENCES:

[1] Paulraj Ponniah: Data Warehousing Fundamentals.

[2] Inmon, W.H.: Building the Data Warehouse

[3] Jiawei Han, Micheline Kamber and Jian Pei : Data Mining Concepts and Techniques.

[4] Kimball, R.: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses.

[5] Surajit Choudhary: Data Warehousing and OLAP technology.

[6] Umeshwar Dayal: An overview of data warehousing and technology.

[7] S. Nagabhushana: Data Mining OLAP and Data Mining.

[8] Pieter Adriaans and Dolf Zanting: Data Mining.

[9] Sam Anahory and Dennis Murray: Data Warehousing In the Real World.

[10] Abiteboul, S.; Clue, S.; Milo, T.; Mogilevsky, P.; Simeon, J.: Tools for Data Translation and Integration.

[11] Bernstein, P.A.; Dayal, U.: An Overview of Repository Technology.

[12] Bouzeghoub, M.; Fabret, F.; Galhardas, H.; Pereira, J; Simon, E.; Matulovic, M.: Data Warehouse Refreshment.

[13] Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology.

[14] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: *Cleansing Data for Mining and Warehousing*.

[15] http://www.carolla.com/wp-dw.htm

[16] http://system-services.com/dwintro.asp

[17] http://www.peterindia.net/DataWarehousingView.html

[18] http://www.cs.sunysb.edu/~cse634/presentations/Data Warehousing-part-1.pdf

[19] http://docs.oracle.com/cd/B10501_01/server.920/a9652 0/concept.htm

[20] http://www.peterindia.net/DataWarehousingView.html

[21] http://research.microsoft.com/pubs/76058/sigrecord.pdf

[22] Subramanian, A., Smith, L. D., Nelson, A. C., Campbell, J. F. and Bird, D. A.: 'Strategic Planning for Data Warehousing'. *Information & Management*.

[23] Summer, E. and Ali, D. L.: 'A Practical Guide for Implementing Data Warehousing'. *Computers and Industrial Engineering*.

[24] Tseng, F. S. C. and Chou, A. Y. H.: 'The concept of document warehousing for multidimensional modelling of textual-based business intelligence'. *Decision Support Systems*.

[25] Vassiliadis, P., Bouzeghoub, M. and Quix, C.: 'Towards Quality-Oriented Data Warehouse Usage and Evolution'. *Information Systems*.

[26] Wand, Y. and Wang, R.: 'Anchoring Data Quality Dimensions in Ontological Foundations'. *Communications of the ACM.*

[27] Waston, H. J., Goodhue, D.L. and Wixom B. H.: 'The benefits of data warehousing: why some organizations realize exceptional payoffs'. *Information & Management*.

[28] Watson, H. J., Haley, B.: 'Data warehousing: A framework and survey of current practices'. *Journal of Data Warehousing*.

[29] Wixom, B. H. and Waston, H. J.: 'An Empirical Investigation of the Factors affecting Data Warehousing Success'.

◈◈◈