International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 10
June 2016

A Survey on Data Warehousing Issues towards Data Store Structure

Dhanunjaya Rao Bachalakuri

ABSTRACT: *Data warehousing is a process for assembling and managing data from various sources for the purpose of ahead a single detailed view of an enterprise. The purpose of work is to study the benefits of implementing a data warehouse are as follows: To provide a single version of truth about enterprise information. This may appear somewhat noticeable but it is not unusual in an enterprise for two database systems to have two different versions of the truth. To provide a system in which managers that do not have a strong technical background are able to run complex queries. If the managers are able to access the information they require, it is likely to reduce the bureaucracy around the managers.*

KEYWORDS: Data warehousing; ODS; ETL; OLTP

I. INTRODUCTION

For an enterprise with branches in many locations, the branches may have their own systems. For example, in a university with only one campus, the library may run its own catalog and borrowing database system while the student administration may have own systems running on another machine. There might be a separate finance system, a property and facilities management system and another for human resources management. A large company might have the following system.· Human Resources · Financials · Billing · Sales leads · Web sales · Customer support such systems are called online transaction processing (OLTP) systems. The OLTP systems are mostly relational database systems designed for transaction processing. The performance of OLTP systems is usually very important since such systems are used to support the users (i.e. staff) that provide service to the customers. A data warehouse is a reporting database that contains relatively recent as well as historical data and may also contain aggregate data. The ODS is *subject-oriented*. That is, it is organized around the major data subjects of an enterprise. In a university, the subjects might be students, lecturers and courses while in company the subjects might be customers, salespersons and products. The ODS is *integrated.* That is, it is a collection of subject-oriented data from a variety of systems to provide an enterprise-wide view of the data. The ODS is current valued. That is, an ODS is up-to-date and reflects the current status of the information. An ODS does not include historical data. Since the OLTP systems data is changing all the time, data from underlying sources refresh the ODS as regularly and frequently as possible. The ODS is *volatile*. That is, the data in the ODS changes frequently as new information refreshes the ODS. The ODS is *detailed*. That is, the ODS is detailed enough to serve the needs of the operational management staff in the enterprise.

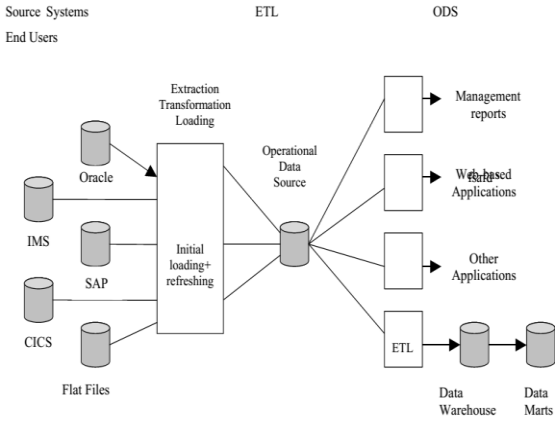

Fig.1 A possible Operational Data Store structure

The granularity of the data in the ODS does not have to be exactly the same as in the source OLTP system. The extraction of information from source databases needs to be efficient and the quality of data needs to be maintained. Populating an ODS involves an acquisition process of extracting, transforming and loading data from OLTP source systems. This process is ETL. Completing populating the database, checking for anomalies and testing for performance are necessary before an ODS system can go online.

In building and ODS, data warehousing is a process of integrating enterprise-wide data, originating from a variety of sources, into a single repository. As shown in Fig.2 the data warehouse may be a central enterprise-wide data warehouse for use by all the decision makers in the enterprise or it may consist of a number of smaller data warehouse (often called data marts or local data

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 10
June 2016

warehouses) . A data mart stores information for a limited number of subject areas. For example, a company might have a data mart about marketing that supports marketing and sales. The data mart approach is attractive since beginning with a single data mart is relatively inexpensive and easier to implement.

II. DATA WAREHOUSES

A centralized warehouse can provide better quality data and minimize data inconsistencies since the data quality is controlled centrally. The tools and procedures for putting data in the warehouse can then be better controlled. Controlling data quality with a decentralized approach is obviously more difficult. As with any centralized function, though, the units or branches of an enterprise may feel no ownership of the centralized warehouse may in some cases not fully cooperate with the administration of the warehouse. Also, maintaining a centralized warehouse would require considerable coordination among the various units if the enterprise is large and this coordination may incur significant costs for the enterprise. As an example of a data warehouse application we consider the telecommunications industry which in most countries has become very competitive during the last few years. If a company is able to identify a market trend before its competitors do, then that can lead to a competitive advantage. What is therefore needed is to analyses customer needs and behavior in an attempt to better understand what the customers want and need. Such understanding might make it easier for a company to identify, develop, and deliver some relevant new products or new pricing schemes to retain and attract customers.

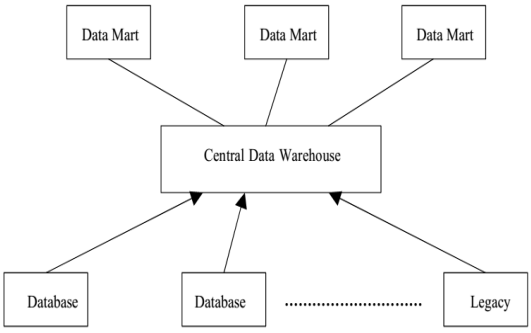

Fig.2 Simple structure of a data warehouse system.

A. ODS and DW Architecture

A typical ODS structure was shown in Fig.1. It involved extracting information from source systems by using ETL processes and then storing the information in the CICS, Flat Files, Oracle the ODS could then be used for producing a variety of reports for management.

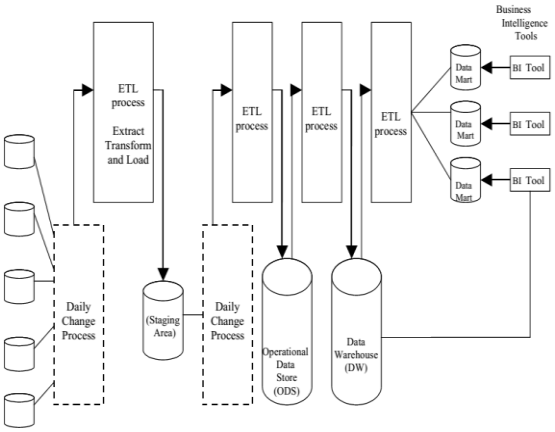

Fig.3 Another structure for ODS and DW

III. DATA WAREHOUSE ARCHITECTURE

The architecture of a system that includes an ODS and a data warehouse shown in Fig.3 is more complex. It involves extracting information from source systems by using an ETL process and then storing the information in a staging database. The daily changes also come to the staging area. Another ETL process is used to transform information from the staging area to populate the ODS. The ODS is then used for supplying information via another ETL process to the area warehouse which in turn feeds a number of data marts that generate the reports required by management. It should be noted that not all ETL processes in this architecture involve data cleaning; some may only involve data extraction and transformation to suit the target systems. There are a number of ways of conceptualizing a data warehouse.

1) One approach is to view it as a three-level structure. The lowest level consists of the OLTP and legacy systems, the middle level consists of the global or central data warehouse while the top level consists of local data warehouses or data marts.
2) Another approach is possible if the enterprise has an ODS. The three levels then might consist of OLTP and legacy systems at the bottom, the ODS in the middle and the data warehouse at the top.

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 10
June 2016

Whatever the architecture, a data warehouse needs to have a data model that can form the basis for implementing it. To develop a data model we view a data warehouse as a multidimensional structure consisting of dimensions, since that is an intuitive model that matches the types of OLAP queries posed by management. A dimension is an ordinate within a multidimensional structure consisting of a list of ordered values (sometimes called members) just like the x-axis and y-axis values on a two-dimensional graph.

A data warehouse model often consists of a central fact table and a set of surrounding dimension tables on which the facts depend. Such a model is called a star schema because of the shape of the model representation. A simple example of such a schema is shown in Fig.4 for a university where we assume that the number of students is given by the four dimensions – degree, year, country and scholarship. These four dimensions were chosen because we are interested in finding out how many students come to each degree program, each year, from each country under each scholarship scheme.

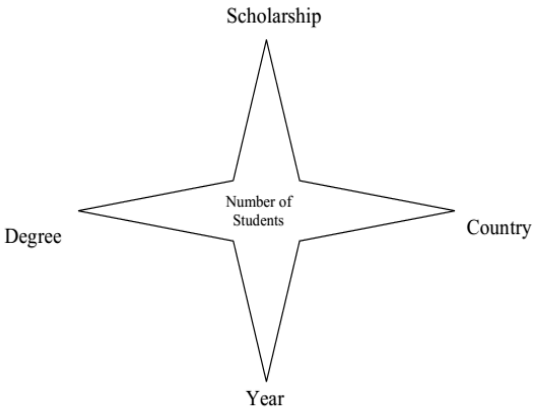

Fig.4 A simple example of a star schema.

A characteristic of a star schema is that all the dimensions directly link to the fact table. The fact table may look like table 1.1 and the dimension tables may look Tables 1.2 to 1.5.

-

Year	*Degree name*	*Country name*	*Scholarship name*	*Number*
200301	BSc	Australia	Govt	35
199902	MBBS	Canada	None	50
200002	LLB	USA	ABC	22
199901	BCom	UK	Commonwealth	7
200102	LLB	Australia	Equity	2

The first dimension is the degree dimension. An example of this dimension table is Table 1.2.

Table 1.2 An example of the degree dimension table

-

Name	*Faculty*	*Scholarship eligibility*	*Number of semesters*
BSc	Science	Yes	6
MBBS	Medicine	No	10
LLB	Law	Yes	8
BCom	Business	No	6
LLB	Arts	No	6

We now present the second dimension, the country dimension. An example of this dimension table is Table 1.3.

Table 1.3 An example of the country dimension table

-

Name	*Continent*	*Education Level*	*Major religion*
Nepal	Asia	Low	Hinduism
Indonesia	Asia	Low	Islam
Norway	Europe	High	Christianity
Singapore	Asia	High	NULL
Colombia	South America	Low	Christianity

The third dimension is the scholarship dimension. The dimension table is given in Table 7.4.

Table 1.4 An example of the scholarship dimension table

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 10
June 2016

-

Name	Amount (%)	Scholarship eligibility	Number
Colombo	100	All	6
Equity	100	Low income	10
Asia	50	Top 5%	8
Merit	75	Top 5%	5
Bursary	25	Low income	12

The fourth dimension is the year dimension. The dimension table is given in Table 1.5.
Table 1.5 An example of the year dimension table

Name	New programs
2001	Journalism
2002	Multimedia

A. Implementation Guidelines

1. *Build incrementally:* Data warehouses must be built incrementally. Generally it is recommended that a data mart may first be built with one particular project in mind and once it is implemented a number of other sections of the enterprise may also wish to implement similar systems. An enterprise data warehouse can then be implemented in an iterative manner allowing all data marts to extract information from the data warehouse. Data warehouse modeling itself is an iterative methodology as users become familiar with the technology and are then able to understand and express their requirements more clearly.

2. *Need a champion:* A data warehouse project must have a champion who is willing to carry out considerable research into expected costs and benefits of the project. Data warehousing projects require inputs from many units in an enterprise and therefore need to be driven by someone who is capable of interaction with people in the enterprise and can actively persuade colleagues. Without the cooperation of other units, the data model for the warehouse and the data required to populate the warehouse may be more complicated than they need to be. Studies have shown that having a champion can help adoption and success of data warehousing projects.

3. *Senior management support:* A data warehouse project must be fully supported by the senior management. Given the resource intensive nature of such projects and the time they can take to implement, a warehouse project calls for a sustained commitment from senior management. This can sometimes be difficult since it may be hard to quantify the benefits of data warehouse technology and the managers may consider it a cost without any explicit return on investment. Data warehousing project studies show that top management support is essential for the success of a data warehousing project.

4. *Ensure quality:* Only data that has been cleaned and is of a quality that is understood by the organization should be loaded in the data warehouse. The data quality in the source systems is not always high and often little effort is made to improve data quality in the source systems. Improved data quality, when recognized by senior managers and stakeholders, is likely to lead to improved support for a data warehouse project.

5. *Corporate strategy:* A data warehouse project must fit with corporate strategy and business objectives. The objectives of the project must be clearly defined before the start of the project. Given the importance of senior management support for a data warehousing project, the fitness of the project with the corporate strategy is essential.

6. *Business plan:* The financial costs (hardware, software, and people ware), expected benefits and a project plan (including an ETL plan) for a data warehouse project must be clearly outlined and understood by all stakeholders. Without such understanding, rumors about expenditure and benefits can become the only source of information, undermining the project.

7. *Training:* A data warehouse project must not overlook data warehouse training requirements. For a data warehouse project to be successful, the users must be trained to use the warehouse and to understand its capabilities. Training of users and professional development of the project team may also be required since data warehousing is a complex task and the skills of the project team are critical to the success of the project.

8. *Adaptability:* The project should build in adaptability so that changes may be made to the data warehouse if

and when required. Like any system, a data warehouse will need to change, as needs of an enterprise change. Furthermore, once the data warehouse is operational, new applications using the data warehouse are almost certain to be proposed. The system should be able to support such new applications.

9. ***Joint management:*** The project must be managed by both IT and business professionals in the enterprise. To ensure good communication with the stakeholders and that the project is focused on assisting the enterprise's business, business professionals must be involved in the project along with technical professionals.

IV. SOFTWARE FOR ODS, ZLE, ETL AND DATA WAREHOUSING

ODS Software · IQ Solutions: Dynamic ODS from Sybase offloads data from OLTP systems and makes if available on a Sybase IQ platform for queries and analysis. · ADH Active Data Hub from Glenridge Solutions is a real-time data integration and reporting solution for PeopleSoft, Oracle and SAP databases. ADH includes an ODS, an enterprise data warehouse, a workflow initiator and a meta library.

ZLE Software HP ZLE framework based on the HP Nonstop platform combines application and data integration to create an enterprise-wide solution for real-time information. The ZLE solution is targeted at retail, telecommunications, healthcare, government and finance.

ETL Software · Aradyme Data Services from Aradyme Corporation provides data migration services for extraction, cleaning, transformation and loading from any source to any destination. Aradyme claims to minimize the risks inherent in many-to-one, many-to-many and similar migration projects. · Data Flux from a company with the same name (acquired by SAS in 2000) provides solutions that help inspect, correct, integrate, enhance, and control data. Solutions include data · Dataset V from Intercon Systems Inc is an integrated suite for data cleaning, matching, positive identification, de-duplication and statistical analysis. · WinPure List Cleaner Pro from WinPure provides a suite consisting of eight modules that clean, correct unwanted punctuation and spelling errors identifies missing data via graphs and a scoring system and removes duplicates from a variety of data sources.

Data Warehousing Software · mySAP Business Intelligence provides facilities of ETL, data warehouse management and business modeling to help build data warehouse, model information architecture and manage data from multiple sources. · SQL Server 2005 from Microsoft provides ETL tools as well as tools for building a relational data warehouse and a multidimensional database. · Sybase IQ is designed for reporting, data warehousing and analytics. It claims to deliver high query performance and storage efficiency for structured and unstructured data. Sybase has partnered with Sun in providing data warehousing solutions.

V. CONCLUSION

To offer a single version of truth about enterprise information. This may appear to some extent noticeable but it is not unusual in an enterprise for two database systems to have two dissimilar versions of the truth. To provide a system in which managers that do not have a strong technical background are able to run complex queries. If the managers are able to access the information they require, it is possible to reduce the bureaucracy around the managers.

REFERENCES

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson Education, 2005.

[2] G. K. Gupta: Introduction to Data Mining with Case Studies, 3rdEdition, PHI, New Delhi, 2009.

[3] Arun K Pujari: Data Mining Techniques, 2nd Edition, Universities Press, 2009.

[4] Jiawei Han and Micheline Kamber: Data Mining - Concepts and Techniques, 2nd Edition, Morgan Kaufmann Publisher, 2006.

[5] Alex Berson and Stephen J. Smith: Data Warehousing,

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 10
June 2016

BIODATA

AUTHOR 1

Dhanunjaya Rao Bachalakuri completed his B.Tech in CSE from Rammappa Engineering College, Warangal in 2004 and M.Tech CSE from Hi-Tech Engineering College, Hyderabad in 2014, Telangana, India. His research areas of interest are Datamining, Software Engineering, Data warehousing.