A Brief Study on Data Warehouse

Baldev Singh, Professor
Department of Computer Science and Engineering, Vivekananda Global University, Jaipur
Email Id- baldev_singh@vgu.ac.in

ABSTRACT: ***Data warehousing and OLAP have become the most important tools for any industry's decision makers. Essentially, data warehousing is the process of gathering and storing historical data in a single repository, referred to as a data warehouse, and then using that warehouse to generate analytical results. Being a helping hand for top-level professionals, it is constantly in the spotlight of the database industry, posing new challenges on a daily basis. We present a critical review of data warehousing, as well as different types of architectures and data modeling for data warehouses, in this paper, author described some of the current data warehousing tools and techniques available. We looked into more problems and issues, as well as some of the research areas in data warehousing.***

KEYWORDS: Data Warehouse, Online Analytical Processing (OLAP).

INTRODUCTION

A data warehouse is a storage location for historical data from various sources. Rather than transaction processing, it is designed for query and analysis. Furthermore, the data warehousing concept encompasses the tools and techniques available for data extraction, transformation, and loading, as well as an OLAP engine, client analysis tools, and other applications used to manage and process data in order to provide decision support to knowledge workers or decision makers.

In any organization, a Data Warehouse is a step toward making a computer system capable of analyzing trends and assisting in critical decision making. From historical data, we can sometimes find very interesting and useful trends that we can use for future planning. The traditional operational databases were designed to aid in the organization's clerical operations, whereas data warehouse and OLAP technologies are designed to aid decision makers (e.g., managers, analysts, etc.). As a result, new challenges in the fields of data warehousing and OLAP emerge on a daily basis to meet the needs of higher-level professionals[1]–[5].

Author presents a critical review of data warehousing technology in this paper. We went through various types of data warehouse designs and data modeling. We looked into the data warehousing tools and techniques that are currently available. Some of the major research issues are also identified.

Foundation of Data Warehousing:

Data warehousing came into prominence as a distinguishable type of computer database during the late 1980 and early 1990s. The definition of Data warehousing emerges to satisfy the demands of the upper management to get analysis results which normal operational database was not providing efficiently. With improvement in technologies and higher demand from the user the concept of Data warehousing has gone through many fundamental stages namely.

- Offline operational Database
- Offline Data warehouse
- Real time Data warehouse
- Integrated Data warehouse.

Architecture of Data Warehousing:

The architecture of Data warehouse depends on the Business process of any organization taking into the account Data consolidation across the organization with security, the level of query requirement management of the Meta, Data modelling and organization, warehouse staging area planning for optimum bandwidth utilization and full technology implementation.

Process architecture, data model architecture, technology architecture, information architecture, and resource architecture are all examples of warehouse architecture[6]–[9].

Architecture of the Process:

It describes the processes involved in turning row data into information. It primarily consists of three subprocesses that are together known as the "ETL" process Extract: Data extraction from a variety of sources using appropriate compression and encryption techniques.

Transform: Converting data collected from many sources into a common format.

Load: Loading the converted data into the data warehouse is one of the steps.

Architecture of the Data Model:

According to Georgia University, there are five different types of data modeling approaches for warehouses:

- Data Mart (independent).
- Dimension-conforming data mart bus architecture.
- The spoke and the hub.
- It is centralized.
- Coordinated.

Architecture of Technology:

It refers to the data warehouse's technical framework. It covers data base connection protocols (ODBC, JDBC, OLE DB, and so on), data base management implementation standards, middleware (based on ORB, RMI, CCOM/DOM, and so on), network protocols (DNS, LDAO, and so on), and associated technologies.

Architecture of Information:

It is the framework that manages the storage, retrieval, alteration, and deletion of data in the Data warehouse by converting information from one form to another step by step.

Architecture of Resources:

It refers to the different resources available, such as software resources for data warehouse maintenance and management. The performance of the data warehouse system is directly related to the quality of the resource architecture.

Typical Data Warehouse Architecture Model:

The above categorization provides an overview of the many types of attributes that should be considered when designing a data warehouse architecture. However, when we discuss the overall design of a data warehouse, we typically refer to a multi-tiered architecture. The architecture of a data warehouse is shown in Figure 1.

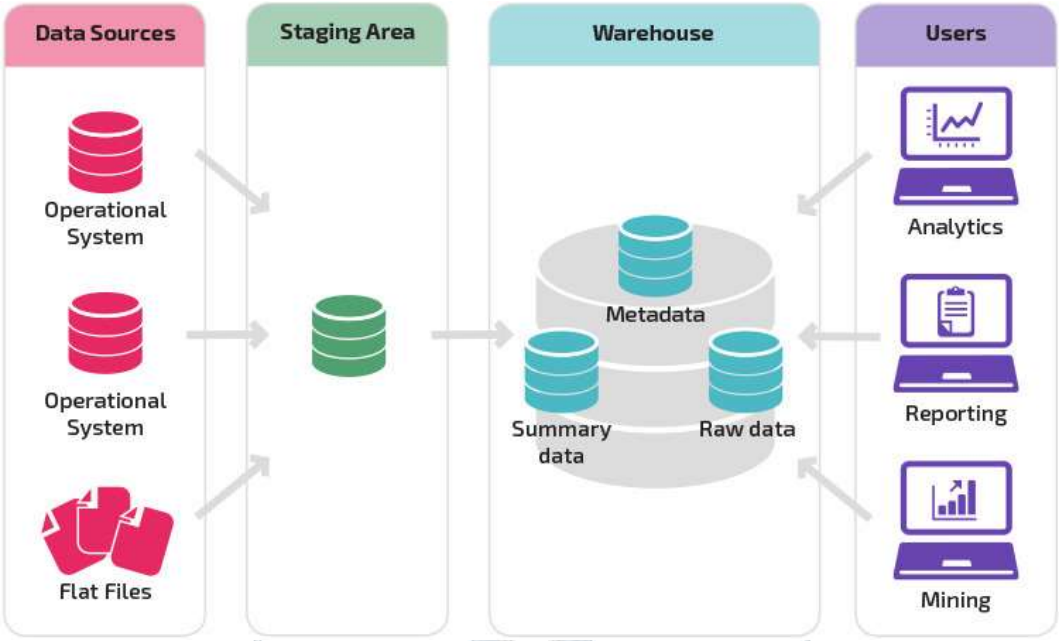

Figure 1: Illustrates the architecture of Data warehouse[10]

Models for Data Warehouses:

Warehouse for businesses:

It's a data warehouse that houses information on all of the company's subjects. It's typically a large data warehouse that requires extensive business modeling. It's a data warehouse that houses information on all of the organization's subjects.

Data mart: It is a subset of the corporate data warehouse that contains data on a specific topic that is useful to a certain group of users. They solely include information on a single topic.

Warehouse in the Cloud:

As a collection of views, it is constructed on top of the operational databases. It's essentially a collection of views on a live database.

Techniques and Tools:

The following are the different types of data warehousing tools.

Utilities & Tools for the Back End:

These tools are also referred to as ETL (Extraction, Transform, and Load) tools, and they are used to carry out the following tasks:

- Extraction of data.
- Cleaning of data.
- Transformation of data.
- Make a load.
- Recharge your batteries.

Oracle Warehouse Builder (OWB), Microsoft Integration Services (SSIS), Telnet Open Studio, IBM Information Server, IBM Cognos Manager, Open Text Integration Centre, Information Builders, ETL Solutions (ETI), and others are some of the most popular products on the market.

Front-End Tools and Conceptual Model:

Front-end tools, often known as OLAP tools, are divided into three categories: multidimensional OLAP (MOLAP), relational OLAP (ROLAP), and hybrid OLAP (HOLAP).

MOLAP: A cube is created by aggregating data from a relational database. Because the data is pre-aggregated inside the cube, it is easier to generate reports.

ROLAP: Unlike MOLAP, no data is pre-aggregated before entering the cube. The ROLAP engine may be thought of as a little SQL generator.

HOLAP: It is a cross between MOLAP and ROLAP. Business Objects, Cognos, Microsoft, Analysis Service, Micro Strategy, and Palo OLAP server are some of the tools offered.

Issues & Problems:

Despite much study over the past decade, data warehouses still have a number of areas to investigate and enhance. The following are some of the main problems to be addressed:

- The first stage in creating a data warehouse is to extract and clean data. The most essential element of any database is the quality of the data in order to obtain the required result as quickly as possible. We now have a variety of tools for data extraction and cleaning, but none of them are as efficient as we would want. We obviously need quality data to obtain quality results, therefore extraction and cleaning of data to acquire quality data is a hot study topic for data warehouses.
- Data transformation and integration is another topic that should be investigated further since data warehouses are built utilizing data from a variety of sources, thus we should have efficient technologies accessible now. Because various databases have different schemas and formats, converting them to a comparable format before loading into the data warehouse is one of the most essential jobs in data warehousing. Data transformation with the least amount of mistake and information loss is still a long way off.
- Data warehouse maintenance is another area where we have a lot of room for improvement. To effectively handle the growing size of the data warehouse, we should search for improved maintenance methods, as well as better software and hardware. Metadata management should also be investigated further.
- The primary goal of any system is to get the result quickly. We have many technologies in the data warehouse for efficient query processing, but they still need to be enhanced significantly to reach the necessary efficiency. More study on query processing is required.

DISCUSSION

In computing, a data warehouse (DW or DWH), sometimes known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis and is regarded a fundamental component of business intelligence. DWs are central repositories of integrated data from one or more diverse sources. A data warehouse contains current and historical data for the whole company and feeds BI and analytics. Data warehouses utilize a database server to bring in data from an organization's databases and offer extra capabilities for data modeling, data lifecycle management, data source integration, and more. Underestimation of data loading resources. Often, we fail to anticipate the time required collecting, clean, and uploading the data to the warehouse, hidden issues in source systems and Data homogeneity are some of the drawbacks of data warehouse.

CONCLUSION

Data warehousing is the foundation of computerized decision support tool. It has been studied a lot in the last decade but yet there are many problems to be addressed in future. Productivity as well as leadership is among the top research topics at today. We have selected some of the newest technologies available for data warehousing and categorized the tools in logical way. The architecture of the data warehouse is also separated logically as well as a typical model of the architecture is also provided. We also examined some of the key study topics including data cleansing, data transformation, maintenance and effective query processing. We recognized key research topics in the data warehousing and the things to be done in future to get the most out of our data warehousing.

REFERENCES:

[1] N. Garcelon *et al.*, “A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse,” *J. Biomed. Inform.*, 2018, doi: 10.1016/j.jbi.2018.02.019.

[2] L. W. Santoso and Yulia, "Data Warehouse with Big Data Technology for Higher Education," 2017, doi: 10.1016/j.procs.2017.12.134.

[3] "Perancangan Dan Implementasi Data Warehouse Untuk Mendukung Sistem Akademik (Studi Kasus Pada STKIP Muhammadiyah Kotabumi)," *J. Teknol. Inf. Magister Darmajaya*, 2016.

[4] N. Jukić, A. Sharma, S. Nestorov, and B. Jukić, "Augmenting Data Warehouses with Big Data," *Inf. Syst. Manag.*, 2015, doi: 10.1080/10580530.2015.1044338.

[5] S. L. Visscher, J. M. Naessens, B. P. Yawn, M. S. Reinalda, S. S. Anderson, and B. J. Borah, "Developing a standardized healthcare cost data warehouse," *BMC Health Serv. Res.*, 2017, doi: 10.1186/s12913-017-2327-8.

[6] S. R. Gardner, "Building the data warehouse," *Commun. ACM*, 1998, doi: 10.1145/285070.285080.

[7] Z. Bicevska and I. Oditis, "Towards NoSQL-based Data Warehouse Solutions," 2016, doi: 10.1016/j.procs.2017.01.080.

[8] I. Moalla, A. Nabli, L. Bouzguenda, and M. Hammami, "Data warehouse design approaches from social media: review and comparison," *Social Network Analysis and Mining*. 2017, doi: 10.1007/s13278-017-0423-8.

[9] N. Garcelon *et al.*, "Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack," *J. Biomed. Inform.*, 2017, doi: 10.1016/j.jbi.2017.07.016.

[10] O. D. Integrator, "(Data Warehouse)," vol. 8, no. 1. pp. 3–5, 2013.