International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT239015

A Review on Data Warehousing Concepts, Challenges and Applications

Dr. L. C. Manikandan*[1], Dr. R. K. Selvakumar[2]

*[1]CSE, Valia Koonambaikulathamma College of Engineering & Technology, Trivandrum, Kerala, India

[2]CSE, CVR College of Engineering, Hyderabad, Telungana, India

Article Info

Publication Issue :
Volume 9, Issue 1
January-February-2023

Page Number : 25-31

Article History
Accepted: 01 Jan 2023
Published: 15 Jan 2023

ABSTRACT

Data warehousing (DW) is a technique for gathering and organizing data from many sources to produce valuable business insights. Business executives can methodically organize, comprehend, and apply their data by adopting data warehousing, which offers structures and tools. This paper's goal is to introduce new learners to the fundamental ideas of DW, as well as its challenges and applications.

Keywords: Data Warehousing, OLAP Server, OLE-DB, ODBC, ROLAP, JDBC, MOLAP

I. INTRODUCTION

A database system created for analytics is called a data warehouse. Data warehouse is frequently used to connect and analyse corporate data from many sources [1]. The central component of the Business Intelligence system, which is designed for data processing and reporting, is the data warehouse. Data cleansing, integration, and consolidation are necessary for the creation of a data warehouse. The following discussion covers important data warehouse characteristics [5].

Subject-oriented: Instead of focusing on the organization's regular operations, it offers information on a certain topic. Products, clients, suppliers, sales, revenue, and other topics are all possible.

Integrated: Data from various sources, including relational databases, flat files, etc., are combined to create the data warehouse. The effective analysis of data is improved by this integration.

Time Variant: A specific time period is assigned to the data collected in a data warehouse. Data in a data warehouse offers information from a historical perspective.

Non-volatile: Non-volatile data does not lose previous information when new information is added to it. Because the operational database and the data warehouse are maintained apart, frequent changes in the operational database do not affect the data warehouse.

II. THREE TIER DATA WAREHOUSE ARCHITECTURE

Typical three-tier architectures [5, 11] for data warehouses include a bottom tier (data warehouse server), middle tier (OLAP server) and top tier (Front end Tools) shown in Figure1.

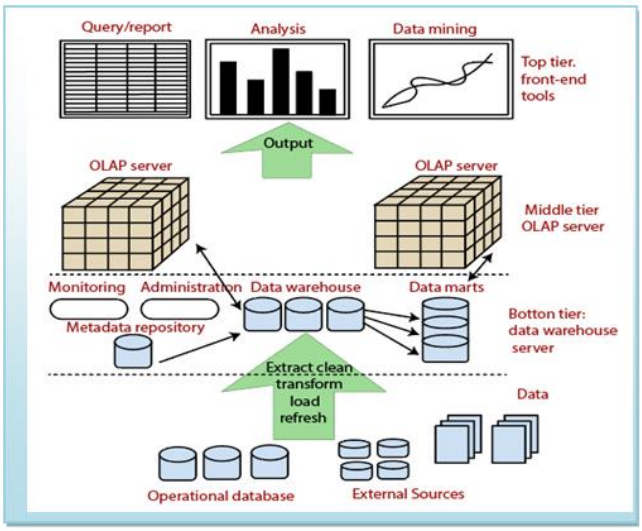

Figure1. Three-tier Architectures for Data Warehouse

A. Bottom-Tier

The Data Warehouse server, which is a constant RDBMS, is considered to be the bottom tier. There may be a metadata repository and a number of specialised data marts included. Application programme interfaces, or "gateways," are used to pull data from operational databases and external sources [11]. The DBMS provides a gateway so that customer programmes can generate SQL code that can be run at a server. Examples of gateways include Microsoft's OLE-DB (Open-Linking and Embedding for Databases) and JDBC (Java Database Connection), as well as ODBC (Open Database Connection) and ODBC (Java Database Connection).

B. Middle-Tier

An OLAP server is part of the middle tier, which allows for quick data warehouse querying. Either the Relational OLAP (ROLAP) or Multidimensional OLAP (MOLAP) model is used to create the OLAP server [2]. A relational database management system (DBMS) extension called OLAP transfers relational operations to functions on multidimensional data. A server designed specifically for multidimensional information and operations, or MOLAP, is used.

C. Top-Tier

A top-tier that includes front-end tools for OLAP results display and extra tools for data mining the OLAP-generated data [3]. The query and reporting tools, analytical tools, and data mining tools are all stored at this layer.

III. DATA WAREHOUSE MODELS

A. Virtual Warehouse:

A virtual warehouse is an overview of a working data warehouse. Construction of a virtual warehouse is simple. Extra space on active database servers is needed to create a virtual warehouse.

B. Data Mart:

Data marts include information that is unique to a certain group. For instance, the marketing data mart can include information about products, clients, and sales. Subject-specific data marts are used [4]. Data marts are implemented using Windows- or Unix/Linux-based servers. The implementation data mart cycles are timed in short increments, i.e., weeks as opposed to months or years.

C. Enterprise warehouse:

Enterprise warehouse compiles all the data on topics spanning the entire company. It offers data integration across the entire company. The information is combined from operating systems and outside data sources. The size of this data can range from a few gigabytes to many terabytes or more.

IV. DATA WAREHOUSING TERMINOLOGIES

A. Metadata

Data about data is the simplest definition of metadata. Metadata are the data that serve as a representation for other data. A book's index, for instance, functions as metadata for the book's contents [5, 11]. In other words, metadata can be thought of as the distilled information that directs us to the full information.

B. Metadata Repository

A data warehouse system's metadata repository is a crucial component. An application that holds descriptive data about the data model used to store and communicate metadata is known as a metadata repository.

- **Business Metadata:** It contains information on data ownership, business definition, and evolving policies.
- **Technical Metadata:** It comprises the names of the database systems, the sizes, types, and names of the tables and columns.
- **Operational Metadata:** Both data lineage and currency are included. Data's state—active, archived, or purged—determines its currency. Data migration and transformation history is referred to as the lineage of the data.

C. Data Cube

We can represent data in various dimensions using a data cube. Facts and dimensions serve to define it. The entities that a company maintains records with regard to are known as the dimensions.

D. Data Mart

Data marts only include information that is unique to a given group. For instance, the marketing data mart might simply have information on products, clients, and sales. Subjects are the only focus of data marts [6].

V. DATA WAREHOUSING SCHEMAS

The database as a whole is logically described by the schema [1,11]. The names and descriptions of all record kinds are included. Data warehouses employ the Star, Snowflake, and Fact Constellation schema, while databases use the relational paradigm.

A. Star Schema

Each dimension in a star schema is represented by a single dimension table. The collections of attributes are contained in this dimension table. The following Figure2 displays sales information for a corporation in relation to the four dimensions of time, item, branch and location.

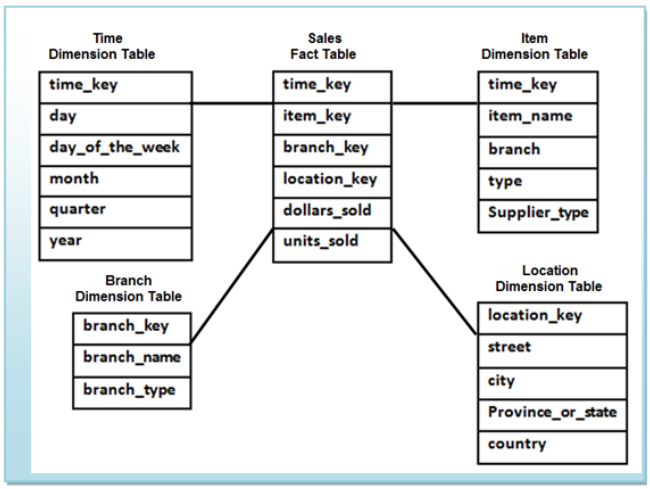

Figure2. Star schema of a data warehouse for sales

A fact table is located in the middle. It has the access codes for each of the four dimensions. The attributes, including dollars sold and units sold, are also included in the fact table [7]. Note that each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set

{location key, street, city, province or state, country}

B. Snowflake Schema

In the Snowflake schema, several dimension tables have been standardised. The data is divided up into more tables during the normalisation process [8]. For instance, the item and supplier dimension tables in the star schema are segregated from the item dimension table and normalised. Snowflake schema of a data warehouse for sales is shown in Figure3.

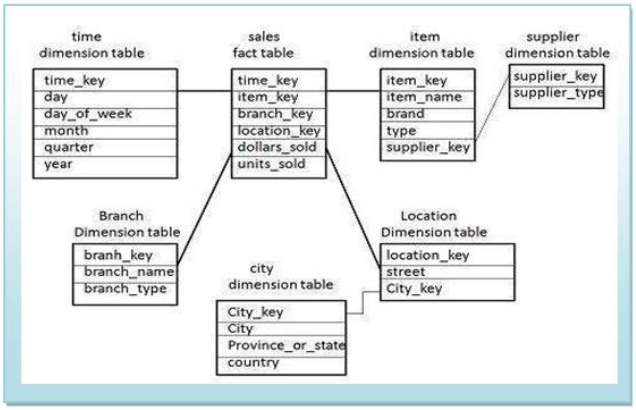

Figure3. Snowflake schema of a data warehouse for sales

The attributes item key, item name, type, brand, and supplier-key are now present in the item dimension table. The supplier dimension table and the supplier key are connected. The properties supplier key and supplier type are included in the supplier dimension table. Note that the Snowflake schema has less redundancy as a result of normalisation, making it easier to maintain and using less storage space.

C. Fact Constellation Schema

There are various fact tables in a fact constellation. Two fact tables, especially the sales and shipping ones, are displayed in the Figure4.

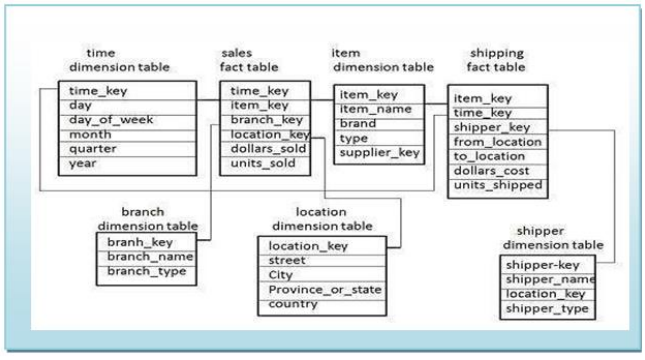

Figure4. Fact constellation schema of a data warehouse for sales and shipping

Similar to the star schema, the sales fact table is also there [9]. The five dimensions of the shipping fact table are item key, time key, shipper key, from location, and to location. Additionally, there are two measurements in the shipping fact table: dollars sold and units sold. Additionally, sharing dimension tables amongst fact tables is an option. For instance, the sales fact table and the shipping fact table both use the same time, item, and location dimension tables.

VI. NEED FOR DATA WAREHOUSE

Need of Data Warehouse is shown in Figure5.

Figure 5. Need of Data Warehouse

A. Business User: To access historical data summaries, business users need a data warehouse. The information might be provided to these non-technical persons in a simple way [10,11].

B. Store historical data: Keeping historical time-variable data used for a variety of reasons in a data warehouse is necessary.

C. Make strategic decisions: On the information in the data warehouse, several techniques might rely. As a result, data warehouse aids in decision-making at the strategic level.

D. For data consistency and quality: By combining data from several sources in one location, the user can successfully work to bring uniformity and consistency in data.

E. High response time: The need for a substantial amount of flexibility and quick response times arises from the need for the data warehouse to be prepared for occasionally unforeseen loads and query patterns.

VII. BENEFITS OF DATA WAREHOUSE

- Recognize market trends and improve your forecasting abilities.
- Data Warehouses are built to handle massive volumes of data..
- Data warehouse structures are easier for end users to navigate, comprehend, and query.
- Data warehouses may make it simpler to design and manage queries that would be complex in multiple normalised databases.
- An effective way to handle the demand for a lot of information from many consumers is through data warehousing.
- The ability to evaluate a sizable volume of historical data is provided by data warehousing [11].

VIII. CHALLENGES FOR DATA WAREHOUSING

Remove any unwanted data immediately. combining several sources into a single common template. For instance, a system might refer to an attribute as "Employee ID" whereas another might use "EID." Missing and incorrect data handling must be done. This can include eliminating some entries altogether or substituting projected or default values for them. Complex data warehouse queries are common. It may be necessary to apply specialised data structure, access, and implementation methods based on multidimensional perspectives since they involve the computation of big groupings of data at summary levels [10].

a. **Data Quality:** Data is originating from various places and from all areas of an organisation. Errors occur when a data warehouse tries to merge contradictory data from different sources. Data quality issues are brought on by inconsistent data, duplicates, logical inconsistencies, and missing data. Analytical and reporting errors are caused by poor data quality.

b. **Understanding Analytics:** Analytics and reporting will need to be taken into account while designing a data warehouse. The business user will need to know precisely what analysis will be conducted in order to accomplish this.

c. **Quality Assurance:** Big Data reporting and analytics are used by a data warehouse's end user to aid in decision-making. As a result, the data must be entirely correct to prevent a credit union leader from making poor choices that could harm the organization's ability to succeed in the future.

d. **Performance:** Similar to how a car is built, a data warehouse is built. To fulfil its intended functions, a car needs to be carefully built from the start. To really tailor the vehicle to a buyer's specific performance requirements, however, there are some options that must be taken into account. To meet demands for overall performance, a data warehouse must also be carefully planned. While the finished item can be altered to meet the organization's performance requirements.

e. **Designing the Data Warehouse:** The majority of people don't want to "waste" their time developing the specifications required for a suitable data warehouse design. Typically, there is a clear understanding of what people want from a data warehouse. As a result, there is a communication breakdown between the technicians developing the data warehouse and the business customers. The usual outcome is a data warehouse that falls short of what the user had hoped for. After first delivery, adjustments and upgrades are required because the data warehouse is insufficient for the end user.

f. **User Acceptance:** People are hesitant to alter their everyday routines, particularly if the new procedure is confusing. To create a data warehouse that is readily adopted by an organisation, various obstacles must be addressed.

g. **Cost:** The idea that credit unions may develop their own data warehouse in-house and save money is a common one. A successful do-it-yourself project is highly expensive, which is the hard fact.

IX. APPLICATIONS OF DATA WAREHOUSE

A. Information Processing: Data stored in a data warehouse can be processed there. Querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs are all methods for processing the data [10].

B. Analytical Processing: Analytical processing of the data housed there is supported by a data warehouse. The data can be examined using fundamental OLAP procedures including slice-and-dice, drill-down, drill-up, and pivoting.

C. Data Mining: Finding hidden patterns and relationships, creating analytical models, performing classification, and making predictions are all methods used in data mining to enhance knowledge discovery. Utilizing visualisation tools, these mining results can be displayed.

Data warehouses are widely used in the following fields:

a. **Finance Industry:** Revolve around the analysis and patterns of customer spending, which helps their clients make the most money possible.

b. **Banking Industry:** The analysis of consumer data, market trends, governmental rules and reports, as well as financial decision making, are all given special attention, along with risk management and policy reversal.

c. **Customer Goods Industry:** They are employed in inventory management, market and advertising research, as well as the forecasting of consumer trends. Additionally, a thorough examination of sales and production is done.

d. **Retailers:** Between manufacturers and consumers, retailers act as intermediaries. To keep track of products, advertising campaigns, and consumer purchasing patterns, they employ warehouses. Additionally, they use an elimination procedure to decide their shelf space after analysing sales to identify rapid and slow-moving product lines.

X. CONCLUSION

The core ideas of data warehousing, the three-tier data warehouse design, terminologies, schemas, advantages, challenges and applications were the main topics of this review study. For young readers and researchers to comprehend the fundamental ideas of data warehousing, this would be very beneficial.

XI. REFERENCES

[1] Vaibhav Singh and Ashwini Ghate,"A Review: Analysis on Data Warehousing and Data Mining", International Journal for Research & Development in Technology, Vol.8, Issue.1. July 2017.

[2] A.R.Arunachalam and S.Srigowthem, "A Study on Data Warehouse Architecture", International Journal of Pure and Applied Mathematics, Vol.116, No. 13, pp.273-275, 2017.

[3] Muhammad Arif, Ghulam Mujtaba, "A Survey: Data Warehouse Architecture", International Journal of Hybrid Information Technology. Vol.8, No. 5, pp. 349-356, 2015.

[4] Dishek Mankad and Preyash Dholakia, "The Study on Data Warehouse Design and Usage", International Journal of Scientific and Research Publications, Vol.3, Issue.3, March 2013.

[5] Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining Concepts and Techniques", Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, 2012.

[6] Song G, Yang D, Lin Z, Tang S, Wang T, Xie K, "Active real time data warehouse concepts, problems and applications", Journal of Computer Research and Development, 44(suppl.), pp.441-446, 2007.

[7] Mohanty.S.," Data warehousing design, development and best practice", New Delhi, Tata McGraw-Hill.

[8] Sandeep Singh and Sona Malhotra, "Data Warehouse and its Methods", Journal of Global Research in Computer Science, Vol.2, No. 5, May 2011.

[9] Ralph Kimball, Margy Ross, "The Data Warehouse Toolkit", Wiley Computer Publishing, 2ndedition, pp.79-85, 2002.

[10] Inmon W.H, "Building the Data Warehouse", John Wiley, 1992.

[11] https://www.javatpoint.com/data-warehouse-architecture

AUTHOR PROFILE

Dr.L.C.Manikandan is working as Professor in Computer Science and Engineering department at Valia Koonambaikulathamma College of Engineering and Technology, Thiruvananthapuram, Kerala INDIA. He has received his Ph.D. and M.Tech. Degree in Computer and Information Technology from Manonmaniam Sundaranar University, M.Sc., and B.Sc. degree in Computer Science from Bharathidasan and Manonmaniam Sundaranar University. He has 18 years of teaching experience in reputed institutions. He has published several patents, textbooks and research papers in various reputed international journals. He carries out research in Digital Image Processing, Video Surveillance and Video coding.

Dr. R. K. Selvakumar received his Ph.D degree in Computer Engineering from Manonmaniam Sundaranar University, India. He has completed the Master degrees M.Sc.(Maths), DOEACC-C(CT), M.C.A., M.Phil.(CS) and M.Tech.(C&IT). He is working as Professor at CVR College of Engineering, Hyderabad, Telungana, INDIA. He carries out research in Digital Image Processing, Video Surveillance, Video coding, Digital Watermarking and Soft Computing.

Cite this article as :

Dr. L. C. Manikandan, Dr. R. K. Selvakumar, "A Review on Data Warehousing Concepts, Challenges and Applications", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9 Issue 1, pp. 25-31, January-February 2023. Available at doi : https://doi.org/10.32628/CSEIT239015
Journal URL : https://ijsrcseit.com/CSEIT239015