Acta Scientific COMPUTER SCIENCES

Volume 5 Issue 3 March 2023

Review Article

Data Warehousing: A Literature Review on Effective Implementation Approaches

Cheryl Ann Alexander[1]* and Lidong Wang[2]

[1]*Institute for IT innovation and Smart Health, Mississippi, USA*

[2]*Institute for Systems Engineering Research, Mississippi state university, Vicksburg, USA*

***Corresponding Author:** Cheryl Ann Alexander, Institute for IT Innovation and Smart Health, Mississippi, USA.

Received: January 16, 2023

Published: February 18, 2023

Abstract

Data Warehousing is data-driven by society. Society is data-driven, making the process of making business decisions both difficult and data-driven. Organizations use many processes to determine what business decisions would be best for them, each of them capable of processing huge amounts of data and assisting the management in decision-making. For decades, the only business intelligence process in play has been data warehouses. Now, however, Big Data has made it possible to expand and modernize to support other data sources capable of centralizing heterogeneous data in a manner that expands the variety and dynamics of data such as data lakes, the cloud, etc.

Keywords: Data Warehousing; Big Data; Data Analysis; Cloud; Machine Learning; Cybersecurity

Introduction

Society is data-driven, making the process of making business decisions both difficult and data-driven. Organizations use many processes to determine what business decisions would be best for them, each of them capable of processing huge amounts of data and assisting the management in decision-making. For decades, the only business intelligence process in play has been data warehouses. Now, however, Big Data has made it possible to expand and modernize to support other data sources capable of centralizing heterogeneous data in a manner that expands the variety and dynamics of data such as data lakes, the cloud, etc. The purpose of this business intelligence is to broaden the strategic decision-making power that the company has, using data analysis and data mining to deepen the strength behind decisions [1]. Thus, the development of data warehouses emphasizes the importance of quality data for successful business decision-making. Although high-quality data is the goal of data-based decision-making using a data warehouse, unless the quality of data is improved in many cases, the decision-making process continues to collapse because many data warehouses fail to effectively clean the data and correct mistakes or data inconsistencies, leaving low-quality, inconsistent, and useless data for use in important business decisions [2].

However, high-quality data is the aim of data warehousing to broaden the evolution of their data-based decision-making processes. The Big Data architecture has indeed fostered the growth and expansion of deep technologies within data warehousing because the velocity, veracity, volume, and variety of data has outgrown traditional technologies and while Big Data has been the primary method by which data warehouses have approached the analytics, data warehouse architecture has continued to evolve to constantly adapt to the evolution of data [1]. To define a data warehouse as a computer system that stores and analyzes data to reveal trends, patterns, and correlations that provide information and insight for successful business intelligence. The typical data warehouse stores and integrates data from transactional databases (i.e., internal sources such as finance, sales, productivity, etc.). The data warehouse emerged when businesses realized that analyzing data

from transactional databases had slowed, some had even crashed from the overload, making some databases and business intelligence solutions obsolete [3].

In this data-driven society, the speed of data, the purity, and the ability to correlate data from transactional databases simultaneously slowed, as the amounts of data increased substantially, making most database solutions obsolete. Organizational changes and technology changes increased the need for business solutions that could process data quickly and clean the data while storing it. As the onset of the big data era began, massive amounts of data could be analyzed, stored, and cleaned simultaneously while delivering accurate and successful decision-making solutions from semi-structured and unstructured data. However, even with the increases in adaptability, solutions for stronger more flexible architectures and more adaptable solutions continued to produce newer deep technologies [3]. In Figure 1 [3], the evolution of a modern data warehouse is time-lined. Over several decades, modern data management has caused the evolution of the data warehouse to reduce the number of errors and increase the success of decision-making for businesses.

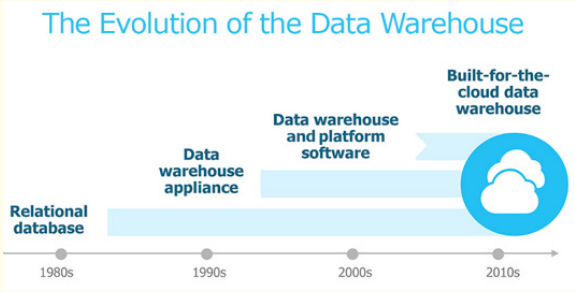

Figure 1: The Evolution of a Modern Data Warehouse.

A data warehouse is a structured, non-volatile historical data storage schema that organizes, analyzes, and orients data by subject for analytical purposes. However, the data warehouse has a limited ability to manage the semi-structured and unstructured data typical of big data. As a result, data lakes are becoming more common as a business intelligence storage solution serving as a heterogenous data storage architecture for data warehouses [1]. Because achieving high-quality data is often challenging for a data warehouse, many fail to compensate, fail to clean data entirely, or do not have an architecture that can adequately perform the functions necessary for a correct model. A data warehouse can be defined as a massive database that functions to support decision-making in each department supporting the business (i.e., administrative, financial, marketing, etc.) [2].

Components of a modern data warehouse

There are multiple components in a data warehouse, however, there are a few more important than others. The prime intent of a data warehouse is to facilitate a singular type of truth for decision-making and potential forecasting. The data warehouse contains both historical and commutative data from single or multiple sources within an organization and can simplify making reports and decisions. Components such as sourcing, acquisitions, clean-up, and transformation tools (i.e., ETL), metadata, data warehouse bus architecture, and query tools, tend to be most important to the final data product [1-3].

Over the last few years, the accumulation of massive amounts of various types of data such as structured, unstructured, and semi-structured data have demanded the development of novel efficient algorithms able to effectively integrate and analyze heterogeneous data within the data warehouse. Traditional tools may not be able to leverage the huge amounts of semi-structured and unstructured data associated with the data warehouse, making it necessary to develop tools and technologies, and frameworks that can manage large-scale data analytics such as Apache Hadoop and Apache HBase database management solutions, primarily aimed at managing data growth, yet leaving the problems associated with data evolution and structure unresolved [4]. Therefore, a great deal of manual manipulation of the data is necessary when solutions do not support automatic or semi-automatic propagation of changes to the data in the warehouse. To solve this problem, data warehouse architecture becomes a pivotal challenge to find a solution that supports data analysis of data integrated within a data warehouse and can discover changes to structured and semi-structured data and can automatically or semi-automatically propagate changes within the data in the system to maintain a continuous system operation [4]. This makes architecture one of the most important components within the data warehouse and the ability to run not only structured data but unstructured and semi-structured data as well. The most unique feature is the adaptation component, which manages changes in data sources or other parts of the unique data highway.

Data warehouse architecture

Data warehouse architecture also contains several key components that contribute to successful management of data manipulation and ease of handling data flow and problems in processing the data. Data sources, data highway, meta-store, and an adaptation component are critical components to a successful analytics data warehouse architecture [4]. From the source, data is loaded into the data highway; data is obtained from several heterogeneous sources at this point and includes big data sources, etc. Because the architecture supports big data sources at this point, it may support any file type or data type including handwritten notes (unstructured data), structured data such as regular files, and semi-structured files such as GIF, JSON, and CSV formats [4].

Data architecture components have a unique format that allows adaptation; the adaptation can be aimed at changing data sources and handling data within the data highway, while users can expect to have more flexibility with the metadata. Figure 2 [4] details a typical data warehouse architecture.

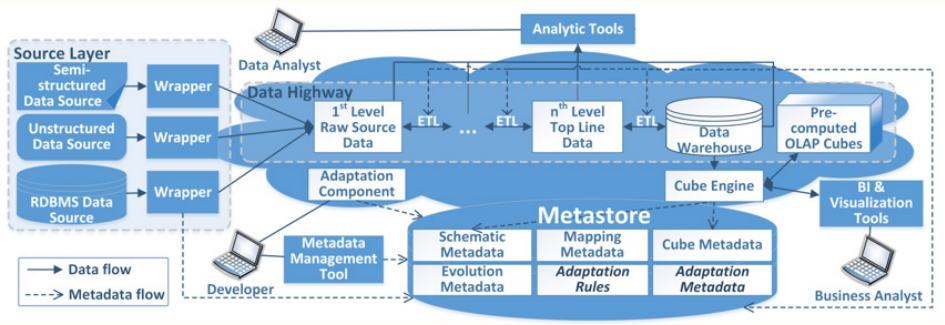

Figure 2: Typical Data Warehouse Architecture.

Data warehousing has evolved over the decades and become an essential means of analytical processing. Capable of performing key metrics, running critical decision models, or ensuring its ability to align with corrective data received from analytics, data warehousing architecture provides vital functions of any data management or business intelligence system [5]. Breaking down current existing architecture for a data warehouse and its ability to collect critical data successfully, apply data successfully in a business intelligence framework, and determine how well or precisely the data architecture is working based on the position in the company (e.g., middle-management, upper-management, data processors, data warehouse staff members, etc.) are displayed in Table 1. In a survey among all levels of employees, all levels of data warehouse architectures were examined, and the findings are related in table 1 [5].

Findings indicate a truly federated system may be dependable for some mergers and acquisitions functions, and the independent data mart has long been unsuitable for data warehouse use, while the data mart bus and hub and spoke architectures seem most likely to achieve good business intelligence and decision-making [5].

Architecture Type	Information Quality	System Quality	Individual Impacts	Organizational Impacts
Independent Data Mart	5	5	5	5
Data Mart Bus	2	2	2	2
Hub and spoke	1	1	1	1
Centralized	3	3	3	3
Federated	4	4	4	4

Table 1: Architecture Type Rated on a Scale of Viability.

A data warehouse can be a complex data storage system that contains consistent historical and semantic data for decision-making. Architecture can serve as the physical implementation for structuring, storing, and processing data by end-users. However, innumerable challenges may interfere with the selection and functioning of a data warehouse architecture (e.g., data modeling, increasing storage volume in cases of increasing data, GPS locations, etc.). Sometimes a hybrid approach is much more effective at providing a tenable solution to business intelligent decision-making

[6]. This distributed approach is primarily suited for data decisions made for online websites, social media sites, and other sites not interested in business decisions or historical data. However, the data on distributed sites (e.g., Twitter, Facebook, Instagram, etc.) must be cleaned with data and stored in a specific format after enrichment with even more data. Because the integrated data all share one model, are often used to clean the data, and are stored in one manner. This process makes the distributed data warehouse is more efficient [6,7].

Data warehouse security

As data is one of the most crucial assets of an organization, protecting that data, data security, and access control, is the only smart option to keep business secrets within the organizational operations. Companies are faced with substantial amounts of data from multiple heterogeneous sources, stored in a separate warehouse for decision-making by organizational management. The concern for unsuspecting persons to access the data warehouse is cause for concern. Measures such as confidentiality, integrity, and the availability of the data have been implemented to prevented cyberattacks or unsuspecting fraudulent theft or access to the data. To control the data, all access must be controlled, all unauthorized access blocked, and the integrity of the data must be assured [8].

One of the most data prolific paradigms, with the highest need for data security and integrity, is the healthcare arena. Healthcare was previously dominated by error-prone documentation, time-consuming manual notes, orders, and other records, resulting in a poorly run, error-filled structured database. Now, after the digitalization of healthcare records, an efficiently run database filled with automated data has replaced the disorganization of the handwritten record. With the introduction of the Big Data era, structured, unstructured, and semi-structured data can all be stored and analyzed in one record. However, the key issue with healthcare records also includes the need for a prominent level of data security. Not only should data integrity be verified, but also data access, and denial of service [9]. One example of data access control is the intrusion detection network system (IDNs). One IDNS, the Role-Based Access Control (RBAC), and RBAC2 secure data from attack and ensure the safety of the data warehouse data and security of the decision-making behind the data warehouse [10]. Also, the secondary access control makes it appealing to those data warehouses that need to protect their data. For healthcare data, the data warehouse can also be used in precision medicine and this alone could have a staggering impact if data access were allowed by unauthorized personnel [9,10]. However, scientific research has other security needs as it is the basis for all medical decision-making, and the integrity must be protected. The BioWH2 tool can bring a third level of security to medical research [11].

A three-layered data warehouse structure was proposed. The role-based permission and security are executed in the access control layer. The data privacy and data security plans are developed based on the types of persons of data collection, the types of users, data transmission, data storage, and data analysis [12]. A blockchain based decentralized integrity verification model for a data warehouse was presented. The model is used to validate the integrity of a data warehouse and replace the existing process. The blockchain confirms the authenticity of files [13].

Extract, transfer, and load (ETL)

Data within an organization can be as small as necessary to run operations, conduct business, and prepare for the future to massive data collected from multiple sources in multiple formats (i.e., structured, semi-structured, and unstructured data). Simply put, data in an organization is typically used to make smart business decisions [14]. Data is fluid. Data can flow as information from a computer program, or data can contain all the information necessary for the organization to make strategic business decisions, or clinical decisions for diagnosis and treatment, making data an invaluable and most important business commodity [15]. According to MySQL (i.e., Structured Query Language), the data in the data warehouse is critical for updating, accessing, inserting, and modifying the data in the dataset. This becomes important when considering another crucial component of data warehousing, ETL (i.e., Extract, Transfer, and Load). While SQL allows the database to communicate within itself and data to organize and update, ETL communicates with important datasets from multiple and integrated sources, making it the most crucial aspect of data warehousing because ETL allows the database to become integrated, gathering data from multiple and variable sources within and outside the datasets so that SQL can communicate within the entire dataset [14].

The first part of ETL is extracted, which is foremost in completing the data warehouse as it allows the data warehouse to gather data from multiple various sources. Data can be drawn from differ-

ent parts of operational datasets and converted to flat files. Afterward, data can be cleaned for the data warehouse and transformed to fit a specific data warehouse schema. For example, charts and historical data can be converted to regular file specifications. After the data has been cleaned appropriately, the final process of loading can be done as the new data has to meet the specifications of the data warehouse schema to make business decisions and use the data for business intelligence [15]. The data must also meet the criteria for historical data, integrated, encapsulated, etc. There is also the modeling schema (i.e., Snowflakes), which is a multidimensional extension of the star scheme in which the hierarchal nature of the star scheme is further broken down into normalizing data for the data warehouse. Figure 3 [15] illustrates how the ETL can break down data, whether it is in tables, graphs, charts, medical imaging, or raw data into an organized data warehouse schema.

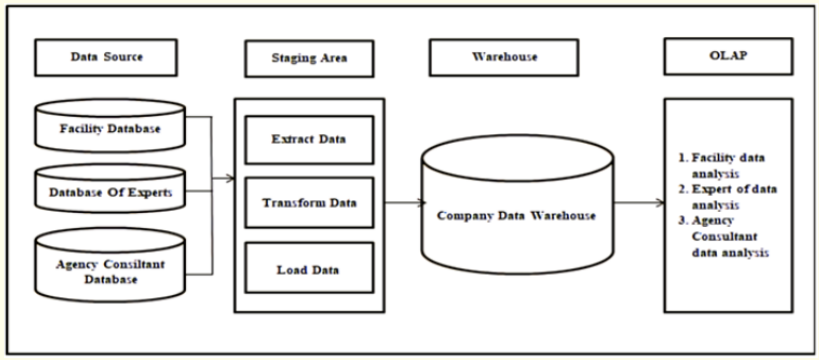

Figure 3: ETL Process Breaking Data into Organized DW Schema.

Conclusion

Society is data-driven. Data is fluid and composed of multiple types such as charts, graphs, structured, unstructured, and semi-structured data within an organization, and intelligent data from outside the organization used to make smart business decisions or provide precision medical care based on the individual's health history. A data warehouse is how organizations will organize historical and other useful data used in making smart business decisions. The data warehouses have several components which are dependent upon distinct characteristics. Broken down, a good data warehouse is consistent with good, viable, cleaned data that comes from multiple sources from inside and outside the organization and has been organized by an ETL process to fit into the data warehouse schema. As with any dataset, however, data security becomes an issue, whether it is from accidental unauthorized access or malicious intentional invasion. Good security is essential to protecting some of the most valuable information in business today.

Acknowledgements

The authors would like to thank Technology and Healthcare Solutions for support.

Conflict of Interest

Any financial interest or any conflict of interest does not exist.

Bibliography

1. Oukhouya L., *et al*. "A generic metadata management model for heterogeneous sources in a data warehouse". In E3S Web of Conferences 297 (2021): 01069.

2. Ali TZ., *et al*. "A Framework for Improving Data Quality in Data Warehouse: A Case Study". In 2020 21st International Arab Conference on Information Technology (ACIT) (2020): 1-8.

3. Kraynak J and Baum D. "Cloud data warehousing". (2nd ed). John Wiley and Sons (2020).

4. Solodovnikova D., *et al*. "Managing Evolution of Heterogeneous Data Sources of a Data Warehouse". In ICEIS 1 (2021): 105-117.

5. Chowdhury R., *et al*. "Proposed Formula Based on Study of Correlation Between Hub and Spoke Architecture and Bus Architecture in Data Warehouse Architecture, Based on Distinct Parameters". *Research Journal of Science and Technology* 3.3 (2011): 143-146.

6. Hadzhiev V and Rashidov A. "A Hybrid Model for Structuring, Storing and Processing Distributed Data on the Internet". In 2021 International Conference Automatics and Informatics (ICAI) (2021): 82-85.

7. Choudhary RK. "Key organizational factors in data warehouse architecture selection". *Vivekananda Journal of Research* 1.1 (2012): 24-32.

8. Kechar M and Bahloul SN. "An access control system architecture for xml data warehouse using xacml". In Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication (2015): 1-6.

9. Pavlenko E., *et al*. "Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies". *BMC Medical Informatics and Decision Making* 20.1 (2020): 1-3.

10. Arora A and Gosain A. "Intrusion detection system for data warehouse with second level authentication". *International Journal of Information Technology* 13.3 (2021): 877-887.

11. Friedrichs M. "BioDWH2: an automated graph-based data warehouse and mapping tool". J*ournal of Integrative Bioinformatics* 18.2 (2021): 167-176.

12. Shahid A., *et al*. "Big data warehouse for healthcare-sensitive data applications". *Sensors* 21.7 (2021): 2353.

13. Bergers J., *et al*. "Dwh-dim: a blockchain based decentralized integrity verification model for data warehouses". In 2021 IEEE International Conference on Blockchain (Blockchain) 6 (2021): 221-228.

14. Wang J and Liu B. "Design of ETL tool for structured data based on data warehouse". In Proceedings of the 4th International Conference on Computer Science and Application Engineering (2020): 1-5.

15. Fana WS., *et al*. "Data Warehouse Design With ETL Method (Extract, Transform, And Load) for Company Information Centre". *International Journal of Artificial Intelligence Research* 5.2 (2021): 132-137.