International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I9P101

Volume 71 Issue 9, 1-6, September 2023
© 2023 Seventh Sense Research Group®

Original Article

Evolution of Enterprise Data Warehouse: Past Trends and Future Prospects

Sivakumar Ponnusamy

Senior Data Engineer, Cognizant Technology Solutions, Richmond, VA, USA.

Received: 03 July 2023 Revised: 18 August 2023 Accepted: 01 September 2023 Published: 15 September 2023

Abstract *- Data Warehousing has evolved over the past few decades primarily due to the exponential growth of data that traditional system is unable to handle and secondly due to technological advancement, which makes it feasible to have real-time data and cloud technology which provides unlimited storage and scalability. The journey for these changes started with the MIS (Management Information system) when data integration from various IT systems was possible. In the next stages, data repositories come into demand, and warehousing modernizes with the assistance of data mart mechanisms. The emergence of new tools and software used for the same has also given rise to Modern cloud-based SaaS data processing systems. Data lakes and data lakehouses have transformed the systems, providing greater autonomy and enabling the processing of larger volumes of data to generate insights for decision-making. The future of Datawarehouse will be based on AI and Machine Learning, which would be helpful with infrastructure scalability, cost savings, and agility, as well as increasing the reliability and usability of the data as well.*

Keywords *- Data warehouse, ETL, DataLakeHouse, Bigdata, Machine learning.*

1. Introduction

Data warehousing has been subjected to alternation and innovation as the world of data has gone through changes in terms of volume and variety of the information collected. Technological advancement is the key, urging users and developers to work on the new data models. One of the common examples of such is the high usage of cloud computing, which has changed the complete IT infrastructure. In the same way, the data warehousing techniques have also changed with the transformation in the methodologies to manage it. According to the research presented by Nambiar and Mundra (2022), advancements in cloud computing, IoT, and data analytics have caused changes in data warehousing over time. The need described by businesses can also be resolved with the help of smart tools and techniques. Businesses require data warehousing solutions that can scale to meet their performance needs. As data volumes increase, systems must handle larger query workloads and promptly deliver results. According to the findings of D. Subotić (2015), multiple factors have led to the evolution of the data warehousing methodologies and techniques. Mainly, data volume and variety of the data needed to be stored is the key element. Data warehousing systems must evolve to accommodate larger and more diversified datasets, including structured, semi-structured and unstructured data, as the amount of data generated by businesses and individuals continues to expand dramatically [2].

Dhaouadi et al. (2022) have said that the need for real-time insights has prompted data warehousing to develop to facilitate real-time data processing and analytics. This requires technologies that process and analyze data as it is generated, enabling businesses to make faster decisions. The adoption of cloud computing has changed data warehousing in addition to the other contributing elements since it offers scalable, affordable, and adaptable solutions. Cloud-based data warehouses eliminate the need for up-front infrastructure investments by enabling organizations to scale their resources up or down in response to demand. Data warehouses can now handle and analyze data more effectively thanks to advancements in hardware, including faster processors, greater memory capacity, and high-speed storage devices [3].

2. Literature Review

2.1. MIS

MIS was the major intervention in the world of data warehousing and IT integration. Businesses relied on MIS systems to generate basic reports and gain insights into their data. These systems were typically file-based and lacked the capabilities to handle large volumes of data. Mishra et al. (2015) described in their research that MIS systems laid the groundwork for structured data management practices. They introduced the idea of arranging data into clearly defined fields and tables, consistent with data warehousing's structured nature. It was simpler to gather, store, and retrieve data from numerous sources using this method. It pulled data from various departments within an organization to generate reports and provide insights. This need for data integration and consolidation paved the way for the data warehousing concept, where data from different sources is brought together into a central repository for analysis [4].

Varajão et al. (2022) have given the justification for the rise of MIS and caused the rise of data warehousing with its assistance. They stated that MIS systems emphasized the importance of historical data for decision-making. Data warehouses built upon this principle by storing historical data over time, enabling trend analysis, historical comparisons, and predictive modeling. This system is primarily focused on generating predefined reports and basic analytics. Data warehousing expanded on this concept by providing more advanced reporting and analytics capabilities, allowing users to perform complex queries, drill down into data, and create custom reports. More sophisticated Business Intelligence (BI) tools were developed with the evolution of data warehousing. These tools allow users to create interactive dashboards and visualizations and perform ad-hoc queries for deeper insights. MIS is also significant in usage, highlighting the need for accurate and consistent data for effective decision-making. Data warehousing solutions adopted data quality practices to ensure data accuracy, integrity, and consistency, improving overall data reliability [5].

2.2. Data Warehouses

After the MIS, the computing technologies further transformed into more modern technologies, which were excessively beneficial for the data warehousing processes. Data warehousing emerged as a new concept in the late 1980s and early 1990s. A Data warehouse is subject-oriented, integrated, time-variant, non-volatile data collection used to support management decision-making processes. Data warehouses are centralized, integrated repositories that store structured data from various sources. They are designed to support analytical processing and provide a historical data view. Technologies like Online Analytical Processing (OLAP) and Extract, Transform, Load (ETL) became essential components of data warehousing solutions. Online analytical processing (OLAP) allows data analysis from different viewpoints.

OLAP provides the benefit of faster decision-making and an integrated view of data. OLAP system operates in 3-main types-MOLAP (Multi-dimensional OLAP), ROLAP (Relational OLAP) and HOLAP (Hybrid OLAP). Data modelling is data representation in a Data warehouse or OLAP cube. It stores multidimensional data as Star or Snowflake schema. Star schema consists of Facts and a Dimension table. The fact table contains numerical facts related to business processes, which refers to the Dimension table via foreign keys. Snowflake schema is an extension of star schema where some dimension table leads to one or more secondary dimension tables [6]. Business people perform basic analytical operations with OLAP cubes, such as slice, dice, Rollup, Drill down and Pivot.

2.3. Data Marts

Data marts are subsets of data warehouses that focus on specific business departments or user groups. They are designed to provide faster access to relevant data for particular use cases, making it easier for end-users to obtain the information they need without querying the entire data warehouse. As David Loshin (2013) stated in his book about data marts, he presented his opinion on data warehousing. Data marts and data warehouses differ mainly because they serve different purposes.

Data warehouses are generally used for exploratory analysis, while data marts are for formalized reporting and specific drill-down investigations. As data marts focus on the goals and needs of a specific department, they contain smaller amounts of data, but that data is highly relevant to the department's operation. Different departments might require different data mart structures due to their unique analytical or reporting needs [9,10].

Following the research presented by Edward M. Leonard (2011), the role of data marts comes after the traditional data warehouse techniques. Data marts are specialized structures developed from a data warehouse and designed to organize data for specific business purposes. This customization makes data marts a vital tool for addressing the unique data needs of different departments or business units.

Three are 3-type of data marts differ based on their relationship to the Data warehouse and source system, which feeds DataMart or Data warehouse. Dependent data marts are subsets which get loaded from the Enterprise data warehouse (top-down approach or Bill Inmon model). Independent data marts are standalone DataMart, which serves specific business domains and are combined to form Enterprise Datawarehouse (Bottom-up Approach or Ralph Kimball model). Hybrid data marts combine data from existing data warehouses and other Operational Data stores (ODS).

Moreover, the availability of numerous reporting tools makes data warehouses user-friendly. These tools empower individuals to extract data by themselves instead of waiting for others to distribute it, enhancing the efficiency of data analysis and decision-making processes. The author further emphasized the critical role of data warehouses and data marts in enabling Business Intelligence by facilitating data access, organization, and analysis. The ability to create data marts from a data warehouse and to use diverse reporting/BI tools makes data warehouses invaluable for the Business and management community to make strategic business decisions [11].

2.4. BigData

As time passed, data volumes exploded, and traditional data warehouses faced challenges in handling the sheer scale and diversity of data. Santoso and Yulia (2017) have stated how combining big data technology with data warehouses can aid the decision-making process for university management by turning raw data into actionable insights. Big Data is defined by 3V's- Volume, velocity and variety. Data generated from social media, IOT sensors, and weblogs are a few examples of Big data. It can be structured, semi-structured or unstructured data. There are valuable insights that can be derived from Bigdata, such as customer sentiments and market insights, which is impossible without implementing a big data solution. Apache Hadoop is mainly for storage, and MapReduce, Spark and other technologies are used for processing. Apache Hive is a distributed, fault-tolerant Data warehouse system that enables big data analytics. The paper concludes by pointing out the need for future developments and the implementation of institutional projects involving Big Data [7,8].

2.5. NoSQL Database

The need for the data fields had increased and required major changes in the data management domain. NoSQL databases became more popular as big data increased and more adaptable data models were required. These databases, like MongoDB, Cassandra, AWS DynamoDB, Couchbase and Hbase, provide horizontal scaling and schema flexibility, making them perfect for some large data applications. The research offered by the article from Sokolova et al. (2019) states a redesign of a database management system for a retail business company. Originally based on a traditional data model, the database system is migrated to a hybrid model that combines SQL and NoSQL databases. Adding the NoSQL database enhances the system's flexibility, scalability, and efficiency. NoSQL databases store data in a single data format, including a JSON document, rather than the traditional table structure of a relational database. NoSQL is also a type of distributed database, meaning that information is copied and stored on various servers, which can be remote or local; hence, it provides data availability and reliability. Some main types of NoSQL databases are key-value, document, graph, In-memory, wide-column and search. NoSQL databases are very good at horizontal scaling with high performance and availability. The paper also discusses the architecture of this redesigned system and its functionality, emphasizing the benefits brought by the hybrid approach, combining both SQL and NoSQL databases [12].

2.6. Modern Cloud Data System

Cloud computing revolutionized data warehousing by offering scalable, flexible, and cost-effective solutions. Cloud-based data warehouses like Snowflake, Amazon Redshift and Azure Synapse Analytics became popular choices for organizations seeking to offload infrastructure management and scale their data processing based on demand. Bhatti and Rad (2017) have described cloud computing in the way that the significant shift in the Information Technology industry from traditional relational databases to cloud databases over the last 40 years. The cloud is particularly suitable for data-intensive applications, such as storing and mining large datasets and commercial data [13].

Applications supported by cloud databases are diverse and adaptable, with many value-based data management applications like banking, online reservation, e-commerce, and inventory management being developed. However, while these databases support key features like Atomicity, Consistency, Isolation, and Durability (ACID), their use in the cloud is not straightforward. The paper's objective was to investigate the pros and cons of databases commonly used in cloud systems and to examine the challenges associated with developing cloud databases. Data security is the main issue with moving into a cloud-based Data warehouse as we cannot store NPI data in plain text; hence, it requires added complexity, such as tokenization, before storing confidential data in the cloud. Some of the challenges stated include the security risks, reliability of the cloud system for operations, and higher costs possess the challenges [14,15].

2.7. Data Lake and Data Lakehouse

Data lakes emerged as an alternative approach to traditional data warehousing. A data lake is a centralized repository that can store both structured and unstructured data in its raw form. The concept of a Data lakehouse combines data lakes with some elements of data warehousing, aiming to bridge the gap between data engineering and data analytics by providing features like ACID transactions, data indexing, and support for SQL queries directly on raw data. The modern data architecture ensures the combination of elements from both a data warehouse and a data lake. It is known as a "data lakehouse," which is a hybrid of data warehouses and data lakes. This new concept seeks to break down silos between data engineers and data scientists to foster a collaborative environment, ultimately enabling more effective data analysis [16].

Further, details were given to explain the procedure on which data lakes and lake houses operate. The data lake house is designed to handle both structured and unstructured data, providing a unified platform for data engineers and data scientists to work together. Previously, these two roles tended to operate in separate domains, with data engineers mostly working with structured data in data warehouses and data scientists preferring data lakes for their versatility in handling both structured and unstructured data. The datalakehouse merges these domains, eliminating duplication of effort and speeding up the process of finding value in the data [17].

Such a mechanism also brings improvements in data management. It is capable of handling diverse types of data, including structured, semi-structured, and unstructured data, all in a cost-effective manner. This ability, combined with data diversity, reduces the risk of data loss and enhances data recovery and availability. The lake house paradigm fosters an environment conducive to not only descriptive and predictive reporting but also prescriptive reporting, which provides advice on potential outcomes and next steps. Faster access to shared, secure, and connected data enables businesses to align with modern analytics and gain insights more quickly.

Moreover, it supports the need for faster development and productization, essential for businesses seeking to extract value from their data scientists. With data scientists spending less time on data preparation, they can focus more on modeling data and deriving insights from it. This agility and speed are key for organizations wishing to mature their business reporting and analytics practices. The lakehouse provides a conducive environment for machine learning and AI operations. With data's increasing volume and diversity, organizations leverage machine learning and AI to analyze and interpret data effectively. The lakehouse offers a "data playground" for data scientists, allowing them to build advanced analytics models using large quantities of structured and unstructured data [18].

2.8. Real-time Data Processing

The need for data processing technologies has been enhanced with the intervention of online businesses and the boom of usage in enterprises. With the demand for real-time analytics, modern data systems have focused on providing real-time data processing capabilities. Technologies like Apache Kafka, Apache Flink, and Apache Spark Streaming allow for real-time data ingestion, processing, and analytics [19].

2.9. Data Governance

As data becomes more critical, the need for proper data governance and security measures has grown. Organizations now focus on implementing robust data governance frameworks to ensure data privacy, compliance, and data quality. The introduction to the concepts related to data governance has enhanced the data warehousing techniques, which has resulted in the minimization of data-related research and also made the way for the governance, its usage, and utility that can be gained. Research presented by Abraham et al. (2019) stated that data governance refers to managing data to enhance its value and minimize data-related costs and risks. In the context of data governance, a data warehouse could be an essential tool since it offers a structured repository for storing and analyzing data, which could be crucial in implementing effective data governance strategies. This can enable organizations to maintain data quality, consistency, security, and privacy, which are significant components of data governance. Nonetheless, for specific references to data warehousing, it might be beneficial to review other parts of the text or a different text [20].

2.10. AI and Machine Learning Integration

The trend of development and installation of modern tools and technologies are a part of daily targets for large firms. As plenty of data is available, and not just commercial enterprises, non-profit organization have made their dependence upon the data. Nowadays, The decision-making process depends entirely on the data available, which could be qualitative or quantitative. There has been a push to integrate AI and machine learning capabilities into data warehousing solutions in recent years. This integration enables businesses to leverage advanced analytics, predictive modeling, and machine learning algorithms to gain deeper insights from their data. A research paper presented by Sizwe et al. (2015) examines the role of knowledge engineering in enhancing organizational capabilities and adapting to unpredictable market environments. It stresses the importance of transforming collected data into real-time information to support successful decision-making [21] and delivers timely results. As data becomes increasingly crucial, data governance practices have gained prominence, ensuring data quality, security, and compliance. The integration The authors emphasize the necessity of integrating artificial intelligence into data warehousing and data mining. The integration of data warehousing holds promise as machine learning and AI algorithms drive advanced analytics, automated data preparation, intelligent query optimization, and even more accessible user interactions through natural language. The authors emphasize the necessity of integrating artificial intelligence into data warehousing and data mining. The integration of AI can help in analyzing and interpreting vast amounts of data, which can be a complex and challenging process. The research aims to explore suitable techniques, technologies, and trends to facilitate this integration, providing an insightful overview of data warehousing and data mining. It also aims to highlight the techniques and limitations of analyzing and interpreting large amounts of data. In this context, AI and machine learning can be crucial tools for making sense of vast, complex data sets and extracting meaningful insights [22].

3. Future Prospects

The potential for future data warehousing relies on implementing advanced analytical steps. Data warehouses will leverage machine learning and AI algorithms to provide more advanced analytics and predictive insights. ML models can be trained on historical data to make predictions and recommendations, enabling organizations to predict trends and make informed decisions. There is also a possibility that Machine learning algorithms can be used to automate the process of data preparation, including data cleaning, transformation, and integration. This will reduce the manual effort required to structure and format data for analysis. The generation and ease of access to AI have made everything easier in this world. So comes the data warehousing as well, as it is expected that AI-powered query optimization will become more sophisticated, automatically selecting the most efficient query execution plans based on data distribution, workload patterns, and system performance. AI and Machine Learning applications in Data warehousing will help optimize resource allocation & usage, thereby reducing operational costs for Datawarehouse. Data warehouses will integrate NLP capabilities to enable users to query and interact with data using natural language. This will make data analysis more accessible to a wider range of users, including business users without much technical expertise, thus reducing the learning curve for querying data [23].

4. Conclusion

The evolution of data warehousing has been a dynamic journey driven by the increasing volume, velocity, and variety of data. Technological advancements, including cloud computing, big data solutions, and AI integration, have played pivotal roles in shaping the modern data warehousing landscape. This evolution has empowered businesses to make more informed decisions, gain competitive advantages, and enhance their overall operations. Research has highlighted the extensive factors influencing data warehousing changes over time, reflecting the growing needs of businesses that are now effectively addressed through smart tools and techniques. The scalability imperative is ever more relevant, with data volumes continually on the rise. In response, data warehousing solutions have adapted to handle larger query workloads efficiently by incorporating AI and machine learning capabilities into data warehousing. This has marked a transformative phase, enabling deeper insights and automation of various processes. The future of data warehousing holds promise as machine learning and AI algorithms drive advanced analytics, automated data preparation, intelligent query optimization, and even more accessible user interactions through natural language processing. As technology continues to advance, data warehousing's evolution is poised to remain aligned with business needs, enabling organizations to harness data's full potential for strategic decision-making and innovation.

References

[1] Athira Nambiar, and Divyansh Mundra, “An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management,” *Big Data and Cognitive Computing*, vol. 6, no. 4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Danijela Subotić, “Data Warehouse Schema Evolution Perspectives,” *New Trends in Database and Information Systems II*, pp. 333-338, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Asma Dhaouadi et al., “Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons,” *Data*, vol. 7, no. 8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Lahar Mishra, Ratna Kendhe, and Janhavi Bhalerao, “Review on Management Information Systems (MIS) and its Role in Decision Making,” *International Journal of Scientific and Research Publications*, vol. 10, no. 5, pp. 1-5, 2015. [Google Scholar] [Publisher Link]

[5] João Varajão, João Carlos Lourenço, and João Gomes, “Models and Methods for Information Systems Project Success Evaluation –A Review and Directions for Research,” *Heliyon*, vol. 8, no. 12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Karwan Jameel, Abdulmajeed Adil, and Maiwan Bahjat, “Analyses the Performance of Data Warehouse Architecture Types," *Journal of Soft Computing and Data Mining*, vol. 3, no. 1, pp. 45-57, 2022. [Google Scholar] [Publisher Link]

[7] Leo Willyanto Santoso, and Yulia, “Data Warehouse with Big Data Technology for Higher Education,” *Procedia Computer Science*, vol. 124, pp. 93-99, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] V. Rathika, and L. Arockiam, “General Aspect of (Big) Data Migration Methodologies,” *SSRG International Journal of Computer Science and Engineering*, vol. 1, no. 9, pp. 1-5, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[9] David Loshin, *Business Intelligence the Savvy Manager's Guide*, 2nd ed., Elsevier, 2012. [Google Scholar] [Publisher Link]

[10] Muhammad Khalid, “Challenges of Dimensional Modeling in Business Intelligence Systems,” *International Journal of Computer & Organization Trends*, vol. 5, no. 3, pp. 30-31, 2015. [CrossRef] [Publisher Link]

[11] Edward M. Leonard, B.S., “*Design and Implementation of an Enterprise Data Warehouse*,” Thesis, Marquette University, 2011. [Google Scholar] [Publisher Link]

[12] Marina V. Sokolova, Francisco J. Gómez, and Larisa N. Borisoglebskaya, “Migration from an SQL to a Hybrid SQL/NoSQL Data Model,” *Journal of Management Analytics*, vol. 7, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Junaid Hassan et al., “The Rise of Cloud Computing: Data Protection, Privacy, and Open Research Challenges—A Systematic Literature Review (SLR),” *Computational Intelligence and Neuroscience*, pp. 1-26, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Harrison John Bhatti, and Babak Bashari Rad, “Databases in Cloud Computing: A Literature Review,” *International Journal of Information Technology and Computer Science*, vol. 9, no. 4, pp. 9-17, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[15] Soukaina Ait Errami et al., “Spatial Big Data Architecture: From Data Warehouses and Data Lakes to the LakeHouse,” *Journal of Parallel and Distributed Computing*, vol. 176, pp. 70-79, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mitesh Athwani, “A Novel Approach to Version XML Data Warehouse,” *SSRG International Journal of Computer Science and Engineering*, vol. 8, no. 9, pp. 5-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Philipp Wieder, and Hendrik Nolte, “Toward Data Lakes as Central Building Blocks for Data Management and Analysis,” *Front Big Data,* vol. 5, pp. 1-18, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Dave Langton, The New Data Lakehouse: An Overdue Paradigm Shift for Data, Database Trends and Application, 2022. [Online]. Available: https://www.dbta.com/BigDataQuarterly/Articles/The-New-Data-Lakehouse-An-Overdue-Paradigm-Shift-for-Data-151318.aspx

[19] Abdul Jabbar, Pervaiz Akhtar, and Samir Dani, “Real-Time Big Data Processing for Instantaneous Marketing Decisions: A Problematization Approach,” *Industrial Marketing Management,* vol. 90, pp. 558-569, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Rene Abraham, Johannes Schneider, and Jan vom Brocke, “Data Governance: A Conceptual Framework, Structured Review, and Research Agenda,” *International Journal of Information Management,* vol. 49, pp. 424-438, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Nelson Sizwe. Madonsela, Paulin. Mbecke, and Charles Mbohwa, “Integrating Artificial Intelligence into Data Warehousing and Data Mining,” *Proceedings of the World Congress on Engineering and Computer Science,* vol. 2, pp. 1-5, 2015. [Google Scholar] [Publisher Link]

[22] Maria F. Chan1, Alon Witztum, and Gilmer Valdes, “Integration of AI and Machine Learning in Radiotherapy QA,” *Frontiers in Artificial Intelligence*, vol. 3, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Gizem Turcan, and Serhat Peker, “A Multidimensional Data Warehouse Design to Combat the Health Pandemics,” *Journal of Data, Information and Management*, vol. 4, pp. 371-386, 2022. [CrossRef] [Google Scholar] [Publisher Link]