Cloud DW Migration Dilemma: Migrate Legacy DW or Build New DW

Hari Prasad Bomma

Data Engineer, USA
haribomma2007@gmail.com

Abstract

As organizations grapple with the ever growing volume and complexity of data, the need for efficient and scalable data management solutions has become increasingly paramount. Traditional data warehouses, while effective in their time, often struggle to keep up with the demands of modern big data analytics. In this research paper, we will explore the potential benefits and challenges of two common approaches one being migrating legacy data warehouse and another building a new era cloud based data warehouse. This paper aims to provide a holistic understanding of the evolving landscape of data warehousing and the potential benefits and challenges associated with adopting new era data warehouse architecture.

Keywords: Data warehouse, Data Migration, Real time Streaming, Data Lake house, Batch Processing, Massive Parallel Processing (MPP), Structured Data, Unstructured data

Introduction:

Traditional data warehouses have served organizations well for decades, providing a centralized repository for structured data and enabling advanced analytics and reporting. However, these systems are often rigid and require significant upfront design and planning before data can be integrated and used [1]. Traditional data warehouses are struggling to keep up with the sheer volume and variety of data generated by modern business operations. As organizations increasingly rely on unstructured data, real time streams, and other non traditional data sources, the limitations of the traditional data warehouse model become more pronounced.

One of the primary challenges associated with traditional data warehouses is their lack of flexibility. These systems are designed to handle structured data, which fits neatly into rows and columns. As a result, they are not well suited to manage unstructured data types, such as text, images, or sensor data, now increasingly prevalent in many industries. Integrating these diverse data sources into a traditional data warehouse requires complex, time consuming transformations and often compromises the speed and efficiency of data processing.

The rigid architecture of traditional data warehouses makes it difficult to adapt to the rapidly changing needs of modern businesses. Customizing these systems to accommodate new data sources or analytics requirements involves lengthy design and implementation processes, hindering organizations' ability to respond quickly to market changes and new opportunities. This limitation is particularly problematic in today’s fast paced business environment, where agility and adaptability are crucial for maintaining a competitive edge.

Additionally, traditional data warehouses often lack the capability to process and analyze data in real time. Modern business operations increasingly rely on real time data streams to make informed decisions quickly. Traditional data warehouses, with their batch processing approach, are not equipped to handle this demand for immediate insights. Real time data processing requires more advanced architectures, such as data lakes or streaming analytics platforms, which can ingest, store, and analyze data with minimal latency.

Another significant limitation is the cost associated with maintaining traditional data warehouses. These systems generally require substantial investment in hardware, software, and manual labor for ongoing maintenance and upgrades. As data volumes and complexity grow, the operational expenses associated with traditional data warehouses can become prohibitively high, prompting organizations to seek more cost effective alternatives.

Being said traditional data warehouses have been instrumental in the evolution of data management and analytics, their inherent limitations are becoming more pronounced in the face of modern data challenges. Organizations are looking towards more flexible, scalable, and cost effective solutions to meet the demands of today's data driven landscape.

Data Migration Considerations:

Data migration is the process of transferring data from one system to another, and it is a common activity in the implementation of business applications [2]. The migration process involves several steps, including data validation, extract, transform, and load processes, and the use of right tools to facilitate the migration. Data migration can be a very complex and time taking process. At the same time it is essential to ensure that the data is accurately transferred and the new system is capable to meet the organization's requirements.

The design of a data warehouse is a critical step in the implementation process. The first step in designing a data warehouse is to analyze the data sources and identify the user's needs [3]. The organization of data within the warehouse is also an important consideration, as it can impact the efficiency and effectiveness of the system [4]. Migration from an existing data warehouse can be a cost effective option, as it leverages the existing infrastructure and data. However, it can also be a complex and time consuming process, and it may not provide the same level of flexibility and scalability as building a new data warehouse.

Building a new data warehouse can provide organizations with a more flexible and scalable solution, offering several key benefits. Modern data warehousing solutions are designed to handle a wide range of data types, including structured, semi structured, and unstructured data. This flexibility allows organizations to integrate various data sources, such as text, images, videos, social media, IoT sensor data, and more. Unlike traditional data warehouses, modern solutions can adapt to different data formats and types without requiring extensive transformations and restructures. This agility ensures organizations can incorporate new data sources quickly and efficiently, keeping pace with the evolving business landscape. Additionally, new data warehouse architectures are designed to scale seamlessly with the growing volume and complexity of data. Cloud based data warehouses, in particular, offer on demand scalability, allowing organizations to adjust their storage and processing capabilities according to current needs. This elastic scalability ensures that the system can handle large datasets and high throughput data streams without compromising performance.

Modern data warehousing solutions leverage advancements in cloud technology to provide cost effective storage and computing solutions. Cloud data warehouses typically operate on a pay as you go

model, allowing organizations to scale their usage based on actual demand. This model helps reduce operational costs by eliminating the need for upfront investments in hardware and maintenance of on premise infrastructure. Additionally, cloud based solutions often include built in redundancy, automated backups, and disaster recovery features, mitigating additional costs associated with data loss and system downtime. New data warehousing solutions incorporate advanced technologies and architectures that significantly improve performance. These include columnar storage formats, massively parallel processing (MPP), and in memory computing. Columnar storage enhances query performance by organizing data in columns instead of rows, making it faster to retrieve and aggregate data. MPP enables the system to distribute complex queries across multiple processors, reducing processing time. In memory computing allows data to be stored and processed within the system's RAM, minimizing latency and promoting faster analysis. Moreover, modern data warehouses come equipped with robust security features to protect sensitive information, such as advanced encryption techniques, user authentication and authorization, and comprehensive auditing capabilities. These solutions are also designed to seamlessly integrate with advanced analytics tools and AI platforms, facilitating the use of machine learning models, predictive analytics, and other data driven techniques to glean insights from the data.

Emergence of Data Lakes:

In contrast to the traditional data warehouse, the "data lake" approach offers a flexible and scalable solution for managing big data. Data lakes ingest raw data from a variety of sources and store it in its native format in a distributed file system, such as the Hadoop Distributed File System. This "schema on read" approach eliminates the need to transform and load data into a pre defined schema, allowing for more agile and iterative data exploration. The flexibility of data lakes is advantageous for organizations dealing with diverse data types. Storing data in its raw form enables the inclusion of various data sources such as text, images, social media posts, and IoT sensor data which can then be processed and analyzed as needed for richer insights. Scalability is another key benefit, as data lakes built on distributed file systems can scale horizontally, adding storage and processing power as data volumes grow, thus handling expansive datasets without performance bottlenecks. This capacity supports real time analytics and big data applications that require rapid data processing. The "schema on read" paradigm facilitates ad hoc queries and on the fly data transformations, enabling analysts to experiment with different models and techniques without extensive upfront design. This dynamic approach allows organizations to uncover patterns and trends, fostering innovation and enabling timely decision making. Additionally, data lakes provide a unified repository that supports various use cases, from batch processing and machine learning to real time analytics and data science, promoting better collaboration and more integrated business intelligence efforts

Methodology:

When it comes to data management, organizations must decide whether to migrate from an existing data warehouse or build a new one. This choice hinges on the specific needs and requirements of the organization, considering different key factors. Each organization must carefully evaluate the aspects to determine the most suitable approach for their unique data management needs. The factors include the organization's budget, the complexity of existing systems, the desired flexibility to handle various data types, and the level of security required to protect sensitive information.

Migrating from an existing data warehouse can be a more cost effective option, as it leverages the existing infrastructure and data [3]. However, building a new data warehouse may require a higher upfront investment, but it can provide a more flexible and scalable solution in the long run.

Parameters to consider in both cases:

Cost: Building a new data warehouse involves high initial investments in hardware, software, and development resources. While migrating a legacy system may lower these costs, the complexity of the data migration can still result in considerable expenses.

Data Model: Existing data models often bring back legacy system issues like duplicity and redundancy. Building new data models can disrupt underlying dependencies and reports. To address these challenges, adopt only clean tables or columns if migrating. Alternatively, design a new data model for the data warehouse to provide futuristic support and solutions.

Data Migration: Migrating complete data from legacy systems sounds easy but is a tedious and laborious task. Matching source and target systems is complex while incremental loads are processing. For building a new data warehouse, prioritize adopting clean tables first.

Development and Testing: During migration, developmental efforts are minimal, but testing efforts are extensive. Conversely, building a new data warehouse involves high development and testing efforts.

Reporting: Migrating legacy systems to the cloud will ensure reports function as expected, but you'll end up with outdated reports. Building a new data model and data warehouse will support futuristic reports, but you'll need to create plan and build new reports alongside the new data warehouse.

Vendor Support: Make sure to check the existing vendors are adaptable to new systems.

Flexibility and Scalability: A new data warehouse can be designed to be more flexible and scalable, allowing for the incorporation of new data sources and the adaptation to changing business requirements.

Security: Data security is a critical consideration in any data management solution. Both migration and building a new data warehouse should prioritize data security, ensuring the integrity and confidentiality of the data.

Future Proofing: It's essential to think about future growth and technological advancements. Building a new data warehouse allows for the incorporation of cutting edge technologies that can adapt to future needs, whereas migrating an existing warehouse might be limited by outdated technology.

Conclusion:

In conclusion, the decision to migrate from an existing data warehouse or to build a new one depends on several factors, including cost, complexity, flexibility, and security. Organizations must carefully evaluate these factors to make an informed decision that aligns with their data management objectives. Ultimately, the right approach will empower organizations to achieve greater agility, enhance their analytics capabilities, and secure their sensitive information. A well planned transition ensures that organizations can capitalize on the benefits of modern data architectures, driving innovation and maintaining a competitive edge in the digital landscape.

Reference:

1. Saddad, E., El-Bastawissy, A., Hoda, M., & Hazman, M. (2020). *"Lake Data Warehouse Architecture for Big Data Solutions"*. In International Journal of Advanced Computer Science and Applications (Vol. 11, Issue 8).https://doi.org/10.14569/ijacsa.2020.0110854
2. Saranya, N., Brindha, R., Aishwariya, N., Kokila, R., Matheswaran, P., & Poongavi, P. (2021). *"Data Migration using ETL Workflow"*. In 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) (p. 1661). https://doi.org/10.1109/icaccs51430.2021.9441840

3. Dicko, A., Barro, S. G., Traoré, Y., & Staccini, P. (2021).*" A Data Warehouse Design for Dangerous Pathogen Monitoring"*. In Studies in health technology and informatics. IOS Press. https://doi.org/10.3233/shti210198

4. Dupor, S., & Jovanović, V. (2014). *"An approach to conceptual modelling of ETL processes "*(p.1485). https://doi.org/10.1109/mipro.2014.6859801

5. Martinho, B., & Santos, M. Y. (2016). *"An Architecture for Data Warehousing in Big Data Environments"*. In Lecture notes in business information processing (p. 237). Springer Science+Business Media. https://doi.org/10.1007/978-3-319-49944-4_18

6. Panwar, A., & Bhatnagar, V. (2019). *"Data Lake Architecture. In International Journal of Organizational and Collective Intelligence"* (Vol. 10, Issue 1, p. 63). IGI Global. https://doi.org/10.4018/ijoci.2020010104

7. Liu, R., Isah, H., & Zulkernine, F. (2020). *"A Big Data Lake for Multilevel Streaming Analytics."* https://doi.org/10.1109/ibdap50342.2020.9245460