DOI: https://doi.org/10.5281/zenodo.10897094

DATA MODELS AND ARCHITECTURES IN DATA WAREHOUSING

Abdurashitova Muniskhon
Department of Control and Computer Engineering, Turin Polytechnic University in Tashkent, Uzbekistan
Email: abdurashitova.munisxon@gmail.com

ABSTRACT

The separation between transactional computing and data analysis ensures that the transactional systems responsible for processing and recording business transactions can operate efficiently and without interruption. Meanwhile, data analysis systems can focus on processing and analyzing large amounts of data to extract valuable insights that can inform business decisions. By employing different data warehousing architectures and models, organizations can separate these two functions, preventing the resource-intensive data analysis processes from interfering with the smooth functioning of the transactional systems.

Keywords: *Data warehousing, Two level architecture, Three level architecture, ETL processes, Data marts.*

I OBJECTIVE

Data warehousing is a crucial aspect of modern-day enterprises that must analyze and process vast amounts of data to drive informed decisions. At the heart of every data warehouse is its architecture that forms the foundation for the entire system. The data warehouse architecture comprises various components, such as data sources, ETL (Extract, Transform, Load) processes, data models, and storage systems, all working to ensure efficient and accurate data processing. A well-designed data warehouse architecture helps businesses store, manage, and retrieve data seamlessly while providing a reliable data analysis, reporting, and visualization platform.

II INTRODUCTION

A data warehouse is a centralized data repository used to support business decision-making activities. It is designed to facilitate the efficient querying and analysis of large volumes of data by consolidating data from multiple sources into a single, coherent data model. A data model is a conceptual representation of data structures used to organize, describe, and define the relationships between different data elements[1]. It provides a way to standardize data across various systems and applications, making it easier to manage and analyze data. OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing) are two different types of data models used in data warehousing. OLAP and OLTP models serve various purposes, and it's essential to understand their differences when designing a data warehouse.

III MAIN PART

Many different database management systems are available, each with its own strengths and weaknesses. These systems organize data in a way that makes it easy to find, retrieve, and use when needed. The choice of which database management system to use will depend on a number of factors, such as the specific needs of the organization, the size of the database, and the level of security required. Several different types of data management systems exist, and we will discuss the most popular ones that are frequently used in different organizations.

Relational database management systems (RDBMS): These are the most common data management systems and are used to store structured data in a tabular format[2]. Examples of RDBMS include Oracle, SQL Server, and MySQL.

NoSQL databases: These are non-relational databases that are designed to handle unstructured data, such as social media posts, images, and videos[3]. Examples of NoSQL databases include MongoDB, Cassandra, and Couchbase.

Data warehouses: These are used to store large volumes of data from multiple sources and are designed to support complex queries and analysis. Examples of data warehouses include Amazon Redshift, Google BigQuery, and Snowflake.

Data lakes: These are used to store large volumes of raw, unstructured data that may not have a defined structure or schema. Examples of data lakes include Amazon S3, Azure Data Lake Storage, and Google Cloud Storage.

Master data management (MDM) systems: These are used to manage and maintain a single, authoritative source of reference data for an organization. Examples of MDM systems include Informatica MDM, Talend MDM, and IBM InfoSphere MDM. These systems are essential for organizations to manage and make sense of the vast amounts of data they generate and collect. They provide a way to store and organize data so that it can be easily accessed, analyzed, and used to support business decision-making activities.

The data warehouse is designed to support OLAP processing and is optimized for querying and analysis rather than transaction processing. Data is typically extracted from various sources and transformed to meet the requirements of the data warehouse schema[4]. The data warehouse is then loaded with the transformed data, making it available for querying and analysis by business analysts and decision-makers. Employed architectures typically involve a combination of hardware and software components that work together to provide a scalable, reliable, and high-performance environment for data storage and processing. There are generally three main types of data warehouse architectures. The first one is Single-tier architecture. It is not frequently used in many systems due to the absence of separation between transactional and analytical processing. In this architecture, all the data warehouse components are located on a single server, including the ETL process, database management system, and data presentation layer. This architecture is simple and easy to implement but can be challenging to scale for large data sets. The second type of architecture is Two-tier architecture. In this architecture, the data warehouse is divided into two layers: the database management system layer or, similarly, the data warehouse level and the source layer. This architecture separates the storage and analysis of data, making it easier to scale and manage. In *Figure 1*, all the components of the two-level architecture are described. Three-tier architecture is an extended version of Two-level architecture. This architecture divides the data warehouse into three layers: the ETL process layer, the database management system layer, and the data presentation layer(source level). Compared to two-level architecture, the ETL process level is added between the data warehouse level and source level, providing greater scalability and flexibility and allowing for more advanced data processing and analysis. The ETL process is a critical component of data warehousing and refers to the process of Extracting, Transforming, and Loading data into a data warehouse. The first step in the ETL process is to extract data from various sources, such as transactional databases, flat files, and other external sources. This data is then transformed into a format that is suitable for analysis and is loaded into the data warehouse. During the transformation process, the data is cleaned, integrated, and standardized to ensure consistency and accuracy of the data. This may involve removing duplicates, filling in missing data, and converting data types[5]. Once the data has been transformed, it is loaded into the data warehouse, where it can be accessed and analyzed by users. The data is organized in a way that makes it easy to query and analyze and may involve creating hierarchies, dimensions, and fact tables. This architecture provides greater scalability and flexibility, allowing for more advanced data processing and analysis. Each architecture discussed above has its own advantages and disadvantages, and the choice of architecture largely depends on the specific needs and requirements of the organization[6].

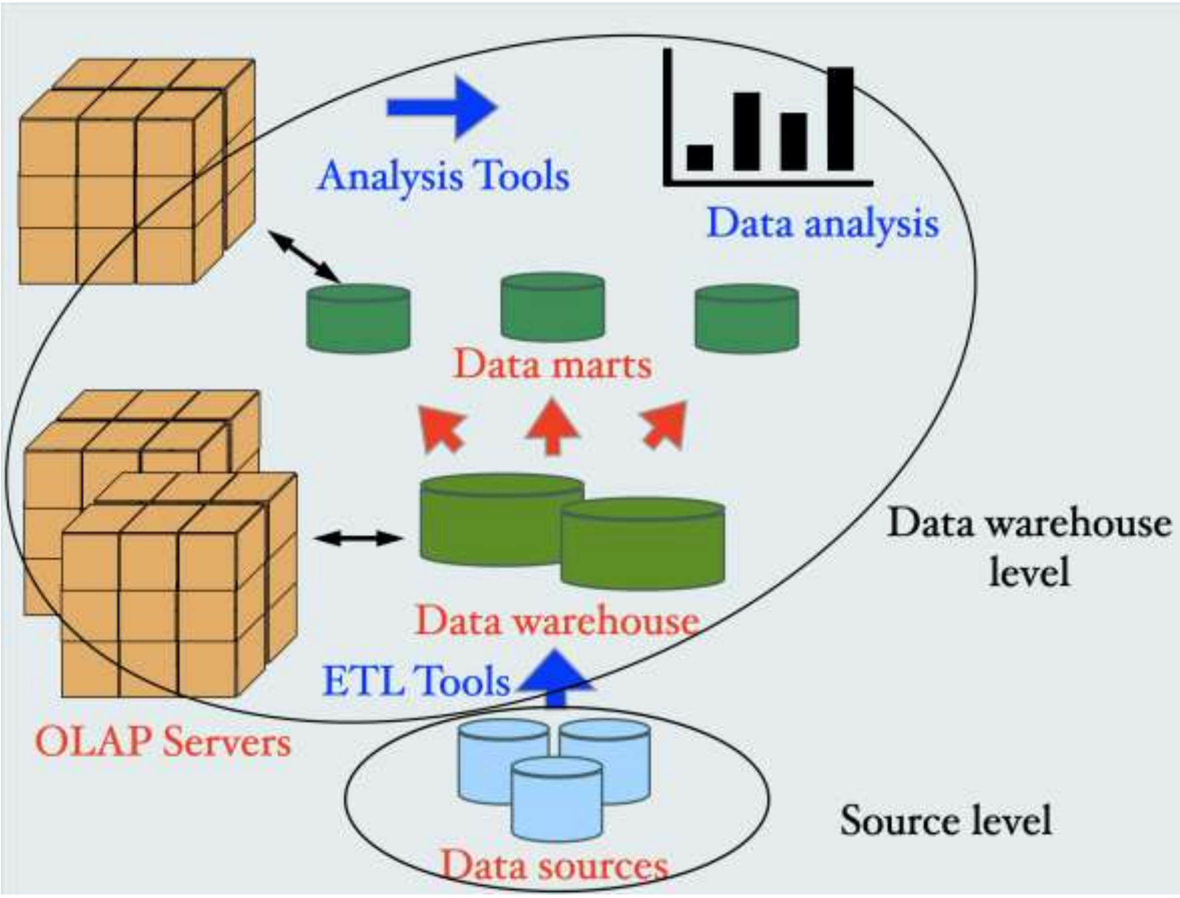

Figure 1: Data warehouse two level architecture: decoupling between Source and Data warehouse data

V CONCLUSION

In conclusion, data warehouse architectures are essential for organizations that need to store and analyze large amounts of data efficiently. While there are different approaches to building a data warehouse, they all share the same goal of providing a centralized source of high-quality data for decision-making. Regardless of the approach, a successful data warehouse architecture requires a clear understanding of the organization’s data needs, as well as strong data management and governance practices. With the right architecture in place, organizations can make informed decisions based on accurate and reliable data, ultimately leading to improved business outcomes.d decisions based on accurate and reliable data, ultimately leading to improved business outcomes.

REFERENCES

[1] https://learndatamodeling.com/blog/category/data-modeling

[2]https://blog.chromia.com/ai-meets-relational-blockchain-the-benefits-of-integrating-with-chromia

[3] https://biztech.dev/deciphering-database-development-demystifying-the-data

[4] https://serverlogic3.com/what-is-role-and-structure-of-the-data-warehousing

[5] https://www.kingswaysoft.com/resources/industry-analysis/unlocking-the-power-of-etl

[6] Aldoseri, Abdulaziz, and Khalifa Al-Khalifa. "Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges." Applied Sciences, vol. 13, no. 12, 2023, p. 7082.