Comparative Datawarehouse vs Datalake

Valdivia Guzman, Alejandra Maria   Pazos Alarcón, Christian Joshua

Farfan Colque, Mathius Omar   Condori Quispe, Yhónn Joel

April 9, 2022

Resumen

Cada día se generan enormes cantidades de datos procedentes de las tecnologías digitales y los sistemas de información. Por ello, el tratamiento de estos datos masivos requiere una arquitectura específica y un buen conocimiento de cómo manejar los datos. Los sistemas tradicionales de gestión de bases de datos ya no pueden utilizarse para este tipo de datos, ya que fueron diseñados originalmente para datos limitados y estructurados. Por otra parte, se ha desarrollado una arquitectura específica conocida como Data Lake con el fin de extraer información valiosa oculta en los datos. El objetivo principal de este artículo es explorar las dos arquitecturas, a saber, el almacén de datos y el lago de datos. Además, describe las principales diferencias y expone los factores clave de cada una.

Abstract

Each day huge quantities of data are generated from digital technologies and information systems. Therefore, processing these massive data requires a specific architecture and a good knowledge on how to handle data. Traditional databases management system can no longer be used for this type of data since they were originally designed for limited and structured data. Moreover, dedicated architecture known as Data Lake has been developed in order to extract valuable information hidden in data. The main objective of this paper is to explore the two architectures, namely, data warehouse and data lake. Furthermore, it describes the main differences and exposes key factors of each one.

I. Introduction

Relational Databases or RDBMS played a key role in making data management mainstream. They are good with highly structured, low quantity data of the pre-internet era. The advent of the internet coincided with the ambitions of large organizations to incorporate a 360-degree view of their customer database. This led to a new type of storage destination known as Data Warehouses. Today, even this storage destination has a smaller subset, popularly known as Data Lakes. When it comes to the field of Database Storage, the Data Warehouse vs Data Lake question is a relatively tough choice.

Even today very few people understand the differences between these 2 types of storage. Although Data Warehouses are good at handling structured Big Data, companies quickly realized that the Data Warehouses might not cater to the rising demand for insights into unstructured data. The 21st century witnessed a deluge of data collection by organizations not only from internal sources but also from public repositories. They needed a technology that could complement the capabilities of Data Warehouses, an extension that can facilitate the immense unstructured aspect of Big Data. As a result, Data Lakes came into existence. In this article we are going to explore both, Data Lakes and Warehouses, unfold their key differences and discuss their usage in the context of an organization.

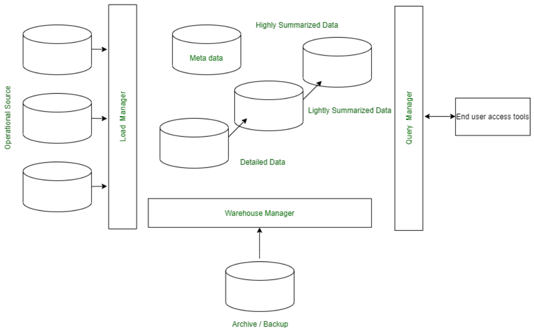
II. State of Art

i. Data Warehouse

Data Warehouse is a set of data produced for decision making, where current and historical data of potential usefulness for decision making by managers throughout the organization is stored. The data is structured and available in a form that allows analytical processing activities: OLAP, data mining, querying, reporting and other DSS applications. In exact terms Data Warehouse is defined as a collection of data, subject-oriented, integrated, time-specific and non-volatile information, to enable the decision making process by management[3].

The Data Warehouse is more than the consolidation of all the company's operational databases, as it takes into account business intelligence, external data and data associated with specific dates, making it a unique type of database. An important aspect of the Data Warehouse is that it is more of an architecture than a technology, and although there is a relationship between Data Warehousing and database technology, they are not the same, and Data Warehousing requires the support of several different types of technology[3].

ii. Data Warehouse Components



Data Warehouse Architecture[5].

ii.1 Operational Source

- An operational Source is a data source consists of Operational Data and External Data.
- Data can come from Relational DBMS like Informix, Oracle[5].

ii.2 Load Manager

- The Load Manager performs all operations associated with the extraction of loading data in the data warehouse.
- These tasks include the simple transformation of data to prepare data for entry into the warehouse[5].

ii.3 Warehouse Manage

- The warehouse manager is responsible for the warehouse management process.
- The operations performed by the warehouse manager are the analysis, aggregation, backup and collection of data, denormalization of the data[5].

ii.4 Query Manager

- Query Manager performs all the tasks associated with the management of user queries.
- The complexity of the query manager is determined by the end-user access operations tool and the features provided by the database[5].

ii.5 Detailed Data

- It is used to store all the detailed data in the database schema.
- Detailed data is loaded into the data warehouse to complement the data collected[5].

ii.6 Summarized Data

- Summarized Data is a part of the data warehouse that stores predefined aggregations.
- These aggregations are generated by the warehouse manager[5].

ii.7 Archive and Backup Data

- The Detailed and Summarized Data are stored for the purpose of archiving and backup.
- The data is relocated to storage archives such as magnetic tapes or optical disks[5].

ii.8 Metadata

- Metadata is basically data stored above data.
- It is used for extraction and loading process, warehouse, management process, and query management process[5].

ii.9 End User Access Tools

- End-User Access Tools consist of Analysis, Reporting, and mining.
- By using end-user access tools users can link with the warehouse[5].

iii. Data Warehouse Examples

In general, DWH is implemented in companies that handle large volumes of data related to customers, products or transactions. Among the sectors that make use of this tool, the following can be mentioned:

iii.1 DWH in the telecom sector

The world of telecommunications is extremely dynamic and competitive. For this reason, organizations resort to tools that allow them to study their internal productivity, the market, its changes and behaviors in the face of new technologies.

Therefore, telecommunications companies use data warehouses to store the data of millions of customers. This involves the backup of invoices, services used, records of calls made, equipment sold, among others. All this information is very useful for activities such as:

- The design of marketing strategies
- Audits in the operations area
- Service delivery analysis
- Forecasts of risks of customer leakage and others

iii.2 DWH in the mass consumer sector

Companies implement data warehousing to stay competitive in the market. In this way they can predict, for example, the amount of production they will need to meet demand in a given time range.

Retail chains can also share certain accesses to their data warehouses with their suppliers. This will give manufacturers information related to the supply of products and their sale to the end consumer.

This whole process allows coordinating management between producers and stores, in addition to accessing data that is decisive for the development of marketing campaigns.

iii.3 DWH in the transport sector

In both the travel and distribution sectors, the use of DWH is an excellent tool for storing customer information, most frequented destinations, freight management, luggage tracking, among others.

Thus, data such as travel reservations to a certain destination, or delivery times of orders, will allow the development of analysis for the creation of promotions or for diagnostics of the organization's logistics processes.

iv. Datalake

A data lake is a scalable storage and analysis system for data of any type, retained in their native format and used mainly by data specialists (statisticians, data scientists or analysts) for knowledge extraction. Its characteristics include[4]:

- a metadata catalog that enforces data quality.
- data governance policies and tools.
- accessibility to various kinds of users.
- integration of any type of data.
- a logical and physical organization.
- scalability in terms of storage and processing.

v. Datalake Components

v.1 Data Ingestion

Temporary loading layer in which data passes through basic checks before being stored in the raw data layer. It can perform[2]:

- Basic quality controls, such as possible filters according to the origin of the data, discarding unknown sources.
- Data encryption processes if required for security reasons.
- Simple metadata and traceability records by tags, storing the origin of the data, date and time of loading, format and other technical characteristics, privacy and security level, encryption algorithm, etc.

v.2 Data Storage

Layer without established schema where all data, structured or unstructured, are stored without undergoing adaptations. It is a layer that requires expert data discovery analysts using big data tools (Hive, Spark, Map Reduce, etc.)[2].

v.3 Data Processing

Once the data analysts have performed data discovery on the raw data, it may be necessary to process and adapt certain datasets to accommodate them in a layer of recurrent use. Advanced data quality, integrity and other adaptations can take place in this layer to provide a trusted layer of data exploration that can be accessed by other users[2].

v.4 Data Access

This is a more advanced layer where, finally, data is made available to business analysts. These analysts will be able to generate reports and analysis to answer business questions and support decision making[2].

vi. Datalake Examples

vi.1 Victoria University

Victoria University, an Australian public university with more than 40,000 students, is using Cazena to collect and manage the massive amounts of data generated by student interactions and their business systems. With Cazena's Instant Data Lake with Cloudera on AWS, VU has improved educational outcomes and business operations, helping to attract and retain students[1].

III. Conclusions

- There's no better way to choose which data storage platform best fits your company than to evaluate it based on your needs and business operation.
- Furthermore, data lakes and data warehouses are two inseparable components that are extremely effective when both are utilized well.
- Not to mention, data lakes are becoming more and more user-friendly while data warehouses continue to prove their worth in terms of data analysis and reporting.

References

[1] Cazena. Victoria university. 2021.

[2] Víctor Dertiano. Data lake - introducción | características y capas básicas. 2021.

[3] Rolando Alberto Gonzales López et al. *Impacto de la Data Warehouse e inteligencia de negocios en el desempeño de las empresas: investigación empírica en Perú, como país en vías de desarrollo*. PhD thesis, Universitat Ramon Llull, 2012.

[4] Pegdwendé Sawadogo and Jérôme Darmont. On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56(1):97–120, 2021.

[5] Tanushree Sharma. Implementation and components in data warehouse. 2021.