Comparative Datawarehouse vs Datalake

Valdivia Guzman, Alejandra Maria   Pazos Alarcón, Christian Joshua

Farfan Colque, Mathius Omar   Condori Quispe, Yhónn Joel

April 8, 2022

Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

I. Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

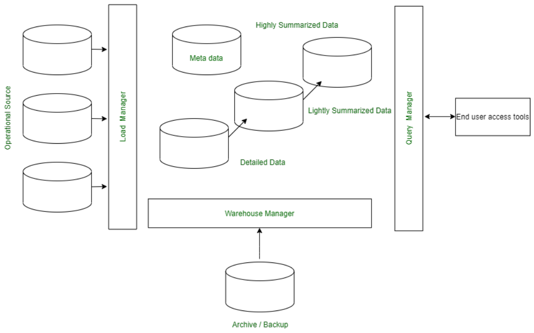
II. State of Art

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. 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.

iv. Datalake Components

iv.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.

iv.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].

iv.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].

iv.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].

v. Datalake Examples

v.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

- Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

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.