Comparativas entre Datawarehouse y Datalake

Villanueva Yucra Josue, Chavez Linares Cesar, Rodrigo Lira Alvarez

September 2022

I. Summary

The discussion of Data Lake vs. Data Warehouse is something very common among those companies that are about to implement big data solutions. The conversation about data and analysis in the field of big data quickly leads us to the Data Lake or data lake, but very often companies do not fully understand what this means and competent are the differences between Data Lake vs Data Deposit. Data lakes and data warehouses are widely used for big data storage, but although they are both data warehouses, they are not interchangeable terms. A data lake or "data lake" is a large set of raw data, which does not yet have a defined purpose. Instead, a data warehouse or "data warehouse" is a repository of data that is already structured and filtered and has been processed for a specific purpose. These two types of data storage are often confused, but they are much more different than a simple view might seem. In fact, the only thing they have in common is that they contain vast amounts of data. It is important to make the distinction, since data lakes and data warehouses serve different purposes, thus requiring a different approach to be properly optimized.

La discusión de Data Lake vs. Data Warehouse es algo muy común entre aquellas empresas que están por implementar soluciones de big data. La conversación sobre datos y análisis en el campo del big data nos lleva rápidamente al Data Lake o lago de datos, pero muy a menudo las empresas no entienden del todo lo que esto significa y competentes son las diferencias entre Data Lake vs Data Deposit. Los lagos de datos y los almacenes de datos se utilizan ampliamente para el almacenamiento de big data, pero aunque ambos son almacenes de datos, no son términos intercambiables. Un lago de datos o "lago de datos" es un gran conjunto de datos sin procesar, que aún no tiene un propósito definido. En cambio, un almacén de datos o "almacén de datos" es un depósito de datos que ya está estructurado y filtrado y ha sido procesado para un propósito específico. Estos dos tipos de almacenamiento de datos a menudo se confunden, pero son mucho más diferentes de lo que podría parecer una vista simple. De hecho, lo único que tienen en común es que contienen grandes cantidades de datos. Es importante hacer la distinción, ya que los lagos de datos y los almacenes de datos tienen diferentes propósitos, por lo que requieren un enfoque diferente para optimizarlos adecuadamente.

II. Introduction

The Internet and new technologies have caused excessive access and storage of information by customers and potential customers.

Companies are increasingly aware of the importance of this data to get to know users better and thus be able to offer them what they really ask for, and not what we think they need. This is what is called applying customer centric strategies.

This requires managing high volumes of data, both in real time and organized. For this, there is nothing better than a Data Warehouse or a Data Lake.

III. Development

i. ¿What is the Data Warehouse?

Data Warehouse is a combination of technologies and components for the strategic use of data. Collect and manage data from various sources to provide meaningful business insights. It is the electronic storage of large amounts of information designed for query and analysis rather than transaction processing. It is a process of transforming data into information.

The important functions to be performed are:

- Data extraction.
- Data cleanup.
- Data transformation.
- Upload and update data.

Advantages of a Data Warehouse

- The data is ready to be used.
- Good performance in data access.
- Most of the users of a company are operational, data warehouse is ideal for them.
- It is very suitable for generating reports and metrics.

Disadvantages of a Data Warehouse

- Higher storage costs, which implies thinking carefully about what data is really necessary.
- Not flexible to change.
- Time investment demand before storage to decide schemes, formats and use cases.

ii. ¿What is Data Lake?

A data lake is a storage repository that can store a large amount of structured, semi-structured, and unstructured data. It's a place to store all types of data in its native format with no hard limits on account or file size. Delivers a wealth of data for higher analytical performance and native integration.

Data Lake is like a big container that is very similar to real lakes and rivers. Just like in a lake, multiple tributaries enter. Similarly, a data lake has structured data, unstructured data, machine to machine, logs flowing in real time.

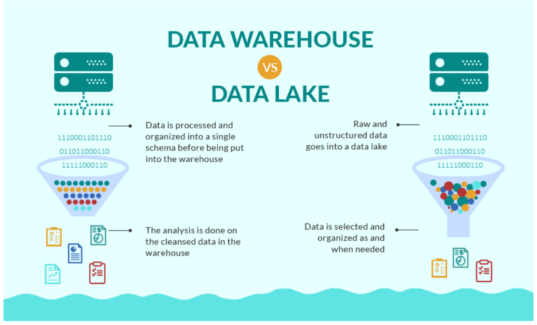
Advantages of a Data Lake

- There is no need to discard data.
- You can nurture multiple users in a company.
- Easily adapts to changes.
- By being able to integrate very different types of data, all kinds of analysis can be carried out.
- Allows you to easily add new data.

Disadvantages of a Data Lake

- It is not intended to access data performantly.
- Every time data is required, it must be transformed and curated for the intended use.
- Demand to invest in generating standards of good practices at the organizational level.

iii. Key difference between Data Lake and Data Warehouse


DATA WAREHOUSE
VS
DATA LAKE
1110001101110
011011000110
11111000110
Data is processed and organized into a single schema before being put into the warehouse
The analysis is done on the cleansed data in the warehouse
Raw and unstructured data goes into a data lake
1110001101110
011011000110
11111000110
Data is selected and organized as and when needed

iv. Comparison of Elaboration Methodologies

One of the differences that marks the comparison between Data Lake vs. Data Warehouse is that the latter generally follows a methodology like those defined by Inmon and Kimball, unlike Data Lakes.

Data Warehouse Methodologies

When designing a Data Warehouse solution, organizations come across different methodologies and approaches to follow, which should be evaluated to select the one that best suits the project requirements.

The following methodologies were designed by Ralph Kimball, Bill Inmon and Dan Linsted accordingly.

1. Dimensional (BUTTON UP / UP)

The dimensional methodology or Ralph Kimball methodology, maintains a bottom-up design, so the Data Marts are the first to be created and then they are integrated into the Data Warehouse creating a more complete storage.

It seeks that the storage of user data runs as quickly as possible. According to Kimball, a data warehouse is a copy of transactional data specifically structured for analytical queries and reporting to support decision making. With its methodology, creating data marts first provides analytical and reporting capabilities for specific business and functional processes.

Major Differentiators

- Maintaining the Data Warehouse requires only a small team of developers and data architects.
- Provides good functionality for department metrics and KPI tracking by targeting Data Marts to reporting on department or business processes.

Kimball's methodology proposes to create a business matrix that contains the common elements that are used by Data Marts, such as conformed-shared dimension, measures, etc., having this information, the user can develop solutions that support the analysis through business processes for cross-selling.

2. Relational (TOP DOWN / DESCENDING)

Bill Inmon's relational methodology shows a top-down design, where the Data Warehouse is built first and therefore the Data Marts. Placing the DW in the center of corporate information, which ensures a logical framework in the data.

Create a structure of entities ensuring that data is not repeated. This model creates a single source of truth for the entire business. Data loading becomes less complex due to the normalized structure of the model. However, using this layout to perform queries is complicated, since it includes a large number of tables and links.

This model proposes the construction of Data Marts separately for each department. All the data that enters the Data Warehouse is integrated. To ensure integrity and consistency across the enterprise, the Data Warehouse acts as a single data source for multiple Data Marts.

Main Advantages

- The Data Warehouse provides a single version of the truth, being the single source of data for the Data Marts.
- It is easier for users to understand business processes, as the logical model represents detailed business entities.
- The ETL process is easier and less prone to failure, since data updating anomalies are avoided by having very low redundancy.

3. Data Vault

It is a detail-oriented historical tracking methodology and set of uniquely linked normalized tables that support one or more functional areas of business. It is a data model that is specifically designed to meet the needs of one or more enterprise data warehouses.

Main Advantages

- Specially designed to store records. Makes the data logging process easier.
- With this methodology, it is easier to add

a new data source without modifying the existing one.
- Easily automate ETL processes.

Data Vault Architecture

Data Vault contains three basic tables:

- Hub
- Links
- Satélites

IV. Conclusions

Both Data Warehouses and Data Lakes are intended to coexist in companies that wish to base their decisions on data. As can be understood, both are complementary, not substitutes, being able to help any business to better understand the market and the consumer, in order to be able to carry out strategies based on their knowledge, with increasingly personalized communications, that is, being more customer-centric.

Tanto los Data Warehouse como los Data Lakes están pensados para coexistir en empresas que deseen basar sus decisiones en datos. Como se puede entender, ambos son complementarios, no sustitutivos, pudiendo ayudar a cualquier negocio a conocer mejor el mercado y al consumidor, para poder realizar estrategias basadas en su conocimiento, con comunicaciones cada vez más personalizadas, es decir, siendo más centrado en el cliente.

V. Recommendations

The data warehouse responds to more mature needs, when we already know what data is the most important and what kind of work we are going to do with it.

On the other hand, data lakes are more agile, cheaper and more flexible alternatives: very desirable attributes for young companies, such as your startup.

Data lakes can be especially effective when we are not sure what we want to do with all our data, but we know that the information is potentially valuable.

El almacén de datos responde a necesidades más maduras, cuando ya sabemos qué datos son los más importantesy qué tipo de trabajo vamos a hacer con ellos.

Por otro lado, los lagos de datos son alternativas más ágiles, económicas y flexibles: atributos muy deseables para empresas jóvenes, como tu startup.

Los lagos de datos pueden ser especialmente efectivos cuando no estamos seguros de qué queremos hacer con todos nuestros datos, pero sabemos que la información es potencialmente valiosa.

References

[1] https://aws.amazon.com/es/data-warehouse/

[2] https://aws.amazon.com/es/big-data/datalakes-and-analytics/what-is-a-data-lake/

[3] https://aws.amazon.com/es/lake-formation/

[4] Data Kitchen (2020). Agile Data Lake and Warehouse. Recuperado de `https://cutt.ly/Tnod6Bj`

[5] https://www.coursera.org/lecture/data-lakes-data-warehouses-gcp/build-a-data-lake-using-cloud-storage-GKMgz

[6] https://www.coursera.org/specializations/data-warehousing `https://cutt.ly/Jnofruq`