Data Lakes vs. Data Warehouses – common arguments

TEXTO COMPLETO

All data-driven organizations use data in three ways: to report on the past, to understand the present and to predict the future. Data storage isn’ t as simple as it once seemed. Intricate machines and technologies now collect an incredible breadth of data —over 2.5 quintillion bytes every day! —from equipment sensors, logs, users, consumers, and elsewhere.

Considering the volume and variety of data available today, quite a few misconceptions exist about the ways in which data can be stored. The common argument is around the two predominant types of data storage –data lakes and data warehouses. Data warehouses support reporting and analytics on historical data while data lakes support newer use cases that leverage data for machine learning, predictions, and real-time analysis.
These are the typical arguments one hears in favour of either of them –

Argument #1: You need either a Data Lake or a Data Warehouse
This gives the impression that data leaders need to choose one over the other. But the reality is that data lakes and data warehouses serve two different purposes. While both provide storage for data, they do so, using different structures, support different formats, and are optimized for different uses. A company may benefit from using a data warehouse as well as a data lake.

Data warehouses best serve businesses looking to analyse operational systems data for business intelligence. Data warehouses work well for this because the stored data is structured, cleaned, and prepped for analysis. Alternatively, data lakes allow businesses to store data in any format for virtually any use, including machine learning (ML) models and big data analysis.

Abhijit Singh, Senior Architect (Data Platform), Pharmeasy, says: “Pharmeasy deals with volume and variety of data and our biggest challenge is variety. We run hundreds of pipelines on a daily basis and this is apart from clickstream data. We use data lake to bring all the raw data into one place. After running complex joins and data modelling using Apache Hive, and then move processed data to a data warehouse to enable ad-hoc queries and analytics for data analysts.

“Data lake maintains historical data, which allows us go back to it for feature validation. We have over 5 years of data. Majority of data analytics happens on the data lake with Apache Hive. Both data warehouse and data lake are equally important. Planning for data is critical. Over time, one might lose data. Start with a data lake, build a data warehouse on top of it.”

Argument #2: Data Lakes Are Niche; Data Warehouses Aren’ t!
Artificial intelligence (AI) and ML represent some of the fastest-growing cloud workloads, and organizations are increasingly turning to data lakes to help ensure the success of these projects. Because data lakes allow you to store virtually any type of data (structured and unstructured) without first prepping or cleansing, you’ re able to retain as

much potential value as possible for future, unspecified use. This setup is ideal for more complex workloads like machine learning models where the specific data types and uses have yet to be determined.

Data warehouses may be the more well-known of the two options, but data lakes (and similar types of storage infrastructure) are likely to continue rising in popularity in conjunction with data workload trends. Data warehouses work well for certain types of workloads and use cases, and data lakes represent another option that serves other types of workloads.

Argument #3: Data Warehouses Are Easy to Use, While Data Lakes Are Complex
It’s true that data lakes require the specific skills of data engineers and data scientists (or experts with similar skill sets) to sort and make use of the data stored within. The unstructured nature of the data makes it less readily accessible to those without a full understanding of how the data lake works.

However, once data scientists and data engineers build data models or pipelines, business users can often leverage integrations (custom or pre-built) with popular business tools to explore the data. Likewise, most business users access data stored within data warehouses through connected business intelligence (BI) tools like Tableau and Looker. With the help of third-party BI tools, business users should be able to access and analyze data, whether that data is stored in a data warehouse or a data lake.

Both these types of data storage have made a strong move to the cloud. With cloud data lakes, companies are able to pay for only the data storage and compute they need. This means they are able to scale up or down as their data requires. This scalability has been a huge breakthrough in Big Data’s adoption driving the increased popularity of cloud data lakes.

Argument #4: Which of the two supports a greater diversity of use cases?
In a cloud data warehouse model, you have to transform the data into the right structure in order to make it usable. In a cloud data lake, you can load raw data, unstructured or structured, from various sources. With a Cloud Data Lake, it’s only when you are ready to process the data that it is transformed and structured.

Data warehouses are purpose-built and optimized for SQL-based access to support BI but offer limited functionality for streaming analytics and machine learning. This makes it impractical, costly, and time-consuming to ingest data in real-time, or streams of data.

Machine learning
While some data warehouses extend their SQL-based access to offer machine learning functionality, they do not offer native support to run widely available, programmatic data processing frameworks such as Apache Spark, Tensorflow and more.

In contrast, data lakes are ideal for machine learning use cases. They not only provide SQL-based access to data but also provide native support for programmatic distributed data processing frameworks like Apache Spark and Tensorflow through languages such as Python, Scala, Java and more.

Streaming analytics
Streaming analytics enables the ingestion, processing, and analysis of data in real-time without requiring data to be stored prior to analysis. Unlike other forms of data, the value of streaming data diminishes with the passage of time.

Data warehouses require sequential ETL to ingest and transform the data prior to its usage for analytics, and hence,

they are inefficient for streaming analytics. Some data warehouses support “micro-batching” to collect data often and in small increments. This stream to batch conversion increases the time between the arrival of data to its use for analytics making data warehouses inadequate for many forms of streaming analytics.

Data lakes support native streaming where streams of data are processed and made available for analytics as it arrives. The data pipelines transform the data as it is received from the data stream and trigger computations required for analytics. The native streaming feature of the data lake makes them highly suitable for streaming analytics.

Continuous data engineering
Data warehouses support sequential ETL operations, where data flows in a waterfall model from the raw data format to a fully transformed set, optimized for fast performance,

In contrast, data lakes are exceptionally strong for use cases that require continuous data engineering. In data lakes, the waterfall approach of ETL is replaced by iterative and continuous data engineering. The raw data that lands in a data lake can be accessed and transformed iteratively via SQL and programmatic interfaces to meet the changing needs of the use case. This support for continuous data engineering is critical for interactive analytics and machine learning.

Argument #5: Which of the two supports a greater diversity of data types?
With the proliferation of new types of data including IoT, social, geo-spatial, multi-media, click-stream and log data, the nature of data that we collect and use has greatly diversified. The data warehouse, invented in late 1980, was designed for highly structured data generated by business apps.

Some newer data warehouses support semi-structured data such as JSON, Parquet and XML files, they provide limited support and diminished performance for such data sets compared to structured data sets. Data warehouses do not support the storage of unstructured data.

Data lakes support native storage of all three data types –structured, semi-structured and unstructured. Structured data is ideally suited for traditional business intelligence, while semi-structured and unstructured data is useful for deeper analytics and machine learning.

Argument #6: Support for Open vs. Proprietary Data Formats
The data warehouse stores the data in a proprietary format. Once the data is stored in the data warehouse, access to this data is limited to SQL and any custom drivers provided by the data warehouse. Some data warehouses can store XML, ORC and Parquet files however these files are vendor locked and available through access mechanisms supported by the data warehouse.

In contrast, the data lake stores data in an open and standard format preventing any proprietary lock-in of data. An open data lake ingests data from sources such as applications, databases, data warehouses, and real-time streams. It stores this data in an open format, such as ORC and Parquet, that is platform-independent, machine-readable, optimized for fast access and analytics and made available to consumers without restrictions that would impede the re-use of the data.

The increase in volume, velocity, and variety of data, combined with new types of analytics and machine learning is creating a greater need for data lakes which in many cases can co-exist with data warehouses. Unlike the data warehouse’s world of proprietary formats, proprietary SQL extensions, proprietary metadata repository and lack of

programmatic access to data, an open data lake prevents vendor lock-in while supporting a diverse range of analytics.

The open data lake provides a robust and future-proof data management paradigm to support a wide range of data processing needs including data exploration, interactive analytics, and machine learning.

* Ashish Dubey
* The author is VP Solutions Architecture, Qubole.

DETALLES

Materia:	Data processing; Machine learning; Big Data; Engineering; Workloads; Proprietary; Business intelligence; Data warehouses; Pipelines; Artificial intelligence
Término de indexación de negocios:	Asunto: Machine learning Big Data Workloads Business intelligence Data warehouses Artificial intelligence
Título:	Data Lakes vs. Data W arehouses – common arguments
Título de publicación:	Dataquest; Gurgaon
Año de publicación:	2020
Fecha de publicación:	Aug 17, 2020
Editorial:	Athena Information Solutions Pvt. Ltd.
Lugar de publicación:	Gurgaon
País de publicación:	India, Gurgaon
Materia de publicación:	Computers--Data Base Management, Computers
ISSN:	0970034X
Tipo de fuente:	Revista especializada
Idioma de la publicación:	English
Tipo de documento:	News
ID del documento de ProQuest:	2434752874
URL del documento:	https://universidadviu.idm.oclc.org/login?url=https://www.proquest.com/trade-journals/data-lakes-vs-warehouses-common-arguments/docview/2434752874/se-2?accountid=198016
Copyright:	Copyright 2020 Cyber Media (India) Ltd., distributed by Contify.com

Última actualización:	2024-11-14
Base de datos:	ProQuest One Academic

ENLACES