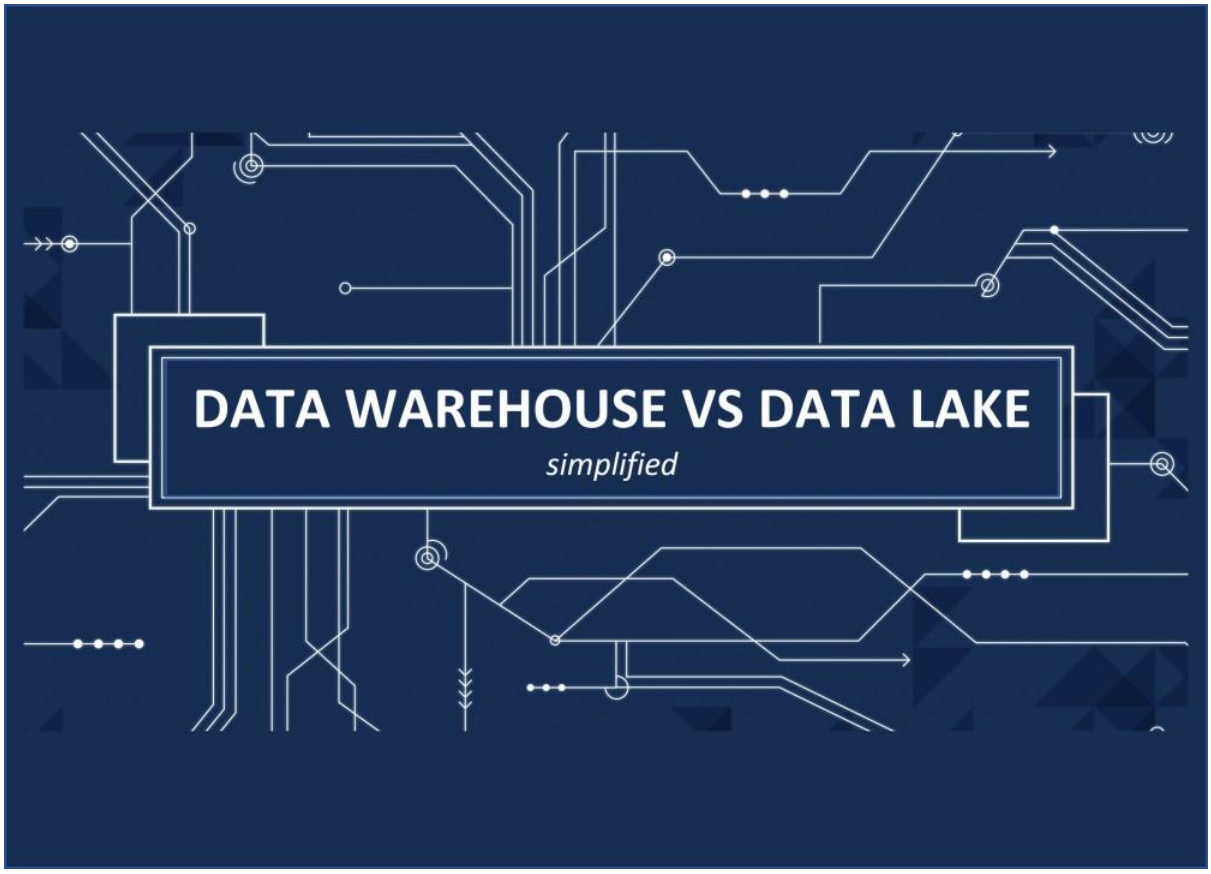
DATA WAREHOUSE VS DATA LAKE

Article by Dylan Tan
4 minute read



"Data is the new oil". "It's valuable, but if unrefined it cannot really be used". The phrase was originally coined by Clive Humby, a British mathematician and data science entrepreneur. The phrase was immortalized over the last decade as it was echoed by Gartner, The Economist, Forbes and in pretty much every article that speaks about the importance of data. From marketing and finance to logistics and product, decision-making of most large private corporations today are driven by data at all levels.

While data warehouses and data lakes may share some similar features and use cases, there are fundamental differences between the two in terms of design characteristics, data management and ideal use conditions. To understand which works best for you, we need to first understand the difference between the two.

What is a Data Warehouse (DW or DWH) or also known as Enterprise Data Warehouse (EDW)?

A data warehouse is a data management system that provides business intelligence for structured operational data. Data warehouses use predefined data schema to ingest structured data usually accomplished through an ETL (Extract-Transform-Load) process. The data source must fit into a predefined structure before it can enter the warehouse. This is also known as schema-on-write data model. The data is later connected to downstream analytical tools for Business Intelligence (BI) initiatives.

In an ETL process, data must first be extracted onto the staging layer of the data warehouse. The data is then cleaned and transformed into Dimensional Models and finally loaded onto a Data Mart (or Data Cube) for reporting, dashboarding and visualizations.

Schema-on-Write Data Model requires clear understanding of use cases as well as more time and compute resources

What is a Data Lake (DL)?

A data lake is a centralized data repository system where data from a variety of sources whether structured, semi-structured or unstructured are stored as-is in their raw format usually accomplished through an ELT (Extract-Load-Transform) process. Data lakes help eliminate data silos as it can act as single landing zone for data from multiple sources. Data lakes ingest all data types in their source format without the need for structures and pre-defined data schemas. Data is aggregated and transformed only at the point of query. This data model is known as schema-on-read.

In an ELT process, data is extracted from data sources and loaded into the data lake in its raw form. The data is only transformed at the point of query.

Schema-on-Read Data Model does not require clear understanding of use cases. Compute resources are consumed at the point of query making it more cost-effective.

When to use Data Warehouse and when to use Data Lake?

There is no hard and fast rule, but the below can be used as a quick guide:

Use Data Warehouse when you:

- have smaller data volumes that do not change too fast
- need periodic reports and dashboards like daily sales report, weekly financial report or monthly performance reports
- deal with structured data
- have full clarity of what you want to accomplish (Data can only be loaded to the data warehouse after the use for it has been defined)
- want fast and easy access to multiple operational business users

Use Data Lake when you:

- have larger data volumes that change rapidly
- need real-time/near real-time data that tells you what happened one minute or five minutes ago
- deal with structured, semi-structured and unstructured data in its raw form
- do not have full clarity of use case (Data is loaded as-is in its raw form without any transformation)
- want fast and easy access to a few Data Scientists and Power Users

If you think of a **data warehouse** as a store of bottled water – **cleansed and packaged and structured for easy consumption,** the **data lake** is a **large body of water in a more natural state**. The contents of the data lake stream in from a source to fill the lake and various users of the lake can come to examine, dive in, or take samples - James Dixon, Founder of Pentaho

Conclusion

Data Warehouse is the better option for companies that have clearly-defined use cases and operational business users who need periodic reports, dashboards and visualizations. However, data warehouse can become very expensive to maintain as the volume of data increases. Data Lake is the perfect solution for a centralized repository to eliminate data silos. It is fast to deploy and perfect for large data volumes. Since data is only transformed at the point of query, only data scientists and superpower users can consume the data. Having said that, this does not mean that data from data lake cannot be consumed by analytical software, business intelligence, and others. To harness the technological potential of both the data warehouse and the data lake, large corporations can choose to build a data lake and then have the data loaded into data warehouses and data marts.