The right house for your data workload: Data Lakehouse or Data Warehouse?

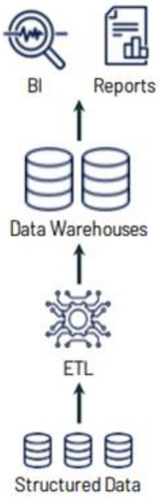

Figure 1: Data Warehouse

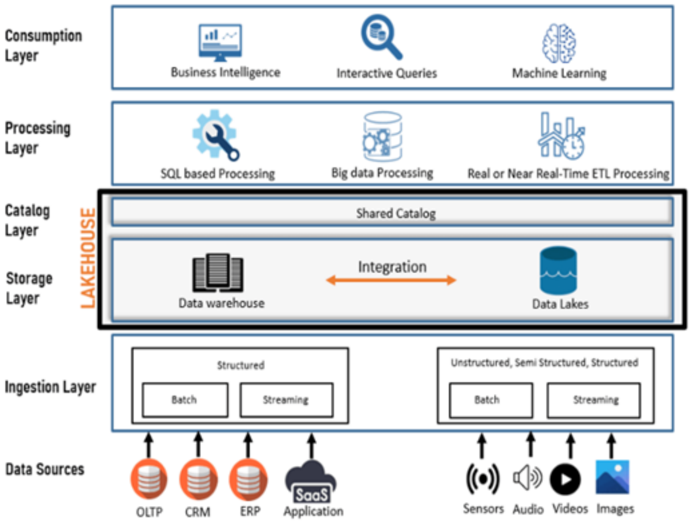

Figure 2: Data Lakehouse

With businesses having an ever increasing volume of data, it is important to carefully consider the specific needs and requirements of your organisation when selecting a data architecture that is well-suited to business needs and can support the long-term growth of your data management and analysis systems.

Both Data Lakehouse and a data warehouse are data architectures that enables organisations to effectively store, manage, and analyse data.

As there are pros and cons in both architectures, it begs the question: how do you choose the right house for the data in your business - Data Warehouse (DW) or Data Lakehouse (DL)?

DW is like a “library”, in that it stores and organises data in a structured way to support efficient access and retrieval. As a library has shelves, catalogues, and reference materials to help users find and use the books it stores, a data warehouse has tables, schemas, and query tools (usually SQL) to help users find and analyse the data it stores.

On the other hand, DL is like a “modern warehouse” which uses a centralised storage to manage and track a wide range of goods and materials. DL provides a centralised and efficient location for storing and accessing a large volume of data and empowers users through the features it has for tracking and managing the data effectively.

It's important to carefully identify your business needs first, as this will help in choosing an appropriate and cost-effective data architecture. In some cases, a DL would be a complex architecture, and could be a waste of money and time, while not providing the desired benefits and value to the organisation. At the same time, a DW may frustrate users when they need transactional processing or real-time analysis.

Some key differences between a DL and a DW are:

- **Data Types**: DL is designed to handle structured, semi-structured and unstructured data, whereas DWs are typically optimised for structured data and support a tabular data model, with data organised into rows and columns.
- **Data Volumes**: DL is well-suited for handling large volumes of data, while DW may struggle to scale to the same level.
- **Data Ingestion**: DL supports both batch and streaming data ingestion, which can be important for certain types of data and applications and enables real-time analysis. DW, on the other hand, has more frequent batch updates. DL like DW apply schemas to data, which helps in the standardisation of high volumes of data.
- **Data Transformations and modelling**: DW typically use SQL and specialised data modelling tools, such as data modelling software or ETL tools, but DL offers a more flexible and scalable approach utilising Apache Spark, SQL/Python, which can be used to create and manage data structures in the data lake.
- **Data Storage**: In DW, the same data may be stored in two or more separate places for different purposes which cause data redundancy; but DL centralises all data in a single tier and can be accessible for data engineers, data scientists and data analytics teams.
- **Data Governance**: DL provides built-in data governance features, such as data versioning and rollback, which can be useful for maintaining data integrity and traceability. DW may require additional tools or complex processes to achieve the same level of data governance.
- **SQL query Performance**: Technologies like Databricks and Spark provide SQL interface at close to interactive speeds over data lake. This has opened the possibility of DL serving analysis and exploratory needs directly, while DW need ETL processes to provide a pre-defined purpose. Moreover, DL supports distributed compute which increase the performance of large-scale data processing significantly, enables the better segmented query performance, more fault-tolerant design, and superior parallel data processing.
- **Cost**: The cost of a DL or DW will depend on various factors, including the scale of the data being managed and the specific hardware and software used; however, the highlighted feature in DL is decoupling the compute and storage. Not only can this functionality allow for substantial cost savings, but it also facilitates parsing and enriching of the data for real-time streaming and querying.

Although a DL can add value to many scenarios, it may be an over-designed model in your organisation. So, when choosing between a DW and a DL, it can be helpful to consider "what-if" design scenarios to evaluate how each option might perform in different circumstances. Some factors to consider when evaluating these scenarios might include:

- What types of data will you be storing and querying? Will you need to handle video, images, json or arrays data files?
- How large is your current dataset, and how quickly is it growing? Which approach will be able to scale to handle your data volume within your timeframe?

- How frequently do you need to update your data? Will real-time ingestion be important, or is batch processing sufficient?
- What types of data transformations will you need to perform? What flexibility and scalability do you need?
- What are your data governance requirements, such as data quality, security, and traceability? Do you already have any governance features and tools in your organisation?
- How will each option perform under different workloads and query patterns? What are the key performance metrics that are important to your use case?
- How will the total cost of ownership compare for each option, including hardware, software, and maintenance costs?

Evaluating these scenarios can help you understand the trade-offs and potential benefits of each option and make an informed decision about which one is best suited for your needs.

If you're interested in building a better data platform or want a review of your current data architecture, please contact a Fujitsu Data & AI specialist now.

Contact

Fujitsu Data & AI

+61 3 9924 3000