Data Lakehouse

A data lake house combines together the best bits of data warehouses and data lakes.

Data Lakehouses could be seen as the natural convergence of the two architectures (see https://cloud.google.com/blog/products/data-analytics/data-lake-and-data-warehouse-convergence)

Data Lake

Data Lake is the name we give to a collection of tools that are often used together to process large amounts of data. Typically it includes a storage system like S3 or HDFS and a processing system like Apache Spark or Hadoop.

- Store lots of data - often in its raw "unprocessed" form in pseudo-real-time
- Process a subset of data in real-time or in batch modes
- Provide language-agnostic language runtimes for data analysis.

Data Warehouse

A data warehouse is usually where data that has been processed and is now structured is stored. It is often used directly by business analysts in downstream applications. Data warehouses don't scale easily and typically have a lot more validation and processing associated with them.

Data Lakehouse

A data lakehouse attempts to combine elements of both Data Lake and Data Warehouse - again it is typically the name given to a group of systems architected together to provide this set of functionality. It normally supports Extract, Load and Transform paradigm.

References

- https://cloud.google.com/learn/what-is-a-data-lake
- https://www.snowflake.com/guides/what-data-lakehouse

- https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake
-

Revision #5
Created 22 November 2023 09:08:00 by James
Updated 21 January 2024 14:54:01 by James