DAT 202 Assignment 5

Colin Bowers - July 9, 2024

Identify the difference between Warehouses and Data Lake and why Lakehouse is trending in the industry?

Data warehouses are optimized for structured data and complex queries, supporting business intelligence and reporting. Data lakes store semi-structured, and unstructured data – mostly in its raw format – providing greater flexibility for data analysis.

The Lakehouse architecture combines the best of both, offering the robust data management capabilities of data warehouses with the scalability and flexibility of data lakes. Here are a few factors contributing to the increased usage of lake houses:

1. *Cost Management:* Lake houses combine low-cost cloud storage for data lakes and the performance optimization of data warehouses providing cost-effective solutions.
2. *Unified Storage and Processing*: A lake house combines the scalability and efficiency of data lakes with the data management and ACID compliance of data warehouses, providing a single platform for both analytics and operational workloads.
3. *Support for Diverse Data Types:* Lake houses can handle structured, semi-structured, and unstructured data, allowing organizations to analyze a broad range of data sources, including logs, multimedia, and sensor data, alongside traditional structured data.

Do organizations need a data warehouse and why?

Not strictly required per se, but any organization dealing with Big Data (defined by the 5 V's: Volume, Velocity, Variety, Veracity, and Value) and interested in complex analytics or predictive modelling will greatly benefit from a warehouse architecture. In these cases, it will enable the necessary data quality, security, availability and dependability for business intelligence.

How does Vector database and Graph databases work and can you give an example on where to use each?

Vector databases store data as high-dimensional numerical vectors and allow for easy queries for other similar vectors using, for example, cosine similarity. These databases are used in tasks such as recommendation systems, where finding similar items (e.g., products, movies) based on user preferences is frequently performed.

Graph databases store data as nodes and sets of connections between nodes making them useful for traversing relationships and querying complex, interconnected data. They are used in applications such as social networks, where modelling relationships and querying for neighbours or the distance between individuals is common.

In your opinion, How is the industry progressing with Hadoop/MapReduce and Spark ? What is the difference between them?

Hadoop/MapReduce and Spark are both frameworks for processing large datasets. Spark is faster and more flexible due to its in-memory processing capabilities and therefore can be more cost-effective for compute-intensive tasks. Hadoop has HDFS (Hadoop Distributed File System) which is designed for fault tolerant, highly available, and scalable storage for large datasets which can be more effective for storage-intensive tasks.

The industry is progressively favoring Spark for its ease of use, speed, and advanced analytics capabilities.

How important is the choice of databases with AI/LLM transformation projects? Give an example.

The choice of database is crucial in AI and LLM transformation projects as it affects data access speed, scalability, and the ability to handle diverse data types.

For example, a vector database can significantly enhance LLM (Large Language Model) projects by efficiently storing, indexing, and querying high-dimensional vectors representing data points, such as text embeddings. Text embeddings numerical representations of text data that capture the semantic meaning of words. These vectors are used by LLMs in their NLP (Natural Language Processing) for fast and accurate retrieval of similar text entries, improving tasks like semantic search, clustering, nearest neighbor search, recommendation systems, and contextual analysis. This accelerates model training and inference, enhancing the overall performance and scalability of AI projects.