Hive

1. Difference between Data warehouse and database?

Database:

OLTP - online transactional processing

Acid properties are fullfilled

Data stored in normalized format

Traditional data

WFS ,LFS

Write many read many

Works on single machines

Works slow

Datawarehouse:

OLAP - online analysis processing

Acid properties are not fullfiled

Bigdata

Write once read many

Works fast

2.Difference between Data warehouse and data mart?

A data warehouse serves as a centralized repository for all organizational data, while data marts offer tailored views of data for specific business areas or user groups within the organization.

3.Difference between OLTP vs OLAP

Online transaction processing (OLTP) captures, stores, and processes data from transactions in real time. Online analytical processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems.

4.Why hive metadata is stored in SQL?

Hive metadata is stored in SQL databases for compatibility, scalability, concurrency, reliability, and security reasons. SQL databases are widely used, scalable, support concurrent access, ensure data integrity, and offer advanced security features, making them a practical choice for managing metadata in large-scale data analytics environments.

5. Which SQL is the default database for hive?

\*\* Default storage to store metadata in hive : Derby. (Admin can configure any other DB)

The default SQL database for Apache Hive is Apache Derby, though in production, organizations often use more robust databases like MySQL, PostgreSQL, or Oracle.

6. What is managed table?

A managed table in Apache Hive is one where Hive handles both metadata and data files. It stores metadata in the metastore and manages the underlying data files on the storage system. When you create, load, or delete data from a managed table, Hive takes care of the associated metadata and data files.

Managed by hive, location of data is default hive location. The data is used only by hive application. when table deleted, both data and metadata is deleted.

7. What is external table?

An external table in Apache Hive is one where Hive manages the metadata but the data files are stored externally, allowing users to specify the location of the data files. Unlike managed tables, Hive does not manage the lifecycle of the data files, providing flexibility and data independence.

8. When do we use external table?

external tables are used when there is a need to access data stored externally, share data across systems, maintain data independence, or work with temporary data without moving it into the Hive warehouse directory.

9. Diff between managed and external table?

managed tables are entirely managed by Hive, while external tables provide flexibility by allowing users to store data files externally and manage them independently.

10. What happens if you don’t provide location to external table?

If you don't provide a location to an external table in Apache Hive, the table creation will fail because the location of the data files must be specified for external tables. Unlike managed tables, where Hive manages both metadata and data files internally within the default Hive warehouse directory, external tables require users to explicitly specify the location of the data files.

11. Performance optimization in hive?

Partition data.

Use bucketing.

Choose efficient file formats like ORC or Parquet.

Apply compression.

Enable vectorized processing.

Utilize caching and materialized views.

Tune queries.

Optimize hardware and cluster configuration.

12. Explain partition table. Give example

A partitioned table in Hive divides data into logical partitions based on specified column values. For example, consider a sales table partitioned by transaction\_date. Data is stored in directories based on transaction dates, improving query performance by minimizing data scanned for each query.

13. Explain bucket table. Give example

A bucketed table in Hive organizes data into a fixed number of buckets based on the hash value of specified columns. For example, a user\_profiles table can be bucketed by user\_id into 4 buckets. This improves query performance by evenly distributing data and facilitating efficient join operations.

14. Diff between partition and bucketed table.

Partitioned Table:

• Data divided into logical partitions based on column values.

• Each partition stored as separate directory.

• Improves query performance by skipping irrelevant partitions.

Bucketed Table:

• Data divided into fixed number of buckets based on hash values.

• Each bucket contains subset of data.

• Improves query performance, especially for joins, by facilitating efficient data retrieval.

In essence, partitioning organizes data into logical divisions, while bucketing evenly distributes data into fixed-size groups. Both techniques enhance query performance in different ways.

15. How is data distributed among buckets?

Data is distributed among buckets in a bucketed table based on the hash value of specified column(s). Hive calculates the hash value for each record and assigns it to a bucket accordingly, ensuring even distribution for efficient query processing.