**Hive Interview Questions**

**1. Difference between Data warehouse and database?**

- **Data Warehouse**: **Used for analytical processing (OLAP**). Contains historical, aggregated, and summarized data from various sources, optimized for complex querying.

- **Database**: **Used for transactional operations (OLTP**). Manages real-time, operational data and is optimized for quick, routine transactions.

**2. Difference between Data warehouse and data mart?**

- **Data Warehouse**: Comprehensive storage of enterprise-wide data, integrated from all subject areas across a company.

- **Data Mart**: A subset of a data warehouse focused on a specific business unit or subject area.

**3. Why hive metadata is stored in SQL?**

- Hive metadata is stored in an RDBMS to maintain structural information (like table schemas) and location pointers to data. SQL databases are reliable, scalable, and can effectively handle the kind of querying and transactional operations required for metadata management.

**4. Which SQL is the default database for hive?**

- The default metastore for Hive is **Derby** (also known as **Apache Derby**).

**5. What is a managed table?**

- A managed table is controlled by Hive. Data deletion, insertion, or table drop actions will result in manipulation of the actual data in the Hadoop Distributed File System (HDFS).

**6. What is an external table?**

- An external table in Hive only references data stored in HDFS. Hive does not manage the lifecycle of external tables' data.

**7. When do we use an external table?**

External tables are used when:

- Hive should not control the data lifecycle.

- Data is used across multiple processing frameworks.

- Data resides in a location other than the default Hive warehouse.

**8. Diff between managed and external table?**

- Managed Table: Data is controlled and managed by Hive. Deleting the table deletes the data.

- External Table: Hive only references the data. Deleting the table won't delete the underlying data in HDFS.

**9. What happens if you don’t provide a location to an external table?**

- If you don't specify a location for an external table, Hive will use the default location set in the Hive warehouse.

**10. Performance optimization in Hive:**

- Utilizing partitioning and bucketing.

- Implementing appropriate file formats like Parquet or ORC.

- Using vectorization to process batch rows simultaneously.

- Caching frequently queried tables in memory.

- Reducing the amount of data scanned via predicate pushdown.

**11. Explain partition table. Give example:**

- Partitioning in Hive is a way to split large datasets into manageable parts based on column values. For instance, if you have sales data for multiple years, you can partition by the 'year' column. Example:

(sql)

CREATE TABLE sales\_partitioned (product STRING, amount INT)

PARTITIONED BY (year INT);

**12. Explain bucket table. Give example:**

- Bucketing in Hive is used to distribute data across multiple fixed buckets based on a column's hash value, example:

(sql)

CREATE TABLE sales\_bucketed (product STRING, amount INT)

CLUSTERED BY (product) INTO 5 BUCKETS;

**13. Diff between partition and bucketed table:**

- Partition: Splits data into multiple folders in HDFS based on column values.

- Bucketed Table: Divides data into a fixed number of buckets (files) based on the hash of a column.

**14. How is data distributed among buckets?**

- Data is distributed amongst buckets based on the hash value of a specified column. All rows having the same hash value for the bucketed column will go into the same bucket. The number of buckets is fixed, so the data distribution depends on the hash function and the number of buckets specified.