**Hive interview Questions**

**1. What is the definition of Hive? What is the present version of Hive?**

ANS : Hive is an open-source data warehousing and SQL-like query language that allows data analysts to query and analyze large datasets stored in Hadoop distributed file system (HDFS). The present version of Hive is Hive 4.0.0.

**2. Is Hive suitable to be used for OLTP systems? Why?**

ANS : Hive is not suitable for OLTP (Online Transaction Processing) systems because it is designed for batch processing of large datasets. It is optimized for read-heavy workloads, where queries are executed on large volumes of data, rather than small, frequent transactions typically seen in OLTP systems.

**3. How is HIVE different from RDBMS? Does hive support ACID transactions. If not then give the proper reason.**

ANS : Hive is different from RDBMS (Relational Database Management System) in several ways. RDBMS is designed for online transaction processing (OLTP) systems, while Hive is designed for online analytical processing (OLAP) systems. Hive is schema-on-read, meaning that the schema is applied at the time of reading data, while RDBMS is schema-on-write, meaning that the schema is applied at the time of writing data. Hive does not support ACID (Atomicity, Consistency, Isolation, and Durability) transactions, which are essential for ensuring data consistency in OLTP systems.

**4. Explain the hive architecture and the different components of a Hive architecture?**

ANS : The Hive architecture consists of three main components:

* Hive Client: A command-line interface or a graphical user interface that allows users to interact with Hive and submit queries.
* Hive Metastore: A relational database that stores metadata information about Hive tables, including their schema and location on HDFS.
* Hive Execution Engine: A component responsible for executing Hive queries, which includes several sub-components such as the Query Compiler, Query Optimizer, and Query Executor.

**5. Mention what Hive query processor does? And Mention what are the components of a Hive query processor?**

ANS : The Hive query processor converts SQL-like queries written in HiveQL (Hive Query Language) into MapReduce or Tez jobs that can be executed on Hadoop. The components of a Hive query processor include:

Parser: Parses the HiveQL query and checks for syntax errors.

Semantic Analyzer: Checks the query for semantic errors and performs query optimization.

Query Planner: Translates the optimized query into an execution plan.

Query Executor: Executes the query plan and returns the results.

**6. What are the three different modes in which we can operate Hive?**

ANS : The three different modes in which we can operate Hive are:

Local Mode: Runs Hive in a single-threaded, standalone mode on a local machine without requiring a Hadoop cluster.

MapReduce Mode: Executes Hive queries on a Hadoop cluster using the MapReduce framework.

Tez Mode: Executes Hive queries on a Hadoop cluster using the Tez framework, which provides better performance than MapReduce for some types of queries.

**7. Features and Limitations of Hive.**

ANS : Features of Hive:

Provides a SQL-like interface for querying large datasets stored in HDFS.

Supports a wide range of file formats, including CSV, Avro, and Parquet.

Integrates with other Hadoop ecosystem tools, such as Pig and Spark.

Provides a flexible schema design that allows for schema-on-read.

Limitations of Hive:

Hive is not suitable for OLTP systems and real-time processing.

Hive has higher latency compared to traditional RDBMS systems.

Hive does not support ACID transactions.

**8. How to create a Database in HIVE?**

ANS : CREATE DATABASE database\_name;

**9. How to create a table in HIVE?**

ANS : CREATE TABLE table\_name (

column\_name1 data\_type,

column\_name2 data\_type,

...

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

**10.What do you mean by describe and describe extended and describe formatted with respect to database and table**

ANS : In a database, "describe" is a command used to display the structure and metadata of a table or a database object. "Describe extended" provides more detailed information on a particular object, while "describe formatted" provides detailed information on the object in a specific format.

**11.How to skip header rows from a table in Hive?**

ANS : In Hive, you can skip header rows from a table using the 'tblproperties' clause. Specifically, you can add the 'skip.header.line.count' property and set it to the number of header rows you want to skip.

**12.What is a hive operator? What are the different types of hive operators?**

ANS : Hive operators are used to perform various operations on data stored in tables. There are three types of Hive operators: relational operators, logical operators, and bitwise operators.

**13.Explain about the Hive Built-In Functions**

ANS : Hive Built-In Functions are functions that come pre-packaged with Hive and can be used to perform various operations on data stored in tables. They include functions for data type conversion, string manipulation,

**14. Write hive DDL and DML commands.**

ANS : Hive DDL (Data Definition Language) commands are used to define the structure of tables and databases. Examples include 'create database', 'create table', and 'alter table'. Hive DML (Data Manipulation Language) commands are used to manipulate data stored in tables. Examples include 'select', 'insert', and 'update'.

**15.Explain about SORT BY, ORDER BY, DISTRIBUTE BY and CLUSTER BY in Hive.**

ANS : SORT BY, ORDER BY, DISTRIBUTE BY, and CLUSTER BY are all used for sorting and grouping data in Hive. SORT BY sorts data based on a specified column, ORDER BY sorts data in ascending or descending order, DISTRIBUTE BY determines how data is distributed across reducer tasks, and CLUSTER BY is a combination of SORT BY and DISTRIBUTE BY.

**16.Difference between "Internal Table" and "External Table" and Mention when to choose “Internal Table” and “External Table” in Hive?**

ANS : An internal table is managed by Hive, which means that Hive manages the table and its data. An external table is not managed by Hive, and the data is stored in a location outside of Hive. Internal tables are preferred when the data is temporary or is not needed outside of Hive, while external tables are preferred when the data needs to be accessed outside of Hive or needs to be shared across different clusters.

**17.Where does the data of a Hive table get stored?**

ANS : The data of a Hive table can be stored either in HDFS (Hadoop Distributed File System) or in a location external to Hadoop, such as an S3 bucket or a local file system.

**18.Is it possible to change the default location of a managed table?**

ANS : Yes, it is possible to change the default location of a managed table using the 'location' clause in the 'alter table' command. However, it is not recommended as it can cause issues with data consistency and metadata management

**19.What is a metastore in Hive? What is the default database provided by Apache Hive for metastore?**

ANS : A metastore in Hive is a central repository that stores metadata information about the structure and schema of data stored in Hive. The default database provided by Apache Hive for metastore is Derby.

**20.Why does Hive not store metadata information in HDFS?**

ANS : Hive does not store metadata information in HDFS because HDFS is not suitable for storing small files and frequently changing data, which is common in metadata. Instead, Hive uses a relational database for storing metadata.

**21.What is a partition in Hive? And Why do we perform partitioning in Hive?**

ANS : A partition in Hive is a way of dividing a table into smaller, more manageable parts based on the values of one or more columns. We perform partitioning in Hive to improve query performance and make it easier to manage large datasets.

**22.What is the difference between dynamic partitioning and static partitioning?**

ANS : Static partitioning involves specifying the partition values while inserting data into a partitioned table, while dynamic partitioning automatically creates partitions based on the values in the input data.

**23.How do you check if a particular partition exists?**

ANS : To check if a particular partition exists, you can use the ‘SHOW PARTITIONS’ command or query the ‘PARTITIONS’ table in the Hive metastore database.

**24.How can you stop a partition form being queried?**

ANS : To stop a partition from being queried, you can use the ‘ALTER TABLE’ command to exclude the partition or drop the partition altogether.

**25.Why do we need buckets? How Hive distributes the rows into buckets?**

ANS : We need buckets in Hive to further partition data within a partition and distribute it evenly across nodes in a cluster. Hive distributes rows into buckets based on a hash function applied to one or more columns.

**26.In Hive, how can you enable buckets?**

ANS : To enable buckets in Hive, you can use the ‘CLUSTERED BY’ clause in the ‘CREATE TABLE’ statement and specify the columns to use for bucketing.

**27.How does bucketing help in the faster execution of queries?**

ANS : Bucketing helps in the faster execution of queries because it reduces the amount of data that needs to be scanned by limiting it to a specific bucket or set of buckets, which can be read in parallel. This can result in significant performance improvements for certain types of queries.

**28.How to optimise Hive Performance? Explain in very detail.**

ANS : To optimize Hive performance, one can partition the data, use bucketing, use efficient file formats, optimize query execution, and use appropriate hardware and infrastructure.

**29. What is the use of Hcatalog?**

ANS : HCatalog is a table and storage management layer for Hadoop that provides a relational view of data stored in Hadoop Distributed File System (HDFS).

**30. Explain about the different types of join in Hive.**

ANS : Hive supports different types of join operations, including inner join, outer join, left join, right join, and full outer join.

**31.Is it possible to create a Cartesian join between 2 tables, using Hive?**

ANS : Yes, it is possible to create a Cartesian join between 2 tables using Hive by using the CROSS JOIN keyword.

**32.Explain the SMB Join in Hive?**

ANS : SMB join or Sort Merge Bucketed join is a type of join in Hive that is optimized for joining large tables that have been bucketed and sorted on the join key.

**33.What is the difference between order by and sort by which one we should use?**

ANS : SORT BY orders the data only in the reducer phase, whereas ORDER BY orders the data both in the mapper and reducer phases. ORDER BY should be used when the result set is small, while SORT BY should be used for large result sets.

**34.What is the usefulness of the DISTRIBUTED BY clause in Hive?**

ANS : The DISTRIBUTED BY clause in Hive is used to distribute the data across different nodes in the cluster based on a particular column or set of columns. This can improve query performance by minimizing data shuffling.

**35.How does data transfer happen from HDFS to Hive?**

ANS : Data is transferred from HDFS to Hive through Hive's HCatalog API, which provides a relational view of data stored in HDFS.

**36.Wherever (Different Directory) I run the hive query, it creates a new metastore\_db, please explain the reason for it?**

ANS : Hive creates a new metastore\_db directory wherever it runs because it stores metadata about the tables in that directory.

**37.What will happen in case you have not issued the command: ‘SET hive.enforce.bucketing=true;’ before bucketing a table in Hive?**

ANS : If the command "SET hive.enforce.bucketing=true;" is not issued before bucketing a table in Hive, the table may not be properly bucketed, which can affect performance.

**38.Can a table be renamed in Hive?**

ANS : Yes, a table can be renamed in Hive using the ALTER TABLE command.

**39.Write a query to insert a new column(new\_col INT) into a hive table at a position before an existing column (x\_col)**

ANS : ALTER TABLE table\_name ADD COLUMNS (new\_col INT) AFTER x\_col;

**40.What is serde operation in HIVE?**

ANS : serde (serialization/deserialization) operation in Hive is used to convert data between internal representation and external representation (such as CSV, JSON, Avro, etc.) for data storage and processing.

**41.Explain how Hive Deserializes and serialises the data?**

ANS : Hive uses built-in or custom serdes to deserialize data from external storage formats into internal formats and serialize data from internal formats into external formats. The serde defines the schema of the data and how to read and write it.

**42.Write the name of the built-in serde in hive.**

ANS : The built-in serde in Hive is called LazySimpleSerDe.

**43.What is the need of custom Serde?**

ANS : Custom serdes are needed when Hive does not support the data format used for data storage or processing. Custom serdes can be used to deserialize and serialize data in custom formats, such as protobuf or Thrift.

**44. Can you write the name of a complex data type(collection data types) in Hive?**

ANS : Examples of complex data types (collection data types) in Hive are ARRAY, MAP, and STRUCT.

**45.Can hive queries be executed from script files? How?**

ANS : Yes, Hive queries can be executed from script files using the -f option followed by the path to the script file.

**46.What are the default record and field delimiter used for hive text files?**

ANS : The default record delimiter for Hive text files is the newline character (\n), and the default field delimiter is the tab character (\t).

**47.How do you list all databases in Hive whose name starts with s?**

ANS : To list all databases in Hive whose name starts with "s", you can use the command "SHOW DATABASES LIKE 's\*'".

**48.What is the difference between LIKE and RLIKE operators in Hive?**

ANS : The LIKE operator is used to match a pattern using simple regular expressions, while the RLIKE (regex-like) operator is used to match a pattern using regular expressions.

**49.How to change the column data type in Hive?**

ANS : To change the column data type in Hive, you can use the ALTER TABLE command with the CHANGE COLUMN clause, followed by the column name, new data type, and other optional attributes.

**50.How will you convert the string ’51.2’ to a float value in the particular column?**

ANS : To convert the string '51.2' to a float value in a particular column, you can use the CAST function like this: CAST(column\_name AS FLOAT).

**51.What will be the result when you cast ‘abc’ (string) as INT?**

ANS : When you cast 'abc' (string) as INT in Hive, the result will be NULL because 'abc' cannot be converted to an integer.

**52.What does the following query do?**

**a. INSERT OVERWRITE TABLE employees**

**b. PARTITION (country, state)**

**c. SELECT ..., se.cnty, se.st**

**d. FROM staged\_employees se;**

ANS : This query will overwrite the data in the "employees" table, partitioned by "country" and "state", with the data from the "staged\_employees" table. It will select specific columns from "staged\_employees" and add "country" and "state" columns to the result set. The "..." in the SELECT statement represents other columns being selected from "staged\_employees".

**53.Write a query where you can overwrite data in a new table from the existing table.**

ANS : To overwrite data in a new table from an existing table in SQL, you can use the INSERT INTO SELECT statement with the OVERWRITE option.

**54.What is the maximum size of a string data type supported by Hive? Explain how Hive supports binary formats.**

ANS : The maximum size of a string data type supported by Hive is 2 GB. Hive supports binary formats through the use of SerDe (Serializer/Deserializer) libraries

**55. What File Formats and Applications Does Hive Support?**

ANS : Hive supports various file formats, including TextFile, SequenceFile, RCFile, ORC, Parquet, and Avro. It also supports applications like Hadoop, Spark, and Pig.

**56.How do ORC format tables help Hive to enhance its performance?**

ANS : ORC (Optimized Row Columnar) format tables help Hive to enhance its performance by providing efficient compression and better query performance due to its columnar storage format.

**57.How can Hive avoid mapreduce while processing the query?**

ANS : Hive can avoid MapReduce while processing queries by using Tez, a more efficient execution engine, or by using LLAP (Low Latency Analytical Processing), a caching and data serving layer.

**58.What is view and indexing in hive?**

ANS : A view is a virtual table that is created from a query. An index is a data structure that improves the speed of data retrieval operations on a table.

**59.Can the name of a view be the same as the name of a hive table?**

ANS : Yes, the name of a view can be the same as the name of a Hive table.

**60.What types of costs are associated in creating indexes on hive tables?**

ANS : The costs associated with creating indexes on Hive tables include increased storage requirements, additional processing time for data modifications, and potential performance overhead for some queries.

**61.Give the command to see the indexes on a table.**

ANS : The command to see the indexes on a table in Hive is "SHOW INDEXES ON table\_name;"

**62. Explain the process to access subdirectories recursively in Hive queries.**

ANS : To access subdirectories recursively in Hive queries, you can use the "RECURSIVE" keyword in the HiveQL command "LOAD DATA."

**63.If you run a select \* query in Hive, why doesn't it run MapReduce?**

ANS : When you run a "SELECT \*" query in Hive, it doesn't run MapReduce because Hive maintains a metadata cache of the table schema, and the query can be executed directly on this metadata cache.

**64.What are the uses of Hive Explode?**

ANS : Hive Explode is used to flatten arrays or maps in Hive tables into separate rows, making it easier to analyze the data.

**65. What is the available mechanism for connecting applications when we run Hive as a server?**

ANS : Hive can be run as a server using Thrift, which provides a mechanism for connecting applications to Hive. Applications can use various programming languages and drivers to communicate with Hive over Thrift.

**66.Can the default location of a managed table be changed in Hive?**

ANS : Yes, the default location of a managed table can be changed in Hive.

**67.What is the Hive ObjectInspector function?**

ANS : The Hive ObjectInspector function is used to inspect objects in Hive.

**68.What is UDF in Hive?**

ANS : UDF stands for User-Defined Function, which allows users to write their own custom functions in Hive.

**69.Write a query to extract data from hdfs to hive.**

ANS : To extract data from HDFS to Hive, you can use the LOAD DATA INPATH command.

**70.What is TextInputFormat and SequenceFileInputFormat in hive.**

ANS : ‘TextInputFormat’ is used for reading plain text files in Hive, while ‘SequenceFileInputFormat’ is used for reading Hadoop Sequence Files.

**71.How can you prevent a large job from running for a long time in a hive?**

ANS : You can prevent a large job from running for a long time in Hive by setting appropriate map and reduce parameters, using bucketing and partitioning, and optimizing queries.

**72.When do we use explode in Hive?**

ANS : explode is used in Hive to transform a column of arrays or maps into multiple rows.

**73.Can Hive process any type of data formats? Why? Explain in very detail**

ANS : Hive can process various data formats, including structured, semi-structured, and unstructured data, due to its ability to use different file formats and serializers/deserializers.

**74.Whenever we run a Hive query, a new metastore\_db is created. Why?**

ANS : A new metastore\_db is created whenever a Hive query is run to store metadata about the table being queried.

**75.Can we change the data type of a column in a hive table? Write a complete query.**

ANS : Yes, the data type of a column in a Hive table can be changed using the ‘ALTER TABLE’ command.

**76.While loading data into a hive table using the LOAD DATA clause, how do you specify it is a hdfs file and not a local file ?**

ANS : To specify that it is a HDFS file and not a local file, you need to specify the full HDFS path to the file in the LOAD DATA INPATH command.

**77.What is the precedence order in Hive configuration?**

ANS : The precedence order in Hive configuration is command line arguments, Hive-site.xml, and Hadoop configuration files.

**78.Which interface is used for accessing the Hive metastore?**

ANS : The JDBC interface is used for accessing the Hive metastore.

**79.Is it possible to compress json in the Hive external table ?**

ANS : Yes, you can compress JSON in a Hive external table using a codec like gzip or snappy.

**80.What is the difference between local and remote metastores?**

ANS : Local metastores are embedded within Hive, while remote metastores are hosted on a separate server.

**81.What is the purpose of archiving tables in Hive?**

ANS : The purpose of archiving tables in Hive is to move infrequently accessed or historical data to a separate location for storage or analysis purposes.

**82.What is DBPROPERTY in Hive?**

ANS : ‘DBPROPERTY’ is a Hive function that returns the value of a specified database propert

**83.Differentiate between local mode and MapReduce mode in Hive.**

ANS : Local mode in Hive runs queries on a single machine, while MapReduce mode runs queries on a Hadoop cluster.