Hive interview Questions

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

Hive is a data warehouse product designed on top of HDFS. Current stable release is 3.1.3.

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

Hive is not suitable for OLTP systems as it is based on HDFS mapreduce batch processing.

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

Hive does not enforce schema on load. It is only enforced while read. Hive supports ACID transactions with orc data.

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

Hive Client – all client, web UI and database drivers

Hive Services – Server, CLI, driver

MapReduce

HDFS

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

Hive query processor converts queries to MapReduce jobs. The components are parser, logical optimizer, physical optimizer and execution engine.

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

Local mode, distributed mode and pseudo-distributed mode.

1. Features and Limitations of Hive.

Hive a datawarehouse based on HDFS used for OLAP processing as batch. SQL type queries can be used with table supporting partitioning, bucketing etc. It offers great performance for properly organized data.

It does not support OLTP processing. Subqueries are not supported. Latency is high as it takes to return the result.

1. How to create a Database in HIVE?

create database <databasename>;

1. How to create a table in HIVE?

create table <tablename> (columnname1 type1, columnname2 type2, etc.) fields terminated by ‘,’ lines terminated by ‘\n’ stored as textfile;

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

Describe – get basic database, table details

Describe extended – get detailed information about the database/table

Describe formatted – describe extended but in a formatted fashion.

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

TBLPROPERTIES(“skip.header.line.count”=”1”);

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

Operators allow different operations to be performed using operands. There are different types like relational, arithmetic, logical, string, complex etc.

1. Explain about the Hive Built-In Functions

Functions available in Hive for performing predefined actions. There are mathematical functions, collection function, type conversion function, date function, conditional function, string function.

1. Write hive DDL and DML commands.

Data Definition Language (DDL) HiveQL statements define the structure of the database. These are used to create, drop, alter and other operations.

Data Manipulation Language (DML) HiveQL statements operates on the data contained with tables. Examples are select, insert, delete, update, etc.

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

Sort by orders output data based on key within reducer. Reducers can have overlapping data based on key.

Order by does global ordering causing only one reducer to be used.

Distribute by distributes data based on key to the reducers. However, data is not ordered.

Cluster by distributes data based on key as well as orders data based on the key.

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

Internal table is default where data is managed internally by Hive. If an internal table is dropped then the table, metadata and data are deleted.

External tables apply metadata on top of data in Hive with explicitly loading the data to the table. When an external table is dropped the table and metadata is removed but not the data.

Internal tables are used for temporary purposes or if ACID transactions need to be performed.

External tables are used when we want to use external data without loading. When data may not be brought to Hive or duplication of data is not advised.

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

Default - /user/hive/warehouse on HDFS

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

Default location can be changed by specifying the desired directory in hive.metastore.warehouse.dir configuration parameter in the hive-site.xml.

For specific cases default location can be changed by using LOCATION keyword in HiveQL DDL statement.

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

The definition of Hive objects such as databases, tables, and functions are stored in the Metastore. Apache Derby RDBMS is the default database provided by Apache Hive for metastore.

1. Why does Hive not store metadata information in HDFS?

HDFS read/write operations are time consuming processes. It will be more efficient if the metadata information is obtained instantaneously for every operation.

1. What is a partition in Hive? And why do we perform partitioning in Hive?

Partition means separating table data based on values so that query performances can be improved.

1. What is the difference between dynamic partitioning and static partitioning?

Static Partitioning - Partitions are manually added, Data Loading is faster, individual files are loaded as per the partition.

Dynamic partitioning – Hive automatically determines the number of partitions to be created based on the unique keys for the partitioned column. Data loading is slower than Static. Good approach if all values of a partitioned column are unknown. Entire table is loaded in a single shot.

1. How do you check if a particular partition exists?

use SHOW PARTITIONS command

1. How can you stop a partition form being queried?

ENABLE OFFLINE clause with ALTER TABLE statement

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

Technique to split the data into more manageable files based on splitting data across the no. of buckets. Data is distributed across the buckets in such a way that a particular value for a bucketed column is not present in more that one bucket.

1. In Hive, how can you enable buckets?

use CLUSTERED BY clause with the column name you wanted to bucket and the count of the buckets. hive.enforce.bucketing property to true.

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

When a query involves a condition on the bucketing column(s), Hive can determine which buckets to read based on the query predicates, resulting in significant performance improvements. When two bucketed tables are joined on the same bucketing column(s), Hive can perform the join operation on a bucket-to-bucket basis, reducing the amount of data shuffling and improving overall performance.

1. How to optimise Hive Performance? Explain in very detail.

Partitioning, Bucketing, Compression, Denormalizing Data, Use simple joins, store data as ORC, fine tune queries, etc.

1. What is the use of Hcatalog?

It is a utility which enables users to use different data processing tools to read and write data on the grid with ease.

1. Explain about the different types of join in Hive.

Inner join, Full join, left join, right join

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

Yes, using CROSS JOIN

1. Explain the SMB Join in Hive?

SMB is a join performed on bucket tables that have the same sorted, bucket, and join condition columns.

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

ORDER BY performs a total ordering of the query result set. SORT BY orders the data only within each reducer, thereby performing a local ordering, where each reducer’s output will be sorted. Better performance is traded for total ordering.

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

Hive uses the columns in Distribute By to distribute the rows among reducers. All rows with the same Distribute By columns will go to the same reducer.

1. How does data transfer happen from HDFS to Hive?

HDFS data can be accessed in Hive using external table.

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

Hive creates local metastores for query execution and by default it creates from wherever query in run.

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

Bucketing will not work properly if the command is not set.

1. Can a table be renamed in Hive?

ALTER TABLE table\_name RENAME TO new\_table\_name;

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

ALTER TABLE table\_name  
  
CHANGE COLUMN new\_col INT  
  
BEFORE x\_col

1. What is serde operation in HIVE?

SerDe means Serializer and Deserializer. Hive uses SerDe and FileFormat to read and write table rows.

1. Explain how Hive Deserializes and serialises the data?

Serialise is done on write, the structured data is serialised into a bit/byte stream for storage. On read, the data is deserialised from the bit/byte storage format to the structure required by the reader.

1. Write the name of the built-in serde in hive.

org.apache.hadoop.hive.serde2

1. What is the need of custom Serde?

Custom Serde is required to serialize and deserialize custom data formats which are not supported by available Serde packages.

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

Array, Map, Struct and union

1. Can hive queries be executed from script files? How?

Write the queries in filename.hql and running the following command from hive shell “hive -f filename.hql”

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

Record delimiter - \n, field delimiter - \001

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

show databases like 's\*';

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

LIKE is used for normal pattern matching. RLIKE is used where RegExp pattern matching is needed.

1. How to change the column data type in Hive?

ALTER TABLE table\_name CHANGE column\_oldname column\_newname new\_datatype

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

select cast(column\_name as float) where column\_name = ’51.2’;

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

NULL value is returned.

1. 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;

The query loads a dynamically partitioned hive table ‘employees’ with composite keys country and state.

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

CREATE TABLE table2 LIKE table1;

INSERT OVERWRITE TABLE table2 SELECT \* FROM table1;

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

2 GB.

1. What File Formats and Applications Does Hive Support?

TEXTFILE, SEQUENCEFILE, RCFILE, ORC, PARQUET, AVRO

Java, PHP, Python, C++, Ruby

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

Hive is customized for columnar data formats especially ORC which achieves better performance.

1. How can Hive avoid mapreduce while processing the query?

setting the hive.exec.mode.local.auto property to ‘true’.

1. What is view and indexing in hive?

Views are subsets of table data presented as read only information.

Indexing is creating keys for a table using which faster data can be obtained.

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

No.

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

Index occupies space and it need maintenance each time data is changed.

1. Give the command to see the indexes on a table.

SHOW INDEX ON table\_name

1. Explain the process to access subdirectories recursively in Hive queries.

Set mapred.input.dir.recursive=true;

Set hive.mapred.supports.subdirectories=true;

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

There is no filtration of data needed.

1. What are the uses of Hive Explode?

Hive's explode() function takes an array (or a map) as input and outputs the elements of the array (map) as separate rows.

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

Thrift Client, JDBC Driver, ODBC Driver

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

Using the LOCATION keyword, we can change the default location of Managed tables while creating the managed table in Hive.

1. What is the Hive ObjectInspector function?

Hive ObjectInspector is a group of flexible APIs to inspect value in different data representation.

1. What is UDF in Hive?

User Defined Functions, also known as UDF, allow you to create custom functions to process records or groups of records.

1. Write a query to extract data from hdfs to hive.\

LOAD DATA LOCAL INPATH '/filename.csv'

OVERWRITE INTO TABLE table\_name;

1. What is TextInputFormat and SequenceFileInputFormat in hive.

TextInputFormat is a class to read and write data in text format.

SequenceFileInputFormat is a class to read and write data in Hadoop SequenceFile format.

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

Setting the MapReduce jobs to execute in strict mode set hive.mapred.mode=strict;

The strict mode ensures that the queries on partitioned tables cannot execute without defining a WHERE clause.

1. When do we use explode in Hive?

Hive's explode() function takes an array (or a map) as input and outputs the elements of the array (map) as separate rows.

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

Hive only processes predefined data types. It needs serde processes which is not available for all data types.

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

Hive creates local metastores for query execution and by default it creates from wherever query in run.

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

ALTER TABLE table\_name CHANGE column\_name column\_name new\_datatype

1. 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?

By not using LOCAL keyword.

1. What is the precedence order in Hive configuration?

Hive SET command has the highest priority

-hiveconf option from Hive Command Line

hive-site.xml file

hive-default.xml file

hadoop-site.xml file

hadoop-default.xml file

1. Which interface is used for accessing the Hive metastore?

WebHCat API web interface

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

Yes, it can be stored after gzip compression in .gz format.

1. What is the difference between local and remote metastores?

Local Metastore:- Here metastore service still runs in the same JVM as Hive but it connects to a database running in a separate process either on same machine or on a remote machine.

Remote Metastore:- Metastore runs in its own separate JVM not on hive service JVM.

1. What is the purpose of archiving tables in Hive?

To reduce the number of hdfs files in the Hive table partition.

1. What is DBPROPERTY in Hive?

It is database properties.

1. Differentiate between local mode and MapReduce mode in Hive.

Use Local mode when:

Hadoop is installed under the pseudo mode, possessing only one data node

The data size is smaller and limited to a single local machine

Users expect faster processing because the local machine contains smaller datasets.

Use Map Reduce mode when:

Hadoop has multiple data nodes, and the data is distributed across these different nodes

Users must deal with more massive data sets

MapReduce is Hive's default mode.