1. How to add partition on existing table

We cannot add partition to existing non partitioned table, to insert partitions we need to create a new partitioned table with static or dynamic partitioning and then insert data in it from non partitioned table

1. What is schema strategy in hive
2. Is there a date datatype in hive

Yes, timestamp and date are 2 datatypes to store dates

1. How to recursively access the directory
2. When do you use explode
3. How to cast string as float - cast(RECURR as float) as RECURR
4. Where does hive store its metadata

By default it stores it in rdbms

1. How to change location of the data for a particular table

Location of data has to be changed manually, table can point out at a different location to read data from

1. Why cartesian join is not possible in hive
2. Can we have insert/update query on views

No, it is only possible in hive tables using acid properties not in views

In table,

UPDATE students SET name = null WHERE gpa <= 1.0;

DELETE FROM students WHERE gpa <= 1,0;

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

No, due to high latency it is not recommended for OLTP platform, t is not meant for real time processing

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

In user/hive/warehouse by default

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

It stores in rdbms due to low latency rate, hdfs is an heavy operation for read and write

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

Matastore is in same jvm in local mode as that of data, in remote both are under different jvm

1. What is the difference between the external table and managed table?

If you drop external table only schema is dropped not data, where else in managed table even data is lost, so in external table you only give hive control over schema and not data

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

Yes, by using location clause

1. When should we use SORT BY instead of ORDER BY?

Sort by uses multiple reducer where else order by uses single reducer

1. Partition vs bucketing

Partition is used to improve query timing by passing the scanning of entire data, bucketing can be used along with partitioning to further enhance the processing time.partitioning can have unequal data. Bucketing uses hashing of keys

1. Dynamic partitioning

When we dont know the value of partitions we use dynamic partitioning, we have to put dynamic partitioning as true and non strict using below queries

hive.exec.dynamic.partitioning=true

hive.exec.dynamic.partitioning.mode=nonstrict

1. Need of bucketing
2. How Hive distributes the rows into buckets?

Using hashing of keys

1. What is the use of Hcatalog?

To connect with external databases and services

1. Where is table data stored in Apache Hive by default?

user>hive>warehouse

1. Are multi-line comments supported in Hive?

No

1. What are the different components of a Hive architecture?
2. Is it possible to use the same metastore by multiple users, in case of the embedded hive?

No, it has to be done in local mode

1. Modes in which hive can run

Distributedmode

Standalone

Pseudodistributed mode

1. What are the types of meta stores?
2. What is the default Apache Hive metastore database?

derby

1. What are the Hive collection data types?

Sruct, arrays

1. Is it possible to run UNIX shell commands in Hive?

Yes, by using ! in start

1. Is it possible to execute Hive queries from a script file?

Yes, using .hql files

1. What is a .hiverc file?
2. How can you check if a specific partition exists?

Show partitions

1. If you had to list all databases that began with the letter ‘c’, how would you do it?

Show databases like “c”

1. When a Hive table partition is pointed to a new directory, what happens to the data?

It doesnt move, it has to be moved manually

1. Do you save space in the HDFS by archiving Hive tables?

No, arcchiving doesnt save any space

1. Strictmode

Allows to have dynamic partitioning

1. How can you stop a partition from being accessed in a query?
2. What is a table generating function on Hive?
3. Can you avoid MapReduce on Hive? - select \* query will not call mapreduce
4. What is view..its existence after termination of session, temp storage in
5. Indexing
6. What is Hcatalog?
7. What is UDF in Hive?

It is used to get additional functionality in hive which is not available by default

1. What does /\*streamtable(table\_name)\*/ do?
2. Name the components of a Hive query processor?
3. How Hive distributes the rows into buckets?
4. How do ORC format tables help Hive to enhance the performance?
5. What are the different components of a Hive architecture?
6. Hbase vs hive
7. Local metastore for concurrent usage
8. Suppose, I create a table that contains details of all the transactions done by the customers of year 2016: CREATE TABLE transaction\_details (cust\_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ; . Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?

Create a new partitioned table with partition column as month. Set dynamic partition mode non strict and insert data from non partitioned to partitioned and then query

### What is the default maximum dynamic partition that can be created by a mapper/reducer? How can you change it?

By default it is 100, we can change it by using num partitions parameter

1. I am inserting data into a table based on partitions dynamically. But, I received an error – FAILED ERROR IN SEMANTIC ANALYSIS: Dynamic partition strict mode requires at least one static partition column. How will you remove this error?

hive.exec.dynamic.partition.mode=nonstrict

hive.exec.dynamic.partition=true

### What will happen in case you have not issued the command: ‘SET hive.enforce.bucketing=true;’ before bucketing a table in Hive in Apache Hive 0.x or 1.x?

1. Need of indexing
2. Suppose, I have a CSV file – ‘sample.csv’ present in ‘/temp’ directory with the following entries:

id first\_name last\_name email gender ip\_address

1 Hugh Jackman hughjackman@cam.ac.uk Male 136.90.241.52

2 David Lawrence dlawrence1@gmail.com Male 101.177.15.130

3 Andy Hall andyhall2@yahoo.com Female 114.123.153.64

4 Samuel Jackson samjackson231@sun.com Male 89.60.227.31

5 Emily Rose rose.emily4@surveymonkey.com Female 119.92.21.19

How will you consume this CSV file into the Hive warehouse using built SerDe?

Using row format=hive.serde2.csv

1. Suppose, I have a lot of small CSV files present in /input directory in HDFS and I want to create a single Hive table corresponding to these files. The data in these files are in the format: {id, name, e-mail, country}. Now, as we know, Hadoop performance degrades when we use lots of small files.So, how will you solve this problem where we want to create a single Hive table for lots of small files without degrading the performance of the system?

Load all data in a normal table stored as csv, create another table stored as sequencefile, insert data from normal to sequence table

1. Acid properties, applicable from?, what commands to set. How to create

Mapreduce will come in only during insert command and not while load and select \*

Load data local inpath for linux data and load data inpath to load data from hdfs

Desc extended tablename

Show functions

All aggregate/join queries in hive will trigger map reduce job like select count(\*)

Bucketing is useful in case of all unique records and unique columns

Sampling uses bucketing

How to decide no of buckets

Data size of table / 128mb -> log of value obtained comma 2 . for eg [log(17,2)]...2 power of value obtained in previous step is bucket size

How to change no of bucket

Example of using only bucketing

Msck

repair table

Required things for acid tables: must be an internal table, must be orc, must be orc format, in tblproperties transactional must be set to true

Olap vs oltp

Map side joins

<https://data-flair.training/blogs/apache-hive-interview-questions/>

https://www.edureka.co/blog/interview-questions/hive-interview-questions/