Hive

1. **What is the hadoop version you are using?**

   apache hadoop 3.2.4

2. **What is the hive version you are using?**

   Hive version - hive 3.1.3

3. **Cloudera version you are using and what is the latest version?**

   Cdh 6.3
   Latest version cdp 7

4. **Explain what is hive?**

   - Hive is an etl and data warehousing tool developed on top of hadoop distributed file system (hdfs).
   - Hive is an open-source-software that lets programmers analyze large data sets on hadoop.
   - Hive is a tool to process structured data in hadoop.
   - However, the apache software foundation took it up, but initially, hive was developed by facebook.

5. **Features and limitations of hive.**
   Features of hive

   - Hive uses a language called hiveql (hql), which is similar to sql.
   - Hiveql automatically translates sql-like queries into mapreduce jobs which will execute on hadoop.
   - Hive is best suited for data warehouse applications.
   - HiveQL facilitates querying and managing large data sets residing in Hadoop.
   - To process data without storing in hdfs
   - hive supports external tables.
   - Hive is close to olap (online analytic processing)

   Limitation of hive

   - Hive does not provide record-level update, insert, nor delete.
   - Hive queries have higher latency than SQL queries, because of start-up overhead for MapReduce jobs submitted for each hive query.

- As Hadoop is a batch-oriented system, hive doesn’t support oltp (online transaction processing).

6. **What is partition in hive?**

- It is one of the ways of dividing huge tables into different small no.of tables based on partition keys.
  Ex: year,month,day, country ,state etc.
- Partition reduces the query latency by scanning only relevant partitioned data instead of the whole data set.
- Physically, a partition is nothing but a sub-directory (folder) in the table directory (folder).
- Partition is based on a hash partitioning algorithm.

7. **Where to use partition? And what are the properties that we must enable?**

- Go with partitioning when there are less number of distinct values in the column.
- The number of partitions will be equal to the number of distinct values.
- Cardinality (no.of distinct values) of the column is low.

```
Set hive.exec.dynamic.partition = true;
Set hive.exec.dynamic.partition.mode = nonstrict;
```

```
Create table zipcodes(recordnumber int,country string,city string,zipcode int)
Partitioned by(state string)
Row format delimited
Fields terminated by ',';
```

8. **How to decide partition column?**

- Let’s say if we have a requirement to scan data department-wise then we can create a partition on the dept column, and it will scan the data from there.

- Let’s say if we have a requirement to scan data (year, month, day, hour) and you have created a partition on 4 columns (year, month, day, hour) .in this case hive will distribute data based on (year, month, day, hour) columns.

9. **What is static & dynamic partitioning and when is it used?**

Hive static partitioning

- We need to manually create each partition before inserting data into a partition.
- We need to know all partitions in advance.

- In static partitioning we need to specify the partition column value in each and every load statement.

  Load data inpath '/user/cloudera/india.txt' into table partition_table partition(country='ind');
  Load data inpath 'user/cloudera/us.txt' into table partition_table partition(country = 'us');

- Static partition saves your time in loading data compared to dynamic partition.
- We can alter the partition in the static partition.
- You can perform static partition on hive manage table or external table.
- If you want to use the static partition in the hive you should set property.this property set by default in hive-site.xml

  Set hive.mapred.mode = strict

- Static partition is in strict mode.

<u>Hive dynamic partitioning</u>

- Partitions will be created dynamically based on input data to the table.
- No need to specify partition column value each and every load statement.

  Insert into table partition_table partition(country) select * from non_partition_table;

- Dynamic partitions are suitable when we have a lot of partitions and we can not predict in advance new partitions, ahead of time.
- Dynamic partition takes more time in loading data compared to static partition.
- Dynamic partition there is no required where clause to use limit.
- We can't perform alter on the dynamic partition.
- You can perform dynamic partition on hive external table and managed table.
- If you want to use the dynamic partition in the hive then the mode is in non-strict mode.
- Here are hive dynamic partition properties you should allow
  Set hive.exec.dynamic.partition = true;
  Set hive.exec.dynamic.partition.mode = nonstrict;

10. How to load data into partition table from non-partition table?

- Create a non-partitioned table **non_partition_table** and load the data.
- Now create a partitioned table **partition_table** and specify the partition column (say state)

- Load data from **non_partition_table** to **partition_table** like below.
  Insert into table partition_table partition(state) select * from non_partition_table ;

- Here you should ensure that the partition column is the last column of the non-partitioned table.

11. How to see all partitions in hive?

SHOW PARTITIONS <TABLE_NAME>

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

SHOW PARTITIONS TABLE_NAME PARTITION(PARTITIONED_COLUMN='PARTITION_VALUE')

13. Can a partition be archived? What are the advantages and disadvantages?

Yes. We can archive a partition.
Advantage
Basically, it decreases the number of files stored in name node and the archived file can be queried using hive.
Disadvantage
Although, it will cause less efficient query and does not offer any space savings.

14. **How to add the partition in existing table without the partition table?**

Basically, we cannot add/create the partition in the existing table, especially which was not partitioned while creation of the table.

If you had partitioned the existing table, then by using the alter table command, you will be allowed to add the partition.

15. **How can you add a new partition for the month December in the above partitioned table?**

For adding a new partition in the above table partitioned_transaction, we will issue the command give below:

ALTER TABLE PARTITIONED_TRANSACTION ADD PARTITION (MONTH='DEC') LOCATION '/PARTITIONED_TRANSACTION';

16. **What is the significance of the line set hive.mapred.mode = strict;**
Basically, in strict mode, it sets the mapreduce jobs. So, by which the queries on partitioned tables cannot run without a where clause. Hence, it prevents very large job running for a long time.

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

- By default the number of maximum partition that can be created by a mapper or reducer is set to 100.

- One can change it by issuing the following command:
  SET HIVE.EXEC.MAX.DYNAMIC.PARTITIONS.PERNODE = <VALUE>

- you can set the total number of dynamic partitions that can be created by one statement by using: **SET HIVE.EXEC.MAX.DYNAMIC.PARTITIONS = <VALUE>**

18. What is a bucketing/ clustering in hive? Where to use? And what are the properties that we have to enable?

- The table is divided into the number of partitions, and these partitions can be further subdivided into more manageable parts known as buckets/clusters.
- Whenever **cardinality (no.of distinct values)** of column is high we go with fixed number of buckets.
  Ex : column_id
- Bucket is a file.
- User need to defined number of buckets needed.
- Bucket is based on hash partitioning algorithm.
- "clustered by" clause is used to divide the table into buckets.

```
SET HIVE.ENFORCE.BUCKETING = TRUE;

Create table zipcodes(recordnumber int,country string,city string,zipcode int)
Partitioned by(state string)
Clustered by (zipcode) into 32 buckets
Row format delimited
Fields terminated by ',';
```

Load data into bucket

```
Load data inpath '/data/zipcodes.csv' into table zipcodes;
```

19. **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?**

- If have not issued the command: **SET HIVE.ENFORCE.BUCKETING=TRUE;**
- you one may find the number of files that will be generated in the table directory to be not equal to the number of buckets.
- As an alternative, one may also set the number of reducer equal to the number of buckets by using **SET MAPRED.REDUCE.TASK = NUM_BUCKET**.

20. Hash partitioning algorithm?

Hash_function (column) % no.of reducers - partitioner in map reduce
Hash_function (partitioned_column) % no.of partitions - partition algorithm
Hash_function (bucketing_column) % no.of buckets - bucketing algorithm

21. Hive functions?

- **Mathematical**
  Round,ceil,floor,sqrt,exp,tan,cos,sin
- **Aggregate**
  Count(*),sum(),avg(),min(),max()
- **In-built functions**
  Length(),reverse(),concat(),substr(),upper(),lower(),trim,ltrim,rtrim

22. Hive vs rdbms?

Hive

- Hive uses hql/sql language.
- Hive uses hdfs as storage.
- Hive used for analysis.
- Metadata stored in meta store --- a database in rdbms.

rdbms

- Oracle db uses sql language.
- Oracle uses ntfs/ext4 etc as storage.
- Used for real time transactions.
- Metadata stored in rdbms.

23. **Where is the hive data stored/ what is the default hive warehouse directory?**

- In an hdfs directory – **/user/hive/warehouse**, the hive table is stored, by default.
- We can manually change default directory to desired directory using **hive.metastore.warehouse.dir** configuration parameter in the **hive-site.xml**.

24. What is the difference between internal table (create table) and external table (create external table) in hive? When to choose "internal table" and "external table"?

Two types of tables in hive. Such as:

- Internal table
- External table

Managed table:

- In the managed table, both the data and schema are under control of hive.also known as the internal table, these types of tables manage the data and move it into its warehouse directory by default.
- If one drops a managed table, the metadata information along with the table data is deleted from the hive warehouse directory.
- To create the internal table, we use the command ‘create table’
- In hive you can choose internal table, if the processing data available in local file system if we want hive to manage the complete lifecycle of data including the deletion.

External table:

- In the external table, only the schema is under the control of hive. External tables in hive refer to the data that is at an existing location outside the warehouse directory
- Hive deletes the metadata information of a table and does not change the table data present in hdfs.
- To create the external table, we use the command ‘create external table’.
- You can choose external table, if processing data available in hdfs useful when the files are being used outside of hive.

25. Can we create hive external table without location and what happens if we drop this table?

- When you create external table without location , the data will be stored in the hive default location : /user/hive/warehouse/<database_name>.db/<table_name>

- If drop this external table, table schema only dropped and data still remains in default path.

26. **Is it possible to change the default location of managed tables in hive, if so, how?**

Yes, by using the **location** keyword while creating the managed table, we can change the default location of managed tables.the user must specify the storage path of the managed table as the value of the **location** keyword.

27. How to delete external table data?

- By using rm -r we delete data table data in
  Hdfs dfs -rm -r user/hive/warehouse/db/order ;

  Or

- Convert external table to internal table using table properties.
  Alter table orders set tablproperties (external = false);

28. Fsck in hdfs?

The fsck hadoop command is used to check the health of the hdfs

Ex : hdfs dfs fsck /user/cloudera

29. Msck in hive?

- Msck :metastore consistency check
- Hive stores a list of partitions for each table in its metastore.
- If partitions are manually added to the distributed file system (dfs), the metastore is not aware of these partitions.
- Running the msck statement ensures that the tables are properly populated.

Ex : msck repair table <tablename>

30. Different types of execution engine in hive

- Map reduce
- Tez
- Spark

31. If you run a select * query in hive, why does it not run mapreduce?

When executing queries like select, filter, limit, it skips mapreduce function and lowers the latency of mapreduce using hive.fetch.task.conversion property.

32. When hive triggers map reduce job?

Hive queries will trigger map reduce jobs only when we specify function at end:
Ex:
Select empid,fname,salary from employee order by salary;
Select sum(salary),dept from employee group by empid,fname,dept;

33. What are the hive performance optimization techniques?

Following are the hive optimization techniques for hive performance tuning,

1. Partitioning
2. Bucketing
3. Optimizing joins
   - ✓ Hive map join /auto map join/map side join/broadcast join
   - ✓ Bucket map-side join
   - ✓ Smb(sort merge bucket join)
4. Types of file formats
   - ✓ Row based file format
   - ✓ Column based file format

5. Cost based optimization.
6. Vectorization
7. Indexing
8. Compression
9. Orc file format
10. Execution engine

34. Cost based optimization (cbo) ?

- Cbo is hive optimization techniques.
- Hive optimizes each query's logical and physical execution plan before submitting for final execution.
- These optimizations based on query cost.
- That results in different potentially decisions: how to order joins, which type of join to perform, the degree of parallelism and others.
- Cbo will work only for orc file formats in hive.
- To use cbo, set the following parameters at the beginning of your query:
  Set hive.cbo.enable=true;
  Set hive.compute.query.using.stats=true;
  Set hive.stats.fetch.column.stats=true;
  Set hive.stats.fetch.partition.stats=true;

35. Indexing?

- Basically, we are creating the pointer to particular column name of the table,
- The user has to manually define the hive index.
- In hive, the index table is different than the main table.
- First, the index of the column is checked and then the operation is performed on that column only.
- Without an index, queries involving filtering with the "where" clause would load an entire table and then process all the rows
- Indexing a table helps in performing any operation faster.
- Indexing in hive is present only for orc file format, as it has a built-in index.
  Create index index_name on table table_name (col_name);

36. Vectorization?

- It process batch of records (1024 rows) instead of an each row.
- This feature is introduced in hive 0.13.
- Columnar format (orc) is must.
- Enabled with two parameters settings:

Set hive.vectorized.execution = true

Set hive.vectorized.execution.enabled = true

37. Compression? How to reduce data transfer? How to compress data for snappy and ggip format file in hive?

Hive queries involve a lot of disks i/o operations, network i/o operations,data transferring between mappers and reducers, which can be easily reduced by reducing the size of the data which is done by compression.

Following are the main situations where compression can save cost:

- Reading data from a local dfs directory
- Reading data from a non-local dfs directory
- Moving data from reducers to the next stage reducers
- Moving the final output back to the dfs.

The above optimizations will save a whole lot of execution cost and will lead to pretty quicker execution of jobs. As a result, the overall hive query will have better performance.

Set hive.exec.compress.output=true;
Set mapred.output.compression.codec=org.apache.hadoop.io.compress.snappycodec;
Set mapred.output.compression.type=block;

38. What are a map join, bucket map join and sort merge bucket join in hive?

Map join in hive

- Map join is a hive feature that is used to speed up hive queries.
- Map join is a type of join where a smaller table is loaded in memory and the join is done in the map phase of the mapreduce job.
- As no reducers are necessary, map joins are way faster than the regular joins.
- Only **hive.auto.convert.join** is a map join option that must be set to true while other options are optional and have default values. So, it is not necessary to change the default values.

Bucket map join

- A bucket map join is used when the tables are large and all the tables used in the join are bucketed on the join columns.
- For example, if one table has 2 buckets then the other table must have either 2 buckets or a multiple of 2 buckets (2, 4, 6, and so on).
- If the preceding condition is satisfied then the joining can be done at the mapper side only.

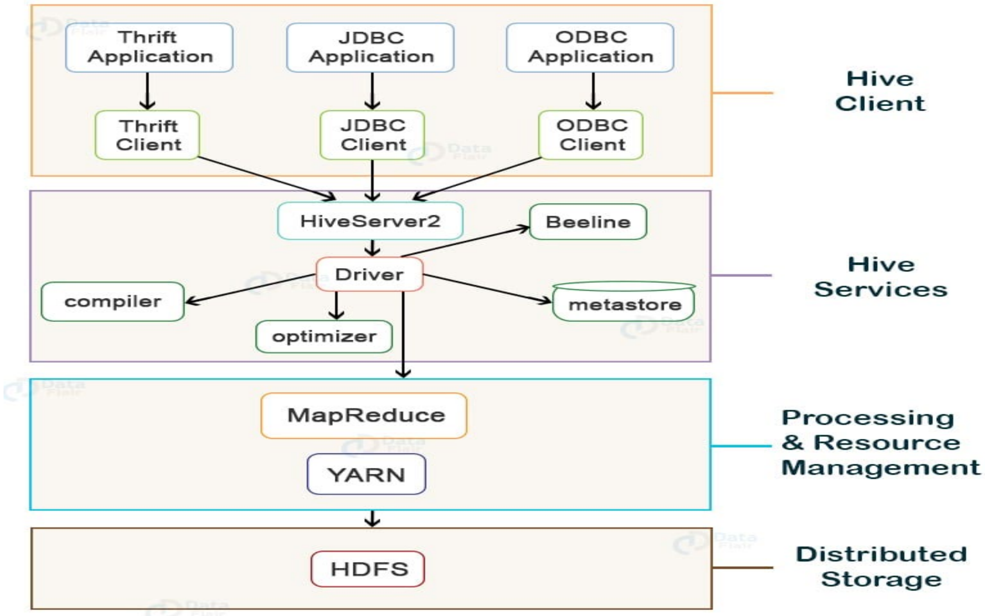
sort merge buket join

- In smb each mapper reads a bucket from the first table and the corresponding bucket from the second table and then a merge, sort,join is performed.
- All tables should have the same number of buckets in smb join.
- Sort merge bucket (smb) join in hive is mainly used as there is no limit on file or partition or table join. Smb join can best be used when the tables are large.

39. What are the different components of hive architecture?

The major components of apache hive are:

- Hive client
- Hive services
- Processing and resource management
- Distributed storage



Hive Architecture & Its Components

40. How can clients interact with hive? (or) what options are available when it comes to attaching applications to the hive server?

There are 3 ways possible in which a client can interact with the hive.

- **Thrift client:**

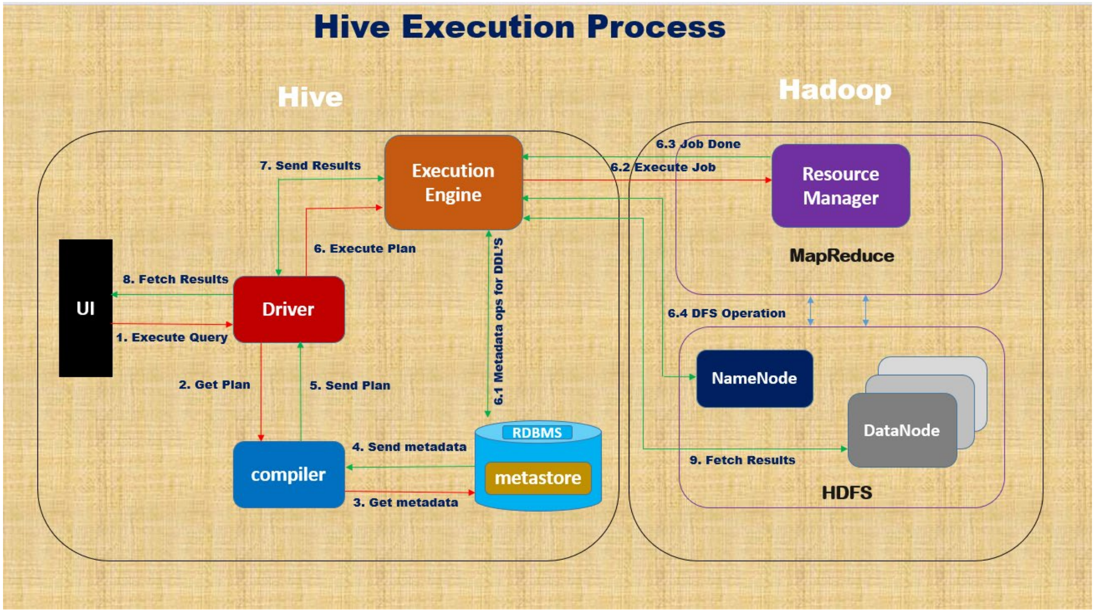
Thrift is a software framework. Basically, that allows access to hive over a single port.it allows clients using languages including java, c++, ruby, and many others, to programmatically access hive remotely.

- **Jdbc client:**
  we can also use an jdbc driver application to connect to the hive server.the beeline cli uses jdbc driver.

- **Odbc client:**
  we can also use an odbc driver application to connect to the hive server.

41. The various services offered by hive are?

- **Driver:**
  The driver manages the life cycle of hive ql queries. It receives queries from ui.it designs a session handle for the query, and then the queries are sent to the compiler for the execution plan.
- **Metastore:**
  It contains the organized data and information on various tables and their partitions in hive warehouse.
- **Compiler:**
  It creates the execution plan for the queries, performs semantic analysis on different query blocks, and generates query expressions.
- **Execution engine:**
  It implements the execution plans created by the compiler. This acts as a bridge between the hive and hadoop to process the query. Execution engine communicates bidirectionally with metastore to perform operations, such as create or drop tables.



42. Explain what is metastore in hive?

Metastore is a central repository in hive.
It is used for storing metadata hive partitions, tables, databases, and so on.
By default, the metastore service runs in the same jvm as the hive service running

43. Why does hive not store metadata information in hdfs?

We know that the hive's data is stored in hdfs. However, the metadata is either stored locally or it is stored in rdbms. The metadata is not stored in hdfs, because hdfs read/write operations are time-consuming. This allows us to achieve low latency and is faster. Hive stores metadata information in the metastore using rdbms instead of hdfs.

44. What are the 3 different ways to setup the metastore server?

There are three modes for metastore deployment which hive offers.

1. Embedded metastore

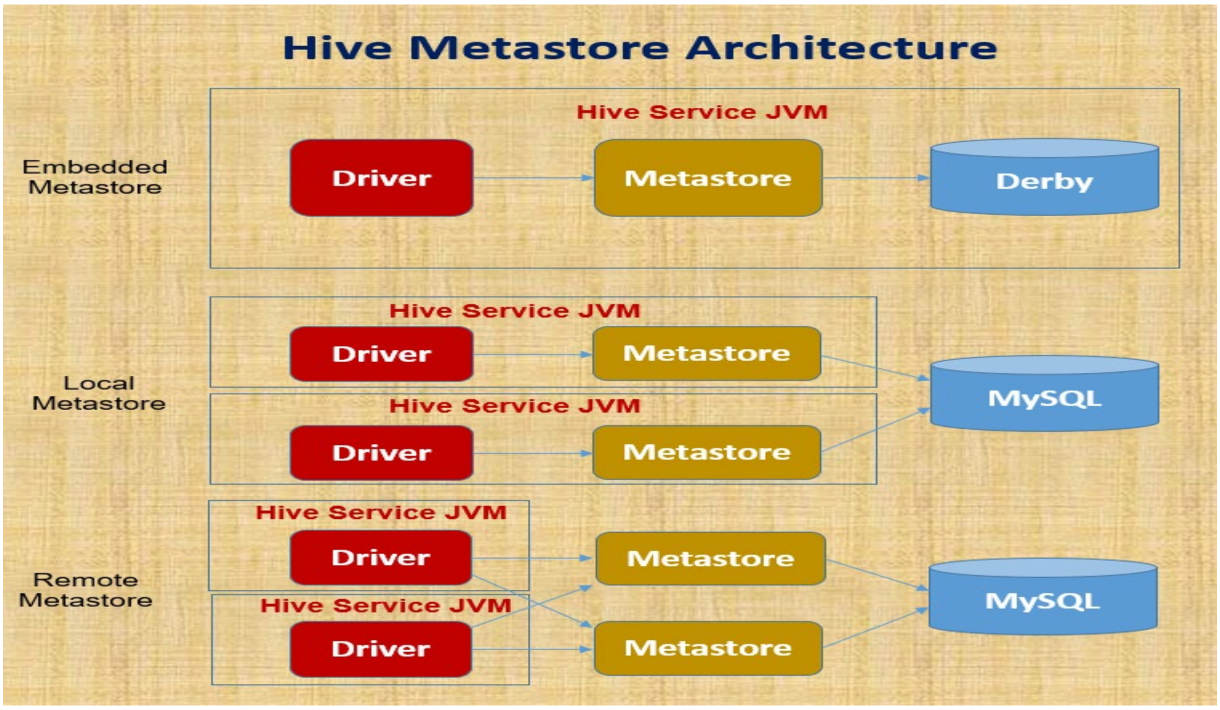
- Derby is the default database for the embedded metastore.
- Both metastore service and hive service runs in the same jvm.
- We can only have one hive session open at a time
- Trying to start a second session gives the error: failed to start database 'metastore_db'

2. Local metastore

- In local metastore, a standalone database (mysql/postgresql /other db with jdbc driver) is used as a metastore.
- In this mode, multiple users can open connection to metastore and make sql queries against it.
- The metastore service runs in the same jvm in which the hive service is running and connects to a database running in a separate jvm.

3. Remote metastore

- In remote metastore, all hive clients will make a connection to a metastore server (mysql or postgresql or oracle) which in turn queries the datastore.
- Here, metastore runs on its own separate jvm, hive service runs on its own separate jvm.



45. Is it possible to use the same metastore by multiple users, in case of embedded hive?

No, it is not possible to use metastore in sharing mode. It is recommended to use Standalone "real" database like mysql or postgresql.

46. Can multiple users use one metastore?

No, metastore sharing is not supported by hive.

47. How can you configure remote metastore mode in hive?

Basically, hive-site.xml file has to be configured with the below property, to configure metastore in hive – **Hive.metastore.uris thrift: //node1 (or ip address):9083 ip address and port of the metastore host**

48. Wherever (different directory) we run hive query, it creates new metastore_db, please explain the reason for it?

Whenever we run the hive in embedded mode, it creates the local metastore. And before creating the metastore it looks whether metastore already exist or not. This property is defined in configuration file hive-site.xml.

Property is "javax.jdo.option.connectionurl" with default value **"jdbc:derby:;databasename=metastore_db;create=true".**

So to change the behavior change the location to absolute path, so metastore will be used from that location.

49. What is the default database provided by apache hive for metastore?

By default, hive provides an embedded derby database instance backed by the local disk for the metastore. This is called the embedded metastore configuration.

50. What is the difference between order by and sort by in hive?

Sort by

- Multiple mappers output data goes through multiple corresponding reducers.
- SORT BY will sort the data within each reducer. We can use any number of reducers for SORT BY operation.
- If there are multiple reducers, the total order would be missing.
- It gives partially ordered result.

Order by

- All mappers output data will pass through the one reducer only.
- It will sort all the data together.
- It gives fully ordered result.
- Order by query takes more time then sort by.

51. Explain about acid transactions in hive? When hive table eligible for acid transactions?

Hive supports acid transactions. Hive acid table allow us to insert, update, delete.

Requirements for hive acid table:

- Table should be internal table.
- Table should use buckets.
- Use only orc file format.
- Table properties (transactional = true)

52. Difference between row based and column based file format?

Row based file format:

- New record enters at end so row based file provides faster writes.
- It has to scan entire row so performance slow.
- Less compression then column based file format due to different data types together.

Ex: avro

File stored as row:

1,sandeep,27,1000,mech 2,satish,30,4000,cse 3,anusha,24,500,degree

Column based file format:

- It provides faster reads.
- Writing is slower
- Very good compression than row based.

  Ex:orc,parquet

File stored as row:

1 2 3 sandeep satish anusha 27 30 24 1000 4000 500 mech cse dgree

53. input file formats in hadoop?

- Text files
- Sequence files
- Avro files
- Orc files
- Parquet files

54. Text files file formats?

A text file is the most basic and a human-readable file. It can be read or written in any programming language and is mostly delimited by comma or tab.
It is also difficult to represent binary data such as an image.

55. sequence files (flat file)? What is the purpose of sequence file?

It is a specific file format which stores data in binary format. They store key-value pairs in a binary container format. However, sequence files are not human- readable.

Purpose of sequence file:

- To enable/store/process binary data
- Sequence file is used in mapreduce as input/output formats.
- The other objective of using sequence file is to pack many small files into a single large sequence file hadoop
- Hadoop and spark are optimized for large files, so packing small files into a sequence file makes storing and processing the smaller files more efficient.

56. **Orc file format?**

- Orc refers to optimized row columnar.
- It reduces the size of the original data up to 75%.
- Hence, data processing speed also increases.
- Orc format improves the performance than other file formats like rc and sequence etc.

57. Difference between orc and parquet and avro file formats?

Orc:

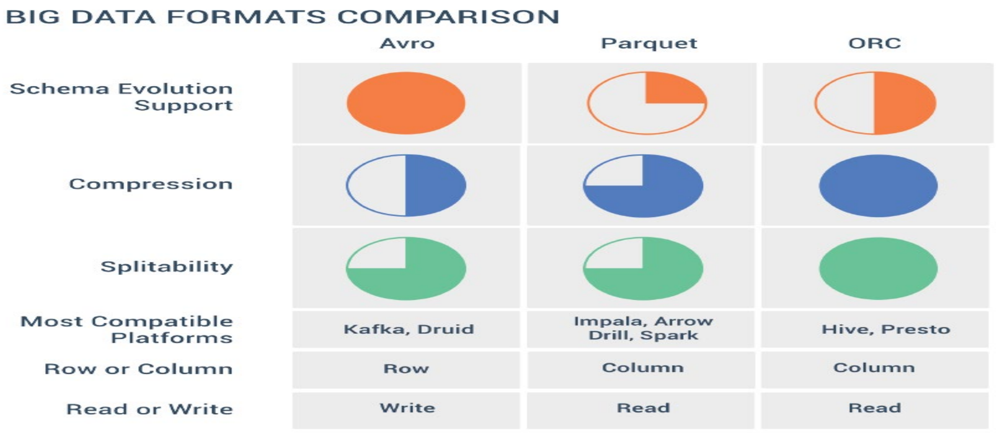
- It belongs to column-based file formats.
- Orc is best suited when you’re working with hive.
- Orc takes less storage than parquet.
- 1gb data stored in 200mb.
- Searching is faster in orc because it uses indexes when compared with parquet.
- It is designed to well work with hive.

Parquet:

- It belongs to column-based file formats.
- Parquet is much better for analytical querying.
- Parquet offers faster reads but slower writes.
- Parquet best suited when you’re working with spark.
- 1gb data stored in 250mb.
- For nested data parquet is a wonderful fit.
- It is designed with general intension (it should support most of the things).
- Compatibility with other platforms parquet is better than orc.
- It just supports at end. You can just append or delete the columns from the end.

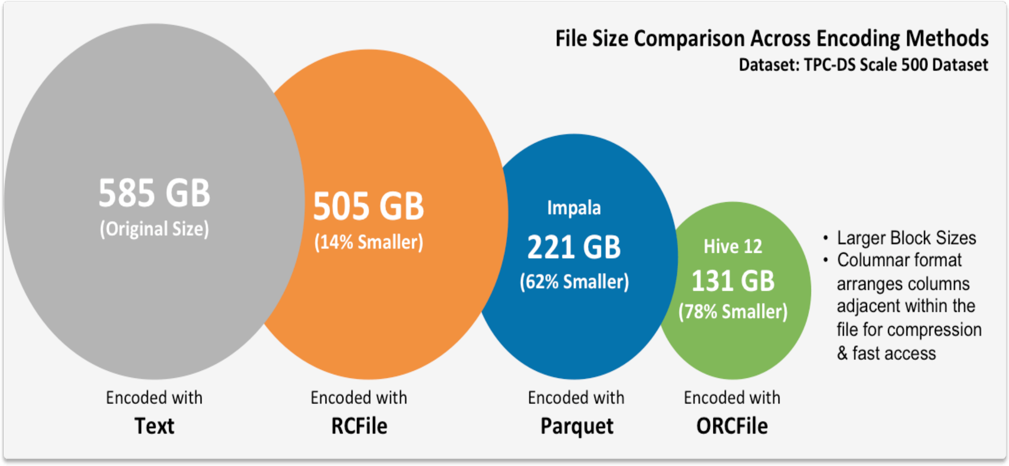
Avro:

- Avro is a row-based file format.
- Avro provides faster writes but slower reads.
- Avro is ideal in the case of etl operations where we need to query all the columns.
- Avro is quite mature at schema evolution i.e adding or modifying columns.



BIG DATA FORMATS COMPARISON

	Avro	Parquet	ORC
Schema Evolution Support	●	◔	◑
Compression	◑	◕	●
Splitability	◕	◕	●
Most Compatible Platforms	Kafka, Druid	Impala, Arrow Drill, Spark	Hive, Presto
Row or Column	Row	Column	Column
Read or Write	Write	Read	Read



58. Does hive provide oltp or olap?

Hive doesn't provide crucial features required for oltp, online transaction processing.it's closer to being an olap tool, online analytic processing. So, hive is best suited for data warehouse applications, where a large data set is maintained and mined for insights,reports, etc.

59. **Why is hive not suitable for oltp systems?**

Hive is not suitable for oltp systems because it does not provide insert and update function at the row level.

60. What kind of applications is supported by apache hive?

Hive supports all those client applications that are written in:

- java
- php
- python
- c++
- ruby

By exposing its thrift server.

61. **Mention what are the different modes of hive?**

Depending on the size of data nodes in hadoop, hive can operate in two modes.
These modes are,

- local mode
- map reduce mode

62. is multi line comment supported in hive script ?

No.

63. Which classes are used by the hive to read and write hdfs files?

Following classes are used by hive to read and write hdfs files

- **Textinputformat/hiveignorekeytextoutputformat:** these 2 classes read/write data in plain text file format.
- **Sequencefileinputformat/sequencefileoutputformat**: these 2 classes read/write data in hadoop sequencefile format.

64. **How to skip header rows from a table in hive?**

```
CREATE EXTERNAL TABLE EMPLOYEE (NAME STRING,
JOB STRING,
DOB STRING,
ID INT,
SALARY INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ‘ ‘
STORED AS TEXTFILE
LOCATION ‘/USER/DATA’ TBLPROPERTIES("SKIP.HEADER.LINE.COUNT"="2”);
```

65. **What is a hive variable? What do we use it for?**

Hive variables are basically created in the hive environment that is referenced by hive scripting languages. They allow to pass some values to a hive query when the query starts executing. They use the source command.

66. Explain the process to access subdirectories recursively in hive queries.

By using the below commands, we can access subdirectories recursively in hive:

```
Hive> set mapred.input.dir.recursive=true;
Hive> set hive.mapred.supports.subdirectories=true;
```

Hive tables can be pointed to the higher level directory, and this is suitable for the directory structure like: /data/country/state/city/

67. **Can we change the settings within a hive session? If yes, how?**

- Yes, we can change the settings within a hive session using the set command. It helps change the hive job settings for an exact query.
  For example : **set hive.enforce.bucketing=true;**
- The below code will list all the properties including the hadoop defaults in the system.
  Set -v

68. Explain rlike in hive?

Rlike: its full form is right-like and it is a special function in hive. It helps examine two substrings, i.e., if the substring of a matches with b, then it evaluates to true.
Example:
‘intellipaat’ rlike ‘tell’ true

69. **Mention when to use map reduce mode?**

Map reduce mode is used when,

- It will perform on large amount of data sets and query going to execute in a parallel way
- Hadoop has multiple data nodes, and data is distributed across different node we use hive in this mode
- Processing large data sets with better performance needs to be achieved

70. Mention what are the type of database does hive support?

For single user metadata storage, hive uses derby database and for multiple user metadata or shared metadata case hive uses mysql.

71. Which java class handles the input record encoding into files which store the tables in hive?
Ans.org.apache.hadoop.mapred.textinputformat

72. **Which java class handles the output record encoding into files which result from hive queries?**
Ans. Org.apache.hadoop.hive.ql.io.hiveignorekeytextoutputformat

73. Which classes are used by the hive to read and write hdfs files?(or) hive default read and write classes?

Ans. Following classes are used by hive to read and write **hdfs** files

Textinputformat/hiveignorekeytextoutputformat: basically, it read/write data in plain text file format.

Sequencefileinputformat/sequencefileoutputformat: however, it read/write data in hadoop sequencefile format.

74. What is the hive objectinspector function?

It helps to analyze the structure of individual columns and rows and provides access to the complex objects that are stored within the database.

75. Mention what is (hs2) hiveserver2?

It is a server interface that performs following functions.

- It allows remote clients to execute queries against hive
- Retrieve the results of mentioned queries

Some advanced features based on thrift rpc in its latest version include

- Multi-client concurrency
- Authentication

76. Mention what hive query processor does?

Hive query processor convert graph of mapreduce jobs with the execution time framework. So that the jobs can be executed in the order of dependencies.

77. WHAT ARE THE COMPONENTS USED IN HIVE QUERY PROCESSOR?

The components of a hive query processor include

- Logical plan of generation.
- Physical plan of generation.
- Execution engine.
- Operators.
- Udf's and udaf's.
- Optimizer.
- Parser.
- Semantic analyzer.
- Type checking

78. Mention what are views in hive?

views are similar to tables, in hive. Basically, they are generated based on the requirements.

Also, we can save any result set data as a view in hive.

Although, its usage is similar to as views used in sql.while we can perform all type of dml operations on a view.

79. Can we use the load or insert command to view?

No, these commands cannot be used with respect to a view in hive.

80. **Can we load data into a view?**

Ans. No.

81. **Mention if we can name view same as the name of a hive table?**

No. The name of a view must be unique compared to all other tables and as views present in the same database.

82. In hive, can you overwrite hadoop mapreduce configuration in hive?

Yes, you can overwrite hadoop mapreduce configuration in hive.

83. **Explain when to use explode in hive?**

Hadoop developers sometimes take an array as input and convert into a separate table row. To convert complex data types into desired table formats, hive use explode.

84. Mention how can you stop a partition form being queried?

You can stop a partition form being queried by using the enable offline clause with alter table statement.

85. Is there a data type in hive to store date information?

The timestamp data type in hive stores all data information in the java.sql.timestamp format.

86. **What are the hive collection data types?**

Array, map, and struct are the three hive collection data types.

87. **Is it possible to run unix shell commands in hive?**

Yes, one can run shell commands in hive by adding a '!' before the command.

88. **Is it possible to execute hive queries from a script file?**

Yes, one can do so with the help of a source command. For example

– hive> source /path/queryfile.hql

89. What is a .hiverc file?

It is a file that consists of a list of commands that need to be run when the command line input is initiated.

Basically, when the hive cli starts, it is a file containing the list of commands needs to run. Like, setting the strict mode to be true etc.

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

Show databases like 'c.*'

91. **How do you list all databases whose name starts with p?**

Show databases like ‘p.*’

92. **Is it possible to delete dbproperty in hive?**

No, there is no way to delete the dbproperty.

93. **When a hive table partition is pointed to a new directory, what happens to the data?**

The data remains in the old directory and needs to be transferred manually.

94. **Do you save space in the hdfs by archiving hive tables?**

No, archiving hive tables only helps reduce the number of files that make for easier management of data.

95. **What is a table generating function on hive?**

Mapreduce is a programming framework that allows hive to divide large datasets into smaller units and process them parallelly.

96. **Can you avoid mapreduce on hive?**

You can make hive avoid mapreduce to return query results by setting the hive.exec.mode.local.auto property to ‘true’.

97. **Can a cartesian join be created between two hive tables?**

This is not possible as it cannot be implemented in mapreduce programming.

98. **Are multi-line comments supported by hive?**

No, multi-line comments are supported by hive.

99. **How can you view the indexes of a hive table?**

By using the following command:

Show index on table_name

100. **Can you specify the name of the table creator in hive?**

Yes, by using the tblproperties clause.

For example – **tblproperties (‘creator’= ‘john’)**

101. **What does /*streamtable(table_name)*/ do?**

It is a query hint that allows for a table to be streamed into memory before a query is executed.

102. How does data transfer happen from hdfs to hive?

Ans. Basically, the user need not load data that moves the files to the /user/hive/warehouse/.

But only if data is already present in hdfs. Hence, using the keyword external that creates the table definition in the hive metastore the user just has to define the table.

```
Create external table table_name (
 id int,
 myfields string
)
Location '/my/location/in/hdfs';
```

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

Ans. The default record delimiter is − \n

and the filed delimiters are − \001,\002,\003

104. What do you mean by schema on reading?

Ans. However, while reading the data and not enforced when writing data, the schema is validated with the data.

105. What does the "use" command in the hive do?

Ans. Basically, fix the database on which all the subsequent hive queries will run we use the "use" command in hive.

106. What is the maximum size of string data type supported by hive?

Ans. Maximum size is 2 gb.

107. What is the significance of 'if exists" clause while dropping a table?

Ans. Since, the table being dropped does not exist in the first place, hive throws an error, when we issue the command drop table if exists table_name.

108. What types of costs are associated with creating the index on hive tables?

Basically, there is a processing cost in arranging the values of the column on which index is created since indexes occupies.