1 What is Hive?

Hive is a data warehouse software which is used for facilitates querying and managing large data sets residing in distributed storage.

Hive language almost look like SQL language called HiveQL. Hive also allows traditional map reduce programs to customize mappers and reducers when it is inconvenient or inefficient to execute the logic in HiveQL (User Defined Functions UDFS)

2 What is Hive Metastore?

Hive metastore is a database that stores metadata about your Hive tables (eg. Table name, column names and types,table location, storage handler being used, number of buckets in the table, sorting columns if any, partition columns if any, etc.).

When you create a table,this metastore gets updated with the information related to the new table which gets queried when you issue queries on that table.

Hive is a central repository of hive metadata. it has 2 parts services and data. by default it uses derby db in local disk. it is referred as embedded metastore configuration. It tends to the limitation that only one session can be served at any given point of time.

3 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.

4 What is Object Inspector functionality?

Hive uses Object Inspector to analyze the internal structure of the row object and also the structure of the individual columns.

Object Inspector provides a uniform way to access complex objects that can be stored in multiple formats in the memory, including:

•Instance of a Java class (Thrift or native Java)

•A standard Java object (we use java.util.List to represent Struct and Array, and use java.util.Map to represent Map)

•A lazily-initialized object (For example, a Struct of string fields stored in a single Java string object with starting offset for each field).

A complex object can be represented by a pair of ObjectInspector and Java Object. The ObjectInspector not only tells us the structure of the Object, but also gives us ways to access the internal fields inside the Object.

5 What is the functionality of Query Processor in Apached Hive?

This component implements the processing framework for converting SQL to a graph of map/reduce jobs and the execution time framework to run those jobs in the order of dependencies.

6 If you run hive as a server, what are the available mechanism for connecting it from application?

There are following ways by which you can connect with the Hive Server:

1. Thrift Client: Using thrift you can call hive commands from a various programming languages e.g. C++,Java, PHP, Python and Ruby.

2. JDBC Driver : It supports the Type 4 (pure Java) JDBC Driver

3. ODBC Driver: It supports ODBC protocol.

7 What kind of data warehouse application is suitable for Hive?

Hive is not a full database. The design constraints and limitations of Hadoop and HDFS impose limits on what Hive can do.

Hive is most suited for data warehouse applications, where

1) Relatively static data is analyzed,

2) Fast response times are not required, and

3) When the data is not changing rapidly.

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.

8 Which database hive used for Metadata store? What are the metastore configuration hive supports?

Hive can use derby by default and can have three type metastore configuration. It supports

- Embedded Metastore

- Local Metastore
- Remote Metastore

Embedded uses derby db to store data backed by file stored in disk. It can't support multi session at same time and services of metastore runs in same JVM as hive.

Local Metastore:

In this case we need to have stand alone db like MySql, which would be communicated by metastore services.Benefit of this approach is, it can support multiple hive session at a time. and service still runs in same process as Hive.

Remote Metastore:

Metastore and Hive service would run in different process. with stand alone Mysql kind db.

9 what are Binary storage formats hive supports?

Hive natively supports text file format, however hive also has support for other binary formats. Hive supports Sequence, Avro, RCFiles.

1. Sequence files :-General binary format. splittable, compressible and row oriented. a typical example can be.if we have lots of small file, we may use sequence file as a container, where file name can be a key andcontent could stored as value. it support compression which enables huge gain in performance.

2. Avro datafiles:-Same as Sequence file splittable, compressible and row oriented except support of schema evolution and multilingual binding support.

3. RCFiles :-Record columnar file, it's a column oriented storage file. it breaks table in row split. in each split stores that value of first row in first column and followed sub subsequently..

10 Is it possible to use 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 PostGresSQL.

11 What is Apache Hcatalog ?

HCatalog is built on top of the Hive metastore and incorporates Hive's DDL. Apache Hcatalog is a table and data management layer for hadoop,we can process the data on Hcatalog by using APache pig,Apache Mapreduce and Apache Hive. There is no need to worry in Hcatalog where data is stored and which format of data generated.

HCatalog displays data from RCFile format, text files, or sequence files in a tabular view. It also provides REST. APIs so that external systems can access these tables' metadata.

12 What is the work of Hive/Hcatalog ?

Hive/HCatalog also enables sharing of data structure with external systems including traditional data management tools.

13 What is WebHCatServer ?

The WebHcatServer provides a REST – like web API for Hcatalog. Applications make HTTP requests to run Pig,Hive, and HCatalog DDL from within applications.

14 What is Hive Present Version ?

Hive-0.13.1

15 What is the stable version of Hive ?

Hive-0.12.0

16 Is it possible to create multiple table in hive for same data?

Hive creates schema and append on top of an existing data file. One can have multiple schema for one data file,schema would be saved in hive’s metastore and data will not be parsed read or serialized to disk in given schema.When s/he will try to retrieve data schema will be used. Lets say if my file have 5 column(Id,Name,Class,Section,Course) we can have multiple schema by choosing any number of column.

17 Wherever (Different Directory) I run hive query, it creates new metastore_db, please explain the reason for it?

Whenever you 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”withdefaultvalue“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.

18 What is SerDe in Apache Hive ?

A SerDe is a short name for a Serializer Deserializer.

Hive uses SerDe (and FileFormat) to read and write data from tables. An important concept behind Hive is that it DOES NOT own the Hadoop File System (HDFS) format that data is stored in. Users are able to write files to HDFS with whatever tools/mechanism takes their fancy(“CREATE EXTERNAL TABLE” or “LOAD DATA INPATH,”) and use Hive to correctly “parse” that file format in a way that can be used by Hive.

A SerDe is a powerful (and customizable) mechanism that Hive uses to “parse” data stored in HDFS to be used by Hive.

19 Give examples of the SerDe classes which hive uses to Serialize and Deserilize data ?

Hive currently use these SerDe classes to serialize and deserialize data:

- MetadataTypedColumnsetSerDe: This SerDe is used to read/write delimited records like CSV, tab-separated control-A separated records (quote is not supported yet.)

- ThriftSerDe: This SerDe is used to read/write thrift serialized objects. The class file for the Thrift object must be loaded first.

- DynamicSerDe: This SerDe also read/write thrift serialized objects, but it understands thrift DDL so the schema of the object can be provided at runtime. Also it supports a lot of different protocols,including TBinaryProtocol, TJSONProtocol, TCTLSeparatedProtocol (which writes data in delimited records).

20 How do you write your own custom SerDe ?

In most cases, users want to write a Deserializer instead of a SerDe, because users just want to read their own data format instead of writing to it.

•For example, the RegexDeserializer will deserialize the data using the configuration parameter 'regex', and possibly a list of column names.

•If your SerDe supports DDL (basically, SerDe with parameterized columns and column types), you probably want to implement a Protocol based on DynamicSerDe, instead of writing a SerDe from scratch. The reason is that the framework passes DDL to SerDe through "thrift DDL" format, and it's non-trivial to write a "thrift DDL" parser.

21 What are the types of tables in Hive?

There are two types of tables.

1. Managed tables.

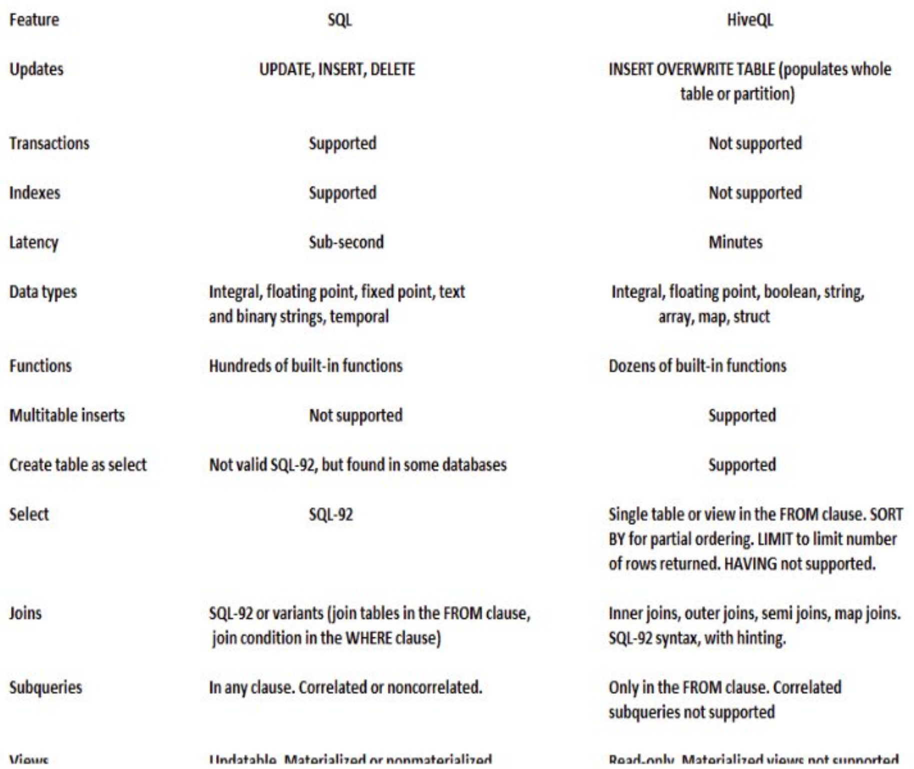
2. External tables.

Only the drop table command differentiates managed and external tables. Otherwise, both type of tables are very similar. I When you drop an internal table, it drops the data, and it also drops the metadata. When you drop an external table, it only drops the meta data. That means hive is ignorant of that data now. It does not touch the data itself.

22 Does Hive support record level Insert, delete or update?

Hive does not provide record-level update, insert, or delete. Henceforth, Hive does not provide transactions too. However, users can go with CASE statements and built in functions of Hive to satisfy the above DML operations. Thus, a complex update query in a RDBMS may need many lines of code in Hive.

23 Difference between SQL and HiveQL ?



Feature	SQL	HiveQL
Updates	UPDATE, INSERT, DELETE	INSERT OVERWRITE TABLE (populates whole table or partition)
Transactions	Supported	Not supported
Indexes	Supported	Not supported
Latency	Sub-second	Minutes
Data types	Integral, floating point, fixed point, text and binary strings, temporal	Integral, floating point, boolean, string, array, map, struct
Functions	Hundreds of built-in functions	Dozens of built-in functions
Multitable inserts	Not supported	Supported
Create table as select	Not valid SQL-92, but found in some databases	Supported
Select	SQL-92	Single table or view in the FROM clause. SORT BY for partial ordering. LIMIT to limit number of rows returned. HAVING not supported.
Joins	SQL-92 or variants (join tables in the FROM clause, join condition in the WHERE clause)	Inner joins, outer joins, semi joins, map joins. SQL-92 syntax, with hinting.
Subqueries	In any clause. Correlated or noncorrelated.	Only in the FROM clause. Correlated subqueries not supported
Views	Updatable. Materialized or nonmaterialized	Read-only. Materialized views not supported

24 what is Partition?

To increase performance Hive has the capability to partition data

1 The values of partitioned column divide a table into segments

2 Entire partitions can be ignored at query time

3 Similar to relational databases' indexes but not as granular

25 what is bucketing?

Mechanism to query and examine random samples of data

- Break data into a set of buckets based on a hash function of a "bucket column" .