What is Hive?
It's an open source project under the Apache Software Foundation, it's a data warehouse software ecosystem in Hadoop. Which manage vast amount of structured data sets, by using HQl language; it's similar to SQL.

Where hive is the best suitable?
When you are doing data warehouse applications,
Where you are getting static data instead of dynamic data,
when the application on high latency (response time high).
where a large data set is maintained and mined for insights, reports.
When we are using queries instead of scripting we use Hive.

When hive is not suitable?
It doesn't provide OLTP transactions supports only OLAP transactions.
If application required OLTP, switch to NoSQL databases.
HQL queries have higher latency, due to the mapreduce.

Hive Support Acid Transactions?
By default it doesn't support record-level update, insert and delete, but recent Hive 1.4 later versions supporting insert, update and delete operations. So hive support ACID transactions. To achieve updates & deletion transactions in 1.4 version, you must change given default values.
hive.support.concurrency – true
hive.enforce.bucketing – true
hive.exec.dynamic.partition.mode – nonstrict
hive.txn.manager – org.apache.hadoop.hive.ql.lockmgr.DbTxnManager
hive.compactor.initiator.on – true (for exactly one instance of the Thrift metastore service)
hive.compactor.worker.threads – a positive number on at least one instance of the Thrift metastore service

What is Hive MetaStore?
MetaStore is a central repository of Hive, that allows to store meta data in external database. By default Hive store meta data in Derby database, but you can store in MySql, Oracle depends on project.

Why I choose Hive instead of MapReduce?
There are Partitions to simplify the data process, Bucketing for sampling the data, sort the data quickly, and simplify the mapreduce process. Partitions and Buckets can segmenting large data sets to improve Query performance in Hive. So It is highly recommendable for structure data.

Can I access Hive without Hadoop?
Hive store and process the data on the top of Hadoop, but it's possible to run in Other data storage systems like Amazon S3, GPFS (IBM) and MapR file systems.

What is the relationship between MapReduce and Hive? or How Mapreduce jobs submits on the cluster?
Hive provides no additional capabilities to MapReduce. The programs are executed as MapReduce jobs via the interpreter. The Interpreter runs on a client machine which rurns HiveQL queries into MapReduce jobs. Framework submits those jobs onto the cluster.

If you run select * query in Hive, why it's not run Mpareduce?
It's an optimization technique. hive.fetch.task.conversion property can (FETCH task) minimize latency of mapreduce overhead. When queried SELECT, FILTER, LIMIT queries, this property skip mapreduce and using FETCH task. As a result Hive can execute query without run mapreduce task. By default it's value "minimal". Which optimize: SELECT STAR, FILTER on partition columns, LIMIT queries only, where as another value is "more" which optimize :
SELECT, FILTER, LIMIT only (+TABLESAMPLE, virtual columns).

What is the importance of Vectorization in Hive?
It's a query optimization technique. Instead of processing multiple rows, Vectorization allows to process process a batch of rows as a unit. Consequently it can optimize query performance. The file must be stored in ORC format to enable this Vectorization. It's disabled by default, but enable this property by run this command.

```
set hive.vectorized.execution.enabled=true;
```

How Hive can improve performance with ORC format tables?
Hive can store the data in highly efficient manner in the Optimized Row Columnar (ORC) file format. It can ease many Hive file format limitations. Using ORC files can improves the performance when reading, writing, and processing data. Enable this format by run this command and create table like this.

```
set hive.compute.query.using.stats=true;
set hive.stats.dbclass=fs;
CREATE TABLE orc_table (
id int,
name string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\;'
LINES TERMINATED BY '\n'
STORED AS ORC;
```

Difference between sort by or order by clause in Hive? Which is the fast?
ORDER BY – sort the data in one reducer. Sort by much faster than order by.
SORT BY – sort the data within each reducer. You can use n number of reducers for sort.
In the first case (order by) maps sends each value to the single reducer and count them all.
In the second case (sort by) maps splits up the values to many reducers and each reduce generates its list and finds the count. So it can sort quickly.
Example:

```
SELECT name, id, cell FROM user_table ORDER BY id, name;
SELECT name, id, cell FROM user_table DISTRIBUTE BY id SORT BY name;
```

Wherever you run hive query, first it creates new metastore_db, why? What is the importance of Metastore_db?
When we run the hive query, first it creates a local metastore, before creates the metastore first Hive checks whether metastore is already exists or not? If presents shows error, else the process goes on. This configuration is set in hive-site.xml like this.

```
<property>
<name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:derby:;databaseName=metastore_db;create=true</value>
<description>JDBC connect string for a JDBC metastore</description>
</property>
```

Hive can process any type of data formats?
Yes, Hive uses the SerDe interface for IO operations. Different SerDe interfaces can read and write any type of data. If normal directly process the data where as different type of data is in the Hadoop, Hive use different SerDe interface to process such data.
Example:
MetadataTypedColumnsetSerDe: used to read/write CSV format data.
JsonSerDe: process Json data.
RejexSerDe: process weblog data.
AvroSerde: Avro format data.

Tell me different Hive metastore configuration.
There are three types of metastores configuration called
1) Embedded metastore
2) Local metastore
3) Remote metastore.
If Hive run any query first it enter into embedded mode, It's default mode. In Command line all operations done in embedded mode only, it can access Hive libraries locally. In the embedded metastore configuration, hive driver, metastore interface and databases use same JVM. It's good for development and testing.
In local metastore the metastore store data in external databases like MYSQL. Here Hive driver and metastore run in the same JVM, but remotely communicate with external Database. For better protection required credentials in Local metastore. Where as in Remote server, use remote mode to run the queries over Thrift server.
In Remote metastore, Hive driver and metastore interface would be running in a different JVM. So for better protection, required credentials such are isolated from Hive users.

What Is the HWI?
The Hive Web Interface is an alternative to the command line interface. HWI is a simple graphical interface, It's hive web interface. The HWI allows start at database level directly. you can get all SerDe, column names and types and simplifies the hive steps. It's session based interface, so you can run multiple hive queries simultaneously. There is no local metastore mode in HWI.

What is the difference between Like and Rlike operators in HIVE?
Like: used to find the substrings within a main string with regular expression %.
Rlike is a special fuction which also finds the sub strings within a main string, but return true or false without using regular expression.
Example: Tablename is table, column is name.
name=VenuKatragadda, venkatesh, venkateswarlu
Select * from table where name like "venu%. //VenuKatragadda.
select * from table where name rlike "venk%". // false, true, true.

What are the Hive default read and write classes?
Hive use 2+2 classes to read and write the files.
1)TextInputFormat/HiveIgnoreKeyTextOutputFormat.
2) SequenceFileInputFormat/SequenceFileOutputFormat:
First class used to read/write the plain text. Second class used for sequence files.

What is Query processor in Hive?
It's a core processing unit in Hive framework, it converting SQL to map/reduce jobs and run in the other dependencies. As a result hive can convert the Hive queries into Hive queries.

What are Views in Hive?
Based on user requirement create and manage view. You can set data as view. It's a logical construct. It's used where query is more complicated and to hide complexity of query and make easy to the users.
Example:
Create view table_name as select * from employee where salary>10000;

What is different between database and data-warehouse?
Typically database is designed for OLTP transactional operations. Where as Data-warehouse is implemented for OLAP (analysis) operations.
OLTP can constrained to a single application. OLAP resists as a layer on the top of several databases.
OLTP process current, streaming and dynamic data where as OLAP process Retired, historic and static data only.
Database completely has normalization concept. DWH is De-normalization concept.

What is the different between Internal and external tables in Hive?
Hive will create a database on the master node to store meta data to keep data in safe. Let example, If you partition table, table schema stores data in the external table.
In Managed table, Schema stored in the local system, but in External table MetaStore separate from the node and stored in a secure database. In Internal Table, Hive reads and loads entire file as it is to process, but in External simply loads depends on the query logic. If user drop the table, Hive drop original data and MetaStore, but in External table, just drop MetaStore, but not original data. Hive by default store in internal table, but it's not recommendable. Store the data in external table.

How to write single and multiple line commands in Hive?
To write single line commands we use –followed by commands.
eg: –It is too important step.
Hive doesn't supports multiple comments now.

What is Thrift server & client, JDBC and ODBC driver importance in Hive?
Thrift is a cross language RPC framework which generate code and cobines a software stack finally execute the Thrift code in remote server. Thrift compiler acts as interpreter between server and client. Thrift server allows a remove client to submit request to Hive, using different programming languages like Python, Ruby and scala.
JDBC driver: A JDBC driver is a software component enabling a Java application to interact with a database.
ODBC driver: ODBC accomplishes DBMS independence by using an ODBC driver as a translation layer between the application and the DBMS

Does Hive support 100% SQL Quries like Insert, Delete and Updates?
Hive doesn't support Updates in record level. To update, It integrate with Hbase.

When you are use Hive?
When the data is structured data, Static data, Low density is not a problem, If the data processed based on the queries, Hive is the best option. Most often data warehouse data processed in the Hive.

What is the use of partition in hive?
To analyze a particular set of data, not required to load entire data, desired data partition is a good approach. To achieve this goal, Hive allows to partition the data based on particular column. Static partition and Dynamic partition, both can optimize the Hive performance. For Instant, required a particular year information, partition based on year.

Is it mandatory to create Schema in Hive?
Yes, It's mandatory to create a table in Database. Hive is schema oriented modal. It store schema information in external database.

How Hive Serialize and DeSerialize the data?
In Hive language, SerDe also called Serialization and DeSerialization. Usually when read/write the data, user first communicate with inputformat, then it connect with Record reader to read/write record.The data is stored in Serialized (binary) format in Record. To serialize the data dat goes to row, here deserialized custem serde use object inspector to deserialize the data in fields. now user see the data in the fields, that deliver to the end user.

How Hive use Java in SerDe?
To insert data into table, Hive create an object by using Java. To transfer java objects over network, the data should be serialized. Each field serialized by using Object inspector and finally serialized data stored in Hive table.

Difference between order by and sort by in hive?
SORT BY -use number of reducers, so it can process quickly.
ORDER BY – use single reducer. If data is too large, it's take a long time to sort the data.

Does Hive Support Insert, delete, or updation?
As of now, Hive doesn't support record level updadation, insert and deletion queries. HQL is subset of SQL, but not equalto SQL. To update Hive integrate with Hbase.
Tell me few function names in Hive
CONTACT('Venu'-'Bigdata'-'analyst'); // Venu-Bigdata-analyst
CONTACT_WS('-', 'venu', 'bigdata', 'analyst'); //venu-bigdata-analyst
REPEAT('venu',3);
TRIM(' VENU '); //VENU (without spaces)
LTRim(' venu '); //venu (trim leftside, but not rightside)
RTRIM(' venu '); // venu(trim rightside only, but not leftside)
REVERSE('venu'); //unev
LOWER('Venu'); //venu
LCASE ""
UPPER OR UCASE('Venu'); //VENU
RLIKE .. return T/F for sub string.
'Venu' RLIKE 'en' //True
'Venu' RLIKE '^V.*' //T

Difference between Internal and External Table?
External table: Schema is stored in Database. Actual data stored in Hive tables. If data lost in External table, it lost only metastore, but not actual data.
Internal table: MetaStore and actual data both stored in local system. If any situation, data lost, both actual data and meta store will be lost.

What is the difference between Hive and Hbase?
Hive allows most of the SQL queries, but Hbase not allows SQL queries directly.
Hive doesn't support record level update, insert, and deletion operations on table, but Hbase can do it.
Hive is a Data warehouse framework where as Hbase is a NoSQL database.
Hive run on the top of Mapreduce, Hbase run on the top of HDFS.

How many ways you can run Hive?
In CLI mode (By using command line inerface).
By using JDBC or ODBC.
By Called Hive Thift client. It allows java, PHP, Python, Ruby and C++ to write commands to run in Hive.

Can you explain different type of SerDe?
By default Hive used Lazy Serde also allows Jeson Serde and most often used RegexSerde to be Serialized and DeSerialized Data.

Why we are using buckets in Hive?
To process many chunks of files, to analyze vast amount of data, sometime burst the process and time. Bucketing is a sampling concept to analyze the data, by using hashing algorithm. set hive.enforce.bucketing=true; can enable the process.

How Hive Organize the data?
Hive organize in three ways such as Tables, Partitions and Buckets. Tables organize based on Arrays, Maps, primitive column types. Partitions has one or more partition keys based on project requirements. Buckets used for analyze the data for sampling purpose. It's good approach to process a pinch of data in the form of buckets instead of process all data.

Can you explain about Hive Architecture?
There are 5 core components there in Hive such as: UI, Driver, Compiler, Metastore, Execute Engine.

What is User Interface (UI)?
UI: This interface is interpreter between users and Driver, which accept queries from User and execute on the Driver. Now two types of interfaces available in Hive such as command line interface and GUI interface. Hadoop provides Thrift interface and JDBC/ODBC for integrating other applications.

What is importance of Driver in Hive?
Driver: It manages life cycle of HiveQL queries. Driver receives the queries from User Interface and fetch on the ODBC/JDBC interfaces to process the query. Driver create separate independent section to handle each query.
Compiler: Compiler accept plans from Drivers and gets the required metadata from MetaStore, to execute Plan.
MetaStore: Hive Store meta data in the table. It means information about data is stored in MetaStore in the form of table, it may be internal or external table. Hive compiler get the meta data information from metastore table.
Execute Engine:
Hive Driver execute the output in the execution Engine. Here, execute engine executes the queries in the MapReduce JobTracker. Based on Required information, Hive queries run in the MapReduce to process the data.

When we are use explode in Hive?
Sometime Hadoop developer takes array as input and convert into a separate table row. To achieve this goal, Hive use explode, it acts as interpreter to convert complex data-types into desired table formats.
Syntax:
```
SELECT explode (arrayName) AS newCol FROM TableName;
SELECT explode(map) AS newCol1, NewCol2 From TableName;
```

What is ObjectInspector functionality in Hive?
Hive uses ObjectInspector to analyze the internal structure of the rows, columns and complex objects. Additionally gives us ways to access the internal fields inside the object. It not only process common data-types like int, bigint, STRING, but also process complex data-types like arrays, maps, structs and union.

Can you overwrite Hadoop Mapreduce configuration in Hive?
Yes, You can overwrite Hive map, reduce steps in hive conf settings. Hive allows to overwrite Hadoop configuration files.

How to display the present database name in the terminal?
There are two ways to know the current database. One temporary in cli and second one is persistently.
1) in CLI just enter this command: set hive.cli.print.current.db=true;
2) In hive-site.xml paste this code:
```
<property>
<name>hive.cli.print.current.db</name>
<value>true</value>
</property>
```
In second scenario, you can automatically display the Hive database name when you open terminal.

Is a job split into map?
No, Hadoop framework can split the data-file, but not Job. This chunks of data stored in blocks. Each split need a map to process. Where as Job is a configurable unit to control execution of the plan/logic. Job is not a physical data-set to split, it’s a logical configuration API to process those split.

What is the difference between Describe and describe extended?
To see table definition in Hive, use describe <table name>; command
Where as To see more detailed information about the table, use describe extended <tablename>; command
Another important command describe formatted <tablename>; also describe all details in a clean manner.

What is Hive?
Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems.
Hive was originally developed at Facebook. It's now a Hadoop subproject with many contributors. Users need to concentrate only on the top level hive language rather than java map reduce programs. One of the main advantages of Hive is its SQLish nature. Thus it leverages the usability to a higher extend.

A hive program will be automatically compiled into map-reduce jobs executed on Hadoop. In addition, HiveQL supports custom map-reduce scripts to be plugged into queries.

Hive example:
selecting the employee names whose salary more than 100 dollars from a hive table called tbl_employee.
SELECT employee_name FROM tbl_employee WHERE salary > 100;
Users are excited to use Hive since it is very similar to SQL.

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.

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" 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.

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.

Is multiline comment supported in Hive Script ?
No.

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.

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.

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.

Give examples of the SerDe classes whihc hive uses to Serializa 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).

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.

What is ObjectInspector functionality ?

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

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.

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.
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.

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.

How can the columns of a table in hive be written to a file?
By using awk command in shell, the output from HiveQL (Describe) can be written to a file.
hive -S -e "describe table_name;" | awk -F" " '{print 1}' > ~/output.

CONCAT function in Hive with Example?
CONCAT function will concat the input strings. You can specify any number of strings separated by comma.

Example:
CONCAT ('Hive','-','performs','-','good','-','in','-','Hadoop');
Output:
Hive-performs-good-in-Hadoop
So, every time you delimit the strings by '-'. If it is common for all the strings, then Hive provides another command CONCAT_WS. Here you have to specify the delimit operator first.
CONCAT_WS ('-','Hive','performs','good','in','Hadoop');
Output: Hive-performs-good-in-Hadoop

REPEAT function in Hive with example?
REPEAT function will repeat the input string n times specified in the command.
Example:
REPEAT('Hadoop',3);
Output:
HadoopHadoopHadoop.
Note: You can add a space with the input string also.

TRIM function in Hive with example?
TRIM function will remove the spaces associated with a string.
Example:
TRIM('Hadoop');
Output:
Hadoop.

Note: If you want to remove only leading or trialing spaces then you can specify the below commands respectively.
LTRIM('Hadoop');
RTRIM('Hadoop');
REVERSE function in Hive with example?
REVERSE function will reverse the characters in a string.
Example:
REVERSE('Hadoop');
Output:
poodaH

LOWER or LCASE function in Hive with example?
LOWER or LCASE function will convert the input string to lower case characters.
Example:
LOWER('Hadoop');
LCASE('Hadoop');
Output:
hadoop

What is the Hive configuration precedence order?
There is a precedence hierarchy to setting properties. In the following list, lower numbers take precedence over higher numbers:

1. The Hive **SET** command
2. The command line **-hiveconf** option
3. hive-site.xml
4. hive-default.xml
5. hadoop-site.xml (or, equivalently, core-site.xml, hdfs-site.xml, and mapred-site.xml)
6. hadoop-default.xml (or, equivalently, core-default.xml, hdfs-default.xml, and mapred-default.xml)

How do change settings within Hive Session?
We can change settings from within a session, too, using the SET command. This is useful for changing Hive or MapReduce job settings for a particular query. For example, the following command ensures buckets are populated according to the table definition.

```
hive> SET hive.enforce.bucketing=true;
```
To see the current value of any property, use SET with just the property name:
```
hive> SET hive.enforce.bucketing;
hive.enforce.bucketing=true
```

By itself, **SET** will list all the properties and their values set by Hive. This list will not include Hadoop defaults, unless they have been explicitly overridden in one of the ways covered in the above answer. Use **SET -v** to list all the properties in the system, including Hadoop defaults.

How to print header on Hive query results?
We need to use following set command before our query to show column headers in STDOUT.
```
hive> set hive.cli.print.header=true;
```

How to get detailed description of a table in Hive?
Use below hive command to get a detailed description of a hive table.
```
hive> describe extended <tablename>;
```

How to access sub directories recursively in Hive queries?
To process directories recursively in Hive, we need to set below two commands in hive session. These two parameters work in conjunction.

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

Now hive tables can be pointed to the higher level directory. This is suitable for a scenario where the directory structure is as following: */data/country/state/city*

How to skip header rows from a table in Hive?
Suppose while processing some log files, we may find header records.
System=….
Version=…
Sub-version=….
Like above, It may have 3 lines of headers that we do not want to include in our Hive query. To skip header lines from our tables in Hive we can set a table property that will allow us to skip the header lines.

```
CREATE EXTERNAL TABLE userdata (
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"="3”);
```

Is it possible to create multiple table in hive for same data?
As hive creates schema and append on top of an existing data file. One can have multiple schema for one data file, schema will be saved in hive’s metastore and data will not be parsed or serialized to disk in given schema. When we will try to retrieve data, schema will be used. For example if we have 5 column (name, job, dob, id, salary) in the data file present in hive metastore then, we can have multiple schema by choosing any number of columns from the above list. (Table with 3 columns or 5 columns or 6 columns).
But while querying, if we specify any column other than above list, will result in NULL values.

What is the maximum size of string data type supported by Hive?
Maximum size is 2 GB.

What are the Binary Storage formats supported in Hive?
By default Hive supports text file format, however hive also supports below binary formats. Sequence Files, Avro Data files, RCFiles, ORC files, Parquet files **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 and content could stored as value. it support compression which enables huge gain in performance.
Avro datafiles: Same as Sequence file splittable, compressible and row oriented except support of schema evolution and multilingual binding support.
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.
ORC Files: Optimized Record Columnar files

Is HQL case sensitive?
HQL is not case sensitive.

Describe CONCAT function in Hive with Example?
CONCAT function will concatenate the input strings. We can specify any number of strings separated by comma.
Example: CONCAT (‘Hive’,’-‘,’is’,’-‘,’a’,’-‘,’data warehouse’,’-‘,’in Hadoop’);
Output: Hive-is-a-data warehouse-in Hadoop
So, every time we delimit the strings by ‘-‘. If it is common for all the strings, then Hive provides another command CONCAT_WS. Here you have to specify the delimit operator first.
Syntax: CONCAT_WS (‘-‘,’Hive’,’is’,’a’,’data warehouse’,’in Hadoop’);
Output: Hive-is-a-data warehouse-in Hadoop

Describe REPEAT function in Hive with example?

REPEAT function will repeat the input string n times specified in the command.
Example: REPEAT('Hive',3);
Output: HiveHiveHive.

Describe REVERSE function in Hive with example?
REVERSE function will reverse the characters in a string.

Example: REVERSE('Hive');
Output: eviH

Describe TRIM function in Hive with example?
TRIM function will remove the spaces associated with a string.
Example: TRIM(' Hadoop ');
Output: Hadoop.
If we want to remove only leading or trailing spaces then we can specify the below commands respectively.
LTRIM(' Hadoop');
RTRIM('Hadoop ');

Describe RLIKE in Hive with an example?
RLIKE (Right-Like) is a special function in Hive where if any substring of A matches with B then it evaluates to true. It also obeys Java regular expression pattern. Users don't need to put % symbol for a simple match in RLIKE.
Examples: 'Express' RLIKE 'Exp' –> True
'Express' RLIKE '^E.*' –> True (Regular expression)
Moreover, RLIKE will come handy when the string has some spaces. Without usingTRIM function, RLIKE satisfies the required scenario. Suppose if A has value 'Express '(2 spaces additionally) and B has value 'Express'.In these situations, RLIKE will work better without using TRIM.
'Express ' RLIKE 'Express' –> True
Note: RLIKE evaluates to NULL if A or B is NULL.

What is Hadoop Hive?
Hadoop Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems.
Hive was originally developed at Facebook. It's now a Hadoop subproject with many contributors. Users need to concentrate only on the top level hive language rather than java map reduce programs. One of the main advantages of Hive is its SQLish nature. Thus it leverages the usability to a higher extend.
A hive program will be automatically compiled into map-reduce jobs executed on Hadoop. In addition, HiveQL supports custom map-reduce scripts to be plugged into queries.

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.

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.

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.

Give examples of the SerDe classes whihc hive uses to Serializa 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).

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.

What is ObjectInspector functionality ?
Hive uses ObjectInspector to analyze the internal structure of the row object and also the structure of the individual columns.

ObjectInspector 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.

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

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

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.

What kind of datawarehouse 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.

How can the columns of a table in hive be written to a file?

By using awk command in shell, the output from HiveQL (Describe) can be written to a file.
hive -S -e "describe table_name;" | awk -F" " '{print 1}' > ~/output.

CONCAT function in Hive with Example?

CONCAT function will concat the input strings. You can specify any number of strings separated by comma.

Example:
CONCAT ('Hive','-','performs','-','good','-','in','-','Hadoop');

Output:
Hive-performs-good-in-Hadoop

So, every time you delimit the strings by '-'. If it is common for all the strings, then Hive provides another command CONCAT_WS. Here you have to specify the delimit operator first.

CONCAT_WS ('-','Hive','performs','good','in','Hadoop');
Output: Hive-performs-good-in-Hadoop

REPEAT function in Hive with example?
REPEAT function will repeat the input string n times specified in the command.
Example:
REPEAT('Hadoop',3);
Output:
HadoopHadoopHadoop.
Note: You can add a space with the input string also.

TRIM function in Hive with example?
TRIM function will remove the spaces associated with a string.
Example:
TRIM(' Hadoop ');
Output:
Hadoop.
Note: If you want to remove only leading or trialing spaces then you can specify the below commands respectively.
LTRIM(' Hadoop');
RTRIM('Hadoop ');

REVERSE function in Hive with example?
REVERSE function will reverse the characters in a string.
Example:
REVERSE(‘Hadoop’);
Output:
poodaH

LOWER or LCASE function in Hive with example?

LOWER or LCASE function will convert the input string to lower case characters.
Example:
LOWER(‘Hadoop’);
LCASE(‘Hadoop’);
Output:
hadoop
Note:
If the characters are already in lower case then they will be preserved.

UPPER or UCASE function in Hive with example?
UPPER or UCASE function will convert the input string to upper case characters.
Example:
UPPER(‘Hadoop’);
UCASE(‘Hadoop’);
Output:
HADOOP
Note:
If the characters are already in upper case then they will be preserved.

Double type in Hive – Important points?
It is important to know about the double type in Hive. Double type in Hive will present the data differently unlike RDBMS.
See the double type data below:
24624.0
32556.0
3.99893E5
4366.0

E5 represents 10^5 here. So, the value 3.99893E5 represents 399893. All the calculations will be accurately performed using double type. The maximum value for a IEEE 754 double is about 2.22E308.
It is crucial while exporting the double type data to any RDBMS since the type may be wrongly interpreted. So, it is advised to cast the double type into appropriate type before exporting.

Rename a table in Hive – How to do it?
Using ALTER command, we can rename a table in Hive.
ALTER TABLE hive_table_name RENAME TO new_name;

There is another way to rename a table in Hive. Sometimes, ALTER may take more time if the underlying table has more partitions/functions. In that case, Import and export options can be utilized. Here you are saving the hive data into HDFS and importing back to new table like below.
EXPORT TABLE tbl_name TO ‘HDFS_location’;
IMPORT TABLE new_tbl_name FROM ‘HDFS_location’;

If you prefer to just preserve the data, you can create a new table from old table like below.
CREATE TABLE new_tbl_name AS SELECT * FROM old_tbl_name;
DROP TABLE old_tbl_name;

How to change a column data type in Hive?
ALTER TABLE table_name CHANGE column_name column_name new_datatype;
Example: If you want to change the data type of ID column from integer to bigint in a table called employee.
ALTER TABLE employee CHANGE id id BIGINT;

Difference between order by and sort by in hive?
SORT BY will sort the data within each reducer. You can use any number of reducers for SORT BY operation.
ORDER BY will sort all of the data together, which has to pass through one reducer. Thus, ORDER BY in hive uses single reducer.
ORDER BY guarantees total order in the output while SORT BY only guarantees ordering of the rows within a reducer. If there is more than one reducer, SORT BY may give partially ordered final results

RLIKE in Hive?
RLIKE (Right-Like) is a special function in Hive where if any substring of A matches with B then it evaluates to true. It also obeys Java regular expression pattern. Users don’t need to put % symbol for a simple match in RLIKE.
Examples:
‘Express’ RLIKE ‘Exp’ –> True
‘Express’ RLIKE ‘^E.*’ –> True (Regular expression)
Moreover, RLIKE will come handy when the string has some spaces. Without using TRIM function, RLIKE satisfies the required scenario. Suppose if A has value ‘Express ‘ (2 spaces additionally) and B has value ‘Express’ RLIKE will work better without using TRIM.
‘Express’ RLIKE ‘Express’ –> True
Note:
RLIKE evaluates to NULL if A or B is NULL.

Difference between external table and internal table in HIVE ?
Hive has a relational database on the master node it uses to keep track of state. For instance, when you CREATE TABLE FOO(foo string) LOCATION ‘hdfs://tmp/’;, this table schema is stored in the database. If you have a partitioned table, the partitions are stored in the database(this allows hive to use lists of partitions without going to the filesystem and finding them, etc). These sorts of things are the ‘metadata’.
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.