Apache Hive

Apache Hive data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.

1) What is the definition of Hive? What is the present version of Hive and explain about ACID transactions in Hive?

Answer) Hive is an open source data warehouse system. We can use Hive for analyzing and querying in large data sets of Hadoop files. Its similar to SQL. Hive supports ACID transactions: The full form of ACID is Atomicity, Consistency, Isolation, and Durability. ACID transactions are provided at the row levels, there are Insert, Delete, and Update options so that Hive supports ACID transaction. Insert
Delete
Update

2)Explain what is a Hive variable. What do we use it for?

Answer)Hive variable is basically created in the Hive environment that is referenced by Hive scripting languages. It provides to pass some values to the hive queries when the query starts executing. It uses the source command.

3)What kind of data warehouse application is suitable for Hive? What are the types of tables in Hive?

Answer) Hive is not considered as a full database. The design rules and regulations of Hadoop and HDFS put restrictions on what Hive can do.Hive is most suitable for data warehouse applications.
Where
Analyzing the relatively static data.
Less Responsive time.
No rapid changes in data.Hive doesnot provide fundamental features required for OLTP, Online Transaction Processing.Hive is suitable for data warehouse applications in large data sets. Two types of tables in Hive
Managed table.
External table.

4)How to change the warehouse.dir location for older tables?

www.smartdatacamp.com

Answer)To change the base location of the Hive tables, edit the hive.metastore.warehouse.dir param. This will not affect the older tables. Metadata needs to be changed in the database (MySQL or Derby). The location of Hive tables is in table SDS and column LOCATION.

5)What is Hive Metastore?

Answer)Hive metastore is a database that stores metadata about your Hive tables (eg. tablename, 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.

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

Answer)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.

7)Is it possible to use same metastore by multiple users, in case of embedded hive?

Answer)No, it is not possible to use metastore in sharing mode. It is recommended to use standalone real database like MySQL or PostGresSQL.

8)Is multiline comment supported in Hive Script ?

Answer) No.

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

Answer)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.

10)What is SerDe in Apache Hive ?

Answer)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 format that data is stored in. Users are able to write files to HDFS with whatever tools or 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.

11)Which classes are used by the Hive to Read and Write HDFS Files

Answer) Following classes are used by Hive to read and write HDFS files

TextInputFormat or HiveIgnoreKeyTextOutputFormat: These 2 classes read/write data in plain text file format.

SequenceFileInputFormat or SequenceFileOutputFormat: These 2 classes read/write data in hadoop SequenceFile format.

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

Answer)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 or write thrift serialized objects. The class file for the Thrift object must be loaded first.

DynamicSerDe: This SerDe also read or 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).

13)How do you write your own custom SerDe ?

Answer)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

www.smartdatacamp.com

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 its non-trivial to write a thrift DDL parser.

14)What is ObjectInspector functionality ?

Answer)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.

15)What is the functionality of Query Processor in Apache Hive ?

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

16)What is the limitation of Derby database for Hive metastore?

Answer)With derby database, you cannot have multiple connections or multiple sessions instantiated at the same time. Derby database runs in the local mode and it creates a log file so that multiple users cannot access Hive simultaneously.

17)What are managed and external tables?

Answer)We have got two things, one of which is data present in the HDFS and the other is the metadata, present in some database.

There are two categories of Hive tables that is Managed and External Tables.

In the Managed tables, both the data and the metadata are managed by Hive and if you drop the managed table, both data and metadata are deleted.

There are some situations where your data will be controlled by some other application and you want to read that data but you must allow Hive to delete that data. In such case, you can create an external table in Hive. In the external table, metadata is controlled by Hive but the actual data will be controlled by some other application. So, when you delete a table accidentally, only the metadata will be lost and the actual data will reside wherever it is.

18)What are the complex data types in Hive?

Answer)MAP: The Map contains a key-value pair where you can search for a value using the key.

STRUCT:A Struct is a collection of elements of different data types. For example, if you take the address, it can have different data types. For example, pin code will be in Integer format.

ARRAY:An Array will have a collection of homogeneous elements. For example, if you take your skillset, you can have N number of skills

UNIONTYPE:It represents a column which can have a value that can belong to any of the data types of your choice.

19)How does partitioning help in the faster execution of queries?

Answer)With the help of partitioning, a subdirectory will be created with the name of the partitioned column and when you perform a query using the WHERE clause, only the particular sub-directory will be scanned instead of scanning the whole table. This gives you faster execution of queries.

20)How to enable dynamic partitioning in Hive?

Answer)Related to partitioning there are two types of partitioning Static and Dynamic. In the static partitioning, you will specify the partition column while loading the data.

Whereas in dynamic partitioning, you push the data into Hive and then Hive decides which value should go into which partition. To enable dynamic partitioning, you have set the below property

set hive.exec.dynamic.parition.mode = nonstrict;

Example: insert overwrite table emp_details_partitioned partition(location)

select * from emp_details;

21)How does bucketing help in the faster execution of queries?

www.smartdatacamp.com

Answer)If you have to join two large tables, you can go for reduce side join. But if both the tables have the same number of buckets or same multiples of buckets and also sorted on the same column there is a possibility of SMBMJ in which all the joins take place in the map phase itself by matching the corresponding buckets. Buckets are basically files that are created inside the HDFS directory.

There are different properties which you need to set for bucket map joins and they are as follows:
set hive.enforce.sortmergebucketmapjoin = false;
set hive.auto.convert.sortmerge.join = false;
set hive.optimize.bucketmapjoin = ture;
set hive.optimize.bucketmapjoin.sortedmerge = true;

22)How to enable bucketing in Hive?

Answer)By default bucketing is disabled in Hive, you can enforce to enable it by setting the below property

set hive.enforce.bucketing = true;

23)Which method has to be overridden when we use custom UDF in Hive?

Answer)Whenever you write a custom UDF in Hive, you have to extend the UDF class and you have to override the evaluate() function.

24)What are the different file formats in Hive?

Answer)There are different file formats supported by Hive
Text File format
Sequence File format
RC file format
Parquet
Avro
ORC

Every file format has its own characteristics and Hive allows you to choose easily the file format which you wanted to use.

25)How is SerDe different from File format in Hive?

Answer)SerDe stands for Serializer and Deserializer. It determines how to encode and decode the field values or the column values from a record that is how you serialize and deserialize the values of a column

But file format determines how records are stored in key value format or how do you retrieve the records from the table

26)What is RegexSerDe?

Answer)Regex stands for a regular expression. Whenever you want to have a kind of pattern matching, based on the pattern matching, you have to store the fields. RegexSerDe is present in org.apache.hadoop.hive.contrib.serde2.RegexSerDe.

In the SerDeproperties, you have to define your input pattern and output fields. For example, you have to get the column values from line xyz/pq@def if you want to take xyz, pq and def separately.

To extract the pattern, you can use:

input.regex = (.*)/(.*)@(.*)

To specify how to store them, you can use

output.format.string = %1$s%2$s%3$s;

27)How is ORC file format optimised for data storage and analysis?

Answer)ORC stores collections of rows in one file and within the collection the row data will be stored in a columnar format. With columnar format, it is very easy to compress, thus reducing a lot of storage cost.

While querying also, it queries the particular column instead of querying the whole row as the records are stored in columnar format.

ORC has got indexing on every block based on the statistics min, max, sum, count on columns so when you query, it will skip the blocks based on the indexing.

28)How to access HBase tables from Hive?

Answer)Using Hive-HBase storage handler, you can access the HBase tables from Hive and once you are connected, you can query HBase using the SQL queries from Hive. You can also join multiple tables in HBase from Hive and retrieve the result.

29)When running a JOIN query, I see out-of-memory errors.?

www.smartdatacamp.com

Answer)This is usually caused by the order of JOIN tables. Instead of [FROM tableA a JOIN tableB b ON], try [FROM tableB b JOIN tableA a ON] NOTE that if you are using LEFT OUTER JOIN, you might want to change to RIGHT OUTER JOIN. This trick usually solve the problem the rule of thumb is, always put the table with a lot of rows having the same value in the join key on the rightmost side of the JOIN.

30)Did you used Mysql as Metatstore and faced errors like com.mysql.jdbc.exceptions.jdbc4. CommunicationsException: Communications link failure ?

Answer)This is usually caused by MySQL servers closing connections after the connection is idling for some time. Run the following command on the MySQL server will solve the problem [set global wait_status=120]

When using MySQL as a metastore I see the error [com.mysql.jdbc.exceptions.MySQLSyntaxErrorException: Specified key was too long; max key length is 767 bytes].

This is a known limitation of MySQL 5.0 and UTF8 databases. One option is to use another character set, such as latin1, which is known to work.

31)Does Hive support Unicode?

Answer)You can use Unicode string on data or comments, but cannot use for database or table or column name.

You can use UTF-8 encoding for Hive data. However, other encodings are not supported (HIVE 7142 introduce encoding for LazySimpleSerDe, however, the implementation is not complete and not address all cases).

32)Are Hive SQL identifiers (e.g. table names, column names, etc) case sensitive?

Answer)No. Hive is case insensitive.

33)What is the best way to load xml data into hive

Answer)The easiest way is to use the Hive XML SerDe (com.ibm.spss.hive.serde2.xml.XmlSerDe), which will allow you to directly import and work with XML data.

34)When Hive is not suitable?

Answer)It does not provide OLTP transactions support only OLAP transactions.If application required OLAP, switch to NoSQL database.HQL queries have higher latency, due to the mapreduce.

35)Mention what are the different modes of Hive?

Answer)Depending on the size of data nodes in Hadoop, Hive can operate in two modes.

These modes are,Local mode and Map reduce mode

36)Mention what is ObjectInspector functionality in Hive?

Answer)ObjectInspector functionality in Hive is used to analyze the internal structure of the columns, rows, and complex objects. It allows to access the internal fields inside the objects.

37)Mention what is (HS2) HiveServer2?

Answer)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

38)Mention what Hive query processor does?

Answer)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.

39)Mention what are the components of a Hive query processor?

Answer)The components of a Hive query processor include,
Logical Plan Generation
Physical Plan Generation
Execution Engine
Operators
UDFs and UDAFs

www.smartdatacamp.com

Optimizer
Parser
Semantic Analyzer
Type Checking

40)Mention if we can name view same as the name of a Hive table?

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

41)Explain how can you change a column data type in Hive?

Answer)You can change a column data type in Hive by using command,

ALTER TABLE table_name CHANGE column_name column_name new_datatype;

42)Mention what is the difference between order by and sort by in Hive?

Answer)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 a single

43)Explain when to use explode in Hive?

Answer)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.

44)Mention how can you stop a partition form being queried?

Answer)You can stop a partition form being queried by using the ENABLE OFFLINE clause with ALTER TABLE statement.

45)What is the need for custom Serde?

Answer)Depending on the nature of data the user has, the inbuilt SerDe may not satisfy the format of the data. SO users need to write their own java code to satisfy their data format requirements.

46)What is the default location where hive stores table data?

Answer)hdfs://namenode_server/user/hive/warehouse

47)Is there a date data type in Hive?

Answer)Yes. The TIMESTAMP data types stores date in java.sql.timestamp format

48)Can we run unix shell commands from hive? Give example?

Answer)Yes, using the ! mark just before the command.For example !pwd at hive prompt will list the current directory.

49)Can hive queries be executed from script files? How?

Answer)Using the source command.
Example
Hive> source /path/to/file/file_with_query.hql

50)What is the importance of .hiverc file?

Answer)It is a file containing list of commands needs to run when the hive CLI starts. For example setting the strict mode to be true etc.

51)What are the default record and field delimiter used for hive text files?

Answer)The default record delimiter is −

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

52)What do you mean by schema on read?

Answer)The schema is validated with the data when reading the data and not enforced when writing data.

www.smartdatacamp.com

53)How do you list all databases whose name starts with p?

Answer)SHOW DATABASES LIKE p.*

54)What does the USE command in hive do?

Answer)With the use command you fix the database on which all the subsequent hive queries will run.

55)How can you delete the DBPROPERTY in Hive?

Answer)There is no way you can delete the DBPROPERTY.

56)What is the significance of the line?

Answer)set hive.mapred.mode = strict;

57)How do you check if a particular partition exists?

Answer)This can be done with following query

SHOW PARTITIONS table_name PARTITION(partitioned_column=partition_value)

58)Which java class handles the Input record encoding into files which store the tables in Hive?

Answer)org.apache.hadoop.mapred.TextInputFormat

59)Which java class handles the output record encoding into files which result from Hive queries?

Answer)org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

60)What is the significance of IF EXISTS clause while dropping a table?

Answer)When we issue the command DROP TABLE IF EXISTS table_name

Hive throws an error if the table being dropped does not exist in the first place.

61)When you point a partition of a hive table to a new directory, what happens to the data?

Answer)The data stays in the old location. It has to be moved manually.

62)Write a query to insert a new column(new_col INT) into a hive table (htab) at a position before an existing column (x_col)

Answer)ALTER TABLE table_name
CHANGE COLUMN new_col INT
BEFORE x_col

63)Does the archiving of Hive tables give any space saving in HDFS?

Answer)No. It only reduces the number of files which becomes easier for namenode to manage.

64)While loading data into a hive table using the LOAD DATA clause, how do you specify it is a hdfs file and not a local file ?

Answer)By Omitting the LOCAL CLAUSE in the LOAD DATA statement.

65)If you omit the OVERWRITE clause while creating a hive table,what happens to file which are new and files which already exist?

Answer)The new incoming files are just added to the target directory and the existing files are simply overwritten. Other files whose name does not match any of the incoming files will continue to exist.

If you add the OVERWRITE clause then all the existing data in the directory will be deleted before new data is written.

66)What does the following query do?

INSERT OVERWRITE TABLE employees

PARTITION (country, state)

www.smartdatacamp.com

SELECT ..., se.cnty, se.st

FROM staged_employees se

Answer)It creates partition on table employees with partition values coming from the columns in the select clause. It is called Dynamic partition insert.

67)What is a Table generating Function on hive?

Answer)A table generating function is a function which takes a single column as argument and expands it to multiple column or rows. Example explode()

68)How can Hive avoid mapreduce?

Answer)If we set the property hive.exec.mode.local.auto to true then hive will avoid mapreduce to fetch query results.

69)What is the difference between LIKE and RLIKE operators in Hive?

Answer)The LIKE operator behaves the same way as the regular SQL operators used in select queries.

Example

street_name like %Chi

But the RLIKE operator uses more advance regular expressions which are available in java

Example

street_name RLIKE .*(Chi|Oho).* which will select any word which has either chi or oho in it.

70)Is it possible to create Cartesian join between 2 tables, using Hive?

Answer)No. As this kind of Join can not be implemented in mapreduce

71)As part of Optimizing the queries in Hive, what should be the order of table size in a join query?

Answer)In a join query the smallest table to be taken in the first position and largest table should be taken in the last position.

72)What is the usefulness of the DISTRIBUTED BY clause in Hive?

Answer)It controls how the map output is reduced among the reducers. It is useful in case of streaming data

73)How will you convert the string 51.2 to a float value in the price column?

Answer)Select cast(price as FLOAT)

74)What will be the result when you do cast(abc as INT)?

Answer)Hive will return NULL

75)Can we LOAD data into a view?

Answer)No. A view can not be the target of a INSERT or LOAD statement.

76)What types of costs are associated in creating index on hive tables?

Answer)Indexes occupies space and there is a processing cost in arranging the values of the column on which index is cerated.

77)Give the command to see the indexes on a table.

Answer)SHOW INDEX ON table_name

This will list all the indexes created on any of the columns in the table table_name.

78)What does /*streamtable(table_name)*/ do?

Answer)It is query hint to stream a table into memory before running the query. It is a query optimization Technique.

79)The following statement failed to execute. What can be the cause?

www.smartdatacamp.com

LOAD DATA LOCAL INPATH ${env:HOME}/country/state/
OVERWRITE INTO TABLE address;

Answer)The local inpath should contain a file and not a directory. The $env:HOME is a valid variable available in the hive environment

80)How do you specify the table creator name when creating a table in Hive?

Answer)The TBLPROPERTIES clause is used to add the creator name while creating a table.
The TBLPROPERTIES is added like
TBLPROPERTIES(creator = Joan)

81)Suppose I have installed Apache Hive on top of my Hadoop cluster using default metastore configuration. Then, what will happen if we have multiple clients trying to access Hive at the same time?

Answer)The default metastore configuration allows only one Hive session to be opened at a time for accessing the metastore. Therefore, if multiple clients try to access the metastore at the same time, they will get an error. One has to use a standalone metastore, i.e. Local or remote metastore configuration in Apache Hive for allowing access to multiple clients concurrently.

Following are the steps to configure MySQL database as the local metastore in Apache Hive:

One should make the following changes in hive-site.xml:
javax.jdo.option.ConnectionURL property should be set to
jdbc:mysql://host/dbname?createDataba
seIfNotExist=true.
javax.jdo.option.ConnectionDriverName property should be set to com.mysql.jdbc.Driver.
One should also set the username and password as:
javax.jdo.option.ConnectionUserName is set to desired username.
javax.jdo.option.ConnectionPassword is set to the desired password.

The JDBC driver JAR file for MySQL must be on the Hive classpath, i.e. The jar file should be copied into the Hive lib directory.

Now, after restarting the Hive shell, it will automatically connect to the MySQL database which is running as a standalone metastore.

82)Is it possible to change the default location of a managed table?

Answer)Yes, it is possible to change the default location of a managed table. It can be achieved by using the clause LOCATION [hdfs_path].

83)When should we use SORT BY instead of ORDER BY?

Answer)We should use SORT BY instead of ORDER BY when we have to sort huge datasets because SORT BY clause sorts the data using multiple reducers whereas ORDER BY sorts all of the data together using a single reducer. Therefore, using ORDER BY against a large number of inputs will take a lot of time to execute.

84)What is dynamic partitioning and when is it used?

Answer)In dynamic partitioning values for partition columns are known in the runtime, i.e. It is known during loading of the data into a Hive table.

One may use dynamic partition in following two cases:

Loading data from an existing non-partitioned table to improve the sampling and therefore, decrease the query latency.

When one does not know all the values of the partitions before hand and therefore, finding these partition values manually from a huge data sets is a tedious task.

85)Suppose, I create a table that contains details of all the transactions done by the customers of year 2016: CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY , ;

Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?

Answer)We can solve this problem of query latency by partitioning the table according to each month. So, for each month we will be scanning only the partitioned data instead of whole data sets.

As we know, we can not partition an existing non-partitioned table directly. So, we will be taking following steps to solve the very problem:

Create a partitioned table, say partitioned_transaction:

CREATE TABLE partitioned_transaction (cust_id INT, amount FLOAT, country STRING) PARTITIONED BY (month STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY , ;

2. Enable dynamic partitioning in Hive:
SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;

3. Transfer the data from the non – partitioned table into the newly created partitioned table:

INSERT OVERWRITE TABLE partitioned_transaction PARTITION (month) SELECT cust_id, amount, country, month FROM transaction_details;

www.smartdatacamp.com

Now, we can perform the query using each partition and therefore, decrease the query time.

86)How can you add a new partition for the month December in the above partitioned table?

Answer)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;

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

Answer)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

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

Answer)To remove this error one has to execute following commands:
SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;

89)Suppose, I have a CSV file sample.csv present in temp directory with the following entries:
id first_name last_name email gender ip_address
1 Hugh Jackman hughjackman@cam.ac.uk Male 136.90.241.52
2 David Lawrence dlawrence1@gmail.com Male 101.177.15.130
3 Andy Hall andyhall2@yahoo.com Female 114.123.153.64
4 Samuel Jackson samjackson231@sun.com Male 89.60.227.31
5 Emily Rose rose.emily4@surveymonkey.com Female 119.92.21.19

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

Answer)SerDe stands for serializer or deserializer. A SerDe allows us to convert the unstructured bytes into a record that we can process using Hive. SerDes are implemented using Java. Hive comes with several built-in SerDes and many other third-party SerDes are also available.

Hive provides a specific SerDe for working with CSV files. We can use this SerDe for the sample.csv by issuing following commands:
CREATE EXTERNAL TABLE sample
(id int, first_name string,
last_name string, email string,
gender string, ip_address string)
ROW FORMAT SERDE org.apache.hadoop.hive.serde2.OpenCSVSerde
STORED AS TEXTFILE LOCATION temp;

Now, we can perform any query on the table sample:

SELECT first_name FROM sample WHERE gender = male;

90)Suppose, I have a lot of small CSV files present in input directory in HDFS and I want to create a single Hive table corresponding to these files. The data in these files are in the format: {id, name, e-mail, country}. Now, as we know, Hadoop performance degrades when we use lots of small files.

So, how will you solve this problem where we want to create a single Hive table for lots of small files without degrading the performance of the system?

Answer)One can use the SequenceFile format which will group these small files together to form a single sequence file. The steps that will be followed in doing so are as follows:
Create a temporary table:
CREATE TABLE temp_table (id INT, name STRING, e-mail STRING, country STRING)
ROW FORMAT FIELDS DELIMITED TERMINATED BY , STORED AS TEXTFILE;
Load the data into temp_table:
LOAD DATA INPATH input INTO TABLE temp_table;
Create a table that will store data in SequenceFile format:
CREATE TABLE sample_seqfile (id INT, name STRING, e-mail STRING, country STRING)
ROW FORMAT FIELDS DELIMITED TERMINATED BY , STORED AS SEQUENCEFILE;
Transfer the data from the temporary table into the sample_seqfile table:
INSERT OVERWRITE TABLE sample SELECT * FROM temp_table;
Hence, a single SequenceFile is generated which contains the data present in all of the input files and therefore, the problem of having lots of small files is finally eliminated.

91)Can We Change settings within Hive Session? If Yes, How?

Answer)Yes we can change the settings within Hive session, using the SET command. It helps to change Hive job settings for an exact query.

Example: The following commands shows buckets are occupied according to the table definition.

hive> SET hive.enforce.bucketing=true;

We can see the current value of any property by using SET with the property name. SET will list all the properties with their values set by Hive.

www.smartdatacamp.com

hive> SET hive.enforce.bucketing;
hive.enforce.bucketing=true

And this list will not include defaults of Hadoop. So we should use the below like

SET -v

It will list all the properties including the Hadoop defaults in the system.

92)Is it possible to add 100 nodes when we have 100 nodes already in Hive? How?

Answer)Yes, we can add the nodes by following the below steps.
Take a new system create a new username and password.
Install the SSH and with master node setup ssh connections.
Add ssh public_rsa id key to the authorized keys file.
Add the new data node host name, IP address and other details in /etc/hosts slaves file
192.168.1.102 slave3.in slave3.
Start the Data Node on New Node.
Login to the new node like suhadoop or ssh -X hadoop@192.168.1.103.
Start HDFS of a newly added slave node by using the following command
./bin/hadoop-daemon.sh start data node.
Check the output of jps command on a new node

93)Explain the concatenation function in Hive with an example?

Answer)Concatenate function will join the input strings.We can specify the N number of strings separated by a comma.
Example:
CONCAT (It,-,is,-,a,-,eLearning,-,provider);
Output:
It-is-a-eLearning-provider
So, every time we set the limits of the strings by -. If it is common for every strings, then Hive provides another command
CONCAT_WS. In this case,we have to specify the set limits of operator first.
CONCAT_WS (-,It,is,a,eLearning,provider);
Output: It-is-a-eLearning-provider.

94)Explain Trim and Reverse function in Hive with examples?

Answer)Trim function will delete the spaces associated with a string.
Example:
TRIM(BHAVESH);
Output:
BHAVESH
To remove the Leading space

LTRIM(BHAVESH);
To remove the trailing space
RTRIM(BHAVESH);
In Reverse function, characters are reversed in the string.
Example:
REVERSE(BHAVESH);
Output:
HSEVAHB

95)How to change the column data type in Hive? Explain RLIKE in Hive?

Answer)We can change the column data type by using ALTER and CHANGE.

The syntax is :

ALTER TABLE table_name CHANGE column_namecolumn_namenew_datatype;

Example: If we want to change the data type of the salary column from integer to bigint in the employee table.

ALTER TABLE employee CHANGE salary salary BIGINT;RLIKE: Its full form is Right-Like and it is a special function in the Hive. It helps to examine the two substrings. i.e, if the substring of A matches with B then it evaluates to true.

Example:
Bhavesh RLIKE ave True
Bhavesh RLIKE ^B.* True (this is a regular expression)

96)Explain process to access sub directories recursively in Hive queries?

Answer)By using below commands we can access sub directories 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 which is like /data/country/state/city/

97)How to skip header rows from a table in Hive?

Answer)Header records in log files
System=
Version=
Sub-version=

In the above three lines of headers that we do not want to include in our Hive query. To skip header lines from our tables in the Hive,set a table property that will allow us to skip the header lines.

www.smartdatacamp.com

```
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);
```

98)What is the maximum size of string data type supported by hive? Mention the Hive support binary formats

Answer)The maximum size of string data type supported by hive is 2 GB.

Hive supports the text file format by default and it supports the binary format Sequence files, ORC files, Avro Data files, Parquet files.

Sequence files: Splittable, compressible and row oriented are the general binary format.

ORC files: Full form of ORC is optimized row columnar format files. It is a Record columnar file and column oriented storage file. It divides the table in row split. In each split stores that value of the first row in the first column and followed sub subsequently.

AVRO datafiles: It is same as a sequence file splittable, compressible and row oriented, but except the support of schema evolution and multilingual binding support.

99)What is the precedence order of HIVE configuration?

Answer)We are using a precedence hierarchy for setting the properties

SET Command in HIVE

The command line –hiveconf option
Hive-site.XML
Hive-default.xml
Hadoop-site.xml
Hadoop-default.xml

100)If you run a select * query in Hive, Why does it not run MapReduce?

Answer)The hive.fetch.task.conversion property of Hive lowers the latency of mapreduce overhead and in effect when executing queries like SELECT, FILTER, LIMIT, etc., it skips mapreduce function

101)How Hive can improve performance with ORC format tables?

Answer)We can store the hive data in highly efficient manner in the Optimized Row Columnar file format. It can simplify many Hive file format limitations. We can improve the performance by using ORC files while reading, writing and processing the data.

```
Set hive.compute.query.using.stats-true;
Set hive.stats.dbclass-fs;
CREATE TABLE orc_table (
idint,
name string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY \:
LINES TERMINATED BY \n
STORES AS ORC;
```

102)What is available mechanism for connecting from applications, when we run hive as a server?

Answer)Thrift Client: Using thrift you can call hive commands from various programming languages. Example: C++, PHP,Java, Python and Ruby.

JDBC Driver: JDBC Driver supports the Type 4 (pure Java) JDBC Driver

ODBC Driver: ODBC Driver supports the ODBC protocol.

103)Explain about the different types of join in Hive?

Answer)HiveQL has 4 different types of joins –

JOIN- Similar to Outer Join in SQL

FULL OUTER JOIN – Combines the records of both the left and right outer tables that fulfil the join condition.

LEFT OUTER JOIN- All the rows from the left table are returned even if there are no matches in the right table.

RIGHT OUTER JOIN-All the rows from the right table are returned even if there are no matches in the left table.

104)How can you configure remote metastore mode in Hive?

Answer)To configure metastore in Hive, hive-site.xml file has to be configured with the below property –

www.smartdatacamp.com

hive.metastore.uris

thrift: //node1 (or IP Address):9083

IP address and port of the metastore host

105)What happens on executing the below query? After executing the below query, if you modify the column how will the changes be tracked?

Answer)Hive> CREATE INDEX index_bonuspay ON TABLE employee (bonus)

AS org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler;

The query creates an index named index_bonuspay which points to the bonus column in the employee table. Whenever the value of bonus is modified it will be stored using an index value.

106)How to load Data from a .txt file to Table Stored as ORC in Hive?

Answer)LOAD DATA just copies the files to hive datafiles. Hive does not do any transformation while loading data into tables.

So, in this case the input file /home/user/test_details.txt needs to be in ORC format if you are loading it into an ORC table.

A possible workaround is to create a temporary table with STORED AS TEXT, then LOAD DATA into it, and then copy data from this table to the ORC table.
Here is an example:
CREATE TABLE test_details_txt(visit_id INT, store_id SMALLINT) STORED AS TEXTFILE;
CREATE TABLE test_details_orc(visit_id INT, store_id SMALLINT) STORED AS ORC;
Load into Text table
LOAD DATA LOCAL INPATH /home/user/test_details.txt INTO TABLE test_details_txt;
Copy to ORC table
INSERT INTO TABLE test_details_orc SELECT * FROM test_details_txt;

107)How to create HIVE Table with multi character delimiter

Answer)FILELDS TERMINATED BY does not support multi-character delimiters. The easiest way to do this is to use RegexSerDe:

```
CREATE EXTERNAL TABLE tableex(id INT, name STRING)
ROW FORMAT org.apache.hadoop.hive.contrib.serde2.RegexSerDe
WITH SERDEPROPERTIES (
input.regex = ^(\\d+)~\\*(.*)$
)
STORED AS TEXTFILE
LOCATION /user/myusername;
```

108)Is there any way to get the column name along with the output while execute any query in Hive?

Answer)If we want to see the columns names of the table in HiveQl, the following hive conf property should be set to true.

hive> set hive.cli.print.header=true;

If you prefer to see the column names always then update the $HOME/.hiverc file with the above setting in the first line..

Hive automatically looks for a file named .hiverc in your HOME directory and runs the commands it contains, if any

109)How to Improve Hive Query Performance With Hadoop?

Answer)Use Tez Engine

Apache Tez Engine is an extensible framework for building high-performance batch processing and interactive data processing. It is coordinated by YARN in Hadoop. Tez improved the MapReduce paradigm by increasing the processing speed and maintaining the MapReduce ability to scale to petabytes of data.

Tez engine can be enabled in your environment by setting hive.execution.engine to tez:

set hive.execution.engine=tez;

Use Vectorization

Vectorization improves the performance by fetching 1,024 rows in a single operation instead of fetching single row each time. It improves the performance for operations like filter, join, aggregation, etc.

Vectorization can be enabled in the environment by executing below commands.
set hive.vectorized.execution.enabled=true;
set hive.vectorized.execution.reduce.enabled=true;

Use ORCFile

Optimized Row Columnar format provides highly efficient ways of storing the hive data by reducing the data storage format by 75% of the original. The ORCFile format is better than the Hive files format when it comes to reading, writing, and processing the data. It uses techniques like predicate push-down, compression, and more to improve the performance of the query.

Use Partitioning

With partitioning, data is stored in separate individual folders on HDFS. Instead of querying the whole dataset, it will query partitioned dataset.
1)Create Temporary Table and Load Data Into Temporary Table
2)Create Partitioned Table
3)Enable Dynamic Hive Partition
4)Import Data From Temporary Table To Partitioned Table

www.smartdatacamp.com

Use Bucketing

The Hive table is divided into a number of partitions and is called Hive Partition. Hive Partition is further subdivided into clusters or buckets and is called bucketing or clustering.

Cost-Based Query Optimization

Hive optimizes each querys logical and physical execution plan before submitting for final execution. However, this is not based on the cost of the query during the initial version of Hive.

During later versions of Hive, query has been optimized according to the cost of the query (like which types of join to be performed, how to order joins, the degree of parallelism, etc.).

www.smartdatacamp.com