1. **What is Apache Hive?**

   Apache Hive is an open-source data warehouse system for querying and analysing large datasets stored in Hadoop files. Hadoop is a framework for handling large datasets in a distributed computing environment.

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.

3. **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.

4. **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.

5. **What is SerDe in Apache Hive?**

   SerDe is short for Serializer/Deserializer. Hive uses the SerDe interface for IO. The interface handles both serialization and deserialization and also interpreting the results of serialization as individual fields for processing. A SerDe allows Hive to read in data from a table, and write it back out to HDFS in any custom format. Anyone can write their own SerDe for their own data formats.

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.

6. **Wherever (Different Directory) you 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.

7. **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.

8. **Is multiline comment supported in Hive Script?**

No.

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

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

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

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

ODBC Driver: It supports ODBC protocol.

10. **Which classes are used by the 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.

11. 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.

12. 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.

13. 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.

14. 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).

15. What are the different types of tables available in Hive?

There are two types.

a. Managed table: in which both the data and schema are under control of Hive.

b. External table: where only the schema is under control of Hive.

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

No, Hive does not provide insert and update at row level. So it is not suitable for OLTP system.

17. Can a table be renamed in Hive?

Alter Table table_name RENAME TO new_name

18. 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.

19. 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.

20. 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.

21. 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.

22. 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 running 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;

23. 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 running this command.

set hive.vectorized.execution.enabled=true;

24. Difference between sort by or order by clause in Hive? Which is the fast?

ORDER BY – sort the data in one reducer.

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.

Sort by is much faster than order by.

25. What are different Hive metastore configurations?

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 Thift 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.

26. Can Hive process any type of databases?

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 whereas different type of data is in the Hadoop, Hive use different SerDe interface to process such data.

27. What Is the HWI?

The Hive Web Interface is an alternative to the Hive command line interface. HWI is a simple graphical interface.

28. What is the difference between Like and Rlike operators in HIVE?

Like is used to find the substrings within a main string with regular expression %.

Rlike is a special function which also finds the sub strings within a main string, but return true or false without using regular expression.

29. What are the Hive default read and write classes?

Hive uses following classes to read and write the files.

TextInputFormat/HiveIgnoreKeyTextOutputFormat

SequenceFileInputFormat/SequenceFileOutputFormat

First class used to read/write the plain text. Second class used for sequence files.

30. 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.

31. 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.

32. What is different between database and data-warehouse?

Typically database is designed for OLTP transactional operations whereas Datawarehouse 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 whereas OLAP process Retired, historic and static data only.

Database completely has normalization concept. DWH is Denormalization concept.

33. 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. For example, if you partition table, table schema stores data in the external table.

In Managed table, Schema is stored in the local system, but in External table MetaStore is 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.

34. What is the use of partition in hive?

To analyse a particular set of data, not required to load entire data, desired data partition is a good approach. To achieve this goal, Hive allows partitioning the data based on particular column. Static partition and Dynamic partition both can optimize the Hive performance.

35. 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.

36. 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 whereas Hbase is a NoSQL database.

Hive run on the top of Mapreduce, Hbase run on the top of HDFS.

37. 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.

38. Why do we use 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.

39. How does 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.

40. 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.

41. When do we use explode in Hive?

Sometimes Hadoop developer takes array as input and converts into a separate table row. To achieve this goal, Hive uses explode, which acts as an 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;

42. 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 datatypes like int, bigint, STRING, but also process complex datatypes like arrays, maps, structs and union.

43. Can you overwrite Hadoop Mapreduce configuration in Hive?

Yes, you can overwrite Hive map, reduce steps in hive configuration settings. Hive allows overwriting Hadoop configuration files.

44. How to display the present database name in the terminal?

There are two ways to know the current database. One is temporary in CLI and second one is persistently.

1) In CLI just enter this command: set hive.cli.print.current.db=true;

2) In hivesite.

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.

45. Is a job split into map?

No, Hadoop framework can split the datafile, but not Job. This chunks of data stored in blocks. Each split needs a map to process whereas Job is a configurable unit to control execution of the plan/logic. Job is not a physical dataset to split, it’s a logical configuration API to process those split.

46. What is the difference between Describe and describe extended?

To see table definition in Hive, use describe <table name>; command

Whereas

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.

47. What is difference between static and dynamic partition of a table?

To prune data during query, partition can minimize the query time. The partition is created when the data is inserted into table. Static partition can insert individual rows whereas Dynamic partition can process entire table based on a particular column. At least one static partition is must to create any (static, dynamic) partition. If you are partitioning a large datasets, doing a sort of ETL flow Dynamic partition is recommendable.

48. What is the difference between partition and bucketing?

The main aim of both Partitioning and Bucketing is execute the query more efficiently. When you are creating a table, the slices are fixed in the partitioning of the table. Bucketing follows Hash algorithm. Based on number of buckets, randomly the data is inserted into the bucket to sampling of the data.

49. What is the default location where hive stores table data?

hdfs://namenode_server/user/hive/warehouse

50. Is there a date data type in Hive?

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

51. What are collection data types in Hive?

There are three collection data types in Hive.

ARRAY

MAP

STRUCT

52. Can we run unix shell commands from hive?

Yes, using the '!' mark just before the command.

For example, !pwd at hive prompt will list the current directory.

53. What is a Hive variable? What for we use it?

The hive variable is variable created in the Hive environment that can be referenced by Hive scripts. It is used to pass some values to the hive queries when the query starts executing.

54. Can hive queries be executed from script files? How?

Using the source command.

Example: Hive> source /path/to/file/file_with_query.hql

55. What is the importance of .hiverc file?

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.

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

The default record delimiter is − \n

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

57. What do you mean by schema on read?

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

58. What does the "USE" command in hive do?

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

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

SHOW DATABASES LIKE 'p.*'

60. Can you delete the DBPROPERTY in Hive?

There is no way you can delete the DBPROPERTY.

61. What is the significance of the line

set hive.mapred.mode = strict; ?

Tutortek
#1569, Opp Navami Hotel,
East End Main Road,
Jayanagar 9th Block, Bangalore,560069.
Website: www.tutortek.com
Call Us:+91 9886635020 / 9686511166
Email: enquiry@tutortek.com

It sets the mapreduce jobs to strict mode. By which the queries on partitioned tables cannot run without a WHERE clause. This prevents very large job running for long time.

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

This can be done with following query

SHOW PARTITIONS table_name PARTITION(partitioned_column='partition_value')

63. Which java class handles the Input and output record encoding into files which store the tables in Hive?

Input: org.apache.hadoop.mapred.TextInputFormat

Output: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

64. What is the significance of 'IF EXISTS" clause while dropping a table?

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.

65. Does the archiving of Hive tables give any space saving in HDFS?

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

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

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

67. How can you stop a partition form being queried?

By using the ENABLE OFFLINE clause with ALTER TABLE statement.

68. 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)

ALTER TABLE table_name

CHANGE COLUMN new_col INT

BEFORE x_col

Tutortek
#1569, Opp Navami Hotel,
East End Main Road,
Jayanagar 9th Block, Bangalore,560069.
Website: www.tutortek.com
Call Us:+91 9886635020 / 9686511166
Email: enquiry@tutortek.com

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

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.

70. **How can Hive avoid mapreduce?**

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

71. **Is it possible to create Cartesian join between 2 tables, using Hive?**

No. This kind of Join cannot be implemented in mapreduce.

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

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

73. **What will be the result when you do cast('abc' as INT)?**

Hive will return NULL.

74. **As part of Optimizing the queries in Hive, what should be the order of table size in a join query?**

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

75. **How will you convert the string '51.2' to a float value in the price column?**

Select cast(price as FLOAT)

76. **Can the name of a view be same as the name of a hive table?**

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

77. **Can we LOAD data into a view?**

No. A view cannot be the target of INSERT or LOAD statement.

78. Give the command to see the indexes on a table.

SHOW INDEX ON table_name

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

79. What types of costs are associated in creating index on hive tables?

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

80. Can a partition be archived? What are the advantages and Disadvantages?

Yes. A partition can be archived. Advantage is it decreases the number of files stored in namenode and the archived file can be queried using hive. The disadvantage is it will cause less efficient query and does not offer any space savings.

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

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

82. What is a generic UDF in hive?

It is a UDF which is created using a java program to server some specific need not covered under the existing functions in Hive. It can detect the type of input argument programmatically and provide appropriate response.

83. How do you specify the table creator name when creating a table in Hive?

The TBLPROPERTIES clause is used to add the creator name while creating a table.

The TBLPROPERTIES is added like −

TBLPROPERTIES(‘creator’= ‘Joan’)

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

LOAD DATA LOCAL INPATH '${env:HOME}/country/state/'

OVERWRITE INTO TABLE address;

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

85. 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)

86. How to 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.

87. 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;

88. 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”);

89. 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>;

90. 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

91. Is HQL case sensitive?

No.

92. What is the maximum size of string data type supported by Hive?

Maximum size is 2GB.

93. 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.

94. 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 be stored as value. It supports 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

95. 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

96. Describe REPEAT function in Hive with example?

REPEAT function will repeat the input string n times specified in the command.

Example: REPEAT(‘Hive’,3);

Tutortek

Output: HiveHiveHive.

97. 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 using TRIM 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.

98. 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 ');

99. Describe REVERSE function in Hive with example?

REVERSE function will reverse the characters in a string.

Example: REVERSE('Hive');

Output: eviH

100. Is there any alternative way to rename a table without ALTER command?

By using Import and export options we can be rename a table as shown below. Here we 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 we prefer to just preserve the data, we 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;

101. How to rename a table in Hive?

Using ALTER command with RENAME, we can rename a table in Hive.

ALTER TABLE hive_table_name RENAME TO new_name;

102. What is Double data type in Hive?

Double data type in Hive will present the data differently unlike RDBMS.

See the double type data below:

14324.0

342556.0

1.28893E4

E4 represents 10^4 here. So, the value1.28893E4 represents 12889.3. All the calculations will be accurately performed using double type

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.

103. How can we change a column data type in Hive?

We can use below command to alter data type of a column in hive.

ALTER TABLE table_name CHANGE column_name column_name new_datatype;

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

ALTER TABLE employee CHANGE empid empid BIGINT;

104. How can we copy the columns of a hive table into a file?

By using awk command in shell, the output from HiveQL Describe command can be written to a file.

$ hive -S -e "describe table_name;" | awk -F" " ’{print 1}’ > ~/output.

105. 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

Relatively static data is analyzed,

Fast response times are not required, and

When the data is not changing rapidly.