**Hive Assignment 4**

• Can we use same name for a TABLE and VIEW in Hive?

The name of a view must be unique, and it cannot be the same as any table or database or view's name.

• How will you load data into a VIEW in Hive?

Views are read-only and may not be used as the target of LOAD/INSERT/ALTER.

• What is Bucketing in Hive?

Bucketing in hive is the concept of breaking data down into ranges, which are known as buckets, to give extra structure to the data so it may be used for more efficient queries. The range for a bucket is determined by the hash value of one or more columns in the dataset (or Hive metastore table)

• What are the pros and cons of archiving a partition in Hive?

The main advantage of archiving is that it will decrease the number of files to be stored in NameNode. We can even query an archived partition in Hive.

The main disadvantage of archiving is that queries become slower and less efficient in Hive.

• How can we specify in Hive to load an HDFS file in LOAD DATA?

LOAD DATA INPATH 'hdfs\_file' INTO TABLE tablename;

• What is a Skewed table in Hive?

A Skewed tables is a special type of table in which some values in a column appear more often. Due to this the distribution in skewed. In Hive, when we specify a table as SKEWED during creation, then skewed values are written into separate files and remaining values go to another file.

• What is the use of CLUSTERED BY clause during table creation in Hive?

“clustered by” clause is used to divide the table into buckets. Each bucket will be saved as a file under table directory.

• What is a Managed table in Hive?

Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive.

• How will you prevent data to be dropped or queried from a partition in Hive?

We can use ALTER TABLE table\_name ENABLE NO\_DROP to prevent a table partition from being dropped.

We can use ALTER TABLE table\_name ENABLE OFFLINE to prevent a table partition from

being queried. In offline mode, we can still access metadata of a table partition.

• What is the use of TOUCH in ALTER statement?

TOUCH clause in ALTER statement is used to read the metadata and write it back. This operation will modify the last accessed time of a partition in Hive.

With TOUCH statement we can also execute the POST and PRE hooks on a table partition.

• How does OVERWRITE clause work in CREATE TABLE statement in Hive?

We use OVERWRITE clause in CREATE TABLE statement to delete the existing data and write new data in a Hive table.

• What are the options to connect an application to a Hive server?

We can use following options to connect an application a Hive server:

JDBC Driver: We can use JDBC Driver with embedded as well as remote access to connect to HiveServer. This is for Java based connectivity.

Python Client: For Python language application there is Python client that can connect to Hive server.

Ruby Client: With Ruby client driver also we can connect to Hive server.

Thrift Client: We can use Beeline command line shell to connect to Hive server over Thrift. For production mode, this is one of the very good options. It is a secure option for production use. Also we do not need to grant HDFS access to users for using Thrift client.

• How TRIM and RPAD functions work in Hive?

With TRIM function we can delete the spaces before and after a String

E.g. TRIM(‘ Smith ’)

Smith

RPAD function is used to add padding (extra spaces) in a String on the right hand side.

E.g. Let say we have a String “Hello”.

LPAD(‘Hello’,8,’ ‘)

Hello

• How will you recursively access sub-directories in Hive?

hive> Set mapred.input.dir.recursive=true;

hive> Set hive.mapred.supports.subdirectories=true;

• What is the optimization that can be done in SELECT \* query in Hive?

Write column names needed only and use filter

• What is the use of ORC format tables in Hive?

The Optimized Row Columnar (ORC) file format provides a highly efficient way to store Hive data. It was designed to overcome limitations of the other Hive file formats. Using ORC files improves performance when Hive is reading, writing, and processing data.

• What are the main use cases for using Hive?

The most predominant use cases for Apache Hive are to batch SQL queries of sizable data sets and to batch process large ETL and ELT jobs.

• What are the different Types of Tables available in Hive?

There are two types of tables in Hive ,one is Managed table and second is external table. the difference is , when you drop a table, if it is managed table hive deletes both data and meta data,if it is external table Hive only deletes metadata.

• Is Hive suitable to be used for Oltp systems?

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

• Can Table be Renamed in Hive?

You can rename the table name in the hive. You need to use the alter command.

• Can we change Data Type of column in Hive Table?

Yes, By using this command below one can change the column data type: ALTER TABLE table\_name CHANGE column\_name column\_name new\_datatype;

• What is Metastore in Hive?

Metastore is the central repository of Apache Hive metadata. It stores metadata for Hive tables (like their schema and location

• What is the need for Custom Serde?

SerDe means Serializer and Deserializer. Hive uses SerDe and FileFormat to read and write table rows. Main use of SerDe interface is for IO operations. A SerDe allows hive to read the data from the table and write it back to the HDFS in any custom format.

• Why do we need Hive?

Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data.

• What is the Default Location where Hive stores Table Data?

**/usr/hive/warehouse** is the default location for all managed tables. External tables may be stored at a different location.

• What are the Three Different Modes in which Hive can be run?

Standalone Mode.

Pseudo-distributed Mode.

Fully-Distributed Mode.

• Is there a Date Data Type in Hive?

Hive provides DATE and TIMESTAMP data types in traditional UNIX time stamp format for date/time related fields in hive.

• What are Collection Data Types in Hive?

**Hive data types can be classified into two parts.**

Primitive Data Types :

Primitive Data Types also divide into 3 types which are as follows.

Numeric Data Type

TINYINT – 1 byte signed integer -128 to 127

SMALLINT – ‘2 byte signed integer -32, 768 to 32, 767

INT – 4 byte signed integer –2,147,483,648 to 2,147,483,647

BIGINT – 8 byte signed integer’ 9,223,372,036,854,775,808 to 9,223,372,036,854,775,807

FLOAT – ‘Single precision floating point Single Precision

DOUBLE – Double precision floating point Double Precision

DECIMAL – Precise decimal type based on Java BigDecimal Object Big Decimal

Date/Time Data Type

TIMESTAMP

String Data Type

Complex data types

Array

MAP

STRUCT

UNION

• Can we run Unix Shell Commands from Hive?

simple rule to run Unix command in Hive shell is to put ! (exclamation) sign before every Unix command

• What is Hive Variable?

Hive variables are key-value pairs that can be set using the set command and they can be used in scripts and Hive SQL. The values of the variables in Hive scripts are substituted during the query construct.

• Can Hive Queries be executed from Script Files?

It is possible by using the source command.

For example :

Hive> source /path/to/file/file\_with\_query.hql

• 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

• 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.

• How do you find list all Databases whose name starts with P?

SHOW DATABASES LIKE ‘p.\*’

• 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

• How can you Delete Dbproperty in Hive?

There is no way you can delete the DBPROPERTY.

• What is the Significance of the Line Set Hive.mapred.mode = Strict.

It sets the mapreduce jobs to strict mode.By which the queries on partitioned tables can not run

without a WHERE clause. This prevents very large job running for long time

• 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’)

• Which Java Class handles the Input Record Encoding into files which store Tables in Hive?

org.apache.hadoop.mapred.TextInputFormat

• Which Java Class handles the Output Record Encoding into files which result from Hive Queries?

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

• What is the significance of if Exists clause while dropping 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.

• When you point a Partition of Hive Table to New Directory so what happens with Data?

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

• Write Query to Insert New Column new\_col Int into Hive Table at a position before an existing Column x\_col.

ALTER TABLE table\_name

CHANGE COLUMN new\_col INT

BEFORE x\_col

• Does 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.

• How can you Stop Partition form being queried?

By using the ENABLE OFFLINE clause with ALTER TABLE statement.

• While loading Data into Hive Table using Load Data Clause so how do you specify it is a Hdfs File and not a Local File?

By Omitting the LOCAL CLAUSE in the LOAD DATA statement.

• If you omit Overwrite clause while creating Hive Table so what happens with 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.

• What is a Table Generating Function on Hive?

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

• 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.

• Is it possible to create Cartesian Join between 2 Tables using Hive?

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

• As part of optimizing Queries in Hive what should be the Order of Table Size in 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.

• What is the Usefulness of the Distributed by clause in Hive?

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

• Can Name of a View be Same as 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.

• Can we load Data into View?

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

• What types of Costs are associated in creating Index in Hive Tables?

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