**Hive Assignment 4**

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

No, the name of the view must be unique.

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

We can’t load or insert data into a view.

• What is Bucketing in Hive?

The values in a column are hashed into a number of buckets which is defined by user. It is a way to avoid too many partitions or nested partitions while ensuring optimizes query output.

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

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.

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

By excluding LOCAL clause in load data statement.

• What is a Skewed table in Hive?

A table that is having values that are present in large numbers in the table compared to other data.

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

Cluster By is a short-cut for both Distribute By and Sort By. Hive uses the columns in Distribute By to distribute the rows among reducers. All rows with the same Distribute By columns will go to the same reducer.

• 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?

By using the ENABLE OFFLINE, ENABLE NO\_DROP clause with ALTER TABLE statement.

• What is the use of TOUCH in ALTER statement?

TOUCH clause in ALTER statement is used to read the metadata.

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

OVERWRITE clause is used in CREATE TABLE statement to delete the existing data and write new data in a Hive table and as the name suggests, OVERWRITE helps in overwriting existing data in a Hive table.

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

We can use Beeline, a JDBC, or an ODBC connection to Hive Server.

• How TRIM and RPAD functions work in Hive?

Both are functions for processing String data type in Hive.

TRIM function we can delete the spaces before and after a String. It is very useful for formatting user input in which user may have entered extra spaces.

RPAD function is used to add padding (extra spaces) in a String on the right-hand side. So that String reaches a specified length.

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

We can use following commands in Hive to recursively access sub-directories:

SET hive.mapred.supports.subdirectories=TRUE;

SET mapred.input.dir.recursive=TRUE;

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

By selecting only the necessary columns not \*.

By using order by with limit.

When using joins place the largest table first, followed by the smallest, and then by decreasing size.

By using WHERE clauses especially within joins, so the tables to be joined are as small as possible.

• 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?

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?

Managed table and external table. In managed table both the data an schema in under control of hive but in external table only the schema is under control of Hive.

• Is Hive suitable to be used for Oltp systems?

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

• Can Table be Renamed in Hive?

Yes, using the following command:

Alter Table table\_name RENAME TO new\_name;

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

Yes, using the following command:

ALTER TABLE <table-name> CHANGE <old-col-name> <new-col-name> <data-type>;

• 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) and partitions in a relational database.

• 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 is a tool in Hadoop ecosystem which allows users to read, write, and manage petabytes of data using SQL. It is suitable for accessing and analyzing data in Hadoop using SQL syntax.

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

hdfs://namenode\_server/user/hive/warehouse

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

Local mode, Distributed mode and Pseudo-distributed mode.

• Is there a Date Data Type in Hive?

Yes, 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?

There are four collection data types in hive:

Array, Map, Struct and Union.

• Can we run Unix Shell Commands from Hive?

Yes, by using ! mark before the command.

• What is Hive Variable?

The hive variable is variable created in the Hive environment that can be referenced by Hive scripts.

• Can Hive Queries be executed from Script Files?

Yes, by using HIVE -F command from shell.

• What are the default Record and Field Delimiter used for Hive Text Files?

The default record delimiter is − \n

And the field 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?

The USE command access the database on which all subsequent hive queries will run.

• How can you Delete Dbproperty in Hive?

DBPROPERTY can’t be deleted.

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

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

• How do you check if a Particular Partition Exists?

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 excluding LOCAL clause in 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.

• What is a Table Generating Function on Hive?

Table Generating Function transform a single row into multiple rows. Example, EXPLODE it’s the only table generated function. This function takes array as an input and outputs the elements of array into separate rows.

• How can Hive avoid Mapreduce?

By setting the property hive.exec.mode.local.auto to true, 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 is useful in case of streaming data.

• Can Name of a View be Same as Name of a Hive Table?

No, the name of the view must be unique.

• Can we load Data into View?

We can’t load or insert data into a view.

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

There is a processing cost in arranging the values of the column on which index is created since Indexes occupies.