**Hive Assignment 4**

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

No , can’t use same name , I try do that and get error



Within a database, base tables and views share the same namespace, so a base table and a view cannot have the same name.

1. How will you load data into a VIEW in Hive?

create view y as select \* from y ;

After created view create as select the view take structure and data from table

And no load or insert in view

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

1. What is Bucketing in Hive?

The bucketing in Hive is a data organizing technique. It is similar to partitioning in Hive with an added functionality that it divides large datasets into more manageable parts known as buckets. So, we can use bucketing in Hive when the implementation of partitioning becomes difficult. However, we can also divide partitions further in buckets.

The concept of bucketing is based on the hashing technique.

1. 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

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

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

1. What is a Skewed table in Hive?

A skewed table is a special type of table where the values that appear very often (heavy skew) are split out into separate files and rest of the values go to some other file

Ex: CREATE TABLE tableName (column1 STRING, column2 STRING) SKEWED BY (column1) on (‘value1’)

1. 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.

1. What is a Managed table in Hive?

Managed Table has full control over its dataset. That is, when you drop the table the table’s dataset or files will also be deleted from HDFS.

1. 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.

1. 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.

1. How does OVERWRITE clause work in CREATE TABLE statement in Hive?

The insert overwrite table query will overwrite the any existing table or partition in Hive. It will delete all the existing records and insert the new records into the table. If the table property set as **‘auto.purge’=’true’**, the previous data of the table is not moved to trash when insert overwrite query is run against the table.

If we not set the ‘auto.purge’=’true’ in the table properties and run the insert overwrite query frequently, it occupy the memory for the previous data in the trash and create the insufficient memory issue after some time.

1. What are the options to connect an application to 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.

1. How TRIM and RPAD functions work in Hive?

**TRIM** : 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. The other variations of TRIM function are LTRIM and RTRIM that remove spaces from left and right side of the string respectively.

**RPAD** : function is used to add padding (extra spaces) in a String on the right hand side. So that String reaches a specified length. LPAD function is same as RPAD but it pads on the left hand side of String.

1. How will you recursively access sub-directories in Hive?

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

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

Once above options are set to true, Hive will recursively access sub-directories of a directory in MapReduce.

1. What is the optimization that can be done in SELECT \* query in Hive?
2. What is the use of ORC format tables in Hive?

The ORC file format provides a highly efficient way to store data in Hive table. This file system was actually designed to overcome limitations of the other Hive file formats. The Use of ORC files improves performance when Hive is reading, writing, and processing data from large tables.

1. 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.

1. What are the different Types of Tables available in Hive?

There are two types. 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.

1. 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.

1. Can Table be Renamed in Hive?

Alter Table table\_name RENAME TO new\_name

1. Can we change Data Type of column in Hive Table?

Using REPLACE column option

ALTER TABLE table\_name REPLACE COLUMNS

1. What is Metastore in Hive?

Metastore is the central repository of Hive metadata. It is divided into 2 pieces: a service and the backing store for the data. By default, the metastore service runs in the same JVM as the Hive service and contains an embedded Derby database instance backed by the local disk. This is called as the embedded metastore configuration.

1. What is the need for Custom Serde?

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.

1. Why do we need Hive?

Hive is a tool in Hadoop ecosystem which provides an interface to organize and query data in a databse like fashion and write SQL like queries. It is suitable for accessing and analyzing data in Hadoop using SQL syntax.

1. What is the Default Location where Hive stores Table Data?

“/user/hive/warehouse

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

Local mode

Distributed mode

Pseudodistributed mode

1. Is there a Date Data Type in Hive?

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

1. What are Collection Data Types in Hive?

Array , Map and struct

1. Can we run Unix Shell Commands from Hive?

Yes , by using ! mark before statement

1. What is Hive Variable?

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.

1. Can Hive Queries be executed from Script Files?

Yes , Using the source command.

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

The default record delimiter is − \n

1. 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.

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

SHOW DATABASES LIKE ‘p.\*’

1. 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.

1. How can you Delete Dbproperty in Hive?

There is no way you can delete the DBPROPERTY.

1. What is the Significance of the Line

Set Hive.mapred.mode = Strict.

1. How do you check if a Particular Partition Exists?

SHOW PARTITIONS table\_name PARTITION(partitioned\_column=’partition\_value’)

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

org.apache.hadoop.mapred.TextInputFormat

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

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

1. 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.

1. 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.

1. Write Query to Insert New Column new\_col Int into Hive Table at a position before an existing Column x\_col.
2. ALTER TABLE table\_name
3. CHANGE COLUMN new\_col INT
4. BEFORE x\_col
5. 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.

1. How can you Stop Partition form being queried?

By using the ENABLE OFFLINE clause with ALTER TABLE atatement.

1. 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.

1. 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

1. 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. Example exploe()

1. 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.

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

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

1. 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.

1. What is the Usefulness of the Distributed by clause in Hive?

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

1. Can Name of a View be Same as Name of a Hive Table?

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

1. Can we load Data into View?

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

1. 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