**Hive Assignment 4**

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

No, we cannot use same name for a table and view in Hive. So we have to select a unique name for a view in Hive.

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

We can only load data into a Hive table not VIEW.

• What is Bucketing in Hive?

Hive Bucketing a.k.a (Clustering) is a technique to split the data into more manageable files, (By specifying the number of buckets to create). The value of the bucketing column will be hashed by a user-defined number into buckets.

Bucketing can be created on just one column, you can also create bucketing on a partitioned table to further split the data which further improves the query performance of the partitioned table.

Each bucket is stored as a file within the table’s directory or the partitions directories. Note that partition creates a directory and you can have a partition on one or more columns; these are some of the differences between Hive partition and bucket.

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

We can archive some less used partitions 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?

We have to remove the LOCAL clause from LOAD DATA statement to avoid loading the local file. Once LOCAL clause is removed, we can load HDFS file.

• 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.

E.g. CREATE TABLE tableName (column1 STRING, column2 STRING) SKEWED BY (column1) on (‘value1’)

During queries, we get better performance in Hive with SKEWED tables.

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

It is used to distribute and sort table into number of buckets specified by the user

• 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. Saved under /user/hive/warehouse

• 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?

In Hive, 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.

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

There is no overwrite clause in CREATE TABLE statement !

• 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. 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:

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.

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

We can convert some of the SELECT queries in Hive into single FETCH task. With this optimization, latency of SELECT query is decreased.

To use this we have to set the value of hive.fetch.task.conversion parameter. The permissible values are:

0: It means FETCH is disabled.

1: It is minimal mode. SELECT \*, FILTER on partition columns (WHERE and HAVING clauses), LIMIT only

2: It is more mode: SELECT, FILTER, LIMIT only (including virtual columns) “more” can even take UDF expressions in the SELECT clause.

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

We use Optimized Row Columnar (ORC) file format to store data efficiently in Hive. It is used for performance improvement in reading, writing and processing of data.

In ORC format, we can overcome the limitations of other Hive file formats. Some of the advantages of ORC format are:

There is single file as the output of each task. This reduces load on NameNode.

It supports date time, decimal, struct, map etc complex types.

It stores light-weight indexes within the file.

We can bound the memory used in read/write of data.

It stores metadata with Protocol Buffers that supports add/remove of fields.

• What are the main use cases for using Hive?

Some of the main use cases for Hive are:

Analysis of static Big data Applications in which less responsive time is acceptable .

Analysis of data that does not change rapidly.

What are the different Types of Tables available in Hive?

Managed Table and External Table

• Is Hive suitable to be used for Oltp systems?

No, because hive does not support insert and update at row level

Hive used for batch processing (OLAP)

Can’t be used for OLTP because real time operation of database

you can’t delete data from a table using DML

• Can Table be Renamed in Hive?

ALTER TABLE old\_name RENAME TO new\_name;

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

ALTER TABLE TABLE\_NAME CHANGE COLUMN\_NAME COLUMN\_NAME NEW\_DATATYPE

• What is Metastore in Hive?

Metastore is a central repository of Apache HIVE metadata, It stores metadata for hive tables like schema and location of tables and partitions of tables in a relational database.

• What is the need for Custom Serde?

it instructs hive how to process the record

Usually I need the custom SerDe to read data in format which is not supported till now in built-in SerDe ,so you will write the schema of what you want to read.

• Why do we need Hive?

* To manage, store and process large datasets and all that using SQL
* Hive is the data warehouse for big data platform you can store in it any structured data and analyze it using sql.

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

* /user/hive/warehouse/database\_name/table\_name

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

1. Local Mode: Map Reduce jobs run locally on user machine. This is default mode which Hadoop uses local file system.
2. Distributed Mode: Hive as well as Hadoop is running in a fully distributed manner. Name Node, Data Node, Job Tracker ... etc run on different machines in this mode
3. Pseudo Distributed Mode: This is mode used by developer to test the code before deploying to production. In this mode all daemons run on same virtual machine.

• Is there a Date Data Type in Hive?

* Hive provides for date -> Date(yyyy-mm-dd), TIMESTAMP(yyyy-mm-dd hh:mm:ss) data types

• What are Collection Data Types in Hive?

* Hive offers several collection data types:

1 – Struct: you define structure with established columns and data types,

Example: address struct<street:string, city:string, state:string, zipcode:int>

2 - Map: less structured, <key,value> for column and define the data type for key and value.

Example: preferences map<pref\_code string, pref\_value string>

3 – Array: allow you to store n number of values of same data type

• Can we run Unix Shell Commands from Hive?

* Yes, by using “!” before the command

• What is Hive Variable?

-Hive variables are key value pairs that can be set using set command and they can be used in hive scripts and hive sql.

- Hive stores variables in four different namespaces:

- hive config

- hive var

- hive system

- env

• Can Hive Queries be executed from Script Files?

* By using this command
  + source /path/to/file/file\_with\_query.hql

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

* The field record delimiter:’ \n’
* The fields delimiter: ‘,’

• 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?

* By using “use” command, you can use specific database then any query you will write applied to that database

• 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’t run without a WHERE clause. This prevents very large job running for long time.

• How do you check if a Particular Partition Exists?

* By using this command:
  + hive> SHOW PARTITIONS songs PARTITION(year='1969',artist\_name='The Box Tops')

• 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 NEW\_COL INT

BEFORE X\_COL ;

• Does Archiving of Hive Tables give any space saving in Hdfs?

* No, it not giving any extra space because it only reduces the number of files stored in name node.

• 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.

• 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: explode

• 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 cannot 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 ensures that is not overlapping exists between data when distributed it among the reducers

• 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’t be the target of an 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 created.