**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?

It is not possible to load data into views, hive view is just a projection of a hive table just a select statement from table.

• What is Bucketing in Hive?

The concept of breaking data down into ranges, to give extra structure to data. (More Organization)

The Range of buckets determined by the hash values of columns used in CLUSTERED BY.

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

Pros: It will decrease the number of files to be stored in the Name Node and we can even query archived partition in hive.

Cons: Queries become slower and less efficient in HIVE.

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

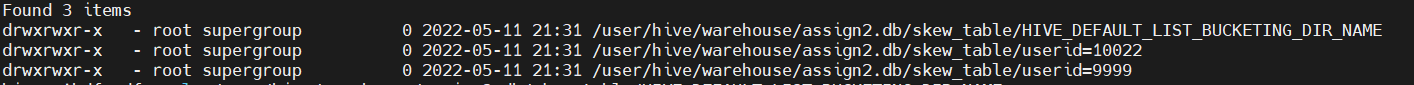
LOAD DATA INPATH ‘HDFS FILE PATH’ INTO TABLE TABLE\_NAME;

• What is a Skewed table in Hive?

Skewed tables are those in which some column values occur more frequently than others. When create a skewed table, Hive automatically separate skewed values into different files or directories, and take this in consideration when searching, thus enhance the performance.

A picture containing text

Description automatically generated



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

CLUSTERED BY is used to divide the table into buckets based on some column or columns, The records with the same clustered column go to same bucket depend on hash function. CLUSTERED BY is consist of DISTRIBUTE BY, SORTED BY.

• What is a Managed table in Hive?

Managed tables are HIVE owned tables where the entire life cycle of the tables data are managed and controlled by Hive.

IF you drop the managed table, the structure of table in metastore and the data will be deleted.

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

* To Prevent Table partition from being dropped:
  + ALTER TABLE table\_name ENABLE NO\_DROP
* To Prevent Table partition from being queried:
  + ALTER TABLE table\_name ENABLE OFFLINE

• What is the use of TOUCH in ALTER statement?

Touch read metadata and writes it back. During the process it updates “TIMESTAMP METADATA creation”

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

OVERWRITE used to delete the existing data and write new data in HIVE table.

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

You can use Beeline, a JDBC, or an ODBC connection to HiveServer.

• How TRIM and RPAD functions work in Hive?

* TRIM: delete the spaces before and after a String
  + EXAMPLE: TRIM (‘ABDOL ‘) -> ‘ABDOL’
* RPAD: adding extra spaces to right of string to reach a defined length
  + EXAMPLE:RPAD(‘ABDOL’,’8’,’ ‘) -> ‘ABDOL ‘

• 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?

SELECT COL1, COL2, COL3, … FROM TABLE;

Because select \* is cost inefficient and can hurt the performance, because the db engines first will scan the metastore to retrieve the column names then select from data.

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

ORC reduces the size of the original data up to 75%.

Using ORC files improves performance when Hive is reading, writing, and processing data comparing to Text,Sequence and Rc.

• What are the main use cases for using Hive?

 batch SQL queries of sizable data sets

batch process large ETL AND ELT jobs

• 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?

* Yes, using

ALTER TABLE db\_name.old\_table\_name RENAME TO db\_name.new\_table\_name;

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

YES, by using this command

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?

* SerDe : instruct 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 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 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 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 created.