**Hive Assignment 4**

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

No, it should be unique.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

* We can’t load a data into a view, a view is a virtual table which Is read only, used to get the data from a parent table.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

• What is Bucketing in Hive?

* 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.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

* main advantage of archiving is that it will decrease the number of files to be stored in NameNode.
* 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 ‘file\_source\_path’ into table table\_name.

Without mention local keyword

• 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.
* Advantage of having such a setup is that for the values that appear more frequently than other values get split out into separate files(or separate directories if we are using STORED AS DIRECTORIES clause). And this information is used by the execution engine during query execution to make processing more efficient.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

• 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.  Bucketed tables will create almost equally distributed data file parts.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

• What is a Managed table in Hive?

* A Managed table is a table which is controled and managed by hive, which means dropping or deletion of the table will cause of deleting the file source also, Also support ACID operations.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

• 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.
* 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. Essentially, as the name suggests.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

• 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?

* 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.

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

* Set mapred.input.dir.recursive=true;
* Set hive.mapred.supports.subdirectories=true;

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

* SELECT fields instead of SELECT \*.
* Mention databasename with tables name (dbname.tbname) for faster accessing.

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

* provides a highly efficient way to storing and compression Hive data.
* Better performance when selecting certain columns from a table

• What are the main use cases for using Hive?

* to 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
* External table
* Virtual table

• Is Hive suitable to be used for Oltp systems?

* No, bcs hive is not a trasncational db, doesn’t support insert and update for a row level.

• Can Table be Renamed in Hive?

* Yes , using ALTER TABLE … RENAME TO .. command.

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

* Yes, by using ALTER TABLE.. CHANGE colname colname data\_type

• What is Metastore in Hive?

* The metastore could be any RDBMS database, such as MySQL or Oracle, which used to store table, database, schema, and other HQL definitions

• What is the need for Custom Serde?

* If we have semi structured data, it 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?

* As hadoop ecosystem is written in java, so any job that need to be submitted in hadoop cluster need to be programmed in java, and also map-reduce framework isn’t friendly to work with, so at the end you will need a java expert to deal with any job,

An alternative to this is Hive, As it has an execution engine that converts hql to java.

* Anyone with w DB background can deal with it.
* Allow Creation of table from already existing data In hdfs
* Hive gives you a dwh capabilities
* You can partition and bucketing your data for a better performance.

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

/user/hive/warehouse/dbname.db/

• 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?

* Yes, Date and timestamp.

• What are Collection Data Types in Hive?

* Struct
* Maps
* Arrays

• Can we run Unix Shell Commands from Hive?

* Yes, put ! (exclamation) sign before every Unix command

• What is Hive Variable?

* key-value pairs that can be set using the set command,

• Can Hive Queries be executed from Script Files?

* Yes, by putting them in a .hql file format
* Executre with hive -f ‘filepath’.

• 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 data is not verified before loading but rather when a query is issued, which means data is checked with the schema when any query is issued on it.

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

* SHOW DATABASES LIKE 'p\*';

• What does the use command in Hive do?

* To use specific database.

• How can you Delete Dbproperty in Hive?

* We cannot delete the DBPROPERTY in Hive.

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

* It will inforce you to create at least static partition column before dymanic parittiions columns.

• 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?

* allows the statement to succeed even if the specified tables does not exist.

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

* If there was a Data with the table it won’t be moved with the table.
* If a new data is loaded it will be stored in that location.

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

ALTER TABLE h\_table  
  
 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?

ALTER TABLE table\_name ENABLE OFFLINE

• 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?

Load data inpath ‘file\_source\_path’ into table table\_name.

Without mention local keyword

• 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

• What is a Table Generating Function on Hive?

* **transform a single row into multiple rows**. EXPLODE is 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?

* When queried SELECT, FILTER, LIMIT queries, this property skip mapreduce and using FETCH task. As a result Hive can execute query without run mapreduce task.

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

* Yes.

• As part of optimizing Queries in Hive what should be the Order of Table Size in Join Query?

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

* **distribute the input rows among reducers**. It ensures that all rows for the same key columns are going to the same reducer. So, if we need to partition the data on some key column, we can use the DISTRIBUTE BY clause in the hive queries

• Can Name of a View be Same as Name of a Hive Table? No,

• Can we load Data into View? No,

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

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