**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?  
 A view can’t be the target of an INSERT or LOAD statement.

• What is Bucketing in Hive?  
Bucketing in hive is the concept of breaking data down into ranges, which are known as buckets, to give extra structure to the data so it may be used for more efficient queries. The range for a bucket is determined by the hash value of one or more columns in the dataset (or Hive metastore table). These columns are referred to as `bucketing` or `clustered by` columns

• What are the pros and cons of archiving a partition in Hive?  
Pros: 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.  
Cons: 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?

if we do not use local keyword, it assumes it as a HDFS Path.

Load data inpath '/data/empnew.csv' into table emp

• What is a Skewed table in Hive?  
A skewed table is a table that is having values that are present in large numbers in the table compared to other data. Skewed data is stored in a separate file while the rest of the data is stored in a separate file

• What is the use of CLUSTERED BY clause during table creation in Hive?  
 Cluster By used as an alternative for both Distribute BY and Sort BY clauses in Hive-QL.

Cluster BY clause used on tables present in Hive. Hive uses the columns in Cluster by to distribute the rows among reducers. Cluster BY columns will go to the multiple reducers.

It ensures sorting orders of values present in multiple reducers

* What is a Managed table in Hive?

In a managed table, both the table data and the table schema are managed by Hive. The data will be located in a folder named after the table within the Hive data warehouse, which is essentially just a file location in HDFS.

The location is user-configurable when Hive is installed. By managed or controlled we mean that if you drop (delete) a managed table, then Hive will delete both the Schema (the description of the table) and the data files associated with the table. Default location is /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.

• 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.  
This statement cannot be used for creating a table or partition if it does not exist yet.

• 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, OVERWRITE helps in overwriting existing data in a Hive table.

• 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.

• How TRIM and RPAD functions work in Hive?  
TRIM and RPAD functions are for processing String data type in Hive.  
With 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?

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?

First, tweak your data through partitioning, bucketing, compression, etc. Improving the execution of a hive query is another Hive query optimization technique. You can do this by using Tez, avoiding skew, and increasing parallel execution. Lastly, sampling and unit testing can help optimize a query by allowing you to see (and solve) problems on a smaller scale, first.

• 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?

Apache Hive not only analyzes data, but also **enables its users to read and write the data in an organized manner**. What's more is that this software defines specific schemas related to data analysis and stores them in Hadoop Distributed File System (HDFS) which helps in future analysis.

• What are the different Types of Tables available in Hive?  
There are two types. Managed table and external table.

In a managed table, both the table data and the table schema are managed by Hive. The data will be located in a folder named after the table within the Hive data warehouse, which is essentially just a file location in HDFS. The location is user-configurable when Hive is installed. By managed or controlled we mean that if you drop (delete) a managed table, then Hive will delete both the Schema (the description of the table) and the data files associated with the table. Default location is /user/hive/warehouse).

An external table is one where only the table schema is controlled by Hive. In most cases, the user will set up the folder location within HDFS and copy the data file(s) there. This location is included as part of the table definition statement. When an external table is deleted, Hive will only delete the schema associated with the table. The data files are not affected.

• 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.

• Can Table be Renamed in Hive?  
 Alter Table table\_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\_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?  
 A SerDe **allows hive to read the data from the table and write it back to the HDFS in any custom format**. If we have unstructured data, then we use RegEx SerDe which will instruct hive how to handle that record. We can also write our own Custom SerDe in any format.

• Why do we need Hive?

Apache Hive not only analyzes data, but also **enables its users to read and write the data in an organized manner**. What's more is that this software defines specific schemas related to data analysis and stores them in Hadoop Distributed File System (HDFS) which helps in future analysis.

• 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  
Pseudodistributed mode

• Is there a Date Data Type in Hive?  
The hive timestamp format is **YEAR-Month-Day-Hours-Minutes-seconds [YYYY-MM-DD HH:MM:SS] with an optional fraction of seconds**. Anything else that will come with the above format should be read as a string and converted later

• What are Collection Data Types in Hive?  
ARRAY  
MAP  
STRUCT

• Can we run Unix Shell Commands from Hive?  
Yes, using the ! mark just before the command. For example !pwd at hive prompt will list the current director

• 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.

• Can Hive Queries be executed from Script Files?  
hive -f script\_file.hql

• 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 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?  
With the use command you fix the database on which all the subsequent hive queries will run.

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

• How do you check if a Particular Partition Exists?  
This can be done with following query   
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?  
The IF EXISTS clause **allows the statement to succeed even if the specified tables does not exist**. If the table does not exist and you do not include the IF EXISTS clause, the statement will return an error.

• 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 atatement.

• 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. If you add the OVERWRITE clause then all the existing data in the directory will be deleted before new data is written.

• 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 controls how the map output is reduced among the reducers. 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 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.