**Hive Assignment 4**

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

No, 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 insert data in a Hive view, Hive view is just a projection of a Hive table.

• 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 megastore table).

• 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 Name Node. 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?

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 skew table is a table that is having values that are present in large numbers in the table compared to other data.

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

CLUSTERED BY in Hive is same as DISTRIBUTE BY and SORT BY. When we specify CLUSTERED BY, it will first distribute the data into different reducers by using a Hash. Once data is distributed, it will sort the data.

• What is a Managed table in Hive?

It is the default table in Hive. When the user creates a table in Hive without specifying it as external, then by default, an internal table gets created in a specific location in HDFS.

By default, an internal table will be created in a folder path similar to /user/hive/warehouse directory of HDFS. We can override the default location by the location property during table creation.

Graphical user interface, text, application

Description automatically generated

• 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.  
  
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.

• 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. It is very useful for formatting user input in which user may have  entered extra  spaces.

E.g. TRIM(‘ Smith ’) Smith.

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?

Use Column Names instead of \* in SELECT Clause

This seems to be odd but it will definitely improve the performance of Hive query on TEXT file format.

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

The *Optimized Row Columnar* ([ORC](https://orc.apache.org/)) 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.

The ORC file format provides the following advantages:

* Efficient compression.
* Fast reads.
* Proven in large-scale deployments.

• What are the main use cases for using Hive?

Hive is an abstraction layer where the SQL like syntax is converted to map reduce jobs - Hive capabilities are extended with the ability to use various execution engines (TEZ) - Hive stores metadata in a conventional SQL database (usually MYSQL) - Hive supports a wide range of SQL syntax and sql analytical queries, Hive supports user defined functions and different file formats.

• What are the different Types of Tables available in Hive?

### 1. Hive Internal Table

It is the default table in Hive. When the user creates a table in Hive without specifying it as external, then by default, an internal table gets created in a specific location in HDFS.

By default, an internal table will be created in a folder path similar to **/user/hive/warehouse** directory of HDFS.

### 2. Hive External Table

We create an external table for external use as when we want to use the data outside the Hive.

External tables are stored outside the warehouse directory. They can access data stored in sources such as remote HDFS locations or Azure Storage Volumes.

We can create the external table by specifying the **EXTERNAL** keyword in the Hive create table statement.

Graphical user interface, text, application, chat or text message

Description automatically generated

• Is Hive suitable to be used for Oltp systems?

No, it is not suitable for OLTP system since it does not offer insert and update at the row level.

• Can Table be Renamed in Hive?

YES: we can rename the table name in the hive. We need to use the alter command.

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

Yes I can change the column data type:

ALTER TABLE table\_name CHANGE column\_name column\_name new\_datatype;

• What is Metastore in Hive?

MetaStore is a central repository of Hive, that allows to store meta data in external database. By default Hive store meta data in Derby database, but you can store in MySql, Oracle depends on project.

Hive metastore (HMS) is a service that stores metadata related to Apache Hive and other services, in a backend RDBMS, such as MySQL or PostgreSQL. Impala, Spark, Hive, and other services share the metastore. The connections to and from HMS include HiveServer, Ranger, and the NameNode that represents HDFS.

• What is the need for Custom Serde?

SerDe is a Serializer DeSerializer. Hive uses SerDe to read and write data from tables. Generally, users prefer to write a Deserializer instead of a SerDe as they want to read their own data format rather than writing to it.

• Why do we need Hive?

 Hive allows users to read, write, and manage petabytes of data using SQL.

Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data.

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

/usr/hive/warehouse is the default location for all managed tables. External tables may be stored at a different location.

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

The three modes in which Hadoop can run are :

Standalone mode: This is the default mode. It uses the local FileSystem and a single Java process to run the Hadoop services.

Pseudo-distributed mode: This uses a single-node Hadoop deployment to execute all Hadoop services.

Fully-distributed mode: This uses separate nodes to run Hadoop master and slave services.

• Is there a Date Data Type in Hive? Yes

• What are Collection Data Types in Hive?

Basically, in Hive, there are three collection data types. Such as;

ARRAY, MAP, STRUCT

• Can we run Unix Shell Commands from Hive?

Yes, when, just before the command we use the! mark, we run UNIX shell commands from the hive.

For example-

! pwd at hive prompt will list the current directory.

• What is Hive Variable?

A variable created in the Hive environment that can be referenced by Hive scripts is what we call a Hive Variable. Basically, when the query starts executing it is used to pass some values to the hive queries.

• Can Hive Queries be executed from Script Files?

It is possible by using the source command.

For example -

Hive> source /path/to/file/file\_with\_query.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?

Schema on read refers to an innovative data analysis strategy in new data-handling tools like Hadoop and other more involved database technologies. In schema on read, data is applied to a plan or schema as it is pulled out of a stored location, rather than as it goes in.

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

show databases like '%<DB\_NAME>'

• What does the use command in Hive do?

Basically, fix the database on which all the subsequent hive queries will run we use the “USE” command in Hive.

• 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.

Basically, in strict mode, it sets the MapReduce jobs. So, by which the queries on partitioned tables cannot run without a WHERE clause. Hence, it prevents very large job running for a long time.

• How do you check if a Particular Partition Exists?

Basically, with the following query, we can check whether a particular partition exists or not

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?

Since, the table being dropped does not exist in the first place, Hive throws an error, when we issue the command DROP TABLE IF EXISTS table\_name.

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

Basically, the data stays in the old location. Hence, 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 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?

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 exploe()

• 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 ho wthe 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 whne compared to all other tables and views present in the same database.

• Can we load Data into View?

No. A view can not 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 cerated.