**Hive Assignment 4**

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

The view must be unique not as table name

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

We can’t load data into view because we select view from table

Create view (column a,column b) as select

view is a purely logical object with no associated storage

• 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. And we can take sample of it to test something and machine learning

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

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

Make directory and put the file in hdfs and load data using load data inpath command

!hdfs load data inpath ‘/employee’ 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. Skew data is stored in a separate file while the rest of the data is stored in a separate file.

If we make analysis on data that meause which gender who uses skin care products

We will find a large number of data go through females

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

It’s combination of distribute by and sort by to make the data ordered and the query will run on a multiple reducers

ensures that each of the N reducers gets non-overlapping sets, then sorts at the reducers by those ranges.

• What is a Managed table in Hive?

* The table that managed and stored by hive and it’s an internal table
* It stored in /hive/warehouse
* If a managed table or partition is dropped, the data and metadata associated with that table or partition are deleted, the data and metadata associated with that table or partition are deleted

• 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 from being queried

• What is the use of TOUCH 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?

OVERWRITE is used to replace any existing data in the table or partition and insert with the new rows.

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

• How TRIM and RPAD functions work in Hive?

-Delete blank spaces after and before string

-Add padding of any string like spaces or anything like regex

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

Set mapred.input.dir.recursive=true;

Set hive.mapred.supports.subdirectories=true

Once above options are set to true,

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

-Select specific columns and column names

-access table by database like assign.emp to avoid searching in data dictionary

-apply where clause if needed to filter data before selection

• 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.

• 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?

Internal and external

• Is Hive suitable to be used for Oltp systems?

We can’t update and insert in row

And hive is considered to be like a datawarehouse or datastore and works on a large datasets and it’s a top level of hive and store a huge amont of data in hdfs

• Can Table be Renamed in Hive?

Yes , Alter Table table\_name RENAME TO new\_name

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

Yes, Using REPLACE column option

• What is Metastore in Hive?

It stores metadata for Hive tables (like their schema and location) and partitions in a relational database. It provides client access to this information by using metastore service API.

• What is the need for Custom Serde?

Yes, SerDe is a Library which is built-in to the Hadoop API

Hive uses Files systems like HDFS or any other storage (FTP) to store data, data here is in the form of tables (which has rows and columns).

SerDe - Serializer, Deserializer instructs hive on how to process a record (Row). Hive enables semi-structured (XML, Email, etc) or unstructured records (Audio, Video, etc) to be processed also. For Example If you have 1000 GB worth of RSS Feeds (RSS XMLs). You can ingest those to a location in HDFS. You would need to write a custom SerDe based on your XML structure so that Hive knows how to load XML files to Hive tables or other way around.

• Why do we need Hive?

To query and use sql statements to get KPIs and analysis of data

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

/user/hive/warehouse

• 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. The TIMESTAMP data types stores date in java.sql.timestamp forma

• What are Collection Data Types in Hive?

Array, map, struct, enum

• Can we run Unix Shell Commands from Hive?

Yes, I can run it by writing ! before command

• 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?

Yes , run it by hive -f txt.csv

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

Record: /n

Field: ;

• 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?

Use and access specific database

• How can you Delete Dbproperty in Hive?

We can’t 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?

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?

Drop tale table\_name if not exist;

• 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?

write ENABLE OFFLINE clause in ALTER TABLE

• 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?

Using Load data inpath ‘/emp/emp.csv’ into table table\_name not load data local

• 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?

These functions transform a single row into multiple rows.

• 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?

Yes

• 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 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?

The view must be unique not as table name

• Can we load Data into View?

We can’t load data into view because we select view from table

Create view (column a,column b) as select

view is a purely logical object with no associated storage

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

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