**Hive Assignment 4**

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

No. The name of a view must be unique when compared to all other tables and views present in the same database.

• 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?

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 disadvantage is it will cause less efficient query and does not offer any space savings.

• 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.

• 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.We have to specify CLUSTERED BY clause during table creation.

• What is a Managed table in Hive?

A managed table is stored under the [hive.metastore.warehouse.dir](https://cwiki.apache.org/confluence/display/Hive/Configuration+Properties" \l "ConfigurationProperties-hive.metastore.warehouse.dir) path property, by default in a folder path similar to /user/hive/warehouse/databasename.db/tablename/. The default location can be overridden by the location property during table creation. If a managed table or partition is dropped, the data and metadata associated with that table or partition are deleted. If the PURGE option is not specified, the data is moved to a trash folder for a defined duration.

Use managed tables when Hive should manage the lifecycle of the table, or when generating temporary tables

• 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?

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?

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. The other variations of TRIM function are LTRIM and RTRIM that remove spaces from left and right side of the string respectively.

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. LPAD function is same as RPAD but it pads on the left hand side of String.

E.g. Let say we have a String “Hello”.

LPAD(‘Hello’,8,’ ‘)  
Hello

We can also specify our optional padding character in RPAD and LPAD functions.

These functions are similar to the ones in SQL.

• 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?

Hive query on TEXT file format. [Hive file formats](https://dwgeek.com/hive-different-file-formats-text-sequence-rc-avro-orc-parquet-file.html/) such as PARQUET, ORC replaces \* with actual column names.

Filtering data before select where , having before select.

Access table with database name to avoid the search headache on the metastore.

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

The ORC file format provides the following advantages:

Efficient compression: Stored as columns and compressed, which leads to smaller disk reads. The columnar format is also ideal for vectorization optimizations in Tez.

Fast reads: ORC has a built-in index, min/max values, and other aggregates that cause entire stripes to be skipped during reads. In addition, predicate pushdown pushes filters into reads so that minimal rows are read. And Bloom filters further reduce the number of rows that are returned.

Proven in large-scale deployments: Facebook uses the ORC file format for a 300+ PB deployment.

• 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?

There are two types. Managed table and external table. In managed table both the data an schema in under control of hive but in external table only the schema is under control of Hive.

• 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?

mention your current table name after the ALTER TABLE command and specify the new table name after the RENAME TO command.

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

By using the command alter table.

• What is Metastore in Hive?

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. Beeline, Hue, JDBC, and Impala shell clients make requests through thrift or JDBC to HiveServer. The HiveServer instance reads/writes data to HMS. By default, redundant HMS operate in active/active mode. The physical data resides in a backend RDBMS, one for HMS. All HMS instances use the same backend database. A separate RDBMS supports the security service, Ranger for example.

• What is the need for Custom Serde?

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?

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?

Hive stores tables files by default at /user/hive/warehouse location on HDFS file system.

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

Local mode: In Hive local mode, Map Reduce jobs related to Hive run locally on a user machine. This is the default mode in which Hadoop uses local file system.

Distributed Mode: In this mode, Hive as well as Hadoop is running in a fully distributed mode. NameNode, DataNode, JobTracker, TaskTracker etc run on different machines in this mode.

Pseudo-distributed Mode: This is the mode used by developers to test the code before deploying to production. In this mode, all the daemons run on same virtual machine. With this mode, we can quickly write scripts and test on limited data sets.

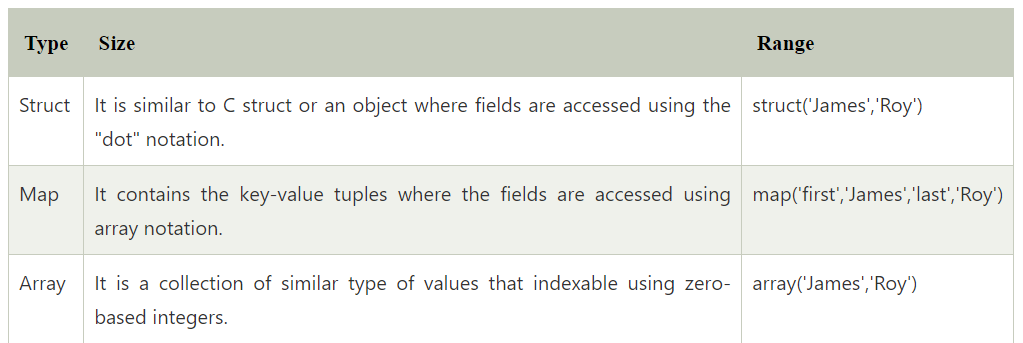
• Is there a Date Data Type in Hive?

yes, hive has date and timestamp as a primitive data type .DATE values are represented in the form YYYY-­MM-­DD. Example: DATE ‘2014-­12-­07’. Date ranges allowed are 0000-­01-­01 to 9999-­12-­31.

• What are Collection Data Types in Hive?

complex data type

Array – Map- Struct- uniontype



• Can we run Unix Shell Commands from Hive?

A simple rule to run [Unix command](https://studywithswati.wordpress.com/category/unix/) in Hive shell is to put ! (exclamation) sign before every Unix command.

• What is Hive Variable?

It stores values, you can either set value for hive variables at session level or while launching hive shell.

Set myvar=2

or

$hive –d myvar=2

• Can Hive Queries be executed from Script Files?

Yes, file saved with extension. hql then run it with hive –f script.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 data is checked with the schema when any query is issued 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 choose 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.

If your partitioned table is very large, you could block any full table scan queries by putting Hive into strict mode using the set hive.mapred.mode=strict command. In this mode, when users submit a query that would result in a full table scan (i.e. queries without any partitioned columns) an error is issued.

• 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?

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 to the new directory.

• 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 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 removing 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?

It 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?

It is not possible to insert data in a Hive view, Hive view is just a projection of a Hive table.

• 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.