**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. But it possible if they are not in the sane scheme.

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

Hive view is just a projection of a Hive table (you can see it as pre saved query). From Hive documentation. view is a purely logical object with no associated storage. (No support for materialized views is currently available in Hive.) When a query references a view, the view's definition is evaluated in order to produce a set of rows for further processing by the query. (This is a conceptual description; in fact, as part of query optimization, Hive may combine the view's definition with the query's, e.g. pushing filters from the query down into the view.)

• 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)

• 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 NameNode. 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?

Create a folder on HDFS 🡪 hadoop fs -mkdir

Move the text file from local file system into newly created folder 🡪 hadoop fs -put

Load Data from HDFS path into HIVE TABLE.

• 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 (the values that appear very often -heavy skew) are split out and stored in a separate file while the rest of the data is stored in some other file.

• 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. Bucketing can be done along with partitioning or without partitioning on Hive tables. Bucketed tables will create almost equally distributed data file.

• What is a Managed table in Hive?

Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. All the write operations to the Managed tables are performed using Hive SQL commands.

• 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?

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?

OVERWRITE is used to replace any existing data in the table or partition and insert with the new rows. When working with the partition you can also specify to overwrite only when the partition exists using the IF NOT EXISTS option.

• 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?

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.

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.

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

Using Hive in HDInsight.

SET hive.mapred.supports.subdirectories=TRUE;

SET mapred.input.dir.recursive=TRUE;

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

Aliasing , use filters on data, access table with database name.

• 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?

The most predominant use cases for Apache Hive are to batch SQL queries of sizable data sets and to batch process large ETL and ELT jobs.

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

There are two types of tables in Hive ,one is Managed table and second is external table. the difference is , when you drop a table, if it is managed table hive deletes both data and meta data, if it is external table Hive only deletes metadata.

• 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?

We can rename the table name in the hive.

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

Yes, By using this command below one can change the column data type: ALTER TABLE table\_name CHANGE column\_name column\_name new\_datatype;

• 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. It provides client access to this information by using metastore service API.

• What is the need for Custom Serde?

We 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?

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. Hadoop Integration: Runs on top of Hadoop, with Apache Tez or MapReduce for processing.

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

/usr/hive/warehouse.

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

Hadoop – Different Modes of Operation

Standalone Mode.

Pseudo-distributed Mode.

Fully-Distributed Mode.

• Is there a Date Data Type in Hive?

Hive provides DATE and TIMESTAMP data types in traditional UNIX time stamp format for date/time related fields in hive. 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.Dec 8, 2014

Cast(string as date): If the string is in the form ‘YYYY-MM-DD’, then a date value corr...

Cast(date as string): Date is formatted as a string in the form ‘YYYY-MM-DD’

Cast(date as timestamp): Midnight of the year/month/day of the date value is returned ...

Cast(timestamp as date): The year/month/day of the timestamp is returned

• What are Collection Data Types in Hive?

Hive Primitive Datatypes

Data type Size

TINYINT – 1 byte signed integer -128 to 127

SMALLINT – '2 byte signed integer -32, 768 to 32, 767

INT – 4 byte signed integer –2,147,483,648 to 2,147,483,647

BIGINT – 8 byte signed integer' 9,223,372,036,854,775,808 to 9,223,372,036,854,775,807

|  |
| --- |
| FLOAT – ‘Single precision floating point |
| DOUBLE – Double precision floating point |
| DECIMAL  – Precise decimal type based on Java Big Decimal Object |

• Can we run Unix Shell Commands from Hive?

Yes.

• What is Hive Variable?

Hive variables are key-value pairs that can be set using the set command and they can be used in scripts and Hive SQL. The values of the variables in Hive scripts are substituted during the query construct.

• Can Hive Queries be executed from Script Files?

Yes , We can.

• 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.

• What do you mean by Schema on Read?

we have Schema on read, which means the data is not verified before loading but rather when a query is issued, here we have very fast initial load as the data is not read.

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

Show databases or Show schemas statement is lists all the database names in Hive metastore. In this statement, will use the optional LIKE clause allows the list of databases to be filtered using a regular expression ‘P \*’.

• What does the use command in Hive do?

Allow to use table in database without accessing with names.

• 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 sets the map reduce 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?

Show partitions table name.

• Which Java Class handles the Input Record Encoding into files which store Tables in Hive?

The following Java class is used to handle the input record encoding into files that store Hive tables: 'org. apache.

• Which Java Class handles the Output Record Encoding into files which result from Hive Queries?

The following Java class is used to handle the output record encoding into Hive query files: 'org. apache.

• 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. Before dropping a table, you must first remove any stored procedures that reference the table.

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

Changing the point of partition will not move the data to the new location.

• 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?

When creating your table set the option **Require partition filter** which will force the users to use partition filter when accessing the 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?

Use load data in path not load data local inpath.

• 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?

These functions 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?

conversion property can (FETCH task) minimize latency of mapreduce overhead. 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 with cross join.

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

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?

to distribute the rows among reducers. All rows with the same Distribute By columns will go to the same reducer. However, Distribute By does not guarantee clustering or sorting properties on the distributed keys.

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

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