Hive Database

In Hive, the database is considered as a catalog or namespace of tables. So, we can maintain multiple tables within a database where a unique name is assigned to each table. Hive also provides a default database with a name **default**.

- **hive> show databases;** **// to check the existing databases**
- **hive> create database demo;** **// to create new database**
- **hive> create database if not exists demo;**
- **hive>create database demo**

  >WITH DBPROPERTIES ('creator' = 'sumit', 'date' = '2019-06-03');
- **hive> describe database extended demo;**
- **hive> drop database demo;**
- **hive> drop database if exists demo;**
- **hive> drop database if exists demo cascade;**

➢ **Hive Tables**

In Hive, we can create a table by using the conventions similar to the SQL. It supports a wide range of flexibility where the data files for tables are stored. It provides two types of table: -

- Internal table
- External table

Internal Table

The internal tables are also called managed tables as the lifecycle of their data is controlled by the Hive. By default, these tables are stored in a subdirectory under the directory defined by hive.metastore.warehouse.dir (i.e. /user/hive/warehouse). The internal tables are not flexible enough to share with other tools like Pig. If we try to drop the internal table, Hive deletes both table schema and data.

hive> create table demo.employee (Id int, Name string , Salary float)
row format delimited
fields terminated by ',' ;

External Table

The external table allows us to create and access a table and a data externally. The **external** keyword is used to specify the external table, whereas the **location** keyword is used to determine the location of loaded data.

As the table is external, the data is not present in the Hive directory. Therefore, if we try to drop the table, the metadata of the table will be deleted, but the data still exists.

$hdfs dfs -mkdir /HiveDirectory
$hdfs dfs -put hive/emp_details /HiveDirectory

hive> create external table emplist (Id int, Name string , Salary float)
row format delimited
fields terminated by ','
location '/HiveDirectory';

select * from emplist;

Hive - Load Data

Once the internal table has been created, the next step is to load the data into it. So, in Hive, we can easily load data from any file to the database.

- Let's load the data of the file into the database by using the following command: -

  hive> load data local inpath '/homehive/empdata' into table demo.emp;

  Here, *empdata* is the file name that contains the data.

  hive> select * from demo.emp;

If we want to add more data into the current database, execute the same query again by just updating the new file name.

hive>load data local inpath '/home/hive/empdata1' into table demo.emp;

hive>select * from demo.emp;

- In Hive, if we try to load unmatched data (i.e., one or more column data doesn't match the data type of specified table columns), it will not throw any exception. However, it stores the Null value at the position of unmatched tuple.

➢ **Hive - Drop Table**

Hive facilitates us to drop a table by using the SQL **drop table** command. Let's follow the below steps to drop the table from the database.

- Let's check the list of existing databases by using the following command:
  hive> show databases;
- Now select the database from which we want to delete the table by using the following command: -
  hive> use demo;
- Let's check the list of existing tables in the corresponding database.
  hive> show tables;
- Now, drop the table by using the following command: -
  hive> drop table demo.emp;

- Let's check whether the table is dropped or not.
  hive> show tables;

Hive - Alter Table

In Hive, we can perform modifications in the existing table like changing the table name, column name, comments, and table properties. It provides SQL like commands to alter the table.

Rename a Table

If we want to change the name of an existing table, we can rename that table by using the following signature: -

hive>ALTER TABLE old_table_name RENAME TO new_table_name;

Let's see the existing tables present in the current database.

hive>SHOW TABLES;

Adding column

In Hive, we can add one or more columns in an existing table by using the following syntax

hive>ALTER TABLE table_name ADD COLUMNS(column_name datatype);

Ex: hive>Alter table demo.emp add columns (age int);

hive>describe demo.emp;
hive>select * from demo.emp;

Change Column

In Hive, we can rename a column, change its type and position. Here, we are changing the name of the column by using the following signature: -

hive>ALTER TABLE table_name CHANGE old_column_name new_column_name datatyp;

Ex: hive>ALTER TABLE demo.emp CHANGE name fname string;
hive>DESCRIBE demo.emp;
hive>SELECT * FROM demo.emp;

Delete or Replace Column

Hive allows us to delete one or more columns by replacing them with the new columns. Thus, we cannot drop the column directly.

Let's see the existing schema of the table.

hive>DESCRIBE demo.emp;

```
OK
id                      int
name                    string
salary                  float
age                     int
Time taken: 0.308 seconds, Fetched: 4 row(s)
hive>
```

Now, drop a column from the table.

hive>ALTER TABLE demp.emp REPLACE COLUMNS(id string, first_name string, age int);

hive>DESCRIBE demo.emp;

```
OK
id                      string
first_name              string
age                     int
Time taken: 0.337 seconds, Fetched: 3 row(s)
hive>
```

Operations in HiveQL

Arithmetic Operations

hive> select id, name, salary + 50 from employee;
hive> select id, name, (salary * 10) /100 from employee; **//find 10% salary from each employee**

Relational Operators

hive> select * from employee where salary >= 25000;
hive> select * from employee where salary < 25000;

Aggregate Functions in Hive

In Hive, the aggregate function returns a single value resulting from computation over many rows. Let"s see some commonly used aggregate functions: -

Return Type	Operator	Description
BIGINT	count(*)	It returns the count of the number of rows present in the file.
DOUBLE	sum(col)	It returns the sum of values.
DOUBLE	sum(DISTINCT col)	It returns the sum of distinct values.
DOUBLE	avg(col)	It returns the average of values.
DOUBLE	avg(DISTINCT col)	It returns the average of distinct values.
DOUBLE	min(col)	It compares the values and returns the minimum one form it.
DOUBLE	max(col)	It compares the values and returns the maximum one form it.

Sorting and Aggregating

- Sorting data in Hive can be achieved by using a standard ORDER BY clause, but there is a catch. ORDER BY produces a result that is totally sorted, as expected, but to do so it sets the number of reducers to one, making it very inefficient for large datasets.
- When a globally sorted result is not required and in many cases it isn't, you can use Hive's nonstandard extension, SORT BY, instead. SORT BY produces a sorted file per reducer.
- In some cases, you want to control which reducer a particular row goes to, typically so you can perform some subsequent aggregation. This is what Hive's DISTRIBUTE BY clause does.

GROUP BY and HAVING Clause

The Hive Query Language provides GROUP BY and HAVING clauses that facilitate similar functionalities as in SQL. Here, we are going to execute these clauses on the records of the below table:

- *Create Table:*

hive> create table emp (Id int, Name string, Salary float, Desig string)

>row format delimited

>fields terminated by ',' ;

- *Load data into emp table:*

hive> load data local inpath '/home/hive/emp_data' into table emp;

ID	NAME	SALARY	DESIG
1	John	40,000	Sr. Manager
2	Harry	30,000	Developer
3	Clark	25,000	Developer
4	Ram	35,000	Manager
5	Imran	50,000	Sr. Manager
6	Lisa	45,000	Manager
7	Syam	35,000	Manager

GROUP BY Clause:

The HQL Group By clause is used to group the data from the multiple records based on one or more column. It is generally used in conjunction with the aggregate functions (like SUM, COUNT, MIN, MAX and AVG) to perform an aggregation over each group.

hive> SELECT desig, sum(salary) from emp GROUP BY desig;
Developer 55,000
Manager 1,15,000
Sr. Manager 90,000

HAVING Clause:

The HQL HAVING clause is used with GROUP BY clause. Its purpose is to apply constraints on the group of data produced by GROUP BY clause. Thus, it always returns the data where the condition is TRUE.

hive> SELECT desig, sum(salary) from emp
>GROUP BY desig HAVING salary>=30,000;
Developer 30,000
Manager 1,15,000
Sr. Manager 90,000

ORDER BY and SORT BY Clause

By using HiveQL ORDER BY and SORT BY clause, we can apply sort on the column. It returns the result set either in ascending or descending order. Here, we are going to execute these clauses on the records of the below table:

ID	NAME	SALARY	DESIG
1	John	40,000	Sr. Manager
2	Harry	30,000	Developer
3	Clark	25,000	Developer
4	Ram	35,000	Manager
5	Imran	50,000	Sr. Manager
6	Lisa	45,000	Manager
7	Syam	35,000	Manager

ORDER BY Clause

In HiveQL, ORDER BY clause performs a complete ordering of the query result set. Hence, the complete data is passed through a single reducer. This may take much time in the execution of large datasets. However, we can use LIMIT to minimize the sorting time.

hive> select * from emp order by salary desc;

ID	NAME	SALARY	DESIG
5	Imran	50,000	Sr. Manager
6	Lisa	45,000	Manager
1	John	40,000	Sr. Manager
4	Ram	35,000	Manager
7	Syam	35,000	Manager
2	Harry	30,000	Developer
3	Clark	25,000	Developer

SORT BY Clause

The HiveQL SORT BY clause is an alternative of ORDER BY clause. It orders the data within each reducer. Hence, it performs the local ordering, where each reducer's output is sorted separately. It may also give a partially ordered result.

hive> select * from emp sort by salary desc;

ID	NAME	SALARY	DESIG
5	Imran	50,000	Sr. Manager
6	Lisa	45,000	Manager
1	John	40,000	Sr. Manager
4	Ram	35,000	Manager
7	Syam	35,000	Manager
2	Harry	30,000	Developer
3	Clark	25,000	Developer

DISTRIBUTE BY:

Distribute BY clause used on tables present in Hive. Hive uses the columns in Distribute by to distribute the rows among reducers. All Distribute BY columns will go to the same reducer.

- It ensures each of N reducers gets non-overlapping ranges of column
- It doesn't sort the output of each reducer

hive>SELECT Id, Name from emp DISTRIBUTE BY Id;