set hive.warehouse.subdir.inherit.perms = false

Creating Hive Table Partitioned by Multiple Columns and Importing Data

We will see how to create a Hive table partitioned by multiple columns and how to import data into the table.

Partitioning

We can use partitioning feature of Hive to divide a table into different partitions. Each partition of a table is associated with a particular value(s) of partition column(s). Partitioning allows Hive to run queries on a specific set of data in the table based on the value of partition column used in the query.

Without partitioning, any query on the table in Hive will read the entire data in the table.  
If we have a large table then queries may take long time to execute on the whole table. We can make Hive to run query only on a specific partition by partitioning the table and running queries on specific partitions. A table can be partitioned on columns like – city, department, year, device etc.

Create Table

We will create an Employee table partitioned by state and department. The partition columns need not be included in the table definition. They are available to be used in the queries. When the table is partitioned using multiple columns, then Hive creates nested sub-directories based on the order of the partition columns.

CREATE TABLE Employee(

ID BIGINT,

NAME STRING,

AGE INT,

SALARY BIGINT

)

COMMENT 'This is Employee table stored as textfile partitioned by STATE and DEPARTMENT'

PARTITIONED BY(STATE STRING, DEPARTMENT STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

Load data into the table Partitions

**Input Files :-**

Suppose we have 2 states – Maharashtra and karnataka and we have 2 departments – HR and BIGDATA.

On our HDFS, we have records for employees from Maharashtra state HR department in the file /user/bigdata01/mh\_hr\_employees.csv‘ and records for employees from Maharashtra state BIGDATA department in the file /user/bigdata01/mh\_bigdata\_employees.csv‘.

We have records for employees from Karnataka state HR department in the file /user/bigdata01/kr\_hr\_employees.csv‘ and records for employees from Karnataka state BIGDATA department in the file /user/bigdata01/kr\_bigdata\_employees.csv‘.

Contents of /user/bigdata01/mh\_hr\_employees.csv‘ :-

1,aarti,28,30000

2,sakshi,22,20000

3,mahesh,25,25000

Contents of /user/bigdata01/mh\_bigdata\_employees.csv‘ :-

10001,rajesh,29,50000

10002,rahul,23,250000

10003,dinesh,35,70000

Contents of /user/bigdata01/kr\_hr\_employees.csv‘ :-

600001,shruti,28,31000

600002,purvi,26,27000

600003,manav,35,54000

Contents of /user/bigdata01/kr\_bigdata\_employees.csv‘ :-

70001,sujay,39,72000

70002,raman,43,850000

70003,ravi,28,60000

**Load Data :-**

Now to import data for employees into their respective partitions in the Hive table, run following queries.

#Load employees data for partition having State as Maharashtra and department as HR

LOAD DATA INPATH ‘/user/bigdata01/mh\_hr\_employees.csv’

INTO TABLE Employee PARTITION (state='Maharashtra', department='HR');

#Load employees data for partition having State as Maharashtra and department as BIGDATA

LOAD DATA INPATH /user/bigdata01/mh\_bigdata\_employees.csv'

INTO TABLE Employee PARTITION (state='Maharashtra', department='BIGDATA');

#Load employees data for partition having State as Karnataka and department as HR

LOAD DATA INPATH /user/bigdata01/kr\_hr\_employees.csv'

INTO TABLE Employee PARTITION (state='Karnataka', department='HR');

#Load employees data for partition having State as Maharashtra and department as BIGDATA

LOAD DATA INPATH /user/bigdata01/kr\_bigdata\_employees.csv'

INTO TABLE Employee PARTITION (state='Karnataka', department='BIGDATA');

Once data is loaded in the table partitions, we can see that Hive has created two directories under the Employee table directory on HDFS – /user/hive/warehouse/employee and two sub-directories under each directory.

/user/hive/warehouse/employee/state=Maharashtra/department=HR

/user/hive/warehouse/employee/state=Maharashtra/department=BIGDATA

/user/hive/warehouse/employee/state=Karnataka/department=HR

/user/hive/warehouse/employee/state=Karnataka/department=BIGDATA

**Verify Data :-**

**select \* from employee where state='Maharashtra' and department='HR';**

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| employee.id  | employee.name  | employee.age  | employee.salary  | employee.state  | employee.department  |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| 1            | aarti          | 28            | 30000            | Maharashtra     | HR                   |

| 2            | sakshi         | 22            | 20000            | Maharashtra     | HR                   |

| 3            | mahesh         | 25            | 25000            | Maharashtra     | HR                   |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

**select \* from employee where state='Maharashtra' and department='BIGDATA';**

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| employee.id  | employee.name  | employee.age  | employee.salary  | employee.state  | employee.department  |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| 10001        | rajesh         | 29            | 50000            | Maharashtra     | BIGDATA              |

| 10002        | rahul          | 23            | 250000           | Maharashtra     | BIGDATA              |

| 10003        | dinesh         | 35            | 70000            | Maharashtra     | BIGDATA              |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

**select \* from employee where state='Maharashtra';**

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| employee.id  | employee.name  | employee.age  | employee.salary  | employee.state  | employee.department  |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| 10001        | rajesh         | 29            | 50000            | Maharashtra     | BIGDATA              |

| 10002        | rahul          | 23            | 250000           | Maharashtra     | BIGDATA              |

| 10003        | dinesh         | 35            | 70000            | Maharashtra     | BIGDATA              |

| 1            | aarti          | 28            | 30000            | Maharashtra     | HR                   |

| 2            | sakshi         | 22            | 20000            | Maharashtra     | HR                   |

| 3            | mahesh         | 25            | 25000            | Maharashtra     | HR                   |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

**select \* from employee;**

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| employee.id  | employee.name  | employee.age  | employee.salary  | employee.state  | employee.department  |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

| 70001        | sujay          | 39            | 72000            | Karnataka       | BIGDATA              |

| 70002        | raman          | 43            | 850000           | Karnataka       | BIGDATA              |

| 70003        | ravi           | 28            | 60000            | Karnataka       | BIGDATA              |

| 600001       | shruti         | 28            | 31000            | Karnataka       | HR                   |

| 600002       | purvi          | 26            | 27000            | Karnataka       | HR                   |

| 600003       | manav          | 35            | 54000            | Karnataka       | HR                   |

| 10001        | rajesh         | 29            | 50000            | Maharashtra     | BIGDATA              |

| 10002        | rahul          | 23            | 250000           | Maharashtra     | BIGDATA              |

| 10003        | dinesh         | 35            | 70000            | Maharashtra     | BIGDATA              |

| 1            | aarti          | 28            | 30000            | Maharashtra     | HR                   |

| 2            | sakshi         | 22            | 20000            | Maharashtra     | HR                   |

| 3            | mahesh         | 25            | 25000            | Maharashtra     | HR                   |

+--------------+----------------+---------------+------------------+-----------------+----------------------+--+

We can see that query for a particular partition reads data from that partition only and therefore the queries on a set of partitions perform fast on partitioned tables. In the query result, the partition columns and their values are added at the last.

If no partitioned columns are used in the query, then all the directories are scanned and partitioning will not have any effect.

Choosing right columns to partition the table is the major task as this will greatly impact the query performance.