Select Query with Group by clause in Hive

We will see how to write a Select query using Group by clause in Hive.

Hive Table

We have a table ‘Employee’ in Hive with the following schema and data.

0: jdbc:hive2://localhost:10000> **desc Employee;**

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

|  col\_name   | data\_type  | comment  |

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

| id          | bigint     |          |

| name        | string     |          |

| age         | int        |          |

| salary      | bigint     |          |

| department  | string     |          |

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

#Data in the table

0: jdbc:hive2://localhost:10000> **select \* from Employee;**

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

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

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

| 60001            | Sudip              | 34                | 62000                | HR                       |

| 70001            | Suresh             | 45                | 76000                | FINANCE                  |

| 80001            | Aarti              | 25                | 37000                | BIGDATA                  |

| 80002            | Neha               | 27                | 39000                | FINANCE                  |

| 80003            | Rajesh             | 29                | 59000                | BIGDATA                  |

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

Group By

Group by is used to group the rows based on the values of some columns and then apply aggregation functions per group on all the records.

While using a Group by clause, the columns in the Select should meet the following conditions.

1. Column can be directly used in Select if it is part of the Group by clause.
2. Else, there should be some aggregation used on the column.

Some basic aggregations available to be used in the Select columns while using a Group by clause are:-

Count

Min

Max

Sum

Avg

To get number of employees per department, we can use Group by clause as shown below :-

select department, count(\*) as num\_employees from Employee group by department;

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

| department  | num\_employees  |

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

| BIGDATA     | 2              |

| FINANCE     | 2              |

| HR          | 1              |

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

To get the maximum Salary per department, we can use Group by clause as shown below :-

select department, max(salary) as max\_salary from Employee group by department;

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

| department  | max\_salary  |

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

| BIGDATA     | 59000       |

| FINANCE     | 76000       |

| HR          | 62000       |

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

To get the average age of employees per department, we can use Group by clause as shown below:-

select department, avg(age) as avg\_age from Employee group by department;

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

| department  | avg\_age  |

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

| BIGDATA     | 27.0     |

| FINANCE     | 36.0     |

| HR          | 34.0     |

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

To get the sum of salaries per department, we can use Group by clause as shown below:-

select department, sum(salary) as sum\_salary from Employee group by department;

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

| department  | sum\_salary  |

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

| BIGDATA     | 96000       |

| FINANCE     | 115000      |

| HR          | 62000       |

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