**Assignment 7.3**

**Problem Statement:**

**● Hive Data Definitions**

**● Hive Data Manipulations**

**● HiveQL Manipulations**

**Solution**

**● Hive Data Definitions**

Hive Data Definition Language (DDL) is a subset of Hive SQL statements that describe the data structure in Hive by creating, deleting, or altering schema objects such as databases, tables, views, partitions, and buckets. Most Hive DDL statements start with the keywords CREATE, DROP, or ALTER. The syntax of Hive DDL is very similar to the DDL in SQL. The comments in Hive start from --

DDL Commands in Hive

1. CREATE Database,Table
2. DROP Database,Table
3. TRUNCATE Table
4. ALTER Database,Table
5. SHOW Databases,Tables,Table Properties,Partitions,Functions,Index
6. DESCRIBE Database, Table ,View

1. Create commands

**Create Database in Hive**

This DDL command in Hive is used for creating databases.

CREATE (DATABASE) [IF NOT EXISTS] database\_name

[COMMENT database\_comment]

[LOCATION hdfs\_path]

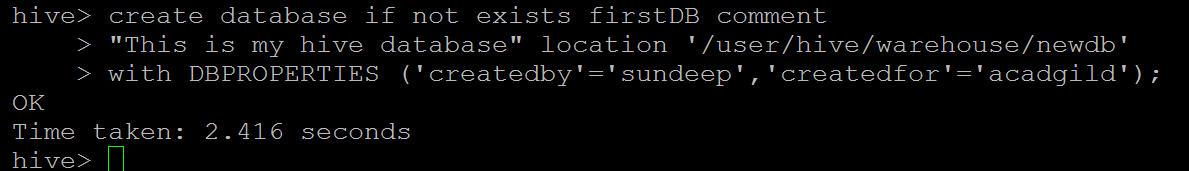
[WITH DBPROPERTIES (property\_name=property\_value, ...)];

for example

create database if not exists firstDB comment

"This is my hive database" location '/user/hive/warehouse/newdb'

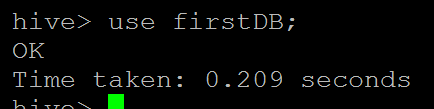
with DBPROPERTIES ('createdby'='sundeep','createdfor'='acadgild');



**Use Database Command in Hive**

This hive command is used to select a specific database for the session on which hive queries would be executed.

use firstDB;



#### ****Create Table Command in Hive****

Hive create table command is used to create a table in the existing database that is in use for a particular session.

CREATE TABLE [IF NOT EXISTS] [db\_name.]table\_name --

[(col\_name data\_type [COMMENT col\_comment], ...)]

[COMMENT table\_comment]

[LOCATION hdfs\_path]

for example

create table employee(

name string,

skill string,

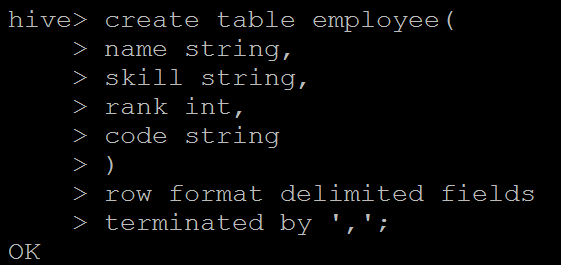
rank int,

code string

)

row format delimited fields terminated by ',';

It will create the table employee with column name



**Create a table in hive by copying an existing table schema**

Hive lets programmers create a new table by replicating the schema of an existing table but remember only the schema of the new table is replicated but not the data. When creating the new table, the location parameter can be specified.

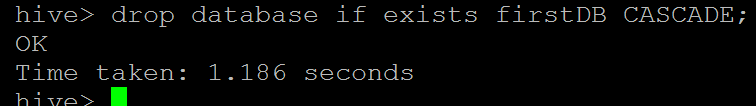
CREATE TABLE [IF NOT EXISTS] [db\_name.]table\_name Like [db\_name].existing\_table [LOCATION hdfs\_path]

2. Drop commands

**Drop Database in Hive**

This command is used for deleting an already created database in Hive and the syntax is as follows -

DROP (DATABASE) [IF EXISTS] database\_name [RESTRICT|CASCADE];

****

**DROP Table Command in Hive**

Drops the table and all the data associated with it in the Hive metastore.

DROP TABLE [IF EXISTS] table\_name [PURGE];

3. Truncate commands

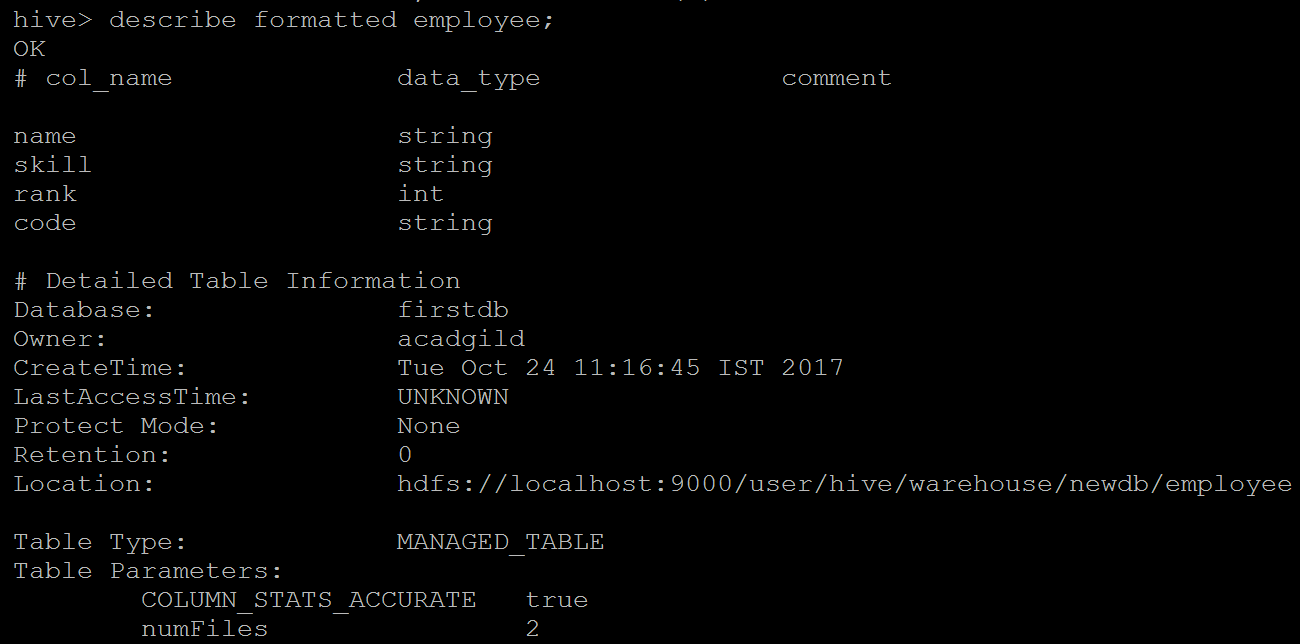
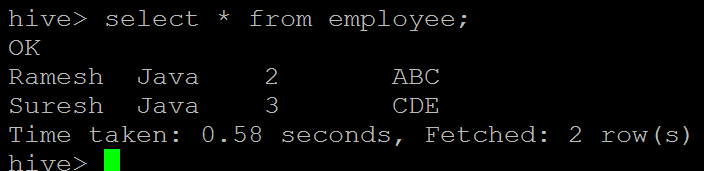
**TRUNCATE Table Command in Hive**

This hive command is used to truncate all the rows present in a table i.e. it deletes all the data from the Hive meta store and the data cannot be restored.

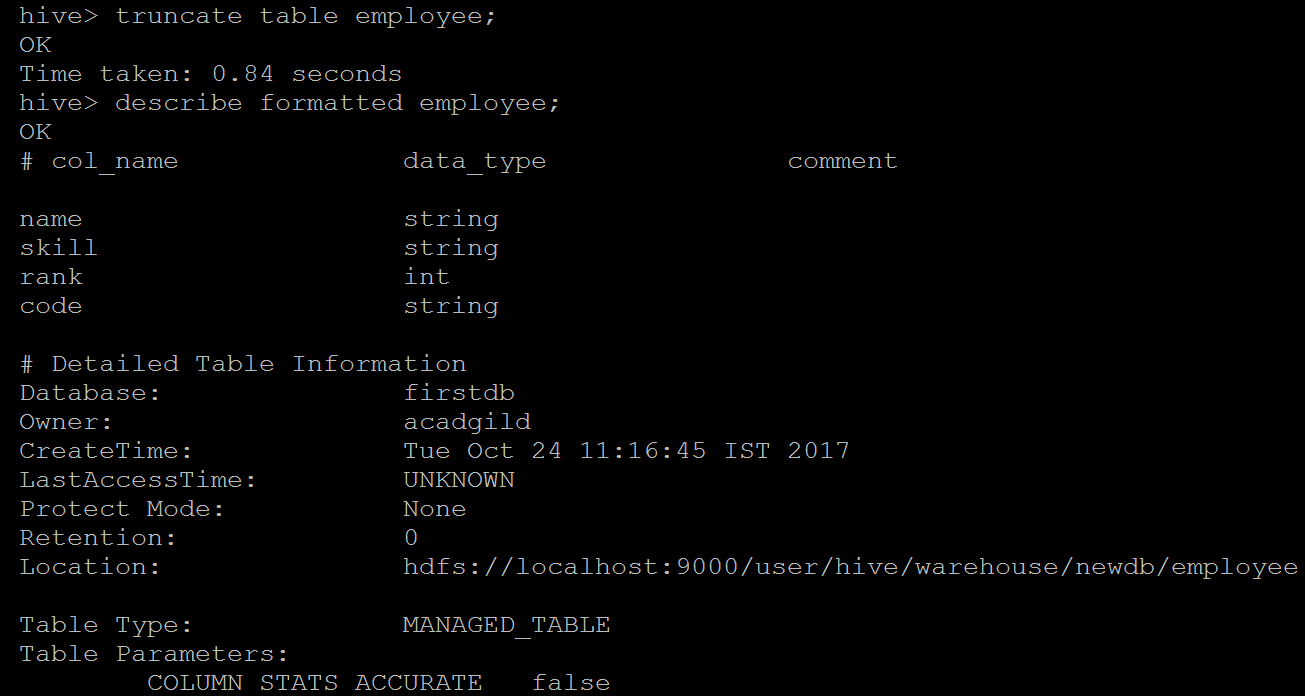
TRUNCATE TABLE [db\_name].table\_name

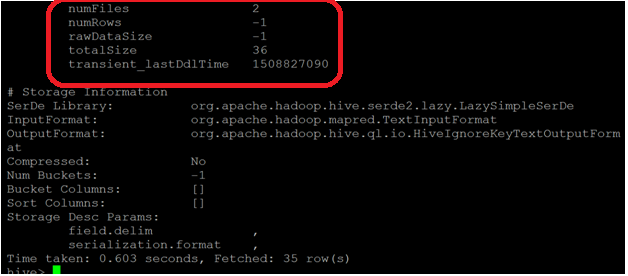
Usage of TRUNCATE Table in Hive

truncate table employee;









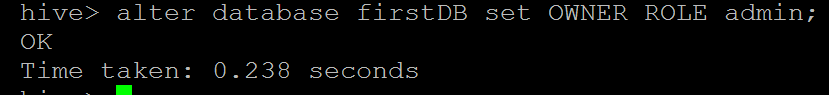
4. Alter commands

**Alter Database Command in Hive**

Whenever the developers need to change the metadata of any of the databases, alter hive DDL command can be used as follows –

ALTER (DATABASE) database\_name SET DBPROPERTIES (property\_name=property\_value, ...);

alter database firstDB set OWNER ROLE admin;

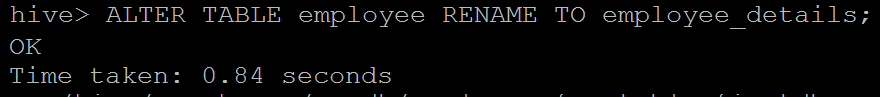
****

**ALTER Table Command in Hive**

Using ALTER Table command, the structure and metadata of the table can be modified even after the table has been created. Let’s try to change the name of an existing table using the ALTER command –

ALTER TABLE [db\_name].old\_table\_name RENAME TO [db\_name].new\_table\_name;

ALTER TABLE employee RENAME TO employee\_details;

****

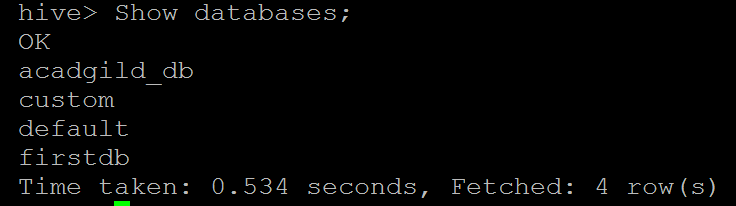
4. Show commands

**Show Database Command in Hive**

Programmers can view the list of existing databases in the current schema.

Usage of Show Database Command

Show databases;

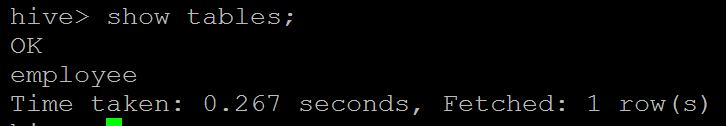


**Show Table Command in Hive**

Gives the list of existing tables in the current database schema.

Usage of Show tables Command

Show tables;



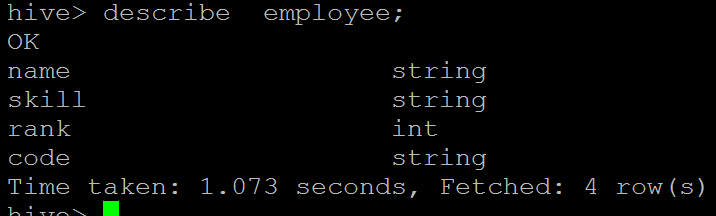
5. Describe commands

**DESCRIBE Table Command in Hive**

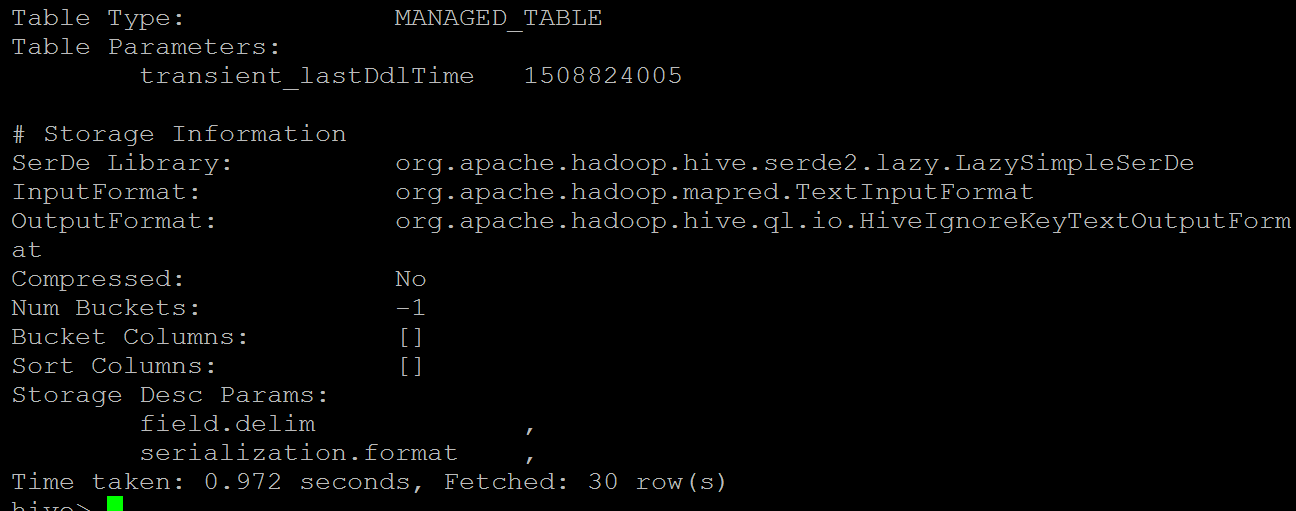
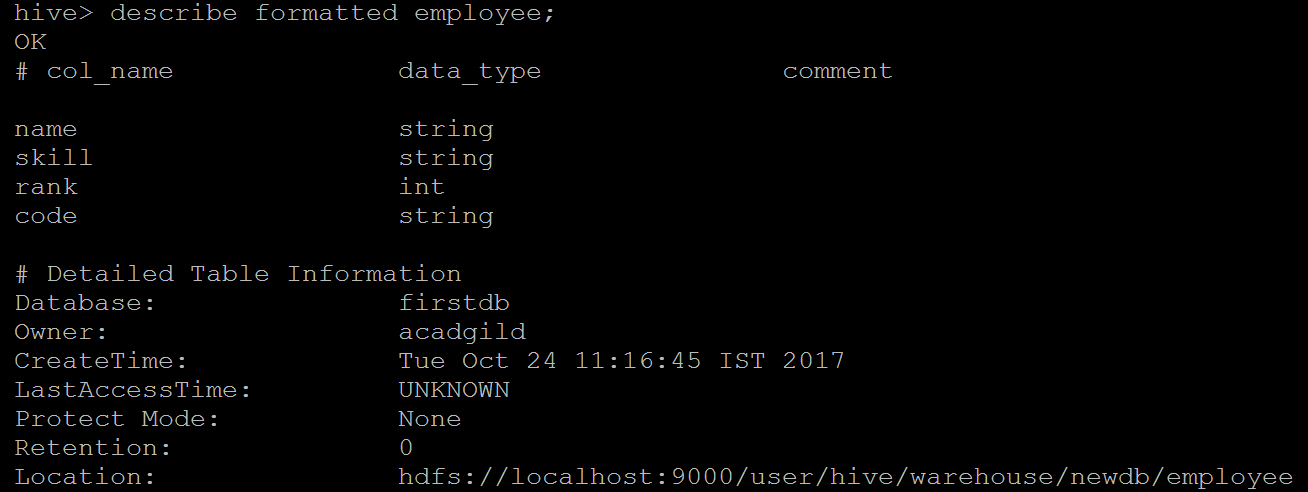
Gives the information of a particular table and the syntax is as follows –

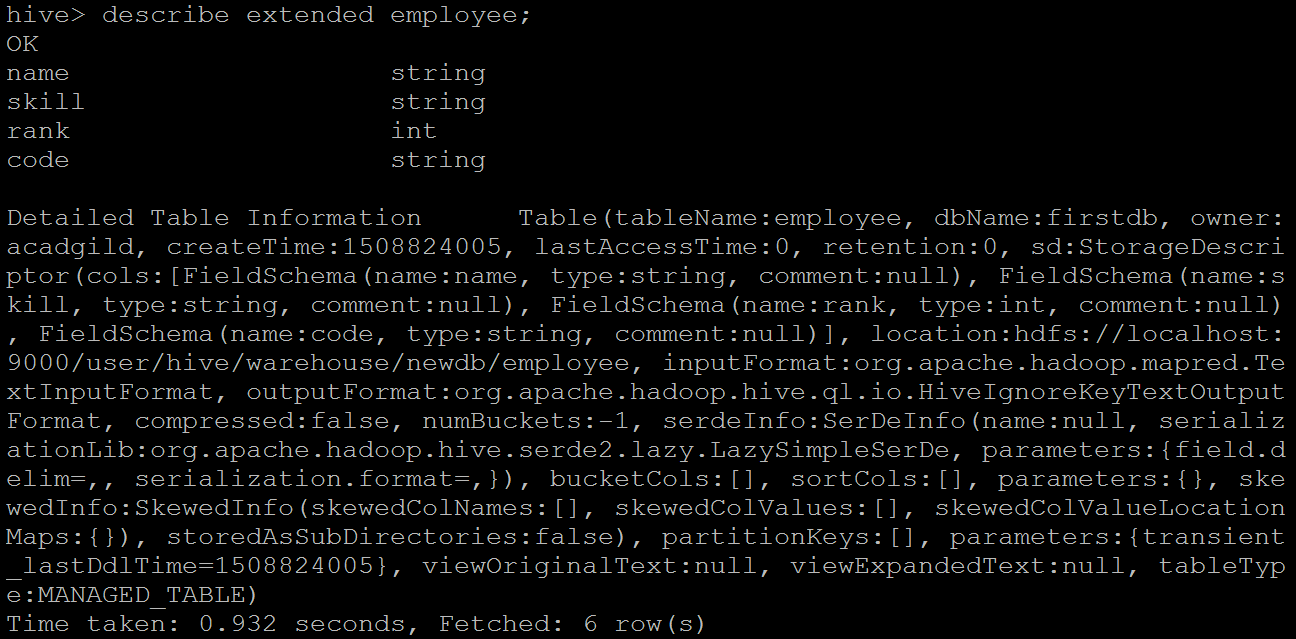
DESCRIBE [EXTENDED|FORMATTED] [db\_name.] table\_name[.col\_name ( [.field\_name]

describe employee



describe formatted employee;





**● Hive Data Manipulations**

DML (Data Manipulation Language) commands in Hive are used for inserting and querying the data from hive tables once the structure and architecture of the database has been defined using the DDL.

Data can be loaded into Hive tables using –

* LOAD command
* Insert command

**Usage of LOAD Command for Inserting Data Into Hive Tables**

Syntax for Load Command in Hive

LOAD DATA [LOCAL] INPATH 'hdfsfilepath/localfilepath' [OVERWRITE] INTO TABLE existing\_table\_name

sample dataset 'emp\_details.txt'

Amit,Big Data,1,BBSR

Venkat,Web Technology,2,BBSR

Aditya,DBA,1,BNG

Ravinder,Java,2,BBSR

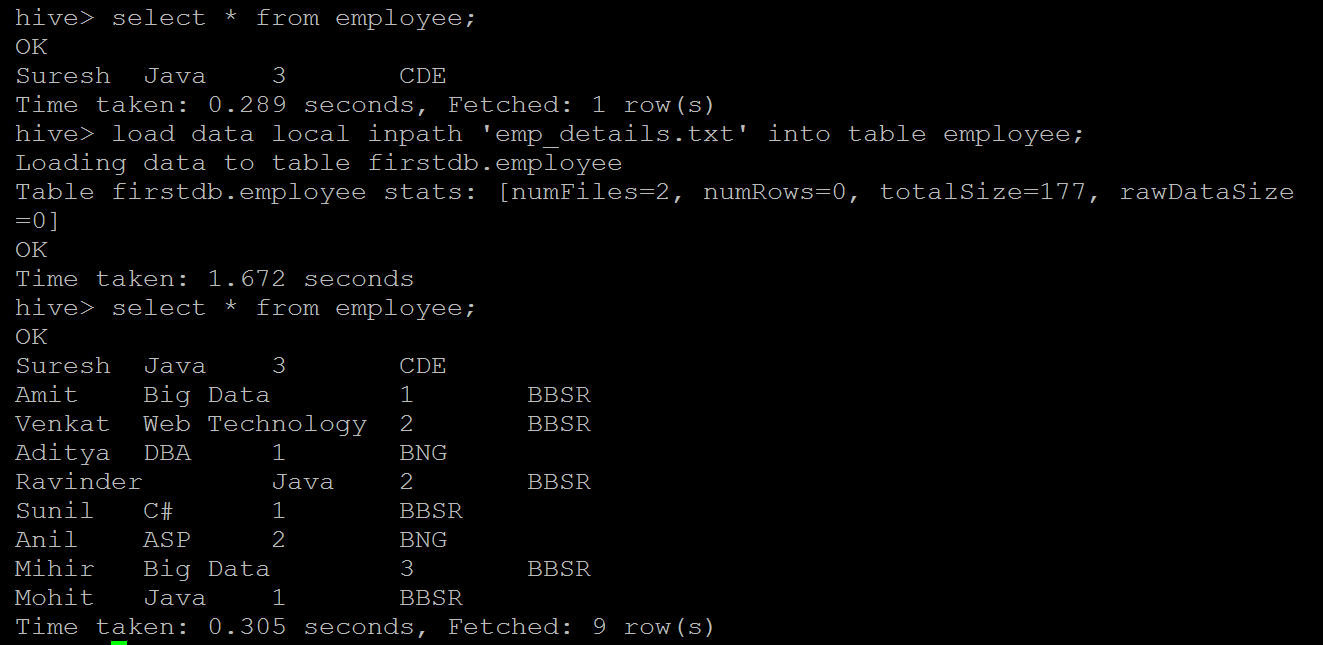
Sunil,C#,1,BBSR

Anil,ASP,2,BNG

Mihir,Big Data,3,BBSR

Mohit,Java,1,BBSR

load data local inpath 'emp\_details.txt' into table employee;



If the keyword LOCAL is not specified, then Hive will need absolute URI of the file. However, if local is specified then it assumes the following rules -

1. It will assume it’s an HDFS path and will try to search for the file in HDFS.
2. If the path is not absolute, then hive will try to locate the file in the /user/ in HDFS.

Using the OVERWRITE keyword while importing means the data will be ingested i.e. it will delete old data and put new data otherwise it would just append the new data. The contents of the target table will be deleted and replaced by the files referred to by file path; otherwise the files referred by file path will be added to the table.

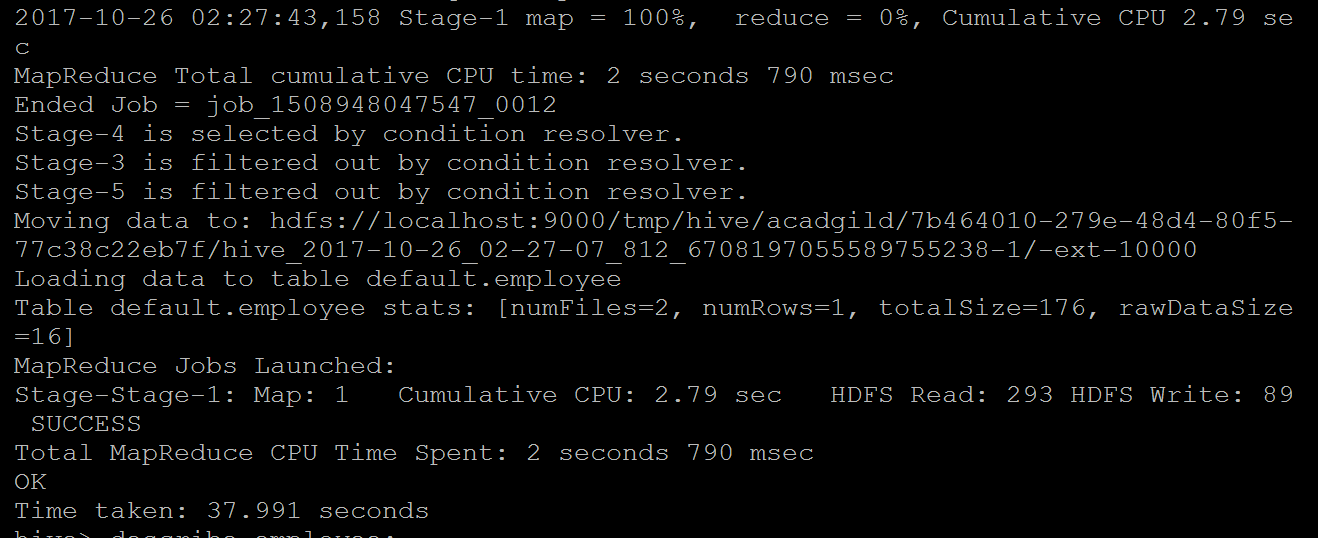
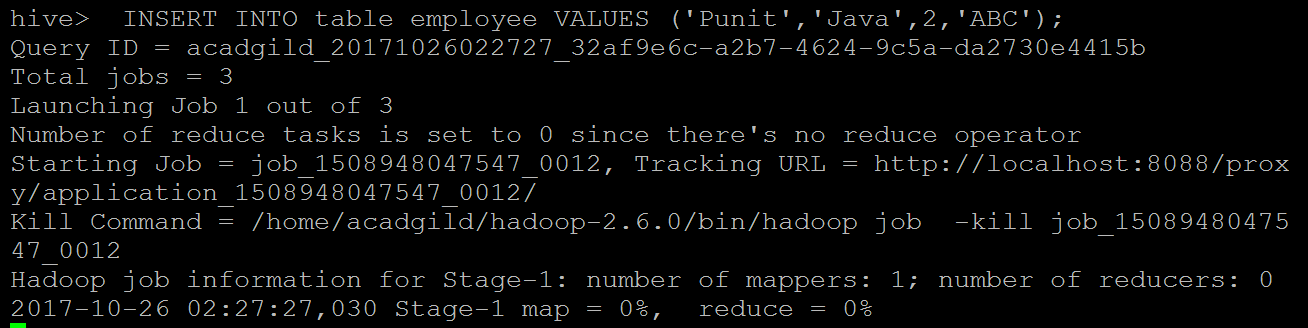
**Usage of INSERT Command for Inserting Data Into Hive Tables**

The INSERT statement lets you load data into a table from a query.

**1.Using Values**

Using Values command ,we can append more rows of data into   existing table.

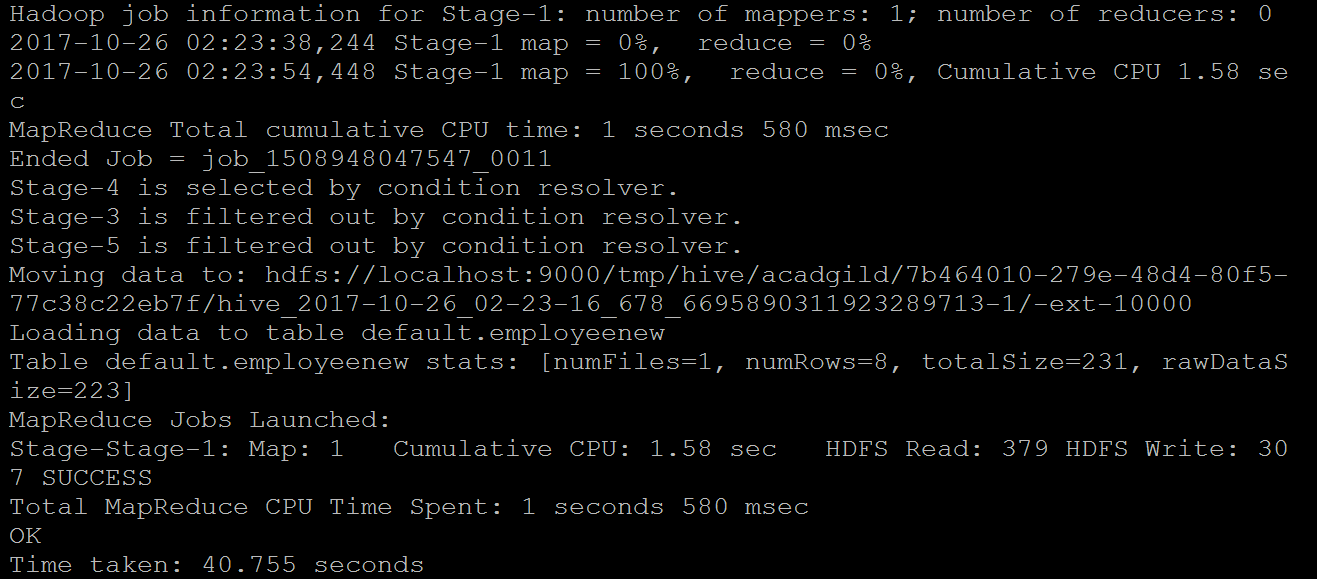
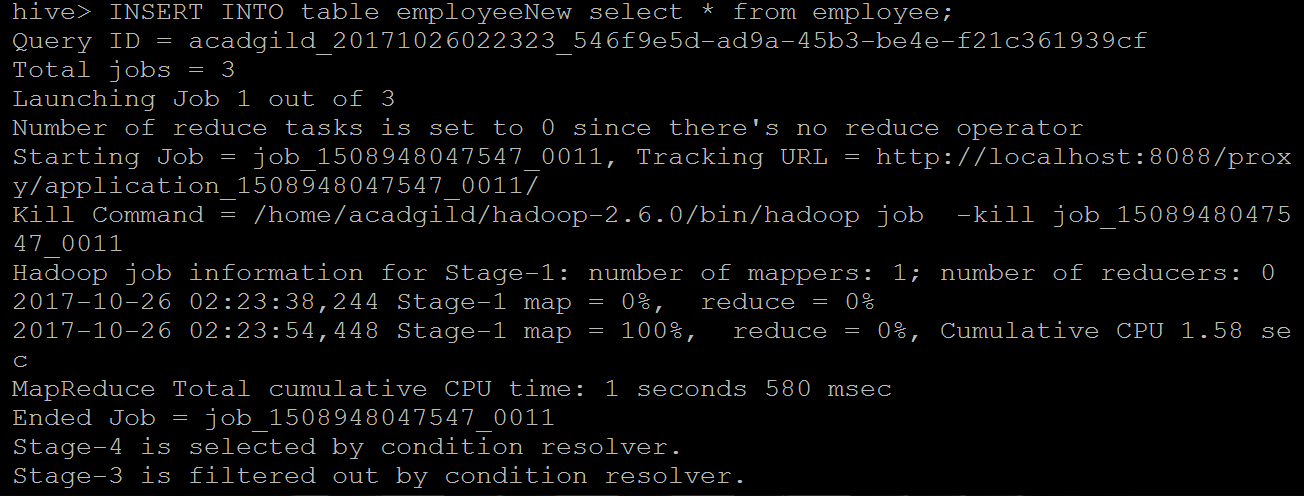
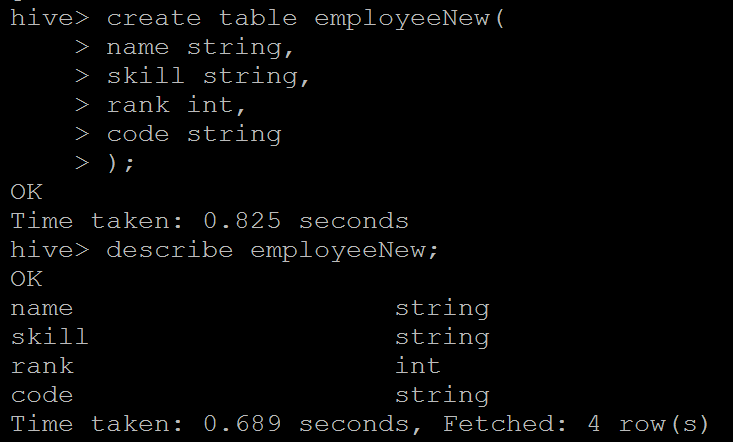
INSERT INTO table employee VALUES ('Punit','Java',2,'ABC');



**2 Using Queries**

You can also insert data using insert with select query output into a table.

INSERT INTO table employeeNew select \* from employee;



**● HiveQL Manipulations**

Hive provides SQL type querying language for the ETL purpose on top of Hadoop file system.

Hive Query language (HiveQL) provides SQL type environment in Hive to work with tables, databases, queries.

We can have a different type of Clauses associated with Hive to perform different type data manipulations and querying. For better connectivity with different nodes outside the environment. HIVE provide JDBC connectivity as well.

Syntax

**SELECT** [ALL | DISTINCT] select\_expr, select\_expr, ...

FROM table\_reference

[WHERE where\_condition]

[**GROUP BY** col\_list]

[**HAVING** having\_condition]

[**CLUSTER BY** col\_list | [**DISTRIBUTE** BY col\_list] [**SORT BY** col\_list] [**ORDER BY** col\_list]]]

[LIMIT number]

SELECT is the projection operator in SQL. The points are:

SELECT scans the table specified by the FROM clause

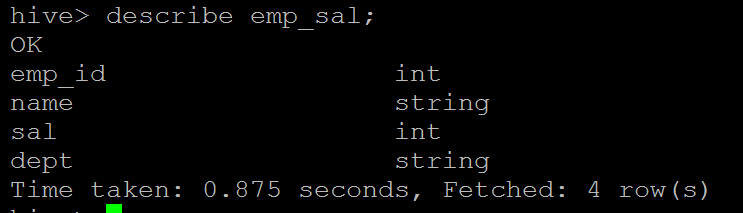
WHERE gives the condition of what to filter

GROUP BY gives a list of columns which specify how to aggregate the records

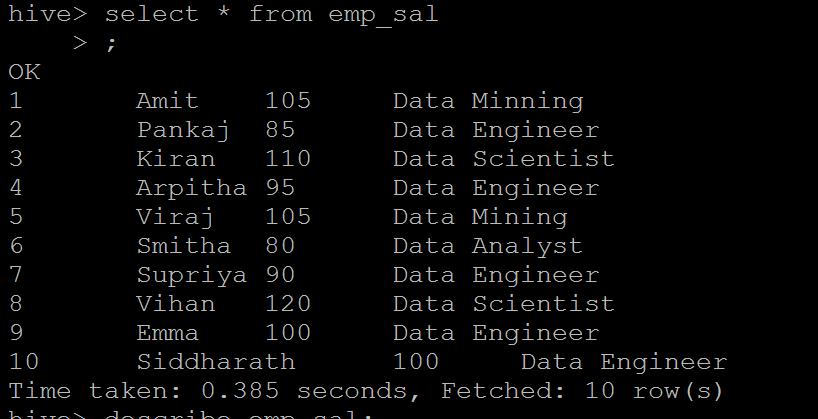
CLUSTER BY, DISTRIBUTE BY, SORT BY specify the sort order and algorithm

LIMIT specifies how many # of records to retrieve

here table used is emp\_sal



simple query to get all column data of table

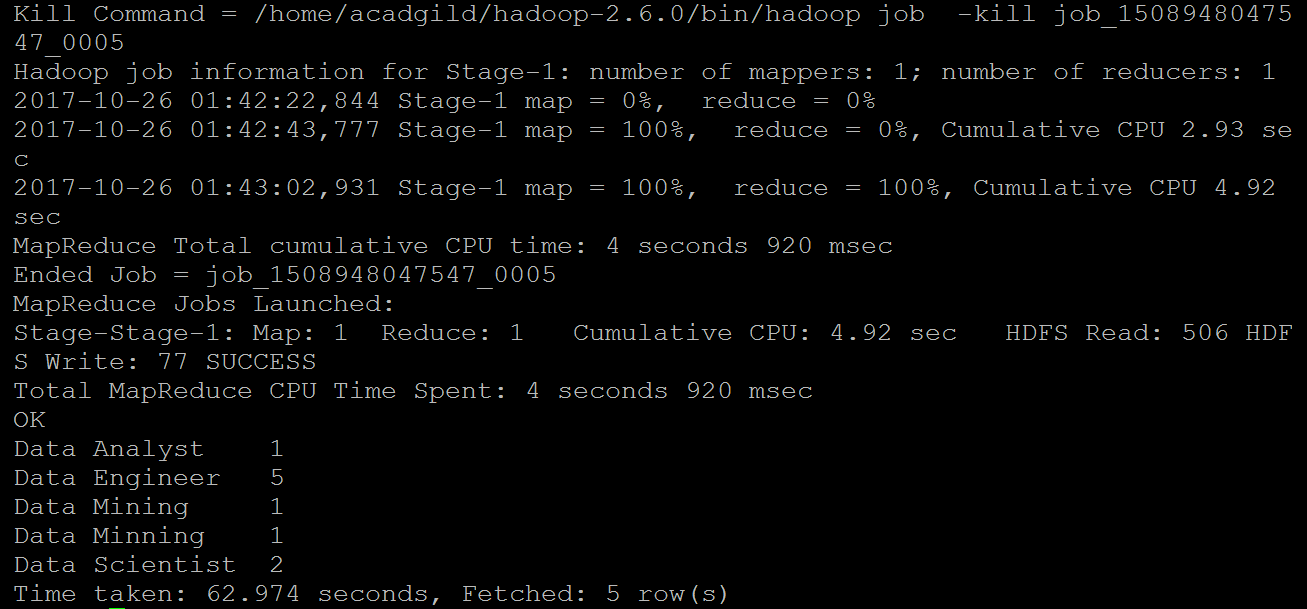
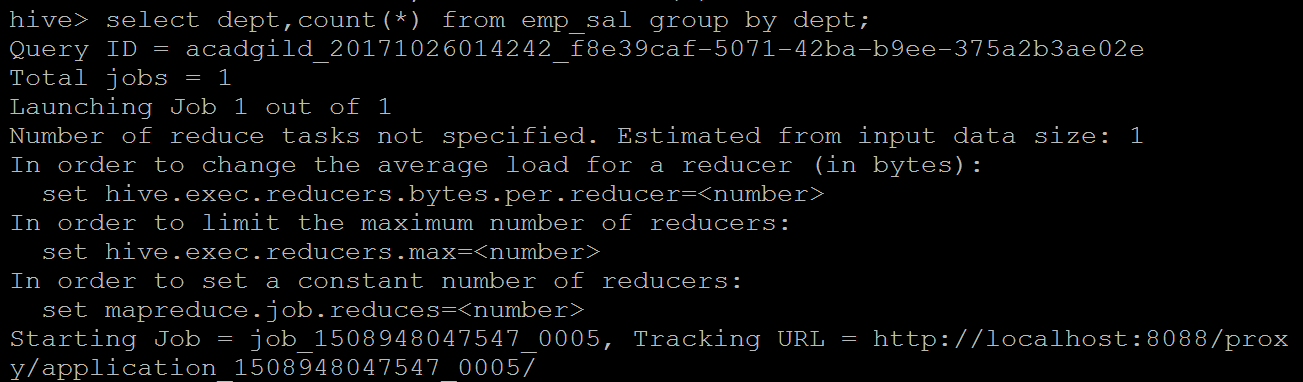


**GROUP BY Clauses**

A GROUP BY clause is frequently used with aggregate functions, to group the result set by columns and apply aggregate functions over each group. Functions link count,max,avg can also be used to compute the grouping key.

To get department wise employee count

select dept,count(\*) from emp\_sal group by dept;

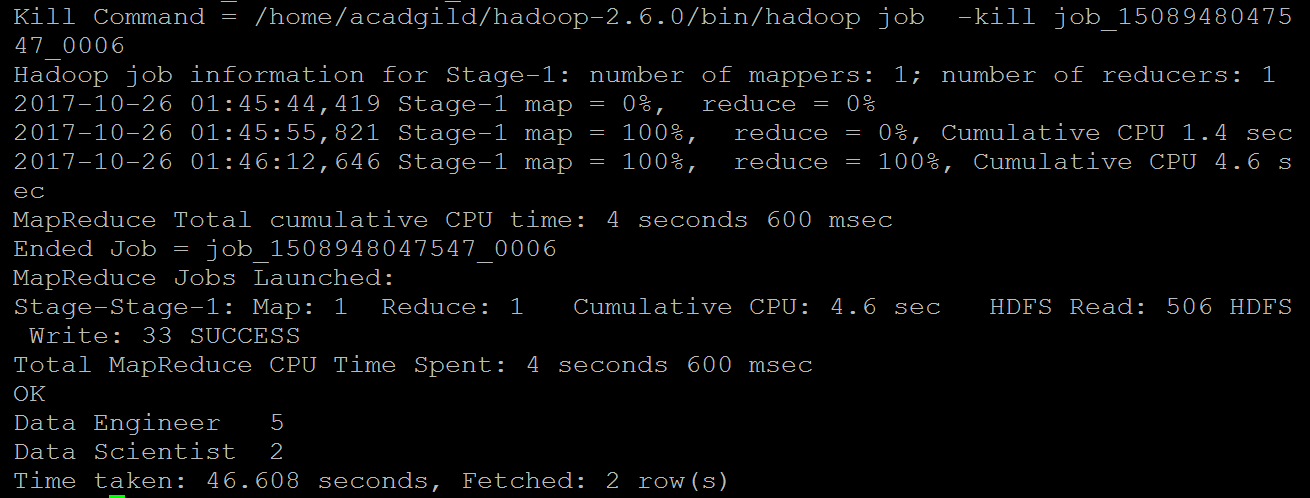
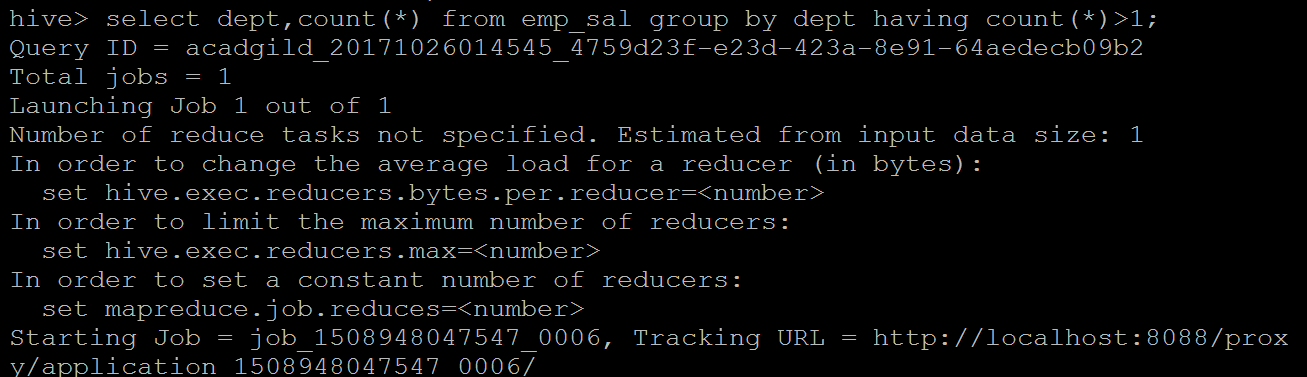


**HAVING Clauses**

A HAVING clause lets you filter the groups produced by GROUP BY, by applying predicate operators to each groups.

To get department with number of employee having number of employee>1

select dept,count(\*) from emp\_sal group by dept having count(\*)>1;

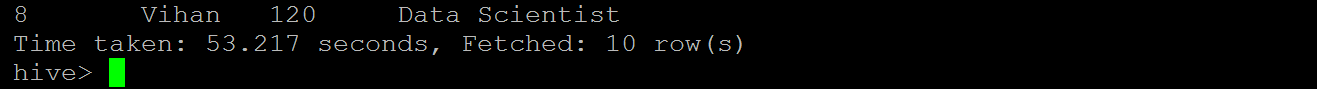
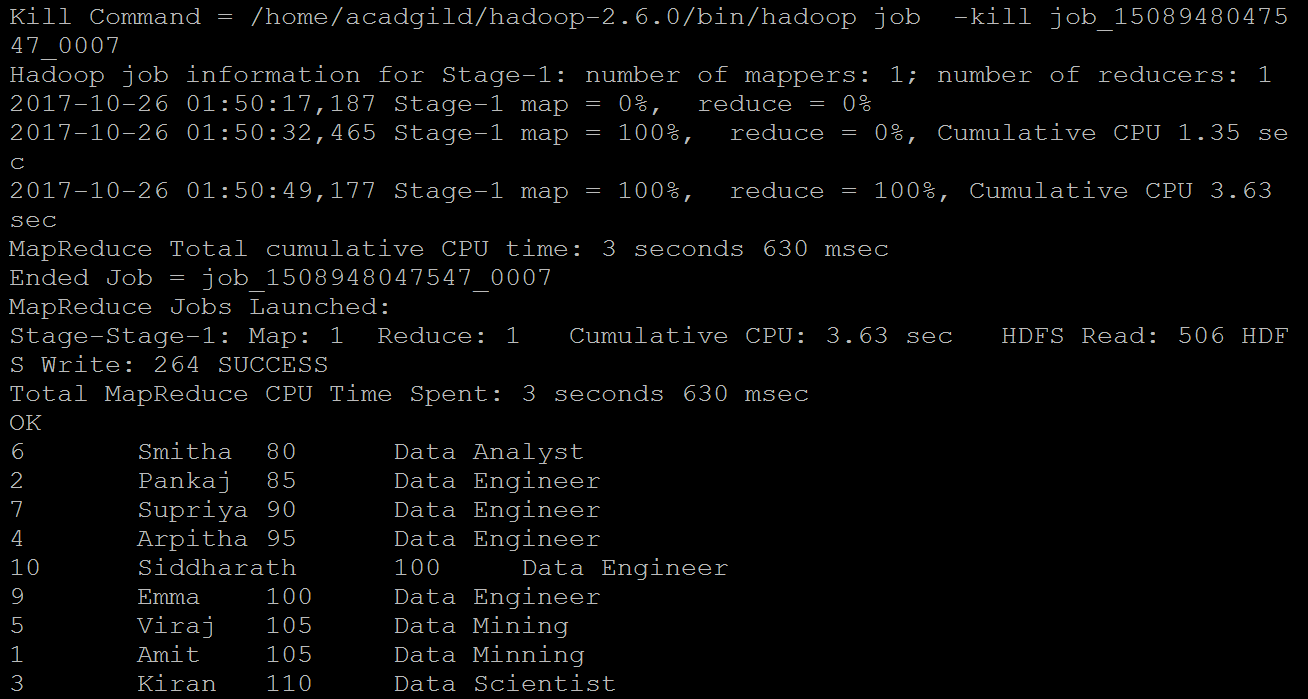
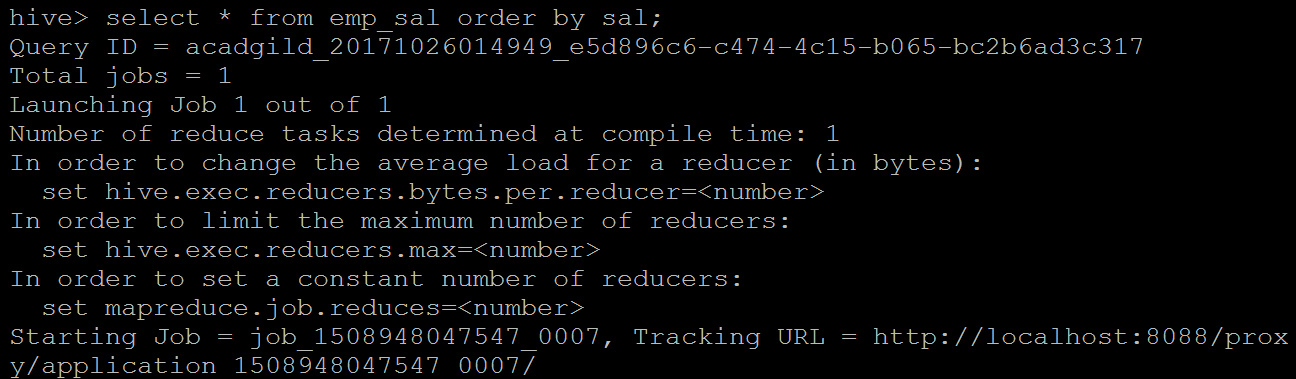


**Order by:**

The ORDER BY syntax in HiveQL is similar to the syntax of ORDER BY in SQL language.

Order by is the clause we use with "SELECT" statement in Hive queries, which guarantees total ordering of data. Order by clause use columns on Hive tables for grouping particular column values mentioned with Order by. For whatever the column name we are defining the order by clause the query will selects and display results by ascending or descending order the particular column values.

select \* from emp\_sal order by sal;

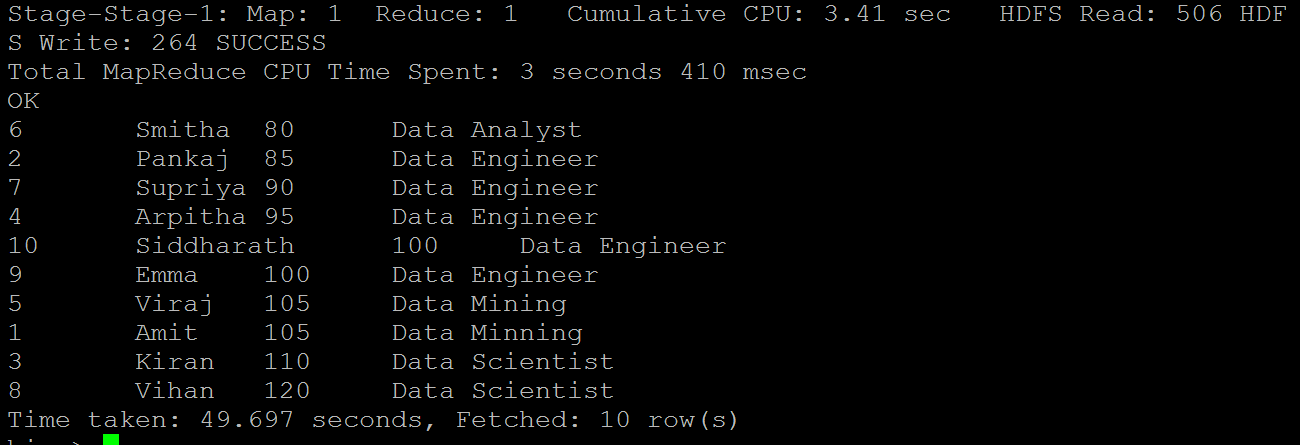
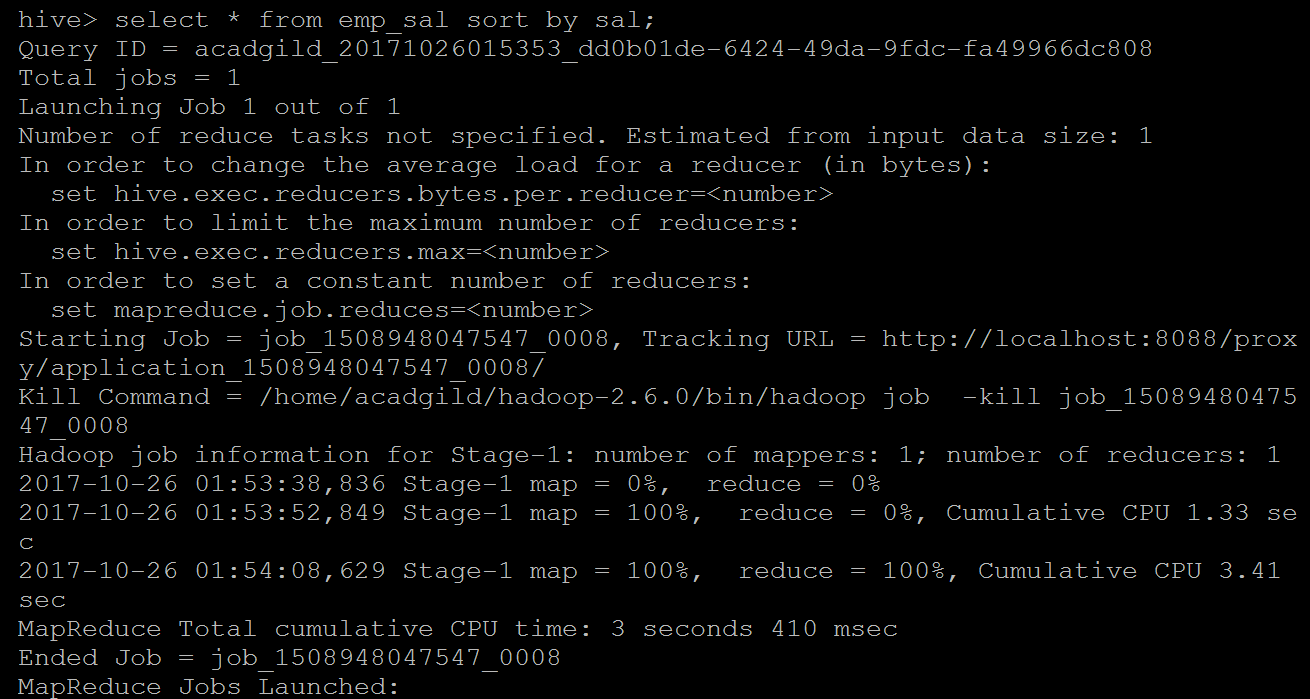


## Sort by:

Sort by clause performs on column names of Hive tables to sort the output. We can mention DESC for sorting the order in descending order and mention ASC for Ascending order of the sort.

In this sort by it will sort the rows before feeding to the reducer. Always sort by depends on column types.

select \* from emp\_sal sort by sal;



**Difference between Sort By and Order By**

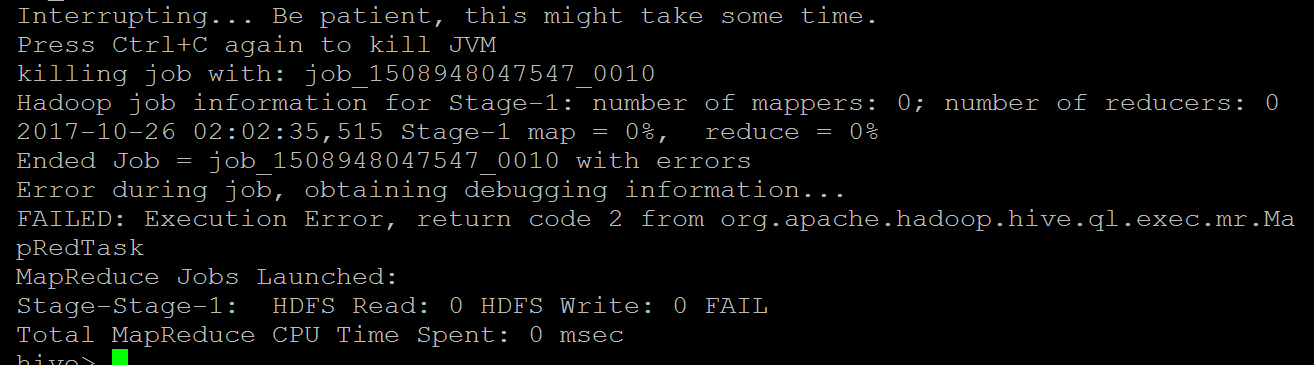
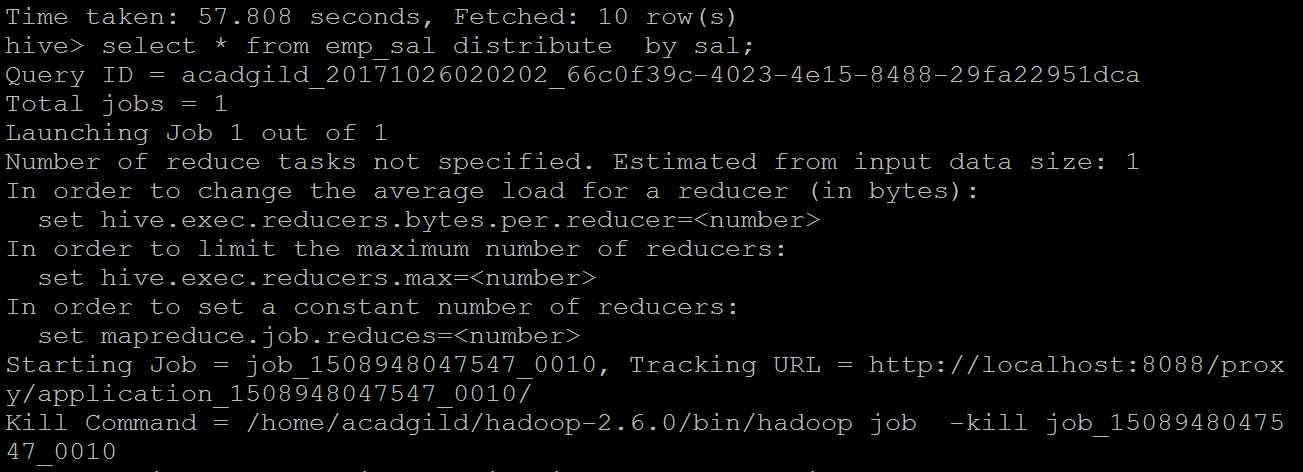
Hive supports SORT BY which sorts the data per reducer. The difference between "order by" and "sort by" is that the former guarantees total order in the output while the latter only guarantees ordering of the rows within a reducer. If there are more than one reducer, "sort by" may give partially ordered final results.

## 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

select \* from emp\_sal distribute by sal;



## Cluster By:

Cluster By used as an alternative for both Distribute BY and Sort BY clauses in Hive-QL.

Cluster BY clause used on tables present in Hive. Hive uses the columns in Cluster by to distribute the rows among reducers. Cluster BY columns will go to the multiple reducers. It ensures sorting orders of values present in multiple reducers

select \* from emp\_sal cluster by sal;

