Big Data And Hadoop

Session 15 – Assignment 1

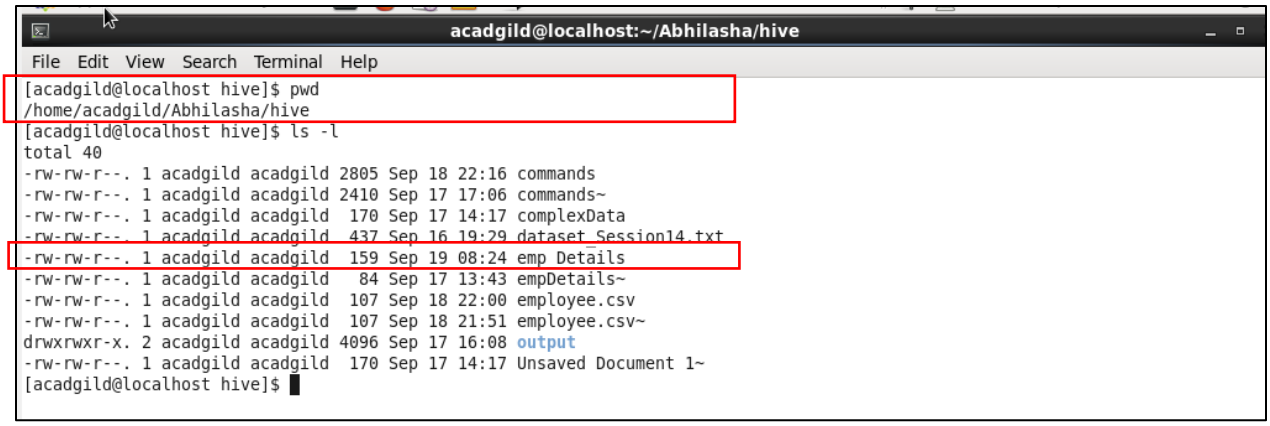
Problem Statement:

Calculate the number of employees corresponding to each skill from the table 'employee' which is loaded in the Demo .

Solution:

Input File

The input file is present on the local file system at **/home/acadgild/Abhilasha/hive** as follows:



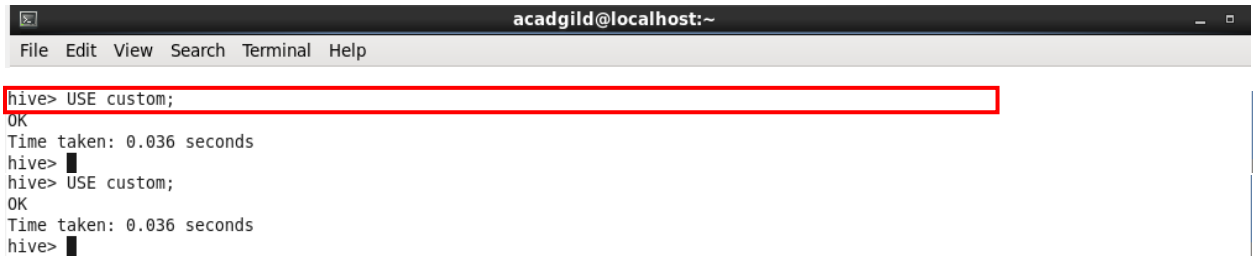
The content of the input file can be seen using the **cat** command as follows:

Start hive: We start the hive command line by executing the command hive as shown below:

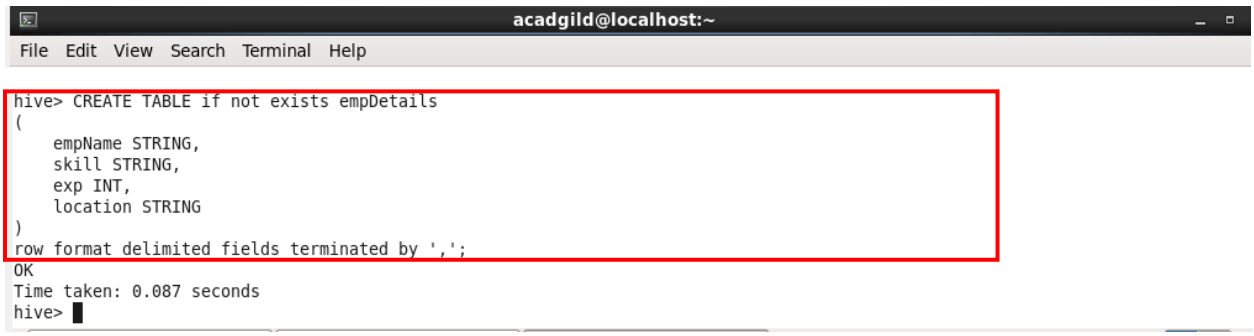
The above snapshot also shows that hive prompt has started. A pre-requisite to use hive is to start mysql server. This was done using the command sudo service mysqld start.

Step 1: We use **SHOW DATABASES** command to list the databases present. The database we will be using is **custom** as shown below:

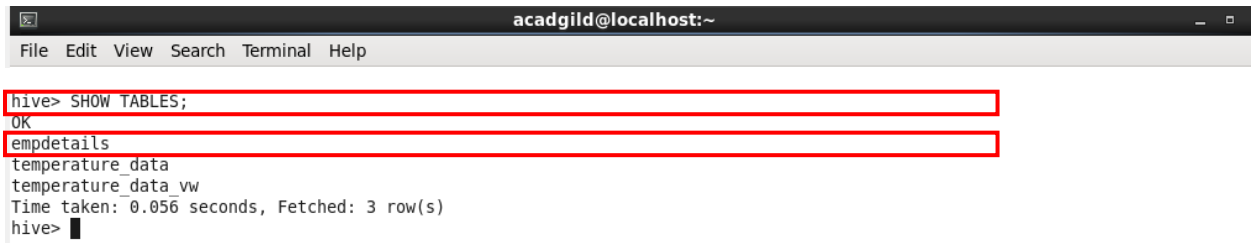
Step 2: We use **USE custom** command to make use of custom database, as shown below:



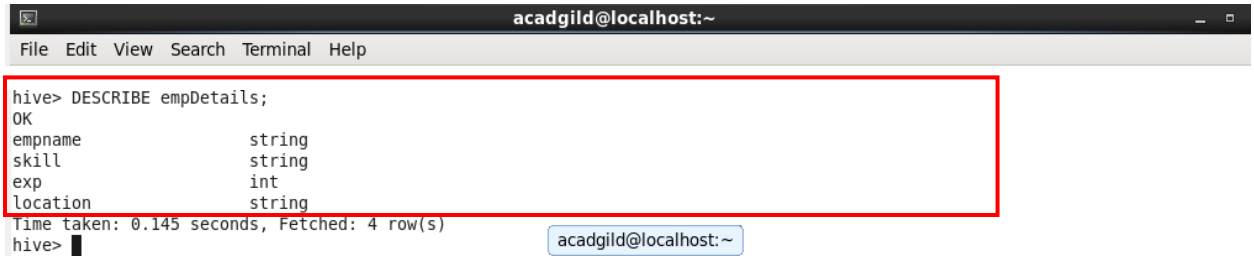
Step 3: We create the table using **CREATE TABLE** command. The fields of the table are: empName, skill, exp and location.



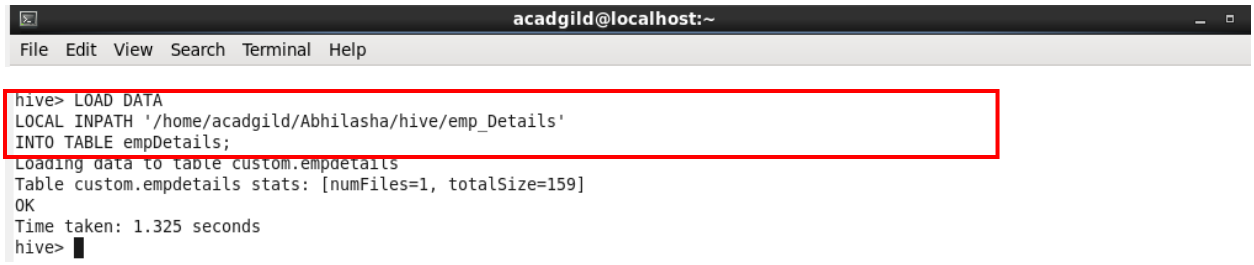
Step 4: **SHOW TABLES** command will help us verify that the table is created.



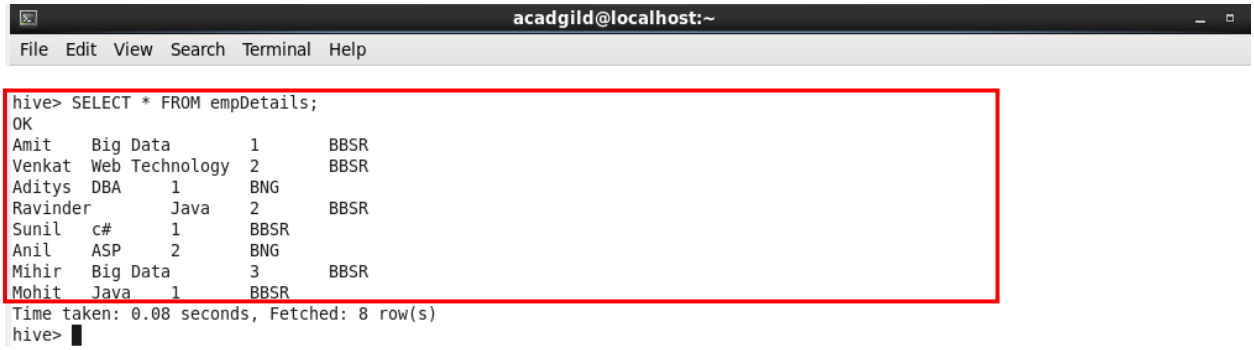
Step 5: **DESCRIBE** command will help us verify the schema of the table as follows:



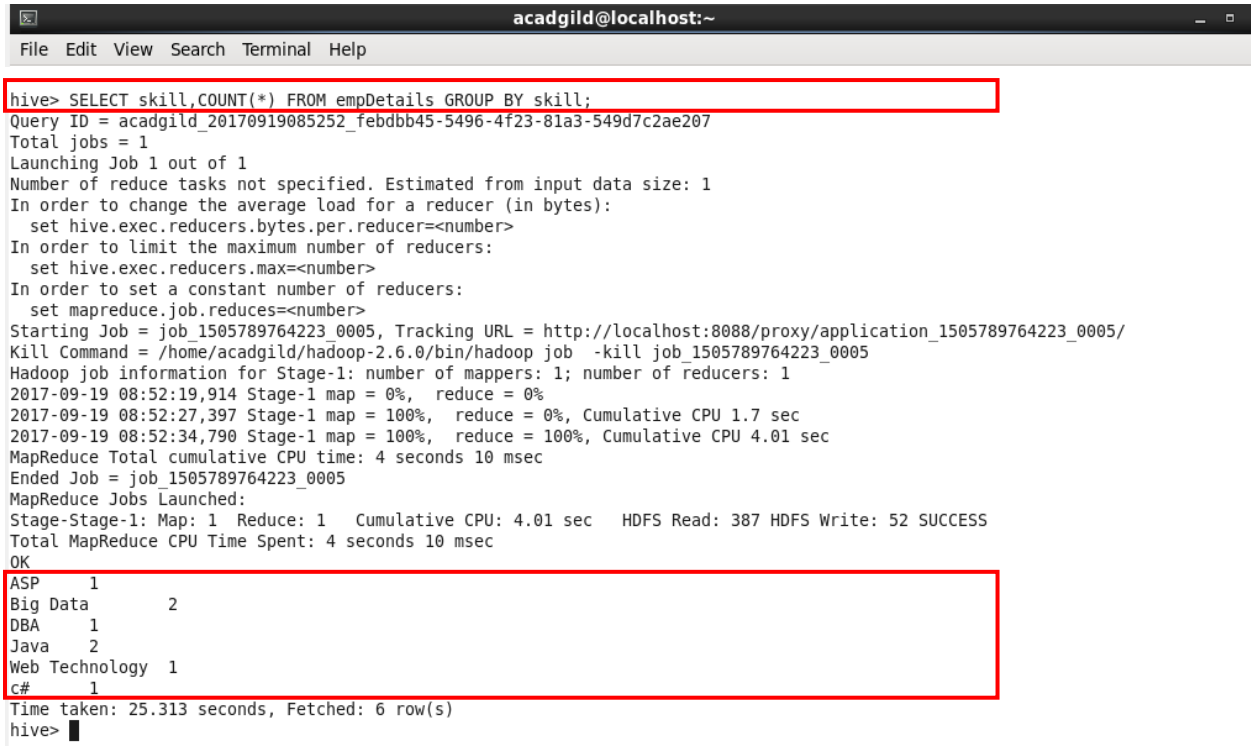
Step 6: Next is to load the data from input file, which is located at **/home/acadgild/Abhilasha/hive** as follows. We use the **LOAD** command and use the keyword **LOCAL** to specify that the file is present in the local file system and not HDFS.



Step 7: Using the **SELECT *** query, we can display the complete data as follows:

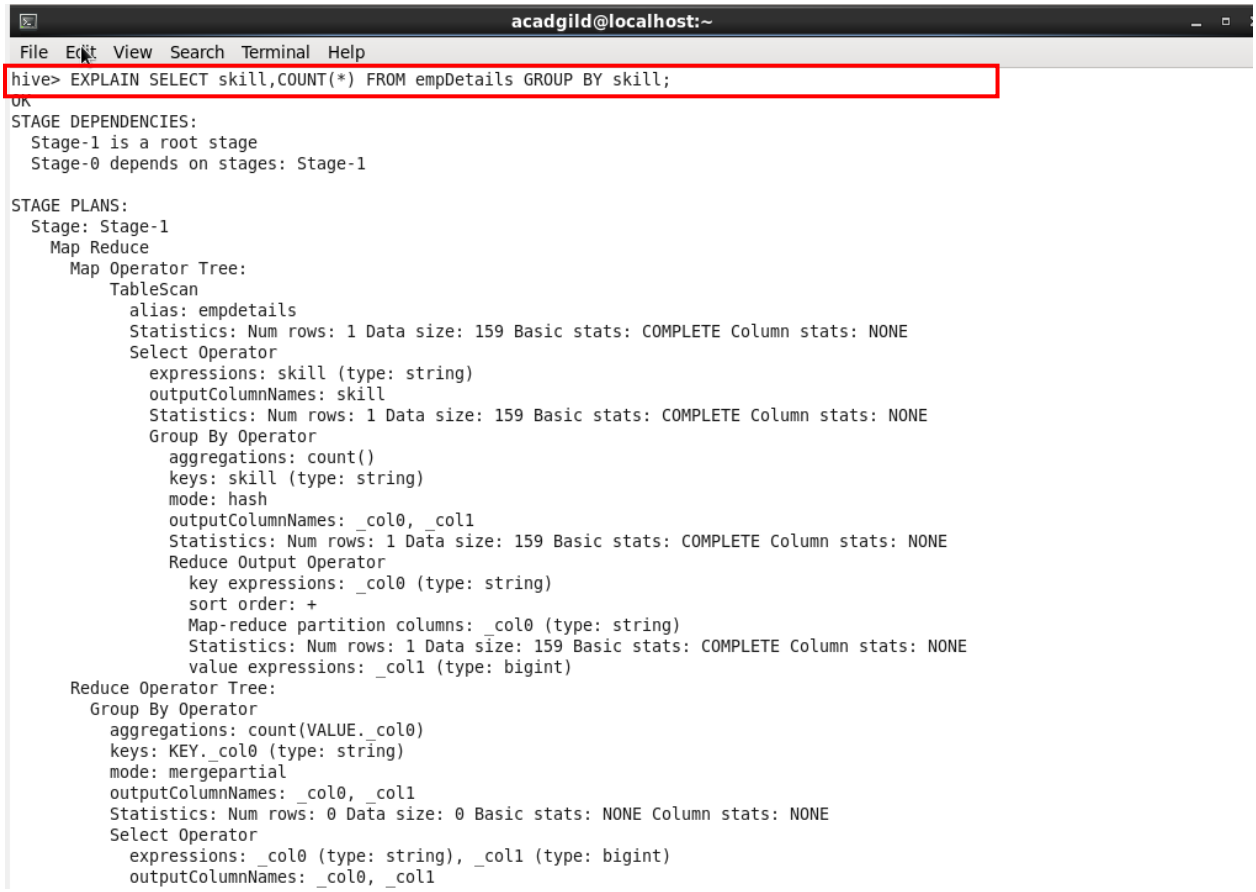


Step 8: In order to calculate the number of employees corresponding to each skill, we perform a **GROUP BY** on the skill column and use **COUNT** function to find the count of employees as follows:



The above screen shot also shows the output of the query.

Step 9: Using **EXPLAIN** command, we can get the plan of execution as follows:



Step 10: We can also store this result into a file using **INSERT** command as follows. The output directory is **/home/acadgild/Abhilasha/hive/output.** The delimiter used to separate the fields in the file is '|'.

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help

hive> insert overwrite local directory '/home/acadgild/Abhilasha/hive/output'
row format delimited
fields terminated by '|'
SELECT skill,COUNT(*) FROM empDetails GROUP BY skill;
Query ID = acadgild_20170919085353_e0708939-5450-4c4a-96b9-f05f7f61d3aa
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1505789764223_0006, Tracking URL = http://localhost:8088/proxy/application_1505789764223_0006/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1505789764223_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-09-19 08:53:21,222 Stage-1 map = 0%,  reduce = 0%
2017-09-19 08:53:28,712 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.76 sec
2017-09-19 08:53:36,119 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.02 sec
MapReduce Total cumulative CPU time: 4 seconds 20 msec
Ended Job = job_1505789764223_0006
Copying data to local directory /home/acadgild/Abhilasha/hive/output
Copying data to local directory /home/acadgild/Abhilasha/hive/output
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.02 sec   HDFS Read: 387 HDFS Write: 52 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 20 msec
OK
Time taken: 27.08 seconds
hive>
```

Step 11: We can see the new directory created named output as follows:

```
acadgild@localhost:~/Abhilasha/hive
File  Edit  View  Search  Terminal  Help

[acadgild@localhost hive]$ pwd
/home/acadgild/Abhilasha/hive
[acadgild@localhost hive]$ ls -l
total 44
-rw-rw-r--. 1 acadgild acadgild 2805 Sep 18 22:16 commands
-rw-rw-r--. 1 acadgild acadgild 2410 Sep 17 17:06 commands~
-rw-rw-r--. 1 acadgild acadgild  170 Sep 17 14:17 complexData
-rw-rw-r--. 1 acadgild acadgild  437 Sep 16 19:29 dataset_Session14.txt
-rw-rw-r--. 1 acadgild acadgild  159 Sep 19 08:49 emp_Details
-rw-rw-r--. 1 acadgild acadgild  159 Sep 19 08:24 emp_Details~
-rw-rw-r--. 1 acadgild acadgild   84 Sep 17 13:43 empDetails~
-rw-rw-r--. 1 acadgild acadgild  107 Sep 18 22:00 employee.csv
-rw-rw-r--. 1 acadgild acadgild  107 Sep 18 21:51 employee.csv~
drwxrwxr-x. 2 acadgild acadgild 4096 Sep 19 08:53 output
-rw-rw-r--. 1 acadgild acadgild  170 Sep 17 14:17 Unsaved Document 1~
[acadgild@localhost hive]$
```

Step 12: The contents of the can be displayed using **CAT** command as follows:

```
acadgild@localhost:~/Abhilasha/hive/output
File  Edit  View  Search  Terminal  Help

[acadgild@localhost hive]$ cd output
[acadgild@localhost output]$ ls -l
total 4
-rw-r--r--. 1 acadgild acadgild 52 Sep 19 08:53 000000_0
[acadgild@localhost output]$ cat 000000_0
ASP|1
Big Data|2
DBA|1
Java|2
Web Technology|1
c#|1
[acadgild@localhost output]$
```