ASSIGNMENT 7.1

Dataset Used:

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

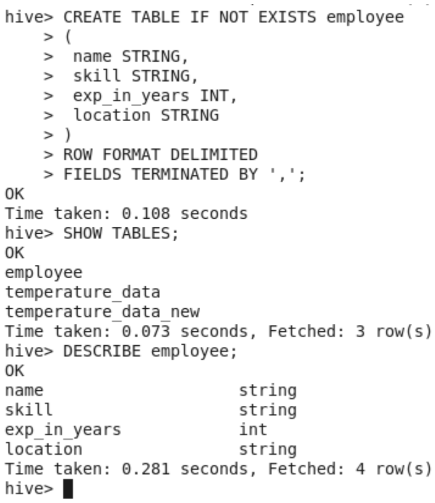
Problem Statement:

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

Solution:

Step 1: Create table with name ‘employee’ using Hive Query Language (HQL) on Hive prompt.

```
CREATE TABLE IF NOT EXISTS employee
(
  name STRING,
  skill STRING,
  exp_in_years INT,
  location STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

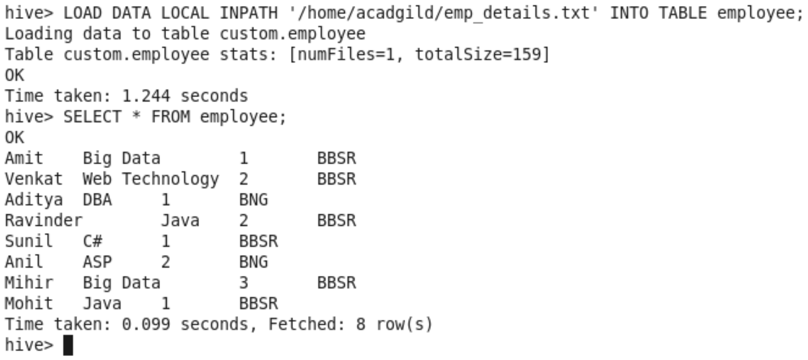


```
hive> CREATE TABLE IF NOT EXISTS employee
    > (
    >  name STRING,
    >  skill STRING,
    >  exp_in_years INT,
    >  location STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.108 seconds
hive> SHOW TABLES;
OK
employee
temperature_data
temperature_data_new
Time taken: 0.073 seconds, Fetched: 3 row(s)
hive> DESCRIBE employee;
OK
name                    string
skill                   string
exp_in_years            int
location                string
Time taken: 0.281 seconds, Fetched: 4 row(s)
hive> 
```

Step 2: Load the data from text file into the table created above.

Data load query:

LOAD DATA LOCAL INPATH '/home/acadgild/emp_details.txt' INTO TABLE employee;



```
hive> LOAD DATA LOCAL INPATH '/home/acadgild/emp_details.txt' INTO TABLE employee;
Loading data to table custom.employee
Table custom.employee stats: [numFiles=1, totalSize=159]
OK
Time taken: 1.244 seconds
hive> SELECT * FROM employee;
OK
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
Time taken: 0.099 seconds, Fetched: 8 row(s)
hive> 
```

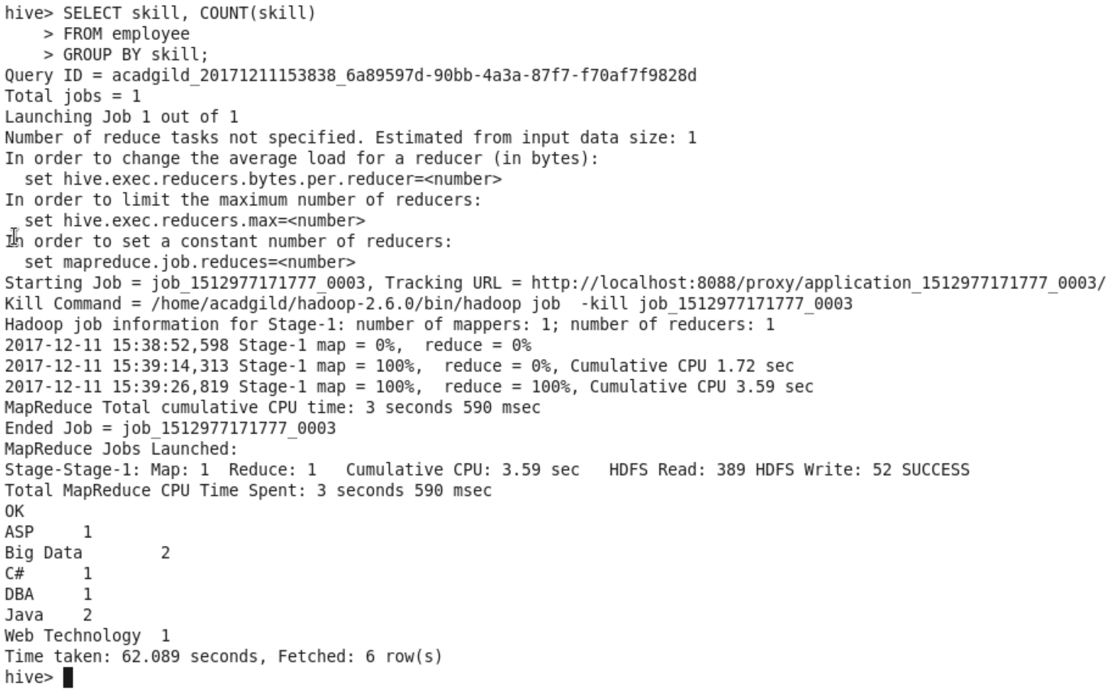
Step 3: Find the number of employees corresponding to each skill from the table 'employee'.

SELECT query:

SELECT skill, COUNT(skill)
FROM employee
GROUP BY skill;

In the above query, GROUP BY clause is used to group employees by their skill and get count of employees in each group.

Output:



```
hive> SELECT skill, COUNT(skill)
    > FROM employee
    > GROUP BY skill;
Query ID = acadgild_20171211153838_6a89597d-90bb-4a3a-87f7-f70af7f9828d
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_1512977171777_0003, Tracking URL = http://localhost:8088/proxy/application_1512977171777_0003/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1512977171777_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-11 15:38:52,598 Stage-1 map = 0%,  reduce = 0%
2017-12-11 15:39:14,313 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.72 sec
2017-12-11 15:39:26,819 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.59 sec
MapReduce Total cumulative CPU time: 3 seconds 590 msec
Ended Job = job_1512977171777_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.59 sec   HDFS Read: 389 HDFS Write: 52 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 590 msec
OK
ASP     1
Big Data        2
C#      1
DBA     1
Java    2
Web Technology  1
Time taken: 62.089 seconds, Fetched: 6 row(s)
hive> 
```