PROBLEM STATEMENT

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

BIG DATA HADOOP & SPARK TRAINING

ACADGILD

ASSIGNMENT 7.1

BY :-

SAHIL KHURANA

Associated Data Files

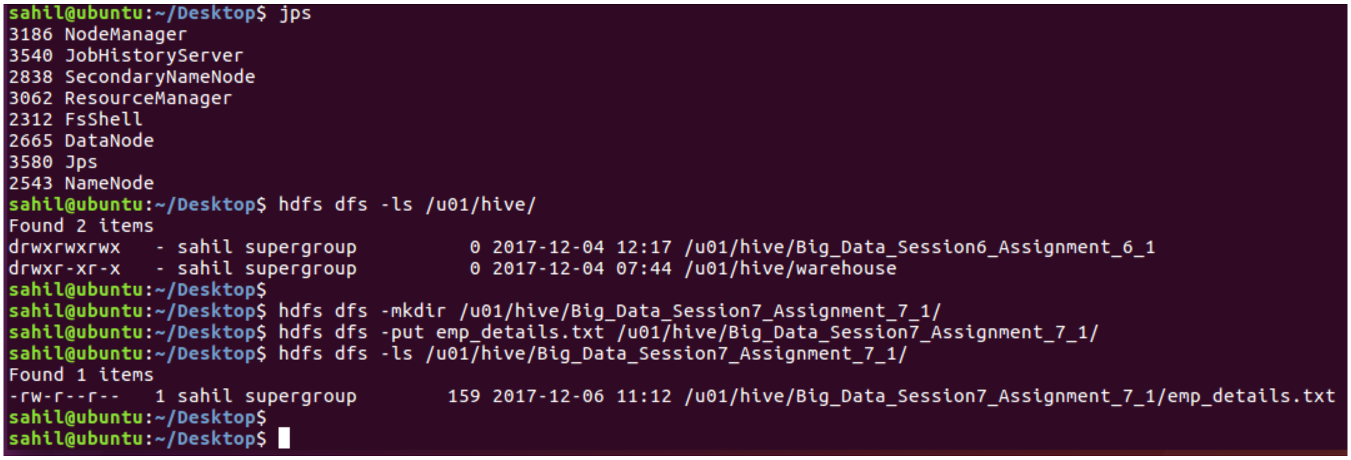
https://drive.google.com/file/d/0Bxr27gVaXO5sQWV4UUpOXzNuZDA/view

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
```

Note: - To solve the Assignment, I have created a VM with Ubuntu 16.04 OS and configured Hadoop 2.8.2 and hive-2.3.2 on the same.

Step 1:- Put the dataset in HDFS location

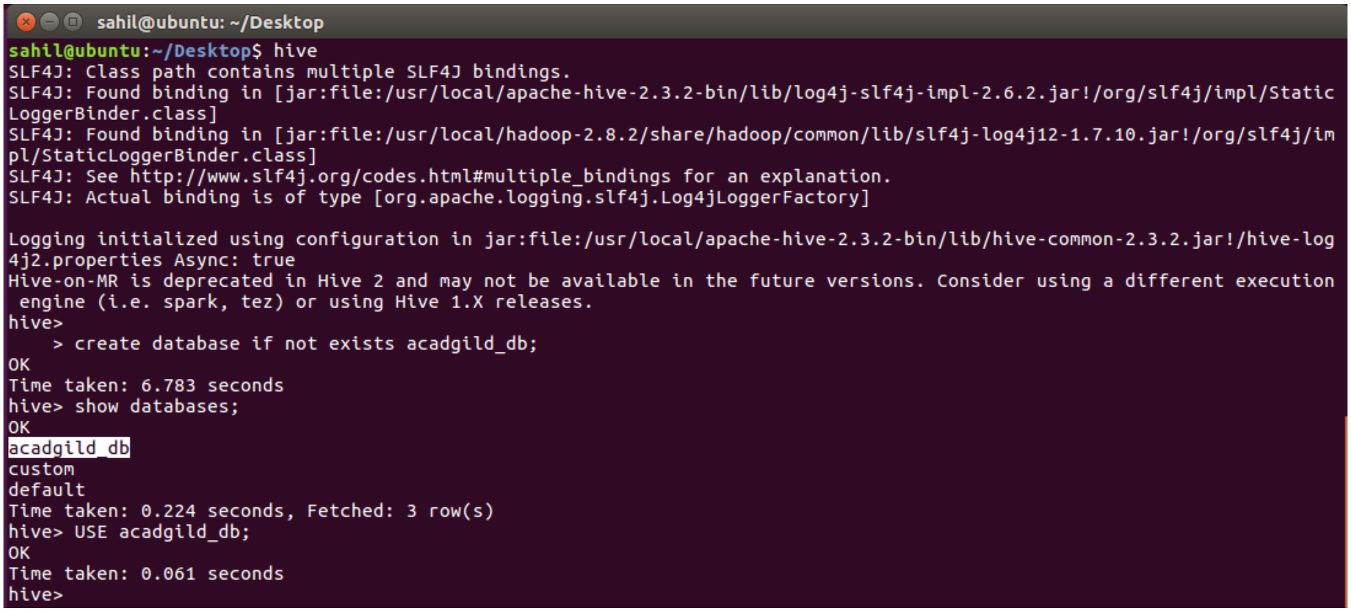


```
sahil@ubuntu:~/Desktop$ jps
3186 NodeManager
3540 JobHistoryServer
2838 SecondaryNameNode
3062 ResourceManager
2312 FsShell
2665 DataNode
3580 Jps
2543 NameNode
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hive/
Found 2 items
drwxrwxrwx   - sahil supergroup          0 2017-12-04 12:17 /u01/hive/Big_Data_Session6_Assignment_6_1
drwxr-xr-x   - sahil supergroup          0 2017-12-04 07:44 /u01/hive/warehouse
sahil@ubuntu:~/Desktop$
sahil@ubuntu:~/Desktop$ hdfs dfs -mkdir /u01/hive/Big_Data_Session7_Assignment_7_1/
sahil@ubuntu:~/Desktop$ hdfs dfs -put emp_details.txt /u01/hive/Big_Data_Session7_Assignment_7_1/
sahil@ubuntu:~/Desktop$ hdfs dfs -ls /u01/hive/Big_Data_Session7_Assignment_7_1/
Found 1 items
-rw-r--r--   1 sahil supergroup        159 2017-12-06 11:12 /u01/hive/Big_Data_Session7_Assignment_7_1/emp_details.txt
sahil@ubuntu:~/Desktop$
sahil@ubuntu:~/Desktop$
```

Step 2:- Open the Hive Shell and CREATE the DATABASE.

Commands used in Step 2

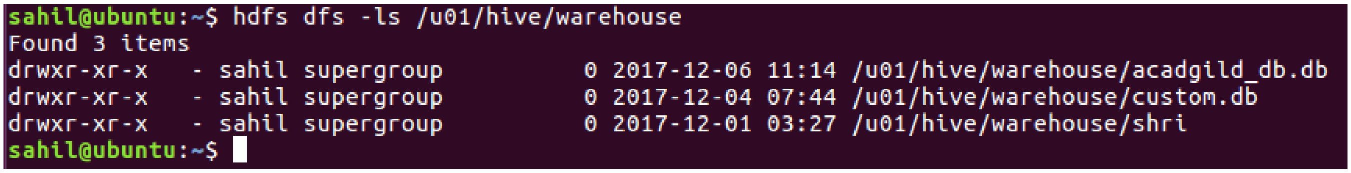
```
hive                                                  -- open the hive shell
hive> create database if not exists acadgild_db;      -- create database
hive> show databases;                                 -- check whether database is created or not
hive> USE acadgild_db;                                -- use database is USE
```



```
sahil@ubuntu:~/Desktop$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.8.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log
4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution
 engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
    > create database if not exists acadgild_db;
OK
Time taken: 6.783 seconds
hive> show databases;
OK
acadgild_db
custom
default
Time taken: 0.224 seconds, Fetched: 3 row(s)
hive> USE acadgild_db;
OK
Time taken: 0.061 seconds
hive>
```

Custom database created in default directory in hive

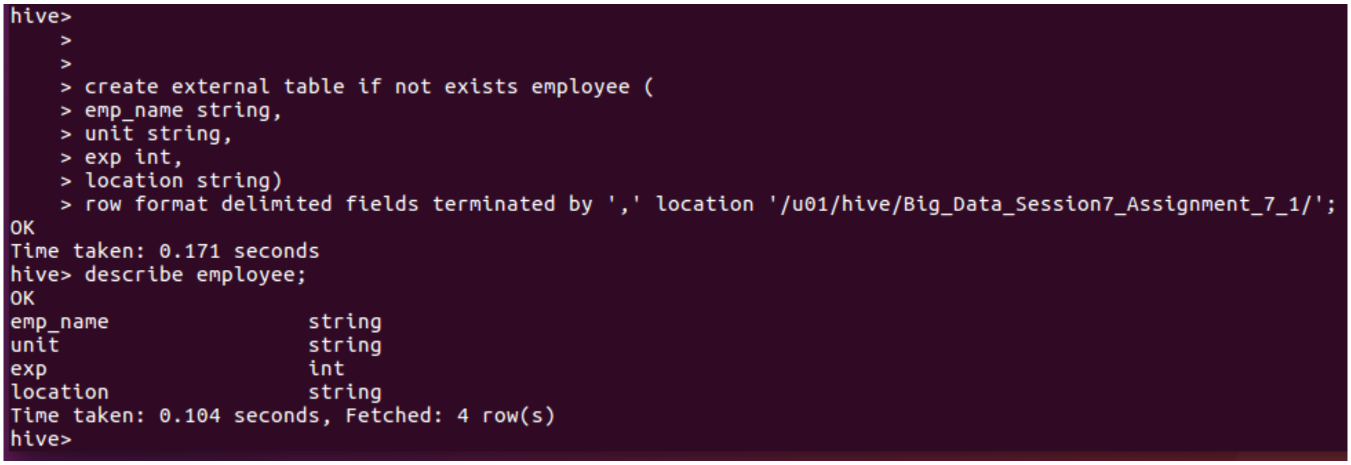


```
sahil@ubuntu:~$ hdfs dfs -ls /u01/hive/warehouse
Found 3 items
drwxr-xr-x   - sahil supergroup          0 2017-12-06 11:14 /u01/hive/warehouse/acadgild_db.db
drwxr-xr-x   - sahil supergroup          0 2017-12-04 07:44 /u01/hive/warehouse/custom.db
drwxr-xr-x   - sahil supergroup          0 2017-12-01 03:27 /u01/hive/warehouse/shri
sahil@ubuntu:~$
```

Step 3:- CREATE EXTERNAL TABLE

Commands used in Step 3

```
hive>
    > create external table if not exists employee_details (
    > emp_name string,
    > unit string,
    > exp int,
    > location string)
    > row format delimited fields terminated by ',' location
'/u01/hive/Big_Data_Session7_Assignment_7_1/';
hive> describe employee;
```



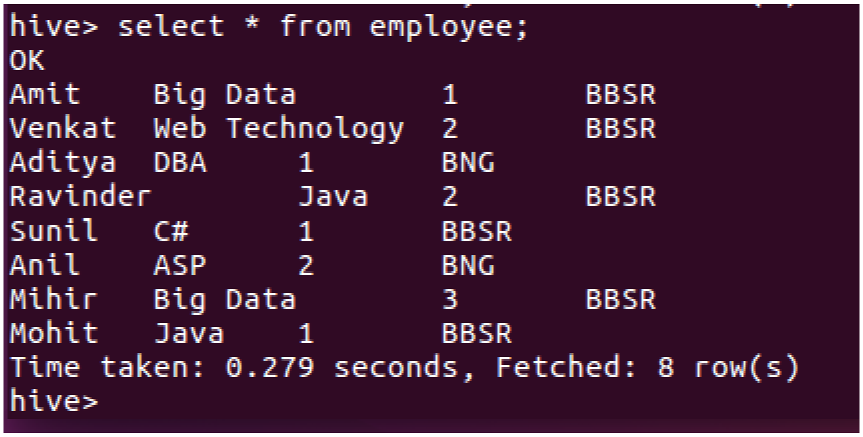
```
hive>
    >
    >
    > create external table if not exists employee (
    > emp_name string,
    > unit string,
    > exp int,
    > location string)
    > row format delimited fields terminated by ',' location '/u01/hive/Big_Data_Session7_Assignment_7_1/';
OK
Time taken: 0.171 seconds
hive> describe employee;
OK
emp_name                string
unit                    string
exp                     int
location                string
Time taken: 0.104 seconds, Fetched: 4 row(s)
hive>
```

Step 4 :- Check whether the dataset is imported in the hive table or not.

Commands used in Step 4

hive>

    > select * from employee;



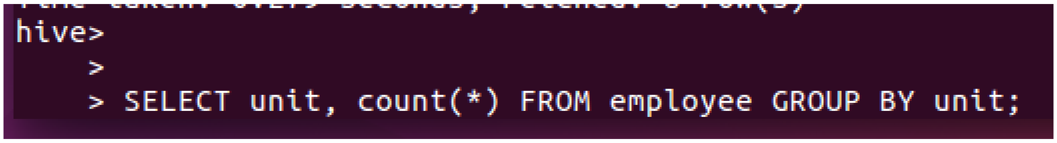
```
hive> select * from employee;
OK
Amit    Big Data        1       BBSR
Venkat  Web Technology  2       BBSR
Aditya  DBA     1       BNG
Ravinder        Java    2       BBSR
Sunil   C#      1       BBSR
Anil    ASP     2       BNG
Mihir   Big Data        3       BBSR
Mohit   Java    1       BBSR
Time taken: 0.279 seconds, Fetched: 8 row(s)
hive>
```

Step 5:- Calculate the number of employees corresponding to each skill from the table 'employee' which is loaded in the Demo.

Commands used in Step 5

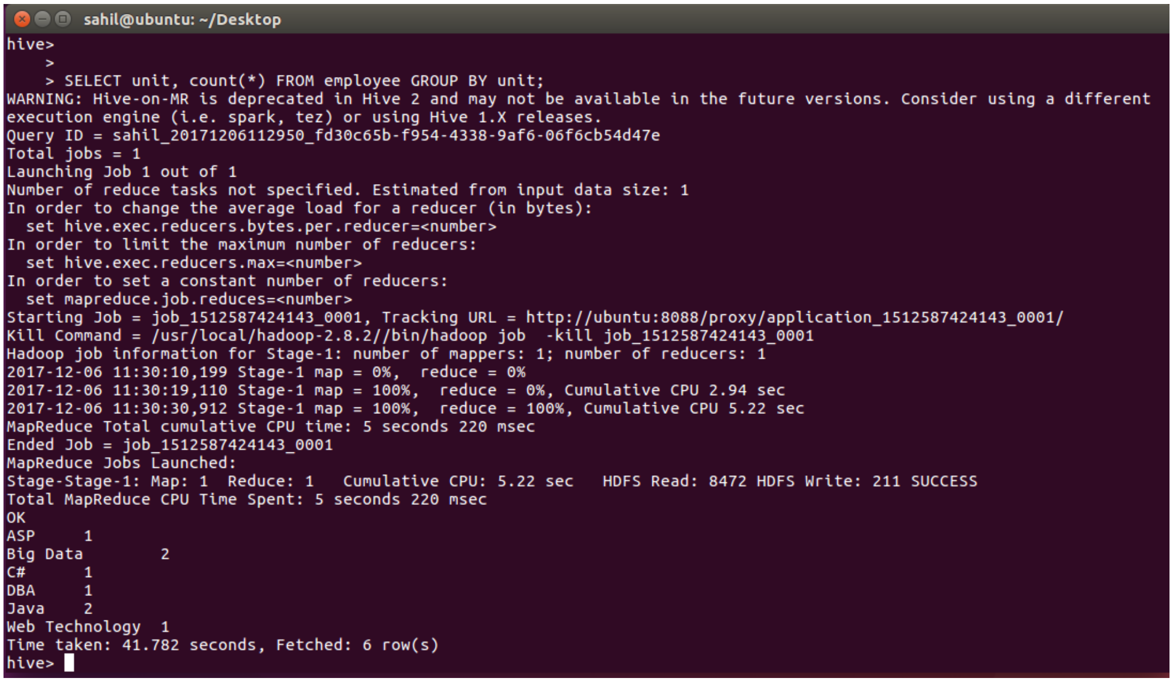
hive>

    > SELECT unit, count(*) FROM employee GROUP BY unit;



```
hive>
    >
    > SELECT unit, count(*) FROM employee GROUP BY unit;
```

Result:-



```
hive>
    >
    > SELECT unit, count(*) FROM employee GROUP BY unit;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different
execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = sahil_20171206112950_fd30c65b-f954-4338-9af6-06f6cb54d47e
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_1512587424143_0001, Tracking URL = http://ubuntu:8088/proxy/application_1512587424143_0001/
Kill Command = /usr/local/hadoop-2.8.2//bin/hadoop job  -kill job_1512587424143_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-06 11:30:10,199 Stage-1 map = 0%,  reduce = 0%
2017-12-06 11:30:19,110 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.94 sec
2017-12-06 11:30:30,912 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.22 sec
MapReduce Total cumulative CPU time: 5 seconds 220 msec
Ended Job = job_1512587424143_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.22 sec   HDFS Read: 8472 HDFS Write: 211 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 220 msec
OK
ASP     1
Big Data        2
C#      1
DBA     1
Java    2
Web Technology  1
Time taken: 41.782 seconds, Fetched: 6 row(s)
hive>
```

ASP	1
Big Data	2
C#	1
DBA	1
Java	2
Web Technology	1