

## 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:

```
acadgild@localhost:~/Abhilasha/hive
File Edit View Search Terminal Help
[acadgild@localhost hive]$ pwd
/home/acadgild/Abhilasha/hive
[acadgild@localhost hive]$ ls -l
total 40
-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 08:24 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 17 16:08 output
-rw-rw-r--. 1 acadgild acadgild 170 Sep 17 14:17 Unsaved Document 1~
[acadgild@localhost hive]$
```

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

```
acadgild@localhost:~/Abhilasha/hive
File Edit View Search Terminal Help
[acadgild@localhost hive]$ cat emp_Details
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
[acadgild@localhost hive]$
```

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

```
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ hive

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-0.14.0.jar!/hive-log4j.properties
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/hive-jdbc-0.14.0-standalone.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hive>
```

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:

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> SHOW DATABASES;  
OK  
acadgild  
h1  
custom  
default  
Time taken: 0.067 seconds, Fetched: 4 row(s)  
hive>
```

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

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> USE custom;  
OK  
Time taken: 0.036 seconds  
hive>  
hive> USE custom;  
OK  
Time taken: 0.036 seconds  
hive>
```

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

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> CREATE TABLE if not exists empDetails  
(  
    empName STRING,  
    skill STRING,  
    exp INT,  
    location STRING  
)  
row format delimited fields terminated by ',';  
OK  
Time taken: 0.087 seconds  
hive>
```

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

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> SHOW TABLES;  
OK  
empdetails  
temperature_data  
temperature_data_vw  
Time taken: 0.056 seconds, Fetched: 3 row(s)  
hive>
```

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

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> DESCRIBE empDetails;  
OK  
empname      string  
skill        string  
exp          int  
location     string  
Time taken: 0.145 seconds, Fetched: 4 row(s)  
hive>
```

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.

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> LOAD DATA  
LOCAL INPATH '/home/acadgild/Abhilasha/hive/emp_Details'  
INTO TABLE empDetails;  
Loading data to table custom.empdetails  
Table custom.empdetails stats: [numFiles=1, totalSize=159]  
OK  
Time taken: 1.325 seconds  
hive>
```

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

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> SELECT * FROM empDetails;  
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: 0.08 seconds, Fetched: 8 row(s)  
hive>
```

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:

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> SELECT skill,COUNT(*) FROM empDetails GROUP BY skill;  
Query ID = acadgild_20170919085252_febdbb45-5496-4f23-81a3-549d7c2ae207  
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_0005, Tracking URL = http://localhost:8088/proxy/application_1505789764223_0005/  
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job -kill job_1505789764223_0005  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2017-09-19 08:52:19,914 Stage-1 map = 0%, reduce = 0%  
2017-09-19 08:52:27,397 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.7 sec  
2017-09-19 08:52:34,790 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.01 sec  
MapReduce Total cumulative CPU time: 4 seconds 10 msec  
Ended Job = job_1505789764223_0005  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.01 sec HDFS Read: 387 HDFS Write: 52 SUCCESS  
Total MapReduce CPU Time Spent: 4 seconds 10 msec  
OK  
ASP 1  
Big Data 2  
DBA 1  
Java 2  
Web Technology 1  
c# 1  
Time taken: 25.313 seconds, Fetched: 6 row(s)  
hive>
```

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:

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
hive> EXPLAIN SELECT skill,COUNT(*) FROM empDetails GROUP BY skill;  
OK  
STAGE DEPENDENCIES:  
  Stage-1 is a root stage  
  Stage-0 depends on stages: Stage-1  
  
STAGE PLANS:  
  Stage: Stage-1  
    Map Reduce  
    Map Operator Tree:  
      TableScan  
        alias: empdetails  
        Statistics: Num rows: 1 Data size: 159 Basic stats: COMPLETE Column stats: NONE  
      Select Operator  
        expressions: skill (type: string)  
        outputColumnNames: skill  
        Statistics: Num rows: 1 Data size: 159 Basic stats: COMPLETE Column stats: NONE  
      Group By Operator  
        aggregations: count()  
        keys: skill (type: string)  
        mode: hash  
        outputColumnNames: _col0, _col1  
        Statistics: Num rows: 1 Data size: 159 Basic stats: COMPLETE Column stats: NONE  
      Reduce Output Operator  
        key expressions: _col0 (type: string)  
        sort order: +  
        Map-reduce partition columns: _col0 (type: string)  
        Statistics: Num rows: 1 Data size: 159 Basic stats: COMPLETE Column stats: NONE  
        value expressions: _col1 (type: bigint)  
    Reduce Operator Tree:  
      Group By Operator  
        aggregations: count(VALUE._col0)  
        keys: KEY._col0 (type: string)  
        mode: mergepartial  
        outputColumnNames: _col0, _col1  
        Statistics: Num rows: 0 Data size: 0 Basic stats: NONE Column stats: NONE  
      Select Operator  
        expressions: _col0 (type: string), _col1 (type: bigint)  
        outputColumnNames: _col0, _col1
```

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]$
```

