C V Raman Global University Big Data Analytics Lab 6 PIG Classwork Assignment on Student Data Analysis

Copy the data files to hadoop: hdfs dfs -copyFromLocal CSE.csv /

hdfs dfs -copyFromLocal Dept.csv /

Explanations: Above commands copy the files from local path (/home/cloudera) to root director (/)

```
root@hp-HP-Pavilion-Notebook:~# hdfs dfs -copyFromLocal CSE.csv /
root@hp-HP-Pavilion-Notebook:~# hdfs dfs -ls /
Found 3 items
-rw-r---- 1 root supergroup
-rw-r---- 1 root supergroup
-rw-r---- 1 root supergroup
-rw-r---- 1 root supergroup
-root@hp-HP-Pavilion-Notebook:~#
```

To check whether file is copied or not: hdfs dfs -ls /

Load the data into pig:

cseData1 = LOAD '/CSE.csv' USING PigStorage(',') AS
(Name:chararray, RollNo:int, Marks:int, City:chararray);

deptData1 = LOAD '/Dept.csv' USING PigStorage(',') AS
(RollNo:int, Department:chararray, City:chararray,
College:chararray);

```
grunt> cseData = LOAD 'CSE.csv' AS (Name:chararray, RollNo:int, Marks:int, City:chararray);
grunt> deptData = LOAD '.csv' AS (RollNo:int, Department:chararray, City:chararray, College:chararray);
grunt>
```

a). To group the tuples of students based on city. Ans:

groupByCity = GROUP cseData1 BY City;

```
grunt> groupByCity = GROUP cseData BY City;
grunt>
```

dump groupByCity;

({(Guneet Bajaj,16103123,175,Pune,,,),(Sanyam Dhawan,16103122,174,Pune,,,),(Karan Chabra,16103121,173,Pune,,,),(Gagandeep Chugh,16103120,172,Pune,,,),(Sanyam Mehta,16103119,117,Pune,,,),(Kantshk Gautan,16103118,170,Pune,,,),(Harsimran Singh,16103117,169,Pune,,,),(Vardaan Bajaj,16103 116,168,Pune,,,),(Shurha Bajaj,16103115,167,Pune,,,),(Prantta,16103113,166,Pune,,,),(Tantshq Mittal,16103112,165,Pune,,,),(Kunal Gupta,16103111,1614,Pune,,,),(Kunal Gupta,16103111,1614,Pune,,,),(Kunal Gupta,16103111,1614,Pune,,,),(Muskan Sharna,16103109,162,Pune,,,),(Shughndeep Singh,16103108,161,Pune,,,),(Mestal Jethila,16103107,160,Pune,,,),(Bagin Aggarval,16103105,Pune,,),(Shill Bhardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103105,Pune,,,),(Shill,Bardwal,16103097,145,Chennal,,,),(Nordeep Kaur,16103097,145,Chennal,,,),(Prabhnoor Kaur,16103096,144,Chennal,,,),(Srishti,16103089,143,Chennal,,,),(Shill,Bardwal,16103081,146,Chennal,,,),(Shill,Bardwal,16103081,143),Chennal,,,),(Churvis tingh,16103085,149,Gulyrat,,),(Akshall,16103084,147,Chennal,,,),(Shill,Bardwal,16103084,149,Chennal,,,),(Shill,Bardwal,16103084,149,Chennal,,,),(Churvis tingh,16103085,136,Chennal,,,),(Shill,Bardwal,16103084,139,Chennal,,,),(Churvis tingh,16103087,149,Chennal,,,),(Shill,Bardwal,16103084,149,Chennal,,,),(Churvis tingh,16103087,149,Chennal,,,),(Shill,Bardwal,16103084,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Shill,Bardwal,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(Curvis tingh,16103087,149,Chennal,,,),(

b). To display the name of all students in upper case. Ans: upperCaseNames = FOREACH cseData1 GENERATE UPPER(Name);

```
grunt> upperCaseNames = FOREACH cseData GENERATE UPPER(Name);
grunt>
```

dump upperCaseNames;

```
(RIYANKA MONDAL)
(GURVIR SINGH)
(ANIL KUMAR)
(DIVYA SEHJAL)
(SRISHTI)
(PRABHNOOR KAUR)
(NOORDEEP KAUR)
(ABHINAV THAKUR)
(KANIKA SHARMA)
(AKSHANT SARMAL)
(DHRUV GULERIA)
(ABHISHEK)
(KSHITTI FORWAL)
(Screenshot
(RACHAY SHARMA)
(SHIVAM JAND )
(ABHISHEK GARG)
(SAHIL BHARDWAJ)
(PRAGUN AGGARWAL)
(LOVEDEEP SINGH)
( KAINAAT )
(BHALINDER SINGH MUNIER)
(NEETI JETHLIA)
(SHAGUNDEEP SINGH)
(MUSKAN SHARMA)
(SHUBHKARMANDEEP KAUR)
(KUNAL GUPTA)
(TANISHQ MITTAL)
(PRANITA)
(SIMRAN BAJAJ)
(VARDAAN BAJAJ)
(HARSIMRAN SINGH)
(KANISHK GAUTAM)
(SANYAM MEHTA)
(GAGANDEEP CHUGH)
(KARAN CHABRA)
(SANYAM DHAWAN)
(GUNEET BAJAJ)
```

c). To find the tuples of those students where the marks is greater than equal to 70.

Ans: marksGT70 = FILTER cseData1 BY Marks >= 70;

```
grunt> marksGT70 = FILTER cseData BY Marks >= 70;
grunt>
```

dump marksGT70;

```
(Abhinav Singla, 16103013, 70, Chandigarh)
(Shounak Chowdhury ,16103014,71,New Delhi)
(Abhinav Agarwal, 16103015, 72, New Delhi)
(Gaurav Ahuja, 16103016, 73, Chandigarh)
(Deepak Tiwari,16103017,74,Chandigarh)
(Pranjali,16103018,75,Chandigarh)
(Muskan Garg, 16103019, 76, Chandigarh)
(Ishneet Kaur, 16103020, 77, Chandigarh)
(Mankeerat Singh ,16103021,78,Chandigarh)
(Nikita Gupta,16103022,79,Chandigarh)
(Pradhuman Goyal, 16103023, 80, Chandigarh)
(Arunav Sharma, 16103024, 81, Chandigarh)
(Rohan Dhadwal, 16103025, 82, Himachal)
(Dhanwant Kaur, 16103026, 83, Himachal)
(Ritesh Goyal, 16103027, 84, Himachal)
(Aakankasha Sharma, 16103028, 85, Himachal)
(Leevansha, 16103029, 86, Himachal)
(Harshima Singla,16103030,87,Himachal)
(Janamdeep Singh, 16103031, 88, Himachal)
(Baljinder Singh, 16103032, 89, Panjab)
(Priyanshu Sharma, 16103033, 90, Panjab)
(Arshdeep Singh Sachdeva, 16103034, 91, Panjab)
(Krittika Chhabra, 16103035, 92, Panjab)
(Mayank Saggar, 16103036, 93, Panjab)
(Shivam Sehgal,16103037,94,Panjab)
(Kartik Kaushal,16103039,95,Panjab)
(Chetna Gupta,16103040,96,Panjab)
(Madhav Pruthi, 16103041, 97, Panjab)
(Nirmaljot Singh, 16103042, 98, Panjab)
(Pritish Gupta, 16103043, 99, Panjab)
(Vaibhav Goel, 16103044, 100, Pune)
(Parikh Oberoi, 16103045, 101, Pune)
(Abhinav Chawla, 16103046, 102, Pune)
(Kushal Dudeja,16103047,103,Pune)
(Abhinav Narang, 16103048, 104, Pune)
(Deepti,16103049,105,Pune)
(Aditya Sharma,16103050,106,Pune)
```

d). To get details of students who belongs to New Delhi only. Ans: NewDelhiStudents = FILTER cseData1 BY City == 'New Delhi';

```
grunt> NewDelhiStudents = FILTER cseData BY City == 'New Delhi';
grunt>
```

dump NewDelhiStudents;

```
(Sanjana Koul,16103006,64,New Delhi)
(Nikhat Parveen,16103007,65,New Delhi)
(Ramesh Chandra Pokharel,16103008,66,New Delhi)
(Pradeep Parajuli,16103010,67,New Delhi)
(Sarthak Vaswani,16103011,68,New Delhi)
(Shounak Chowdhury ,16103014,71,New Delhi)
(Abhinav Agarwal,16103015,72,New Delhi)
grunt>
```

e). To display the results where the city name Chandigarh is replaced with shorter form CHD and New Delhi with NDLS.

replacedData = FOREACH cseData GENERATE REPLACE(REPLACE(City, 'Chandigarh', 'CHD'), 'New Delhi', 'NDLS');

dump replacedData;

```
(Hyderabad)
(Hyderabad)
(Hyderabad)
(NDLS)
(NDLS)
(NDLS)
(NDLS)
(NDLS)
(CHD)
(CHD)
(NDLS)
(NDLS)
(CHD)
(CHD)
(CHD)
(CHD)
(CHD)
(CHD)
(CHD)
(CHD)
(CHD)
(Himachal)
(Himachal)
(Himachal)
(Himachal)
(Himachal)
(Himachal)
(Himachal)
(Panjab)
(Panjab)
(Panjab)
```

f). To calculate average marks for each student.

To do so execute the following

Group the students by Roll number: marksByStudent = GROUP cseData1 BY RollNo;

```
grunt> marksByStudent = GROUP cseData BY RollNo;
grunt>
```

dump marksByStudent;

```
(16103085,{(Gurvir singh,16103085,140,Chennai)})
(16103086,{(Anil Kumar,16103086,141,Chennai)})
(16103087, {(Divya Sehjal, 16103087, 142, Chennai)})
(16103089,{(Srishti,16103089,143,Chennai)})
(16103090, {(Prabhnoor Kaur, 16103090, 144, Chennai)})
(16103091, {(Noordeep Kaur, 16103091, 145, Chennai)})
(16103093,{(Abhinav Thakur,16103093,146,Chennai)})
(16103094,{(Kanika Sharma,16103094,147,Himachal)})
(16103095,{(Akshant Sarmal,16103095,148,Chennai)})
(16103096,{(Dhruv Guleria,16103096,149,Gujrat)})
(16103097,{(Abhishek,16103097,150,Gujrat)})
(16103098,{(Kshitij Porwal,16103098,151,Gujrat)})
  (Raghav Sharma,16103099,152,Gujrat)})
Screenshot (Shivam Jand ,16103100,153,Gujrat)})
(16103101,{(Abhishek Garg,16103101,154,Gujrat)})
(16103102,{(Sahil Bhardwaj,16103102,155,Pune)})
(16103103,{(Pragun Aggarwal,16103103,156,Pune)})
(16103104,{(Lovedeep Singh,16103104,157,Pune)})
(16103105,{( Kainaat ,16103105,158,Pune)})
(16103106, {(Bhalinder Singh Munier, 16103106, 159, Pune)})
(16103107,{(Neeti Jethlia,16103107,160,Pune)})
(16103108,{(Shagundeep Singh,16103108,161,Pune)})
(16103109, {(Muskan Sharma, 16103109, 162, Pune)})
(16103110, {(Shubhkarmandeep Kaur, 16103110, 163, Pune)})
(16103111,{(Kunal Gupta,16103111,164,Pune)})
(16103112,{(Tanishq Mittal,16103112,165,Pune)})
(16103113,{(Pranita,16103113,166,Pune)})
(16103115,{(Simran Bajaj,16103115,167,Pune)})
(16103116,{(Vardaan Bajaj,16103116,168,Pune)})
(16103117,{(Harsimran Singh,16103117,169,Pune)})
(16103118,{(Kanishk Gautam,16103118,170,Pune)})
(16103119,{(Sanyam Mehta,16103119,171,Pune)})
(16103120, {(Gagandeep Chugh, 16103120, 172, Pune)})
(16103121,{(Karan Chabra,16103121,173,Pune)})
(16103122, {(Sanyam Dhawan, 16103122, 174, Pune)})
(16103123,{(Guneet Bajaj,16103123,175,Pune)})
(,{(Name,,,City)})
```

Calculate average for each group: avgMarks = FOREACH marksByStudent GENERATE group AS RollNo, AVG(cseData1.Marks) AS avgMark;

```
grunt> avgMarks = FOREACH marksByStudent GENERATE group AS RollNo, AVG(cseData.Marks) AS avgMark;
grunt>
grunt>
```

dump avgMarks;

```
(16103085,140.0)
(16103086,141.0)
(16103087,142.0)
(16103089,143.0)
(16103090,144.0)
(16103091,145.0)
(16103093,146.0)
(16103094,147.0)
(16103095,148.0)
(16103096,149.0)
(16103097,150.0)
(16103098,151.0)
(16103099,152.0)
(16103100,153.0)
(16103101,154.0)
(16103102,155.0)
(16103103,156.0)
(16103104,157.0)
(16103105,158.0)
(16103106,159.0)
(16103107,160.0)
(16103108,161.0)
(16103109,162.0)
(16103110,163.0)
(16103111,164.0)
(16103112,165.0)
(16103113,166.0)
(16103115,167.0)
(16103116,168.0)
(16103117,169.0)
(16103118,170.0)
(16103119,171.0)
(16103120,172.0)
(16103121,173.0)
(16103122,174.0)
(16103123,175.0)
```

i). To merge contents of two relation CSE.csv and Dept.csv. Ans: mergeRelation = JOIN cseData1 BY RollNo, deptData1 BY RollNo;

```
grunt> mergeRelation = JOIN cseData BY RollNo, deptData BY RollNo;
grunt>
```

dump mergeRelation;

```
(Riyanka Mondal, 16103084, 139, Chennai, 16103084, Computer Science, Chennai, PEC)
(Gurvir singh, 16103085, 140, Chennai, 16103085, Computer Science, Chennai, PEC)
(Anil Kumar, 16103086, 141, Chennai, 16103086, Computer Science, Chennai, PEC)
(Divya Sehjal, 16103087, 142, Chennai, 16103087, Computer Science, Chennai, PEC)
(Srishti,16103089,143,Chennai,16103089,Computer Science,Chennai,PEC)
(Prabhnoor Kaur, 16103090, 144, Chennai, 16103090, Electronics, Chennai, PEC)
(Noordeep Kaur, 16103091, 145, Chennai, 16103091, Electronics, Chennai, PEC)
(Abhinav Thakur,16103093,146,Chennai,16103093,Electronics,Chennai,PEC)
(Kanika Sharma,16103094,147,Himachal,16103094,Electronics,Himachal,PEC)
(Akshant Sarmal, 16103095, 148, Chennai, 16103095, Electronics, Chennai, PEC)
(Dhruv Guleria,16103096,149,Gujrat,16103096,Electronics,Gujrat,PEC)
(Abhishek, 16103097, 150, Gujrat, 16103097, Electronics, Gujrat, PEC)
 (shitli Porwal,16103098,151,Gujrat,16103098,Electronics,Gujrat,PEC)
Screenshot
Ragnav Sharma,16103099,152,Gujrat,16103099,Electronics,Gujrat,PEC)
(Shivam Jand ,16103100,153,Gujrat,16103100,Electronics,Gujrat,PEC)
(Abhishek Garg, 16103101, 154, Gujrat, 16103101, Electronics, Gujrat, PEC)
(Sahil Bhardwaj,16103102,155,Pune,16103102,Electronics,Pune,PEC)
(Pragun Aggarwal, 16103103, 156, Pune, 16103103, Computer Science, Pune, PEC)
(Lovedeep Singh, 16103104, 157, Pune, 16103104, Computer Science, Pune, PEC)
( Kainaat ,16103105,158,Pune,16103105,Computer Science,Pune,PEC)
(Bhalinder Singh Munier, 16103106, 159, Pune, 16103106, Computer Science, Pune, PEC)
(Neeti Jethlia,16103107,160,Pune,16103107,Computer Science,Pune,PEC)
(Shagundeep Singh,16103108,161,Pune,16103108,Computer Science,Pune,PEC)
(Muskan Sharma,16103109,162,Pune,16103109,Computer Science,Pune,PEC)
(Shubhkarmandeep Kaur,16103110,163,Pune,16103110,Computer Science,Pune,PEC)
(Kunal Gupta,16103111,164,Pune,16103111,Computer Science,Pune,PEC)
(Tanishq Mittal,16103112,165,Pune,16103112,Computer Science,Pune,PEC)
(Pranita,16103113,166,Pune,16103113,Computer Science,Pune,PEC)
(Simran Bajaj,16103115,167,Pune,16103115,Computer Science,Pune,PEC)
(Vardaan Bajaj, 16103116, 168, Pune, 16103116, Mechanical, Pune, PEC)
(Harsimran Singh, 16103117, 169, Pune, 16103117, Mechanical, Pune, PEC)
(Kanishk Gautam, 16103118, 170, Pune, 16103118, Mechanical, Pune, PEC)
(Sanyam Mehta, 16103119, 171, Pune, 16103119, Mechanical, Pune, PEC)
(Gagandeep Chugh, 16103120, 172, Pune, 16103120, Mechanical, Pune, PEC)
(Karan Chabra, 16103121, 173, Pune, 16103121, Civil, Pune, PEC)
(Sanyam Dhawan,16103122,174,Pune,16103122,Civil,Pune,PEC)
(Guneet Bajaj,16103123,175,Pune,16103123,Civil,Pune,PEC)
```

- j). To display the details of students who belongs to computer Science only.
- : CSEstudents = FILTER mergeRelation BY Department == 'Computer Science';

```
grunt> CSEstudents = FILTER mergeRelation BY Department == 'Computer Science';
grunt> ■
```

dump CSEstudents;

```
(Sidharth Bhandart,16103001,61,Hyderabad,16103001,Computer Science,Hyderabad,PEC)
(Prayag Panta,16103003,C2,Hyderabad,16103003,Computer Science,Hyderabad,IIT Roorkee)
(Rajesh Jha,16103004,63,Hyderabad,16103003,Computer Science,Hyderabad,IIT Roorkee)
(Sanjana Koul,16103006,64,New Delht,16103006,Computer Science,New Delht,IIT Roorkee)
(Rikhat Parveen,16103007,65,New Delht,16103007,Computer Science,New Delht,IIT Roorkee)
(Ramesh Chandra Pokharel,16103008,66,New Delht,16103008,Computer Science,New Delht,IIT Roorkee)
(Pradeep Parajult,16103010,67,New Delht,16103010,Computer Science,New Delht,IIT Roorkee)
(Sarthak Vaswant,16103011,68,New Delht,16103011,Computer Science,How Delht,IIT Roorkee)
(Sarsh Garg,16103012,69,Chandigarh,16103012,Computer Science,Chandigarh,IIT Roorkee)
(Saprash Garg,16103012,69,Chandigarh,16103013,Computer Science,Chandigarh,IIT Roorkee)
(Shounak chowdhury,16103014,71,New Delht,16103013,Computer Science,Chandigarh,IIT Roorkee)
(Arshdeep Singh Sachdeva,16103034,91,Panjab,16103034,Computer Science,Panjab,NITTR Chandigarh)
(Krittika Chhabra,16103035,92,Panjab,16103035,Computer Science,Panjab,NITTR Chandigarh)
(Krittika Chhabra,16103035,92,Panjab,16103037,Computer Science,Panjab,NITTR Chandigarh)
(Ksivam Sehgal,16103037,94,Panjab,16103037,Computer Science,Panjab,NITTR Chandigarh)
(Ksitvam Sehgal,16103037,94,Panjab,16103037,Computer Science,Panjab,NITTR Chandigarh)
(Kartik Kaushal,16103034,95,Panjab,16103037,Computer Science,Panjab,NITTR Chandigarh)
(Kartik Kaushal,16103041,97,Panjab,16103041,Computer Science,Panjab,NITTR Chandigarh)
(Kartik Gupta,16103044,96,Panjab,16103047,Computer Science,Panjab,NITTR Chandigarh)
(Kadhav Pruthi,16103041,97,Panjab,16103041,Computer Science,Panjab,IIT Delht)
(Pritish Gupta,16103043,99,Panjab,16103043,Computer Science,Pune,PEC)
(Abhinav Chaula,16103044,100,Pune,16103045,Computer Science,Pune,PEC)
(Kushal Dudeja,16103044,100,Pune,16103045,Computer Science,Pune,PEC)
(Abhinav Chaula,16103053,107,Pune,16103045,Computer Science,Pune,PEC)
(Shirish Gupta,16103053,107,Pune,
```