



Session 4: Schedulers in YARN & Introduction to Pig

Assignment 2 Question

*Create a sample dataset and implement the below Pig commands on the same dataset.*

1. *Concat :-*

Input file :-

(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)

(2,siddarth,Battacharya,22,9848022338,Kolkata,78)

(3,Rajesh,Khanna,22,9848022339,Delhi,90)

(4,Preethi,Agarwal,21,9848022330,Pune,93)

(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75)

(6,Archana,Mishra,23,9848022335,Chennai,87)

(7,Komal,Nayak,24,9848022334,trivendram,83

Command:-

student\_details = LOAD '/user/cloudera/student\_details.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray, gpa:int);

grunt> Dump student\_details;

grunt> student\_name\_concat = foreach student\_details Generate CONCAT (firstname, lastname);

grunt> student\_name\_concat = foreach student\_details Generate CONCAT (firstname, lastname);

grunt> Dump student\_name\_concat;

Output :-

(RajivReddy)

(siddarthBattacharya)

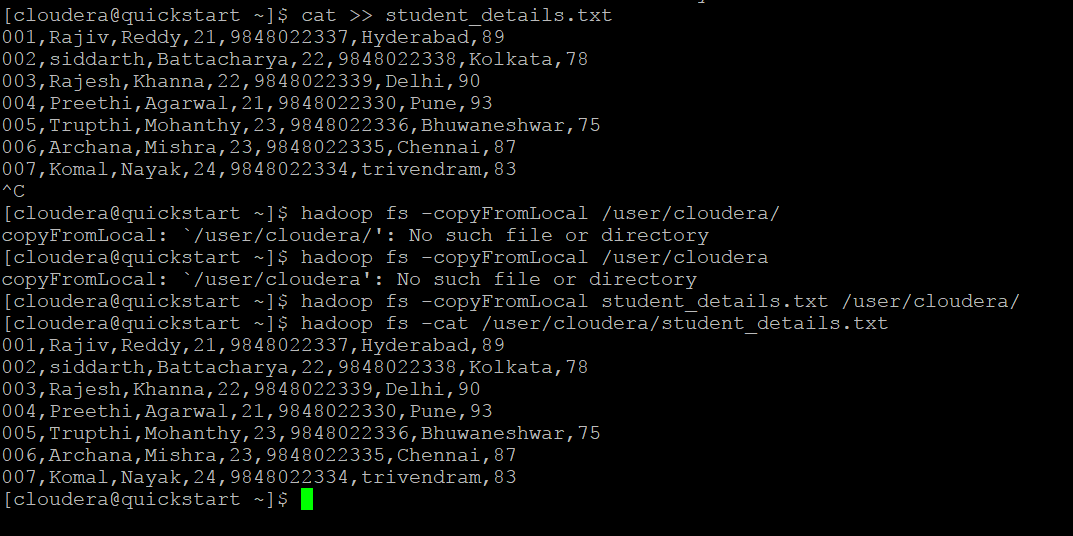
(RajeshKhanna)

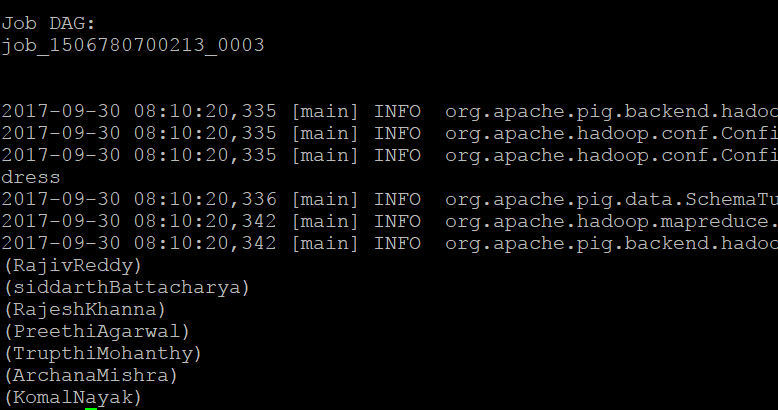
(PreethiAgarwal)

(TrupthiMohanthy)

(ArchanaMishra)

(KomalNayak)





1. *Tokenize :-*

Input file:-

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/student\_details1.txt

001,Rajiv Reddy,21,Hyderabad

002,siddarth Battacharya,22,Kolkata

003,Rajesh Khanna,22,Delhi

004,Preethi Agarwal,21,Pune

005,Trupthi Mohanthy,23,Bhuwaneshwar

006,Archana Mishra,23 ,Chennai

007,Komal Nayak,24,trivendram

008,Bharathi Nambiayar,24,Chennai

grunt> student\_details = LOAD '/user/cloudera/student\_details1.txt' USING PigStorage(',') as (id:int, name:chararray, age:int, city:chararray);

grunt> student\_name\_tokenize = foreach student\_details Generate TOKENIZE(name);

grunt> Dump student\_name\_tokenize;

Output:-

({(Rajiv),(Reddy)})

({(siddarth),(Battacharya)})

({(Rajesh),(Khanna)})

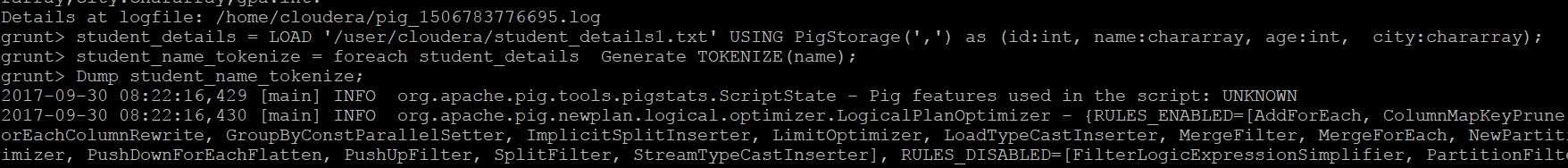
({(Preethi),(Agarwal)})

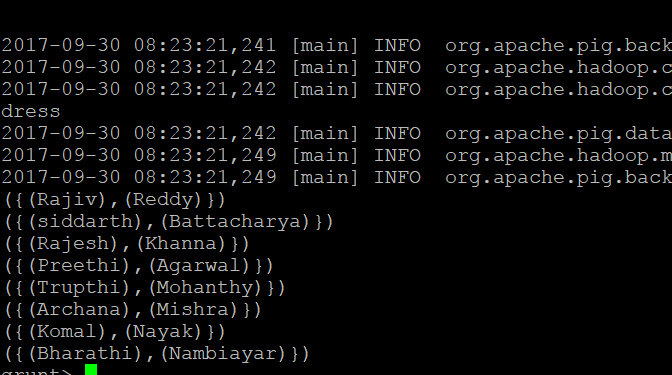
({(Trupthi),(Mohanthy)})

({(Archana),(Mishra)})

({(Komal),(Nayak)})

({(Bharathi),(Nambiayar)})





1. *Sum:-*

Input file :-

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/employee.txt

1,John,2007-01-24,250

2,Ram,2007-05-27,220

3,Jack,2007-05-06,170

3,Jack,2007-04-06,100

4,Jill,2007-04-06,220

5,Zara,2007-06-06,300

Command:

grunt> employee\_data = LOAD '/user/cloudera/employee.txt' USING PigStorage(',') as (id:int, name:chararray, workdate:chararray, daily\_typing\_pages:int);

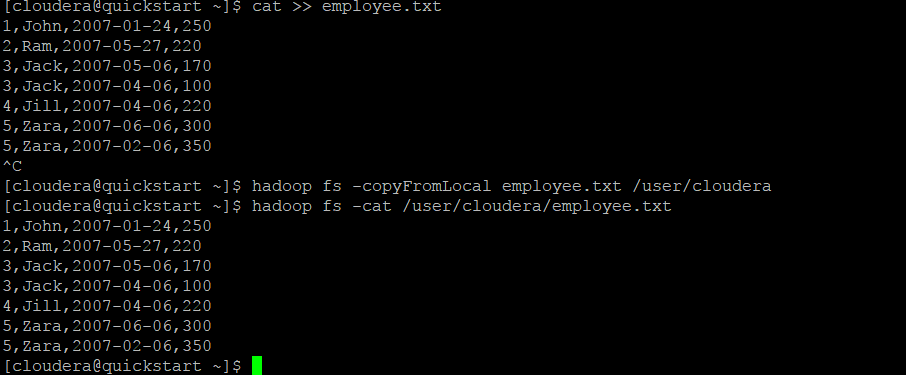
grunt> employee\_group = Group employee\_data all;

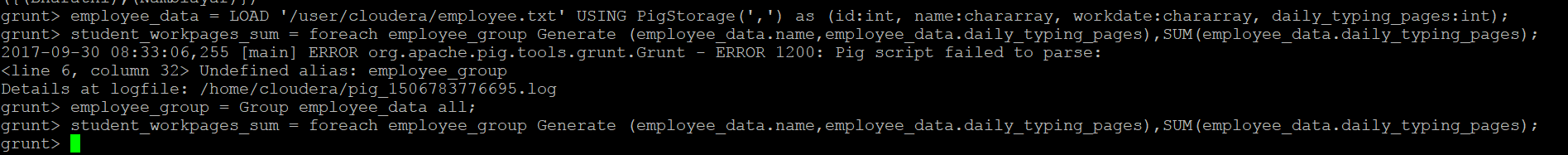
grunt> student\_workpages\_sum = foreach employee\_group Generate (employee\_data.name,employee\_data.daily\_typing\_pages),SUM(employee\_data.daily\_typing\_pages);

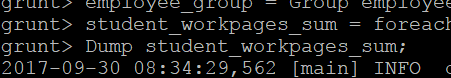
grunt> Dump student\_workpages\_sum;

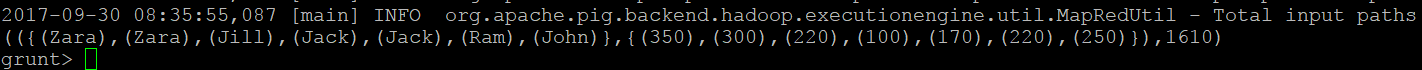
output file:-

(({(Zara),(Zara),(Jill),(Jack),(Jack),(Ram),(John)},{(350),(300),(220),(100),(170),(220),(250)}),1610)









1. *Min: -*

Input file: -

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/student\_details.txt

001,Rajiv,Reddy,21,9848022337,Hyderabad,89

002,siddarth,Battacharya,22,9848022338,Kolkata,78

003,Rajesh,Khanna,22,9848022339,Delhi,90

004,Preethi,Agarwal,21,9848022330,Pune,93

005,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75

006,Archana,Mishra,23,9848022335,Chennai,87

007,Komal,Nayak,24,9848022334,trivendram,83

Command:-

grunt> student\_details = LOAD '/user/cloudera/student\_details.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray, gpa:int);

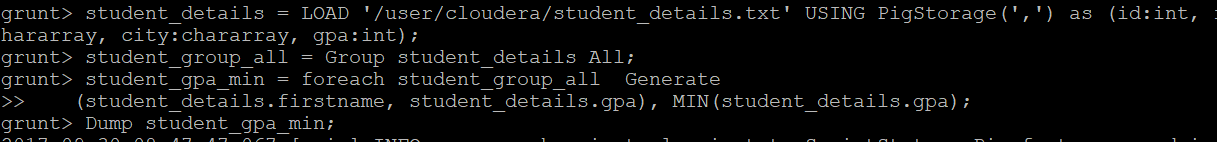
grunt> student\_group\_all = Group student\_details All;

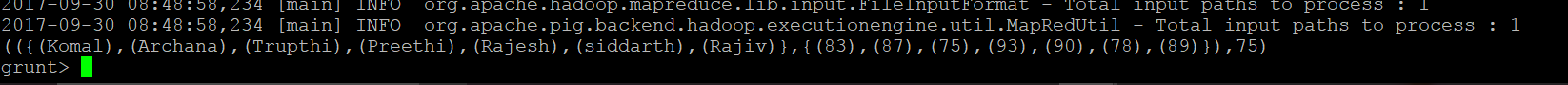
grunt> student\_gpa\_min = foreach student\_group\_all Generate

>> (student\_details.firstname, student\_details.gpa), MIN(student\_details.gpa);

grunt> Dump student\_gpa\_min;

output file: (({(Komal),(Archana),(Trupthi),(Preethi),(Rajesh),(siddarth),(Rajiv)},{(83),(87),(75),(93),(90),(78),(89)}),75)





1. *Max:-*

Input file: -

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/student\_details.txt

001,Rajiv,Reddy,21,9848022337,Hyderabad,89

002,siddarth,Battacharya,22,9848022338,Kolkata,78

003,Rajesh,Khanna,22,9848022339,Delhi,90

004,Preethi,Agarwal,21,9848022330,Pune,93

005,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75

006,Archana,Mishra,23,9848022335,Chennai,87

007,Komal,Nayak,24,9848022334,trivendram,83

Command: -

grunt> student\_details = LOAD '/user/cloudera/student\_details.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray, gpa:int);

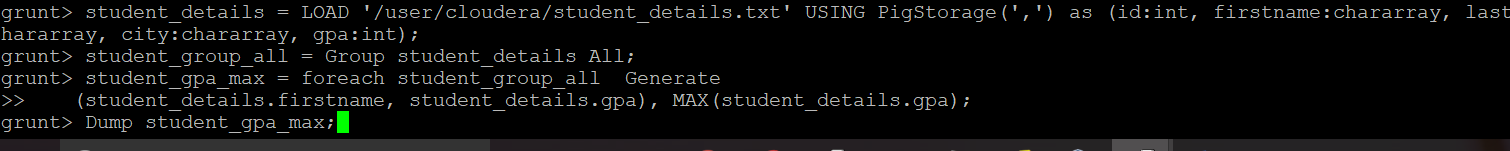
grunt> student\_group\_all = Group student\_details All;

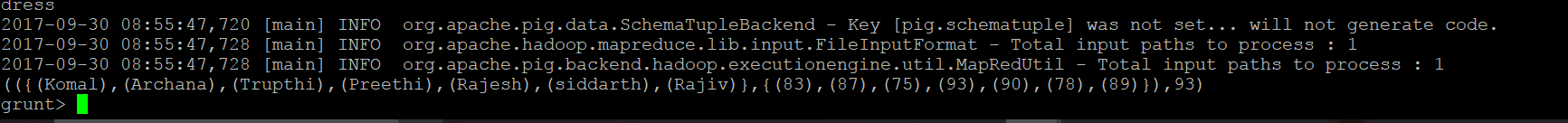
grunt> student\_gpa\_max = foreach student\_group\_all Generate

>> (student\_details.firstname, student\_details.gpa), MAX(student\_details.gpa);

grunt> Dump student\_gpa\_max;

output file: - (({(Komal),(Archana),(Trupthi),(Preethi),(Rajesh),(siddarth),(Rajiv)},{(83),(87),(75),(93),(90),(78),(89)}),93)





1. *Limit: -*

Inputfile :-

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/student\_details.txt

001,Rajiv,Reddy,21,9848022337,Hyderabad,89

002,siddarth,Battacharya,22,9848022338,Kolkata,78

003,Rajesh,Khanna,22,9848022339,Delhi,90

004,Preethi,Agarwal,21,9848022330,Pune,93

005,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75

006,Archana,Mishra,23,9848022335,Chennai,87

007,Komal,Nayak,24,9848022334,trivendram,

Command:-

grunt> student\_details = LOAD '/user/cloudera/student\_details.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray, gpa:int);

grunt> limit\_data = LIMIT student\_details 4;

grunt> Dump limit\_data;

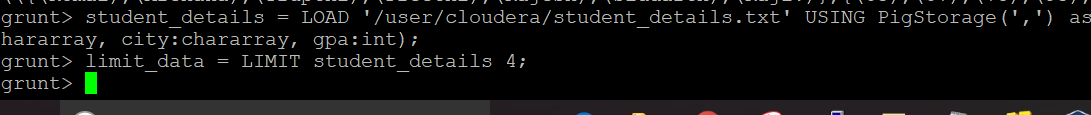
output: -

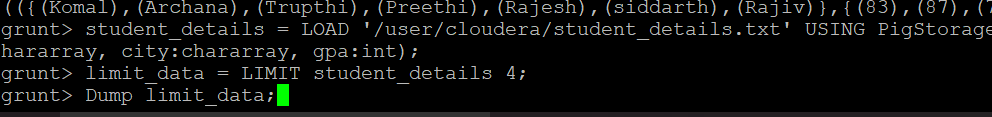
(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)

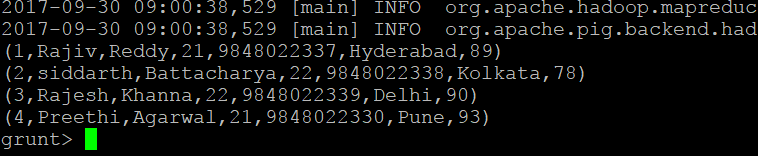
(2,siddarth,Battacharya,22,9848022338,Kolkata,78)

(3,Rajesh,Khanna,22,9848022339,Delhi,90)

(4,Preethi,Agarwal,21,9848022330,Pune,93)







1. *Store*: -

Input file: -

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/student\_data.txt

001,Rajiv,Reddy,9848022337,Hyderabad

002,siddarth,Battacharya,9848022338,Kolkata

003,Rajesh,Khanna,9848022339,Delhi

004,Preethi,Agarwal,9848022330,Pune

005,Trupthi,Mohanthy,9848022336,Bhuwaneshwar

006,Archana,Mishra,9848022335,Chennai.

Command: -

grunt> student = LOAD '/user/cloudera/student\_data.txt' USING PigStorage(',') as ( id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray );

grunt> STORE student INTO '/user/cloudera/pig\_Output/ ' USING PigStorage (',');

Output file: -

[cloudera@quickstart ~]$ hdfs dfs -ls '/user/cloudera/pig\_Output/'

Found 2 items

-rw-r--r-- 1 cloudera cloudera 0 2017-09-30 09:07 /user/cloudera/pig\_Output/\_SUCCESS

-rw-r--r-- 1 cloudera cloudera 224 2017-09-30 09:07 /user/cloudera/pig\_Output/part-m-00000

[cloudera@quickstart ~]$ hdfs dfs -cat '/user/cloudera/pig\_Output/part-m-00000'

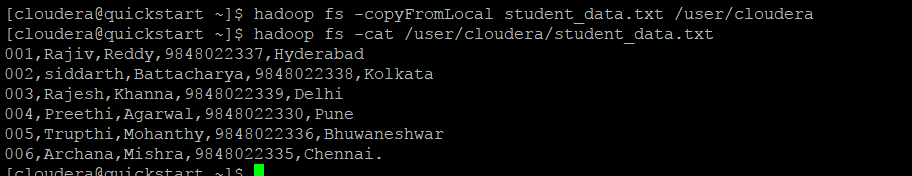
1,Rajiv,Reddy,9848022337,Hyderabad

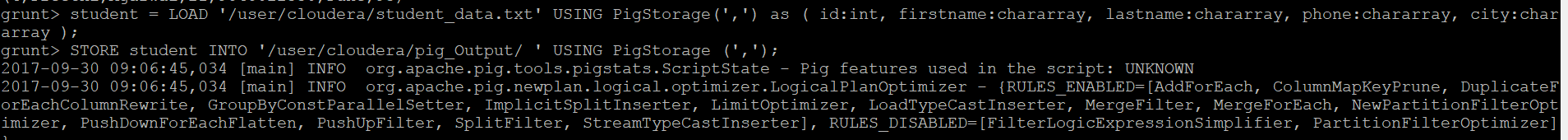
2,siddarth,Battacharya,9848022338,Kolkata

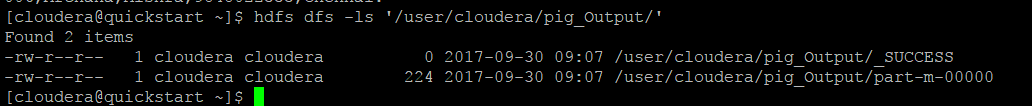
3,Rajesh,Khanna,9848022339,Delhi

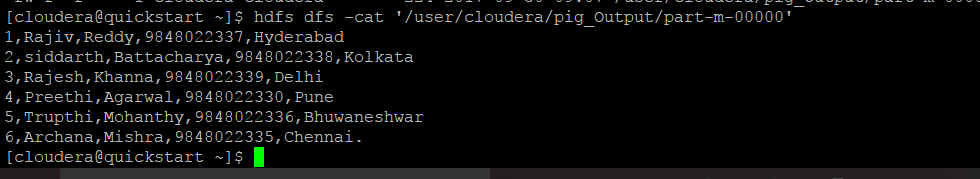
4,Preethi,Agarwal,9848022330,Pune

5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar









1. *Distinct: -*

Input file: -

cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/student.txt

001,Rajiv,Reddy,9848022337,Hyderabad

002,siddarth,Battacharya,9848022338,Kolkata

002,siddarth,Battacharya,9848022338,Kolkata

003,Rajesh,Khanna,9848022339,Delhi

003,Rajesh,Khanna,9848022339,Delhi

004,Preethi,Agarwal,9848022330,Pune

005,Trupthi,Mohanthy,9848022336,Bhuwaneshwar

006,Archana,Mishra,9848022335,Chennai

006,Archana,Mishra,9848022335,Chennai

Command: -

grunt> student = LOAD '/user/cloudera/student.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray);

grunt> distinct\_data = DISTINCT student;

grunt> Dump distinct\_data;

output: -

(1,Rajiv,Reddy,9848022337,Hyderabad)

(2,siddarth,Battacharya,9848022338,Kolkata )

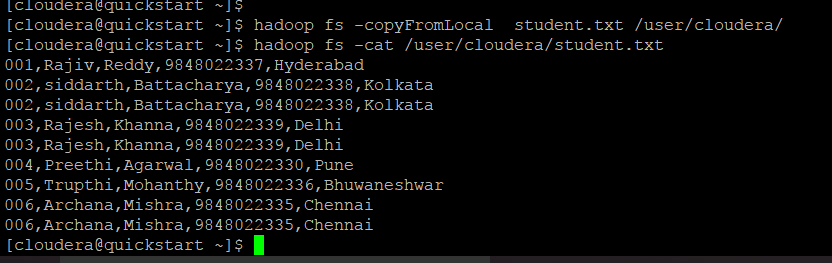
(3,Rajesh,Khanna,9848022339,Delhi )

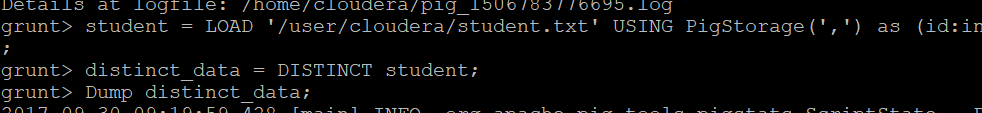
(4,Preethi,Agarwal,9848022330,Pune )

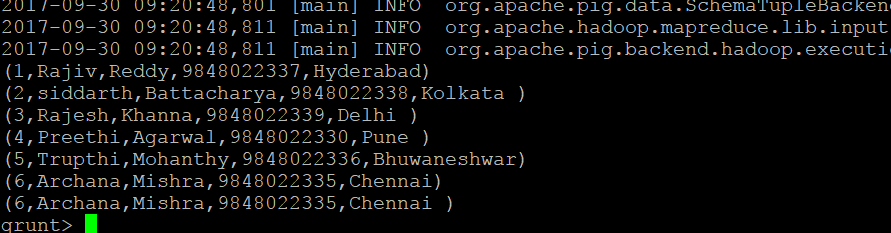
(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)

(6,Archana,Mishra,9848022335,Chennai)

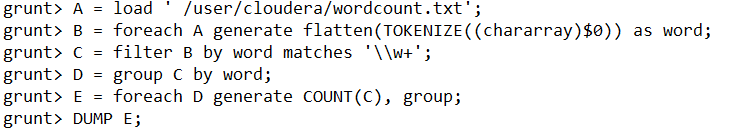
(6,Archana,Mishra,9848022335,Chennai )







1. *Flatten:-*





1. *Is empty:-*

Input file: -

[cloudera@quickstart ~]$ cat >> empsales.txt

1,Robin,22,25000,sales

2,BOB,23,30000,sales

3,Maya,23,25000,sales

4,Sara,25,40000,sales

5,David,23,45000,sales

6,Maggy,22,35000,sales

^C

[cloudera@quickstart ~]$ cat >>empbonus.txt

1,Robin,22,25000,sales

2,Jaya,23,20000,admin

3,Maya,23,25000,sales

4,Alia,25,50000,admin

5,David,23,45000,sales

6,Omar,30,30000,admin

Command: -

grunt> empsales = LOAD '/user/cloudera/empsales.txt' USING PigStorage(',') as (sno:int, name:chararray, age:int, salary:int, dept:chararray);

grunt> empbonus = LOAD '/user/cloudera/empbonus.txt' USING PigStorage(',') as (sno:int, name:chararray, age:int, salary:int, dept:chararray);

grunt> cogroup\_data = COGROUP empsales by age, empbonus by age;

grunt> isempty\_data = filter cogroup\_data by IsEmpty(empsales);

grunt> Dump isempty\_data;

Output file: -

(30,{},{(6,Omar,30,30000,admin)})

