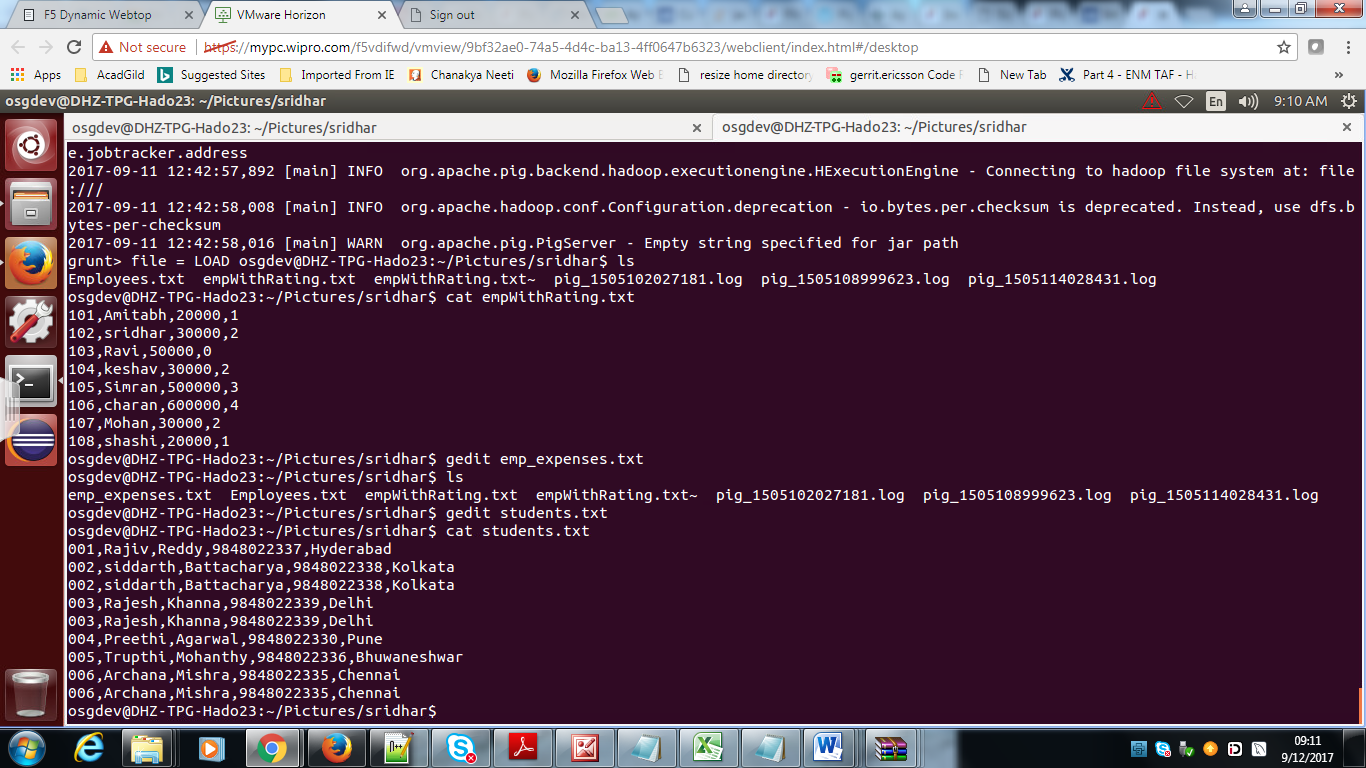
**Session 4: Schedulers in YARN & Introductionto Pig**

**Assignment 2 Question**

**Concat** : The CONCAT() function of Pig Latin is used to concatenate two or more expressions of the same type.

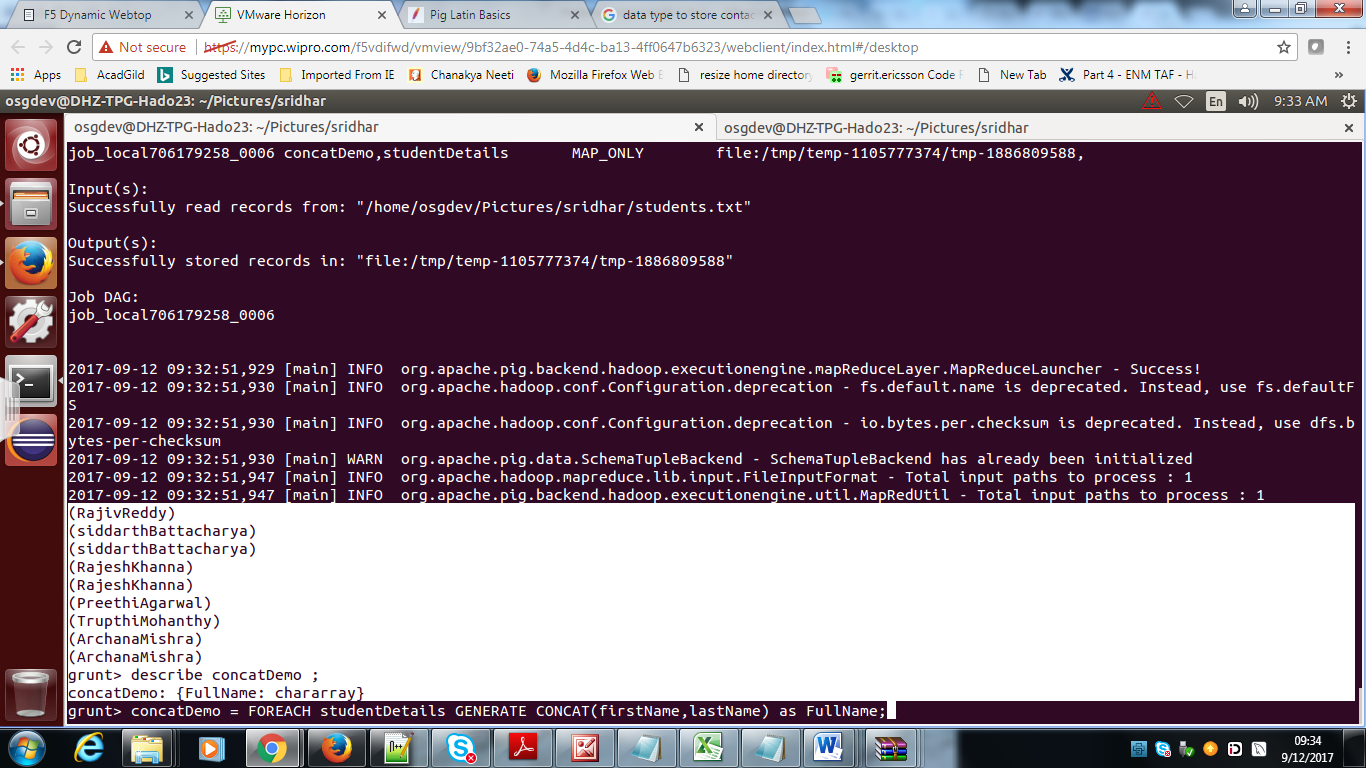
**Command** : concatDemo = FOREACH studentDetails GENERATE CONCAT(firstName,lastName) as FullName;

**Dataset**: students.txt



First we will load the file to a relation using command :

studentDetails = LOAD '/home/osgdev/Pictures/sridhar/students.txt' USING PigStorage(',') as (id : int , firstName : chararray , lastName : chararray , contactNo :bytearray , location : chararray);describe studentDetails

**Output**:

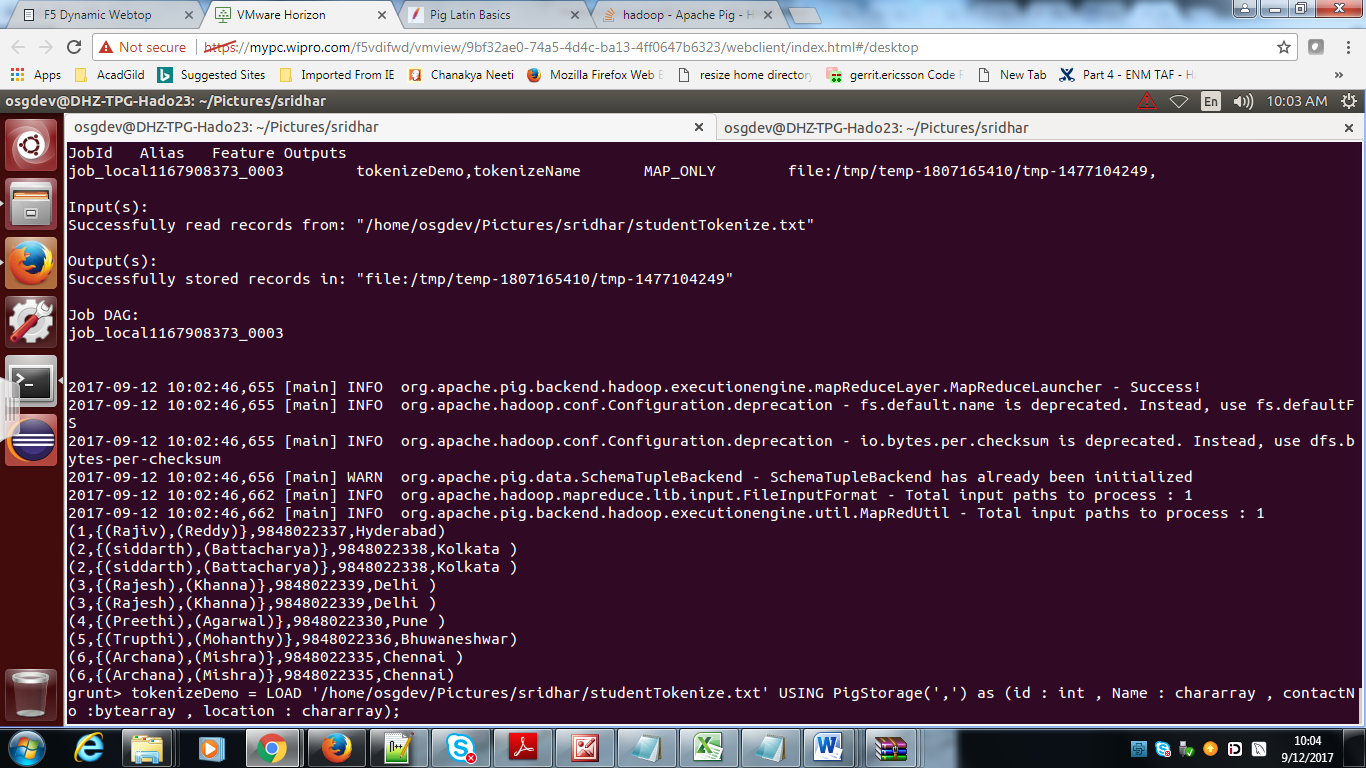
**Tokenize** : The TOKENIZE() function of Pig Latin is used to split a string (which contains a group of words) in a single tuple and returns a bag which contains the output of the split operation.

**Dataset** used:

osgdev@DHZ-TPG-Hado23:~/Pictures/sridhar$ cat studentTokenize.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

tokenizeDemo = LOAD '/home/osgdev/Pictures/sridhar/studentTokenize.txt' USING PigStorage(',') as (id : int , Name : chararray , contactNo :bytearray , location : chararray);describe studentDetails

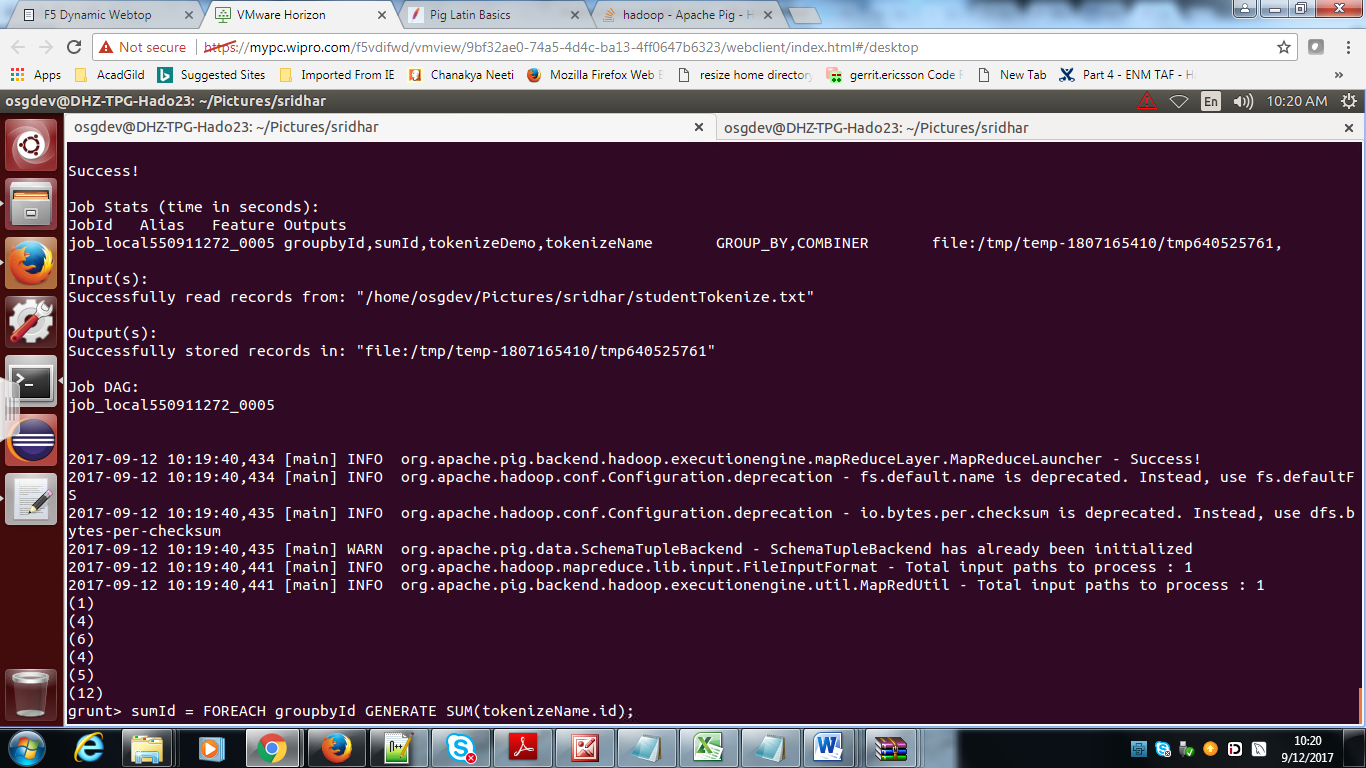
**Command** : tokenizeName = FOREACH tokenizeDemo GENERATE id , TOKENIZE(Name), contactNo , location;

**Output**:

**Sum**: SUM is used to get the total of the numeric values of a column in a single-column bag. While computing the total, the SUM() function ignores the NULL values.

**Command** : sumId = FOREACH groupId GENERATE SUM(tokenizeName.id)

**Output**:



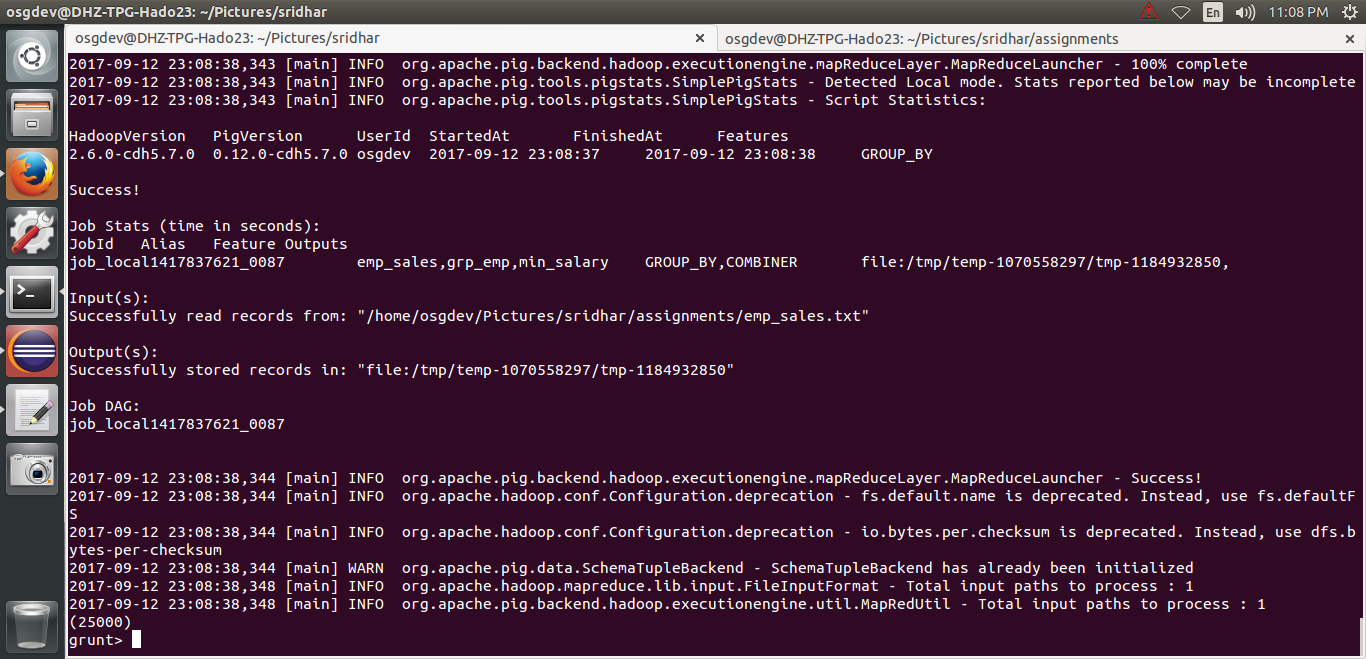
**Min:** The MIN() function of Pig Latin is used to get the minimum (lowest) value (numeric or chararray) for a certain column in a single-column bag. While calculating the minimum value, the MIN() function ignores the NULL values.

**commands**

emp\_sales = LOAD '/home/osgdev/Pictures/sridhar/assignments/emp\_sales.txt' USING PigStorage(',')  
   as (sno:int, name:chararray, age:int, salary:int, dept:chararray);

grp\_emp = GROUP emp\_sales all;

min\_salary = FOREACH grp\_emp GENERATE MIN(emp\_sales.salary);



**Max** **:** The MAX() function of Pig Latin is used to get the maximum (highest) value (numeric or chararray) for a certain column in a single-column bag. While calculating the minimum value, the MIN() function ignores the NULL values.

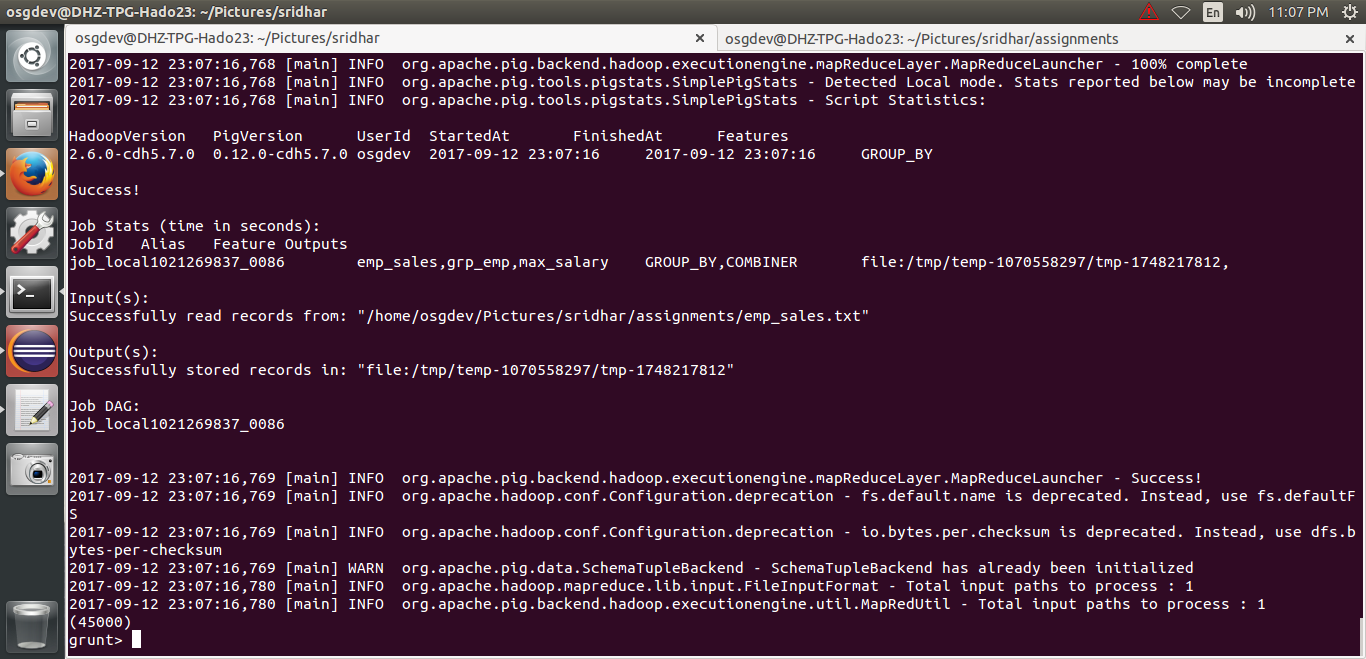
**commands**

emp\_sales = LOAD '/home/osgdev/Pictures/sridhar/assignments/emp\_sales.txt' USING PigStorage(',')  
   as (sno:int, name:chararray, age:int, salary:int, dept:chararray);

grp\_emp = GROUP emp\_sales all;

max\_salary = FOREACH grp\_emp GENERATE MAX(emp\_sales.salary);

**Output**:



**Limit :** The LIMIT operator is used to get a limited number of tuples from a relation.

**Command** : LimitResult = LIMIT countWord 5;

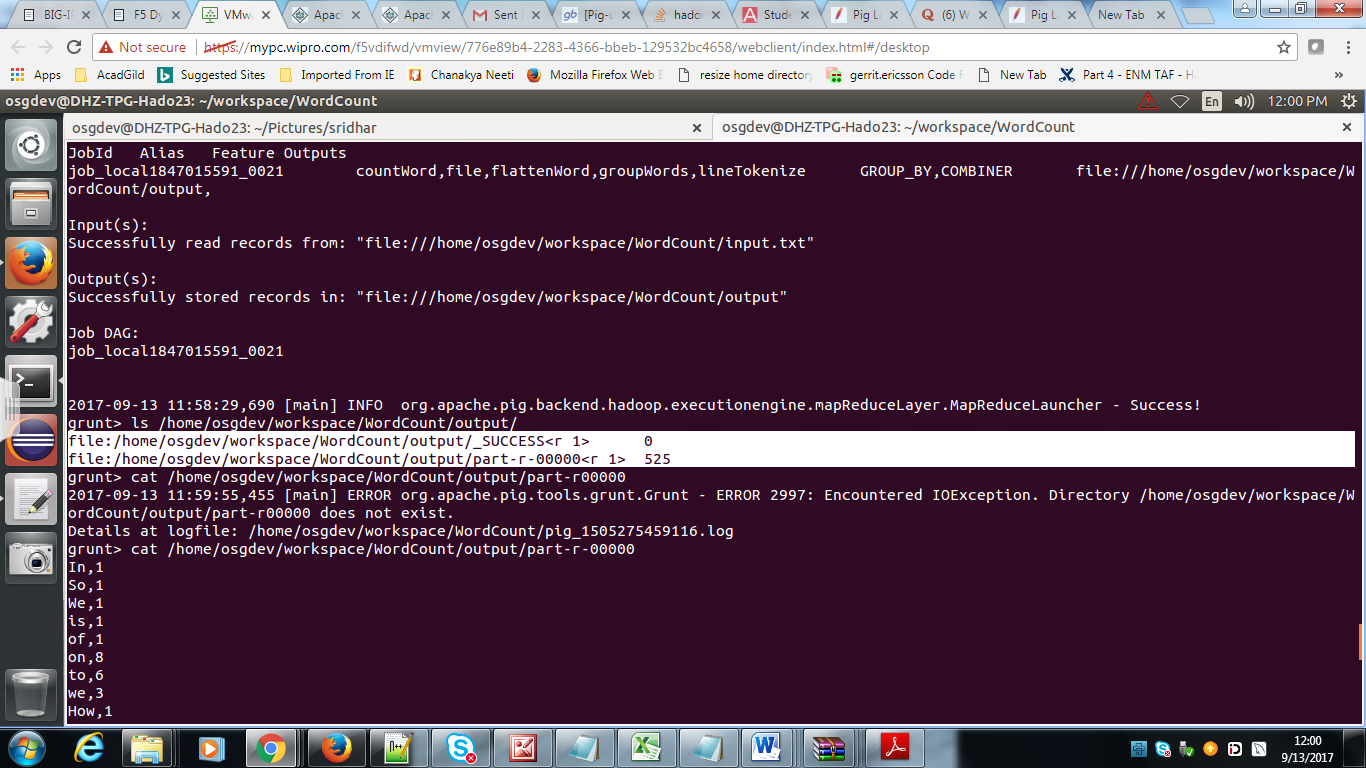
**Output**



**Store**: It is used to store the loaded data in the file system

**Command**: STORE countWord into ‘output/’ USING PigStorage;

**Output**:



**Distinct** : The DISTINCT operator is used to remove redundant (duplicate) tuples from a relation.

**Syntax** : Relation\_name2 = DISTINCT Reklation\_name1;

Data set :

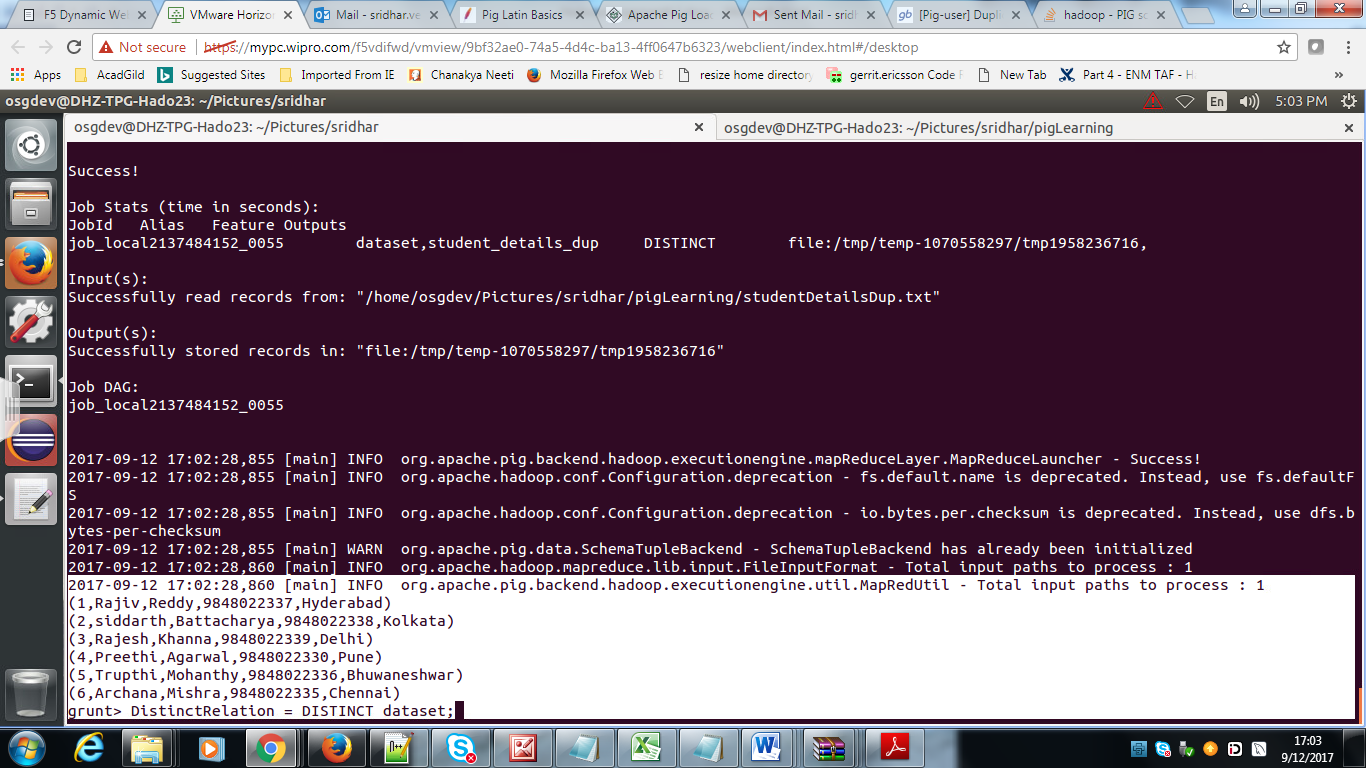
(1,Rajiv,Reddy,9848022337,Hyderabad)  
(2,siddarth,Battacharya,9848022338,Kolkata)  
(2,siddarth,Battacharya,9848022338,Kolkata)  
(3,Rajesh,Khanna,9848022339,Delhi)  
(3,Rajesh,Khanna,9848022339,Delhi)  
(4,Preethi,Agarwal,9848022330,Pune)  
(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)  
(6,Archana,Mishra,9848022335,Chennai)  
(6,Archana,Mishra,9848022335,Chennai)

grunt> describe dataset;  
dataset: {id: int,firstname: chararray,lastname: chararray,phone: chararray,city: chararray}

--**Command** to eliminated duplicate tuples from a relation

DistinctRelation = DISTINCT dataset;

**Output:**

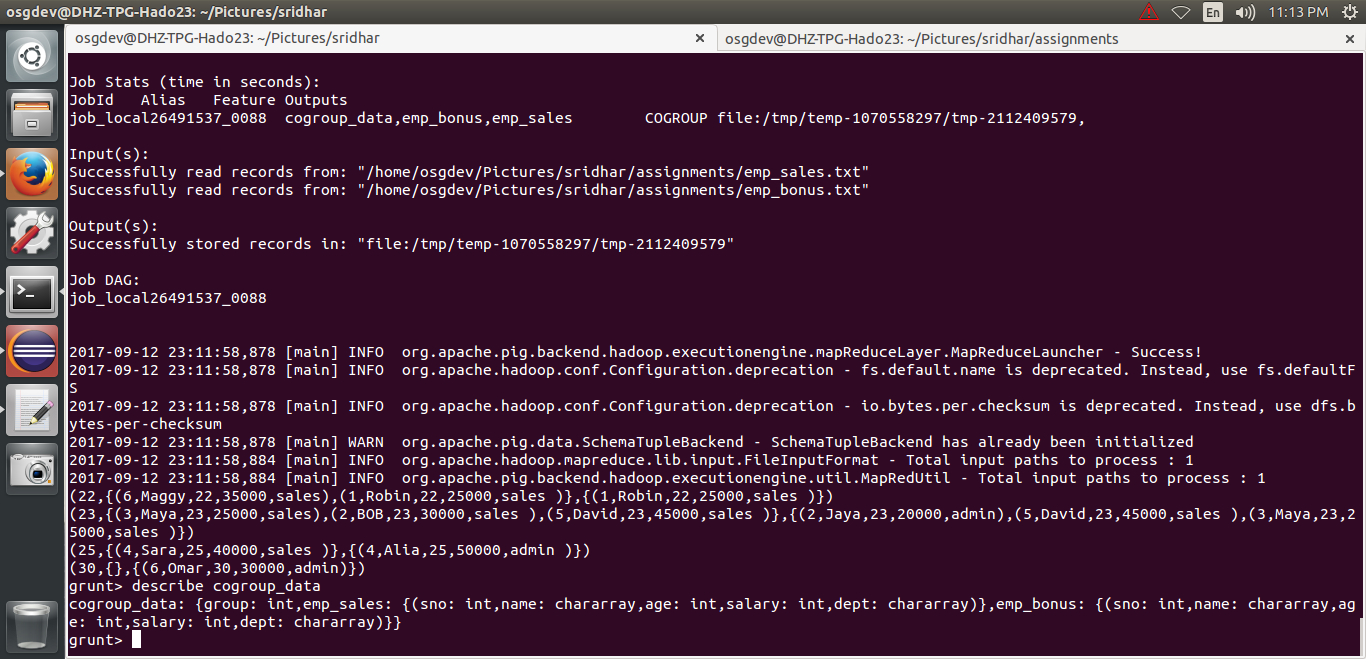


:

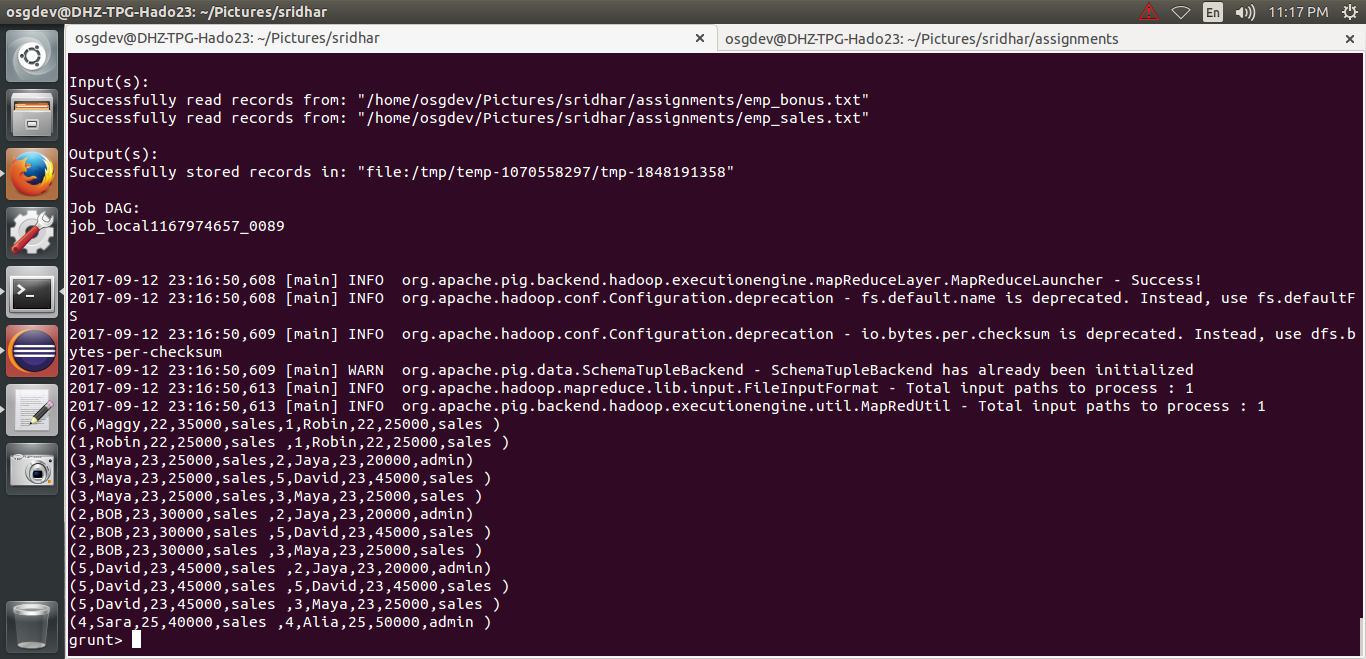
**Flatten** : FLATTEN is used to un-nest tuple or bag.

**command**

 flattenDemo = FOREACH cogroup\_data GENERATE FLATTEN(emp\_sales), FLATTEN(emp\_bonus);

**Dataset**:

**OutPut**:



**IsEmpty** : The IsEmpty() function of Pig Latin is used to check if a bag or map is empty.

The emp\_sales relation holds the tuples that are not there in the relation emp\_bonus.

   emp\_sales = LOAD '/home/osgdev/Pictures/sridhar/assignments/emp\_sales.txt' USING PigStorage(',')  
   as (sno:int, name:chararray, age:int, salary:int, dept:chararray);

emp\_bonus = LOAD '/home/osgdev/Pictures/sridhar/assignments/emp\_bonus.txt' USING PigStorage(',')  
   as (sno:int, name:chararray, age:int, salary:int, dept:chararray);

  cogroup\_data = COGROUP emp\_sales by age, emp\_bonus by age;  
   f1 = FILTER cogroup\_data by IsEmpty(emp\_sales);

