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

1. **concat**

The concat () function of pig latin is used to concatenate two or more expressions of same type

Assume that we have a file named student\_details.txt with fields like

**001,Rajiv,Reddy,21,9875627192,Hyderabad,89**



**Syntax:**

Student\_name\_concat = foreach student\_deatils Generate CONCAT (firstname,lastname);



1. **Tokenize**

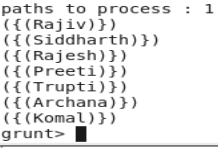
Tokenize function can be used to split a string. For example, let us split the above student\_details.txt using the following command.



**Syntax:**

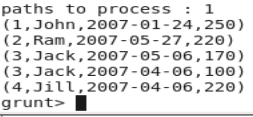
**student\_name\_tokenize = foreach student\_details Generate TOKENIZE(firstname);**

**dump student\_name\_tokenize**



1. **Sum**

We have a file name employee.txt under /home/acadgild/pig. Running dump on employee.txt will give the following output



**Load Syntax:**

employee\_data = LOAD '/home/acadgild/pig/employee.txt' USING PigStorage(',') as (id:int, name:chararray, workdate:chararray, daily\_typing\_pages:int);

**Group Syntax:**

employee\_group = Group employee\_data all;

**Sum Syntax:**

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

C:\Users\p.muthurajaiah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\sum.png

1. **Min**

The min function of pig Latin is used to get the minimum value for a certain column in a single column bag. When calculating the min value NULL is ignored.

Student Details dataset:



**Min() Syntax:**

**student\_group\_all = Group student\_details All;**

**student\_gpa\_min = foreach student\_group\_all Generate**

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

C:\Users\p.muthurajaiah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\min.png

**5. Max**

The max function of pig Latin is used to get the maximum value for a certain column in a single column bag. When calculating the max value NULL is ignored.

Student Details dataset:



**Max() Syntax:**

**student\_group\_all = Group student\_details All;**

**student\_gpa\_max = foreach student\_group\_all Generate**

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

C:\Users\p.muthurajaiah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\max.png

**6. Limit**

The limit operator is used to get a limited number of tuples from a relation.

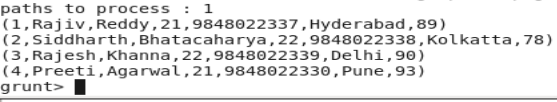
**Syntax:**

Result = LIMIT Relation\_name required number of tuples;

Student Details dataset:



C:\Users\p.muthurajaiah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Limit.png



**7. Store**

We can store the loaded data in the file system using the store operator.

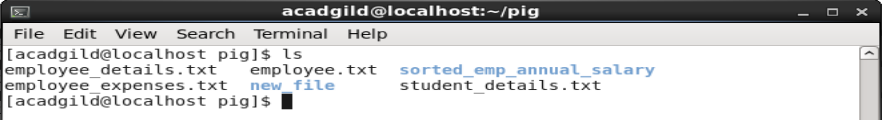
Given below is the syntax of store operator

STORE Relation\_name INTO ‘required directory path’ [using function];

Student\_details.txt dataset



STORE student\_details INTO ' /home/acadgild/new\_file ' USING PigStorage (',');



**8. Distinct**

The distinct operator is used to remove redundant information (duplicate) tuples from a relation.

**Syntax:**

Relation\_name2 = DISTINCT Relatin\_name1;

We have a file named student\_details



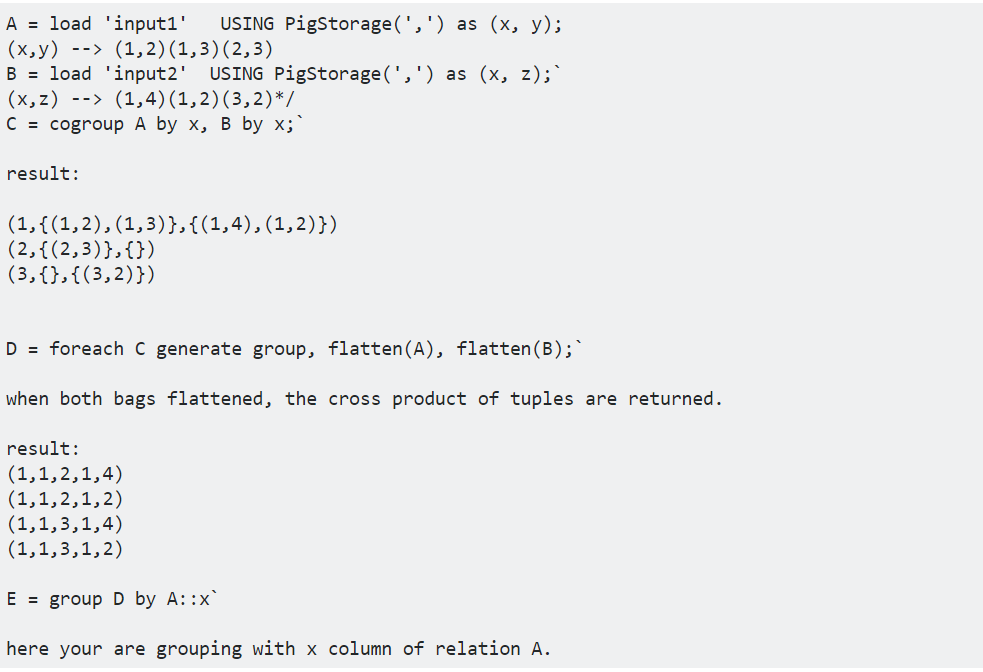
C:\Users\p.muthurajaiah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\distinct.png

**Since there are no duplicate entries in the relation all the rows/entries are displayed.**



**9. Flatten**

The FLATTEN operator looks like a UDF syntactically, but it is an operator that changes the structure of tuples and bags in a way that a UDF cannot. Flatten un-nests tuples as well as bags. The idea is the same, but the operation and result is different for each type of structure



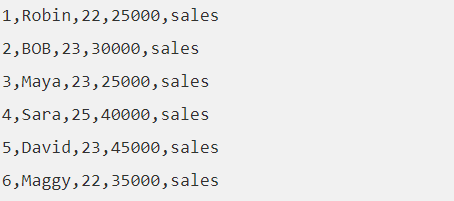
**10. IsEmpty**

The IsEmpty() function of pig latin is used to check if a bag or map is empty.

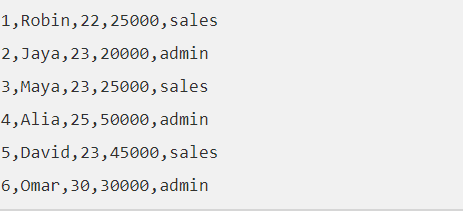
**Syntax:**

IsEmpty(Expression)

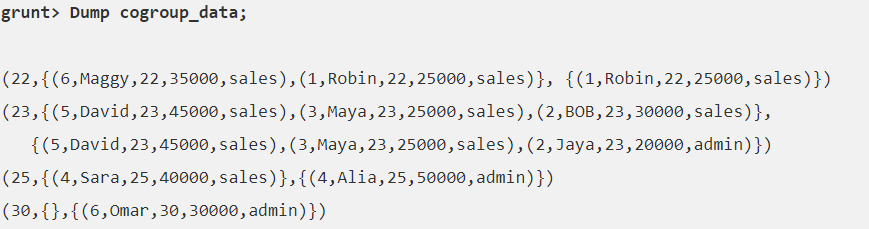
Emp\_sales.txt



Emp\_bonus.txt



**cogroup\_data = COGROUP emp\_sales by age, emp\_bonus by age;**



**grunt> isempty\_data = filter cogroup\_data by IsEmpty(emp\_sales);**

