Consider a file named “names.txt” having entries of few names, these names are repeated as shown in the following image:



The problem statement is to find out how many times each word(names) have appeared in the file.

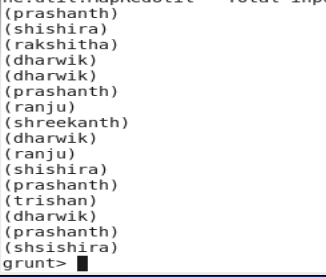
This can be done by writing a pig script which in turn converts it into a MapReduce program

The first step is to load the file into a relation. This can be done using the Load command.

Syntax: LOAD ‘/home/acadgild/pig/names.txt’ AS (line:chararaay);

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

The above command loads the names.txt into a relation named “Lines” .The relation will be formed as following:



Flatten command is applied to each line in the relation. The bag is converted to a tuple by this command, The syntax is as follows:

Words = FOREACH LINES GENERATE FLATTEN (TOKENIZE(line)) as word;

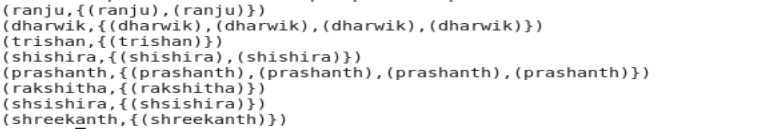
C:\Users\p.muthurajaiah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\linesnew.png

In the next step the words are grouped together so the count logic can be applied on them.

Syntax: grouped = GROUP Words by word;

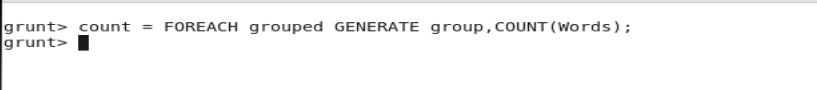
C:\Users\p.muthurajaiah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\group.png

Group command output is as follows:



Now we can apply count logic on the grouped words.

Syntax: count = FOREACH grouped GENERATE group, COUNT(Words);



The final output shows the count of each words in the file.

