Word Count in Pig Latin

hadooplessons.info/2015/01/word-count-in-pig-latin.html

In this Post, we learn how to write word count program using Pig Latin.

Assume we have data in the file like below.

This is a hadoop post
hadoop is a bigdata technology
and we want to generate output for count of each word like below

```
(a,2)
(is,2)
(This,1)
(class,1)
(hadoop,2)
(bigdata,1)
(technology,1)
```

Now we will see in steps how to generate the same using Pig latin.

1.Load the data from HDFS

Use Load statement to load the data into a relation .
As keyword used to declare column names, as we dont have any columns, we declared only one column named line.

```
input = LOAD '/path/to/file/' AS(line:Chararray);
```

2. Convert the Sentence into words.

The data we have is in sentences. So we have to convert that data into words using
TOKENIZE Function.

```
(TOKENIZE(line));
```

(or)
If we have any delimeter like space we can specify as

```
(TOKENIZE(line,' '));
```

Output will be like this:

```
({(This),(is),(a),(hadoop),(class)})
({(hadoop),(is),(a),(bigdata),(technology)})
```

but we have to convert it into multiple rows like below

```
(This)
(is)
(a)
(hadoop)
```

```
(class)
(hadoop)
(is)
(a)
(bigdata)
(technology)
```

3.Convert Column into Rows

I mean we have to convert every line of data into multiple rows ,for this we have function called FLATTEN in pig.

Using FLATTEN function the bag is converted into tuple, means the array of strings converted into multiple rows.

```
Words = FOREACH input GENERATE FLATTEN(TOKENIZE(line,' ')) AS word;
```

Then the ouput is like below

```
(This)
(is)
(a)
(hadoop)
(class)
(hadoop)
(is)
(a)
(bigdata)
(technology)
```

3. Apply GROUP BY

We have to count each word occurance, for that we have to group all the words.

```
Grouped = GROUP words BY word;
```

4. Generate word count

```
wordcount = FOREACH Grouped GENERATE group, COUNT(words);
```

We can print the word count on console using Dump.

```
DUMP wordcount;
```

Output will be like below.

```
(a,2)
(is,2)
(This,1)
(class,1)
(hadoop,2)
(bigdata,1)
(technology,1)
```

Below is the complete program for the same.

```
input = LOAD '/path/to/file/' AS(line:Chararray);
Words = FOREACH input GENERATE FLATTEN(TOKENIZE(line,' ')) AS word;
Grouped = GROUP words BY word;
wordcount = FOREACH Grouped GENERATE group, COUNT(words);
```

You may check same word count using Hive .