

Assignment 6

In the next two assignments the codes that are already developed in JAVA for map-reduce will be developed in Pig-Latin

Q1) WordCount Problem

```
A1 = load 'hdfs://localhost/user/cloudera/pig/UN.txt' as (line:chararray);
A2 = foreach A1 generate TOKENIZE(line) as tokens;
A3 = foreach A2 generate flatten(tokens) as words;
A4 = group A3 by words;
A5 = foreach A4 generate group, COUNT(A3);
A5 = order A5 by $1;
dump A5;
store A5 into 'hdfs://localhost/user/cloudera/pig/results/WordCount';|
```

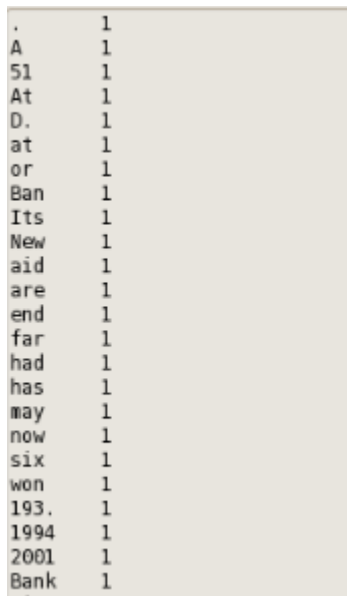
A1 -> it loads the data from given path into Relation A1.

A2 -> It tokenizes all the words based on default space delimiter.

A3 -> Flatten as discussed removes the nesting. In this case it removes individual elements from the tuples returned by TOKENIZE

A4 -> Grouping of same words (basically \$0).

A5 -> Count is generated for each output of group and then ordered based on count such that word with least count is on top.

A screenshot of a terminal window showing the output of the Pig script. The output consists of a list of words and their corresponding counts, ordered by count in descending order. The words and counts are: . 1, A 1, 51 1, At 1, D. 1, at 1, or 1, Ban 1, Its 1, New 1, aid 1, are 1, end 1, far 1, had 1, has 1, may 1, now 1, six 1, won 1, 193. 1, 1994 1, 2001 1, Bank 1.

.	1
A	1
51	1
At	1
D.	1
at	1
or	1
Ban	1
Its	1
New	1
aid	1
are	1
end	1
far	1
had	1
has	1
may	1
now	1
six	1
won	1
193.	1
1994	1
2001	1
Bank	1

Q2) Calculating Subscriber's downloaded bytes for each individual id.

```
A = load 'hdfs://localhost/user/cloudera/pig/Data_File.txt' as (line:chararray);
B = foreach A generate (chararray)SUBSTRING(line,15,26) as id, (double)SUBSTRING(line,87,97) as bytes;
C = group B by id;
D = foreach C generate group, SUM(B.bytes);
dump D;
store D into 'hdfs://localhost/user/cloudera/pig/results/Subscribers_Data';|
```

A ->It loads the data from given path into Relation A.

B ->Foreach output of A (chararray) , SUBSTRING is taken out.

Substring from 15-26 represents customer id while substring from 87-97 represents downloaded bytes.

C ->Grouping of output from B is done based on id.

D ->Group outputs a bag with a collection of tuples. These tuples have same id but different data downloaded. Sum of these downloaded data is done.

11128052609	4.4564564562E11
11128052610	4.09569569546E11
11128052611	3.98958958936E11
11128052612	4.18058058034E11
11128052613	4.11691691668E11
11128052614	4.28668668644E11
11128052615	4.47767767742E11
11128052616	4.20180180156E11
11128052617	4.244244244E11
11128052618	4.0320320318E11
11128052619	3.86226226204E11
11128052620	4.05325325302E11
11128052621	4.1381381379E11
11128052622	4.30790790766E11
11128052623	3.56516516496E11
11128052624	4.26546546522E11
11128052625	3.84104104082E11
11128052626	3.94714714692E11
11128052627	3.9259259257E11
11128052628	4.43523523498E11
11128052629	3.65005004984E11
11128052630	4.22302302278E11
11128052631	4.43523523498E11
11128052632	3.98958958936E11
11128052633	4.30790790766E11

Q3) Sorting above collected data as per Data Downloaded

Basically above output has to be ordered using order-by command on second field.

```
A = load 'hdfs://localhost/user/cloudera/pig/Data_File.txt' as (line:chararray);
B = foreach A generate (chararray)SUBSTRING(line,15,26) as id, (double)SUBSTRING(line,87,97) as bytes;
C = group B by id;
D = foreach C generate group, SUM(B.bytes);
E = foreach D generate $1 as bytes,$0 as id;
F = order E by bytes desc;
dump F;
store F into 'hdfs://localhost/user/cloudera/pig/results/Subscribers_Data_bytes';|
```

Ouput:

```
2014-02-03 22:00:39,442 [main] INFO org.apache.pig
2014-02-03 22:00:39,447 [main] INFO org.apache.hadoop
2014-02-03 22:00:39,447 [main] INFO org.apache.pig
(4.79599599572E11,11128052635)
(4.6686686684E11,11128052645)
(4.64744744718E11,11128052649)
(4.64744744718E11,11128052646)
(4.58378378352E11,11128052657)
(4.58378378352E11,11128052650)
(4.5625625623E11,11128052647)
(4.52012011986E11,11128052641)
(4.47767767742E11,11128052615)
(4.4564564562E11,11128052609)
(4.43523523498E11,11128052628)
(4.43523523498E11,11128052631)
(4.3503503501E11,11128052639)
(4.30790790766E11,11128052622)
(4.30790790766E11,11128052633)
(4.28668668644E11,11128052614)
(4.26546546522E11,11128052644)
(4.26546546522E11,11128052624)
(4.26546546522E11,11128052637)
(4.244244244E11,11128052653)
(4.244244244E11,11128052656)
(4.244244244E11,11128052617)
(4.22302302278E11,11128052630)
(4.20180180156E11,11128052616)
(4.20180180156E11,11128052654)
(4.18058058034E11,11128052612)
(4.1381381379E11,11128052621)
(4.11691691668E11,11128052634)
(4.11691691668E11,11128052613)
(4.09569569546E11,11128052651)
(4.09569569546E11,11128052610)
```

Q4) Relational join

```
ID = load 'hdfs://localhost/user/cloudera/pig/id' using PigStorage('\t') as (id:int, name:chararray, designation:chararray);
TRIPS = load 'hdfs://localhost/user/cloudera/pig/trips' using PigStorage('\t') as (id:int, place:chararray, roundtrips:int);
JOINED = join ID by id, TRIPS by id;
FINAL = foreach JOINED generate $0,$1,$2,$4,$5;
dump FINAL;
store FINAL into 'hdfs://localhost/user/cloudera/pig/results/Rel_Join';|
```

- 1) First Data is loaded in 2 relations from 2 files.
- 2) These relations are then joined by id.
- 3) From the joined data 4th column is excluded because it's a repetition of first column.

Output:

```
2014-02-03 21:59:17,601 [main] INFO org.apache.pig.backend.hado
2014-02-03 21:59:17,606 [main] INFO org.apache.hadoop.mapreduce
2014-02-03 21:59:17,606 [main] INFO org.apache.pig.backend.hado
(101,aaa,executive,pune,1)
(101,aaa,executive,hyd,2)
(102,bbb,manager,pune,2)
(102,bbb,manager,hyd,3)
(102,bbb,manager,bang,4)
(103,ddd,manager,bang,5)
(103,ddd,manager,pune,2)
(103,ddd,manager,hyd,3)
(104,hhh,manager,hyd,4)
(104,hhh,manager,bang,5)
(104,hhh,manager,pune,2)
(105,bbb,executive,pune,2)
(105,bbb,executive,hyd,4)
(106,ccc,trainee,pune,1)
(107,eee,trainee,pune,2)
(108,ggg,president,chen,1)
(108,ggg,president,del,1)
(108,ggg,president,pune,2)
(108,ggg,president,hyd,3)
(108,ggg,president,bang,4)
(109,hhh,trainee,pune,1)
(110,fff,vice-president,hyd,3)
(110,fff,vice-president,bang,5)
(110,fff,vice-president,del,2)
(110,fff,vice-president,pune,2)
```