Exercise 04: Word Count using Pig Grouping

Here, we will be running Apache Pig Sample scripts using grunts. It is to just see the power of Apache Pig.

Step 1A: Start Grunt shell.

Open terminal and type *pig*

Step 1B: Create a file at /user/cloudera/pigfile.txt With following content.

I am learning Pig Using cloudera
I am learning Spark Using cloudera
I am learning Java Using cloudera

Step 2 : Load the file stored in hdfs with variable 'in1' and each line have to store in 'line' (Space separated file)

(I am learning Pig Using cloudera)
(I am learning Spark Using cloudera)
(I am learning Java Using cloudera)

Step 3: flatten the words in each line from variable 'in1' and save separated words into variable 'wordsinline'

grunt>wordsinline = FOREACH input1 GENERATE flatten(TOKENIZE(line, ' ')) as word;
grunt>DUMP wordsinline;

Step 4: Group the similar words and save into variable 'groupwords'

grunt>groupwords = _____ wordsinline by word;
grunt>dump groupwords;
grunt>describe groupwords;

Step 5: Count Words in the group.

grunt>countwords = foreach _____;
grunt>DUMP countwords;