EXPERIMENT 6

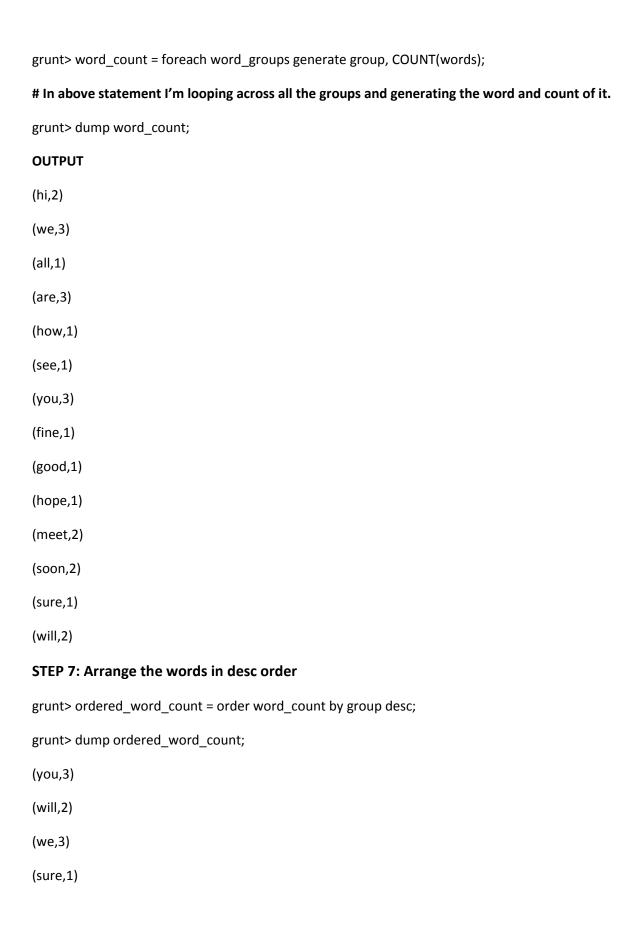
WORD COUNT IN PIG

Aim: To perform word count on a text file using pig latin commands
Objective:
To perform word count on a text file using functions like tokenize and flatten
STEP 1: Create a file in local on which you want to perform word count
In my case I have made a file called pig_word with the following contents
hi how are you
hope you are fine
we are all good
hi we will meet
see you soon
sure soon we will meet
STEP 2: Copy the file in HDFS
[cloudera@quickstart Desktop]\$ hadoop fs -put /user/cloudera/ pig_word
STEP 3: Load the data in pig
#Use load operator to load the text file into pig and I will name each line as line and its data type is chararray
grunt> input1 = load '/user/cloudera/pig_word' as (line:chararray);
grunt> dump input1;
OUTPUT
(hi how are you)
(hope you are fine)
(we are all good)

(hi we will meet)
(see you soon)
(sure soon we will meet)
STEP 4: Condense all the tuples in each line to one single line using function FLATTEN and then break the line into words using TOKENIZE function
grunt> words= foreach input1 generate FLATTEN(TOKENIZE(line)) as word;
grunt> dump words;
(hi)
(how)
(are)
(you)
(hope)
(you)
(are)
(fine)
(we)
(are)
(all)
(good)
(hi)
(we)
(will)
(meet)
(see)
(you)
(soon)

```
(sure)
(soon)
(we)
(will)
(meet)
STEP 5: Now group the collection of words based on word
grunt> word_groups = group words by word;
grunt> dump word_groups;
OUTPUT
(hi,{(hi),(hi)})
(we,{(we),(we),(we)})
(all,{(all)})
(are,{(are),(are),(are)})
(how,{(how)})
(see,{(see)})
(you,{(you),(you),(you)})
(fine,{(fine)})
(good,{(good)})
(hope,{(hope)})
(meet,{(meet),(meet)})
(soon,{(soon),(soon)})
(sure,{(sure)})
(will,{(will),(will)})
```

STEP 6: Determine the count of each word



```
(soon,2)
(see,1)
(meet,2)
(how,1)
(hope,1)
(hi,2)
(good,1)
(fine,1)
(are,3)
(all,1)
STEP 8: Store the above result in HDFS
grunt> store ordered_word_count into '/user/cloudera/pig_word_output1';
[cloudera@quickstart Desktop]$ hadoop fs -ls /user/cloudera/pig_word_output1/
OUTPUT
Found 2 items
-rw-r--r-- 1 cloudera cloudera
                                 0 2018-09-19 22:47 /user/cloudera/pig_word_output1/_SUCCESS
                                 89 2018-09-19 22:47 /user/cloudera/pig_word_output1/part-r-
-rw-r--r-- 1 cloudera cloudera
00000
[cloudera@quickstart Desktop]$ hadoop fs -cat /user/cloudera/pig_word_output1/part-r-00000
OUTPUT
(you,3)
(will,2)
(we,3)
(sure,1)
(soon,2)
```

(see,1)

(meet,2)

(how,1)

(hope,1)

(hi,2)

(good,1)

(fine,1)

(are,3)

(all,1)