Word Count Using Pig Command

Input file: Assume we have data input data

This is a hadoop post  
hadoop is a bigdata technology

and we want to generate output for count of each word like below  
 Output file:  
(a,2)  
(is,2)  
(This,1)  
(class,1)  
(hadoop,2)  
(bigdata,1)  
(technology,1)

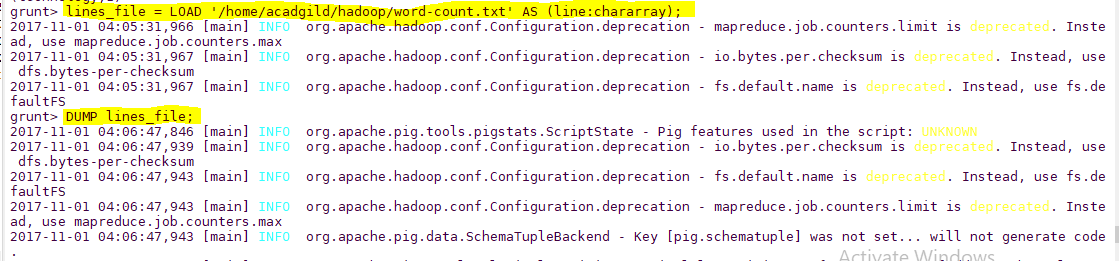
Step1: Load the data

Command: lines\_file = LOAD '/home/acadgild/hadoop/word-count.txt' AS (line:chararray);

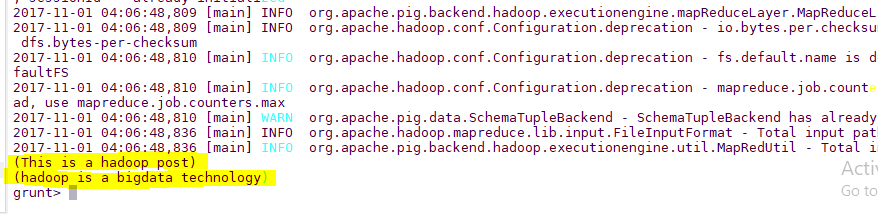
DUMP lines\_file;

Screen-shot:

Command Screen-shot



Output



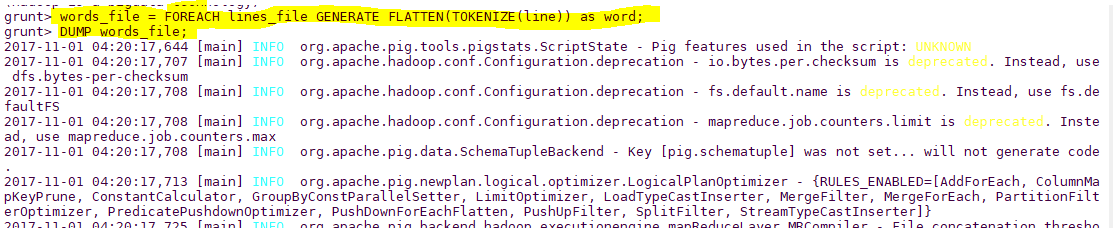
Step2:

Command: words\_file = FOREACH lines\_file GENERATE FLATTEN(TOKENIZE(line)) as word;

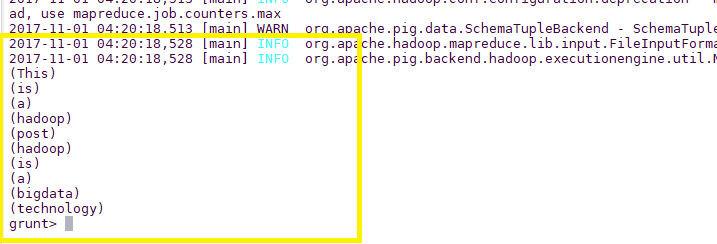
DUMP words\_file;

Screen-shot

Command



Output



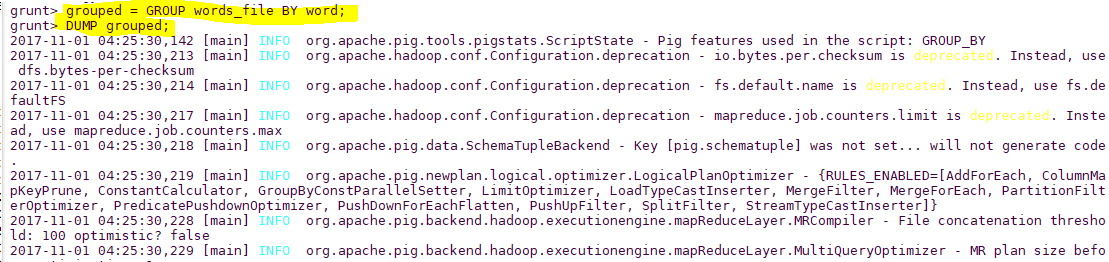
Step3:

Command: grouped = GROUP words\_file BY word;

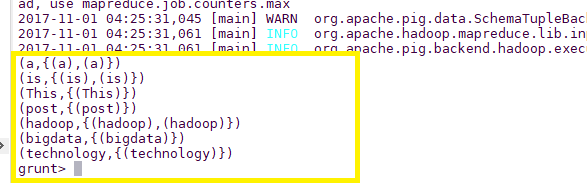
DUMP grouped;

Screen-shot:

Commands Screen-shot



Output



Step4:

Command:

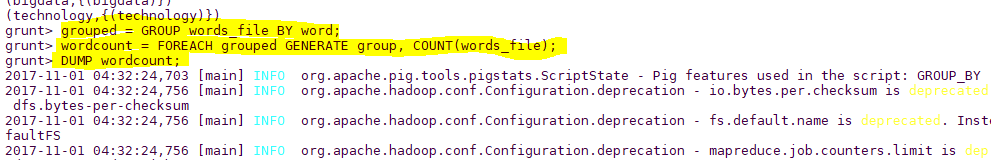
grouped = GROUP words\_file BY word;

wordcount = FOREACH grouped GENERATE group, COUNT(words\_file);

DUMP wordcount;

Screen-shot

Command Screen-shot



Output

