

Assignment 7.1

Program to find word count using pig

```
grunt> Line = LOAD '/Assignment1.txt' using PigStorage AS (Line:chararray);
//Loading the file Assigmenmt1.txt to the bag Lines

grunt> words = FOREACH Line GENERATE TOKENIZE(Line);

// creating a bag called words to tokenize the Lines

grunt> words = FOREACH Line GENERATE FLATTEN(TOKENIZE(Line));

//to consider each line as a token.

grunt> words = FOREACH Line GENERATE FLATTEN(TOKENIZE(Line)) As wording;

//give alias to the flattened output

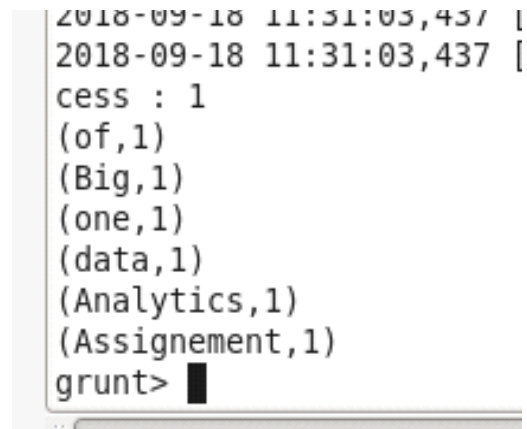
grunt> grouped = Group words by wording;

// grouping the scatterd words

grunt> wordcount = FOREACH grouped GENERATE group, COUNT(words);

// command for generating the count

grunt> dump wordcount;
```



```
2018-09-18 11:31:03,437 [
2018-09-18 11:31:03,437 [
cess : 1
(of,1)
(Big,1)
(one,1)
(data,1)
(Analytics,1)
(Assigment,1)
grunt> █
```

Task 2:

Data sets are not accessible the link gives error 404.

TASK 3 - getting an error - not able to read delayedflights.csv file.. tried multiple ways but the same error. was not able to debug

```
grunt> REGISTER '/home/acadgild/piggybank.jar';
grunt> A = LOAD '/home/acadgild/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-09-18 15:29:20,625 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. In:
d, use fs.defaultFS
grunt> B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17 as origin,(chararray)$18 as dest;
grunt> C = filter B by dest is not null;
grunt> D = group C by dest;
grunt> E = for each D generate group, COUNT(C.dest);
2018-09-18 15:36:22,260 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 13, column 0> Syntax error, un-
ceted symbol at or near 'E'
Details at logfile: /home/acadgild/pig_1537258086637.log
grunt> E = foreach D generate group, COUNT(C.dest);
grunt> F = order E by $1 DESC;
grunt> RESULT = LIMIT F 5;
grunt> A1 = LOAD '/home/acadgild/airplanes.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','NO_MULTILINE'
IX','SKIP_INPUT_HEADER');
2018-09-18 15:39:46,856 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. In:
d, use fs.defaultFS
grunt> A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;
grunt> joined_table = join Result by $0, A2 by dest;
2018-09-18 15:43:06,314 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse:
<line 18, column 20> Undefined alias: Result
Details at logfile: /home/acadgild/pig_1537258086637.log
grunt> joined_table = join RESULT by $0, A2 by dest;
grunt> dump joined_table;
```

```
Input(s):
Failed to read data from "/home/acadgild/DelayedFlights.csv"
```

```
Output(s):
```

```
Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
Job DAG:
job_1537247019915_0013 -> null,
null -> null,
null -> null,
null -> null,
null
```

