



Session 4: Schedulers in YARN & Introduction to Pig

Assignment 3 Question

# Problem Statement

*Write a program to implement wordcount using Pig.*

grunt> A = load ' /user/cloudera/wordcount.txt';

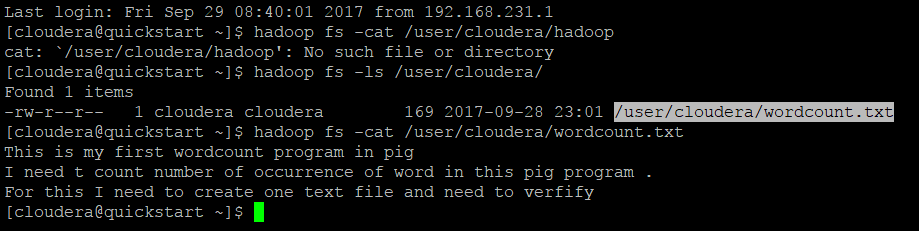
grunt> B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;

grunt> C = filter B by word matches '\\w+';

grunt> D = group C by word;

grunt> E = foreach D generate COUNT(C), group;

grunt> DUMP E;



# Output

HadoopVersion PigVersion UserId StartedAt FinishedAt Features

2.6.0-cdh5.12.0 0.12.0-cdh5.12.0 cloudera 2017-09-29 08:43:39 2017-09-29 08:48:05 GROUP\_BY,FILTER

Success!

Job Stats (time in seconds):

JobId Maps Reduces MaxMapTime MinMapTIme AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs

job\_1506689790344\_0001 1 1 76 76 76 76 18 18 18 18 A,B,C,D,E GROUP\_BY,COMBINER hdfs://quickstart.cloudera:8020/tmp/temp-154493844/tmp1324684462,

Input(s):

Successfully read 3 records (547 bytes) from: "/user/cloudera/wordcount.txt"

Output(s):

Successfully stored 25 records (311 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-154493844/tmp1324684462"

Counters:

Total records written : 25

Total bytes written : 311

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job\_1506689790344\_0001

2017-09-29 08:48:06,034 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2017-09-29 08:48:06,105 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

2017-09-29 08:48:06,106 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address

2017-09-29 08:48:06,107 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.

2017-09-29 08:48:06,168 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2017-09-29 08:48:06,168 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(2,I)

(1,t)

(2,in)

(1,is)

(1,my)

(2,of)

(2,to)

(1,For)

(1,and)

(1,one)

(2,pig)

(1,This)

(1,file)

(3,need)

(1,text)

(2,this)

(1,word)

(1,count)

(1,first)

(1,create)

(1,number)

(2,program)

(1,verfify)

(1,wordcount)

(1,occurrence)

grunt>