Homework 11 By Eric Brown
Course: DAAN 825

Analyzing Text Data with Apache Pig and Hadoop

In this project, I use **Apache Pig** and **Hadoop** to analyze a set of text data (a collection of poems). This assignment showcases how Pig, a high-level language for processing big data, simplifies data analysis on large datasets without needing to write complex code.

What This Project Does

1. **Counting Words and Grouping by Length**: Using a scripting language called Pig Latin, I was able to count words in each poem based on their lengths. For example, I was able to find out how many 3-letter, 4-letter, or 5-letter words appear in each poem. This type of information can be useful for text analysis and understanding the structure of the language in each poem.

2. **How Pig Transforms Data Efficiently**: Apache Pig automatically converts the Pig Latin script into a series of **MapReduce** tasks. This means that instead of writing detailed code for data processing, Pig takes care of the complexity. Each task (like grouping words by length) runs across multiple computers in a Hadoop cluster, allowing it to handle large datasets efficiently.

Why Apache Pig?

Apache Pig is a powerful tool for processing big data because:

- It makes complex data transformations simpler with a scripting language.
- It can process massive amounts of data quickly by distributing tasks over many computers.
- It’s widely used for analyzing data in industries like web analytics, social media, and scientific research.

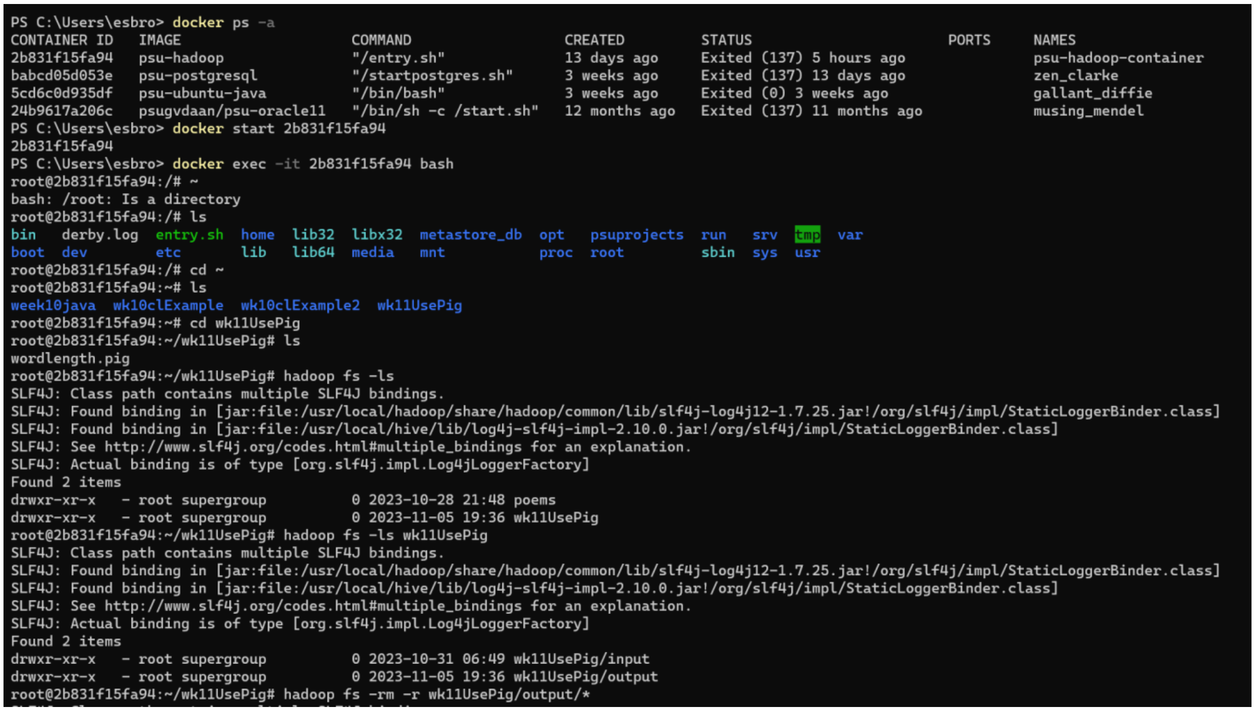
This project provides a look at how Apache Pig can be used to analyze text data easily and quickly, demonstrating the power of high-level data processing on Hadoop.

1.) Run the example on the poems data in your Hadoop server

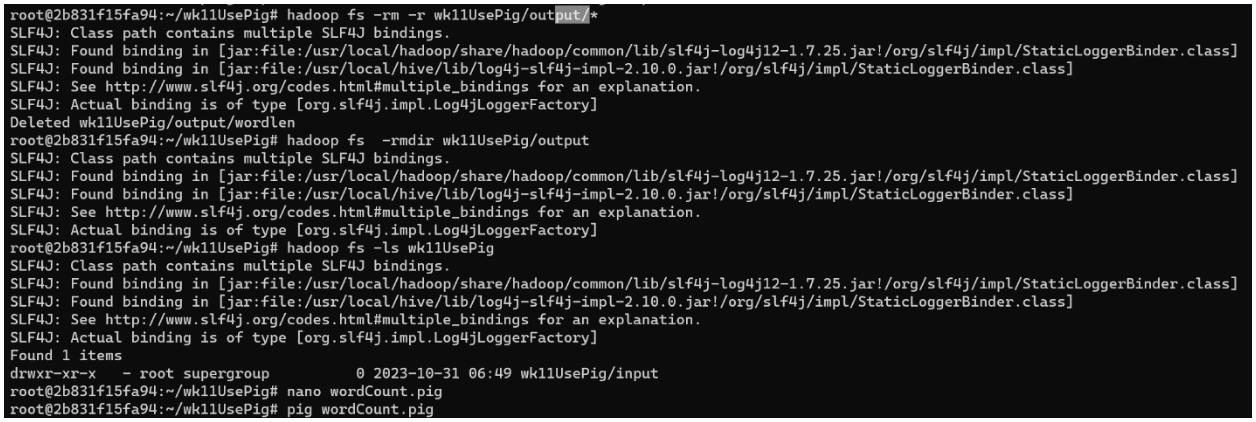
Program Code:

```
p1 = LOAD 'poems/input/Poem1.txt' USING TextLoader AS(line:Chararray);
p2 = LOAD 'poems/input/Poem2.txt' USING TextLoader AS(line:Chararray);
p3 = LOAD 'poems/input/Poem3.txt' USING TextLoader AS(line:Chararray);
p4 = LOAD 'poems/input/Poem4.txt' USING TextLoader AS(line:Chararray);
p5 = LOAD 'poems/input/Poem5.txt' USING TextLoader AS(line:Chararray);
p6 = LOAD 'poems/input/Poem6.txt' USING TextLoader AS(line:Chararray);
p = UNION p1, p2, p3, p4, p5, p6;

words = foreach p generate flatten(TOKENIZE(line , ' ,!?\t\n\r\f\\.\\-')) as word;
words_lower = foreach words generate LOWER(word) as word_lower;
words_unique = group words_lower by word_lower;

words_with_count = foreach words_unique generate COUNT(words_lower) as cnt, group;
words_with_count_sorted = ORDER words_with_count BY cnt DESC, group;

final_result = RANK words_with_count_sorted;
store final_result into 'poems/output/wordcount1';
```

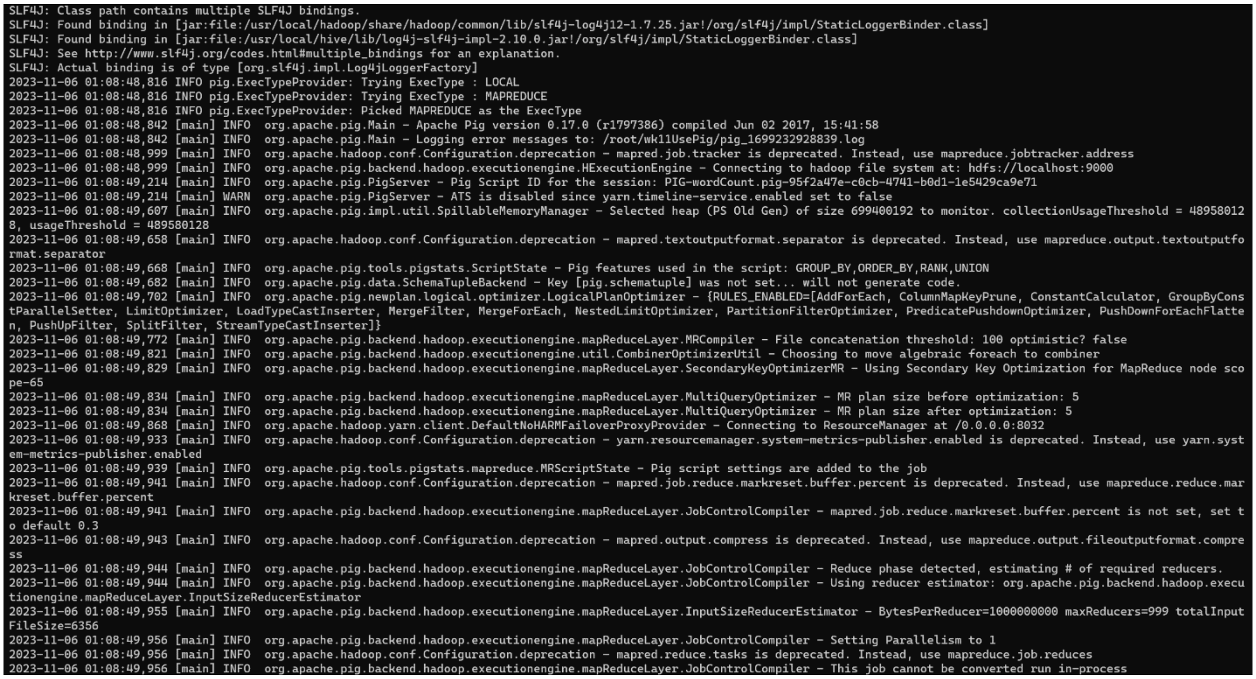
Output:



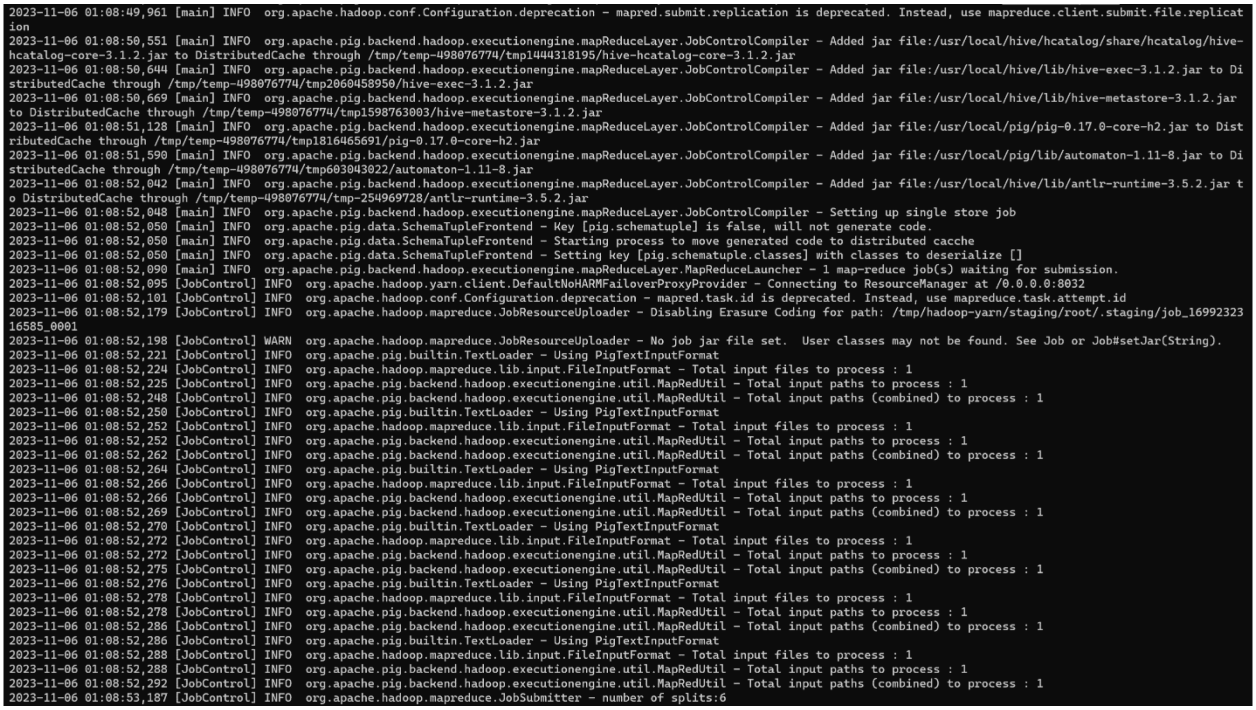
```
PS C:\Users\esbro> docker ps -a
CONTAINER ID   IMAGE                      COMMAND                  CREATED         STATUS                         PORTS     NAMES
2b831f15fa94   psu-hadoop                 "/entry.sh"              13 days ago     Exited (137) 5 hours ago                 psu-hadoop-container
babcd05d053e   psu-postgresql             "/startpostgres.sh"      3 weeks ago     Exited (137) 13 days ago                 zen_clarke
5cd6c0d935df   psu-ubuntu-java            "/bin/bash"              3 weeks ago     Exited (0) 3 weeks ago                   gallant_diffie
24b9617a206c   psugvdaan/psu-oracle11     "/bin/sh -c /start.sh"   12 months ago   Exited (137) 11 months ago               musing_mendel
PS C:\Users\esbro> docker start 2b831f15fa94
2b831f15fa94
PS C:\Users\esbro> docker exec -it 2b831f15fa94 bash
root@2b831f15fa94:/# ~
bash: /root: Is a directory
root@2b831f15fa94:/# ls
bin   derby.log  entry.sh  home  lib32  libx32  metastore_db  opt   psuprojects  run   srv  tmp  var
boot  dev        etc       lib   lib64  media   mnt           proc  root         sbin  sys  usr
root@2b831f15fa94:/# cd ~
root@2b831f15fa94:~# ls
week10java  wk10clExample  wk10clExample2  wk11UsePig
root@2b831f15fa94:~# cd wk11UsePig
root@2b831f15fa94:~/wk11UsePig# ls
wordlength.pig
root@2b831f15fa94:~/wk11UsePig# hadoop fs -ls
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
drwxr-xr-x   - root supergroup          0 2023-10-28 21:48 poems
drwxr-xr-x   - root supergroup          0 2023-11-05 19:36 wk11UsePig
root@2b831f15fa94:~/wk11UsePig# hadoop fs -ls wk11UsePig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
drwxr-xr-x   - root supergroup          0 2023-10-31 06:49 wk11UsePig/input
drwxr-xr-x   - root supergroup          0 2023-11-05 19:36 wk11UsePig/output
root@2b831f15fa94:~/wk11UsePig# hadoop fs -rm -r wk11UsePig/output/*
```



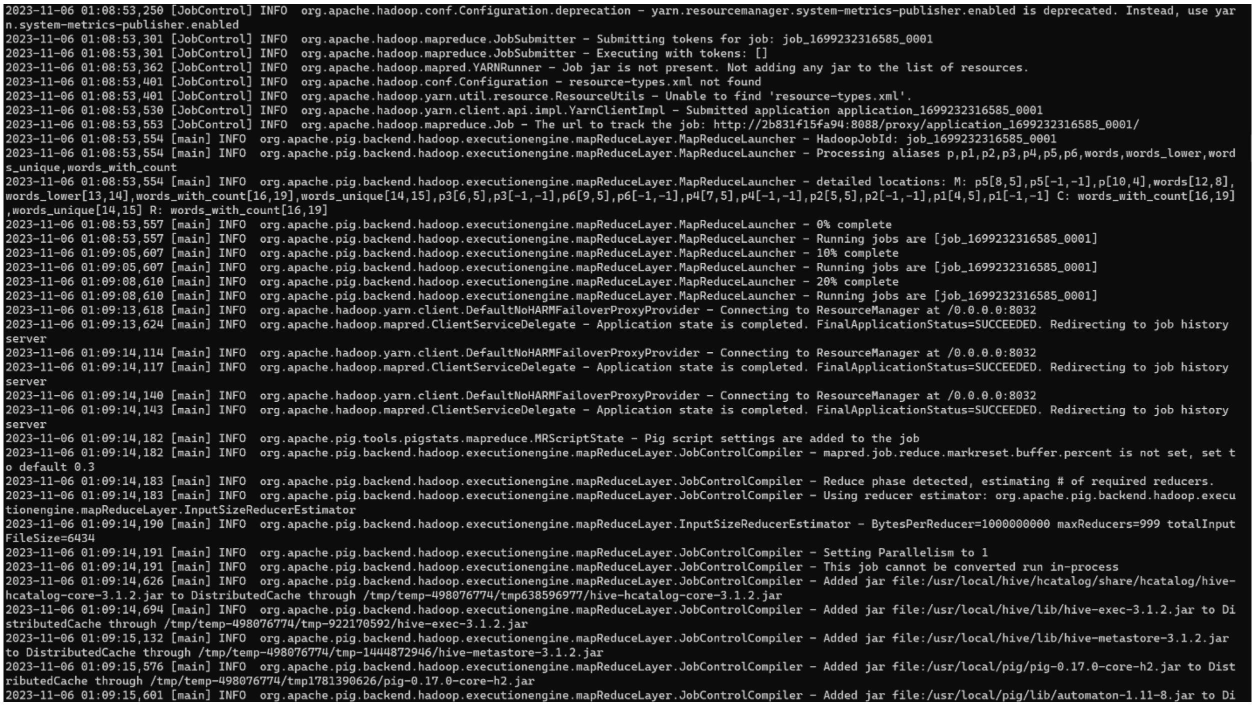
```
root@2b831f15fa94:~/wk11UsePig# hadoop fs -rm -r wk11UsePig/output/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Deleted wk11UsePig/output/wordlen
root@2b831f15fa94:~/wk11UsePig# hadoop fs  -rmdir wk11UsePig/output
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
root@2b831f15fa94:~/wk11UsePig# hadoop fs -ls wk11UsePig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 1 items
drwxr-xr-x   - root supergroup          0 2023-10-31 06:49 wk11UsePig/input
root@2b831f15fa94:~/wk11UsePig# nano wordCount.pig
root@2b831f15fa94:~/wk11UsePig# pig wordCount.pig
```



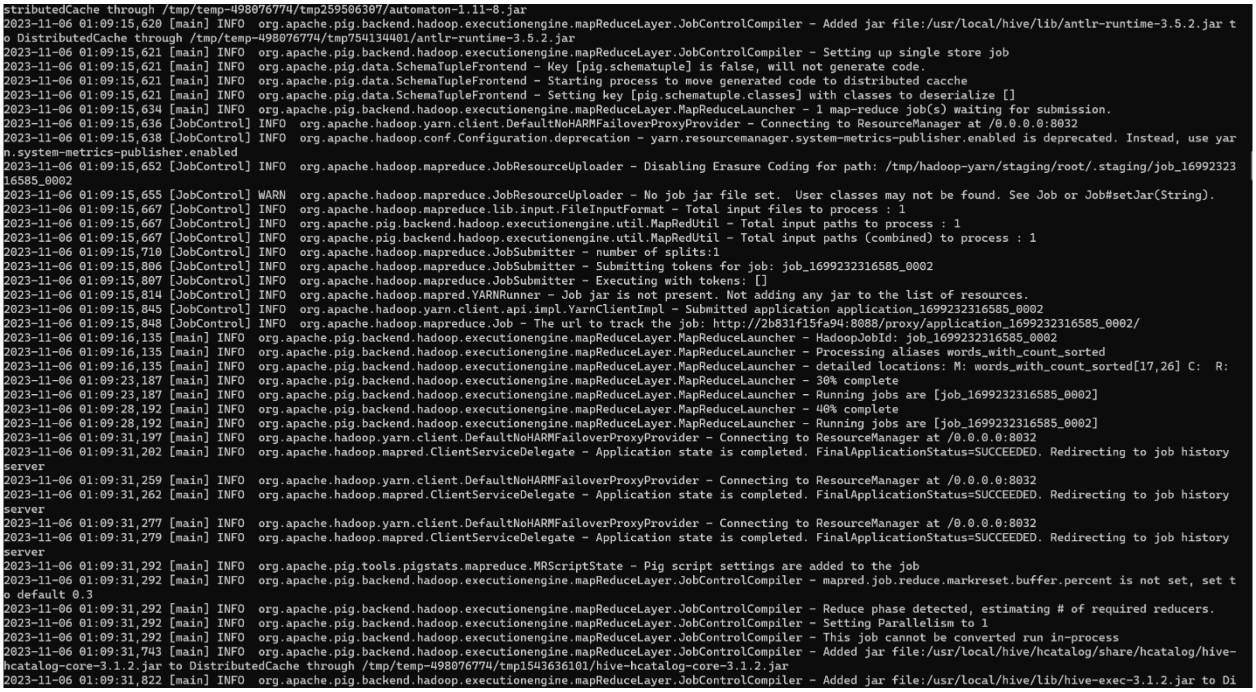
```
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2023-11-06 01:08:48,816 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2023-11-06 01:08:48,816 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2023-11-06 01:08:48,816 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-11-06 01:08:48,842 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2023-11-06 01:08:48,842 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/wk11UsePig/pig_1699232928839.log
2023-11-06 01:08:48,999 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-11-06 01:08:48,999 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2023-11-06 01:08:49,214 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-wordCount.pig-95f2a47e-c0cb-4741-b0d1-1e5429ca9e71
2023-11-06 01:08:49,214 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2023-11-06 01:08:49,607 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 48958012
8, usageThreshold = 489580128
2023-11-06 01:08:49,658 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputfo
rmat.separator
2023-11-06 01:08:49,668 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,RANK,UNION
2023-11-06 01:08:49,682 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2023-11-06 01:08:49,702 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByCons
tParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatte
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2023-11-06 01:08:49,772 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2023-11-06 01:08:49,821 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2023-11-06 01:08:49,829 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node sco
pe-65
2023-11-06 01:08:49,834 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 5
2023-11-06 01:08:49,834 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 5
2023-11-06 01:08:49,868 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:08:49,933 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.syst
em-metrics-publisher.enabled
2023-11-06 01:08:49,939 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2023-11-06 01:08:49,941 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.mar
kreset.buffer.percent
2023-11-06 01:08:49,941 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set t
o default 0.3
2023-11-06 01:08:49,943 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compre
ss
2023-11-06 01:08:49,944 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2023-11-06 01:08:49,944 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.execu
tionengine.mapReduceLayer.InputSizeReducerEstimator
2023-11-06 01:08:49,955 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInput
FileSize=6356
2023-11-06 01:08:49,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2023-11-06 01:08:49,956 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2023-11-06 01:08:49,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
```



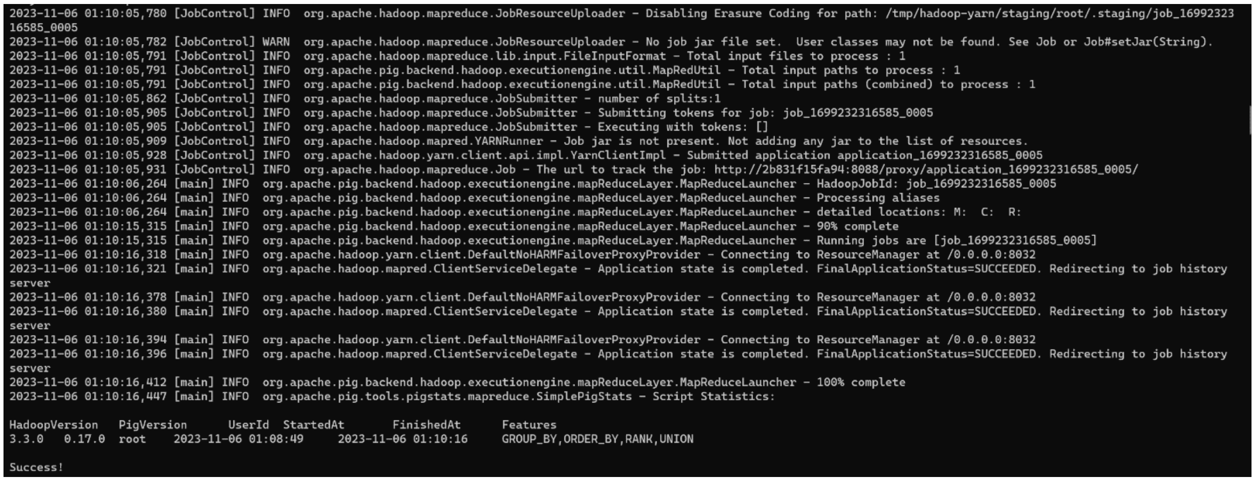
```
2023-11-06 01:08:49,961 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replicat
ion
2023-11-06 01:08:50,551 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/hcatalog/share/hcatalog/hive-
hcatalog-core-3.1.2.jar to DistributedCache through /tmp/temp-498076774/tmp1444318195/hive-hcatalog-core-3.1.2.jar
2023-11-06 01:08:50,644 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/lib/hive-exec-3.1.2.jar to Di
stributedCache through /tmp/temp-498076774/tmp2060458950/hive-exec-3.1.2.jar
2023-11-06 01:08:50,669 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/lib/hive-metastore-3.1.2.jar
to DistributedCache through /tmp/temp-498076774/tmp1598763003/hive-metastore-3.1.2.jar
2023-11-06 01:08:51,128 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/pig/pig-0.17.0-core-h2.jar to Dist
ributedCache through /tmp/temp-498076774/tmp1816465691/pig-0.17.0-core-h2.jar
2023-11-06 01:08:51,590 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/pig/lib/automaton-1.11-8.jar to Di
stributedCache through /tmp/temp-498076774/tmp603043022/automaton-1.11-8.jar
2023-11-06 01:08:52,042 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/lib/antlr-runtime-3.5.2.jar t
o DistributedCache through /tmp/temp-498076774/tmp-254969728/antlr-runtime-3.5.2.jar
2023-11-06 01:08:52,048 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2023-11-06 01:08:52,050 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2023-11-06 01:08:52,050 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2023-11-06 01:08:52,050 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2023-11-06 01:08:52,090 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2023-11-06 01:08:52,095 [JobControl] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:08:52,101 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2023-11-06 01:08:52,179 [JobControl] INFO  org.apache.hadoop.mapreduce.JobResourceUploader - Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_16992323
16585_0001
2023-11-06 01:08:52,198 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2023-11-06 01:08:52,221 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat
2023-11-06 01:08:52,224 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:08:52,225 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:08:52,248 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:08:52,250 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat
2023-11-06 01:08:52,252 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:08:52,252 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:08:52,262 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:08:52,264 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat
2023-11-06 01:08:52,266 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:08:52,266 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:08:52,269 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:08:52,270 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat
2023-11-06 01:08:52,272 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:08:52,272 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:08:52,275 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:08:52,276 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat
2023-11-06 01:08:52,278 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:08:52,278 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:08:52,286 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:08:52,286 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat
2023-11-06 01:08:52,288 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:08:52,288 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:08:52,292 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:08:53,187 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:6
```



```
2023-11-06 01:08:53,250 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yar
n.system-metrics-publisher.enabled
2023-11-06 01:08:53,301 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1699232316585_0001
2023-11-06 01:08:53,301 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2023-11-06 01:08:53,362 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2023-11-06 01:08:53,401 [JobControl] INFO  org.apache.hadoop.conf.Configuration - resource-types.xml not found
2023-11-06 01:08:53,401 [JobControl] INFO  org.apache.hadoop.yarn.util.resource.ResourceUtils - Unable to find 'resource-types.xml'.
2023-11-06 01:08:53,530 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1699232316585_0001
2023-11-06 01:08:53,553 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://2b831f15fa94:8088/proxy/application_1699232316585_0001/
2023-11-06 01:08:53,554 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1699232316585_0001
2023-11-06 01:08:53,554 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases p,p1,p2,p3,p4,p5,p6,words,words_lower,word
s_unique,words_with_count
2023-11-06 01:08:53,554 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: p5[8,5],p5[-1,-1],p[10,4],words[12,8],
words_lower[13,14],words_with_count[16,19],words_unique[14,15],p3[6,5],p3[-1,-1],p6[9,5],p6[-1,-1],p4[7,5],p4[-1,-1],p2[5,5],p2[-1,-1],p1[4,5],p1[-1,-1] C: words_with_count[16,19]
,words_unique[14,15] R: words_with_count[16,19]
2023-11-06 01:08:53,557 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2023-11-06 01:08:53,557 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1699232316585_0001]
2023-11-06 01:09:05,607 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 10% complete
2023-11-06 01:09:05,607 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1699232316585_0001]
2023-11-06 01:09:08,610 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 20% complete
2023-11-06 01:09:08,610 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1699232316585_0001]
2023-11-06 01:09:13,618 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:09:13,624 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:09:14,114 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:09:14,117 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:09:14,140 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:09:14,143 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:09:14,182 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2023-11-06 01:09:14,182 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set t
o default 0.3
2023-11-06 01:09:14,183 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2023-11-06 01:09:14,183 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.execu
tionengine.mapReduceLayer.InputSizeReducerEstimator
2023-11-06 01:09:14,190 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInput
FileSize=6434
2023-11-06 01:09:14,191 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2023-11-06 01:09:14,191 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2023-11-06 01:09:14,626 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/hcatalog/share/hcatalog/hive-
hcatalog-core-3.1.2.jar to DistributedCache through /tmp/temp-498076774/tmp638596977/hive-hcatalog-core-3.1.2.jar
2023-11-06 01:09:14,694 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/lib/hive-exec-3.1.2.jar to Di
stributedCache through /tmp/temp-498076774/tmp-922170592/hive-exec-3.1.2.jar
2023-11-06 01:09:15,132 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/lib/hive-metastore-3.1.2.jar
to DistributedCache through /tmp/temp-498076774/tmp-1444872946/hive-metastore-3.1.2.jar
2023-11-06 01:09:15,576 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/pig/pig-0.17.0-core-h2.jar to Dist
ributedCache through /tmp/temp-498076774/tmp1781390626/pig-0.17.0-core-h2.jar
2023-11-06 01:09:15,601 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/pig/lib/automaton-1.11-8.jar to Di
```



```
stributedCache through /tmp/temp-498076774/tmp259506307/automaton-1.11-8.jar
2023-11-06 01:09:15,620 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/lib/antlr-runtime-3.5.2.jar t
o DistributedCache through /tmp/temp-498076774/tmp754134401/antlr-runtime-3.5.2.jar
2023-11-06 01:09:15,621 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2023-11-06 01:09:15,621 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2023-11-06 01:09:15,621 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2023-11-06 01:09:15,621 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2023-11-06 01:09:15,634 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2023-11-06 01:09:15,636 [JobControl] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:09:15,638 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yar
n.system-metrics-publisher.enabled
2023-11-06 01:09:15,652 [JobControl] INFO  org.apache.hadoop.mapreduce.JobResourceUploader - Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_16992323
16585_0002
2023-11-06 01:09:15,655 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2023-11-06 01:09:15,667 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:09:15,667 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:09:15,667 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:09:15,710 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2023-11-06 01:09:15,806 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1699232316585_0002
2023-11-06 01:09:15,807 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2023-11-06 01:09:15,814 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2023-11-06 01:09:15,845 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1699232316585_0002
2023-11-06 01:09:15,848 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://2b831f15fa94:8088/proxy/application_1699232316585_0002/
2023-11-06 01:09:16,135 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1699232316585_0002
2023-11-06 01:09:16,135 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases words_with_count_sorted
2023-11-06 01:09:16,135 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: words_with_count_sorted[17,26] C:  R:
2023-11-06 01:09:23,187 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 30% complete
2023-11-06 01:09:23,187 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1699232316585_0002]
2023-11-06 01:09:28,192 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 40% complete
2023-11-06 01:09:28,192 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1699232316585_0002]
2023-11-06 01:09:31,197 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:09:31,202 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:09:31,259 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:09:31,262 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:09:31,277 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:09:31,279 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:09:31,292 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2023-11-06 01:09:31,292 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set t
o default 0.3
2023-11-06 01:09:31,292 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2023-11-06 01:09:31,292 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2023-11-06 01:09:31,292 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2023-11-06 01:09:31,743 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/hcatalog/share/hcatalog/hive-
hcatalog-core-3.1.2.jar to DistributedCache through /tmp/temp-498076774/tmp1543636101/hive-hcatalog-core-3.1.2.jar
2023-11-06 01:09:31,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/lib/hive-exec-3.1.2.jar to Di
```

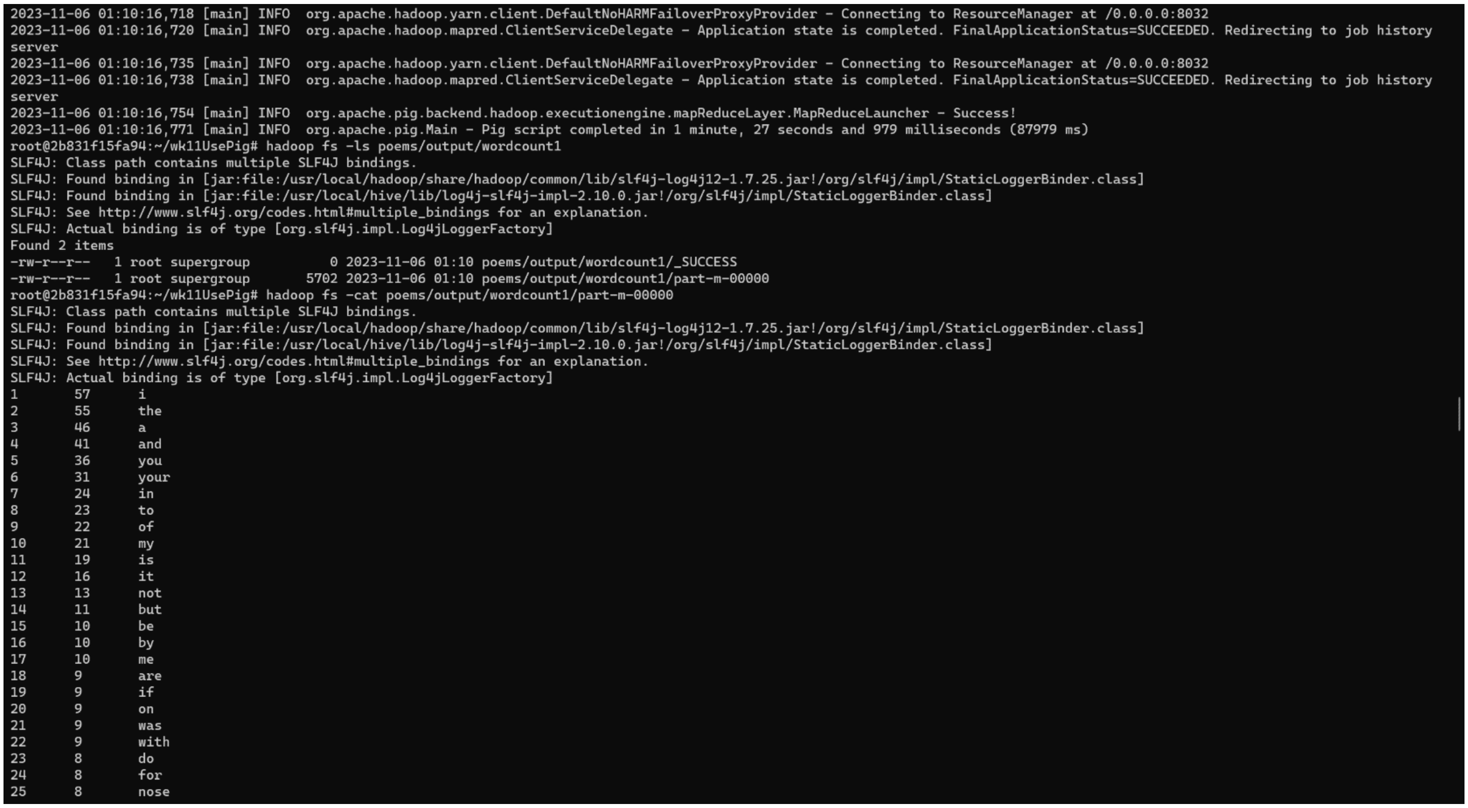


```
2023-11-06 01:10:05,780 [JobControl] INFO  org.apache.hadoop.mapreduce.JobResourceUploader - Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_16992323
16585_0005
2023-11-06 01:10:05,782 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2023-11-06 01:10:05,791 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-11-06 01:10:05,791 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-11-06 01:10:05,791 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-11-06 01:10:05,862 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2023-11-06 01:10:05,905 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1699232316585_0005
2023-11-06 01:10:05,905 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2023-11-06 01:10:05,909 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to the list of resources.
2023-11-06 01:10:05,928 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1699232316585_0005
2023-11-06 01:10:05,931 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://2b831f15fa94:8088/proxy/application_1699232316585_0005/
2023-11-06 01:10:06,264 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1699232316585_0005
2023-11-06 01:10:06,264 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases
2023-11-06 01:10:06,264 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M:  C:  R:
2023-11-06 01:10:15,315 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 90% complete
2023-11-06 01:10:15,315 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1699232316585_0005]
2023-11-06 01:10:16,318 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:10:16,321 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:10:16,378 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:10:16,380 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:10:16,394 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:10:16,396 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:10:16,412 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2023-11-06 01:10:16,447 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.0   0.17.0  root    2023-11-06 01:08:49     2023-11-06 01:10:16     GROUP_BY,ORDER_BY,RANK,UNION

Success!
```

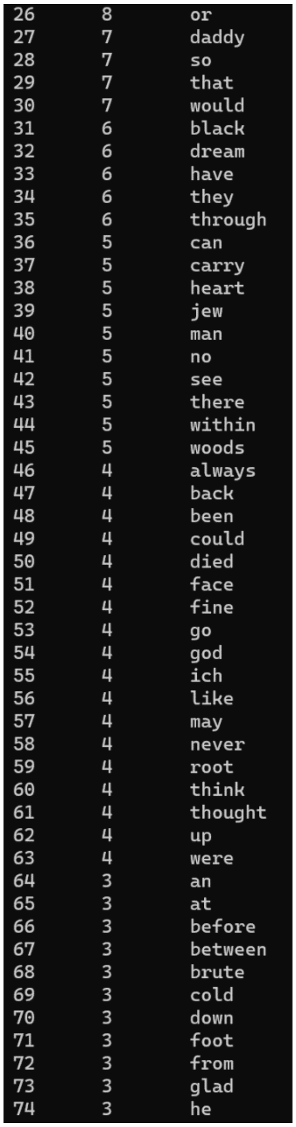
Displayed Below are Additional Output from Pig Latin script transformed into a set of MapReduce jobs

↓

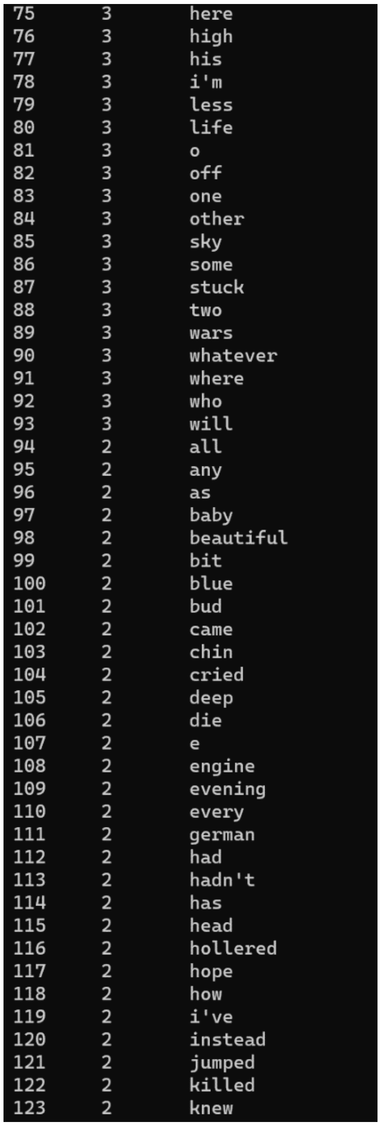
```
Job Stats (time in seconds):

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

job_1699232316585_0001  6       1       5       4       4       4       1       1       1       1
p,p1,p2,p3,p4,p5,p6,words,words_lower,words_unique,words_with_count     GROUP_BY,COMBINER

job_1699232316585_0002  1       1       1       1       1       1       1       1       1       1       words_with_count_sorted SAMPLER

job_1699232316585_0003  1       1       2       2       2       2       2       2       2       2       words_with_count_sorted ORDER_BY

job_1699232316585_0004  1       0       1       1       1       1       0       0       0       0               MAP_ONLYjob_1699232316585_0005  1       0
2       2       2       2       0       0       0       0               MAP_ONLYhdfs://localhost:9000/user/root/poems/output/wordcount1,


Input(s):

Successfully read 20 records from: "hdfs://localhost:9000/user/root/poems/input/Poem6.txt"

Successfully read 98 records from: "hdfs://localhost:9000/user/root/poems/input/Poem4.txt"

Successfully read 28 records from: "hdfs://localhost:9000/user/root/poems/input/Poem1.txt"

Successfully read 27 records from: "hdfs://localhost:9000/user/root/poems/input/Poem5.txt"

Successfully read 20 records from: "hdfs://localhost:9000/user/root/poems/input/Poem2.txt"

Successfully read 38 records from: "hdfs://localhost:9000/user/root/poems/input/Poem3.txt"


Output(s):

Successfully stored 486 records (5702 bytes) in: "hdfs://localhost:9000/user/root/poems/output/wordcount1"


Counters:

Total records written : 486

Total bytes written : 5702

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0


Job DAG:

job_1699232316585_0001  ->      job_1699232316585_0002,

job_1699232316585_0002  ->      job_1699232316585_0003,

job_1699232316585_0003  ->      job_1699232316585_0004,

job_1699232316585_0004  ->      job_1699232316585_0005,

job_1699232316585_0005
```



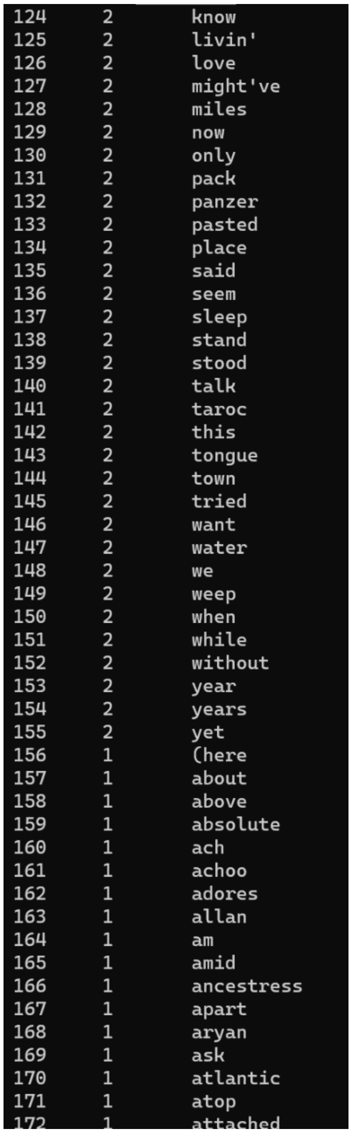
```
2023-11-06 01:10:16,718 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:10:16,720 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:10:16,735 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-06 01:10:16,738 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history
server
2023-11-06 01:10:16,754 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-11-06 01:10:16,771 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 27 seconds and 979 milliseconds (87979 ms)
root@2b831f15fa94:~/wk11UsePig# hadoop fs -ls poems/output/wordcount1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 root supergroup          0 2023-11-06 01:10 poems/output/wordcount1/_SUCCESS
-rw-r--r--   1 root supergroup       5702 2023-11-06 01:10 poems/output/wordcount1/part-m-00000
root@2b831f15fa94:~/wk11UsePig# hadoop fs -cat poems/output/wordcount1/part-m-00000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
1       57      i
2       55      the
3       46      a
4       41      and
5       36      you
6       31      your
7       24      in
8       23      to
9       22      of
10      21      my
11      19      is
12      16      it
13      13      not
14      11      but
15      10      be
16      10      by
17      10      me
18      9       are
19      9       if
20      9       on
21      9       was
22      9       with
23      8       do
24      8       for
25      8       nose
```



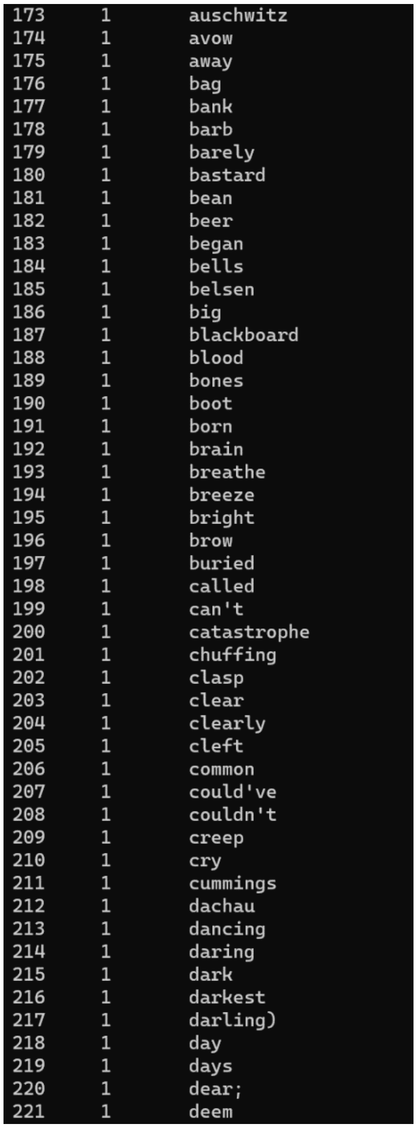
```
26      8       or
27      7       daddy
28      7       so
29      7       that
30      7       would
31      6       black
32      6       dream
33      6       have
34      6       they
35      6       through
36      5       can
37      5       carry
38      5       heart
39      5       jew
40      5       man
41      5       no
42      5       see
43      5       there
44      5       within
45      5       woods
46      4       always
47      4       back
48      4       been
49      4       could
50      4       died
51      4       face
52      4       fine
53      4       go
54      4       god
55      4       ich
56      4       like
57      4       may
58      4       never
59      4       root
60      4       think
61      4       thought
62      4       up
63      4       were
64      3       an
65      3       at
66      3       before
67      3       between
68      3       brute
69      3       cold
70      3       down
71      3       foot
72      3       from
73      3       glad
74      3       he
```



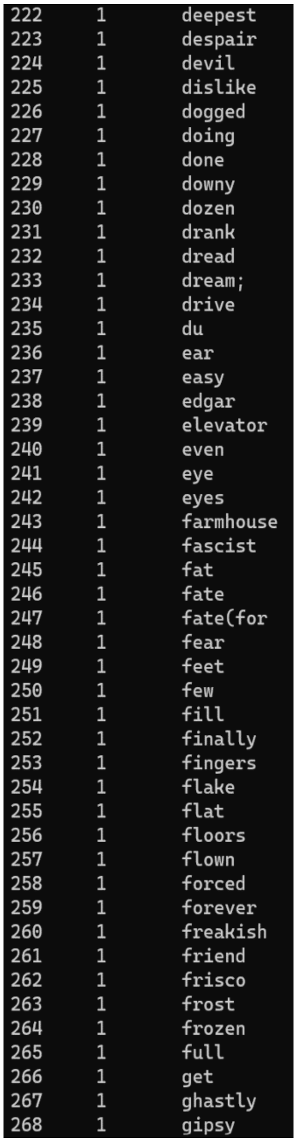
```
75      3       here
76      3       high
77      3       his
78      3       i'm
79      3       less
80      3       life
81      3       o
82      3       off
83      3       one
84      3       other
85      3       sky
86      3       some
87      3       stuck
88      3       two
89      3       wars
90      3       whatever
91      3       where
92      3       who
93      3       will
94      2       all
95      2       any
96      2       as
97      2       baby
98      2       beautiful
99      2       bit
100     2       blue
101     2       bud
102     2       came
103     2       chin
104     2       cried
105     2       deep
106     2       die
107     2       e
108     2       engine
109     2       evening
110     2       every
111     2       german
112     2       had
113     2       hadn't
114     2       has
115     2       head
116     2       hollered
117     2       hope
118     2       how
119     2       i've
120     2       instead
121     2       jumped
122     2       killed
123     2       knew
```



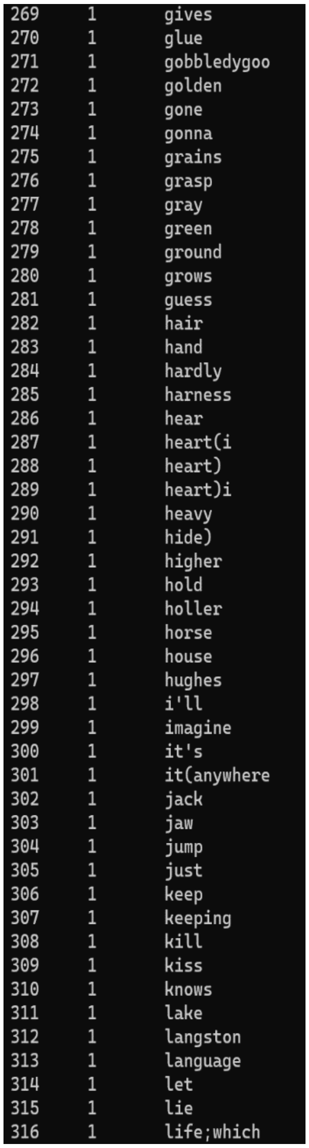
```
124     2       know
125     2       livin'
126     2       love
127     2       might've
128     2       miles
129     2       now
130     2       only
131     2       pack
132     2       panzer
133     2       pasted
134     2       place
135     2       said
136     2       seem
137     2       sleep
138     2       stand
139     2       stood
140     2       talk
141     2       taroc
142     2       this
143     2       tongue
144     2       town
145     2       tried
146     2       want
147     2       water
148     2       we
149     2       weep
150     2       when
151     2       while
152     2       without
153     2       year
154     2       years
155     2       yet
156     1       (here
157     1       about
158     1       above
159     1       absolute
160     1       ach
161     1       achoo
162     1       adores
163     1       allan
164     1       am
165     1       amid
166     1       ancestress
167     1       apart
168     1       aryan
169     1       ask
170     1       atlantic
171     1       atop
172     1       attached
```



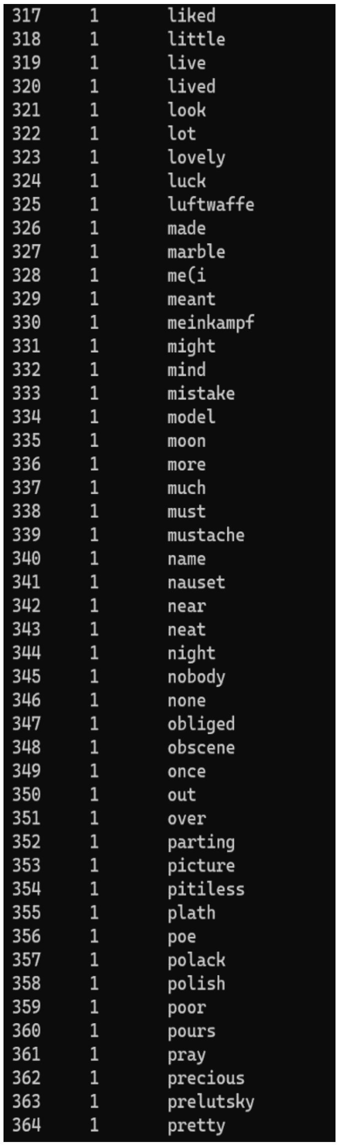
```
173     1       auschwitz
174     1       avow
175     1       away
176     1       bag
177     1       bank
178     1       barb
179     1       barely
180     1       bastard
181     1       bean
182     1       beer
183     1       began
184     1       bells
185     1       belsen
186     1       big
187     1       blackboard
188     1       blood
189     1       bones
190     1       boot
191     1       born
192     1       brain
193     1       breathe
194     1       breeze
195     1       bright
196     1       brow
197     1       buried
198     1       called
199     1       can't
200     1       catastrophe
201     1       chuffing
202     1       clasp
203     1       clear
204     1       clearly
205     1       cleft
206     1       common
207     1       could've
208     1       couldn't
209     1       creep
210     1       cry
211     1       cummings
212     1       dachau
213     1       dancing
214     1       daring
215     1       dark
216     1       darkest
217     1       darling)
218     1       day
219     1       days
220     1       dear;
221     1       deem
```



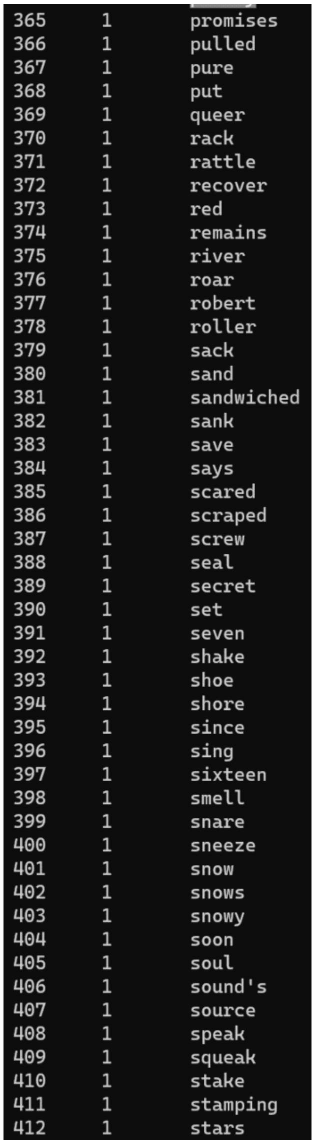
```
222     1       deepest
223     1       despair
224     1       devil
225     1       dislike
226     1       dogged
227     1       doing
228     1       done
229     1       downy
230     1       dozen
231     1       drank
232     1       dread
233     1       dream;
234     1       drive
235     1       du
236     1       ear
237     1       easy
238     1       edgar
239     1       elevator
240     1       even
241     1       eye
242     1       eyes
243     1       farmhouse
244     1       fascist
245     1       fat
246     1       fate
247     1       fate(for
248     1       fear
249     1       feet
250     1       few
251     1       fill
252     1       finally
253     1       fingers
254     1       flake
255     1       flat
256     1       floors
257     1       flown
258     1       forced
259     1       forever
260     1       freakish
261     1       friend
262     1       frisco
263     1       frost
264     1       frozen
265     1       full
266     1       get
267     1       ghastly
268     1       gipsy
```



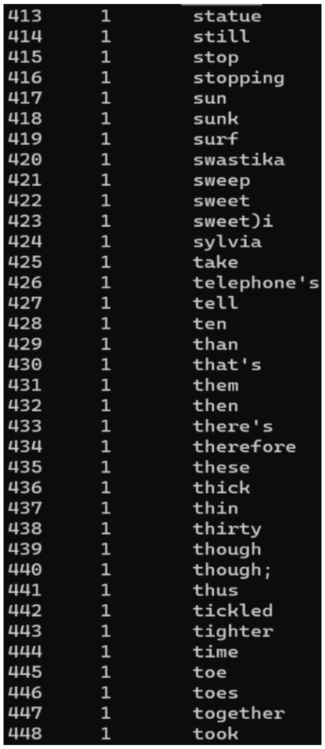
```
269     1       gives
270     1       glue
271     1       gobbledygoo
272     1       golden
273     1       gone
274     1       gonna
275     1       grains
276     1       grasp
277     1       gray
278     1       green
279     1       ground
280     1       grows
281     1       guess
282     1       hair
283     1       hand
284     1       hardly
285     1       harness
286     1       hear
287     1       heart(i
288     1       heart)
289     1       heart)i
290     1       heavy
291     1       hide)
292     1       higher
293     1       hold
294     1       holler
295     1       horse
296     1       house
297     1       hughes
298     1       i'll
299     1       imagine
300     1       it's
301     1       it(anywhere
302     1       jack
303     1       jaw
304     1       jump
305     1       just
306     1       keep
307     1       keeping
308     1       kill
309     1       kiss
310     1       knows
311     1       lake
312     1       langston
313     1       language
314     1       let
315     1       lie
316     1       life;which
```



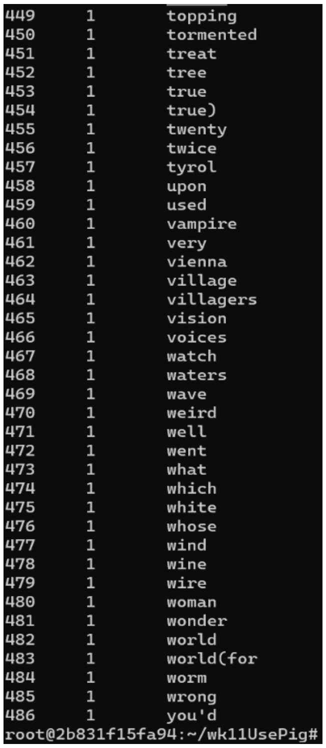
```
317     1       liked
318     1       little
319     1       live
320     1       lived
321     1       look
322     1       lot
323     1       lovely
324     1       luck
325     1       luftwaffe
326     1       made
327     1       marble
328     1       me(i
329     1       meant
330     1       meinkampf
331     1       might
332     1       mind
333     1       mistake
334     1       model
335     1       moon
336     1       more
337     1       much
338     1       must
339     1       mustache
340     1       name
341     1       nauset
342     1       near
343     1       neat
344     1       night
345     1       nobody
346     1       none
347     1       obliged
348     1       obscene
349     1       once
350     1       out
351     1       over
352     1       parting
353     1       picture
354     1       pitiless
355     1       plath
356     1       poe
357     1       polack
358     1       polish
359     1       poor
360     1       pours
361     1       pray
362     1       precious
363     1       prelutsky
364     1       pretty
```



```
365     1       promises
366     1       pulled
367     1       pure
368     1       put
369     1       queer
370     1       rack
371     1       rattle
372     1       recover
373     1       red
374     1       remains
375     1       river
376     1       roar
377     1       robert
378     1       roller
379     1       sack
380     1       sand
381     1       sandwiched
382     1       sank
383     1       save
384     1       says
385     1       scared
386     1       scraped
387     1       screw
388     1       seal
389     1       secret
390     1       set
391     1       seven
392     1       shake
393     1       shoe
394     1       shore
395     1       since
396     1       sing
397     1       sixteen
398     1       smell
399     1       snare
400     1       sneeze
401     1       snow
402     1       snows
403     1       snowy
404     1       soon
405     1       soul
406     1       sound's
407     1       source
408     1       speak
409     1       squeak
410     1       stake
411     1       stamping
412     1       stars
```



```
413     1       statue
414     1       still
415     1       stop
416     1       stopping
417     1       sun
418     1       sunk
419     1       surf
420     1       swastika
421     1       sweep
422     1       sweet
423     1       sweet)i
424     1       sylvia
425     1       take
426     1       telephone's
427     1       tell
428     1       ten
429     1       than
430     1       that's
431     1       them
432     1       then
433     1       there's
434     1       therefore
435     1       these
436     1       thick
437     1       thin
438     1       thirty
439     1       though
440     1       though;
441     1       thus
442     1       tickled
443     1       tighter
444     1       time
445     1       toe
446     1       toes
447     1       together
448     1       took
```



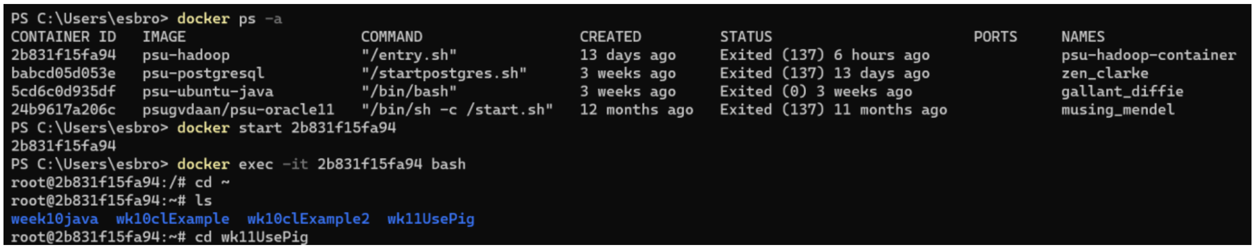
```
449     1       topping
450     1       tormented
451     1       treat
452     1       tree
453     1       true
454     1       true)
455     1       twenty
456     1       twice
457     1       tyrol
458     1       upon
459     1       used
460     1       vampire
461     1       very
462     1       vienna
463     1       village
464     1       villagers
465     1       vision
466     1       voices
467     1       watch
468     1       waters
469     1       wave
470     1       weird
471     1       well
472     1       went
473     1       what
474     1       which
475     1       white
476     1       whose
477     1       wind
478     1       wine
479     1       wire
480     1       woman
481     1       wonder
482     1       world
483     1       world(for
484     1       worm
485     1       wrong
486     1       you'd
root@2b831f15fa94:~/wk11UsePig#
```

2.) Change the code to count how many words of different lengths are in the poems data.

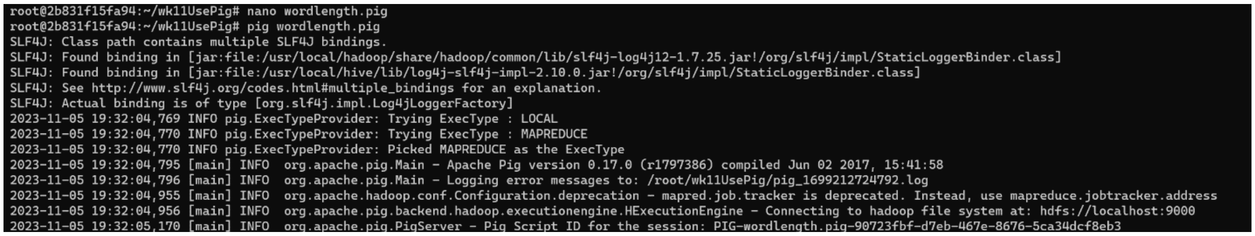
Program Code:

```
p1 = LOAD '/user/root/poems/input/Poem1.txt' USING TextLoader AS (line: chararray);
p2 = LOAD '/user/root/poems/input/Poem2.txt' USING TextLoader AS (line: chararray);
p3 = LOAD '/user/root/poems/input/Poem3.txt' USING TextLoader AS (line: chararray);
p4 = LOAD '/user/root/poems/input/Poem4.txt' USING TextLoader AS (line: chararray);
p5 = LOAD '/user/root/poems/input/Poem5.txt' USING TextLoader AS (line: chararray);
p6 = LOAD '/user/root/poems/input/Poem6.txt' USING TextLoader AS (line: chararray);
p = UNION p1, p2, p3, p4, p5, p6;

words = FOREACH p GENERATE FLATTEN(TOKENIZE(line, ' ,!?\t\n\r\f\\.\\-')) as word;

words_lower = FOREACH words GENERATE LOWER(word) as word_lower;
words_unique = GROUP words_lower BY word_lower;
words_len = FOREACH words_unique GENERATE SIZE(words_lower) as word_length;

groupWords = GROUP words_len BY word_length;
wordCount = FOREACH groupWords GENERATE CONCAT('Length ', (chararray) group) as word_length, COUNT(words_len) as word_count;

words_lenSorted = ORDER wordCount BY word_count DESC, word_length;

STORE words_lenSorted INTO 'wk11UsePig/output/wordlen';
```

Output:



```
PS C:\Users\esbro> docker ps -a
CONTAINER ID   IMAGE                   COMMAND                CREATED         STATUS                      PORTS     NAMES
2b831f15fa94   psu-hadoop              "/entry.sh"            13 days ago     Exited (137) 6 hours ago              psu-hadoop-container
babcd05d053e   psu-postgresql          "/startpostgres.sh"    3 weeks ago     Exited (137) 13 days ago              zen_clarke
5cd6c0d935df   psu-ubuntu-java         "/bin/bash"            3 weeks ago     Exited (0) 3 weeks ago                gallant_diffie
24b9617a206c   psugvdaan/psu-oracle11  "/bin/sh -c /start.sh" 12 months ago   Exited (137) 11 months ago            musing_mendel
PS C:\Users\esbro> docker start 2b831f15fa94
2b831f15fa94
PS C:\Users\esbro> docker exec -it 2b831f15fa94 bash
root@2b831f15fa94:/# cd ~
root@2b831f15fa94:~# ls
week10java  wk10clExample  wk10clExample2  wk11UsePig
root@2b831f15fa94:~# cd wk11UsePig
```



```
root@2b831f15fa94:~/wk11UsePig# nano wordlength.pig
root@2b831f15fa94:~/wk11UsePig# pig wordlength.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2023-11-05 19:32:04,769 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2023-11-05 19:32:04,770 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2023-11-05 19:32:04,770 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-11-05 19:32:04,795 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2023-11-05 19:32:04,796 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/wk11UsePig/pig_1699212724792.log
2023-11-05 19:32:04,955 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-11-05 19:32:04,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2023-11-05 19:32:05,170 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-wordlength.pig-90723fbf-d7eb-467e-8676-5ca34dcf8eb3
```

Displayed Below are Additional Output from Pig Latin script transformed into a set of MapReduce jobs

```
2023-11-05 19:32:05,170 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2023-11-05 19:32:05,566 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size
699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2023-11-05 19:32:05,613 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1045: <file wordlength.pig, line 19, column 40>
Could not infer the matching function for org.apache.pig.builtin.CONCAT as multiple or none of them fit. Please use an explicit cast.
Details at logfile: /root/wk11UsePig/pig_1699212724792.log
2023-11-05 19:32:05,624 [main] INFO  org.apache.pig.Main - Pig script completed in 883 milliseconds (883 ms)
root@2b831f15fa94:~/wk11UsePig# nano wordlength.pig
root@2b831f15fa94:~/wk11UsePig# pig wordlength.pig
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-
1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2023-11-05 19:35:04,099 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2023-11-05 19:35:04,100 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2023-11-05 19:35:04,100 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-11-05 19:35:04,123 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2023-11-05 19:35:04,123 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/wk11UsePig/pig_1699212904120.log
2023-11-05 19:35:04,280 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2023-11-05 19:35:04,280 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop
file system at: hdfs://localhost:9000
2023-11-05 19:35:04,494 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-wordlength.pig-61af50fb-9468-
43c7-9c87-d97d3af2307e
2023-11-05 19:35:04,494 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2023-11-05 19:35:04,885 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size
699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2023-11-05 19:35:04,939 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is
deprecated. Instead, use mapreduce.output.textoutputformat.separator
2023-11-05 19:35:04,950 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script:
GROUP_BY,ORDER_BY,UNION
2023-11-05 19:35:04,965 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not
generate code.
```

```
2023-11-05 19:35:04,984 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}

2023-11-05 19:35:05,052 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false

2023-11-05 19:35:05,105 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner

2023-11-05 19:35:05,114 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-77

2023-11-05 19:35:05,119 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4

2023-11-05 19:35:05,119 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4

2023-11-05 19:35:05,154 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:05,231 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled

2023-11-05 19:35:05,238 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job

2023-11-05 19:35:05,241 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent

2023-11-05 19:35:05,241 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3

2023-11-05 19:35:05,242 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress

2023-11-05 19:35:05,243 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.

2023-11-05 19:35:05,244 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator

2023-11-05 19:35:05,255 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=6356

2023-11-05 19:35:05,255 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1

2023-11-05 19:35:05,256 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces

2023-11-05 19:35:05,256 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process

2023-11-05 19:35:05,262 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication

2023-11-05 19:35:05,876 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/usr/local/hive/hcatalog/share/hcatalog/hive-hcatalog-core-3.1.2.jar to DistributedCache through /tmp/temp-1663939476/tmp1561475714/hive-hcatalog-core-3.1.2.jar
```

```
2023-11-05 19:35:06,978 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job

2023-11-05 19:35:06,981 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.

2023-11-05 19:35:06,981 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche

2023-11-05 19:35:06,981 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []

2023-11-05 19:35:07,019 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.

2023-11-05 19:35:07,023 [JobControl] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:07,029 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id

2023-11-05 19:35:07,108 [JobControl] INFO  org.apache.hadoop.mapreduce.JobResourceUploader - Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1699210896959_0001

2023-11-05 19:35:07,125 [JobControl] WARN  org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).

2023-11-05 19:35:07,143 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat

2023-11-05 19:35:07,149 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1

2023-11-05 19:35:07,149 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

2023-11-05 19:35:07,176 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1

2023-11-05 19:35:07,178 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat

2023-11-05 19:35:07,180 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1

2023-11-05 19:35:07,180 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

2023-11-05 19:35:07,190 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1

2023-11-05 19:35:07,191 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat

2023-11-05 19:35:07,192 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1

2023-11-05 19:35:07,193 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

2023-11-05 19:35:07,198 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1

2023-11-05 19:35:07,199 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat

2023-11-05 19:35:07,200 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1

2023-11-05 19:35:07,200 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
```

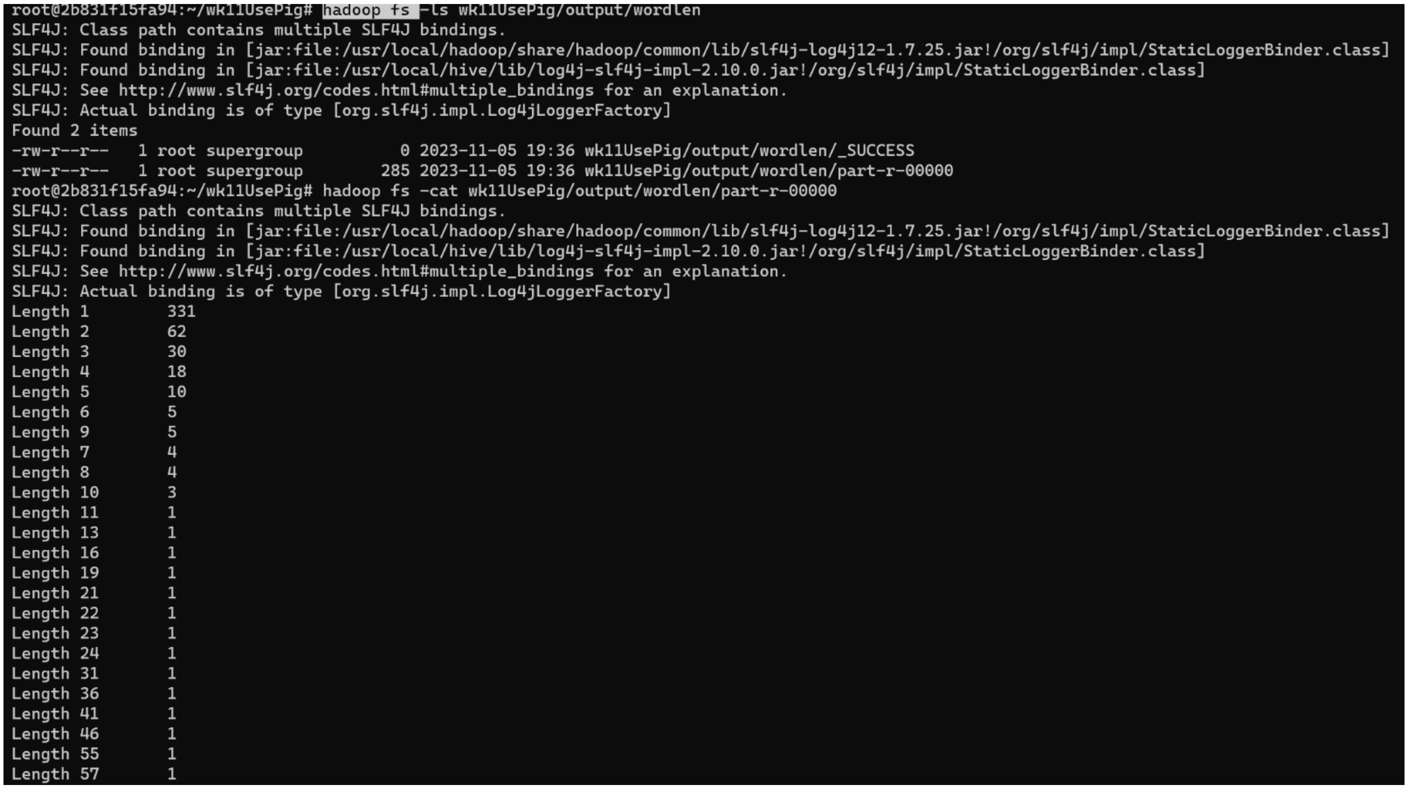
```
2023-11-05 19:35:07,207 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process :
1

2023-11-05 19:35:07,207 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1

2023-11-05 19:35:07,216 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths
(combined) to process : 1

2023-11-05 19:35:07,217 [JobControl] INFO  org.apache.pig.builtin.TextLoader - Using PigTextInputFormat

2023-11-05 19:35:07,218 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process :
1

2023-11-05 19:35:07,218 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 1

2023-11-05 19:35:07,222 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths
(combined) to process : 1

2023-11-05 19:35:07,734 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:6

2023-11-05 19:35:07,797 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-
metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled

2023-11-05 19:35:07,865 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job:
job_1699210896959_0001

2023-11-05 19:35:07,865 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []

2023-11-05 19:35:07,928 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar to
the list of resources.

2023-11-05 19:35:07,968 [JobControl] INFO  org.apache.hadoop.conf.Configuration - resource-types.xml not found

2023-11-05 19:35:07,969 [JobControl] INFO  org.apache.hadoop.yarn.util.resource.ResourceUtils - Unable to find 'resource-
types.xml'.

2023-11-05 19:35:08,301 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application
application_1699210896959_0001

2023-11-05 19:35:08,325 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job:
http://2b831f15fa94:8088/proxy/application_1699210896959_0001/

2023-11-05 19:35:08,326 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
HadoopJobId: job_1699210896959_0001

2023-11-05 19:35:08,326 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Processing aliases p,p1,p2,p3,p4,p5,p6,words,words_len,words_lower,words_unique

2023-11-05 19:35:08,326 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
detailed locations: M: p5[8,5],p5[-1,-1],p[10,4],words[12,8],words_lower[14,14],words_unique[15,15],p3[6,5],p3[-1,-1],p6[9,5],p6[-
1,-1],p4[7,5],p4[-1,-1],p2[5,5],p2[-1,-1],p1[4,5],p1[-1,-1] C:  R: words_len[16,12]

2023-11-05 19:35:08,329 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
0% complete
```

```
2023-11-05 19:35:08,329 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Running jobs are [job_1699210896959_0001]

2023-11-05 19:35:18,373 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 4%
complete

2023-11-05 19:35:18,374 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Running jobs are [job_1699210896959_0001]

2023-11-05 19:35:20,379 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
12% complete

2023-11-05 19:35:20,379 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Running jobs are [job_1699210896959_0001]

2023-11-05 19:35:22,382 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
25% complete

2023-11-05 19:35:22,383 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Running jobs are [job_1699210896959_0001]

2023-11-05 19:35:28,397 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:28,403 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:35:28,886 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:28,888 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:35:28,908 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:28,910 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:35:28,945 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the
job

2023-11-05 19:35:28,945 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3

2023-11-05 19:35:28,946 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
Reduce phase detected, estimating # of required reducers.

2023-11-05 19:35:28,946 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
```

```
2023-11-05 19:35:37,944 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 37%
complete

2023-11-05 19:35:37,944 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Runnin
jobs are [job_1699210896959_0002]

2023-11-05 19:35:42,952 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50%
complete

2023-11-05 19:35:42,952 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Runnin
jobs are [job_1699210896959_0002]

2023-11-05 19:35:45,960 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:45,964 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:35:46,034 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:46,037 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:35:46,052 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:35:46,054 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:35:46,068 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job

2023-11-05 19:35:46,068 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3

2023-11-05 19:35:46,069 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce
detected, estimating # of required reducers.

2023-11-05 19:35:46,069 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using re
estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator

023-11-05 19:36:18,181 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceMa
at /0.0.0.0:8032

2023-11-05 19:36:18,185 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:36:18,238 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:36:18,241 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:36:18,253 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to
ResourceManager at /0.0.0.0:8032

2023-11-05 19:36:18,255 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:36:18,268 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100%
complete
```

```
2023-11-05 19:36:18,294 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:


HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features

3.3.0   0.17.0  root    2023-11-05 19:35:05     2023-11-05 19:36:18     GROUP_BY,ORDER_BY,UNION


Success!


Job Stats (time in seconds):

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

job_1699210896959_0001  6       1       4       4       4       4       1       1       1       1
p,p1,p2,p3,p4,p5,p6,words,words_len,words_lower,words_unique    GROUP_BY

job_1699210896959_0002  1       1       1       1       1       1       1       1       1       1       groupWords,wordCount    GROUP_BY,COMBINER

job_1699210896959_0003  1       1       1       1       1       1       1       1       1       1       words_lenSorted SAMPLER

job_1699210896959_0004  1       1       1       1       1       1       1       1       1       1       words_lenSorted ORDER_BY
hdfs://localhost:9000/user/root/wk11UsePig/output/wordlen,


Input(s):

Successfully read 20 records from: "/user/root/poems/input/Poem6.txt"

Successfully read 98 records from: "/user/root/poems/input/Poem4.txt"

Successfully read 28 records from: "/user/root/poems/input/Poem1.txt"

Successfully read 27 records from: "/user/root/poems/input/Poem5.txt"

Successfully read 20 records from: "/user/root/poems/input/Poem2.txt"

Successfully read 38 records from: "/user/root/poems/input/Poem3.txt"


Output(s):

Successfully stored 24 records (285 bytes) in: "hdfs://localhost:9000/user/root/wk11UsePig/output/wordlen"


Counters:

Total records written : 24

Total bytes written : 285

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0
```

```
2023-11-05 19:36:18,296 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032

2023-11-05 19:36:18,304 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:36:18,322 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032

2023-11-05 19:36:18,324 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:36:18,338 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032

2023-11-05 19:36:18,340 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:36:18,358 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032

2023-11-05 19:36:18,361 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

2023-11-05 19:36:18,375 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032

2023-11-05 19:36:18,493 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2023-11-05 19:36:18,506 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 14 seconds and 433 milliseconds (74433 ms)
```



```
root@2b831f15fa94:~/wk11UsePig# hadoop fs -ls wk11UsePig/output/wordlen
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 root supergroup          0 2023-11-05 19:36 wk11UsePig/output/wordlen/_SUCCESS
-rw-r--r--   1 root supergroup        285 2023-11-05 19:36 wk11UsePig/output/wordlen/part-r-00000
root@2b831f15fa94:~/wk11UsePig# hadoop fs -cat wk11UsePig/output/wordlen/part-r-00000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Length 1        331
Length 2        62
Length 3        30
Length 4        18
Length 5        10
Length 6        5
Length 9        5
Length 7        4
Length 8        4
Length 10       3
Length 11       1
Length 13       1
Length 16       1
Length 19       1
Length 21       1
Length 22       1
Length 23       1
Length 24       1
Length 31       1
Length 36       1
Length 41       1
Length 46       1
Length 55       1
Length 57       1
```