Module 4: Assignment – 1

1. Find out the count of each word in the ‘Shakespeare.txt’ dataset using Pig UDF

2. Store the output in a specified output directory of your choice

Solution:

```
grunt> input_lines = LOAD '/home/bitnami/pigdata/Shakespeare.txt' AS (line:chararray);
2021-11-05 18:51:09,348 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
grunt> filtered_words = FILTER words BY word MATCHES '\\w+';
grunt> word_groups = GROUP filtered_words BY word;
grunt> word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS word;
grunt> ordered_word_count = ORDER word_count BY count DESC;
```

```
grunt> store ordered_word_count into '/home/bitnami/pigoutput/wordcount_output';
2021-11-05 18:53:17,028 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
2021-11-05 18:53:17,078 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutpu
tformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2021-11-05 18:53:17,098 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in th
e script: GROUP_BY,ORDER_BY,FILTER
2021-11-05 18:53:17,139 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
2021-11-05 18:53:17,205 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RU
LES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimiz
er, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, Pre
dicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2021-11-05 18:53:17,423 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRComp
iler - File concatenation threshold: 100 optimistic? false
2021-11-05 18:53:17,491 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
2021-11-05 18:53:17,509 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimize
rUtil - Choosing to move algebraic foreach to combiner
2021-11-05 18:53:17,549 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Second
aryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-26
2021-11-05 18:53:17,567 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQ
ueryOptimizer - MR plan size before optimization: 3
2021-11-05 18:53:17,567 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQ
ueryOptimizer - MR plan size after optimization: 3
2021-11-05 18:53:17,610 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.che
cksum is deprecated. Instead, use dfs.bytes-per-checksum
2021-11-05 18:53:17,890 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties f
rom hadoop-metrics2.properties
2021-11-05 18:53:18,098 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metri
c snapshot period at 10 second(s).
```

Job Details

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime  M
inReduceTime    AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1302465955_0002        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n
/a      ordered_word_count      SAMPLER
job_local404951707_0001 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     f
iltered_words,input_lines,word_count,word_groups,words  GROUP_BY,COMBINER
job_local40574420_0003  1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     o
rdered_word_count       ORDER_BY        /home/bitnami/pigoutput/wordcount_output,

Input(s):
Successfully read 135281 records from: "/home/bitnami/pigdata/Shakespeare.txt"

Output(s):
Successfully stored 32325 records in: "/home/bitnami/pigoutput/wordcount_output"

Counters:
Total records written : 32325
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local404951707_0001 ->      job_local1302465955_0002,
job_local1302465955_0002        ->      job_local40574420_0003,
job_local40574420_0003
```

Output:

```
1       fines
1       finem
1       final
1       filme
1       files
1       filch
1       feuer
1       fells
1       feith
1       feine
1       feild
1       feest
1       feend
1       feate
1       fears
1       fayth
1       fayne
1       fause
1       fanne
1       faite
1       fairy
1       fadge
1       exist
1       exalt
1       ewers
1       every
1       euade
1       estre
1       essay
1       escus
1       equal
1       envie
1       entre
1       endue
1       endes
1       emmew
1       elboe
1       eiect
1       egall
1       eeues
1       ebbes
1       eaues
1       eauen
1       earle
1       dwelt
1       dusty
1       dusky
1       dugge
1       ducat
1       dryed
1       drudg
```