YARN – MapReduce Lab 2

Link to the git repository:

https://github.com/JinJinB/Data_Engineering_Project.git

1. MapReduce JAVA

1.6.3 Run the job

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHOT-j
ar-with-dependencies.jar wordcount /user/jin-young.bae/ebook.txt /user/jin-young
.bae/wordcount
23/07/05 15:25:27 INFO impl.TimelineReaderClientImpl: Initialized TimelineReader
 URI=https://localhost:8199/ws/v2/timeline/, clusterId=yarn_cluster
23/07/05 15:25:28 INFO client.RMProxy: Connecting to ResourceManager at master01
.hadoop.efrei.clemlab.com/163.172.76.182:8050
23/07/05 15:25:28 INFO client.AHSProxy: Connecting to Application History server
 at master01.hadoop.efrei.clemlab.com/163.172.76.182:10200
23/07/05 15:25:28 INFO hdfs.DFSClient: Created token for jin-young.bae: HDFS_DEL
EGATION_TOKEN owner=jin-young.bae@HADOOP.EFREI, renewer=yarn, realUser=, issueDa
te=1688563528333, maxDate=1689168328333, sequenceNumber=459, masterKeyId=11 on 1
63.172.76.182:8020
23/07/05 15:25:28 INFO security.TokenCache: Got dt for hdfs://master01.hadoop.ef
rei.clemlab.com:8020; Kind: HDFS_DELEGATION_TOKEN, Service: 163.172.76.182:8020,
 Ident: (token for jin-young.bae: HDFS_DELEGATION_TOKEN owner=jin-young.bae@HADO
OP.EFREI, renewer=yarn, realUser=, issueDate=1688563528333, maxDate=168916832833
3, sequenceNumber=459, masterKeyId=11)
23/07/05 15:25:28 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding f
or path: /user/jin-young.bae/.staging/job_1688376829169_0217
23/07/05 15:25:29 INFO input.FileInputFormat: Total input files to process : 1
23/07/05 15:25:29 INFO mapreduce.JobSubmitter: number of splits:1
23/07/05 15:25:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
```

(…)

```
23/07/05 15:25:30 INFO mapreduce.Job: Running job: job_1688376829169_0217
23/07/05 15:25:47 INFO mapreduce.Job: Job job_1688376829169_0217 running in uber
 mode : false
23/07/05 15:25:47 INFO mapreduce.Job:  map 0% reduce 0%
23/07/05 15:25:57 INFO mapreduce.Job:  map 100% reduce 0%
23/07/05 15:26:07 INFO mapreduce.Job:  map 100% reduce 100%
23/07/05 15:26:08 INFO mapreduce.Job: Job job_1688376829169_0217 completed succe
ssfully
23/07/05 15:26:08 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=93158
                FILE: Number of bytes written=703343
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=229101
                HDFS: Number of bytes written=68003
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
```

(…)

```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=228962
        File Output Format Counters
                Bytes Written=68003
-bash-4.2$
```

Result : with $ hdfs dfs -cat wordcount/part-r-00000

(…)

```
"fits   1
"get    3
"grief,"        1
"loathsome      1
"terror,"       1
"the    2
"with   1
"__     1
"'But   1
"'Dear  1
"'Dearest!      1
"'God   1
"'HUSBAND.'     1
"'I     6
"'In    1
"'It    1
"'MY    2
"'My    3
"'Now   1
"'Talk  1
"'These 1
"'Though        1
"'True, 1
"'We    1
"'YOUR  1
"'You   4
"'_Can  1
•       4
-bash-4.2$
```

2. Remarkable trees of Paris

In this part, we will only show the end of the programs to shorten the report.

a) Districts containing trees

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar
-with-dependencies.jar district /user/jin-young.bae/trees.csv /user/jin-young.bae/
district
```

End of District

```
File Input Format Counters
        Bytes Read=16680
File Output Format Counters
        Bytes Written=44
```

Result of District

```
-bash-4.2$ hdfs dfs -cat district/part-r-00000
3
4
5
6
7
8
9
11
12
13
14
15
16
17
18
19
20
```

b) Show all existing species

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHO
T-jar-with-dependencies.jar species /user/jin-young.bae/trees.csv /user/jin-y
oung.bae/species
```

End of Species

```
File Input Format Counters
        Bytes Read=16680
File Output Format Counters
        Bytes Written=317
```

Result of Species

```
-bash-4.2$ hdfs dfs -cat species/part-r-00000
Acer
Aesculus
Ailanthus
Alnus
Araucaria
Broussonetia
Calocedrus
Catalpa
Cedrus
Celtis
Corylus
Davidia
Diospyros
Eucommia
Fagus
Fraxinus
Ginkgo
Gymnocladus
Juglans
Liriodendron
Maclura
Magnolia
Paulownia
Pinus
Platanus
Pterocarya
Quercus
Robinia
Sequoia
Sequoiadendron
Styphnolobium
Taxodium
Taxus
Tilia
Ulmus
Zelkova
```

c) Number of trees by kinds

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHO
T-jar-with-dependencies.jar speciescount /user/jin-young.bae/trees.csv /user/
jin-young.bae/speciescount
```

End of SpeciesCount

```
File Input Format Counters
        Bytes Read=16680
File Output Format Counters
        Bytes Written=390
```

Result of SpeciesCount

```
-bash-4.2$ hdfs dfs -cat speciescount/part-r-00000
Acer    3
Aesculus        3
Ailanthus       1
Alnus   1
Araucaria       1
Broussonetia    1
Calocedrus      1
Catalpa 1
Cedrus  4
Celtis  1
Corylus 3
Davidia 1
Diospyros       4
Eucommia        1
Fagus   8
Fraxinus        1
Ginkgo  5
Gymnocladus     1
Juglans 1
Liriodendron    2
Maclura 1
Magnolia        1
Paulownia       1
Pinus   5
Platanus        19
Pterocarya      3
Quercus 4
Robinia 1
Sequoia 1
Sequoiadendron  5
Styphnolobium   1
Taxodium        3
Taxus   2
Tilia   1
Ulmus   1
Zelkova 4
```

d) Maximum height per kind of tree

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHO
T-jar-with-dependencies.jar maxheight /user/jin-young.bae/trees.csv /user/jin
-young.bae/maxheight
```

End of MaxHeight

```
File Input Format Counters
        Bytes Read=16680
File Output Format Counters
        Bytes Written=496
```

Result of MaxHeight

```
-bash-4.2$ hdfs dfs -cat maxheight/part-r-00000
Acer    16.0
Aesculus        30.0
Ailanthus       35.0
Alnus   16.0
Araucaria       9.0
Broussonetia    12.0
Calocedrus      20.0
Catalpa 15.0
Cedrus  30.0
Celtis  16.0
Corylus 20.0
Davidia 12.0
Diospyros       14.0
Eucommia        12.0
Fagus   30.0
Fraxinus        30.0
Ginkgo  33.0
Gymnocladus     10.0
Juglans 28.0
Liriodendron    35.0
Maclura 13.0
Magnolia        12.0
Paulownia       20.0
Pinus   30.0
Platanus        45.0
Pterocarya      30.0
Quercus 31.0
Robinia 11.0
Sequoia 30.0
Sequoiadendron  35.0
Styphnolobium   10.0
Taxodium        35.0
Taxus   13.0
Tilia   20.0
Ulmus   15.0
Zelkova 30.0
```

e) Sort the trees height from smallest to largest

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHO
T-jar-with-dependencies.jar sortheight /user/jin-young.bae/trees.csv /user/ji
n-young.bae/sortheight
```

End of SortHeight

```
File Input Format Counters
        Bytes Read=16680
File Output Format Counters
        Bytes Written=1296
```

Result of SortHeight

```
-bash-4.2$ hdfs dfs -cat sortheight/part-r-00000
2.0     Fagus
5.0     Taxus
6.0     Cedrus
9.0     Araucaria
10.0    Styphnolobium
10.0    Quercus
10.0    Pinus
10.0    Gymnocladus
10.0    Fagus
11.0    Robinia
12.0    Diospyros
12.0    Magnolia
12.0    Zelkova
12.0    Eucommia
12.0    Acer
12.0    Diospyros
12.0    Broussonetia
12.0    Davidia
13.0    Taxus
13.0    Maclura
14.0    Diospyros
14.0    Pinus
14.0    Diospyros
15.0    Acer
15.0    Catalpa
15.0    Fagus
```

(…)

```
31.0    Platanus
32.0    Platanus
33.0    Ginkgo
34.0    Platanus
35.0    Taxodium
35.0    Liriodendron
35.0    Platanus
35.0    Ailanthus
35.0    Sequoiadendron
40.0    Platanus
40.0    Platanus
40.0    Platanus
42.0    Platanus
45.0    Platanus
```

f) District containing the oldest tree

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar
-with-dependencies.jar oldesttree /user/jin-young.bae/trees.csv /user/jin-young.ba
e/oldesttree
```

End of OldestTree

```
File Input Format Counters
        Bytes Read=16680
File Output Format Counters
        Bytes Written=2
```

Result of OldestTree

```
-bash-4.2$ hdfs dfs -cat oldesttree/part-r-00000
5
```

g) District containing the most trees

```
-bash-4.2$ yarn jar /home/jin-young.bae/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-w
ith-dependencies.jar districtcount /user/jin-young.bae/trees.csv /user/jin-young.bae
/districtcount
```

End of 1st MapReduce: TreeCount

```
File Input Format Counters
        Bytes Read=16680
File Output Format Counters
        Bytes Written=80
```

End of 2nd MapReduce: MaxCount

```
File Input Format Counters
        Bytes Read=80
File Output Format Counters
        Bytes Written=6
```

Result of 1st MapReduce: TreeCount

```
-bash-4.2$ hdfs dfs -cat temp/part-r-00000
3       1
4       1
5       2
6       1
7       3
8       5
9       1
11      1
12      29
13      2
14      3
15      1
16      36
17      1
18      1
19      6
20      3
```

Result of 2nd MapReduce: MaxCount

```
-bash-4.2$ hdfs dfs -cat districtcount/part-r-00000
16      36
```