YARN & MapReduce 2 - Karim ABED/Mohamed ZENATI

Lien du Github: https://github.com/Krimsooo/lab2_map-reduce-Karim-Mohamed

MapReduce JAVA

1.6 Send the JAR to the edge node

1.6.3 Run the job

```
[karim.abed@hadoop-edge01 ~]$ yarn jar hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-
with-dependencies.jar wordcount inputJar outputTest/
21/11/04 10:06:45 INFO impl.TimelineReaderClientImpl: Initialized TimelineReader
URI=https://hadoop-master03.efrei.online:8199/ws/v2/timeline/, clusterId=yarn-
cluster
21/11/04 10:06:
45 INFO client.AHSProxy: Connecting to Application History server at hadoop-
master03.efrei.online/163.172.102.23:10200
21/11/04 10:06:45 INFO hdfs.DFSClient: Created token for karim.abed:
HDFS_DELEGATION_TOKEN owner=karim.abed@EFREI.ONLINE, renewer=yarn, realUser=,
issueDate=1636016805927, maxDate=1636621605927, sequenceNumber=6620,
masterKeyId=77 on ha-hdfs:efrei
21/11/04 10:06:45 INFO security.TokenCache: Got dt for hdfs://efrei; Kind:
HDFS_DELEGATION_TOKEN, Service: ha-hdfs:efrei, Ident: (token for karim.abed:
HDFS_DELEGATION_TOKEN owner=karim.abed@EFREI.ONLINE, renewer=yarn, realUser=,
issueDate=1636016805927, maxDate=1636621605927, sequenceNumber=6620,
masterKeyId=77)
21/11/04 10:06:46 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to
rm2
21/11/04 10:06:46 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for
path: /user/karim.abed/.staging/job_1630864376208_4522
21/11/04 10:06:47 INFO input.FileInputFormat: Total input files to process : 0
21/11/04 10:06:47 INFO mapreduce.JobSubmitter: number of splits:0
21/11/04 10:06:47 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1630864376208_4522
21/11/04 10:06:47 INFO mapreduce.JobSubmitter: Executing with tokens: [Kind:
HDFS_DELEGATION_TOKEN, Service: ha-hdfs:efrei, Ident: (token for karim.abed:
HDFS_DELEGATION_TOKEN owner=karim.abed@EFREI.ONLINE, renewer=yarn, realUser=,
issueDate=1636016805927, maxDate=1636621605927, sequenceNumber=6620,
masterKeyId=77)]
21/11/04 10:06:47 INFO conf.Configuration: found resource resource-types.xml at
file:/etc/hadoop/1.0.3.0-223/0/resource-types.xml
21/11/04 10:06:47 INFO impl.TimelineClientImpl: Timeline service address: hadoop-
master03.efrei.online:8190
21/11/04 10:06:48 INFO impl.YarnClientImpl: Submitted application
application_1630864376208_4522
```

```
21/11/04 10:06:48 INFO mapreduce.Job: The url to track the job: https://hadoop-
master02.efrei.online:8090/proxy/application_1630864376208_4522/
21/11/04 10:06:48 INFO mapreduce.Job: Running job: job_1630864376208_4522
21/11/04 10:06:58 INFO mapreduce.Job: Job job_1630864376208_4522 running in uber
mode : false
21/11/04 10:06:58 INFO mapreduce.Job:  map 0% reduce 0%
21/11/04 10:07:08 INFO mapreduce.Job:  map 0% reduce 100%
21/11/04 10:07:08 INFO mapreduce.Job: Job job_1630864376208_4522 completed
successfully
21/11/04 10:07:08 INFO mapreduce.Job: Counters: 41
```

1.8 Remarkable trees of Paris

```
[karim.abed@hadoop-edge01 ~]$ wget https://github.com/makayel/hadoop-examples-
mapreduce/blob/main/src/test/resources/data/trees.csv
--2021-11-04 11:17:13--  https://github.com/makayel/hadoop-examples-
mapreduce/blob/main/src/test/resources/data/trees.csv
Resolving github.com (github.com)... 140.82.121.4
Connecting to github.com (github.com)|140.82.121.4|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: ‘trees.csv’

    [ <=>
] 178,973     --.-K/s   in 0.03s

2021-11-04 11:17:13 (5.61 MB/s) - ‘trees.csv’ saved [178973]

[karim.abed@hadoop-edge01 ~]$ ls
davinci.txt                                                    local.txt
message     science.txt             sudoku.dta  trees.csv
hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-with-dependencies.jar  mapper.py
reducer.py  secret-of-the-universe  test        ulysses.txt
[karim.abed@hadoop-edge01 ~]$ hdfs dfs -copyFromLocal trees.csv
```

1.8.1 Districts containing trees (very easy)

```
[karim.abed@hadoop-edge01 ~]$ yarn jar hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-
with-dependencies.jar ex1 dataset/trees.csv 1.8.1_output/
...
21/11/07 21:44:18 INFO mapreduce.Job:  map 0% reduce 0%
21/11/07 21:44:27 INFO mapreduce.Job:  map 100% reduce 0%
21/11/07 21:44:32 INFO mapreduce.Job:  map 100% reduce 100%
21/11/07 21:44:32 INFO mapreduce.Job: Job job_1630864376208_5065 completed
successfully
...

[karim.abed@hadoop-edge01 ~]$ hdfs dfs -ls 1.8.1_output
Found 2 items
```

```
-rw-r--r--   3 karim.abed karim.abed          0 2021-11-07 21:44
1.8.1_output/_SUCCESS
-rw-r--r--   3 karim.abed karim.abed         80 2021-11-07 21:44
1.8.1_output/part-r-00000
[karim.abed@hadoop-edge01 ~]$ hdfs dfs -cat 1.8.1_output/part-r-00000
3
4
5
6
7
8
9
11
12
13
14
15
16
17
18
19
20
```

1.8.2 Show all existing species (very easy)

```
[karim.abed@hadoop-edge01 ~]$ yarn jar hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-
with-dependencies.jar ex2 dataset/trees.csv 1.8.2_output/
...
21/11/07 23:04:47 INFO mapreduce.Job:  map 0% reduce 0%
21/11/07 23:04:56 INFO mapreduce.Job:  map 100% reduce 0%
21/11/07 23:05:06 INFO mapreduce.Job:  map 100% reduce 100%
21/11/07 23:05:06 INFO mapreduce.Job: Job job_1630864376208_5066 completed
successfully
...

[karim.abed@hadoop-edge01 ~]$ hdfs dfs -ls -R
drwx------   - karim.abed karim.abed          0 2021-11-07 22:33 .Trash
drwx------   - karim.abed karim.abed          0 2021-11-07 22:33 .Trash/Current
drwx------   - karim.abed karim.abed          0 2021-11-07 22:33
.Trash/Current/user
drwx------   - karim.abed karim.abed          0 2021-11-07 23:04
.Trash/Current/user/karim.abed
-rw-r--r--   3 karim.abed karim.abed   68851475 2021-11-04 11:10
.Trash/Current/user/karim.abed/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-with-
dependencies.jar
-rw-r--r--   3 karim.abed karim.abed   68851339 2021-11-07 22:33
.Trash/Current/user/karim.abed/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-with-
dependencies.jar1636320918902
-rw-r--r--   3 karim.abed karim.abed   68851339 2021-11-07 22:35
.Trash/Current/user/karim.abed/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-with-
dependencies.jar1636322004491
```

```
-rw-r--r--   3 karim.abed karim.abed   68851339 2021-11-07 22:53
.Trash/Current/user/karim.abed/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-with-
dependencies.jar1636322365091
-rw-r--r--   3 karim.abed karim.abed   68851339 2021-11-07 22:59
.Trash/Current/user/karim.abed/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-with-
dependencies.jar1636322651596
drwx------   - karim.abed karim.abed          0 2021-11-07 23:05 .staging
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-07 21:44 1.8.1_output
-rw-r--r--   3 karim.abed karim.abed          0 2021-11-07 21:44
1.8.1_output/_SUCCESS
-rw-r--r--   3 karim.abed karim.abed         80 2021-11-07 21:44
1.8.1_output/part-r-00000
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-07 23:05 1.8.2_output
-rw-r--r--   3 karim.abed karim.abed          0 2021-11-07 23:05
1.8.2_output/_SUCCESS
-rw-r--r--   3 karim.abed karim.abed        451 2021-11-07 23:05
1.8.2_output/part-r-00000
drwxr-xr-x   - karim.abed karim.abed          0 2021-10-23 21:10 data
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-01 18:54 data/10GB-sort-
input
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-01 18:55 data/10GB-sort-
output
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-01 18:55 data/10GB-sort-
validate
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-06 14:33 dataset
-rw-r--r--   3 karim.abed karim.abed      16680 2021-11-06 14:33 dataset/trees.csv
drwxr-xr-x   - karim.abed karim.abed          0 2021-10-26 23:01 gutenberg
-rw-r--r--   1 karim.abed karim.abed    1428843 2021-10-23 22:15
gutenberg/davinci.txt
-rw-r--r--   1 karim.abed karim.abed        305 2021-10-23 22:15
gutenberg/science.txt
-rw-r--r--   1 karim.abed karim.abed    1586336 2021-10-23 22:15
gutenberg/ulysses.txt
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-01 16:42 gutenberg-output
-rw-r--r--   3 karim.abed karim.abed   68851337 2021-11-07 23:04 hadoop-examples-
mapreduce-1.0-SNAPSHOT-jar-with-dependencies.jar
drwxr-xr-x   - karim.abed karim.abed          0 2021-10-28 21:29 inputJar
-rw-r--r--   1 karim.abed karim.abed     448821 2021-10-21 14:48 local.txt
drwxr-xr-x   - karim.abed karim.abed          0 2021-11-04 10:07 outputTest
-rw-r--r--   3 karim.abed karim.abed          0 2021-11-04 10:07
outputTest/_SUCCESS
-rw-r--r--   3 karim.abed karim.abed          0 2021-11-04 10:07 outputTest/part-
r-00000
drwxr-xr-x   - karim.abed karim.abed          0 2021-09-30 10:53 raw
-rw-r--r--   1 karim.abed karim.abed     448821 2021-09-30 10:52 raw/84-0.txt
-rw-r--r--   1 karim.abed karim.abed        524 2021-10-27 20:57 reducer.py
-rw-r--r--   1 karim.abed karim.abed        162 2021-10-23 20:36 sudoku.dta

[karim.abed@hadoop-edge01 ~]$ hdfs dfs -cat 1.8.2_output/part-r-00000
araucana
atlantica
australis
baccata
bignonioides
```

```
biloba
bungeana
cappadocicum
carpinifolia
colurna
coulteri
decurrens
dioicus
distichum
excelsior
fraxinifolia
giganteum
giraldii
glutinosa
grandiflora
hippocastanum
ilex
involucrata
japonicum
kaki
libanii
monspessulanum
nigra
nigra laricio
opalus
orientalis
papyrifera
petraea
pomifera
pseudoacacia
sempervirens
serrata
stenoptera
suber
sylvatica
tomentosa
tulipifera
ulmoides
virginiana
x acerifolia
```

1.8.3 Number of trees by kinds (easy)

```
[karim.abed@hadoop-edge01 ~]$ yarn jar hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-
with-dependencies.jar ex3 dataset/trees.csv 1.8.3_output/
...
21/11/07 23:30:36 INFO mapreduce.Job:  map 0% reduce 0%
21/11/07 23:30:45 INFO mapreduce.Job:  map 100% reduce 0%
21/11/07 23:30:54 INFO mapreduce.Job:  map 100% reduce 100%
21/11/07 23:30:56 INFO mapreduce.Job: Job job_1630864376208_5067 completed
successfully
```

```
...

[karim.abed@hadoop-edge01 ~]$ hdfs dfs -cat 1.8.3_output/part-r-00000
araucana        1
atlantica       2
australis       1
baccata         2
bignonioides    1
biloba          5
bungeana        1
cappadocicum    1
carpinifolia    4
colurna         3
coulteri        1
decurrens       1
dioicus         1
distichum       3
excelsior       1
fraxinifolia    2
giganteum       5
giraldii        1
glutinosa       1
grandiflora     1
hippocastanum   3
ilex            1
involucrata     1
japonicum       1
kaki            2
libanii         2
monspessulanum  1
nigra           3
nigra laricio   1
opalus          1
orientalis      8
papyrifera      1
petraea         2
pomifera        1
pseudoacacia    1
sempervirens    1
serrata         1
stenoptera      1
suber           1
sylvatica       8
tomentosa       2
tulipifera      2
ulmoides        1
virginiana      2
x acerifolia    11
```

1.8.4 Maximum height per kind of tree (average)

```
[karim.abed@hadoop-edge01 ~]$ yarn jar hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-
with-dependencies.jar ex4 dataset/trees.csv 1.8.4_output/
...
21/11/07 23:46:18 INFO mapreduce.Job:  map 0% reduce 0%
21/11/07 23:46:27 INFO mapreduce.Job:  map 100% reduce 0%
21/11/07 23:46:36 INFO mapreduce.Job:  map 100% reduce 100%
21/11/07 23:46:36 INFO mapreduce.Job: Job job_1630864376208_5070 completed
successfully
21/11/07 23:46:37 INFO mapreduce.Job: Counters: 54
...

[karim.abed@hadoop-edge01 ~]$ hdfs dfs -cat 1.8.4_output/part-r-00000
araucana        9.0
atlantica       25.0
australis       16.0
baccata         13.0
bignonioides    15.0
biloba          33.0
bungeana        10.0
cappadocicum    16.0
carpinifolia    30.0
colurna         20.0
coulteri        14.0
decurrens       20.0
dioicus         10.0
distichum       35.0
excelsior       30.0
fraxinifolia    27.0
giganteum       35.0
giraldii        35.0
glutinosa       16.0
grandiflora     12.0
hippocastanum   30.0
ilex            15.0
involucrata     12.0
japonicum       10.0
kaki            14.0
libanii         30.0
monspessulanum  12.0
nigra           30.0
nigra laricio   30.0
opalus          15.0
orientalis      34.0
papyrifera      12.0
petraea         31.0
pomifera        13.0
pseudoacacia    11.0
sempervirens    30.0
serrata         18.0
stenoptera      30.0
suber           10.0
sylvatica       30.0
tomentosa       20.0
```

```
tulipifera      35.0
ulmoides        12.0
virginiana      14.0
x acerifolia    45.0
```

1.8.5 Sort the trees height from smallest to largest (average)

```
[karim.abed@hadoop-edge01 ~]$ yarn jar hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-
with-dependencies.jar ex5 dataset/trees.csv 1.8.5_output/
...
21/11/08 22:32:00 INFO mapreduce.Job:  map 0% reduce 0%
21/11/08 22:32:09 INFO mapreduce.Job:  map 100% reduce 0%
21/11/08 22:32:18 INFO mapreduce.Job:  map 100% reduce 100%
21/11/08 22:32:19 INFO mapreduce.Job: Job job_1630864376208_5099 completed
successfully
21/11/08 22:32:19 INFO mapreduce.Job: Counters: 54
...

[karim.abed@hadoop-edge01 ~]$ hdfs dfs -cat 1.8.5_output/part-r-00000
3 - Fagus sylvatica (Fagaceae)                  2.0
89 - Taxus baccata (Taxaceae)                   5.0
62 - Cedrus atlantica (Pinaceae)                6.0
39 - Araucaria araucana (Araucariaceae)         9.0
44 - Styphnolobium japonicum (Fabaceae)         10.0
32 - Quercus suber (Fagaceae)                   10.0
95 - Pinus bungeana (Pinaceae)                  10.0
61 - Gymnocladus dioicus (Fabaceae)             10.0
63 - Fagus sylvatica (Fagaceae)                 10.0
4 - Robinia pseudoacacia (Fabaceae)             11.0
93 - Diospyros virginiana (Ebenaceae)           12.0
66 - Magnolia grandiflora (Magnoliaceae)        12.0
50 - Zelkova carpinifolia (Ulmaceae)            12.0
7 - Eucommia ulmoides (Eucomiaceae)             12.0
48 - Acer monspessulanum (Sapindacaees)         12.0
58 - Diospyros kaki (Ebenaceae)                 12.0
33 - Broussonetia papyrifera (Moraceae)         12.0
71 - Davidia involucrata (Cornaceae)            12.0
36 - Taxus baccata (Taxaceae)                   13.0
6 - Maclura pomifera (Moraceae)                 13.0
68 - Diospyros kaki (Ebenaceae)                 14.0
96 - Pinus coulteri (Pinaceae)                  14.0
94 - Diospyros virginiana (Ebenaceae)           14.0
91 - Acer opalus (Sapindaceae)                  15.0
5 - Catalpa bignonioides (Bignoniaceae)         15.0
70 - Fagus sylvatica (Fagaceae)                 15.0
2 - Ulmus carpinifolia (Ulmaceae)               15.0
98 - Quercus ilex (Fagaceae)                    15.0
28 - Alnus glutinosa (Betulaceae)               16.0
78 - Acer cappadocicum (Sapindaceae)            16.0
75 - Zelkova carpinifolia (Ulmaceae)            16.0
16 - Celtis australis (Cannabaceae)             16.0
```

```
64 - Ginkgo biloba (Ginkgoaceae)                18.0
83 - Zelkova serrata (Ulmaceae)                 18.0
23 - Aesculus hippocastanum (Sapindaceae)       18.0
60 - Fagus sylvatica (Fagaceae)                 18.0
34 - Corylus colurna (Betulaceae)               20.0
51 - Platanus x acerifolia (Platanaceae)        20.0
43 - Tilia tomentosa (Malvaceae)                20.0
15 - Corylus colurna (Betulaceae)               20.0
11 - Calocedrus decurrens (Cupressaceae)        20.0
1 - Corylus colurna (Betulaceae)                20.0
8 - Platanus orientalis (Platanaceae)           20.0
20 - Fagus sylvatica (Fagaceae)                 20.0
35 - Paulownia tomentosa (Paulowniaceae)        20.0
12 - Sequoiadendron giganteum (Taxodiaceae)     20.0
87 - Taxodium distichum (Taxodiaceae)           20.0
13 - Platanus orientalis (Platanaceae)          20.0
10 - Ginkgo biloba (Ginkgoaceae)                22.0
47 - Aesculus hippocastanum (Sapindaceae)       22.0
86 - Platanus orientalis (Platanaceae)          22.0
14 - Pterocarya fraxinifolia (Juglandaceae)     22.0
88 - Liriodendron tulipifera (Magnoliaceae)     22.0
18 - Fagus sylvatica (Fagaceae)                 23.0
24 - Cedrus atlantica (Pinaceae)                25.0
31 - Ginkgo biloba (Ginkgoaceae)                25.0
92 - Platanus x acerifolia (Platanaceae)        25.0
49 - Platanus orientalis (Platanaceae)          25.0
97 - Pinus nigra (Pinaceae)                     25.0
84 - Ginkgo biloba (Ginkgoaceae)                25.0
73 - Platanus orientalis (Platanaceae)          26.0
65 - Pterocarya fraxinifolia (Juglandaceae)     27.0
42 - Platanus orientalis (Platanaceae)          27.0
85 - Juglans nigra (Juglandaceae)               28.0
76 - Pinus nigra laricio (Pinaceae)             30.0
19 - Quercus petraea (Fagaceae)                 30.0
72 - Sequoiadendron giganteum (Taxodiaceae)     30.0
54 - Pterocarya stenoptera (Juglandaceae)       30.0
29 - Zelkova carpinifolia (Ulmaceae)            30.0
27 - Sequoia sempervirens (Taxodiaceae)         30.0
25 - Fagus sylvatica (Fagaceae)                 30.0
41 - Platanus x acerifolia (Platanaceae)        30.0
77 - Taxodium distichum (Taxodiaceae)           30.0
55 - Platanus x acerifolia (Platanaceae)        30.0
69 - Pinus nigra (Pinaceae)                     30.0
38 - Fagus sylvatica (Fagaceae)                 30.0
59 - Sequoiadendron giganteum (Taxodiaceae)     30.0
52 - Fraxinus excelsior (Oleaceae)              30.0
37 - Cedrus libanii (Pinaceae)                  30.0
22 - Cedrus libanii (Pinaceae)                  30.0
30 - Aesculus hippocastanum (Sapindaceae)       30.0
80 - Quercus petraea (Fagaceae)                 31.0
9 - Platanus orientalis (Platanaceae)           31.0
82 - Platanus x acerifolia (Platanaceae)        32.0
46 - Ginkgo biloba (Ginkgoaceae)                33.0
45 - Platanus orientalis (Platanaceae)          34.0
```

```
56 - Taxodium distichum (Taxodiaceae)           35.0
81 - Liriodendron tulipifera (Magnoliaceae)     35.0
17 - Platanus x acerifolia (Platanaceae)        35.0
53 - Ailanthus giraldii (Simaroubaceae)         35.0
57 - Sequoiadendron giganteum (Taxodiaceae)     35.0
26 - Platanus x acerifolia (Platanaceae)        40.0
74 - Platanus x acerifolia (Platanaceae)        40.0
40 - Platanus x acerifolia (Platanaceae)        40.0
90 - Platanus x acerifolia (Platanaceae)        42.0
21 - Platanus x acerifolia (Platanaceae)        45.0
```

1.8.6 District containing the oldest tree (diffcult)

```
[karim.abed@hadoop-edge01 ~]$ yarn jar hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-
with-dependencies.jar ex6_1 dataset/trees.csv 1.8.6_1_output/
...
21/11/08 22:39:04 INFO mapreduce.Job:  map 0% reduce 0%
21/11/08 22:39:13 INFO mapreduce.Job:  map 100% reduce 0%
21/11/08 22:39:17 INFO mapreduce.Job:  map 100% reduce 100%
21/11/08 22:39:18 INFO mapreduce.Job: Job job_1630864376208_5100 completed
successfully
21/11/08 22:39:18 INFO mapreduce.Job: Counters: 54
...

[karim.abed@hadoop-edge01 ~]$ hdfs dfs -cat 1.8.6_1_output/part-r-00000
1601    5
```