

Sprawozdanie

MapReduce 1

Oscar Teeninga

1. Konfiguracja testowa.

Cluster: MPR_cluster Waiting Cluster ready after last step completed.

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

ID: j-2E8YZJ0SZ6HAX

Creation date: 2021-04-12 00:11 (UTC+2)

Elapsed time: 19 hours

After last step completes: Cluster waits

Termination protection: Off [Change](#)

Tags: -- [View All](#) / [Edit](#)

Master public DNS: ec2-3-238-101-28.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-5.32.0

Hadoop distribution: Amazon 2.10.1

Applications: Hive 2.3.7, Hue 4.8.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2

Log URI: s3://aws-logs-343636106896-us-east-1/elasticmapreduce/

EMRFS consistent view: Disabled

Custom AMI ID: --

Application user interfaces

Persistent user interfaces: [YARN timeline server](#), [Tez UI](#)

On-cluster user Not Enabled [Enable an SSH Connection](#)

interfaces: [View All](#)

Network and hardware

Availability zone: us-east-1d

Subnet ID: [subnet-51701a37](#) [View](#)

Master: Running 1 m4.large

Core: Running 2 m4.large

Task: --

Cluster scaling: Not enabled

Security and access

Key name: teeninga_oscar_mpr

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master: [sg-0f9118605c18288bd](#) [View](#) (ElasticMapReduce-master)

Security groups for Core & Task: [sg-0a81c41dc8f8980ff](#) [View](#) (ElasticMapReduce-slave)

2. Zbiór pobrałem za pomocą download.sh w przedziale 1300..1400, a więc 100 tekstów.

```
[hadoop@ip-172-31-9-118 books]$ ls
1300-0.txt 1307.txt 1315.txt 1324.txt 1332-0.txt 1341.txt 1348.txt 1356.txt 1366-0.txt 1373.txt 1379-0.txt 1386.txt 1393.txt
1301-0.txt 1308-0.txt 1316-0.txt 1325.txt 1332.txt 1342-0.txt 1349.txt 1357-0.txt 1366.txt 1374-0.txt 1379.txt 1387-0.txt 1394.txt
1302.txt 1309.txt 1316.txt 1326.txt 1333.txt 1343-0.txt 1350-0.txt 1357.txt 1367.txt 1374.txt 1380-0.txt 1387.txt 1395.txt
1303-0.txt 1310-0.txt 1317-0.txt 1327-0.txt 1334-0.txt 1343.txt 1351.txt 1358.txt 1368-0.txt 1375-0.txt 1380.txt 1388.txt 1396-0.txt
1303.txt 1311.txt 1317.txt 1327.txt 1334.txt 1344-0.txt 1352-0.txt 1359.txt 1369-0.txt 1375.txt 1381-0.txt 1389-0.txt 1396.txt
1304.txt 1312-0.txt 1318.txt 1328-0.txt 1335.txt 1344.txt 1352.txt 1360.txt 1369.txt 1376-0.txt 1382-0.txt 1389.txt 1397.txt
1305-0.txt 1312.txt 1319.txt 1328.txt 1336.txt 1345-0.txt 1353-0.txt 1361.txt 1370.txt 1376.txt 1383-0.txt 1390-0.txt 1398.txt
1305.txt 1313-0.txt 1320-0.txt 1329-0.txt 1337.txt 1345.txt 1353.txt 1362.txt 1371-0.txt 1377-0.txt 1384.txt 1390.txt 1399-0.txt
1306-0.txt 1313.txt 1321-0.txt 1330-0.txt 1338.txt 1346.txt 1354-0.txt 1363.txt 1371.txt 1377.txt 1385-0.txt 1391-0.txt 1400-0.txt
1306.txt 1314-0.txt 1322-0.txt 1330.txt 1339-0.txt 1347.txt 1354.txt 1364.txt 1372-0.txt 1378-0.txt 1385.txt 1391.txt
```

3. Uruchomiłem przykładową implementację Word Count w postaci mapper.py oraz reducer.py zgodnie z instrukcją. Otrzymany czas to 603,8 s.

```
21/04/12 08:05:37 INFO mapreduce.Job: Counters: 52
  File System Counters
    FILE: Number of bytes read=8285246
    FILE: Number of bytes written=60574685
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=64876854
    HDFS: Number of bytes written=4494043
    HDFS: Number of read operations=435
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=6
  Job Counters
    Killed map tasks=1
    Killed reduce tasks=1
    Launched map tasks=142
    Launched reduce tasks=4
    Data-local map tasks=98
    Rack-local map tasks=44
    Total time spent by all maps in occupied slots (ms)=111059136
    Total time spent by all reduces in occupied slots (ms)=47329248
    Total time spent by all map tasks (ms)=2313732
    Total time spent by all reduce tasks (ms)=493013
    Total vcore-milliseconds taken by all map tasks=2313732
    Total vcore-milliseconds taken by all reduce tasks=493013
    Total megabyte-milliseconds taken by all map tasks=3553892352
    Total megabyte-milliseconds taken by all reduce tasks=1514535936
  Map-Reduce Framework
    Map input records=1279689
    Map output records=10965196
    Map output bytes=84611156
    Map output materialized bytes=19988214
    Input split bytes=17868
    Combine input records=0
    Combine output records=0
    Reduce input groups=370108
    Reduce shuffle bytes=19988214
    Reduce input records=10965196
    Reduce output records=370108
    Spilled Records=21930392
    Shuffled Maps =426
    Failed Shuffles=0
    Merged Map outputs=426
    GC time elapsed (ms)=42993
    CPU time spent (ms)=306170
    Physical memory (bytes) snapshot=67977322496
    Virtual memory (bytes) snapshot=484520374272
    Total committed heap usage (bytes)=55018258432
  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=64858986
  File Output Format Counters
    Bytes Written=4494043
21/04/12 08:05:37 INFO streaming.StreamJob: Output directory: books-output

real    10m3.814s
user    0m9.806s
sys     0m0.725s
[hadoop@ip-172-31-9-118 lab5]$ ls
```

4. Uruchomiłem własną sekwencyjną wersję Word Count oraz zmierzyłem czasy.

```
[hadoop@ip-172-31-9-118 lab5]$ strings books/*.txt | ./wordcount.py  
6.20159101486
```

5. Podsumowanie

Czas sekwencyjne był znacznie krótszy (6.2s) względem MapReduce (603s) o mniej-więcej dwa rzędy wielkości.