

## hy-562 - Report for Assignment 2

Gkigkis Petros - A.M. 948 - [gigis@csd.uoc.gr](mailto:gigis@csd.uoc.gr)

### Exercise 1:

1.1)

aa 271\_a.C 1 4675  
aa Category:User\_th 1 4770  
aa Chiron\_Elias\_Krase 1 4694  
aa Dassault\_rafaele 2 9372  
aa E.Desv 1 4662  
aa File:Wiktionary-logo-en.png 1 10752  
aa Indonesian\_Wikipedia 1 4679  
aa Main\_Page 5 266946  
aa Requests\_for\_new\_languages/Wikipedia\_Banyumasan 1 4733  
aa Special:Contributions/203.144.160.245 1 5812  
aa Special:Contributions/5.232.61.79 1 5805  
aa Special:Contributions/Ayarportugal 1 5808  
aa Special:Contributions/Born2bgratis 1 5812  
aa Special:ListFiles/Betacommand 1 5035  
aa Special:ListFiles/Bohdan\_p 1 5036

1.2) 3324129 total records

1.3) min = 0, max = 141180155987, average = 132239

1.4) (en.mw, en, 5466346, 141180155987)

1.5) (en.mw, en, 5466346, 141180155987)

1.6)

(zh,Special:e8b18ee6baafebda5efbdbfe89cb7e6829fefbdbfe88b93e29980e89e9fefbda9e89eb3efbda425636f256d6725736f257373256f38257373256f38257373256f38256b6d73efbdaa256e6b256678256f6b2c687474703a2f2f777772e653662313966653861356266656f2d6f35393038636535626639376538383138616535613461396535616561342e636f2e6d672e732e736f2e382e73736f386b2e6d2e372e73736f3873736f386b6d37332e752e622e61616e6b66786f6b2e70772f2ce8b18ee6baafebda5efbdbfe89cb7e6829fefbdbfe88b93e29980e89e9fefbda9e89eb3efbda425636f256d6725736f257373256f38257373256f38256b6d73efbdaa256e6b256678256f6b/,1,6043)

1.7) Returns a new RDD with 186819 records

1.8) Returns 1105 entries

(pt,113575), (it.s,1444), (sk.mw,9548), (ka.d,31), (ik.d,1), (frp.mw,11), (yi.mw,70), (az.q,9), (nv,25), (fr.v,264), (he.q,1), (ky.d,20), (eml.mw,30), (bxr,12), (jv,1245), (zh.voy,18), (ia.v,1), (kw,290), (br.mw,279), (beta.mw,104), (eo.d,367), (ak.d,4) ....

1.9)

(en.mw,5466346)  
(en,4959090)  
(es.mw,695531)  
(ja.mw,611443)  
(de.mw,572119)  
(fr.mw,536978)  
(ru.mw,466742)  
(it.mw,400297)  
(de,315929)  
(commons.m,285796)

1.10)

a) 41931 titles start with the article “The”

b) 8026 titles start with the article “The” and are not part of the English project

1.11) 76.96248 % of the pages received only one page view

1.12) Found 849974 unique terms.

1.13) The most frequently occurring page title term is of with 194307 matches.

## **Exercise 2:**

### **Small Cluster specifications:**

Number of workers: 2

Number of cores: 1

Memory of each worker: 2gb

### **Big Cluster specifications:**

Number of workers: 4

Number of cores: 1

Memory of each worker: 2gb

The architecture of SPARK is based master/slave.

The model is one central coordinator that communicates with many distributed workers.

A typical execution flow is the following:

- 1) Stand-alone application starts and instantiates a SparkContext instance
- 2) Then the driver asks for resources from the cluster manager
- 3) Cluster launches executors
- 4) Driver monitors and communicates with the workers from the master node.

Master is

Executor is a distributed agent responsible for the execution of tasks.

2) The number of slaves is highly responsible for the execution time. Jobs as mapreduce can run in parallel using the data collection RDD. However, the higher the better number of slaves is not always true. The number of slaves on a job depends on the size of data and the kind of processing. Finally, transferring data using the network can add significant processing delay.

3)

I would go for the b option. The range of n can be from 1 to 14. (keeping two cores for the master node) In case of 14 the available memory will be about 4.5 GB which would still work due to the fact the slave will take a part of the RDD and not the full table. I would go for 7 slaves to avoid delay on synchronizing all slaves and also network data transfer.

Also based on the following “white” paper from facebook.

### **Apache Spark @Scale: A 60 TB+ production use case**

<https://code.facebook.com/posts/1671373793181703/apache-spark-scale-a-60-tb-production-use-case/>

They mention the following:

**Configuring number of tasks:** Since our input size is 60 T and each HDFS block size is 256 M, we were spawning more than 250,000 tasks for the job. Although we were able to run the Spark job with such a high number of tasks, we found that there is significant performance degradation when the number of tasks is too high. We introduced a configuration parameter to make the map input size configurable, so we can reduce that number by 8x by setting the input split size to 2 GB.

So with a master (2 cores) and 7 slaves (2 cores per slave) it seems as the best option.

Based on observation from the Small and Big cluster the big cluster seems faster on processing. However, I am using one core per slave and the datafile is too small to get a clean picture.

Execution times:

#	Ex:3	Ex:5	Ex:6	Ex:7	Ex:12	Ex:13
Small Cluster	6 sec	2 sec	0.9 sec	1,9 sec	8 sec	9 sec
Big Cluster	7,9 sec	1,5 sec	0,7 sec	1,3 sec	8 sec	7 sec

The given google cloud will perform better from my emulated topologies.