



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Πανεπιστήμιο Πειραιώς

Τμήμα Πληροφορικής

Κατανεμημένα και πολυπεξεργαστικά υπολογιστικά συστήματα

7ο εξάμηνο

Τσολάκης Σταμάτιος p18161

Κατέβας Χρήστος p18068

Θέμα 1:

Το MapReduce είναι ένα μοντέλο προγραμματισμού για την επεξεργασία μεγάλων συνόλων δεδομένων με έναν παράλληλο, κατανεμημένο αλγόριθμο σε ένα σύμπλεγμα υπολογιστών. Ένα πρόγραμμα MapReduce αποτελείται από μια διαδικασία χαρτογράφησης, η οποία εκτελεί φιλτράρισμα και ταξινόμηση και μια μέθοδο μείωσης, η οποία εκτελεί μια λειτουργία σύνοψης. Το "MapReduce System" ενορχηστρώνει την επεξεργασία ομαδοποιώντας τους κατανεμημένους διακομιστές, εκτελώντας τις διάφορες εργασίες παράλληλα, διαχειρίζοντας όλες τις επικοινωνίες και τις μεταφορές δεδομένων μεταξύ των διαφόρων τμημάτων του συστήματος και παρέχοντας πλεονασμό και ανοχή σφαλμάτων. Το μοντέλο είναι μια εξειδίκευση της στρατηγικής split-apply-combine για ανάλυση δεδομένων. Οι βασικές συνεισφορές του πλαισίου MapReduce δεν είναι ο πραγματικός χάρτης και οι λειτουργίες μείωσης, αλλά η επεκτασιμότητα και η ανοχή σφαλμάτων που επιτυγχάνεται για μια ποικιλία εφαρμογών με τη βελτιστοποίηση της μηχανής εκτέλεσης. Ως εκ τούτου, μια εφαρμογή με ένα νήμα του MapReduce δεν είναι συνήθως ταχύτερη από μια παραδοσιακή (μη-MapReduce) υλοποίηση. Τα “πλεονεκτήματα” παρατηρούνται μόνο με εφαρμογές πολλαπλών νημάτων σε υλικό πολλών επεξεργαστών. Η χρήση αυτού του μοντέλου είναι επωφελής μόνο όταν η βελτιστοποιημένη λειτουργία κατανεμημένης τυχαίας αναπαραγωγής (η οποία μειώνει το κόστος επικοινωνίας δικτύου) και τα χαρακτηριστικά ανοχής σφαλμάτων του πλαισίου MapReduce “μπαίνουν στο παιχνίδι”. Η βελτιστοποίηση του κόστους επικοινωνίας είναι απαραίτητη για έναν καλό αλγόριθμο MapReduce. Οι βιβλιοθήκες MapReduce έχουν γραφτεί σε πολλές γλώσσες προγραμματισμού, με διαφορετικά επίπεδα βελτιστοποίησης. Μερικές δημοφιλείς υλοποιήσεις ανοιχτού κώδικα που έχει υποστήριξη για κατανεμημένες ανακατέματα είναι:

- 1)Apache Hadoop
- 2)Apache Spark
- 3)Google BigQuery

Θέμα 2:

Η υλοποίηση και η εκτέλεση της εφαρμογής πραγματοποιήθηκε σε Ubuntu 20.04. Χρησιμοποιήθηκαν οι παρακάτω τεχνολογίες:

- Apache Hadoop v.3.2.1
- Python3
- Java jdk (javac1.8.0_312)

Σημείωση

Δεν διαθέτουμε “ισχυρούς” υπολογιστές για να σηκώσουμε παράλληλα πολλά virtual machines και να έχουμε μια πλήρη προσομοίωση ενός κατανεμημένου συστήματος. Συνεπώς η όλη διαδικασία θα εκτελεσθεί σε 1 node (master και slave(s) είναι ένα μηχάνημα), στο hadoop.

Ακολουθούν screenshots από την εκτέλεση του προγράμματος:

```

cmark0@mark0:~$ cd Downloads/
cmark0@mark0:~/Downloads$ ~/Downloads/hadoop-3.2.1/bin/hdfs namenode -format
2022-02-26 09:01:30,085 INFO namenode.NameNode: STARTUP_MSG:
/*XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*/
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = mark0/127.0.0.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.2.1
STARTUP_MSG: classpath = /home/c/Downloads/hadoop-3.2.1/etc/hadoop:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/zookeeper-3.4.13.jar:/ho
me/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/log4j-1.2.17.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/kerb-simplekdc-1.0.1.ja
r:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/audience-annota
tions-0.5.0.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jcip-annotations-1.0-1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common
lib/ljsr311-api-1.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/kerb-client-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common
lib/lib/kerb-common-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/kerby-config-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/h
adoop/common/lib/jsch-0.1.54.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jetty-io-9.3.24.v20180605.jar:/home/c/Downloads/hadoop-3.2.1/
share/hadoop/common/lib/kerb-admin-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/kerby-dn-1.0.1.jar:/home/c/Downloads/hadoop-3.2
.1/share/hadoop/common/lib/jsp-api-2.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3
.2.1/share/hadoop/common/lib/curator-framework-2.13.0.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jersey-json-1.19.jar:/home/c/Downloa
ds/hadoop-3.2.1/share/hadoop/common/lib/asm-5.0.4.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jaxb-api-2.2.11.jar:/home/c/Downloads/h
adoop-3.2.1/share/hadoop/common/lib/kerby-asn1-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/woodstox-core-5.0.3.jar:/home/c/Downd
loads/hadoop-3.2.1/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/snappy-java-1.0.5.jar:/home/c
/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jetty-xml-9.3.24.v20180605.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-math3
-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/kerb-identit
y-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jersey-servlet-1.19.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/co
mmon/lib/netty-3.10.5.Final.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jetty-security-9.3.24.v20180605.jar:/home/c/Downloads/hadoop-3
.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/curator-client-2.13.0.jar:/home/c/
Downloads/hadoop-3.2.1/share/hadoop/common/lib/jackson-core-2.9.8.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jetty-webapp-9.3.24.v20
180605.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/htrace-core4-1.0-incubating.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/comon
lib/kerb-core-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/json-smart-2.3.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common
lib/ListenableFuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-text-1.4.jar:/h
ome/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/kerby-pkix-1
.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jul-to-slf4j-1.7.25.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/slf4j
-log4j12-1.7.25.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/httpclient-4.5.6.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/li
b/slf4j-api-1.7.25.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jersey-core-1.19.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common
lib/gson-2.2.4.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-beanutils-1.9.3.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop
common/lib/hadoop-annotations-3.2.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jetty-server-9.3.24.v20180605.jar:/home/c/Downloads/h
adoop-3.2.1/share/hadoop/common/lib/hadoop-auth-3.2.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-io-2.5.jar:/home/c/Download
s/hadoop-3.2.1/share/hadoop/common/lib/token-provider-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jackson-databind-2.9.8.jar:/h
ome/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jetty-util-9.3.24.v20180605.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/paraname
r-2.3.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/jsr305-3.0.0.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/accessors-s
mart-1.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/error_prone_annotations-2.0.0.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/co
mmon/lib/kerby-util-1.0.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/javax-servlet-api-1.0.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-collections4-4.0.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-lang3-3.0.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-math3-3.1.1.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-codec-1.11.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-io-2.5.jar:/home/c/Downloads/hadoop-3.2.1/share/hadoop/common/lib/commons-net-3.0.9.jar:/home/c/Downloads/hadoop-3.2.1/shar
e/hadoop/common/lib/commons-pool2-2.4.2.jar:/home/c/Downloads/hadoop-3.
```

```
2022-02-26 09:01:33,774 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-cdfs/name/current/fsimage.ckpt_00000000000000000000 of size 390 bytes saved in 0 seconds.
2022-02-26 09:01:33,854 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2022-02-26 09:01:33,872 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2022-02-26 09:01:33,874 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at mark0/127.0.0.1
*****/
c@mark0:~/Downloads$ ~/Downloads/hadoop-3.2.1/sbin/start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as c in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [mark0]
Starting resourcemanager
Starting nodemanagers
c@mark0:~/Downloads$ hdfs dfs -ls /
c@mark0:~/Downloads$ hdfs dfs -mkdir /map
c@mark0:~/Downloads$ hdfs dfs -copyFromLocal /home/c/Downloads/measurements.csv /map/
2022-02-26 09:08:52,116 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
c@mark0:~/Downloads$ hadoop jar /home/c/Downloads/hadoop-streaming-2.7.3.jar -input /map/measurements.csv -output /map/output -mapper "python3 mapper.py" -file ./mapper.py -reducer "python3 reducer.py" -file ./reducer.py
2022-02-26 09:20:10,691 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [./mapper.py, ./reducer.py, /tmp/hadoop-unjar2347479977087507987/] [] /tmp/streamjob5109776314818093494.jar tmpDir=null
2022-02-26 09:20:13,767 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2022-02-26 09:20:14,539 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2022-02-26 09:20:15,348 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/c/.staging/job_1645858955261_0001
2022-02-26 09:20:15,788 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2022-02-26 09:20:16,160 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2022-02-26 09:20:16,246 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2022-02-26 09:20:16,488 INFO mapred.FileInputFormat: Total input files to process : 1
2022-02-26 09:20:16,638 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2022-02-26 09:20:16,710 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2022-02-26 09:20:16,752 INFO mapreduce.JobSubmitter: number of splits:2
2022-02-26 09:20:17,235 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2022-02-26 09:20:17,358 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1645858955261_0001
2022-02-26 09:20:17,359 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-02-26 09:20:18,179 INFO conf.Configuration: resource-types.xml not found
2022-02-26 09:20:18,180 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-02-26 09:20:19,622 INFO impl.YarnClientImpl: Submitted application application_1645858955261_0001
2022-02-26 09:20:19,805 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1645858955261_0001/
```

```
2022-02-26 09:20:18,180 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-02-26 09:20:19,622 INFO impl.YarnClientImpl: Submitted application application_1645858955261_0001
2022-02-26 09:20:19,805 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1645858955261_0001/
2022-02-26 09:20:19,811 INFO mapreduce.Job: Running job: job_1645858955261_0001
2022-02-26 09:20:44,834 INFO mapreduce.Job: Job job_1645858955261_0001 running in uber mode : false
2022-02-26 09:20:44,837 INFO mapreduce.Job: map 0% reduce 0%
2022-02-26 09:21:06,691 INFO mapreduce.Job: map 100% reduce 0%
2022-02-26 09:21:20,948 INFO mapreduce.Job: map 100% reduce 100%
2022-02-26 09:21:22,022 INFO mapreduce.Job: Job job_1645858955261_0001 completed successfully
2022-02-26 09:21:22,368 INFO mapreduce.Job: Counters: 54
```

File System Counters

```
FILE: Number of bytes read=8928076
FILE: Number of bytes written=18544813
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=24502341
HDFS: Number of bytes written=49
HDFS: Number of read operations=11
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=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=36191
Total time spent by all reduces in occupied slots (ms)=12233
Total time spent by all map tasks (ms)=36191
Total time spent by all reduce tasks (ms)=12233
Total vcore-milliseconds taken by all map tasks=36191
Total vcore-milliseconds taken by all reduce tasks=12233
Total megabyte-milliseconds taken by all map tasks=37059584
Total megabyte-milliseconds taken by all reduce tasks=12526592
```

Map-Reduce Framework

```
Map input records=782958
Map output records=782958
Map output bytes=7362154
Map output materialized bytes=8928082
Input split bytes=188
Combine input records=0
Combine output records=0
Reduce input groups=4
```

```
Total vcore-milliseconds taken by all map tasks=36191
Total vcore-milliseconds taken by all reduce tasks=12233
Total megabyte-milliseconds taken by all map tasks=37059584
Total megabyte-milliseconds taken by all reduce tasks=12526592
Map-Reduce Framework
  Map input records=782958
  Map output records=782958
  Map output bytes=7362154
  Map output materialized bytes=8928082
  Input split bytes=188
  Combine input records=0
  Combine output records=0
  Reduce input groups=4
  Reduce shuffle bytes=8928082
  Reduce input records=782958
  Reduce output records=4
  Spilled Records=1565916
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=648
  CPU time spent (ms)=17120
  Physical memory (bytes) snapshot=790638592
  Virtual memory (bytes) snapshot=7704297472
  Total committed heap usage (bytes)=640679936
  Peak Map Physical memory (bytes)=303820800
  Peak Map Virtual memory (bytes)=2566344704
  Peak Reduce Physical memory (bytes)=187219968
  Peak Reduce Virtual memory (bytes)=2572345344
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=24502153
File Output Format Counters
  Bytes Written=49
```

```
2022-02-26 09:21:22,369 INFO streaming.StreamJob: Output directory: /map/output
c@marko:~/Downloads$
```

```

Reduce output records=4
Spilled Records=1565916
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=648
CPU time spent (ms)=17120
Physical memory (bytes) snapshot=790638592
Virtual memory (bytes) snapshot=7704297472
Total committed heap usage (bytes)=640679936
Peak Map Physical memory (bytes)=303820800
Peak Map Virtual memory (bytes)=2566344704
Peak Reduce Physical memory (bytes)=187219968
Peak Reduce Virtual memory (bytes)=2572345344

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=24502153
File Output Format Counters
  Bytes Written=49
2022-02-26 09:21:22,369 INFO streaming.StreamJob: Output directory: /map/output
c@mark0:~/Downloads$ hdfs dfs -copyToLocal /map/output /home/c/Desktop/
copyToLocal: '/home/c/Desktop/output/_SUCCESS': File exists
copyToLocal: '/home/c/Desktop/output/part-00000': File exists
c@mark0:~/Downloads$ hdfs dfs -copyToLocal /map/output /home/c/Desktop/
2022-02-26 09:34:04,550 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
c@mark0:~/Downloads$ cd ..
c@mark0:~$ cd Desktop/
c@mark0:~/Desktop$ cd output/
c@mark0:~/Desktop/output$ ls
part-00000 _SUCCESS
c@mark0:~/Desktop/output$ cat part-00000
axaia1 58138
axaia2 257393
pos_id 1
unipi 467426
c@mark0:~/Desktop/output$ 

```

Τα τελικά αποτελέσματα είναι:

axaia1: 58138

axaia2: 257393

unipi: 467426

Security is off.

Safemode is off.

17 files and directories, 5 blocks (5 replicated blocks, 0 erasure coded block groups) = 22 total filesystem object(s).

Heap Memory used 119.19 MB of 222 MB Heap Memory. Max Heap Memory is 850 MB.

Non Heap Memory used 52.48 MB of 53.65 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	18.66 GB
Configured Remote Capacity:	0 B
DFS Used:	23.87 MB (0.12%)
Non DFS Used:	16.67 GB
DFS Remaining:	1019.16 MB (5.33%)
Block Pool Used:	23.87 MB (0.12%)
DataNodes usages% (Min/Median/Max/stdDev):	0.12% / 0.12% / 0.12% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)

Για την διαδικασία του MapReduce απαιτούνται 2 προγράμματα.
mapper.py :

```
import sys

for line in sys.stdin:
    line=line.strip()
    words=line.split(',')
    print(words[0], 1)
```

Ο mapper “διαβάζει” από το standard input(χρήστη της βιβλιοθήκης sys) τα δεδομένα (γραμμές του αρχείου csv). Αφαιρούμε όλους τους κενούς χαρακτήρες (με την βοήθεια της συνάρτησης strip()) και χωρίζουμε τη γραμμή (string) σε πίνακα με συμβολοσειρών με κριτήριο τον χαρακτήρα ‘,’ (συνάρτηση split()). Τέλος εκτυπώνει από τον πίνακα words το πρώτο κελί (δηλαδή το pos_id) με τον αριθμό ‘1’, που υποδηλώνει ότι μετρήθηκε.

```
reducer.py
import sys
prev_word = None
prev_count = 0

for line in sys.stdin:
    line = line.strip()
    word, count = line.split(' ',1)

    count = int(count)

    if prev_word == word:
        prev_count +=count
    else:
        if prev_word:
            print('%s\t%s' % (prev_word, prev_count))
            prev_count = count
            prev_word = word
if prev_word == word:
    print('%s\t%s' % (prev_word, prev_count))
```

Ο reducer “διαβάζει” τα αποτελέσματα που εκτύπωσε ο mapper. Γίνεται ο διαχωρισμός των γραμμών με βάση αυτά που ορίσαμε στον mapper. Η συνθήκη if-else δουλεύει μόνο επειδή το hadoop ταξινομεί αλφαβητικά τα αποτελέσματα που βγάζει ο mapper (προσθέτει διαδοχικά τους άσους που συναντά μετά από τα pos_id που γράφτηκαν στο standard output).

Για να τρέξει το hadoop πρέπει να γίνουν οι ακόλουθες τροποποιήσεις στα αρχεία του hadoop(core, yarn, hdfs, mapred). Αν υποθέσουμε ότι έχουμε κατεβάσει το hadoop στο φάκελο Downloads τα αρχεία που απαιτούν τροποποίηση βρίσκονται στο path “~/Downloads/hadoop-3.2.1/etc/hadoop\$ “

```
4 you may not use this file except in compliance with the License.
5 You may obtain a copy of the License at
6
7 http://www.apache.org/licenses/LICENSE-2.0
8
9 Unless required by applicable law or agreed to in writing, software
10 distributed under the License is distributed on an "AS IS" BASIS,
11 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12 See the License for the specific language governing permissions and
13 limitations under the License. See accompanying LICENSE file.
14 -->
15
16
17
18 <configuration>
19   <property>
20     <name>yarn.nodemanager.aux-services</name>
21     <value>mapreduce_shuffle</value>
22   </property>
23   <property>
24     <name>yarn.nodemanager.env-whitelist</name>
25     <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_HOME,PATH,LANG,TZ,HADOOP_M/
26   </property>
27   <property>
28     <name>mapreduce.map.memory.mb</name>
29     <value>1024</value>
30   </property>
31   <property>
32     <name>yarn.scheduler.capacity.root.support.user-limit-factor</name>
33     <value>2</value>
34   </property>
35   <property>
36     <name>yarn.nodemanager.disk-health-checker.min-healthy-disks</name>
37     <value>0.0</value>
38   </property>
39   <property>
40     <name>yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-percentage</name>
41     <value>100.0</value>
42   </property>
43 </configuration>
```

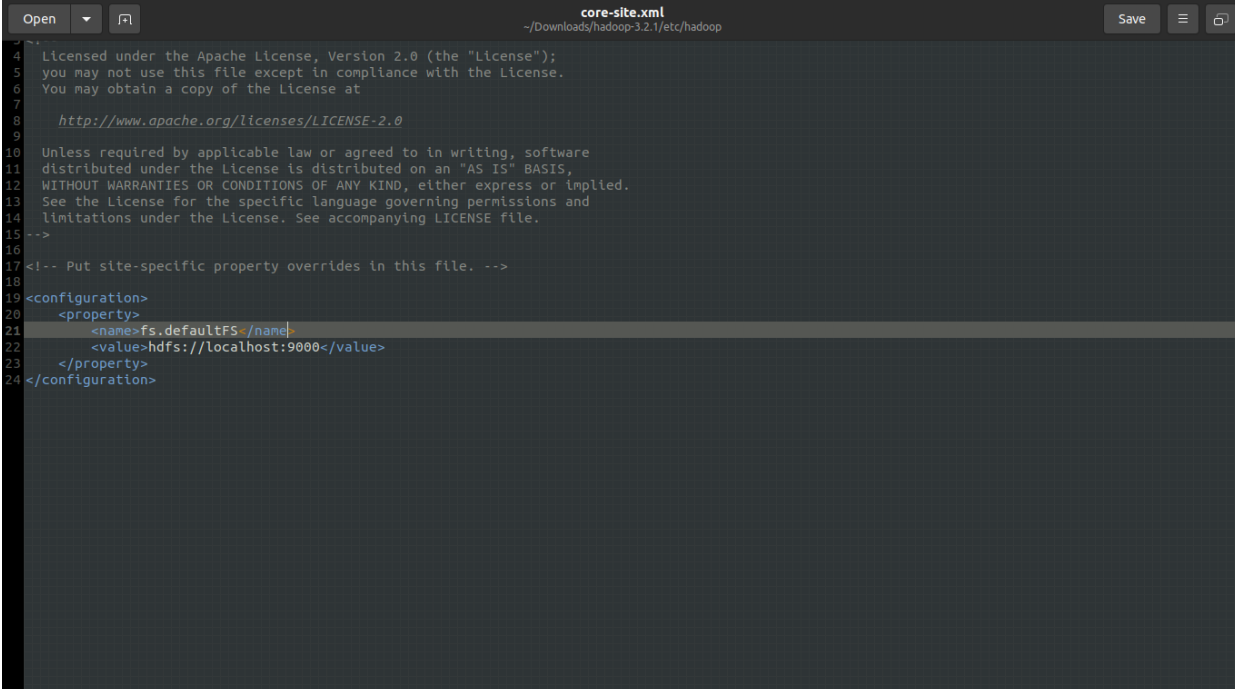
yarn-site.xml

```

1 <?xml version="1.0"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>mapreduce.framework.name</name>
22     <value>yarn</value>
23   </property>
24   <property>
25     <name>mapreduce.application.classpath</name>
26     <value>${HADOOP_MAPRED_HOME}/share/hadoop/mapreduce/*:${HADOOP_MAPRED_HOME}/share/hadoop/mapreduce/lib/*</value>
27   </property>
28   <property>
29     <name>yarn.scheduler.maximum-allocation-mb</name>
30     <value>1024</value>
31   </property>
32 </configuration>

```

mapred.site.xml (εδώ ορίζουμε και το μέγεθος της μνήμης που θα χρειαστούν τα slaves nodes πχ 1024mb)



```

1 ~
2 Licensed under the Apache License, Version 2.0 (the "License");
3 you may not use this file except in compliance with the License.
4 You may obtain a copy of the License at
5
6 http://www.apache.org/licenses/LICENSE-2.0
7
8 Unless required by applicable law or agreed to in writing, software
9 distributed under the License is distributed on an "AS IS" BASIS,
10 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
11 See the License for the specific language governing permissions and
12 limitations under the License. See accompanying LICENSE file.
13 -->
14
15 <!-- Put site-specific property overrides in this file. -->
16
17 <configuration>
18   <property>
19     <name>fs.defaultFS</name>
20     <value>hdfs://localhost:9000</value>
21   </property>
22 </configuration>

```

core-site.xml

```
1<?xml version="1.0" encoding="UTF-8"?>
2<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3<!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8   http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17<!-- Put site-specific property overrides in this file. -->
18
19<configuration>
20   <property>
21     <name>dfs.replication</name>
22     <value>1</value>
23   </property>
24</configuration>
```

hdfs-site.xml