

Big Data Εργασία 1

## Dai16054 Νικόλαος Στεφανίδης

## Dai15057 Δημήτριος Τουργαίδης

## Dai16060 Κωσταντίνος Τσίωλης

## Dai16067 Μιχούλης Γεώργιος

3.2.1 Τον αλγόριθμο που χρησιμοποιήσαμε για να λύσουμε το πρόβλημα

Για να λύσουμε την άσκησηχρησιμοποιήσαμε δυο αλγορίθμους τον Mapper και τον Reducer. Ο Mapperέπαιρνε τα δεδομένα που μας δώσατε και τα έφερνε στη μορφή που θέλαμε, δηλαδήκρατούσε τα συγκεκριμέναπεδία τα οποία μας ενδιέφεραν και τα έκανεprint σε μορφήkey και value, οπού ως Keyέστελνε τα πεδίαlast\_name, first\_name, middle\_name , rest\_citu\_desc και ως value το NCID. Έπειτα o Reducer παίρνει τα key value ταξινομημένα και για τις εγγραφές με τα ίδιαkey τις κρατάει και παίρνει τα Value τους και τα ενώνει οπότε στο τέλος εμφανίζει εγγραφές που έχουν ίδιο value και όλα τα διαφορετικά τους key ενωμένα. Στα αποτελέσματα του reducer στην περιπτώσει που κάποιος δεν έχει κάποιο από το πεδία στην θέση του εμφανίζει ‘-‘.

3.2.2 Τα προβλήματα που αντιμετωπίσαμε κατά την επίλυση της άσκησης

Τα προβλήματα που είχαμε να αντιμετωπίσουμε ήταν αφού τα στήσαμε τα μηχανήματα και τρέχαμε τους αλγορίθμους μας συνεχώς μας έβγαζεerror ότι δεν είχαμε διαθέσιμη μνήμη στα temporary μας φακέλους και ο λόγος ήταν καθώς μας είχαν“hackarei”οπότε τρέχανε παράλληλα διεργασίες. Ο τρόπος που το αντιμετωπίσαμε αυτό ήταν πως αρχικά διαγράψαμε τους φακέλους myhdfs από όλα τα μηχανάκια μας και ξανά κάναμε την εντολή για format και τέλος τρέξαμε τις εντολές που μας είχατε δώσει ώστε να ασφαλίσουμε τα μηχανάκια μας. Έπειτα το πρόβλημα που αντιμετωπίσαμε είναι αφού τρέξαμε τους αλγορίθμους μας για τον μέγιστο αριθμό των κόμβων μας έπειτα όταν πήγαμε να τρέξουμε για ένα κόμβο μας έβγαζε το εξής error:

user@snf-867379:~$ ~/hadoop/bin/hadoop jar ~/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.9.2.jar -mapper ~/mapper.py -reducer ~/reducer.py -input -outputegw11packageJobJar: [/tmp/hadoop-unjar5945154745261588749/] [] /tmp/streamjob4550025453817180041.jar tmpDir=null19/04/22 18:42:27 INFO client.RMProxy: Connecting to ResourceManager at master/83.212.102.217:803219/04/22 18:42:27 INFO client.RMProxy: Connecting to ResourceManager at master/83.212.102.217:803219/04/22 18:42:28 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/user/.staging/job\_1555947652137\_000219/04/22 18:42:28 ERROR streaming.StreamJob: Error Launching job : Cannot delete /tmp/hadoop-yarn/staging/user/.staging/job\_1555947652137\_0002. Name node is in safe mode.The reported blocks 118 needs additional 43 blocks to reach the threshold 0.9990 of total blocks 162.The number of live datanodes 1 has reached the minimum number 0. Safe mode will be turned off automatically once the thresholds have been reached. NamenodeHostName:master

at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.newSafemodeException(FSNamesystem.java:1412

at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkNameNodeSafeMode(FSNamesystem.java:1400)

at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.delete(FSNamesystem.java:2850)

at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.delete(NameNodeRpcServer.java:1065)

at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.delete(ClientNamenodeProtocolServerSideTranslatorPB.java:641)

at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)

at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:503)

at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:989)

at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:871)

at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:817)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1893)

at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2606)

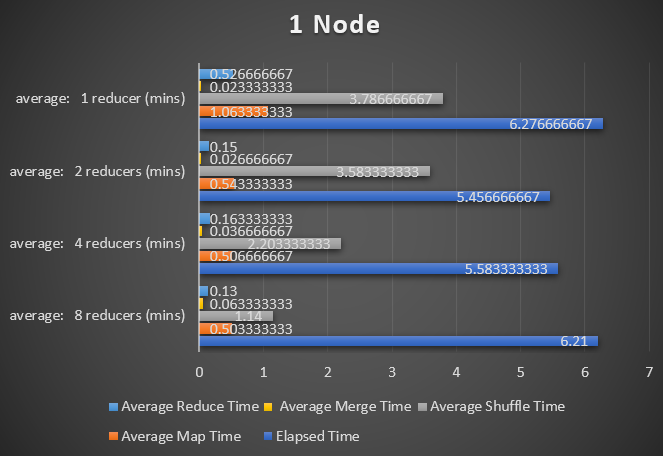
StreamingCommandFailed!

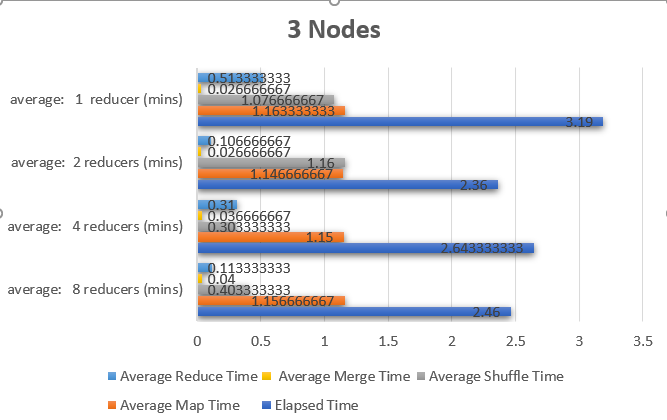
Και ο τρόπος που το αντιμετωπίσαμε είναι τρέχοντας την εξήςεντολής: SafeModeisOFF.

Ένα ακόμη πρόβλημα που είχαμε και δεν μπορέσαμε να αντιμετωπίσουμε είναι στον Reducer πως τα ενωμέναNCID δεν μπορούσαμε να τα εμφανίσουμε στην σειράαλλάεμφανίζει το πρώτοNCID στη σειρά με τα υπόλοιπαπεδία και κάθε επόμενοNCID το εμφανίζει σε επόμενη σειρά ξεχωριστά. Τέλος όταν πήγαμε να κατεβάσουμε τα αποτελέσματα από το dfs στην την περίπτωση του ενός κόμβου κατέβηκαν όλα κανονικά αλλά στην περίπτωση με τους παραπάνω κόμβους μας έβγαζε το εξής error για κάποια αρχεία: get: Zero blocklocations for /user/user/pyout22/part-00000. Name node is in safe mode. Όποτε σε κάποιους φακέλους δεν υπάρχει ο αντίστοιχος αριθμός αποτελεσμάτων του reducer που θα έπρεπε αλλά λιγότερος, αυτό γίνεται στους φακέλους που αφορούν τους τρεις κόμβους. Μπούμε να σας τους δείξουμε όμως κανονικά άμα μπούμε από τον Browser στο dfs.

3.2.4 Γραφικές παραστάσεις και πίνακες με τους χρόνους που μετρήσαμε







3.2.5. Σχολιασμό των χρόνων που παρατηρήσαμε

Σε ότι αφορά τους χρόνους παρατηρούμε αρχικά την εμφανή διαφορά του elapsetime στην περίπτωση τους ενός κόμβου σε σχέση με τους τρεις κόμβους όπου στην πρώτη περίπτωση θέλει περίπου 3 λεπτά παραπάνω. Στη συνέχεια στο elapse time παρατηρείται μια διαφορά στο χρόνο δηλαδή αύξησης στην περιπτώσεις που υλοποιείται με ένανreducer σε σχέση όταν υλοποιείται με δυο, τέσσερείς ή οχτώ ανεξάρτητος των αριθμών των κόμβων. Το ίδιο ισχύεις και στις περιπτώσεις Average Reduce Time και ο λόγος είναι καθώς τρέχουν ποιο πολλή reducer παράλληλα οπότε η δουλειά θα γίνει γρηγορότερα προφανώς.

Τώρα ότι αφορά τη περίπτωση του Average Shuffle Timeόσο αυξάνεται ο αριθμός των reducer μειώνονται η χρόνοι και υπάρχει εμφανής διαφορά στους χρόνους τις περίπτωσης με έναν κόμβο σε σχέση με τρεις καθώς η δουλειά γίνεται παράλληλα άρα γρηγορότερα. Τώρα ότι αφορά τους χρόνους για το AverageMapTime στην περίπτωση του ενός κόμβου παρατηρούμε ότι η δουλειά γίνεται ελαφρώς γρηγορότερα. Τέλος για το Average Merge Timeόσο αυξάνεται ο αριθμός των Reducer αυξάνεται και το Average Merge Time και παρατηρούμε πως στην περίπτωση του ενός κόμβου ο ρυθμός αύξησης του χρόνου είναι μεγαλύτερος το οποίο παρατηρείται κυρίως στους οχτώ Reducer λόγου ότι δεν παραλληλοποιουνται διεργασίες οπότε δεν έχει την ίδια υπολογιστική ισχύεις.