ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάθημα: Big Data/Ανάλυση Δεδομένων Μεγάλου Όγκου Εξάμηνο: Z'

Φοιτητές: Θεμιστοκλής Κουκουτζέλας-dai19022,Θεόφιλος Κιαπίδης-dai19079

1η Εργασία Μαθήματος

Εισαγωγή

Αντικείμενο της παρούσας εργασίας ήταν η κατασκευή ενός προγράμματος MapReduce, το οποίο θα ανέλυε ένα μεγάλο σύνολο δεδομένων, έχοντας ως αποτέλεσμα την λίστα των IP που προσπέλασαν ένα αρχείο σε διαφορετικές ημερομηνίες. Για τις ανάγκες της εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, χωρίς την χρήση επιπλέον modules, πάρα μόνο του sys.

1)Αλγόριθμοι Προγράμματος

Mapper:

For each line in document:

Key=IP,Filename

Value=DateofAccess

Emit(key,value)

Το πρόγραμμα του Mapper διαβάζει κάθε γραμμή του csv αρχείου με τα δεδομένα. Το filename στο πρόγραμμα παράγεται από την στήλη extension.Αν στη στήλη extension,υπάρχει μόνο η κατάληξη του αρχείου τότε προκύπτει από το συνδυασμό των στηλών accesion & extension του αρχείου. Το κάθε key είναι μονοδιάστατος πίνακας. Αφού δηλωθεί και στο value η ημερομηνία, το ζευγος key,value εκπέμπεται.

Reducer:

For each line:

If IP,Filename =previousIP,previousFilename and date!=previousDate:

If key not already printed:

emit(key)

Το πρόγραμμα του Reducer λαμβάνει ζευγάρια key,value, ομαδοποιημένα με βάση το key,ώστε να γίνει η σύγκριση των ημερομηνιών μεταξύ τους. Για ίδιο key, αν βρεθούν τουλάχιστον δύο διαφορετικές ημερομηνίες προσπέλασης τότε εκπέμπεται ως αποτέλεσμα το key. Το κάθε key πρέπει να εκτυπώνεται το πολύ μια φορά.

Ο mapper και ο reducer παρατίθενται στα αντίστοιχα αρχεία στο φάκελο Source Code που παραδίδεται μαζί με το έγγραφο.

2)Χρόνοι εκτέλεσης Προγράμματος

Οι αναλυτικοί πίνακες εκτέλεσης του προγράμματος βρίσκονται στο αρχείο MapReduceTimes.xlsx, που παραδίδεται μαζί με το παρών έγγραφο.

Οι παρατηρήσεις μας όσον αφορά τους χρόνους είναι οι εξής:

Α)Η μεγαλύτερη διαφορά στο Elapsed Time, υπάρχει στη χρήση 2 reducers από 1 και 2 node, προς όφελος των 2 nodes κατά περίπου 4 λεπτά, παρότι το average map time είναι αυξημένο από τα 20s περίπου στα 30s, ενώ το average reduce time παραμένει το ίδιο.

Β)Από τις παραπάνω εκτελέσεις προκύπτει πως ο συνδυασμός 2 nodes-2 reducers πέτυχε την μακράν καλύτερη απόδοση από όλες τις εκτελέσεις που πραγματοποιήθηκαν καθώς ο μέσος χρόνος εκτέλεσης ήταν 7 λεπτά και 3 δευτερόλεπτα ενώ σε όλες τις υπόλοιπες ήταν σχεδόν 3 λεπτά πάνω.

Γ) Οι 4 Reducers τόσο σε 1 όσο και σε 2 nodes, παρότι έχουν σταθερά τις μικρότερες μέσες τιμές σε Map time και Reduce time, έχουν συνολικά τους μεγαλύτερους χρόνους εκτέλεσης, από όλα τα υπόλοιπα σενάρια εκτέλεσης.

Δ) Τα Average Reduce Times, είναι αντιστρόφως ανάλογα του αριθμού των χρησιμοποιούμενων reducers τόσο 1 όσο και σε 2 nodes. Όσο αυξάνονται οι reducers, μειώνονται οι χρόνοι που απαιτούνται για reduce. Κατά κάποιο τρόπο μπορούμε να πούμε ότι τα 4 λεπτά του 1 reducer “σπάνε” σε κατά μέσο όρο 2 λεπτά στους 2 reducers και 1 λεπτό στους 4 reducers. Κάθε φορά που διπλασιάζονται δηλαδή οι reducers, υποδιπλασιάζονται οι χρόνοι που απαιτούνται για την αντίστοιχη διεργασία.

Ε) Υπάρχουν 2 “ομάδες” ως προς το Average Map Time. Η μια κυμαίνεται σε χρόνους κοντά στα 30 δευτερόλεπτα και η άλλη στα 20. Κοντά στα 30 κυμαίνονται τα σενάρια:

1node-1reducer,2 nodes-1reducer και 2 nodes-2 reducers.

Κοντά στα 20 κυμαίνονται τα σενάρια:

1 node-2 reducers,1 node-4 reducers,2 nodes-4 reducers.

Από αυτή την παρατήρηση προκύπτει επίσης ότι δεν υπάρχουν μεγάλες διακυμάνσεις στο Average Map Time,σε αντίθεση με το Average Reduce Time, οπου οι διαφορές μεταξύ 4 και 1 Reducer φτάνουν μέχρι και περίπου τα 3 λεπτά.

3)Οδηγίες εκτέλεσης Προγράμματος

Η εκτέλεση γίνεται με τη χρήση του Hadoop σε Linux Server.Αφού αποθηκευτούν ο mapper και ο reducer σε τοπικό φάκελο, ο χρήστης μεταβαίνει στον συγκεκριμένο φάκελο και εκτελεί την παρακάτω εντολή:

~/hadoop/bin/hadoop jar ~/hadoop/share/hadoop/tools/lib/hadoop-streaming-*.jar -files mapper.py,reducer.py -mapper mapper.py -reducer reducer.py -input CustomInputName -output CustomOutputName –numReduceTasks (αριθμός Reducers π.χ 4)

Η παραπάνω εντολή προϋποθέτει την αποθήκευση των log files σε συγκεκριμένο φάκελο μέσα στο hdfs του χρήστη. Επίσης, στην παραπάνω εντολή εφόσον θέλουμε να χρησιμοποιήσουμε μόνο έναν reducer, παραλείπουμε το argument numReduceTasks.

ΓΙΑ ΕΚΤΕΛΕΣΗ ΣΕ 1 ΚΟΜΒΟ ΠΡΕΠΕΙ ΝΑ ΕΚΤΕΛΕΣΤΕΙ Η ΠΑΡΑΚΑΤΩ ΕΝΤΟΛΗ ΣΤΟΥΣ ΚΟΜΒΟΥΣ ΠΟΥ ΘΑ ΜΕΙΝΟΥΝ ΑΝΕΝΕΡΓΟΙ:

~/hadoop/sbin/yarn-daemon.sh stop nodemanager

Τέλος, μαζί με τα παραπάνω, παραδίδονται και οι έξοδοι του προγράμματος για κάθε αριθμό reducers που χρησιμοποιήσαμε.