Τμήμα Εφαρμοσμένης Πληροφορικής
Ανάλυση Δεδομένων Μεγάλου Όγκου
Ακαδημαϊκό έτος 2022-2023

Πρώτη ομαδική εργασία
Ημερομηνία παράδοσης: Τετάρτη 7 Δεκεμβρίου 2022

Σας δίνεται ένα σύνολο δεδομένων, το οποίο προέρχεται από το Kaggle και αφορά συναλλαγές καταναλωτών. Κάθε γραμμή αφορά μία συναλλαγή και περιέχει κωδικούς προϊόντων που αγοράστηκαν. Το σύνολο δεδομένων μπορείτε να το κατεβάσετε ώς εξής:

wget snf-803830.vm.okeanos.grnet.gr/data.zip

Στόχος σας είναι να κατασκευάσετε πρόγραμμα MapReduce το οποίο θα εντοπίζει υποσύνολα προϊόντων τα οποία αγοράζονται συχνά, δηλαδή εμφανίζονται σε τουλάχιστον Ν συναλλαγές, όπου το Ν θα δίνεται ως παράμετρος κατά την εκτέλεση.

Αντιμετωπίστε το παραπάνω πρόβλημα με Hadoop & MapReduce σε κατανεμημένη διαμόρφωση, για α) έναν κόμβο, β) για δύο κόμβους. Βάλετε ως κατώφλια τις παρακάτω τιμές συναλλαγών: 5000, 10000, 50000

Η λύση σας μπορεί να αποτελείται από πλέον του ενός κύκλου MapReduce.

Παράδειγμα εκτέλεσης για υποσύνολα προϊόντων που αγοράζονται τουλάχιστον 2 φορές:

Αρχείο εισόδου:	Έξοδος:	
1 2 3 4	[1, 2, 3]	2
1 2 3	[1, 2]	2
2 2	[1, 3]	2
1	[1]	3
	[2, 3]	2
	[2]	3
	[3]	2

Για κάθε μία από τις παραπάνω διαμορφώσεις, πραγματοποιήστε 5 εκτελέσεις και αφού αφαιρέσετε τη μεγαλύτερη και μικρότερη τιμή, καταγράψτε το μέσο όρο των 3 υπολοίπων εκτελέσεων για τον κάθε έναν από τους παρακάτω χρόνους:
Elapsed Time, Average Map Time, Average Reduce Time.

Οδηγίες

1 **Η εργασία πραγματοποιείται αποκλειστικά σε ομάδες των 2 φοιτητών. Αν δεν μπορείτε να βρείτε συνεργάτη, χρησιμοποιήστε αυτόν τον σύνδεσμο.**
2 Γλώσσα προγραμματισμού είναι η Java.

3 Θα παραδώσετε:
   3.1 Τον κώδικά σας.
   3.2 Μία αναφορά που θα περιλαμβάνει:
      3.2.1 Τον αλγόριθμο που χρησιμοποιήσατε για να λύσετε το πρόβλημα.
      3.2.2 Τυχόν περιορισμούς και θεωρήσεις της λύσης σας.
      3.2.3 Επιπλέον πακέτα, αν χρησιμοποιήσατε.
      3.2.4 Τις παραμέτρους συστήματος, αν τροποποιήσατε.
      3.2.5 Γραφικές παραστάσεις και πίνακες με τους χρόνους που μετρήσατε.
      3.2.6 Σχολιασμό των χρόνων που παρατηρήσατε.
   3.3 Τα αρχεία εξόδου με το αποτέλεσμα του προγράμματός σας για κάθε αριθμό reducers που χρησιμοποιήσατε.
   3.4 Το .jar που θα δημιουργήσετε.
4 Η εργασία θα παραδοθεί **ηλεκτρονικά αποκλειστικά μέσω openeclass** σε μορφή **ενός αρχείου** zip.
5 Για να μετρήσετε τους χρόνους, θα χρησιμοποιήσετε τον job history server. Ξεκινάει με την εντολή:

```
~/hadoop/sbin/mr-jobhistory-daemon.sh start historyserver
```

Και τερματίζει αν αντίστοιχα όπου start βάλουμε stop.

Μπορείτε να τον βρείτε στην εξής διεύθυνση για τη μηχανή σας:

<MASTER_NAME>:19888

6 Για την εκτέλεση σε έναν κόμβο, μπορείτε να χρησιμοποιήσετε το παρακάτω script (χωρίς τις προαιρετικές παραμέτρους) για να τερματίσετε την κατάλληλη υπηρεσία:

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

Ενδεικτικές Αναφορές:

1 Κεφάλαια 2.1-2.3 από το βιβλίο «Εξόρυξη από Μεγάλα Σύνολα Δεδομένων», διαθέσιμο και online στο http://www.mmds.org/ (αγγλικά)
2 «Data-Intensive Text Processing with MapReduce». Διαθέσιμο online στο https://lintool.github.io/MapReduceAlgorithms/
3 Σημειώσεις του μαθήματος