ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

Τμήμα Πληροφορικής

ΕΠΛ 646 – Προχωρημένα Θέματα Βάσεων Δεδομένων

ΑΣΚΗΣΗ 5 – Υλοποίηση Σχεσιακών Τελεστών με Μεγάλα Δεδομένα στο MapReduce

Ημερομηνία Ανάθεσης: Παρασκευή, 05/04/2024
Ημερομηνία Παράδοσης: Παρασκευή, 19/04/2024 (14 ημέρες)

Διδάσκων: Δημήτρης Ζεϊναλιπούρ

(Να υποβληθεί η λύση ως ένα συμπιεσμένο αρχείο μέσω του **Moodle**)

Ι. Στόχος Άσκησης

Ο στόχος της άσκησης είναι η εξοικείωση με μια πλατφόρμα επεξεργασίας μεγάλων δεδομένων. Ειδικότερα, θα κληθείτε να χρησιμοποιήσετε το *Apache Hadoop* ή *Apache Spark* για να διατυπώσετε στο Προγραμματιστικό Μοντέλο *Map Reduce* την λύση για κάθε ένα από τα ερωτήματα που αναφέρονται πιο κάτω. Η επίλυση του θέματος θα πρέπει να γίνει στα πλαίσια του εικονικού μηχανήματος το οποίο σας έχει δοθεί στο εργαστήριο.

ΙΙ. Προεργασία και Εκκίνηση του Apache Hadoop

Α) Εκκινήστε το HDFS και το YARN του Hadoop τρέχοντας απλά `start-all.sh`, όπου θα παίρνατε κάτι της μορφής:

```
epl646:/home/epl646>start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as epl646 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 [epl646vm]
Starting resourcemanager
Starting nodemanagers
```

III. Εισαγωγή Δεδομένων στο HDFS

Τα απαραίτητα δεδομένα θα τα κατεβάσετε από τη σελίδα του μαθήματος με όνομα αρχείου as5-supplementary.zip. Μεταφέρετε τα αρχεία στο VM σας και στη συνέχεια να τα εισάγετε τα στο HDFS (Hadoop Distributed File System) το οποίο βρίσκεται ρυθμισμένο μέσα στο VM σας. Το συμπιεσμένο αρχείο as5-supplementary.zip περιέχει τα ακόλουθα CSV αρχεία:

```
r.500000.data.txt       (30.8 MB)        s.500000.data.txt       (30.8 MB)
r.50000.data.txt        (3.03 MB)        s.50000.data.txt        (3.03 MB)
r.5000.data.txt          (306 KB)        s.5000.data.txt          (306 KB)
r.5.data.txt               (302 B)       s.5.data.txt               (298 B)
```

Ο αριθμός στο όνομα του αρχείου υποδηλώνει πόσες πλειάδες (**n**) έχει κάθε αρχείο.
Το περιεχόμενο των αρχείων αποτελείται από εγγραφές της ακόλουθης μορφής:

```
tupleKey_n,columnA,columnB,columnC
…
tupleKey_n,columnA,columnB,columnC
```

Μπορείτε να ανεβάσετε ένα αρχείο στο HDFS με την εντολή:

```
hadoop fs -put <localsrc> ... <HDFS_dest_Path>
```

π.χ.,

```
hadoop fs -put ./r.5.data.txt /user/epl646/
hadoop fs -put ./s.5.data.txt /user/epl646/
```

IV. Ζητούμενα Άσκησης

Καλείστε να γράψετε 6 διακριτά προγράμματα που υλοποιούν τις ακόλουθες πράξεις:

1. $\sigma_{R.a>0.5 \wedge R.b<0.5}$ – Επιλογή γραμμών από το πίνακα (αρχείο) R, δείχνουμε ολόκληρη τη γραμμή ως αποτέλεσμα
2. $\prod_{R.a,R.b}$R – Προβολή στηλών από τον πίνακα (αρχείο) R με απαλοιφή των διπλότυπων
3. $R \cup S$ – Ένωση των γραμμών του πίνακα (αρχείο) R με τις γραμμές του πίνακα (αρχείο) S με απαλοιφή διπλότυπων (όχι union all δηλαδή)
4. $R - S$ – Αφαίρεση των γραμμών από τον πίνακα (αρχείο) R που υπάρχουν και στον πίνακα (αρχείο) S
5. $R \bowtie_{R.key=S.key} S$ – Join του πίνακα (αρχείο) R με τον πίνακα (αρχείο) S
6. $\mathfrak{F}_{Median(R.a)}$ – Εύρεση της γραμμής που περιέχει την ως τιμή στη στήλη A τον μέσο (median) των τιμών της στήλης A από τον πίνακα (αρχείο) R.

V. Δημιουργία MapReduce Εργασιών

Για να μπορέσετε να δημιουργήσετε μια εργασία MapReduce θα πρέπει να υλοποιήσετε μια κλάση που κληρονομεί την **Mapper** και υπερκαλύπτει-μεταβάλλει (override) την συνάρτηση **map**. Το αντίστοιχο θα πρέπει να γίνει με την **Reducer** και **reduce**. Η εκτέλεση του προγράμματος σας θα πρέπει να γίνει ως ακολούθως:

```
hadoop jar <filename>.jar <main-class>
```

VI. Παραδοτέα

Παραδώστε **όλα** τα πηγαία αρχεία σας μέσω του Moodle σε 1 συμπιεσμένο αρχείο (as5.zip):

Καλή Επιτυχία!