**ΠΡΟΧΩΡΗΜΕΝΑ ΘΕΜΑΤΑ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ**

Ακ. έτος 2013-2014, 9ο Εξάμηνο ΗΜ&ΜΥ

Εξαμηνιαία Εργασία

*Εισαγωγή στο MapReduce και στις βάσεις NoSQL*

# 1. Εισαγωγή

Σκοπός της συγκεκριμένης εργασίας είναι η εισαγωγή και η εξοικείωση με το προγραμματιστικό μοντέλο MapReduce και τις βάσεις NoSQL. Συγκεκριμένα, το MapReduce θα μελετηθεί με τη βοήθεια του framework Hadoop, ενώ οι NoSQL βάσεις μέσω τις HBase. Και τα 2 αυτά εργαλεία τρέχουν σε ένα κατανεμημένο υπολογιστικό σύστημα το οποίο μπορεί να αυξομειώνεται ανάλογα με τις ανάγκες τις εφαρμογής. Για τις ανάγκες τις άσκησης χρησιμοποιήθηκα 2 virtual machines από τον ~okeanos τα οποία έτρεχαν Debian Wheezy.

Σαν αντικείμενο μελέτης είχαμε στη διάθεσή μας 2 διαφορετικά datasets, ένα από την America On Line με 20 εκατομμύρια ερωτήματα 650.000 χρηστών και ένα από την Wikipedia με τίτλους άρθρων.

Τα datasets αυτά ανέβηκαν αρχικά στο HDFS (κατανεμημένο σύστημα αρχείων που χρησιμοποιεί το Hadoop) ώστε να μπορεί να γίνει η κατανεμημένη επεξεργασία.

# 2. Εισαγωγή στο MapReduce

**2.1 Yπολογισμός αριθμού αναζητήσεων ανά ημέρα**

Στο ερώτημα αυτό πρέπει να εξάγουμε ένα διάγραμμα από το dataset της AOL το οποίο θα δείχνει το πλήθος των αναζητήσεων που έγιναν ανά μέρα.

Από το αρχείο των αναζητήσεων οι mappers για κάθε γραμμή θα κάνουν emit ένα keyvalue της μορφής <data, 1>. Στη συνέχεια οι reducers θα μαζεύουν τα key-values ανά μέρα. Συνοπτικά:

MAP(key, value) // value is the file line

date = getDate(line)

emit(date,1)

REDUCE(key, value\_list) //

sum = 0

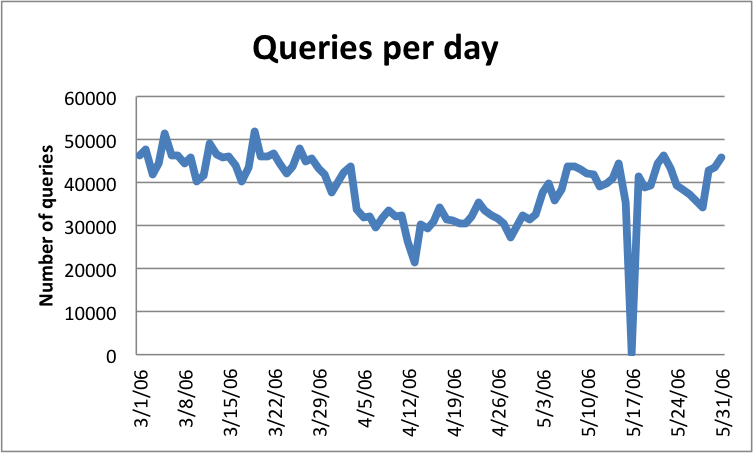
for v in value\_list

sum += v

emit(key, sum)

Μπορούμε να χρησιμοποιήσουμε μόνο έναν reduce ή και περισσότερους. Στην περίπτωση των περισσότερων θα πρέπει να μοιράσουμε τα key-values ανάλογα διατηρώντας τη σειρά. Επιλέξαμε 2 reducers και γνωρίζοντας το εύρος των ημερομηνιών κατασκευάσαμε έναν custom partitioner που μοιράζει τα key-values στα 2.

Παρακάτω βλέπουμε το διάγραμμα του αποτελέσματος.



*Διαγραμμα 1: Ερωτήματα ΑΟL ανά μέρα*

## 2.2 Υπολογισμός ποσοστού «επιτυχών» και «ανεπιτυχών» αναζητήσεων

Στο ερώτημα αυτό ζητείται να υπολογιστούν τα ποσοστά επιτυχών και ανεπιτυχών αναζητήσεων. Επιτυχής αναζήτηση θεωρείται αυτή στην οποία ο χρήστης επέλεξε να μεταβεί σε κάποιο URL, δηλαδή εκείνα τα entries του dataset τα οποία έχουν 4ο και 5ο πεδίο.

Οι mappers διαβάζουν γραμμή-γραμμή το dataset και αν η γραμμή έχει επιτυχή αναζήτηση κάνουν emit keyvalue της μορφής <1,success>, διαφορετικά <1, failure>. Όπως φαίνεται θα χρησιμοποιήσουμε έναν reducer στον οποίο θα καταλήξουν όλα τα αποτελέσματα αφού θέλουμε ποσοστά και χρειαζόμαστε το συνολικό πλήθος αναζητήσεων. O reducer θα μετρήσει τις συνολικές αναζητήσεις και θα υπολογίσει τα ποσοστά επιτυχών και ανεπιτυχών αναζητήσεων. Συνοπτικά:

MAP(key, value):

if check\_success(value):

emit(1,success)

else:

emit(1, failure)

REDUCE(key, value\_list):

success\_count = 0

failure\_count = 0

count = 0

for v in value\_list:

if v==success:

success\_count++

else:

failure\_count++

count++

emit(“Success”, success\_count/count)

emit(“Failure”, failure\_count/count)

Τα αποτελέσματα είναι

Successful searches (%): 53.13

Unsuccessful searches (%): 46.87

## 2.3 Ιστοσελίδων που επισκέφτηκαν πάνω από 10 διαφορετικοί χρήστες

Στο ερώτημα αυτό με βάση το AOL dataset θα πρέπει να να βρούμε τις ιστοσελίδες με πάνω από 10 unique visits και για κάθε μια από αυτές να τυπώσουμε τον αριθμό των επισκέψεων.