ΚΑΤΑΝΕΜΗΜΕΝΑ ΔΙΑΔΙΚΤΥΑΚΑ ΣΥΣΤΗΜΑΤΑ

6ο Εξάμηνο - Ακαδημαϊκό έτος 2015-16

ΕΚΦΩΝΗΣΗ 2ης ΕΡΓΑΣΙΑΣ

ΥΠΕΥΘΥΝΟΙ ΕΡΓΑΣΙΑΣ

Καλτιριμίδου Έφη (kaltirim@csd.auth.gr)
Ιωάννης Δεμάτης (icdematis@csd.auth.gr)

ΕΙΣΑΓΩΓΗ

Η ραγδαία αύξηση των διαθέσιμων δεδομένων στον Ιστό θέτει νέες προκλήσεις και δυσκολίες στην επεξεργασία τόσο μεγάλου όγκου πληροροφορίας από έναν υπολογιστικό πόρο. Λύσεις βασιζόμενες σε έννοιες των κατανεμημένων και παράλληλων συστημάτων έχουν προταθεί ώστε να επιτευχθεί καλύτερη απόδοση και να γίνει εφικτή η ανάλυση μεγάλου όγκου δεδομένων. Μια τεχνολογία που τα τελευταία χρόνια κερδίζει ολοένα και περισσότερο δημοτικότητα είναι το MapReduce της Google [1] αλλά και το framework Apache Hadoop [2] που βασίζεται στη φιλοσοφία του MapReduce. Το Hadoop παρέχει τρόπους δημιουργίας ομάδων από CPUs (clusters) και τεχνικές επικοινωνίας αυτών των ομάδων. Ουσιαστικά, αποτελείται από ένα σύνολο από ενότητες λογισμικού (προγράμματα) που συνεργάζονται μεταξύ τους για το διαμερισμό εργασιών και μεταφοράς δεδομένων.

ΠΕΡΙΓΡΑΦΗ - ΣΚΟΠΟΣ

Θεωρείστε ότι είστε ένας απο τους προγραμματιστές που δουλευούν στο backend του κοινωνικού δικτύου Facebook. Το κομμάτι που σας ζητά το Facebook να υλοποιήσετε είναι η δυνατότητα του να κάνει προτάσεις σε κάθε χρήστη για νέους φίλους που πιθανόν γνωρίζει στην πραγματική του ζωή. Το Facebook διαθέτει μια λίστα με τους φίλους του κάθε χρήστη (σημειώστε ότι η σχέση “φίλος” είναι μια αμφίδρομη σχέση στο Facebook : “Εαν ο Α είναι φίλος του Β , τότε κι ο Β είναι φίλος του Α”).
Χρησιμοποιώντας αυτες τις λίστες μπορεί να βρει τον αριθμό των κοινών φίλων μεταξύ δύο χρηστών, και πιθανές συσχετίσεις μεταξύ των μη κοινών τους φίλων, για παράδειγμα δύο φίλοι που έχουν τουλάχιστον 10 κοινούς φίλους είναι πολύ πιθανό να γνωρίζουν και τους μη κοινούς τους φίλους. Με αυτήν την πληροφορία στη συνέχεια έχει την δυνατότητα να κάνει προβλέψεις και προτάσεις για νέους φίλους. Λόγω του όγκου των δεδομενων που θα χρησιμοποιηθούν έχει επιλεχθεί για την υλοποίηση της εργασίας το framework Hadoop MapReduce για την εκτέλεση των υπολογισμών.

ΑΝΑΛΥΤΙΚΑ - ΒΗΜΑΤΑ ΥΛΟΠΟΙΗΣΗΣ ΕΡΓΑΣΙΑΣ

1. Δεδομένα

   Οι λίστες που αποθηκεύονται για κάθε χρήστη είναι της μορφής:

   Alice → (John, Leo, Mary, Paul, Rebecca, Steve, Veronica)
   John → (Alice, Kevin, Leo, Mary, Rebecca, Steve, Veronica)
   Leo → (Alice, John)
   ……

2. Έυρεση κοινών φίλων

   Απο τις λίστες αυτές πρέπει να βρείτε τους κοινούς φίλους κάθε χρήστη με όλους τους φίλους του. Για παράδειγμα οι κοινοί φίλοι των ζευγαριών της Alice:

   Alice - John : (Leo, Mary, Rebecca, Steve, Veronica)
   Alice - Leo : (John)
   ……

3. Όριο και πιθανοί φίλοι

   Εαν οι κοινοί φίλοι δύο χρηστών ξεπερνούν κάποιο όριο που έχουμε θέσει, για παράδειγμα με 5 κοινούς φίλους, υπάρχει μεγάλη πιθανότητα κάποιος απο τους μη κοινούς φίλους τους να γνωρίζει και τους δύο χρήστες. Βρίσκουμε τους μη κοινούς τους φίλους και τους εξάγουμε ως αποτέλεσμα για να γίνει η πρόταση προς τον χρήστη. Για παράδειγμα, η Alice με τον John έχουν τουλάχιστον 5 κοινούς φίλους, οπότε ως αποτέλεσμα θα εξάγουμε οτι η Alice μπορεί να γνωρίζει και τον Kevin, και αντίστοιχα ο John μπορεί να γνωρίζει τον Paul, αλλά δεν είναι στους κοινούς τους φίλους.

ΥΛΟΠΟΙΗΣΗ

1. Θα πρέπει να υπολοιήσετε τα παραπάνω βήματα σε ένα πρόγραμμα Map Reduce σε java το οποίο να μπορεί να εκτελεστεί σε Apache Hadoop framework. Κατά την εκτέλεση του Map θα πρέπει να χωρίσετε τις λίστες στα επιμέρους ζευγάρια και να τα στείλετε στην Reduce. Κατα την εκτέλεση της Reduce θα πρέπει να έχετε μαζί τους φίλους και των δύο χρηστών και να εξάγετε τους πιθανούς γνωστούς για τον κάθε έναν ξεχωριστά.
2. Για την εκτέλεση του προγράμματος θα πρέπει να εγκαταστήσετε το Apache Hadoop σε pseudo distributed mode.
3. Τα δεδομένα εισόδου σας θα είναι αρχεία txt αποθηκευμένα στον HDFS του Hadoop που θα εκτελέσετε το πρόγραμμα σας, και θα είναι της μορφής:

   ```
   Alice   John, Leo, Mary, Paul, Rebecca, Steve, Veronica
   John   Alice, Kevin, Leo, Mary,  Rebecca, Steve, Veronica
   ```

   Όπου η πρώτη λέξη είναι ο χρήστης και ακολουθούν τα ονόματα των φίλων του χωρισμένα μεταξύ τους με ένα κόμμα.
4. Τα δεδομένα εξόδου σας θα πρέπει να είναι αρχείο(α) της μορφής :

   ```
   Alice  Kevin
   John   Paul
   ```

ΠΑΡΑΔΟΤΕΑ - ΔΙΕΥΚΡΙΝΗΣΕΙΣ

1. Η εργασία θα υλοποιηθεί από ομάδες φοιτητών με κάθε ομάδα να αποτελείται το πολύ από 4 άτομα.
2. Κάθε ομάδα θα παραδώσει ένα αρχείο zip με όνομα AEM1_AEM2_AEM3_AEM4.zip
3. Μέσα στο αρχείο .zip θα πρέπει να υπάρχουν:
    a. Ο κώδικας της εργασίας.
    b. Το εκτελέσιμο jar της εργασίας.
    c. Μια σύντομη αναφορά σε pdf με τα ονοματεπώνυμα της ομάδας και τον τρόπο υλοποίησης.
    d. Ο φάκελος logs που βρίσκεται στο path $HADOOP_HOME/logs αφού πρώτα έχει καθαρισθεί πριν την τελική εκτέλεση του προγράμματος.

Bonus

Στο bonus κομμάτι της εργασίας, θα πρέπει να εγκαταστήσετε έναν κατανεμημένο Hadoop cluster σε τουλάχιστον δύο μηχανήματα, όπου ο Master node θα είναι διαφορετικό μηχάνημα από τους slaves. Θα πρέπει να εκτελέσετε την εργασία και στο κατανεμημένο περιβάλλον.
Για την αξιολόγηση του bonus θα πρέπει να στείλετε επίσης:

1. Όλα τα configuration files που βρίσκονται στο path $HADOOP_HOME/etc/hadoop .
2. Τα καινούργια logs μετά την εκτέλεση στο κατανεμημένο σύστημα.
3. Να συμπεριλάβετε στην αναφορά σας τον χρόνο εκτέλεσης σε ψεύδο-κατανεμημένο και κατανεμημένο περιβάλλον και να βγάλετε συμπεράσματα σχετικά με αυτό. Επίσης να αναφέρετε τον αριθμό, το Linux distribution, και τα τεχνικά χαρακτηριστικά των μηχανημάτων του κατανεμημένου cluster.

Το bonus αντιστοιχεί σε 1 επιπλέον μονάδα στον τελικό βαθμό.

Τα παραπάνω να αποσταλούν με email στους υπεύθυνους της εργασίας, καθώς και οποιαδήποτε απορία προκύψει κατα την υλοποίηση να γίνεται αρχικά με αποστολή email το οποίο να έχει θέμα “ ΚΔΣ 2016 ”.

Ημερομηνία παράδοσης είναι η Δευτέρα 30/5/2016.

ΑΝΑΦΟΡΕΣ

[1]Jeffrey Dean and Sanjay Ghemawat - MapReduce: Simplified Data Processing on Large Clusters
[2] Apache Hadoop