Απαλλακτική Εργασία στα Κατανεμημένα Συστήματα

Θέμα 1ο: Το μοντέλο MapReduce [3 μονάδες]

Περιγράψτε το προγραμματιστικό μοντέλο MapReduce [1]. Ενδεικτικά, αναφέρεται τον τρόπο λειτουργίας του, τις εφαρμογές του, τα πλεονεκτήματα και τα μειονεκτήματα του καθώς επίσης και τα δημοφιλή πλαίσια (frameworks) λογισμικού που έχουν αναπτυχθεί για την υποστήριξη του.

Θέμα 2ο: Υλοποίηση Εφαρμογής Κατανεμημένης Επεξεργασίας [7 μονάδες]

Θεωρείστε ότι είστε ο διαχειριστής ενός δικτύου σταθμών μέτρησης της ηλεκτρομαγνητικής ακτινοβολίας στο περιβάλλον. Οι σταθμοί μέτρησης είναι τοποθετημένοι σε συγκεκριμένες θέσεις και συλλέγουν έναν πολύ μεγάλο αριθμό μετρήσεων. Οι μετρήσεις από όλους τους σταθμούς καταγράφονται σε ένα αρχείο του οποίου το μέγεθος μπορεί να ανέλθει σε εκατοντάδες GB. Για κάθε μέτρηση αποθηκεύεται η θέση (pos_id) και η ημερομηνία και ώρα (date_time) που έγινε η δειγματοληψία καθώς επίσης και η μέση τιμή (avg_value) της ακτινοβολίας που καταγράφτηκε. Ως διαχειριστής του συστήματος επιθυμείτε να αναπτύξετε μια εφαρμογή που θα διαβάζει το αρχείο των μετρήσεων και θα υπολογίζει το πλήθος των μετρήσεων που έχουν καταγραφεί για κάθε θέση μέτρησης (δηλαδή, για κάθε διαφορετικό pos_id). Για την επιτάχυνση της επεξεργασίας του πολύ μεγάλου αρχείου των μετρήσεων έχετε στη διάθεση σας μια ομάδα από *n* κόμβους-εργάτες (worker nodes). Εκμεταλλευόμενοι την παρουσία των κόμβων-εργατών, να αναπτύξετε τη ζητούμενη εφαρμογή ακολουθώντας την παρακάτω κατανεμημένη λογική επεξεργασίας:

- Η εφαρμογή θα δέχεται ως είσοδο το αρχείο των μετρήσεων και θα το χωρίζει σε τμήματα που θα ισομοιράζονται (κατά το δυνατόν) στους κόμβους εργάτες.
- Κάθε κόμβος-εργάτης θα επεξεργάζεται το τμήμα που λαμβάνει και θα παράγει τα μερικά αποτελέσματα που αφορούν αυτό το τμήμα, δηλαδή θα επιστρέφει για κάθε θέση μέτρησης που υπάρχει στο τμήμα το πλήθος των μετρήσεων που υπάρχουν για αυτήν μέσα στο τμήμα.
- Η εφαρμογή θα συλλέγει τα μερικά αποτελέσματα από τους κόμβους εργάτες και θα παράγει τα τελικά, ολικά, αποτελέσματα.

Για την επικοινωνία μεταξύ του σταθμού εργασίας του χρήστη της εφαρμογής και των κόμβων-εργατών, να χρησιμοποιήσετε το πρωτόκολλο https. Η υλοποίηση της εφαρμογής να γίνει σε μια από τις ακόλουθες γλώσσες προγραμματισμού: JAVA, Python 3 ή C++. Οι διευθύνσεις IP των κόμβων-εργατών θα πρέπει να ορίζονται σε ένα αρχείο διαμόρφωσης (configuration file) της εφαρμογής.

Για την υποστήριξη της ανάπτυξης και της εκτέλεσης της εφαρμογής, σας παρέχεται ένα ενδεικτικό αρχείο μετρήσεων σε μορφές csv και json που είναι διαθέσιμες μέσω των συνδέσμων https://pithos.okeanos.grnet.gr/public/WzsbBccghfRz6Hr3PgUsH και https://pithos.okeanos.grnet.gr/public/hlEBKTRWVQLUS44Urj7IL5 αντίστοιχα.

Σημείωση: Εναλλακτικά της παραπάνω υλοποίησης μπορείτε να κατασκευάσετε την εφαρμογή χρησιμοποιώντας το Hadoop [2] framework.

Εκπόνηση και Παραδοτέα Εργασίας

Η εκπόνηση της εργασίας μπορεί να γίνει σε ομάδες των τεσσάρων ατόμων το πολύ. Το παραδοτέο θα πρέπει να είναι ένα συμπιεσμένο αρχείο zip αποτελούμενο από (α) την αναφορά που θα περιλαμβάνει τις απαντήσεις στα θέματα 1 και 2 και (β) τον κώδικα του θέματος 2. Επισημαίνεται ότι, για το θέμα 2 θα πρέπει να γίνει επεξήγηση του κώδικα στην αναφορά της εργασίας.

Τρόπος και προθεσμία παράδοσης

Η παράδοση των εργασιών θα γίνεται ηλεκτρονικά μέσω της ασύγχρονης πλατφόρμας

τηλεκπαίδευσης (https://gunet2.cs.unipi.gr/courses/TMD131) έως και την **ημέρα εξέτασης του μαθήματος**. Η αποστολή της εργασίας να γίνει μόνο από ένα μέλος κάθε ομάδας.

Απορίες, σχόλια και παρατηρήσεις

Για ζητήματα που σχετίζονται με την παρούσα εργασία, μπορείτε να επικοινωνήσετε με τον Δρ. Απόστολο Καραλή (akaralis@unipi.gr).

Ενδεικτική Βιβλιογραφία

[1] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107–113. DOI: https://doi.org/10.1145/1327452.1327492

[2] Apache Software Foundation. *Hadoop*. Διαθέσιμο στον ιστότοπο https://hadoop.apache.org (τελευταία πρόσβαση 2/12/2022)