**Κατανεμημένα Συστήματα – Εργασία Χειμερινού Εξαμήνου**

**Παπαγεωργίου Ιωάννης**

**Λάκκα Αλεξάνδρα**

**Μέρος Α’(Batch):**

Στην περίπτωση των batch queries, μόλις τρέξει το πρόγραμμα, ζητάτε από το χρήστη να δώσει ένα αριθμό(1-7) ως input, ώστε να εκτελεστεί το αντίστοιχο query. Με την εκτέλεση του query, εμφανίζονται οι πρώτες 20 γραμμές από το query και τα αποτελέσματα γράφονται στο HDFS, στο μονοπάτι που έχουμε ορίσει. Πρέπει να σημειωθεί, ότι για το query 6, ο χρήστης θα πρέπει να δώσει ως input ένα pickup\_latitude, pickup\_latitude και ένα dropoff\_latitude, dropoff\_latitude ζευγάρι, καθώς και μια συγκεκριμένη ώρα(ενδεικτικές είσοδοι υπάρχουν σαν σχόλια στον κώδικα).

**Μέρος Β’(Streaming):**

Στο κομμάτι του streaming, αυτό που κάναμε είναι να διαβάζουμε ένα directory στο HDFS, στο οποίο βρίσκεται αποθηκευμένο το csv αρχείο με τα δεδομένα, σπασμένο σε μικρότερα αρχεία(mini-batches). Αυτό που κάνουμε είναι κάθε φορά να διαβάζουμε ένα μονοπάτι του mini-batch αρχείου και να εκτελούμε το query στο dataframe που φτιάχνουμε από αυτό. Συνεχίζουμε, προσθέτοντας στο file path από το οποίο χτίζουμε το dataframe, και τα υπόλοιπα μονοπάτια. Στο τέλος, θα προκύψει ένα dataframe που θα περιέχει όλες τις εγγραφές και, επομένως, το τελικό query θα εκτελεστεί σε όλες. Τα αποτελέσματα αποθηκεύονται και πάλι στο HDFS.

Μια άλλη προσέγγιση είναι να διαβάσουμε από ένα directory στο HDFS όλα τα αρχεία με τα δεδομένα, να τα συνενώσουμε σε ένα αρχείο και να το διαβάζουμε γραμμή-γραμμή(mini-batches), εκτελώντας, πάλι, το query για τη νέα γραμμή, καθώς και για τις προηγούμενες εγγραφές, που έχουμε φροντίσει να κρατήσουμε αποθηκευμένες.

**VM Connection:**

Για τη σύνδεση των VM με Docker Swarm, πρέπει πρώτα και στα 2 VM να γίνει ρύθμιση δικτύου με Bridge. Μετά γίνεται docker swarm init σε ένα από τα 2 μηχανήματα, το οποίο θα είναι και ο manager και θα κάνει deploy και το docker-compose.yml, και το άλλο θα κάνει join με το token που αντιστοιχεί στον manager του swarm. Στον κώδικα, θα πρέπει η IP του Spark master να είναι αυτή του μηχανήματος στο οποίο θα τρέχει το service για το master.