Αναφορά Ατομικής Εργασίας

Στην παρούσα εργασία μας ζητήθηκε ο χειρισμός δεδομένων ενός αρχείου csv που περιλαμβάνει δεδομένα συναλλαγών λιανικής, με την χρήση του Apache Spark. Η εργασία υλοποιήθηκε με την χρήση της εικονικής μηχανής του μαθήματος όπου είναι εγκατεστημένο το apache spark, με εντολές γραμμένες στο shell του spark. Μετά από δοκιμές των εντολών μία προς μία συντάχθηκε ένα script σε python που εκτελεί τις εντολές και παράγει τα ζητούμενα.

Σχετικά με το Ζητούμενο 1 και το Ζητούμενο 2, δημιουργήθηκε ένα python script με όνομα **task1-2.py** . Το script πέρα από τις κατάλληλες εντολές για την απάντηση των ζητούμενων, συμπεριλαμβάνει τα απαραίτητα imports συναρτήσεων από τις βιβλιοθήκες του pyspark, και δημιουργεί ένα Spark Session με προκειμένου να τρέξει τις εντολές που ακολουθούν. Για να εκτελεστεί αρκεί να βάλουμε το script στην διαδρομή /home/bigdata , να ανοίξουμε ένα terminal σε αυτό το directory και να τρέξουμε την εντολή

```
bigdata@bigdata-virtualbox:~$ spark-submit task1-2.py
```

Τρέχοντας το script, γίνονται τα imports, δημιουργείται ένα Spark Session και απαντώνται τα ζητούμενα ένα προς ένα. Με το τέλος εκτέλεσης ενός ζητούμενου εμφανίζεται ένα μήνυμα ολοκλήρωσης στο terminal.

Κατά το ζητούμενο 1.5 δημιουργείται ένα directory με όνομα mai25067 όπου εντός δημιουργείται φάκελος **transactions_by_quantity** που περιέχει το εξαγόμενο αρχείο csv με το πλήθος των συναλλαγών ανά πλήθος αντικειμένων που αγοράστηκαν. Κατά το ζητούμενο 1.6 εντός του directory mai25067, δημιουργείται φάκελος με όνομα **total_price_by_transactions** που περιέχει το εξαγόμενο αρχείο csv με το συνολικό ποσό που δαπανήθηκε ανά συναλλαγή. Κατά το ζητούμενο 2 εντός του directory mai25067, δημιουργείται φάκελος με όνομα **statistics_table** που περιέχει το εξαγόμενο αρχείο csv με τον ζητούμενο Πίνακα Στατιστικών. Στα 2 πρώτα ζητούμενα, δεν αντιμετωπίστηκαν ιδιαίτερες δυσκολίες, καθώς η ύλη του μαθήματος καθώς και το documentation του Apache Spark, βοήθησαν στην επίλυσή τους.

Σχετικά με το Ζητούμενο 3, δημιουργήθηκε ένα python script με όνομα **task3.py** . Ως csv για το stream είναι το αρχείο customer_shopping_data_new.csv , στο οποίο προστέθηκαν ορισμένες εγγραφές στο τέλος. Το ζητούμενο **δεν** κατάφερα να το απαντήσω. Προσπάθησα με διάφορα resources στο internet καθώς και το documentation του Apache Spark, ωστόσο όποια δοκιμή και να έκανα δεν κατάφερνα να το βγάλω σωστά. Δυσκολεύτηκα να καταλάβω πως να εισάγω και να χρησιμοποιήσω σωστά την έννοια του χρόνου (timestamp) έτσι ώστε το stream να ανανεώνεται με βάση μια χρονική διάρκεια, και το απαραίτητο watermark που χρειάζεται για να δουλέψει.