Τμήμα Εφαρμοσμένης Πληροφορικής
ΠΜΣ στην Εφαρμοσμένη Πληροφορική
Ανάπτυξη Εφαρμογών & Μεγάλα Δεδομένα
Ακαδημαϊκό έτος 2024-2025
Θ. Κασκάλης – Α. Καρακασίδης

Α. Καρακασίδης: Ατομική εργασία μαθήματος (2024-2025)

Ημερομηνία παράδοσης: Τετάρτη, 12 Φεβρουαρίου 2025

Σας δίδεται ένα αρχείο δεδομένων σε μορφή csv το οποίο περιλαμβάνει δεδομένα συναλλαγών λιανικής. Καλείστε να επεξεργαστείτε τα δεδομένα αυτά χρησιμοποιώντας το Apache Spark και συγκεκριμένα το pyspark, Η εργασία αποτελείται από 3 ζητούμενα:

Ζητούμενο 1ο (Μονάδες 3): Βασική Επεξεργασία Δεδομένων

1. Για τις στήλες payment_method και shopping_mall αντικαταστήστε το κενό με κάτω παύλα (underscore)
2. Μετατρέψτε τα ονόματα των shopping_mall σε κεφαλαία.
3. Υπολογίστε και αποθηκεύστε σε μεταβλητή το πλήθος των συναλλαγών που βρίσκονται στο αρχείο που διαβάσατε.
4. Μετατρέψτε από TL που είναι το νόμισμα στις συναλλαγές σε ευρώ (ισοτιμία 1TL = 0.1E) το πεδίο price
5. Υπολογίστε και αποθηκεύστε σε αρχείο csv το πλήθος των συναλλαγών ανά πλήθος αντικειμένων που αγοράστηκαν.
6. Υπολογίστε και αποθηκεύστε σε αρχείο csv σε κάθε συναλλαγή το συνολικό ποσό που δαπανήθηκε υπολογίζοντας σε νέα στήλη total το γινόμενο price*quantity

Ζητούμενο 2ο (Μονάδες 6): Δημιουργία Πίνακα Στατιστικών

Δημιουργήστε πίνακα στατιστικών για όλες τις συναλλαγές που εμφανίζονται στο αρχείο csv, ο οποίος θα αποθηκευτεί επίσης σε αρχείο csv, Υπολογίστε κατά εμπορικό κέντρο:

1. Το συνολικό ποσό που έχει δαπανηθεί
2. Το πλήθος των τεμαχίων προϊόντων που έχουν πωληθεί
3. Το πλήθος των συναλλαγών που έχουν πραγματοποιηθεί

Ζητούμενο 3ο (Μονάδα 1): Χρήση του Apache Spark Structured Streaming API

Για τις συναλλαγές που έχετε τροποποιήσει/προσθέσει, υπολογίστε και πάλι τα στατιστικά του ζητουμένου 2. Για τα δεδομένα που θα αποθηκεύσετε από την εφαρμογή της εργασίας του κ. Κασκάλη, χρησιμοποιώντας Apache Spark Structured Streaming, να τοποθετήσετε το csv αρχείο που προκύπτει από τη web εφαρμογή σε κατάλληλο φάκελο, να ενημερώσετε με το συγκεκριμένο μηχανισμό τον πίνακα στατιστικών του ζητουμένου 2 και να αποθηκεύσετε τον νέο πίνακα στατιστικών σε csv αρχείο.

Οδηγίες

Α. Το csv αρχείο που θα χρησιμοποιήσετε ως είσοδο έχει τα παρακάτω πεδία.

invoice_no: Αριθμός Τιμολογίου – μοναδικό για κάθε συναλλαγή
customer_id: Αναγνωριστικό Πελάτη – μοναδικό για κάθε συναλλαγή
gender: Φύλλο του πελάτη.
age: Ηλικία του πελάτη
category: Κατηγορία του προϊόντος που αγοράστηκε
quantity: Η ποσότητα του προϊόντος που αγοράστηκε
price: Τιμή μονάδος σε Τουρκικές Λίρες (TL).
payment_method: Τρόπος πληρωμής (cash, credit card, debit card).
invoice_date: Ημερομηνία συναλλαγής
shopping_mall: Πολυκατάστημα στο οποίο έγινε η συναλλαγή

Β. Το csv αρχείο που θα προκύψει από την εφαρμογή του κ. Κασκάλη, θα έχει τα ίδια πεδία με τα παραπάνω

Διευκρινίσεις

1. **Η εργασία είναι ατομική**
2. **Για την υλοποίησή της θα πρέπει να χρησιμοποιήσετε Python και κάποιο από τα APIs που προσφέρει το Apache Spark (RDD, Spark Dataframe, Spark SQL).**
3. **Δεν θα πρέπει να χρησιμοποιήσετε εξωτερικές βιβλιοθήκες (π.χ. Pandas). Σε αυτή την περίπτωση δεν θα βαθμολογηθείτε.**
4. Θα παραδώσετε:
   4.1. Τα αρχεία κώδικά σας
   4.2. Μία αναφορά, έως 3 σελίδες, με οδηγίες εκτέλεσης της εφαρμογής σας, τα προβλήματα που αντιμετωπίσετε και τους τρόπους επίλυσής τους.
5. Η εργασία θα πρέπει να παραδοθεί **ηλεκτρονικά** μέσω openeclass σε μορφή **ενός αρχείου** zip.