Αποκεντρωμένα Συστήματα Διαχείρισης Μεγάλου Όγκου Δεδομένων

Τμήμα Μηχανικών Η/Υ & Πληροφορικής - Πανεπιστήμιο Πατρών

ΔΠΜΣ ΥΔΑ

Διδάσκοντας: Σπυρίδων Σιούτας

Εργαστηριακή Άσκηση 2023

Στην ιστοσελίδα https://www.stats.govt.nz/large-datasets/csv-files-for-download/ υπάρχουν δεδομένα για ανάλυση/επεξεργασία σε αρχεία csv που αφορούν τους ακόλουθους τομείς: Business, Census, Economy, Effects of COVID-19 on trade, Environment, Government finance, Health, Industries, Labour market, Population, Society.

Στη συγκεκριμένη εργαστηριακή άσκηση θα ασχοληθούμε με δεδομένα του Effects of COVID-19 on trade, και συγκεκριμένα το "**effects-of-covid-19-on-trade-at-15-december-2021-provisional.csv**", που περιέχει συνολικά 111.438 εγγραφές, με την ακόλουθη δομή:

Direction | Year | Date | Weekday | Country | Commodity | Transport_Mode | Measure | Value | Cumulative

Το συγκεκριμένο αρχείο περιέχει τα ακόλουθα πεδία:

- Direction (imports, exports ,reimports)
- Year (2015,…., 2021)
- Date (01/01/2015, 01/02/2015,……, 15/12/2021),
- Weekday (Monday,…., Sunday)
- Country (All, China, European Union, Asia, Australia, USA,…..)
- Commodity ("All", "Milk powder, butter, and cheese", "Fish, crustaceans, and molluscs", "Non-food manufactured goods", "Electrical machinery and equip",…….)
- Transport_Mode {All, sea, air, …)
- Measure ($, Tonnes)
- Value (long integer)
- Cumulative (long integer)

Αρχικά σας ζητείται να υλοποιήσετε πρόγραμμα στο περιβάλλον του Apache Spark, που θα ενσωματώνει τα παραπάνω δεδομένα και στη συνέχεια να απαντήσετε στα ακόλουθα συνδυαστικά ερωτήματα:

1. Συνολική παρουσίαση του τζίρου (στήλη value) ανά μήνα (στις αντίστοιχες μονάδες μέτρησης)
2. Συνολική παρουσίαση του τζίρου (στήλη value) για κάθε χώρα (στις αντίστοιχες μονάδες μέτρησης)

3. Συνολική παρουσίαση του τζίρου (στήλη value) για κάθε μέσο μεταφοράς (στις αντίστοιχες μονάδες μέτρησης)
4. Συνολική παρουσίαση του τζίρου (στήλη value) για κάθε μέρα της εβδομάδας (στις αντίστοιχες μονάδες μέτρησης)
5. Συνολική παρουσίαση του τζίρου (στήλη value) για κάθε κατηγορία εμπορεύματος (στις αντίστοιχες μονάδες μέτρησης)

6. Παρουσίαση των 5 μηνών με το μεγαλύτερο τζίρο, ανεξαρτήτως μέσου μεταφοράς και είδους
7. Παρουσίαση των 5 κατηγοριών εμπορευμάτων με το μεγαλύτερο τζίρο, για κάθε χώρα
8. Παρουσίαση της ημέρας με το μεγαλύτερο τζίρο, για κάθε κατηγορία εμπορεύματος

Μπορείτε να επιλέξετε ως γλώσσα υλοποίησης είτε τη Scala είτε την Python, ωστόσο πρέπει να **χρησιμοποιηθεί η ίδια γλώσσα προγραμματισμού** για όλα τα ερωτήματα. Ο κώδικας που απαντά σε κάθε ερώτημα **θα πρέπει να περιέχει αναλυτικό σχολιασμό** και να βρίσκεται σε ένα μόνο αρχείο.

Η ονομασία του να ακολουθεί την σύμβαση:
Query[αριθμός_ερωτήματος].[επέκταση_γλώσσας] (π.χ. Query1.scala ή Query2.py).

Επιπλέον, στα παραδοτέα της άσκησης πρέπει να περιλαμβάνεται αναφορά της διαδικασίας σε μορφή word ή Pdf, στην οποία θα αποσαφηνίζονται τα στάδια της εγκατάστασης, τα βασικά σημεία του κώδικά σας, καθώς και screenshots από τα αποτελέσματα, ξεχωριστά για κάθε ερώτημα.

Σημείωση: Για την επίλυση της άσκησης είναι απαραίτητη η χρήση **Spark SQL και Dataframes** σε περιβάλλον Apache Spark, μέσω της πλατφόρμας Databricks (www.databricks.com) ή αν επιθυμείτε μπορείτε να εγκαταστήσετε το Apache Spark σε ένα vm.

Παραδοτέα

1. **Γραπτή Αναφορά** (σε αρχείο pdf ή word) που θα περιλαμβάνει:

    - **Αναλυτική περιγραφή της διαδικασίας που ακολουθήσατε (και για την εγκατάσταση – ενσωμάτωση των δεδομένων στη πλατφόρμα Databricks (ή VM) στο Apache Spark**
    - **Τον κώδικα σε γλώσσα python ή scala εμπλουτισμένο με αναλυτικό σχολιασμό**
    - **Screenshots παραδειγμάτων της εφαρμογής καθώς και της εγκατάστασης στο Databricks ή se vm (σε κάθε βήμα να υπάρχει αναλυτική περιγραφή)**
    - **Σχόλια - Παραδοχές που τυχόν έγιναν κατά την ανάπτυξη της εργασίας**

- **Αρχείο powerpoint παρουσίασης της εργασίας (για 10 λεπτά παρουσίαση)**

2. Συμπιεσμένα σε ένα αρχείο zip:
   - **Την πιο πάνω γραπτή αναφορά**
   - **Τον ΤΕΛΙΚΟ κώδικα σε python ή scala.**
   - **Το αρχείο powerpoint**

Το αρχείο zip πρέπει να έχει όνομα τον **αριθμό μητρώου** του φοιτητή (π.χ. 3972.zip), και να ανεβεί (**ΥΠΟΧΡΕΩΤΙΚΑ**) **στο e-class.**
Σε ξεχωριστό αρχείο .txt μέσα στο zip να αναφέρεται το **ονοματεπώνυμο, ο αριθμός μητρώου και η e-mail διεύθυνση του φοιτητή.**

Διευκρινήσεις

1. Η άσκηση θα γίνει σε ομάδες από 1 έως 3 άτομα.
2. Οριστική ημερομηνία παράδοσης είναι η τελευταία εβδομάδα της εξεταστικής εξέτασης περιόδου Ιουνίου 2023 **ΜΟΝΟ**! Αναλόγως θα καθοριστεί και η ημερομηνία της παρουσίασης της εργασίας.
3. Για τυχόν απορίες ή υποδείξεις μπορείτε να απευθύνεστε με e-mail στο mvonitsanos@ceid.upatras.gr ή στις Συζητήσεις στο e-class του μαθήματος https://eclass.upatras.gr/courses/CEID1175/

Παράρτημα

Apache Spark Documentation

Ο σημαντικότερος βοηθός σας κατά την εκπόνηση της εργασίας δεν είναι άλλος από τη τεκμηρίωση που θα βρείτε στην ιστοσελίδα του Apache Spark. Ένα καλό σημείο για να ξεκινήσετε την ενασχόληση σας είναι ο παρακάτω οδηγός (δείτε την ενότητα “Self-Contained Applications” για τη δημιουργία αυτόνομων προγραμμάτων):

https://spark.apache.org/docs/latest/quick-start.html

Τις κλάσεις και τις συναρτήσεις που θα χρειαστείτε μπορείτε να τις αναζητήσετε στα αντίστοιχα API docs:

https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.package

https://spark.apache.org/docs/latest/api/python/index.html

Για την υλοποίηση της εργασίας ιδιαίτερα χρήσιμες θα σας φανούν οι συναρτήσεις που προσφέρει η Spark SQL και μπορείτε να βρείτε στα παρακάτω links:

https://spark.apache.org/docs/latest/api/python/pyspark.sql.html#module-pyspark.sql.functions

Περισσότερες πληροφορίες καθώς και παραδείγματα κώδικα σχετικά με την Spark SQL και τα DataFrames υπάρχουν εδώ:

https://spark.apache.org/docs/latest/sql-programming-guide.html