PROJECT2_1[40%] Decentralized Query processing and Optimization in Apache Spark

Υπεύθυνοι Καθηγητές: Σπύρος Σιούτας (Καθηγητής ΤΜΗΥΠ), Γεράσιμος Βονιτσάνος (Συμβασιούχος Π.Δ. 407/80 Επίκουρος Καθηγητής)

- Θέμα 1

Στο αρχείο tempm.txt περιέχονται μετρήσεις για τη θερμοκρασία (σε °C) σε μια ευρωπαϊκή πόλη, την περίοδο από 13/2/2014 έως και 8/6/2014. Όμοια, στο αρχείο hum.txt υπάρχουν οι καταγραφές των τιμών της υγρασίας (σε ποσοστό %) για την ίδια χρονική περίοδο. Κάθε γραμμή των αρχείων αυτών αντιστοιχεί και σε μία ημέρα μετρήσεων ενώ οι γραμμές έχουν την παρακάτω μορφή:

{“timestamp1”: “value1”, “timestamp2”: “value2”, … , “timestampN”: “valueN”}

Σας ζητείται να υλοποιήσετε διαφορετικά προγράμματα στο περιβάλλον του Apache Spark, για να απαντήσετε σε 2 τουλάχιστον από τα παρακάτω ερωτήματα:

- Σε πόσες μέρες η θερμοκρασία κυμάνθηκε από 18 °C έως 22 °C;
- Ποιες ήταν οι 10 πιο κρύες μέρες και οι 10 πιο ζεστές μέρες;
- Ποιος μήνας είχε την υψηλότερη τυπική απόκλιση στις τιμές της υγρασίας;
- Ποια ήταν η μέγιστη και ποια η ελάχιστη τιμή του δείκτη δυσφορίας;

Σημείωση: Για τον υπολογισμό του δείκτη δυσφορίας χρησιμοποιείστε το τύπο:

$$DI = T - 0{,}55\,(1 - 0{,}01*RH)\,(T - 14{,}5)$$

T: θερμοκρασία, RH: υγρασία

- Θέμα 2

Τα αρχεία agn.us.txt, ainv.us.txt, ale.us.txt περιέχουν πλήρη ιστορικά δεδομένα με τις ημερήσιες τιμές για τρεις μετοχές που διαπραγματεύονται στα χρηματιστήρια αξιών NYSE και NASDAQ από το 2005 έως το 2017. Κάθε γραμμή των αρχείων αυτών αντιστοιχεί και σε μία ημέρα καταγραφής της κίνησης της κάθε μετοχής, και έχει την παρακάτω μορφή:

Date, Open, High, Low, Close, Volume, OpenInt

Σας ζητείται να υλοποιήσετε διαφορετικά προγράμματα στο περιβάλλον του Apache Spark, για να απαντήσετε σε 2 τουλάχιστον από τα παρακάτω ερωτήματα:

- Ποιος ήταν ο μέσος όρος για τις τιμές του ανοίγματος, κλεισίματος και όγκου συναλλαγών για κάθε ημερολογιακό μήνα για κάθε μετοχή

- Για πόσες ημέρες ήταν η τιμή του ανοίγματος της κάθε μετοχής πάνω από 35 δολάρια
- Ποιες ήταν οι μέρες με την υψηλότερη τιμή στο άνοιγμα και ποιες με την υψηλότερη τιμή στον όγκο συναλλαγών για κάθε μετοχή
- Ποιες ήταν οι χρονιές που κάθε μετοχή σημείωσε την υψηλότερη τιμή στο άνοιγμα και ποιες και τη χαμηλότερη τιμή στο κλείσιμο

- Θέμα 3

Το αρχείο tour_occ_ninat.xls περιέχει δεδομένα της Eurostat σχετικά με τις διανυκτερεύσεις τουριστών σε 36 ευρωπαικές χώρες για το διάστημα 2006-2015.

Σας ζητείται να υλοποιήσετε διαφορετικά προγράμματα στο περιβάλλον του Apache Spark, για να απαντήσετε σε 2 τουλάχιστον από τα παρακάτω ερωτήματα:

- Ποιος ήταν ο μέσος όρος διανυκτερεύσεων για κάθε χώρα για το χρονικό διάστημα 2007-2014
- Για πόσες και ποιες χρονιές ήταν ο αριθμός διανυκτερεύσεων της χώρας Ελλάδα υψηλότερος από 5 άλλες ευρωπαικές χώρες (της επιλογής σας)
- Ποιες ήταν οι χώρες με το μεγαλύτερο αριθμό διανυκτερεύσεων ανά έτος
- Ποια ήταν η χρονιά που η κάθε χώρα είχε το μικρότερο αριθμό διανυκτερεύσεων σε σχέση με όλες τις υπόλοιπες ευρωπαϊκές χώρες.

Οδηγίες

Μπορείτε να επιλέξετε ως γλώσσα υλοποίησης την Python, ή να γράψετε εντολές SQL ωστόσο πρέπει να χρησιμοποιηθεί η ίδια γλώσσα σε όλη την άσκηση. **Ο κώδικας που απαντά σε κάθε ερώτημα θα πρέπει να βρίσκεται σε ένα μόνο αρχείο και η ονομασία του να ακολουθεί την σύμβαση: *Query[αριθμός_ερωτήματος].[txt]*** **(π.χ. Query2.txt).** Τέλος, στα παραδοτέα της άσκησης πρέπει να περιλαμβάνεται μια σύντομη αναφορά σε μορφή pdf ή word στην οποία θα υπάρχουν σε μορφή screenshot τα αποτελέσματα από τη πλατφόρμα του Databricks, και θα αποσαφηνίζονται τα βασικά σημεία του κώδικά σας (αν κρίνετε εσείς σκόπιμο) και οι παραδοχές σας (αν υπάρχουν).

Η προαιρετική εργασία μπορεί να γίνει ατομικά ή σε ομάδα (αλλά θα δηλώσετε τα μέλη της ομάδας με email στο mvonitsanos@ceid.upatras.gr)

Ημερομηνία παράδοσης στο e-class του μαθήματος: ΕΞΕΤΑΣΤΙΚΗ ΙΑΝ-ΦΕΒΡ
Ημερομηνία **προαιρετικής** παρουσίασης στο zoom link του μαθήματος: ΕΞΕΤΑΣΤΙΚΗ ΙΑΝ-ΦΕΒΡ