Πανεπιστήμιο Πειραιώς, Τμήμα Πληροφορικής
ΠΜΣ «ΚΑΤΑΝΕΜΗΜΕΝΑ ΣΥΣΤΗΜΑΤΑ, ΑΣΦΑΛΕΙΑ ΚΑΙ ΑΝΑΔΥΟΜΕΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ»,
2023-2024

ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ

Διδάσκοντες: Νίκος Πελέκης, Γιώργος Παπαστεφανάτος.
Εργαστηριακοί βοηθοί: Γ. Αλεξίου , Σ. Μαρούλης

2η Εργαστηριακή Άσκηση
(ατομική)

Αντικείμενο της άσκησης

Σκοπός της άσκησης είναι η εξοικείωση με κατανεμημένα συστήματα διαχείρισης μεγάλων δεδομένων και πιο συγκεκριμένα με το Spark.

Το σύνολο δεδομένων που θα χρησιμοποιηθεί βρίσκεται στο docker image που σας έχει ήδη δοθεί στον φάκελο /labdata όπου βρισκόταν και τα δεδομένα για όλα τα παραδείγματα μέχρι τώρα.

Οι οδηγίες για το πως να εκκινήσουμε και το για το πως να συνδεθούμε στο docker image βρίσκονται στις διαφάνειες του μαθήματος καθώς και εδώ.

Για την υλοποίηση της εργασίας θα πρέπει να γίνει χρήση της μηχανής Spark και οι προτεινόμενες γλώσσες υλοποίησης είναι: Scala ,Java, Python.

Μεθοδολογία υλοποίησης της άσκησης

Μέρος 1

Ερώτημα 1. Εισαγωγή Δεδομένων και RDD operations.

(a) Create an RDD by loading the txt file from:. "/labdata/shakespear.txt"

(b) Create an RDD1 with each element being a single word from the whole text

(c) Create an RDD2 containing key-value pairs of the form: (line_number, line_text)

(d) Create an RDD3 containing key-value pairs of the form: (line_number, Array_of_line_words)

(e) Create an RDD4 containing key-value pairs of the form: (line_number, (word, frequency))

(f) Create an RDD5 containing key-value pairs of the form: (word, (line_number, frequency))

(g) Create an RDD6 containing key-value pairs of the form: (word, list_of_(line_number, frequency)). That is, for each word, gather all line-frequency pairs in a list.

(h) Create an RDD7 containing key-value pairs of the form: (word, list_of_lines). That is, for each word, gather all lines in a list.

(i) Create an RDD8 containing key-value pairs of the form: (word, total_frequency_of_word_in_text).

(j) Create an RDD9 containing key-value pairs of the form: ((word,line), frequency). That is, the key is the combination of word and line.

(k) Create an RDD10 which counts the above word-line combinations.

(l) Create an RDD11 containing key-value pairs of the form: (word, (line, total_frequency_of_word_in_line))

(m) Create an RDD12 that aggregates the contents of RRD11 into records of the form: (word, (number_of_lines, total_frequency_of_word_in_text))

Ερώτημα 2. Ανάκτηση δεδομένων

(a) Create an RDD by loading the /labdata/nyctaxisub.txt (**1st line is header**)

(b) Create a Dataframe from the above RDD by Inferring the schema using reflection or programmatically.

(c) Find the medallions with the most passengers between 2013-02-09 and 2013-02-11.

(d) Count all the medallion between the dates in (c)

(e) Find the medallions that lasted over 900sec and order them by passenger count in descending order. **(Ημερομηνία και τόπος παράδοσης)**

Απορίες σχετικά με την άσκηση

Για οποιαδήποτε απορία σχετικά με την άσκηση μπορείτε να απευθύνεστε στον κ. Γιώργο Αλεξίου (galexiou@athenarc.gr) με email.

Ημερομηνία και τόπος παράδοσης

Η εργασία να αποσταλεί με email (στο galexiou@athenarc.gr) μέχρι και τις **22/01/2024**. **Το τελικό παραδοτέο της εργασίας θα είναι ένα PDF report με το κομμάτι κώδικα για το κάθε ερώτημα/υποερώτημα συνοδευόμενο με μια μικρή περιγραφή / αιτιολόγηση**. Προαιρετικά μπορείτε να επισυνάψετε και όλο τον κώδικα σας. Παρακαλείσθε να έχετε την παρακάτω πρόταση ως θέμα στο email που θα στείλετε: "CDS110 - Εργασία Spark 2024".