Θέμα: Δημιουργία ολοκληρωμένης κατανεμημένης εφαρμογής με τη χρήση των Apache Airflow και Apache Spark

Επιβλέπων: Βασίλειος Ταμπακάς | **e-mail: tampakas@uop.gr**
τηλ:

Μέλη:

Ακαδημαϊκό Έτος: 2023-2024

Στόχοι

1. Αναλυτική παρουσίαση του Airflow καθώς και μιας συγκριτικής μελέτης εναλλακτικών του Airflow λύσεων, που θα παρουσιάζει:
    a. τα χαρακτηριστικά,
    b. τα προτερήματα και μειονεκτήματά των λύσεων σε σχέση με το Airflow και
    c. τις προτεινόμενες περιοχές εφαρμογής των λύσεων αυτών.
2. Υλοποίηση μιας ολοκληρωμένης κατανεμημένης εφαρμογής που θα συνδυάζει το Apache Airflow με το Apache Spark. Προτείνεται - αλλά υπάρχει διακριτική ευχέρεια επιλογής -να υλοποιηθεί μια εφαρμογή υβριδικού συστήματος συστάσεων (hybrid recommender/recommendation system),)

Αντικείμενο

Το Apache Airflow είναι ένα ανοικτού κώδικα εργαλείο που επιτρέπει, μέσω προγραμματισμού σε γλώσσα Python, τη δημιουργία, προγραμματισμό, αυτοματοποίηση και παρακολούθηση ροής εργασιών (workflow), και χρησιμοποιείται για την ενορχήστρωση κατανεμημένων εφαρμογών. Επιτρέπει τη δημιουργία πολύπλοκων διαδικασιών, γνωστών ως "Directed Acyclic Graphs" (DAGs), όπου οι εργασίες εκτελούνται σε πολλούς διακομιστές ή κόμβους. Οι εργασίες αυτές εκτελούνται σε συγκεκριμένη σειρά, με δυνατότητα συγχρονισμού και επανεκκίνησης σε περίπτωση αποτυχίας. Το Apache Airflow διαθέτει μια διεπαφή χρήστη που καθιστά εύκολη την παρακολούθηση της ροής δεδομένων μέσω διασωλήνωσης (pipeline). Επιπλέον, μπορεί κανείς να προβάλει άμεσα τις εξαρτήσεις (dependencies), τη πρόοδο των εργασιών, τα αρχεία καταγραφής (logfiles), τον κώδικα, τις εργασίες ενεργοποίησης (trigger tasks) και την κατάσταση επιτυχίας (status) των Data Pipelines. Επίσης υποστηρίζει τη λειτουργία backfilling (η διαδικασία επανυπολογισμού συνόλων δεδομένων από ακατέργαστα ιστορικά δεδομένα), η οποία μπορεί επίσης να χρησιμοποιηθεί για τον επανυπολογισμό οποιουδήποτε συνόλου δεδομένων μετά την πραγματοποίηση προσθηκών ή αλλαγών στον κώδικα.
Το Apache Airflow έχει εξελιχθεί σε μια από τις πιο ισχυρές λύσεις ανοικτού κώδικα για τη διαχείριση σωληνώσεων δεδομένων. Όλα αυτά τα χαρακτηριστικά διευκολύνουν τον χρήστη στην άμεση αντιμετώπιση και επίλυση τυχόν προβλημάτων.
Το Apache Spark είναι ένα ανοιχτού κώδικα πλαίσιο επεξεργασίας και ανάλυσης δεδομένων που προσφέρει υψηλή απόδοση και κλιμάκωση. Υποστηρίζει επεξεργασία μεγάλων όγκων δεδομένων, ανάλυση γράφων, μηχανική μάθηση και επεξεργασία ροής δεδομένων. Είναι ευέλικτο και χρησιμοποιείται ευρέως σε ποικίλες εφαρμογές, από ανάλυση δεδομένων έως μεγάλες επιχειρηματικές επεξεργασίες.
Ένα σημαντικό μέρος της πτυχιακής θα αφιερωθεί στην εγκατάσταση/μελέτη /διερεύνηση των δυο συστημάτων. Στη συνέχεια θα επιλεγεί ένα πεδίο δοκιμής για την ανάπτυξη μιας εφαρμογής.
Ένα τέτοιο πεδίο δοκιμής θα μπορούσε να είναι τα Συστήματα Συστάσεων

(Recommendation Systems). Τα συστήματα συστάσεων είναι συστήματα φιλτραρίσματος πληροφοριών που βοηθούν στην αντιμετώπιση του προβλήματος της υπερφόρτωσης πληροφοριών, φιλτράροντας και διαχωρίζοντας πληροφορίες και δημιουργώντας αποσπάσματα από μεγάλες ποσότητες δυναμικά δημιουργούμενων πληροφοριών σύμφωνα με τις προτιμήσεις, τα ενδιαφέροντα ή την παρατηρούμενη συμπεριφορά του χρήστη σχετικά με ένα συγκεκριμένο αντικείμενο ή αντικείμενα.
Οι υβριδικές μέθοδοι μπορούν να εφαρμοστούν με πολλούς τρόπους συνδυάζοντας προσεγγίσεις με βάση το περιεχόμενο (content based) και τη συνεργασία (collaborative), μετριάζοντας έτσι τα κοινά μειονεκτήματα αυτών των τεχνικών και ταυτόχρονα αυξάνοντας την απόδοση του συστήματος κατά την πραγματοποίηση προβλέψεων.
Η εφαρμογή θα μπορούσε να είναι ένα υβριδικό σύστημα συστάσεων όπου η επιλογή των αλγορίθμων του Spark καθώς και τα σύνολα δεδομένων (datasets) θα μπορούν να επιλεγούν από τον φοιτητή με βάση τα ερευνητικά του ενδιαφέροντα υπό την τελική έγκριση του επιβλέποντα καθηγητή. Θα είναι στην ευχέρεια της πτυχιακής ο εμπλουτισμός των δεδομένων με web scraping.

Ακολουθεί μια λίστα με Ιστοσελίδες με Σύνολα Δεδομένων προκειμένου να διευκολυνθεί ο φοιτητής στην επιλογή του κατάλληλου συνόλου δεδομένων:

1. Google Dataset Search (μηχανή αναζήτησης για σύνολα δεδομένων)
2. Kaggle
3. Data.Gov
4. Datahub.io
5. Global Health Observatory Data Repository

Η εργασία περιλαμβάνει: (π.χ. σχεδιασμό συστήματος, ανάπτυξη συστήματος, διερεύνηση βιβλιογραφίας κ.λ.π)
[√] Σχεδιασμό και ανάπτυξη συστήματος
[√] Συγκριτική μελέτη και πλαίσιο αξιολόγησης
[√] Ανάλυση και σχεδιασμό μοντέλου