Συστήματα Διαχείρισης Δεδομένων Μεγάλου Όγκου

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

PROJECT ΕΑΡΙΝΟΥ ΕΞΑΜΗΝΟΥ 2023
Έκδοση εγγράφου 1.1 02/05/2023

ΥΠΕΥΘΥΝΟΙ ΚΑΘΗΓΗΤΕΣ: Μεγαλοοικονόμου Βασίλειος, Κομνηνός Ανδρέας

Big Data Applications

Σκοπός εργασίας

Η παρούσα εργασία αποσκοπεί στην εξοικείωση των φοιτητών με τις τρέχουσες τεχνολογίες αποθήκευσης, ανάκτησης και ανάλυσης των Big Data. Τα Big Data εμφανίζονται με μια πληθώρα από μορφές όπως τα web logs, τα internet clickstreams, τα αδόμητα ή ημιδομημένα δεδομένα. Η ανάλυση των Big Data χρησιμοποιεί πηγές δεδομένων οι οποίες παρέμεναν ανεκμετάλλευτες από τις συμβατικές λύσεις. Ζητήματα όπως η διαχείριση ετερογενών κι ανομοιόμορφων σημαντικά μεγάλων δεδομένων από σχεσιακές βάσεις δεδομένων, η ανάκτηση μεγάλων data sets που είναι διάσπαρτα σε ομογενή ή ετερογενή συστήματα, η επεξεργασία φυσικής γλώσσας, η μηχανική μάθηση και τεχνητή νοημοσύνη καθώς και η πρόβλεψη που στηρίζεται σε αδόμητα δεδομένα, ικανοποιούνται πλέον από τα συστήματα ανάλυσης Big Data. Ειδικότερα στην παρούσα εργασία θα ασχοληθούμε και θα δούμε στην πράξη (hands-on) την διαχείριση δεδομένων με την χρήση εργαλείων ανοικτού κώδικα, και ειδικότερα του Cassandra.

Ομάδες

Η εργασία μπορεί να εκπονηθεί από μεμονωμένα άτομα ή ομάδες το πολύ δύο ατόμων.

Προετοιμασία & Documentation

Προαπαιτούμενο SW

Τα ερωτήματα 1-3 μπορούν να υλοποιηθούν τοπικά στον υπολογιστή σας ή σε cluster εργασίας (DataStax AstraDB). Για την υλοποίηση του project στον τοπικό υπολογιστή θα χρειαστείτε τα παρακάτω βοηθητικά εργαλεία:

- Δημιουργία μιας εικονικής μηχανής με Linux (π.χ. ubuntu) – προαιρετικό αλλά συστήνεται ανεπιφύλακτα.
- Εγκατάσταση Java 8 (11 με μικρές αλλαγές στο config)
- Εγκατάσταση Python 3.6+
- Εγκατάσταση Jupyter Notebooks (προαιρετικό αλλά συνιστάται)
- Εγκατάσταση Cassandra

Πρόσβαση στο cluster εργασίας

Για το 4ο ερώτημα της εργασίας θα χρειαστεί να δημιουργήσετε ένα instance της AstraDB (παραλλαγή της Cassandra) στο cloud. Το ίδιο instance μπορεί να αξιοποιηθεί και για τα ερωτήματα 1-3 αν δε θέλετε να εγκαταστήσετε τοπικά την Cassandra. Για οδηγίες, δείτε το παράρτημα 1.

Περιγραφή datasets κι εργασιών

Δεδομένα

Στην συγκεκριμένη εργασία, το dataset αποτελείται από τέσσερα αρχεία .csv. Τα αρχεία μπορείτε να κατεβάσετε από το Kaggle κι αφορούν το **Food.com – Recipes and Interactions**, κι ειδικότερα τα:

- **RAW_interactions.csv:** raw interaction data (reviews, ratings) for the recipes
- **RAW_recipes.csv:** raw data (steps, ingredients etc) for the recipes
- **PP_recipes.csv:** pre-processed recipe data
- **PP_users.csv:** pre-processed user data
- **Ingr.map.pkl:** αρχείο pickle που περιέχει ένα pandas dataframe με τα ids και τα συστατικά (ονόματα) κάθε συνταγής

URL δεδομένων: https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions

Συμπληρωματικά μπορείτε να διαβάσετε το paper στο οποίο έγινε χρήση του dataset για να κατανοήσετε περισσότερο τα pre-processed αρχεία, αλλά και να δείτε και το σχετικό κώδικα εδώ https://paperswithcode.com/paper/generating-personalized-recipes-from

Ερώτημα 1: Σχεδιασμός βάσης δεδομένων

Σε αντίθεση με τα RDBMS όπου η μοντελοποίηση των δεδομένων (data modelling) γίνεται με γνώμονα την αποφυγή πλεονασμών, στην Cassandra ο πλεονασμός δεδομένων είναι αναπόφευκτος καθώς δεν υποστηρίζονται joins. Συνεπώς η διαδικασία κατασκευής μιας βάσης ξεκινά αναλογιζόμενοι τα πιθανά queries που μπορεί να κάνει ένας χρήστης, και κατασκευάζοντας κατάλληλους πίνακες ώστε να μπορούν να εξάγονται γρήγορα τα σχετικά αποτελέσματα.

Φανταστείτε ότι η βάση που θα φτιάξετε υποστηρίζει ένα user-interface αναζήτησης και επιλογής συνταγών. Σκεφτείτε λοιπόν τις δυνατές ανάγκες ενός χρήστη καθώς πλοηγείται στην εφαρμογή:

1) Εμφάνιση των συνταγών που είναι δημοφιλείς (έχουν καλή βαθμολογία) εντός ενός χρονικού διαστήματος (π.χ. τους τελευταίους 3 μήνες) – αυτό μπορεί να αποτελεί κάλλιστα την αρχική οθόνη «προτάσεων» προς το χρήστη
2) Να αναζητήσει την/τις συνταγές που περιέχουν κάποιες λέξεις – κλειδιά στον τίτλο
3) Να αναζητήσει ταινίες με βάση την κατηγορία δυσκολίας (π.χ. χρόνος παρασκευής ή πλήθος βημάτων) και να τις λάβει με βάση κάποια ταξινόμηση (π.χ. μέση βαθμολογία)
4) Να δει τις λεπτομερείς πληροφορίες για κάποια συνταγή (κατηγορία δυσκολίας, μέση βαθμολογία, ετικέτες, υλικά, διαδικασία παρασκευής)
5) Να δει τις top-n συνταγές που σχετίζονται με κάποια ετικέτα

Με βάση αυτά τα είδη ερωτημάτων το 1ο ερώτημα για την εργασία είναι να φτιάξετε το κατάλληλο σχήμα της βάσης, με το κατάλληλο πλήθος πινάκων, ώστε να υποστηρίζονται τα ανωτέρω είδη ερωτημάτων.

- Σχεδιάστε το εννοιολογικό μοντέλο της βάσης.
- Σχεδιάστε το application workflow διάγραμμα για την εφαρμογή.

- Εφαρμόζοντας τα κατάλληλα mapping rules, σχεδιάστε το διάγραμμα Chebotko για το σύστημά μας.

Ερώτημα 2: Υλοποίηση συστήματος

Υλοποιήστε τη βάση δεδομένων που σχεδιάσατε στην Cassandra, με χρήση κατάλληλων DDL statements. Μπορείτε να το κάνετε είτε με τα κατάλληλα shell statements είτε με κάποιο Python script. Σε κάθε περίπτωση, παραθέστε τα statements που χρησιμοποιήσατε για την κατασκευή κάθε keyspace.

Στη συνέχεια, γράψτε τα κατάλληλα Python scripts ώστε να εισάγετε τα δεδομένα που κατεβάσατε από το Kaggle στα keyspaces που δημιουργήσατε. Παραθέστε, μαζί με τα scripts, και ενδεικτικά screenshots που τεκμηριώνουν την επιτυχή εισαγωγή των δεδομένων (π.χ. αποτελέσματα εκτέλεσης SELECT * statements).

Ερώτημα 3: Εκτέλεση ερωτημάτων

Το ερώτημα αυτό Στον υπολογιστή σας, θα πρέπει να τρέξετε τα παρακάτω ερωτήματα με κατάλληλα Python scripts. Σημειώστε την απάντηση, και το χρόνο εκτέλεσης των ερωτημάτων.

Ερωτήματα προς υλοποίηση:

1. Εμφάνιση των 30 συνταγών με την υψηλότερη μέση βαθμολογία μεταξύ 01/01/2012 και 31/05/2012
2. Εμφάνιση όλων των λεπτομερειών για την ταινία «chic greek salad» (κατηγορία δυσκολίας, διατροφικές αξίες, βήματα, περιγραφή, μέση βαθμολογία)
3. Εμφάνιση των ταινιών της κατηγορίας «εύκολη» ταξινομημένες ως προς τη μέση βαθμολογία τους (εδώ θα πρέπει να φτιάξετε μόνοι σας την κατηγορία για κάθε ταινία – προτείνεται ένας απλός αλγόριθμος όπου η κάθε συνταγή λαμβάνει ένα σκορ *πλήθος_βημάτων* x *χρόνος_παρασκευής* και στη συνέχεια διαχωρισμός των ταινιών σε 4 κατηγορίες ανάλογα με το εύρος των τιμών που θα παραχθούν)
4. Εμφάνιση των συνταγών που περιέχουν την ετικέτα “slow-cooker” με ταξινόμηση ανά ημερομηνία προσθήκης (πιο πρόσφατες πρώτα)
5. Εμφάνιση των 20 συνταγών με την υψηλότερη μέση βαθμολογία για την ετικέτα “cocktail”.

Ερώτημα 4: Σύγκριση επιδόσεων με βάση τα επίπεδα συνέπειας

Για να αξιολογήσουμε τη διαφορά στις επιδόσεις με βάση τα επίπεδα συνέπειας που μπορούν να ρυθμιστούν στην Cassandra, θα αξιοποιήσουμε το cluster εργασίας. Για τη χρήση του θα πρέπει να προσέλθετε στο εργαστήριο, όπου θα σας δοθεί πρόσβαση σε microcluster φτιαγμένο από Raspberry Pi 4 – η πρόσβαση θα γίνει κατόπιν ραντεβού που θα ανακοινωθεί σε επόμενο χρόνο. Για την διαδικασία θα πρέπει να έχετε φροντίσει να έχετε έτοιμα όλα τα ερωτήματα και τα δεδομένα σας, ώστε απλά να τα εκτελέσετε και να καταγράψετε τις επιδόσεις του συστήματος.

A. Στο microcluster, δημιουργήστε ένα αντίγραφο της βάσης δεδομένων και των keyspaces που έχετε φτιάξει στα προηγούμενα ερωτήματα στον τοπικό σας υπολογιστή.
B. Επαναλάβετε την εκτέλεση των scripts εισαγωγής δεδομένων, και σημειώστε το χρόνο εκτέλεσης τους, ορίζοντας για τα sessions το επίπεδο του write consistency σε:
    a. ALL

b. QUORUM
c. ONE

Προφανώς, πριν εκτελέσετε τα ερωτήματα εισαγωγής δεδομένων για κάθε επίπεδο, θα πρέπει προηγουμένως να διαγράψετε όλα τα δεδομένα που έχουν μπει από προηγούμενη διαδικασία εισόδου.

Γ. Επαναλάβετε την εκτέλεση των προηγούμενων ερωτημάτων και σημειώστε το χρόνο εκτέλεσης, ορίζοντας για κάθε ερώτημα τα read consistency levels σε:
   a. ALL
   b. QUORUM
   c. ONE

   Κάθε ένα από τα ερωτήματα πρέπει να επαναληφθεί 10 φορές, και για κάθε εκτέλεση, χρησιμοποιήστε διαφορετικές παράμετρους της επιλογής σας. Για παράδειγμα, για το ερώτημα 5, δημιουργήστε μια λίστα με 10 tags της επιλογής σας, ή αν προτιμάτε με τυχαία επιλογή από τα υπάρχοντα (π.χ. cocktail, slow-cooker, breakfast, …) και εκτελέστε το ερώτημα για κάθε tag. Συνεπώς το ερώτημα 5 πρέπει να εκτελεστεί 30 φορές (10 φορές για κάθε επίπεδο read consistency). Για κάθε εκτέλεση, σημειώνετε το χρόνο που χρειάστηκε.

Δ. Αναλύστε τα αποτελέσματα εκτέλεσης των ερωτημάτων εισαγωγής δεδομένων παραθέτοντας συγκριτικά γραφήματα των μέσων όρων του χρόνου που χρειάστηκε για την εκτέλεση των ερωτημάτων, για την είσοδο των δεδομένων, και για την ανάκτηση δεδομένων, ώστε να φαίνεται η διαφορά στην απόδοση ανάλογα με το consistency level.

Παραδοτέα

Για την εργασία θα πρέπει να παραδώσετε:

1) Τον κώδικα που γράψατε για την εκτέλεση των ερωτημάτων στο τοπικό σας μηχάνημα και στο cluster, **αποκλειστικά σε γλώσσα Python (αρχεία .py)**.
2) Αναφορά στην οποία περιλαμβάνονται
   a. Ο σχεδιασμός της ΒΔ (εννοιολογικό μοντέλο, application workflow, Chebotko diagrams).
   b. Ερωτήματα DDL και εισαγωγής δεδομένων, με screenshots από την επιτυχή εκτέλεση των τελευταίων (SELECT * FROM [keyspace] LIMIT 5)
   c. Ερωτήματα DML για την ανάκτηση δεδομένων και παραγόμενα αποτελέσματα από την εκτέλεση ερωτημάτων
   d. Πίνακας με τους χρόνους εκτέλεσης των ερωτημάτων εισαγωγής των δεδομένων για κάθε write consistency level.
   e. Πίνακας με τους μέσους όρους χρόνου εκτέλεσης των ερωτημάτων ανάκτησης των δεδομένων για κάθε read consistency level
   f. Γραφική παράσταση που απεικονίζει το ΜΕΣΟ ΟΡΟ του χρόνου εκτέλεσης των ερωτημάτων ανάκτησης δεδομένων ανά read consistency level και σχολιασμός των ευρημάτων σας.
3) Τα ωμά δεδομένα μετρήσεων σε αρχείο λογιστικού φύλλου (spreadsheet) που καταγράψατε για τις μετρήσεις σας (αρχεία .xlsx, ods).

Να χρησιμοποιηθεί αποκλειστικά το template υποβολής που επισυνάπτεται στην εκφώνηση.
Παρέχεται template σε αρχείο excel για την καταγραφή των δεδομένων σας και την αυτόματη παραγωγή των γραφικών παραστάσεων (δεν είναι υποχρεωτικό να το χρησιμοποιήσετε)

Καταληκτική ημερομηνία υποβολής: 3/7/2023

Επικοινωνία

Για την επιτυχία σας στο project θα χρειαστείτε καθοδήγηση καθώς κι απαντήσεις σε ερωτήματα που ίσως δεν έχουν καλυφθεί στο παρόν κείμενο. Για απορίες μπορείτε να αποστέλλετε στο akomninos@ceid.upatras.gr

Παράρτημα 1 - Χρήσιμες πληροφορίες

Οδηγοί και χρήσιμο documentation

Εγκατάσταση και documentation Cassandra

- https://cassandra.apache.org/doc/latest/cassandra/getting_started/installing.html

Cassandra Data Modelling

- Introduction to Cassandra Data Modelling https://cassandra.apache.org/doc/latest/cassandra/data_modeling/intro.html
- A. Chebotko, A. Kashlev and S. Lu, "A Big Data Modeling Methodology for Apache Cassandra," 2015 IEEE International Congress on Big Data, 2015, pp. 238-245, doi: 10.1109/BigDataCongress.2015.41.
    - https://ieeexplore.ieee.org/document/7207225
    - http://shiyong.eng.wayne.edu/papers/bigdata2015_andrey.pdf

DataStax driver documentation

- https://docs.astra.datastax.com/en/landing_page/doc/landing_page/apiDocs.html

CQL reference

- https://docs.datastax.com/en/cql-oss/3.3/cql/cqlIntro.html

Ορισμός επιπέδων consistency και παρακολούθηση στατιστικών εκτέλεσης

- Με χρήση CQL, δείτε τον οδηγό εδώ
    - https://docs.datastax.com/en/cql-oss/3.3/cql/cql_using/useTracingTrace.html
- Με χρήση του Python driver δείτε τον οδηγό εδώ
    - Ορισμός consistency level https://docs.datastax.com/en/developer/python-driver/3.25/getting_started/#setting-a-consistency-level
    - Εξαγωγή στατιστικών https://docs.datastax.com/en/developer/python-driver/3.25/faq/#how-do-i-trace-a-request

Παράρτημα 2 – Πρότυπο αναφοράς άσκησης

Συστήματα Διαχείρισης Δεδομένων Μεγάλου Όγκου

Εργαστηριακή Άσκηση 2022/23

Ονομα	**Επώνυμο**	**ΑΜ**

Βεβαιώνω ότι είμαι συγγραφέας της παρούσας εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτήν, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για το συγκεκριμένο μάθημα/σεμινάριο/πρόγραμμα σπουδών.

Έχω ενημερωθεί ότι σύμφωνα με τον εσωτερικό κανονισμό λειτουργίας του Πανεπιστημίου Πατρών άρθρο 50§6, τυχόν προσπάθεια αντιγραφής ή εν γένει φαλκίδευσης της εξεταστικής και εκπαιδευτικής διαδικασίας από οιονδήποτε εξεταζόμενο, πέραν του μηδενισμού, συνιστά βαρύ πειθαρχικό παράπτωμα.

Υπογραφή

___ / ___ / 2023

Υπογραφή

___ / ___ / 2023

Συνημμένα αρχεία κώδικα

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

Αρχείο	Αφορά το ερώτημα	Περιγραφή/Σχόλιο
Erotima1.py	*1*	*Περιέχει όλα τα ερωτήματα για το ερ. 1*

Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

[Τεχνικά χαρακτηριστικά φυσικού Η/Υ που χρησιμοποιήθηκε για την εργασία, αν χρησιμοποιήθηκε μόνο το Astra DB μπορείτε απλά να αναφέρετε αυτό αντί για τον πίνακα]

Χαρακτηριστικό	**Τιμή**
CPU model	*Intel i5-10400F*
CPU clock speed	*2.9GHz*
Physical CPU cores	*6*
Logical CPU cores	*12*
RAM	*16*
Secondary Storage Type	*~~HDD~~/SSD*

Ερώτημα 1: Σχεδιασμός ΒΔ

[δώστε το εννοιολογικό μοντέλο, το application workflow και το Chebotko diagram μαζί με τυχόν επεξηγήσεις που θέλετε να γράψετε για να εξηγήσετε τη φιλοσοφία του καθενός και να το περιγράψετε]

Ερώτημα 2: Ερωτήματα DDL

[επαναλαμβάνετε τον παρακάτω πίνακα για κάθε keyspace στη ΒΔ σας]

Keyspace	*[δώστε το όνομα του keyspace προς δημιουργία]*
DDL statement	*[δώστε το DDL statement για τη δημιουργία του keyspace]*
Screenshot	*[δώστε ένα screenshot που δείχνει δεδομένα μέσα στο keyspace με ως αποτέλεσμα του ερωτήματος SELECT * FROM [keyspace] LIMIT 5]*

Ερώτημα 3: Απαντήσεις ερωτημάτων

[Μην παραθέσετε στο έντυπο όλες τις επιστρεφόμενες εγγραφές! Να καταγράψετε μόνο αυτές που αναφέρει το πρότυπο.]

Ερώτημα	**Απάντηση**
Εμφάνιση των 30 συνταγών με την υψηλότερη μέση βαθμολογία μεταξύ 01/01/2012 και 31/05/2012	*[παραθέστε τις 5 πρώτες μόνο]*
Εμφάνιση όλων των λεπτομερειών για την ταινία «chic greek salad» (κατηγορία δυσκολίας, διατροφικές αξίες, βήματα, περιγραφή, μέση βαθμολογία)	*[όλες τις λεπτομέρειες]*
Εμφάνιση των ταινιών της κατηγορίας «εύκολη» ταξινομημένες ως προς τη μέση βαθμολογία τους	*[παραθέστε τις 5 πρώτες μόνο]*
Εμφάνιση των συνταγών που περιέχουν την ετικέτα “slow-cooker” με ταξινόμηση ανά	*[παραθέστε τις 5 πρώτες μόνο]*

ημερομηνία προσθήκης (πιο πρόσφατες πρώτα)	
Εμφάνιση των 20 συνταγών με την υψηλότερη μέση βαθμολογία για την ετικέτα “cocktail”.	*[παραθέστε τις 5 πρώτες μόνο]*

Ερώτημα 4Α: Χρόνοι εισαγωγής δεδομένων

	Επίπεδο write consistency		
	ALL	**QUORUM**	**ONE**
[Keyspace 1]	*[χρόνος εκτέλεσης]*	*[χρόνος εκτέλεσης]*	*[χρόνος εκτέλεσης]*
[Keyspace 2]	*[χρόνος εκτέλεσης]*	*[χρόνος εκτέλεσης]*	*[χρόνος εκτέλεσης]*
…	…	…	…
[Keyspace n]	*[χρόνος εκτέλεσης]*	*[χρόνος εκτέλεσης]*	*[χρόνος εκτέλεσης]*
Μέσος όρος			

Ερώτημα 4Β: Χρόνοι ανάκτησης δεδομένων

	Επίπεδο write consistency		
	ALL	**QUORUM**	**ONE**
Ερώτημα 1	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*
Ερώτημα 2	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*
Ερώτημα 3	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*
Ερώτημα 4	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*
Ερώτημα 5	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*	*[μ.ο. για τις 10 επαναλήψεις]*
Μέσος όρος			

Ερώτημα 4Γ: Σχολιασμός αποτελεσμάτων

[Συνοψίστε τα αποτελέσματα των χρόνων εισαγωγής δεδομένων και ανάκτησης δεδομένων με κατάλληλες γραφικές παραστάσεις (δύο) και σχολιάστε τα ευρήματά σας – γιατί παρατηρούνται οι όποιες διαφορές στο χρόνο εκτέλεσης; Σε ποια στοιχεία της αρχιτεκτονικής της ΒΔ και του θεωρήματος CAP οφείλονται;]

Βιβλιογραφία

[πηγές που χρησιμοποιήσατε για την εργασία]