ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

Ομάδες το πολύ 2 ατόμων

ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΙΙ
PROJECT ΕΑΡΙΝΟΥ ΕΞΑΜΗΝΟΥ 2020
ΥΠΕΥΘΥΝΟΣ ΚΑΘΗΓΗΤΗΣ : κος Μεγαλοοικονόμου Βασίλης

BigData Analytics

Σκοπός εργασίας

Η παρούσα εργασία αποσκοπεί στην εξοικείωση των φοιτητών με τις τρέχουσες τεχνολογίες αποθήκευσης, ανάκτησης και ανάλυσης των Bigdata. Τα Bigdata εμφανίζονται με μια πληθώρα από μορφές όπως τα web logs, τα internet clickstreams, τα αδόμητα ή ημιδομημένα δεδομένα. Η ανάλυση των Bigdata 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 Bigdata.

Ειδικότερα στην παρούσα εργασία θα ασχοληθούμε και θα δούμε στην πράξη (hands-on) την διαχείριση δεδομένων με την χρήση του οικοσυστήματος Hadoop και ειδικότερα των frameworks Apache HBASE και Apache PHOENIX. Τέλος θα θέσουμε ερωτήματα σε ένα ειδικά διαμορφωμένο cluster με εργαλεία της HortonWorks (πλέον απόκτημα της Cloudera).

Προαπαιτούμενο SW

Για την υλοποίηση του project θα χρειαστείτε τα παρακάτω βοηθητικά εργαλεία στον HY σας:

- Εγκατάσταση του OpenVPN client (για πρόσβαση στο vpn του project). Αντιγραφή του CEID19DBII.ovpn που θα σας σταλεί στο κατάλογο config.
- Εγκατάσταση του putty (ssh client) για telnet από τον υπολογιστή σας στο Okeanos

Προετοιμασία & Documentation

Κατά την διάρκεια της υλοποίησης θα χρειαστεί να μελετήσετε:

τις βασικές αρχές λειτουργίας του HBASE framework στο https://hbase.apache.org/book.html . Ειδικότερα χρήσιμες για το Project και την **εξέταση** του είναι οι ενότητες

Για το Hadoop

- A Brief Summary of Apache Hadoop

Για την HBASE

- Getting Started,
- The Apache HBase Shell & HBase shell commands,
- Data Model, Architecture,
- Apache HBase Coprocessors
- HBase Tools and Utilities ειδικά το ImportTsv
- Appendix E: Compression and Data Block Encoding In HBase,
- Appendix H: HFile format,

Για το PHOENIX

- Overview
- F.A.Q.
- Secondary Indexing
- Views
- Grammar
- Joins
- Subqueries

Πρόσβαση στο cluster εργασίας

Για το συγκεκριμένο Project έχει δημιουργηθεί ειδική υποδομή στο Okeanos. Οδηγίες πρόσβασης θα δοθούν σε κάθε ομάδα ξεχωριστά κατόπιν της σχετικής δήλωσης σας με email στο sergiang@ceid.upatras.gr .

Επικοινωνία

Για την επιτυχία σας στο project θα χρειαστείτε καθοδήγηση καθώς κι απαντήσεις σε ερωτήματα που ίσως δεν έχουν καλυφθεί στο παρόν κείμενο. Για τον λόγο αυτό θα είμαστε σε άμεση επικοινωνία με την χρήση του collaboration tool slack. Θα αποσταλεί πρόσκληση στις ομάδες που θα δηλώσουν το project.

Περιγραφή datasets κι εργασιών

Στην συγκεκριμένη εργασία, θα εισάγετε στην HBASE ένα dataset που αποτελείται από τέσσερα αρχεία csv (θα σας δοθούν οδηγίες για την ακριβή τοποθεσία τους στο cluster) με την χρήση του εργαλείου ImportTsv. Τα αρχεία έχουν ληφθεί από το Kaggle κι αφορούν το Yelp Dataset, κι ειδικότερα τα

yelp_business.csv (περίπου 175.000 επιχειρήσεις)

yelp_business_attributes.csv

yelp_business_hours.csv και

yelp_business_checkin.csv (περίπου 4.000.000 καταγραφές). Το συγκεκριμένο έχει υποστεί μερική επεξεργασία σε σχέση με το αρχικό κι έχει ήδη αποθηκευτεί στον Hadoop cluster

Κατά την εισαγωγή των αρχείων στην HBASE, θα ακολουθήσετε τα εξής:

Θα δημιουργήσετε το USERxx.YELP_BUSINESS table με families a) BASE για τα στοιχεία του yelp_business.csv, b) ATTRIBUTES για τα στοιχεία του yelp_business_attributes.csv και c) HOURS για τα στοιχεία του yelp_business_checkin.csv.

Θα δημιουργήσετε το USERxx.YELP_CHECKIN table με family PERHOUR

ΠΡΟΣΟΧΗ σε όλα τα ονόματα (tables, families, qualifiers) να χρησιμοποιήσετε ΚΕΦΑΛΑΙΑ ΑΓΓΛΙΚΑ από τα A ως το Z & αριθμούς μόνο (μην χρησιμοποιήσετε underscore ή dash). Το USERxx αφορά τον χρήστη που θα δοθεί σε κάθε ομάδα (δηλαδή η ομάδα που θα έχει τον χρήστη user04 θα δημιουργήσει το table με πρόθεμα USER04).

Στην συνέχεια θα χρησιμοποιήσετε το εργαλείο sqlline.py στον folder /usr/hdp/2.6.3.0-235/phoenix/bin/ ώστε να εκτελέσετε τα ερωτήματά σας κι όποιες εργασίες βελτιστοποίησης ή διαμόρφωσης των tables. Το εργαλείο sqlline.py σας δίνει πρόσβαση αφορά στο command line περιβάλλον του apache Phoenix. Θα διασυνδέσετε το apache Phoenix με τα tables που έχετε δημιουργήσει ώστε να μπορέσετε να δώσετε απάντηση στα παρακάτω ερωτήματα:

Q1. Δώστε τα ονόματα, την πολιτεία, το πλήθος των αστεριών των πρώτων 1000 επιχειρήσεων που είναι ενεργές

Q2. Δώστε τα ονόματα, την διεύθυνση, την πόλη και το πλήθος των reviews των επιχειρήσεων που ανήκουν στην κατηγορία ‘Drugstores’ ταξινομημένα κατά φθίνουσα σειρά των reviews.

Q3. Δώστε το σύνολο του πλήθους των reviews ανά κατηγορία για τις επιχειρήσεις που είναι ενεργές και λειτουργούν όλες τι ημέρες της εβδομάδας όλο το εικοσιτετράωρο.

Q4. Δώστε το πλήθος των επιχειρήσεων ανά πολιτεία που δεν επιτρέπεται το κάπνισμα και λειτουργούν την Κυριακή.

Q5. Δώστε το σύνολο των checkin ανά ημέρα κι αντίστοιχη ώρα.

Q6. Δώστε το σύνολο των checkin ανά κατηγορία των ενεργών επιχειρήσεων από την 14:00 ως και την 16:00 τις καθημερινές (εκτός Σαββατοκύριακου).

Q7. Εντοπίστε τις 100 πρώτες επιχειρήσεις (με παράθεση όλων των στοιχείων της family BASIC) με τα περισσότερα checkin το Σάββατο.

Παραδοτέα

Με καταληκτική ημερομηνία την Πέμπτη 27 Ιουνίου 2020 23:59 θα αποστείλετε ηλεκτρονικά στις διευθύνσεις gsergian@ceid.upatras.gr με cc vasilis@ceid.upatras.gr τα παρακάτω:

1. Αναφορά με κατάλληλες επεξηγήσεις σε κάθε φάση των εργασιών σας και απαραίτητα
    a. τις εντολές εισαγωγής των δεδομένων στην HBASE
    b. τις εντολές που εφαρμόσατε στο sqlline.py για
        i. την διασύνδεση της HBASE με το PHOENIX
        ii. την βελτιστοποίηση στο PHOENIX ώστε να εξυπηρετούνται αποδοτικά τα ερωτήματα
        iii. τις εντολές για την λήψη των αποτελεσμάτων των 7 ερωτημάτων
2. τα αποτελέσματα σε μορφή csv με ονόματα USERxxQy.CSV όπου xx ο αριθμός του χρήστη σας στο σύστημα και y ο αριθμός του ερωτήματος.

Εξέταση

Η εξέταση θα είναι προφορική διάρκειας 30’ ανά ομάδα και θα συμπεριλαμβάνει ερωτήματα στο υλικό μελέτης, και στην εργασία σας. Ημερομηνία κι ώρα θα καθοριστεί κατόπιν συνεννόησης.

Για κάθε απορία στην διάθεσή σας.

Με εκτίμηση, Γιώργος Σεργιάννης