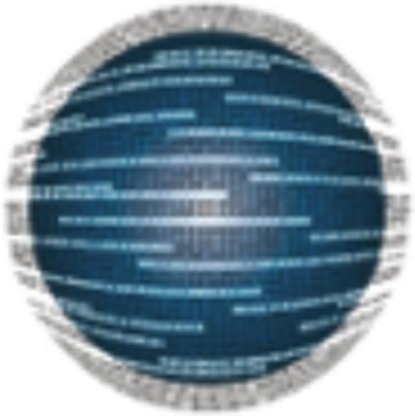

Πανεπιστήμιο Πατρών Πολυτεχνική Σχολή Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Μάθημα Βάσεις Δεδομένων 2

Επιβλέπων καθηγητές: Β. Μεγαλοοικονόμου, Γ. Σεργιάννης

Φοιτητές που εκπονήσαν την εργασία: Βασδάρης Όμηρος ΑΜ 1054429

Δελημιχάλης Αλέξανδρος ΑΜ 1054324

Περιεχόμενα

Εντολές Hbase

Σύνδεση στο Hbase:

Εντολή: hbase shell

Δημιουργία πινάκων

Χρησιμοποιούμε τις παρακάτω εντολές για να δημιουργήσουμε έναν πίνακα YELPBUSINESS και ένα πίνακα YELPCHEKIN, με τις αντίστοιχες, column families, BASE, ATTRIBUTES, HOURS και PERHOUR, αντίστοιχα.

```
create 'USER03.YELPBUSINESS','BASE','ATTRIBUTES','HOURS'
```

```
create 'USER03.YELPCHECKIN','PERHOUR'
```

Φόρτωση δεδομένων από τα αρχεία στους πίνακες

Στο αρχικό terminal:

Για τον Πίνακα YELPBUSINESS:

Από το yelp_business.csv :

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv '-Dimporttsv.separator=,' -
Dimporttsv.columns=HBASE_ROW_KEY,BASE:NAME,BASE:NEIGHBORHOOD,BASE:ADDRESS,BAS
E:CITY,BASE:STATE,BASE:POSTALCODE,BASE:LATITUDE,BASE:LONGTITUDE,BASE:STARS,BASE:RE
VIEWCOUNT,BASE:ISOPEN,BASE:CATEGORIES USER03.YELPBUSINESS
../hbase/dataset/yelp_business.csv
```

Από το yelp_business_atrributes.csv:

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv '-Dimporttsv.separator=,' -
Dimporttsv.columns=HBASE_ROW_KEY,ATTRIBUTES:ACCEPTSINSURANCE,ATTRIBUTES:BYAPPOI
NTMENTONLY,ATTRIBUTES:BUSINESSACCEPTSCREDITCARDS,ATTRIBUTES:BUSINESSPARKINGGA
RAGE,ATTRIBUTES:BUSINESSPARKINGSTREET,ATTRIBUTES:BUSINESSPARKINGVALIDATED,ATTRIB
UTES:BUSINESSPARKINGLOT,ATTRIBUTES:BUSINESSPARKINGVALET,ATTRIBUTES:HAIRSPECIALIZE
SINCOLORING,ATTRIBUTES:HAIRSPECIALIZESINAFRICANAMERICAN,ATTRIBUTES:HAIRSPECIALIZE
SINCURLY,ATTRIBUTES:HAIRSPECIALIZESINPERMS,ATTRIBUTES:HAIRSPECIALIZESINKIDS,ATTRIBU
TES:HAIRSPECIALIZESINEXTENSIONS,ATTRIBUTES:HAIRSPECIALIZESINASIAN,ATTRIBUTES:HAIRSP
ECIALIZESINSTRAIGHTPERMS,ATTRIBUTES:RESTAURANTSPRICERANGETWO,ATTRIBUTES:GOODF
ORKIDS,ATTRIBUTES:WHEELCHAIRACCESSIBLE,ATTRIBUTES:BIKEPARKING,ATTRIBUTES:ALCOHOL
,ATTRIBUTES:HASTV,ATTRIBUTES:NOISELEVEL,ATTRIBUTES:RESTAURANTSATTIRE,ATTRIBUTES:M
USICDJ,ATTRIBUTES:MUSICBACKGROUNDMUSIC,ATTRIBUTES:MUSICNOMUSIC,ATTRIBUTES:MU
SICKARAOKE,ATTRIBUTES:MUSICLIVE,ATTRIBUTES:MUSICVIDEO,ATTRIBUTES:MUSICJUKEBOX,AT
TRIBUTES:AMBIENCEROMANTIC,ATTRIBUTES:AMBIENCEINTIMATE,ATTRIBUTES:AMBIENCECLAS
SY,ATTRIBUTES:AMBIENCEHIPSTER,ATTRIBUTES:AMBIENCEDIVEY,ATTRIBUTES:AMBIENCETOURI
```

STY,ATTRIBUTES:AMBIENCETRENDY,ATTRIBUTES:AMBIENCEUPSCALE,ATTRIBUTES:AMBIENCECASUAL,ATTRIBUTES:RESTAURANTSGOODFORGROUPS,ATTRIBUTES:CATERS,ATTRIBUTES:WIFI,ATTRIBUTES:RESTAURANTSRESERVATIONS,ATTRIBUTES:RESTAURANTSTAKEOUT,ATTRIBUTES:HAPPYHOUR,ATTRIBUTES:GOODFORDANCING,ATTRIBUTES:RESTAURANTSTABLESERVICE,ATTRIBUTES:OUTDOORSEATING,ATTRIBUTES:RESTAURANTSDELIVERY,ATTRIBUTES:BESTNIGHTSMONDAY,ATTRIBUTES:BESTNIGHTSTUESDAY,ATTRIBUTES:BESTNIGHTSFRIDAY,ATTRIBUTES:BESTNIGHTSWEDNESDAY,ATTRIBUTES:BESTNIGHTSTHURSDAY,ATTRIBUTES:BESTNIGHTSSUNDAY,ATTRIBUTES:BESTNIGHTSSATURDAY,ATTRIBUTES:GOODFORMEALDESSERT,ATTRIBUTES:GOODFORMEALLATENIGHT,ATTRIBUTES:GOODFORMEALLUNCH,ATTRIBUTES:GOODFORMEALDINNER,ATTRIBUTES:GOODFORMEALBREAKFAST,ATTRIBUTES:GOODFORMEALBRUNCH,ATTRIBUTES:COATCHECK,ATTRIBUTES:SMOKING,ATTRIBUTES:DRIVETHRU,ATTRIBUTES:DOGSALLOWED,ATTRIBUTES:BUSINESSACCEPTSBITCOIN,ATTRIBUTES:OPEN24HOURS,ATTRIBUTES:BYOBCORKAGE,ATTRIBUTES:BYOB,ATTRIBUTES:CORKAGE,ATTRIBUTES:DIETARYRESTRICTIONSDAIRYFREE,ATTRIBUTES:DIETARYRESTRICTIONSGLUTENFREE,ATTRIBUTES:DIETARYRESTRICTIONSVEGAN,ATTRIBUTES:DIETARYRESTRICTIONSKOSHER,ATTRIBUTES:DIETARYRESTRICTIONSHALAL,ATTRIBUTES:DIETARYRESTRICTIONSSOYFREE,ATTRIBUTES:DIETARYRESTRICTIONSVEGETARIAN,ATTRIBUTES:AGESALLOWED,ATTRIBUTES:RESTAURANTSCOUNTERSERVICE USER03.YELPBUSINESS
../hbase/dataset/yelp_business_attributes.csv

Από το yelp_business_hours.csv:

hbase org.apache.hadoop.hbase.mapreduce.ImportTsv '-Dimporttsv.separator=,' -Dimporttsv.columns=HBASE_ROW_KEY,HOURS:MONDAY,HOURS:TUESDAY,HOURS:WEDNESDAY,HOURS:THURSDAY,HOURS:FRIDAY,HOURS:SATURDAY,HOURS:SUNDAY USER03.YELPBUSINESS ../hbase/dataset/yelp_business_hours.csv

Σβήνουμε την πρώτη εγράφη, με τα ονόματα των στηλών και ελέγχουμε τον τελικό πίνακα :

Στο shell της hbase:

deleteall 'USER03.YELPBUSINESS', 'business_id'

και για να ελέγξουμε ότι αφαιρέθηκε επιτυχώς:

scan 'USER03.YELPBUSINESS', {FILTER=>"(RowFilter(=,'regexstring:business_id'))"}

Για το πίνακα YELPCHECKIN:

Από το yelp_checkin.csv:

hbase org.apache.hadoop.hbase.mapreduce.ImportTsv '-Dimporttsv.separator=,' -Dimporttsv.columns=HBASE_ROW_KEY,PERHOUR:BUSINESSID,PERHOUR:WEEKDAY,PERHOUR:HOUR,PERHOUR:CHECKIN USER03.YELPCHECKIN ../hbase/dataset/yelp_checkin.csv

Σβήνουμε την πρώτη εγράφη, με τα ονόματα των στηλών και ελέγχουμε τον τελικό πίνακα :

Στο shell της hbase:

deleteall 'USER03.YELPCHECKIN' , 'business_id'

και για να ελέγξουμε ότι αφαιρέθηκε επιτυχώς:

scan 'USER03.YELPCHECKIN' , {FILTER=>"(RowFilter(=,'regexstring:business_id'))"}

Διασύνδεση με PHOENIX

Σύνδεση στο Phoenix:

Τρέχουμε την εντολή:

/usr/hdp/2.6.3.0-235/phoenix/bin/sqlline.py

Δημιουργία VIEWS για την σύνδεση της HBASE με την PHOENIX:

Σημείωση: Τα πεδία στο view είναι όλα VARCHAR, διότι ήθελε συγκεκριμένη κωδικοποίηση για τις υπόλοιπες μορφές, οπότε απλώς μέσα στα SELECT κάνουμε το απαραίτητο typecast.

Για τον πίνακα YELPBUSINESS:

CREATE VIEW USER03.YELPBUSINESS (ROWKEY VARCHAR PRIMARY KEY,BASE.NAME VARCHAR,BASE.NEIGHBORHOOD VARCHAR,BASE.ADDRESS VARCHAR,BASE.CITY VARCHAR,BASE.STATE VARCHAR,BASE.POSTALCODE VARCHAR,BASE.LATITUDE VARCHAR,BASE.LONGTITUDE VARCHAR,BASE.STARS VARCHAR,BASE.REVIEWCOUNT VARCHAR,BASE.ISOPEN VARCHAR,BASE.CATEGORIES VARCHAR,
ATTRIBUTES.ACCEPTSINSURANCE VARCHAR,ATTRIBUTES.BYAPPOINTMENTONLY VARCHAR,ATTRIBUTES.BUSINESSACCEPTSCREDITCARDS VARCHAR,ATTRIBUTES.BUSINESSPARKINGGARAGE VARCHAR,ATTRIBUTES.BUSINESSPARKINGSTREET VARCHAR,ATTRIBUTES.BUSINESSPARKINGVALIDATED VARCHAR,ATTRIBUTES.BUSINESSPARKINGLOT VARCHAR,ATTRIBUTES.BUSINESSPARKINGVALET VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINCOLORING VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINAFRICANAMERICAN VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINCURLY VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINPERMS VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINKIDS VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINEXTENSIONS VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINASIAN VARCHAR,ATTRIBUTES.HAIRSPECIALIZESINSTRAIGHTPERMS VARCHAR,ATTRIBUTES.RESTAURANTSPRICERANGETWO VARCHAR,ATTRIBUTES.GOODFORKIDS VARCHAR,ATTRIBUTES.WHEELCHAIRACCESSIBLE VARCHAR,ATTRIBUTES.BIKEPARKING VARCHAR,ATTRIBUTES.ALCOHOL VARCHAR,ATTRIBUTES.HASTV VARCHAR,ATTRIBUTES.NOISELEVEL VARCHAR,ATTRIBUTES.RESTAURANTSATTIRE

VARCHAR,ATTRIBUTES.MUSICDJ VARCHAR,ATTRIBUTES.MUSICBACKGROUNDMUSIC
VARCHAR,ATTRIBUTES.MUSICNOMUSIC VARCHAR,ATTRIBUTES.MUSICKARAOKE
VARCHAR,ATTRIBUTES.MUSICLIVE VARCHAR,ATTRIBUTES.MUSICVIDEO
VARCHAR,ATTRIBUTES.MUSICJUKEBOX VARCHAR,ATTRIBUTES.AMBIENCEROMANTIC
VARCHAR,ATTRIBUTES.AMBIENCEINTIMATE VARCHAR,ATTRIBUTES.AMBIENCECLASSY
VARCHAR,ATTRIBUTES.AMBIENCEHIPSTER VARCHAR,ATTRIBUTES.AMBIENCEDIVEY
VARCHAR,ATTRIBUTES.AMBIENCETOURISTY VARCHAR,ATTRIBUTES.AMBIENCETRENDY
VARCHAR,ATTRIBUTES.AMBIENCEUPSCALE VARCHAR,ATTRIBUTES.AMBIENCECASUAL
VARCHAR,ATTRIBUTES.RESTAURANTSGOODFORGROUPS VARCHAR,ATTRIBUTES.CATERS
VARCHAR,ATTRIBUTES.WIFI VARCHAR,ATTRIBUTES.RESTAURANTSRESERVATIONS
VARCHAR,ATTRIBUTES.RESTAURANTSTAKEOUT VARCHAR,ATTRIBUTES.HAPPYHOUR
VARCHAR,ATTRIBUTES.GOODFORDANCING VARCHAR,ATTRIBUTES.RESTAURANTSTABLESERVICE
VARCHAR,ATTRIBUTES.OUTDOORSEATING VARCHAR,ATTRIBUTES.RESTAURANTSDELIVERY
VARCHAR,ATTRIBUTES.BESTNIGHTSMONDAY VARCHAR,ATTRIBUTES.BESTNIGHTSTUESDAY
VARCHAR,ATTRIBUTES.BESTNIGHTSFRIDAY VARCHAR,ATTRIBUTES.BESTNIGHTSWEDNESDAY
VARCHAR,ATTRIBUTES.BESTNIGHTSTHURSDAY VARCHAR,ATTRIBUTES.BESTNIGHTSSUNDAY
VARCHAR,ATTRIBUTES.BESTNIGHTSSATURDAY VARCHAR,ATTRIBUTES.GOODFORMEALDESSERT
VARCHAR,ATTRIBUTES.GOODFORMEALLATENIGHT
VARCHAR,ATTRIBUTES.GOODFORMEALLUNCH VARCHAR,ATTRIBUTES.GOODFORMEALDINNER
VARCHAR,ATTRIBUTES.GOODFORMEALBREAKFAST
VARCHAR,ATTRIBUTES.GOODFORMEALBRUNCH VARCHAR,ATTRIBUTES.COATCHECK
VARCHAR,ATTRIBUTES.SMOKING VARCHAR,ATTRIBUTES.DRIVETHRU
VARCHAR,ATTRIBUTES.DOGSALLOWED VARCHAR,ATTRIBUTES.BUSINESSACCEPTSBITCOIN
VARCHAR,ATTRIBUTES.OPEN24HOURS VARCHAR,ATTRIBUTES.BYOBCORKAGE
VARCHAR,ATTRIBUTES.BYOB VARCHAR,ATTRIBUTES.CORKAGE
VARCHAR,ATTRIBUTES.DIETARYRESTRICTIONSDAIRYFREE
VARCHAR,ATTRIBUTES.DIETARYRESTRICTIONSGLUTENFREE
VARCHAR,ATTRIBUTES.DIETARYRESTRICTIONSVEGAN
VARCHAR,ATTRIBUTES.DIETARYRESTRICTIONSKOSHER
VARCHAR,ATTRIBUTES.DIETARYRESTRICTIONSHALAL
VARCHAR,ATTRIBUTES.DIETARYRESTRICTIONSSOYFREE
VARCHAR,ATTRIBUTES.DIETARYRESTRICTIONSVEGETARIAN
VARCHAR,ATTRIBUTES.AGESALLOWED VARCHAR,ATTRIBUTES.RESTAURANTSCOUNTERSERVICE
VARCHAR,HOURS.MONDAY VARCHAR,HOURS.TUESDAY VARCHAR,HOURS.WEDNESDAY
VARCHAR,HOURS.THURSDAY VARCHAR,HOURS.FRIDAY VARCHAR,HOURS.SATURDAY
VARCHAR,HOURS.SUNDAY VARCHAR);

Για τον πίνακα YELPCHECKIN:

CREATE VIEW USER03.YELPCHECKIN (PERHOUR.ID VARCHAR,ROWKEY VARCHAR PRIMARY
KEY,PERHOUR.WEEKDAY VARCHAR,PERHOUR.HOUR VARCHAR,PERHOUR.CHECKIN VARCHAR);

Αποθήκευση των αποτελεσμάτων των SELECT σε μορφή csv:

Τρέχουμε τις εντολές:

Για παράδειγμα για το ερώτημα 1, θα κάνουμε:

!outputformat csv

!record USER03Q1.csv

Q1;

!record

Έτσι, για κάθε ερώτημα, αποθηκεύουμε τα αποτελέσματα του σε ένα αρχείο csv.

Queries

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

```
SELECT BASE.NAME,BASE.STATE,BASE.STARS FROM USER03.YELPBUSINESS WHERE BASE.ISOPEN
= '1' LIMIT 1000;
```

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.

```
SELECT BASE.NAME,BASE.ADDRESS,BASE.CITY,BASE.REVIEWCOUNT FROM
USER03.YELPBUSINESS WHERE BASE.CATEGORIES LIKE 'Drugstores' ORDER BY
BASE.REVIEWCOUNT DESC;
```

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.

```
SELECT SUM(BASE.REVIEWCOUNT),CATEGORIES FROM (SELECT
COUNT(BASE.REVIEWCOUNT),BASE.CATEGORIES AS CATEGORIES FROM USER03.YELPBUSINESS
WHERE BASE.ISOPEN = '1' AND HOURS.MONDAY = '0:0-0:0' AND HOURS.TUESDAY = '0:0-0:0'
AND HOURS.WEDNESDAY = '0:0-0:0' AND HOURS.THURSDAY = '0:0-0:0' AND HOURS.FRIDAY =
'0:0-0:0' AND HOURS.SATURDAY = '0:0-0:0' AND HOURS.SUNDAY = '0:0-0:0' GROUP BY
BASE.CATEGORIES) GROUP BY CATEGORIES;
```

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

```
SELECT COUNT(BASE.NAME),BASE.STATE FROM USER03.YELPBUSINESS WHERE
ATTRIBUTES.SMOKING = 'False' AND HOURS.SUNDAY != 'None' GROUP BY BASE.STATE;
```

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

(Σχόλιο: Το COALESCE(TO_NUMBER(REGEXP_SUBSTR) χρησιμοποιείται για τι typecast των πεδίων)

```
SELECT SUM(COALESCE(TO_NUMBER(REGEXP_SUBSTR(PERHOUR.CHECKIN, '^\d+(\.\d+)?')), 0)),
PERHOUR.WEEKDAY, PERHOUR.HOUR FROM USER03.YELPCHECKIN GROUP BY
PERHOUR.WEEKDAY,PERHOUR.HOUR;
```

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

```
SELECT SUM(COALESCE(TO_NUMBER(REGEXP_SUBSTR(CHECKIN, '^\d+(.\d+)?')), 0)),
BASE.CATEGORIES FROM USER03.YELPBUSINESS FULL JOIN USER03.YELPCHECKIN ON
USER03.YELPBUSINESS.ROWKEY = USER03.YELPCHECKIN.ROWKEY WHERE BASE.ISOPEN ='1'
AND PERHOUR.WEEKDAY != 'Sat' AND PERHOUR.WEEKDAY != 'Sun' AND (PERHOUR.HOUR =
'14:00' OR PERHOUR.HOUR = '15:00' OR PERHOUR.HOUR = '16:00') GROUP BY
BASE.CATEGORIES;
```

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

```
SELECT
BASE.NAME,BASE.NEIGHBORHOOD,BASE.ADDRESS,BASE.CITY,BASE.STATE,BASE.POSTALCODE,B
ASE.LATITUDE,BASE.LONGTITUDE,BASE.STARS,BASE.REVIEWCOUNT,BASE.ISOPEN,BASE.CATEGO
RIES FROM USER03.YELPCHECKIN FULL JOIN USER03.YELPBUSINESS ON
USER03.YELPBUSINESS.ROWKEY = USER03.YELPCHECKIN.ROWKEY WHERE PERHOUR.WEEKDAY
= 'Sat' ORDER BY PERHOUR.CHECKIN DESC LIMIT 100;
```