ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΙΙ
PROJECT 2019-2020

Ελευθεριάδης Πέτρος 1041741

1.a

Αρχικά έτρεξα την εντολή “hbase shell ”, δημιούργησα τα tables όπως ζητήθηκαν στην εκφώνηση

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

create 'USER03.YELP_CHECKIN', 'PERHOUR'
```

Ακολούθησε η εισαγωγή των δεδομένων της family “BASE” στο table με τη χρήση της ImportTSV:

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -Dimporttsv.columns="HBASE_ROW_KEY, BASE:NAME, BASE:NEIGHBORHOOD, BASE:ADDRESS, BASE:CITY, BASE:STATE, BASE:POSTALCODE, BASE:LATITUDE, BASE:LONGITUDE, BASE:STARS, BASE:REVIEWCOUNT, BASE:ISOPEN, BASE:CATEGORIES" "USER03.YELP_BUSINESS" /user/hbase/dataset/yelp_business.csv
```

Ακολούθησε η εισαγωγή των δεδομένων της family “ATTRIBUTES” με την εντολή:

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -Dimporttsv.columns="HBASE_ROW_KEY, ATTRIBUTES:ACCEPTSINSURANCE, ATTRIBUTES:BYAPPOINTMENTONLY, ATTRIBUTES:BUSINESSACCEPTSCREDITCARDS, ATTRIBUTES:BUSINESSPARKINGGARAGE, ATTRIBUTES:BUSINESSPARKINGSTREET, ATTRIBUTES:BUSINESSPARKINGVALIDATED, ATTRIBUTES:BUSINESSPARKINGLOT, ATTRIBUTES:BUSINESSPARKINGVALET, ATTRIBUTES:HAIRSPECIALIZESINCOLORING, ATTRIBUTES:HAIRSPECIALIZESINAFRICANAMERICAN, ATTRIBUTES:HAIRSPECIALIZESINCURLY, ATTRIBUTES:HAIRSPECIALIZESINPERMS, ATTRIBUTES:HAIRSPECIALIZESINKIDS, ATTRIBUTES:HAIRSPECIALIZESINEXTENSIONS, ATTRIBUTES:HAIRSPECIALIZESINASIAN, ATTRIBUTES:HAIRSPECIALIZESINSTRAIGHTPERMS,ATTRIBUTES:RESTAURANTSPRICERANGE2, ATTRIBUTES:GOODFORKIDS, ATTRIBUTES:WHEELCHAIRACCESIBLE, ATTRIBUTES:BIKEPARKING, ATTRIBUTES:ALCOHOL, ATTRIBUTES:HASTV, ATTRIBUTES:NOISELEVEL, ATTRIBUTES:RESTAURANTSATTIRE, ATTRIBUTES:MUSICDJ, ATTRIBUTES:MUSICBACKGROUNDMUSIC, ATTRIBUTES:MUSICNOMUSIC, ATTRIBUTES:MUSICKARAOKE, ATTRIBUTES:MUSICLIVE, ATTRIBUTES:MUSICVIDEO, ATTRIBUTES:MUSICJUKEBOX, ATTRIBUTES:AMBIENCEROMANTIC, ATTRIBUTES:AMBIENCEINTIMATE, ATTRIBUTES:AMBIENCECLASSY, ATTRIBUTES:AMBIENCEHIPSTER, ATTRIBUTES:AMBIENCEDIVEY, ATTRIBUTES:AMBIENCETOURISTY, ATTRIBUTES:AMBIENCETRENDY, ATTRIBUTES:AMBIENCEUPSCALE, ATTRIBUTES:AMBIENCECASUAL, ATTRIBUTES:RESTAURANTGOODFORGROUPS, ATTRIBUTES:CATERS, ATTRIBUTES:WIFI, ATTRIBUTES:RESTAURANTRESERVATIONS, 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:DIETARYRESTRICTIONSVEGETERIAN, ATTRIBUTES:AGESALLOWED, ATTRIBUTES:RESTAURANTSCOUNTERSERVICE" "USER03.YELP_BUSINESS" /user/hbase/dataset/yelp_business_attributes.csv
```

Τέλος τα δεδομένα της family “HOURS"

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.YELP_BUSINESS" /user/hbase/dataset/yelp_business_hours.csv

Παρόμοια έγινε και το table CHECKIN με τα δεδομένα της family PERHOUR

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

1.b.i

Εφόσον τελείωσα με την εισαγωγή των δεδομένων στην hbase, πήγα στο bin folder του phoenix και έτρεξε την sqlline.py

Το πρώτο πράμα που έκανα εκεί ήταν η σύνδεση του phoenix με τα tables στην hbase

Η εντολή για τη σύνδεση του table Business είναι η εξής:

CREATE VIEW "USER03.YELP_BUSINESS"(pk 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".LONGITUDE 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".RESTAURANTSPRICERANGE2 VARCHAR, "ATTRIBUTES".GOODFORKIDS VARCHAR, "ATTRIBUTES".WHEELCHAIRACCESIBLE 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".RESTAURANTGOODFORGROUPS VARCHAR, "ATTRIBUTES".CATERS VARCHAR, "ATTRIBUTES".WIFI VARCHAR, "ATTRIBUTES".RESTAURANTRESERVATIONS 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".DIETARYRESTRICTIONSVEGETERIAN 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)

Σε όλες τις στήλες έδωσα τύπο δεδομένων VARCHAR.

Η εντολή για τη σύνδεση του Checkin είναι η εξής:

CREATE VIEW "USER03.YELP_CHECKIN"(pk VARCHAR PRIMARY KEY, "PERHOUR".BUSINESSID VARCHAR, "PERHOUR".WEEKDAY VARCHAR, "PERHOUR".HOUR VARCHAR, "PERHOUR".CHECKINS VARCHAR)

Επίσης όλα τα columns με τύπο VARCHAR

1.b.iii

Query 1

```
1.  SELECT "BASE".NAME, "BASE".CITY, "BASE".STATE, "BASE".STARS
2.  FROM "USER03.YELP_BUSINESS2"
3.  WHERE "BASE".ISOPEN='1'
4.  LIMIT 1000;
```

Query 2

```
5.  SELECT "BASE".NAME, "BASE".ADDRESS, "BASE".CITY, "BASE".REVIEWCOUNT
6.  FROM "USER03.YELP_BUSINESS2"
7.  WHERE "BASE".CATEGORIES='Drugstores'
8.  ORDER BY TO_NUMBER("BASE".REVIEWCOUNT) DESC;
```

Query 3

```
9.  SELECT "BASE".CATEGORIES, SUM(TO_NUMBER("BASE".REVIEWCOUNT)) AS SUM
10. FROM "USER03.YELP_BUSINESS2"
11. WHERE "BASE".ISOPEN='1'
12. AND "ATTRIBUTES".OPEN24HOURS = 'True'
13. GROUP BY "BASE".CATEGORIES;
```

Query 4

```
14. SELECT "BASE".STATE, COUNT(*) AS BUSINESS_COUNT
15. FROM "USER03.YELP_BUSINESS2"
16. WHERE "ATTRIBUTES".SMOKING='False'
17. AND "HOURS".SUNDAY!='NONE'
18. GROUP BY "BASE".STATE;
```

Query 5

```
19. SELECT "PERHOUR".WEEKDAY, "PERHOUR".HOUR, SUM("PERHOUR".CHECKINS) AS CHEKIN_SUM
20. FROM "USER03.YELP_CHECKIN6"
21. GROUP BY "PERHOUR".WEEKDAY, "PERHOUR".HOUR;
```

Σχόλια: Στο συγκεκριμένο query δε δούλευε το TO_NUMBER() όπως δούλευε στα προηγούμενα και στα επόμενα. Γι'αυτο έφτιαξα άλλη μια βάση checkin δηλώνοντας τα CHECKINS ως BIGINT. Δοκίμασα και άλλα numerical types αλλά αυτό είχε τα καλύτερα αποτελέσματα. Προφανώς σε κάποια sum υπήρξε overflow καθώς βγήκαν αρνητικά.

Query 6

```
22. SELECT "BASE".CATEGORIES, SUM(TO_NUMBER("PERHOUR".CHECKINS)) AS CHEKIN_SUM
```

```
23. FROM "USER03.YELP_BUSINESS3" INNER JOIN "USER03.YELP_CHECKIN5"
24. ON  "USER03.YELP_BUSINESS3".PK = "USER03.YELP_CHECKIN5"."PERHOUR".BUSINESSID
25. WHERE ("PERHOUR".HOUR = '14:00'
26. OR "PERHOUR".HOUR = '15:00'
27. OR "PERHOUR".HOUR = '16:00')
28. AND "PERHOUR".WEEKDAY!='Sat'
29. AND "PERHOUR".WEEKDAY!='Sun'
30. AND "BASE".ISOPEN='1'
31. GROUP BY "BASE".CATEGORIES;
```

Σχόλια: Χρησιμοποίησα inner join γιατί κάνω select columns κι απ’τα δυο tables

Query 7

```
32. SELECT /*+ USE_SORT_MERGE_JOIN*/ "BASE".NAME, "BASE".NEIGHBORHOOD,
33. "BASE".ADDRESS, "BASE".CITY, "BASE".STATE, "BASE".POSTALCODE, "BASE".LATITUDE,
34. "BASE".LONGITUDE, "BASE".STARS, "BASE".REVIEWCOUNT, "BASE".ISOPEN,
35. "BASE".CATEGORIES, SUM(TO_NUMBER("PERHOUR".CHECKINS)) AS CHEKIN_SUM
36. FROM "USER03.YELP_BUSINESS2" INNER JOIN "USER03.YELP_CHECKIN5"
37. ON  "USER03.YELP_BUSINESS2".PK = "USER03.YELP_CHECKIN5"."PERHOUR".BUSINESSID
38. WHERE "PERHOUR".WEEKDAY='Sat'
39. GROUP BY "BASE".NAME, "BASE".NEIGHBORHOOD, "BASE".ADDRESS, "BASE".CITY,
40. "BASE".STATE, "BASE".POSTALCODE, "BASE".LATITUDE, "BASE".LONGITUDE,
41. "BASE".STARS, "BASE".REVIEWCOUNT, "BASE".ISOPEN, "BASE".CATEGORIES
42. ORDER BY SUM(TO_NUMBER("PERHOUR".CHECKINS)) DESC
43. LIMIT 100;
```

Σχόλια: Εδώ χρησιμοποίησα το Sort Merge Join επειδή όταν το έτρεχα κανονικά (με hash join) εβγαζε error πως δεν έφτανε η μνήμη. Έτσι λύνεται το πρόβλημα και τρέχει κανονικά.