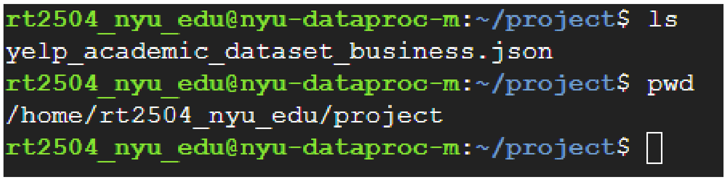
Data Ingestion

My dataset is the “Yelp Business dataset”, found on kaggle.

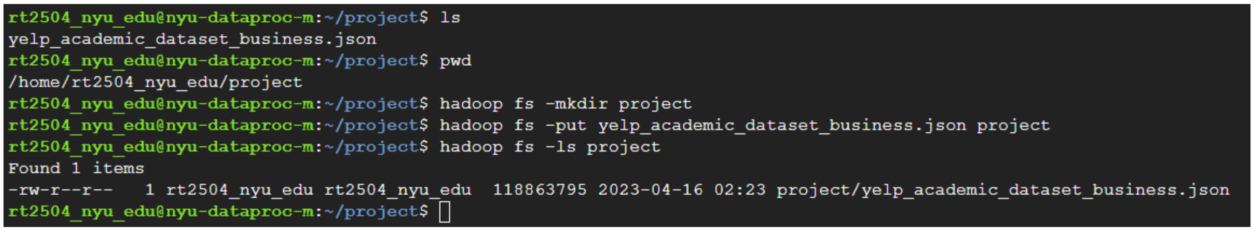
Source: https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_business.json

I exported the json dataset from kaggle into my nyu dataproc



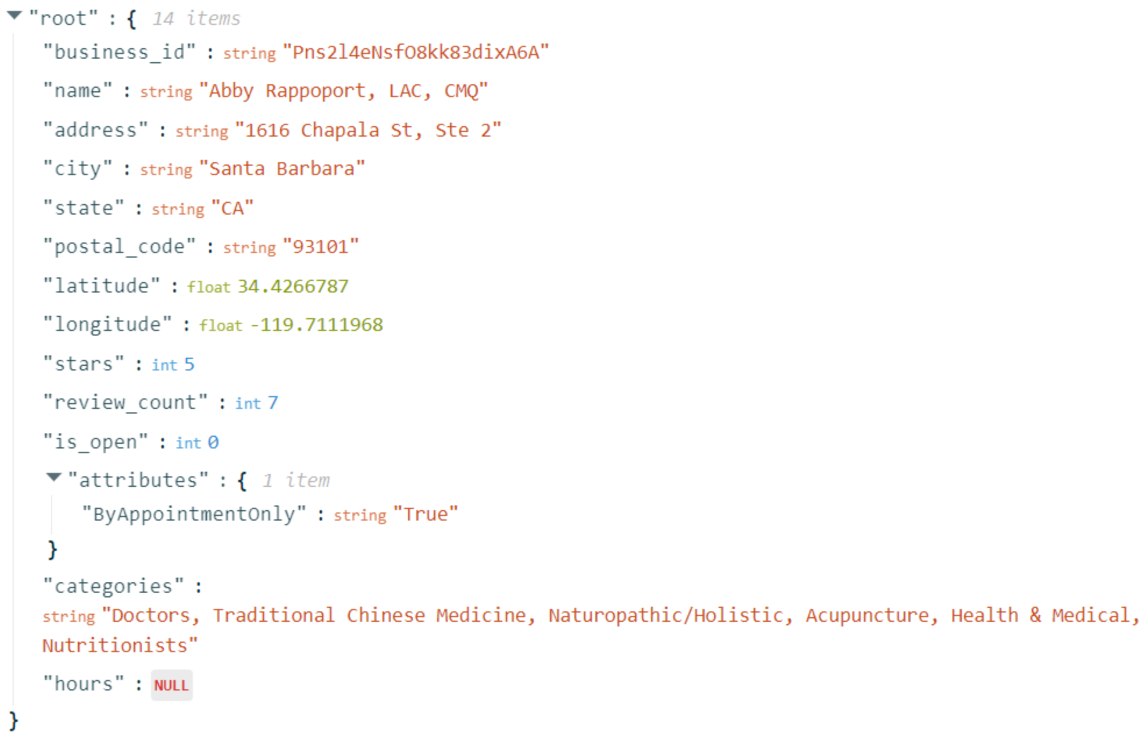
```
rt2504_nyu_edu@nyu-dataproc-m:~/project$ ls
yelp_academic_dataset_business.json
rt2504_nyu_edu@nyu-dataproc-m:~/project$ pwd
/home/rt2504_nyu_edu/project
rt2504_nyu_edu@nyu-dataproc-m:~/project$
```

Move the dataset to HDFS



```
rt2504_nyu_edu@nyu-dataproc-m:~/project$ ls
yelp_academic_dataset_business.json
rt2504_nyu_edu@nyu-dataproc-m:~/project$ pwd
/home/rt2504_nyu_edu/project
rt2504_nyu_edu@nyu-dataproc-m:~/project$ hadoop fs -mkdir project
rt2504_nyu_edu@nyu-dataproc-m:~/project$ hadoop fs -put yelp_academic_dataset_business.json project
rt2504_nyu_edu@nyu-dataproc-m:~/project$ hadoop fs -ls project
Found 1 items
-rw-r--r--   1 rt2504_nyu_edu rt2504_nyu_edu  118863795 2023-04-16 02:23 project/yelp_academic_dataset_business.json
rt2504_nyu_edu@nyu-dataproc-m:~/project$
```

The datacard defines the attribute model for this dataset:



```
▼ "root" : { 14 items
    "business_id" : string "Pns2l4eNsfO8kk83dixA6A"
    "name" : string "Abby Rappoport, LAC, CMQ"
    "address" : string "1616 Chapala St, Ste 2"
    "city" : string "Santa Barbara"
    "state" : string "CA"
    "postal_code" : string "93101"
    "latitude" : float 34.4266787
    "longitude" : float -119.7111968
    "stars" : int 5
    "review_count" : int 7
    "is_open" : int 0
  ▼ "attributes" : { 1 item
      "ByAppointmentOnly" : string "True"
    }
    "categories" :
    string "Doctors, Traditional Chinese Medicine, Naturopathic/Holistic, Acupuncture, Health & Medical,
    Nutritionists"
    "hours" : NULL
}
```

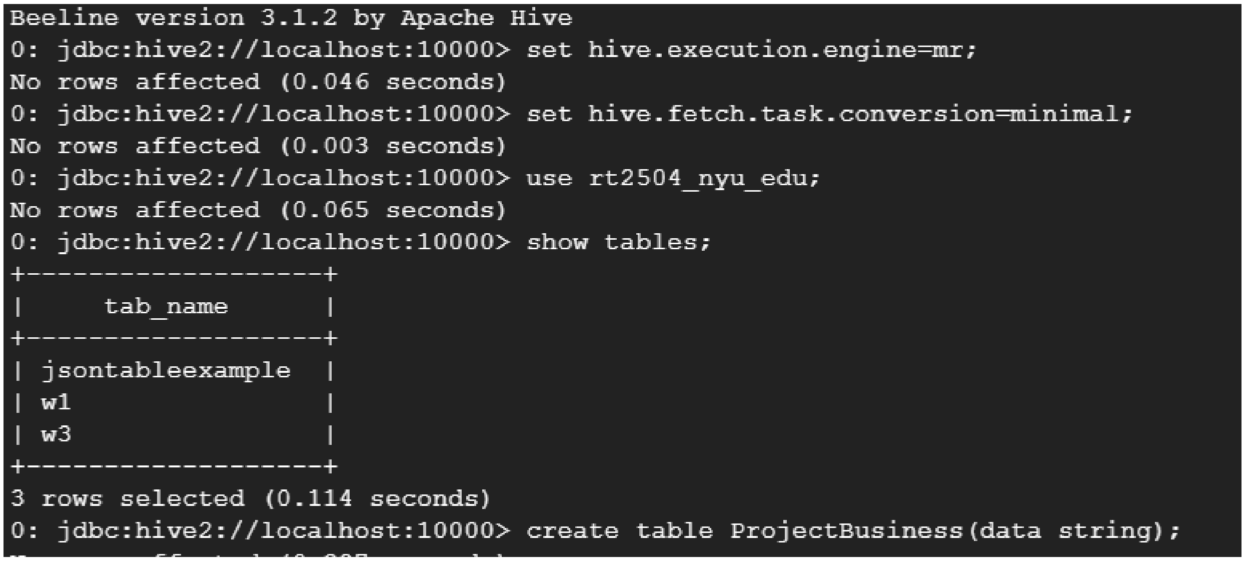
The data is relatively clean, most properties are useful, except “is_open” which is only gives correct information when the api is called, so it is wrong for a saved dataset, and since we will not be using any time related analysis, we can ignore that and the “hours” attributes.

Business_id will be useful to join with other tables, and the localization data can help us with a heatmap or geographical related analysis. The main rating attributes will be the reviews, stars, and we can use categories to classify them.

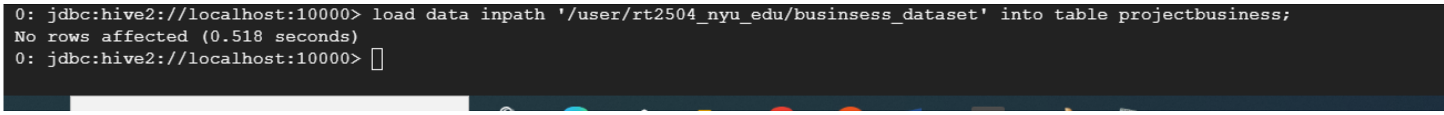
There are many special properties in “Attributes”, but those depend on the business, so the picture above is mainly what is shared by all businesses, the actual data may have a lot more, but we won’t be using them.

We can put the data in hive to be able to easily query and analyze our data later on.

PUTTING IN HIVE:

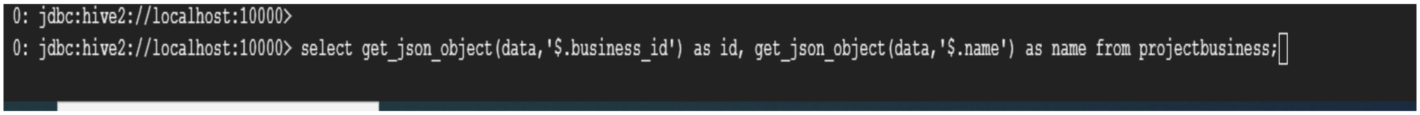


```
Beeline version 3.1.2 by Apache Hive
0: jdbc:hive2://localhost:10000> set hive.execution.engine=mr;
No rows affected (0.046 seconds)
0: jdbc:hive2://localhost:10000> set hive.fetch.task.conversion=minimal;
No rows affected (0.003 seconds)
0: jdbc:hive2://localhost:10000> use rt2504_nyu_edu;
No rows affected (0.065 seconds)
0: jdbc:hive2://localhost:10000> show tables;
+-------------------+
|     tab_name      |
+-------------------+
| jsontableexample  |
| w1                |
| w3                |
+-------------------+
3 rows selected (0.114 seconds)
0: jdbc:hive2://localhost:10000> create table ProjectBusiness(data string);
```



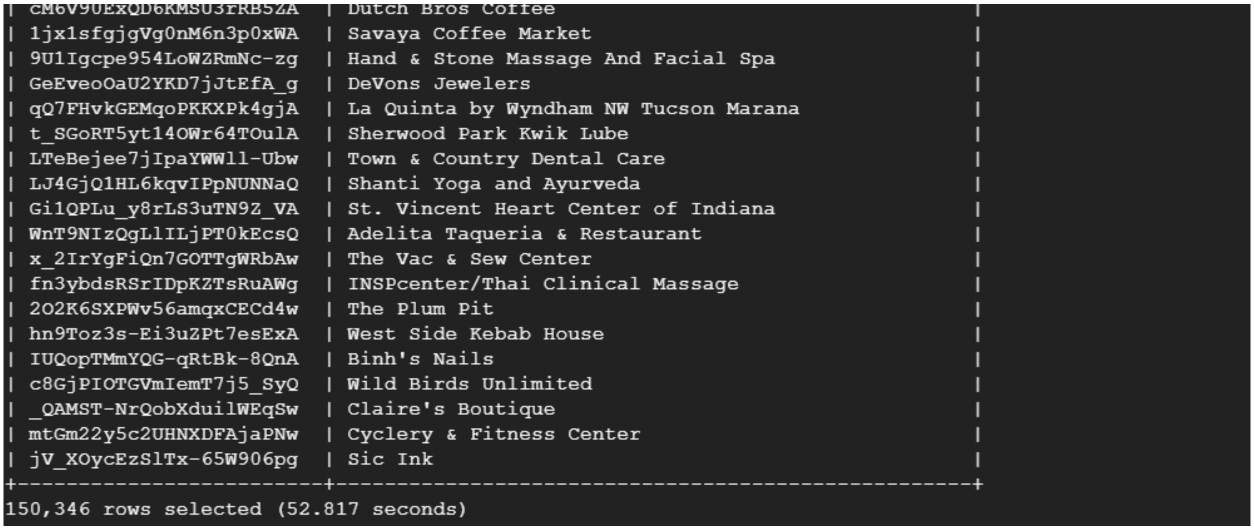
```
0: jdbc:hive2://localhost:10000> load data inpath '/user/rt2504_nyu_edu/businsess_dataset' into table projectbusiness;
No rows affected (0.518 seconds)
0: jdbc:hive2://localhost:10000> 
```

Now that we have it in hive, we can select only the columns we want as such:



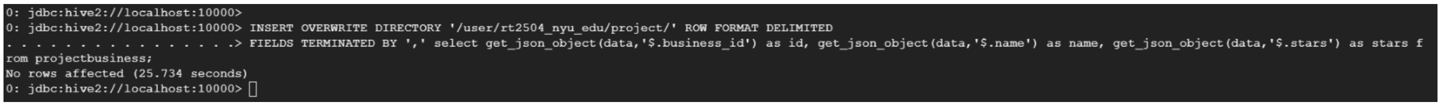
```
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> select get_json_object(data,'$.business_id') as id, get_json_object(data,'$.name') as name from projectbusiness;
```

For example here we can get business_id and name



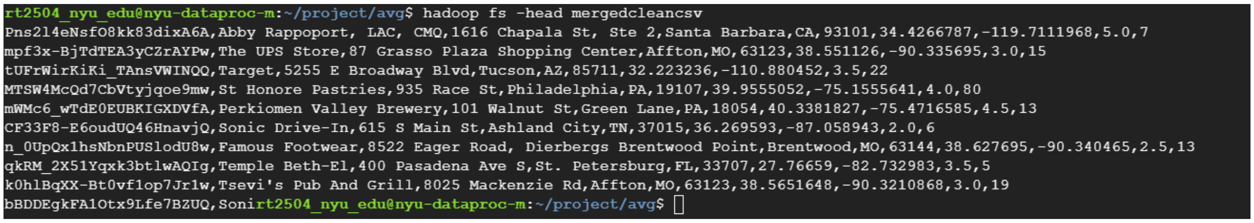
```
| cM6V90ExQD6KMSU3rRB5ZA  | Dutch Bros Coffee                                  |
| 1jx1sfgjgVg0nM6n3p0xWA  | Savaya Coffee Market                               |
| 9U1Igcpe954LoWZRmNc-zg  | Hand & Stone Massage And Facial Spa                |
| GeEveoOaU2YKD7jJtEfA_g  | DeVons Jewelers                                    |
| qQ7FHvkGEMqoPKKXPk4gjA  | La Quinta by Wyndham NW Tucson Marana              |
| t_SGoRT5yt14OWr64TOulA  | Sherwood Park Kwik Lube                            |
| LTeBejee7jIpaYWWll-Ubw  | Town & Country Dental Care                         |
| LJ4GjQ1HL6kqvIPpNUNNaQ  | Shanti Yoga and Ayurveda                           |
| Gi1QPLu_y8rLS3uTN9Z_VA  | St. Vincent Heart Center of Indiana                |
| WnT9NIzQgLlILjPT0kEcsQ  | Adelita Taqueria & Restaurant                      |
| x_2IrYgFiQn7GOTTgWRbAw  | The Vac & Sew Center                               |
| fn3ybdsRSrIDpKZTsRuAWg  | INSPcenter/Thai Clinical Massage                   |
| 2O2K6SXPWv56amqxCECd4w  | The Plum Pit                                       |
| hn9Toz3s-Ei3uZPt7esExA  | West Side Kebab House                              |
| IUQopTMmYQG-qRtBk-8QnA  | Binh's Nails                                       |
| c8GjPIOTGVmIemT7j5_SyQ  | Wild Birds Unlimited                               |
| _QAMST-NrQobXduilWEqSw  | Claire's Boutique                                  |
| mtGm22y5c2UHNXDFAjaPNw  | Cyclery & Fitness Center                           |
| jV_XOycEzSlTx-65W906pg  | Sic Ink                                            |
+-------------------------+----------------------------------------------------+
150,346 rows selected (52.817 seconds)
```

We can also create subdata with only desired columns using



```
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> INSERT OVERWRITE DIRECTORY '/user/rt2504_nyu_edu/project/' ROW FORMAT DELIMITED
. . . . . . . . . . . . . . . .> FIELDS TERMINATED BY ',' select get_json_object(data,'$.business_id') as id, get_json_object(data,'$.name') as name, get_json_object(data,'$.stars') as stars f
rom projectbusiness;
No rows affected (25.734 seconds)
0: jdbc:hive2://localhost:10000> 
```

We can make a csv dataset, and then use mapreduce for data profiling, The data looks like this now:



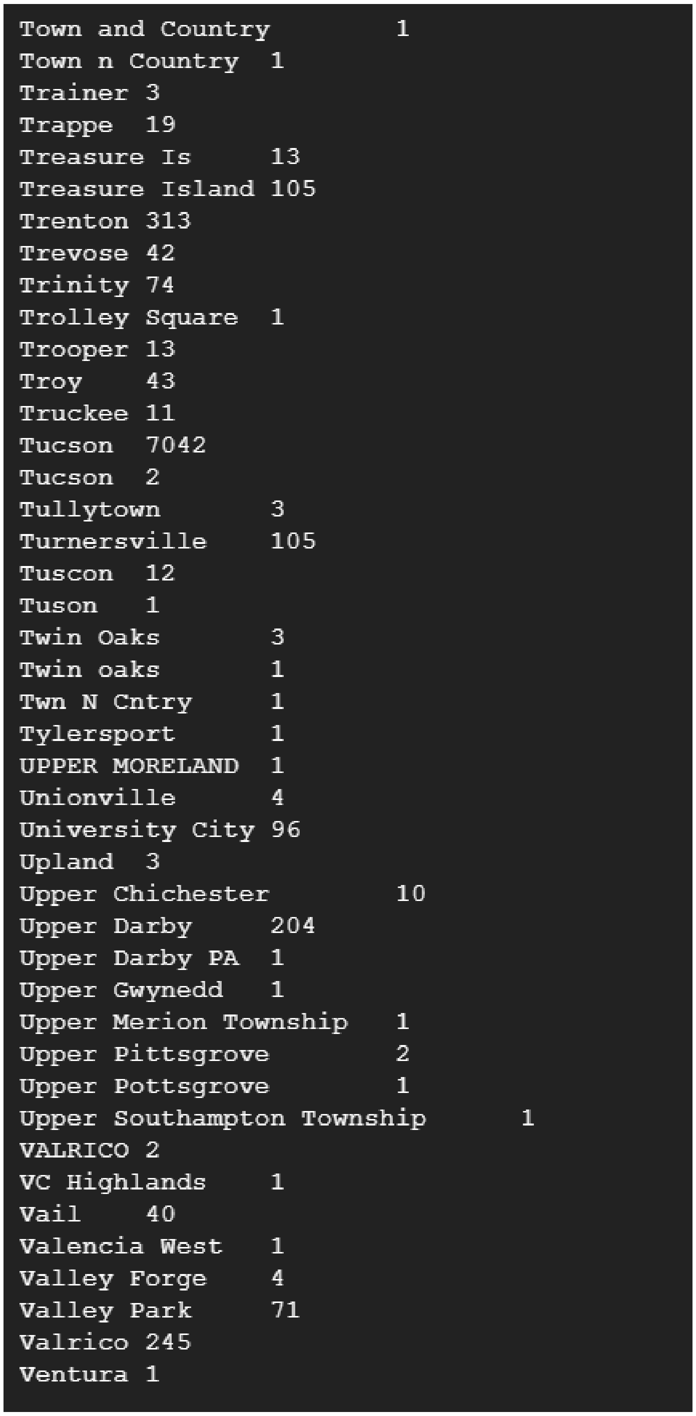
```
rt2504_nyu_edu@nyu-dataproc-m:~/project/avg$ hadoop fs -head mergedcleancsv
Pns2l4eNsfO8kk83dixA6A,Abby Rappoport, LAC, CMQ,1616 Chapala St, Ste 2,Santa Barbara,CA,93101,34.4266787,-119.7111968,5.0,7
mpf3x-BjTdTEA3yCZrAYPw,The UPS Store,87 Grasso Plaza Shopping Center,Affton,MO,63123,38.551126,-90.335695,3.0,15
tUFrWirKiKi_TAnsVWINQQ,Target,5255 E Broadway Blvd,Tucson,AZ,85711,32.223236,-110.880452,3.5,22
MTSW4McQd7CbVtyjqoe9mw,St Honore Pastries,935 Race St,Philadelphia,PA,19107,39.9555052,-75.1555641,4.0,80
mWMc6_wTdE0EUBKIGXDVfA,Perkiomen Valley Brewery,101 Walnut St,Green Lane,PA,18054,40.3381827,-75.4716585,4.5,13
CF33F8-E6oudUQ46HnavjQ,Sonic Drive-In,615 S Main St,Ashland City,TN,37015,36.269593,-87.058943,2.0,6
n_0UpQx1hsNbnPUSlodU8w,Famous Footwear,8522 Eager Road, Dierbergs Brentwood Point,Brentwood,MO,63144,38.627695,-90.340465,2.5,13
qkRM_2X51Yqxk3btlwAQIg,Temple Beth-El,400 Pasadena Ave S,St. Petersburg,FL,33707,27.76659,-82.732983,3.5,5
k0hlBqXX-Bt0vflop7Jr1w,Tsevi's Pub And Grill,8025 Mackenzie Rd,Affton,MO,63123,38.5651648,-90.3210868,3.0,19
bBDDEgkFA1Otx9Lfe7BZUQ,Sonirt2504_nyu_edu@nyu-dataproc-m:~/project/avg$ 
```

Data Profiling:

Because some business have low reviews, the min stars are 0 and the max stars is 5, so these simple statistics aren't very meaningful; let's instead look at profiling location data.

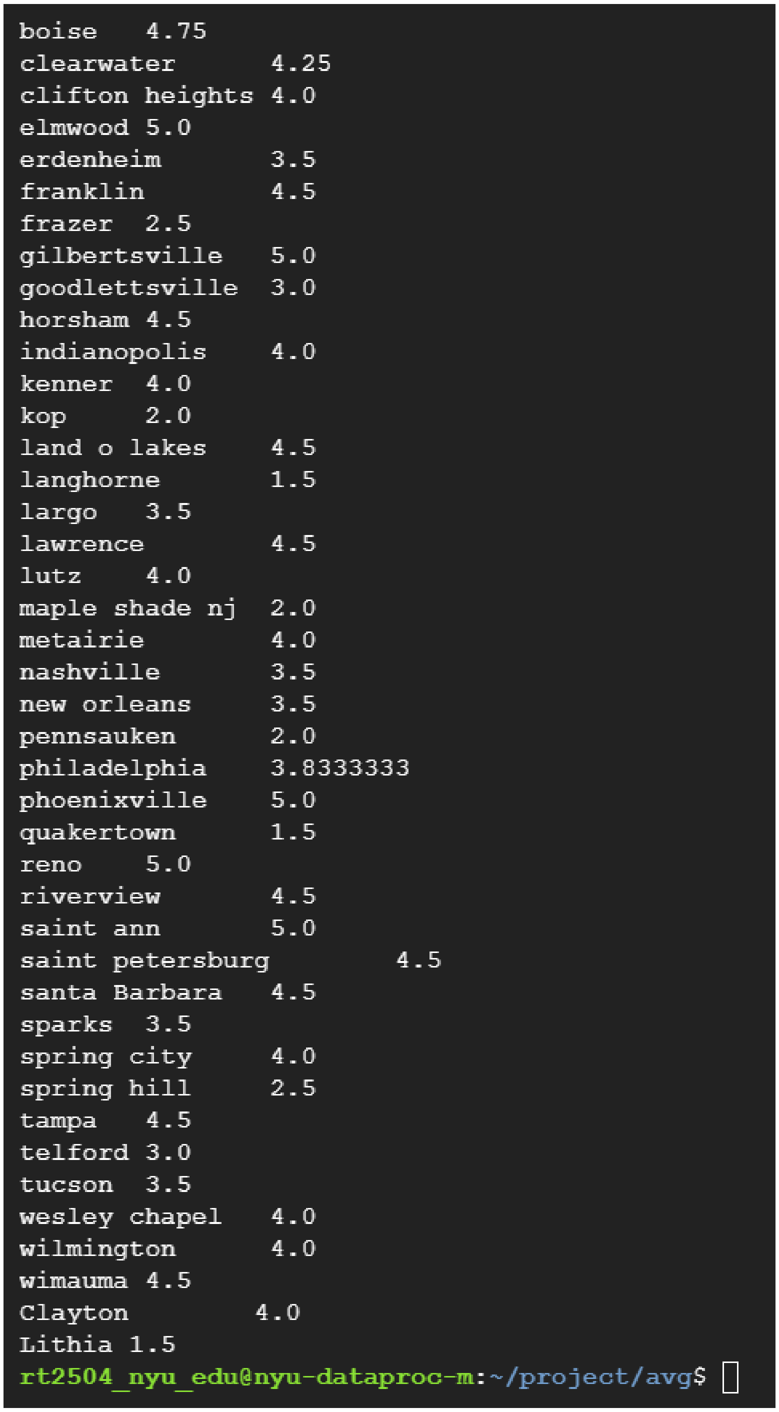
Based on the data format, we can use mapreduce to profile things such as:

Number of (yelp) businesses in each city, using the Count mapreduce files:



```
Town and Country        1
Town n Country  1
Trainer 3
Trappe  19
Treasure Is     13
Treasure Island 105
Trenton 313
Trevose 42
Trinity 74
Trolley Square  1
Trooper 13
Troy    43
Truckee 11
Tucson  7042
Tucson  2
Tullytown       3
Turnersville    105
Tuscon  12
Tuson   1
Twin Oaks       3
Twin oaks       1
Twn N Cntry     1
Tylersport      1
UPPER MORELAND  1
Unionville      4
University City 96
Upland  3
Upper Chichester        10
Upper Darby     204
Upper Darby PA  1
Upper Gwynedd   1
Upper Merion Township   1
Upper Pittsgrove        2
Upper Pottsgrove        1
Upper Southampton Township      1
VALRICO 2
VC Highlands    1
Vail    40
Valencia West   1
Valley Forge    4
Valley Park     71
Valrico 245
Ventura 1
```

-Average stars per city in the dataset, using the average mapreduce files.



```
boise   4.75
clearwater      4.25
clifton heights 4.0
elmwood 5.0
erdenheim       3.5
franklin        4.5
frazer  2.5
gilbertsville   5.0
goodlettsville  3.0
horsham 4.5
indianopolis    4.0
kenner  4.0
kop     2.0
land o lakes    4.5
langhorne       1.5
largo   3.5
lawrence        4.5
lutz    4.0
maple shade nj  2.0
metairie        4.0
nashville       3.5
new orleans     3.5
pennsauken      2.0
philadelphia    3.8333333
phoenixville    5.0
quakertown      1.5
reno    5.0
riverview       4.5
saint ann       5.0
saint petersburg        4.5
santa Barbara   4.5
sparks  3.5
spring city     4.0
spring hill     2.5
tampa   4.5
telford 3.0
tucson  3.5
wesley chapel   4.0
wilmington      4.0
wimauma 4.5
Clayton        4.0
Lithia 1.5
rt2504_nyu_edu@nyu-dataproc-m:~/project/avg$ 
```