

RBDA Project Data Ingestion Report

Reuben Abraham (rca9751)

About the Data

- The dataset I'm working with is the tips dataset in *tip.json* - this dataset contains tips written by users for a business. Tips are short reviews and contain quick suggestions for the business.
- The original dataset is a json flat file with the following fields:

```
{  
    Text: content of the tip,  
    Date: date the tip was posted,  
    Compliment_Count: How many compliments/upvotes it has,  
    Business_ID: Links to the business.json dataset,  
    User_ID: Links to the user.json  
}
```

- Here is a snippet of the dataset.

```
reubenabraham@Reubens-Air: ~/Documents/NYUDocuments/RBDA/Project head yelp_a  
cademic_dataset_tip.json  
{  
  "user_id": "AGNUGVwnZUey3gcPCJ76iw", "business_id": "3uLqwr0qeCNMjKenHJwPGQ", "text": "Avengers time with the ladies.", "date": "2012-05-18 02:17:21", "compliment_count": 0  
}  
{  
  "user_id": "NBNA4MgHP9D3cw--SnaUTkA", "business_id": "QoezRbYQncpRqyrLH6Iqjg", "text": "They have lots of good deserts and tasty cuban sandwiches", "date": "2013-02-05 18:35:10", "compliment_count": 0  
}  
{  
  "user_id": "-cop0vldyKh1qr-vzkDEvw", "business_id": "MYoRNLb5chwjQe3c_k37Gg", "text": "It's open even when you think it isn't", "date": "2013-08-18 00:56:08", "compliment_count": 0  
}  
{  
  "user_id": "FjMQVZjSgY8syIO-53KFKw", "business_id": "hV-bABTK-glhwj3lpsJw", "text": "Very decent fried chicken", "date": "2017-06-27 23:05:38", "compliment_count": 0  
}  
{  
  "user_id": "ld0AperBXk1h6UbqmM80zw", "business_id": "uN00udeJ3Z1_tf6nxg5ww", "text": "Appetizers.. platter special for lunch", "date": "2012-10-06 19:43:09", "compliment_count": 0  
}  
{  
  "user_id": "trf3Qcz8qvCDKXiTgjUcEg", "business_id": "7Rm9Ba50bw23KTA8RedZYg", "text": "Chili Cup + Single Cheeseburger with onion, pickle, and relish + Vanilla Co ca-Cola... so far.", "date": "2012-03-13 04:00:52", "compliment_count": 0  
}  
{  
  "user_id": "SMGA1RjyfuYu-c-22ziY0g", "business_id": "kH-0iXqkL7b8UXNpguBMKg", "text": "Saturday, Dec 7th 2013, ride Patco's Silver Sleigh w/ Santa & his elves on a decorated train into Center City. Trains leave from Lindenwold at 10am, 11:15am, & 12:30pm, and make all stops. Great for kids!", "date": "2013-12-03 23:42:15", "compliment_count": 0  
}  
{  
  "user_id": "YVBB9g23nuVJ0u44zK0pSA", "business_id": "jtri188kuhe_AuE0J51U_A", "text": "This is probably the best place in the cool Springs area to watch a game and eat", "date": "2016-11-22 22:14:58", "compliment_count": 0  
}  
{  
  "user_id": "VL12EHdT40WqSg0nIqkzw", "business_id": "x0DBZmX4Em1VvbqtKN7YKq", "text": "Tacos", "date": "2012-07-27 01:48:24", "compliment_count": 0  
}  
{  
  "user_id": "4ay-fdVks5WMerYL_htkG0", "business_id": "pICJRcyqW1cF96Q3XhLSbw", "text": "Starbucks substitute in boring downtown Tampa. Ugh. Never again!", "date": "2012-06-09 22:57:04", "compliment_count": 0  
}
```

- Here are some general stats about the data:
 - There are 908,915 rows and 5 columns if the data is converted into tabular form
 - The dataset has **no** missing data in any fields across the whole dataset.
 - 301,758 unique users left comments in this dataset
 - user_id : "fCvMnJU1Z-XhAjKg99wK3Q" left 4071 tips - about 4 times the next highest which was 1385 tips
 - tips were left for 106193 unique businesses and 2571 tips were left for business_id 'FEXhWNCMkv22qG04E83Qjg' which was the highest- this was more than double of even the next highest at 1011

- In general tips are disproportionately left by a select few people, for a select few businesses. The majority of the people leaving tips are one-time/few-time tippers.

Data Cleaning

- The data was already very clean to begin with - the map-reduce job I wrote ingests the data, parses the json , cleans the text field to retain only alpha-numeric characters, and ensures a standard date format. There's not much else to do with regard to cleaning.
- The cleaned data is written by the reducer in a comma-separated format. A Snippet is shown below.

```
File Output Format Counters
Bytes Written=117195066
rca9751_nyu_edu@nyu-dataproc-m:~/project$ hadoop fs -cat project/output/part-r-00000 | head
4tF1CWdMxvvpwUIgGsDygA, cb1Vg1NIWry8UA0jyuXnQ,Food is good value but a bit hot,2021-12-07 22:30:00,0
ckqKGM2h17I9Chp5IpAhkw,s2eyoTuJrcP7I_XyjdhUHQ,Great pizza great price,2021-11-20 16:11:44,0
v48Spe6WEpgehsF2xQADpg,hYnMeAO77RGyTtIzUSKYzQ,Love their Cubans,2021-11-05 13:18:56,0
luxtQAuJ2T5Xwa_wp7kUnA,OaGf0Dp56ARhQwIDT90w_g,Great food and service,2021-10-30 11:54:36,0
eYodOTF8pkqKPzHkcxZs-Q,3lHTewuKFt5IImbXJoFeDQ,Disappointed in one of your managers,2021-09-11 19:18:57,0
FowxkbAixI3hlREeCgIa_Q,kfNv-JZpuN6TVNSO6hHdkw,Great experience with a phenomenal food with a lot of flavor and a affordable price,2021-05-09 23:21:10,0
5hJR71jJbhFgOaLi8iz5pQ,AXC_4yZrn-N3BT7-2bV_Q,BOMB food Super delicious great outdoor space and kid friendly,2021-05-04 21:44:53,0
2-vAo2Ufkd7QHA5TG8kwmg,wQUBiBqlzC6cbdKX-GaBqQ,The food was delicious,2021-03-12 00:15:07,0
Apfz4xUeBOObgXkOhOnMxQ,90pJu2O7fIEm_N3lFyue7A,Great food cocktail ambience and service,2021-02-14 15:02:38,0
cXfyVy34hqDgygyJBtme4w,peomsQ8wg84wKlJ2RjZnJQ,This place has gone to shit Food fucking sucks Dont eat here,2021-01-01 00:03:19,0
cat: Unable to write to output stream.
rca9751_nyu_edu@nyu-dataproc-m:~/project$
```

- **All code will be available in the zip file.**
- I will later move this data into a Hive table, so it can be used for further processing and combining with other datasets.