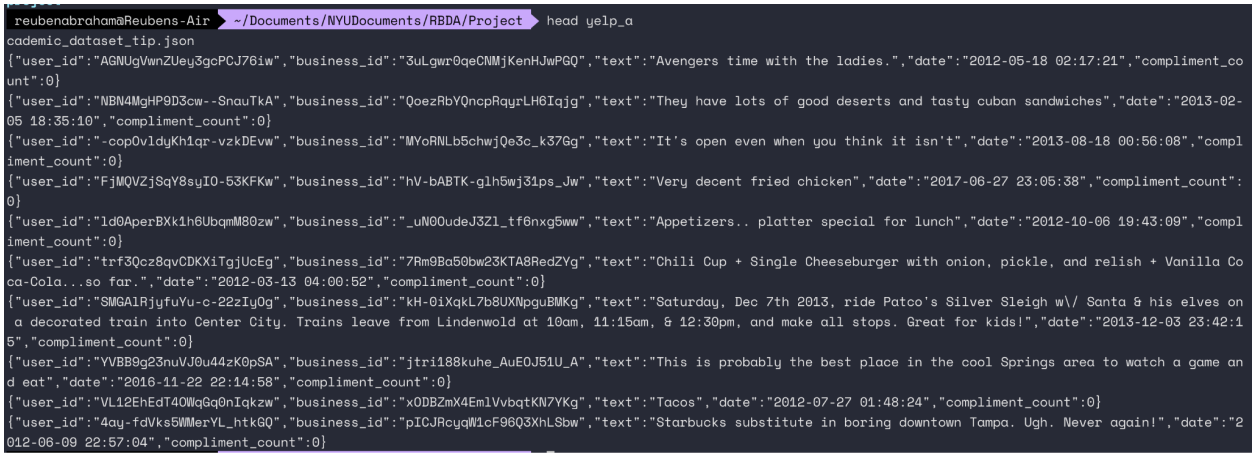
RBDA Project Data Ingestion Report
Reuben Abraham (rca9751)

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.

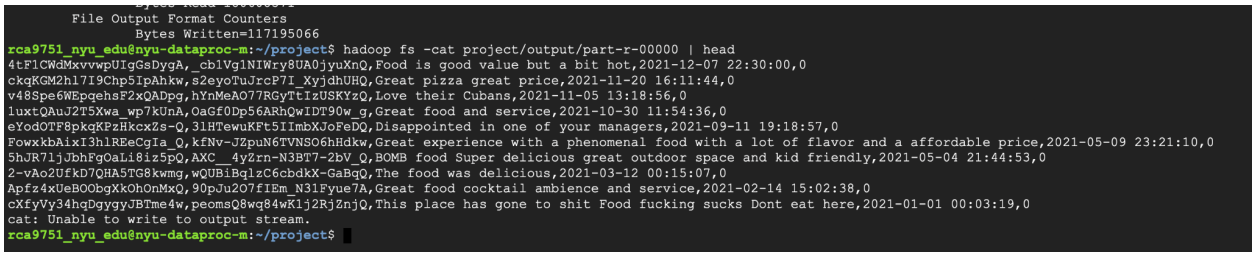


- Here are some general stats about the data:
  - There are 908,915 rows and 5 columns if the data is converted into tabular form
  - The dataset has **no** missing data in any fields across the whole dataset.
  - 301,758 unique users left comments in this dataset
  - user_id : ‘‘fCvMnJU1Z-XhAjKg99wK3Q’’ 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.



- **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.