Data Ingestion Report

Project - Yelp Dataset ETL Analysis
Team Member - Ajeeta Asthana (aa9381)
Group Number - 12

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

Data size: 3.36 GB

Existing Schema


```
{"user_id":"AUi8MPWJ0mLkMfwbui27lg",
 "name":"John",
 "review_count":109,
 "yelping_since":"2010-01-07 18:32:04",
 "useful":154,
 "funny":20,
 "cool":23,
 "elite":"",
 "friends":"gy5fWeSv3Gamuq9Ox4MV4g, lMr3LWU6kPFLTmCpDkACxg,
_5280JO3WzgiVTJUYahCcg, tCqYnhAdQhPO3JAAnc09ig, t903_es-gp3abvdrIQutQA,
Y34Qb5SqxH1kvbyQDKoLFg, 2Dr-IrL_eDjVpr7XYtdLig, cGbvA6fyzh8nxNvTLgmCSg,
hLvInByM9UGS13NEaWnH8g, mC-nfEt4NguxtpFCGfa_Dw, dF8vnPMa3lrwEImjb5CqTw,
CWcUVbQYdmHIn-adDejQpw, zXLn1jPx2DhbSlvsf84cAg, soV1SJe-uq02gs6D6Gl06w,
TiFSHq1fMo0Sq22Yh7RGkw, AvLWKARqvnVsw0oTRiApDg, a4-HCawx-wjSKY1nMLGRpg,
rkqsvHX8ai065o0iqyUHdg, GqzoKYXv6F1O85bhywyoSA, 6XOzQvb_MEZyL0wi_HYsag,
Kc4-kG_UMsCqugSkOGsZ-g, XxY5_KAtNjzf_OyQpAKxtQ, vhNo5fA1bx2NTg_pbafWdw,
cLePLju2DvpZDw6PNDbMOg, NbywhxCnZYuzY3UX9g1vuw, tyL6q25sDONJqqsr9bm2_A,
lwSLBEWTV9B-iQuPuknfAg, RGiLXCIUMCaU2HZcYtdpFA, gWcad62XMJC1FY4ybqsKhg,
yQ4LY_I_0IIP7mkftKf3rg, QAkbXGDZCHAVfgJK0ZrpDw, aUWk-epQVLanHGzL4bmj-Q,
VAZHcCMDx-7ypihJg0g4VQ, jy0v-qoPfc4-fPn3APXwJA, g9Z5MxZ0rw9lAOKx1WfGRQ,
Pox6gRJ5vm7m9DO4nuU-UA, UMHC0kzV6TS8O6HLb5mpYw, _byETu62m9YAJHG1yY837w,
qo_qrGNv5Gjejye-4d-u2w, GEMQDwvvCcMxB9hTDfWElg, ywOCgWH0FRG7Hw-Urb9hSw,
zlQizsViK9gPfxIW85ICBA, MekKt8a9XLFLYF6s2pV_cw, -pgAIxEyUu5rspr93sYskg,
Hx6Aaw2aOhr-3xSKe7scng, 0n3zQFQZgfdpjG4drF812g, YogWNBUTdnLKB9WfiLYZPw,
DLU5G_7fG6VK6ZH8oVJSZQ, 8EXDnMKdEEXbU2AXp3y5-A, Tj8SuJiN4h9WZNs0f2eypg,
gbx3IYFKB-36HhV5wM3G1w, rcU7ysY41qGppbw4pQgjqg, 69a9Aygpj-fW3KxnREpaAw,
xazq3kc_pbjo7z6pZUwPgw, 5It1pYl-OcpW2xFlZS5-dw, XHrKXZQXWdMFEKrlxdarVw,
dccfgKtgbk_WLaC0Mlfjyw, dPzJCGsLOLNozrJ01UnTXQ, PadmV2GEoA6mWpQUpPh7Ig,
whKV8TxcoT7NR5jhsCjNKg, beCVDc6aud9eLsP3PZ90ZQ, 2PqFngvKvApLmX5OdxCLyw,
oNd5bP0MiB-wBkKZejxOag, piTn3h6zP0ZPLHD3L63VWQ",
 "fans":4,
 "average_stars":3.4,
```

```
"compliment_hot":0,
"compliment_more":0,
"compliment_profile":0,
"compliment_cute":0,
"compliment_list":0,
"compliment_note":1,
"compliment_plain":6,
"compliment_cool":3,
"compliment_funny":3,
"compliment_writer":0,
"Compliment_photos":0}
```


Columns interested : user_id, name, review_count, average_stars

Steps:

1: Downloaded the yelp dataset(json) in hadoop using kaggle package

- pip install kaggle
- vi
- vi ~/.kaggle/kaggle.json
- pwd
- ls ~/
- ls -a ~/
- mkdir ~/.kaggle
- vi ~/.kaggle/kaggle.json
- clear
- chmod 600 ~/.kaggle/kaggle.json
- kaggle datasets list
- which kaggle
- pip uninstall kaggle
- ~/.local/bin/kaggle datasets list
- ~/.local/bin/kaggle datasets download yelp-dataset/yelp-dataset
- ls
- unzip yelp-dataset.zip
- ls

2. Wrote a UserDataTuple class, to define columns we are interested
3. Wrote a Mapper class to get required 4 fields out of all the fields in the original dataset.
4. Added Counter for Missing data fields.
5. Future analysis to aggregate users with most reviews, and their average star rating.

Cleaned Schema

```
aa9381_nyu_edu@nyu-dataproc-m:~$ hadoop fs -cat project/output/part-m-00000 | head
mIShuuzz7qH0TDmYM0EOcg  UserDataTuple{name='Pet', review_count='26', average_stars='4.93'}
uwdId9Nn44m2nxZ_2-EzFg  UserDataTuple{name='Julie', review_count='1', average_stars='5.0'}
2url v5gVWCKlHRVl7FOERA  UserDataTuple{name='Laura', review_count='1', average_stars='5.0'}
UelmhHQqO0j_V8dN96UVMw  UserDataTuple{name='Emma', review_count='1', average_stars='5.0'}
2Qs9PUgkLkufo_NUE87nqA  UserDataTuple{name='Tony', review_count='1', average_stars='5.0'}
EsKhwXkBueAJha5atFkUDQ  UserDataTuple{name='Tony', review_count='4', average_stars='4.83'}
F3-lXH7DKTEeAHXnLJOBuA  UserDataTuple{name='Sierra', review_count='6', average_stars='2.33'}
xTbq--M19hh8Az7eI5clNQ  UserDataTuple{name='Tania', review_count='17', average_stars='4.1'}
fCl1ly8V6R_ZipOaGhKenw  UserDataTuple{name='Emily', review_count='11', average_stars='4.18'}
qTypJvwbisolbMxIfeAkFg  UserDataTuple{name='Ryan', review_count='3', average_stars='2.33'}
cat: Unable to write to output stream.
aa9381_nyu_edu@nyu-dataproc-m:~$
```

```
aa9381_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls project
Found 2 items
drwxr-xr-x   - aa9381_nyu_edu aa9381_nyu_edu          0 2023-04-16 17:42 project/output
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu 3363329011 2023-04-16 15:35 project/yelp_academic_dataset_user.json
aa9381_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls project/output
Found 26 items
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu          0 2023-04-16 17:42 project/output/_SUCCESS
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu   14337094 2023-04-16 17:42 project/output/part-m-00000
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    2718010 2023-04-16 17:42 project/output/part-m-00001
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    3527108 2023-04-16 17:42 project/output/part-m-00002
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    9079683 2023-04-16 17:42 project/output/part-m-00003
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    5070801 2023-04-16 17:42 project/output/part-m-00004
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    2499859 2023-04-16 17:42 project/output/part-m-00005
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu   10972809 2023-04-16 17:42 project/output/part-m-00006
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    5697752 2023-04-16 17:42 project/output/part-m-00007
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    4107334 2023-04-16 17:42 project/output/part-m-00008
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu   11961431 2023-04-16 17:42 project/output/part-m-00009
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    3298015 2023-04-16 17:42 project/output/part-m-00010
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    9288036 2023-04-16 17:42 project/output/part-m-00011
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    6985166 2023-04-16 17:42 project/output/part-m-00012
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    4530669 2023-04-16 17:42 project/output/part-m-00013
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu   12604371 2023-04-16 17:42 project/output/part-m-00014
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    3619136 2023-04-16 17:42 project/output/part-m-00015
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu   13175150 2023-04-16 17:42 project/output/part-m-00016
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    3399283 2023-04-16 17:42 project/output/part-m-00017
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    7161348 2023-04-16 17:42 project/output/part-m-00018
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    9756045 2023-04-16 17:42 project/output/part-m-00019
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    4382299 2023-04-16 17:42 project/output/part-m-00020
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu   13254837 2023-04-16 17:42 project/output/part-m-00021
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    3029592 2023-04-16 17:42 project/output/part-m-00022
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu   13687796 2023-04-16 17:42 project/output/part-m-00023
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu    3196001 2023-04-16 17:42 project/output/part-m-00024
aa9381_nyu_edu@nyu-dataproc-m:~$
```

```
aa9381_nyu_edu@nyu-dataproc-m:~/rbdaproject$ ls -l
total 12
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1044 Apr 20 15:14 Clean.java
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1778 Apr 20 15:15 UserCleanMapper.java
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1433 Apr 20 15:15 UserDataTuple.java
aa9381_nyu_edu@nyu-dataproc-m:~/rbdaproject$ javac -classpath `hadoop classpath` *.java
aa9381_nyu_edu@nyu-dataproc-m:~/rbdaproject$ jar cvf userData.jar *.class
added manifest
adding: ANOMALY.class(in = 723) (out= 439)(deflated 39%)
adding: Clean.class(in = 1377) (out= 796)(deflated 42%)
adding: UserCleanMapper.class(in = 2625) (out= 1180)(deflated 55%)
adding: UserDataTuple.class(in = 1523) (out= 733)(deflated 51%)
aa9381_nyu_edu@nyu-dataproc-m:~/rbdaproject$ ls -l
total 32
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu  723 Apr 20  2023 ANOMALY.class
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1377 Apr 20  2023 Clean.class
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1044 Apr 20 15:14 Clean.java
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 2625 Apr 20  2023 UserCleanMapper.class
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1778 Apr 20 15:15 UserCleanMapper.java
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1523 Apr 20  2023 UserDataTuple.class
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 1433 Apr 20 15:15 UserDataTuple.java
-rw-r--r-- 1 aa9381_nyu_edu aa9381_nyu_edu 3976 Apr 20  2023 userData.jar
aa9381_nyu_edu@nyu-dataproc-m:~/rbdaproject$ cd ..
```

```
aa9381_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls project
Found 1 items
-rw-r--r--   1 aa9381_nyu_edu aa9381_nyu_edu 3363329011 2023-04-16 15:35 project/yelp_academic_dataset_user.json
aa9381_nyu_edu@nyu-dataproc-m:~$ hadoop jar userData.jar Clean project/yelp_academic_dataset_user.json project/output
JAR does not exist or is not a normal file: /home/aa9381_nyu_edu/userData.jar
aa9381_nyu_edu@nyu-dataproc-m:~$ hadoop jar rbdaproject/userData.jar Clean project/yelp_academic_dataset_user.json project/output
2023-04-20 15:23:55,141 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.24:8032
2023-04-20 15:23:55,330 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.24:10200
2023-04-20 15:23:55,500 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-04-20 15:23:55,517 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/aa9381_nyu_edu/.staging/job_1679692348553_19657
2023-04-20 15:23:55,754 INFO input.FileInputFormat: Total input files to process : 1
2023-04-20 15:23:55,838 INFO mapreduce.JobSubmitter: number of splits:25
2023-04-20 15:23:56,017 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1679692348553_19657
2023-04-20 15:23:56,019 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-20 15:23:56,200 INFO conf.Configuration: resource-types.xml not found
2023-04-20 15:23:56,200 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-20 15:23:56,415 INFO impl.YarnClientImpl: Submitted application application_1679692348553_19657
2023-04-20 15:23:56,453 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1679692348553_19657/
2023-04-20 15:23:56,454 INFO mapreduce.Job: Running job: job_1679692348553_19657
2023-04-20 15:24:03,546 INFO mapreduce.Job: Job job_1679692348553_19657 running in uber mode : false
2023-04-20 15:24:03,547 INFO mapreduce.Job:  map 0% reduce 0%
2023-04-20 15:24:18,664 INFO mapreduce.Job:  map 12% reduce 0%
2023-04-20 15:24:19,672 INFO mapreduce.Job:  map 48% reduce 0%
2023-04-20 15:24:21,683 INFO mapreduce.Job:  map 96% reduce 0%
2023-04-20 15:24:22,689 INFO mapreduce.Job:  map 100% reduce 0%
2023-04-20 15:24:23,702 INFO mapreduce.Job: Job job_1679692348553_19657 completed successfully
2023-04-20 15:24:23,787 INFO mapreduce.Job: Counters: 34
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=6144794
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3363455541
                HDFS: Number of bytes written=181339625
                HDFS: Number of read operations=175
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=75
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
```

```
  650  hadoop fs -ls
  651  hadoop fs -ls project
  652  hadoop fs -rm  -r project/output
  653  hadoop fs -ls
  654  hadoop fs -ls project
  655  pwd
  656  cd rbdaproject
  657  ls -l
  658  rm -r *.class
  659  rm -r maxTemp.jar
  660  ls -l
  661  javac -classpath `hadoop classpath` *.java
  662  jar cvf userData.jar *.class
  663  ls -l
  664  cd ..
  665  hadoop fs -ls project
  666  hadoop jar userData.jar Clean project/yelp_academic_dataset_user.json project/output
  667  hadoop jar rbdaproject/userData.jar Clean project/yelp_academic_dataset_user.json project/output
  668  hadoop fs -ls project
  669  hadoop fs -ls project/ouput
  670  hadoop fs -ls project/output
  671  hadoop fs -cat project/output/part-m-00000
  672  B
  673  clear
  674  hadoop fs project/output/part-m-00000 -head
  675  hadoop fs -cat project/output/part-m-00000 | head
  676  pwd
  677  hadoop fs -ls project
  678  hadoop fs -cat project/yelp_academic_dataset_user | head
  679  hadoop fs -cat project/yelp_academic_dataset_user.json | head
  680  hostory
  681  history
aa9381_nyu_edu@nyu-dataproc-m:~$
```