***PART 1***

a) **Final\_1a.py**

3 tables created –

Twitter\_data

User\_data

Geo\_data

b) **Final\_1b.py**

Time to save tweets from web URL to a file:

*#loadTweets for part1-b took* ***3134.432579278946*** *seconds. ~87 minutes*

*#500000 tweets or rows*

c) **Final\_1c.py**

Time to load tweets from web URL to the database:

*#loadTweets took* ***5829.0454025268555*** *seconds - 90+ minutes.*

No of rows in each of the tables:

*#Geo table - 11787*

*#Twitter table - 499776*

*#User table – 447304*

d) **Final\_1d.py**

Time to load tweets from local file to the database:

*#loadTweets took* ***271.4044873714447*** *seconds.*

*#User\_data - 447304*

*#Twitter\_data - 499776*

*#Geo\_data – 8467*

*Compared to part 1c – loading tweets from locally saved file is much faster.*

e) **Final\_1e.py**

Time to load tweets via batches from local file to the database

*#Part 1e-Loading tweets via batch took* ***215.38358235359192*** *seconds*

*#Batch Count: 1000.0*

*#User\_data-447304*

*#Twitter\_data-499776*

*#Geo\_data-6993*

***PART 2***

a) **Final\_Part2.py**

1. select \* from twitter\_data where id\_str like '%44%' or id\_str like '%77%';

*First run: Query 2-i took 9.844617366790771 seconds.*

*Second run: Query 2-i took 2.8754050731658936 seconds.*

1. select count(distinct in\_reply\_to\_user\_id) from twitter\_data;

*First run: Query 2-ii took 1.9656035900115967 seconds.*

*Second run: Query 2-ii took 1.6858031749725342 seconds.*

1. select \* from twitter\_data where length(text) = (select max(length(text)) from twitter\_data);

*First run: Query 2-iii took 1.5756027698516846 seconds.*

*Second run: Query 2-iii took 1.4362025260925293 seconds.*

1. select u.name,avg(latitude),avg(longitude)

from twitter\_data t, user\_data u,geo\_data g

where t.coordinate\_id = g.coordinate\_id and t.user\_id = u.id

group by u.name;

*First run: Query 2-iv took 4.461607933044434 seconds.*

*Second run: Query 2-iv took 4.277407646179199 seconds.*

1. query-iv run in loop

*Looping 10 times:*

*First run: Query 2-v-a took 41.84127354621887 seconds.*

*Second run: Query 2-v-a took 41.617273807525635 seconds.*

*Looping 100 times:*

*First run: Query 2-v-b took 795.8962574005127 seconds.*

b) **Final\_Part2b.py**

*Trial Run with 100000 rows:*

*i. Part2b-i took 11.51282024383545 seconds.*

*Count of unique values 18433*

*ii. Part 2b-ii took 15.069626569747925 seconds.*

*Python equivalents of the query clearly take more time than SQL queries.*

**c) Final\_2c.py**

***PART 3***

a) **Final\_3a.py**

*#Part 3a took* ***73.54536151885986*** *seconds*

b) **Final\_3b.py**

*#Part 3-b took* ***159.92888593673706*** *seconds.*

*Part 3a i.e. exporting the contents from database and writing it to a local file was faster.*

***PART 4***

a) **Final\_4a.py**

b) **Final\_4b.py**

My system became unresponsive (after ~90 minutes) when I tried running this program with 500000 rows. Please consider my sample output.

*Output: For a local file with 5000 tweets-*

*Known locations- 103*

*Unknown locations- 4895*

c) **Final\_4c.py**