**Steps to ingest the data:**

**Data Source Used:** 2018 Yellow Taxi Trip Data

<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243>

Click on ‘View Data’ then Export CSV

1) Download the CSV from the website using export as CSV

2) Setup hpc tunnel

3) Copy file to dumbo using WinSCP client

4) Put the data in HDFS

Connect through cisco VPN client

ssh as12578@dumbo.hpc.nyu.edu

hdfs dfs -ls /user/as12578

hdfs dfs -mkdir /user/as12578/project

hdfs dfs -put taxi.csv /user/as12578/project

6) Use the dataset stored in hadoop for cleaning and profiling code