**Steps to ingest the data:**

**Data Source Used:** NYPD Complaint Data Current (Year To Date), Historic

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

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

2) Setup hpc tunnel

3) Run from cmd prompt

hpc gateway

4) scp NYPD\_Complaint\_Data\_Current\_\_Year\_To\_Date\_.csv [asb862@dumbo](mailto:asb862@dumbo):projectRBDA

5) Put the data in HDFS

Log into dumbo cluster

ssh [asb862@gw.hpc.nyu.edu](mailto:asb862@gw.hpc.nyu.edu)

ssh [asb862@dumbo.hpc.nyu.edu](mailto:asb862@dumbo.hpc.nyu.edu)

cd projectRDBA

hdfs dfs -put NYPD\_Complaint\_Data\_Current\_\_Year\_To\_Date\_.csv

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