Methods used to ingest the data.

1) Traffic Speed Data

https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq

This dataset is over 12GB and to avoid downloading to a local PC and then to move to the local drive on Dumbo. We copied directly to the local drive on Dumbo using the following command:

curl -O https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq

We named the file ‘inputs.csv’

Then, to copy to HDFS, we used the following command:

hdfs dfs -put inputs.csv sharedData/InputDataSets/trafficSpeed