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:



We named the file ‘inputs.csv’

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



2) weather dataset: JFK

https://www7.ncdc.noaa.gov/CDO/cdopoemain.cmd?datasetabbv=DS3505

Step 1: I downloaded the data from the website of National Climatic Data Center, U.S. Department of Commerce.<https://www.ncdc.noaa.gov/>

Step 2: I uploaded the data to my local file system using FileZilla

Step 3: I put the data into HDFS using the following commands:

hdfs dfs -mkdir /user/yy1420/project

hdfs dfs -put WeatherData\_CentralPark.txt /user/yy1420/project

3) Speed limits:

https://data.cityofnewyork.us/Transportation/VZV\_Speed-Limits/7n5j-865y