The Steps for Data Ingest

1. Login to Dumbo

2. Download the dataset (csv format) through the URL:

curl -O https://data.cityofnewyork.us/api/views/8h9b-rp9u/rows.csv?accessType=DOWNLOAD

3. Change the file name to a shorter one:

mv rows.csv?accessType=DOWNLOAD rows.csv

4. Make a new directory in HDFS:

hdfs dfs -mkdir /user/syc574/project

5. Put the dataset file into HDFS:

hdfs dfs -put rows.csv /user/syc574/project

6. Check if the dataset is added successfully:

hdfs dfs -cat /user/syc574/project/rows.csv