**Steps to ingest the data:**

**Data Source Used:** National Centers for Environmental Information, National Oceanic and Atmospheric Administration Link

<https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/>

Weather data consists of datasets for daily weather summaries of the following cities:

New York (JFK)

Chicago (ORD)

Dallas (DFW)

San Francisco (SFO)

Los Angeles (LAX)

1) Download the Data from website for specific airports for specific years and all or required features and extract as CSV files in folder weather\_datasets

2) Setup hpc tunnel

3) Run from cmd prompt

hpc gateway

4) scp -r /home/ vsh231/weather\_datasets  [vsh231@dumbo](mailto:asb862@dumbo): weather\_datasets

5) Put the data in HDFS

Log into dumbo cluster

ssh  [vsh231@gw.hpc.nyu.edu](mailto:asb862@gw.hpc.nyu.edu)

ssh  [vsh231@dumbo.hpc.nyu.edu](mailto:asb862@dumbo.hpc.nyu.edu)

mkdir project

hdfs dfs -put weather\_datasets project

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