**Steps to ingest the data:**

**Data Source Used:** Bureau of Transportation Statistics – Airline On-Time Performance Data [Reporting Carrier On-Time Performance (1987-present)](https://transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time)

https://transtats.bts.gov/Tables.asp?DB\_ID=120&DB\_Name=Airline%20On-Time%20Performance%20Data&DB\_Short\_Name=On-Time

1) Download the Data from website for specific years month-wise and all or required states and extract as CSV files in folder flightData

2) Setup hpc tunnel

3) Run from cmd prompt

hpc gateway

4) scp -r /home/asb862/fightData [asb862@dumbo](mailto:asb862@dumbo): flightData

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)

hdfs dfs -put fightData flightData

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