

DATA INGESTING APPROACH

I tried using two ways, one was the curl command, the other one is the transferring it from local machine to Hadoop node.

Step1-

I used the curl command to ingest data initially,

```
[hbr244@login-2-1 ~]$ curl -o NYCweatherdata.txt (https://www1.ncdc.noaa.gov/pub/orders/CD05084358047522.txt)
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total     Spent    Left  Speed
100 1328k  100 1328k    0     0  2954k      0  --:--:-- --:--:-- --:--:--  7101k
[hbr244@login-2-1 ~]$ vi NYCweatherdata.txt
-bash: vi: clean: command not found
[hbr244@login-2-1 ~]$ hdfs dfs -mkdir /user/hbr244/Project
mkdir: `/user/hbr244/Project': File exists
[hbr244@login-2-1 ~]$ hdfs dfs -put NYCweatherdata.txt /user/hbr244/Project
put: `/user/hbr244/Project/NYCweatherdata.txt': File exists
[hbr244@login-2-1 ~]$ hdfs dfs -cat /user/hbr244/Project/NYCweatherdata.txt
```

Step-2-

Downloaded the csv version and transferred it to Hadoop using WinSCP. This way the entire data is now comma separated and easy to operate on.

To transfer using WinSCP

- 1) Connect to VPN
- 2) Open winscp
- 3) Login using your nyu creds
- 4) Drag and drop the file from your local to PUTTY