The link of the original dataset is: https://www.kaggle.com/imdevskp/corona-virus-report?select=country_wise_latest.csv

Steps I take to ingest data:

1. Download the dataset as a zip folder from the link
2. Unzip the folder and use the csv file named “country_wise_latest.csv”
3. Use scp to move this csv file from my laptop to NYU Dumbo by typing “scp ./country_wise_latest.csv yt1324@dumbo.es.its.nyu.edu:/home/yt1324” into a Windows terminal.
4. Open another terminal (I use Cygwin and Putty) to log in to my account on NYU dumbo
5. Type “ls” to make sure the file has successfully been transferred from my local laptop to the NYU Dumbo.
6. Make a new directory for the project on HDFS by typing “hdfs dfs -mkdir /user/yt1324/pbdaa_project
7. Put the csv file to HDFS by “hdfs dfs -put country_wise_latest.csv /user/yt1324/pbdaa_project”
8. Use the command “hdfs dfs -cat /user/yt1324/pbdaa_project/country_wise_latest.csv” to verify the data has been successfully transferred to HDFS.