

## Assignment 2: Big Data Management and Analytics

----- Authors -----

NOVONIL DAS - NXD180022

PRITOM DAS RADHESHYAM - PXR180033

----- Files Included -----

Assignment2.jar

Tweets.csv

AirportData.csv

CoverPage.pdf

ReadMe.pdf

----- Running the program on AWS -----

1) Transfer the jar file: Assignment2.jar from the Target folder to AWS S3 Storage.

### PART 1

Copy **Tweets.csv** to AWS S3 and note the path to the File.

Run the file on the AWS EMR cluster by:

#### Add Steps

Enter in Spark Submit options: **--class "TweetProcessing"**

Jar File: **s3://Bucket\_Name/Assignment2.jar**

Arguments: **s3://Bucket\_Name/Tweets.csv s3://Bucket\_Name/outputTweetProcessing** (space separated arguments)

Output files will be generated in **outputTweetProcessing** folder on S3

## PART 2

Copy **AirportData.csv** to AWS S3 and note the path to the File.

Run the file on the AWS EMR cluster by:

### Add Steps

Enter in Spark Submit options: `--class "PageRank"`

Jar file: `s3://Bucket_Name/Assignment2.jar`

Arguments: `s3://Bucket_Name/AirportData.csv 10 s3://Bucket_Name/outputPageRank` (space separated arguments)

Output files will be generated in **outputPageRank** folder on S3