Name: Pramod Padmakar Nagare

NUID: 001858910

Email: Nagare.p@husky.neu.edu

PROJECT: Amazon Customer Reviews – Analytics/Streaming Pipeline/Sentiment Analysis

Dataset:

- https://registry.opendata.aws/amazon-reviews/
- https://github.com/awslabs/open-data-registry/blob/master/datasets/amazon-reviews.yaml
- https://s3.amazonaws.com/amazon-reviews-pds/readme.html
- 34+ GB Dataset
- 130+ millions amazon customer reviews

Deliverables:

1. Write a script to download dataset to local machine and optimize it
2. Write a script to move data from local machine to HDFS
3. Create a Hadoop cluster on cloud
4. Create a script to load HDFS without downloading dataset to local machine and directly to HDFS
5. Considering a data to be streaming data, create a streaming pipeline using AWS/GCP, Redshift/BigTable
6. Implement analysis using MapReduce that covers, simple MapReduce, Customer Writable objects, chaining map reduce jobs, filtering, secondary sorting, joins
7. Create a test environment to test the map reduce program
8. Implement the sentiment analysis machine learning model
9. Data Visualization with the map reduce output