**Big Data Processing of MovieLens Ratings Dataset Using MapReduce on AWS**

**Milestone 1 – Submit Title Page, Dataset Introduction, and Hadoop Cluster Setup Successful Screenshots**

Student Names: Anastasiya Kotelnikova

Course and Section Number: DS-644-854

NJIT UCID: 31708551

Email Address: ak3384@njit.edu

Date: 4/2/2025

**Dataset Introduction**

**Dataset Name & Source**

Name: MovieLens 1M Dataset

Source: <https://grouplens.org/datasets/movielens/1m/>

**Dataset Description**

The MovieLens 1M dataset is a collection of 1 million ratings collected from 6,040 users on approximately 3,900 movies. The dataset is divided into three main files:  
- ratings.dat – Contains user ID, movie ID, rating (from 1 to 5), and a timestamp.  
- movies.dat – Contains movie ID, movie title, and genres.  
- users.dat – Contains user ID, gender, age group, occupation, and zip code.  
  
Each file uses a double colon (::) as a delimiter and can be easily parsed and converted to CSV for use with Hadoop MapReduce.

**Size & Format**

Size: ~5 MB  
Format: Plain text files (.dat) with :: delimiter, easily convertible to CSV

**Reason for Selection**

The MovieLens dataset is structured, publicly available, and easy to work with. It is large enough to demonstrate the advantages of distributed computing but small enough to efficiently process using a basic Hadoop setup on AWS. It is ideal for MapReduce tasks such as calculating the average rating per movie, identifying the most-rated movies, or analyzing user rating behavior. The simple structure and real-world relevance make it an excellent choice for demonstrating big data processing techniques.

**Hadoop Cluster Setup Screenshots**

**1. AWS EC2 Instances Running**

This screenshot shows both Master and Slave EC2 instances running in the AWS management console. Each instance is in a "running" state and has passed the required status checks, confirming proper initialization.

**2. Passphraseless SSH Login from Master to Slave**

This screenshot demonstrates successful SSH login from the Master node to the Slave node without requiring a password. This confirms that passwordless SSH is correctly configured, which is essential for Hadoop cluster communication.

**3. JPS Output on Master and Slave Nodes**

This screenshot shows the output of the `jps` command on both the Master and Slave nodes. On the Master node, services such as NameNode and SecondaryNameNode are visible. On the Slave node, the DataNode service is running. This verifies that Hadoop daemons are functioning correctly on both nodes.