BU.330.740 Large Scale Computing on the Cloud

**Lab 3 Extension. Movie Recommender using Apache Pig**

Learning Goal: use Apache Pig to implement a movie recommender on MovieLens dataset and deploy on AWS Hadoop cluster

Required Skills: understand basics of collaborative filtering based recommender and Apache Pig

1. Download MovieLens 100K Dataset from <http://files.grouplens.org/datasets/movielens/ml-100k.zip>, and use the original dataset named “*u.data”*.
2. Download Apache Pig script file from Canvas.
3. login into AWS Educate account, go to **AWS Management Console**->**EMR**, choose the cluster you set up and then **Clone**,and choose **DO NOT** **include the steps**.
4. While waiting for the cluster to be provisioned, go to **AWS Management Console**->**S3,** create a bucket for lab6. Create 2 folders in your bucket**,** 1 for input file and 1 for your Apache Pig script. Upload u.data into your input folder, and movie\_recommender.pig into your script folder.
5. Wait till the cluster is ready, add a step of type **Pig program**. Name your Apache Pig program. Point Script to movie\_recommender.pig on your S3; Input to u.data on your S3; and Output to a folder on your S3 instance. **Note that this output folder should not pre-exist**.

A screenshot of a computer

Description automatically generated

Add this step and then wait for your program to complete. After it’s completed, you can check and download results from your S3 bucket -> your output folder.

1. Last but not least, **DO NOT FORGET TO terminate the cluster.**

Reference:

<https://github.com/alanfgates/programmingpig>