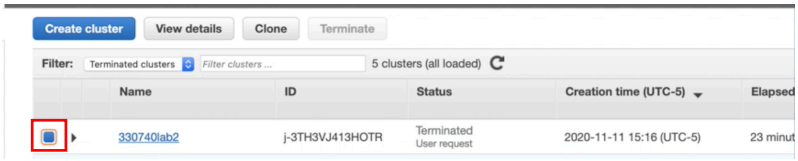
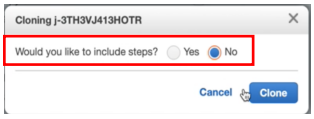
BU.330.740 Large Scale Computing with Hadoop
Lab 6. Movie Recommender using Apache Pig on AWS

Learning Goal: use Apache Pig to implement a movie recommender system on MovieLens movie review dataset, and deploy on AWS Hadoop cluster
Required Skills: understand basics of collaborative filtering based recommendation engine, understand Apache Pig basics

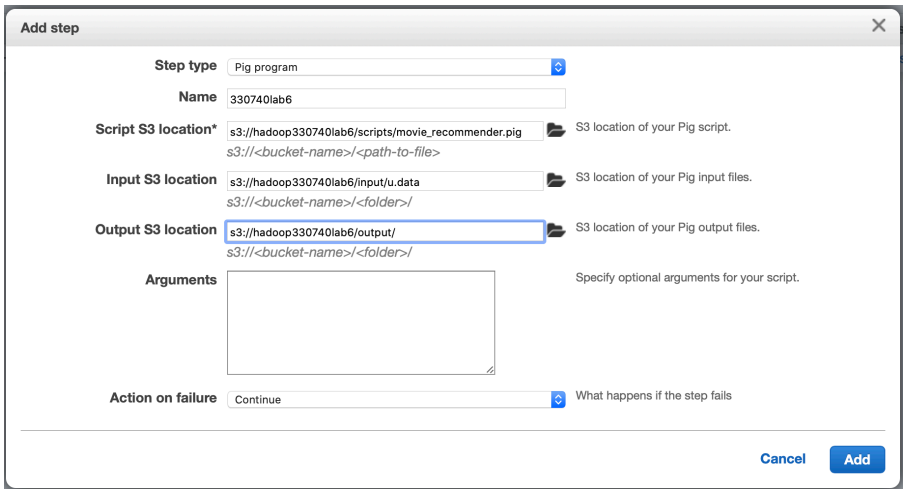
1. Download MovieLens 100K Dataset from http://files.grouplens.org/datasets/movielens/ml-100k.zip, and use the original dataset named "*u.data*".
   Here is an index of unzipped files: http://files.grouplens.org/datasets/movielens/ml-100k/. You can also take a look at u.data file in the browser at http://files.grouplens.org/datasets/movielens/ml-100k/u.data
2. Download Apache Pig script file from Blackboard.

3. login into AWS Educate account, go to **AWS Management Console**->**EMR**, choose the cluster you set up in lab2 or lab3 or lab5 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 scripts. Upload u.data into your input folder, and movie_recommender.pig into your scripts 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**.



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.

6. Last but not least, **DO NOT FORGET TO CLEAN UP RESOURCSES!!** Terminate the cluster, delete all S3 buckets under your account, and always double check.

Reference:

https://github.com/alanfgates/programmingpig
https://grouplens.org/datasets/movielens/