BU.330.740 Large Scale Computing with Hadoop

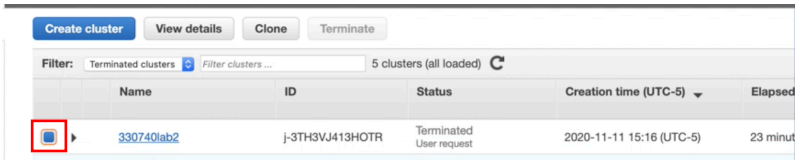
Lab 5. Twitter Sentiment Analysis using Hive on AWS

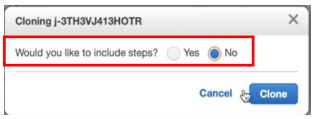
Learning Goal: use Hive to implement Twitter sentiment analysis, and deploy it on AWS Hadoop cluster

Required Skills: understand basics of sentiment analysis using dictionary, understand Hive basics

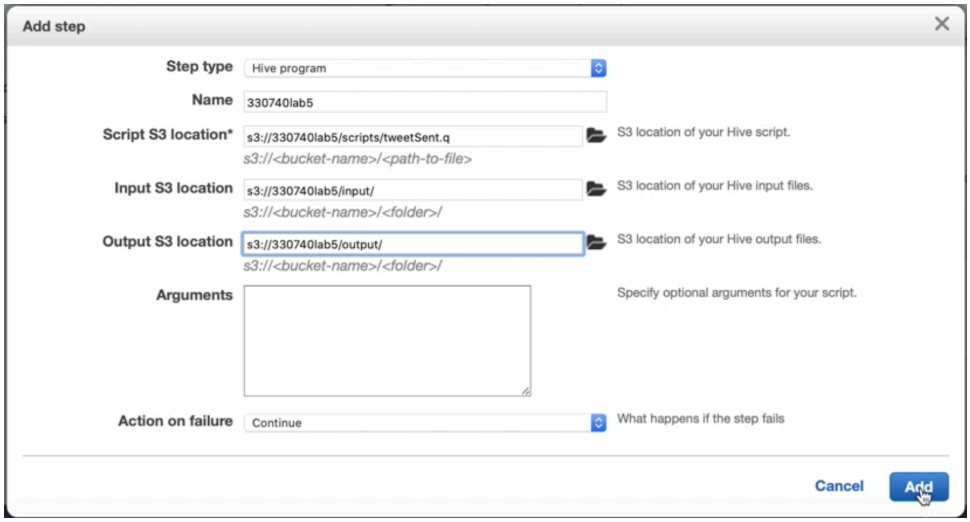
1. Download the zip file of inputs and scripts, unzip and store the three files in your local folder.

2. login into AWS Educate account, go to **AWS Management Console**->**EMR**, choose the cluster you set up in lab2 or lab3 and then **Clone**, and choose **DO NOT include the steps**.





3. While waiting for the cluster to be provisioned, go to **AWS Management Console**->**S3,** create a bucket for lab5. Create 2 folders in your bucket**,** 1 for input files and 1 for your Hive scripts. Under the input folder, create a sub-folder named **tweets** for tweets file and another sub-folder named **dictionary** for dictionary file.
   Upload tweets.csv into **tweets** folder; dictionary.csv into **dictionary** folder; and tweetSent.q into your scripts folder.
   Please note that if you do not use **tweets** and **dictionary** as the folder names, you need to modify the script file, tweetSent.q, accordingly.
4. Wait till the cluster is ready, add a step of type **Hive program**. Name your Hive program. Point Script to tweetSent.q on your S3; Input to the input folder 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.



5. 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://aws.amazon.com/getting-started/projects/analyze-big-data/
https://www.kaggle.com/crowdflower/twitter-airline-sentiment