README

Files Submitted:

The following files/folders were submitted in the zip folder:

- README.pdf (this file)
- CoverPage.pdf
- PageRank.scala
- tweets.scala
- assignment-2_2.11-0.1.jar
- June2020_Origin_Dest_Reporting.csv
- Tweets.csv
- outputDir_PageRank
- outputDir_Tweets

The directories are the expected directories and files outputted by the programs in an AWS cluster.

How to Run:

1) Create an AWS Cluster.

   For instructions in how to create an AWS cluster, please see "Step II: Launching AWS Cluster" in the professor's guide called "Getting Started with AWS" on eLearning.

   https://elearning.utdallas.edu/bbcswebdav/pid-3522631-dt-content-rid-80601566_1/courses/2202-UTDAL-CS-6350-SEC002-24563/GettingStartedAWS.pdf

   Note: The cluster MUST be compatible with Scala version 2.11.8 and Spark version 2.2.1. My programs were ran on emr-5.12.3 and is the ideal version for the AWS cluster when running my programs.

   For instructions to how to change the software configuration of the AWS cluster, please see the guide of the professor called "Getting Started with AWS".

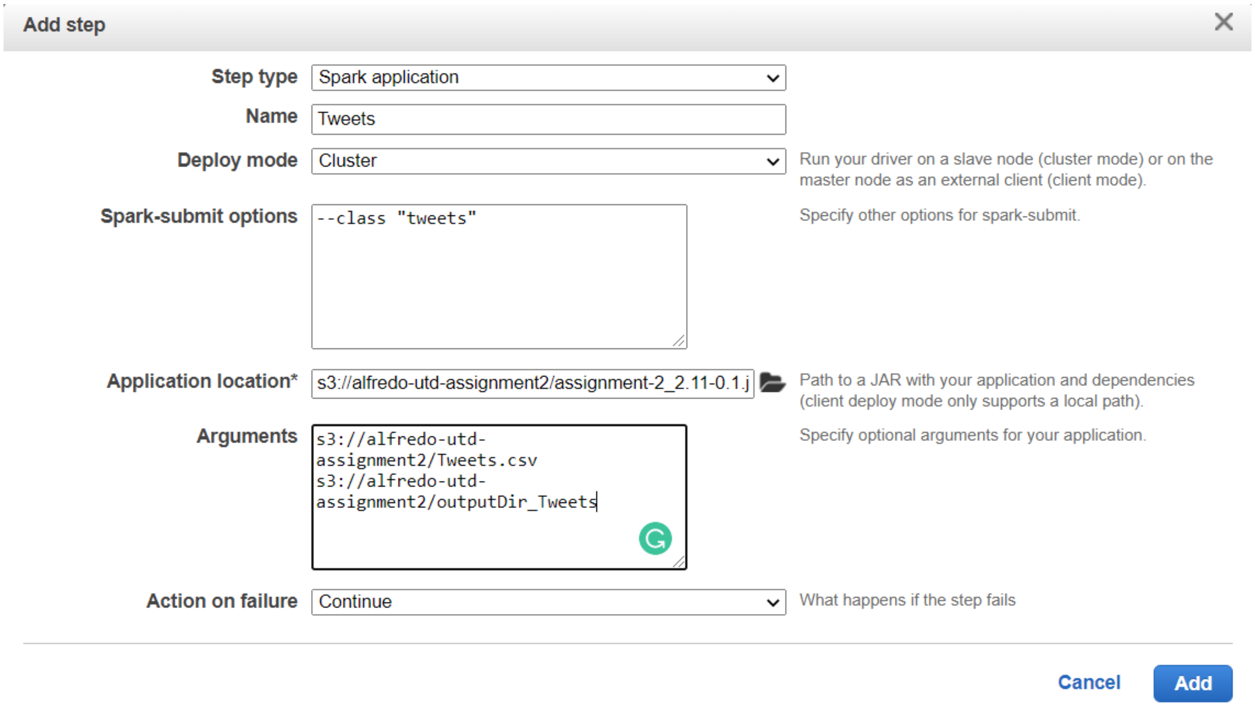
2) Run the JAR File

   - Login into your S3 account: https://aws.amazon.com/s3/ and create bucket in the same region as your cluster.

   - Upload all the input files and the jar file

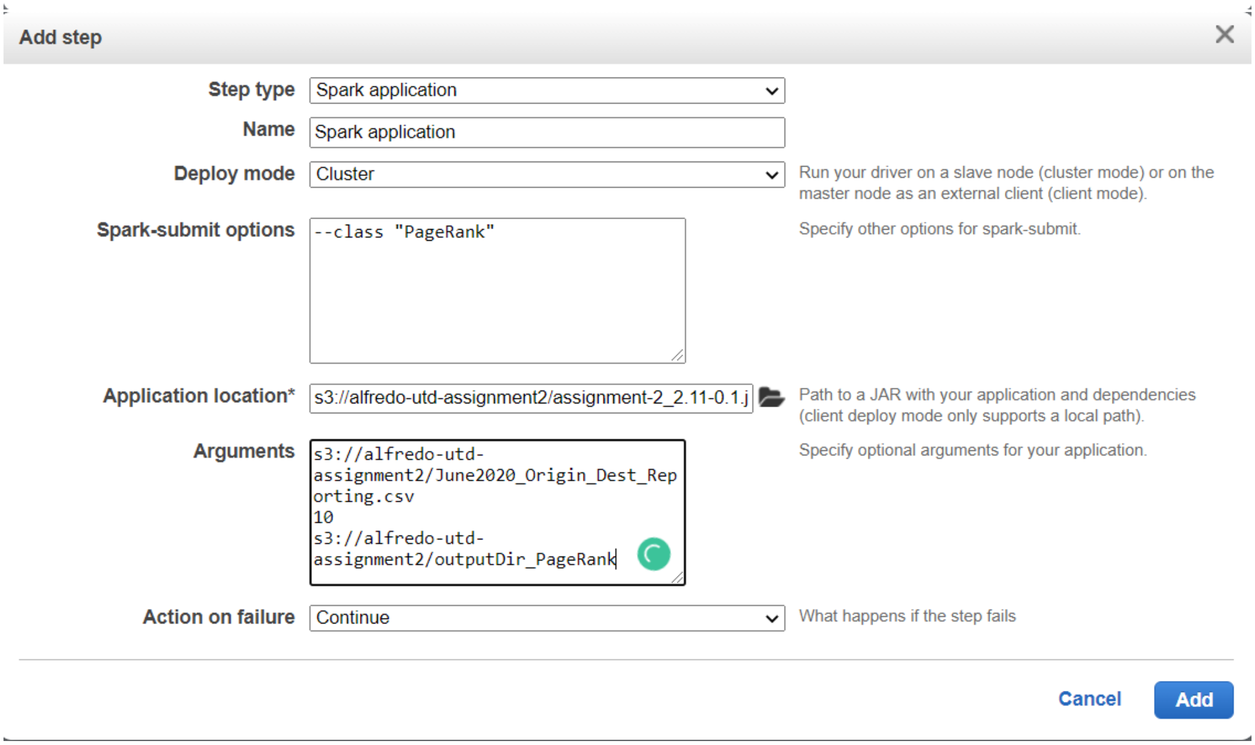
- Now go back to your EMR cluster, click on the cluster name and then click on the "Steps" tab.

- Click on "Add Step" and then fill in the values as shown below:



Add step

Step type	Spark application	
Name	Tweets	
Deploy mode	Cluster	Run your driver on a slave node (cluster mode) or on the master node as an external client (client mode).
Spark-submit options	--class "tweets"	Specify other options for spark-submit.
Application location*	s3://alfredo-utd-assignment2/assignment-2_2.11-0.1.j	Path to a JAR with your application and dependencies (client deploy mode only supports a local path).
Arguments	s3://alfredo-utd-assignment2/Tweets.csv s3://alfredo-utd-assignment2/outputDir_Tweets	Specify optional arguments for your application.
Action on failure	Continue	What happens if the step fails



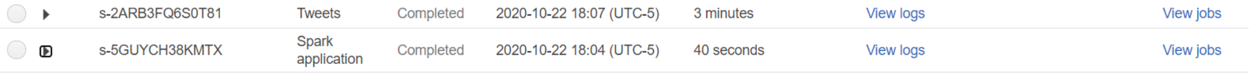
Add step

Step type	Spark application	
Name	Spark application	
Deploy mode	Cluster	Run your driver on a slave node (cluster mode) or on the master node as an external client (client mode).
Spark-submit options	--class "PageRank"	Specify other options for spark-submit.
Application location*	s3://alfredo-utd-assignment2/assignment-2_2.11-0.1.j	Path to a JAR with your application and dependencies (client deploy mode only supports a local path).
Arguments	s3://alfredo-utd-assignment2/June2020_Origin_Dest_Reporting.csv 10 s3://alfredo-utd-assignment2/outputDir_PageRank	Specify optional arguments for your application.
Action on failure	Continue	What happens if the step fails

You are going to have two steps, one for each class in the jar file. Also keep in mind the order of the arguments matters so make sure you have correct order.

Note: The application location and arguments might be a little different for you as the location would be different if you are using your bucket.

After the steps are done then it should look something like this:

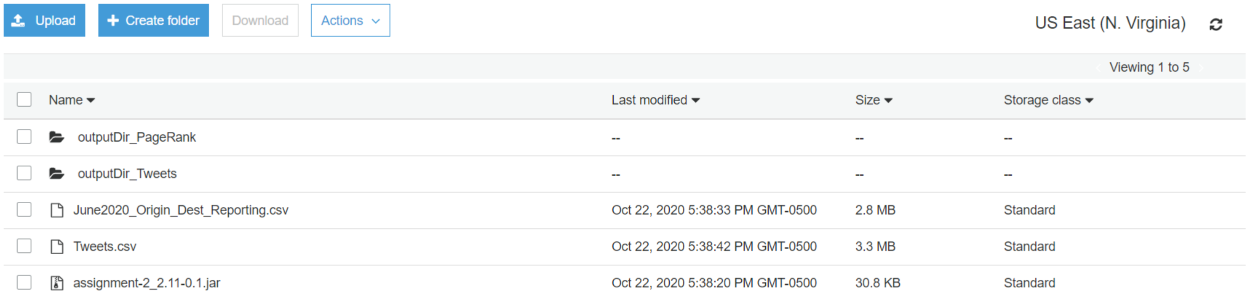


○	s-2ARB3FQ6S0T81	Tweets	Completed	2020-10-22 18:07 (UTC-5)	3 minutes	View logs	View jobs
○	s-5GUYCH38KMTX	Spark application	Completed	2020-10-22 18:04 (UTC-5)	40 seconds	View logs	View jobs

Notice the "Tweets" takes a couple minutes.

3) Look at the output files

Once the programs are completed you, go back to your bucket and you shall see the directories created by the page rank and tweets class. It should look something like this:



□ Name	Last modified	Size	Storage class
□ outputDir_PageRank	--	--	--
□ outputDir_Tweets	--	--	--
□ June2020_Origin_Dest_Reporting.csv	Oct 22, 2020 5:38:33 PM GMT-0500	2.8 MB	Standard
□ Tweets.csv	Oct 22, 2020 5:38:42 PM GMT-0500	3.3 MB	Standard
□ assignment-2_2.11-0.1.jar	Oct 22, 2020 5:38:20 PM GMT-0500	30.8 KB	Standard

Then once you go into the directory the output files should be there.

Finally, note that if you try to run these programs multiple times with an output directory that already exists then an error will occur. So, make sure to delete the directory if you plan on running it again or giving it out a different output name for the next run.

All the programs are working as expected, if you have any questions regarding my programs or any other topic please do not hesitate to contact me. Thank you.