The City College of New York | Department of Computer Science

CSC 445: Big Data Management & Analysis
FALL 2021

Lab 13 – Spark @ AWS

In this lab, we're going to setup an Amazon Elastic MapReduce cluster and run a Spark job on top of it. You will need an Amazon Web Services (AWS) account before proceeding. If you have not requested your AWS account, please do so at http://aws.amazon.com/, and selecting the AWS Educate Student account. Please also make sure that you have a Terminal with SSH capability on your machine (e.g. a Terminal on Linux/Mac OS X or Putty on Windows).

TASK 1 – Setting up a Development Environment in AWS

First, we need to setup a proper Key Pair (or IAM Role for other tasks) to access any . This will be used to enable us logging into future AWS instances. For setting up Access Key, please follow this tutorial (only the first step – *Preparing a key pair*).
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair

Note: if you have another SSH Key, and would like to use it, you could follow Option 2.

Alternatively, you could go through the simple tutorial below to have your key created as well as familiarizing yourself with AWS Elastic Compute Cloud (EC2):
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html
Please be sure to stop your instance after the last step. Otherwise, you credit will be charged for the instance. You can simply through this through the GUI (AWS Console) or through the command line (AWS CLI), e.g.:

```
aws ec2 stop-instances --instance-ids <YOUR_INSTANCE_ID>
```

TASK 2 – Creating an S3 Storage Bucket

Please follow the following tutorial to create an S3 storage bucket:
http://docs.aws.amazon.com/AmazonS3/latest/UG/CreatingaBucket.html

After creating a bucket with a name of your choice, please create a subfolder in that bucket and upload all **Lab 6** materials (aka. the content of our git repo) to the folder. You can do that through the AWS S3 Console at:
https://console.aws.amazon.com/s3/

TASK 3 – Creating an Elastic MapReduce (EMR) Cluster

We can use the following tutorial as a guideline for creating an EMR Cluster on AWS:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html

Please note that you only need to follow Step 1 and Step 2 on that page, and when creating your cluster, let's use only **m4.large** instances to avoid unnecessary charge.

TASK 4 – Running the WordCount example using Spark

Please follow these steps to run the WordCount example using Spark.

1. SSH into your cluster master node (using your SSH key above).

2. Create a lab13 folder, and synchronize that with your bucket S3 lab5 folder

```
mkdir lab13
aws s3 sync s3://YOUR_BUCKET/lab13 lab13
```

3. Upload data onto HDFS and run the Word Count Example:

```
cd lab13
hadoop fs -put book.txt .
spark-submit wordcount.py book.txt output
```

4. Verify your counts.txt to contain the list of words and counts.

5. We could also run the example above but with fetching data directly from our s3 buckets

```
spark-submit wordcount.py \
      s3a://BUCKET/lab13/book.txt \
      s3a://YOUR_BUCKET/lab13/output
```

6. counts.txt is still being copied to local disk since we would like to inspect it from the console.

7. In Step 5, since all of our files and input/output are stored on S3, there is no need for us to login to the master node and submit our job. Indeed, we can add a step to do this through the Console. You can try it out by adding the following Spark Application step when creating your cluster:

   https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark-submit-step.html

   with the following step details:

   Application location: `s3://YOUR_BUCKET/lab13/wordcount.py`
   Arguments: `s3a://BUCKET/lab13/book.txt s3a://YOUR_BUCKET/lab13/output`
   Action on failure: Terminate cluster

8. Wait for finish, then download and checkout your results on S3.

IMPORTANT: always TERMINATE your cluster after your use to avoid additional charges.