BU.330.740 Large Scale Computing on the Cloud

**Lab 1. Word Count using MapReduce**

Learning Goal: practice using AWS S3, EC2 and EMR to implement the word count problem

1. Prepare. Download scripts.zip from Canvas, unzip and store them in a local folder. download an e-book *Pride and Prejudice* by Jane Austen in txt format from

<http://www.gutenberg.org/ebooks/1342> and save in your local folder.

1. Login into your AWS Academy account at <https://awsacademy.instructure.com/login/canvas>. Then please follow “*AWS Academy Learner Lab – Student Guide.pdf”* provided by AWS to launch **AWS Management Console**.

Note: “Start Lab” step may take long especially at the first time.

1. Launch the EMR Cluster. Go back to **AWS Management Console**, search for EMR to open the EMR console. Choose **Create cluster**.

A screenshot of a computer

Description automatically generated

On the **Create Cluster** page:

* 1. Enter a **Cluster name** that helps you identify the cluster.
  2. Under **Application bundle**, choose **Core Hadoop**.
  3. **Cluster configuration**: Instance groups, Primary: **m4.large**, Core: **m4.large**, Task -1: **m4.large**

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

* 1. Under **Security configuration and EC2 key pair**, choose the EC2 key pair, **vockey**.

A screenshot of a browser

Description automatically generated

* 1. Under **IAM roles**, choose **EMR\_DefaultRole**; and **EC2 instance profile for Amazon EMR**, choose **EMR\_EC2\_DefaultRole**.

A screenshot of a computer screen

Description automatically generated

Then click **Create Cluster**.

Now the cluster is starting, which means AWS will provision your cluster (find available nodes to fulfill your job request).

1. Store your files on S3. Go back to **AWS Management Console**, search for S3, or locate it under **Storage** section.

At S3 console, create a S3 bucket. Name your S3 bucket, and leave all other settings unchanged. Because of Hadoop requirements, bucket and folder names that you use with Amazon EMR must contain only letters in lower case, numbers, periods (.), and hyphens (-). After you specify the name, scroll down to the bottom and click **Create bucket**.

A screenshot of a bucket

Description automatically generated

After you create the bucket, choose it from the list and then choose **Create folder**. Create two folders, input and scripts, one by one. After specifying the name of the folder, do not change anything else, but choose **Create folder**.

Then click input folder, choose **Upload** in the upper left corner. Then click **Add files**, to add the ebook file to the folder. Do not change other settings, but click **Upload** directly. Once you see the “Upload succeeded” message, click **Close** and go back to your bucket, click scripts folder, and upload mapper.py and reducer.py to the folder.A screenshot of a computer

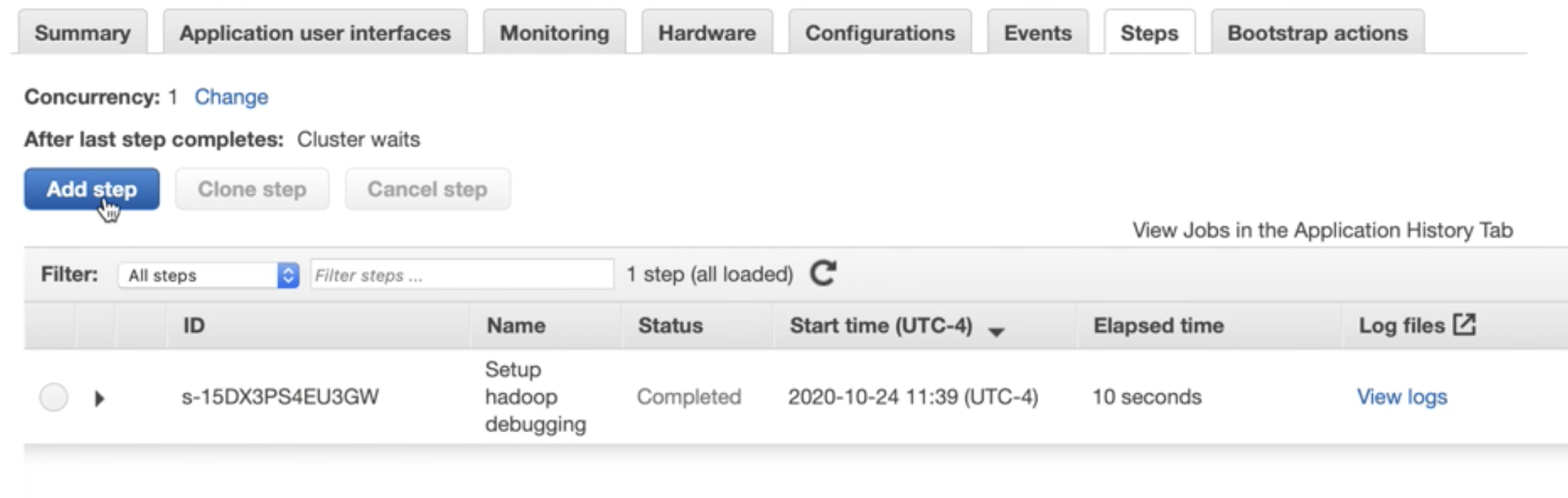
Description automatically generated

1. Wait for cluster to be ready. This process may take more than 10 minutes. After it’s provisioned. Your cluster status will show as “Waiting”. You can always click icon to refresh the status.

A screenshot of a message

Description automatically generated

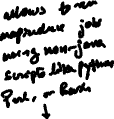
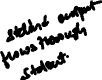
1. Add a step in cluster. After the cluster is ready, click **Steps**, and then **Add step**.



Choose **Streaming program** for **Step type**. **Name** your Streaming program; point **Mapper** to your mapper.py and **Reducer** to your reducer.py on your S3 instance; point **Input** **S3 location** to bigdata.txt; designate **Output S3 location** to a folder called output on your S3 instance, by typing in your S3 bucket and “output/”. **Note that this output folder should not be created before**, and Hadoop will create it for you. You can use the **View** icon (shown in red rectangle) to have access to folders on your own S3 instance. Click **Add** to execute the step.

A screenshot of a step settings

Description automatically generated



Type in output folder name directly, which should not exist in your S3 bucket

Wait while the step is going from **Pending** to **Running** to **Completed**. You can always use icon to refresh the status.

A screenshot of a computer

Description automatically generated

1. Download results. After it’s completed, you can check and download results from your S3 bucket -> output folder. Download these files to your local folder. And you can also check log files in your EMR cluster. Being able to read log files is a skill to acquire, especially if your task has a “failure” or “error” condition. You will get more practice as we go.

A screenshot of a computer

Description automatically generated

1. Once you are done with lab 1. **Do not forget to terminate your cluster!** Otherwise, you will incur unnecessary charges to your account.

A screenshot of a computer

Description automatically generated

Take some time to navigate your AWS Management Console. Double check that your EMR cluster is terminated, and no running instance in your EC2. And go back to AWS Academy LMS to “End Lab”.

Reference:

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-setting-up.html>

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html>

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-master-core-task-nodes.html>