

Launching an AWS EMR Cluster with RStudio, Hive, Pig, and Hue

Feb 9, 2015 • TheCoatlessProfessor • Tags: [rstudio](#), [rstudio-server](#), [aws](#), [emr](#), [ec2](#), [hive](#), [pig](#), [hue](#), [cygwin](#)

Intro

In the last post, [AWS CLI](#) was setup, we authorized a user account, generated a keypair, and assigned default roles for the cluster to take on. We continue onward now to obtain the perfect AWS EMR Cluster that conforms as much as possible to the UIUC Big Data Image. We do this by first obtaining an S3 bucket and uploading install files. Then, we launch the [AWS EMR](#) cluster via [AWS CLI](#).

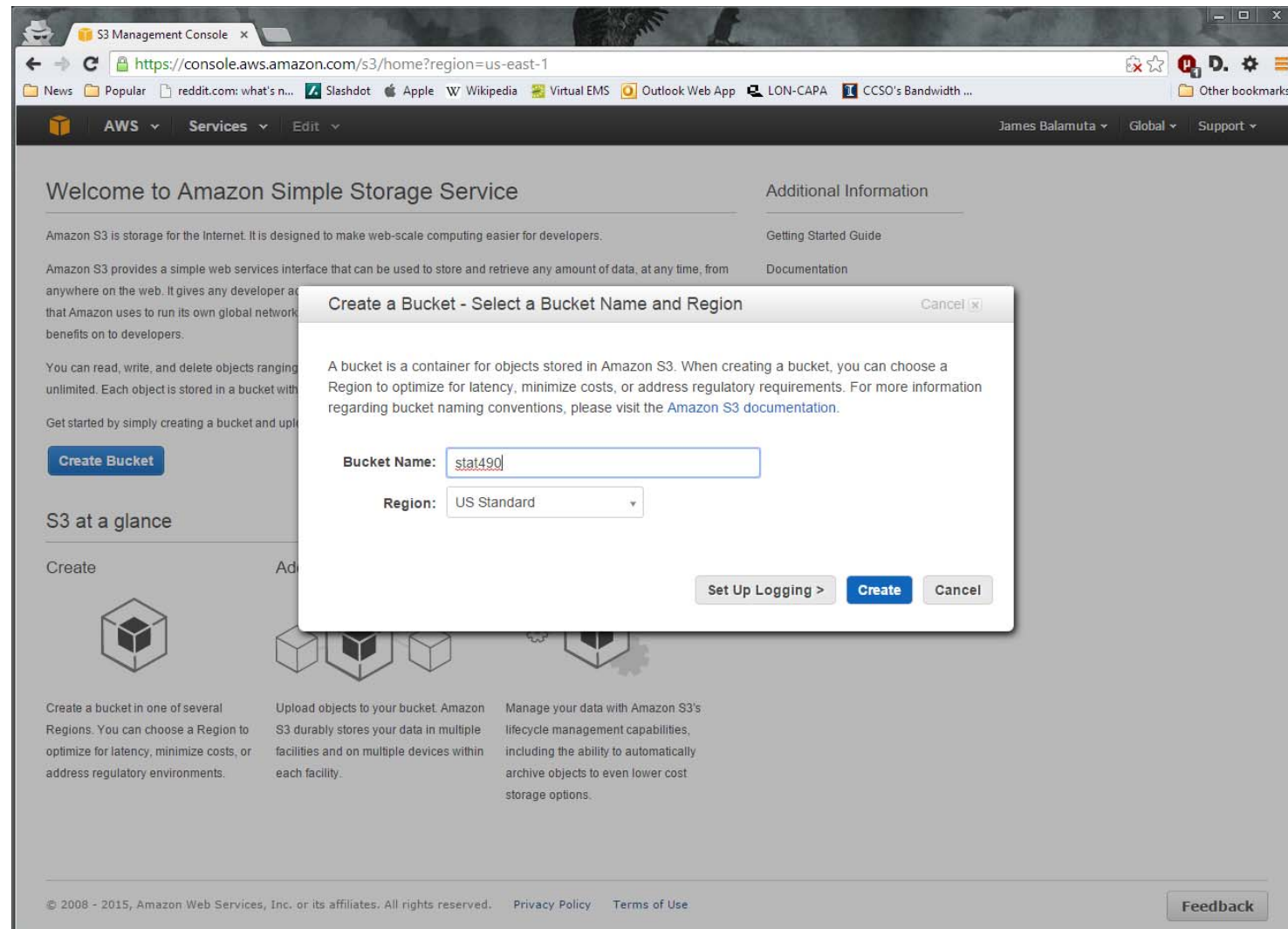
Create a Bucket and Upload files to S3

Now, we need to create a bucket on S3. Amazon's S3 service can be thought of as a hard drive. That is, everything that you wish to keep forever and ever, you will want to save to your bucket. If the files are not placed within the bucket when the cluster is killed, those files will be lost.

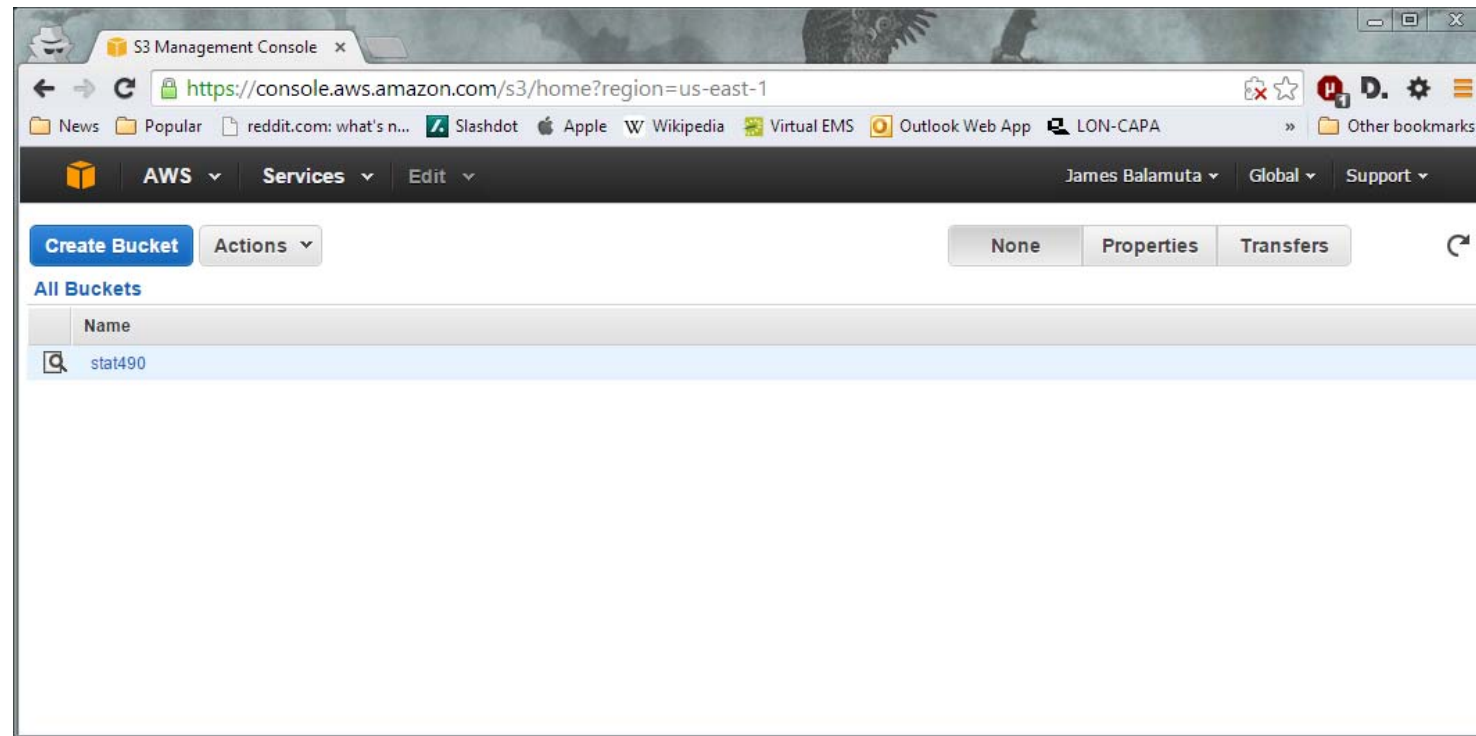
To create a bucket go to the [S3 Console](#) and press

The screenshot shows the Amazon S3 Management Console. The browser address bar displays the URL `https://console.aws.amazon.com/s3/home?region=us-east-1`. The page header includes the AWS logo, navigation tabs for 'AWS', 'Services', and 'Edit', and user information for 'James Balamuta'. The main content area is titled 'Welcome to Amazon Simple Storage Service' and contains several paragraphs of introductory text about S3. A prominent blue 'Create Bucket' button is visible. To the right, an 'Additional Information' sidebar lists links for 'Getting Started Guide', 'Documentation', and 'All S3 Resources'. Below the welcome message, a 'S3 at a glance' section provides a quick overview of S3 capabilities, organized into three columns: 'Create' (creating a bucket), 'Add' (uploading objects), and 'Manage' (managing data with lifecycle policies). Each column includes an icon and a short description. The footer contains copyright information for 2008-2015, links to 'Privacy Policy' and 'Terms of Use', and a 'Feedback' button.

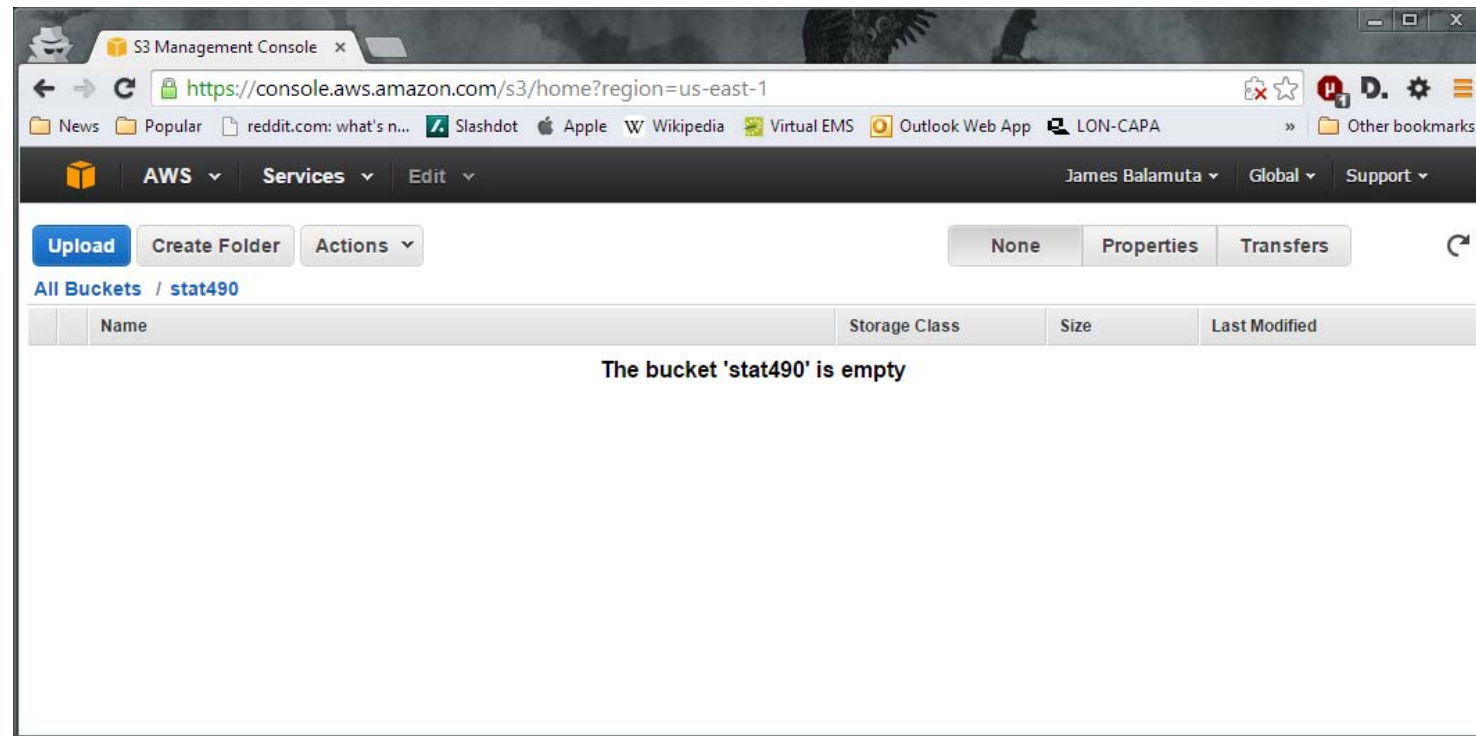
Name the bucket something short and reasonable. You will have to type it out many times.



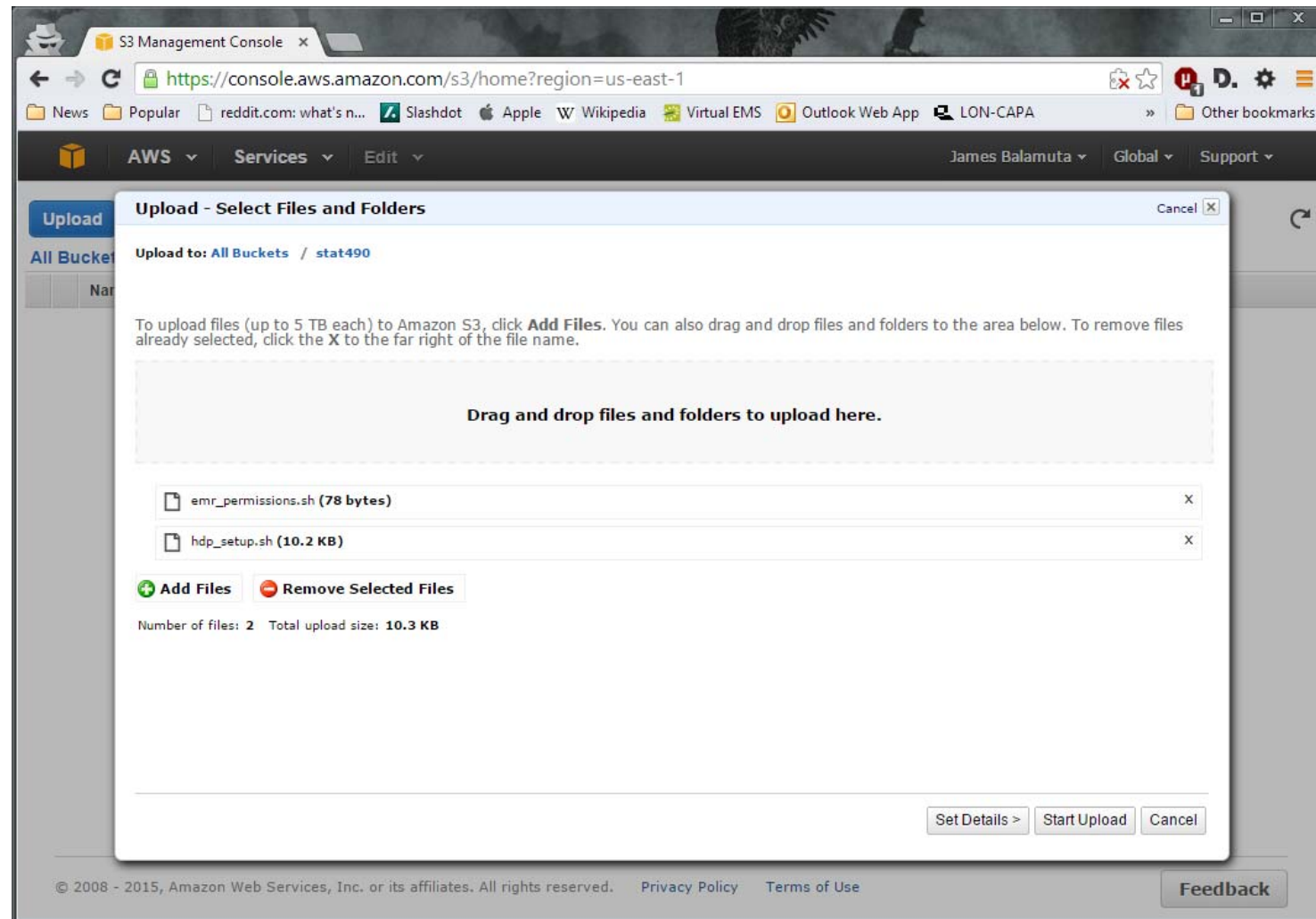
If the bucket was successfully created, the page will update to show its presence:



Click on the name of the new bucket that you created to enter the bucket. Inside the bucket, press the `Upload` button.

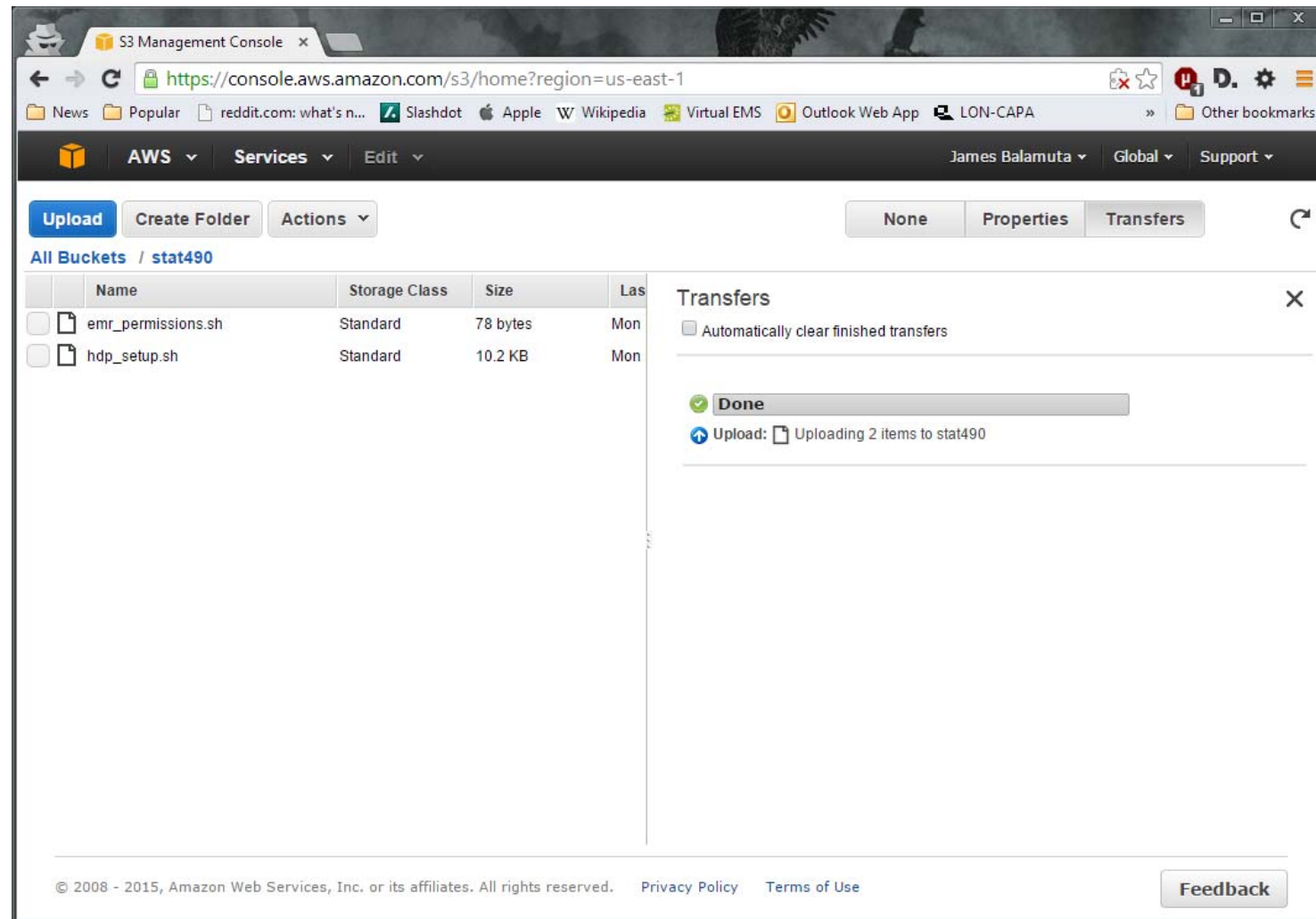


Upload into the bucket these two scripts: [hdp_setup.sh](#) and [emr_permissions.sh](#)



Note: [hdp_setup.sh](#) is the script created to automatically modify the [Hortonworks' Data Platform \(HDP\)](#) Virtual Box Image for UIUC. It's been modified slightly to account for EMR specific install.

If all is well, your bucket should now look like:



Create the cluster via the bootstrap file

Before we launch the cluster, we need to talk about pricing on AWS. [Specifically, the EMR pricing](#). For our purposes, we only need to use the m1 instance type. Anything larger will be overkill and will be costly. By costly, the assignments every month in the course should require no more than the amount of 3 Starbucks single shot grande skim 2 pumps Mocha with whip, no lid, a little cinnamon, and a little nutmeg (note: last 2 ingredients...self serve)!!

For each instance we spawn of the m1 instance type we must pay the following per hour:

General Purpose - Previous Generation

Instance Type	EC2 Cost per Hour	EMR Cost per Hour	Total Cost per Hour
m1.small	\$0.044	\$0.011	\$0.055
m1.medium	\$0.087	\$0.022	\$0.109
m1.large	\$0.175	\$0.044	\$0.219
m1.xlarge	\$0.350	\$0.088	\$0.438

(Rates as of 2/6/15)

The configuration script given below then would yield a cost of \$0.657 per hour (0.219×3).

Note, if you issue the command below, you will be charged monies depending on how long the cluster is active.

Prior to running the below command, replace `<YOUR-X>` with your information. Upon running this command, a hadoop cluster will be created.

```
bucket="<YOUR_BUCKET>"
region="<YOUR_REGION>"
keypair="<YOUR_KEYPAIR>"
master_instance="m1.large"
slave_instance="m1.large"
num_slaves=2

aws emr create-cluster --name emr_cluster \
--ami-version=3.3.0 \
--applications Name=Hue Name=Hive Name=Pig \
--region $region \
--use-default-roles --ec2-attributes KeyName=$keypair \
```



```
--no-auto-terminate \  
--instance-groups \  
InstanceGroupType=MASTER, InstanceCount=1, InstanceType=$master_instance \  
InstanceGroupType=CORE, InstanceCount=$num_slaves, InstanceType=$slave_instance  
--bootstrap-actions \  
Name=emR_bootstrap, \  
Path="s3://$bucket/hdp_setup.sh", \  
Args=[--emrinstall, --rstudio, --hpaths, --rhadoop, --createuser, --sudouser, --sshu  
--steps \  
Name=HDFS_tmp_permission, \  
Jar="s3://elasticmapreduce/libs/script-runner/script-runner.jar", \  
Args="s3://$bucket/emr_permissions.sh"
```

The above command has been modified slightly from the [AWS Lab post on EMR](#) to match the Big Data image implementation.

If the command issued above is successful, you should receive a string back that identifies the cluster ID.

```
James@Calypso ~
$ bucket="stat490"
--ami-version=3.3.0 \
--applications Name=Hue Name=Hive Name=Pi
gJames@Calypso ~
$ \
region="us-east-1"
InstanceGroupType=CORE,InstanceCount=$num_slaves,InstanceType=$s
lJames@Calypso ~
a$ ve_instance \k
eypair="jbb_keypair"
Jar="s3://elasticmapreduce/libs/script-run
nJames@Calypso ~
e$ r/script-runner.jar",\a
ster_instance="m1.large"

James@Calypso ~
$ slave_instance="m1.large"

James@Calypso ~
$ num_slaves=2

James@Calypso ~
$

James@Calypso ~
$ aws emr create-cluster --name emr_cluster \
> --ami-version=3.3.0 \
> --applications Name=Hue Name=Hive Name=Pig \
> --region $region \
> --use-default-roles --ec2-attributes KeyName=$keypair \
> --no-auto-terminate \
> --instance-groups \
> InstanceGroupType=MASTER,InstanceCount=1,InstanceType=$master_instance \
> InstanceGroupType=CORE,InstanceCount=$num_slaves,InstanceType=$slave_instance \
> --bootstrap-actions \
> Name=emR_bootstrap,\
> Path="s3://$bucket/hdp_setup.sh",\
> Args=[--emrinstall,--rstudio,--hpaths,--rhadoop,--createuser,--sudouser,--sshuser
] \
> --steps \
> Name=HDFS_tmp_permission,\
> Jar="s3://elasticmapreduce/libs/script-runner/script-runner.jar",\
> Args="s3://$bucket/emr_permissions.sh"
j-3997TH9PIW6DB

James@Calypso ~
$ |
```

You can monitor the status of the cluster (e.g. from provisioning, to installing, to using, and to terminating) at the [EMR Console](#).

Terminate Cluster

When you are done using the cluster, it is very important that you terminate it. Leaving the cluster active will rack up services fees on your AWS account.

To terminate the cluster via AWS CLI, we first need to get the [list of clusters](#):

```
aws emr list-clusters
```

Say that in our case the cluster ID returned was: `j-STATSatUIUCRocks1`

Then, we need to issue one of the following [termination commands](#) on the cluster ID:

```
# Not protected
aws emr terminate-clusters --cluster-ids j-STATSatUIUCRocks1

# Protected
aws emr terminate-clusters --cluster-ids j-STATSatUIUCRocks1 --no-termination-
```

The same can be done through the [EMR](#) or [EC2](#) consoles.

0 Comments

TheCoatlessProfessor

 Login ▾

 Recommend

 Share

Sort by Best ▾



Start the discussion...

Be the first to comment.

ALSO ON THECOATLESSPROFESSOR

WHAT'S THIS?

Setting up Jekyll with Dreamhost

2 comments • 7 months ago



James Balamuta — Greetings and Salutations
Wiley, The command to use for this is: jekyll build
--source GIT_REPO --destination ...

Accessing RStudio and Hue on AWS EMR + SSH

1 comment • 8 months ago



Vira Semenova — I have followed your instructions
1-5, which resulted at my cluster terminated with
step error "Run Hue". (The ...

Rcpp, RcppArmadillo and OS X Mavericks "-lgfortran" and "-lquadmath" error

2 comments • 8 months ago



Nodp53 — It's was the solution. Thnx!

Installing Amazon Web Services Command Line Interface (AWS CLI) for Windows, OS ...

3 comments • 8 months ago



Shaun Barker — awesome! my pip was corrupting
the aws install so this helped HUUGe!




 Subscribe

 Add Disqus to your site Add Disqus Add

 Privacy

The Coatless Professor

The Coatless Professor © 2016
jjb@thecoatlessprofessor.com
subscribe [via RSS](#)

 [coatless](#)
 [axiomsofxyz](#)
 [jamesbalamuta](#)

TheCoatlessProfessor is a website that strives to bring statistical prowess to the masses through useful articles for the stumbleponer and googler. Our goal is to make readily available helpful tips, tutorials, and resources that the students of Statistics and Computer Science will appreciate.