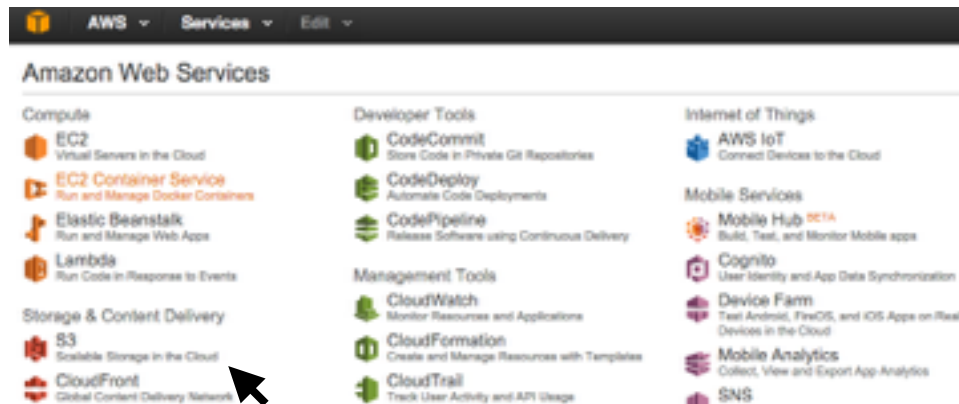


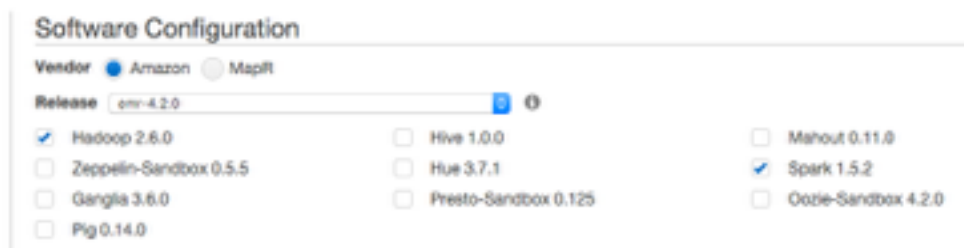
Running stmpy on Amazon Web Services

This guide illustrates the configuration process needed to run stmpy on a Spark cluster via Amazon Web Services. It assumes familiarity with the bash shell and the Jupyter/iPython notebook server.

- I. Sign up for an account with [AWS](#) and log into the console. Set up an RSA key pair for authentication as described in [this guide](#). Save your certificate in a safe place on your computer.
- II. Clone the contents of the [stmpy repository](#) locally. From the AWS console, enter the S3 screen. Using the S3 interface upload the file `aws_bootstrap.bash` from the stmpy repository to one of your S3 buckets. This bash script contains the bootstrap instructions for the cluster.

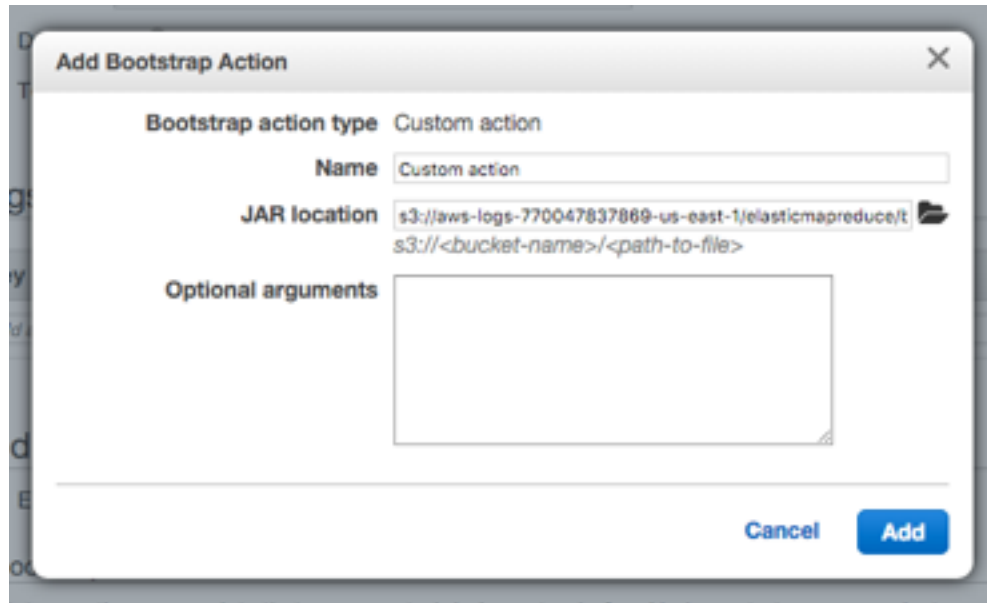


- III. Now go back to the console and enter the EMR (Elastic MapReduce) screen. Click *Create Cluster* and then *Go to advanced options*. In the *Software and Steps* tab, make sure to flag Hadoop and Spark for installation, then click *Next*. Everything else is optional.



- IV. In the *Hardware* tab choose the number of machines in your cluster as well as their specifications according to your computing needs. Click *Next*.

- V. Now we are in the *General Cluster Settings* tab. Expand the *Bootstrap Actions* section and choose to add a custom bootstrap action from the drop-down menu. Click on the folder icon to select a JAR location. A navigation menu showing your S3 buckets will appear. Use the navigation menu to select the *aws_bootstrap.bash* file that was previously saved. Click *Select*, then *Add*.



- VI. In the *Security* tab, make sure to select the key pair you previously created from the *EC2 key pair* drop-down menu. If you don't do this, you will not be able to log into your cluster. Leave all other settings to their defaults.

Security Options

EC2 key pair ⓘ ⓘ

☒ Cluster visible to all IAM users in account ⓘ

- VII. Now click *Create cluster*. The setup should take fifteen minutes or so. The cluster is ready once it is in a *Waiting* state.
- VIII. While the cluster gets set up, follow the instructions in the *Enable Web Connection* tab to establish an SSH tunnel to the AWS cluster. This will involve installing FoxyProxy in Google Chrome. The tunnel will allow you to access the iPython notebook server and the Spark console from your browser.
- IX. Once the cluster is ready, open a bash terminal window and navigate to the folder containing the RSA certificate you previously created. Note the master

public DNS from the EMR web interface. You can now SSH into the cluster via the following command (substitute values appropriately): `ssh -i <MY_KEY>.pem hadoop@<MY_DNS>`

- X. Once you have established an SSH connection to your master node, run the following commands in the terminal:

```
wget https://dl.dropboxusercontent.com/u/113867121/stmpy/jupyter_setup.bash
bash jupyter_setup.bash
```

- XI. Now type `jupyter notebook` in the terminal. This will start the iPython server. Now use your local browser to navigate to `https://<MY_DNS>:8888`. You can now use the iPython server to start a new notebook in which to conduct your analysis. To access Spark and stmpy, run the following commands at the top of the notebook:

```
import findspark; findspark.init()

import pyspark, stmpy
sc = pyspark.SparkContext(appName="myAppName")
```

- XII. You are now ready to use stmpy on AWS! Remember to terminate your cluster when finished in order not to incur unnecessary charges.