EMR on AWS

Elastic MapReduce on Amazon Web Services

Cosmin Stamate

Cloud Computing

Birkbeck
UNIVERSITY OF LONDON

Hadoop Streaming

- Streaming application reads input from *standard input* and then runs a script or executable (called a mapper) against each input.
- The result from each of the inputs is saved locally, typically on a Hadoop Distributed File System (HDFS) partition.
- After the input is processed by the mapper, a second script, called a reducer processes the mapper results.
- The results from the reducer are sent to *standard output*. You can chain together a series of Streaming steps, where the output of one step becomes the input of another step.

Hadoop Streaming

- ❏ Easier to implement, allows you to focus more on algorithm design and less on re-creating boilerplate code.
- ❏ Allows for easy local testing of MapReduce applications
- ❏ On AWS EMR we can write MapReduce applications in many languages if we use the streaming program interface.
- ❏ We can code mappers, reducers and combiners, not only Java, but also in other languages like Python, Perl, Ruby, PHP, or Bash.
- ❏ The only essential thing to remember is that we are using *standard input* and *standard output* to feed our MapReduce streaming functions.

Datasets

Please download the following books in plain text format, which have been sourced from the Gutenberg Project

❑ http://www.dcs.bbk.ac.uk/~cosmin/cc/data/pg27827.txt
❑ http://www.dcs.bbk.ac.uk/~cosmin/cc/data/pg3207.txt
❑ http://www.dcs.bbk.ac.uk/~cosmin/cc/data/pg5200.txt

And the following google 1-grams which have been sourced from the Google Books Ngram Viewer

❑ http://www.dcs.bbk.ac.uk/~cosmin/cc/data/ngrams.txt

Mappers and reducers

We will be doing a word count on the books downloaded from the gutenberg project. We will use only a mapper and the **aggregate** function *(If you look in the wordcount-map.py file you will see that we are using the **LongValueSum** aggregate function)* instead of a actual reducer:

- ❑ http://www.dcs.bbk.ac.uk/~cosmin/cc/wordcount-map.py

We are going to use google ngrams (1-gram) to look for words which were coined in the year 1999

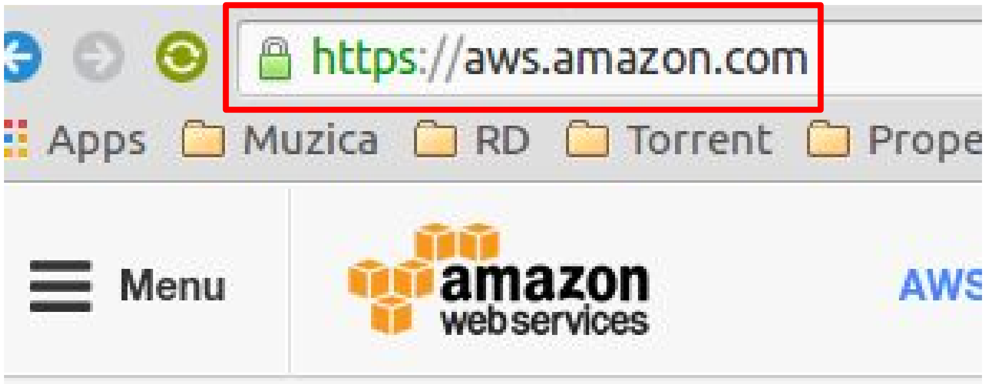
- ❑ http://www.dcs.bbk.ac.uk/~cosmin/cc/ngram-map.py
- ❑ http://www.dcs.bbk.ac.uk/~cosmin/cc/ngram-reduce.p

NOTE: python mapper and reducer

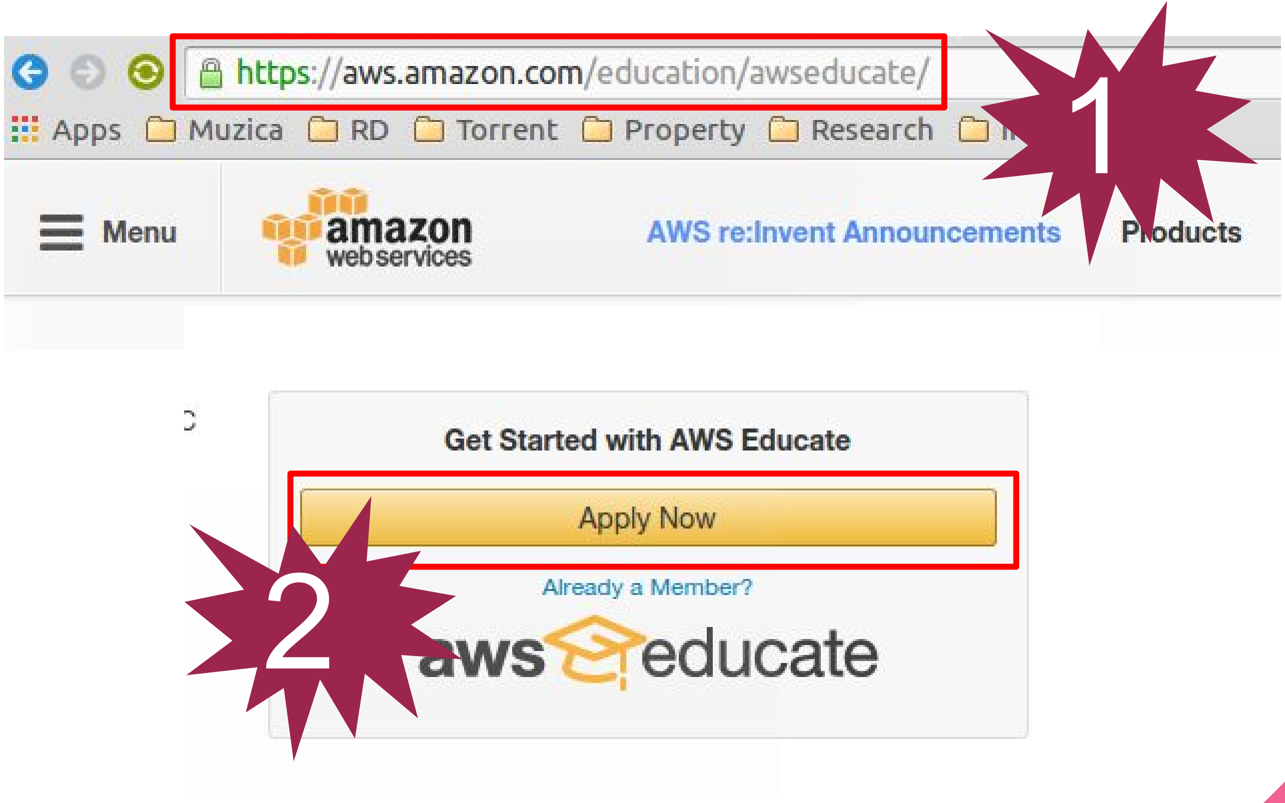
Please note that if you are using the Streaming API and are writing python mappers and reducers you have to add **#!/usr/bin/python** at the top of your python scripts as this is the path to the executable python. This needs to be there for all the scripts you have. The same applies for all other languages, for example if you will use ruby please add **#!/usr/bin/ruby** at the top of your script.

You can also use **#!/usr/bin/env python** depending on your environment setup.

Create a free **AWS** account



Apply for the Academic discount



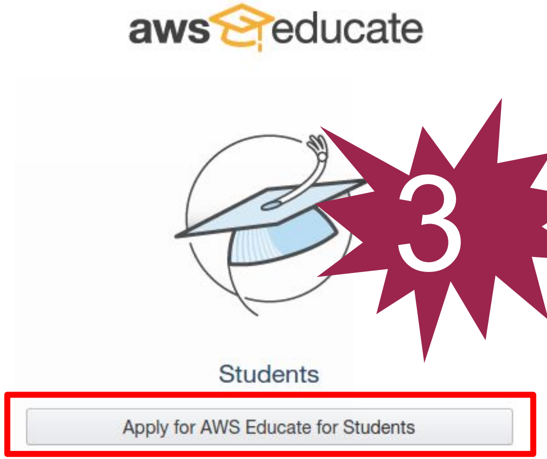
1

https://aws.amazon.com/education/awseducate/

Get Started with AWS Educate

2

Apply Now



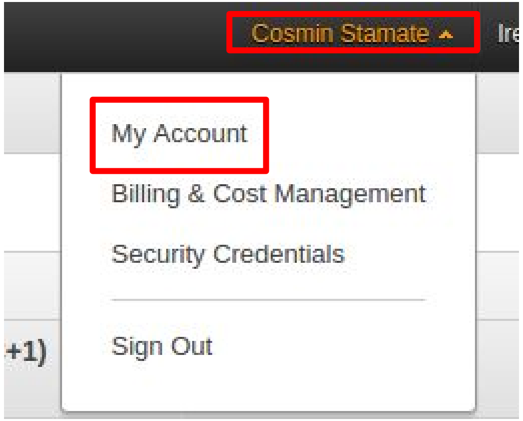
3

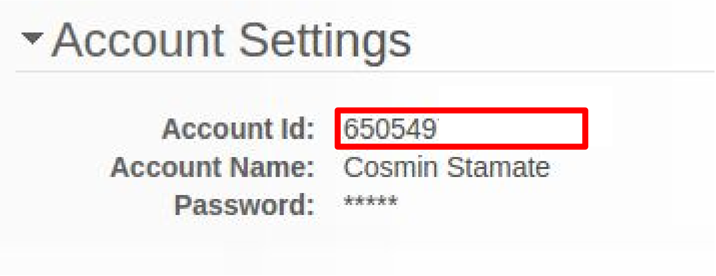
Students

Apply for AWS Educate for Students

Fill in the form and use your university email

- You can get your **AWS account ID** (12 digit number) by loggin in to your AWS console and going to **My Account**, under your name.





Account Settings

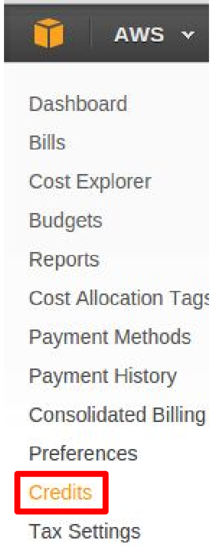
Account Id: 650549

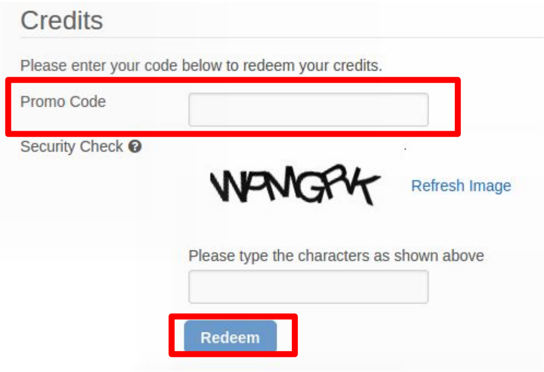
Account Name: Cosmin Stamate

Password: *****

After you receive the **AWS Educate** Application Approved email

- ❑ Go to **My Account > Credits**
- ❑ Paste the promo-code from the approval email and redeem the credits

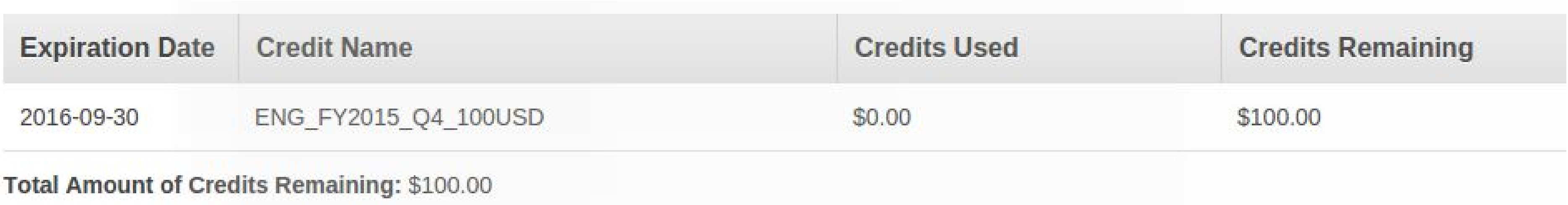



Credits

Please enter your code below to redeem your credits.

Please type the characters as shown above

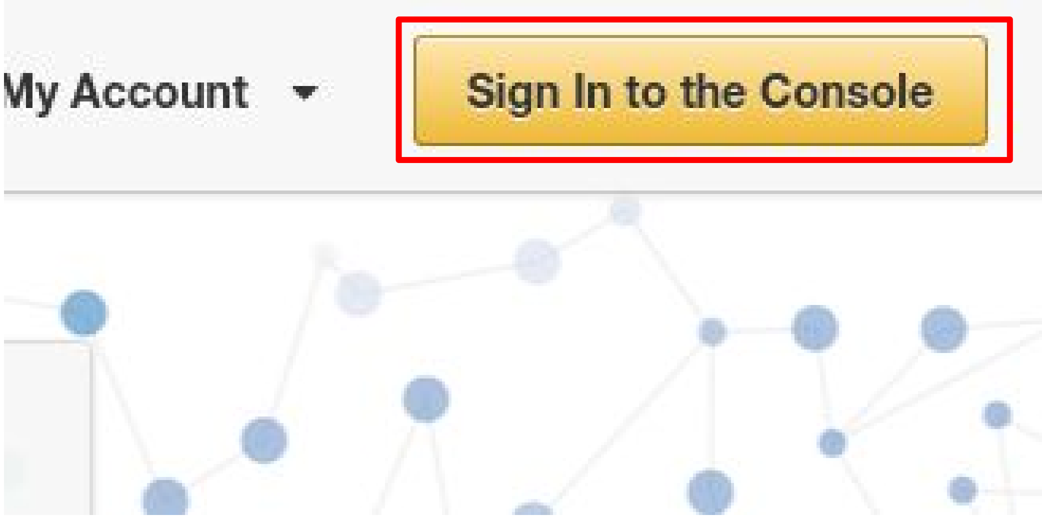
Congratulations, you now have $100 credits!



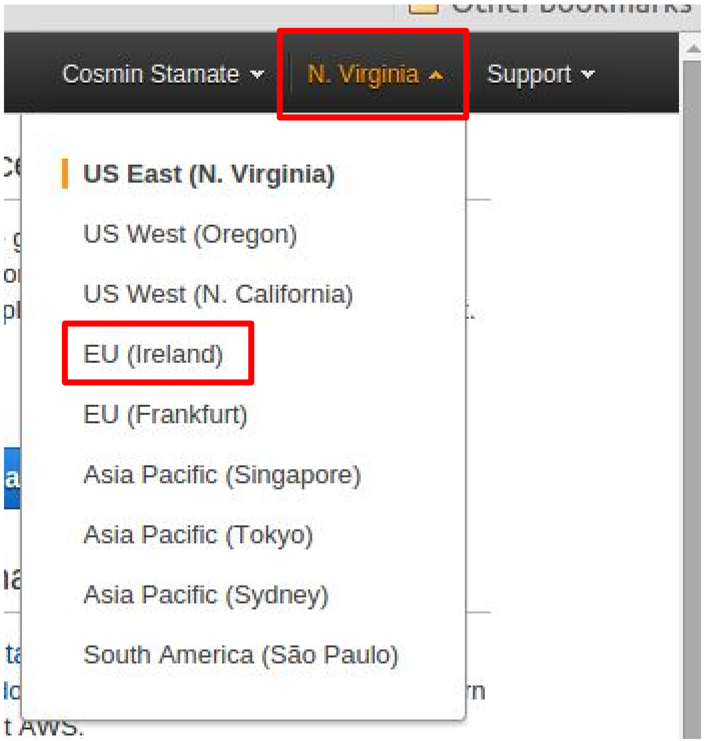
Expiration Date	Credit Name	Credits Used	Credits Remaining
2016-09-30	ENG_FY2015_Q4_100USD	$0.00	$100.00

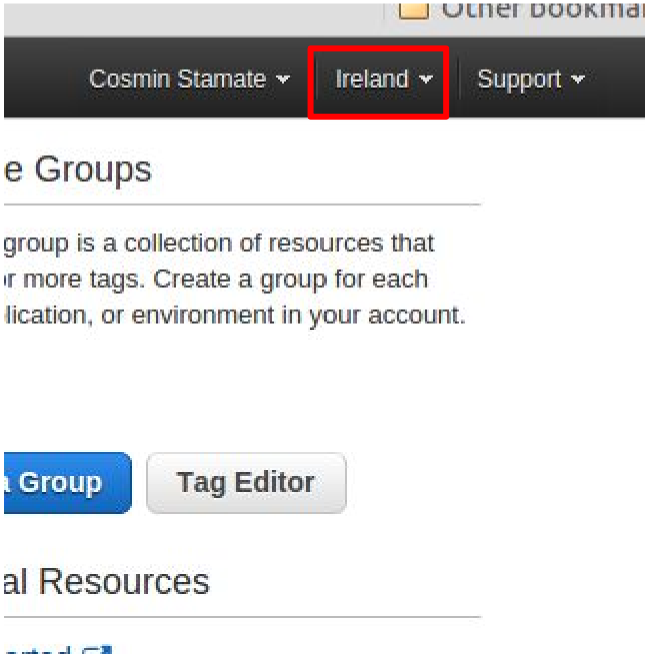
Total Amount of Credits Remaining: $100.00

Sign in to the AWS Console



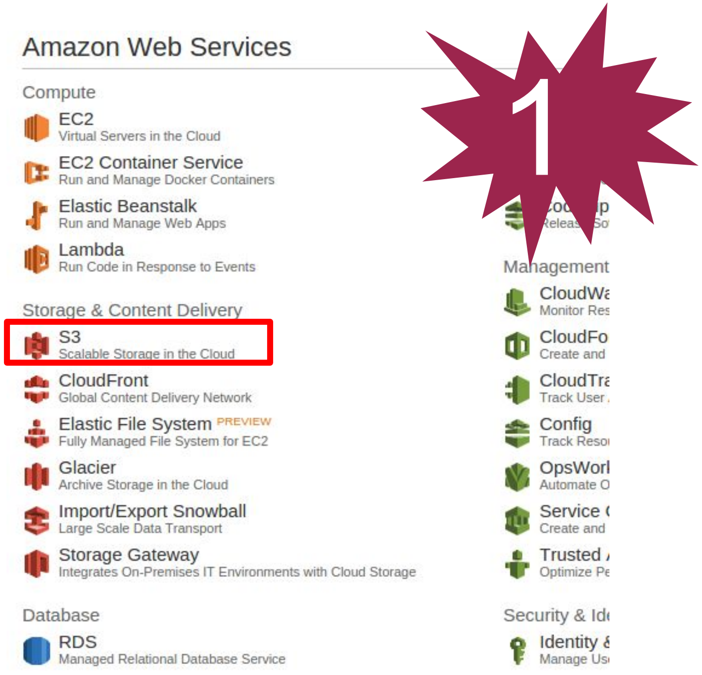
Chose **EU (Ireland)** region





group is a collection of resources that
r more tags. Create a group for each
lication, or environment in your account.

Create a **bucket** under **Storage > S3**



Amazon Web Services

Compute

EC2
Virtual Servers in the Cloud

EC2 Container Service
Run and Manage Docker Containers

Elastic Beanstalk
Run and Manage Web Apps

Lambda
Run Code in Response to Events

Storage & Content Delivery

S3
Scalable Storage in the Cloud

CloudFront
Global Content Delivery Network

Elastic File System PREVIEW
Fully Managed File System for EC2

Glacier
Archive Storage in the Cloud

Import/Export Snowball
Large Scale Data Transport

Storage Gateway
Integrates On-Premises IT Environments with Cloud Storage

Database

RDS
Managed Relational Database Service

Management

CloudWa
Monitor Res

CloudFo
Create and

CloudTra
Track User

Config
Track Reso

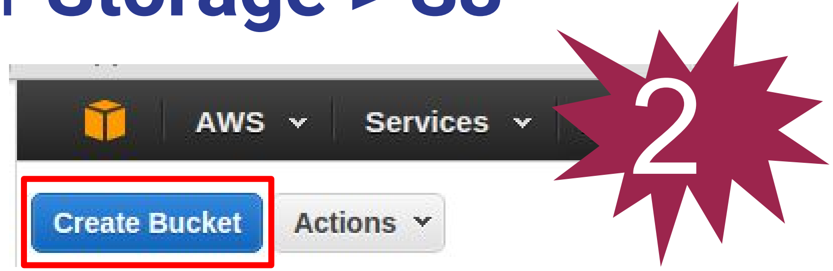
OpsWorl
Automate O

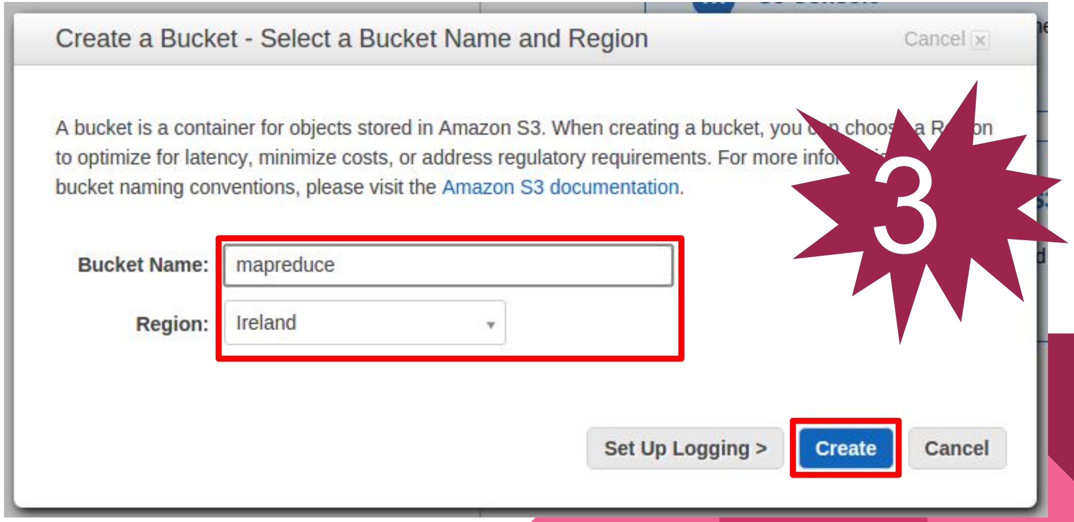
Service
Create and

Trusted
Optimize Pe

Security & Ide

Identity &
Manage Us





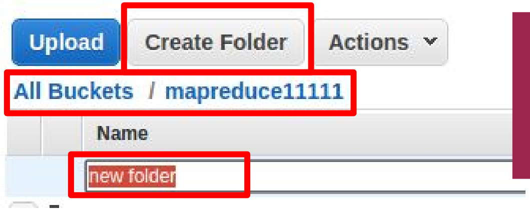
Create a Bucket - Select a Bucket Name and Region

A bucket is a container for objects stored in Amazon S3. When creating a bucket, you c n choos a R on to optimize for latency, minimize costs, or address regulatory requirements. For more infor bucket naming conventions, please visit the Amazon S3 documentation.

Bucket Name: mapreduce

Region: Ireland

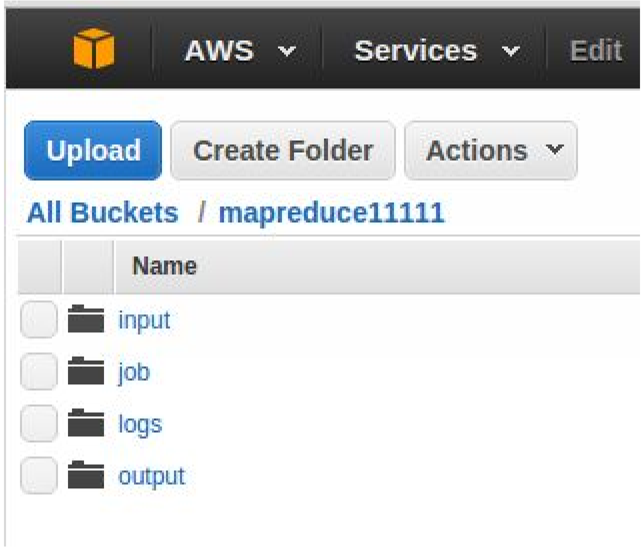
Add folders to the newly created **S3 bucket**



All Buckets / mapreduce11111

new folder

four folders

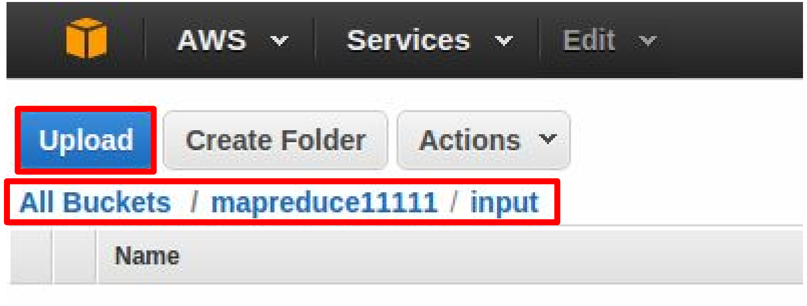


All Buckets / mapreduce11111

- input
- job
- logs
- output

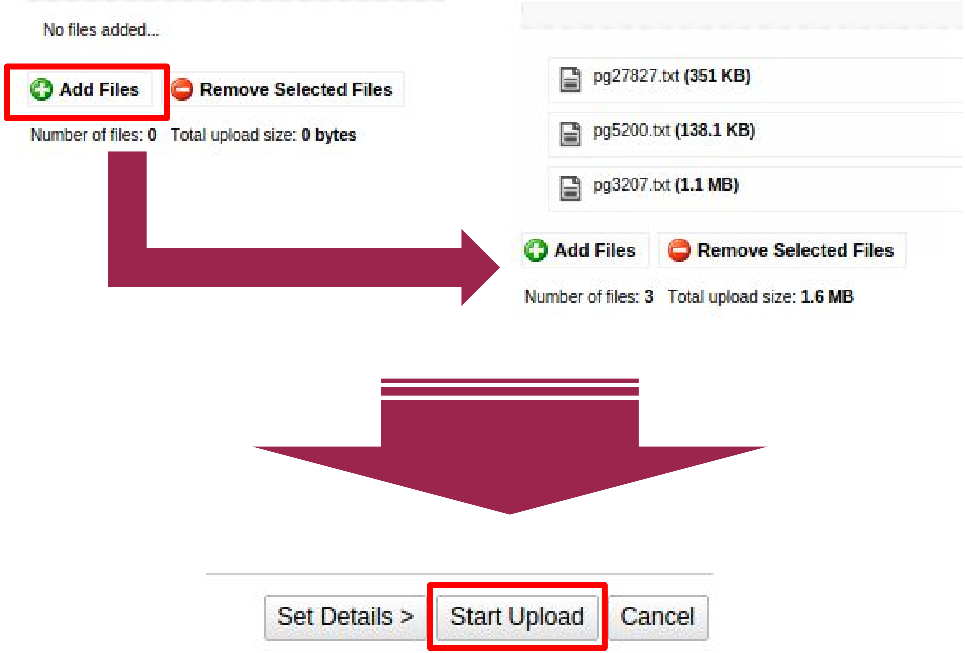
Upload wordcount books into their **S3 bucket** folder

Select the input folder and press Upload



All Buckets / mapreduce11111 / input

Add all your input files and upload



No files added...

Number of files: **0** Total upload size: **0 bytes**

pg27827.txt **(351 KB)**

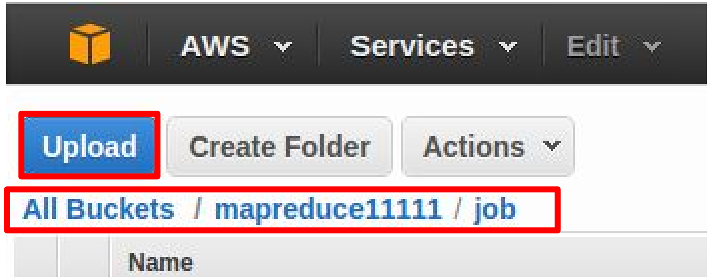
pg5200.txt **(138.1 KB)**

pg3207.txt **(1.1 MB)**

Number of files: **3** Total upload size: **1.6 MB**

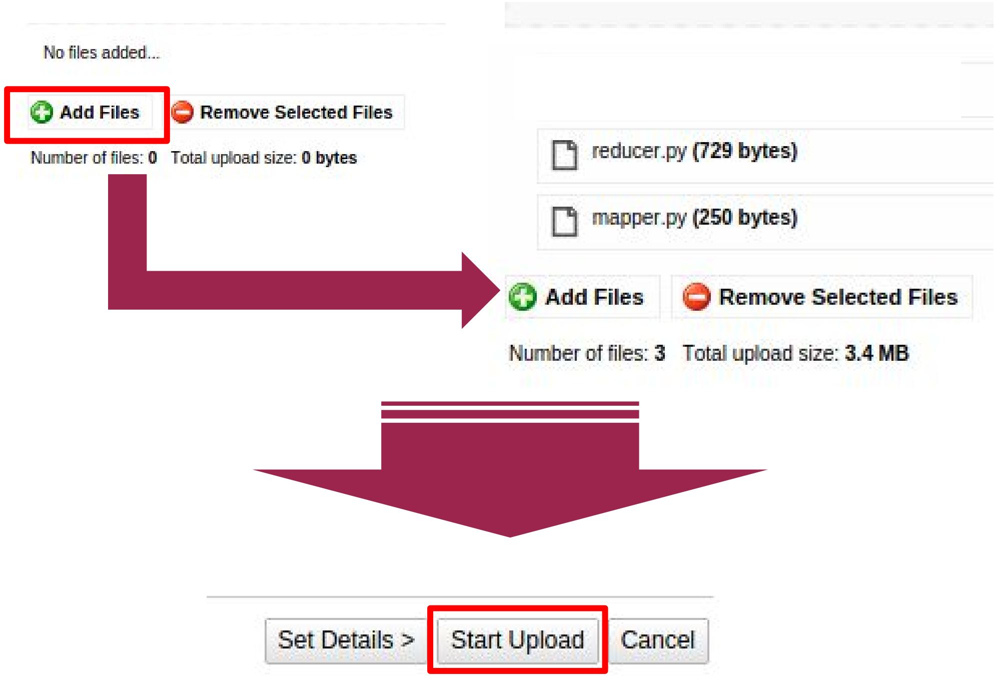
Upload your wordcount MapReduce files

Go to your **job** folder and press upload



All Buckets / mapreduce11111 / job

Select your files and upload them



No files added...

Number of files: 0 Total upload size: 0 bytes

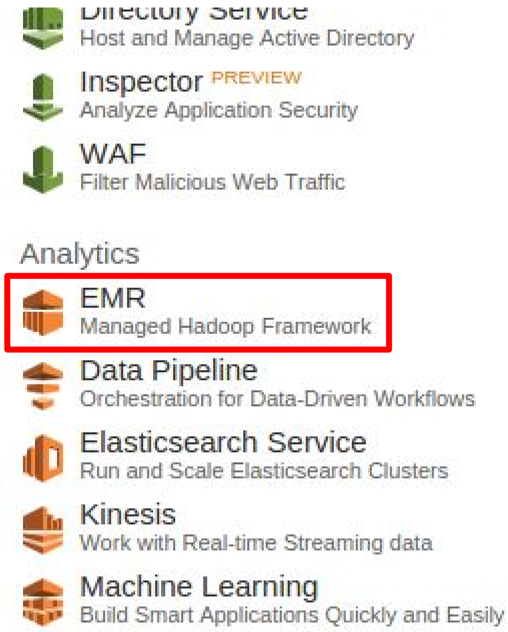
reducer.py (729 bytes)

mapper.py (250 bytes)

Number of files: 3 Total upload size: 3.4 MB

Create an **EMR cluster**

Under Analytics, press on **EMR**



Directory Service
Host and Manage Active Directory

Inspector PREVIEW
Analyze Application Security

WAF
Filter Malicious Web Traffic

Analytics

EMR
Managed Hadoop Framework

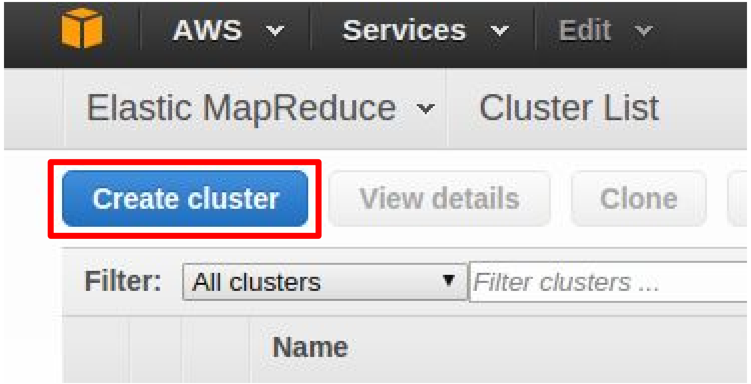
Data Pipeline
Orchestration for Data-Driven Workflows

Elasticsearch Service
Run and Scale Elasticsearch Clusters

Kinesis
Work with Real-time Streaming data

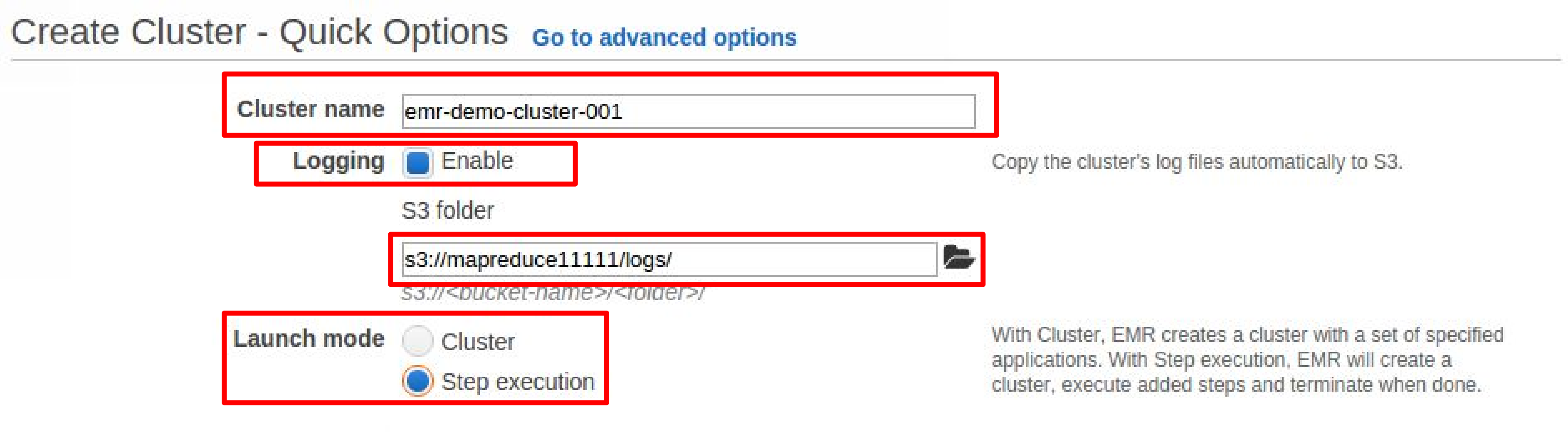
Machine Learning
Build Smart Applications Quickly and Easily

On the next page, press **Create cluster**



Create Cluster - Quick Options

- ❑ Choose a unique cluster name
- ❑ Enable logging and select the **logs/** S3 folder that you created earlier
- ❑ Choose step execution under **Launch mode**



Create Cluster - Quick Options

Cluster name: emr-demo-cluster-001

Logging ☑ Enable

Copy the cluster's log files automatically to S3.

S3 folder

s3://mapreduce11111/logs/

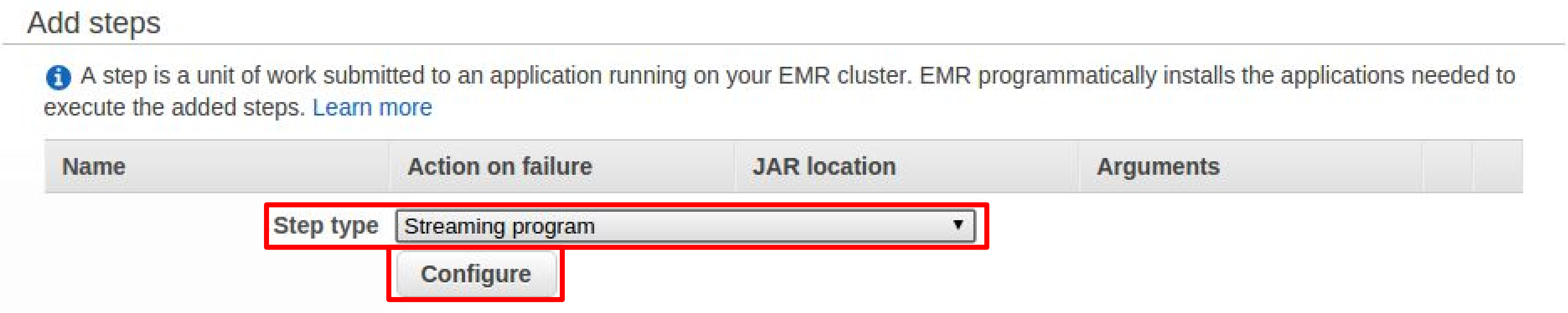
s3://<bucket-name>/<folder>/

Launch mode ○ Cluster ● Step execution

With Cluster, EMR creates a cluster with a set of specified applications. With Step execution, EMR will create a cluster, execute added steps and terminate when done.

Streaming step: Ruby, Perl, Python, PHP, or Bash

Chose **Streaming program** from the **Step type** dropdown



Add steps

A step is a unit of work submitted to an application running on your EMR cluster. EMR programmatically installs the applications needed to execute the added steps. Learn more

Name	Action on failure	JAR location	Arguments

Step type: Streaming program

Streaming step: Ruby, Perl, Python, PHP, or Bash

❑ Chose a unique name

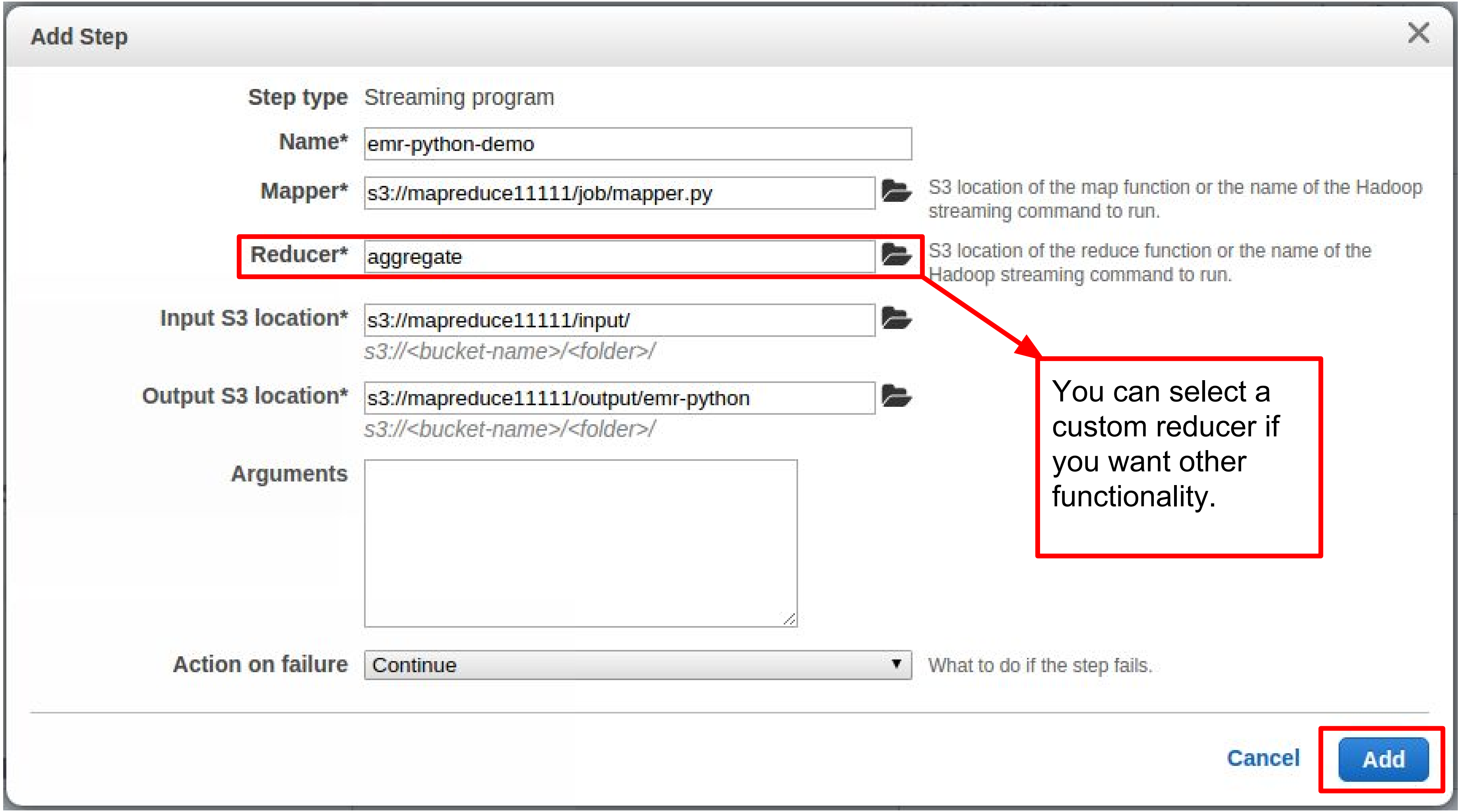
❑ Select the mapper program from your **job** folder, the one that you just uploaded

❑ Select your reducer from the same location, or you can use the keyword **aggregate**. Amazon EMR supports the special aggregate keyword. For more information, go to the Aggregate library supplied by Hadoop.

❑ Next choose your input S3 location, which is the **input** folder from your bucket

❑ The output location is the **output** folder from your bucket followed by a **unique name** that you have to type in. In this case you can use **erm-python** after the output folder: **s3://MapReduce11111/output/emr-python**

❑ Press **Add**



Add Step

Step type	Streaming program
Name*	emr-python-demo
Mapper*	s3://mapreduce11111/job/mapper.py
Reducer*	aggregate
Input S3 location*	s3://mapreduce11111/input/ s3://<bucket-name>/<folder>/
Output S3 location*	s3://mapreduce11111/output/emr-python s3://<bucket-name>/<folder>/
Arguments	
Action on failure	Continue

S3 location of the map function or the name of the Hadoop streaming command to run.

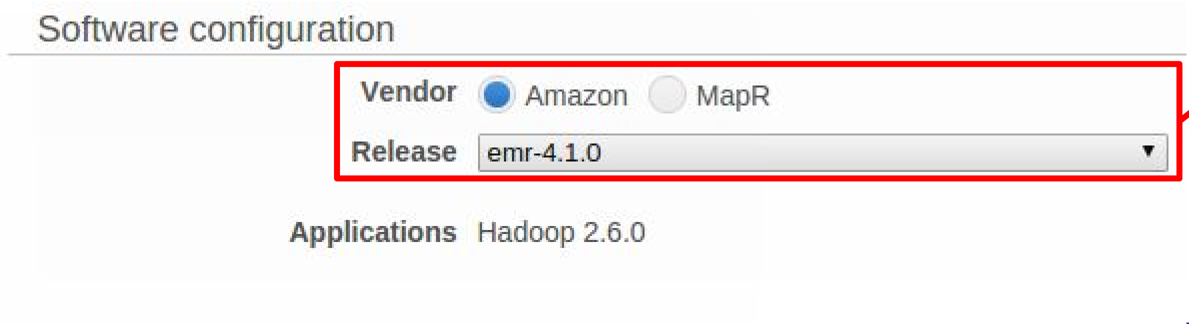
S3 location of the reduce function or the name of the Hadoop streaming command to run.

What to do if the step fails.

You can select a custom reducer if you want other functionality.

Software configuration

- ❏ Depending on how you chose to develop your MapReduce application, choose the appropriate **Vendor** and **Release**. For more info, please visit: http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-plan-hadoop-differences.html
- ❏ If you don't know what to choose, leave the default.



Software configuration

Vendor ● Amazon ○ MapR

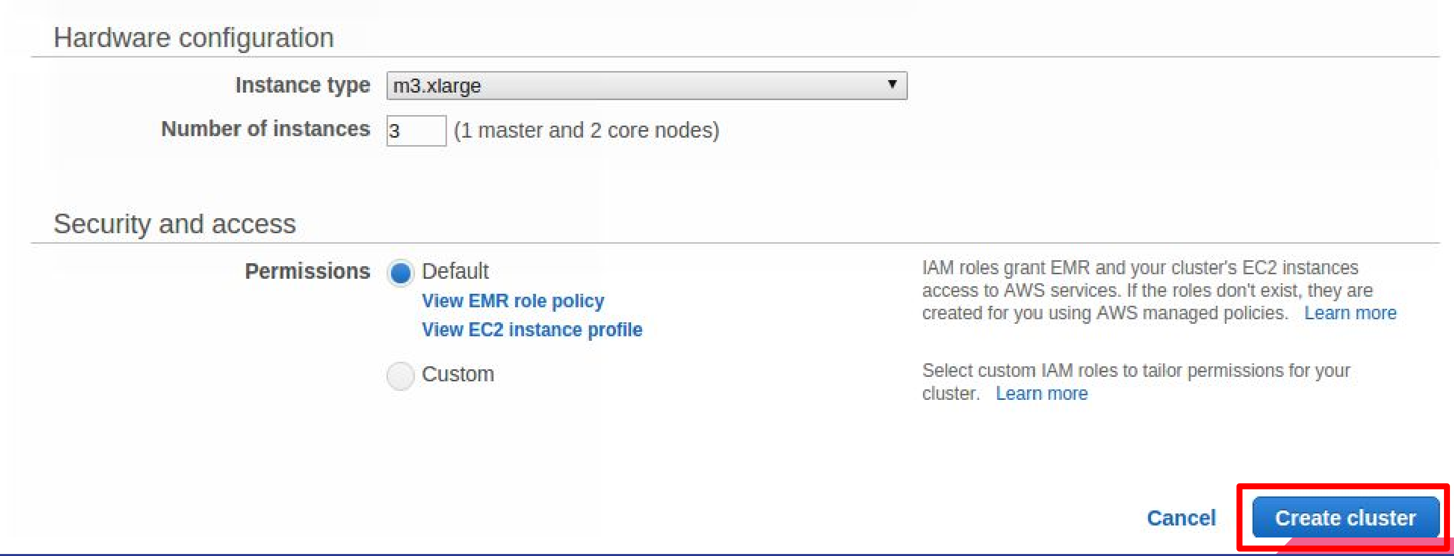
Release emr-4.1.0 ▾

Applications Hadoop 2.6.0

You can use the default (latest) version here, thus you do not need to change it with the version on this slide.

Hardware configuration and **Security**

- Choose your instance types, depending on your application requirements.
- Here is some information on the new instance types introduced in EMR: https://aws.amazon.com/blogs/aws/new-instance-types-for-amazon...



Hardware configuration

Instance type m3.xlarge

Number of instances 3 (1 master and 2 core nodes)

Security and access

Permissions

● Default

View EMR role policy

View EC2 instance profile

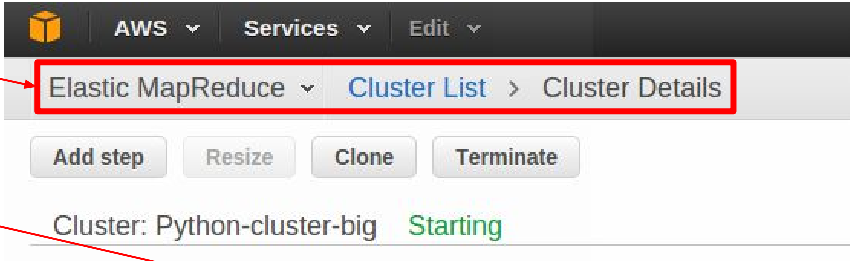
IAM roles grant EMR and your cluster's EC2 instances access to AWS services. If the roles don't exist, they are created for you using AWS managed policies. Learn more

○ Custom

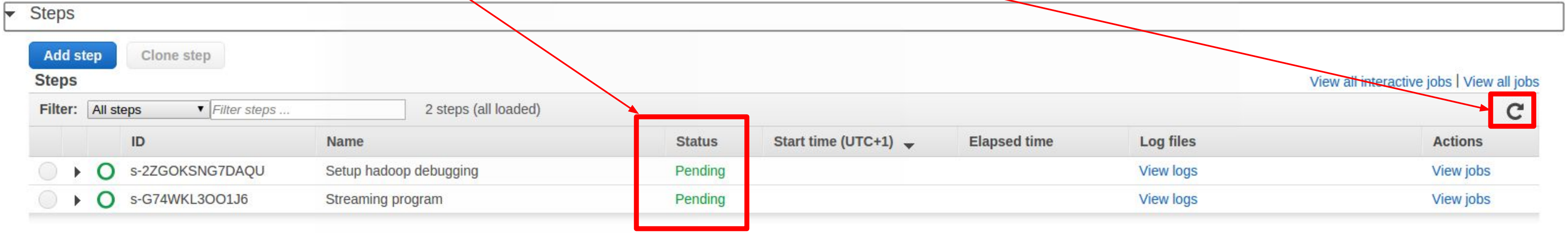
Select custom IAM roles to tailor permissions for your cluster. Learn more

Check status and wait for completion

- Go to **Steps** under **Cluster details**, for the cluster that you just created
- Here you can see the progress of the cluster
- Refresh to update the status



Cluster: Python-cluster-big Starting

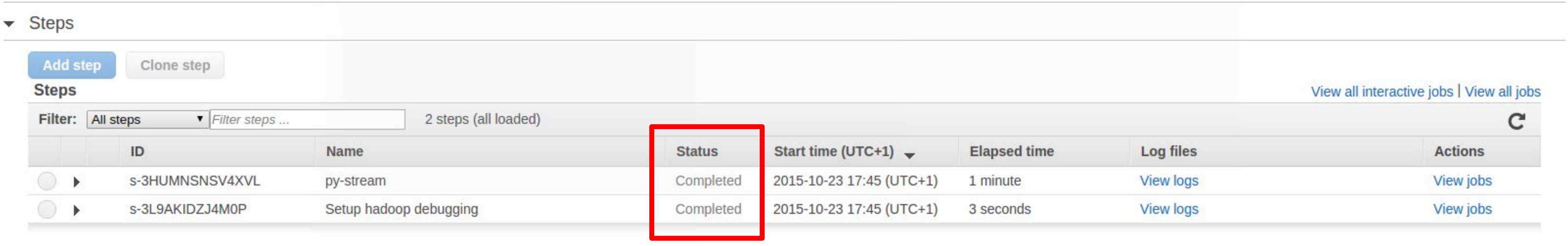


Steps

Filter: All steps ▾ Filter steps ... 2 steps (all loaded)

			ID	Name	Status	Start time (UTC+1)	Elapsed time	Log files	Actions
○	▸	○	s-2ZGOKSNG7DAQU	Setup hadoop debugging	Pending			View logs	View jobs
○	▸	○	s-G74WKL3OO1J6	Streaming program	Pending			View logs	View jobs

Wait until status changes to Completed

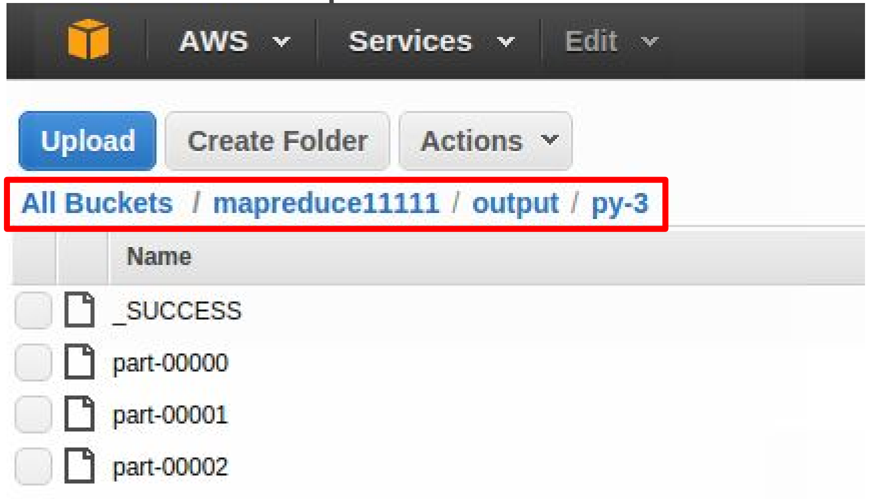


Steps

		ID	Name	Status	Start time (UTC+1)	Elapsed time	Log files	Actions
○	▸	s-3HUMNSNSV4XVL	py-stream	Completed	2015-10-23 17:45 (UTC+1)	1 minute	View logs	View jobs
○	▸	s-3L9AKIDZJ4M0P	Setup hadoop debugging	Completed	2015-10-23 17:45 (UTC+1)	3 seconds	View logs	View jobs

SUCCESS !!!

- ❑ Go back to the **Storage** and **Content Delivery > S3**
- ❑ Select your **S3 bucket**, the output folder and the unique name you chose when you created the cluster
- ❑ You should see here all output files

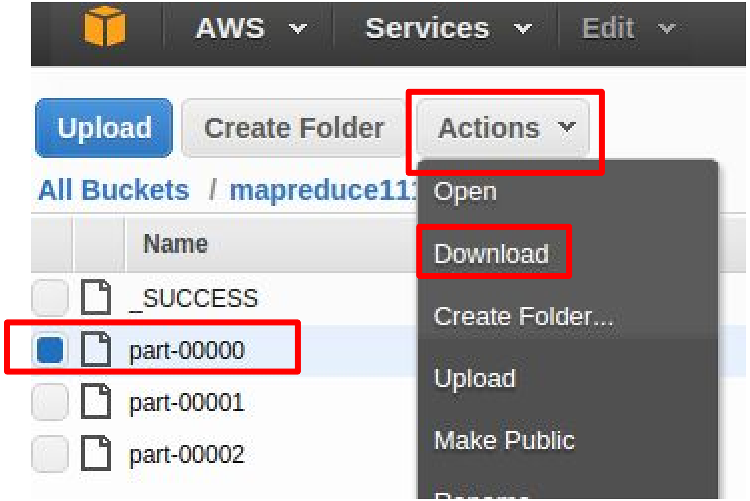


All Buckets / mapreduce11111 / output / py-3

- _SUCCESS
- part-00000
- part-00001
- part-00002

Download or process further

- You can now select the file that you want, press on Actions and choose Download
- Or you can reuse these in a new MapReduce program



Congratulations, you have just successfully executed your first EMR program on AWS

Now please do the same for the google ngrams dataset, using the provided mapper and reducer.

For detailed EMR documentation, please visit: http://docs.aws.amazon.com/ElasticMap...

Java local development (your machine) in Eclipse

- ❏ For Java lovers, the following tutorial can help you get started: http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-common-programming-sample.html
  Please follow exactly all the steps to have the desired outcome.
- ❏ To carry out JUnit testing for your MapReduce code, please have a look at: https://cwiki.apache.org/confluence/display/MRUNIT/MRUnit+Tutorial
  which is a handy tip given by a fellow student Pavel Reich.

Useful links

- https://pythonhosted.org/mrjob/
- https://boto3.readthedocs.io/en/latest/
- http://hortonworks.com/products/sandbox/#downloads
- https://www.javacodegeeks.com/2015/03/running-pagerank-hadoop-job-on-aws-elastic-mapreduce.html