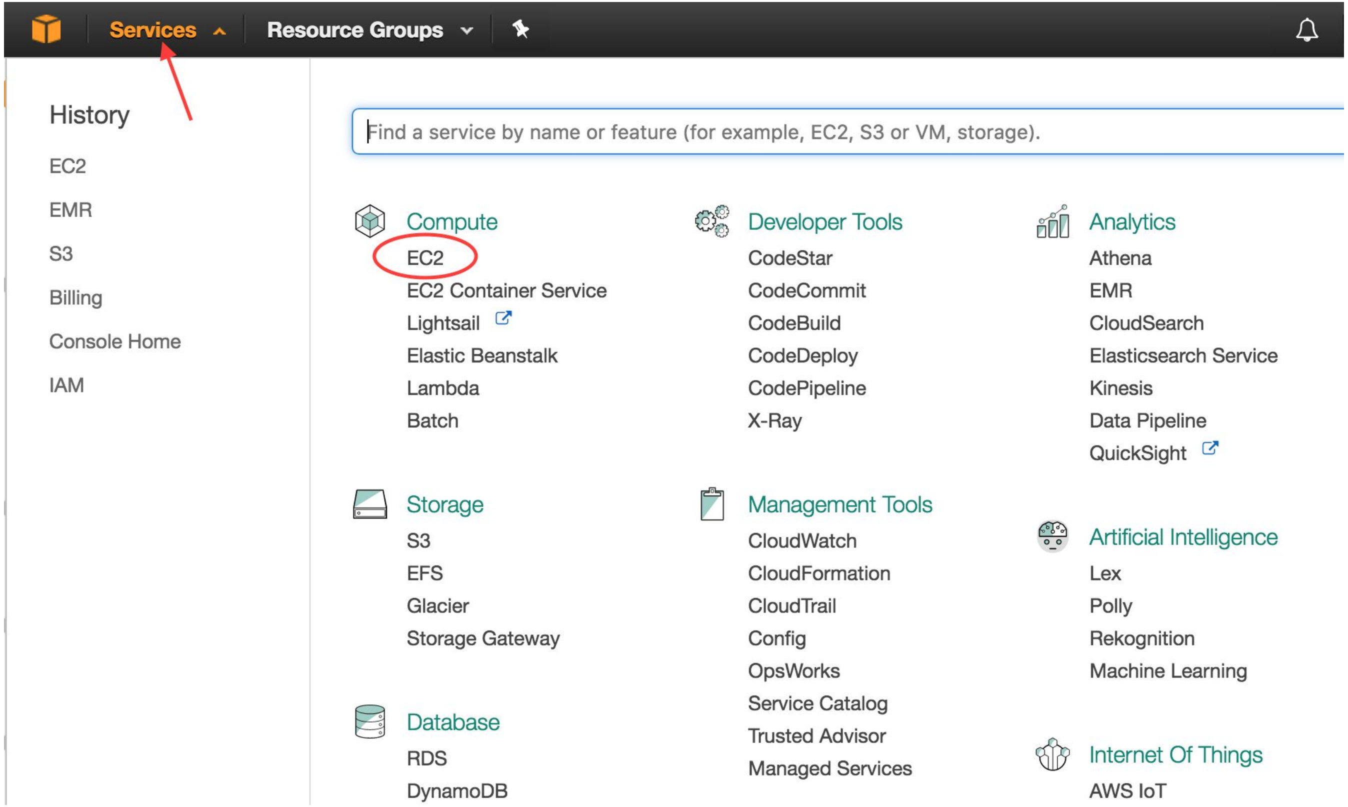
Запуск приложения Apache Spark на облачном сервисе Amazon EC2

Мурашкин Вячеслав

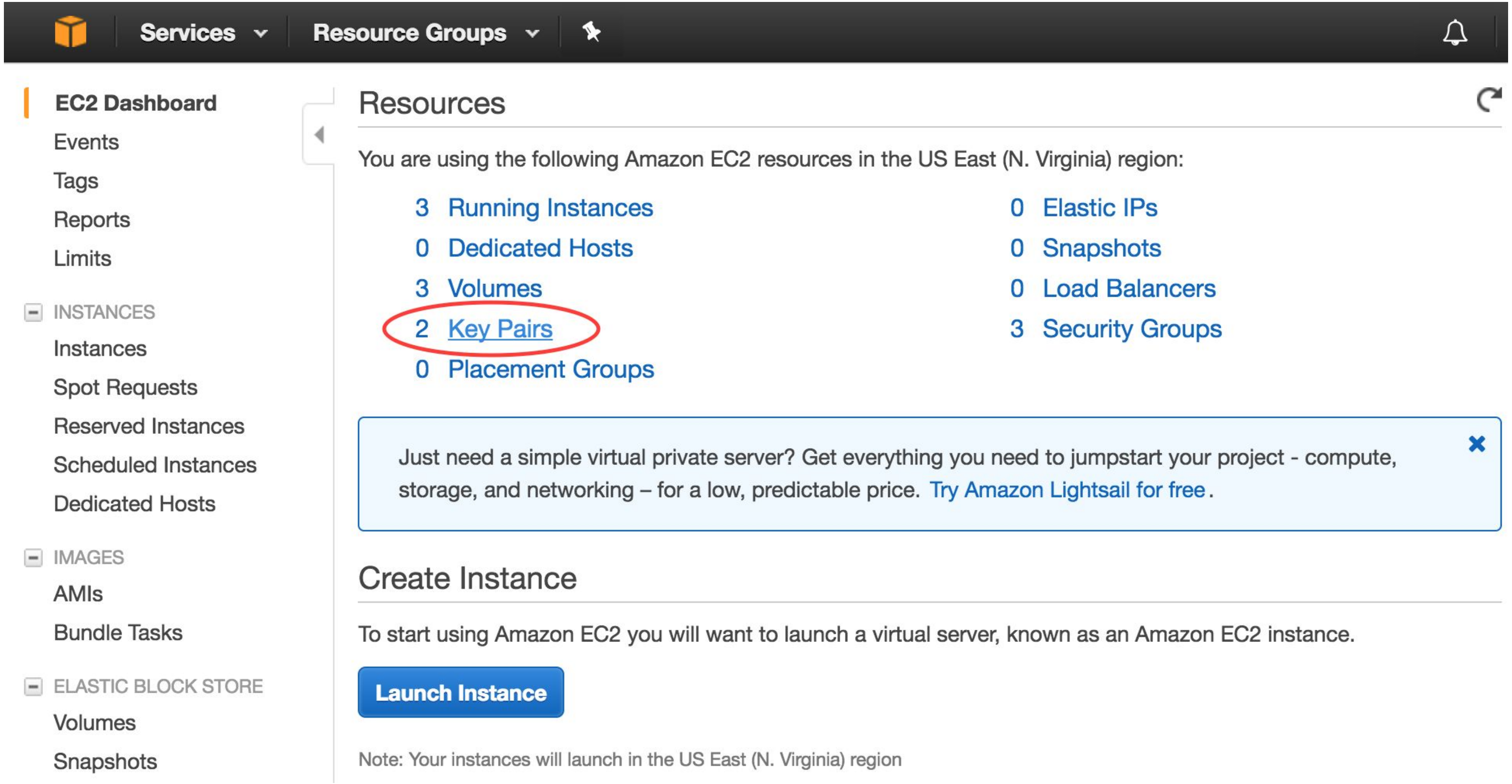
2017

https://github.com/a4tunado/lectures-hse-spark/tree/master/aws

1. Создание ключа для авторизации



1. Создание ключа для авторизации



Resources

You are using the following Amazon EC2 resources in the US East (N. Virginia) region:

3 Running Instances	0 Elastic IPs
0 Dedicated Hosts	0 Snapshots
3 Volumes	0 Load Balancers
2 Key Pairs	3 Security Groups
0 Placement Groups	

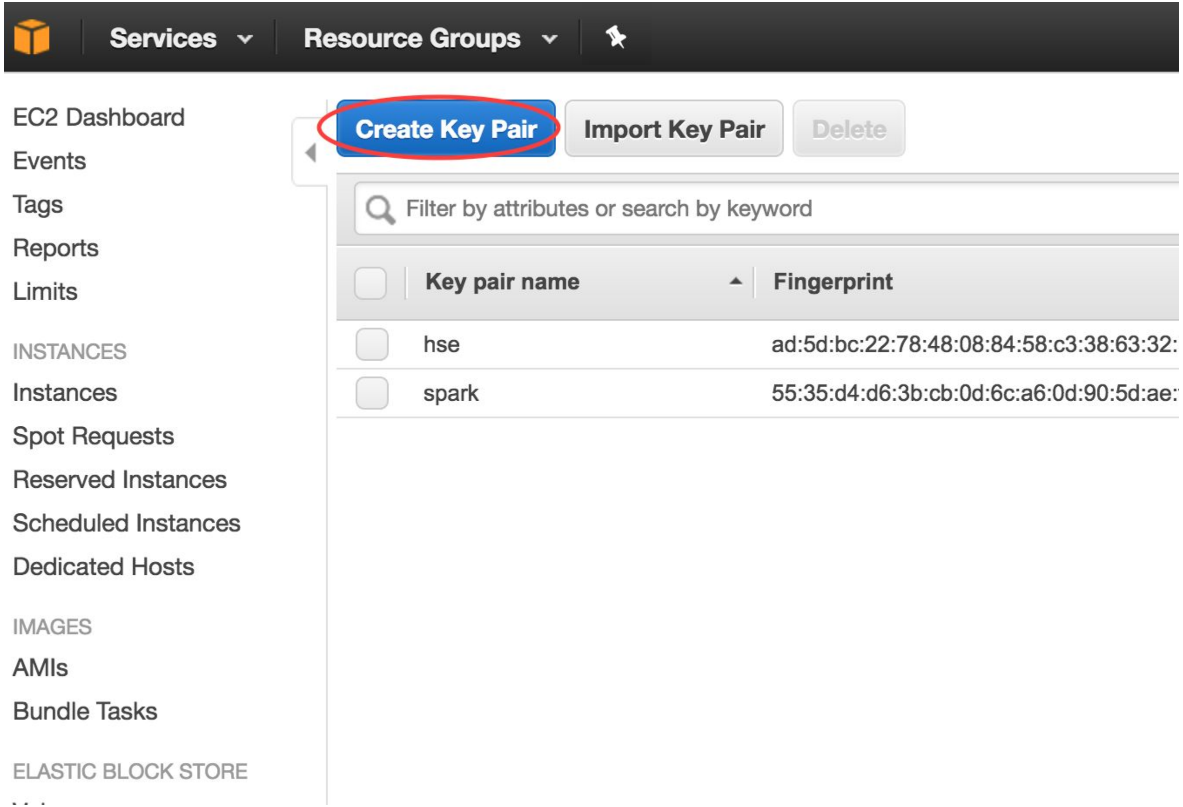
Just need a simple virtual private server? Get everything you need to jumpstart your project - compute, storage, and networking – for a low, predictable price. Try Amazon Lightsail for free.

Create Instance

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

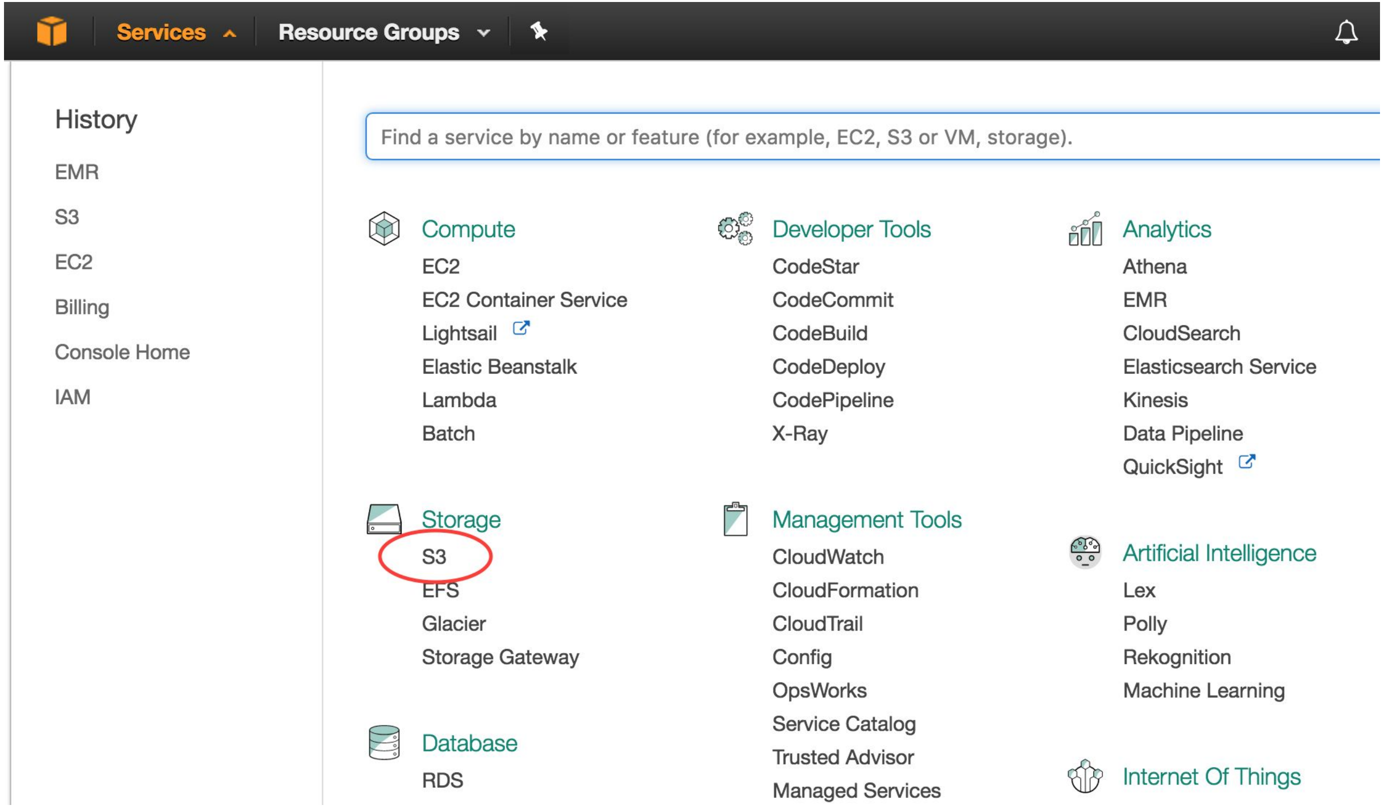
Note: Your instances will launch in the US East (N. Virginia) region

1. Создание ключа для авторизации



	Key pair name	Fingerprint
□	hse	ad:5d:bc:22:78:48:08:84:58:c3:38:63:32:
□	spark	55:35:d4:d6:3b:cb:0d:6c:a6:0d:90:5d:ae:

2. Создание хранилища для загрузки данных



Services ^ | Resource Groups ⌄

History

EMR

S3

EC2

Billing

Console Home

IAM

Find a service by name or feature (for example, EC2, S3 or VM, storage).

Compute

EC2

EC2 Container Service

Lightsail

Elastic Beanstalk

Lambda

Batch

Storage

S3

EFS

Glacier

Storage Gateway

Database

RDS

Developer Tools

CodeStar

CodeCommit

CodeBuild

CodeDeploy

CodePipeline

X-Ray

Management Tools

CloudWatch

CloudFormation

CloudTrail

Config

OpsWorks

Service Catalog

Trusted Advisor

Managed Services

Analytics

Athena

EMR

CloudSearch

Elasticsearch Service

Kinesis

Data Pipeline

QuickSight

Artificial Intelligence

Lex

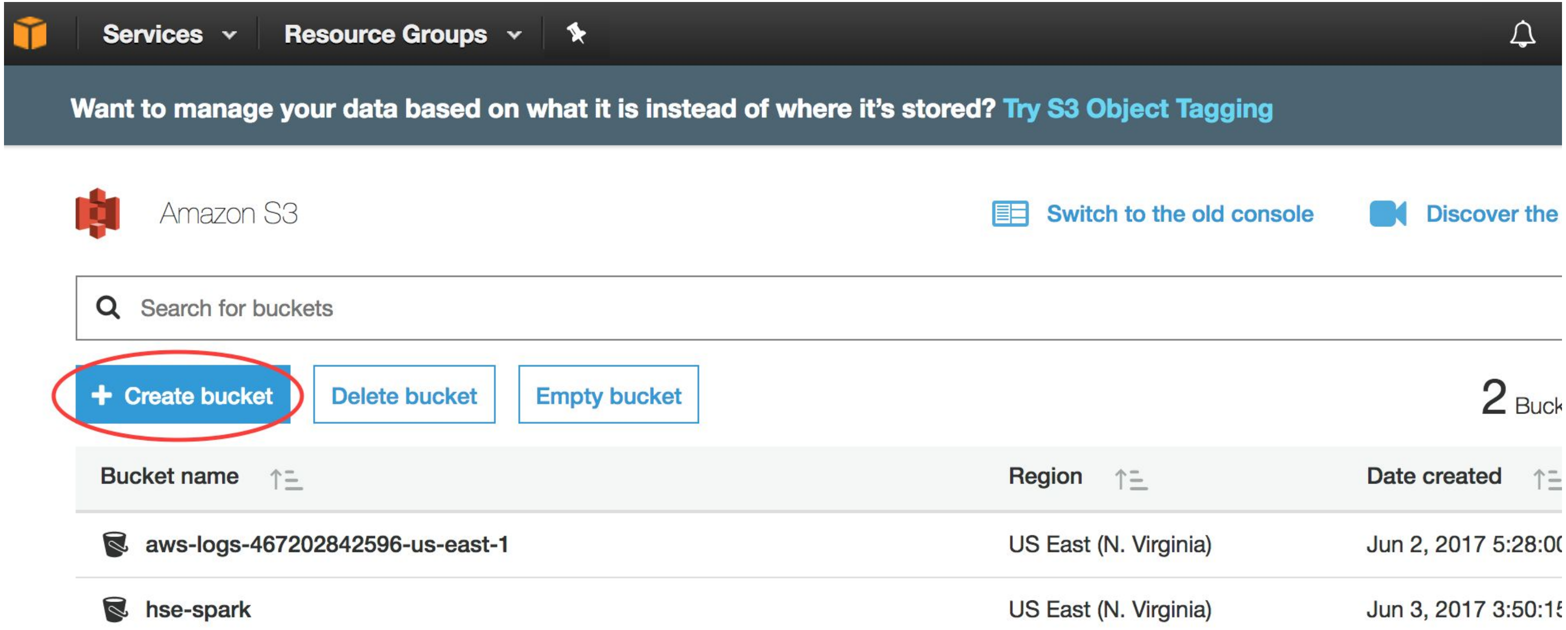
Polly

Rekognition

Machine Learning

Internet Of Things

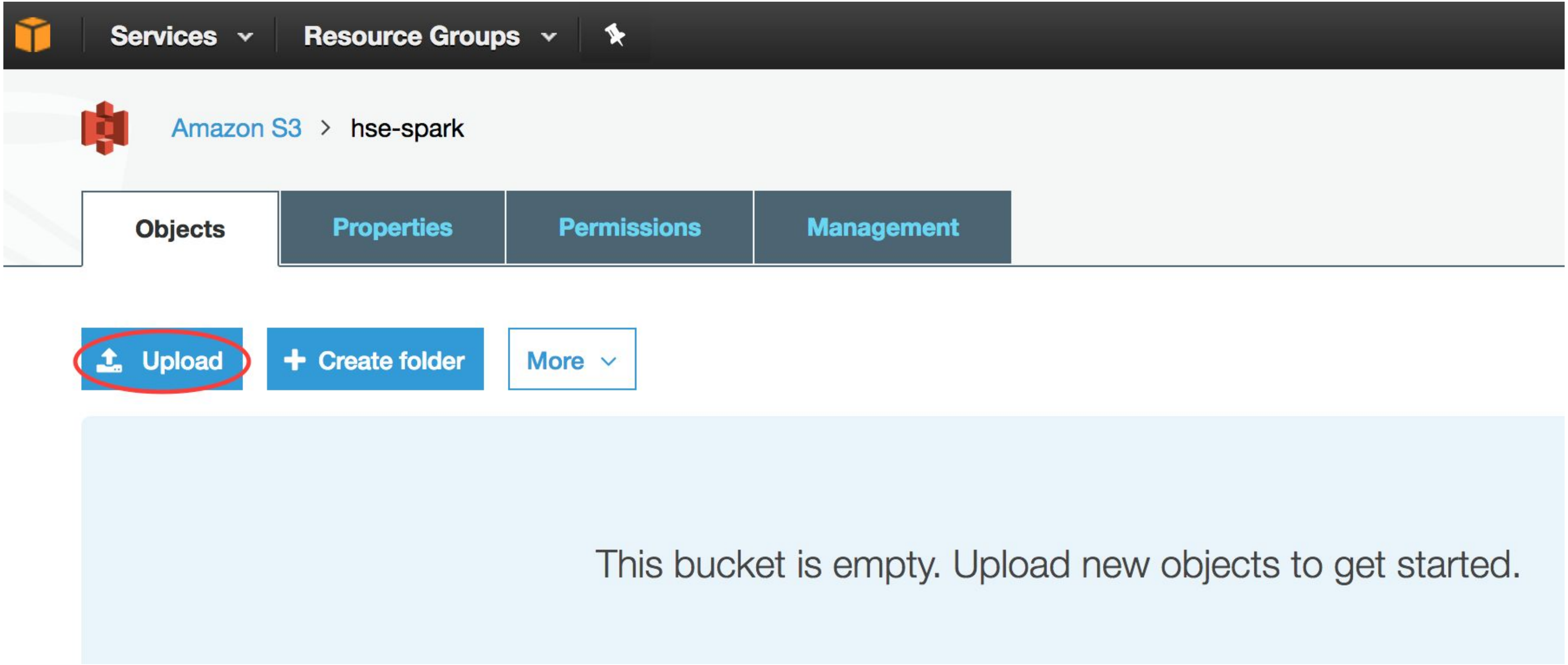
2. Создание хранилища для загрузки данных



Want to manage your data based on what it is instead of where it's stored? Try S3 Object Tagging

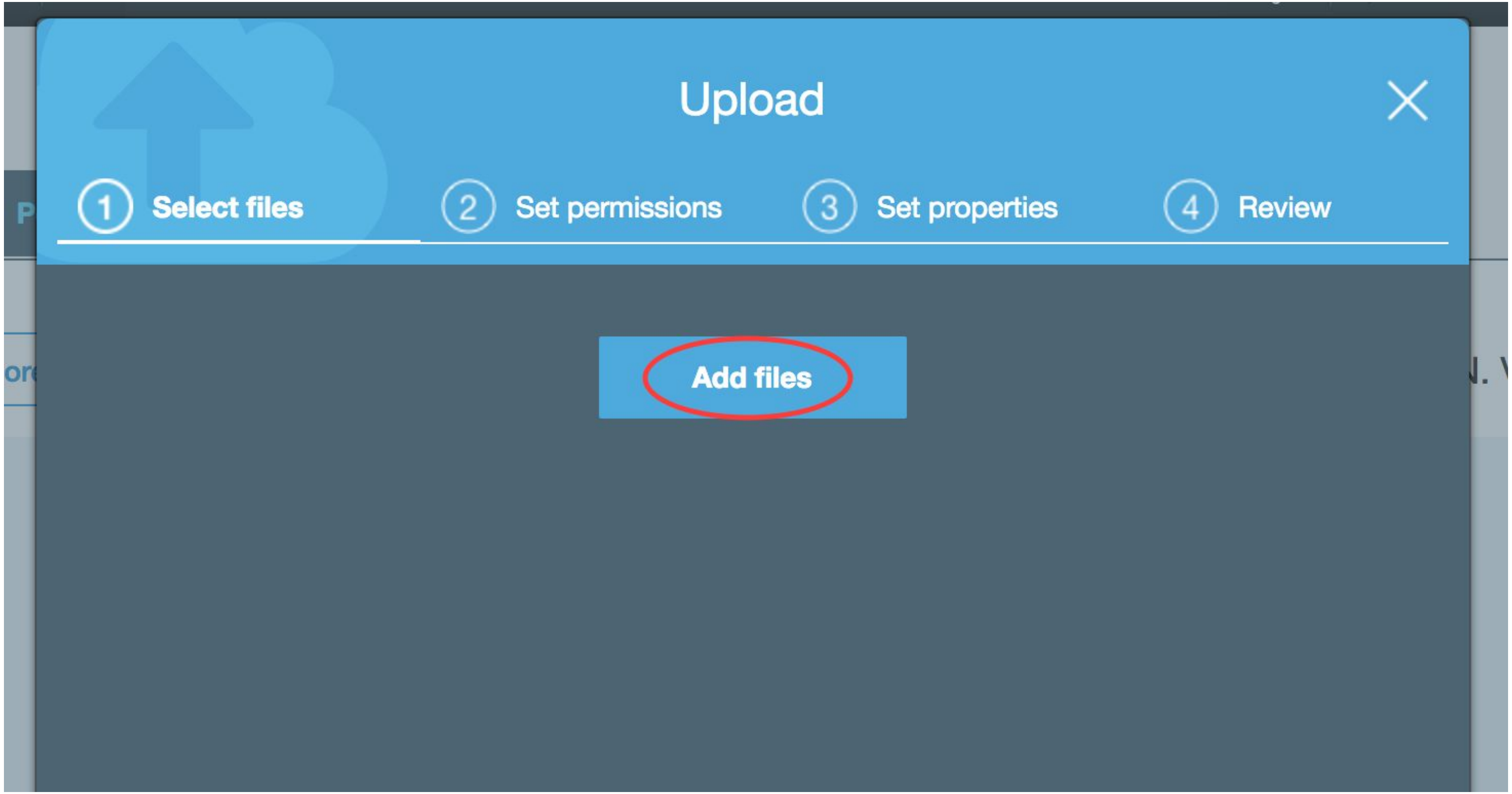
Bucket name	Region	Date created
aws-logs-467202842596-us-east-1	US East (N. Virginia)	Jun 2, 2017 5:28:0
hse-spark	US East (N. Virginia)	Jun 3, 2017 3:50:1

2. Создание хранилища для загрузки данных

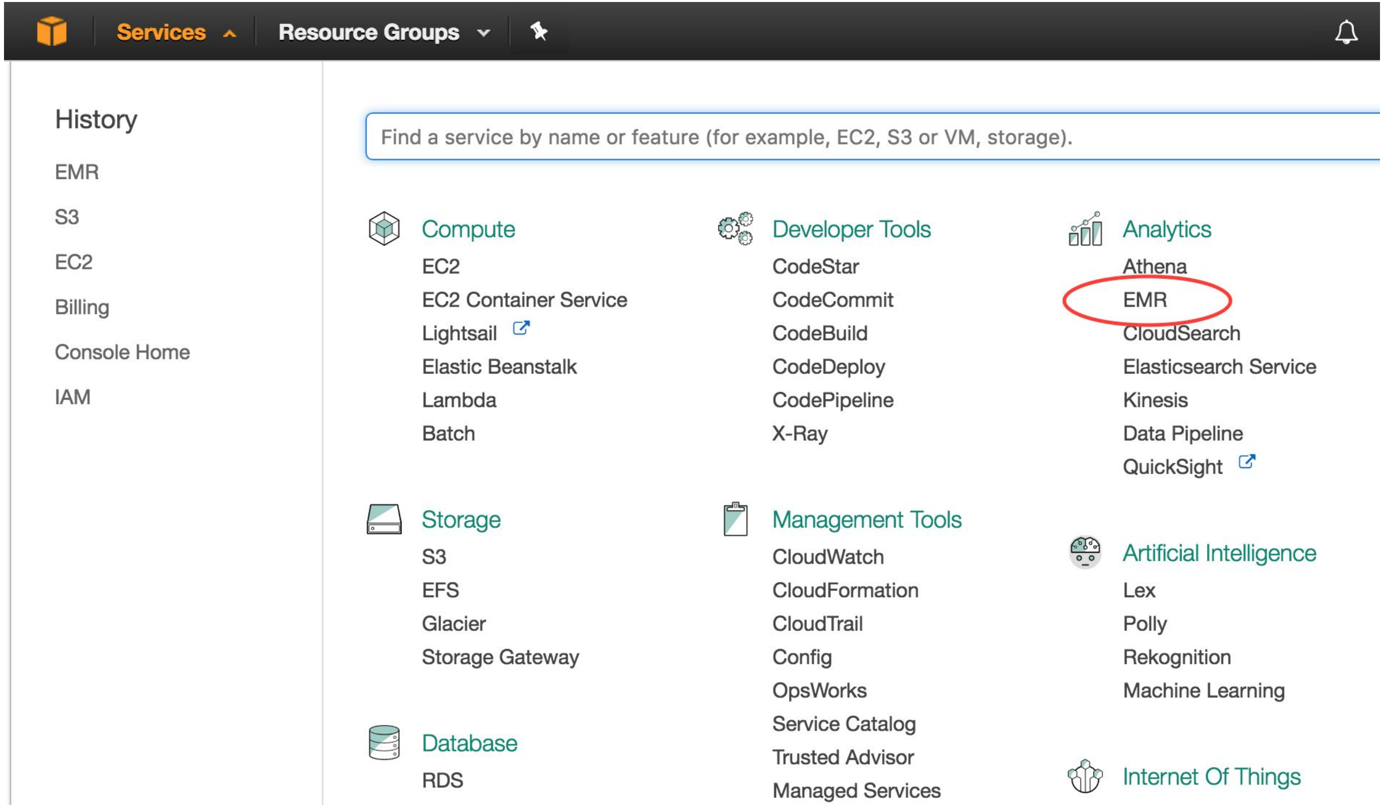


This bucket is empty. Upload new objects to get started.

2. Создание хранилища для загрузки данных



3. Создание EMR кластера



Services
Resource Groups

History
EMR
S3
EC2
Billing
Console Home
IAM

Find a service by name or feature (for example, EC2, S3 or VM, storage).

Compute
EC2
EC2 Container Service
Lightsail
Elastic Beanstalk
Lambda
Batch

Storage
S3
EFS
Glacier
Storage Gateway

Database
RDS

Developer Tools
CodeStar
CodeCommit
CodeBuild
CodeDeploy
CodePipeline
X-Ray

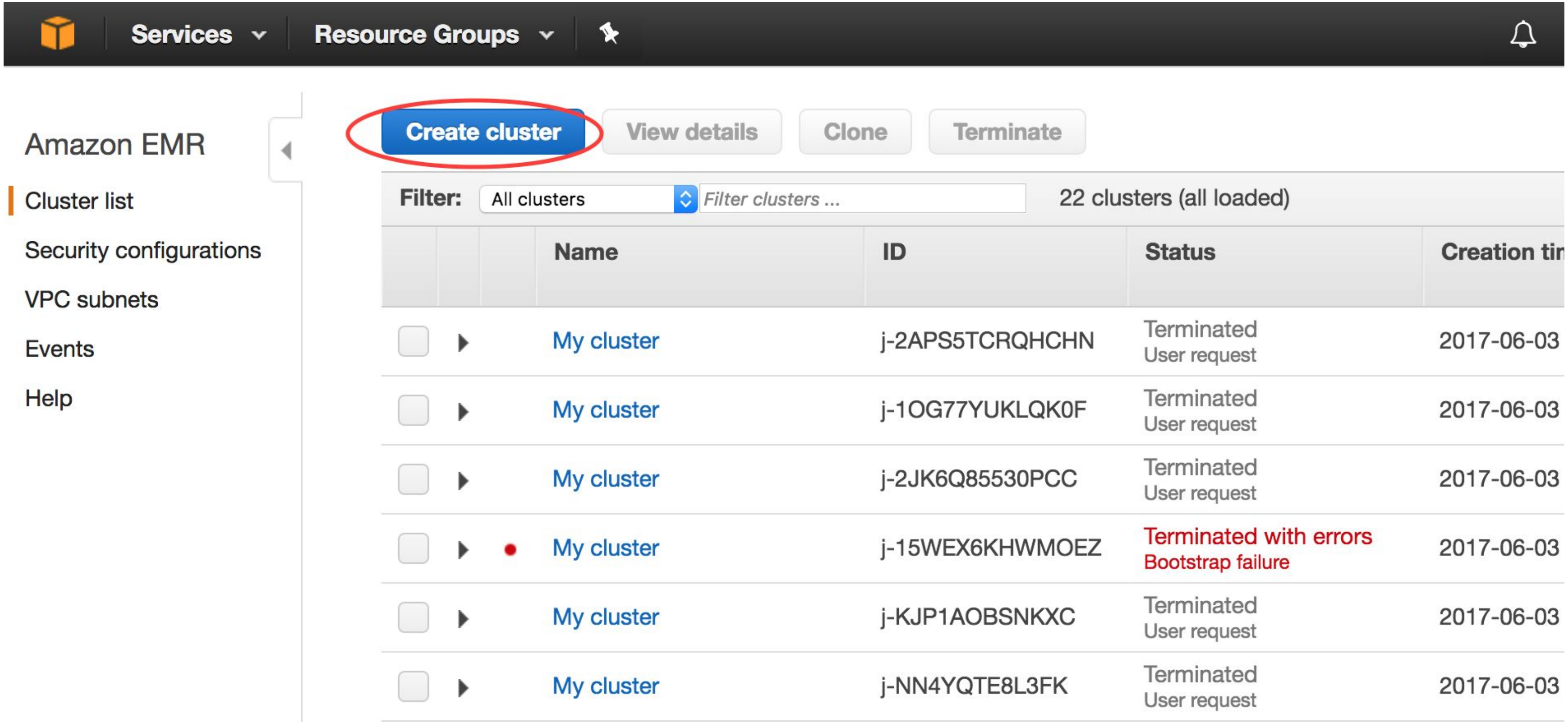
Management Tools
CloudWatch
CloudFormation
CloudTrail
Config
OpsWorks
Service Catalog
Trusted Advisor
Managed Services

Analytics
Athena
EMR
CloudSearch
Elasticsearch Service
Kinesis
Data Pipeline
QuickSight

Artificial Intelligence
Lex
Polly
Rekognition
Machine Learning

Internet Of Things

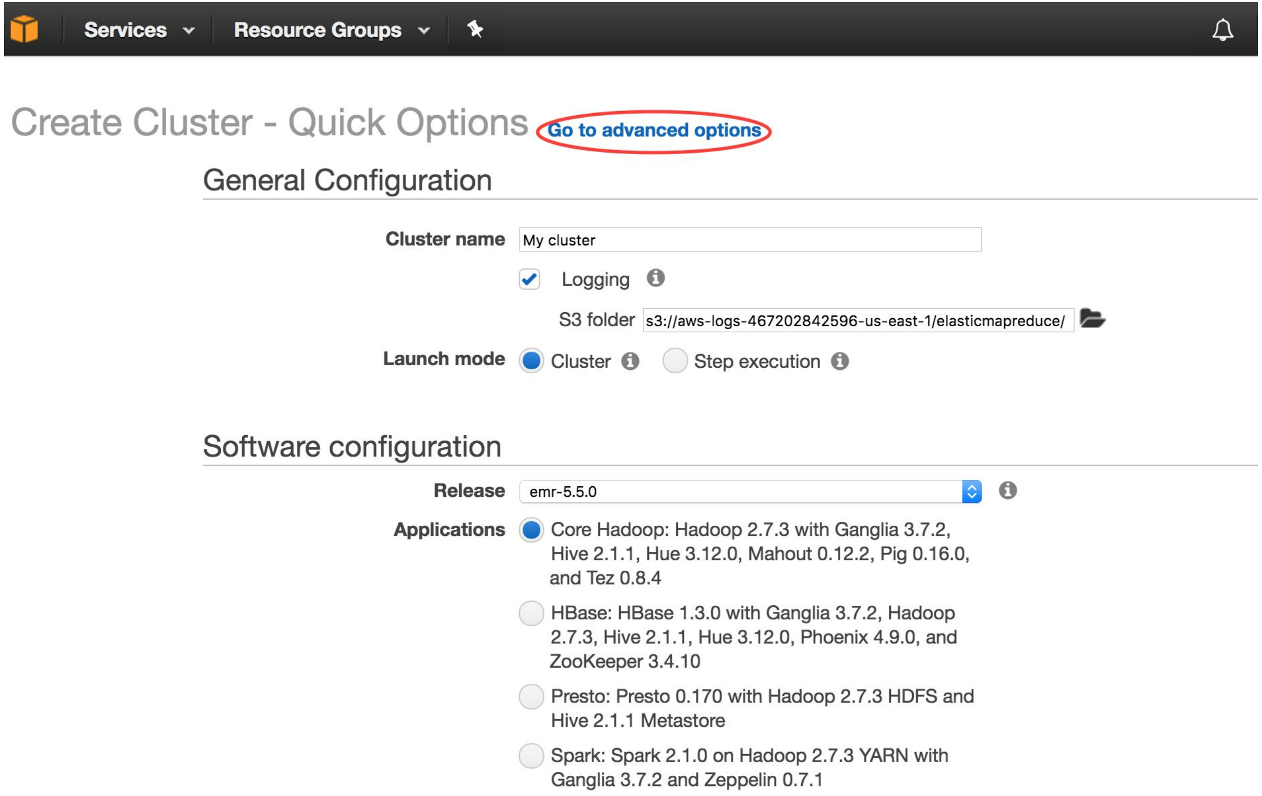
3. Создание EMR кластера



Filter: All clusters Filter clusters ... 22 clusters (all loaded)

			Name	ID	Status	Creation tir
□	▶		My cluster	j-2APS5TCRQHCHN	Terminated User request	2017-06-03
□	▶		My cluster	j-1OG77YUKLQK0F	Terminated User request	2017-06-03
□	▶		My cluster	j-2JK6Q85530PCC	Terminated User request	2017-06-03
□	▶	●	My cluster	j-15WEX6KHWMOEZ	Terminated with errors Bootstrap failure	2017-06-03
□	▶		My cluster	j-KJP1AOBSNKXC	Terminated User request	2017-06-03
□	▶		My cluster	j-NN4YQTE8L3FK	Terminated User request	2017-06-03

3. Создание EMR кластера



Create Cluster - Quick Options

Go to advanced options

General Configuration

Cluster name My cluster

☑ Logging

S3 folder s3://aws-logs-467202842596-us-east-1/elasticmapreduce/

Launch mode ● Cluster ○ Step execution

Software configuration

Release emr-5.5.0

Applications

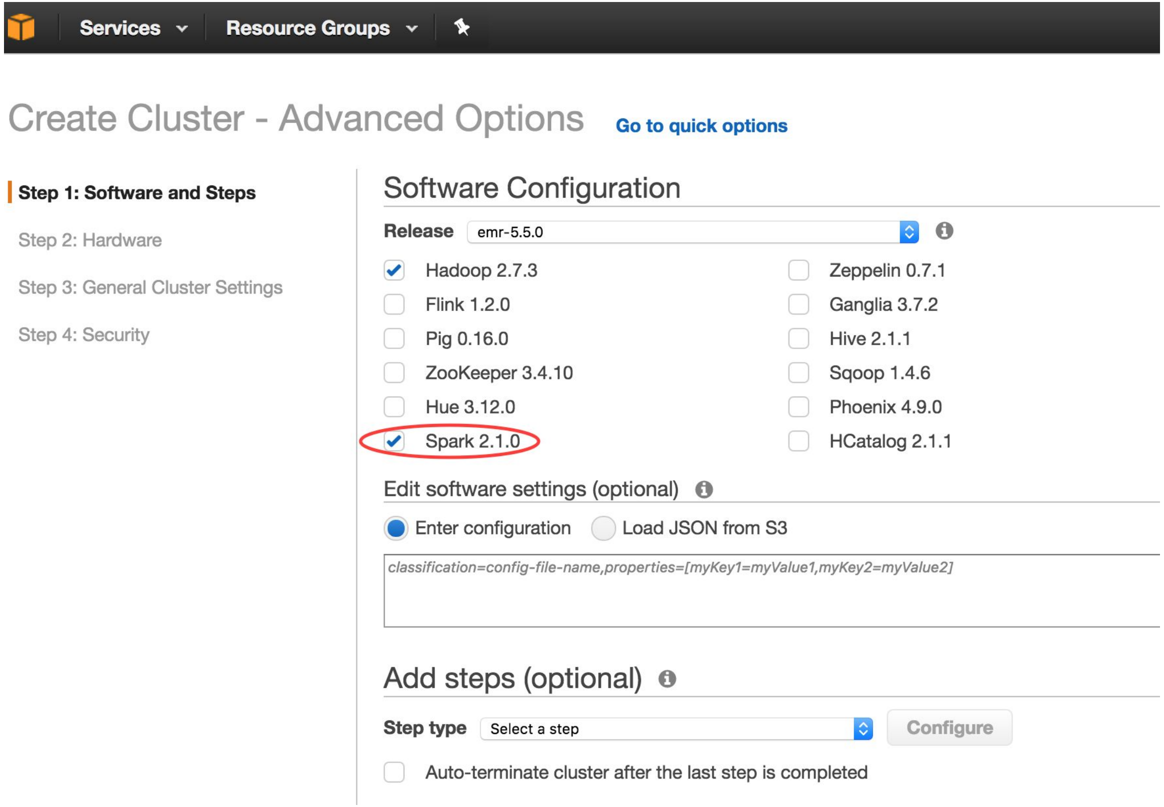
● Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.1.1, Hue 3.12.0, Mahout 0.12.2, Pig 0.16.0, and Tez 0.8.4

○ HBase: HBase 1.3.0 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 2.1.1, Hue 3.12.0, Phoenix 4.9.0, and ZooKeeper 3.4.10

○ Presto: Presto 0.170 with Hadoop 2.7.3 HDFS and Hive 2.1.1 Metastore

○ Spark: Spark 2.1.0 on Hadoop 2.7.3 YARN with Ganglia 3.7.2 and Zeppelin 0.7.1

3. Создание EMR кластера



Create Cluster - Advanced Options

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release: emr-5.5.0

☑ Hadoop 2.7.3
□ Flink 1.2.0
□ Pig 0.16.0
□ ZooKeeper 3.4.10
□ Hue 3.12.0
☑ Spark 2.1.0

□ Zeppelin 0.7.1
□ Ganglia 3.7.2
□ Hive 2.1.1
□ Sqoop 1.4.6
□ Phoenix 4.9.0
□ HCatalog 2.1.1

Edit software settings (optional)

● Enter configuration ○ Load JSON from S3

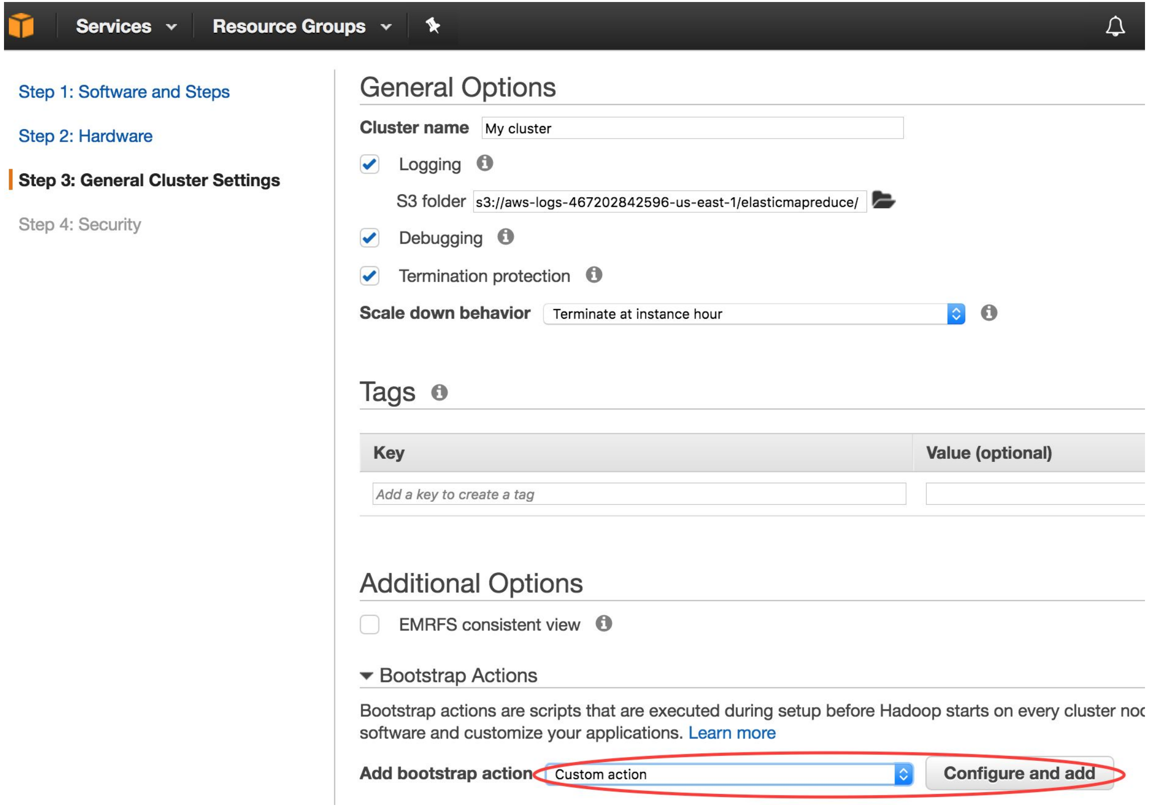
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]

Add steps (optional)

Step type: Select a step

□ Auto-terminate cluster after the last step is completed

3. Создание EMR кластера



Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

General Options

Cluster name My cluster

☑ Logging

S3 folder s3://aws-logs-467202842596-us-east-1/elasticmapreduce/

☑ Debugging

☑ Termination protection

Scale down behavior Terminate at instance hour

Tags

Key	Value (optional)
Add a key to create a tag	

Additional Options

□ EMRFS consistent view

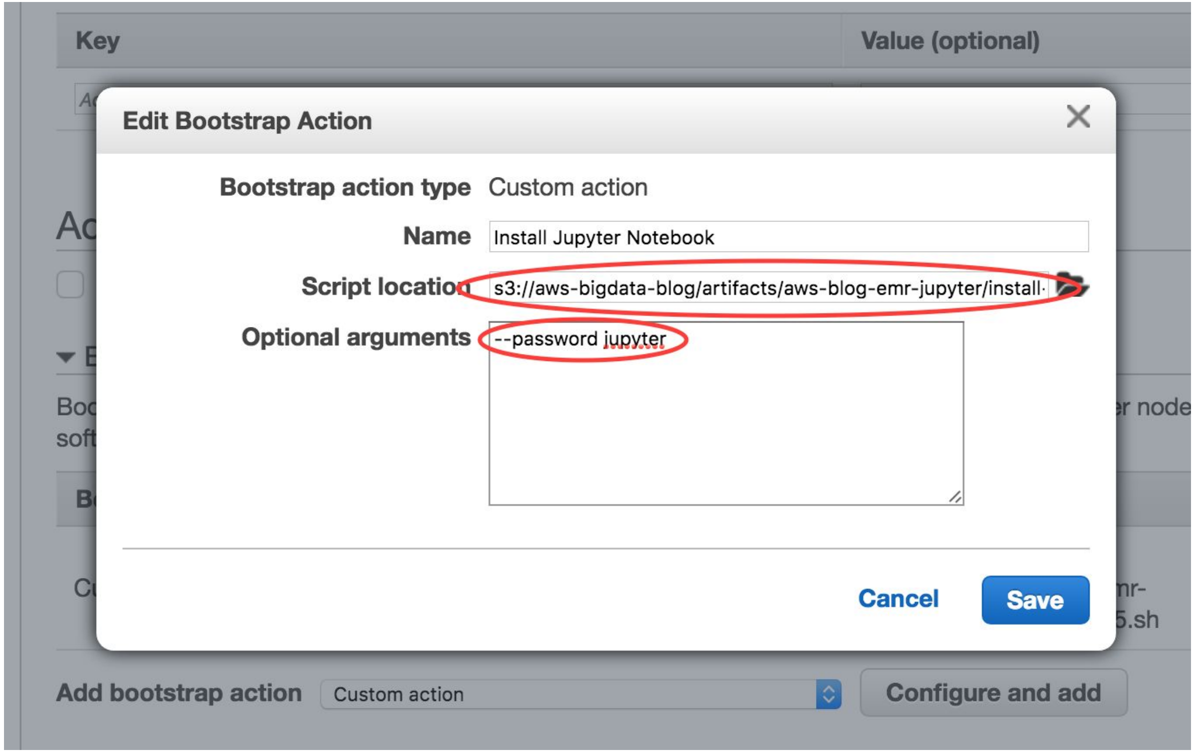
Bootstrap Actions

Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster noc
software and customize your applications. Learn more

Add bootstrap action Custom action **Configure and add**

3. Создание EMR кластера

s3://aws-bigdata-blog/artifacts/aws-blog-emr-jupyter/install-jupyter-emr5.sh



Bootstrap action type Custom action

Name: Install Jupyter Notebook

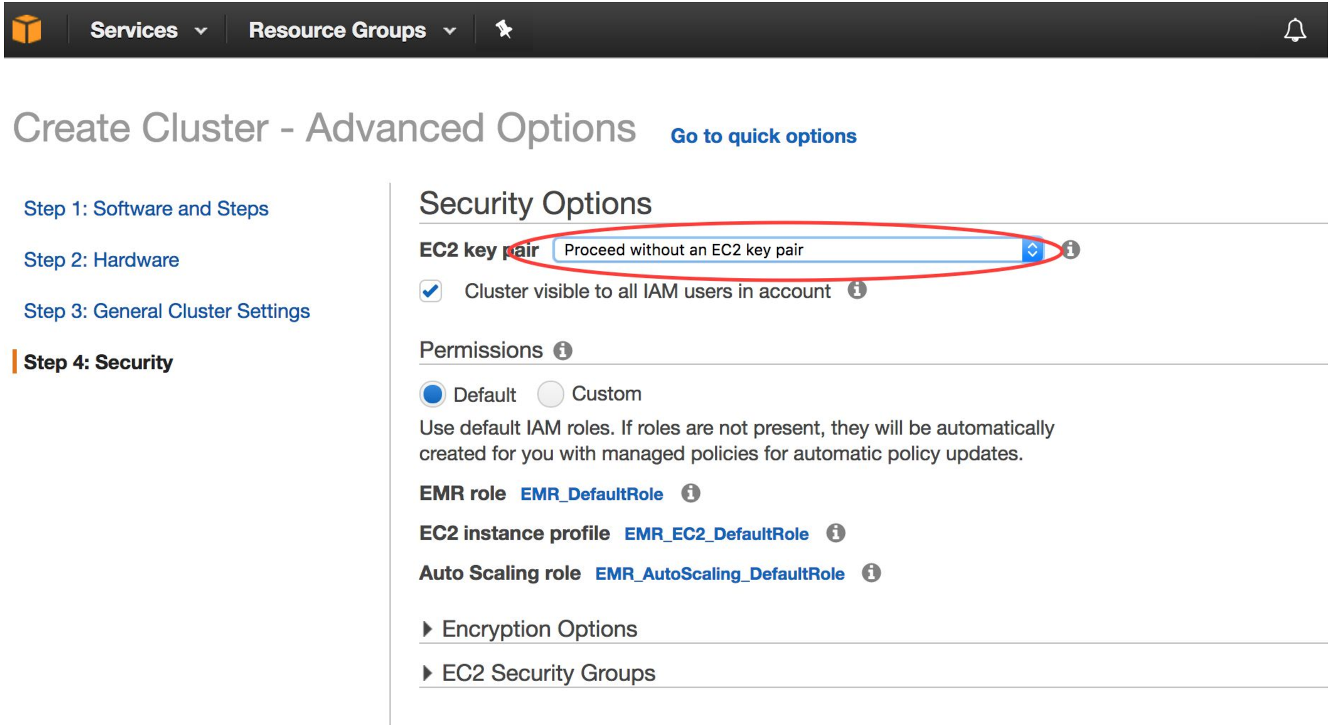
Script location: s3://aws-bigdata-blog/artifacts/aws-blog-emr-jupyter/install-

Optional arguments: --password jupyter

Add bootstrap action: Custom action

3. Создание EMR кластера

- Укажите имя ключа, созданного на первом шаге



Create Cluster - Advanced Options

Security Options

EC2 key pair: Proceed without an EC2 key pair

☑ Cluster visible to all IAM users in account

Permissions

● Default ○ Custom

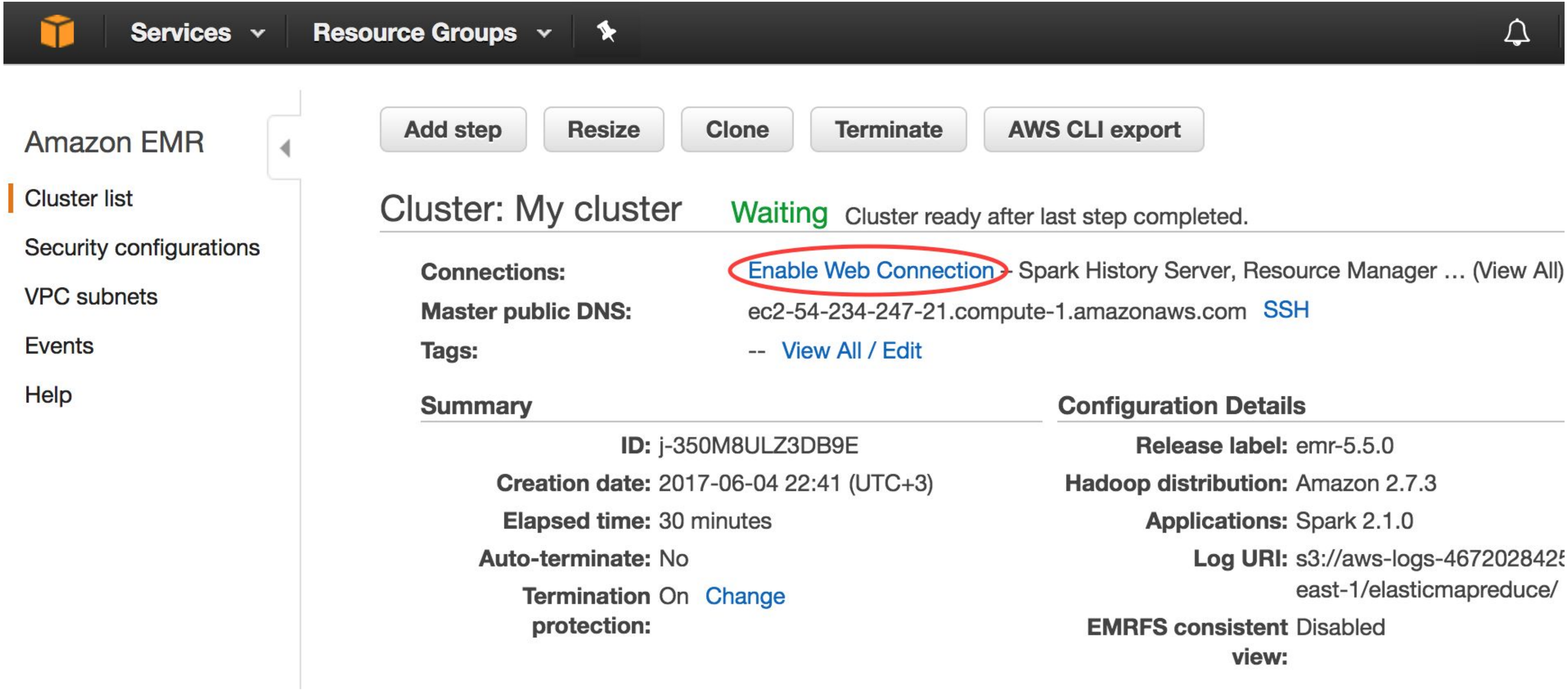
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role EMR_DefaultRole

EC2 instance profile EMR_EC2_DefaultRole

Auto Scaling role EMR_AutoScaling_DefaultRole

4. Настройка ssh прокси для подключения к EMR



Cluster: My cluster Waiting Cluster ready after last step completed.

Connections: Enable Web Connection – Spark History Server, Resource Manager ... (View All)

Master public DNS: ec2-54-234-247-21.compute-1.amazonaws.com SSH

Tags: -- View All / Edit

Summary

ID: j-350M8ULZ3DB9E

Creation date: 2017-06-04 22:41 (UTC+3)

Elapsed time: 30 minutes

Auto-terminate: No

Termination protection: On Change

Configuration Details

Release label: emr-5.5.0

Hadoop distribution: Amazon 2.7.3

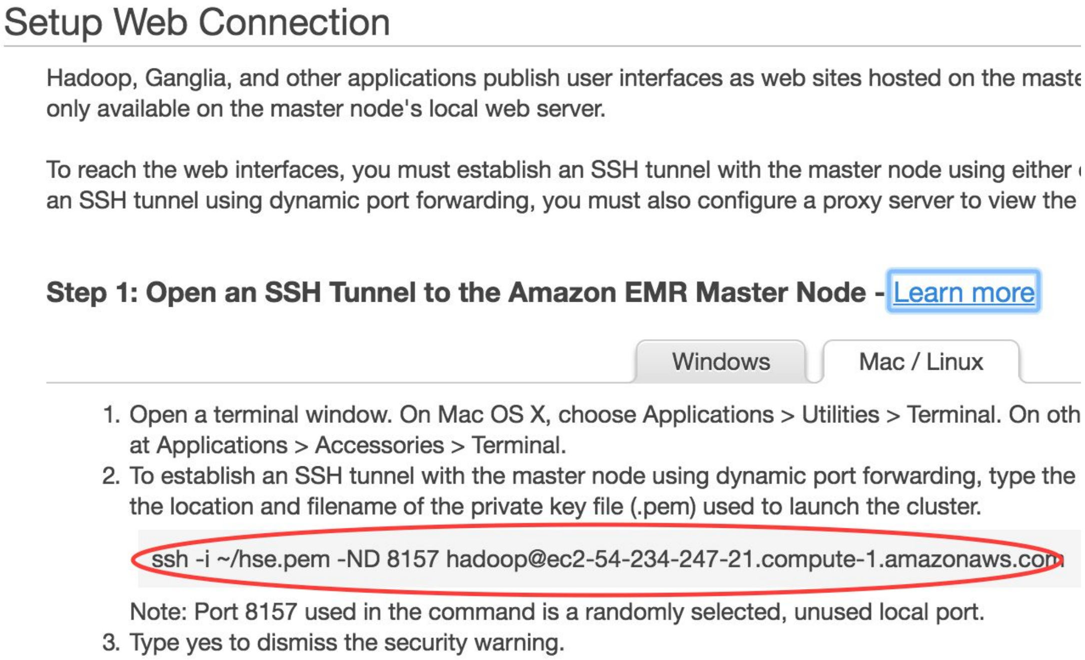
Applications: Spark 2.1.0

Log URI: s3://aws-logs-4672028425
east-1/elasticmapreduce/

EMRFS consistent view: Disabled

4. Настройка ssh прокси для подключения к EMR

- Перед запуском необходимо обновить настройки *.pem файла, выполнив команду: chmod 400 <key-file>.pem



Setup Web Connection

Hadoop, Ganglia, and other applications publish user interfaces as web sites hosted on the maste
only available on the master node's local web server.

To reach the web interfaces, you must establish an SSH tunnel with the master node using either
an SSH tunnel using dynamic port forwarding, you must also configure a proxy server to view the

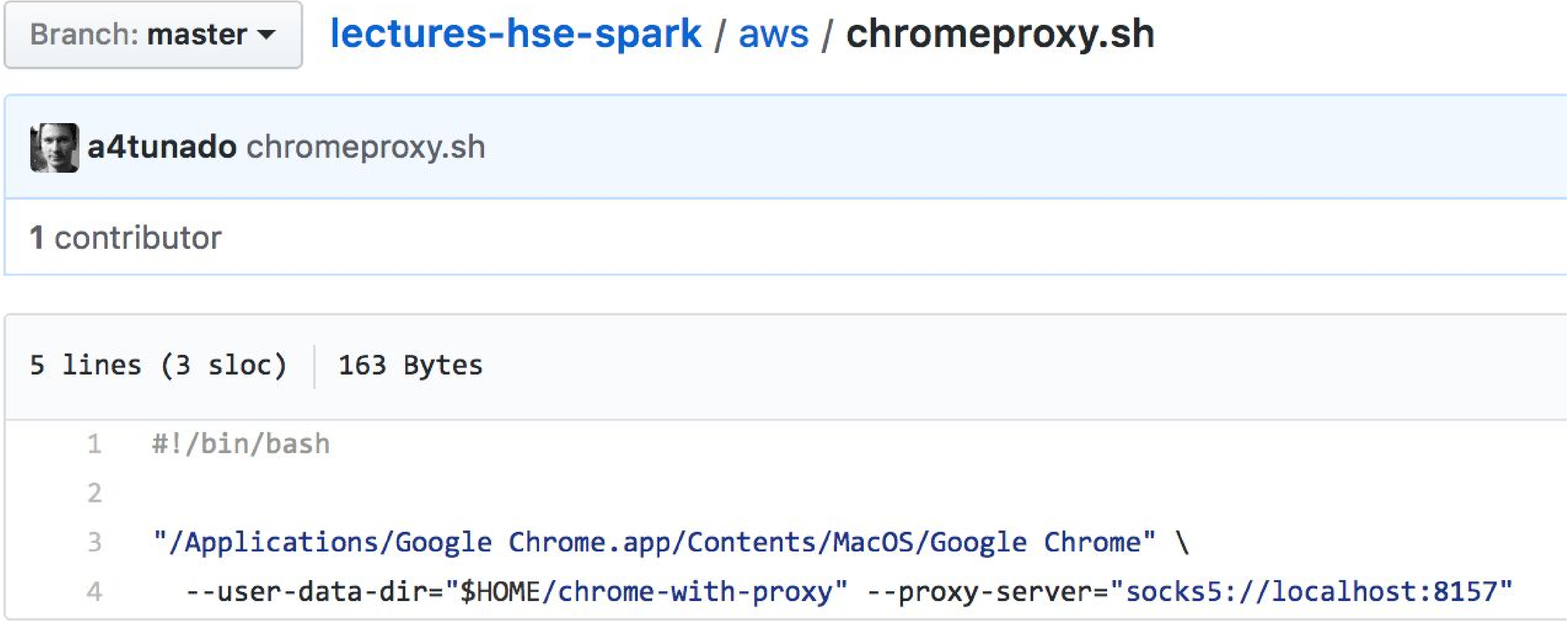
Step 1: Open an SSH Tunnel to the Amazon EMR Master Node - Learn more

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On oth
   at Applications > Accessories > Terminal.
2. To establish an SSH tunnel with the master node using dynamic port forwarding, type the
   the location and filename of the private key file (.pem) used to launch the cluster.

   ```
   ssh -i ~/hse.pem -ND 8157 hadoop@ec2-54-234-247-21.compute-1.amazonaws.com
   ```

   Note: Port 8157 used in the command is a randomly selected, unused local port.
3. Type yes to dismiss the security warning.

4. Настройка ssh прокси для подключения к EMR



lectures-hse-spark / aws / chromeproxy.sh

a4tunado chromeproxy.sh

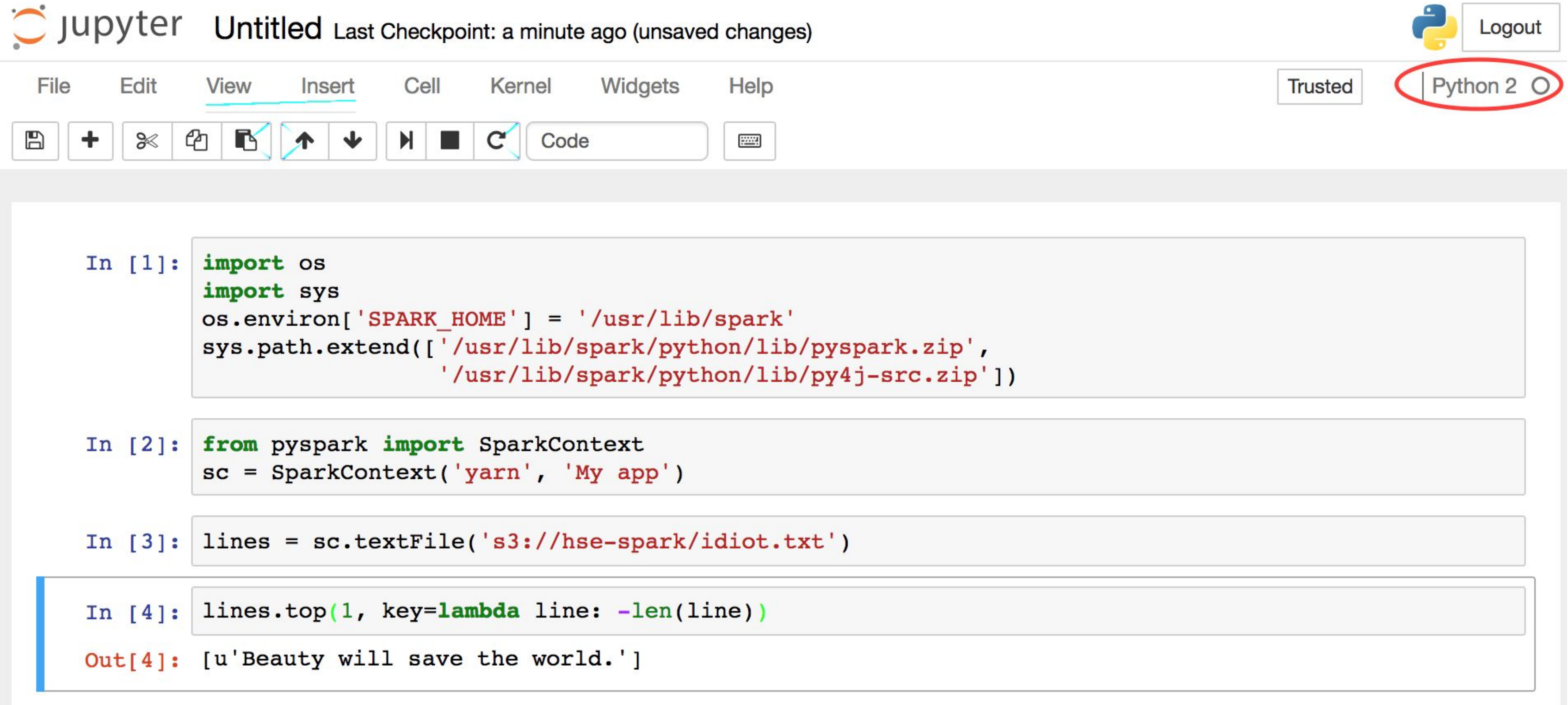
1 contributor

5 lines (3 sloc) | 163 Bytes

```
#!/bin/bash

"/Applications/Google Chrome.app/Contents/MacOS/Google Chrome" \
  --user-data-dir="$HOME/chrome-with-proxy" --proxy-server="socks5://localhost:8157"
```

5. Jupyter notebook



```
In [1]: import os
        import sys
        os.environ['SPARK_HOME'] = '/usr/lib/spark'
        sys.path.extend(['/usr/lib/spark/python/lib/pyspark.zip',
                         '/usr/lib/spark/python/lib/py4j-src.zip'])

In [2]: from pyspark import SparkContext
        sc = SparkContext('yarn', 'My app')

In [3]: lines = sc.textFile('s3://hse-spark/idiot.txt')

In [4]: lines.top(1, key=lambda line: -len(line))

Out[4]: [u'Beauty will save the world.']
```