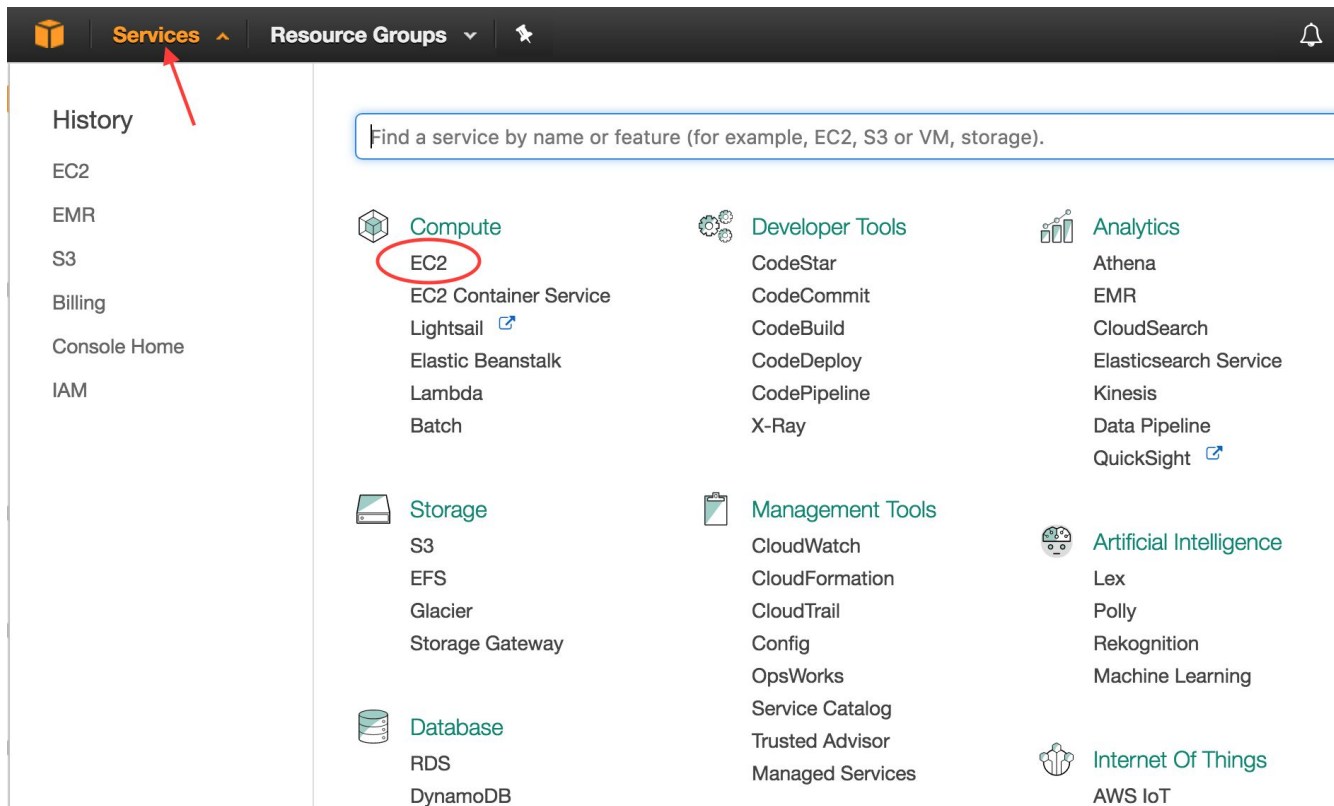


Запуск приложения Apache Spark на облачном сервисе Amazon EC2

Мурашкин Вячеслав
2017

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

1. Создание ключа для авторизации



1. Создание ключа для авторизации

The screenshot shows the Amazon Management Console interface. At the top, there's a navigation bar with 'Services' and 'Resource Groups' dropdowns, and a notification bell. On the left, the 'EC2 Dashboard' sidebar is visible, listing various EC2-related services. The main content area is titled 'Resources' and displays a list of EC2 resources in the US East (N. Virginia) region. The 'Key Pairs' resource is highlighted with a red circle. Below the resources list, there's a promotional banner for Amazon Lightsail. At the bottom, there's a 'Create Instance' section with a 'Launch Instance' button and a note about the default region.

Services ▾ **Resource Groups** ▾ 🔔

EC2 Dashboard

- Events
- Tags
- Reports
- Limits
- INSTANCES
 - Instances
 - Spot Requests
 - Reserved Instances
 - Scheduled Instances
 - Dedicated Hosts
- IMAGES
 - AMIs
 - Bundle Tasks
- ELASTIC BLOCK STORE
 - Volumes
 - Snapshots

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](#). ✕

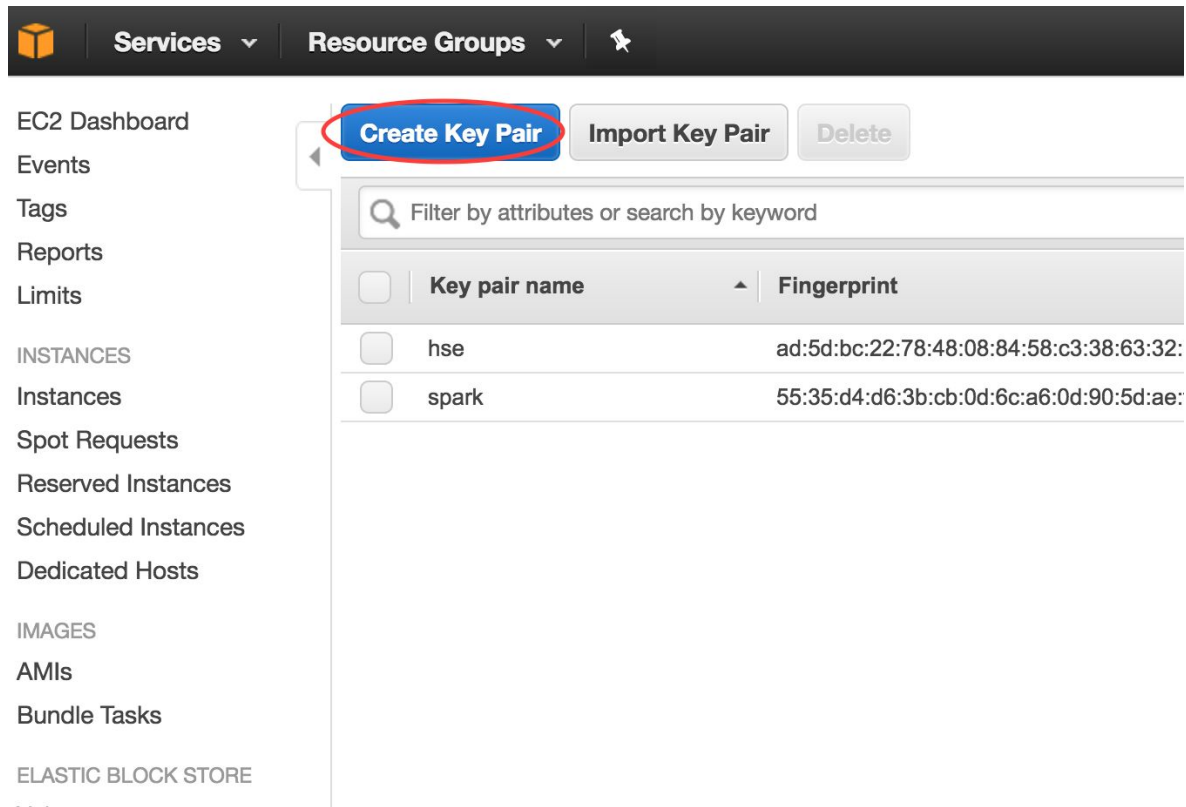
Create Instance

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

Launch Instance

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

1. Создание ключа для авторизации



The screenshot displays the AWS Management Console interface. At the top, there is a navigation bar with the AWS logo, 'Services' dropdown, 'Resource Groups' dropdown, and a pin icon. On the left side, a sidebar lists various services: EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, Instances, Spot Requests, Reserved Instances, Scheduled Instances, Dedicated Hosts, IMAGES, AMIs, Bundle Tasks, and ELASTIC BLOCK STORE. The main content area shows the 'Key Pairs' page. At the top of this page are three buttons: 'Create Key Pair' (highlighted with a red circle), 'Import Key Pair', and 'Delete'. Below these buttons is a search bar with the placeholder text 'Filter by attributes or search by keyword'. Underneath the search bar is a table with two columns: 'Key pair name' and 'Fingerprint'. The table contains two entries: 'hse' with fingerprint 'ad:5d:bc:22:78:48:08:84:58:c3:38:63:32:' and 'spark' with fingerprint '55:35:d4:d6:3b:cb:0d:6c:a6:0d:90:5d:ae:'.

<input type="checkbox"/>	Key pair name	Fingerprint
<input type="checkbox"/>	hse	ad:5d:bc:22:78:48:08:84:58:c3:38:63:32:
<input type="checkbox"/>	spark	55:35:d4:d6:3b:cb:0d:6c:a6:0d:90:5d:ae:

2. Создание хранилища для загрузки данных

The screenshot shows the AWS Management Console interface. At the top, there is a navigation bar with the AWS logo, the word "Services" with an upward arrow, "Resource Groups" with a downward arrow, and a notification bell icon. On the left side, there is a "History" sidebar with links to EMR, S3, EC2, Billing, Console Home, and IAM. The main content area features a search bar with the placeholder text "Find a service by name or feature (for example, EC2, S3 or VM, storage)." Below the search bar, services are organized into categories: Compute, Developer Tools, Analytics, Storage, Management Tools, Database, Artificial Intelligence, and Internet Of Things. The "Storage" category is highlighted, and within it, the "S3" service is circled in red. Other services listed include EC2, EC2 Container Service, Lightsail, Elastic Beanstalk, Lambda, Batch, CodeStar, CodeCommit, CodeBuild, CodeDeploy, CodePipeline, X-Ray, Athena, EMR, CloudSearch, Elasticsearch Service, Kinesis, Data Pipeline, QuickSight, CloudWatch, CloudFormation, CloudTrail, Config, OpsWorks, Service Catalog, Trusted Advisor, Managed Services, Lex, Polly, Rekognition, Machine Learning, and Internet Of Things.

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

Developer Tools

- CodeStar
- CodeCommit
- CodeBuild
- CodeDeploy
- CodePipeline
- X-Ray

Analytics

- Athena
- EMR
- CloudSearch
- Elasticsearch Service
- Kinesis
- Data Pipeline
- QuickSight

Storage

- S3**
- EFS
- Glacier
- Storage Gateway

Management Tools

- CloudWatch
- CloudFormation
- CloudTrail
- Config
- OpsWorks
- Service Catalog
- Trusted Advisor
- Managed Services

Database

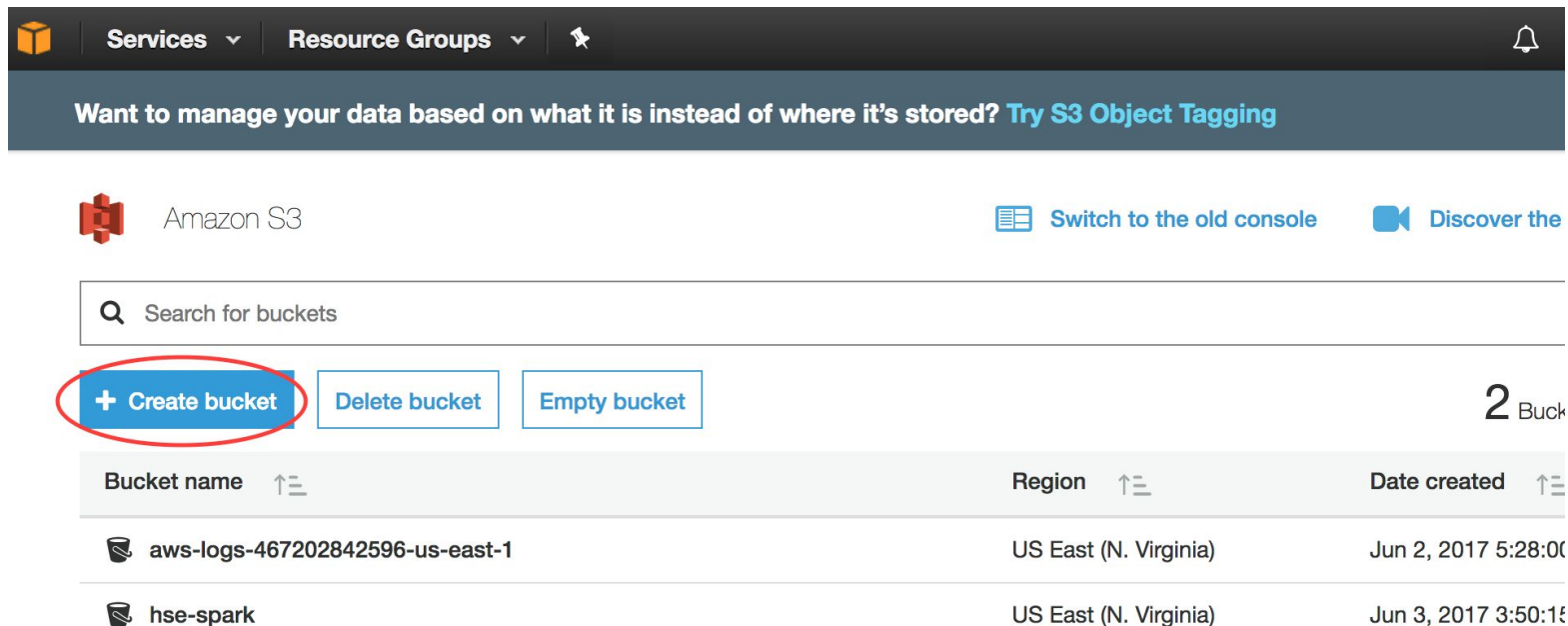
- RDS

Artificial Intelligence

- Lex
- Polly
- Rekognition
- Machine Learning

Internet Of Things

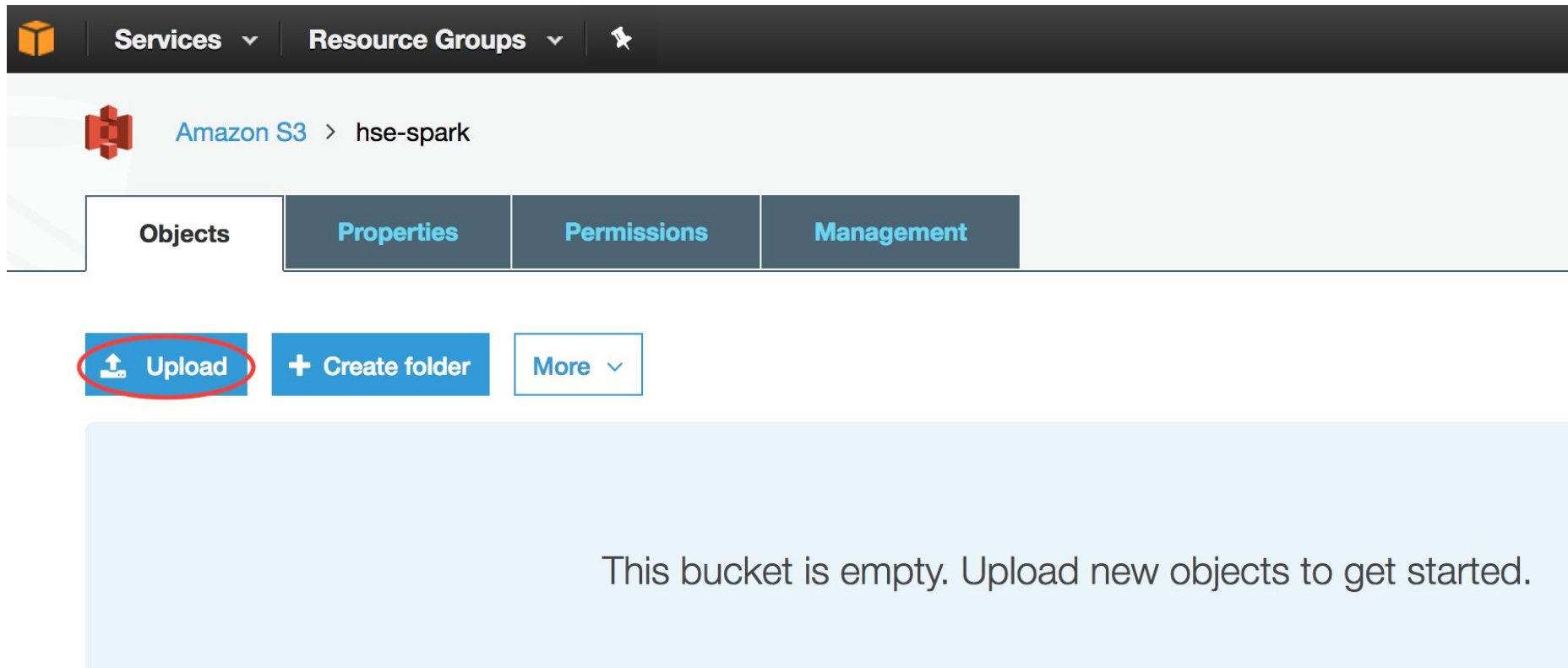
2. Создание хранилища для загрузки данных



The screenshot shows the Amazon S3 console interface. At the top, there's a navigation bar with 'Services' and 'Resource Groups' dropdowns, and a notification bell. Below this is a banner for 'S3 Object Tagging'. The main header area includes the 'Amazon S3' logo, a 'Switch to the old console' link, and a 'Discover the' link. A search bar labeled 'Search for buckets' is present. Below the search bar, three buttons are visible: '+ Create bucket' (highlighted with a red circle), 'Delete bucket', and 'Empty bucket'. On the right side, it says '2 Buck'. Below these buttons is a table listing existing buckets.

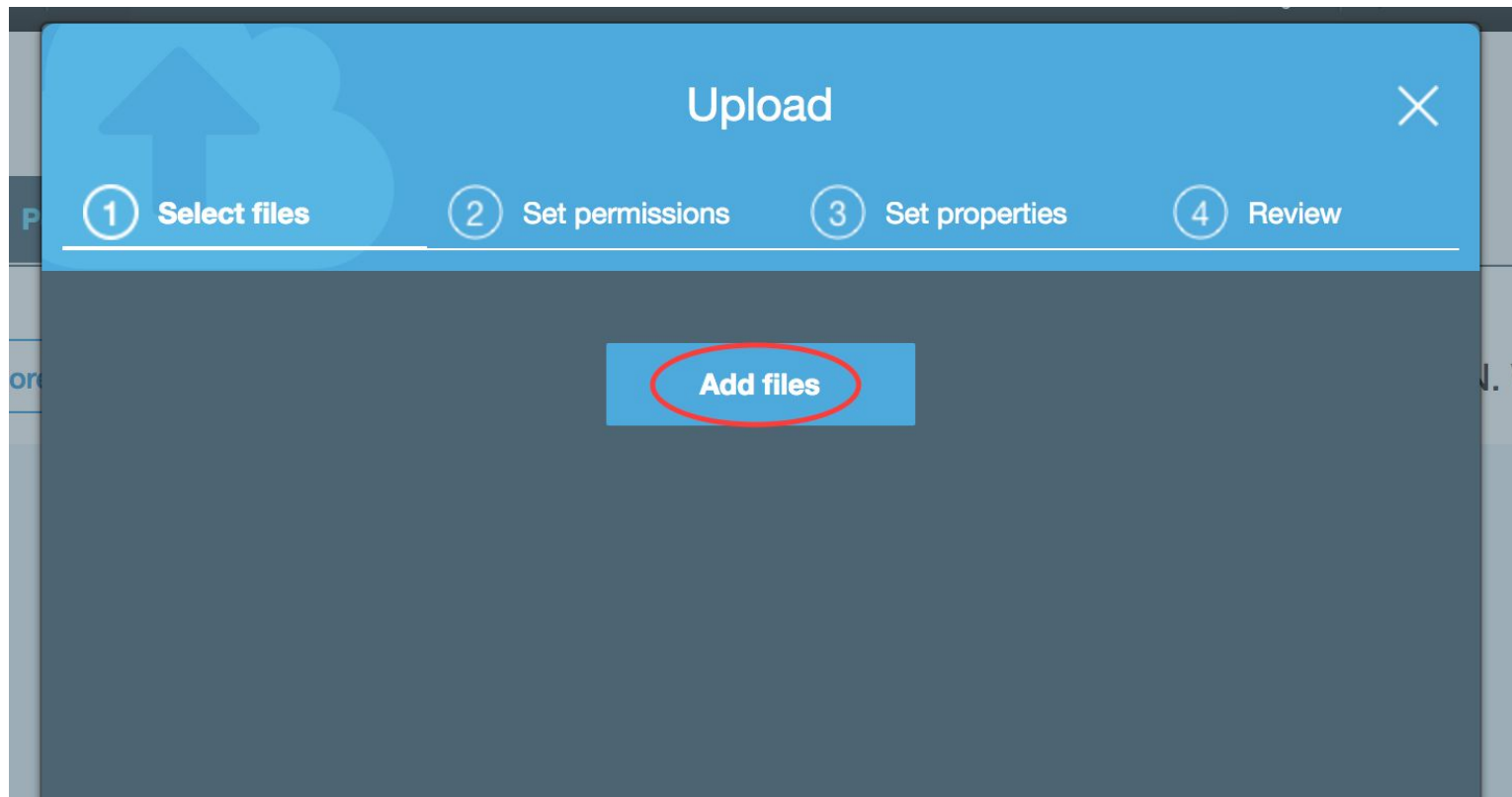
Bucket name	Region	Date created
aws-logs-467202842596-us-east-1	US East (N. Virginia)	Jun 2, 2017 5:28:00
hse-spark	US East (N. Virginia)	Jun 3, 2017 3:50:15

2. Создание хранилища для загрузки данных

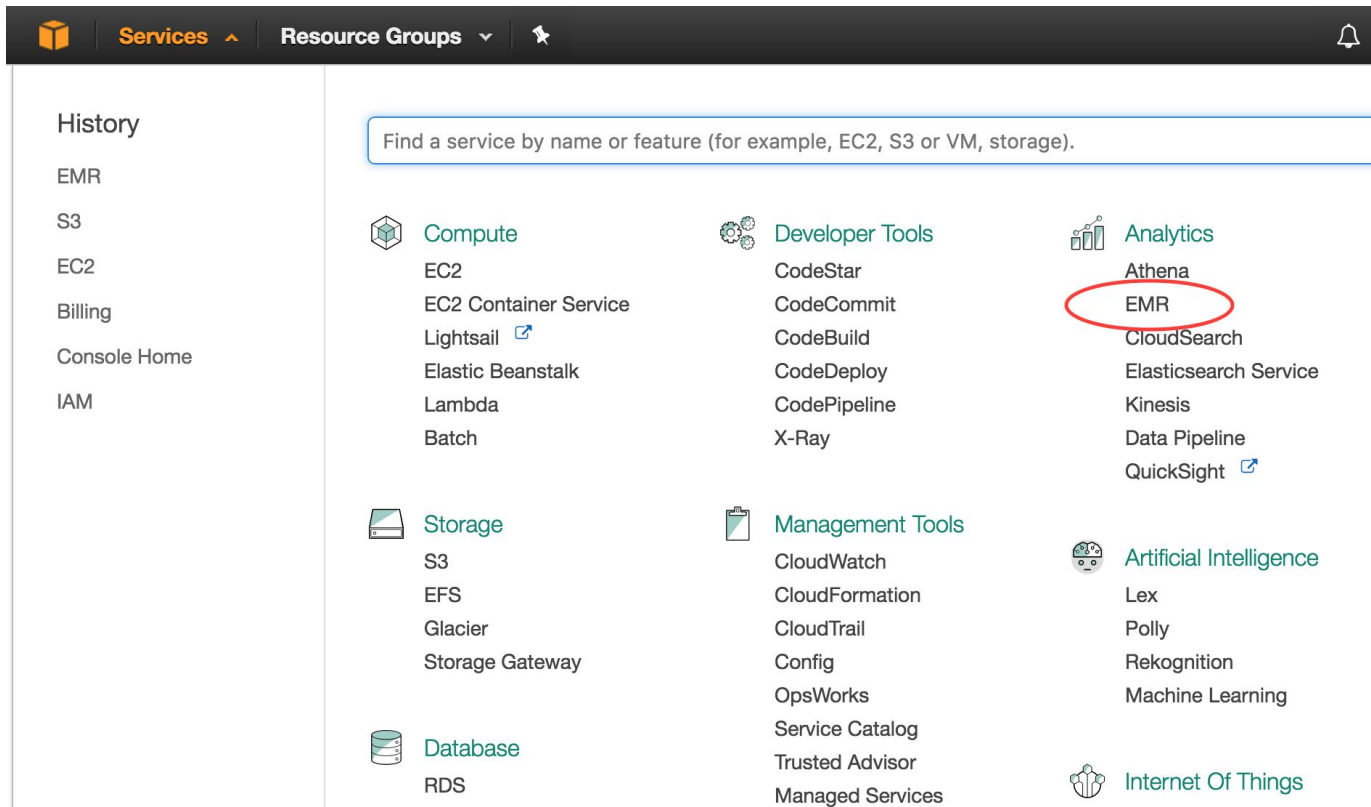


The screenshot displays the Amazon S3 console interface. At the top, there is a navigation bar with the AWS logo, 'Services' dropdown, 'Resource Groups' dropdown, and a search icon. Below this, the breadcrumb path 'Amazon S3 > hse-spark' is shown. A tabbed interface contains four tabs: 'Objects', 'Properties', 'Permissions', and 'Management'. The 'Objects' tab is active. Below the tabs, there are three buttons: 'Upload' (with an upload icon and circled in red), '+ Create folder', and 'More' with a dropdown arrow. The main content area is a light blue box with the text: 'This bucket is empty. Upload new objects to get started.'

2. Создание хранилища для загрузки данных



3. Создание EMR кластера



The screenshot shows the AWS Management Console interface. At the top, there is a navigation bar with the AWS logo, 'Services' (with an upward arrow), 'Resource Groups' (with a downward arrow), and a notification bell. On the left side, there is a 'History' sidebar with links to EMR, S3, EC2, Billing, Console Home, and IAM. The main area displays a search bar with the text 'Find a service by name or feature (for example, EC2, S3 or VM, storage)'. Below the search bar, services are organized into categories: Compute, Developer Tools, Analytics, Storage, Management Tools, and Database. The 'Analytics' category is highlighted, and 'EMR' is circled in red. Other services listed include Athena, CloudSearch, Elasticsearch Service, Kinesis, Data Pipeline, and QuickSight. The 'Storage' category includes S3, EFS, Glacier, and Storage Gateway. The 'Database' category includes RDS. The 'Developer Tools' category includes CodeStar, CodeCommit, CodeBuild, CodeDeploy, CodePipeline, and X-Ray. The 'Management Tools' category includes CloudWatch, CloudFormation, CloudTrail, Config, OpsWorks, Service Catalog, Trusted Advisor, and Managed Services. The 'Artificial Intelligence' category includes Lex, Polly, Rekognition, and Machine Learning. The 'Internet Of Things' category is also visible at the bottom.

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

Developer Tools

- CodeStar
- CodeCommit
- CodeBuild
- CodeDeploy
- CodePipeline
- X-Ray

Analytics

- Athena
- EMR**
- CloudSearch
- Elasticsearch Service
- Kinesis
- Data Pipeline
- QuickSight

Storage

- S3
- EFS
- Glacier
- Storage Gateway

Management Tools

- CloudWatch
- CloudFormation
- CloudTrail
- Config
- OpsWorks
- Service Catalog
- Trusted Advisor
- Managed Services

Database



- RDS

Artificial Intelligence

- Lex
- Polly
- Rekognition
- Machine Learning

Internet Of Things

3. Создание EMR кластера

 Services ▾ Resource Groups ▾ 

Amazon EMR

Cluster list

Security configurations

VPC subnets

Events

Help

Create cluster


View details

Clone

Terminate

Filter: All clusters ▾ Filter clusters ...

22 clusters (all loaded)

		Name	ID	Status	Creation time
<input type="checkbox"/>	▶	My cluster	j-2APS5TCRQHCHN	Terminated User request	2017-06-03
<input type="checkbox"/>	▶	My cluster	j-1OG77YUKLQK0F	Terminated User request	2017-06-03
<input type="checkbox"/>	▶	My cluster	j-2JK6Q85530PCC	Terminated User request	2017-06-03
<input type="checkbox"/>	▶	 My cluster	j-15WEX6KHWMOEZ	Terminated with errors Bootstrap failure	2017-06-03
<input type="checkbox"/>	▶	My cluster	j-KJP1AOBSNKXC	Terminated User request	2017-06-03
<input type="checkbox"/>	▶	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

☒ **Logging** ⓘ

S3 folder

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ


Software configuration

Release ⓘ

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

 Services ▾ Resource Groups ▾ ⌵

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Software Configuration

Release ⓘ

<input checked="" type="checkbox"/> Hadoop 2.7.3	<input type="checkbox"/> Zeppelin 0.7.1
<input type="checkbox"/> Flink 1.2.0	<input type="checkbox"/> Ganglia 3.7.2
<input type="checkbox"/> Pig 0.16.0	<input type="checkbox"/> Hive 2.1.1
<input type="checkbox"/> ZooKeeper 3.4.10	<input type="checkbox"/> Sqoop 1.4.6
<input type="checkbox"/> Hue 3.12.0	<input type="checkbox"/> Phoenix 4.9.0
<input checked="" type="checkbox"/> Spark 2.1.0	<input type="checkbox"/> 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 ⓘ

☐ Auto-terminate cluster after the last step is completed

3. Создание EMR кластера

Services

Resource Groups

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

General Options

Cluster name

☒ Logging

S3 folder

☒ Debugging

☒ Termination protection

Scale down behavior

Tags

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	

Additional Options

☐ EMRFS consistent view

Bootstrap Actions

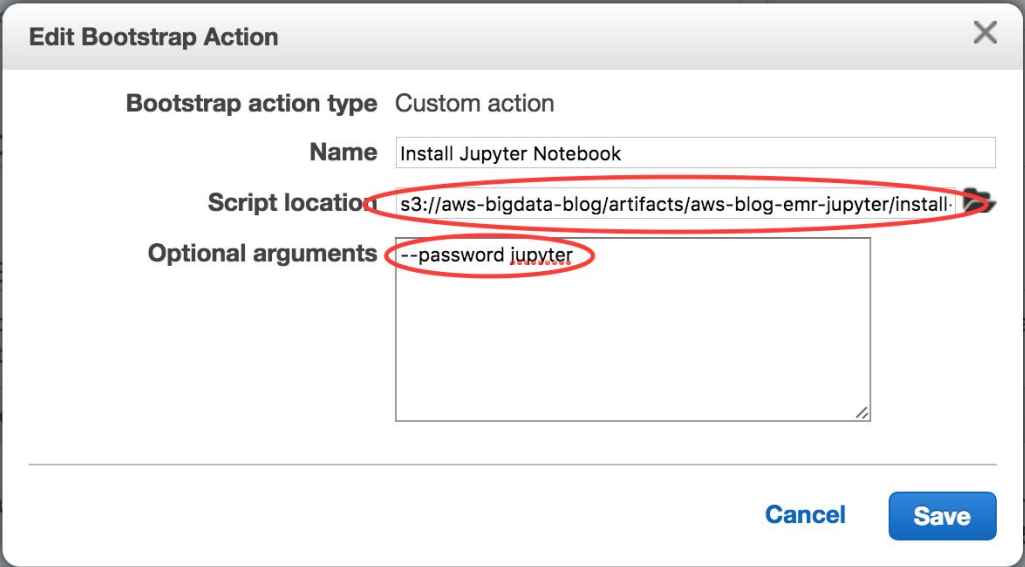
Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node software and customize your applications. [Learn more](#)

Add bootstrap action

Configure and add

3. Создание EMR кластера

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



Edit Bootstrap Action

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`

Buttons: Cancel, Save

Footer: Add bootstrap action Custom action Configure and add

3. Создание EMR кластера

- Укажите имя ключа, созданного на первом шаге

 Services ▾ Resource Groups ▾ 

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Security Options

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](#)

► Encryption Options


► EC2 Security Groups


4. Настройка ssh прокси для подключения к EMR



Services ▾

Resource Groups ▾





Amazon EMR

Cluster list

Security configurations

VPC subnets

Events

Help

Add step

Resize

Clone

Terminate

AWS CLI export

Cluster: My cluster

Waiting

Cluster ready after last step completed.

Connections:

Master public DNS:

Tags:

[Enable Web Connection](#)

Spark History Server, Resource Manager ... (View All)

ec2-54-234-247-21.compute-1.amazonaws.com [SSH](#)

-- [View All](#) / [Edit](#)

Summary

Configuration Details

ID: j-350M8ULZ3DB9E

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

Elapsed time: 30 minutes

Auto-terminate: No

Termination protection: On [Change](#)

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 master node 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](#)

Windows

Mac / Linux

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other operating systems, choose Applications > Accessories > Terminal.
2. To establish an SSH tunnel with the master node using dynamic port forwarding, type 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

Branch: master ▾

[lectures-hse-spark](#) / [aws](#) / [chromeproxy.sh](#)





a4tunado chromeproxy.sh

1 contributor











5 lines (3 sloc) | 163 Bytes

```
1  #!/bin/bash
2
3  "/Applications/Google Chrome.app/Contents/MacOS/Google Chrome" \
4  --user-data-dir="$HOME/chrome-with-proxy" --proxy-server="socks5://localhost:8157"
```

5. Jupyter notebook

 jupyter **Untitled** Last Checkpoint: a minute ago (unsaved changes)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted **Python 2**

         Code 

```
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.']
```