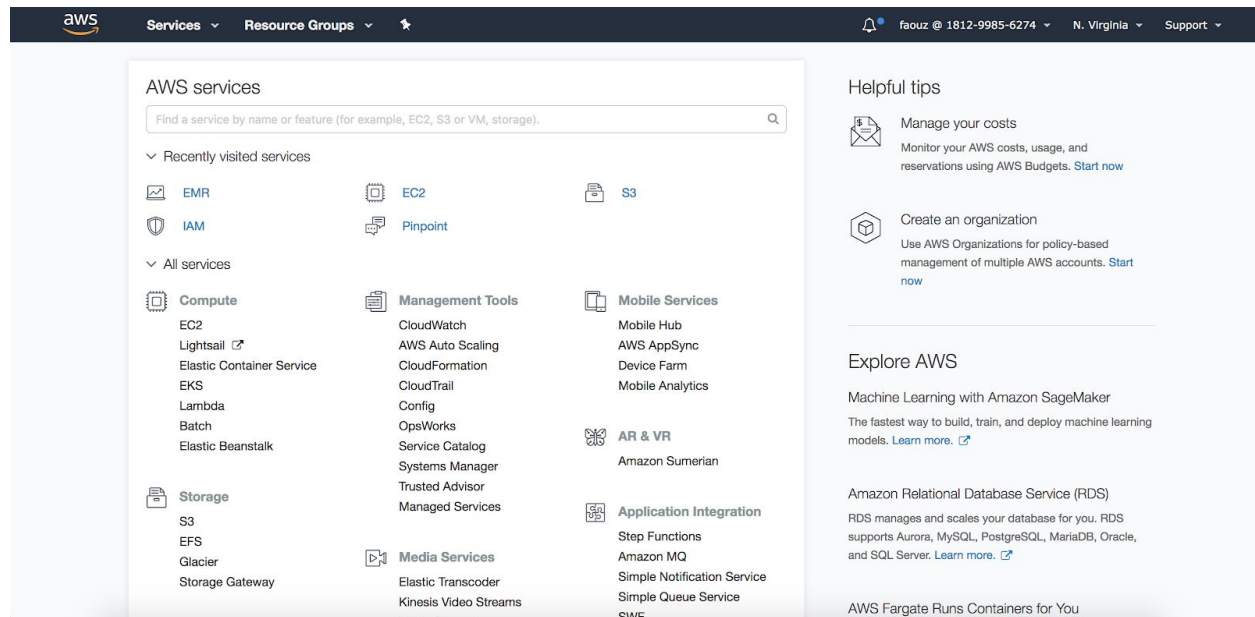
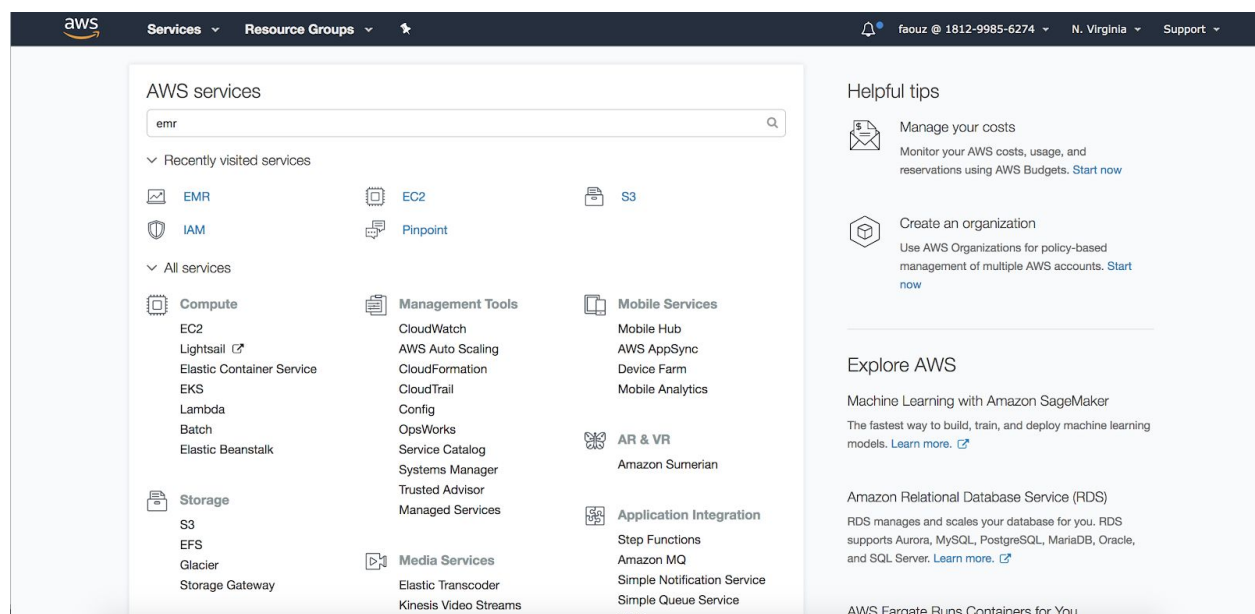


Tutoriel : Créer une cluster Spark 2.3 en 5 minutes sur Amazon Web Services

- Connectez-vous sur la console AWS via votre compte Rosetta Hub.



- Recherchez EMR (Elastic MapReduce)



- Créez un cluster

aws Services Resource Groups

faouz @ 1812-9985-6274 N. Virginia Support

Amazon EMR

- Clusters
- Security configurations
- VPC subnets
- Events
- Help
- What's new

Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

[Create cluster](#)

How Elastic MapReduce Works

Upload

Upload your data and processing application to S3.

[Learn more](#)

Create

Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.

[Learn more](#)

Monitor

Monitor the health and progress of your cluster. Retrieve the output in S3.

[Learn more](#)

Additional Information

More about Elastic MapReduce

- [EMR overview](#)
- [FAQs](#)
- [Pricing](#)

More Help Using Elastic MapReduce

- [Forum](#)
- [Documentation](#)
- [Developer Guide](#)
- [API Reference](#)
- [EMR on GitHub](#)
- [Help portal](#)

- Sélectionnez Spark dans la partie “Applications”

aws Services Resource Groups

faouz @ 1812-9985-6274 N. Virginia Support

Create Cluster - Quick Options [Go to advanced options](#)

General Configuration

Cluster name

☒ **Logging** ⓘ

S3 folder

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

Software configuration

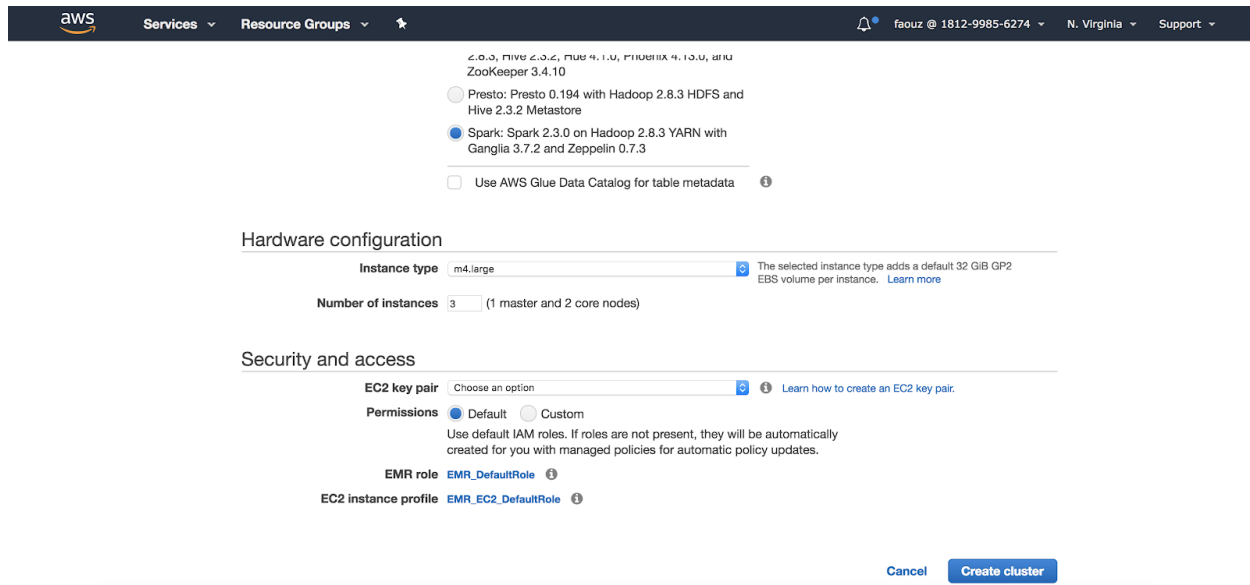
Release ⓘ

Applications

- ☐ Core Hadoop: Hadoop 2.8.3 with Ganglia 3.7.2, Hive 2.3.2, Hue 4.1.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4
- ☐ HBase: HBase 1.4.2 with Ganglia 3.7.2, Hadoop 2.8.3, Hive 2.3.2, Hue 4.1.0, Phoenix 4.13.0, and ZooKeeper 3.4.10
- ☐ Presto: Presto 0.194 with Hadoop 2.8.3 HDFS and Hive 2.3.2 Metastore
- ☒ Spark: Spark 2.3.0 on Hadoop 2.8.3 YARN with Ganglia 3.7.2 and Zeppelin 0.7.3

☐ Use AWS Glue Data Catalog for table metadata ⓘ

- Puis, dans la partie type d'instance sélectionnez **m4.large**, au nombre de 3.



Hardware configuration

Instance type: **m4.large** The selected instance type adds a default 32 GiB GP2 EBS volume per instance. [Learn more](#)

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

Security and access

EC2 key pair: **Choose an option** [Learn how to create an EC2 key pair.](#)

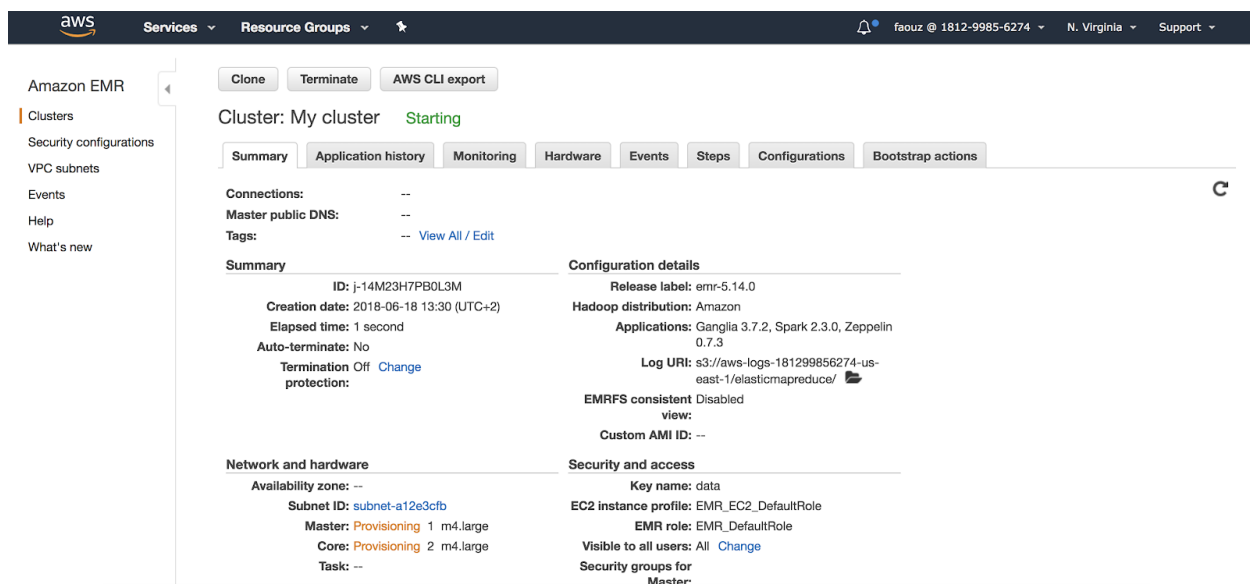
Permissions: **Default** Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: **EMR_DefaultRole**

EC2 instance profile: **EMR_EC2_DefaultRole**

[Cancel](#) [Create cluster](#)

- Attendez quelques minutes pour que votre cluster démarre.



Amazon EMR

Clusters

Security configurations

VPC subnets

Events

Help

What's new

Clone Terminate AWS CLI export

Cluster: My cluster **Starting**

Summary Application history Monitoring Hardware Events Steps Configurations Bootstrap actions

Connections: --

Master public DNS: --

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

Summary

ID: j-14M23H7PB0L3M

Creation date: 2018-06-18 13:30 (UTC+2)

Elapsed time: 1 second

Auto-terminate: No

Termination protection: [Change](#)

Network and hardware

Availability zone: --

Subnet ID: **subnet-a12e3cfb**

Master: **Provisioning** 1 m4.large

Core: **Provisioning** 2 m4.large


Task: --

Configuration details

Release label: emr-5.14.0

Hadoop distribution: Amazon

Applications: Ganglia 3.7.2, Spark 2.3.0, Zeppelin 0.7.3

Log URI: s3://aws-logs-181299856274-us-east-1/elasticmapreduce/ 

EMRFS consistent view: Disabled

Custom AMI ID: --

Security and access

Key name: data

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master:

- Une fois démarrer, le cluster se met en attente

The screenshot shows the AWS Management Console interface for an Amazon EMR cluster. The cluster is named 'My cluster' and is in a 'Waiting' state. The console displays various tabs for cluster management, including Summary, Application history, Monitoring, Hardware, Events, Steps, Configurations, and Bootstrap actions. The Summary tab is selected, showing details such as the cluster ID (j-14M23H7PB0L3M), creation date (2018-06-18 13:30 UTC+2), and the fact that it is not set to auto-terminate. It also lists the master public DNS, tags, and configuration details like the release label (emr-5.14.0) and Hadoop distribution (Amazon). Network and hardware details show the cluster is in the us-east-1b availability zone with a subnet ID of subnet-a12e3cfb. Security and access details include the key name 'data', EC2 instance profile 'EMR_EC2_DefaultRole', and EMR role 'EMR_DefaultRole'.

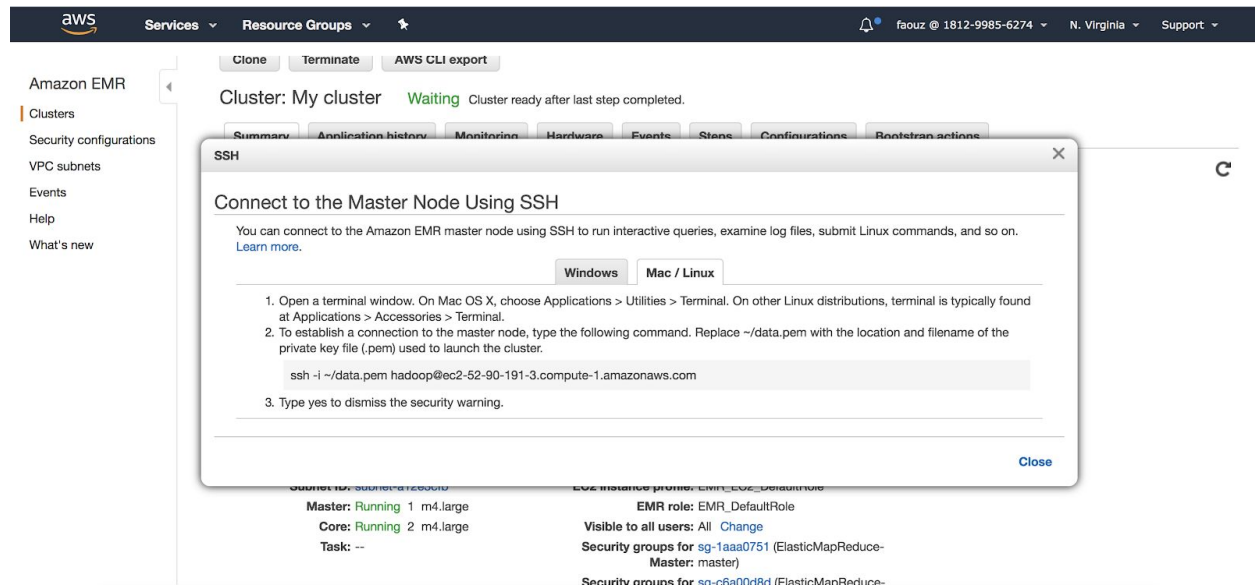
- Cliquez sur **Security groups for Master**, puis dans l'onglet Inbound

The screenshot shows the AWS Management Console interface for the 'Security Groups' section. The 'Create Security Group' button is visible at the top. A search filter 'sg-1aaa0751' is applied. The table below lists the security groups: 'sg-1aaa0751' (ElasticMapReduce-master) and 'sg-c6a00d8d' (ElasticMapReduce-slave). The 'sg-1aaa0751' group is selected, and the 'Inbound' tab is active. The 'Edit' button is visible, and the table below shows the inbound rules for this security group.

Type	Protocol	Port Range	Source	Description
All TCP	TCP	0 - 65535	62.23.191.42/32	
All TCP	TCP	0 - 65535	sg-1aaa0751 (ElasticMapReduce-m	
All TCP	TCP	0 - 65535	sg-c6a00d8d (ElasticMapReduce-sl	
Custom TCP Rule	TCP	8443	207.171.167.25/32	
Custom TCP Rule	TCP	8443	54.240.217.8/29	
Custom TCP Rule	TCP	8443	72.21.196.64/29	

Cliquez sur edit, puis ajouter une règle pour autoriser tous les ports (0 - 65535) à votre adresse IP.

- Enfin, cliquez sur le lien SSH pour avoir les instructions pour pouvoir s'y connecter en SSH.



En copiant cette commande dans votre terminal vous pourrez vous-y connecter pour utiliser le client **pyspark** ou **spark-shell**.