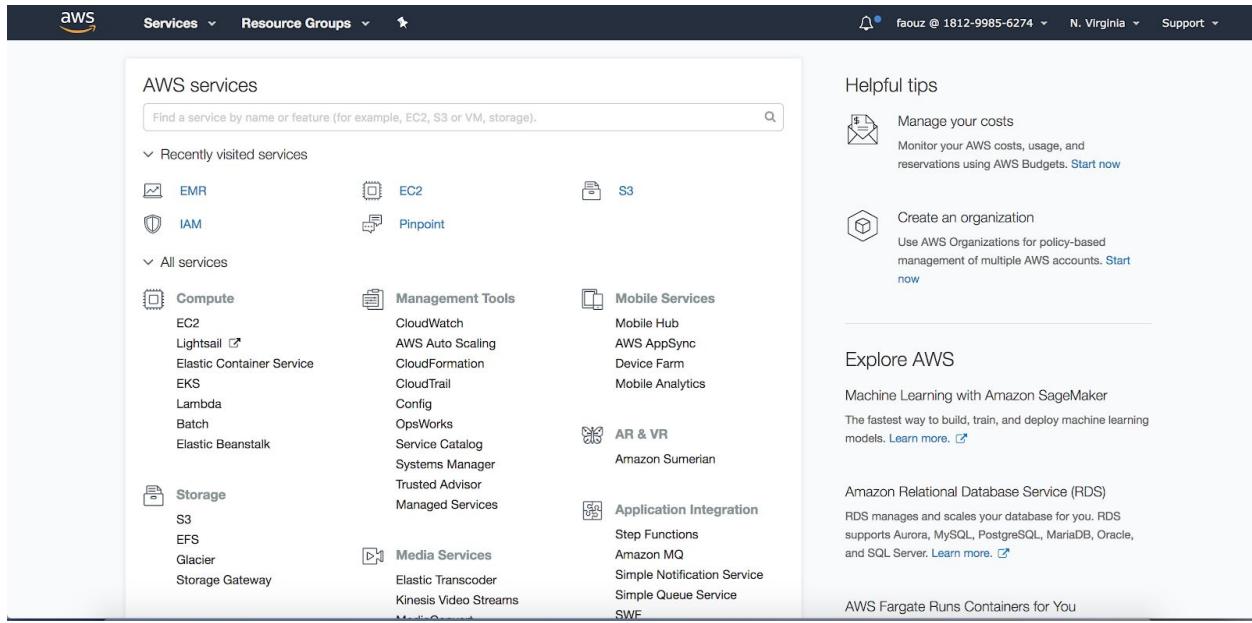


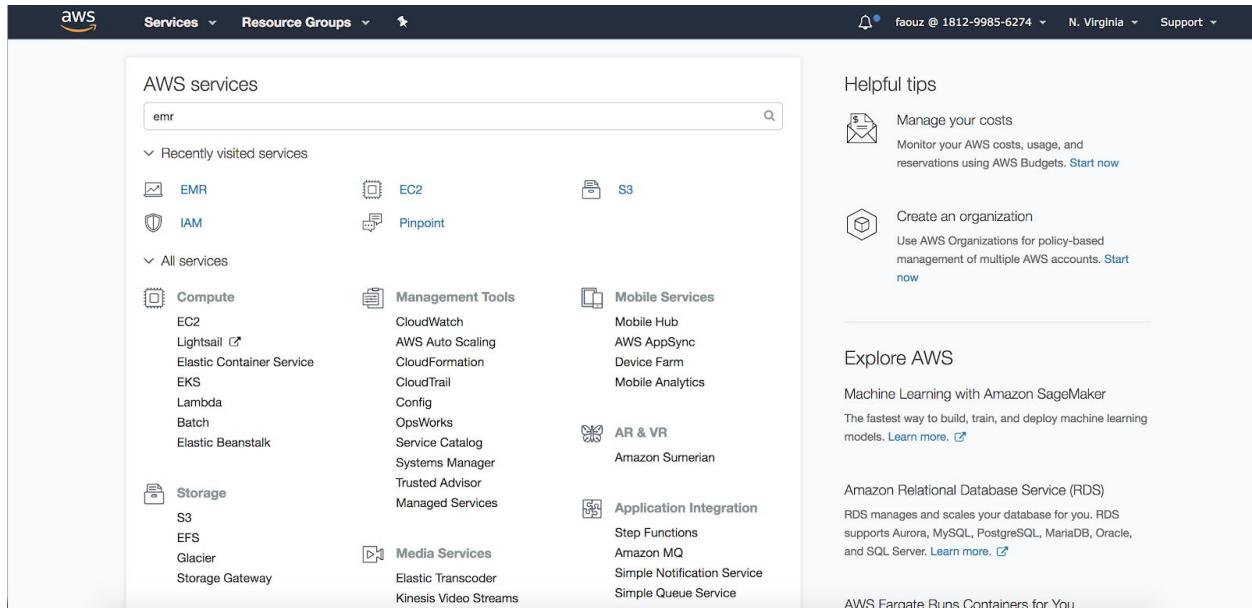
## Tutoriel : Créez une cluster Spark 2.3 en 5 minutes sur Amazon Web Services

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



The screenshot shows the AWS Management Console homepage. The top navigation bar includes the AWS logo, a "Services" dropdown, a "Resource Groups" dropdown, and user information (faouz @ 1812-9985-6274, N. Virginia, Support). The main content area is titled "AWS services" and features a search bar. Below the search bar are sections for "Recently visited services" (EMR, IAM, EC2, Pinpoint, S3) and "All services" categorized into Compute, Storage, Management Tools, Mobile Services, AR & VR, Application Integration, Media Services, and others like CloudWatch, AWS Auto Scaling, and Step Functions. To the right, there's a "Helpful tips" sidebar with links to "Manage your costs" and "Create an organization", and an "Explore AWS" sidebar with links to "Machine Learning with Amazon SageMaker", "Amazon Relational Database Service (RDS)", and "AWS Fargate Runs Containers for You".

- Recherchez EMR (Elastic MapReduce)



This screenshot is identical to the one above, but with the search bar containing the text "emr". The results for the search are not yet visible, but the search term is clearly shown in the search bar.

- Créez un cluster

Screenshot of the AWS Amazon EMR service dashboard:

**Left sidebar:**

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

**Main content area:**

## 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

Three main steps illustrated with icons:

- Upload:** Cloud icon with an orange arrow pointing up.
- Create:** Cluster icon with a gear and an orange arrow pointing down.
- Monitor:** Monitor icon with a graph and an orange arrow pointing down.

Descriptions for each step:

- Upload: Upload your data and processing application to S3.
- Create: Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.
- Monitor: Monitor the health and progress of your cluster. Retrieve the output in S3.

[Learn more](#) (under each step)

**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"

Screenshot of the AWS Create Cluster - Quick Options page:

**General Configuration:**

- Cluster name: My cluster
- Logging
- S3 folder: s3://aws-logs-181299856274-us-east-1/elasticmapreduce/
- Launch mode: Cluster (selected)

**Software configuration:**

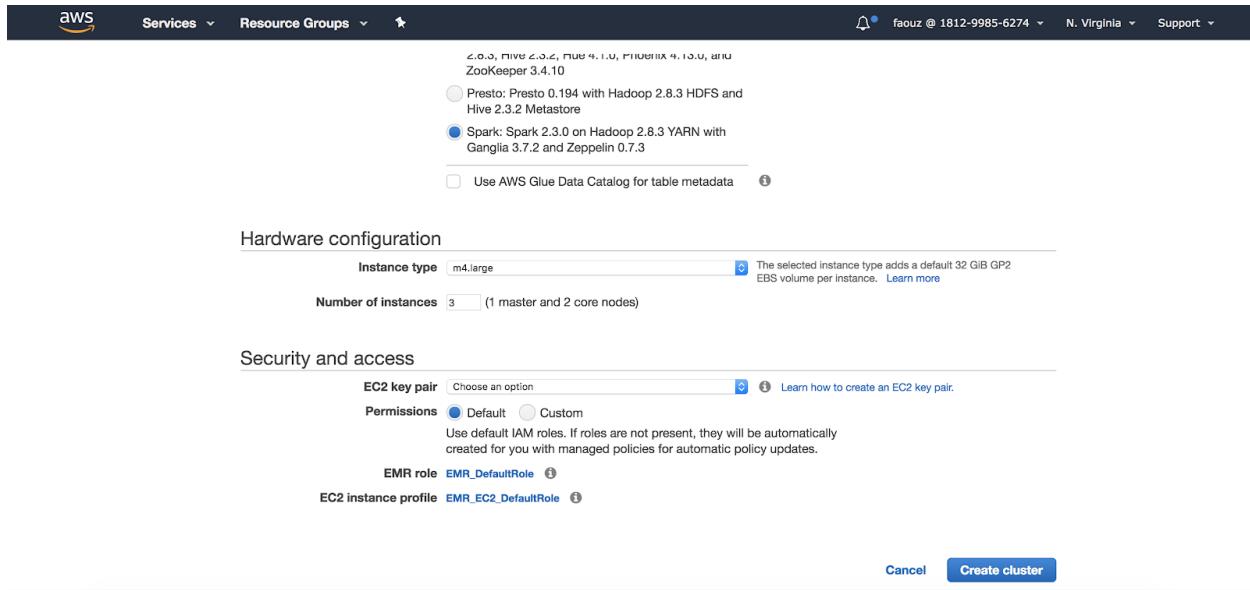
Release: emr-5.14.0

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

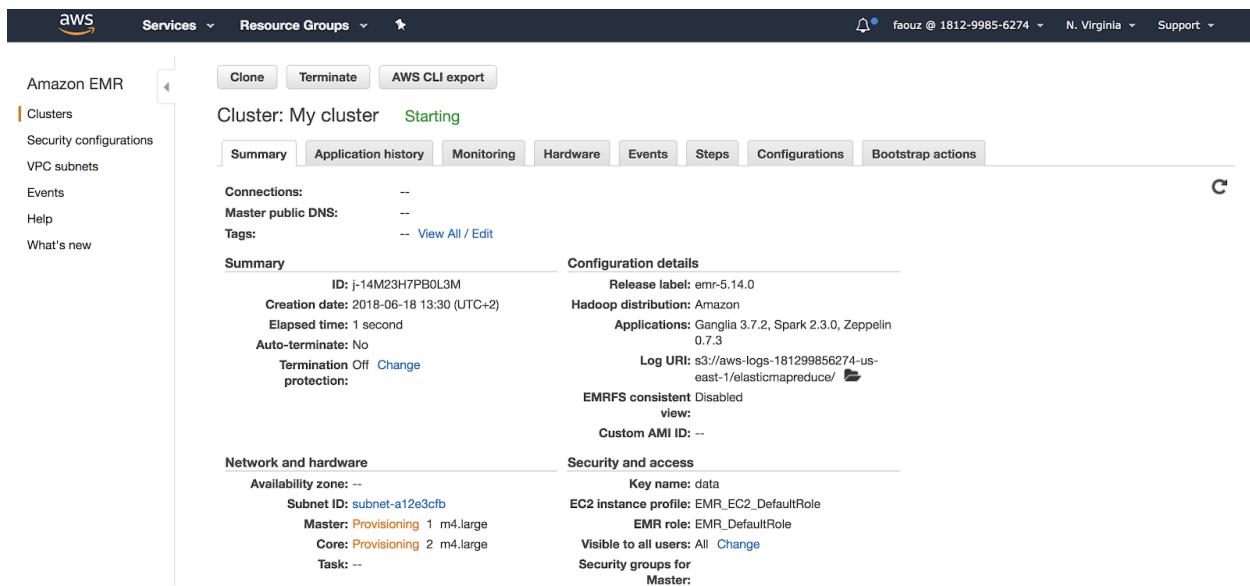
Use AWS Glue Data Catalog for table metadata

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



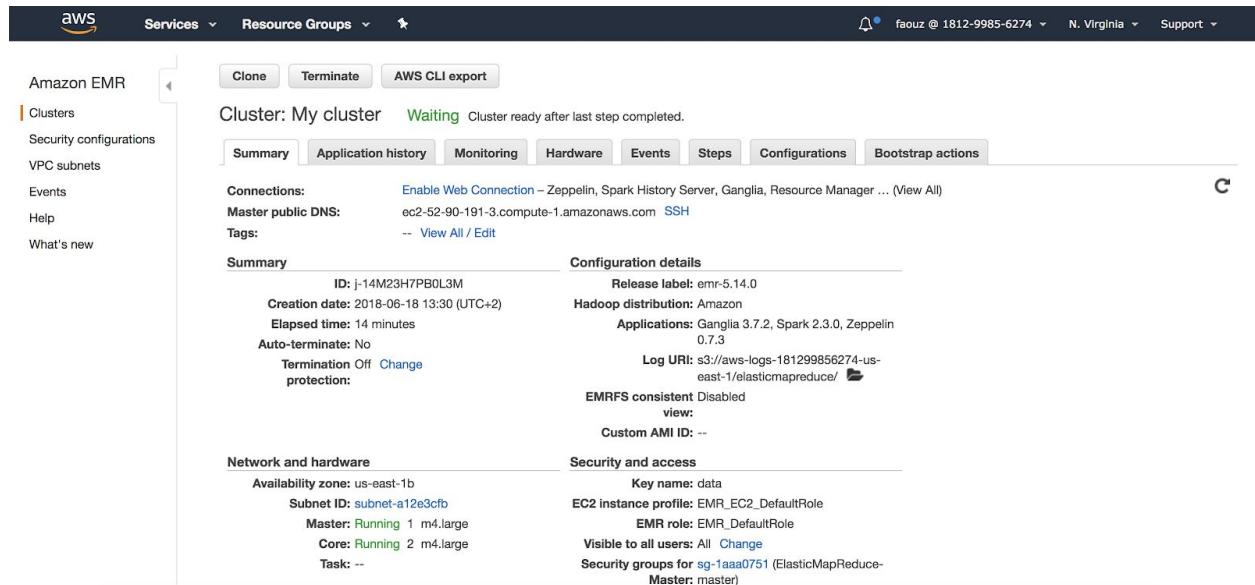
The screenshot shows the AWS Glue Create Cluster wizard. In the 'Hardware configuration' section, the instance type is set to 'm4.large' and the number of instances is set to '3'. In the 'Security and access' section, the EC2 key pair dropdown is empty, and the permissions are set to 'Default'. The EMR role is set to 'EMR\_DefaultRole'. At the bottom right, there are 'Cancel' and 'Create cluster' buttons.

- Attendez quelques minutes pour que votre cluster démarre.



The screenshot shows the AWS Glue Cluster details page for 'Cluster: My cluster'. The cluster status is 'Starting'. The 'Summary' tab shows the cluster ID (j-14M23H7PB0L3M), creation date (2018-06-18 13:30 (UTC+2)), and other basic information. The 'Configuration details' section includes the release label (mr-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/), and custom AMI ID. The 'Network and hardware' section shows the availability zone (--), subnet ID (subnet-a12e3cfb), and task configuration. The 'Security and access' section lists the EC2 instance profile (EMR\_EC2\_DefaultRole), EMR role (EMR\_DefaultRole), and security groups. The left sidebar shows the navigation menu for Amazon EMR, with 'Clusters' selected.

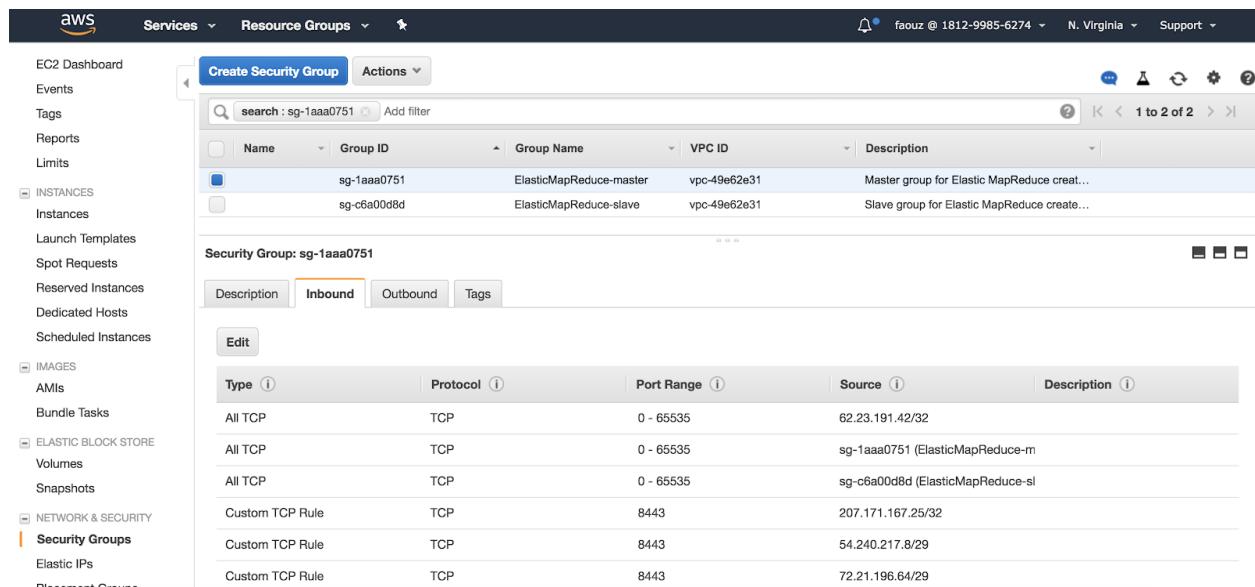
- Une fois démarrer, le cluster se met en attente



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

Summary	Configuration details
ID: j-14M23H7PB0L3M Creation date: 2018-06-18 13:30 (UTC+2) Elapsed time: 14 minutes Auto-terminate: No Termination Off Change protection:	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: --
<b>Network and hardware</b> Availability zone: us-east-1b Subnet ID: subnet-a12e3cfb Master: Running 1 m4.large Core: Running 2 m4.large Task: --	<b>Security and access</b> Key name: data EC2 instance profile: EMR_EC2_DefaultRole EMR role: EMR_DefaultRole Visible to all users: All Change Security groups for sg-1aaa0751 (ElasticMapReduce-Master: master)

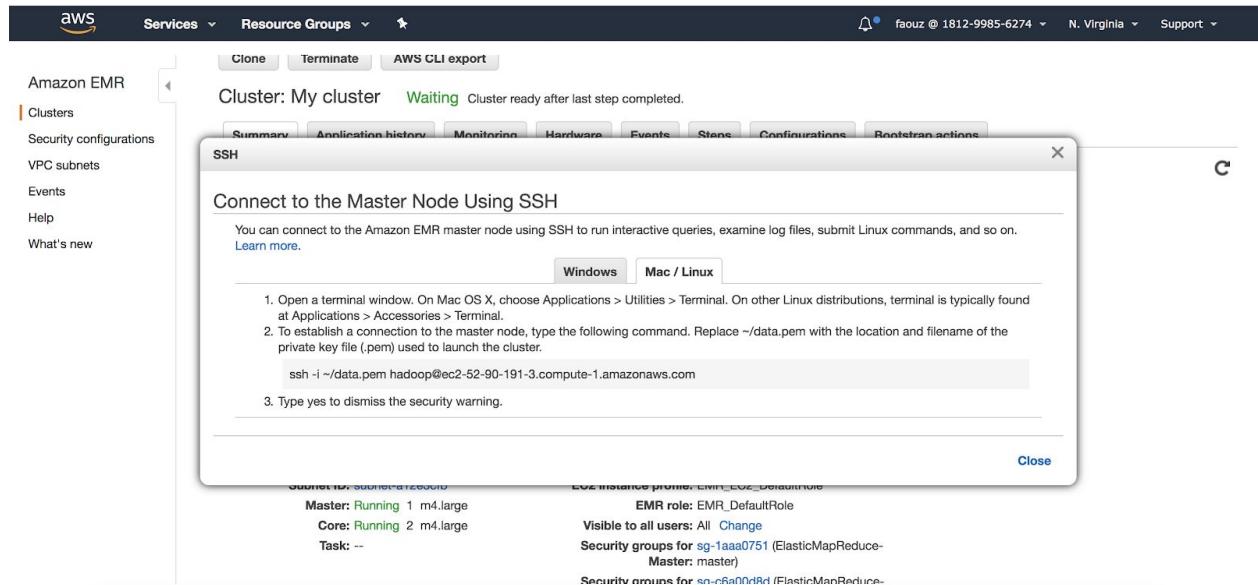
- Cliquez sur **Security groups for Master**, puis dans l'onglet Inbound



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.



The screenshot shows the AWS EMR service page with a cluster named "My cluster" in the "Waiting" state. A modal window titled "SSH" is open, prompting the user to "Connect to the Master Node Using SSH". It provides instructions for connecting via SSH from a Mac OS X terminal or a Linux terminal. Below the instructions, it shows the cluster configuration details:

Cluster ID: Subnet-01234567890123456		EC2 instance profile: EMR_EC2_DefaultRole	
<b>Master:</b> Running 1 m4.large		<b>EMR role:</b> EMR_DefaultRole	
<b>Core:</b> Running 2 m4.large		<b>Visible to all users:</b> All Change	
<b>Task:</b> --		<b>Security groups for sg-1aaa0751 (ElasticMapReduce-Master: master)</b>	
		<b>Security groups for sg-0fa00f1f (ElasticMapReduce-</b>	

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