

Spark on AWS EMR

v1.0

Table of Contents

Objectifs

- Creation d'un bucket AWS S3
 - Démarrage d'un cluster sur AWS EMR
 - Installation de AWS CLI
-

Objectifs

1. Dans ce TP nous allons commencer par creer un bucket AWS S3 dans lequel nous allons stoquer un sous-ensemble de donnees GDELT qu'on va explorer.
2. Dans une deuxieme etape nous allons lancer un cluster Spark sur AWS EMR.
3. Nous allons utiliser ce cluster pour stocker des donnees de GDELT dans notre bucket S3 via un ETL distribue lance depuis un notebook Spark (Zeppelin).
4. A la fin du TP, nous allons utiliser Spark pour explorer ce jeu de donnees qui sera utilise plus tard pour le projet.

Creation d'un bucket AWS S3

Nous allons sauvegarder une copie sur AWS du jeu de donnees utilisees. En utilisant un bucket AWS S3 dans ce but nous nous protegeons d'une eventuelle panne du site web de GDELT et nous allons avoir une source de donnees distribuee et repliquee avec une tres grande tolerance aux pannes.

1. Allez sur la console AWS S3 (<https://s3.console.aws.amazon.com/s3/home?region=us-east-1#>) (<https://s3.console.aws.amazon.com/s3/home?region=us-east-1#>)

The screenshot shows the Amazon S3 console interface. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', and user information. The left sidebar shows 'Amazon S3' and 'Buckets'. The main content area is titled 'S3 buckets' and includes a search bar, an 'All access types' dropdown, and buttons for '+ Create bucket', 'Edit public access settings', 'Empty', and 'Delete'. The '+ Create bucket' button is highlighted with a red box. Below these buttons, it shows '0 Regions' and '0 Buckets'. A large light blue box contains the text 'You do not have any buckets. Here is how to get started with Amazon S3.' and three steps: 'Create a new bucket', 'Upload your data', and 'Set up your permissions'. Each step has an icon, a brief description, and a 'Learn more' link. A 'Get started' button is at the bottom of the light blue box.

1. Créez un bucket ***nom-prenom-telecom-gdelt2018*** (! le nom du bucket doit être unique sur l'ensemble des utilisateurs S3). Assurez-vous que le bucket créé est dans la région ***US-East(N. Virginia)***. Gardez tous les autres paramètres de configuration du bucket à leur valeur par défaut.

← → ↺ https://s3.console.aws.amazon.com/s3/home?region=us-east-1# 🔍 ☆ 🔴 ⋮

aws

Services

Resource Groups

☆

🔔

Andrei Arion

Global

Support

1 Name and region

2 Configure options

3 Set permissions

4 Review

Name and region

Bucket name ⓘ

john-doe-telecom-gdelt2018

Region

US East (N. Virginia) ▾

Copy settings from an existing bucket

You have no buckets0 Buckets ▾

Create

Cancel

Next

← → ↻ <https://s3.console.aws.amazon.com/s3/home?region=us-east-1#> 🔍 ☆ ⚙️

aws Services Resource Groups Andrei Arion Global Support

Create bucket

✓ Name and region ✓ Configure options ✓ Set permissions 4 Review

Name and region [Edit](#)

Bucket name john-doe-telecom-gdelt2018 **Region** US East (N. Virginia)

Options [Edit](#)

Versioning	Disabled
Server access logging	Disabled
Tagging	0 Tags
Object-level logging	Disabled
Default encryption	None
CloudWatch request metrics	Disabled
Object lock	Disabled

Permissions [Edit](#)

Block new public ACLs and uploading public objects	True
Remove public access granted through public ACLs	True
Block new public bucket policies	True
Block public and cross-account access if bucket has public policies	True
System permissions	Disabled

[Previous](#) [Create bucket](#)

← → ↻ <https://s3.console.aws.amazon.com/s3/home?region=us-east-1#> 🔍 ☆ ⚙️

aws Services Resource Groups Andrei Arion Global Support

Welcome to Amazon S3. Create new buckets or select an existing bucket to view and configure properties. [Documentation](#)

Amazon S3

Buckets

Public access settings for this account

S3 buckets

[Discover the new console](#) [Quick tips](#)

🔍 Search for buckets

All access types

[+ Create bucket](#) [Edit public access settings](#) [Empty](#) [Delete](#)

1 Buckets 1 Regions

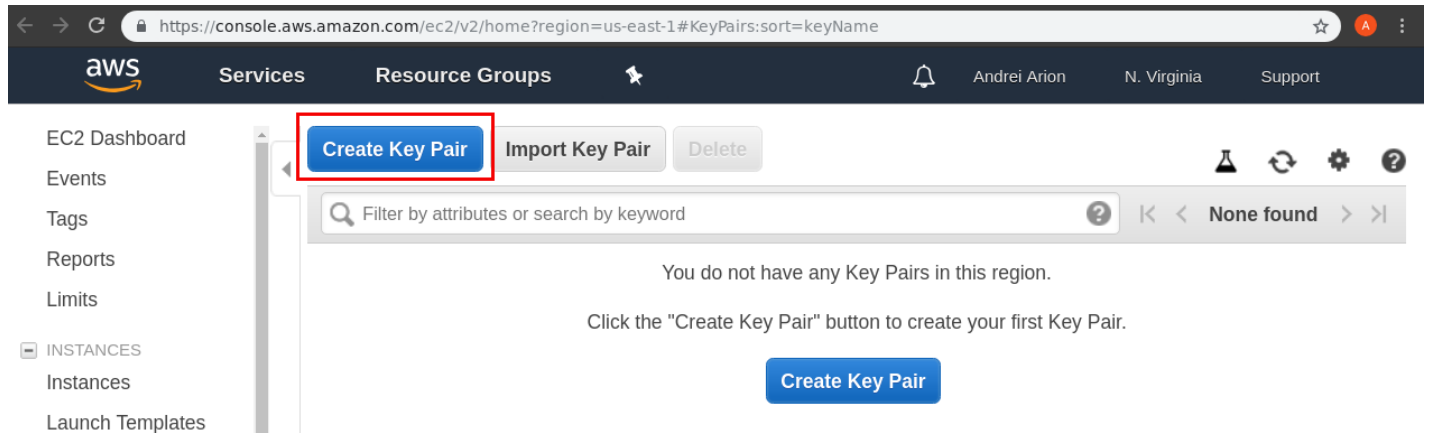
<input type="checkbox"/>	Bucket name	Access	Region	Date created
<input type="checkbox"/>	john-doe-telecom-gdelt2018	Bucket and objects not public	US East (N. Virginia)	Dec 12, 2018 8:31:39 AM GMT+0100

Démarrage d'un cluster sur AWS EMR

Configuration du votre compte AWS

1) Rendez-vous sur la console AWS et créez une paire de clefs *gdeltKeyPair*:

<https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#KeyPairs:sort=keyName>



2) la clé privée sera automatiquement sauvegardée après la création dans un fichier *gdeltKeyPair.pem* . Notez l'emplacement de ce fichier, vous en aurez besoin pour plus tard

La cle est telechargee automatiquement

Key Pair: gdeltKeyPair

Key pair name	Fingerprint
gdeltKeyPair	62:d5:36:ca:ef:b4:43:56:5d:58:de:05:b6:3f:79:c3:a6:d2:74:c0

gdeltKeyPair....pem

3) Allez sur <https://console.aws.amazon.com/iam/home?region=us-east-1#/groups> et creez un group *admin* avec une AdministratorAccess Policy

Create New Group

Group Name	Users	Inline Policy	Creation Time

The screenshot shows the AWS IAM console interface. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', and user information (Andrei Arion, Global, Support). The left sidebar contains the 'Create New Group Wizard' with three steps: 'Step 1: Group Name', 'Step 2: Attach Policy', and 'Step 3: Review'. The main content area is titled 'Set Group Name' and includes a text input field for 'Group Name' with the value 'admin'. Below this, the 'Attach Policy' section is active, displaying a table of available policies. The table has columns for 'Policy Name', 'Attached Entities', 'Creation Time', and 'Edited Time'. The 'AdministratorAccess' policy is highlighted with a red border and its checkbox is checked. At the bottom right, there are 'Cancel', 'Previous', and 'Next Step' buttons.

Set Group Name

Specify a group name. Group names can be edited any time.

Group Name:

Example: Developers or ProjectAlpha
Maximum 128 characters

Attach Policy

Select one or more policies to attach. Each group can have up to 10 policies attached.

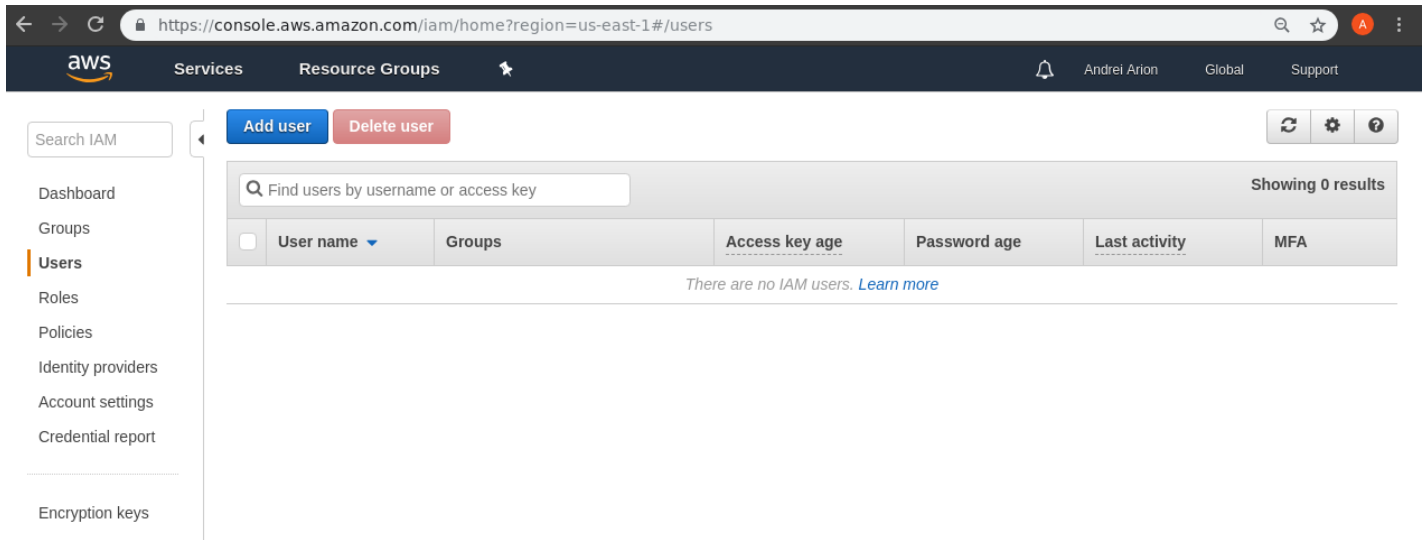
Filter: Policy Type Showing 403 results

	Policy Name	Attached Entities	Creation Time	Edited Time
<input type="checkbox"/>	AmazonElasticMapRed...	1	2015-02-06 19:41 UTC+0...	2017-08-12 01:57 U...
<input type="checkbox"/>	AmazonElasticMapRed...	1	2015-02-06 19:41 UTC+0...	2017-12-12 01:47 U...
<input checked="" type="checkbox"/>	AdministratorAccess	0	2015-02-06 19:39 UTC+0...	2015-02-06 19:39 U...
<input type="checkbox"/>	AlexaForBusinessDevic...	0	2017-11-30 17:47 UTC+0...	2017-11-30 17:47 U...
<input type="checkbox"/>	AlexaForBusinessFullAc...	0	2017-11-30 17:47 UTC+0...	2018-06-26 01:53 U...
<input type="checkbox"/>	AlexaForBusinessGate...	0	2017-11-30 17:47 UTC+0...	2017-11-30 17:47 U...
<input type="checkbox"/>	AlexaForBusinessRead...	0	2017-11-30 17:47 UTC+0...	2018-06-26 01:52 U...
<input type="checkbox"/>	AmazonAPIGatewayAd...	0	2015-07-09 19:34 UTC+0...	2015-07-09 19:34 U...
<input type="checkbox"/>	AmazonAPIGatewayInv...	0	2015-07-09 19:36 UTC+0...	2015-07-09 19:36 U...
<input type="checkbox"/>	AmazonAPIGatewayPu...	0	2015-11-12 00:41 UTC+0...	2015-11-12 00:41 U...
<input type="checkbox"/>	AmazonAppStreamFull...	0	2015-02-06 19:40 UTC+0...	2018-09-10 19:29 U...
<input type="checkbox"/>	AmazonAppStreamRea...	0	2015-02-06 19:40 UTC+0...	2016-12-07 22:00 U...
<input type="checkbox"/>	AmazonAppStreamServ...	0	2016-11-19 05:17 UTC+0...	2018-08-13 20:19 U...
<input type="checkbox"/>	AmazonAthenaFullAccess	0	2016-11-30 17:46 UTC+0...	2018-05-21 20:47 U...
<input type="checkbox"/>	AmazonChimeFullAccess	0	2017-11-01 23:15 UTC+0...	2017-11-01 23:15 U...
<input type="checkbox"/>	AmazonChimeReadOnly	0	2017-11-01 23:04 UTC+0...	2018-03-30 18:24 U...
<input type="checkbox"/>	AmazonChimeUserMan...	0	2017-11-01 23:17 UTC+0...	2018-10-29 23:40 U...
<input type="checkbox"/>	AmazonCloudDirectory...	0	2017-02-25 01:41 UTC+0...	2017-02-25 01:41 U...
<input type="checkbox"/>	AmazonCloudDirectory...	0	2017-03-01 00:42 UTC+0...	2017-03-01 00:42 U...
<input type="checkbox"/>	AmazonCognitoDevelop...	0	2015-03-24 18:22 UTC+0...	2015-03-24 18:22 U...
<input type="checkbox"/>	AmazonCognitoPowerU...	0	2015-03-24 18:14 UTC+0...	2016-06-02 18:57 U...
<input type="checkbox"/>	AmazonCognitoReadOnly	0	2015-03-24 18:06 UTC+0...	2016-06-02 19:30 U...
<input type="checkbox"/>	AmazonConnectFullAcc...	0	2018-10-17 22:59 UTC+0...	2018-10-18 00:28 U...
<input type="checkbox"/>	AmazonConnectReadO...	0	2018-10-17 23:00 UTC+0...	2018-10-17 23:00 U...
<input type="checkbox"/>	AmazonDMSCloudWatc...	0	2016-01-08 00:44 UTC+0...	2016-01-08 00:44 U...

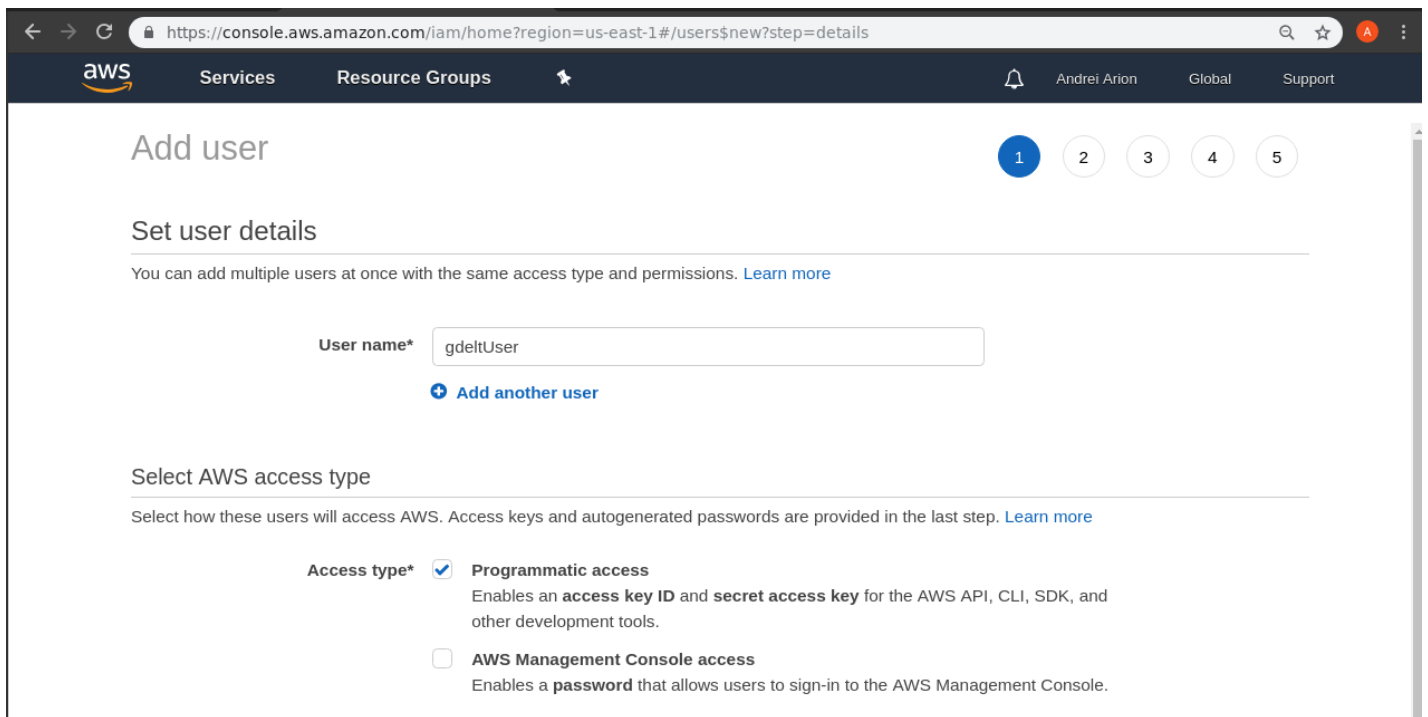
[Cancel](#) [Previous](#) [Next Step](#)

4) Créez un utilisateur pour la gestion de vos clusters:

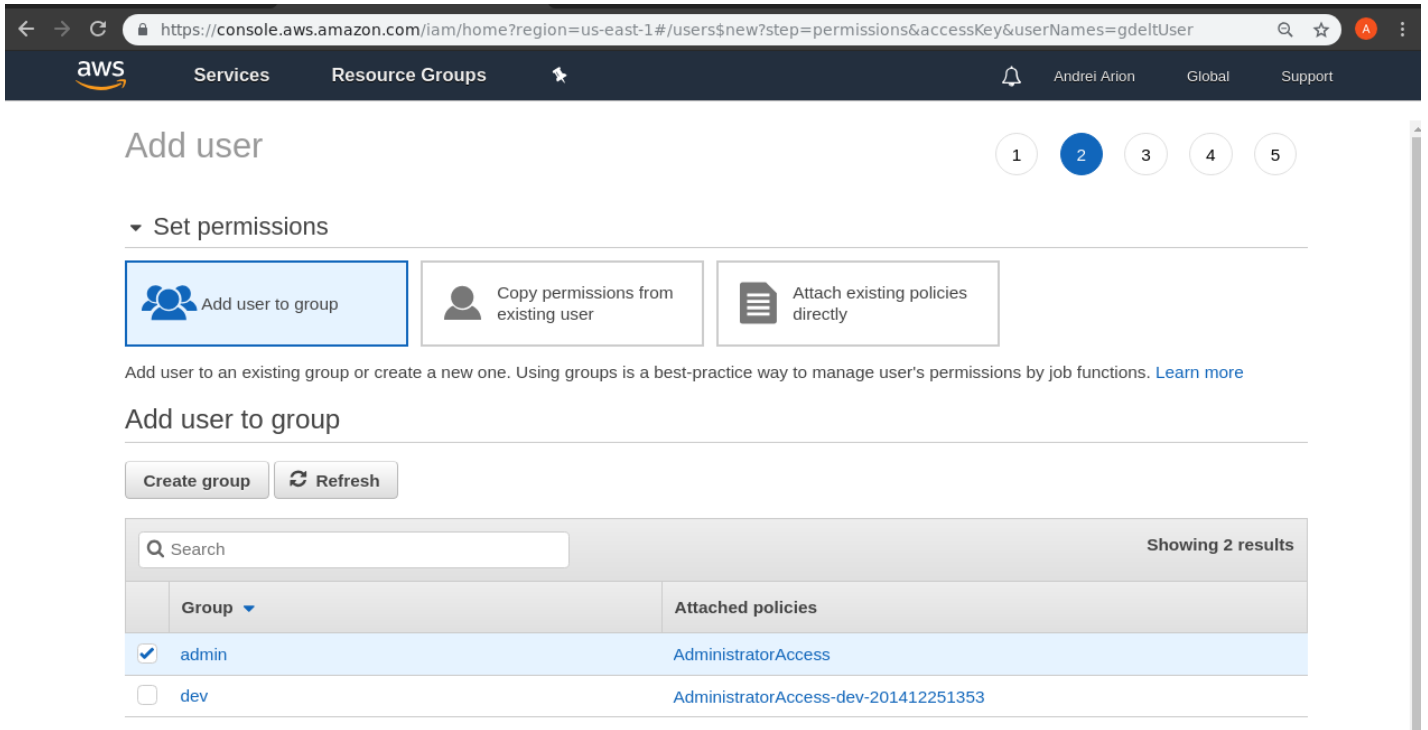
a) Allez sur la console IAM (<https://console.aws.amazon.com/iam/home?region=us-east-1#/users>) et cliquez sur *Add User*



b) Mettez comme nom d'utilisateur: *gdeltUser* et cochez la case **Programmatic access** pour créer un identifiant d'accès et une clé de sécurité (*access key ID* and *secret access key*)

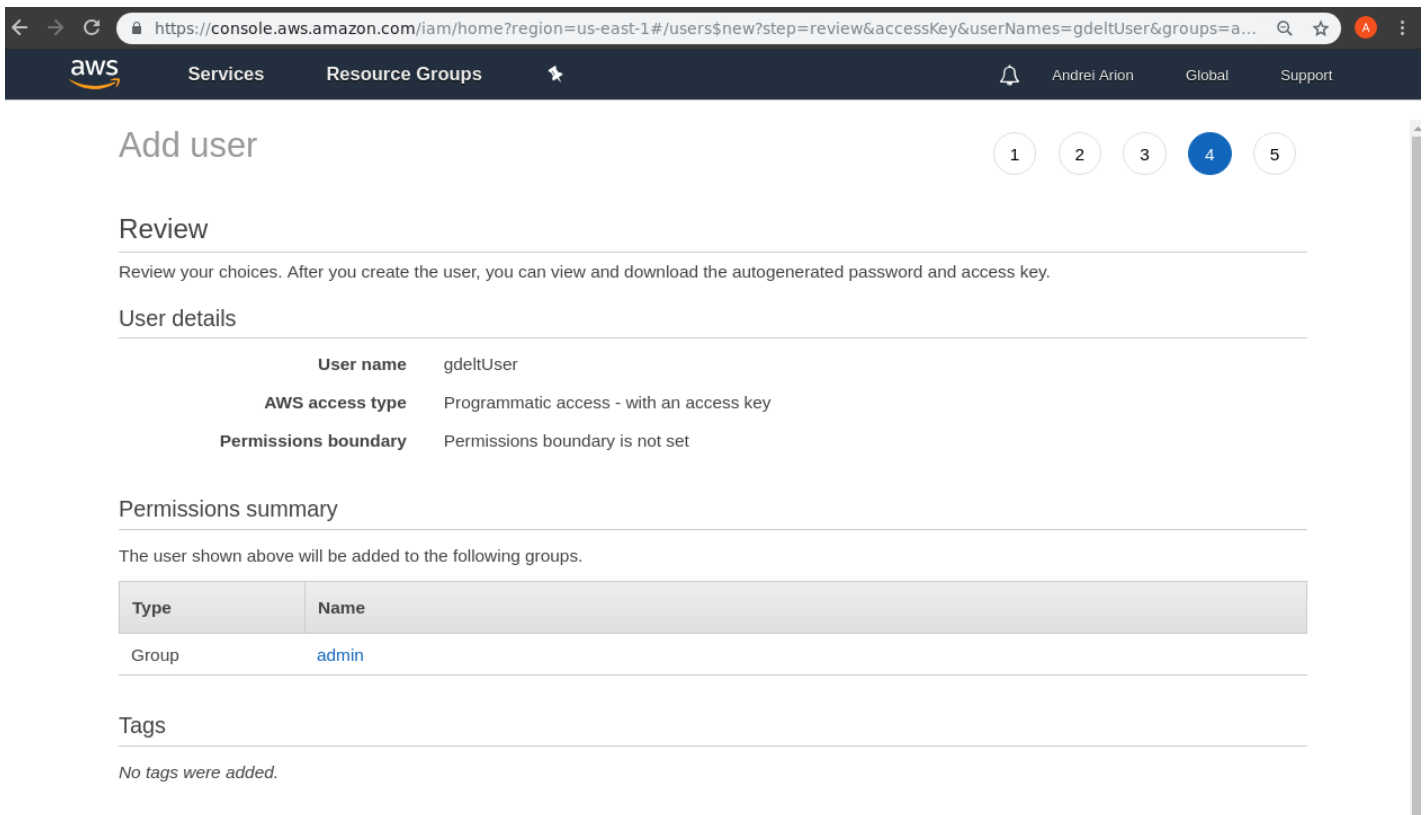


c) A l'étape suivante mettez votre utilisateur dans le groupe administrateur et validez la création de l'utilisateur



The screenshot shows the AWS IAM console 'Add user' page, step 2: Set permissions. The breadcrumb trail shows steps 1 through 5, with step 2 highlighted. The 'Set permissions' section has three options: 'Add user to group' (selected), 'Copy permissions from existing user', and 'Attach existing policies directly'. Below this, a message states: 'Add user to an existing group or create a new one. Using groups is a best-practice way to manage user's permissions by job functions. [Learn more](#)'. The 'Add user to group' section includes a 'Create group' button and a 'Refresh' button. A search bar is present, and a table shows 'Showing 2 results'.

Group	Attached policies
<input checked="" type="checkbox"/> admin	AdministratorAccess
<input type="checkbox"/> dev	AdministratorAccess-dev-201412251353



The screenshot shows the AWS IAM console 'Add user' page, step 4: Review. The breadcrumb trail shows steps 1 through 5, with step 4 highlighted. The 'Review' section contains the text: 'Review your choices. After you create the user, you can view and download the autogenerated password and access key.' Below this is the 'User details' section:

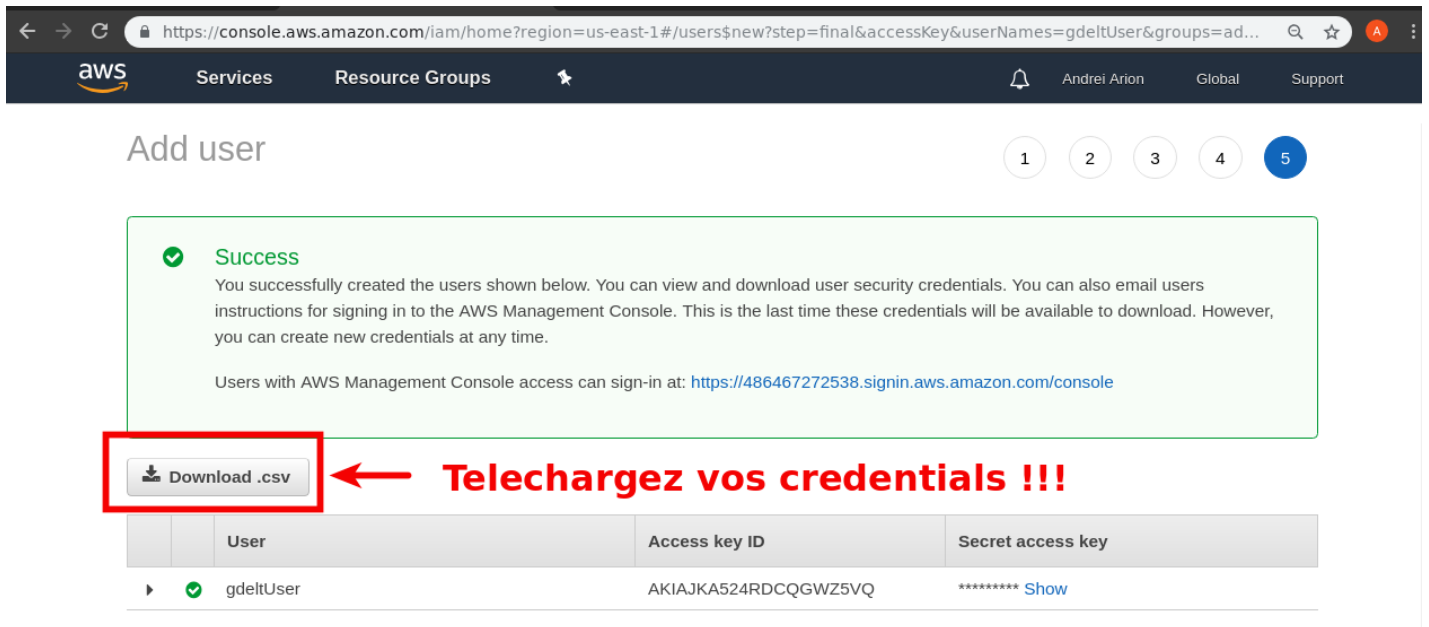
User name	gdeltUser
AWS access type	Programmatic access - with an access key
Permissions boundary	Permissions boundary is not set

The 'Permissions summary' section states: 'The user shown above will be added to the following groups.'

Type	Name
Group	admin

The 'Tags' section states: 'No tags were added.'

d) Sur la page de confirmation de la création de votre utilisateur, cliquez sur *Download csv* pour sauvegarder dans un fichier *credentials.csv* l'ID et la clé de sécurité



The screenshot shows the AWS IAM console 'Add user' page. A green success message states: 'You successfully created the users shown below. You can view and download user security credentials. You can also email users instructions for signing in to the AWS Management Console. This is the last time these credentials will be available to download. However, you can create new credentials at any time. Users with AWS Management Console access can sign-in at: <https://486467272538.signin.aws.amazon.com/console>'. Below the message is a 'Download .csv' button, which is highlighted with a red box and a red arrow pointing to it. To the right of the button, the text 'Telechargez vos credentials !!!' is written in red. Below the message and button is a table with the following data:

User	Access key ID	Secret access key
gdeltUser	AKIAJKA524RDCQGWZ5VQ	***** Show



Pour simplifier les procédures nous allons utiliser un utilisateur avec des droits d'admin. En général ce n'est pas recommandé, pour des raisons de sécurité d'utiliser des comptes avec trop de droits (si quelqu'un arrive à mettre la main sur votre identifiant d'accès et votre clé de sécurité il pourra démarrer des machines en votre nom). Nous vous conseillons de désactiver cet utilisateur à la fin du TP et créer un avec des droits plus spécifiques.

Démarrage d'un cluster de 3 noeuds via AWS EMR

Nous allons utiliser la console *AWS EMR* pour démarrer notre cluster:

1. Allez dans la console *AWS EMR*: <https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#> et cliquer sur *Create cluster*

Amazon EMR

- Clusters
- Security configurations
- VPC subnets
- Events
- Notebooks
- 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](#)

2. Modifiez les paramètres suivants puis validez:

- GdeltCluster* pour le nom du cluster
- sélectionnez dans Applications \Rightarrow *Spark*
- instance_type* \Rightarrow *m3.large*
- key pair* \Rightarrow *gdeltKeyPair*

← → ↻ <https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#quick-create:> 🔍 ☆

aws Services Resource Groups 🔍

🔔 Andrei Arion 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 2.8.5 with Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4
- ☐ HBase: HBase 1.4.7 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.3, Hue 4.2.0, Phoenix 4.14.0, and ZooKeeper 3.4.13
- ☐ Presto: Presto 0.212 with Hadoop 2.8.5 HDFS and Hive 2.3.3 Metastore
- ☒ Spark: Spark 2.3.2 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.0

☐ Use AWS Glue Data Catalog for table metadata ⓘ

Hardware configuration

Instance type

Number of instances (1 master and 2 core nodes)

Security and access

EC2 key pair ⓘ [Learn how to create an EC2 key pair.](#)

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](#) ⓘ

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

3. Votre cluster est en train de démarrer:

The screenshot shows the AWS Management Console for an Amazon EMR cluster named 'GdeltCluster'. The cluster is in the 'Starting' state. The left sidebar shows the navigation menu with 'Clusters' selected. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', and user information (Andrei Arion, N. Virginia, Support).

Cluster: GdeltCluster Starting

Buttons: Clone, Terminate, AWS CLI export

Tabs: Summary, Application history, Monitoring, Hardware, Events, Steps, Configurations, Bootstrap actions

Connections: --
Master public DNS: --
Tags: -- [View All / Edit](#)

Summary

ID: j-27WRR2WQCJ8OQ
 Creation date: 2018-12-12 09:52 (UTC+1)
 Elapsed time: 0 seconds
 Auto-terminate: No
 Termination protection: Off [Change](#)

Configuration details

Release label: emr-5.19.0
 Hadoop distribution: Amazon
 Applications: Ganglia 3.7.2, Spark 2.3.2, Zeppelin 0.8.0
 Log URI: s3://aws-logs-486467272538-us-east-1/elasticmapreduce/
 EMRFS consistent view: Disabled
 Custom AMI ID: --

Network and hardware

Availability zone: --
 Subnet ID: [subnet-9a19d5c3](#)
 Master: Provisioning 1 m1.large
 Core: Provisioning 2 m1.large
 Task: --

Security and access

Key name: gdeltKeyPair
 EC2 instance profile: EMR_EC2_DefaultRole
 EMR role: EMR_DefaultRole
 Visible to all users: All [Change](#)
 Security groups for Master:
 Security groups for Core & Task:

4. Dans quelques minutes votre cluster aura démarré, notez l'adresse de votre master Spark:

The screenshot shows the AWS Management Console for the same Amazon EMR cluster 'GdeltCluster', now in the 'Waiting' state. The cluster is ready after the last step completed. The left sidebar and top navigation bar are the same as in the previous screenshot.

Cluster: GdeltCluster Waiting Cluster ready after last step completed.

Buttons: Clone, Terminate, AWS CLI export

Tabs: Summary, Application history, Monitoring, Hardware, Events, Steps, Configurations, Bootstrap actions

Connections: [Enable Web Connection](#) – Zeppelin, Spark History Server, Ganglia, Resource Manager ... (View All)
Master public DNS: ec2-18-232-58-202.compute-1.amazonaws.com [SSH](#)
Tags: -- [View All / Edit](#)

Summary

ID: j-27WRR2WQCJ8OQ
 Creation date: 2018-12-12 09:52 (UTC+1)
 Elapsed time: 19 minutes
 Auto-terminate: No
 Termination protection: Off [Change](#)

Configuration details

Release label: emr-5.19.0
 Hadoop distribution: Amazon
 Applications: Ganglia 3.7.2, Spark 2.3.2, Zeppelin 0.8.0
 Log URI: s3://aws-logs-486467272538-us-east-1/elasticmapreduce/
 EMRFS consistent view: Disabled
 Custom AMI ID: --

Network and hardware

Availability zone: us-east-1a
 Subnet ID: [subnet-9a19d5c3](#)
 Master: Running 1 m1.large
 Core: Running 2 m1.large
 Task: --

Security and access

Key name: gdeltKeyPair
 EC2 instance profile: EMR_EC2_DefaultRole
 EMR role: EMR_DefaultRole
 Visible to all users: All [Change](#)
 Security groups for [sg-1307c177](#) [\(ElasticMapReduce-Master\)](#): master
 Security groups for [sg-1c07c178](#) [\(ElasticMapReduce-Core & Task\)](#): slave

5. Rajoutez une regle de firewall dans le security group du master pour permettre l'accès SSH:

The screenshot shows the AWS Management Console interface. The top navigation bar includes 'Services', 'Resource Groups', and user information. The main content area displays the details for an EMR cluster named 'test' in the 'Starting' state. The 'Summary' tab is selected, showing cluster ID, creation date, and other metadata. The 'Security and access' section highlights the security groups associated with the cluster: 'sg-1307c177' for the master node and 'sg-1c07c178' for the slave nodes. A red box highlights the 'sg-1307c177' link.

Below the cluster details, the 'Edit inbound rules' modal is open for the 'sg-1307c177' security group. The modal shows a table of existing inbound rules, including several 'Custom TCP' rules for SSH access from specific IP ranges. A new rule is being added at the bottom, highlighted with a red box. The new rule is named 'SSH', uses 'TCP' protocol, allows traffic on port '22' from 'Anywhere' (0.0.0.0/0).

Protocol	Port Range	Source	Description
Custom TCP	8443	72.21.196.64/29	e.g. SSH for Admin Desktop
Custom TCP	8443	72.21.198.64/29	e.g. SSH for Admin Desktop
Custom TCP	8443	72.21.217.0/24	e.g. SSH for Admin Desktop
Custom TCP	8443	54.240.217.8/29	e.g. SSH for Admin Desktop
Custom TCP	8443	54.240.217.16/29	e.g. SSH for Admin Desktop
Custom TCP	8443	54.240.217.64/28	e.g. SSH for Admin Desktop
Custom TCP	8443	54.240.217.80/29	e.g. SSH for Admin Desktop
Custom TCP	8443	54.239.98.0/24	e.g. SSH for Admin Desktop
All UDP	0 - 65535	sg-1307c177	e.g. SSH for Admin Desktop
All UDP	0 - 65535	sg-1c07c178	e.g. SSH for Admin Desktop
All ICMP - IPv4	0 - 65535	sg-1307c177	e.g. SSH for Admin Desktop
All ICMP - IPv4	0 - 65535	sg-1c07c178	e.g. SSH for Admin Desktop
SSH	22	Anywhere (0.0.0.0/0)	SSH from anywhere

NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Buttons: Cancel, Save

6. Pour accéder aux services qui tournent sur le master (Zeppelin -notebook spark, Ganglia - monitoring de ressources, Spark History Server / SparkUI etc) on doit passer par un tunnel SSH.

Vous pouvez utiliser la directions des ports via un tunnel `ssh` comme montre en TP pour accéder aux services via `http://localhost:PORT`.

Cependant on peut rediriger tous les ports a la fois via la meme commande SSH et on peut utiliser un proxy web (ex: FoxProxy) pour faire directement la translation *master-public-dns-name:PORT* \Rightarrow `localhost:PORT`. Pour cela cliquez sur **Enable Web Connection** et suivez la documentation fournie.

La liste complete des services/ports est la suivante:

Name of interface	URI
YARN ResourceManager	<code>http://<i>master-public-dns-name</i>:8088/</code>
YARN NodeManager	<code>http://<i>coretask-public-dns-name</i>:8042/</code>
Hadoop HDFS NameNode	<code>http://<i>master-public-dns-name</i>:50070/</code>
Hadoop HDFS DataNode	<code>http://<i>coretask-public-dns-name</i>:50075/</code>
Spark HistoryServer	<code>http://<i>master-public-dns-name</i>:18080/</code>
Zeppelin	<code>http://<i>master-public-dns-name</i>:8890/</code>
Hue	<code>http://<i>master-public-dns-name</i>:8888/</code>
Ganglia	<code>http://<i>master-public-dns-name</i>/ganglia/</code>
HBase UI	<code>http://<i>master-public-dns-name</i>:16010/</code>



Si vous avez un probleme lors de l'établissement du tunnel `ssh` vérifiez/corrigez les permissions sur votre cle:

```
[aar@wifibridge 2018]# chmod og-rwx ~/Downloads/gdeltKeyPair.pem
```

BASH

Installation de AWS CLI

1. Pour certaines operation ca peut etre pratique d'installer sur votre machine le client AWS([awscli](https://docs.aws.amazon.com/cli/latest/userguide/cli-chap-install.html) (<https://docs.aws.amazon.com/cli/latest/userguide/cli-chap-install.html>)).
2. Une fois le client installe, utilisez la commande ***aws configure*** pour configurer votre installation (inserez votre *Access Key ID* et votre *Secret Access Key* (que vous avez sauvegardé dans `credentials.csv`), spécifiez la région par default à *us-east-1*, et le type de log par défaut à *text*:
3. Pour verifier que la configuration est correcte on peut essayer d'afficher le contenu du bucket S3 precedemment cree:

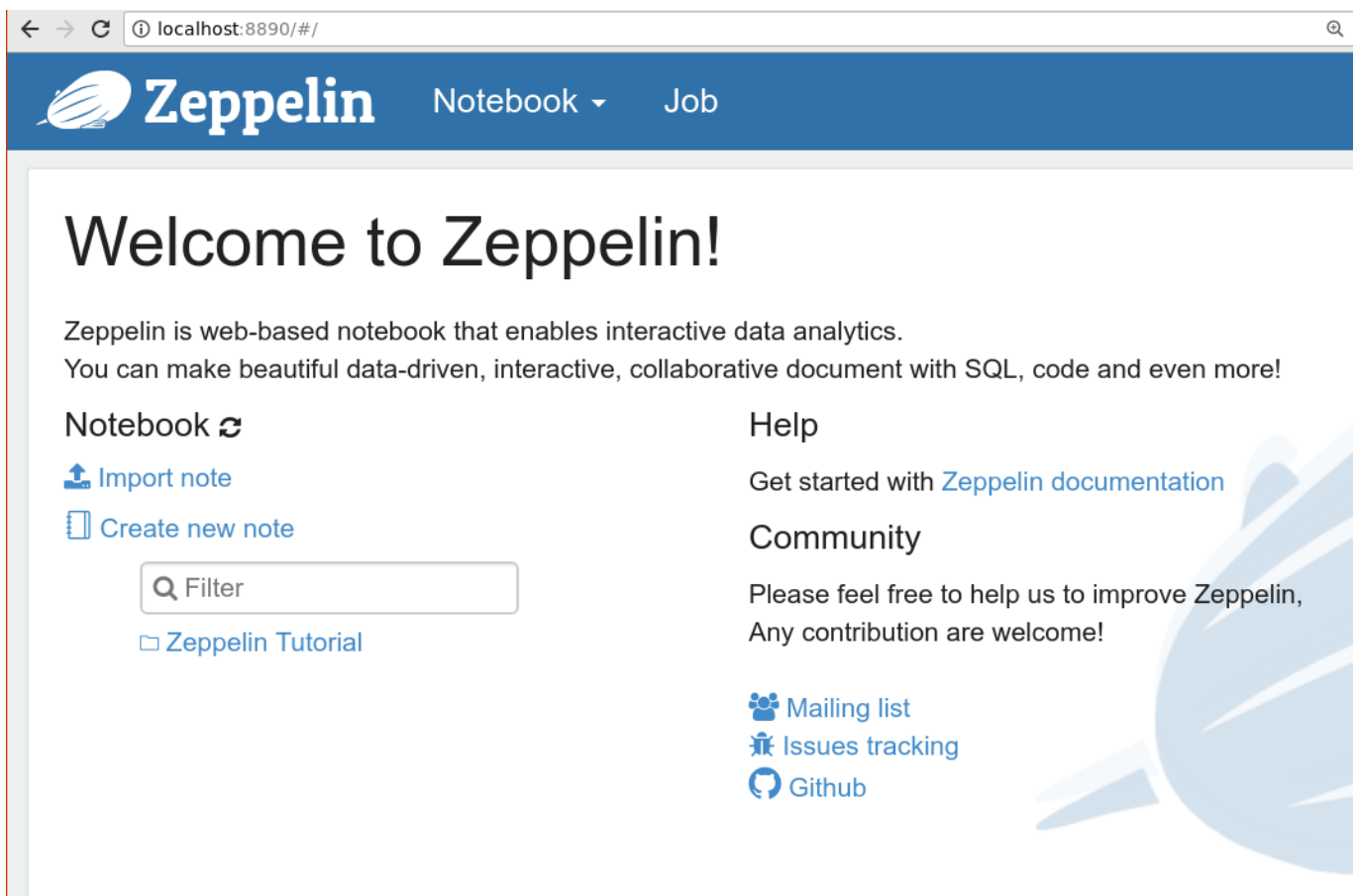
```
[aar@wifibridge 2018]# aws configure 1
AWS Access Key ID [None]: *****JVBA
AWS Secret Access Key [None]: *****EiQv
Default region name [None]: us-east-1
Default output format [None]: text
```

```
[aar@wifibridge 2018]# aws s3 ls --summarize --human-readable --recursive s3://john-doe-telecom-gdelt2018/
```

```
Total Objects: 0
Total Size: 0 Bytes
```

Connexion à l'interface du Zeppelin

1. Ouvrez un navigateur vers <http://master-public-dns-name:8890> (avec la configuration FoxProxy active!) et vous aurez accès à l'interface du Zeppelin



2. Importer le notebooks suivants dans Zeppelin et suivez les instructions

- a. [gdeltETL.json](http://andreiarion.github.com/gdeltETL.json) (<http://andreiarion.github.com/gdeltETL.json>)
- b. [gdeltExploration.json](http://andreiarion.github.com/gdeltExploration.json) (<http://andreiarion.github.com/gdeltExploration.json>)

Éteindre votre cluster

Allez a <https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1>, sélectionner votre cluster et cliquez sur Terminate.

Cluster: GdeltCluster **Waiting** Cluster ready after last step completed.

Connections: [Zeppelin](#), [Spark History Server](#), [Ganglia](#), [Resource Manager](#) ... (View All)

Master public DNS: ec2-18-232-58-202.compute-1.amazonaws.com [SSH](#)

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

Summary

- ID: j-27WRR2WQCJ8OQ
- Creation date: 2018-12-12 09:52 (UTC+1)
- Elapsed time: 1 hour, 37 minutes
- Auto-terminate: No
- Termination protection: Off [Change](#)

Configuration details

- Release label: emr-5.19.0
- Hadoop distribution: Amazon
- Applications: Ganglia 3.7.2, Spark 2.3.2, Zeppelin 0.8.0
- Log URI: s3://aws-logs-486467272538-us-east-1/elasticmapreduce/
- EMRFS consistent view: Disabled
- Custom AMI ID: --

Network and hardware

- Availability zone: us-east-1a
- Subnet ID: [subnet-9a19d5c3](#)
- Master: **Running** 1 m1.large
- Core: **Running** 2 m1.large
- Task: --

Security and access

- Key name: gdeltKeyPair
- EC2 instance profile: EMR_EC2_DefaultRole
- EMR role: EMR_DefaultRole
- Visible to all users: All [Change](#)
- Security groups for [sg-1307c177](#) (ElasticMapReduce-Master: master)
- Security groups for [sg-1c07c178](#) (ElasticMapReduce-Core & Task: slave)

Assurez-vous que votre cluster est arrete:

You can use the AWS Glue Data Catalog as your external Hive metastore for [Apache Spark](#), [Apache Hive](#), and [Presto](#) workloads on Amazon EMR release 5.10.0 and later. To get started, simply select the AWS Glue Data Catalog for table metadata when creating your cluster.

[Create cluster](#) [View details](#) [Clone](#) [Terminate](#)

Filter: [All clusters](#) 1 cluster (all loaded)

	Name	ID	Status	Creation time (UTC+1)	Elap:
<input type="checkbox"/>	GdeltCluster	j-27WRR2WQCJ8OQ	Terminated User request	2018-12-12 09:52 (UTC+1)	1 hou

Last updated 2018-12-12 16:29:21 CET