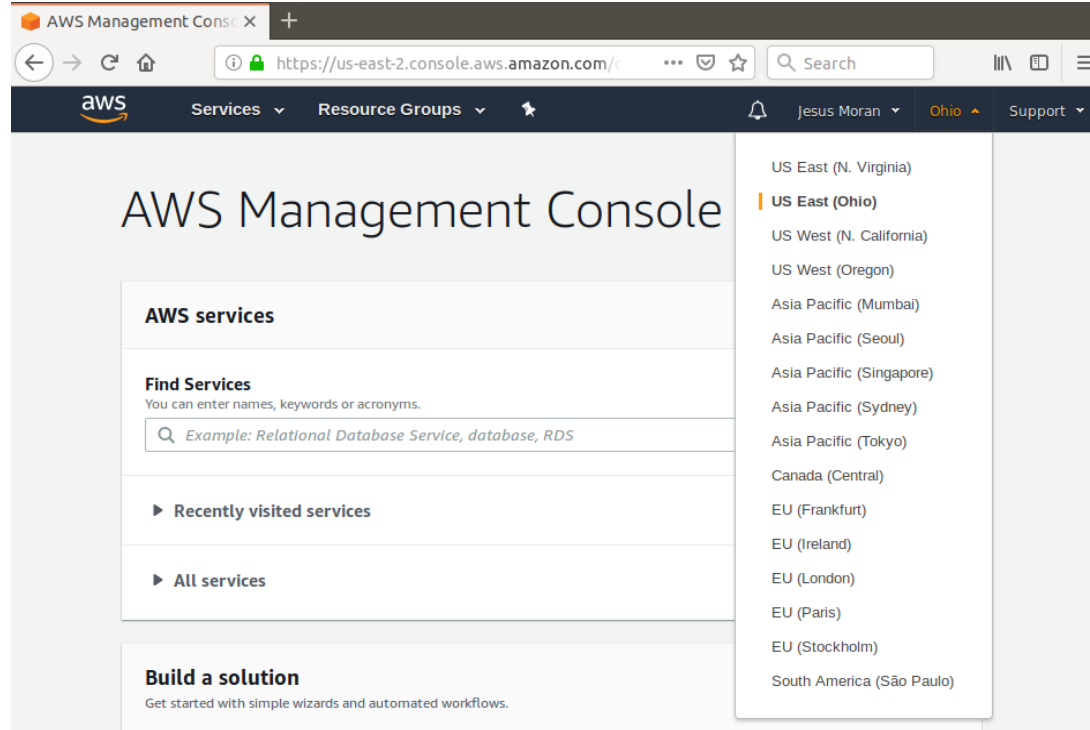


# Procesamiento de datos masivos

Jesús Morán

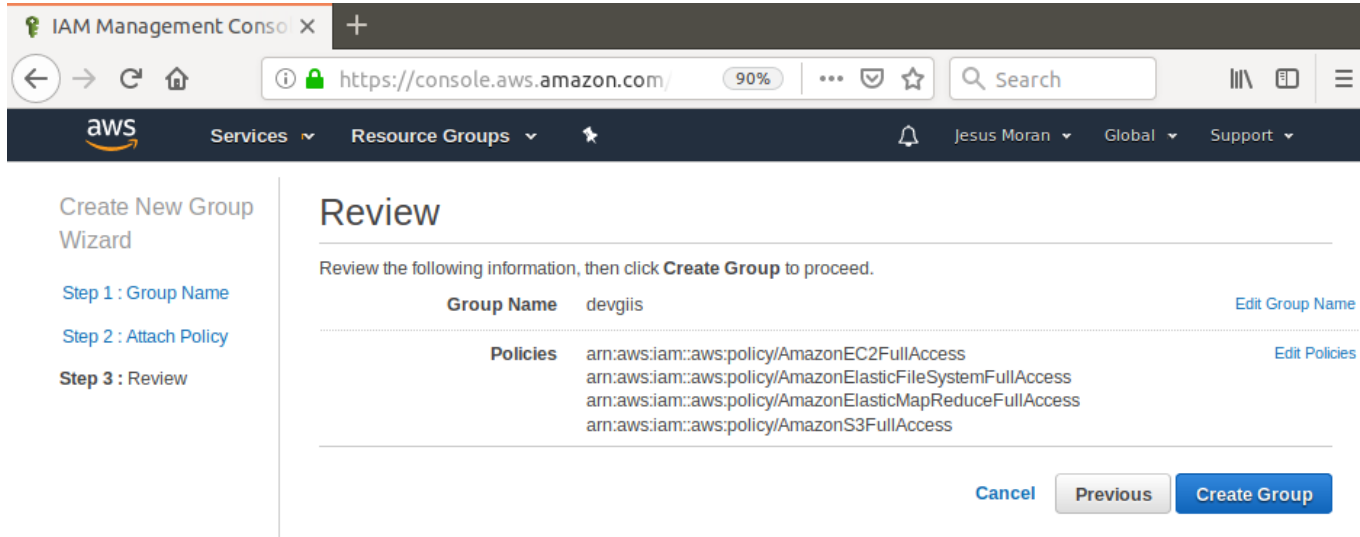
- Crear cuentas root y delegada
  - Servicio IAM
- Almacenamiento masivo de datos
  - Servicio S3
- Crear cluster de computación Big Data
  - Servicio EMR

## 1. Crear cuenta root



1. Crear cuenta root
2. **Crear grupo y permisos (políticas)**

Ej. Permitir al grupo acceso completo a EC2, EMR y S3



1. Crear cuenta root
2. Crear grupo y permisos (políticas)
3. **Crear usuario y asignarlo a grupo**

User name	jesus
AWS access type	Programmatic access and AWS Management Console access
Console password type	Custom
Require password reset	No
Permissions boundary	Permissions boundary is not set

### Permissions summary

The user shown above will be added to the following groups.


Type	Name
Group	<a href="#">devgiis</a>

### Tags

The new user will receive the following tag


Key	Value
e-mail	moranjesus@uniovi.es


1. Crear cuenta root
2. Crear grupo y permisos (políticas)
3. **Crear usuario y asignarlo a grupo**
  1. Se logea desde una URL específica con su usuario-contraseña
  2. Se puede logear desde CLI y API con Access.key.id (usuario) y secret.key (contraseña)

 **Success**

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://31\[REDACTED\]0.signin.aws.amazon.com/console](https://31[REDACTED]0.signin.aws.amazon.com/console)

 Download .csv

	User	Access key ID	Secret access key	Email login instructions
▶	✓ jesu	AK[REDACTED]NI	***** Show	Send email 

1. Crear cuenta root
2. Crear grupo y permisos (políticas)
3. Crear usuario y asignarlo a grupo
4. **Crear roles (permisos a los servicios)**

Ej. EMR\_DefaultRole para y EMR\_EC2\_DefaultRole para permitir al servicio EMR lanzar instancias EC2 y S3

	Role name ▼	Description	Trusted entities
<input type="checkbox"/>	<a href="#">AWSServiceRole...</a>	Enables resource access for AWS to ...	<b>AWS service:</b> support (Service-Linked role)
<input type="checkbox"/>	<a href="#">AWSServiceRole...</a>	Access for the AWS Trusted Advisor ...	<b>AWS service:</b> trustedadvisor (Service-Linked...
<input type="checkbox"/>	<a href="#">EMR_DefaultRole</a>	Allows Elastic MapReduce to call AW...	<b>AWS service:</b> elasticmapreduce
<input type="checkbox"/>	<a href="#">EMR_EC2_Defa...</a>	Allows EC2 instances in an Elastic M...	<b>AWS service:</b> ec2

1. Crear cuenta root
2. Crear grupo y permisos (políticas)
3. Crear usuario y asignarlo a grupo
4. Crear roles (permisos a los servicios)
5. **Logearnos con el nuevo usuario**
  1. Entrar en la URL de login del usuario (no del root)
  2. Poner usuario-contraseña
  3. El usuario sólo tendrá acceso a los servicios que le permitió el usuario root (EC2, EFS, S3 y EMR)



### 1. Crear bucket

- Hay que asignarle un nombre único (no pueden existir buckets de otros usuarios con ese nombre)
- Indicamos si lo queremos privado, público, etc.
- Permisos

+ Create bucket

Edit public access settings


Empty

Delete


1 Regions


↺


1 Buckets


<input type="checkbox"/>	Bucket name ▼	Access ⓘ ▼	Region ▼	Date created ▼
<input type="checkbox"/>	 inhtest-giis-bucket-1	Bucket and objects not public	US East (Ohio)	Apr 17, 2019 5:24:56 PM GMT+0200


1. Crear bucket
2. **Crear carpetas**

 Upload



 Create folder

 Download

 Actions ▾

US East (Ohio) 

Viewing 1 to 2

<input type="checkbox"/>	Name ▾	Last modified ▾	Size ▾	Storage class ▾
<input type="checkbox"/>	 Data	--	--	--
<input type="checkbox"/>	 Programs	--	--	--

Viewing 1 to 2

1. Crear bucket
2. Crear carpetas
3. **Añadir datos**

Aunque en la imagen sólo se añade 1GB, se pueden añadir datos de forma masiva

The screenshot shows the 'Upload' interface for an AWS S3 bucket. The top bar is blue with a large blue arrow icon on the left and a close button (X) on the right. Below the bar, there are four steps: 'Select files' (checked), 'Set permissions' (checked), 'Set properties' (checked), and 'Review' (active, indicated by a circled 4). The main content area is dark blue and contains several sections: 'Files' (with '3036 Files' and 'Size: 1.1 GB'), 'Permissions' (with '1 grantees'), 'Properties' (with 'Encryption No' and 'Storage class Standard'), 'Metadata', and 'Tag'. Each section has an 'Edit' link on the right.

# Crear cluster Big Data (EMR)

## 1. Crear cluster

### General Configuration

Cluster name

☐ Logging ⓘ

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

### Software configuration

Release  ⓘ

- Applications
- ☒ Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.4, Hue 4.3.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.1
  - ☐ HBase: HBase 1.4.9 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.4, Hue 4.3.0, Phoenix 4.14.1, and ZooKeeper 3.4.13
  - ☐ Presto: Presto 0.214 with Hadoop 2.8.5 HDFS and Hive 2.3.4 Metastore
  - ☐ Spark: Spark 2.4.0 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.1

☐ Use AWS Glue Data Catalog for table metadata ⓘ

### Hardware configuration

Instance type  The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. [Learn more](#)

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](#) ⓘ

### 1. Crear cluster

Crea automáticamente instancias EC2 y volúmenes

<input type="checkbox"/>	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public IP
<input type="checkbox"/>		i-04[REDACTED]	m4.large	us-east-2c	 running	 2/2 checks ...	None	 ec2-18-2
<input type="checkbox"/>		i-06[REDACTED]	m4.large	us-east-2c	 running	 2/2 checks ...	None	 ec2-18-1
<input type="checkbox"/>		i-08[REDACTED]	m4.large	us-east-2c	 running	 2/2 checks ...	None	 ec2-3-14

Crea automáticamente grupos de seguridad

<input type="checkbox"/>	Name	Group ID	Group Name	VPC ID	Owner	Description
<input type="checkbox"/>		sg-075[REDACTED]	ElasticMapReduce-master	vpc-e9[REDACTED]	31[REDACTED]	Master
<input type="checkbox"/>		sg-08[REDACTED]	ElasticMapReduce-slave	vpc-e9[REDACTED]	31[REDACTED]	Slave

(tienen distintas reglas de cortafuegos)

1. Crear cluster
2. Ejecutamos programas (steps)

Add step

Step type

Custom JAR

Name\*

Wordcount in EMR

JAR location\*

grams/hadoop-mapreduce-examples-2.4.0.jar

JAR location maybe a path into S3 or a fully qualified java class in the classpath.

Arguments

`wordcount  
s3://inhtest-giis-bucket-1/Data/Books/  
s3://inhtest-giis-bucket-1/OutputWC1`

These are passed to the main function in the JAR. If the JAR does not specify a main class in its manifest file you can specify another class name as the first argument.

Action on failure

Continue

What to do if the step fails.

Cancel





Add

1. Crear cluster
2. Ejecutamos programas (steps)
3. Esperamos hasta que finalice

The screenshot shows the AWS EMR console interface. At the top, there's a navigation bar with 'Resource Groups', a user profile 'jesus @ 3', and regional settings 'Ohio' and 'Support'. Below this, there are three buttons: 'Clone', 'Terminate', and 'AWS CLI export'. The main heading is 'Cluster: inhtest\_giis' followed by a green 'Waiting' status and the text 'Cluster ready after last step completed.' Below this, there are several tabs: 'Summary', 'Application history', 'Monitoring', 'Hardware', 'Configurations', 'Events', 'Steps', and 'Bootstrap actions'. A paragraph explains that Amazon EMR collects information from YARN applications and keeps historical information for up to seven days. Below this, the 'YARN applications (1)' section is shown, featuring a filter dropdown set to 'All applications', a search bar, and a count of '1 applications (all loaded)'. A table lists the application details.

Application ID	Type	Action	Status	Start time (UTC+2)	Duration
▶ application_1555511454980_0001	MapReduce	word count	Succeeded	2019-04-17 19:17 (UTC+2)	1.1 h

1. Crear cluster
2. Ejecutamos programas (steps)
3. Esperamos hasta que finalice
4. **Observamos el resultado almacenado en S3**

<input type="checkbox"/> Name ▼	Last modified ▼	Size ▼	Storage class ▼
<input type="checkbox"/>  _SUCCESS	Apr 17, 2019 8:26:26 PM GMT+0200	0 B	Standard
<input type="checkbox"/>  part-r-00000	Apr 17, 2019 8:26:11 PM GMT+0200	4.3 MB	Standard
<input type="checkbox"/>  part-r-00001	Apr 17, 2019 8:26:22 PM GMT+0200	4.3 MB	Standard
<input type="checkbox"/>  part-r-00002	Apr 17, 2019 8:26:25 PM GMT+0200	4.3 MB	Standard

Viewing 1 to 4



## 1. Permitimos tráfico ssh entrante al maestro

### Edit inbound rules

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ	Description ⓘ	
All TCP ▾	TCP	0 - 65535	Custom ▾ sg-07 <b>Maestro</b> 2	e.g. SSH for Admin Desktop	✕
All TCP ▾	TCP	0 - 65535	Custom ▾ sg-08 <b>Esclavos</b> e	e.g. SSH for Admin Desktop	✕
Custom TCP ▾	TCP	8443	Custom ▾ 52 <b>EMR</b>	e.g. SSH for Admin Desktop	✕
All UDP ▾	UDP	0 - 65535	Custom ▾ sg-07 <b>Maestro</b> 2	e.g. SSH for Admin Desktop	✕
All UDP ▾	UDP	0 - 65535	Custom ▾ sg-08 <b>Esclavos</b> e	e.g. SSH for Admin Desktop	✕
All ICMP - IP ▾	ICMP	0 - 65535	Custom ▾ sg-07 <b>Maestro</b> 2	e.g. SSH for Admin Desktop	✕
All ICMP - IP ▾	ICMP	0 - 65535	Custom ▾ sg-08 <b>Esclavos</b> e	e.g. SSH for Admin Desktop	✕
SSH ▾	TCP	22	Anywhere ▾ 0.0.0.0/0, ::/0	e.g. SSH for Admin Desktop	✕

Add Rule

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.

Cancel

Save

Es mejor poner nuestra IP o un rango de IPs de nuestra empresa

1. Permitimos tráfico ssh entrante al maestro
2. **Conexión ssh con clave privada**

```
ssh -i CLAVE_PRIVADA NOMBRE_DNS_PÚBLICO
```

1.

2.

```

hadoop@ip-172-3[REDACTED]:~
jesus@jesus-PC:~/Escritorio/introAWS$ ssh -i giis1key.pem hadoop@ec2-18-1[REDACTED].us-east-2.
compute.amazonaws.com
The authenticity of host 'ec2-18-1[REDACTED].us-east-2.compute.amazonaws.com (18.1[REDACTED])'
can't be established.
ECDSA key fingerprint is SHA256:P[REDACTED]s.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-18-1[REDACTED].us-east-2.compute.amazonaws.com,18.1[REDACTED]'
(ECDSA) to the list of known hosts.
Last login: Wed Apr 17 17:35:44 2019

  _ | _ | _ )
 _ | ( _ /   Amazon Linux AMI
__| \__| __|

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
7 package(s) needed for security, out of 12 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M          M::::::::M R:::::::::R
EE::::::::EEEEEEEE::::E M::::::::M          M::::::::M R::::RRRRRR:::::R
  E::::E          EEEE M::::::::M          M::::::::M RR::::R          R::::R
  E::::E          M::::M:M::M          M::M::::M          R:::R          R::::R
  E::::EEEEEEEEEE M::::M M::M M::M M::::M          R::RRRRRR:::::R
  E::::::::::::E M::::M M::M:M::M M::::M          R:::::::::RR
  E::::EEEEEEEEEE M::::M M::::M M::::M          R::RRRRRR:::::R
  E::::E          M::::M M::M M::::M          R:::R          R::::R
  E::::E          EEEE M::::M          MMM M::::M          R:::R          R::::R
EE::::::::EEEEEEEE::::E M::::M          M::::M          R:::R          R::::R
E::::::::::::E M::::M          M::::M RR::::R          R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRR          RRRRRR

[hadoop@ip-172-3[REDACTED] ~]$

```

1. Permitimos tráfico ssh entrante al maestro
2. Conexión ssh con clave privada
3. **Observamos la salida del programa**

```
hadoop fs -ls s3a://inhtest-giis-bucket-1/OutputWC1/
```

```
hadoop@ip-172-3[redacted]:~$  
[hadoop@ip-172-3[redacted] ~]$ hadoop fs -ls s3a://inhtest-giis-bucket-1/OutputWC1/  
Found 4 items  
-rw-rw-rw-  1 hadoop hadoop      0 2019-04-17 18:26 s3a://inhtest-giis-bucket-1/OutputWC1/_SUCCESS  
-rw-rw-rw-  1 hadoop hadoop 4496933 2019-04-17 18:26 s3a://inhtest-giis-bucket-1/OutputWC1/part-r-00000  
-rw-rw-rw-  1 hadoop hadoop 4508986 2019-04-17 18:26 s3a://inhtest-giis-bucket-1/OutputWC1/part-r-00001  
-rw-rw-rw-  1 hadoop hadoop 4508681 2019-04-17 18:26 s3a://inhtest-giis-bucket-1/OutputWC1/part-r-00002  
[hadoop@ip-172-3[redacted] ~]$
```

1. Permitimos tráfico ssh entrante al maestro
2. Conexión ssh con clave privada
3. **Observamos la salida del programa**

```
hadoop fs -ls s3a://inhtest-giis-bucket-1/OutputWC1/
```

Hay comandos que nos requieren el access.key y secret.key

```
hadoop fs -Dfs.s3a.access.key=<NuestroAccessKey> -  
Dfs.s3a.secret.key=<NuestroSecret.Key> -ls  
s3a://inhtest-giis-bucket-1/OutputWC1/
```

# Gracias