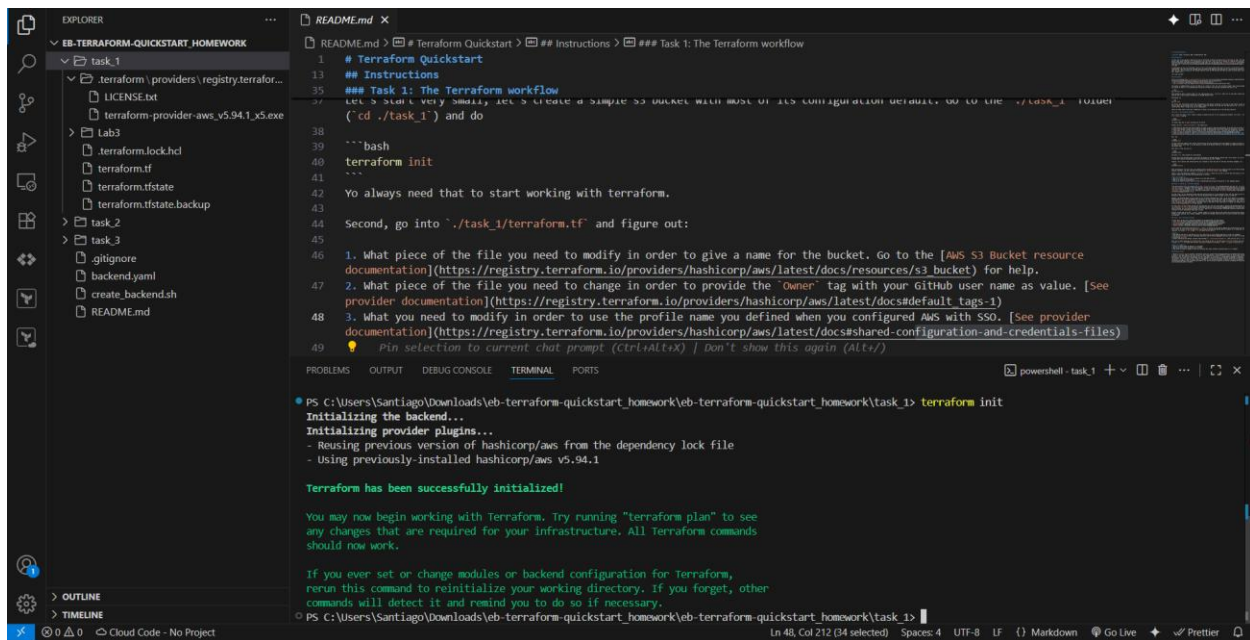


First command: aws configure

Tener en cuenta la ruta del task

Go to the task\_1 path and execute: terraform init

Terraform apply



## PANDAS

```
Location: /usr/local/lib/python3.10/dist-packages
Requires: numpy, python-dateutil, pytz, tzdata
Required-by:
$ Santiago
sh: 5: Santiago: not found
$ python3
Python 3.10.12 (main, Aug 15 2025, 14:32:43) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import pandas as pd
print(pd.__version__)

df = pd.DataFrame({"A": [1, 2, 3], "B": ["x", "y", "z"]})
print(df)
>>> print(pd.__version__)
2.3.3
>>>
>>> df = pd.DataFrame({"A": [1, 2, 3], "B": ["x", "y", "z"]})
>>> print(df)
   A  B
0  1  x
1  2  y
2  3  z
>>>
```

## POLARS

```
>>>
$ print(df.to_pandas())
sh: 20: Syntax error: word unexpected (expecting ")")
$ python3
Python 3.10.12 (main, Aug 15 2025, 14:32:43) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import polars as pl

df = pl.DataFrame({
    "id": [1, 2, 3],
    >>>
>>> df = pl.DataFrame({
...     "id": [1, 2, 3],
...     "nombre": ["Alice", "Bob", "Charlie"],
...     "edad": [25, 30, 35]
... })
>>>
>>> pl.Config.set_tbl_formatting("ASCII_FULL") # salida más limpia
<class 'polars.config.Config'>i
>>> print(df)
shape: (3, 3)
+-----+-----+-----+
| id | nombre | edad |
| --- | --- | --- |
| i64 | str | i64 |
+=====+
| 1 | Alice | 25 |
+-----+-----+-----+
| 2 | Bob | 30 |
+-----+-----+-----+
| 3 | Charlie | 35 |
+-----+-----+-----+
>>> 
```

## DUCK

```
> print(resultado)
> EOF
> cat > test_duckdb.py << 'EOF'
> import duckdb
> import pandas as pd
>
df> = pd.DataFrame({
>     "id": [1, 2, 3, 4],
>     "nombre": ["Alice", "Bob", "Charlie", "Diana"],
>     "edad": [25, 30, 35, 40]
}> )
>
> con = duckdb.connect()
> con.register("personas", df)
>
> resultado = con.execute("SELECT nombre, edad FROM personas WHERE edad > 30").fetchdf()
>
> print("Resultado de DuckDB:")
> print(resultado)
E> OF
> python3 test_duckdb.py
> apt-get update -y
> apt-get install -y python3 python3-pip
> python3 -m pip install --upgrade pip setuptools wheel
p> ython3 -m pip install duckdb
> apt-get update -y
> apt-get install -y python3 python3-pip
p> ython3 -m pip install --upgrade pip setuptools wheel
> python3 -m pip install duckdb
> apt-get update -y
> apt-get install -y python3 python3-pip
> python3 -m pip install --upgrade pip setuptools wheel
> python3 -m pip install duckdb
> apt-get update -y
> apt-get install -y python3 python3-pip
p> ython3 -m pip install --upgrade pip setuptools wheel
> python3 -m pip install duckdb
> python3 -c "import duckdb; print(duckdb.__version__)"
```

## SPARK

```
PS C:\Users\Santiago\Documents\EAFIT\Grandes Volúmenes de Datos\Taller_3\task_3\Polars copy 2> terraform apply -auto-approve
instance_public_ip = "18.212.86.183"
PS C:\Users\Santiago\Documents\EAFIT\Grandes Volúmenes de Datos\Taller_3\task_3\Polars copy 2> aws ssm start-session --target i-0c3dfcf4bd80e9c33

Starting session with SessionId: root-q8j2sx6zetyijubxjfo8zevte
$ pyspark --version
sh: 1: pyspark: not found
$ python3 -m pip show pyspark
WARNING: Package(s) not found: pyspark
$ python3 -m pyspark
/usr/bin/python3: No module named pyspark
$ sudo apt-get update -y
sudo apt-get install -y python3-pip
pip3 install pyspark
sudo apt-get install -y python3-pip
pip3 install pyspark
Hit:1 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:3 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:4 http://security.ubuntu.com/ubuntu jammy-security InRelease
Reading package lists... Done
$ python3 -m pip show pyspark
Name: pyspark
Version: 4.0.1
Summary: Apache Spark Python API
Home-page: https://github.com/apache/spark/tree/master/python
Author: Spark Developers
Author-email: dev@spark.apache.org
License: http://www.apache.org/licenses/LICENSE-2.0
Location: /usr/local/lib/python3.10/dist-packages
Requires: py4j
Required-by:
$ python3 -m pyspark --version
/usr/bin/python3: No module named pyspark.__main__; 'pyspark' is a package and cannot be directly executed
$ python3
Python 3.10.12 (main, Aug 15 2025, 14:32:43) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> from pyspark.sql import SparkSession
```

## SPARK

```
PS C:\Users\Santiago\Documents\EAFIT\Grandes Volúmenes de Datos\Taller_3\task_3\EMR> aws emr describe-cluster --cluster-id j-3617LDYAXQIF8 --query "Cluster.State"
● us.State
"WAITING"

PS C:\Users\Santiago\Documents\EAFIT\Grandes Volúmenes de Datos\Taller_3\task_3\EMR> aws emr list-instances --cluster-id j-3617LDYAXQIF8 --instance-group-types
● MASTER --query "Instances[*].Ec2InstanceId" --output text
i-0a6ef27cf4be79182

PS C:\Users\Santiago\Documents\EAFIT\Grandes Volúmenes de Datos\Taller_3\task_3\EMR> aws ssm start-session --target i-0a6ef27cf4be79182

Starting session with SessionId: root-hol3ijoyzhoxl3jbvilsid888
sh-4.2$ python3 --version
pyspark --vePython 3.7.16
sh-4.2$ pyspark --version
Welcome to

  ____      _
 / ___|  __| | | |
| |  | | | | | | | |
| |  | | | | | | |
| |  | | | | | | |
| |  | | | | | | |
|_|  |_| |_| |_| |_|

version 3.4.1-amzn-2

Using Scala version 2.12.15, OpenJDK 64-Bit Server VM, 1.8.0_462
Branch
Compiled by user release on 2024-03-14T04:28:38Z
Revision
Url
Type --help for more information.
sh-4.2$ █
```

AWS SCREENSHOTS

S3

General purpose buckets

All AWS Regions

Directory buckets

General purpose buckets (2) Info

Refresh

Copy ARN

Empty

Delete

Create bucket

Buckets are containers for data stored in S3.

Find buckets by name

< 1 >

Settings

	Name	AWS Region	Creation date
<input type="radio"/>	<a href="#">emr-logs-407b3f</a>	US East (N. Virginia) us-east-1	September 30, 2025, 21:12:32 (UTC-05:00)
<input type="radio"/>	<a href="#">miprimeracana</a>	US East (N. Virginia) us-east-1	September 28, 2025, 20:23:36 (UTC-05:00)

EC2

Instances (1/4) Info

Refresh

Connect

Instance state

Actions

Launch instances

Find Instance by attribute or tag (case-sensitive)

All states

Instance state = running

Clear filters

< 1 >

Settings

	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IP
<input type="checkbox"/>		i-01b34c64c164e69b7	Running	t3.micro	3/3 checks passed	View alarms +	us-east-1a	ec2-18-2...
<input type="checkbox"/>	EMR-Spark-CL...	i-0a6ef27cf4be79182	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1d	ec2-52-9...
<input type="checkbox"/>	EMR-Spark-CL...	i-09db53e4da650179e	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1d	ec2-18-2...
<input checked="" type="checkbox"/>	EMR-Spark-CL...	i-0fbcf2a1e89d9a06a	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1d	ec2-18-2...

3 INSTANCIAS PARA EL EMR

Instances (1/4) Info

Refresh

Connect

Instance state

Actions

Launch instances

Find Instance by attribute or tag (case-sensitive)

All states

Instance state = running

Clear filters

< 1 >

Settings

Public IPv4 DNS	Public IPv4 ...	Elastic IP	IPv6 IPs	Monitoring	Security group name	Key name	Launch time
:2-18-212-209-163.co...	18.212.209.163	-	-	disabled	allow_tls	-	2025/09/28 21
:2-52-91-48-130.com...	52.91.48.130	-	-	disabled	emr-master-sg-407b3f	emr-key	2025/09/30 22
:2-18-233-155-29.co...	18.233.155.29	-	-	disabled	emr-slave-sg-407b3f	emr-key	2025/09/30 22
:2-18-204-216-173.co...	18.204.216.173	-	-	disabled	emr-slave-sg-407b3f	emr-key	2025/09/30 22

IAM ROLES

Identity and Access Management (IAM)

Search IAM

Dashboard

Access management

User groups

Users

Roles

Policies

Identity providers

Account settings

Root access management

Access reports

Access Analyzer

Resource analysis

Unused access

Analyzer settings

Roles (10)

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Search

Role name

Trusted entities

Last activity

<input type="checkbox"/>	<a href="#">AWSServiceRoleForEMRCleanup</a>	AWS Service: elasticmapreduce (Service-Linked Role)	38 minutes ago
<input type="checkbox"/>	<a href="#">AWSServiceRoleForOrganizations</a>	AWS Service: organizations (Service-Linked Role)	-
<input type="checkbox"/>	<a href="#">AWSServiceRoleForSSO</a>	AWS Service: sso (Service-Linked Role)	8 hours ago
<input type="checkbox"/>	<a href="#">AWSServiceRoleForSupport</a>	AWS Service: support (Service-Linked Role)	-
<input type="checkbox"/>	<a href="#">AWSServiceRoleForTrustedAdvisor</a>	AWS Service: trustedadvisor (Service-Linked Role)	-
<input type="checkbox"/>	<a href="#">EMR_AutoScaling_DefaultRole</a>	AWS Service: application-autoscaling (Service-Linked Role)	-
<input type="checkbox"/>	<a href="#">EMR_DefaultRole</a>	AWS Service: elasticmapreduce (Service-Linked Role)	-
<input type="checkbox"/>	<a href="#">EMR_EC2_DefaultRole</a>	AWS Service: ec2 (Service-Linked Role)	-
<input type="checkbox"/>	<a href="#">emr-ec2-role-407b3f</a>	AWS Service: ec2	8 minutes ago
<input type="checkbox"/>	<a href="#">emr-service-role-407b3f</a>	AWS Service: elasticmapreduce	5 minutes ago

EMR

spark-emr-cluster

Updated less than a minute ago

Terminate

Clone in AWS CLI

Clone

This EMR release reaches End of Support on Jan-24-2026 and will no longer be eligible for technical support. AWS strongly recommends that you run your workloads on the latest Amazon EMR release to receive security-critical updates and fixes. To learn more, see [EMR Standard Support policy](#).

▼ Summary

Cluster info

Cluster ID  
j-3617LDYAXQIF8

Cluster ARN  
[arn:aws:elasticmapreduce:us-east-1:915449291988:cluster/j-3617LDYAXQIF8](#)

Cluster configuration  
Instance groups

Capacity  
1 Primary | 2 Core | 0 Task

Applications

Amazon EMR version  
emr-6.15.0

Installed applications  
Hadoop 3.3.6, Spark 3.4.1

Cluster management

Log destination in Amazon S3  
[emr-logs-407b3f/logs](#)

Persistent application UIs  
[Spark History Server](#)  
[YARN Timeline Server](#)

Primary node public DNS  
[ec2-52-91-48-130.compute-1.amazonaws.com](#)  
Connect to the Primary node using SSH  
Connect to the Primary node using SSM

Status and time

Status  
Waiting

Creation time  
September 30, 2025, 22:33 (UTC-05:00)

Elapsed time  
19 minutes, 58 seconds