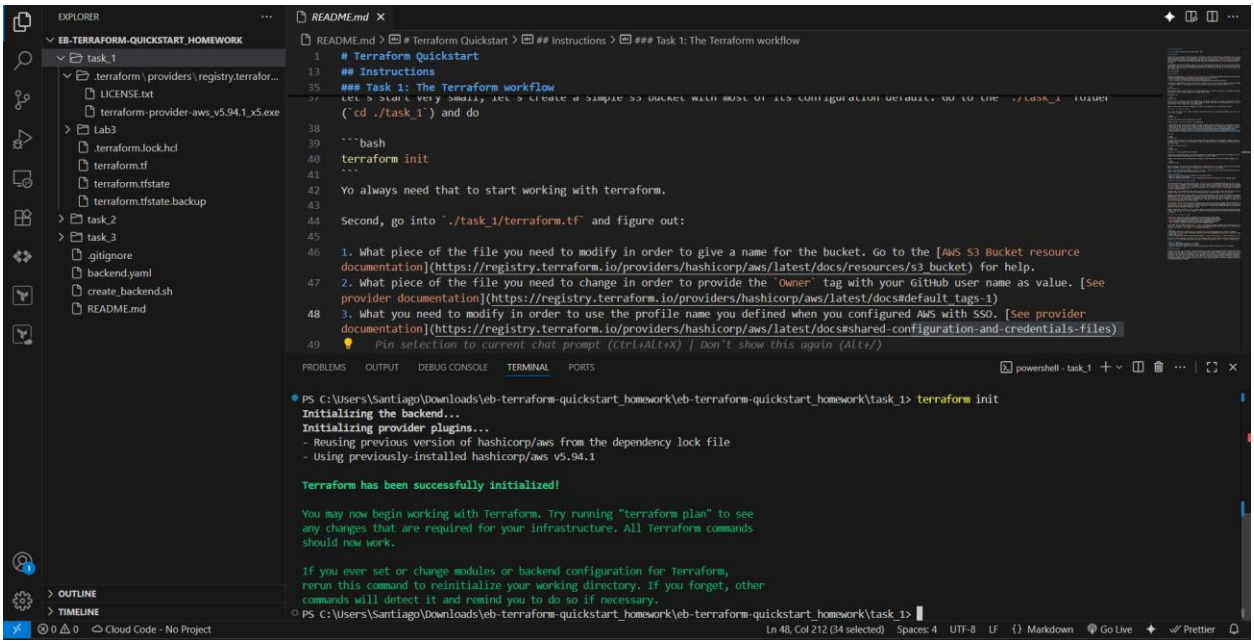
First command: aws configure

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

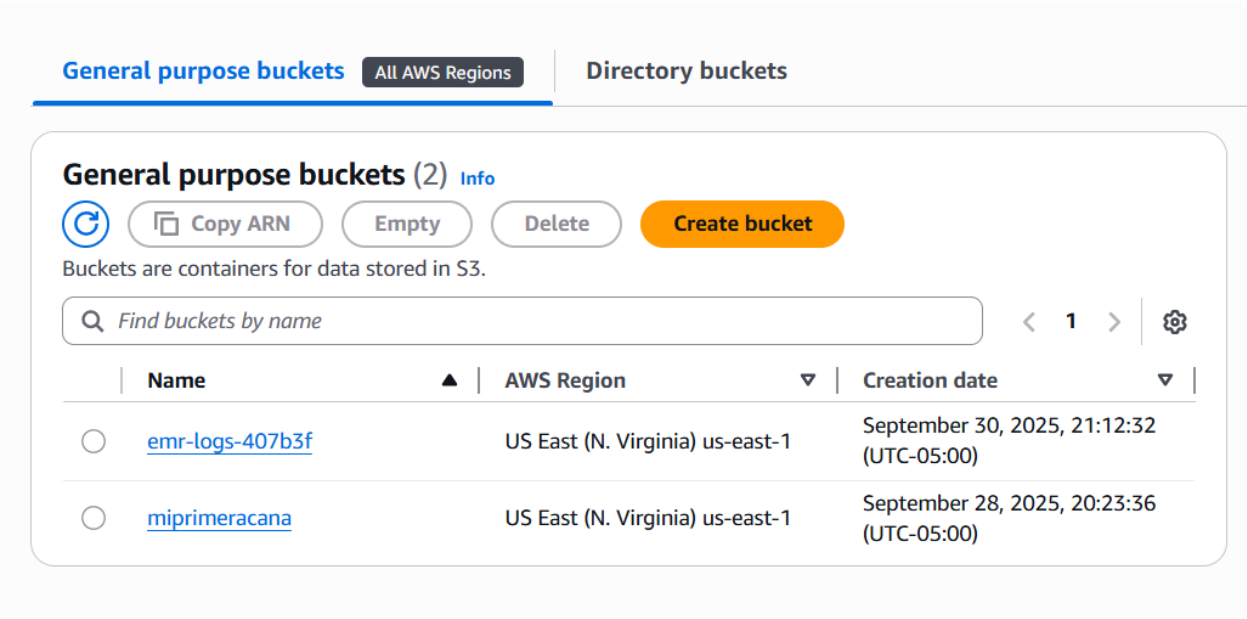
```
PS C:\Users\Santiago\Documents\EAFIT\Grandes Volumenes de Datos\Taller_3\task_3\Polars copy 2> terraform apply -auto-approve
instance_public_ip = "18.212.86.183"
PS C:\Users\Santiago\Documents\EAFIT\Grandes Volumenes de Datos\Taller_3\task_3\Polars copy 2> aws ssm start-session --target i-0c3dfcf4bd80e9c33

Starting session with SessionId: root-q8j2sx6zetyijubjxjfo8zevte
$ 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

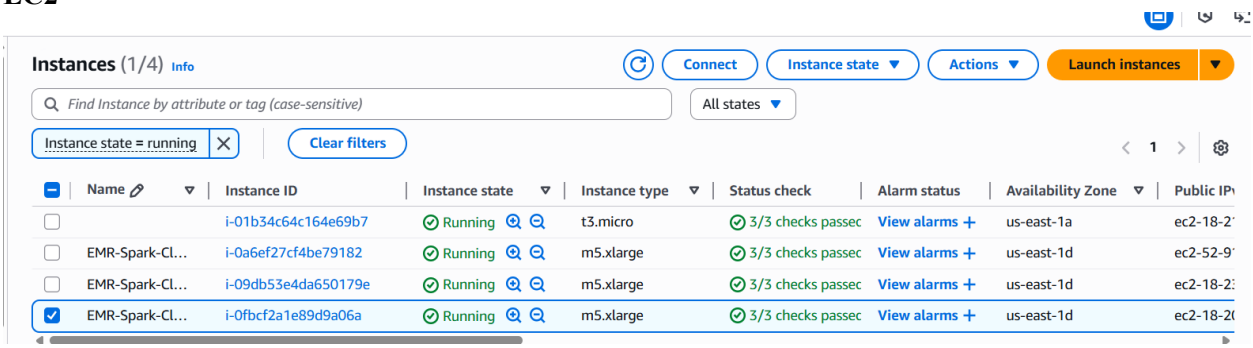
```
PS C:\Users\Santiago\Documents\EAFIT\Grandes Volumenes de Datos\Taller_3\task_3\EMR> aws emr describe-cluster --cluster-id j-3617LDYAXQIF8 --query "Cluster.Stat
us.State"
"WAITING"

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

Starting session with SessionId: root-hol3ijojy7hoxl3jbvil5id888
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

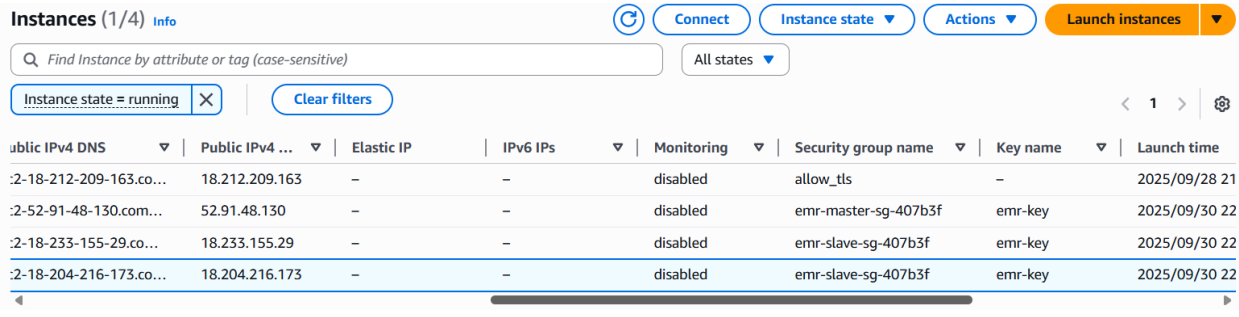
S3



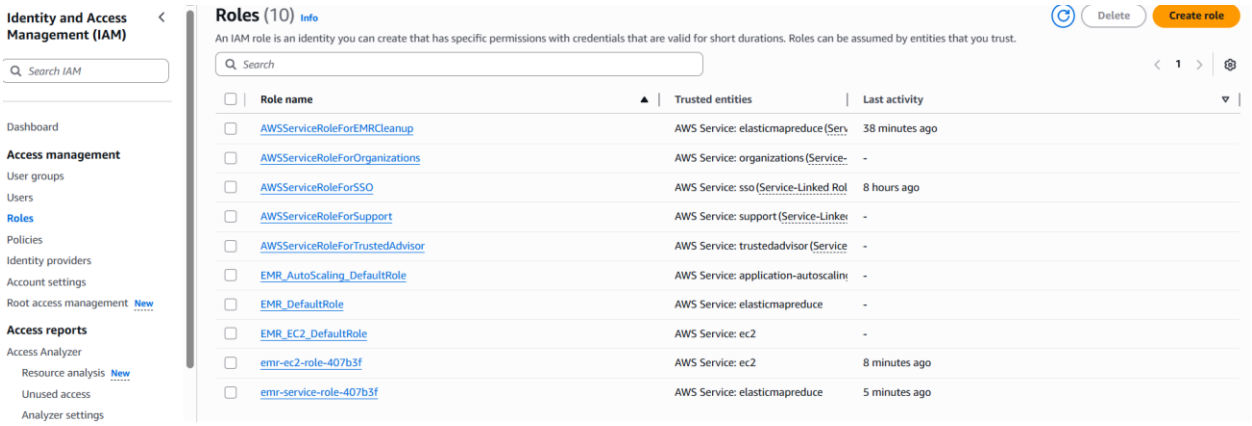
EC2



3 INSTANCIAS PARA EL EMR



IAM ROLES



EMR