



Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 501)

Instalación de Spark en AWS

Juan Pablo Castañeda Serrano

A01752030

1.- Impresión de pantalla del listado de instancias de EC2 de AWS en donde se muestre la instancia creada.

Instances (1/4) Info						
<input type="text" value="Find Instance by attribute or tag (case-sensitive)"/>						
<input type="checkbox"/>	Name ✎	Instance ID	Instance state ▼	Instance type ▼	Status ch	
<input type="checkbox"/>		i-05bc3faaaa1c113f0	Stopped	t2.micro	-	
<input checked="" type="checkbox"/>	IA_AvanzadaV2	i-03f50d48e992983f4	Running	t2.micro	-	
<input type="checkbox"/>		i-052968da6d919f636	Stopped	t2.micro	-	
<input type="checkbox"/>	AI_Avanzada_...	i-00100dcc8b3742fe6	Stopped	t2.micro	-	

2.- Impresión de pantalla conectado al servidor ya sea por Terminal o Putty, ya una vez dentro, ejecutar el comando `ls -l` para la toma de la impresión de pantalla.

```
(base) alfredogarcia@Alfredos-MacBook-Pro-2 ~ % clc
zsh: command not found: clc
(base) alfredogarcia@Alfredos-MacBook-Pro-2 ~ % chmod 400 Desktop/llaves_ec2_bigdata.pem

(base) alfredogarcia@Alfredos-MacBook-Pro-2 ~ % ssh -i Desktop/llaves_ec2_bigdata.pem ubuntu@3.141.21.101

The authenticity of host '3.141.21.101 (3.141.21.101)' can't be established.
ED25519 key fingerprint is SHA256:CcKrNzq5pvuX/Ib9YblplTsxjIww7qmrXysCcxcktVJw.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '3.141.21.101' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 6.2.0-1012-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

System information as of Tue Oct 31 03:44:26 UTC 2023

System load:  0.30126953125   Processes:            102
Usage of /:   5.4% of 28.89GB   Users logged in:      0
Memory usage: 22%             IPv4 address for eth0: 172.31.5.115
Swap usage:   0%

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

The list of available updates is more than a week old.
To check for new updates run: sudo apt update

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ip-172-31-5-115:~$
```

3.- Impresión de pantalla de la pestaña Detalles para que se vea la ip pública, la ip privada y el DNS público de la instancia (es necesario que la instancia esté Running).

Details

Security

Networking

Storage


Status checks

Monitoring

Tags

▼ Instance summary Info

Instance ID

 i-03f50d48e992983f4 (IA_AvanzadaV2)

IPv6 address

–


Hostname type

IP name: ip-172-31-5-115.us-east-2.compute.internal

Answer private resource DNS name

IPv4 (A)

Auto-assigned IP address

 3.141.21.101 [Public IP]


IAM Role

–


IMDSv2

Optional


Public IPv4 address

 3.141.21.101 [\[open address\]](#)

Instance state

 Running


Private IP DNS name (IPv4 only)

 ip-172-31-5-115.us-east-2.compute.internal


Instance type

t2.micro


VPC ID

 vpc-0949ac8bffa46deb [\[Public IP\]](#)


Subnet ID

 subnet-0a6bb23dd9b488bb8 [\[Public IP\]](#)

Private IPv4 addresses

 172.31.5.115


Public IPv4 DNS

 ec2-3-141-21-101.us-east-2.compute.amazonaws.com [\[open address\]](#)

Elastic IP addresses

–

AWS Compute Optimizer finding

 [Opt-in to AWS Compute Optimizer for recommendations.](#) | [Learn more](#)

Auto Scaling Group name

–

▼ Instance details Info

4.- Impresión de pantalla de la terminal o putty una vez que se ejecuta el comando jupyter notebook.

```
ubuntu@ip-172-31-5-115:~$ ls -l
total 0
ubuntu@ip-172-31-5-115:~$ jupyter notebook
Command 'jupyter' not found, but can be installed with:
sudo snap install jupyter # version 1.0.0, or
sudo apt install jupyter-core # version 4.9.1-1
See 'snap info jupyter' for additional versions.
ubuntu@ip-172-31-5-115:~$ sudo snap install jupyter
jupyter 1.0.0 from Jupyter Project (projectjupyter) installed
ubuntu@ip-172-31-5-115:~$ jupyter notebook
[I 03:48:26.492 NotebookApp] Writing notebook server cookie secret to /run/user/1000/snap.jupyter/jupyter/notebook_cookie_secret
[I 03:48:28.252 NotebookApp] Serving notebooks from local directory: /home/ubuntu
[I 03:48:28.252 NotebookApp] The Jupyter Notebook is running at:
http://localhost:8888/?token=5411352600fb32d709af2ff6e1710d5c84139f7a562ed11b
[I 03:48:28.253 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[W 03:48:28.258 NotebookApp] No web browser found: could not locate runnable browser.
[C 03:48:28.258 NotebookApp]

To access the notebook, open this file in a browser:
file:///run/user/1000/snap.jupyter/jupyter/nbserver-1427-open.html
Or copy and paste one of these URLs:
http://localhost:8888/?token=5411352600fb32d709af2ff6e1710d5c84139f7a562ed11b
```

5.- Impresión de pantalla de jupyter notebook visualizando el listado de los notebooks que se proporcionaron como ejemplos.

```
[1]: from pyspark import SparkContext
    sc = SparkContext()

    Setting default log level to "WARN".
    To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
    23/10/09 01:08:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

[2]: rdd = sc.textFile("LaCelestina.txt")

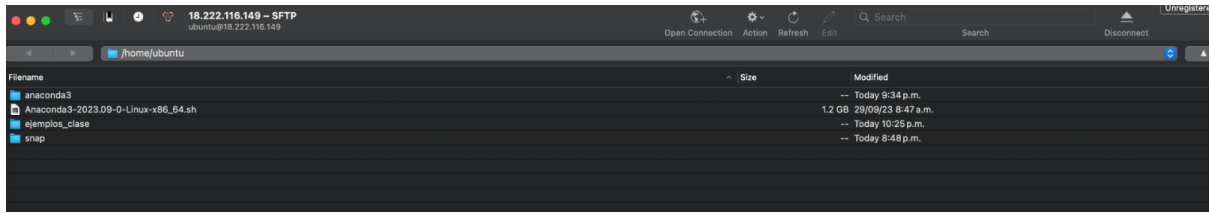
[3]: rdd.takeSample(False, 10)

[4]: palabras = rdd.map(lambda l: l.split(" "))

[5]: palabras.take(4)

[6]: [['**This',
      'is',
      'a',
      'COPYRIGHTED',
      'Project',
      'Gutenberg',
      'Etetx',
      'Details',
      'Below**'],
      ['The',
      'Project',
      'Gutenberg',
      'Etetx',
```

6.- Impresión de pantalla de la conexión abierta al servidor utilizando Cyberduck o Filezilla (ver listado de archivos).



7.- Crear un notebook con su nombre y colocar el llamado a Pyspark para visualizar la versión instalada.

