

Campus Monterrey

Materia

Inteligencia artificial avanzada para la ciencia de datos II

Módulo

Big Data

Tarea

1.- Instalación de Spark en AWS

Estudiante

Juan Pablo Bernal Lafarga - A01742342

Profesor Felipe Castillo Rendón

5 de noviembre del 2024

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

The screenshot shows the AWS Management Console with the 'Instances' section selected. There are two instances listed:

Name	Instance ID	Instance state	Instance type	Status check
Spark	i-0cf458b014bf6700f	Stopped	t2.micro	-
spark3	i-05d0390a641f29e04	Running	t2.micro	2/2 checks passed

Below the table, a detailed view for the 'spark3' instance is shown, listing its security group rules:

Name	Security group rule ID	Port range	Protocol
-	sgr-00c6ea1e6a054efa8	443	TCP
-	sgr-0dfbc1511e0f6d6c5	4040	TCP
-	sgr-015b473ecbd661708	80	TCP
-	sgr-05c5374aefcf41413	8888	TCP
-	sgr-007aa1f9fd8e2aafe	22	TCP

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) juanbernal@Juans-MacBook-Air Big_Data % ssh -i spark3.pem ubuntu@18.117.196.231
The authenticity of host '18.117.196.231' (18.117.196.231) can't be established.
ED25519 key fingerprint is SHA256:XNrjUdQJNesiW+OJ9H9iFYHs5qNNssGPsbG/WkHlxmQ.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '18.117.196.231' (ED25519) to the list of known hosts.
Welcome to Ubuntu 24.04 LTS (GNU/Linux 6.8.0-1017-aws x86_64)

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

System information as of Tue Nov  5 13:42:27 UTC 2024

  System load:  0.16           Processes:      104
  Usage of /:  70.4% of 28.02GB   Users logged in:    0
  Memory usage: 19%            IPv4 address for enx0: 172.31.17.156
  Swap usage:  0%

 * Ubuntu Pro delivers the most comprehensive open source security and
   compliance features.

  https://ubuntu.com/aws/pro

Expanded Security Maintenance for Applications is not enabled.

116 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

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

Last login: Tue Oct 15 14:32:31 2024 from 3.16.146.5
(base) ubuntu@ip-172-31-17-156:~$ ls -l
total 2062888
-rw-rw-r-- 1 ubuntu ubuntu 1064920017 Aug  4  2023 Anaconda3-2023.07-2-Linux-x86_64.sh
drwxrwxr-x 28 ubuntu ubuntu  4096 Oct  8 13:43 anaconda3
drwxrwxr-x  2 ubuntu ubuntu  4096 Sep 17 14:32 certs
drwxrwxr-x  3 ubuntu ubuntu  4096 Oct 15 14:14 notebooks
drwxr-xr-x 13 ubuntu ubuntu  4096 Sep  9  2023 spark-3.5.0-bin-hadoop3
-rw-rw-r--  1 ubuntu ubuntu 400395283 Sep  9  2023 spark-3.5.0-bin-hadoop3.tgz
-rw-rw-r--  1 ubuntu ubuntu 246639881 Sep 24 14:51 spark-3.5.0-bin-hadoop3.tgz.1
-rw-rw-r--  1 ubuntu ubuntu 400395283 Sep  9  2023 spark-3.5.0-bin-hadoop3.tgz.2
drwxr-xr-x  2 ubuntu ubuntu  4096 Sep 17 14:21 wget
-rw-rw-r--  1 ubuntu ubuntu  6288 Sep 24 14:51 wget-log
(base) ubuntu@ip-172-31-17-156:~$
```

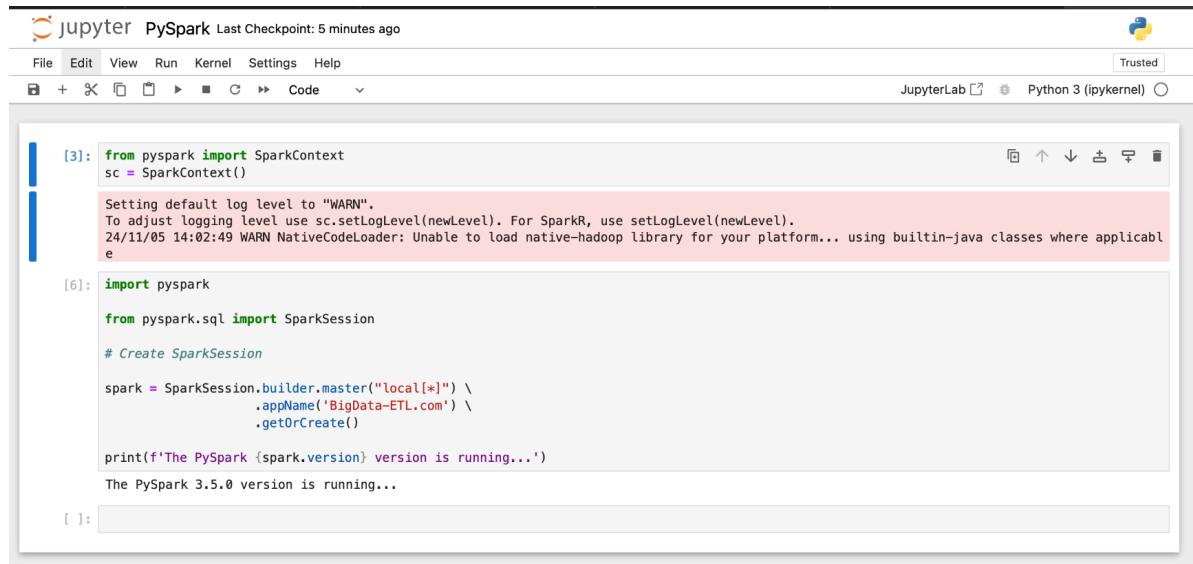
3. Impresión de pantalla de la pestaña Detalles para que se vea la ip pública, la ip privada y el DNS público de la instancia (es necesario que la instancia esté Running).

The screenshot shows the AWS CloudWatch Instances console. At the top, there's a header with 'Instances (1/2) Info' and various filters like 'Name', 'Instance ID', 'Instance state', 'Instance type', and 'Status check'. Below the header is a search bar and a dropdown for 'All states'. The main table lists two instances: 'Spark' (Stopped) and 'spark3' (Running). The 'spark3' row is selected. A detailed view for 'spark3' is shown below, including its Instance ID (i-05d0390a641f29e04), Public IPv4 address (18.117.196.231), Private IPv4 address (172.31.17.156), Public IPv4 DNS (ec2-18-117-196-231.us-east-2.compute.amazonaws.com), Hostname type (IP name: ip-172-31-17-156.us-east-2.compute.internal), and Private IP DNS name (ip-172-31-17-156.us-east-2.compute.internal).

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

The screenshot shows the Jupyter Notebook interface. The top navigation bar includes 'File', 'View', 'Settings', 'Help', and a logo. Below the navigation is a toolbar with buttons for 'Open', 'Download', 'Rename', 'Duplicate', 'Delete', 'New', 'Upload', and a refresh icon. The main area displays a list of notebooks and files. The 'Running' tab is selected. The list includes '1.- spark.ipynb' (selected), '10.- persistencia.ipynb', '11.- spark_sql.ipynb', '12.- spark_sql_fecha.ipynb', '13.- spark_sql_agrupaciones.ipynb', '14.- mllib.ipynb', '15.- regresion_lineal.ipynb', '2.- intro_spark.ipynb', '3.- expresiones_lambda.ipynb', '4.- acciones.ipynb', '5.- transformaciones.ipynb', '6.- ejemplo.ipynb', '7.- valores_nulos.ipynb', '8.- pair_rdd.ipynb', '9.- particionado.ipynb', 'PySpark.ipynb', 'AAPL.csv', 'customers.csv', 'ejemplo.txt', 'LaCelestina.txt', 'Null.csv', 'personas.json', 'sample_linear_regression_data.txt', and 'ventas.csv'. The table columns are 'Name', 'Last Modified', and 'File Size'.

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



The screenshot shows a Jupyter Notebook interface titled "jupyter PySpark". The top menu bar includes File, Edit, View, Run, Kernel, Settings, Help, and a Trusted button. The toolbar below has icons for file operations like New, Open, Save, and Run. On the right, it shows "JupyterLab" and "Python 3 (ipykernel)". The main area contains two code cells:

```
[3]: 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).  
24/11/05 14:02:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
  
[6]: import pyspark  
  
from pyspark.sql import SparkSession  
  
# Create SparkSession  
  
spark = SparkSession.builder.master("local[*]") \  
    .appName('BigData-ETL.com') \  
    .getOrCreate()  
  
print(f'The PySpark {spark.version} version is running...')  
The PySpark 3.5.0 version is running...
```

The output of the first cell is displayed in a pink box at the top of the cell area. The output of the second cell is shown as text at the bottom of the cell area.