



Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Instalación de Spark en AWS

TC3007C.501 Inteligencia Artificial Avanzada para la Ciencia de Datos II

Profesores:

Iván Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernández

Félix Ricardo Botello Urrutia

Edgar Covantes Osuna

Felipe Castillo Rendón

Hugo Terashima Marín

Integrantes:

Julian Lawrence Gil Soares – Aoo832272

10 de Octubre del 2023

1.

Instances (2) Info								
<input type="text" value="Find Instance by attribute or tag (case-sensitive)"/>								
<input type="checkbox"/>	Name <input type="text" value="hw2"/>	Instance ID <input type="text" value="i-0c528cbd855191dfd"/>	Instance state <input checked="" type="checkbox"/> Running	Instance type <input type="text" value="t2.micro"/>	Status check <input type="text" value="Initializing"/>	Alarm status <input type="text" value="No alarms"/>	Availability Zone <input type="text" value="us-east-2a"/>	Public IPv4 DNS <input type="text" value="ec2-3-143-237-2"/>
<input type="checkbox"/>	hw2	i-04dffa81685494e9a	<input type="checkbox"/> Terminated	t2.micro	-	No alarms	us-east-2a	-

2.

```
ubuntu@ip-172-31-3-58: ~
login as: ubuntu
Authenticating with public key "Spark"
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 00:02:11 UTC 2023

System load:  0.0908203125      Processes:           102
Usage of /:   20.4% of 7.57GB   Users logged in:     0
Memory usage: 22%              IPv4 address for eth0: 172.31.3.58
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
```

3.


Instance summary for i-0c528cbd855191dfd Info		
Updated less than a minute ago		
Instance ID <input type="text" value="i-0c528cbd855191dfd"/>	Public IPv4 address <input type="text" value="3.143.237.237"/> open address	Private IPv4 addresses <input type="text" value="172.31.3.58"/>
IPv6 address -	Instance state <input checked="" type="checkbox"/> Running	Public IPv4 DNS <input type="text" value="ec2-3-143-237-237.us-east-2.compute.amazonaws.com"/> open address
Hostname type IP name: ip-172-31-3-58.us-east-2.compute.internal	Private IP DNS name (IPv4 only) <input type="text" value="ip-172-31-3-58.us-east-2.compute.internal"/>	Elastic IP addresses -
Answer private resource DNS name IPv4 (A)	Instance type t2.micro	AWS Compute Optimizer finding Opt-in to AWS Compute Optimizer for recommendations. Learn more
Auto-assigned IP address <input type="text" value="3.143.237.237 [Public IP]"/>	VPC ID <input type="text" value="vpc-0cddb88660ade1836"/> open address	Auto Scaling Group name -
IAM Role -	Subnet ID <input type="text" value="subnet-07e7fc8b8ee45465"/> open address	
IMDSv2 Optional		
Details Security Networking Storage Status checks Monitoring Tags		

4.

```
ubuntu@ip-172-31-3-58:~$ sudo snap install jupyter
Fetch and check assertions for snap "jupyter" (6)
jupyter 1.0.0 from Jupyter Project (projectjupyter✓) installed
ubuntu@ip-172-31-3-58:~$ jupyter notebook
[I 00:22:28.286 NotebookApp] Writing notebook server cookie secret to /run/user/1000/snap.jupyter/jupyter/notebook_cookie_secret
[I 00:22:30.064 NotebookApp] Serving notebooks from local directory: /home/ubuntu
[I 00:22:30.064 NotebookApp] The Jupyter Notebook is running at:
[I 00:22:30.064 NotebookApp] http://localhost:8888/?token=52c7eadbe3dc7c3df1561618068a05d7e0ac73c132e8b728
[I 00:22:30.064 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[W 00:22:30.070 NotebookApp] No web browser found: could not locate runnable browser.
[C 00:22:30.070 NotebookApp]

To access the notebook, open this file in a browser:
    file:///run/user/1000/snap.jupyter/jupyter/nbserver-1521-open.html
Or copy and paste one of these URLs:
    http://localhost:8888/?token=52c7eadbe3dc7c3df1561618068a05d7e0ac73c132e8b728
```

5.

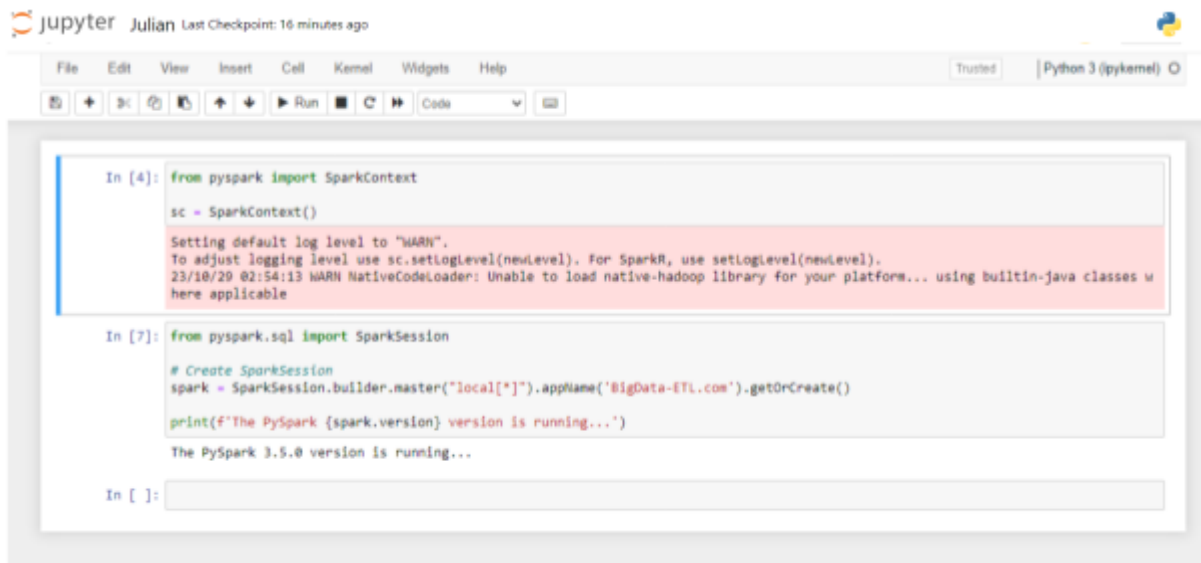

Quit Logout

Files Running Clusters

Duplicate Rename Move Download View Edit +
Upload New ↺

1	Name	Last Modified	File size
<input checked="" type="checkbox"/>	acciones.ipynb	hace unos segundos	16.9 kB
<input type="checkbox"/>	ejemplo.ipynb	hace unos segundos	44.6 kB
<input type="checkbox"/>	expresiones_lambda.ipynb	hace unos segundos	4.24 kB
<input type="checkbox"/>	intro_spark.ipynb	hace unos segundos	4.33 kB
<input type="checkbox"/>	mllib.ipynb	hace unos segundos	18.2 kB
<input type="checkbox"/>	pair_rdd.ipynb	hace unos segundos	26.7 kB
<input type="checkbox"/>	particionado.ipynb	hace unos segundos	8.94 kB
<input type="checkbox"/>	persistencia.ipynb	hace unos segundos	5.92 kB
<input type="checkbox"/>	regresion_lineal.ipynb	hace unos segundos	17.1 kB
<input type="checkbox"/>	spark.ipynb	hace unos segundos	1.82 kB
<input type="checkbox"/>	spark_sql.ipynb	hace unos segundos	24.8 kB
<input type="checkbox"/>	spark_sql_agrupaciones.ipynb	hace unos segundos	9.91 kB
<input type="checkbox"/>	spark_sql_fecha.ipynb	hace unos segundos	13.1 kB
<input type="checkbox"/>	transformaciones.ipynb	hace unos segundos	12.5 kB
<input type="checkbox"/>	valores_nulos.ipynb	hace unos segundos	10.1 kB
<input type="checkbox"/>	AAPL.csv	hace unos segundos	192 kB
<input type="checkbox"/>	Configuración Python con Spark.docx	hace unos segundos	224 kB
<input type="checkbox"/>	Configuración Python con Spark.pdf	hace unos segundos	308 kB
<input type="checkbox"/>	customers.csv	hace unos segundos	86.9 kB
<input type="checkbox"/>	ejemplo.txt	hace unos segundos	55 B
<input type="checkbox"/>	jupyter_notebook_config.py	hace 20 minutos	56.4 kB
<input type="checkbox"/>	LaCelestina.txt	hace unos segundos	688 kB
<input type="checkbox"/>	minitrans	hace una hora	17 B

7.



Jupyter Julian Last Checkpoint: 16 minutes ago

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (pykernel)

In [4]: `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).
23/10/20 @2:54:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

In [7]: `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...

In []: