ejemplo-emr-jupyterhub-python-paquetes

November 14, 2025

[ ]:
```
# ejemplo de notebook con librerias python preinstaladas en el bootstrap del␣
 ↪cluster EMR
# al momento de crear el cluster AWS EMR, en Bootstrap Action:
# Add -> Script location: s3://bucket-name/install-my-jupyter-libraries.sh
#
# contenido de install-my-jupyter-libraries.sh
# #!/bin/bash
# sudo python3 -m pip install boto3 nltk scipy scikit-learn pandas
```

[1]:
```
spark
```

```
Starting Spark application
```

```
<IPython.core.display.HTML object>
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',␣
 ↪layout=Layout(height='25px', width='50%'),…
```

```
SparkSession available as 'spark'.
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',␣
 ↪layout=Layout(height='25px', width='50%'),…
```

```
<pyspark.sql.session.SparkSession object at 0x7fca16effca0>
```

[2]:
```
sc
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',␣
 ↪layout=Layout(height='25px', width='50%'),…
```

```
<SparkContext master=yarn appName=livy-session-0>
```

[3]:
```
import pandas as pd
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',␣
 ↪layout=Layout(height='25px', width='50%'),…
```

[4]:
```
import matplotlib.pyplot as plt

fig, ax = plt.subplots()

fruits = ['apple', 'blueberry', 'cherry', 'orange']
```

```
counts = [40, 100, 30, 55]
bar_labels = ['red', 'blue', '_red', 'orange']
bar_colors = ['tab:red', 'tab:blue', 'tab:red', 'tab:orange']

ax.bar(fruits, counts, label=bar_labels, color=bar_colors)

ax.set_ylabel('fruit supply')
ax.set_title('Fruit supply by kind and color')
ax.legend(title='Fruit color')

plt.show()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',␣
 ↪layout=Layout(height='25px', width='50%'),…
```

[5]:
```
%pip list | grep scikit-learn
```

```
scikit-learn                 1.6.1
Note: you may need to restart the kernel to use updated packages.
```

[6]:
```
import pandas as pd

usuarios = {
    "Nombre": ["Juan", "María", "Carlos", "Ana", "Luis"],
    "Apellido": ["Gómez", "López", "Rodríguez", "Pérez", "Martínez"],
    "Email": ["juan@example.com", "maria@example.com", "carlos@example.com",␣
 ↪"ana@example.com", "luis@example.com"],
    "Telefono": ["123-123-4567", "456-987-6543", "789-567-8901",␣
 ↪"654-234-5678", "963-678-9012"],
    "Edad": [12, 27, 22, 30, 16]
}

usuarios_df = pd.DataFrame(usuarios)
mayores_de_edad = usuarios_df[usuarios_df["Edad"] >= 18]

print(mayores_de_edad)
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',␣
 ↪layout=Layout(height='25px', width='50%'),…
```

	Nombre	Apellido	Email	Telefono	Edad
1	Mar?a	L?pez	maria@example.com	456-987-6543	27
2	Carlos	Rodr?guez	carlos@example.com	789-567-8901	22
3	Ana	P?rez	ana@example.com	654-234-5678	30

[7]:
```
from sklearn import datasets
iris = datasets.load_iris()
digits = datasets.load_digits()
print(digits.data)
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:',␣
  ↪layout=Layout(height='25px', width='50%'),…

[[ 0.  0.  5. … 0.  0.  0.]
 [ 0.  0.  0. … 10.  0.  0.]
 [ 0.  0.  0. … 16.  9.  0.]
 …
 [ 0.  0.  1. …  6.  0.  0.]
 [ 0.  0.  2. … 12.  0.  0.]
 [ 0.  0. 10. … 12.  1.  0.]]
```

[ ]: