2.Create_csv_import-Copy1

March 4, 2025

[2]:
```
from pyspark.sql import SparkSession
import json
```

[9]:
```
# Criar uma sessão Spark
spark = SparkSession.builder.appName("CSVExample").getOrCreate()

# Criar um ficheiro CSV
data = [
    [1, "Alice", 25],
    [2, "Bob", 30],
    [3, "Carlos", 22],
    [4, "Diana", 28],
    [5, "Eva", 35],
    [6, "Fernando", 40]
]
```

[16]:
```
# Cria um CVS
csv_filename = "dados.csv"
with open(csv_filename, mode="w", newline="", encoding="utf-8") as file:
    writer = csv.writer(file)
    writer.writerow(["id", "nome", "idade"])
    writer.writerows(data)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[16], line 4
      2 csv_filename = "dados.csv"
      3 with open(csv_filename, mode="w", newline="", encoding="utf-8") as file
----> 4     writer = csv.writer(file)
      5     writer.writerow(["id", "nome", "idade"])
      6     writer.writerows(data)

NameError: name 'csv' is not defined
```

[15]:
```
# Carregar o CSV com PySpark
df = spark.read.option("header", "true").csv(csv_filename, inferSchema=True)
```

[ ]:
```
# Mostrar os dados
df.show()
```

[ ]:
```
# Mostrar o top 5
df.limit(5).show()
```

[23]:
```
# Mostrar o top 5
df.sort("id").show(5)
```

```
+---+-----+------+
| id|idade|  nome|
+---+-----+------+
|  1|   25| Alice|
|  2|   30|   Bob|
|  3|   22|Carlos|
|  4|   28| Diana|
|  5|   35|   Eva|
+---+-----+------+
only showing top 5 rows
```

[ ]:

[ ]: