google-colab-setup-PySpark

November 14, 2025

1 Data Processing using Pyspark

[ ]:

```
#configuración en google colab de spark y pyspark
from google.colab import drive
drive.mount('/content/gdrive')
```

[ ]:

```
#instalar java y spark
!apt-get install openjdk-17-jdk-headless -qq > /dev/null
!wget -q https://downloads.apache.org/spark/spark-4.0.1/spark-4.0.1-bin-hadoop3.
 ↪tgz
!tar xf spark-4.0.1-bin-hadoop3.tgz
!pip install -q findspark
```

[ ]:

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-17-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-4.0.1-bin-hadoop3"
```

[ ]:

```
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
sc = spark.sparkContext
```

[ ]:

```
# Load csv Dataset
df=spark.read.csv('sample_data/california_housing_test.
 ↪csv',inferSchema=True,header=True)
# desde S3
# df=spark.read.csv('s3a://bucke_name/datasets/sample_data.
 ↪csv',inferSchema=True,header=True)
```

[ ]:

```
#columns of dataframe
df
```

[ ]:

```
#check number of columns
len(df.columns)
```

[ ]:

```
#number of records in dataframe
df.columns
```

[ ]:

```
#shape of dataset
print((df.count(),len(df.columns)))
```

[ ]:

```
#printSchema
df.printSchema()
```

[ ]:

```
#fisrt few rows of dataframe
df.show(5)
```