google-colab-setup-PySpark2AWS-S3

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-3.5.7/spark-3.5.7-bin-hadoop3.
 ↪tgz
!tar xf spark-3.5.7-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-3.5.7-bin-hadoop3"
```

[ ]:
```
import findspark
findspark.init()
```

[ ]:
```
!mkdir -p /content/jars
!wget -q https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.4/
 ↪hadoop-aws-3.3.4.jar -P /content/jars
!wget -q https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.12.
 ↪367/aws-java-sdk-bundle-1.12.367.jar -P /content/jars
```

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

jars = "/content/jars/hadoop-aws-3.3.4.jar,/content/jars/aws-java-sdk-bundle-1.
 ↪12.367.jar"

spark = SparkSession.builder \
    .appName("S3Connection") \
    .master("local[*]") \
    .config("spark.jars", jars) \
    .config('fs.s3a.access.key', "AWS_ACCESS_KEY") \
    .config('fs.s3a.secret.key', "AWS_SECRET_KEY") \
```

```
    .config('fs.s3a.session.token',"AWS_SESSION_TOKEN") \
    .config("spark.hadoop.fs.s3a.impl", "org.apache.hadoop.fs.s3a.
 ↪S3AFileSystem") \
    .config("spark.hadoop.fs.s3a.path.style.access", "true") \
    .config("spark.hadoop.fs.s3a.endpoint", "s3.amazonaws.com") \
    .getOrCreate()

sc = spark.sparkContext
```

[ ]:
```
# Load csv Dataset
# 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)
```