PREGUNTAS Y RESPUESTAS

August 7, 2024

1. Explique qué es información en un formato delta

   El formato Delta es un formato de almacenamiento basado en Apache Parquet que permite transacciones ACID (Atomicity, Consistency, Isolation, Durability) en Apache Spark. Esto significa que puedes realizar operaciones de escritura y lectura concurrentes sin correr el riesgo de corrupción de datos. Delta Lake, que utiliza el formato Delta, proporciona versiones de datos, lo que permite hacer consultas a versiones anteriores y facilita la gestión de datos a gran escala con alta confiabilidad.

2. Construya una sentencia de código que almacene un DataFrame en una montura de Databricks en formato delta

[ ]:
```
# crear una montura en Databricks que apunte a un bucket de S3
dbutils.fs.mount(
  source = "s3a://<bucket-name>",
  mount_point = "/mnt/<mount-name>",
  extra_configs = {"fs.s3a.access.key": "<ACCESS_KEY>", "fs.s3a.secret.key":␣
 ↪"<SECRET_KEY>"}
)
```

[ ]:
```
test_df.write.format("delta").save("/mnt/<mount-name>/test-delta")
```

3. Construya mediante la sentencia withColumn la variable ALTURA_CUADRADO que contenga la variable ALTURA elevada al cuadrado

[ ]:
```
from pyspark.sql.functions import col # importando la función col

test_df = test_df.withColumn("ALTURA_CUADRADO", col("ALTURA")**2)
```

4. Construya mediante la sentencia withColumn una variable que se llame LLAVE_PAIS, que contenga la concatenación de la columna PAIS y NUMERO_ID en una misma columna

[ ]:
```
from pyspark.sql.functions import concat, col # importando las funciones concat␣
 ↪y col

test_df = test_df.withColumn("LLAVE_PAIS", concat(col("PAIS"),␣
 ↪col("NUMERO_ID")))
```

5. Bonus: Escriba mediante el método write la información en una tabla delta ubicada en un delta lake, sobreescribiendo la tabla y el esquema de los datos

[ ]:
```
test_df.write.format("delta").mode("overwrite").option("overwriteSchema",␣
↪"true").save("/mnt/<mount-name>/delta-table")
```