🔰Delta Lake (Bronze → Silver → Gold)

Author: Sandeep Kumar Sharma

This is the simplest possible hands-on lab to understand the **Delta Lake workflow**:
Bronze → Silver → Gold

We take a small dataset, write it to Bronze, clean it in Silver, and then aggregate it in Gold.

✅Prerequisites

- Databricks Notebook
- Spark session (`spark`) available

🔴Step 1: Create Sample Raw Data (Bronze Input)

```
# Creating a simple in-memory dataset
raw_data = [
    (1, "Alice", 50),
    (2, "Bob", 70),
    (3, "Charlie", None)  # Contains missing value
]

columns = ["id", "name", "score"]

df_raw = spark.createDataFrame(raw_data, columns)
```

🔴Step 2: Write Data to Bronze Layer

```
bronze_path = "/mnt/delta/bronze/simple_demo"

df_raw.write.format("delta").mode("overwrite").save(bronze_path)
```

⬛Step 3: Read Bronze → Clean Data → Write to Silver Layer

We remove rows where **score is NULL**.

```
silver_path = "/mnt/delta/silver/simple_demo"

# Read Bronze
df_bronze = spark.read.format("delta").load(bronze_path)

# Simple cleaning: remove null score rows
df_silver = df_bronze.filter(df_bronze.score.isNotNull())

# Write to Silver
df_silver.write.format("delta").mode("overwrite").save(silver_path)
```

🔲Step 4: Read Silver → Aggregate → Write to Gold Layer

Very simple aggregation: **Average score**.

```
gold_path = "/mnt/delta/gold/simple_demo"
from pyspark.sql import functions as F

# Read Silver
df_silver = spark.read.format("delta").load(silver_path)

# Aggregation: Average score
df_gold = df_silver.agg(F.avg("score").alias("avg_score"))

# Write to Gold
df_gold.write.format("delta").mode("overwrite").save(gold_path)
```

🛞 Step 5: Verify Output

```
print("BRONZE Data:")
display(spark.read.format("delta").load(bronze_path))

print("SILVER Data:")
display(spark.read.format("delta").load(silver_path))

print("GOLD Data:")
display(spark.read.format("delta").load(gold_path))
```

‼️ Lab Completed!

You built a complete **Delta Lake Medallion Architecture** pipeline:

- **Bronze:** Raw data
- **Silver:** Cleaned data
- **Gold:** Aggregated business-level data