Lab 1 — Basic Transformation

Cleaning Nulls & Exporting Clean CSV

Author: Sandeep Kumar Sharma

🎯Learning Objective

This lab teaches beginners how to:

- Load a CSV file in Databricks using Python (PySpark)
- Identify and remove null/blank rows
- Save the cleaned results back as a CSV file

This version is intentionally kept **simple**, **clean**, and **beginner-friendly**.

What You Will Learn (Learning Outcomes)

By completing this lab, you will be able to:

- Read data from a CSV file
- View schema and sample records
- Replace blank values with NULL
- Drop rows containing NULL values
- Save cleaned records as a CSV

Dataset Used

Upload the dataset into the following path in Databricks:

```
/Volumes/workspace/default/test/customer_raw.csv
```

Make sure your file name is exactly: **customer_raw.csv**.

Step-by-Step Hands-On Lab

Follow these steps one by one inside your Databricks notebook.

Step 1 — Load the CSV File

```
# Path to input CSV file
csv_path = "/Volumes/workspace/default/test/customer_raw.csv"

# Load CSV file with header
df_raw = spark.read.option("header", True).csv(csv_path)

# Show first 20 records
display(df_raw)
```

Step 2 — Check Schema and Sample Rows

```
# Print schema of the dataframe
df_raw.printSchema()

# Show sample data
df_raw.show(10, truncate=False)
```

This helps learners understand what kind of data the CSV contains.

Step 3 — Replace Blank Values with NULL

Before removing any rows, we convert blank strings ("") to NULL.

```
from pyspark.sql import functions as F

df_clean = df_raw.replace("", None)
```

Step 4 — Drop Rows Containing ANY Null Values

```
df_clean = df_clean.dropna()
```

This step keeps only those rows where **all columns have valid data**.

Step 5 — Preview Cleaned Data

```
# Display cleaned output
df_clean.show(20, truncate=False)

# Show final record count
print("Final Count After Cleaning:", df_clean.count())
```

Step 6 — Save the Cleaned Data as a CSV

```
output_path = "/Volumes/workspace/default/test/customer_cleaned_csv"

df_clean.write.mode("overwrite").option("header", True).csv(output_path)
```

This will generate a clean CSV output in the provided folder.

🎉Lab Completed Successfully!

In this simple transformation lab, you learned how to:

- Read a CSV file
- Inspect and understand its structure
- Clean basic null/blank data issues
- Save your improved dataset as a new CSV

This forms the **foundation of all upcoming transformation labs**.
