**Dataset: Customers**

**customers.csv**

id,name,address

1,Alice,Address A

2,Bob,Address B

3,Charlie,Address X

4,David,Address D

5,Eva,Address E

**customers\_updates.csv (for SCD Type 2 / MERGE Lab)**

id,name,address

1,Alice,New Address A

3,Charlie,Address C

6,Frank,Address F

**Steps to Prepare the Dataset in Your Environment**

1. **Upload CSVs to DBFS (Databricks) or ADLS (Azure):**

dbutils.fs.mkdirs("/mnt/data/parquet/customers")

dbutils.fs.put("/mnt/data/parquet/customers/customers.csv", """

id,name,address

1,Alice,Address A

2,Bob,Address B

3,Charlie,Address X

4,David,Address D

5,Eva,Address E

""", overwrite=True)

dbutils.fs.put("/mnt/data/parquet/customers/customers\_updates.csv", """

id,name,address

1,Alice,New Address A

3,Charlie,Address C

6,Frank,Address F

""", overwrite=True)

1. **Convert CSV to Parquet (for Lab 1 start point):**

customers\_df = spark.read.option("header", "true").csv("/mnt/data/parquet/customers/customers.csv")

customers\_df.write.mode("overwrite").parquet("/mnt/data/parquet/customers")

1. **Prepare Update DataFrame (for MERGE lab):**

updates\_df = spark.read.option("header", "true").csv("/mnt/data/parquet/customers/customers\_updates.csv")

updates\_df.createOrReplaceTempView("updates")

**Where Each Dataset Fits**

* **Lab 1 (Convert Parquet to Delta):** Use customers.csv → convert to Parquet → then Delta.
* **Lab 2 (MERGE / SCD Type 2):** Use customers\_updates.csv as the source updates.
* **Lab 3 (Schema Evolution):** Start with minimal dataset (id, name) → evolve to add country.
* **Lab 4 (Time Travel):** Any table after updates → run time travel queries.
* **Lab 5 (Optimize & Vacuum):** Run on the Delta tables you’ve created.