**Dataset Pack for Delta Lake Labs**

**1. Customers Dataset (for Lab 1, 2, 4)**

customers.csv

id,name,city

1,Alice,New York

2,Bob,Los Angeles

3,Charlie,Chicago

4,Diana,Houston

5,Eva,Miami

customers\_updates.csv (for Lab 4 Upserts)

id,name,city

1,Alice,Boston

3,Charlie,Seattle

6,Frank,San Francisco

**2. Orders Dataset (for Lab 3: Parquet → Delta)**

orders.csv

order\_id,customer\_id,amount,order\_date

101,1,250.00,2025-09-01

102,2,180.50,2025-09-02

103,3,99.99,2025-09-02

104,1,300.00,2025-09-03

105,4,450.00,2025-09-04

**3. Sales Dataset (for Lab 6: Partitioning, Z-Ordering, Caching)**

sales.csv

sale\_id,customer\_id,region,amount

1001,1,East,200.0

1002,2,West,150.0

1003,3,East,300.0

1004,4,South,500.0

1005,5,North,120.0

1006,1,East,220.0

1007,2,West,180.0

1008,3,East,330.0

**4. Streaming Input Dataset (for Lab 7)**

Folder: /mnt/stream/input/ → each file simulates new streaming events.  
orders\_stream\_1.json

{"order\_id":201,"customer\_id":1,"amount":75.5,"order\_date":"2025-09-05"}

{"order\_id":202,"customer\_id":3,"amount":125.0,"order\_date":"2025-09-05"}

orders\_stream\_2.json

{"order\_id":203,"customer\_id":2,"amount":200.0,"order\_date":"2025-09-06"}

{"order\_id":204,"customer\_id":5,"amount":80.0,"order\_date":"2025-09-06"}

**How to Load These Datasets**

1. **Upload CSVs into DBFS/ADLS path** /mnt/data/parquet/

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

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

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

1. **Convert to Parquet (for baseline labs)**

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

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

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

1. **Streaming schema (Lab 7)**

from pyspark.sql.types import StructType, StructField, IntegerType, DoubleType, StringType

schema = StructType([

StructField("order\_id", IntegerType(), True),

StructField("customer\_id", IntegerType(), True),

StructField("amount", DoubleType(), True),

StructField("order\_date", StringType(), True)

])

**Mapping Datasets → Labs**

* **Lab 1 & 2:** customers.csv → Parquet → Delta.
* **Lab 3:** orders.csv → Parquet → Delta table.
* **Lab 4:** customers.csv (base) + customers\_updates.csv (for MERGE).
* **Lab 5:** Minimal DataFrame (id, name) created in code.
* **Lab 6:** sales.csv → for partitioning, Z-Ordering, caching.
* **Lab 7:** orders\_stream\_\*.json → simulate streaming ingestion.
* **Lab 8:** Works on any Delta table created above (orders\_delta / customers\_delta).