
Scenario: E-Commerce Transactions + Returns + Inventory

We'll create **3 datasets** to simulate a more realistic analytics pipeline:

`orders.csv`

```
OrderID,CustomerID,ProductID,Quantity,Price,OrderDate,Status
3001,C001,P1001,1,75000,2024-05-01,Delivered
3002,C002,P1002,2,50000,2024-05-02,Returned
3003,C003,P1003,1,30000,2024-05-03,Delivered
3004,C001,P1002,1,50000,2024-05-04,Delivered
3005,C004,P1004,3,10000,2024-05-05,Pending
```

`customers.csv`

```
CustomerID,CustomerName,Region,SignupDate
C001,Amit,North,2023-11-12
C002,Sara,South,2024-01-08
C003,John,West,2023-06-20
C004,Priya,East,2024-03-15
```

`products.csv`

```
ProductID,ProductName,Category,Stock,ReorderLevel
P1001,Laptop,Electronics,5,2
P1002,Phone,Electronics,10,3
P1003,Tablet,Electronics,7,2
P1004,Keyboard,Accessories,15,5
```

New Set of Tasks

PySpark + Delta

1. Ingest all 3 CSVs as Delta Tables.
2. Write SQL to get the total revenue per Product.
3. Join Orders + Customers to find revenue by Region.
4. Update the Status of Pending orders to 'Cancelled'.
5. Merge a new return record into Orders.

DLT Pipeline

6. Create raw → cleaned → aggregated tables:
   - Clean: Remove rows with NULLs
   - Aggregated: Total revenue per Category

Time Travel

7. View data before the Status update.
8. Restore to an older version of the orders table.

Vacuum + Retention

9. Run `VACUUM` after changing default retention.

Expectations

10. `Quantity > 0`, `Price > 0`, `OrderDate is not null`

Bonus

11. Use `when-otherwise` to create a new column: `OrderType = "Return" if Status == 'Returned'`
