
Exercise Title: Customer Orders Analysis

Objective:

Participants will:

1. **Prepare two datasets** manually
2. **Ingest** the data using PySpark
3. **Clean**, **join**, **transform**, and **analyze** the data

---

Part 1: Data Preparation (manually create and save as CSV)

1. `customers.csv`

CustomerID	Name	City	Age
101	Aditi	Mumbai	28
102	Rohan	Delhi	35
103	Meena	Bangalore	41
104	Kabir	Hyderabad	30
105	Zoya	Chennai	25

2. `orders.csv`

OrderID	CustomerID	Product	Quantity	Price	OrderDate
1001	101	Laptop	1	70000	2024-01-05
1002	102	Mobile	2	25000	2024-02-10
1003	103	Desk	1	10000	2024-03-15
1004	101	Mouse	3	1000	2024-04-01
1005	104	Monitor	1	12000	2024-04-25

Ask participants to:

- Save these two datasets as **CSV files** (e.g., in `dbfs:/FileStore/` or local drive)

Part 2: Spark Tasks (

1. **Ingest the CSV files** into two PySpark DataFrames
2. **Infer schema** and print the schema for both
3. Add a column `TotalAmount = Quantity * Price` to `orders`
4. Join both DataFrames on `CustomerID`

5. Filter orders where `TotalAmount > 20000`
6. Show customers who placed more than 1 order
7. Group orders by `City` and get average order value
8. Sort orders by `OrderDate` in descending order
9. Write the final result as a **Parquet** file partitioned by `City`
10. Create a **temporary view** and run Spark SQL:
    - Total sales by customer
    - Count of products per city
    - Top 2 cities by revenue
