This dataset contains all transactions occurring for a UK-based online retail store between December 2009 and December 2011. It includes fields such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. ​[Kaggle+1Kaggle+1](https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset?utm_source=chatgpt.com)

**Project Problem Statement: Customer Purchasing Behavior Analysis Using Spark**

**Objective:** Analyze customer purchasing behavior to identify trends, popular products, and customer segments, ultimately aiding in strategic decision-making for marketing and inventory management.​

**Problem Statement:** An online retail company seeks to gain insights into its customer purchasing patterns over a two-year period. They have provided a dataset containing detailed transaction records, including:​

* **InvoiceNo:** Unique identifier for each transaction​
* **StockCode:** Product identifier​
* **Description:** Product description​
* **Quantity:** Number of units purchased​
* **InvoiceDate:** Date and time of the transaction​
* **UnitPrice:** Price per unit of the product
* **CustomerID:** Unique identifier for each customer​
* **Country:** Country where the customer resides​

Your tasks are to:

1. **Load the dataset** into an RDD from the CSV file.
2. **Apply transformations** such as:
   * Filtering out transactions with negative or zero quantities.
   * Mapping data to key-value pairs (e.g., CustomerID → Total Spend).
   * Aggregating data to compute metrics like total spending per customer.
3. **Convert the RDD to a DataFrame** with a well-defined schema.
4. **Perform analyses**, including:
   * Identifying the top 10 customers by total spending.
   * Determining the most frequently purchased products.
   * Analyzing purchasing trends over time (e.g., monthly sales trends).
   * Segmenting customers based on their purchasing behavior.
5. **Store the final processed data** in Parquet format for optimized storage and future analysis.

**Expected Deliverables:**

* A Spark application implementing the above workflow.
* Insights and visualizations derived from the data analysis.
* The processed dataset stored in Parquet format.