Retail Sales and Customer Behavior Analytics

Sushila Sipai

Satyanarayana Reddy Muttana

Yanjie Liu

Department of Computer Science, University of the Cumberlands

MSDS-632-M51: Big Data

June 15, 2025

Introduction

The retail industry generates large volumes of transaction data that, when analyzed effectively, can provide deep insights into customer behavior, product preferences, and seasonal purchasing trends. This research project focuses on analyzing retail sales using the Online Retail dataset from the UCI Machine Learning Repository (Chen, 2012). The dataset includes over 500,000 transactions made between December 2010 and December 2011 by a UK-based online retailer. Using this dataset, we aim to understand buying behavior, perform customer segmentation, and identify seasonal patterns. Apache Spark and its MLlib library will be leveraged for distributed data processing and machine learning.

Big Data Characteristics

The Online Retail dataset exhibits several of the defining characteristics of big data, often referred to as the "5 Vs": volume, velocity, variety, veracity, and value. These characteristics are critical in determining how data should be stored, processed, and analyzed in a big data environment.

Volume refers to the sheer size of data being generated. Although the Online Retail dataset contains approximately 500,000 records, it simulates the volume that a mid-sized e-commerce platform might handle on a daily or weekly basis. In real-world applications, this volume would scale dramatically, especially with the inclusion of multi-year data, user clickstreams, or product browsing history, making distributed data processing systems such as Apache Spark essential for analysis (EMC Education Services, 2015).

Velocity pertains to the speed at which data is generated and processed. While this dataset is historical, it simulates transactions that could occur in real time on a modern e-commerce site. In practice, new sales data is created continuously, requiring near real-time processing to update sales dashboards, generate personalized offers, or detect fraud. Big data platforms must support high-velocity data ingestion and analysis pipelines to keep pace with such dynamic environments (Hashem et al., 2015).

Variety highlights the different types of data sources and formats. The Online Retail dataset includes structured data fields such as invoice numbers, product codes, and timestamps. Though limited in scope, this variety reflects typical transactional databases. In larger implementations, unstructured data like customer reviews, product images, and social media interactions would also be integrated, adding complexity to data integration and processing workflows.

Veracity addresses the quality and trustworthiness of data. The Online Retail dataset presents challenges typical in real-world data, such as missing CustomerID values, inconsistent product descriptions, and negative quantities due to product returns. These issues necessitate thorough data cleaning and preprocessing to ensure that subsequent analyses are accurate and meaningful (Russom, 2011). Poor data veracity can undermine model performance and lead to faulty business conclusions.

Finally, Value represents the actionable insights that can be extracted from data. By applying clustering techniques to segment customers or analyzing time-based purchase patterns, the Online Retail dataset demonstrates how meaningful value can be derived from properly prepared and analyzed transactional data. When analyzed using the big data lifecycle, these insights empower retail companies to make data-driven decisions that enhance customer retention, inventory management, and marketing strategies.

Business Motivations and Drivers

The motivations for adopting big data in the retail industry include:

a) Personalized Marketing: Target customers with personalized promotions based on historical purchasing behavior.

b) Sales Optimization: Identify top-selling products and seasonal demand trends, thereby enabling smarter stock management.

c) Customer Segmentation: Group customers based on purchasing behavior, geolocation, and revenue contribution.

d) Inventory Management: Forecast product demand to optimize stock levels, reduce excess, and prevent stockouts.

e) Revenue Analysis: Track revenue by country and customer segments to focus on profitable regions. This is critical for retailers with a global footprint.

Retailers face constant pressure to differentiate themselves in a competitive landscape. By harnessing big data, companies can make informed, strategic decisions that contribute to long-term growth.

Big Data Adoption and Planning

Our team established a carefully planned technical infrastructure to enable effective big data analytics while working within project constraints. The implementation combined robust data processing capabilities with flexible analytical tools through a hybrid architecture.

Technical Environment Configuration

We deployed a local big data development environment centered around Apache Spark as our core processing engine. This required:

a) Foundation Layer

  - Java 11 JDK was installed to support Spark's JVM-based execution.
  - Hadoop winutils binaries configured for Windows filesystem compatibility

  - Proper environment variables were set up (JAVA_HOME, SPARK_HOME, HADOOP_HOME)

b) Processing Layer

  - Apache Spark 3.5.6 configured in local mode
  - PySpark API for Python integration
  - Spark SQL for structured data operations

c) Analytical Extension Layer

  - Python 3.13.4 with key data science packages (pandas, matplotlib)
  - Google Colab cloud notebooks were used for collaborative analysis
  - Seamless data exchange between Spark and Python environments

The planning phase focused on thoroughly understanding the data schema and identifying key entities such as products, customers, and transactions. Spark code was developed to clean the data and create reusable views, enabling streamlined access and analysis. The team adopted a hybrid approach, seamlessly switching between Spark for scalable data processing and Python for effective visualization and interpretation. This combination allowed us to harness the strengths of both platforms, ensuring efficient analysis while maintaining clarity and depth in the insights derived.

Big Data Analytics Lifecycle

The project adhered to the standard Big Data Analytics Lifecycle, ensuring a structured and goal-driven approach:

a) Business Case Evaluation: The primary objective was to analyze sales performance and understand customer behavior to inform strategic decisions.

b) Data Identification: The Online Retail CSV dataset was selected, containing detailed invoice records, product information, and customer metadata.

c) Data Acquisition and Filtering: Data quality was ensured by removing null values and irrelevant entries, such as rows with missing product descriptions or customer IDs.

d) Data Extraction and Transformation: The dataset was cleaned, trimmed, and enriched with derived metrics such as revenue per transaction to enhance its analytical value.

e) Data Analysis:

   - Top 10 Selling Products: Identified products driving the highest sales volumes.
   - Revenue by Country: Aggregated revenue figures by country to assess geographic performance.
   - Sales Trend by Month: Analyzed monthly trends to support seasonal planning and demand forecasting.
   - Top Customers by Revenue: Highlighted the most valuable customers for targeted marketing and engagement.

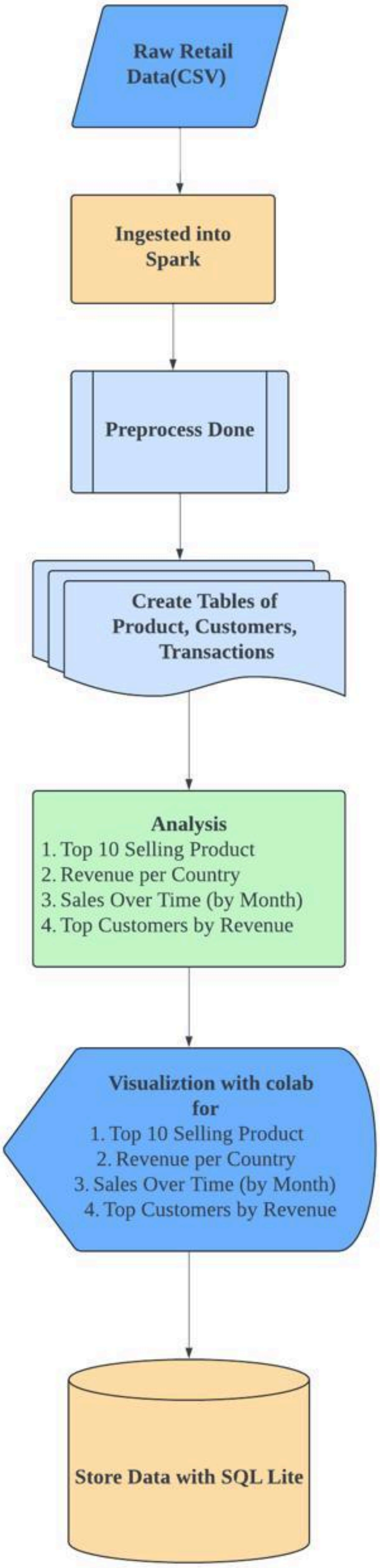
f) Data Visualization:

   - Daily Quantity Sold (Line Chart): Illustrates trends in transaction volume over time.
   - Top 10 Products (Bar Chart): Compared the performance of the best-selling products.
   - Country Revenue (Bar Chart): Shows revenue contribution by country.
   - Revenue Distribution (Histogram): Revealed skewness in customer spending behavior.

g) Interpretation of Results: Enabled identification of key sales drivers, high-performing regions, and top customer segments.

h) Deployment: The final insights were stored in an SQLite database for efficient reporting and querying.

Figure 1: Architecture Diagram for Retail Sales and Customer Behavior Analytics Project


Raw Retail Data(CSV)
Ingested into Spark
Preprocess Done
Create Tables of Product, Customers, Transactions
Analysis
1. Top 10 Selling Product
2. Revenue per Country
3. Sales Over Time (by Month)
4. Top Customers by Revenue
Visualiztion with colab for
1. Top 10 Selling Product
2. Revenue per Country
3. Sales Over Time (by Month)
4. Top Customers by Revenue
Store Data with SQL Lite

This lifecycle ensured that each phase contributed meaningfully to actionable insights and a scalable, data-driven solution.

Results Observed

Figure 2: Line Chart to represent the Total Quantity sold over time

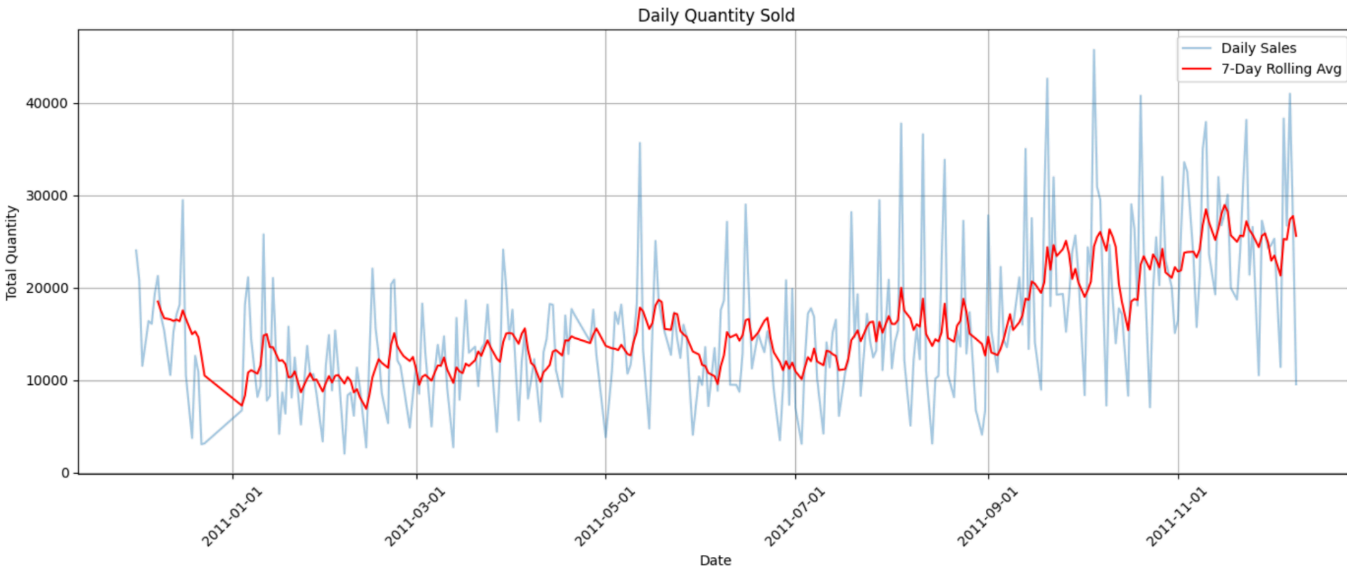


Figure 3: Bar Chart Illustrating Top 10 Best-Selling Products

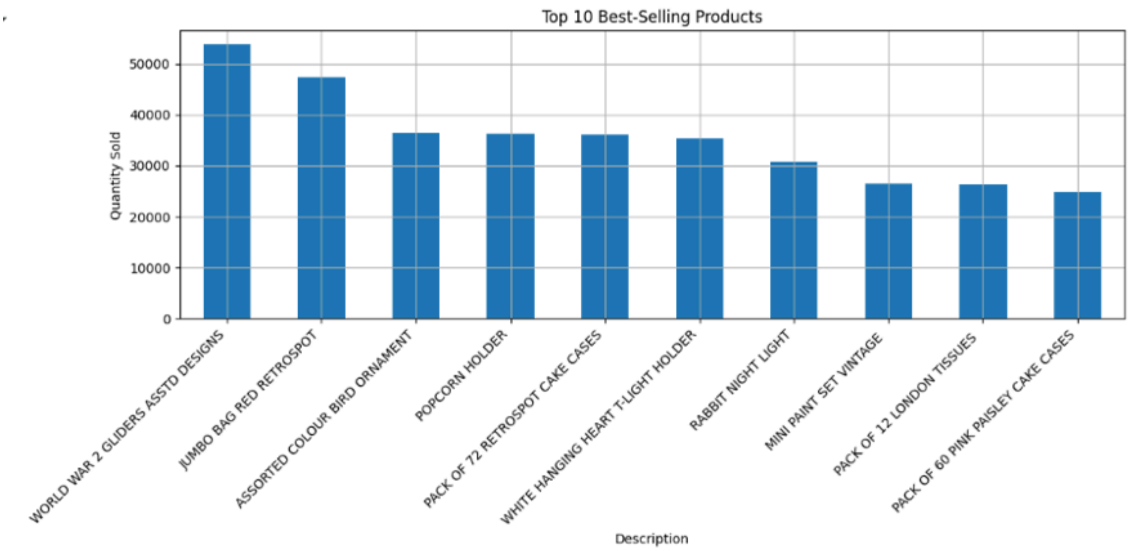


Figure 4: Bar Chart representing the Top 10 Countries by Revenue

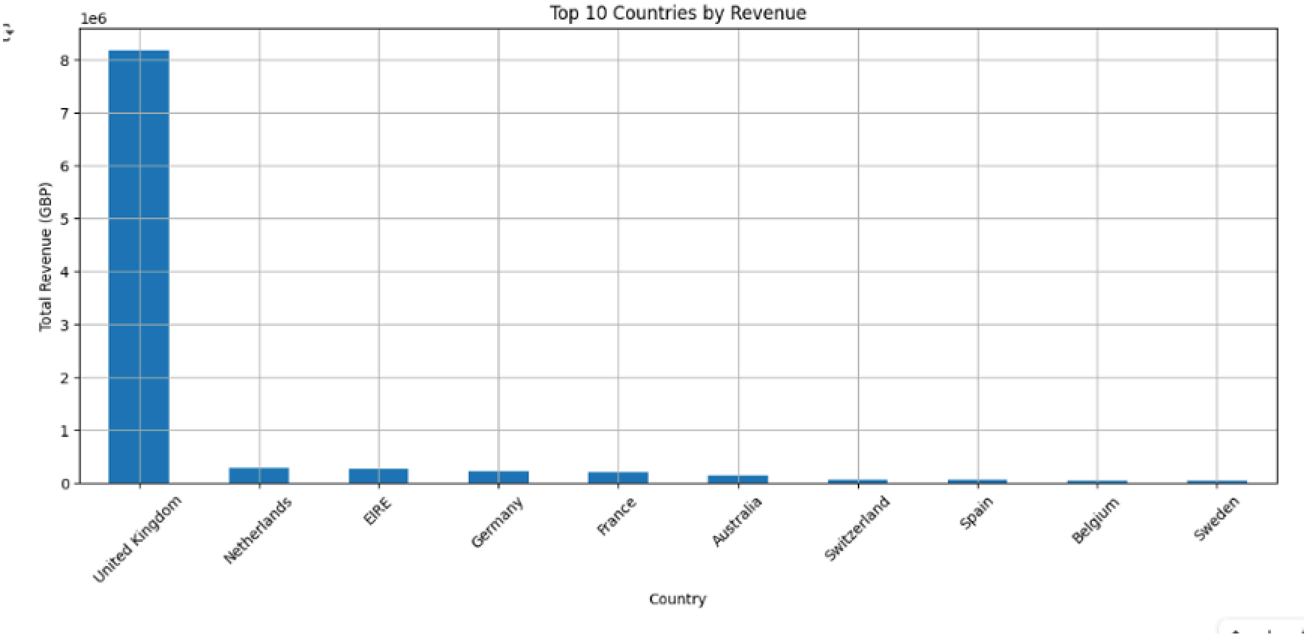
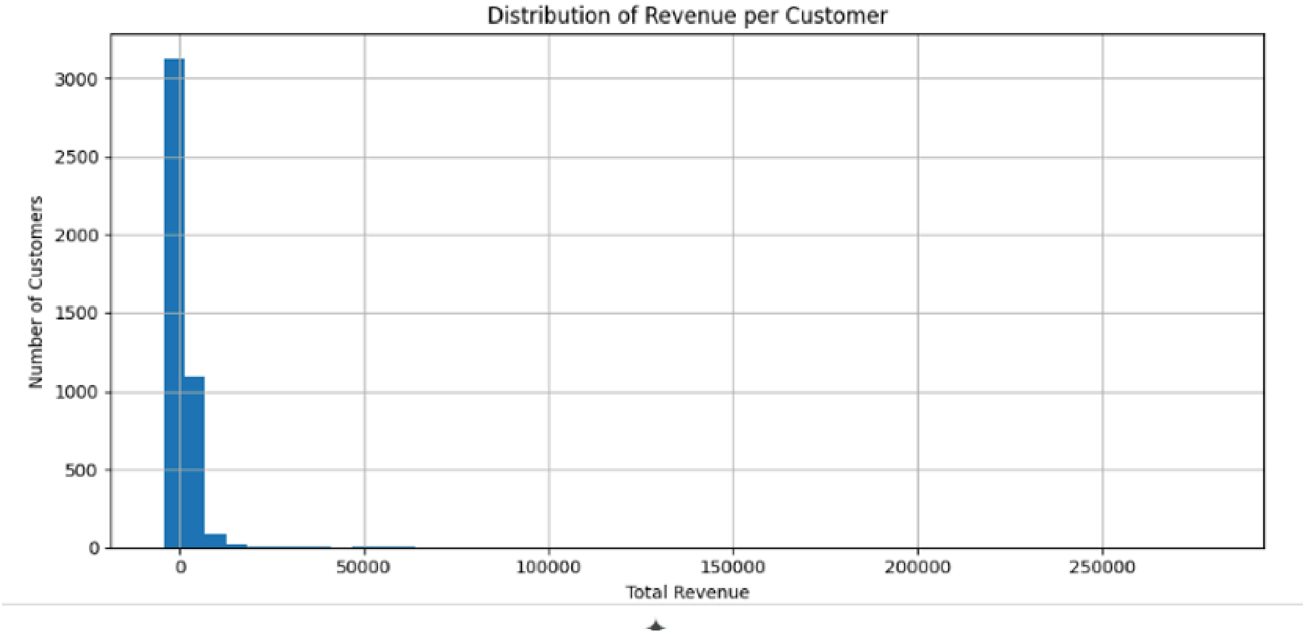


Figure 5: Histogram Representing Distribution of Revenue Per Customer



Available Technologies in Big Data

This project used a suite of technologies across the data processing pipeline:

a) Apache Spark (Scala API): Core engine for data transformation, SQL querying, and performance at scale.

b) Pandas: Used for filtering, grouping, and exploratory data analysis.

c) SQLite: Provided lightweight persistent storage for logical tables.

d) Matplotlib: The Charting library was used to present findings visually.

e) Google Colab: Enabled fast prototyping, cloud execution, and sharing of results.

These technologies are widely used in industry, making them relevant for both academic and enterprise-grade big data projects.

Challenges Faced in the Project

Every big data project brings inherent challenges:

a) Data Volume: Even a moderately sized retail dataset can contain hundreds of thousands of rows.

b) Data Quality: Missing values and irrelevant entries must be filtered for meaningful results.

c) Infrastructure Setup: Configuring Spark, Java, and winutils on a Windows system took significant time.

d) Tool Interoperability: Spark outputs had to be exported and re-ingested into Python/Pandas.

e) Visualization Accuracy: Ensuring that charts reflect the correct time intervals and quantities was vital.

f) Learning Curve: Team members had to work across Scala, SQL, and Python.

Addressing these challenges required iterative development, frequent debugging, and a thorough understanding of the technologies involved. The phase-specific challenges that we faced are as below:

Table 1

Phase-Specific Challenges Faced During the Project

Phases	Challenges
Data Acquisition	Handling missing or malformed records in the CSV dataset
Data Preprocessing	String trimming, data type conversions, and filtering irrelevant entries
Data Modeling	Creating consistent views (products, customers, transactions)
Data Analysis	SQL optimization for joins, aggregations, and reducing compute time
Data Visualization	Formatting labels, choosing appropriate chart types, and color coding
Interpretation & Delivery	Explaining findings to a non-technical audience

Each of these phases required careful planning and precise execution to prevent cascading issues.

Technology-Specific Issues

a) Spark: Required manual installation of Java and Winutils. Memory constraints needed tuning for large datasets.

b) Google Colab: Spark shell in Scala is not supported. Required switching to Python or local notebooks.

c) SQLite: Lacks the ability to handle very large datasets or parallel queries, but is sufficient for prototyping.

d) Matplotlib: Can become cluttered with overlapping labels, requiring careful layout management.

e) Pandas: Performance degraded on large joins and groupby operations required chunking for efficiency.

f) Understanding the limitations of each tool helped us design workarounds that maintained the quality and integrity of the analysis.

Conclusion

This project demonstrates how big data tools can be used effectively in the retail sector. From setting up a hybrid processing environment using Spark and Pandas to storing and analyzing data in SQLite, the project encapsulates real-world challenges and practical solutions. By combining foundational big data principles with hands-on tools, we gained valuable insights into retail performance and customer behavior.

Retail organizations that invest in scalable data pipelines, flexible tools, and skilled data teams are better positioned to respond to market changes and customer needs. With continued advances in cloud computing, machine learning, and data storage, the future of big data in retail holds even greater potential.

References

Chen, D. (2012). Online Retail Data Set. UCI Machine Learning Repository. University of California, Irvine. https://archive.ics.uci.edu/dataset/352/online+retail

EMC Education Services. (2015). Data science and big data analytics: Discovering, analyzing, visualizing and presenting data. Wiley.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, 98–115. https://doi.org/10.1016/j.is.2014.07.006

Russom, P. (2011). Big data analytics. TDWI Best Practices Report. https://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx