International Journal of Modern Engineering and Management | IJMEM
ISSN No: 3048-8230
Volume 2 Issue 6 June - 2025

Apache Spark-Based Big Data Analytics on E-commerce Trends

Ayushika Singh
Dept. Of Computer Science and Engineering(Data Science)
Noida Institute of Engineering and Technology
Greater Noida, India
ayushika793@gmail.com

***Abstract*— In the era of digital transformation, e-commerce platforms generate massive volumes of data from user transactions, behaviors, and trends. This paper presents a scalable and interactive system using Apache Spark for big data processing, Streamlit for user interface design, and Gemini AI for AI-driven insights. The proposed solution allows users to upload e-commerce datasets, perform real-time preprocessing and visualizations, and obtain intelligent insights powered by generative AI. With support for CSV export, PDF reporting, and theme-based user experience, the system aims to bridge the gap between technical analytics and business decision-making. The paper evaluates the system's performance, usability, and scope for future enhancements.**

Keywords— Apache Spark, E-commerce Analytics, Generative AI, Streamlit Dashboard, Big Data Processing
Key Words: Artificial Intelligence, Transportation, Safety.

***Abstract*— In the era of digital transformation, e-commerce platforms generate massive volumes of data from user transactions, behaviors, and trends. This paper presents a scalable and interactive system using Apache Spark for big data processing, Streamlit for user interface design, and Gemini AI for AI-driven insights. The proposed solution allows users to upload e-commerce datasets, perform real-time preprocessing and visualizations, and obtain intelligent insights powered by generative AI. With support for CSV export, PDF reporting, and theme-based user experience, the system aims to bridge the gap between technical analytics and business decision-making. The paper evaluates the system's performance, usability, and scope for future enhancements.**
Keywords— Apache Spark, E-commerce Analytics, Generative AI, Streamlit Dashboard, Big Data Processing

INTRODUCTION

In recent years, the global expansion of e-commerce has led to an unprecedented surge in digital transactions, user interactions, and behavioral data. With the proliferation of smartphones, secure payment gateways, and personalized shopping experiences, online platforms now collect vast and complex datasets on a daily basis. These include customer purchase histories, browsing behavior, seasonal trends, product reviews, and regional sales data—collectively referred to as "e-commerce big data." Harnessing the value embedded in this data is critical for businesses seeking to remain competitive in a fast-evolving digital marketplace.

Despite the abundance of data, many organizations still rely on traditional analytics tools such as spreadsheets or small-scale processing engines like standalone Python or R scripts. These approaches often fail to scale efficiently and are limited in handling high-velocity or high-volume data. Moreover, they lack interactivity and fail to support real-time decision-making, especially when dealing with datasets exceeding hundreds of thousands of records. The result is a significant gap between data collection and actionable insight generation. To address these challenges, this research proposes an end-to-end solution that combines the distributed computing power of Apache Spark with the accessibility of Streamlit, an open-source Python framework for building interactive web apps. In addition, the system integrates Gemini AI, Google's generative language model, to assist in interpreting analytics outcomes through natural language responses. This hybrid system not only supports the scalable processing of large datasets but also improves usability by offering chart suggestions, automated insights, and downloadable report features—all accessible via a lightweight web interface.The primary objective of this research is to design and implement a system capable of:

1. Efficiently processing large-scale e-commerce data using Apache Spark.
2. Providing a user-friendly web interface for data exploration and visualization.
3. Generating contextual insights through the integration of generative AI (Gemini).
4. Supporting output in multiple formats including CSV and PDF.
5. Enhancing user experience through customizable themes and interactive feedback.

The contributions of this paper are multifold. First, it presents a practical implementation of a big data pipeline tailored for e-commerce analytics. Second, it demonstrates the value of combining traditional analytics with generative AI to bridge the interpretability gap. Lastly, it showcases how open-source tools can be orchestrated to deliver a scalable, intelligent, and accessible analytics platform suitable for academic, research, and commercial applications.

LITERATURE REVIEW

Traditional vs Modern Big Data Analytics Tools

Traditional data analytics solutions, such as Excel, SQL-based BI tools, and standalone Python scripts, have historically served organizations for summarizing and visualizing structured datasets. These tools are often intuitive but limited in scalability and real-time processing capability. As data complexity and volume increased, especially in sectors like e-commerce, their inability to efficiently handle high-dimensional, high-velocity data became evident.

Modern big data analytics frameworks emerged to address these limitations. Technologies such as Hadoop introduced distributed storage and batch processing, while Apache Spark revolutionized the field by enabling in-memory computations, fault tolerance, and support for multiple programming languages. These tools are better suited for handling terabyte-scale datasets in real time, facilitating faster data transformation and analysis pipelines.

Apache Spark and Its Relevance in Big Data

Apache Spark, introduced by Zaharia et al. [1], is a unified analytics engine known for its high-speed, in-memory cluster computing capabilities. Unlike Hadoop's MapReduce model, Spark allows iterative operations and interactive querying via Spark SQL and DataFrames. Spark also supports streaming, machine learning (MLlib), and graph processing (GraphX), making it an ideal choice for end-to-end data analysis workflows.

Its relevance in e-commerce analytics lies in its ability to scale horizontally, handle structured and unstructured data, and integrate seamlessly with real-time applications. Spark's ecosystem is widely adopted across industry verticals and serves as the backbone for many enterprise data lakes and AI platforms.

Existing work on E-commerce Trend Analytics

Numerous studies have investigated methods for understanding and forecasting e-commerce trends. Researchers have applied techniques like linear regression, decision trees, and clustering to model customer behavior, sales fluctuations, and product popularity. Malik and Khan [4] explored machine learning models for sales prediction, while other works focused on customer segmentation and recommendation systems. However, most of these implementations rely on static datasets and lack integration with scalable platforms or interactive dashboards. The need for a system that can handle both the volume and velocity of e-commerce data while providing interpretive support remains largely unmet.

Role of AI in Augmenting Data Interpretation

Generative AI models like Google's Gemini or OpenAI's GPT are transforming the way humans interact with data. By interpreting data patterns and summarizing findings in natural language, these models reduce the cognitive barrier for non-technical users. AI-powered assistants can generate hypotheses, suggest visualizations, and even provide predictive recommendations. Recent advances in AI have been applied to healthcare, finance, and education, but their application in exploratory data analytics—particularly in self-service BI tools—is still emerging. This research pioneers such integration in the context of big data analytics for e-commerce.

Gaps in Current Research

Despite advancements in both big data frameworks and AI, a major gap remains in their practical convergence within real-world analytics workflows. Few systems integrate Spark-based processing with intelligent front-end solutions that guide users through visual data exploration. Similarly, the use of generative AI for contextual interpretation of trends is not widely implemented in scalable dashboards. This research addresses these gaps by building a platform that combines the processing power of Apache Spark, the accessibility of Streamlit, and the interpretability of Gemini AI—all within a single pipeline designed for dynamic e-commerce trend analysis.

SYSTEM ARCHITECTURE AND DESIGN

The proposed system is designed to facilitate scalable and interactive big data analytics on e-commerce datasets by integrating distributed computing, data visualization, and AI-based interpretation. This section provides an overview of the system's architecture, a breakdown of its functional modules, and the design principles that guided its development.

Overall Architecture Diagram

The system is structured into three primary layers:

1.Presentation Layer – Implements the front-end using Streamlit, allowing users to upload datasets, explore visualizations, and receive AI-generated insights.

2.Processing Layer – Handles data ingestion, transformation, and aggregation using Apache Spark and PySpark.

3.Service Layer – Integrates generative AI through Gemini for insights and FPDF for report generation.

The flow begins with the user uploading a CSV or Parquet dataset. Apache Spark ingests and preprocesses the data, which is then converted to a Pandas DataFrame for lightweight operations. Based on user input, the system generates visualizations, AI summaries, and exportable reports.

Module Breakdown

- Data Ingestion (Apache Spark)
  Data ingestion is initiated when a user uploads a dataset through the Streamlit interface. Apache Spark reads the data using SparkSession and stores it as a Spark DataFrame. This enables the system to manage large datasets without memory bottlenecks and prepares the data for transformation.

- Data Processing (PySpark, Pandas)
  Once ingested, the data undergoes cleaning and preprocessing using PySpark's SQL and DataFrame APIs. Operations include null handling, type casting, and filtering. For visualization and user interaction, Spark DataFrames are converted into Pandas DataFrames, which are lightweight and better suited for UI-based tasks.

- Visualization (Matplotlib, Seaborn)
  The Pandas DataFrame is used to generate visual representations such as bar charts, scatter plots, and heatmaps. These are implemented using Python's Matplotlib and Seaborn libraries. The visualization engine automatically suggests chart types based on the selected column data types, ensuring context-aware outputs.

- AI Insight Generation (Gemini API)
  This module uses Google's Gemini API to interpret relationships between data columns. A sample of the dataset is passed as a prompt, and Gemini returns a natural language insight. This feature helps users—especially non-technical ones—understand trends without manual data analysis.
- Report Generation (FPDF)
  FPDF is used to compile selected charts and AI-generated insights into a downloadable PDF report. The module allows users to export both numerical and visual analysis, enabling sharing and offline review.

Design Principles: Modularity, Scalability, Usability

- Modularity: Each module (data ingestion, AI generation, visualization, export) is developed independently, allowing easy maintenance and future upgrades.
- Scalability: Apache Spark ensures the platform can handle large-scale datasets without performance degradation. The design supports cloud-based deployment for further scalability.
- Usability: The Streamlit interface is intuitive and responsive. Features such as light/dark themes, column selection, and download buttons improve user experience. The inclusion of AI insights adds accessibility for users with limited data literacy.

METHODOLOGY

A. Dataset Acquisition and Characteristics The datasets used for testing were derived from open e-commerce platforms and contained attributes like product categories, sales values, customer locations, order dates, and ratings. Most datasets exceeded 100,000 rows to test performance at scale.

B. Data Preprocessing Techniques Upon ingestion via Spark, data was cleaned using PySpark functions such as dropna() for null values and type casting for consistency. Unnecessary columns were filtered out, and categorical columns were encoded as needed.

C. Chart Suggestion Logic The chart recommendation engine uses a rules-based approach. For example, if column A is categorical and column B is numerical, the system suggests a bar chart. If both are numerical, it suggests scatter plots or heatmaps.

D. AI Prompt Engineering for Gemini The system uses sampled rows (first 5–10) from selected columns and formats them into text prompts. Prompts are then passed to Gemini's API for interpretation. An example: "Analyze the relationship between Product Category and Sales using the sample below..."

E. Integration Flow: From File Upload to Downloadable Report The user uploads a dataset via Streamlit. Spark processes it, visualizations are generated, Gemini provides insights, and users can download results as CSV or PDF—all within a single session.

F. PERT/Gantt Overview for Project Phases The project followed a Gantt timeline with five major phases: Requirements Analysis, Design, Development, Testing, and Documentation. PERT analysis identified data integration and AI insight generation as critical paths.

IMPLEMENTATION

- Environment Setup (Spark, Streamlit, Python 3.10+) :The system was built using Python 3.10+, Apache Spark 3.x, and Streamlit 1.25+. Other packages include google-generativeai, fpdf, pandas, matplotlib, and seaborn.
- Key Code Components : The SparkProcessor class initializes the SparkSession and reads files. Visualization functions generate charts using Matplotlib/Seaborn. The Gemini integration wraps the API call with a sample prompt for text generation.
- UI Interaction Features : The Streamlit interface offers toggling between light and dark themes, selection of data columns, and dynamic chart rendering based on user input.
- Session Management (Stateful PDF/Chart Export) : Using Streamlit's st.session_state, the system stores charts and processed data temporarily to allow export to PDF or CSV upon user request.
- Error Handling and Performance Tuning : The system includes try-except blocks around all major operations including file uploads, API calls, and data parsing. Spark's in-memory processing ensures reduced lag on large datasets.

RESULT AND ANALYSIS

The developed analytics system was evaluated on multiple dimensions, including data processing speed, visualization quality, and the effectiveness of AI-generated insights. The primary focus was on how Apache Spark improves the handling of large-scale datasets and how the integration of Gemini AI supports interpretability. In terms of processing performance, Apache Spark demonstrated substantial gains over traditional tools such as Pandas. For example, loading and preprocessing a dataset containing over 100,000 rows took less than half the time with Spark. Its in-memory computation and distributed architecture enabled efficient execution of filtering, aggregation, and transformation operations without overwhelming system resources.

The visualization engine, built using Matplotlib and Seaborn, accurately reflected the relationships between user-selected columns. The system's ability to automatically suggest chart types—such as bar charts for categorical-numeric pairs or scatter plots for numeric-numeric combinations—was found to be highly effective. These visuals not only improved user comprehension but also highlighted patterns such as sales peaks, product demand trends, and geographic variations.

The inclusion of Gemini AI significantly enhanced the interpretability of data. By analyzing small, representative samples of the dataset, Gemini generated human-readable insights that complemented visual analysis. For instance, in one case, the model identified a seasonal dip in sales within specific product categories, which was not immediately obvious from charts alone. Such insights bridge the gap between raw analytics and strategic decision-making.

Overall, the system's results underscore its potential as a scalable, user-friendly analytics platform for e-commerce trend analysis. The combination of Spark's performance, Streamlit's interactivity, and Gemini's contextual intelligence makes the system both powerful and accessible.

CONCLUSION

This research presents a comprehensive and scalable system for conducting big data analytics on e-commerce trends by integrating Apache Spark, Streamlit, and Gemini AI. As e-commerce platforms continue to generate massive volumes of data, there is a growing need for solutions that are not only computationally powerful but also intuitive and interpretable. Traditional tools often fall short in processing speed, interactivity, and accessibility, especially for non-technical users. The system developed in this project addresses these challenges by leveraging Apache Spark for distributed data processing, ensuring efficient handling of large datasets without compromising performance. The interactive interface built with Streamlit empowers users to upload data, select analytical parameters, visualize results, and generate downloadable reports in real time. Furthermore, the integration of Gemini AI introduces a layer of cognitive intelligence, transforming raw data into actionable insights through natural language interpretation.

The outcomes of the project validate the system's ability to significantly reduce data processing time, enhance visualization accuracy, and provide meaningful AI-generated interpretations. The modular architecture and seamless user experience make this platform suitable for academic research, business analytics, and educational use cases. In essence, this work demonstrates how combining open-source technologies with generative AI can democratize big data analytics. It sets a foundation for future innovations where complex analytical tasks can be performed and understood with minimal technical overhead, making data-driven decision-making more inclusive and efficient.

FUTURE SCOPE

While the current system effectively integrates Apache Spark, Streamlit, and Gemini AI for scalable and intelligent e-commerce analytics, several enhancements can be made to extend its functionality and impact. Future iterations could incorporate **real-time data streaming** using Spark Structured Streaming to handle continuous data inflow from live e-commerce platforms. Additionally, deploying the application on cloud platforms like AWS or GCP would enable **horizontal scalability and multi-user access**. Incorporating **role-based dashboards** and **user authentication** would further align the system with enterprise needs. Moreover, integrating **voice-based query input** and **multilingual support** could improve accessibility for non-technical and diverse users. The use of reinforcement learning or adaptive AI to personalize insights based on user behavior is another promising area. These future developments would transform the platform from a static analytics tool into a dynamic, real-time, AI-powered decision support system.

ACKNOWLEDGEMENT

I would like to express a deep sense of gratitude and thanks profusely to Prof. Sovers Singh Bisht, Asst. HOD(Data Science) project guide, without the wise counsel and able guidance, it would have been impossible to complete the report in this manner.

I express gratitude to other faculty members of Data Science department of NIET for their intellectual support throughout the course of this work.

REFERENCES

M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Apache Spark: A unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, Nov. 2016.

A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, Apr. 2015.

M. Minelli, M. Chambers, and A. Dhiraj, *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, 1st ed. Hoboken, NJ: Wiley, 2013.

M. Malik and A. Khan, "E-commerce trend prediction using machine learning techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 5, pp. 292–298, 2019.

Google Cloud, "Vertex AI and Gemini Integration Guide," [Online]. Available: https://cloud.google.com/vertex-ai/docs/generative-ai/overview.