Predictive Modeling of E-Commerce Purchase Behavior using Clustering techniques and Spark

Peta Sandeep
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidypeetham, India
BL.EN.U4CSE21156@bl.students.amrita.edu

Rejeti Kartik
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidypeetham, India
BL.EN.U4CSE21170@bl.students.amrita.edu

Sanjay Baitha
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidypeetham, India
BL.EN.U4CSE21185@bl.students.amrita.edu

Vanipenta Pavan Kumar Reddy
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidypeetham, India
BL.EN.U4CSE21215@bl.students.amrita.edu

Dr. Nidhin Prabhakar T V
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidypeetham, India
tv_nidhin@blr.amrita.edu

***Abstract*—The ever-growing scale of e-commerce has caused enormous quantities of records, necessitating advanced strategies for extracting actionable insights into patron conduct. This challenge makes a speciality of predictive modeling of e-commerce buy behavior the use of clustering strategies and Spark's allotted data processing framework. The technique integrates Spark's in-memory computing with superior clustering algorithms, along with K-Means, K-Means++, and Gaussian Mixture Model (GMM). By reducing functions, handling lacking values, and imposing the RFM (Recency, Frequency, Monetary) version, this method segments customers effectively. A comparative evaluation is achieved among Spark-based and non-Spark-primarily based methodologies to assess scalability, performance, and accuracy. The integration of Spark ensures fault tolerance, parallel processing, and improved clustering performance, optimizing consumer insights and enterprise techniques.**

***Index Terms*—Predictive Modeling, E-Commerce Purchase Behavior, Apache Spark, Clustering Techniques, K-Means, K-Means++, Gaussian Mixture Model, RFM Method, Big Data Analytics, Customer Segmentation, Machine Learning Integration.**

I. Introduction

In the era of digital transformation, e-trade structures are rapidly redefining worldwide retail landscapes, offering exceptional convenience and personalised reviews to purchasers. The exponential increase of online transactions has made understanding and predicting consumer buy conduct a crucial focus for agencies. Leveraging machine learning and huge facts analytics, companies can discover actionable insights to enhance patron satisfaction, optimize inventory management, and pressure sales boom.

The venture of as it should be predicting e-trade buy behavior lies in the large and heterogeneous nature of the records generated through person interactions. Clickstream statistics, demographic info, browsing styles, and purchase histories produce wealthy but unstructured datasets. Harnessing this statistics to build correct predictive models needs sturdy computational frameworks and superior algorithms capable of figuring out intricate patterns and delivering scalable answers.

This project offers a novel method to e-trade purchase conduct prediction the usage of clustering strategies inside the Spark environment. Spark's distributed computing abilities permit green processing and analysis of huge-scale e-trade datasets, overcoming traditional garage and computational barriers. By integrating system getting to know techniques along with clustering algorithms—K-Means, K-Means++ , and Gaussian Mixture Model (GMM)—with Spark, this framework enables patron segmentation and offers insights into purchase patterns.

The method includes information preprocessing, characteristic engineering, and making use of clustering strategies to phase clients effectively. By leveraging Spark's in-memory processing and scalability, the proposed solution guarantees more desirable performance in managing extensive datasets, enabling groups to are expecting consumer alternatives, optimize choice-making, and supply personalized user stories. This work bridges the distance among information technological know-how and retail analytics, empowering e-commerce platforms with predictive equipment that beautify the client journey and foster lengthy-time period growth.

II. LITERATURE REVIEW

The upward push of e-commerce has created great datasets that require superior strategies for evaluation and prediction. Leveraging Hadoop's disbursed architecture and machine learning fashions offers a robust framework to investigate customer purchase behavior, permitting scalable and correct predictions.

Amirhossein Jamarani et al. [1] provides a systematic literature review of big data and predictive analytics, examining the main research approaches, applications, evaluation metrics, tools, and challenges in this field. The key models used in Industrial include machine learning, deep learning, and structural equation modeling. The advantages are improved accuracy in predicting carbon emissions, identifying industry trends, and understanding the impact of IT and HR capabilities on organizational performance. The disadvantages are not explicitly mentioned, but the papers highlight the need to address important issues and future directions in these models. Similarly The key models used in E-commerce include machine learning ensembles, text-matching algorithms, and quantitative models. The advantages are improved accuracy in predicting financial time-series data, increased transaction chances, and personalized service. The disadvantages include loss of precision when tested on commodity search and limitations due to the impact of COVID-19 on travel frequency.

Xiaodong Zhang et al. [2] introduces a multimodal gaining knowledge of framework designed to beautify the information and prediction of purchaser behavior inside the cross-border e-commerce industry by way of integrating various statistics assets. The framework consists of three fashions: Unimodal Model I, which utilizes marketplace positioning data, demonstrating mild performance with progressively lowering education loss and improving accuracy; and Unimodal Model II, primarily based on product attribute statistics, which plays in addition to the primary model. The spotlight of the have a look at is the Multimodal Model, using tensor fusion to integrate both statistics sorts, drastically surpassing the unimodal models in performance, attaining decrease education loss and better accuracy in both schooling and checking out stages. While the multimodal approach demonstrates significant advantages, such as a extra comprehensive expertise of consumer conduct and stepped forward predictive capability, it additionally poses challenges, together with elevated computational complexity and potential integration problems associated with multimodal information fusion.

Hussain Saleem et al. [3] explores how information technology and machine studying strategies enhance e-trade income overall performance at the social web. Supervised Learning identifies patterns and predicts target attributes however is based closely on categorised schooling statistics. Unsupervised Learning uncovers intrinsic data structures without requiring labels, even though the problem depends on the algorithm used. Semi-Supervised Learning improves classifier accuracy by way of leveraging unlabeled statistics however still calls for some categorized examples. Finally, Reinforcement Learning enables mastering via rewards and studies with out complete environmental understanding however entails better implementation complexity. These procedures provide numerous blessings and challenges, depending at the context in their utility.

Shahriar Akter et al. [4] systematically opinions large statistics analytics (BDA) in e-commerce, that specialize in its definitions, traits, types, business value, and challenges. Using the Resource-Based View (RBV), BDA is framed as a distinctive competence aiding key e-trade features like client loyalty, pricing, and stock management. From a Sociomaterialism Perspective, the integration of control, technology, and talent in BDA is highlighted as important for boosting corporation performance. Leveraging Service Marketing Theory, BDA is proven to permit carrier innovation and beautify provider transport models, emphasizing its transformative ability inside the e-trade region.

Deborah Osinachi et al. [5] explores the use of device studying algorithms to decorate predictive analytics in consumer behavior studies, focusing on their blessings and packages. The key model used like Decision Tree, SVM, Clustering algorithms. Decision Trees provide interpretable models for decision-making, even as Neural Networks excel at shooting complex patterns and relationships. Support Vector Machines (SVMs) provide sturdy type skills, and Clustering Algorithms like ok-manner are effective for consumer segmentation. Regression Analysis predicts continuous consequences, making it valuable for fashion forecasting. Although the paper does not element the negative aspects, it emphasizes that choosing the appropriate set of rules depends on the look at's objectives, data traits, and model complexity.

V V Satyanarayana Tallapragada et al. [6] proposes an IoT-integrated big data analytics system called EMOMETRIC that can track and analyze real-time retail customer emotions to provide business intelligence to retailers.The Emometric system uses a facial model constructed from annotated face data, where the key variations are extracted using Principal Component Analysis (PCA). The model is able to efficiently capture key facial expressions and poses like yaw, pitch, mouth opening/closing, and smiling. The key advantages of this model-based approach are that it is easy to construct since faces don't vary much in geometry across people, and the PCA-based parameter extraction allows for a compact and efficient representation of the facial features. The paper does not mention any significant disadvantages of this model.

Kirti Mahajan et al. [7] examines how e-commerce companies can use Big Data and Cloud Computing to cater to customers' needs through predictive analysis and data-driven insights. The model used in the study is a Random Forest regression model. The key advantages of this model are that it allows for personalized customer experiences and better understanding of customer feedback, which can improve customer service and satisfaction. The paper does not explicitly mention any disadvantages of the Random Forest model, but suggests that future research could explore the use of advanced deep learning techniques and edge computing to improve the

model's accuracy and real-time responsiveness.

Gautam Pal et al. [8] introduces a Multi-Agent Lambda Architecture (MALA) system for e-commerce analytics, combining real-time and batch processing to supply deeper insights at low latency. Key benefits consist of clever collaboration between components, faster education instances, adaptability, horizontal scalability, and decreased infrastructure prices compared to unmarried-layer tactics. However, preserving separate codebases for move and batch layers can reason synchronization challenges, and for less difficult use instances, the MALA model may not beautify accuracy over batch-only fashions. Additionally, its complexity can be excessive for e-commerce situations with smaller statistics volumes.

Hua Wang et al. [9] examines strategies for analyzing and predicting e-commerce user behavior the use of information mining and synthetic neural networks. Data mining technology successfully uncovers styles and rules in user conduct data, assisting in user profiling, personalised hints, and churn prediction. Artificial neural networks excel at figuring out complicated patterns and dealing with big-scale datasets, making them ideal for building sturdy user behavior prediction models. These techniques are further leveraged to advise optimization strategies and actionable suggestions for enhancing e-commerce platform performance.

Sedat Usluoglu et al. [10] offers an correct product categorization machine for e-commerce platforms, leveraging huge information analytics and machine studying algorithms along with Naive Bayes (NB), Stochastic Gradient Descent (SGD), and Support Vector Machines (SVM). Naive Bayes is fast, smooth to put in force, and powerful but assumes independence amongst times, which won't preserve in actual-world records. SGD, a variant of trendy gradient descent, addresses challenges like sluggish convergence and the inability to discover a international minimal. SVM gives sturdy supervised getting to know by using identifying an top-rated hyperplane for sophistication separation with most margin. These algorithms enhance the efficiency and accuracy of product categorization.

Dheeraj Malhotra et al. [11] shows limitations of traditional seek systems in assisting massive facts analytics for E-Commerce environments. It highlights the need for customized seek and ranking systems by means of studying patron preferences and browsing behavior. The observe employs the Hadoop-based totally HDFS-MapReduce model to expand the Relevancy Vector (RV) set of rules, included into the Intelligent Meta Search System for Advanced E-Commerce (IMSS-AE). Advantages: It offers scalable, extensible, and failure-tolerant customized search outputs. Disadvantages: Implementation complexities and dependency on high computational assets for dealing with large facts analytics.

Afifah Farhanah Akadji et al. [12] emphasizes leveraging huge statistics analytics for product demand prediction within the Indonesian e-commerce region, identifying themes like records high-quality, analytical equipment, and organizational lifestyle. The research employs a qualitative analysis version, gathering insights through interviews with enterprise experts. Advantages: Enhanced predictive accuracy, operational performance, and choice-making competencies. Disadvantages: Challenges in information great, integration, and resistance to cultural trade inside companies.

Omorinsola Bibire Seyi- Lande et al. [13] highlights the enhancement of Business Intelligence (BI) in e-trade through superior records integration techniques like extraction, transform and load operation, records virtualization, API integration, and streaming data integration. These strategies allow actual-time insights for dynamic pricing, personalised purchaser reviews, and fraud detection. Model used: Data integration and real-time analytics frameworks blended with AI and gadget learning. Advantages: Improved operational performance, client delight, and selection-making. Disadvantages: Challenges consist of records silos, scalability, technical complexity, and protection worries. Future traits including aspect computing and blockchain promise similarly improvements in e-commerce BI.

Babasaheb Jadhav et al. [14] delves into the transformative effect of large records on choice-making methods, highlighting its integration with IoT, cloud computing, and mobile gadgets. Model used: Conceptual frameworks incorporating facts warehousing, OLAP, and analytics techniques. Advantages: Improved selection-making via actual-time insights, competitive intelligence, and operational efficiency. Disadvantages: High implementation complexity and useful resource demands for small and medium businesses. The take a look at underscores the want for adaptable analytics frameworks to help companies of all scales leverage huge facts successfully.

El Falah Zineb et al. [15] explores the role of Big Data Analytics in enhancing e-commerce decision-making by applying machine learning models to analyze customer behavior, segmentation, and reviews. Model used: Algorithms like k-means for segmentation and classification techniques for review analysis. Advantages: Improved personalization, dynamic pricing, better customer support, and actionable insights for marketing strategies. Disadvantages: Challenges include managing data volume, velocity, and variety, as well as the need for advanced tools to address technological limitations.

III. Dataset Description

The dataset used for this assignment comprises **118,307 rows** and **39 columns** of e-commerce transaction data. Each row represents a transaction, and the columns capture various attributes, including order details, customer information, product specifications, and payment methods. Below are some of the key columns from the dataset:

- **order_id:** Unique identifier for a transaction.
- **customer_id:** Unique identifier for a particular customer.
- **order_status:** The current status of the order (e.g., delivered, shipped, processing, or canceled).
- **payment_type:** The method used for payment (e.g., credit card, voucher, boleto).
- **customer_city** and **customer_state:** Information about the customer's location.

- **product_category_name:** The category to which the purchased product belongs.
- **product_category_name_english:** The translated category name in English.
- **price:** The amount spent per transaction.
- **seller_city:** The city where the transaction has occurred.

IV. Methodology

A. Spark

Apache Spark is a powerful open source software, based on distributed file system that is very useful to handle big data processing datasets. It is built on the Resilient Distributed Dataset architecture which is an immutable distributed collection of elements of the data that is being used partitioned across nodes in the cluster. Its core engine allows for in-memory processing, allowing for better reading and writing immediate to disk. It can handle big data because it can keep many distributed clusters and spread the volume of data across these clusters along with fault tolerance and parallel processing. It also allows for integration with Hadoop, Hive, Machine learning, Structured Query language, streaming and many more tools to enhance the working of the machinery.

B. KMeans clustering

K-Means clustering is a clustering algorithm that partitions the data into K different clusters based on similarity of features. The clusters are initially chosen at random and are updated with the mean value of the cluster points. It is done until the centroids stabilize minimizing the euclidean distance between data points. The clusters are usually equal in size.

C. KMeans++ clustering

K-Means++ clustering is an enhanced version of the K-Means clustering algorithm in which the clusters are chosen using a probabilistic method instead of randomly being chosen, that ensures that they are spread out across the dataset. The first centroid is randomly chosen but subsequent centroids are chosen based on distance from chosen points. This ensures that there are fewer chances of getting stuck in the local minima.

D. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) clustering is a probabilistic clustering method which assumes the data is in Gaussian form, which each of them represent a cluster. Unlike K-Means clustering, it assigns each data point to a probability of multiple clusters. This may cause clusters to overlap or have odd shapes. It uses the expectation maximization algorithm to refine the parameters such as mean, covariance, and weights to get the maximum likelihood of the observed data.

E. Proposed Methodology

The dataset having 39 columns, needs to reduced as there are a lot of unimportant features. The review features, totaling 6, are not important for the context of getting user insights and hence are removed. There are many missing values present in each feature and they are handled as such. The categorical features are filled with mode value and the numerical features are filled with the mean value. We find for each customer_id the recency, frequency and the monetary value using the RFM method for each customer. Then methodology is split into two ways, one working with Spark and another without Spark. Establishing the Spark session, we apply label encoding on the categorical features and then applied standard scaling to normalize the entire data. We remove, in both, the most correlated columns, bringing down the features to 28 features. Finally both sections undergo the clustering methods and their spread of the data along with metrics are compared across all methodologies. The clustering methods have been cross validated to get the best result. The labels have been compared to the price insights. The Fig. 1 shows the overall workflow.

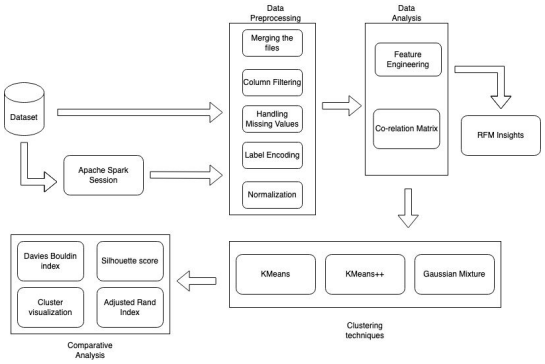

Fig. 1. System workflow

V. Results

A. Metrics Used

The clustering with or without Spark has to be measured to allow for a comparative analysis. The metrics used in the clustering methods are Silhouette score, Davies-Bouldin index (DB), and Adjusted rand index (ARI). The Silhouette score is defined for the data point i:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{1}$$

where $a(i)$ is the average intra-cluster distance and $b(i)$ is the average inter-cluster distance. The DB index is defined as:

$$DBI = \frac{1}{k}\Sigma_{i=1}^{k} max_{j!=i}\frac{S_i + S_j}{D_{ij}} \tag{2}$$

where S_i is the average intra-cluster distance for cluster i, S_j is the average intra-cluster distance for cluster j, D_{ij} is the distance between the centroids of clusters i, and j and k is the total number of clusters. The ARI score is defined as:

$$ARI = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \tag{3}$$

where Index is the number of pair agreements, the expected index is the expected number of agreements and max index is the maximum possible of agreements.

	customer_id	recency	frequency	monetary
0	00012a2ce6f8dcda20d059ce98491703	-436	1	89.80
1	000161a058600d5901f007fab4c27140	-315	1	54.90
2	0001fd6190edaaf884bcaf3d49edf079	-177	1	179.99
3	0002414f95344307404f0ace7a26f1d5	-346	1	149.90
4	000379cdec625522490c315e70c7a9fb	-575	1	93.00
...	...	...	...	...
98660	fffcb937e9dd47a13f05ecb8290f4d3e	-559	1	78.00
98661	fffecc9f79fd8c764f843e9951b11341	-571	3	164.70
98662	fffeda5b6d849fbd39689bb92087f431	-625	1	47.90
98663	ffff42319e9b2d713724ae527742af25	-647	1	199.90
98664	ffffa3172527f765de70084a7e53aae8	-363	2	21.80

Fig. 2. The Recency Frequency and Monetary Metrics for a customer ID

B. RFM Insights

The Fig. 2 shows the recency, frequency and monetary values for each customer_id where for example ID: 9ef432eb6251297304e76186b10a928d has the latest transaction at 436 days, totally done 1 time and amount of 89.80 Real and also ID: fffecc9f79fd8c764f843e9951b11341 has the latest transaction at 571 days, totally done 3 times and amounting to 164.70 Real.

C. Clustering methods

The clustering methods applied is shown in Fig. 3 to apply cross validation to get the best hyperparameter, where K for K-Means and K-Means++ is best at 2 and n_samples for GMM is 2. The Fig. 4 shows the GMM clustering running without

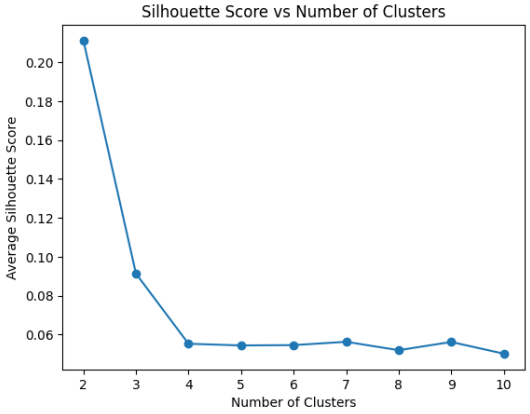

Fig. 3. Clusters cross validation curve against silhouette score

Spark, and Fig. 5 shows the GMM clustering running with Spark. Both show the two clusters being formed have a lot of overlap.

On the other hand, The Fig. 6 shows the K-Means clustering running without Spark, and the Fig. 7 shows the K-means clustering running with Spark. The clusters have a lot less overlap.

The Fig. 8 shows the K-Means++ clustering running without Spark, and the Fig. 9 shows the K-Means++ clustering running with Spark. The clusters also show a lot less overlap.

The Table I shows overall the clustering methods compared using Spark with the metrics. Among these methods the GMM performs the worst on all metrics. The K-Means and K-Means++ have the same silhouette score and ARI, however,

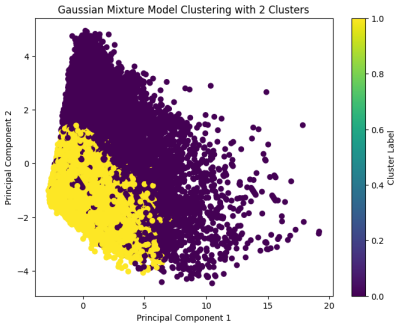

Fig. 4. Visualization of GMM without Spark

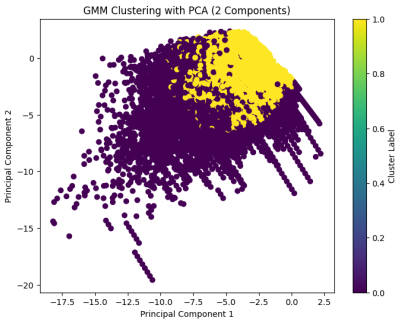

Fig. 5. Visualization of GMM with Spark

K-means++ performs the best by having the lower amount of DB index.

The Table II shows overall the clustering methods compared without using Spark with the metrics. Among these methods the GMM performs the worst on all metrics. The K-Means++ however outperforms the other methods for having lowest silhouette score, highest ARI and second lowest DB index.

VI. Conclusion

The current paper has shown that conducting research using the big data tools, especially the Hadoop, coupled with sophisticated algorithms in the machine learning can enhance the prediction of the e-commerce buy behavior adequately. Compared to traditional databases, e-commerce Hadoop system scales the amorphous and diverse data on distributed storage

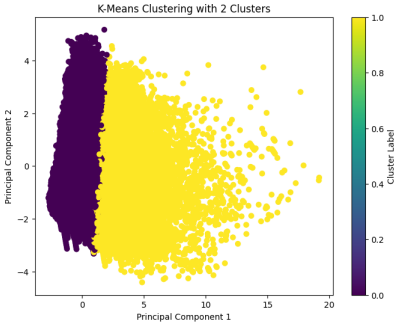

Fig. 6. Visualization of K-Means without Spark

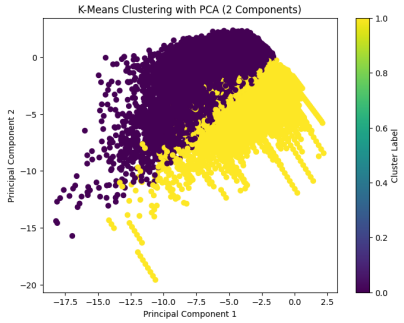

Fig. 7. Visualization of K-Means with Spark

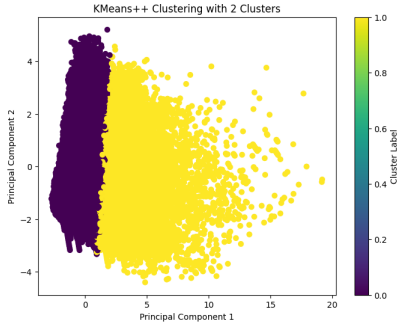

Fig. 8. Visualization of K-Means++ without Spark

of Hadoop and handles wide range of parallel computational problems. The subsequent processed data to predictive models enable machine learning models to unveil subtle patterns and provide accurate prognosis of results, making improvements the decision-making procedures in business innovation such as inventory control, customer categorization, and tailored marketing solutions. The findings confirm the effectiveness of the suggested framework to enhance business progress and increase customers' satisfaction.

Furthermore, the findings of this research about interface between big data technologies and predictive analytics support the notion of applicability of such combinations in today's e-commerce context. In real-time analysis and prediction tools that enable a business to respond quickly to shifts in the market or customer demand, and use data-driven strategies to promote techniques for retailing. With regards to the role of this work as a bridge between computational frameworks and retail analytics, it identifies the nature of subsequent progress in e-commerce systems, focusing on the necessity of creating scalable intelligent solutions, ultimately defining the strategies towards successful formation of superior customer experiences and building the long-terms growth in the field of the digital economy.

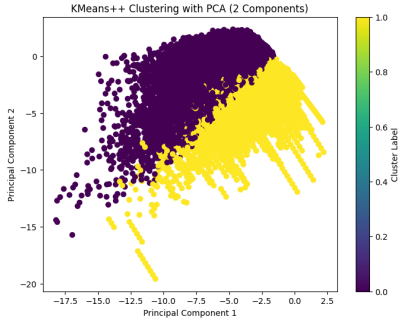

Fig. 9. Visualization of K-Means++ with Spark

TABLE I
COMPARISON AMONG CLUSTERING METHODS WITH SPARK

Clustering methods	Silhouette Score	Adjusted Rand Index	Davies Bouldin Index
KMeans	0.21017	0.00108	2.7116677
GMM	4.4948	0.013619	4.18876
KMeans++	0.21017	0.001080	2.17166

TABLE II
COMPARISON AMONG CLUSTERING METHODS WITHOUT SPARK

Clustering methods	Silhouette Score	Adjusted Rand Index	Davies Bouldin Index
KMeans	0.19957	0.00175	2.64995
GMM	0.11575	0.000399	4.038177
KMeans++	0.19357	0.00175	2.738248

REFERENCES

[1] A. Jamarani, S. Haddadi, R. Sarvizadeh, M. Haghi Kashani, M. Akbari, and S. Moradi, "Big data and predictive analytics: A sytematic review of applications," *Artificial Intelligence Review*, vol. 57, no. 7, p. 176, 2024.

[2] X. Zhang and C. Guo, "Research on multimodal prediction of e-commerce customer satisfaction driven by big data," *Applied Sciences*, vol. 14, no. 18, p. 8181, 2024.

[3] H. Saleem, K. B. Muhammad, A. H. Nizamani, S. Saleem, and A. Aslam, "Data science and machine learning approach to improve e-commerce sales performance on social web," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 19, 2019.

[4] S. Akter and S. F. Wamba, "Big data analytics in e-commerce: a systematic review and agenda for future research," *Electronic Markets*, vol. 26, pp. 173–194, 2016.

[5] O. D. Segun-Falade, O. S. Osundare, W. E. Kedi, P. A. Okeleke, T. I. Ijomah, and O. Y. Abdul-Azeez, "Utilizing machine learning algorithms to enhance predictive analytics in customer behavior studies," 2024.

[6] V. Tallapragada, N. A. Rao, and S. Kanapala, "Emometric: An iot integrated big data analytic system for real time retail customer's emotion tracking and analysis," *International Journal of Computational Intelligence Research*, vol. 13, no. 5, pp. 673–695, 2017.

[7] K. Mahajan, D. Bordoloi, C. Barboza, D. Bansal, B. M. Rao, *et al.*, "Big data with cloud computing model for customer need identification in e-commerce industry," in *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, pp. 155–159, IEEE, 2024.

[8] G. Pal, G. Li, and K. Atkinson, "Multi-agent big-data lambda architecture model for e-commerce analytics," *Data*, vol. 3, no. 4, p. 58, 2018.

[9] H. Wang, L. Wang, and F. Zhu, "E-commerce user behavior analysis and prediction based on artificial neural network and data mining," in *2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 7, pp. 583–586, IEEE, 2024.

[10] S. Usluoğlu, D. Kılınç, and F. Bozyiğit, "E-commerce product categorization using big data analytics," *Artificial Intelligence Theory and Applications*, vol. 1, no. 2, pp. 1–8, 2021.

[11] D. Malhotra and O. Rishi, "An intelligent approach to design of e-commerce metasearch and ranking system using next-generation big data analytics," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 2, pp. 183–194, 2021.

[12] A. F. Akadji and R. Dewantara, "Big data analysis for product demand prediction in indonesian e-commerce," *West Science Information System and Technology*, vol. 2, no. 01, pp. 9–17, 2024.

[13] O. B. Seyi-Lande, E. Johnson, G. S. Adeleke, C. P. Amajuoyi, and B. D. Simpson, "Enhancing business intelligence in e-commerce: Utilizing advanced data integration for real-time insights," *International Journal of Management & Entrepreneurship Research*, vol. 6, no. 6, pp. 1936–1953, 2024.

[14] B. Jadhav, "The role of data science and analytics in predictive modelling and decision-making," 2023.

[15] E. F. Zineb, R. Najat, and A. Jaafar, "An intelligent approach for data analysis and decision making in big data: a case study on e-commerce industry," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.