University "St. Kliment Ohridski"
Bitola
Faculty of Information and
Communication Technology - Bitola
Republic of North Macedonia

PROCEEDINGS

15th International Conference on APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2025

AIIT International Conference

Bitola, November 7, 2025

University “St. Kliment Ohridski” Bitola
Faculty of Information and Communication Technology - Bitola
Republic of North Macedonia

PROCEEDINGS
15th International Conference on
APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2025

AIIT
International Conference

November 7, 2025 Bitola

Proceedings publisher and organizer of the conference:
University "St. Kliment Ohridski", Bitola, Faculty of Information and Communication Technology – Bitola, Republic of North Macedonia

For publisher:
Blagoj Ristevski, PhD, Full Professor, Dean of the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Proceedings editors:
Kostandina Veljanovska, PhD
Željko Stojanov, PhD

Conference Chairmans:
Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia – chair
Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia – co-chair
Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia – co-chair

Technical preparation of the proceedings:

Kostandina Veljanovska, PhD
Marija Apostoloska Kondoska, MSc
Darko Pajkovski, MSc

Cover design:
Kostandina Veljanovska, PhD
Hristina Dimova Popovska, MSc

e-Proceedings
ISBN 978-608-5003-06-8

CIP - Каталогизација во публикација

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

004-049.8(062)

INTERNATIONAL conference on applied internet and information technologies AIIT 2025 (15 ; 2025 ; Bitola, Republic of North Macedonia)
Proceedings / 15th International conference on applied internet and information technologies AIIT 2025, November 7 2025, Bitola, Republic of North Macedonia ; [editors Kostandina Veljanovska, Zeljko Stojanov]. - Bitola : University "St. Kliment Ohridski", Bitola Faculty of information and communication technologies, 2025. - 477 стр. : илустр. ; 30 см
Библиографија кон трудовите
ISBN 978-608-5003-06-8

а) Информатичка технологија -- Примена -- Собири
COBISS.MK-ID 67608325

Introduction

As organizing partners of 15th International Conference on Applied Internet and Information Technologies AIIT 2025, we warmly welcome all participants, researchers, and colleagues joining us from various countries and universities, united by our shared commitment to advancing knowledge in the fields of computer science, applied Internet, and information technologies.

The AIIT conference has become a long-standing tradition of excellence and collaboration, co-organized by the Faculty of Information and Communication Technologies – Bitola, University "St. Kliment Ohridski," and the Technical Faculty "Mihajlo Pupin" – Zrenjanin, University of Novi Sad, Serbia. Over the past fifteen years, this partnership has fostered not only strong academic cooperation but also genuine friendship among our institutions and scholars.

This year's conference proudly continues that tradition, bringing together innovative research, diverse perspectives, and new insights into technologies that are shaping our digital future. The Scientific Program Committee once again faced the demanding task of selecting the highest-quality papers from more than sixty submissions spanning a wide range of topics—including Artificial Intelligence, Immersive Technologies, Mathematical Simulations, Data Science and Big Data Analytics, Knowledge and IT Management, Cybersecurity, Software Engineering, Data Mining, Digital Transformation, Behavioral Economics and Business, Social Engineering, Digital Humanities, Augmented Humanity, and Hybrid Intelligence. This ensures that the program reflects both scientific rigor and creative originality.

We would like to express our sincere gratitude to all reviewers for their dedicated work, as well as to the members of the Organizing Committee for their professionalism, commitment, and enthusiasm in preparing this event.

We are confident that these proceedings will provide an enriching and thought-provoking reading experience.

Conference chairs:

Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (chair)

Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (co - chair)

Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (co – chair)

MAIN ORGANIZERS:

Faculty of Information and Communication Technologies - Bitola
University "St. Kliment Ohridski" University - Bitola
NORTH MACEDONIA
http://fikt.uklo.edu.mk/

Technical Faculty "Mihajlo Pupin" Zrenjanin
University of Novi Sad SERBIA
http://www.tfzr.uns.ac.rs/

ORGANIZATION PARTNERS:

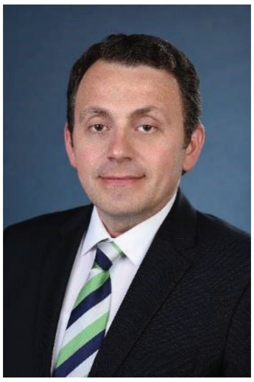
Faculty of Computer Science
Irkutsk National Research Technical University
Institute of Informational Technologies and Data Analysis
Irkutsk, RUSSIA
http://www.istu.edu/

Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences, Irkutsk, RUSSIA
http://idstu.irk.ru/

Irkutsk State Transport University (IrGUPS)
Irkutsk, RUSSIA
https://www.irgups.ru/

Faculty of Engineering South-west
University "Neophyte Rilsky"-Blagoevgrad
BULGARIA
http://www.swu.bg/

Conference Chairs



Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (chair)

Prof. Dr. Blagoj Ristevski is a Full Professor at the Faculty of Information and Communication Technologies (FICT) at the University "St. Kliment Ohridski" - Bitola, where he currently serves as Dean. He holds a PhD in Technical Sciences from the Faculty of Electrical Engineering and Information Technologies, Institute of Computer Science and Informatics, at Ss. Cyril and Methodius University in Skopje. His research interests span Databases, Data Science, Data Mining, Big Data Analytics, Bioinformatics, Computer Graphics, and Cybersecurity. Prof. Ristevski has supervised numerous BSc, MSc, and PhD theses and has led several international research projects. He has served on the management committees of multiple COST actions, reviewed for numerous high-impact journals, and evaluated project proposals for the Horizon 2020 and Horizon Europe programs. Prof. Ristevski is also a senior member of IEEE.

Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (co – chair)

Kostandina Veljanovska, Ph.D. completed her education at the University "Sts. Kiril i Metodi", Skopje (BSc in Computer Science), at the University of Toronto, Toronto (MASc in Applied Engineering) and got her MSc and also her PhD in Technical Sciences at the University "St. Kliment Ohridski", Bitola, R. Macedonia. She has completed postdoc in Artificial Intelligence at the Laboratory of Informatics, Robotics and Microelectronics at the University of Montpellier, Montpellier, France. She worked as a Research assistant at the Faculty of Applied Science, University of Toronto, Canada. She also, worked at research team for Constraints, Learning and Agents at LIRMM, University of Montpellier. Currently, she works as a Full Professor in Artificial Intelligence and Systems, Computer Science and Computer Engineering at the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola and serves as a Vice-dean for Science and Collaboration. Her research work is focused on artificial intelligence, machine learning techniques, intelligent systems and human - computer interaction. She participated in several international and domestic scientific projects. She has published numerous scientific papers in the area of interest, as well as several monographic items. She is a reviewing referee for well-known publishing house, journals with significant impact factor in science and also, member of editorial board of several international conferences.

Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (co – chair)

Željko Stojanov, Ph.D. received PhD degree in Computer science and applied informatics at University of Novi Sad, Serbia. He works as a full professor at University of Novi Sad, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia. His research interests are in the fields of software engineering, software architecture, software life cycle, business informatics, learning and knowledge management, engineering education, and human aspects of software engineering. He is author of scientific papers published in refereed journals and in the proceedings of international conferences. He participated in several research and industrial projects at national and international levels. He has over fifteen years of experience working with small software companies as a consultant in the fields of software development, software maintenance and software process improvement.

Organizing Committee

Chairs

Kostandina Veljanovska (President), Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (vice-president)

Members

Blagoj Ristevski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Višnja Ognjenović, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Eleonora Brtka, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dalibor Dobrilovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dragica Radosav, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Biljana Radulovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Božidar Milenkovski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Snežana Savoska, Faculty of Information and Communication Technologies - Bitola,N. Macedonia
Vladimir Brtka, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Zoltan Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Siniša Mihajlović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Velibor Premčevski, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nikola Rendevski , Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Aleksandra Stojkov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Maja Gaborov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Milica Mazalica, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Igor Vecštejn, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Marko Blažić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vuk Amižić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Natasa Blazeska-Tabakovska, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Jovana Borovina, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dalibor Šeljmeši, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Mimoza Bogdanoska Jovanovska, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Vladimir Šinik, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nadežda Ljubojev, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Marina Blažeković Toshevski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Hristina Dimova Popovska, Faculty of Information and Communication Technologies, University "St.Kliment Ohridski" - Bitola, N. Macedonia
Darko Pajkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Marija Apostoloska Kondoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Milcho Prisagjanec, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Ilche Dimovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Zoran Pavlovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Vladimir Karuović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Evgeny Cherkashin, Institute of System Dynamic and Control Theory SB RAS, Russia
Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Russia
Filip Tsvetanov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria

Program Committee

Blagoj Ristevski (president), Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia,

Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (vice-president)

Kostandina Veljanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia (vice-president)

Eleonora Brtka, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Višnja Ognjenović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Dalibor Dobrilović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Dragica Radosav, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Dragana Glušac, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Borislav Odadžić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Miodrag Ivković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Biljana Radulović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Vladimir Brtka, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Zoltan Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Jelena Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Vesna Makitan, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Nadežda Ljubojev, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Vladimir Šinik, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Igor Nedelkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Aleksandar Markoski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Violeta Manevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Pece Mitrevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Ilija Jolevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Dragan Grueski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Monika Markovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Snežana Savoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Sonja Mančevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Mimoza Bogdanoska Jovanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Nataša Blažeska Tabakovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Božidar Milenkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Zoran Kotevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Nikola Rendevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski",

Bitola, North Macedonia
Andrijana Bocevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Tome Dimovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Marina Blažeković Toševski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Lela Ivanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Ilija Hristoski, Faculty of Economics - Prilep, North Macedonia
Elena Vlahu-Gjorgievska, Univerisity of Wollongong, Australia
Mimoza Mijoska, Internatonal Slavic University GAVRILA ROMANOVICH DERZHAVIN, Faculty of Technical Sciences and Informatics
Blagoj Nenovski, University "St. Kliment Ohridski", Bitola, North Macedonia
Nora Pireci Sejdiu, University "St. Kliment Ohridski", Bitola, North Macedonia
Saso Nikolovski, AUE University, Faculty of Informatics-Skopje, North Macedonia
Aybeyan Selim, International Vision University, Gostivar, North Macedonia
İlker Ali, International Vision University, Gostivar, North Macedonia
Fehmi Skender, International Vision University, Gostivar, North Macedonia
Ming Chen, Zhejiang University, China
Alexander Feoktistov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Alexander Yurin, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Igor Bychkov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Andrey Gachenko, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Andrey Mikhailov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Alexey Daneev, Irkutsk State Transport University, Irkutsk, Russia
Denis Sidorov, Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Viacheslav Paramonov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Andrey Dorofeev, Institute of High Technologies, Irkutsk National Research Technical University, Irkutsk, Russia
Gogolák László, Subotica Tech - College of Applied Sciences, Subotica, Serbia
Zlatko Čović, Subotica Tech - College of Applied Sciences, Department of Informatics, Subotica, Serbia
Zora Konjović, University Singidunum, Centar Novi Sad, Serbia
Siniša Nešković, Faculty of organizational sciences, University of Belgrade, Serbia
Nataša Gospić, Faculty of transport and traffic engineering, Belgrade, Serbia
Branko Markoski, Faculty of technical Sciences, Novi Sad, Serbia
Željen Trpovski, Faculty of technical Sciences, Novi Sad, Serbia
Branimir Đorđević, Megatrend University, Belgrade, Serbia
Slobodan Jovanović, Faculty of Information Technology, Belgrade, Serbia
Željko Eremić, College of Technical Sciences - Zrenjanin, Serbia
Rajnai Zoltán, Obuda University, Budapest, Hungary
Tünde Anna Kovács, PhD, Óbuda University, Hungary
Zoltán Nyikes, PhD, Milton Friedman University, Hungary
Mirjana Pejic Bach, University of Zagreb, Croatia
Androklis Mavridis, Aristotel University of Thessaloniki, Greece

CONTENT

Optimizing Real-Time Data Processing with Kafka and Databricks Integration for Scalable Machine Learning Solutions

Aneta Trajkovska[1], Blagoj Ristevski[2], Kostandina Veljanovska[3], Trajche Trajkov[4], Nikola Rendevski[5]

[1, 2, 3, 4, 5] *University "St. Kliment Ohridski" – Bitola, Faculty of Information and Communication Technologies - Bitola, ul. Partizanska bb, Bitola, Republic of Macedonia*

[1]aneta.trajkovska@uklo.edu.mk [2]blagoj.ristevski@uklo.edu.mk [3]kostandina.veljanovska@uklo.edu.mk [4]trajkov.trajche@uklo.edu.mk [5]nikola.rendevski@uklo.edu.mk

Abstract:
Integrating Apache Kafka with Databricks enables seamless streaming, analytics, and model deployment, bridging the gap between raw data and actionable insights. Recent trends focus on optimizing performance, reducing latency, and enhancing scalability, allowing organizations to build intelligent, responsive systems. Advances such as event-driven architectures, real-time feature engineering, and automated model orchestration are driving new levels of efficiency in data processing pipelines. Moreover, cloud-native deployments and containerized environments are making these integrations more flexible and resilient, supporting diverse workloads across industries. This paper examines these advancements and their implications for efficient, large-scale machine learning solutions, highlighting both current best practices and emerging innovations that can transform how organizations leverage their data assets.

Keywords:
Scalable machine learning, Apache Kafka, Azure Databricks, real-time data processing.

1. Introduction

Leveraging powerful tools like Apache Kafka and Databricks allows enterprises not only to handle high-volume data efficiently but also to derive actionable insights through advanced analytics and machine learning. The integration of these platforms is reshaping how data pipelines are designed, enabling greater flexibility, scalability, and responsiveness. As data volumes continue to grow, real-time processing and scalable machine learning have become critical for modern organizations.

Data processing is rapidly evolving to meet the demands of real-time analytics and large-scale machine learning [1]. Key trends include the adoption of streaming architectures, which allow continuous ingestion and analysis of data as it is generated, and the integration of artificial intelligence and machine learning for automated insights and anomaly detection. Cloud-native and serverless technologies are increasing scalability and flexibility, while distributed computing frameworks are optimizing performance and resource utilization. Additionally, organizations are focusing on end-to-end pipeline automation, data governance, and low-latency processing to ensure reliable, actionable insights across diverse business applications [2], [3].

The remainder of this paper is structured as follows: Section 2 discusses enhancements in Kafka and Databricks integration. The trends utilized for data processing in real-time are described in Section 3. Section 4 describes optimizing system architecture, while Section 5 presents the implementation of various machine learning solutions. Section 6 provides analytical results and performance evaluation. Finally, Section 7 concludes the paper and provides aspects for future research and development.

2. Kafka and Databricks integration

Optimizing the integration between Apache Kafka and Databricks is critical for building scalable and efficient data processing pipelines capable of supporting real-time analytics and machine learning workloads. Kafka, with its distributed streaming architecture, provides a reliable mechanism for ingesting high-throughput data streams, while Databricks offers a unified analytics platform for processing, analyzing, and operationalizing this data, as shown in Figure 1. Effective integration requires careful consideration of data serialization, partitioning strategies, and throughput optimization to minimize latency and ensure consistent data flow [4], [5]. Additionally, leveraging features such as structured streaming, delta lake storage, and parallelized processing within Databricks can enhance performance and fault tolerance [6]. Recent research and industry practices emphasize the importance of automated pipeline orchestration, real-time monitoring, and scalable infrastructure to fully exploit the capabilities of both platforms. By optimizing this integration, organizations can accelerate the deployment of machine learning models, improve predictive accuracy, and enable more responsive, data-driven decision-making [7].

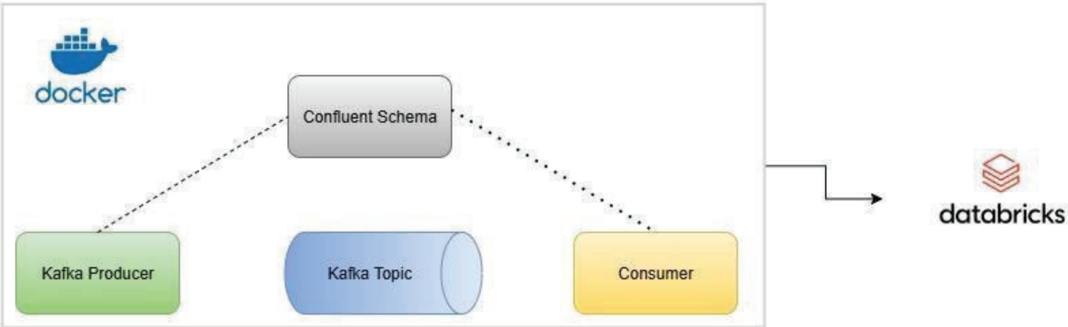


Figure 1: High-level integration of Kafka with the Databricks platform

3. Trends in Real-Time Data Processing

The field of real-time data processing has experienced significant advancements in recent years, driven by the growing demand for instantaneous insights from large-scale and continuously generated datasets. Modern architectures emphasize low-latency streaming, enabling organizations to process and analyze data as it is produced, rather than relying solely on batch processing. Key trends include the adoption of distributed streaming platforms such as Apache Kafka, which provide fault-tolerant and scalable mechanisms for data ingestion, and the integration of cloud-native and serverless technologies that enhance flexibility and resource efficiency. Additionally, real-time feature engineering, automated data pipelines, and machine learning model deployment are increasingly incorporated directly into streaming workflows, allowing for immediate predictive analytics and anomaly detection. These developments not only improve the responsiveness and intelligence of data-driven systems but also enable organizations to handle complex workloads with greater reliability, scalability, and operational efficiency [8]. Real-time data analytics processes involve continuous ingestion, processing, and analysis of data as it is generated, enabling organizations to make instant, data-driven decisions, as shown in Figure 2.

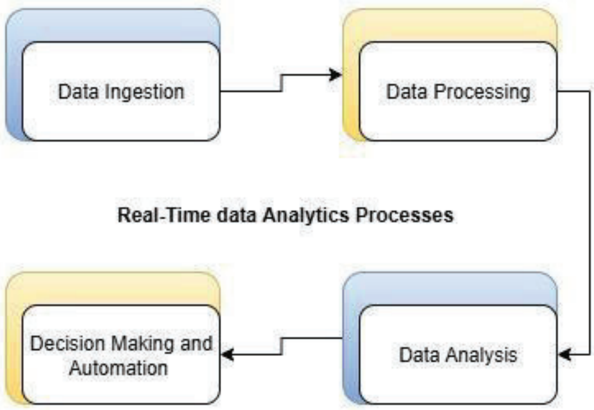


Figure 2: Cycle of the Real-Time analytics process

4. Optimizing System Architecture

The system architecture designed for this study leverages a combination of distributed streaming, cloud-based analytics, and scalable machine learning frameworks to ensure efficient data processing and analysis, which is an extension of our previous work [9]. At its core, Apache Kafka serves as the primary data ingestion and streaming platform, enabling reliable, high-throughput handling of real-time data streams. Databricks is employed as the unified analytics and processing environment, supporting both batch and streaming workloads through its integration with Apache Spark and Delta Lake for optimized storage and query performance [10]. The architecture is further enhanced with containerized deployment strategies and cloud-native orchestration tools to provide flexibility, fault tolerance, and scalability across diverse workloads. The infrastructure as code is deployed as shown in Figure 3. Additional technologies, such as structured streaming, automated pipeline scheduling, and monitoring frameworks, are incorporated to ensure seamless data flow, low latency, and efficient resource utilization. By combining these components, the system architecture provides a robust foundation for implementing scalable machine learning solutions and performing real-time analytics on large and heterogeneous datasets.

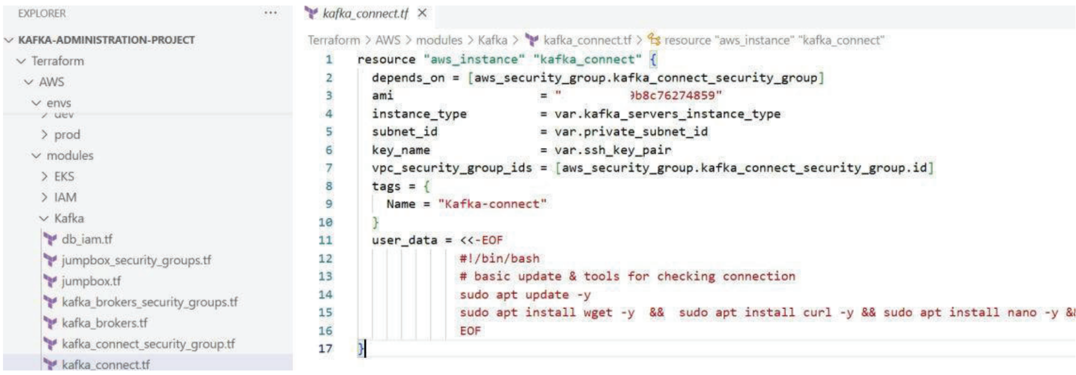


Figure 3. Infrastructure scripts for deploying resources

The solution is highly scalable, providing flexibility to easily add new resources or remove existing ones as requirements evolve. This adaptability ensures that the system can efficiently handle growth, changes in workload, or shifts business priorities without significant rework or downtime. Furthermore, it introduces a clear and well-structured approach to managing the codebase, which simplifies future maintenance and enhancements.

5. Implementation of Machine Learning Solutions in Databricks

The implementation of machine learning solutions within Databricks leverages its unified analytics platform to streamline the entire data-to-insight workflow. Databricks provides an environment that integrates data engineering, feature engineering, model development, training, and deployment, enabling scalable and collaborative machine learning pipelines [11]. By utilizing Apache Spark's distributed computing capabilities, large volumes of structured and unstructured data can be processed efficiently, facilitating both batch and real-time analytics. The platform supports a wide range of machine learning frameworks and libraries, such as MLlib, TensorFlow, and PyTorch, allowing the development of predictive models tailored to specific business needs. Additionally, features like Delta Lake ensure data consistency and reliability, while MLflow provides robust experiment tracking, model versioning, and deployment management. This integration of tools and frameworks within Databricks not only accelerates the development of machine learning models but also ensures reproducibility, scalability, and seamless operationalization in production environments, making it a comprehensive solution for advanced analytics.

We utilized a dataset originating from the automotive domain, encompassing features such as car_name, company_name, car_price, discount, and various customer and company location attributes [14]. Prior to implementing any machine learning techniques, the dataset underwent a comprehensive validation process to ensure data quality, consistency, and completeness. This included verifying the accuracy of numerical variables, examining categorical attributes for missing or inconsistent values, and performing initial exploratory analysis to detect potential outliers or anomalies. Only after this rigorous validation step was completed did we proceed with the development and application of machine learning models, ensuring the reliability and robustness of subsequent analytical results.

Firstly, a linear regression model was employed to analyze the relationship between the selected predictor variables and the target variable. Linear regression, as a widely used statistical method, estimates the dependent variable as a linear combination of independent variables, allowing for the quantification of the influence of each predictor [12].

```
Model Coefficients: [ 9.94489579e-01 -1.06947871e+02 -1.16760661e-01 -8.22234880e-01
 -1.17586645e+00 -4.79181110e-02]
Model Intercept: 183.7967266474443
```

Figure 4. Linear regression model results

These results indicate that, holding all other variables constant, the target variable increases by approximately 0.995 units for each one-unit increase in the first predictor, while the remaining predictors contribute negative adjustments of varying magnitude, as shown in Figure 4. The intercept represents the expected value of the target variable when all predictors are equal to zero.

The visualization of the linear regression coefficients provides a clear representation of the relative impact of each feature on the prediction of discounted car prices, as shown in Figure 5. The results reveal that discount and car price are the most dominant factors influencing the target variable. The negative coefficient of discount -106.94 suggests that higher discounts lead to a substantial reduction in the final price, which aligns with business expectations. On the other hand, the car price exhibits a strong positive coefficient of 0.99, indicating a nearly direct proportional relationship with the discounted price. From a methodological perspective, these findings highlight the need for improved feature engineering techniques to better represent categorical data. Additionally, the dominance of numerical features suggests that future research should explore non-linear models (e.g., Random Forest or Gradient Boosted Trees) to capture potential complex interactions between features.

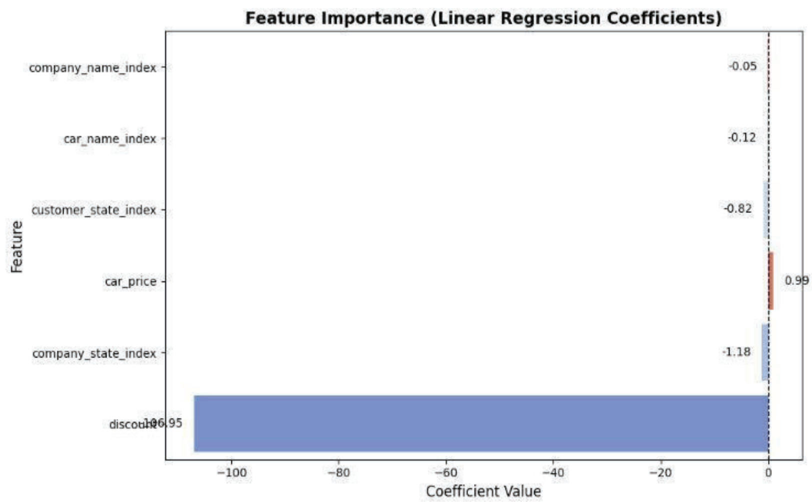


Figure 5. Visualization of feature importance based on the coefficients

To evaluate the linear relationships among variables relevant to car pricing, a correlation heatmap was constructed. This visualization illustrates the correlation coefficients between key features, including car price, discount, company, and customer identifiers, and the discounted price. The heatmap serves as a diagnostic tool to inform feature selection and assess multicollinearity in the context of a linear regression model, as shown in Figure 6. The variable car_price exhibits a strong positive correlation with discounted_price r = 0.73, indicating that the discounted price is largely dependent on the original car price. This relationship is expected, given that the discounted price is typically derived by subtracting the discount from the base price. Conversely, discount shows a weak negative correlation with discounted_price r = –0.14, suggesting that the magnitude of the discount does not linearly scale with the final price, potentially due to pricing strategies or categorical discounting mechanisms. Among categorical indices, company_name_index and company_state_index demonstrate a moderate positive correlation (r = 0.45), implying a non-negligible association between company identity and geographic location. However, other categorical variables, such as car_name_index, customer_state_index, and company_name_index, show weak or negligible correlations with pricing variables, indicating limited predictive value in a linear framework.

The diagonal of the heatmap, representing self-correlations r = 1.0, confirms the integrity of the matrix. Overall, the heatmap highlights that while some variables are strongly linearly related and suitable for inclusion in a regression model, others may contribute noise or redundancy. These insights are critical for optimizing model performance and ensuring interpretability in predictive analytics.

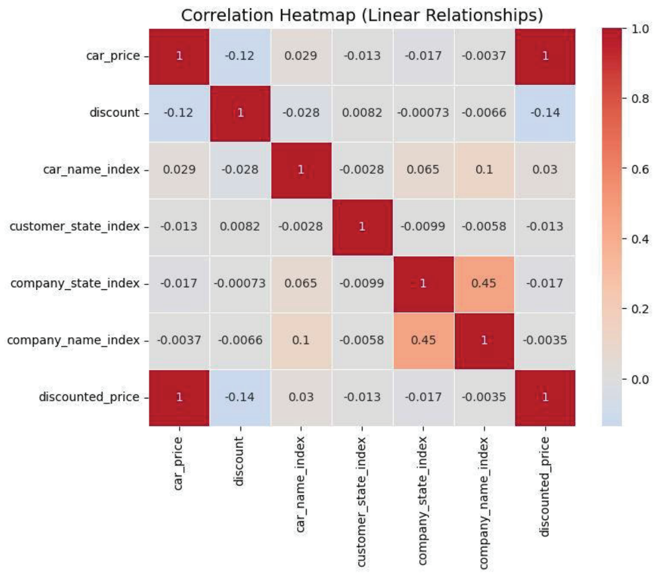


Figure 6. Correlation Heatmap while using linear regression

The Random Forest regression model was implemented to predict the discounted_price of vehicles using features such as car_price, discount, car_name_index, etc [13]. After training the model, feature importance was computed to understand the contribution of each variable to the predictive task. The model identified car_price as the most influential predictor, which aligns with domain expectations since the base price directly impacts the discounted price. Other features, such as company_name_index and customer_state_index, exhibited significantly lower importance, indicating their comparatively smaller role in determining pricing outcomes.

To evaluate the model's performance, metrics such as Root Mean Square Error (RMSE) and R-squared (R^2) were calculated. A low RMSE value suggests that the model can predict discounted prices with reasonable accuracy. The R^2 score indicated that a substantial proportion of variance in the target variable (discounted price) was explained by the model, validating its predictive effectiveness.

The scatter plot indicates that the model performs adequately in predicting lower car prices but exhibits reduced accuracy for high-value cases, presented in Figure 7. Most of the data points are concentrated near the origin, whereas outliers with actual prices reaching up to 800,000 are systematically underpredicted, with predictions plateauing around 600,000. This pattern suggests that the model may be underfitting extreme values, potentially due to their scarcity in the dataset or insufficient representation in the feature set. Enhancing feature engineering or applying a logarithmic transformation to the target variable could mitigate this issue and improve predictive performance for high-priced vehicles.

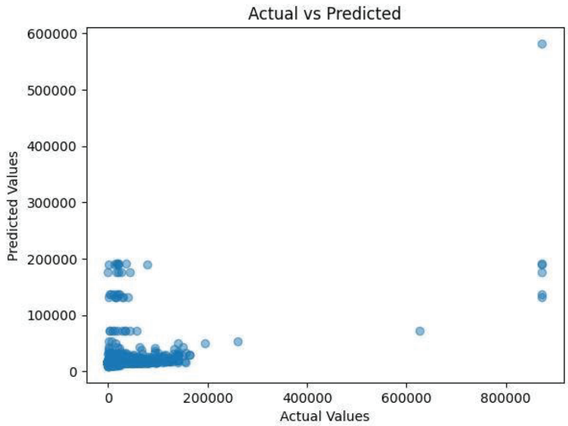


Figure 7. Model evaluation through an actual or predicted scatter plot

6. Performance Evaluation

Performance evaluation is essential for assessing the efficiency and effectiveness of integrated data processing and machine learning systems [13]. Key metrics include throughput, latency, scalability, and model accuracy, which provide insights into system responsiveness and reliability. In Kafka- Databricks pipelines, evaluating data ingestion rates, processing times, and machine learning outcomes helps identify bottlenecks and optimization opportunities. Rigorous performance assessment ensures that the system can deliver real-time, actionable insights while supporting scalable and robust analytics workflows. Below is explained how the performances were improved after the implementation of the solution:

Reduced latency: real-time data ingestion and processing through the optimized Kafka–Databricks pipeline significantly decreases end-to-end latency, enabling faster availability of actionable insights.

Improved throughput: the integrated solution efficiently handles higher data volumes, ensuring consistent processing rates and preventing bottlenecks even under peak load conditions.

Enhanced model accuracy: By providing timely and clean streaming data to machine learning models, predictions are more accurate and reflect the latest trends in the data.

Scalability: the architecture supports horizontal scaling, allowing the system to accommodate growing data streams and expanding computational requirements without degradation in performance.

Robust fault tolerance: Kafka’s distributed messaging combined with Databricks’ resilient processing framework improves system reliability, minimizing data loss and ensuring continuous operation.

Comprehensive monitoring and evaluation: integration enables end-to-end observability of data flow and model performance, allowing precise performance evaluations and informed decision-making for further optimization, as shown in Figure 8.

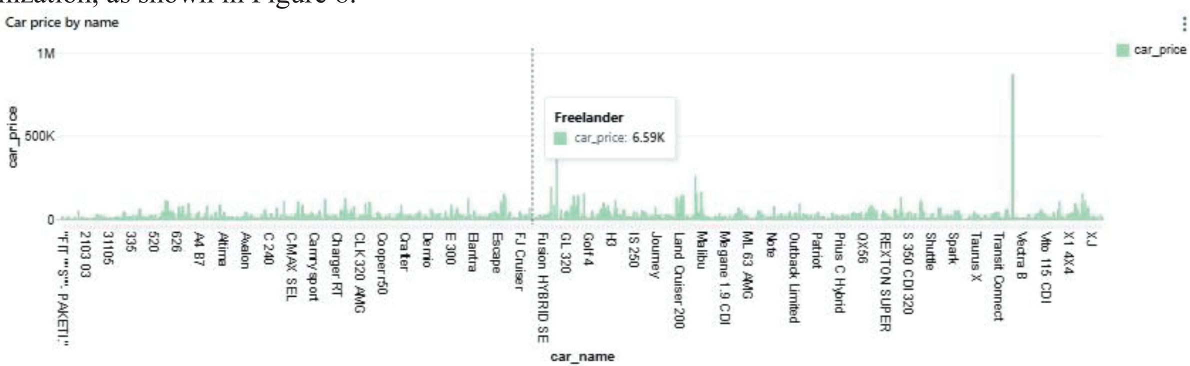


Figure 8. Visualization of the data on Databricks dashboard

7. Conclusions

In conclusion, this paper synthesizes the key findings regarding the integration of Kafka and Databricks for scalable machine learning solutions, highlighting the performance improvements, architectural considerations, and practical implementation strategies identified throughout the study. Furthermore, it outlines several directions for future research and development, including the exploration of advanced optimization techniques, real-time feature engineering, automated model orchestration, and the integration of emerging cloud-native and edge computing technologies. By addressing these areas, future work can further enhance the efficiency, scalability, and robustness of data processing and machine learning systems, supporting more intelligent, responsive, and data-driven decision-making across diverse applications and industries. Moreover, continuous experimentation with emerging stream processing technologies can further amplify performance, scalability, and model accuracy.

References:

[1] Z.Zheng, P. Wang, J. Liu and S. Sun. Real-Time Big Data Processing Framework: Challenges and Solutions. Applied Mathematics & Information Sciences, International Journal, Sci 9, No.6 , 3169-3190 (2015). http://dx.doi.org/10.12785/amis/090646

[2] W. Yang, X. Liu, L. Zhang and L. T. Yang, "Big Data Real-Time Processing Based on Storm," 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Melbourne, VIC, Australia, 2013, pp. 1784-1787, doi: 10.1109/TrustCom.2013.247.

[3] Zheng, T., Chen, G., Wang, X. et al. Real-time intelligent big data processing: technology, platform, and applications. Sci. China Inf. Sci. 62, 82101 (2019). https://doi.org/10.1007/s11432-018-9834-8

[4] S. Vyas, R. K. Tyagi, C. Jain and S. Sahu, "Literature Review : A Comparative Study of Real Time Streaming Technologies and Apache Kafka," 2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, 2021, pp. 146-153, doi: 10.1109/CCICT53244.2021.00038.

[5] B. R. Hiraman, C. Viresh M. and K. Abhijeet C., "A Study of Apache Kafka in Big Data Stream Processing," 2018 International Conference on Information , Communication, Engineering and Technology (ICICET), Pune, India, 2018, pp. 1-3, doi: 10.1109/ICICET.2018.8533771.

[6] Etaati, L. (2019). Azure Databricks. In: Machine Learning with Microsoft Technologies. Apress, Berkeley,
CA. https://doi.org/10.1007/978-1-4842-3658-1_10

[7] A. Amruth, R. Ramanan, R. Paul, C. Vimal and B. M. Beena, "Cloud Based Big Data Solution for Cancer Classification: Using Databricks on Large Scale Genomic Data," 2024 1st International Conference on Communications and Computer Science (InCCCS), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/InCCCS60947.2024.10593322.

[8] Shubhodip Sasmal. Real-time Data Processing with Machine Learning Algorithms. International Research Journal of Engineering and Applied Sciences, Vol. 11, Issue 4, (2023). Doi: 10.55083/irjeas.2023.v11i04012

[9] A. Trajkovska, B. Ristevski, T. Trajkov, N. Rendevski and K. Veljanovska, "From Data to Decisions: Real- Time Analytics and ML with Kafka and Databricks," *2025 60th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, Ohrid, North Macedonia, 2025, pp. 1-4, doi: 10.1109/ICEST66328.2025.11098429.

[10] Ravi Shankar Koppula. Implementing Data Lakes with Databricks for Advanced Analytics. North American Journal of Engineering Research, Vol. 3 No. 2 (2022), https://najer.org/najer/article/view/33

[11] L'Esteve, R. (2022). Databricks. In: The Azure Data Lakehouse Toolkit. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-8233-5_3

[12] James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J. (2023). Linear Regression. In: An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-031- 38747-0_3

[13] Henry Lucas. 1971. Performance Evaluation and Monitoring. ACM Comput. Surv. 3, 3 (Sept. 1971), 79–
91. https://doi.org/10.1145/356589.356590
[14] https://www.kaggle.com/datasets/abdullahkhanuet22/olx-cars-dataset