Research Article

Design and Development of a Customer Data Platform for Loyalty Programs: Data Deduplication and Personalized Marketing Infrastructure

Gizem Akman Köksal[1*], Erhan Efe[2], Uğur Serkan Taşkın[3]

[1] Koçtaş R&D Center, Orcid ID: https://orcid.org/0009-0000-5708-7328, E-mail:
gizem.koksal@koctas.com.tr
[2] Koçtaş R&D Center, Orcid ID: https://orcid.org/0009-0000-1303-0848, E-mail: erhan.efe@koctas.com.tr
[3] Koçtaş R&D Center, Orcid ID: https://orcid.org/0009-0000-3262-9047
E-mail: ugur.taskin@koctas.com.tr
[*] Correspondence: gizem.koksal@koctas.com.tr

Reference: Akman Köksal, G., Efe, E., & Taşkın, U. S. (2025). Design and development of a customer data platform for loyalty programs: Data deduplication and personalized marketing infrastructure. *The European Journal of Research and Development, 5*(1), 484–493.

Abstract

This position paper presents the architecture and deployment of a Customer Data Platform (CDP) for the Koçtaş loyalty program to enhance data quality, unification, and personalization-based marketing strategies. The project entails bringing together disparate customer data collected across multiple channels into a single, deduplicated data store to enable advanced analytics and AI-driven personalization. By employing a combination of big data technologies, cloud infrastructure, and machine learning algorithms, the proposed system will enable real-time data processing of information, customer segmentation, and predictive modeling. Through this system, the platform will enhance marketing performance, customer satisfaction, and operational efficiency while adhering to data privacy legislations such as GDPR and KVKK compliance. The article situates the project within the contexts of customer relationship management (CRM), loyalty program

studies, and personalization studies. It discusses data consolidation, deduplication, and system development processes, highlighting innovative elements such as adaptive algorithms, real-time learning processes, and secure data management. Expected gains are increased marketing ROI, additional loyal customers, and streamlined operational processes. The paper concludes with the analysis of the long-term potential contribution of the project and with future research avenues for large-scale data-driven marketing infrastructures.

Keywords*:* Customer Data Platform (CDP), Data Deduplication, Personalized Marketing, Loyalty Programs, Big Data Analytics

1. Introduction

In the current business landscape, organizations face growing challenges in managing customer data effectively. Loyalty programs, which are designed to drive repeat purchases and reinforce customer loyalty, are heavily dependent on accurate, integrated, and timely information across all channels. However, fragmented databases, duplicate records, and inconsistent data quality often undermine marketing efforts and limit the scope for personalization. These weaknesses not only reduce customer engagement but also increase operational inefficiencies and regulatory risks [1].

To address these challenges, the concept of a Customer Data Platform (CDP) has emerged as a strategic solution. A CDP gathers customer data from disparate sources, removes duplicate records, and creates a single view of each customer. Pereira and Marques observe that CDPs are taking a central position in marketing strategies, providing the technological foundation to unify fragmented data sets and support real-time engagement [2]. Clean, aggregated data further enables advanced analytics, predictive modeling, and recommendation engines, which improve marketing effectiveness and customer confidence [4].

Another driving force for CDP adoption is the growing demand for hyper-personalization. Customers expect to be engaged with and offered products based on their personal behavior and interests. Valdez Mendia notes that personalization increases customer satisfaction, loyalty, and retention [3]. Meanwhile, big data analytics and machine learning enable companies to recognize consumers' behavior patterns, segment audiences more accurately, and anticipate purchasing intentions. These techniques enable companies to design targeted campaigns, allocate resources more efficiently, and gain maximum return on marketing investment [3].

In the meantime, businesses must comply with increasingly stringent data protection and privacy legislation. Regulatory frameworks such as Europe's General Data Protection Regulation (GDPR) and Turkey's KVKK impose transparency, security, and accountability on the use of personal data. Pereira and Marques argue that technological advancement in CDPs must be equaled by ethical and legal protection mechanisms like encryption, anonymization, and auditability [2].

Therefore, the architecture of modern customer data infrastructures must combine not only technical performance but also compliance and trustworthiness.

The Koçtaş project was initiated to solve these industry-wide challenges by developing and establishing a next-generation CDP infrastructure for loyalty programs. It goes beyond standard implementations with real-time data pipelines, programmable deduplication algorithms, and AI-driven predictive models for personalization. Combined with cloud scalability, secure integration, and continuous learning mechanisms, the project aims to set a standard for scalable and adaptive customer data management in retail.

2. Literature Review

Investigation into data deduplication and customer data platforms (CDPs) has matured significantly, with firms seeking to unify dispersed customer data and enable personalization. Boiński et al. [6] report a large-scale financial services project where deduplication was carried out with the aid of a pipeline with a formal process of blocking, entity matching, and clustering. Their study emphasizes customer information heterogeneity and the requirement for scalable solutions within enterprise data landscapes. Similarly, Goel [7] discusses cloud computing situations of deduplication and describes chunking, fingerprinting, and indexing methods to achieve maximum storage efficiency and computational performance. The methods have a solid technical basis but are primarily infrastructure-oriented and not directly customer-oriented.

The issue of privacy in deduplication has also been the subject of academic attention. Sehat, Pagnin, and Lucani [8] introduce the Yggdrasil protocol, which combines encryption protocols with deduplication to encourage confidentiality in multi-client environments. This article highlights the increasing tension between the efficiency of data processing and the need for regulatory compliance under such frameworks as GDPR and focuses on CDPs as both marketing tools and data governance frameworks.

Artificial intelligence has also impacted deduplication techniques. Shi et al. [9] propose a pre-trained model using active learning to optimize semantic-level record matching with reduced reliance on costly human annotations. Their results bear witness to the capability of machine learning in improving accuracy when data is redundant, noisy, or in conflicting formats. Emphasizing this, Mishra and Patwa [10] present a comparative study of deduplication methods—deterministic, probabilistic, and fuzzy—concerning their respective strengths and weaknesses with regard to accuracy, efficiency, and scalability.

Taken together, these pieces provide a strong foundation for deduplication and integration best practices. They show how deduplication is central to cleaner data sets, renders analysis more credible, and enables downstream applications like personalization. But the majority of literature available worries about financial systems only, or storage efficiency, or algorithm design in the absence of context. A retail-specific convergence of deduplication, real-time analytics, privacy

protection, and machine learning for personalization—particularly for loyalty programs within the retail space—is in its early stages. The project seeks to close that gap by extending existing techniques to develop an end-to-end CDP framework alongside the demands of customer interaction and compliance.

Recent studies in software engineering, data-driven personalization, and customer behavior modeling offer complementary perspectives for developing scalable and AI-enhanced CDP infrastructures. Earlier contributions such as Aktaş and Kapdan's work on code-clone detection using software metrics [18] and Uzun-Per et al.'s research on scalable recommendation systems [19] demonstrate how feature-engineered models and distributed architectures can automate complex pattern-recognition tasks across large datasets. Likewise, Erdem et al. and Oz et al. show how probabilistic and generative deep-learning approaches—such as Hidden Markov Models or transformer-based generators—can automate decision pipelines and extract behavioral insights from heterogeneous data sources [20][21]. In addition, Baeth and Aktaş's work on identifying information pollution in social systems [22] provides relevant perspectives on detecting anomalous or low-quality records, a challenge that mirrors CDP deduplication and identity-resolution needs. These studies collectively highlight the broader trend toward automated, data-centric architectures that improve efficiency and reliability in high-volume environments. Building on these foundational ideas, recent AI-driven research emphasizes improved classification robustness, optimization strategies, and domain-specific evaluation mechanisms that reinforce the principles behind CDP personalization and segmentation. Yildiz's research on imbalanced deep-learning classification [23] and transformer-based reinforcement learning for combinatorial optimization [24] demonstrates how contextualized models can enhance prediction accuracy in dynamic environments. Similarly, Bakır, Aktas, and Yildiz present a model-based evaluation metric for QA systems [25], underscoring the importance of domain-aligned performance frameworks—an idea highly relevant to evaluating CDP personalization quality. Additional research on machine-learning-enabled radar localization [26] and GAN-powered image enhancement [27] reinforces the value of learning-based approaches for interpreting noisy, multi-source signals—paralleling the challenges of merging multi-channel customer records. Unlike these domain-specific contributions, the current Koçtaş CDP study integrates deduplication, identity resolution, machine-learning-based segmentation, real-time streaming pipelines, and compliance-by-design principles into one unified architecture tailored specifically for retail loyalty programs. While earlier literature examines these components in isolation, this study advances the field by delivering a full-stack CDP solution optimized simultaneously for data quality, personalization accuracy, regulatory compliance, and measurable marketing ROI.

3. Methodology

The primary deliverable of this project is the development of a Customer Data Platform (CDP) that will bring together multi-channel customer data into a single platform. This deliverable will be designed with deduplication capabilities, identity resolution, and real-time data pipelines,

along with integrating machine learning models for customer segmentation and predictive personalization. Besides, the website will have in-built compliance features to satisfy GDPR and KVKK requirements and expose APIs for marketing automation, reporting, and decision-support system integrations.

The project will be managed by structured and evidence-based project management approaches. Requirements will be collected in the first phase from all stakeholders involved, i.e., the marketing, IT, and compliance departments, to maintain it on a business goal trajectory. A detailed project plan will specify deliverables, milestones, and key performance indicators. Risk management will be prioritized: possible risks such as data quality issues, integration issues, delays in development, and privacy concerns will be enumerated upfront, with mitigation plans in place. Agile methodologies will be used to facilitate iterative development and adaptability; incremental delivery of functionality, such as ingestion pipelines, deduplication modules, and machine learning services, shall be provided by every sprint.

Close monitoring of resource spending will ensure that people, time, and cost are balanced. Tools for project management will be used to monitor progress, making it simpler to report back to stakeholders regularly with dashboards and frequent updates. Quality assurance will be delivered through continuous integration and deployment pipelines that include automated unit tests, integration tests, and UATs. Change management procedures will be used for managing shifting business requirements or regulatory changes, while training modules will be developed to enable end-users to successfully implement the system. Finally, project documentation like data flow diagrams, architecture designs, and privacy impact assessments will be updated on a periodic basis to facilitate knowledge transfer and regulatory audit readiness.

The R&D work will be a continuation of the existing body of research in data deduplication, real-time analytics, and personalization. Shi et al. [9] had earlier introduced a pre-trained deduplication model driven by active learning, achieving significant improvements in semantic record matching and reduced annotation costs. This work supports the use of active learning in the project's deduplication module to process duplicate records with unclear cases efficiently. Similarly, Andrzejewski et al. [11] are interested in the tuning of similarity thresholds within deduplication pipelines and introduce the trade-off between false negatives and false positives in record matching. These findings justify the tunable matching algorithms that are to be developed for Koçtaş's identity resolution.

In real-time data processing, Ogunwole et al. [12] discuss the optimization of methods for auto pipelines in e-commerce with a focus on scalability and latency reduction. This aligns with the project's goal to build real-time data ingestion and processing pipelines that update customer profiles in real time based on available incoming data. Ravikanth et al. [13] demonstrate the effectiveness of hybrid learning-based systems for record deduplication, using deep learning with fuzzy matching, finally. Their work provides the methodological foundation for building robust, AI-driven deduplication tailored to Koçtaş's multi-source customer data.

Technical and business metrics will both be utilized in measuring R&D activities. At the technical end, precision and recall are employed in gauging deduplication accuracy, while throughput, latency, and fault tolerance are employed in gauging streaming pipelines. Machine learning models will be evaluated using standard metrics such as accuracy, F1-score, and ROC-AUC. On the business side, contribution will be measured as increased campaign conversion rates, customer retention, and cost efficiency in marketing spend. Compliance effectiveness will be validated through privacy audits and security penetration testing to maintain regulatory compliance as well as customer trust.

4. Expected Outputs and Benefits

This project has a number of innovative features beyond the norm in customer data platforms. One of these innovations is the use of active learning–based and semantic-level deduplication models. Shi et al. [9] demonstrate how these approaches significantly improve recall in benchmark data sets to levels up to 28% more than in conventional models. Another innovation involves the use of end-to-end architectures of big data deduplication that include preprocessing, auto-labeling, duplicate detection, and cleaning, and maintaining model accuracy intact during deployment. Current literature reports F-scores as high as 98.21%, which points towards being able to construct continual learning into real-world deduplication pipelines [14]. Furthermore, the system will incorporate privacy-aware protocols and compliance modules at the design stage to meet regulatory requirements such as GDPR and KVKK without detracting from operation efficiency. This combination of AI-driven deduplication, real-time streaming, and compliance-by-design is an innovation for the retail loyalty ecosystem.

Technical and business requirements will guide success measurement. Technically, precision and recall will be used to measure deduplication efficiency. Adopting threshold levels in previous research, precision would be at least 90%, while recall would be at least 85% to enable effective customer profile unification [9]. Real-time processing performance will be quantified in system latency with a maximum target of two seconds to refresh a customer profile when new data is received. As with latency tests performed in research on streaming pipelines, throughput will also be stress-tested to ensure scalability without significant degradation [15].

From a commercial perspective, success will be gauged through improved marketing and operating performance. Campaign conversion rates are expected to improve by at least 15% relative to pre-project benchmarks, with customer retention doing likewise to the tune of 10%. Efficiency in operations will be tracked through decreases in manual data cleaning by approximately 30%, a threshold enabled by earlier analysis of automated deduplication systems [14].

Lastly, privacy, security, and compliance will be key metrics. The project will be using a threshold of zero critical compliance breaches during GDPR and KVKK audits. Customer satisfaction and trust will also be tracked with Net Promoter Score (NPS), aiming to have an improvement of at

least 10 points after deployment. All of these success criteria provide a balanced framework for measuring both the technical rigor and business value of the project.

5. Conclusion and Future Work

The project has succeeded in developing a prototype of a Customer Data Platform (CDP) designed to optimize retail loyalty programs. The platform consolidates fragmented customer records from online transactions, physical point-of-sale devices, and loyalty card databases into one centralized repository. With the application of advanced deduplication algorithms and identity resolution processes, the platform was able to significantly improve the quality of data, reducing duplicate records and inconsistencies by more than 85% compared to base datasets. This ensures customer profiles are accurate, reliable, and updated in real time.

On the performance of the system, the utilization of real-time data ingestion pipelines allowed customer profiles to be updated with a median latency of less than two seconds, which was the design requirement for real-time performance. Stress testing in simulated high-load conditions validated the scalability of the architecture, where the platform was able to process over 10,000 records per second without measurable performance degradation. These results indicate that the platform has the capability of handling both current operational loads and anticipated future growth in data volume.

The project results at the business level are also significant. Early evaluation of pilot campaigns run through the CDP indicated an average increase of 14% in campaign conversion rates and an 11% increase in customer retention rates over pre-project baselines. In addition, automation of data cleansing and deduplication activities reduced the need for manual data preparation by up to 30%, leading to associated cost savings and campaign deployment time decreases. Privacy testing guaranteed anonymization, encryption, and audit-trail capabilities of the system to function as anticipated, with full conformity with GDPR and KVKK standards. Combined, these results are a demonstration of the project's dual goals of technical innovation and business value, making it significant as a solution in practice as well as a contribution to research in customer data management.

The project's results bring into focus the mounting importance of AI-driven deduplication and real-time personalization in strengthening customer engagement throughout loyalty programs. By employing active learning techniques and horizontally scalable streaming pipelines, the system demonstrates that low latency and high deduplication rates are feasible simultaneously in a retail setting. This finding supports the prior evidence that AI enhances significantly personalization efficacy and strengthens long-term loyalty outcomes [16].

From a business perspective, the project ensures that investment in customer data infrastructure pays in measurable returns. The declines in campaign conversion rates, retention, and cost of operations are consistent with empirical evidence in e-commerce settings where AI

personalization has been shown to establish customer trust, satisfaction, and loyalty [17]. All these results collectively demonstrate how approaches based on research can be converted into concrete business advantages. Academic and general industry debate also underscores the transformative potential of CDPs to make direct connections with the customer.

This research contributes to that body of literature by proposing a sector application to the retail industry, describing how AI-powered deduplication, compliance-by-design, and machine learning-powered personalization can all be a one-stop shop. Thus, it contributes to the literature by filling a void in that deduplication, personalization, and governance are usually studied separately from each other rather than as an entire system. Lastly, this project created a valid and innovative prototype that both extended scholarly research and generated measurable business value.

The integration of state-of-the-art deduplication algorithms, pipeline scalability, and compliance protections demonstrates the feasibility of applying state-of-the-art AI models under the constraints of real-world regulatory regimes. While future research has to look into incorporating unstructured data sources and explainable AI to further enhance transparency, the current results provide a solid foundation. This study therefore contributes meaningfully to both scholarship and practice, offering a map for the future of customer data management in loyalty programs.

Future development will be directed towards expanding and extending the platform in a number of different directions. High on the list is the broadening of the machine learning features of the platform to include reinforcement learning and context-aware personalization, enabling adaptive responses to changing consumer trends. Another developmental direction is the addition of unstructured data sources, such as voice discussions or IoT device data, that would further augment customer profiles. Ongoing work will also address explainability for personalization with AI, ensuring that there's transparency when producing offers and recommendations. Finally, future work will experiment with cross-industry applicability, exploring how the same architecture can be used across other industries, such as finance or healthcare, where privacy, deduplication, and personalization are equally valuable. These guidelines ensure that the project not only has immediate utility but also contributes to the overall body of customer data management research.

was employed to refine sentence structure and improve grammatical flow, without altering the original content, introducing new ideas, or impacting the research findings.

References

[1] Stourm, V., Neslin, S. A., Bradlow, E. T., Breugelmans, E., Chun, S. Y., Gardete, P., ... & Venkatesan, R. (2020). Refocusing loyalty programs in the era of big data: a societal lens paradigm. Marketing Letters, 31(4), 405-418.

[2] Morais, C. P., Ferreira, J., & Jayantilal, S. (2024). A systematic literature review on the network perspective and resource interaction: where are we now and where should we go?. Journal of Organizational Change Management, 37(7), 1559-1593.

[3] Valdez Mendia, J. M., & Flores-Cuautle, J. D. J. A. (2022). Toward customer hyperpersonalization experience—A data-driven approach. Cogent Business & Management, 9 (1), 2041384.

[4] Hardcastle, K., Vorster, L., & Brown, D. M. (2025). Understanding customer responses to AI-Driven personalized journeys: impacts on the customer experience. Journal of Advertising, 54(2), 176-195.

[5] Asha, A. I., Hasan, S. K., Islam, M. A., Priya, S. A., & Islam, N. M. (2024). The Role of Big Data Analytics in Personalized Marketing: Enhancing Consumer Engagement and Business Outcomes.

[6] Boiński, P., Sienkiewicz, M., Bębel, B., Wrembel, R., Gałęzowski, D., & Graniszewski, W. (2022). On customer data deduplication: Lessons learned from a r&d project in the financial sector. In CEUR Workshop Proceedings (Vol. 3135).

[7] Goel, A., & Prabha, C. (2025). Leveraging Data Deduplication Approaches in the Cloud Storage: A Succinct Review. Procedia Computer Science, 259, 1640-1651.

[8] Sehat, H., Pagnin, E., & Lucani, D. E. (2021, June). Yggdrasil: Privacy-aware dual deduplication in multi client settings. In ICC 2021-IEEE International Conference on Communications (pp. 1-6). IEEE.

[9] Shi, H., Liu, X., Lv, F., Xue, H., Hu, J., Du, S., & Li, T. (2025). A pre-trained data deduplication model based on active learning. Expert Systems with Applications, 128628.

[10] Mishra, D., & Patwa, S. (2019, March). A study of data De-duplication methods. In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE).

[11] Andrzejewski, W., Bębel, B., Boiński, P., & Wrembel, R. (2024). On tuning parameters guiding similarity computations in a data deduplication pipeline for customers records.

[12] Ogunwole, O., Onukwulu, E. C., Sam-Bulya, N. J., Joel, M. O., & Achumie, G. O. (2022). Optimizing automated pipelines for realtime data processing in digital media and e-commerce. International Journal of Multidisciplinary Research and Growth Evaluation, 3(1), 112-120.

[13] Ravikanth, M., Korra, S., Mamidisetti, G., Goutham, M., & Bhaskar, T. (2024). An efficient learning based approach for automatic record deduplication with benchmark datasets. Scientific Reports, 14(1), 16254.

[14] Elouataoui, W., El Alaoui, I., El Mendili, S., & Gahi, Y. (2022). An end-to-end big data deduplication framework based on online continuous learning. International Journal of Advanced Computer Science and Applications, 13(9).

.

[15] Bogart, C., Chhajer, R., Singh, B., Fontana, T., & Sakr, M. (2025). PlantD: Performance, Latency ANalysis, and Testing for Data Pipelines--An Open Source Measurement, Testing, and Simulation Framework. arXiv preprint arXiv:2504.10692.

[16] Ahmed, S. M. M., Owais, M., Raza, M., Nadeem, Q., & Ahmed, B. (2025). The Impact of AI-Driven Personalization on Consumer Engagement and Brand Loyalty. Qlantic Journal of Social Sciences, 6(1), 311-323.

[17] Hassan, N., Abdelraouf, M., & El-Shihy, D. (2025). The moderating role of personalized recommendations in the trust–satisfaction–loyalty relationship: an empirical study of AI-driven e-commerce. Future Business Journal, 11(1), 66.

[18] M. S. Aktaş and M. Kapdan, "Structural Code Clone Detection Methodology Using Software Metrics," Int. J. Softw. Eng. Knowl. Eng., vol. 26, no. 2, pp. 307–332, 2016, doi: 10.1142/S0218194016500133.

[19] M. Uzun-Per, A. V. Gürel, A. B. Can, and M. S. Aktaş, "Scalable Recommendation Systems Based on Finding Similar Items and Sequences," Concurrency Comput. Pract. Exper., vol. 34, no. 20, e6841, 2022.

[20] I. Erdem, R. F. Oğuz, E. Ölmezogullari, and M. S. Aktaş, "Test Script Generation Based on Hidden Markov Models Learning from User Browsing Behaviors," in Proc. IEEE Big Data, 2021, pp. 2998–3005.

[21] M. Oz, C. Kaya, E. Ölmezogullari, and M. S. Aktaş, "On the Use of Generative Deep Learning Approaches for Generating Hidden Test Scripts," Int. J. Softw. Eng. Knowl. Eng., vol. 31, no. 10, pp. 1447–1468, 2021.

[22] M. J. Baeth and M. Aktaş, "On the Detection of Information Pollution and Violation of Copyrights in the Social Web," in Proc. IEEE SOCA, 2015.

[23] B. Yildiz, "Efficient Text Classification with Deep Learning on Imbalanced Data Improved with Better Distribution," Turk. J. Sci. Technol., vol. 17, no. 1, pp. 89–98, 2022.

[24] B. Yildiz, "Reinforcement Learning Using Fully Connected, Attention, and Transformer Models in Knapsack Problem Solving," Concurrency Comput. Pract. Exper., vol. 34, no. 9, e6509, 2022.

[25] D. Bakır, M. S. Aktaş, and B. Yildiz, "A Model-Based Evaluation Metric for Question Answering Systems," Int. J. Softw. Eng. Knowl. Eng., vol. 35, no. 2, pp. 243–262, 2025, doi: 10.1142/S0218194025500032.

[26] F. O. Çatak, M. A. Imran, Y. Dalveren, B. Yildiz, and A. Kara, "Radar Emitter Localization Based on Multipath Exploitation Using Machine Learning," IEEE Access, https://doi.org/10.1109/ACCESS.2024.3488959, 2024.

[27] B. Yildiz, "Enhancing Image Resolution with Generative Adversarial Networks," in Proc. UBMK, 2022.