International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com
Volume 4, Issue 4, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i4.983

Master Data Solution: Customer Master Data Management with LLMs

Ravikant Singh[1]

[1]Sr. Data Engineering Manager"

Publication Date: 2025/04/13

Abstract

Customer Master Data Management (CMDM) serves as a critical operational foundation for enterprises because it enables consistent accurate customer data management across all organizational systems. The current CMDM systems face multiple problems because they fail to sustain high data quality standards and handle expanding operations and unorganized data entries. The research develops an advanced CMDM system which uses domain-specific Large Language Models (LLMs) that receive training from organizational customer data. The system unites LLM-based intelligence with a multi-model framework to deliver superior standardization and validation and enrichment functions. The Human-in-the-Loop (HITL) governance model maintains compliance and accuracy standards through ongoing improvement processes. The system unites better data quality with intelligent customer analytics to create foundations for AI-driven business growth.

Keywords: *Customer Master Data Management (CMDM), Large Language Models (LLMs), Data Quality, Unstructured Data, Scalability, Multi-Model Architecture, Data Standardization, Data Validation, Data Enrichment, Human-in-the-Loop (HITL), Data Governance, AI-Powered Enterprise.*

I. INTRODUCTION

Organizations in the digital economy rely on growing volumes of customer data to drive operations, analytics, compliance, and personalization (Vilminko-Heikkinen & Pekkola, 2017). The discipline of Customer Master Data Management (CMDM) functions as the core of data ecosystems by establishing and sustaining unified accurate complete customer information throughout all systems and business units. Traditional CMDM systems increasingly struggle with data quality, limited scalability, and poor handling of unstructured or semi-structured data (Vilminko-Heikkinen & Pekkola, 2017).

Business expansion leads to customer data distribution across different platforms and geographical locations. Conventional rule-based CMDM solutions encounter difficulties in resolving duplicate records and data validation across sources and finding useful information from inconsistent records which demand manual intervention and substantial IT resources (Otto et al., 2011). The inefficient process hinders organizational decision-making while blocking digital transformation initiatives.

The implementation of Large Language Models (LLMs) with Artificial Intelligence brings a revolutionary chance to transform Customer Master Data Management (Li, 2024). The training of domain-specific LLMs with organizational customer datasets enables them to understand context for performing intelligent data standardization and validation and enrichment of ambiguous records. (Li, 2024). Organizations can enhance customer master record accuracy and completeness and governance through the combination of LLM capabilities with multi-model architecture and Human-in-the-Loop (HITL) oversight framework (Li, 2024).

- *Challenges in Traditional Customer Master Data Management (CMDM):*

Traditional Customer Master Data Management (CMDM) systems involved creating centralized customer information repositories which maintained consistency between business units and applications (Pansara, 2021). The growing scale of organizations along with digital transformation reveals limitations in legacy systems which damage data reliability and restrict operational flexibility. Few key challenges are mentioned below.

Singh, R. (2025). Master Data Solution: Customer Master Data Management with LLMs. *International Journal of Scientific Research and Modern Technology*, *4*(4), 64–69. https://doi.org/10.38124/ijsrmt.v4i4.983

- Data Quality Issues.
- Scalability Limitations.
- Inability to Handle Unstructured Data.
- Integration Complexity and Siloed Systems.
- High IT Dependency.

II. LLMS IN CUSTOMER MASTER DATA HANDLING

Organizations will see change when handling customer master data through Large Language Models (LLMs) (Zhao, et al., 2024). Instead of using the traditional rule-based systems, LLMs use semantic reasoning, adaptive learning and contextual understanding to data operations. LLMs designed for customer data domains improve accuracy, uniformity and data integrity of CMDM (Li, 2024).

- Contextual Data Understanding: LLMs' helps to identify the customer information accurately in context. It also helps to detect duplicates, identify anomalies and standardize entries with less manual intervention (Zhao, et al., 2024).
- Intelligent Data Standardization: LLMs are used to normalize customer data formats like phone numbers and postal codes in real-time. LLMs can predict missing information and convert records to match company-wide standards by using pattern detection and without manual programming (Li, 2024).
- Unstructured Data Processing: Traditional CMDM systems have difficulties processing unstructured data including emails, PDFs, customer service logs and social media interactions. LLMs are really good at extracting structured data from unstructured sources (Zhou, Zhao, & & Li, 2024).
- Multi-Language and Regional Adaptation: By using natural language processing, LLMs process data in multiple languages and understand regional characteristics, which benefits global organizations in managing customer data across jurisdictions. This improves compliance standards and localization precision (Zhao, et al., 2024).
- Enhanced Matching and Deduplication: LLMs' semantic similarity approach is more capable than traditional string matching for better customer record unification. An LLM system uses address, transaction history and associated accounts to determine that "Jonathan A. Smith" and "Jon Smith" are the same person (Li, 2024).
- Continuous Learning and Adaptability: Fine-tuning LLMs with data allows these models to learn and adapt continuously. This adaptive nature makes the system to learn business rules, new data types and compliance requirements (Zhou, Zhao, & & Li, 2024).

III. MULTI-MODEL ARCHITECTURE FOR MASTER DATA QUALITY

Modern CMDM solutions implement multi-model architecture to achieve high-quality consistent and actionable customer master data.

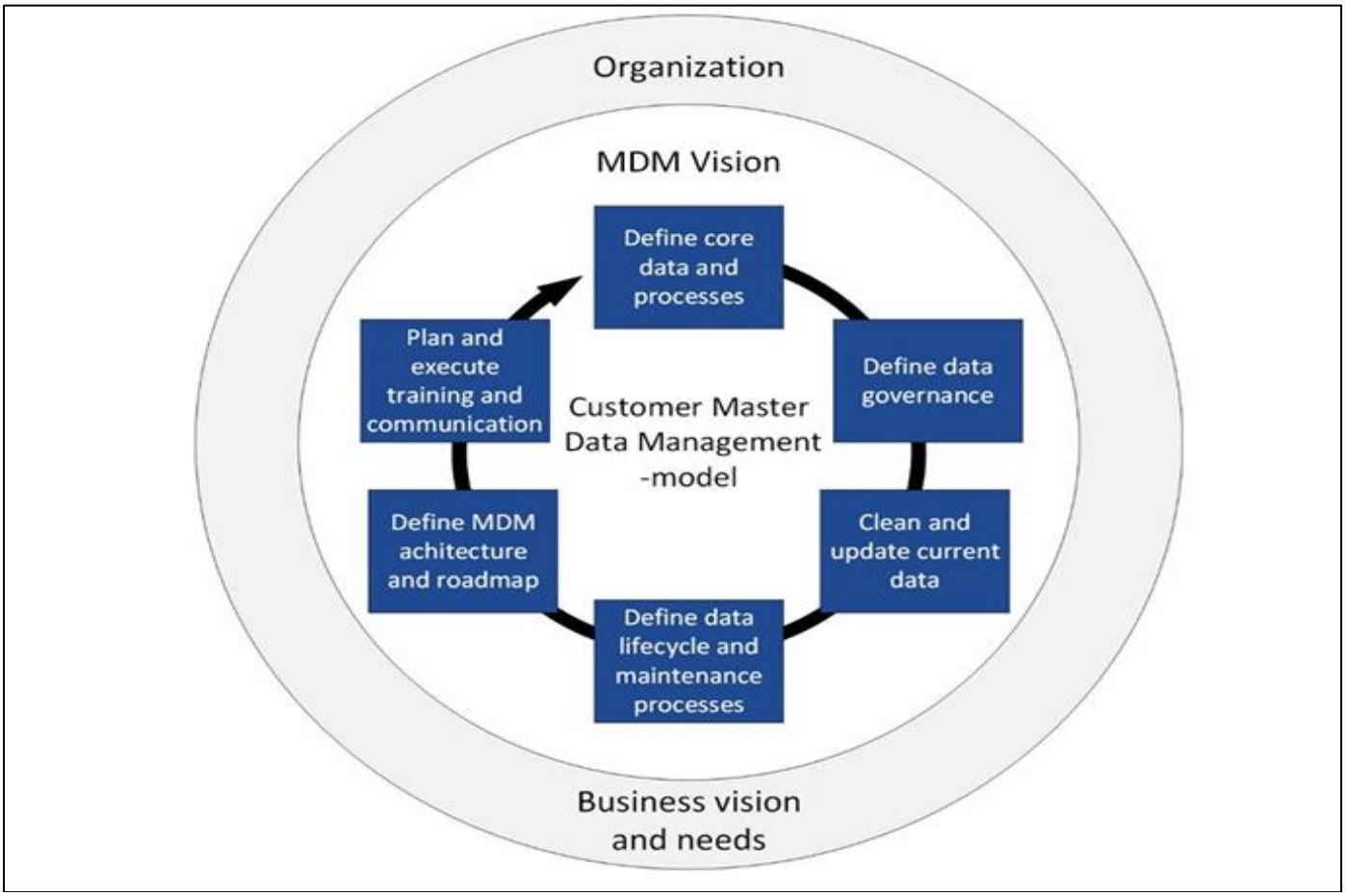

Fig 1 Customer Master Data Management Model (Pansara, 2021).

The multi-model approach differs from monolithic systems because it unifies different data processing techniques and machine learning models for structured, semi-structured and unstructured data into one framework (Gartner, 2022).

Table 1 Components of the Customer Master Data Management (Credency, 2023).

Component	Function
LLM Module	Semantic understanding, classification, entity resolution
Rules Engine	Deterministic validation (e.g., formatting, field checks)
Knowledge Graph	Entity linking and relationship mapping
Search Index	Fast retrieval for matching and enrichment
Data Pipeline	Stream/batch ingestion, transformation, and monitoring

The architecture provides enhanced flexibility and scalability and intelligence for managing complex customer data ecosystems. From Table.1 the CMDM components and its function are explained in detail (Credency, 2023).

IV. IMPLEMENTATION STRATEGIES: LLM INTEGRATION APPROACHES

A strategic approach must be followed when integrating Large Language Models (LLMs) into Customer Master Data Management (CMDM) systems to achieve scalability and data security and effectiveness (Li, 2024). The selection of integration approaches depends on the current infrastructure and data sensitivity levels and use-case complexity of organizations.

➢ *Embedded LLMs within CMDM Platforms:*

LLMs function as part of CMDM systems which operate from private cloud or on-premises environments.

- Use Case: Suitable for enterprises with high regulatory compliance needs (e.g., healthcare, finance) (Zhao, et al., 2024).
- Advantages: Full data control, fine-tuning operations on its proprietary customer information, Reduced latency (Zhou, Zhao, & & Li, 2024).
- Challenges: Requires internal ML expertise, High infrastructure and maintenance costs (Zhao, et al., 2024).

➢ *API-Based Integration with Public LLMs:*

The CMDM tools connect to external LLMs (e.g., OpenAI, Anthropic, Cohere) through secure APIs for interaction.

- Use Case: Fast deployment for organizations with lower sensitivity data or early-stage LLM adoption (Arora & & Singh, 2023).
- Advantages: Rapid implementation, access to advanced models without needing to manage infrastructure (Arora & & Singh, 2023).
- Challenges: Data privacy concerns (mitigated by encryption/anonymization), Potential API rate limits or cost scaling (Arora & & Singh, 2023).

➢ *Hybrid Approach (LLM + Traditional Rules-Based Engines):*

The system merges deterministic rules (e.g., regex, matching logic) with probabilistic LLM reasoning.

- Use Case: The approach works best for organizations to move from traditional systems while they test AI enhancement capabilities (Vertsel & & Rumiantsau, 2024).
- Advantages: Gradual adoption, High interpretability (Vertsel & & Rumiantsau, 2024).
- Challenges: Integration complexity requires strong governance to prevent conflicts between different logic layers (Vertsel & & Rumiantsau, 2024).

➢ *Federated or Edge LLMs:*

The system deploys LLMs at data source locations or edge devices to process information without moving sensitive data to a central point.

- Use Case: Industries with strict data residency or distributed architecture (Zhou, Zhao, & & Li, 2024).
- Advantages: Maximum data sovereignty, Scalable across geographies (Zhou, Zhao, & & Li, 2024).
- Challenges: High coordination complexity, Model versioning and updates are more difficult (Zhou, Zhao, & & Li, 2024).

V. HUMAN-IN-THE-LOOP (HITL) FOR GOVERNANCE AND ACCURACY

Organizations need to establish governance systems, standards and accountability measures when integrating AI and Large Language Models (LLMs) into Customer Master Data Management (CMDM). The Human-in-the-Loop (HITL) model serves as an oversight system maintaining balance between automated processes and human decision-making in critical data environments (Retzlaff, 2024).

➢ *HITL in CMDM:*

The Human-in-the-Loop system is a framework enabling experts to monitor, validate and enhance AI model outputs (Kumar, 2024). HITL serves as a fundamental element in CMDM to maintain quality standards and compliance requirements of customer master records during these processes:

- Entity resolution and de-duplication.
- Standardization of non-obvious data.
- Interpretation of unstructured inputs.
- Approval of automated data enrichment.
- Flagging and reviewing anomalies (Kumar, 2024).

➢ *Key Functions of HITL in CMDM:*

- Exception Handling: Data stewards handle complex records. Validation of AI Outputs: Reviewers validate LLM suggestions when proposing entity unification between "ABC Inc." and "ABC Corporation" (Retzlaff, 2024).Complex records are escalated to data stewards for expert resolution (Kumar, 2024).
- Feedback Loops: LLM models receive human feedback to improve behavior through time (Retzlaff, 2024).
- Policy Enforcement: The system verifies CMDM operations following GDPR and HIPAA regulations and organizational standards (Kumar, 2024).
- Auditable Decision Trails: AI and human decisions are tracked through auditable trails for transparency (Retzlaff, 2024).

➢ *Benefits of HITL in CMDM:*

- Trust and Reliability: Builds confidence in AI-assisted CMDM through human oversight.
- Regulatory Compliance: Ensures outputs adhere to data privacy laws.
- Continuous Improvement: Facilitates model learning through human feedback.
- Error Reduction: Minimizes risks of incorrect merges or enrichments (Retzlaff, 2024).

➢ *Benefits and Use Cases of LLMs in Customer Master Data Management:*

Below Table 2. gives us more detailed information on the benefits of using LLMs in Customer Master Data Management.

Table 2 Benefits of LLMs in Customer Master Data Management (Zhao, et al., 2024).

Benefit	Description
Improved Data Quality	Higher precision in matching, deduplication, and enrichment
Scalable and Adaptive	LLMs adapt to new data types, sources, and formats with less manual tuning
Faster Onboarding	Reduced time to onboard customer data from multiple systems
Regulatory Compliance	Enforced data governance, auditability, and traceability of records
Enhanced Customer Experience	Unified 360° view of the customer enables personalized engagement

➢ *Use Cases:*

Below Table 3. gives us more detailed information on the Use cases in which LLMs can be used in Customer Master Data Management.

Table 3 Use Cases of LLMs in Customer Master Data Management (Forbes, 2024).

Use Case	Description
Customer Data Unification	Consolidating fragmented customer profiles across CRM, ERP, and support systems using LLM-based entity resolution.
Automated Onboarding Validation	Validating customer entries from forms, emails, and documents during onboarding using NLP and LLM parsing.
Predictive Field Completion	Using LLMs to suggest or autofill missing address or contact fields based on known attributes.
Chatbot Data Capture	Extracting and storing accurate customer details from live chat interactions.
Support Ticket Enrichment	Identifying key customer issues and tagging metadata automatically from support logs.
Multilingual Data Standardization	Translating and harmonizing customer information submitted in various languages.

VI. LIMITATIONS AND RISKS OF LLM-BASED CMDM SOLUTIONS

Although implementation of Large Language Models (LLMs) within Customer Master Data Management (CMDM) gives us advantages, organizations must handle limitations and risks to maintain reliability and compliance.

- *Model Drift:*

- Description: Due to changes in customer conduct, data structures and business regulations, LLM performance decreases over time which results in model drift (Raymond, 2025).
- Impact: This reduces the matching accuracy with outdated predictions and misaligned classifications (Raymond, 2025).
- Mitigation: The system needs to be retrained and using current business data (Raymond, 2025).

- *Data Privacy and Security:*

- Description: Fine-tuning or querying LLMs with sensitive customer data creates privacy risks violating GDPR, HIPAA and CCPA regulations (Raymond, 2025).
- Impact: Legal compliance issues and reputation damage occur when personal data is exposed during model operations (Raymond, 2025).
- Mitigation: Privacy methods including differential privacy, on-premises models and encryption should be implemented (Raymond, 2025).

- *Bias and Hallucination:*

- Description: LLM training data allows biases to spread while enabling generations of believable yet incorrect outputs (Raymond, 2025).
- Impact: This leads to wrong customer merges, classification errors and misinterpretations affecting business decisions and fairness standards (Raymond, 2025).
- Mitigation: HITL validation and explainability tools along with bias audits help to understand validate system outputs (Raymond, 2025).

- *Cost and Complexity:*

- Description: Maintaining and deploying LLM-powered CMDM systems requires major investments in infrastructure, governance frameworks and talent acquisition (Raymond, 2025).
- Impact: Technical debt and high computational expenses occur in LLM deployment using the existing systems (Raymond, 2025).
- Mitigation: Using cloud-based APIs and fine-tuned smaller models with phased deployment helps balance costs and value (Raymond, 2025).

VII. FUTURE TRENDS

Self-learning CMDM systems that use large language models (LLMs) have become popular because they improve both matching precision and operational speed throughout time. They require industries to develop flexible systems which maintain their competitive edge (Rane et al., 2024; Arévalo and Jurado, 2024).

Real-time customer data fusion represents an advancing trend that merges transactional information with behavioral data to generate complete customer understanding. Businesses can create personalized customer interactions through this method which leads to enhanced customer experiences through timely targeted responses customer relationship management (Arévalo and Jurado, 2024; Aslam, 2023).

The adoption of blockchain technology with LLMs has started to increase because it enables organizations to build secure audit trails system which support CMDM systems. The combination of blockchain with AI technology creates a secure system that protects AI and machine learning technology continues to transform business operations through enhanced efficiency and market responsiveness and sustainable practices in multiple industries (Rane et al., 2024; Gulyamov, 2024).

VIII. CONCLUSION

Large Language Models (LLMs) along with Customer Master Data Management (MDM) play important role for data management strategies and advancing personalization. The use of LLMs can improve customer experiences by integrating and processing large amounts of customer data in real time, enabling personalized strategies that rely on complete data about who's involved, what is agreed and how its traced (Valdez Mendia & Flores-Cuautle, 2022). Large Language Models (LLMs) along with Customer Master Data Management (MDM) play important role for data management strategies and advancing personalization (Saarijärvi et al., 2013).

Master Data Management (MDM) success depends on more than technology because organizations need defined data ownership and excellent data quality standards and team-level data alignment. The system requirements indicate a need for technology that connects business units with IT departments through effective communication channels (Vilminko-Heikkinen & Pekkola, 2017). Large language models (LLMs) help organizations achieve process simplification and data precision improvement while delivering immediate analytical results. The solution enables e-commerce businesses to handle integration problems effectively while making expedited strategic decisions (Lande et al., 2024).

LLMs with customer data management frameworks can help companies to work more efficiently, improve their day-to-day work by making data management

available to both IT and managerial departments. Thus, resulting in both organizational efficiency and customer satisfaction (Otto et al., 2011; Silvola et al., 2011).

REFERENCES

[1]. Vilminko-Heikkinen, R., & Pekkola, S. (2017). Master data management and its organizational implementation. Journal of Enterprise Information Management, 30(3), 454–475. https://doi.org/10.1108/jeim-07-2015-0070.

[2]. Otto, B., Hüner, K. M., & Österle, H. (2011). Toward a functional reference model for master data quality management. Information Systems and E-Business Management, 10(3), 395–425. https://doi.org/10.1007/s10257-011-0178-0.

[3]. Li, G. (2024). LLM for data management. Tsinghua University. https://dbgroup.cs.tsinghua.edu.cn/ligl/papers/p2031-li-vldb2024.pdf.

[4]. Pansara, R. (2021). Master data management challenges. International Journal of Computer Science and Mobile Computing, 10(10), 47–49. Retrieved from https://www.researchgate.net/publication/355780526_Master_Data_Management_Challenges.

[5]. Zhou, X., Zhao, X., & Li, G. (2024). LLM-enhanced data management. arXiv. https://arxiv.org/abs/2402.02643.

[6]. Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., ... & Li, Q. (2024). Recommender systems in the era of large language models (LLMs). IEEE Transactions on Knowledge and Data Engineering, 36(11), 6889–6907. https://arxiv.org/abs/2307.02046.

[7]. Gartner. (2022). Master data management (MDM). Gartner Glossary. https://www.gartner.com/en/information-technology/glossary/master-data-management-mdm.

[8]. Credencys. (2023). What is customer master data management? https://www.credencys.com/blog/what-is-customer-master-data-management/.

[9]. Arora, D., & Singh, H. G. (2023). Have LLMs advanced enough? A challenging problem-solving benchmark for large language models. arXiv preprint arXiv:2305.15074. https://arxiv.org/abs/2305.15074.

[10]. Vertsel, A., & Rumiantsau, M. (2024). Hybrid LLM-rule-based data extraction. arXiv. https://arxiv.org/pdf/2404.15604.

[11]. Retzlaff, C. O., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., ... & Holzinger, A. (2024). Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. Journal of Artificial Intelligence Research, 79, 359-415.

[12]. Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S. K., & Sharma, R. (2024). Applications, challenges, and future directions of human-in-the-loop learning.

[13]. Forbes Technology Council. (2024). Successful real-world use cases for LLMs (and lessons they teach). Forbes https://www.forbes.com/councils/forbestechcouncil/2024/.

[14]. Raymond, D. (2025). Top 10 cons & disadvantages of large language models (LLM). Project Managers. Retrieved from https://projectmanagers.net/top-10-disadvantages-of-large-language-models-llm/.

[15]. Rane, N. L., Desai, P., Paramesha, M., & Rane, J. (2024). Artificial intelligence, machine learning, and deep learning for sustainable and resilient supply chain and logistics management. deep science. https://doi.org/10.70593/978-81-981367-4-9_5.

[16]. Gulyamov, S. (2024). Intelligent waste management using IoT, blockchain technology and data analytics. E3S Web of Conferences, 501, 01010. https://doi.org/10.1051/e3sconf/202450101010.

[17]. Aslam, F. (2023). The Impact of Artificial Intelligence on Chatbot Technology: A Study on the Current Advancements and Leading Innovations. European Journal of Technology, 7(3), 62–72. https://doi.org/10.47672/ejt.1561.

[18]. Arévalo, P., & Jurado, F. (2024). Impact of Artificial Intelligence on the Planning and Operation of Distributed Energy Systems in Smart Grids. Energies, 17(17), 4501. https://doi.org/10.3390/en17174501.

[19]. Rane, N. L., Kaya, & Rane, J. (2024). Advancing the Sustainable Development Goals (SDGs) through artificial intelligence, machine learning, and deep learning. deep science. https://doi.org/10.70593/978-81-981271-8-1_4.

[20]. Saarijärvi, H., Karjaluoto, H., & Kuusela, H. (2013). Customer relationship management: the evolving role of customer data. Marketing Intelligence & Planning, 31(6), 584–600. https://doi.org/10.1108/mip-05-2012-0055.

[21]. Silvola, R., Kropsu-Vehkapera, H., Haapasalo, H., & Jaaskelainen, O. (2011). Managing one master data – challenges and preconditions. Industrial Management & Data Systems, 111(1), 146–162. https://doi.org/10.1108/02635571111099776.

[22]. Lande, O., Simpson, B., Adeleke, G., Amajuoyi, C., & Johnson, E. (2024). Enhancing business intelligence in e-commerce: Utilizing advanced data integration for real-time insights. International Journal of Management & Entrepreneurship Research, 6(6), 1936–1953. https://doi.org/10.51594/ijmer.v6i6.1207.

[23]. Valdez Mendia, J. M., & Flores-Cuautle, J. J. A. (2022). Toward customer hyper-personalization experience — A data-driven approach. Cogent Business & Management, 9(1). https://doi.org/10.1080/23311975.2022.2041384.