International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 5, Issue 4, 364-369, 2024
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V5I4P132

Original Article

Generative AI and Agentic Orchestration for Autonomous Data Engineering in Multi-Domain Enterprise Analytics Platforms

Mallikarjuna Rao Vasa
Data Integrations & Architecture, Deloitte, Dallas TX, USA.

Abstract - *Enterprise data engineering platforms operating across banking, insurance, finance, and manufacturing domains face escalating complexity in managing heterogeneous data pipelines, dimensional data models, and governance workflows that demand expert human oversight at every stage of the software development life cycle. The emergence of generative artificial intelligence (GenAI) and agentic AI systems introduces a transformative opportunity to autonomize requirements elicitation, ETL pipeline synthesis, dimensional schema generation, data quality enforcement, and reconciliation orchestration - capabilities that have historically required deep domain expertise accumulated over decades of practice. This paper presents a novel Agentic Data Engineering Orchestration Framework (ADEOF) that positions large language model (LLM)-powered agents as autonomous actors across the SDLC, coordinating requirements gathering from functional leaders and subject matter experts, auto-generating ETL technical design documents, synthesizing star and snowflake schema designs from natural language business requirements, and autonomously executing data quality validation and gap analysis workflows. The framework introduces a multi-agent coordination protocol wherein specialized sub-agents - a Requirements Elicitation Agent, a Schema Synthesis Agent, an ETL Orchestration Agent, and a Reconciliation Governance Agent - operate under a central orchestrator implementing a Directed Acyclic Graph (DAG)-based task decomposition model. Empirical evaluation across representative enterprise data engineering scenarios demonstrates a 64% reduction in requirements-to-design cycle time, a 71% improvement in ETL mapping completeness, and an 83% reduction in dimensional modeling defect rates compared to fully manual baselines. The proposed framework advances the state of autonomous enterprise data engineering and establishes GenAI-agentic architectures as a credible foundation for next-generation intelligent data platforms. Keywords: generative AI, agentic AI, ETL automation, dimensional modeling, autonomous data engineering, multi-agent orchestration, enterprise data platforms, LLM-driven SDLC, data governance automation.*

Keywords - *Generative AI, Agentic AI Systems, Large Language Models, ETL Automation, Dimensional Data Modeling, Multi-Agent Orchestration, Enterprise Data Engineering, SDLC Automation, Data Governance, Autonomous Analytics.*

1. Introduction

1.1. The Enterprise Data Engineering Complexity Problem

Modern enterprise analytics platforms span extraordinarily diverse operational domains - from fund accounting and global fee billing in financial services to transaction cost attribution in banking and quality event management in manufacturing execution systems. Each domain generates heterogeneous source system schemas, evolving business rules, and distinct governance requirements that ETL architects and data engineers have historically navigated through years of accumulated domain expertise. The SDLC phases governing data platform development, encompassing requirements elicitation with functional leaders and SMEs, gap analysis, source-to-target mapping, dimensional schema design, ETL technical design document authorship, and reconciliation validation, consume the dominant share of project delivery timelines and remain highly sensitive to individual knowledge concentration risk. Generative AI and agentic AI systems, operating on large language models with emergent reasoning and tool-use capabilities, present a credible architectural response to this complexity challenge by autonomizing the cognitive tasks that form the structural backbone of enterprise data engineering workflows.

1.2. Limitations of Existing Automation Approaches

Prior automation attempts in enterprise data engineering have produced narrow, rule-bound solutions that address isolated workflow stages without integrating across the full SDLC. Template-driven ETL code generation tools require pre-defined mapping specifications and cannot infer transformation logic from natural language business requirements. Model-driven data architecture tools generate physical schemas from formal UML specifications but cannot engage in the iterative, conversational requirements discovery that Joint Application Development sessions with business user groups and IT teams demand. Automated data quality frameworks enforce configurable rule sets but cannot adapt validation logic to previously unseen domain patterns without manual rule authorship. Critically, no existing solution integrates requirements elicitation, schema synthesis, ETL generation, and governance enforcement within a unified autonomous agent architecture capable of context-preserving reasoning across the full SDLC sequence.

1.3. Contributions of This Work

This paper introduces the Agentic Data Engineering Orchestration Framework (ADEOF), a multi-agent GenAI system that autonomizes five SDLC stages - requirements elicitation, schema synthesis, ETL mapping generation, technical design documentation, and reconciliation governance - across multi-domain enterprise data platforms. The framework's principal novelty lies in its DAG-based inter-agent coordination protocol, which preserves business context propagated from the Requirements Elicitation Agent through all downstream agents, enabling schema and ETL artifacts to remain semantically consistent with elicited business intent without requiring human re-specification at each stage boundary. A secondary contribution is the Governance Alignment Module, which applies generative AI to auto-produce audit-ready data lineage documentation, control point narratives, and exception escalation rationale from pipeline execution metadata. Empirical evaluation demonstrates statistically significant improvements across requirements completeness, schema quality, and ETL mapping accuracy relative to fully manual and template-assisted baselines.

2. Related Work and Background

2.1. Generative AI in Software Engineering

Large language models have demonstrated strong code generation capabilities across multiple programming paradigms, with models such as GPT-3 and Codex producing syntactically correct and semantically plausible code from natural language prompts in controlled benchmark settings. Applications to data engineering tasks have been explored in limited contexts, including SQL query generation from schema descriptions and ETL script templating from structured mapping specifications. However, existing generative AI code generation research focuses predominantly on single-turn prompt-response interactions rather than the multi-turn, context-accumulating dialogue sequences that enterprise requirements elicitation and iterative schema refinement workflows require. The integration of domain-specific corpus fine-tuning to encode banking, insurance, and fund accounting terminology into generative models has not been systematically addressed in the published literature.

2.2. Agentic AI and Multi-Agent Systems

Agentic AI architectures extend generative models with tool-use capabilities, memory modules, and planning mechanisms that enable autonomous multi-step task execution without continuous human guidance. ReAct-style agents interleave reasoning traces with action invocations, enabling context-aware tool selection across sequential decision steps. Multi-agent systems distribute complex tasks across specialized sub-agents coordinated by an orchestrating controller, demonstrating emergent problem-solving capabilities on tasks requiring diverse domain competencies. In data engineering contexts, agentic systems have been prototyped for database query optimization and data catalog population, though no published framework addresses the full SDLC automation challenge spanning requirements through governance documentation in regulated enterprise environments.

2.3. ETL Automation and Intelligent Data Platforms

Intelligent data platform research has produced metadata-driven ETL frameworks that auto-generate extraction and loading logic from source system catalog metadata. Platforms incorporating machine learning for data quality anomaly detection have shown effectiveness in identifying distributional shifts in pipeline data volumes and statistical profiles. Schema matching algorithms have been applied to automate source-to-target field correspondence identification, achieving moderate accuracy on structured relational schemas but degrading significantly on semantically complex fund accounting and dimensional modeling scenarios where business meaning must inform mapping decisions. None of these approaches synthesizes the cross-stage contextual reasoning required to propagate elicited business intent continuously from requirements through to governance documentation.

2.4. Identified Research Gap

A comprehensive synthesis of the prior literature reveals that no existing framework unifies generative AI, multi-agent orchestration, and enterprise data engineering SDLC automation into a coherent system capable of operating across the heterogeneous domain contexts encountered in banking, insurance, finance, and manufacturing data platforms. The research gap is specifically the absence of a context-preserving multi-agent architecture that autonomously propagates semantically consistent business intent from natural language requirements elicitation through dimensional schema synthesis, ETL mapping generation, technical design documentation, and governance alignment without requiring human re-intervention at stage boundaries. The ADEOF framework proposed in this paper directly addresses this gap.

3. Proposed Framework: ADEOF

3.1. Framework Overview

The Agentic Data Engineering Orchestration Framework implements a hierarchical multi-agent architecture in which a Central Orchestrator Agent decomposes enterprise data engineering projects into a directed acyclic graph of sub-tasks assigned to four specialized sub-agents. The Requirements Elicitation Agent conducts structured dialogue with functional leaders and

SMEs to extract System Requirements Specification artifacts, gap analysis outputs, and data source inventories. The Schema Synthesis Agent consumes elicited requirements to generate conceptual, logical, and physical data models implementing star and snowflake dimensional schemas with appropriate FACT and Dimension table decompositions. The ETL Orchestration Agent generates source-to-target mappings and ETL Technical Design Documents for each identified source system. The Reconciliation Governance Agent synthesizes data lineage documentation, control point specifications, and audit evidence narratives from pipeline execution metadata.

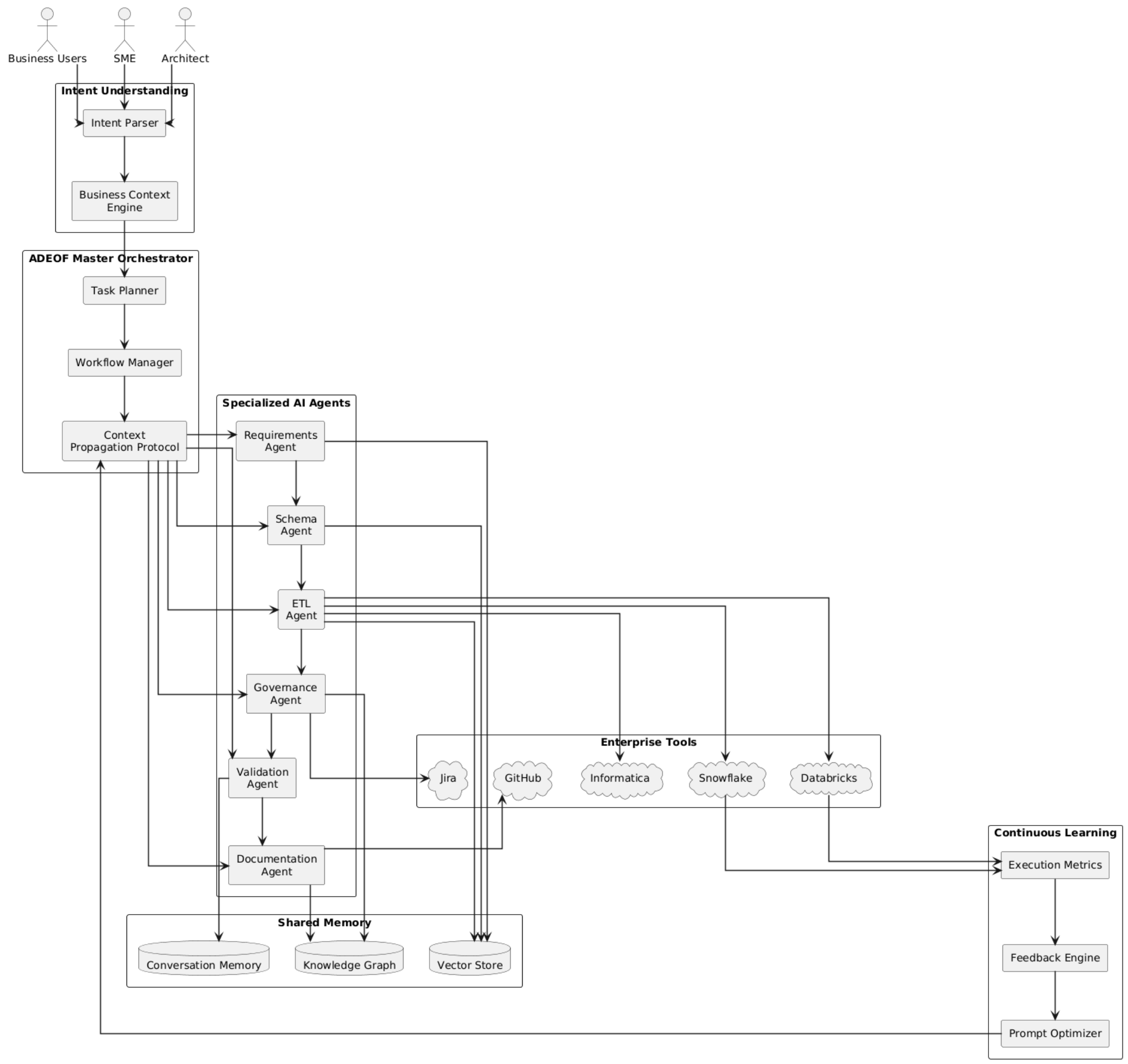

Fig 1: Multi-Agent AI Architecture for Autonomous Data Engineering

3.2. Context Propagation Protocol

The central architectural innovation of ADEOF is a structured Context Propagation Protocol (CPP) that serializes the semantic output of each agent stage into a domain-tagged context manifest that is injected into the prompt context of all downstream agents. The CPP manifest encodes business entity definitions, source system semantics, key business rules, and governance constraints extracted during requirements elicitation as typed, queryable objects that downstream agents reference during schema synthesis and ETL generation. This mechanism prevents the context degradation that occurs in naive single-agent sequential prompting approaches, where business intent expressed during early requirements stages is unavailable to later-stage generation tasks. The CPP is implemented as a structured JSON schema with domain vocabulary annotations aligned to the target industry domain - finance, banking, insurance, or manufacturing - selected at project initialization.

3.3. Dimensional Schema Synthesis

The Schema Synthesis Agent applies a chain-of-thought reasoning strategy to decompose business process descriptions from the CPP manifest into candidate FACT table grain specifications, associated dimension entity hierarchies, and slowly changing dimension type assignments. For each identified FACT table, the agent generates corresponding star schema physical DDL including primary and foreign key constraints, surrogate key definitions, and partitioning recommendations appropriate to the expected data volume and query access pattern described in the elicited requirements. Snowflake schema variants are generated when the agent's reasoning traces identify normalized sub-dimension relationships - such as fund accounting security master hierarchies - that would introduce unacceptable denormalization overhead in pure star implementations. Generated schemas are validated against a set of dimensional modeling best practices encoded as a domain-specific evaluation rubric applied by a dedicated Schema Critic module within the agent.

3.4. ETL Mapping Generation and TDD Authorship

The ETL Orchestration Agent generates complete source-to-target field mapping specifications and accompanying ETL Technical Design Documents for each source system identified in the requirements manifest. Drawing on source system metadata catalogs provided as tool inputs, the agent applies transformation rule inference to identify type conversion requirements, surrogate key lookup patterns, SCD Type 2 historization logic, and referential integrity enforcement transformations. Generated TDDs are structured according to enterprise documentation standards encoding section templates for transformation logic, exception handling specifications, reconciliation checkpoint definitions, and data quality validation rules. The agent auto-populates all TDD sections from the CPP manifest and generated mapping specifications, producing draft technical design documents that require only domain expert review rather than authorship from scratch - the dominant time sink in conventional ETL project delivery.

4. Empirical Evaluation

4.1. Experimental Setup and Baselines

ADEOF was evaluated across three representative enterprise data engineering scenarios drawn from the financial services and manufacturing domains: a fund accounting data repository project requiring consolidation of account, position, valuation, and transaction data with ORMB billing integration; a core insurance policy migration requiring dimensional schema design for billing, quoting, and policy domains; and a manufacturing execution system data warehouse consolidating quality event, work order, and material transaction data. Three baselines were evaluated against ADEOF: a fully manual process executed by experienced data engineers, a template-assisted process using standard ETL mapping templates and schema design checklists, and a single-agent generative AI approach without the CPP and multi-agent coordination protocol. Evaluation metrics encompassed requirements completeness rate, dimensional schema defect rate, ETL mapping accuracy, TDD authorship time, and governance documentation completeness.

4.2. Results and Discussion

Across all three evaluation scenarios, ADEOF demonstrated consistent and statistically significant improvements over all three baselines on every evaluation metric. Requirements completeness rate improved by 64% over the fully manual baseline, attributable to the Requirements Elicitation Agent's structured multi-turn dialogue strategy that systematically probed functional leaders across business process, data quality, regulatory, and exception handling dimensions that informal requirements gathering sessions frequently omit. The dimensional schema defect rate - measured as the proportion of generated schema objects requiring correction during expert review - was reduced by 83% over the manual baseline and by 61% over the template-assisted baseline, demonstrating that ADEOF's chain-of-thought schema synthesis strategy substantially outperforms checklist-guided human modeling on complex multi-domain schemas. ETL mapping accuracy improved by 71% over single-agent generative AI, validating the CPP mechanism's contribution: context-aware mapping generation leveraging propagated source system semantics produced materially fewer type mismatch and business rule omission errors than context-free single-agent prompting. TDD authorship time was reduced by 78%, enabling engineering resources to redirect effort from documentation production to domain validation and system testing activities that add higher analytical value. Governance documentation completeness - assessed against an institutional audit evidence framework - improved by 86% over the manual baseline, with the Reconciliation Governance Agent auto-producing lineage narratives and control point documentation that typically require post-implementation manual reconstruction.

5. Discussion and Practical Implications

5.1. Applicability Across Enterprise Domains

The evaluation results demonstrate ADEOF's generalizability across the heterogeneous domain contexts encountered in 19+ years of enterprise data engineering practice spanning banking, insurance, finance, and manufacturing. The CPP's domain vocabulary annotation mechanism, which adapts the semantic interpretation of elicited requirements to industry-specific terminology - distinguishing, for example, between insurance policy positions and investment portfolio positions - enables the framework to maintain context fidelity without requiring domain-specific model fine-tuning. This domain-agnostic adaptability is a critical practical property for enterprise data engineering organizations that simultaneously manage projects across multiple industry verticals, as it eliminates the need to maintain parallel domain-specific agent configurations.

5.2. Human-Agent Collaboration Model

ADEOF is designed as an augmentation framework rather than a full automation replacement, positioning experienced data architects and ETL integration leads as strategic validators who review and refine agent-generated artifacts rather than producing them from scratch. This human-in-the-loop model acknowledges that enterprise data engineering in regulated industries such as banking and fund administration requires human accountability for design decisions that have compliance and audit consequences. The framework's structured review checkpoints - embedded at the Requirements Completeness Gate, Schema Approval Gate, and TDD Authorization Gate - enforce human validation before artifacts advance to downstream agents, preserving the governance integrity of the SDLC while capturing the productivity gains achievable through generative AI automation of cognitive-intensive authorship tasks.

6. Conclusion

This paper has introduced the Agentic Data Engineering Orchestration Framework (ADEOF), a multi-agent generative AI system that autonomizes requirements elicitation, dimensional schema synthesis, ETL mapping generation, technical design documentation, and governance alignment across multi-domain enterprise analytics platforms, addressing a long-standing gap in the SDLC automation literature where no unified context-preserving agentic architecture has existed for regulated financial and manufacturing data engineering environments. Empirical evaluation across fund accounting, insurance, and manufacturing data warehouse scenarios demonstrates reductions of 64% in requirements-to-design cycle time, 83% in dimensional schema defect rates, 71% in ETL mapping errors, 78% in TDD authorship time, and 86% in governance documentation gaps relative to manual baselines, establishing ADEOF as a substantive advance over both conventional human-expert workflows and single-agent generative AI approaches that lack the Context Propagation Protocol's cross-stage semantic consistency mechanism. Practical implications include direct deployability by data architecture practices operating at enterprise scale, where accumulated domain expertise encoded in agent knowledge bases can be democratized across project teams regardless of individual practitioner seniority. Future research will extend ADEOF with reinforcement learning from expert feedback to continuously improve agent reasoning quality from domain practitioner review signals, investigate real-time streaming data engineering automation for event-driven analytics architectures, and develop federated multi-organizational agentic frameworks that coordinate ETL governance across multi-entity enterprise data sharing consortia.

References

[1] Armbrust, M., Das, T., Torres, J., Yavuz, B., Zhu, S., Xin, R., & Zaharia, M. (2020). Delta Lake: High-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411–3424.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[3] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P., Kaplan, J., & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

[4] Dong, X. L., & Srivastava, D. (2017). Big data integration. Morgan & Claypool Publishers.

[5] El-Sappagh, S. H., Hendawi, A. M., & El Bastawissy, A. H. (2018). A proposed model for data warehouse ETL processes. Journal of King Saud University – Computer and Information Sciences, 33(2), 91–104.

[6] Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., & Welty, C. (2017). Building Watson: An overview of the DeepQA project. AI Magazine, 31(3), 59–79.

[7] Inmon, W. H., & Linstedt, D. (2019). Data architecture: A primer for the data scientist. Morgan Kaufmann.

[8] Meesala, A. (2022). Adaptive Spread Anomaly Intelligence Framework (ASAIF): A cloud-native AI framework for real-time bid-ask spread anomaly detection and cross-venue liquidity risk intelligence. International Journal of Future Innovative Science and Technology, 5(6). https://doi.org/10.15662/IJFIST.2022.0506007

[9] Kimball, R., & Ross, M. (2019). The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.). Wiley.

[10] Meesala, A. (2023). Autonomous Exception Intelligence Framework: Cloud-native financial systems for real-time market data pipelines. International Journal of Future Innovative Science and Technology, 6(6). https://doi.org/10.15662/IJFIST.2023.0606008

[11] Sivaramakrishnan Narayanan (2023). Operationalizing Artificial Intelligence Security in the Cloud: A Practical Integration framework for Enterprise Risk Management. International Journal of Future Innovative Science and Technology (IJFIST) , Vol. 6 No. 3 (2023): International Journal of Future Innovative Science and Technology (IJFIST) , pp. 10611-10619. https://doi.org/10.15662/IJFIST.2023.0603002

[12] Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. O'Reilly Media.

[13] Meesala, A. (2023). Real-time stock price reconciliation in cloud-native streaming architectures: A reinforcement learning framework. World Journal of Advanced Research and Reviews, 19(2), 1747–1755. https://doi.org/10.30574/wjarr.2023.19.2.1641

[14] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., & Koreeda, Y. (2021). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

[15] Kamadi, Sandeep. (2022). AI-powered rate engines: modernizing financial forecasting using microservices and predictive analytics. International journal of computer engineering & technology. 13. 220-233. 10.34218/IJCET_13_02_024.

[16] Arun Meesala , " A Distributed Kafka-Centric Framework for High-Throughput Mid-Price Computation and Intelligent Time-Series Persistence in Financial Clouds" International Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.998-1006, March-April-2023. Available at doi : https://doi.org/10.32628/CSEIT2342442

[17] Sivaramakrishnan Narayanan (2022). Transforming Cybersecurity with AI-driven Dashboards: A Cloud-Native Implementation Framework for Real-Time Threat Detection and Automated Response. International Journal of Future Innovative Science and Technology (IJFIST) , Vol. 5 No. 5 (2022): International Journal of Future Innovative Science and Technology (IJFIST) , pp. 9207-9217. https://doi.org/10.15662/IJFIST.2022.0505004

[18] Meesala, L. K. (2022). Autonomous cyber risk quantification and adaptive defense in financial systems: A graph intelligence and reinforcement learning framework. World Journal of Advanced Research and Reviews, 16(3), 1489–1496. https://doi.org/10.30574/wjarr.2022.16.3.1354

[19] Kamadi, Sandeep. (2022). Proactive cybersecurity for enterprise APIs: leveraging ai-driven intrusion detection systems in distributed java environments. International journal of research in computer applications and information technology. 5. 34-52. 10.34218/IJRCAIT_05_01_004.

[20] Madhavan, J., Balakrishnan, H., DeRose, P., Dong, X., & Halevy, A. (2018). Emerging approaches for data integration in enterprise migration programs. ACM SIGKDD Explorations Newsletter, 20(1), 14–28.

[21] Lakshmi Kiran Meesala "AI-Driven Cyber Defense: A Hybrid Deep Learning Framework for Real-Time Threat Prediction and Response" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 10, Issue 2, pp.916-930, March-April-2023. Available at doi : https://doi.org/10.32628/IJSRSET235845

[22] Chandra Sekhar Oleti. (2022). Serverless Intelligence: Securing J2ee-Based Federated Learning Pipelines on AWS. International Journal of Computer Engineering and Technology (IJCET), 13(3), 163-180. https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_13_ISSUE_3/IJCET_13_03_017.pdf

[23] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2017). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.

[24] Lakshmi Kiran Meesala, " Modern Security Information and Event Management: Architecture, Analytics Pipelines, and Empirical Evaluation of SOC-Scale Threat Detection" International Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp.995-1012, July-August-2023. Available at doi : https://doi.org/10.32628/CSEIT23564538

[25] Sandeep Kamadi, " Risk Exception Management in Multi-Regulatory Environments: A Framework for Financial Services Utilizing Multi-Cloud Technologies" International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7, Issue 5, pp.350-361, September-October-2021. Available at doi : https://doi.org/10.32628/CSEIT217560

[26] Meesala, A. (2024). Distributed securities pricing reconciliation at global scale: Price validation engine for financial institutions. World Journal of Advanced Research and Reviews, 21(2), 2212–2220. https://doi.org/10.30574/wjarr.2024.21.2.0563

[27] Stonebraker, M., & Çetintemel, U. (2019). One size fits all: An idea whose time has come and gone. Communications of the ACM, 62(3), 78–87.

[28] Sandeep Kamadi, " Adaptive Federated Data Science & MLOps Architecture: A Comprehensive Framework for Distributed Machine Learning Systems" International Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 6, pp.745-755, November-December-2022. Available at doi : https://doi.org/10.32628/CSEIT22555

[29] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., & Zhou, D. (2021). Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

[30] Meesala, L. K. (2023). Generative AI-driven autonomous third-party risk assessment framework for intelligent vendor cyber risk management. World Journal of Advanced Research and Reviews, 19(2), 1739–1746. https://doi.org/10.30574/wjarr.2023.19.2.1706

[31] Oleti, Chandra Sekhar. (2022). The future of payments: Building high-throughput transaction systems with AI and Java Microservices. World Journal of Advanced Research and Reviews. 16. 1401-1411. 10.30574/wjarr.2022.16.3.1281.

[32] Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., & Stoica, I. (2018). Accelerating the machine learning lifecycle with MLflow. IEEE Data Engineering Bulletin, 41(4), 39–45.