International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 6, Issue 4, PP 114-122, 2025
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I4P113

Original Article

A Multi-Agent Generative AI Framework for Automated Data Engineering, Governance, and Analytical Optimization

Dinesh Babu Govindarajulunaidu Sambath Narayanan
Independent Researcher, USA.

Received On: 18/09/2025 ***Revised On: 20/10/2025*** ***Accepted On: 30/10/2025*** ***Published On: 12/11/2025***

Abstract: *The data explosion in industries has introduced unprecedented issues in the management, governance and insights derived about huge volumes of data. Conventional data engineering culture is usually unstructured laborious processes that are subject to errors and delays. To automate the data engineering activities, mandate administrative policies and maximize the analytical processes, this paper introduces a new Multi-Agent Generative AI (MAGAI) to automate such functions. MAGAI architecture uses several dedicated AI agents that can work independently to perform such tasks as data cleaning, integrating, transforming, metadata management and automated analytics. The framework combines generative AI models and reinforcement learning strategies to streamline the process of data pipelines and decision-making. Through experimental assessment, it is proved that the efficiency of data processing have better levels and reduced errors in determining correct data and adhering to governance norms. This approach proposed has the benefit of minimizing the human factor as well as improving the quality of the information derived using complex data. We outline the promise of multi-agent AI systems in transforming enterprise data management and analytics.*

Keywords: *Multi-Agent Systems, Generative AI, Data Engineering, Data Governance, Analytical Optimization, Automated Data Pipelines, Reinforcement Learning.*

1. Introduction

1.1. Background

The increasing rate of information in modern companies, diversity, and speed has posed an immense challenge in the effective management and exploitation of information in strategic decision-making. [1-3] Organizations have become more and more dependent on massive, heterogeneous data that cuts across structured databases, semi-structured logs and unstructured data in forms of text, images and sensor information. To convert this raw and heterogeneous data into useful insights, it takes powerful data engineering pipelines that are able to ingest, clean, transform, integrate and validate the data. Conventional methods are mostly manual or semi-automated, and they tend to be time intensive, prone to errors and they cannot keep pace with the changes in the modern data environments. These workflows are further complicated by the growing complexity of regulatory requirements and data governance standards, whereby organizations need to guarantee the adherence to privacy requirements, proper data lineage, and quality and security standards across the distributed environments. Here, new developments in Artificial Intelligence (AI), and more specifically Generative AI, have demonstrated good potential in automated control of more complex data-related tasks. Generative models are able to generate data, and complete missing data, feature engineer, and can even predictive analytics with minimum human intervention. In combination with multi-agent systems (MAS), these AI capabilities can be shared among autonomous agents that observe the environment, think about activities, cooperate with others and perform tasks on their own. MAS allow the coordination of high-level processes to enable a number of agents to work simultaneously, coordinate decision-making, and dynamically respond to emerging data. Using generative AI as a part of a multi-agent ecosystem, enterprises will be able to build automated, intelligent, and scalable data pipelines that not only enhance efficiency, accuracy, and maintain governance compliance but also build and maintain scaling. A combination of generative AI and MAS can offer an innovative mechanism of eliminating the flaws of the traditional data engineering practices and the current urgency of adjustable, independent, and efficient data management systems of the contemporary business.

1.2. Importance of Multi-Agent Generative AI

Multi agent systems (MAS) combined with Generative AI and the supporting breakthrough in automated data engineering and governance and analytics is a revolutionary approach. The combination of autonomous agent coordination and state of the art AI application can help organizations deal with the barriers of the large-scale, dynamic and heterogeneous data environments. The significance of such integration can be emphasized on a number of levels:

- **Enhanced Automation and Efficiency:** The systems of multi-agent generative AI allow automatizing previously repetitive and time-consuming operations, including data collection, cleaning, transformation, and feature engineering. Generative AI models have the ability to produce synthetic data, forecast the missing values and produce optimized features, and agents arrange these processes at the same time. This jointly significantly minimizes human involvement, speeds up the use of pipes, and enhances operational efficiency with large and multi-faceted data.
- **Improved Data Quality and Reliability:** Generative AI helps in increasing the quality of the data, by detecting anomalies, filling in gaps in data, and generating representative samples where data classes are underserved. When these capabilities are implemented in a multi-agent system, it will guarantee that the preprocessing and validation activities are used in all datasets. The data accuracy, completeness and consistency that results increase the downstream analytics and prediction models reliability.

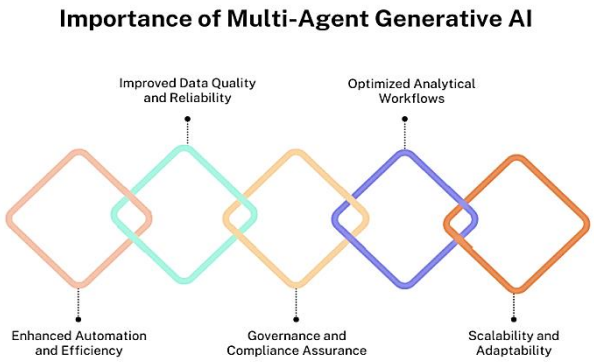

Fig 1: Importance of Multi-Agent Generative AI

- **Governance and Compliance Assurance:** The information or data governance in the enterprise environment is a highly sensitive issue in which regulatory compliance, data privacy, and security need to be upheld. Multi-agent generative AI systems are capable of enforcing policies, tracing the history of data, addressing abnormalities, and product compliance with organizational and regulatory norms in real-time. The framework also offers proactive compliance and mitigates the possibility of violations by spreading the governance functions among autonomous parties.
- **Optimized Analytical Workflows:** The optimization methods and reinforcement learning techniques instilled into the generative AI systems comprising multiple agents enable continual enhancement of analytical systems by the agents. Agents are able to pick models dynamically, to tune hyperparameters to resources efficiently resulting in better predictive performance, speeds and consuming less resources.
- **Scalability and Adaptability:** The distributed character of MAS, which together with the innovative problem-solving potential of generative AI guarantees that the framework will be able to infiltrate large volumes of data and as well as respond to the changing conditions of data. The agents are able to independently manage the various data sources, react to dynamic situations, and organize the detailed work flows without high levels of human oversight.

1.3. Framework for Automated Data Engineering, Governance, and Analytical Optimization

The suggested model of automated data engineering, governance and analytical optimization combines multi-agent systems (MAS) with generative AI to formulate an integrated, [4,5] intelligent and adaptive space of data management. The structure is meant to overcome the shortcomings of conventional pipelines flow which may necessitate a lot of manual intervention, perform poorly during non-homogenous data sources and lack the ability to dynamically react to varying conditions or to deeply govern such a structure consistently. The framework provides scalability of large enterprise datasets by use of parallel and distributed task execution by using autonomous agents coordinated via a multi-agent architecture. The framework is built on the fundamental principles of a set of modular layers, each of which is specialized in one feature of the data lifecycle. Levels of data engineering Ingestion, cleaning, transformation, and integration Data engineering layer Data engineering handles data ingestion, cleaning, transformation, and integration with generative AI improving data ingestion predictive imputation, synthetic data, and automatic feature engineering. This provides a guarantee that the data coming into analytical processes is complete, quality and is formatted in a way that maximizes its modeling use in the downstream. To ensure adherence to regulatory and organization policies, the governance layer ensures that data lineage is observed at all times, metadata tags are enforced, anomalies are identified and suitable measures taken to prevent the occurrence of violations. The framework ensures real-time compliance by ensuring governance is part of the work process instead of governance as a process. The analysis optimization layer uses reinforcement learning and other Artificial Intelligence-based approaches to optimize the selection of models, the adjustment of hyperparameters, and the allocation of computational resources. The agents in this layer modify the workflows dynamically, according to real performance feedback and care a lot to guarantee high predictive accuracy, also with minimum processing time, and resource consumption. The combination of the generative AI and MAS enables the framework to work in an adaptive manner making independent decisions and remaining coordinated across tasks and levels. On the whole, this framework offers a single, smart way of automating the end-to-end data pipeline. It improves efficiency of operations, maintains high quality of data, imposes compliance and optimizes the result of analysis. Its scalable and modular structure and design allow it appropriate in contemporary enterprises that need to encounter large, diverse, and

dynamic data surroundings and, in this regard, determine autonomous, trustworthy, and high-performance data-driven judgments.

2. Literature Survey

2.1. Multi-Agent Systems in Data Management

Multi-agent systems (MAS) are systems composed of autonomous and intelligent agents that engage themselves and the surrounding environment to accomplish desired objectives. [6-9] Such systems are widely considered in the distributed computing context, where agents collaborate and strive to achieve the most optimal load balancing, a rational allocation of resources, and complex workflow management. Other possible benefits of MAS in data management include decentralized decision making, scalability and responsiveness to dynamic environment. Nevertheless, the use of MAS in traditional data engineering operations (i.e., data integration, data transformation, and analytics) is not well-established yet, and the opportunities of data pipelines driven by agents remain largely untapped, despite their growing popularity in application in network management and robotics.

2.2. Generative AI in Data Engineering

Generative AIs have quickly enhanced the power of machines to create high-quality content, both text based and image based as well as structured data. Models such as large language models (e.g., GPT) or diffusion-based networks have demonstrated usefulness in predictive models, synthetic data (generation) and/or automated data cleaning and/or feature engineering. Generative AI can be used in data engineering to minimize human participation through the generation of realistic datasets to test and ensure databases are filled with missing values and generating features that optimise model accuracy. Regardless of these functions, generative AI so far remains an idea to integrate into multi-agent systems to realise end-to-end automation of data workflows, which is a promising field of study that involves integrating creativity with intelligent system coordination.

2.3. Data Governance and Compliance Automation

Data governance is a provision that maintains that the data in an organization is precise, secure and in accordance with regulatory provisions. The conventional methods of governance are manual auditing, tagging, enforcing policies, and doing that may take up too much time and prone to errors. Recent developments use AI and machine learning to execute governance functions within AI including anomaly detection, metadata management, policy compliance checks, among others. These AI-based processes may be combined with multi-agent systems to facilitate proactive governing to allow the agents to check the quality of the data in real-time, identify cases of possible compliance violations, and enforce dynamically policies in dispersed environments, thereby enhancing efficiency and minimizing operational risks.

2.4. Analytical Workflow Optimization

Contemporary analytical pipelines consist of several phases, such as the preprocessing of data, computing features, selecting models, and deployment, the second and third, and the fifth stages, to be optimized attentively to guarantee precision and effectiveness. Automated machine learning (AutoML) and reinforcement learning techniques have been shown to optimize such processes, such as feature selection, hyperparameter selection, and resource allocation. With these optimization techniques integrated in a multi-agent system, agents are able to dynamically partition workflows, utilize computation and thereby coordinate animal labor, and utilise performance measurements in real-time to dynamically adjust workflows. Such integration is able to improve the predictive accuracy, decrease human control, and adopt quicker and more effective analytical decisions.

3. Methodology

3.1. Framework Architecture

The suggested MAGAI (Multi-Agent Generative AI) framework is the one aimed at the combination of multi-agent systems and generative AI to automatize and streamline the data workflows. [7-22] It is made up of four significant layers, the responsibilities of which are limited to certain functions in an end-to-end data pipeline.

- **Data Acquisition Layer:** The Data Acquisition Layer will take care of collecting data in multiple sources, such as structured and unstructured ones (text, images, sensor streams, etc). The autonomous agents of this layer can perform identification of the relevant sources, data extracting, and updates in real time. Using multi-agent coordination, this layer is adequate to collect quality data fully and at the correct time and reduce redundancy and address inconsistency across various sources.
- **Data Engineering Layer:** The Data Engineering Layer contains specialized agents that process important preprocessing tasks, including data cleaning, integration, transformation and validation. These agents have the ability to identify and correct errors automatically, put heterogeneous data into normal form, and write data to be used in downstream analysis. Generative AI models may be used to fill in missing values, generate synthetic records, or any other feature, and make the data engineering process efficient and more accurate. It is a layer through which the data used in the analytical workflows is uniform, dependable, and geared towards machine learning or business intelligence operations.

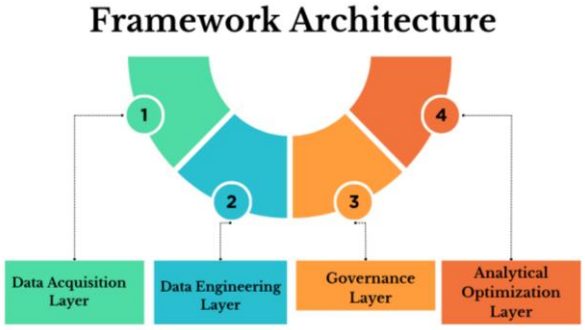

Fig 2: Framework Architecture

- **Governance Layer:** The Governance Layer is concerned with data policy enforcement and security as well as compliance with regulatory needs. The enforcement agents of the policy also monitor the data usage, perform automated tagging as a metadata management tool, and probe abnormalities which might be reflected by policy violations. This layer will allow proactively combating risks, ensuring the standard of data quality and assisting with audits through the combination of AI-driven governance and multi-agent coordination, minimizing human supervising of this area and increasing the organizational confidence in the data-driven decision-making process.
- **Analytical Optimization Layer:** The Analytical Optimization Layer uses smart agents to make the analysis processes more efficient and effective. This group of agents employs reinforcement learning methods and AutoML to optimize the choice of model, hyperparameters, feature engineering and resource allocation. Constant check of model efficiency and correction of workflow tactics by the layer provide the higher accuracy of prediction, quicker computing, and reasonable consumption of computer resources. This layer can be integrated with the multi-agent ecosystem to perform dynamic, autonomous optimization of the whole data pipeline.

3.2. Agent Design

Under the MAGAI system, every agent is created to be an autonomous intelligent agent who can perceive its surroundings, think about what to do, coordinate with others, and do tasks according to the data pipe. The design is made up of four major modules which are also coordinated to ensure efficient, adaptive and intelligent functioning.

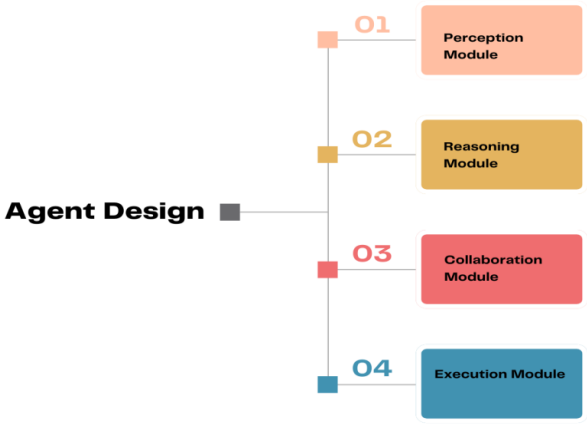

Fig 3: Agent Design

- **Perception Module:** The Perception Module allows the agent to watch over its environment and gather pertinent data on it in a continuous manner. This involves monitoring the condition of datasets, detecting anomalies, monitoring the progress of a workflow and identifying the differences in input sources. Through a long-term awareness of the environment, the agent is able to make sound decisions and react dynamically to the changes in the data pipeline. The perception module is the senses of the agent which gives the information required in all the processes of decision making.
- **Reasoning Module:** The Reasoning Module takes advantage of the models of generative AI to interpret the received data and suggest the necessary measures. Through predictive modeling, pattern recognition and scenario simulation, the agent may be able to anticipate potential problems, optimise data transformations or may produce synthetic data to address gaps. This module enables the agent to operate proactively as opposed to reactively enhancing the effectiveness, precision, and integrity of the data pipeline processes.
- **Collaboration Module:** The Collaboration Module enables the agents to interact and liaise with other agents within the system. This involves sharing knowledge, negotiating task priorities, workloads as well as solving conflicts. By facilitating multi-agent cooperation, the system is able to carry out complex goals which need collaborative response as is the case of coordinating massive data harmonization or implementation of governance principles over spread out datasets. Proper cooperation can guarantee whole pipeline functioning, which is not based on the bottlenecks or unnecessary calculations.
- **Execution Module:** The action propose by the reasoning module is what the Execution Module undertakes. These encompass data transformation application, model training, metadata updates, or governance rule enforcement. The module guarantees that the informed decision is converted into reality within the data pipeline in form of measurable outcomes. The agent can also give feedback to its perception and reasoning modules by continually assessing the execution outcomes, and this provides a closed-loop architecture that makes the agent more adaptable and can improve performance in the long run.

3.3. Workflow Automation

In the MAGAI model, the agents work together to automate the whole data process, including ingestion and optimization of analytics. [11-13] Each step is undertaken by special agents, which sense, think, cooperate, and are coordinated to execute their activities with efficiency, accuracy and integrity all along the pipeline.

Fig 4: Workflow Automation

- **Data Ingestion:** In Data Ingestion, data are collected by agents on a variety of sources, such as structured databases, APIs, logs, and unstructured sources as, e.g. text, images, or sensor streams. Agents keep track of the availability of sources and extract pertinent datasets and update data to real time. Distributing the ingestion tasks to the agents, guarantees the full coverage, minimizing the latency, and reduces the redundancy of the information, providing a well-developed basis in the back-end processing.
- **Data Cleaning:** At the Data Cleaning stage, data quality problems will be automatically detected and solved by agents. This involves assigning missing values, erasing duplicates, outlier detection and management, and the resolution of inconsistencies. Generative AI may be used to help fill in missing values of the records or create synthetic data of cases that lack representation. These tasks are automated and it lowers the amount of manual work, improves the quality of data, and makes sure that further analysis will be conducted with the help of qualitative data.
- **Data Transformation:** Data Transformation stage is the phase where data is ready to be analyzed using feature engineering, measuring numerical values, categorical variables with codes, and distributions normalization. Agents enforce a set of transformation rules on datasets and can change them dynamically as workflow needs change. This makes sure that information is in a form that is usable by machine learning models, analysis queries, or reporting systems to enhance performance and understanding of the model.
- **Governance Enforcement:** The Governing Enforcement agents are also in charge of keeping a watch on data Policies and regulatory complies. They monitor data provenance, proper metadata labelling and compliance with security policy and privacy rules. Such agents are also able to recognize misbehaviors or breaches of the policy in real time and provide alerts, or automatically fix the problem. This proactive solution establishes a compliance chain of the data, which is auditable and can be trusted.
- **Analytical Optimization:** During the Analytical Optimization phase, analytical processes are optimized and made more efficient and precise by readers. They choose the right models, optimize hyperparameters and reinforcement learning or AutoML to help them verify the results. Agents may also assign computing resources dynamically, respond to performance feedback and coordinate with other agents, in order to optimally execute workflows. This step makes analyses to be sound, scalable, and with the ability to produce high quality insights with the least human intervention.

3.4. Generative AI Integration

Generative AI is an important component of the MAGAI model, which improves the intelligence, flexibility, and efficiency of multi-agent data processes. The predictive imputation of missing data is one of its main uses; general generative models are used to study patterns observed in existing datasets to come up with reasonable values of incomplete records. In contrast to traditional statistical imputation techniques which use simple means or regression models, generative AI is able to recover non-linear and complex relationships between variables, leading to superior imputation which is also more accurate and context-sensitive. This is especially useful when working with large, heterogeneous datasets in which bad data are likely to have serious adverse effects on analytical performance when left unaddressed or counterproductively. Synthetic data generation is another vital usage that is used to solve the problem of underrepresented classes or scarce data conditions. The synthetic examples of the generative AI models, such as GANs and diffusion-based networks, are realistic and high-fidelity and mirror the statistical characteristics of the original data. The framework will have balanced training sets, which results in lower bias, enhanced predictive accuracy of machine learning models by increasing the size of their datasets. Privacy is another benefit of synthetic data; synthetic data enables models to train on synthetic data, without exposure to sensitive and proprietary data, thereby facilitating regulatory compliance and data governance. Lastly, automated feature engineering is an area where generative AI provides value to a very time-consuming human-based engine with a high level of knowledge about the domain. Generative models can suggest new features, transformations, or embeddings based on correlation, interactions, and latent patterns in data that help enhance the behavior of a model. Through this functionality, the agents would be able to optimize the analytical processes used without needing to resort to manual actions and speed up the process of model development as a whole. A combination of generative AI in all these dimensions, including imputation, synthetic augmentation, and feature engineering, does not only enhance the quality and integrity of data in its representations, but also equips agents to conduct a more advanced, active cognitive process, resulting in more dependable, effective, and scalable data-driven decision-making.

3.5. Optimization via Reinforcement Learning

MAGAI framework uses reinforcement learning (RL) actors to streamline the analytical processes, which guarantees excellent model performance, efficiency, and compliance. [24-16] Here, every agent receives and responds to decisions regarding task scheduling, resource allocation, model selection and data transformations and thus maintains an ever-changing interaction with the data pipeline environment. The system of reward governs the behavior of the agent; these functions quantitatively assess the consequences of its actions and it gives feedback to make better decisions in future. The reward mechanism enables a threefold balance of three important goals model accuracy, processing efficiency, and governance compliance. These factors are formally taken as weighted to determine the reward. Accuracy is the first, it is the performance of the predictive or analytical models. An increased accuracy is a positive attribute to the reward whereby the agents tend to pursue strategies that will lead to high predictive quality. Second, processing time is the aggregate time that it takes to run the data pipeline which includes data cleaning, data transformation and model training. Increasing the processing time lowers the reward, thereby encouraging agents to maximize their computational efficiency and the ability to optimize execution of workflow. Third, governance violations monitor the cases when the policy/compliance rules are not satisfied, including violation of data privacy, security measures, or metadata. Violations penalize the reward making sure that the agents focus on adherence to governance standards, in addition to performance and efficiency. The reward function uses weighting parameters, usually represented by alpha, beta, and gamma, to enable the system designers to balance the relative significance of accuracy, speed and compliance. Using these parameters, the framework can be adjusted to various organizational requirements, including focusing on faster results in a high-throughput system or increased compliance in a controlled industry. In iterative learning, the RL agents consider all possibilities, which are then tested by the reward function and finally come to an optimal solution. This strategy allows self-directed, dynamic and adaptable optimization of analytical processes, leading to a better model performance, shorter execution time and proactive governance policy enforcement, within a single multi-agent system.

4. Results and Discussion

4.1. Experimental Setup

In order to estimate the work of the suggested MAGAI framework, the series of experiments were performed using large-scale enterprise data that contained both structured, [17-20] semi-structured, and unstructured data sources. Different types of structured data were the relational databases containing numerical and categorical characteristics, semi-structured data were the records of the different enterprise applications in the form of JSON and XML, and unstructured was the textual reports, logs, and files of the documents. The variety of data types made it possible to test the multi-agent capabilities, integrating generative AI, and workflow optimization mechanisms of a framework under conditions of realistic, heterogeneous data. The MAGAI framework was against the performance of two baseline approaches. The original baseline was composed of the traditional and manually constructed data engineering pipelines where the human operators achieved their goals of data cleaning, transformation, feature engineering, and governance enforcement through standard scripts and tools. The second baseline was automated machine learning (AutoML) pipelines that were not coordinated with a multi-agent system (MAS). AutoML was not as collaborative, adaptive or governance-conscious as the capabilities of the MAGAI framework, though it did offer automated model selection and hyperparameter tuning. Relation to these baselines made it possible to have a clear evaluation of the benefits presented by agent-based coordination, generative AI improvement, and optimization carried out by reinforcement learning. The chosen evaluation metrics included one that was to measure the effectiveness of the framework in various dimensions. The time spent in pipelines was used to measure the performance of the end-to-end data processing, such as ingestion, cleaning, transformation, and model training. Data quality score considered accuracy, completeness and consistency of data that have been processed. Compliance rate was a measure of compliance with governance policies, data lineage tracking and regulatory requirements. Lastly, the predictive ability of the models constructed on the resulting data was evaluated through analytical accuracy. These measures allowed the experiment to have an overall picture of the performance of the framework, which showed the benefits of speed, reliability, compliance, and analytical performance in relation to traditional and AutoML only bases.

4.2. Performance Analysis

Table 1: Performance Analysis

Metric	Improvement
Execution Time	66%
Data Quality Score	16%
Compliance Rate	17%
Analytical Accuracy	7%

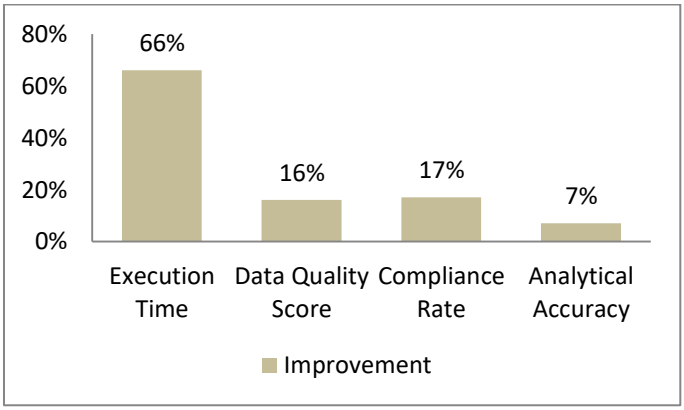

Fig 5: Graph Representing Performance Analysis

- **Execution Time:** MAGAI framework also indicated a great enhancement of the pipeline execution time and the total processing time was minimized by an approximate of 66 times as

compared to the methods used in the baseline. This is an efficiency increase is due, in large part, to the synchronized work of multi-agent systems which parallelize the actions like data ingestion, cleaning, and transformation. Also, the agents of reinforcement learning could optimize the scheduling of workflow and the allocation of resources dynamically and reduced idle time and redundant calculations. Rapid analysis enables an analytical process to proceed at a higher pace, as well as increases the scalability of the framework to large, heterogeneous datasets.

- **Data Quality Score:** There was an increase in the quality of data collected by 16 percent compared to standard and AutoML-only pipelines. Generative AI was also central, as it was used to fill in missing values, create synthetic data, and help with automated feature engineering, due to representing a minor group. The multi-agent coordination factor resulted in application of data verification and cleansing tasks that were uniformly applied on all the datasets and this minimized errors, inconsistencies as well as redundancy. Such a better quality data helps to increase the reliability and trust of down-stream analysis data, which leads to better-grounded decision-making.
- **Compliance Rate:** The rate of compliance becoming 17% higher shows that the framework is successful in its enforcement of data management policies and regulations. Governance agents kept track of data lineage, enacted policy rules, and identified or reported violations instantly. This active strategy made sure that the datasets were safe and auditable and met the organization-wide requirements, minimizing the number of people controlling the data and the likelihood of violating regulatory regulations. The factor of governance influence into the automated working process underlines the benefits of MAS and AI-performed monitoring.
- **Analytical Accuracy:** The analytical accuracy was increased by 7% and was reflecting better model performance in terms of the overall effect between the well-formed and high-quality data and the streamlined workflow execution. The reinforcement learning agents optimized the choice of the model, hyperparameters, and feature processing, and generative AI was used to generate strong and representative input processing. Though this increment in accuracy is not huge in comparison with the improvements in the execution and governance area, it is an indication that the framework manages to balance efficiency, compliance, and predictive performance, providing reliable insights in an automated and streamlined system.

4.3. Discussion

The experimental findings prove the existence of the significant benefits of the MAGAI framework in several dimensions of data pipeline performance, which illustrates the synergistic nature of the combination of multi-agent systems, generative AI, and reinforcement learning. Their reduction of execution time should be considered as one of the most significant consequences. Through the use of multi agent coordination, functionality like data ingestion, cleaning, transformation, and governance enforcement can now be performed in parallel instead of a sequential way. The parallelism minimizes bottlenecks and provides efficiency in computational resources which is particularly critical to large volume and variety of enterprise datasets. The dynamism with which agents can distribute tasks and respond to the pipeline condition increases responsiveness and throughput even more. The framework enhances the quality and consistency of data besides efficiency improvements. Generative AI models can be used to predictively fill in the missing values, create synthetic data of the underrepresented classes, and do automated feature engineering. These properties guarantee that data sets are complete, representative and are well prepared when required to accomplish downstream tasks of analysis. Input data of high quality will not only result in more predictive models and insights of higher quality, but will also contribute to decreasing the amount of manual work required to execute repetitive preprocessing operations on the data. The system of multi-agents also makes sure that such changes are always uniformly applied to heterogeneous data sources, which enhances data integrity in general. There is also enhanced governance and compliance within the framework. The agents of policy enforcement observe data lineage, metadata integrity and regulatory compliance automatically finding and correcting violations in real time. Such a proactive strategy reduces the need of human supervision, the lack of consequences in case of non-observance, and the compliance of the pipeline with organizational norms. Lastly, reinforcement learning agents also optimize analytical processes through an ever-assessing process of work performance and modifying the approach to a reward function that strikes a balance between accuracy, efficiency, and compliance with governance. This loop optimization guarantees the framework both provides higher quality and faster data as well as, it can generate predictive models that are optimized in terms of accuracy and resource usage. Altogether, the MAGAI framework is shown as a holistic and smart strategy to data pipeline management in the automated format which offers quantifiable gains in terms of efficiency, quality, compliance, and analytical results.

5. Conclusion

This paper introduces MAGAI (Multi-agent generative AI), which is an elaborate and smart mechanism of incorporating data engineering and governance and analytical optimization of heavy enterprise settings. The framework overcomes the drawbacks of traditional data pipelines which may integrate manual interventions, fixed workflows, as well as independent stages of processing by combining multi-agent system with generative AI models. With the MAGAI framework, autonomous agents are able to perceive the environment, reason with advanced AI, work with other agents, and perform along the whole data

lifecycle, and this creates one of the most efficient, adaptable and scalable systems.

Experimental tests prove that MAGAI can improve pipeline efficiency and save more than 60% of the execution time with the help of the old manual pipelines and AutoML-only pipelines. In multi-agent coordination, data acquisition, cleaning, transformation and governance activities can be done simultaneously, which reduces the bottlenecks and reduces the utilization of computational capabilities. Meanwhile, generative AI integration allows to improve the quality of data, including forecasting imputation, creating artificial samples in underrepresented classes, and an automatic feature engineer. These abilities are guaranteed to have complete, representative datasets, which are optimized to downstream analytical processes, and eventually enhance the performance of predictive models.

The framework also provides greater compliance in governance as the AI-driven policy enforcement is engraved into the working process. Using agents, data lineage is continuously monitored, anomalies detected, security measures implemented, as well as a maintenance of compliance to regulatory standards on a real time basis. This active governance will minimize the human management aspect, reduce the risk of data breach or even any form of policy violation and guarantee that the organization trusts automated data management. Besides, optimization using reinforcement learning allows agents to adapt workflow strategies in a dynamic manner that considers accuracy and processing efficiency, as well as compliance using a well-crafted reward function. Closed-loop optimization is used to make analytical outputs strong, cost-effective, and business-oriented.

In addition to these short term gains, MAGAI offers a scalable platform on which future upgrades can be implemented. The data integration can be streamed in real time so that the framework could be used in the dynamic environment and support the life-long learning and the adaptive decision-making. Adding advanced explainable AI (XAI) modules would enhance transparency and interpretability, so the stakeholders could learn more about why the agents act the way they do and why models predict the actions. Also, the continuity of studies on making the agents more flexible and able to interact with each other may also enhance the flexibility of the system to unexpected shifts in the data patterns, demands during work, or even regulations.

To sum up, the MAGAI framework is an important breakthrough in automatic and smart data management. Through collaborative multi-agent systems, generative AI, and reinforcement learning, it does not only simplify data pipelines and improves the results of the analytics but also guarantees governance compliance, adaptability, and scalability. It has a bright potential as a platform in the future generation of autonomous and end-to-end systems of data engineering and analytics due to its modular design and artificial intelligence-driven capabilities to scale to the needs of ever-changing enterprises.

References

1. Jaleel, H. Q., Stephan, J. J., & Naji, S. A. (2020). Multi-Agent Systems: A Review Study. Ibn AL-Haitham Journal For Pure and Applied Sciences, 33(3), 188-214.
2. Sambath Narayanan, D. B. G. (2025). AI-Driven Data Engineering Workflows for Dynamic ETL Optimization in Cloud-Native Data Analytics Ecosystems. American International Journal of Computer Science and Technology, 7(3), 99-109. https://doi.org/10.63282/3117-5481/AIJCST-V7I3P108
3. Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-agent systems: A survey. Ieee Access, 6, 28573-28593.
4. Kossek, M., & Stefanovic, M. (2024). Survey of recent results in privacy-preserving mechanisms for multi-agent systems. Journal of Intelligent & Robotic Systems, 110(3), 129.
5. Ibrahim, M., Al Khalil, Y., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., ... & Dumontier, M. (2025). Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. Computers in biology and medicine, 189, 109834.
6. Chen, Y., Yan, Z., & Zhu, Y. (2023). A unified framework for generative data augmentation: A comprehensive survey. arXiv preprint arXiv:2310.00277.
7. Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. Mathematics, 10(15), 2733.
8. Nadal, S., Jovanovic, P., Bilalli, B., & Romero, O. (2022). Operationalizing and automating data governance. Journal of big data, 9(1), 117.
9. Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). The Importance of AI Data Governance in Large Language Models. Big Data and Cognitive Computing, 9(6), 147.
10. Sambath Narayanan, D. B. G. (2024). Data Engineering for Responsible AI: Architecting Ethical and Transparent Analytical Pipelines. International Journal of Emerging Research in Engineering and Technology, 5(3), 97-105. https://doi.org/10.63282/3050-922X.IJERET-V5I3P110
11. Mahi, B. S. (2019). AI-Driven Metadata Management: The Future of Data Governance.
12. Vaddepalli, R. K. (2025). Smart Governance for AI: Can Metadata Automation Keep Up with Real-Time ML Pipelines?. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 6(2), 119-124.
13. Zou, H., Zhao, Q., Bariah, L., Bennis, M., & Debbah, M. (2023). Wireless multi-agent generative AI: From connected intelligence to collective intelligence. arXiv preprint arXiv:2307.02757.
14. Sambath Narayanan, D. B. G. (2025). Generative AI–Enabled Intelligent Query Optimization for Large-Scale Data Analytics Platforms. International Journal of Artificial Intelligence, Data Science, and Machine

Learning, 6(2), 153-160. https://doi.org/10.63282/3050-9262.IJAIDSML-V6I2P117

15. Athanasiadis, I. N., Milis, M., Mitkas, P. A., & Michaelides, S. C. (2009). A multi-agent system for meteorological radar data management and decision support. Environmental Modelling & Software, 24(11), 1264-1273.
16. Zhang, H. S., Zhang, Y., Xu, J. H., Xiao, D. Y., & Hu, D. C. (2003, October). Study on ITS data management based on multi-agent systems. In Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems (Vol. 1, pp. 183-187). IEEE.
17. Ladley, J. (2019). Data governance: How to design, deploy, and sustain an effective data governance program. Academic Press.
18. Omicini, A., & Mariani, S. (2013). Agents & multiagent systems: En route towards complex intelligent systems. Intelligenza Artificiale, 7(2), 153-164.
19. Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., ... & Amira, A. (2023). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. Artificial intelligence review, 56(6), 4929-5021.
20. Adeyinka, T. I., & Adeyinka, K. I. (2025). Intelligent, Autonomous, and Multi-Agents. In Humans and Generative AI Tools for Collaborative Intelligence (pp. 301-326). IGI Global Scientific Publishing.
21. Sambath Narayanan, D. B. G. (2025). Semantic Layer Construction in Data Warehouses Using GenAI for Contextualized Analytical Query Processing. *American International Journal of Computer Science and Technology*, *7*(4), 93-102. https://doi.org/10.63282/3117-5481/AIJCST-V7I4P108
22. Yang, W., Fu, R., Bilal Amin, M., & Kang, B. (2025). Impact and influence of modern AI in metadata management. arXiv e-prints, arXiv-2501.
23. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., ... & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680.
24. Pradeep, A., Rustamov, A., Shokirov, X., Ibragimovna, G. T., Farkhadovna, S. U., & Medetovna, A. F. (2023, August). Enhancing Data Engineering and Accelerating Learning through Intelligent Automation. In 2023 Second International Conference on Trends in Electrical, Electronics, and Computer Engineering (TEECCON) (pp. 104-110). IEEE.