AI-Driven Federated Data Governance: Building Trustworthy and Sustainable Digital Ecosystems

Ravindra Putchakayala[1*], Rohit Yallavula[2]

[1] Sr. Software Engineer, U.S. Bank, Dallas, TX, UNITED STATES
[2] Data Governance Analyst Kemper, Dallas, TX, UNITED STATES

*Corresponding Author Email: ravindra.putchakayala25@gmail.com

ABSTRACT

Through a data utilisation value chain, AI/Data commons will promote the construction of an open collaborative ecosystem based on PCI (Participation- Collaboration-Incentives). This will be accomplished by preserving data sovereignty and protecting sensitive data. Furthermore, it is capable of resolving a wide range of user-defined and customised AI issues. This study presents the design of a high-level functional architecture for trustworthy artificial intelligence and data commons resources.

Keywords: AI-Driven Data Governance; Federated Governance; Privacy-Preserving Architecture; Responsible AI Systems; API-Centric Controls

Introduction

As the paradigm of data management transforms through human, machine, and algorithmic contributions, data creators and proprietors must exercise enhanced control. The volume of information linked to individuals via various digital devices will increase significantly. According to IDC (International Data Corporation), the volume of data is projected to increase tenfold to 162 ZB between 2016 and 2025. 1 User devices generate the largest share of data, primarily consisting of personal information and relevant legislation. It is imperative to improve the beneficial cycle of data utilization by reallocating data management rights from platform operators to data producers/owners, alongside the distribution of values and responsibilities for data usage [1].

Moreover, it is essential to create IoT data and AI management tools that adhere to the EU General Data Protection Regulations (GDPR) to access international markets. The EU seeks to promote the unimpeded transfer of personal data through GDPR while strengthening individuals' rights to protect their personal information in response to changes in personal information management paradigms. Various aspects have been legalized to protect the rights of the data subject, with severe penalties for violators, amounting to 4% of global annual revenue or 20 million euros (e.g., Right to be informed, Right of access, Right to rectification, Right to restrict processing, Right to erasure, Right to data portability, Right to object, Right concerning automated decision-making and profiling).

Furthermore, a marketplace is essential for creating a data commons ecosystem that enables the exchange and utilization of IoT data between data owners and users. Personal information is emerging. To provide intelligent services such as smart cities, the deployment of extensive data analysis and management via AI is crucial; however, practical limitations emerge due to the sensitivity of personal information in data collection, sharing, and utilization. The growing perception of online personal data as economic assets requires the

shared economy model to facilitate transactions, going beyond the mere protection of personal information [2].

Establishing an AI/Data sandbox environment is crucial for facilitating confident collaboration between data owners and AI providers to implement intelligent services. An AI marketplace is necessary to enable the discovery and application of AI modules or the custom development of AI modules within a PCI (Partition-Collaboration-Incentives) framework, specifically designed for data owners limited by data transfer restrictions, encompassing personal and corporate information. Furthermore, it necessitates Data-Network-AI (DNA) matchmaking technology to ensure that network and computing resources are readily available in an incentivized manner to meet the needs of IoT data owners and AI providers.

In late January 2020, ITU-T held a meeting to formulate technical definitions and standards for AI/Data commons. At this summit, ITU-T articulated that AI/Data commons signify a collaborative approach in which various entities share their data, AI modules, and computational resources to collectively tackle challenges. Data commons enable sharing, whereas AI commons foster collaboration. Furthermore, they have considered the establishment of focus groups consisting of four working groups: The Working Group on Repositories, the Working Group on Marketplaces, the Working Group on Infrastructure, and the Working Group on Benchmarks. However, a precise technical definition of AI/Data commons has not yet been formulated [3].

This paper delineates a comprehensive functional architecture for reliable AI and data commons, aimed at fostering an open collaborative ecosystem predicated on Participation-Collaboration-Incentives (PCI). This framework tackles various user-defined AI challenges via a data utilization value chain, ensuring data sovereignty and protecting sensitive information.

AI-Data Commons

AI/Data commons denote a technological framework for AI/Data sandboxes that enables dependable AI analysis and intelligent services via collaboration with IoT data owners and AI providers, while addressing data sovereignty, privacy, and incentives within network and computational infrastructure.

AI commons serve as a foundational technology that creates a marketplace for AI providers to interact and collaborate incentivally, enabling the deployment of dependable AI modules in environments like edge nodes for IoT data analysis. Data commons represent a user-centric, decentralized framework for data management and transactions that amplifies the control rights of IoT data proprietors, facilitating the establishment and dissemination of privacy levels and transmission/use parameters, along with the distribution of value to owners. AI/Data commons are collaborative ecosystems that enable interaction between IoT data owners and AI providers, promoting the creation of various innovative services, generating significant revenue, and ensuring compatibility through ultra-low latency and high-capacity network and computing infrastructure [4].

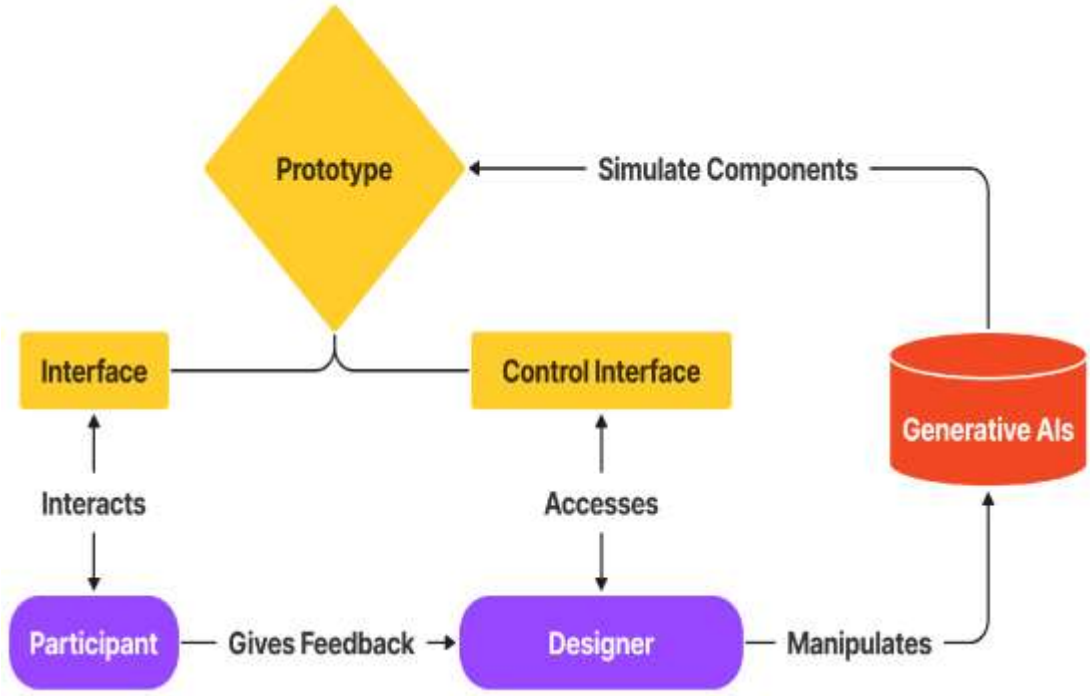

Figure 1 depicts the concept of the AI/Data commons framework.

At present, data cannot be effectively utilized due to issues like data sovereignty and privacy infringements; therefore, there is a need for ecosystems that facilitate data sharing and utilization, allowing data providers to engage with confidence. Additionally, we necessitate an open, collaborative ecosystem (Participation-Collaboration-Incentives) that collectively tackles various user-defined customized AI challenges through voluntary engagement, sharing of personal experiences and resources, and equitable rewards proportional to individual contributions [5].

Projects and solutions related to AI/Data commons are as follows. DECODE(EU) is a project that utilizes blockchain technology to enforce General Privacy Regulations. This solution is a decentralized data ecosystem that emphasizes transparency, privacy, and the safeguarding of citizens' data rights. This framework is insufficient for addressing problems and the notion of PCI. The Open Commons Consortium (USA) provides essential environments for individuals lacking data or computational resources through cloud platforms for open research data. This demonstrates insufficient data sensitivity protection, an inadequate problem-solving approach, and a flawed comprehension of PCI. OCEAN Protocol (Singapore) is a decentralized platform for artificial intelligence services. This solution is a decentralized data ecosystem that employs blockchain technology, fosters a collaborative environment, and offers incentives via Ocean Token. Nonetheless, this demonstrates insufficient data sensitivity protection and lacks a comprehensive framework for addressing issues [6].

The notion of PCI [6]. Acumos AI (USA) functions as the artificial intelligence repository of the Linux Foundation. This solution is a platform that provides all the necessary environments for the entire life cycle of developing AI-related services. Nevertheless, this demonstrates shortcomings in data sovereignty guarantees, safeguarding of data sensitivity, and the concept of PCI.

From a comprehensive perspective, existing technologies do not guarantee data sovereignty and sensitivity, as they are primarily controlled by a singular entity. In contrast, to resolve this issue, we propose the creation of PCI-based open collaborative ecosystem technologies via a data utilization value chain, ensuring data sovereignty and protecting sensitive information [7].

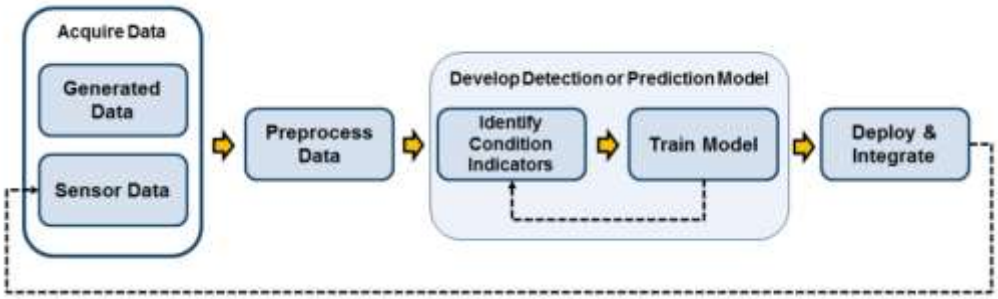

Figure 2. Technical workflow for the AI/Data commons framework.

The AI/Data commons framework consists of the AI/Data commons front-end, the AI/Data commons, and the AI/Data commons networks. The front-end of the AI/Data commons includes problem owners (problem definition tool), problem solvers (AI collaboration sandbox), data/AI modules, compute providers, and consumers (data commons, marketplace). The AI/Data commons encompass an open collaborative environment and marketplace. The PCI-based collaborative execution environment includes data providers, AI module providers, computation providers, storage providers, problem solvers, and outcome evaluators [8].

Our objective is to promote engagement in problem-solving, resource sharing, and collaboration, while offering incentives for contributions, thereby allowing individuals to tackle user-defined, customized AI challenges. Additionally, we intend to improve the ecosystem by distributing a variety of finalized solutions through PCI-based marketplaces.

Advanced Functional Architecture for AI-Data Commons

AI/Data commons technology consists of AI/Data commons participants, AI/Data commons front-end, AI/Data commons framework (including open collaboration and marketplace frameworks), AI/Data commons network, and AI/Data commons infrastructure.

Numerous open collaborative spaces and PCI-based markets have been developed using the commons architecture and the commons marketplace framework. Consequently, a varied group of participants (e.g., problem owners employing a problem definition tool, problem solvers within an AI collaboration sandbox, and data/AI modules alongside compute

providers and consumers in data commons and marketplaces) will tackle user-defined, tailored AI challenges [9].

The open collaboration (PCI) framework execution space is a crucial component of the framework, where participants, including problem owners and solvers, engage in a collaborative process based on Participation-Collaboration-Incentives to tackle various user-defined customized AI challenges.

The marketplace framework is a platform that distributes credible information in compliance with PCI standards. This is a marketplace that distributes data, AI modules, computational resources, and integrated solutions. It includes a user DID (Decentralized Identification)-based asset specification technology, a reliable (i.e., immutable) asset transaction technology, and an asset transaction technology based on user sovereignty control [10].

Common core software is a framework that provides the fundamental functionalities required for developing AI/Data commons. It includes a qualification policy (access control) mechanism for user sovereignty-based data, a data sensitivity protection mechanism, a framework for PCI policy-based open AI collaboration, and virtual environment (sandbox) technology for user sovereignty-based decentralized AI collaboration.

An effective commons network provides functionalities including decentralized identification and the necessary reliability for higher tiers. It encompasses decentralized identification authentication technology for user sovereignty, networking management technology for reliability, token management technology for executing reward policies, and infrastructure interworking technology for computing and storage.

Ultimately, an AI/Data commons infrastructure exists that includes storage, metadata, and tokens, analyzed through the perspective of virtual infrastructure internalization.

Federated Data: Advancing A New Era Of Credible And Reliable Artificial Intelligence

Federated ecology provides a feasible solution to the substantial issues created by isolated data islands due to data privacy and information security regulations in the era of artificial intelligence (AI). Federated data, as the cornerstone of federated ecology, includes data from all nodes within the federation, along with their associated storage, computational, and communication resources. To safeguard privacy, federated data is classified into private and non-private categories, facilitating the implementation of data federalisation via federated management of these datasets.

In data-driven AI technologies, federated data are essential, enabling efficient data retrieval, pre-processing, processing, mining, and visualization for AI applications. It can provide effective solutions for the challenges faced by AI technologies, such as training AI models with insufficient data, improving the generalizability of AI models across various application contexts, and establishing a unified processing workflow for data security and privacy management in AI-driven applications.

Architecture Of Federated Data

Federated data represents a model of parallel data that enables the conversion of extensive datasets into data intelligence. In federated data, data experimentation is essential for determining optimal solutions in alignment with Morton's law, attainable via reinforcement learning and parallel reinforcement learning. Federated data provides an effective solution for data security and privacy issues, regardless of whether in a centralized or distributed framework. Based on federated ecology and federated learning, we propose an architecture for federated data consisting of six components: real data/physical objects, virtual data/digital twins, federated data experimentation, federated fusion, federated security, and credible federated wisdom, as depicted in Fig. 3.

The first two components are employed for handling tangible data from physical entities and virtual data generated by the Digital twins associated with that specific equipment, as well as virtual data produced by other means. The digital twins component is responsible for acquiring real-time data, including the monitoring of operational conditions [11].

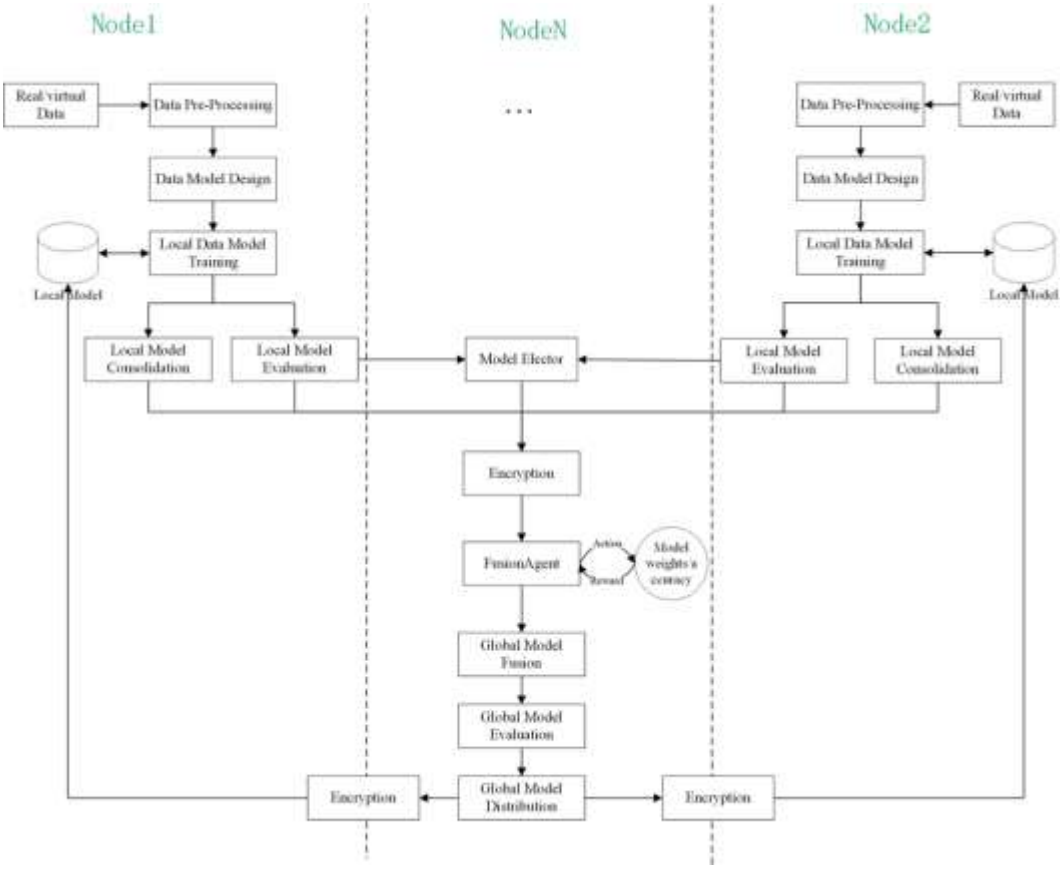

Figure 3. Comprehensive architecture of federated data

Federated data experimentation is an essential method for exploring the optimal model by traversing the solution space, utilizing techniques such as reinforcement learning and parallel reinforcement learning to determine the best parameters for local models.

The federated fusion component integrates local models into a globally optimal model using specialized model fusion techniques, including the dynamic fusion mechanism. The local nodes will subsequently employ the globally optimal model. All transmission operations of these models will utilize diverse encryption technologies or interface with the blockchain to enable model sharing and guarantee the security of the transmission processes.

The federated security component is responsible for executing data security and privacy measures. Data ownership and usage rights are separate; specifically, data for a specific node is stored at its local address, while models trained on this data are transmitted to the federated fusion and federated data experimentation components, similar to federated learning. Moreover, security can be augmented through blockchain technology or alternative encryption techniques.

Reliable federated intelligence can be obtained through federated data experimentation and federation fusion, where data is stored locally and only models are communicated to ensure data security. Federated intelligence can be utilized to govern the functioning of physical entities. Federated data improves AI's self-learning abilities through the feedback loop in its architecture, facilitating the extraction of optimal solutions via experiments with both virtual and real data exchanges.

Federated Data Decomposition

The complex processes of federated data are outlined according to the federated learning standard approved by IEEE, as depicted in Fig. 4, in which multiple nodes can engage in the federated data ecosystem. Each node, designated as client nodes, is capable of participating in the global integration of optimal models, featuring one or more nodes operating as model fusion nodes (referred to as server nodes). Both virtual and real data can function as input in the previously mentioned process. The data is processed to eliminate noise and normalized using standard methods, including min-max normalization and Z-score normalization. A suitable data model will subsequently be developed for the specified dataset using various machine learning techniques, including clustering, classification, and association analysis. During the training process, client nodes will utilize local data to train the model, employing diverse techniques, including model consolidation methods, to improve the model's accuracy by synchronizing its weights layer by layer with those of other nodes. Thereafter, each client node conveys its model parameters to the server nodes for model integration. Before model fusion, the local model trained by each node must be evaluated, after which the model selector will ascertain the permissible parameters for the model in each iterative training session. If the parameters obtained in this training round improve the model's accuracy, they will be integrated into the model fusion; otherwise, they will be rejected by the model selector, as these parameters may compromise the accuracy of the fused model. Since federated data functions in a decentralized manner, the choice of diverse nodes to contribute their parameters to the model is essential for model integration.

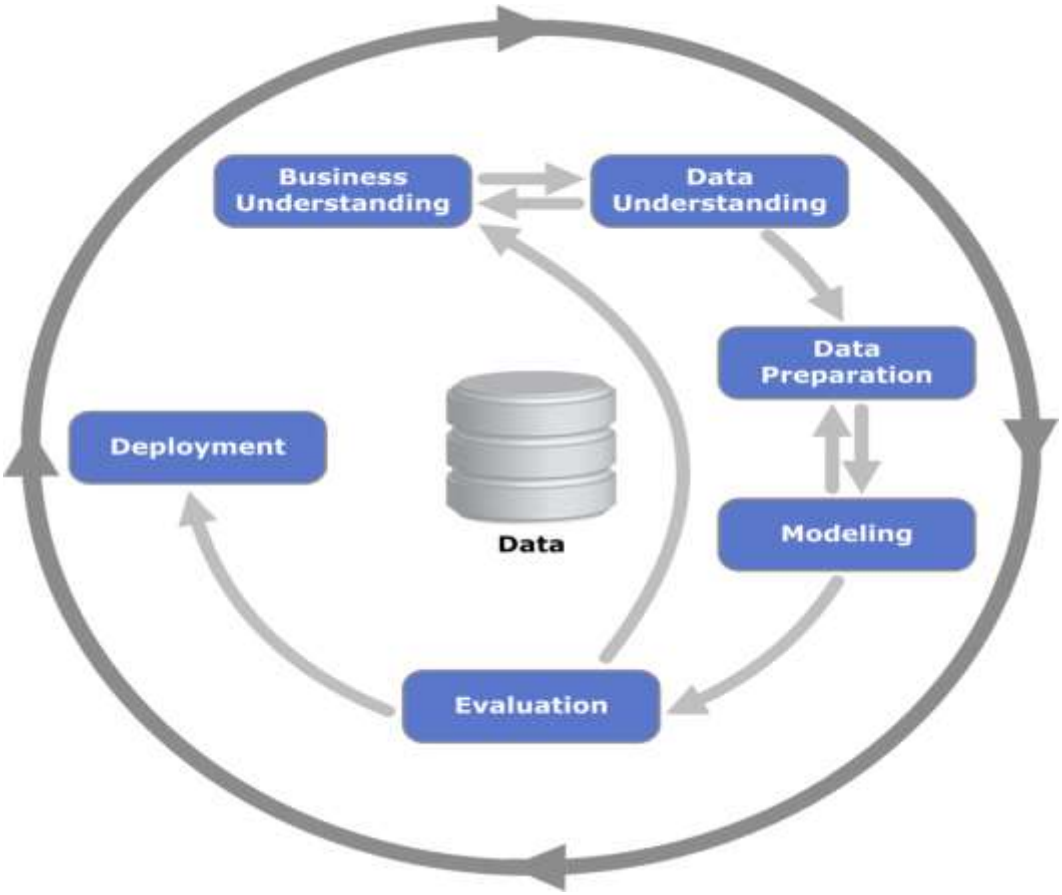

Figure 4. Detailed protocols for federated data.

In federated data experimentation, both reinforcement learning and parallel reinforcement learning have proven effective. The inputs of the models at different nodes serve as the reinforcement learning environment, while the associated outputs denote the optimized parameters of these local models. These factors are subsequently integrated into the global model fusion. Subsequently, the global model's accuracy will be evaluated, and if an improvement is detected, the global model will be distributed to all participating nodes. Similarly, the federated data operational process creates a self-perpetuating cycle of independent learning.

All transmission protocols for these models between client and server nodes are subject to encryption. These nodes can also be incorporated into the blockchain to augment data security and user privacy. In addition to improving security and privacy, federated data fosters the advancement of reliable and trustworthy artificial intelligence.

Conclusion

The global AI/Data policy is developing to simultaneously consider the rights of data providers and the benefits of data usage in the context of privacy infringements. This paper introduces a comprehensive functional architecture for reliable AI/Data commons, aimed at

fostering an open collaborative ecosystem grounded in Participation-Collaboration-Incentives (PCI). This framework tackles various user-defined AI challenges via a data utilization value chain, guaranteeing data sovereignty and protecting sensitive information. The creation of a dependable AI/Data commons ecosystem will bolster national technological competitiveness in personal information utilization through secure data transactions, and promote the expansion of AI applications and the development of new industries across diverse sectors, including smart city environments. It is expected that diverse application services originating from the AI/Data ecosystem will be provided to enhance health, safety, and convenience regarding various societal challenges.

References

[1] Jaladi, D.S. and S. Vutla. (2017) Harnessing the Potential of Artificial Intelligence and Big Data in Healthcare. The Computertech. 31-39.

[2] Gudepu, B.K. (2016) The Foundation of Data-Driven Decisions: Why Data Quality Matters. The Computertech. 1-5.

[3] Pasham, S.D. (2019) Energy-Efficient Task Scheduling in Distributed Edge Networks Using Reinforcement Learning. The Computertech. 1-23.

[4] Jaladi, D.S. and S. Vutla. (2018) The Use of AI and Big Data in Health Care. The Computertech. 45-53.

[5] Tulli, S.K.C. (2022) Technologies that Support Pavement Management Decisions Through the Use of Artificial Intelligence. International Journal of Modern Computing. 5(1): 44-60.

[6] Pasham, S.D. (2017) AI-Driven Cloud Cost Optimization for Small and Medium Enterprises (SMEs). The Computertech. 1-24.

[7] Nersu, S., S. Kathram, and N. Mandaloju. (2020) Cybersecurity Challenges in Data Integration: A Case Study of ETL Pipelines. Revista de Inteligencia Artificial en Medicina. 11(1): 422-439.

[8] Pasham, S.D. (2018) Dynamic Resource Provisioning in Cloud Environments Using Predictive Analytics. The Computertech. 1-28.

[9] Gudepu, B.K. (2017) Data Cleansing Strategies, Enabling Reliable Insights from Big Data. The Computertech. 19-24.

[10] Mandaloju, N. kumar Karne, V., Srinivas, N., & Nadimpalli, SV (2021). Overcoming Challenges in Salesforce Lightning Testing with AI Solutions. ESP Journal of Engineering &

[11] Gudepu, B.K. and O. Gellago. (2018) Data Profiling, The First Step Toward Achieving High Data Quality. International Journal of Modern Computing. 1(1): 38-50.