Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Edge-Driven Multi-Agent Intelligence for Automated Data Validation: A Pioneering Approach to Data Quality at Ingestion

Raghu K Para

Independent Researcher, Artificial Intelligence, Ontario, United States

Data-driven decision-making relies heavily on the quality and integrity of incoming data streams. While centralized data validation has been a mainstay, emerging paradigms in edge computing offer the potential to enforce data quality earlier in the data lifecycle. This paper introduces a multi-agent system (MAS) architecture, executed at the network edge, that automates data validation and quality grading during ingestion. By distributing intelligence among multiple agents embedded near data sources, the framework performs real-time checks on structural coherence, semantic consistency, and contextual reliability. Building on artificial intelligence (AI) principles, these agents classify incoming streams into standardized "bronze," "silver," and "gold" quality tiers, laying the foundation for more reliable data lakes and warehouses downstream. We present a conceptual model for edge-based data validation, analyze key challenges such as resource constraints, scalability, and agent coordination, and demonstrate the potential for improved data quality with minimal latency overhead.
Finally, we discuss how an edge-based multi-agent paradigm can drive future developments in IoT ecosystems, industrial automation, and mission-critical domains seeking robust, near-real-time data vetting.
Keywords: Data Validation, Data Quality, Edge Intelligence, Multi-Agent Systems, Automated Ingestion, Bronze-Silver-Gold Classification, AI at the Edge.

1. Introduction

Modern organizations increasingly depend on real-time data streams—from IoT sensors, electric vehicle modems, social media platforms, transactional systems, and other high-

velocity sources—to drive decision-making (Hilbert, 2020; Jagadish et al., 2014). Yet, the quality of this data can vary wildly, affecting the reliability of analytics, machine learning (ML) pipelines downstream, and operational processes (Batini et al., 2009). Traditional data validation often occurs much after ingestion, within centralized data lakes or warehousing environments (Inmon, 2005; Kimball & Ross, 2013). However, such a post hoc approach can allow low-quality, incomplete, or erroneous data to propagate downstream—adding cost, slowing time to insight, and causing potential business risks due to inaccuracies (Zhu et al., 2019).

Advances in edge computing—the practice of placing compute resources and intelligence near the data source (Shi & Dustdar, 2016)—offer an opportunity to shift quality checks upstream. By executing data validation at or near the point of ingestion, organizations can filter, highlight or correct defective data before it becomes embedded in mission-critical systems (Satyanarayanan, 2017; Varghese & Simmhan, 2017). Meanwhile, breakthroughs in multi-agent systems (MAS), where multiple autonomous entities coordinate to solve complex tasks (Weiss, 2013), enable decentralized control and distributed intelligence (Wooldridge, 2009). In tandem, these developments hint a new paradigm: edge-based multi-agent intelligence that automates data validation in real time, incrementally building trust in the incoming data streams.

This paper proposes an integrated framework for automated data validation at ingestion using multi-agent systems deployed at the edge. Our approach is designed to:

- Adapt to resource-constrained environments (e.g., embedded devices, industrial gateways) by employing lightweight AI models for real-time checks.
- Coordinate among multiple agents that can parse diverse data formats and parquet files, handle dynamic workloads, and apply domain-specific rules.
- Categorize data into bronze, silver, or gold tiers based on completeness, accuracy, consistency, and other quality dimensions (Olshannikova et al., 2020).
- Provide immediate feedback for data producers, enabling quick fixes or corrections to upstream data sources.

Section 2 reviews the relevant literature on data quality frameworks, edge computing, and multi-agent architectures. Section 3 outlines the conceptual model and layers for edge-based data validation. Section 4 explores the role of AI techniques in pattern recognition, anomaly detection, and rule-based inference. Section 5 delves into agent coordination mechanisms, resource constraints, and deployment considerations. Section 6 highlights use cases and potential benefits. Section 7 discusses open challenges, including privacy and integration issues, while Section 8 concludes with future directions.

By establishing robust data validation at the edge, organizations stand to reduce data contamination, improve trust in analytics-driven business, and optimize storage and processing resources. The synergy of multi-agent intelligence and edge computing can drive a new era of proactive data quality management—empowering next-generation IoT, streaming analytics, and real-time control systems with data that is accurate, reliable, and promptly verified.

2. BACKGROUND AND RELATED WORK

A. Data Quality and Validation

Data quality research has long emphasized dimensions such as accuracy, completeness, timeliness, consistency, and relevance (Wand & Wang, 1996; Batini et al., 2009). Typical workflows involve post-ingestion cleaning in data warehouses, applying transformations like deduplication, type checks, referential integrity checks, and outlier removal (Kimball & Ross, 2013). Though effective for batch-oriented systems, these methods often yield latency or overhead for real-time analytics (Cichy & Rass, 2020).

Frameworks for grading data quality frequently adopt multi-level taxonomies—bronze for raw or unverified data, silver for partially curated data, and gold for thoroughly validated or integrated datasets (Lake & Quintero, 2020). While widely accepted in data engineering spheres, these tiered strategies typically require substantial offline, sometimes on-premises processing. This paper focuses on how to automate such categorization at the edge, reducing the pipeline burden in centralized systems including on cloud architecture.

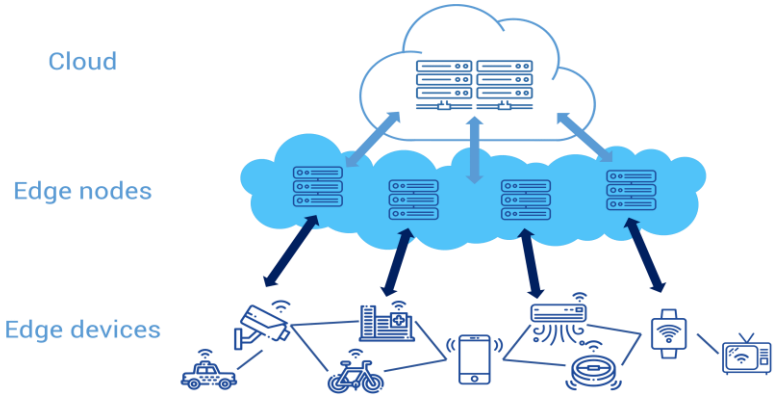

Fig 1. Distributed Edge Computing Paradigm

B. Edge Computing Paradigms

Edge computing aims to push computation and intelligence away from centralized clouds to geographically distributed edge devices (Satyanarayanan, 2017). This shift is motivated by latency sensitivities, bandwidth constraints, privacy requirements, and the need for independence in remote or mobile scenarios (Shi & Dustdar, 2016). Real-time analytics on the edge has been explored largely for video processing, sensor fusion, and local ML inference (Xu & Helal, 2018; Varghese & Simmhan, 2017). Data validation at the edge, however, remains relatively underexplored, often limited to basic range or null checks or incomplete rule-based scripts (Liu et al., 2022). Our approach extends beyond these simpler schemes by employing a multi-agent framework that can dynamically scale, handle and execute complex data validation logic.

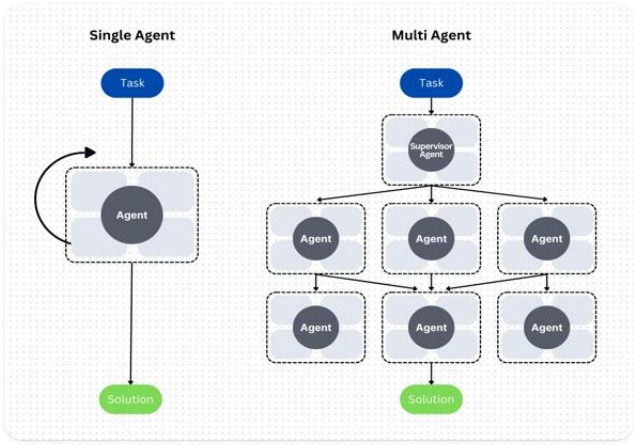

Fig 2. Multi-Agent Systems Flow

C. Multi-Agent Systems (MAS)

A multi-agent system comprises multiple autonomous or semi-autonomous agents coming together to achieve common, overlapping or coordinated goals (Weiss, 2013). Agents can sense the environment, reason or learn from data, and act to fulfill their local objectives while coordinating via communication protocols (Wooldridge, 2009). MAS has been applied in robotics, supply chain optimization, and distributed sensor networks among many fields (Zambonelli et al., 2003). The concept of agent-based data management is gaining traction, particularly for contextual adaptation and real-time decision-making (Gou et al., 2013; Bravetti et al., 2021).

Building on these principles, an MAS for data validation could partition the workload across specialized agents: one for structural validation to check on schema formats, another for semantic coherence to check on domain rules, and others for anomaly detection or metadata enrichment. This layered structure may allow parallel execution and robust fault tolerance at the edge.

D. AI-Assisted Data Checks

While rule-based validation remains common (e.g., regex or DSL-based checks), AI-based approaches can detect subtle patterns or anomalies (Aggarwal, 2015). Supervised models may classify records as acceptable or suspicious based on labeled examples, while unsupervised methods like clustering or autoencoders can flag outliers in streaming data (Chandola et al., 2009; Para, 2024). With the edge's resource constraints, however, model selection and optimization become critical (Zhang et al., 2019). Researchers have explored tiny ML or model compression techniques for on-device inference (Lane et al., 2015). Similarly, reinforcement learning can help agents adapt validation policies over time (Zhu et al., 2021). We integrate these methods into our architecture, ensuring performance viability in edge scenarios.

In summary, the literature suggests that edge computing and multi-agent intelligence present a promising approach to tackling the challenges of data validation and quality grading at the

edge. This synergy not only preserves systemic performance but also contributes to higher-quality data integrity.

3. CONCEPTUAL MODEL AND ARCHITECTURE

A. Overview

We propose an Edge-Driven Multi-Agent Intelligence (EDMAI) framework that performs automated data validation on ingest streams. Figures 1 and 2 conceptually depict data flowing from IoT sensors or external APIs to an edge gateway, where multiple agents examine quality metrics in near real time. Data validated at the edge is then forwarded to downstream systems, tagged with quality metadata (bronze, silver, gold).

Each agent operates semi-autonomously, focusing on specialized checks: schema adherence, domain constraints, semantic context, or anomaly detection (Para, 2024). The MAS environment includes communication channels for agent-to-agent messaging (Weiss, 2013). If an agent flags data as suspicious, it can query other agents or request user input. This distributed design ensures no single point of failure or bottleneck.

B. Layered Components

- Data Ingestion Layer: Responsible for interfacing with data sources, staging incoming records, and distributing them to agent modules. This layer may handle fundamental transformations or audits, such as format normalization.
- Agent Coordination Layer: Orchestrates the lifecycle of individual agents (monitors registration, implements scheduling) and routes data among them. It is a driving service, which ensures dynamic discovery of specialized agents for tasks (e.g., a sensor agent vs. a transaction agent).
- Validation/Quality Agent Layer: This layer of core intelligence, ensures reading from a local message bus or queue, applying their validations and checks in parallel. They produce a combined quality score or classification (e.g., bronze, silver, gold).
- Feedback and Control Layer: Aggregates the outcomes from the validation layer, updates logs, triggers alerts where necessary, and escalates data for manual review if confidence is low. This layer also feeds user or domain expert corrections back into the agent knowledge base.

C. Data Quality Tiers

- Bronze: Data that passes foundational or structural checks but is incomplete or uncertain.
- Silver: Data that meets additional semantic or domain rules, sufficiently reliable for certain analytics or operational usage.
- Gold: Data thoroughly validated across all relevant criteria, comprising minimal anomalies, duplicates, or inconsistencies

Agents cumulatively assign these tier labels, storing them in metadata for subsequent analysis

or processing. This immediate labeling enforces quality gating, ensures only "Gold" or higher-tier data flows into mission-critical pipelines, and systematically organizes partial or questionable data for corrective actions.

4. AI-DRIVEN DATA VALIDATION

A. Rule-Based and Symbolic Checks

At the edge, some validations remain well-served by direct domain-specific rules.

- Schema Matches: Ensuring required fields, data types, permissible ranges (Kimball & Ross, 2013)
- Constraint/Rule-based Satisfaction: For instance, date fields must follow ISO standards, numeric fields must be within known domain boundaries, or mandatory foreign key references must exist (Batini et al., 2009).
- Regex Pattern Matches: For email addresses, phone numbers, or ID formats.

Such rules can be stored in knowledge bases, triggered by specialized "schema agents" or "domain agents." While these methods are quick or real-time, they do not detect subtle anomalies or emerging data drifts (Cichy & Rass, 2020).

B. Anomaly Detection

Anomaly detection leverages statistical or ML-based methods to flag unusual data points (Chandola et al., 2009; Aggarwal, 2015; Para, 2024). In the edge-driven MAS context, an "anomaly agent" may run a simplified isolation forest or local outlier factor algorithm. Alternatively, small neural autoencoders can learn typical data patterns and measure reconstruction error in real time (Zhang et al., 2019; Para, 2024). If the error exceeds a certain defined threshold, the agent marks the record as suspicious.

Despite resource limitations, careful hyperparameter tuning or model compression can sustain near-real-time performance (Lane et al., 2015). The agent's local memory can store recent samples, allowing incremental updates. Over time, drift detection modules identify if the data distribution shifts significantly requiring a retraining or recalibration step at that point (Zhu et al., 2021).

C. Semantic and Contextual Checks

In scenarios like smart agriculture, automotive, ecommerce or manufacturing, domain knowledge can be crucial. A "semantic agent" might cross-reference readings from the sensors or vehicle modems with environmental conditions or business logic (e.g., temperature must not exceed X if machine status is set to Y, electric vehicle charging cannot have readings reported when turned off) (Gou et al., 2013). Knowledge graphs or ontologies can aid in verifying relationships between entities or events (Bravetti et al., 2021).

Contextual signals, such as time, location, or correlated sensor streams, allow the agent to interpret data within a broader situational scope. For instance, if an accelerometer reading spikes but no motion is recorded by a correlated camera sensor, a potential data inconsistency arises. Combining these signals requires advanced inference engines or well-designed and

simplified semantic reasoners tuned for edge devices (Wooldridge, 2009).

D. Reinforcement Learning for Policy Adaptation

Rather than a static set of rules maintained heuristically, some agents could leverage reinforcement learning (RL) to adapt validation policies. If a certain rule triggers false positives often, the RL-based agent adjusts threshold parameters to reduce them (Zhu et al., 2021). The agent receives feedback signals from downstream systems or user interventions. Over time, it refines its detection strategy, balancing sensitivity (to catch errors) with specificity (to avoid any false alarms). RL is especially pertinent where data evolves unpredictably—e.g., dynamic IoT networks or markets (Aggarwal, 2015).

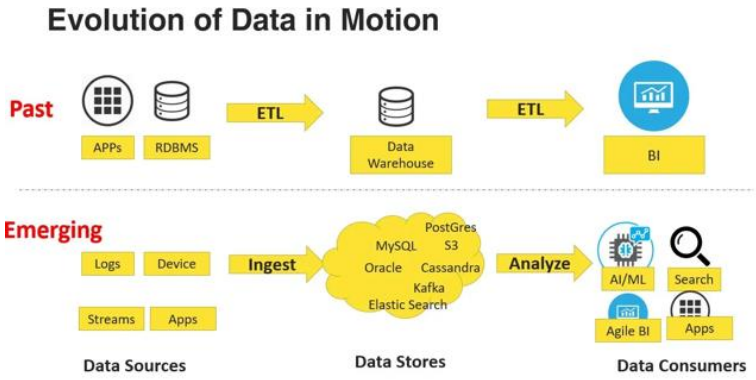

Fig 3. Evolution of Data Streaming

5. MULTI-AGENT COORDINATION & DEPLOYMENT

A. Agent Roles and Hierarchies

We propose the following arrangement of agents.

- Ingestion Agents: Interface with data sources, perform quick syntax checks, route data to specialized validators.
- Validation Agents: Focus on domain logic, anomalies, or semantic checks. Possibly multiple parallel validator agents exist for different data types or use cases.
- Coordinator or Manager Agents: Facilitate workload distribution, aggregate results, resolve conflicts among validators, and assign final quality scores.
- Supervisor Agents (optional): Oversee RL policy updates or domain rule modifications, handle user feedback, track performance metrics.

B. Communications and Scalability

The MAS can use publish-subscribe topics or message queues (e.g., MQTT, RabbitMQ) for asynchronous data exchange. Each validator agent subscribes to relevant data topics. If an agent classifies data as suspicious, it can publish an event, prompting other agents or requesting a manager or a coordinator to re-check or quarantine that record (Weiss, 2013).

Scalability arises from the ability to spin up additional validator agents or distribute them across multiple edge nodes. Agents can register or deregister with the manager dynamically, allowing ephemeral edge devices to join or leave the system (Zambonelli et al., 2003).

C. Resource Constraints and Edge Deployment

Edge devices typically have limited CPU, memory, and power budgets (Shi & Dustdar, 2016). Strategies to address these constraints include:

- Model Compression: Pruning or quantization for neural-based anomaly detection (Lane et al., 2015).
- Thin-Client Rule Engines: Minimizing overhead in symbolic checks.
- Agent Migration: Offloading busier computations to a nearby fog node or micro-cloud when local resources become insufficient (Satyanarayanan, 2017).
- Adaptive Frequencies: Adjusting the frequency of deep checks vs. basic checks to manage workloads (Zhang et al., 2019).

Hence, an MAS at the edge must diligently orchestrate local computations and synergy with more powerful nodes to achieve an overall robust performance.

6. CHALLENGES AND RESEARCH DIRECTIONS

A. Privacy and Security

Edge devices often operate in volatile environments. Ensuring data confidentiality during processing—especially if personally identifiable or sensitive data is validated—requires encryption, secure enclaves, or privacy-preserving computations (Acquisti et al., 2016). Agents must also authenticate themselves to avoid impersonation or any sort of malicious infiltration (Mosenia & Jha, 2017).

B. Knowledge Assessment

Domain rules, data schemas, and anomaly signatures can evolve over time. Agents must maintain up-to-date knowledge bases over time as well. Automatic or semi-automatic rule updates—potentially triggered by new domain knowledge or user feedback—are critical to keep validation consistent over time (Gou et al., 2013). Handling versioning and rollback also becomes a challenge (Zhu et al., 2021)

C. Conflict Resolution

Multiple validator agents might have a disagreement on a record's quality rating. The system must define conflict resolution strategies—majority voting, confidence weighting, or a predefined hierarchical priority. If disagreements persist, the data might remain in a "gray zone" awaiting manual intervention or curation (Weiss, 2013). Designing stable consensus algorithms is nontrivial in dynamic, multi-agent contexts (Zambonelli et al., 2003).

D. Complexity vs. Latency Trade-Off

More computational or more sophisticated checks yield better detection but risk incurring too

much latency, especially at scale (Cichy & Rass, 2020). Striking a balance between thorough validation and real-time throughput is a key concern. Further research might explore dynamic validation that downgrades check under higher loads or selectively applies advanced methods only to suspicious data (Zhang et al., 2019).

E. Standardization and Interoperability

As edge computing environments expand, the lack of standard frameworks for agent deployment, rule specification, and data labeling complicates adoption (Bravetti et al., 2021). Industry consortia or open-source platforms could define common APIs and templates for multi-agent data validation. However, without consensus, prolonged fragmentation may restrict cross-domain synergy.

7. CONCLUSION AND FUTURE OUTLOOK

The necessity for real-time, automated data validation is rising in parallel with the growth of data-centric systems and IoT networks. This paper outlined how multi-agent intelligence at the edge can bring robust, context-aware validation logic closer to data sources, ensuring faster detection of errors or anomalies and higher fidelity in downstream analytics. By combining symbolic rules with AI-driven anomaly detection, a layered MAS can concurrently handle semantic, structural, and contextual checks—tagging data with “bronze,” “silver,” or “gold” tiers right at ingestion.

We highlight several benefits: fewer data inaccuracies flowing into centralized data warehouses or lakes, lighter onus on big data pipelines, improved accuracy timeliness of insights, and “more-heightened-than-ever” trust in real-time applications. While edge-based systems confront unique resource, security, and scalability hurdles, the synergy between edge intelligence and multi-agent design provides a promising blueprint for future data quality and management solutions.

Looking ahead, further research on integrated reinforcement learning for agent policy evolution, advanced privacy-preserving models for sensitive data, and standardized frameworks for agent-based data quality will further refine this paradigm. As industries embrace devices with sensors and pervasive IoT and real-time decision-making, edge intelligence for data validation could become an essential and a pioneering fixture—ensuring data integrity from the starting bytes ingested into the digital pipeline. By deploying robust, decentralized, and adaptive solutions, we inch closer to a future where erroneous or low-value data is systematically filtered out, quarantined and even corrected, leaving only consistent, contextualized, and higher-quality information to power next-generation data-driven enterprises.

References

1. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2016). Privacy and human behavior in the age of information. Science, 347(6221), 509–514.
2. Aggarwal, C. C. (2015). Outlier analysis (2nd ed.). Springer.

3. Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52.
4. Bravetti, M., Montesi, F., & Zavattaro, G. (2021). Towards agent-based data management in modern distributed systems. Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 1564–1566.
5. Cichy, K., & Rass, S. (2020). A data validation approach for streaming data ingestion. Computers & Security, 96, 101850.
6. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58.
7. Gou, Y., Zhang, F., & Gao, X. (2013). A multi-agent framework for data validation in supply chain management. International Journal of Production Research, 51(13), 3946–3960.
8. Hilbert, M. (2020). Digital technology and social change: The digital transformation of society from a historical perspective. Dialogues in Clinical Neuroscience, 22(2), 189–194.
9. Inmon, W. H. (2005). Building the data warehouse (4th ed.). John Wiley & Sons.
10. Jagadish, H. V., Gehrke, J., Labrinidis, A., et al. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86–94.
11. Para, R. K. (2024)." Using Autoencoders for Anomaly and Drift Detection in Linguistic Segmentation on Product Review Platforms and Recommendation Systems." Nanotechnology Perceptions (2024): 3333-3345
12. Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.). John Wiley & Sons.
13. Lake, P., & Quintero, R. (2020). Data lake architecture: Designing the data lake and avoiding the garbage dump. Apress.
14. Lane, N. D., Bhattacharya, S., Georgiev, P., & Ju, Y. (2015). Advancing adaptive and context-aware computing: Mobile crowdsensing, wearables, and big data integration. IEEE Pervasive Computing, 14(1), 38–46.
15. Lee, C. S., & Siau, K. (2018). A review of data mining techniques. Industrial Management & Data Systems, 118(1), 178–193.
16. Liu, C., Huang, J., & Li, X. (2022). An edge-based data validation framework for IoT sensor data streams. Future Generation Computer Systems, 130, 12–21.
17. Mosenia, A., & Jha, N. K. (2017). A comprehensive study of security of internet-of-things. IEEE Transactions on Emerging Topics in Computing, 5(4), 586–602.
18. Olshannikova, E., Morales, C. A., & Granados, D. (2020). A taxonomy-based approach for data quality classification in data lakes. Information Systems Management, 37(3), 184–196.
19. O'Sullivan, D., O'Connor, Y., & Fraccascia, F. (2017). Edge analytics for healthcare IoT: A real-time wellness data pipeline architecture. Journal of Sensor and Actuator Networks, 6(4), 26.
20. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.
21. Shi, W., & Dustdar, S. (2016). The promise of edge computing. Computer, 49(5), 78–81.
22. Varghese, B., & Simmhan, Y. (2017). Demystifying fog computing: Characterizing architectures, applications and abstractions. Proceedings of the 2nd IEEE International Conference on Fog and Edge Computing, 115–124.
23. Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86–95.
24. Weiss, G. (2013). Multiagent systems (2nd ed.). MIT Press.
25. Wooldridge, M. (2009). An introduction to multiagent systems (2nd ed.). John Wiley & Sons.
26. Xu, W., & Helal, A. (2018). Scalable cloud-edge computing for IoT-based analytics. IEEE Internet of Things Journal, 5(3), 2013–2021.
27. Zambonelli, F., Omicini, A., & Papadopoulos, G. A. (2003). Developing multiagent systems: The Gaia methodology. ACM Transactions on Software Engineering and Methodology, 12(3), 317–370.
28. Zhang, D., Li, B., Li, L., & Verma, D. (2019). Adaptive data quality validation in IoT-based data streams. IEEE Access, 7, 112593–112607.
29. Zhu, Y., Li, M., & Gao, J. (2019). A big data cleaning framework for quality improvement of open data. Future Generation Computer Systems, 95, 662–681.
30. Zhu, Y., Qiu, K., & Wu, X. (2021). Reinforcement learning for dynamic data quality validation in stream processing. IEEE Transactions on Knowledge and Data Engineering, 33(12), 3491–3507.