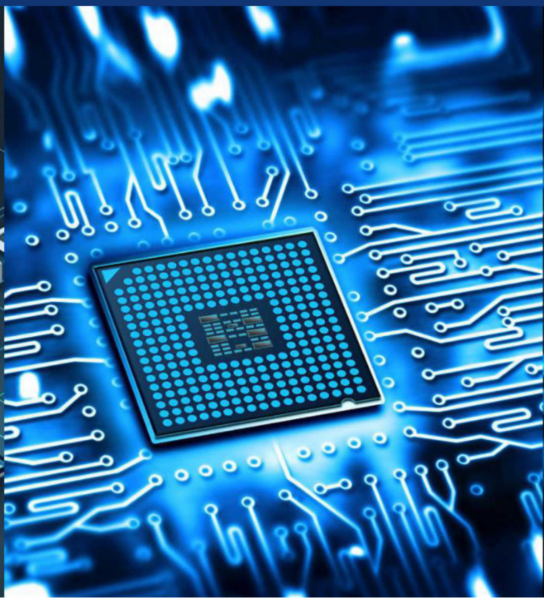
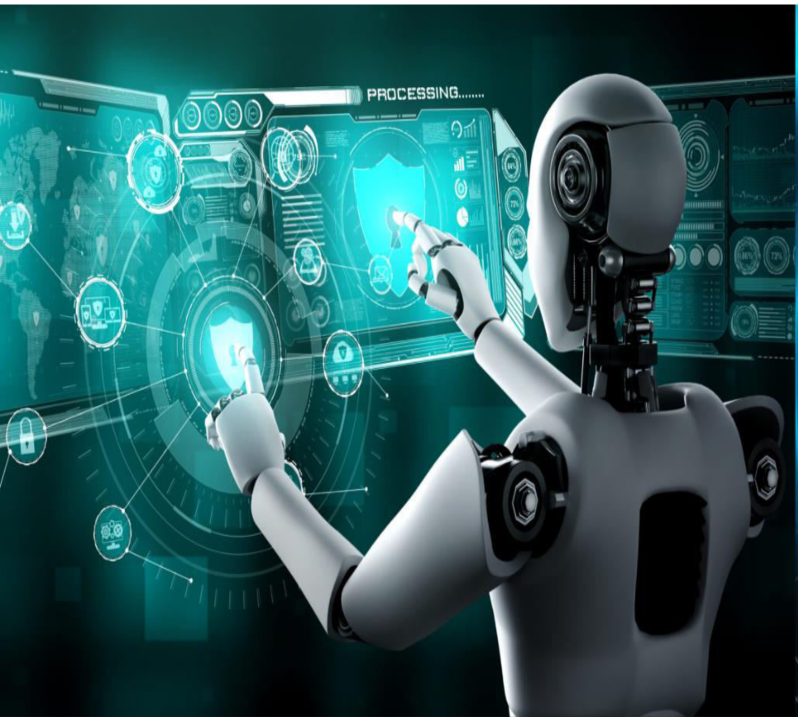
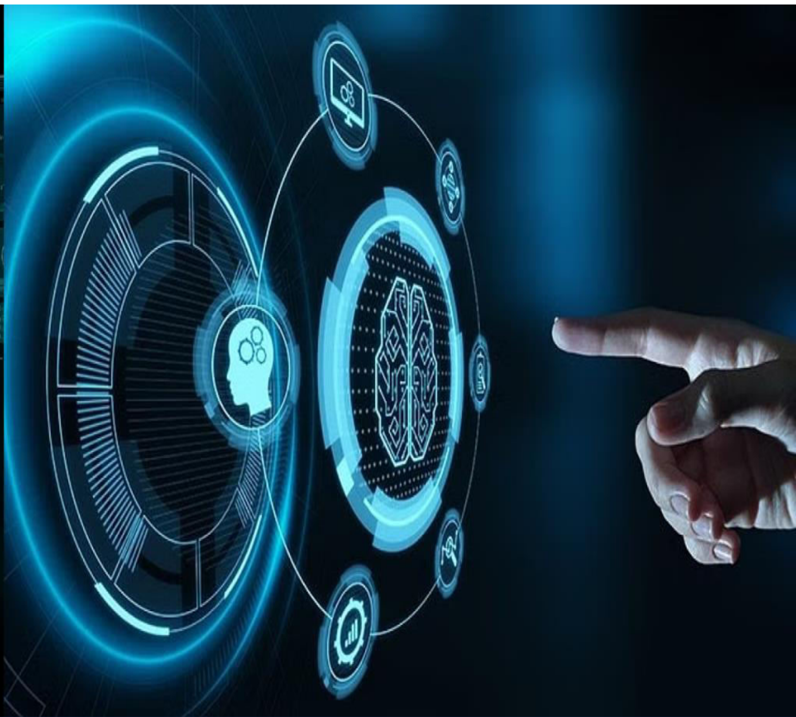
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Impact Factor: 8.625

Volume 13, Issue 1, January 2025

www.ijircce.com ijircce@gmail.com +91-9940572462 +91 63819 07438

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

An Event-Driven Architecture for Autonomous Supply Chain Risk Detection and Decision Automation

Kartheek Chandra Ambati

Sr Systems Engineer, CSCS, USA

ABSTRACT: The primary goal of the supply chain risk detection and alerting MVP architecture is to support critical functionality with limited user input with the aid of various data sources (e.g., ERP, WMS, IoT sensors) to synthesize the data amplified as a real-time visualization for a user. Autonomous AI agents support a modular and event-driven architecture that adds awareness of a situation for predictive and risk mitigation purpose. Advantages of this architecture are to provide the capability of identifying risk through machine learning and anomaly detection, a scalable microservices architecture to build or incorporate sophisticated AI agents that emphasize logistics and compliance, and allow for the automation of work processes, aiding in creating a feedback loop for complicated decision-making capability. The architecture is also touching upon governance regarding security and ethical AI (both required aspects of any successful operation) while monitoring real-time KPI status, which is beneficial for governance of system health. Further, the architecture provides an additional useful capability for continuous iterative improvements to the automated processes, which aids in supply chain risk governance and reinforces oversight of supplier risk networks by using a process for implementing reinforcement learning and risk governance of operational functions. There is also a strong opportunity for companies to leverage semi-advanced technologies (i.e., edge computing) to intelligently build AI based systems to support real-time proactive risk management practices in supply chain operations, while creating potential implications to enhance agility and the quality of strategic decision-making.

KEYWORDS: Supply Chain Risk Detection, MVP Architecture, Microservices, Supplier Risk Networks, AI-Based Systems

I. INTRODUCTION

The MVP for end-to-end risk detection and alerting will be developed to include a set of fundamental features, with minimal manual effort required, and to provide a base for future AI, scalability and security capabilities. This MVP will primarily focus on identifying, analyzing and communicating risk in a holistic way to allow action to take place to lessen risk exposure. Additionally, the MVP would allow for the automation of the risk process to drive efficiency and lead times for the organization. The MVP for the solution will include only the components necessary for a usable solution to provide information and value to allow for fast testing and feedback. The MVP for the solution will also, have a measurable and scalable architecture to be able to handle additional requirements, growing user capacity demands, and increasing complexity of risk data without hindering performance. Furthermore, the MVP will include a set of best in class security protocols, including data encryption and compliance, to promote user confidence.

Future AI capabilities, including predictive risk modeling and dynamic alerting protocols are considered during the manufacturing of the MVP for users to see and learn from a set of sample data collection protocols. Ultimately, This process of building the MVP enhances risk associated with the risk project- both from technical and financial aspects by including foundational features, providing measurable short term aspects and value to product and market speed, as well as enabling and confirming compliance and security. In conjunction, the foundational MVP architecture allows for business development growth, which enables gradual building of features based on real and confirmed user need and behavior. As stated the MVP provides a short-term risk detection capability and a good solid platform to grow the business.

An end-to-end risk detection MVP must consider flexible configurations to provide users with various risks (i.e. technical risks related to integration points of risk with existing components or systems, scalability of risk associated with features, performance contingencies of reduced performance of risk detection and alerting capabilities and associated risks due to vulnerabilities within the solution featuring sensitive data, and i.e. technical debt from rushing the coding process and its risks).User and market risks relate to a disconnection between perceived user needs, inadequate information collection on feedback due to insufficient engagement, and a lack of confirmatory means of detecting resolution of detection. Operational risks include over-reliance on manual reviewing moves the solution away from its goal of automation, issues with data quality that might affect the accuracy of detection, and ambiguity of alerting processes that may lead to alerts being missed or fatigue or overload.

Project and resource risks can include scope creep, a lack of resources or deadlines that may lead to incomplete (or partial) features, and spending over-budget on the entry to the market. Regulatory and compliance risks encompass non-compliance with industry regulations involving a lack of transparency in estimation of potential risks. In mitigating various risks, testing, stakeholder engagement/focus on appropriately limited features, continuous feedback, security best practices (and reasons why) to respond to regulation, transparency of project management, in terms of the MVP to be delivered against risks, will contribute positively towards the MVP being able to deliver effective, scalable, secure and validated risk detection and alerting functionality [2].

It is a next generation, AI-enabled solution for end-to-end supply chain risk detection and management. It provides automated data ingestion with always-on access to transactional or master data, as well as other email and attachments for consumption by the user. The solution uses rule based pre-screening methodology to prioritize the most critical events through filtering and extracting the relevant information. The AI-enabled risk assessment allows for search and evaluation of risks in an independent manner, while each specialist component can contribute to problem solving and resolution consistently together. There is a real-time notification function to inform both suppliers and internal departments with alerts, as well as the user experience actional insight and reporting systems, and a perpetual and precise audit trail for compliance assurance. Additionally, the system is built to be scalable and secure so that it can accommodate and respond to AI developments in the future and provide high performance for applications and operations for the data and notifications. By combining intelligent risk rating and automated data extraction, the solution allows for early intervention in the supply chain and surmounts reliance on manual oversight while improving detection accuracy. Overall, it demonstrates Forward Lean's modern methodology of supply chain risk management. It leverages artificial intelligence to automate and enhance visibility and control of risk within complicated supply chains [3].

Many autonomous email ingestion pipelines fail for a variety of reasons. Data-formatting and/or data parsing errors are frequent when the format of the email and/or attachments are inconsistent or broken leading to partial ingestion of data. Process and orchestration errors occur when jobs stall due to misconfigured tools or broken dependencies leading to partial availability of data. Data duplication and quality issues result from human and/or processing error, which results in duplicate or stale data. Network failures can result in the unavailability of data. Security and privacy compliance risks arise from lapses in data protection and/or significant differences in privacy compliance across different jurisdictions. Additionally, gaps in visibility and monitoring capabilities can hamper the detection or remediation of ingestion issues. Multi-agent coordination failures can also occur in complex processes that impact the entire pipeline. As a general rule, good error handling, a retry mechanism, monitoring, security measures, and a modular pipeline design can provide good error isolation and remediation processes when incurring ingestion failures [4].

The mean ingestion failures refer to the number of ingestion paths that incur a degree of data loss due to one type of ingestion failure or failure type and event. Job failures and orchestration failures are two clear common ingestion failures and pipeline breakdowns with processes such as scheduled jobs or workload management tools such as Apache Airflow, which can result in data loss or delay of a full data batch with gaps impacting analytics and reporting.Schema modifications can also disrupt ingestion workflows resulting in data loss when source data structures change unexpectedly. Poor data quality is a significant source of ingestion failures because ingesting corrupted or inconsistent source data can lead to incomplete ingestion and silent data loss. Network interruptions during data movement are often irreversible unless robust error handling exists. Crashing resources such as CPU, memory, or storage capacity can also affect ingestion workflows and lead to data loss of unsaved data. Alerting and monitoring misconfigured or manual checking may leave failures unnoticed or undiagnosed longer than it would normally take if the data would not be lost. Consequently, data loss explores the failure of data quality, schema incompatibility, task failures, infrastructure limitations, and inadequate operational observability may contribute to data loss, so mitigate or prevent these incidents require robust schema versioning, robust monitoring with alerts, retry logic, and scoping resources [5].

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The ingestion process is uniquely vulnerable to silent data loss at all stages. From Acquisition of data or sign-up and collection, data losses may go unnoticed due to incomplete email gathering or the network may go down, both instances are data loss. During parsing and extraction, many errors may occur. From decompression to incorrectly converting a field's format among many other reasons complete data may be lost. Data transformation or at ingestion works can incur silent data loss that sometimes goes unseen as well. Ingestion may invoke filtering rules resulting in inducing data loss of good records never realizing until after ingestion is complete. Orchestration tools may result in incorrect executions that are inconspicuous continuing to fork failed ingestion but not identify what records were completely missed ingestion or simply duped.Lastly, data storage can also facilitate silent loss via incomplete writes or corrupted files, guided by a later query or report. These vulnerabilities emphasize the importance of complete ingestion validation, checksums, and monitoring in order to address challenges related to silent data loss.

Ingestion validation provides a number of mechanisms by which to identify and manage silent data loss. Schema validation and enforcement attempts to detect silent corruption or fields missing altogether by validating incoming data against an expected schema. Checksum and hash verification identified abnormalities or changes in the data pre-processing that indicated silent corruption or flows that exceeded expectations. Monitoring counts of records and the overall volume of data can illuminate silent losses or drops in data that would otherwise go undetected. Invariant-based validation checks for defined business rules to mitigate the introduction of aberrant data submitted as silent errors. Anomaly and outlier detection utilizes AI and statistical methods to monitor metrics collected over time regarding the data entering a system to detect deviation of a more or less normal trend indicating silent corruption. Data can also be made addressable through end-to-end data lineage and auditing procedures, to confirm that any inputs expected to be present actually reach that point. Regression testing, by its very nature, checks for silent data loss by way of comparison of current output to a baseline set of values -- to verify that changes have not generated new loss of data. Once the aforementioned mechanisms are built into the ingestion pipeline you address silent data loss before it happens, allowing data validity to be preserved.

Utilizing automated AI-based agents for outputting email analysis and information extraction presents some architectural considerations. These architectural considerations include designing modularized agents, including an agent for recommendations, an agent for extraction, an agent for cross-referencing and an ingestion agent moderated by a parent orchestration agent - who own the inbound emails review process. The ingestion agent own the responsibility for automatically downloading emails and any attached documents through a process of a single and constant monitoring of the email server. The ingestion agent receives the incoming email and potential attachment, undergoing pre-processing for a managed consistency and then data and information extraction through a NLP agent. The important information extracted from the body of the email, is stored in a database and will check for breaks in the data submitted against prior databases. The AI modeling for risk application and recommendation can either be executed solely automatically or proposed recommendations for action can be derived from prior AI modeling and generated by action recommendations from a solvable issues. The architecture designed to build through the adoption of scalable microservices, whereby agents are deployed independently from one another but utilize exchange of API calls and other communications, meaning agents do not have to interact directly with one another. The entire process allows for continuous feedback to models that adjust for change in the supply chain and formalizes compliance, identifying and ensuring secure actions over sensitive data. It is an architecture characterizes a modern supply chain AI system, where AI employs autonomous agents in a real time, managing risk in the increasingly complex tasks comprising workflows.

II. RELATED WORK

In supply chain management, artificial intelligence agents typically operate using multiple modalities, such as natural language programming and computer vision, to structure data extraction from communications. These autonomous agents are coordinated within a multi-agent orchestration and management system that orchestrates activities such as data extraction and ingestion, and other elements of advanced automation around decision-making. There are also significant machine learning models tied to risk assessment and root cause analysis, delivering learning mechanisms by applying supply chain data and enhancing auditability and assurance, trust and transparency, with blockchain integration. Autonomous or agentic decision-making or optimization forms allow for intelligent responses to logistics and disruption, improved by scalable, commercial microservices architectures. The main advantages of using AI agents is productivity increases by automating decision-making, real-time identification of risk, adapting responses to dynamic circumstances, coordination of decision making across multiple supply chain nodes, and agility and reslilience. Challenges remain with multi-agent orientito the potential for data security and privacy concerns, integration into legacy systems, reliance on funded data, and for the heightened governance frameworks for autonomous decision-making potentially required. Key articles in the literature examine the efficiency of AI agents for supply chain logistics, synergies of AI with blockchain for optimization, and implementable uses of agentic AI for resilient supply chain operations. These articles identify and recognize operational and technological issues related to utilizing AI agents in intelligent supply chain management [8].

Recent articles evaluating agentic AI in supply chains share a handful of primary strategies. One evaluative strategy is the Loop of Sense-Plan-Act-Learn, where autonomous or agentic AI agents have the capacity to make real time decisions and ultimately learn, though achieving these capacities relies on tremendous data and sustains sufficient computing ability to perform the extensiveMulti-Agent Systems (MAS) consist of collaborations among specialized agents to achieve broader objectives. They provide scalability and resilience; however, coordinating and communicating such collaboration can be complicated. When agents engage in tasks of coordination, communication, and/or decision-making, those processes may be entirely autonomous and highly context-aware. Alternatively, Reinforcement Learning (RL) enables agents to learn and improve decision-making in dynamic and unknown environments. However, this increases the training data required before an agent can demonstrate useful decision-making, and there is a risk that the agent's decision-making will be erratic when placed in an environment outside of its training condition. Heuristic and graph-based planning methods can produce quick and clearly defined decisions for constraint satisfaction; however, they do not account for rapid environmental changes that require agility or adaptation.

Real-time data insertion can provide more situational awareness and ultimately improve decision-making, but there are well-known barriers to quality data and the cost of the infrastructure required to implement it. AI models for predictive maintenance can facilitate pre-emptive interventions leading to better use of resources and resilience; however, effective predictive maintenance depends on the quality of sensor data and integrating multiple disparate pieces of equipment. Overall, agentic AI greatly focuses on autonomy and context-aware operations to pre-emptively manage processes, and represents a complete opposite to traditional AI in terms of predicting activities, and possible actions to preserve resource usage. However, the higher complexity increases challenges for governance and integration before full benefits of agentic AI become available in supply chain management [9].

Current studies are highlighting the importance of specific datasets and standards for agentic AI research in testing supply chains; The Polytechnic University of Turin’s Document-Intensive Supply Chain Workflows Dataset has around 80 document types from various stakeholders involved in the assessment of document intelligence and autonomous, AI-enhanced workflow decision-making. Transportation Supply Chain Planning, real-life data in retail, can improve operational efficiency and demand forecasting. Evaluation datasets, like Databricks Mosaic AI Agent Evaluation can help set reference points in real-time, authentic situations to assess agentic, AI systems.Use cases that are contingent upon the vendor rely on datasets unique to the specific environment to synthesize demand forecasting and inventory optimization or utilize synthetic datasets and influencer mapping simulations to stress-test supply chain parameters. Open datasets are typically limited to smaller datasets to help with training and evaluation. Overall, the diversity of datasets and benchmarking systems is important in defining and designing agentic AI systems in complex supply chain environments[10].

To enhance demand forecasting datasets for the purpose of evaluating agentic AI, a number of essential enhancements would need to be made. First, real-time data from a variety of sources will be critical to help agents rapidly track to market opportunities outside of schedule inputs, such as pull from social media and economic indicators. Demand data must be organized according to specific product types, geolocation related to sales channels, and demographics of the customer base to develop accurate modeling. Secondly, relevant learning outcome data and feedback will be essential to provide the agents with the ability to learn from prior predictions and modify their approach. Thirdly, the data must be structured for the purposes of representing interconnected nodes throughout the supply chain to help evaluate agent cooperation and negotiation skills. Fourth, simulated or other kinds of disruption may be used to assess the resilience and planning capabilities of the AI in unpredictable events. Finally, as requiring continuous stream processing, to be always updated in real-time, demanding forecasting will continue to need to be provided for speedy iterative improvements. In summary, after enhancing demand forecasting datasets as described, traditional datasets will become dynamic environments for benchmarking autonomous and coordinated AI systems in supply chain settings [11].

It is important to remember that key performance indicators for agentic forecasting extend beyond accuracy to critical areas of focus. Action Success Rate is a key indicator of how effective autonomous action achieves the expected result

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

through the agent decision. Time to Resolution is measured as the length of time to plan and complete and confirm a task, a performance concern for operations. Concurrency and Throughput assess the number of forecasting activities completed concurrently by agents and could be important for risk assessment during scaled-out execution. User satisfaction and trust are related to user survey feedback on independent accuracy and agent transparency and understanding impact willingness to use this product. Robustness and Consistency assess the degree the agents are dependable during unpredictable scenarios in the real world, e.g., across multiple stores.

Learning Rate and Adaptation assess how efficiently the agents are collecting new knowledge, and how swiftly they reconfigure routine situations. Cost Effectiveness is also important for sustainable operations, measured in resource use and operational costs compared to overall agent value. Error Rate and Recovery assess frequency of errors and, how agents manage errors, or escalate requests for assistance. Task Adherence and Workflow Accuracy assess the accuracy of agents toward prescribed decision situations. Predictive risk detection captures the likelihood agents can engage prediction processes and accurately provide mitigations for anticipated failures in forecasts or supply chain disruptions. Collectively these assess the agentic forecasting systems as a whole beyond standard levels of accuracy, and broadly assess the continued operational value of agent decision making systems, potential to scale and trust from users assessing agentic reliability and actual user experience in a production supply chain environment [12].

III. SYSTEM ARCHITECTURE

The agentic AI supply chain architecture is composed of numerous key modules and data streams that act in concert towards improved operational efficiency. The Perception Module provides contemporaneous data collection from several sources, such as social media, as well as IoT devices, and processes the data through such tasks as cleaning and feature extraction. The Cognitive Module is focused on planning and reasoning functions, which are designed to meet the supply chain's goals, employing advanced methodologies such as machine learning for logistics and risk management decision-making. The Memory and Knowledge Layer is responsible for storing learned policies and models, introducing adaptability and long-term learning through effective data retrieval mechanisms. The Orchestrator Layer equally manages tasks among the agents, workflows, and communication mechanisms among module types.

The Action Module applies decisions that were identified through the interfaces of a number of systems and monitors success or failure and follows up on corrective actions as necessary. The Feedback and Learning Loops are contemporary and engage with data received to continuously glean insights and improve the original plan and retrain the model if needed. Governance, Security and Compliance, refer to actual or stakeholder required assurances many organizations establish for the protection of data, such as privacy and compliance with regulations.

Data streams through the architecture, beginning with the ingestion of data in the Perception Module, providing justification in the Cognitive Module, organizing the workforce activity, executing the action, and returning the results back to the Learning Loop for improvement. The modular form allows the system to accommodate adaptable and scalable AI operations for functioning supply chain forecasting and risk management, yielding each possible end-to-end intelligence aspect, and allowing for the continuous improvement of the system and agentic agents.

The system is designed as to include autonomous AI in various domain experts for forecasting, risk mitigation practices, procurement, or logistics and coordination, fully engaging the sense-plan-act-learn practices framework, by design. The functions of organizations utilizing such a layered architecture (without this loyalty to modular solutions) can be leveraged to improve scalability, security, and continuous improvement of their logistics processes and risk management, while an orchestrator termed as a parent conductor is identified to synchronize activity to organizational governance objectives. In addition to real-time data sources given access, reliable and repeatable event-driven APIs and feedback loops have been established, providing a comprehensive, robust, and intelligent supply chain ecosystem than traditional analytical approaches, that encourages efficient and autonomous decision-making and proactive problem solving without the instructions of the sourcing organization. Figure 1, below Architecture Layers for Agentic AI in Supply Chain Forecasting:

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

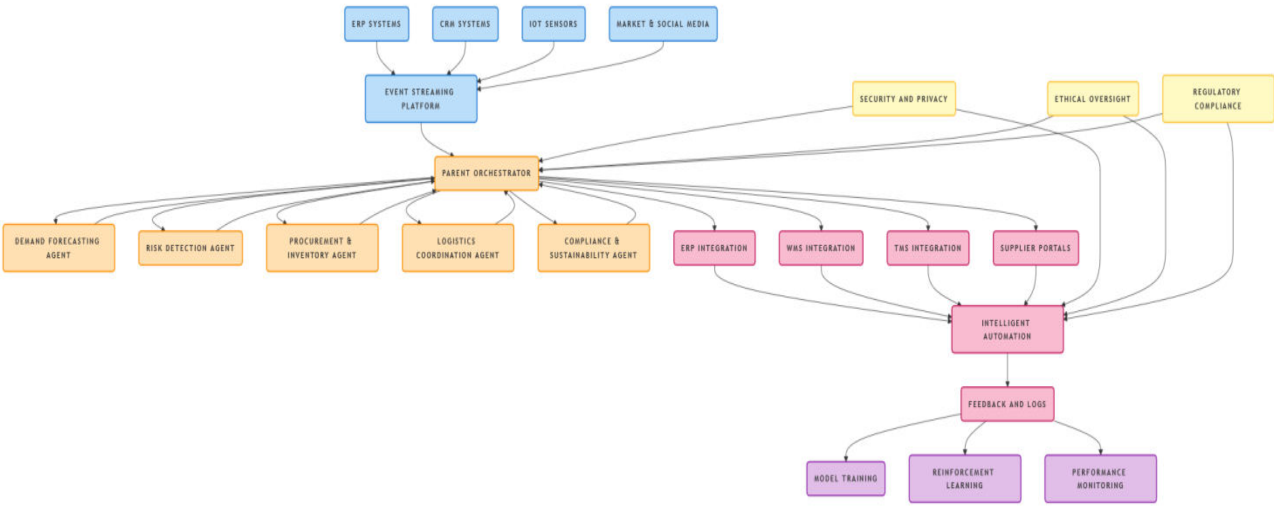


Figure 1: Architecture Layers for Agentic AI in Supply Chain Forecasting

1. **Data Sensing Layer:** The data constitutes various inputs, including market trends, social media, weather APIs, ERP, CRM, inventories, IoT sensors, and supplier feeds. To support this ingestion and normalization, it uses streaming event platforms such as AWS Kinesis and Apache Kafka.
2. **Multi-agent orchestration layer:** The Parent Orchestrator Agent manages delegation of tasks while enabling communication among the agents to ensure alignment with corporate objectives. Each of the specialized agents features a description of its functions or objectives; for example, the Demand Forecasting Agent uses machine learning to improve forecasts.
While the Risk Detection Agent focuses on risk detection through simulation and cause analysis of threats through a risk monitoring consideration for different inter-organization actors involved. The Procurement and Inventory Agent modifies strategy dynamically, the Logistics Coordination Agent determines an optimal delivery route, and the Compliance & Sustainability Agent considers ESG data and regulations. All agents communicate via low-latency cooridation through event-driven APIs.
3. **Decision and Action Layer:** Execution engines automate decision-making via apps through which they integrate with an ERP and transport management system. Intelligent automation manages activities like rerouting orders, renegotiating contracts, and scheduling. All order routing and scheduling depends on the agent recommendations. There is an opportunity to capture feedback on past activities and results in a knowledge base.
4. **Learning and Adaptation Layer:** Continuous model training utilizes real time data for model update, and reinforcement learning and simulation improve agent policy through interaction with the environment. Performance evaluation focuses on value for cost, responsiveness, and accuracy.
5. **Governance and Compliance Layer:** Security and privacy considerations in function can include audit trails of who actively used the technology, encryption, and access restrictions.Ethical oversight emphasizes explainability and human involvement in high-risk decision making. Moreover, regulations could modified for continued sustainability and adherence to regulatory norms.

The intended middleware solutions and APIs for real-time connectivity of an agentic AI supply chain architecture consist of several protocols and platforms. RESTful APIs are very successful in real-time transfer of information with JSON/XML protocols for orders and inventories well sampled across the range of current ERP, WMS, TMS, and IoT systems. GraphQL APIs can also improve data access . Message Queues and Pub/Sub protocols like Apache Kafka and MQTT enable many key features like scalable streaming of data asynchronously within supply chain contexts. Although newer technologies are evolving, EDI continues to function to standardized document transfers across supply chain operations. Integration platforms such as Enterprise Service Bus (ESB) and Integration Platform as a Service (iPaaS) act as middleware frameworks as both can facilitate secure communications and low-code integrations

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301168

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

respectively. API Gateways serve a helpful, optional role in assisting regulators manage and protect API traffic and statistic usage conveniently through measure analytics summation. Stream processing and data streaming technologies like Apache Kafka and AWS Kinesis support and manage streams of execution of applications and across IoT sensors with acceptable mechanisms. Next-generation best practices for integrations focus on staying consistent with data and method frameworks across integrations (e.g. API endpoints), logging and monitoring data and method usage tracking, and developing strategies of committed rollback whenever either practice integration development is established. Middle-ware solutions and APIs all together provide the enabling infrastructure to achieve real-time connectivity to the supply chain systems enhancing visibility and autonomy for AI to make effective decisions off of.

The Agentic AI Supply Chain Forecasting Architecture consists of five-different layers where varying technologies build a structure that enables functionality. The Data Sensing layer, for processors, will provision event streaming solutions such as REST APIs, MQTT, Amazon Kinesis, and Apache Kafka solution will provide means to seamlessly ingest data in real time on multiple products and the various `sources of that data. This all allows AI to continuously "sense" through the multi-sourced `awareness. The Multi-Agent Orchestration layer will host actor frameworks, gRPC define a western-like REST interface, and message queues like RabbitMQ and NATS to allow any specialized AI agents to communicate while coordinating tasking in real-time relative to alignment for goals. The Decision and Action layer will run automated flows connecting between corporate systems solutions via API interface and RPAs between stakeholder suppliers on orders. The Learning and Adaptation layer will improve decision-making over the decision-making processes through MLOps and monitor capability to lead planning for practices using Reinforcement Learning libraries and previous tools while allowing practice for more accurate decision making. Finally, the Governance and Compliance layer will add audit logging and check, encryption, of learning from human oversight with identity management (a.k.a security) with tools that build into proper levels of regulatory support, compliance and oversight relative to humans.

All together these elements and views will be and combine as hybridization to create a fundamentally cohesive intelligent supply chain agentic AI forecasting system capability of engineered planning and accuracy routine making for OC in me on risk management and predictive actions viability (see Figure 2):

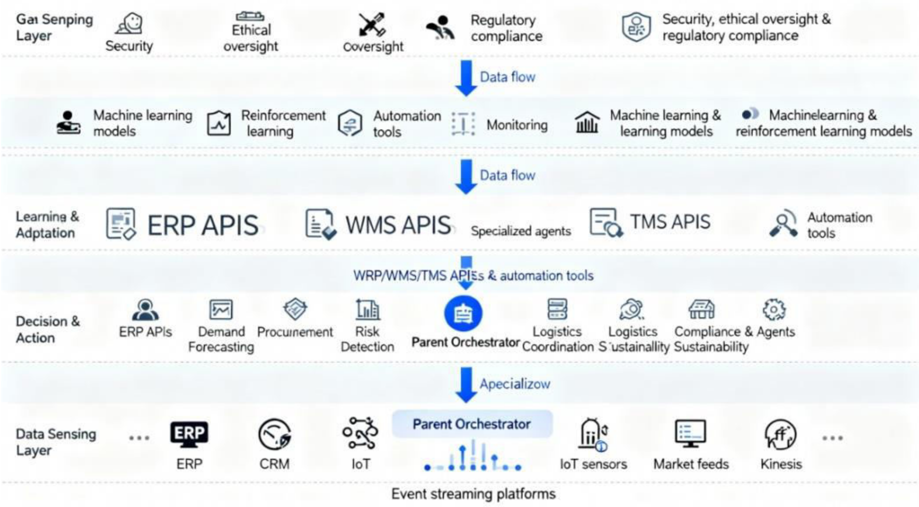


Figure 2: Intelligent Supply Chain Agentic AI Forecasting System

The key real-time KPIs for monitoring AI supply chain models in production can be divided in to technical and business KPIs. Technical KPIs include prediction latency which monitors the time frame from the moment data is taken to the predicted output, and confidence interval accuracy that helps ensure that expected uncertainty in the prediction aligns with outcomes. It is also imperative to have model drift detection to observe if there is a change in underlying data distributions that will require a retrain of the model. Additionally, we must track the anomaly detection rate which tracks unusual patterns identified by the AI and its severity and frequency, as well as automation success/fallback rate, which measures the effectiveness of the autonomous decision making. Moving to the business KPIs, indicators include

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the on time delivery rate communicated as a percentage of shipments that are on schedule, and inventory turnover ratio, which tracks the speed of inventory movement. The perfect order rate indicates an order that was error free, and supply chain cycle time accounts for the total amount of time from order to delivered. The order fill rate tracks the present inventory in comparison to potential customer demand. Include the cost to serve to track all related costs in transportation, warehousing, and procurement. Although for all of these KPIs there are numerous monitoring and visualization technologies to show live visualizations and alerting for these KPIS (e.g. Tableau or Power BI dashboards), a deploying model to monitor cloud-native AI performance where analysis can be completed in real-time using the managed service will increase coverage, e.g. AWS Cloud Native AI Monitoring Integration into Kinesis, leading to improved root cause and causal analysis. Overall monitoring of an AI supply chain forecasting system requires all these KPIs to quickly identify risk, optimize processes, and maintain high standards of service.

Next are the tools and best practices for tracking latency percentiles on streaming systems like Kafka and Kinesis. The recommended compass to tracking is the: Prometheus for metrics recording data. Grafana to visualize and alert on metrics. AWS CloudWatch for native Kinesis metrics with thresholds. Confluent Control Center for streaming Kafka metrics. Obkio for tracking network latency. And For testing performance - Netperf, k6, and JMeter. The KPIs for monitoring latency are producer and consumer latency measured at p50, p90, p99, consumer lag, and tail latency which should focus on high percentile delays such as on the 99th and 99.9th percentile that can be tracked in Table 1 below [14].

Tool/Platform	Key Query Example	Metric Tracked
Prometheus/Grafana	histogram_quantile(0.99, rate(latency_bucket[5m]))	p99 latency
CloudWatch	Percentile (p50, p95, p99) charts for data stream latency	Kinesis stream latency
Kafka	kafka_consumer_lag	Consumer lag
Netperf/k6/JMeter	Custom scripts for percentiles	Network/API latency

Table 1: Tools To monitor latency percentiles

We introduce a new idea to create a visual aid that tracks latency percentiles in streaming systems like Amazon Kinesis and Apache Kafka. The goal is to visually represent the data related to measuring latency metrics to understand how to characterize system performance and efficiency. The Latency Percentiles Chart tracks latency at different percentiles over time to provide visibility in a shift, or outlier behavior. It should be used to complement consumer lag (delay in processing messages) and the throughput, which indicates the system load. The Alarms & Thresholds Panel provides the user with a simple mechanism to assist with decision-making by indicating latency boundaries that alert them.

The Latency Histogram and CDF Curve will give background on latency percentile as it shows outliers and tail behavior. Percentile data is a better indication of user experienced latency than averages, especially in reporting distributed streaming systems all of which indicate tail latency is where to look when responding to delays in processing. Consumer latency is measures to ascertain timely consumptions of stream data, while throughput measures help link the surge in latency with the higher the load on the system. Alerts are established to inform the user with respect to business SLA. This dashboard idea can be used with Prometheus, Grafana, AWS CloudWatch, or a Kafka-like solution such as Confluent Control Center, thereby allowing real-time monitoring and alerts about the latency health of the streaming system is shown in below Figure 3:

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

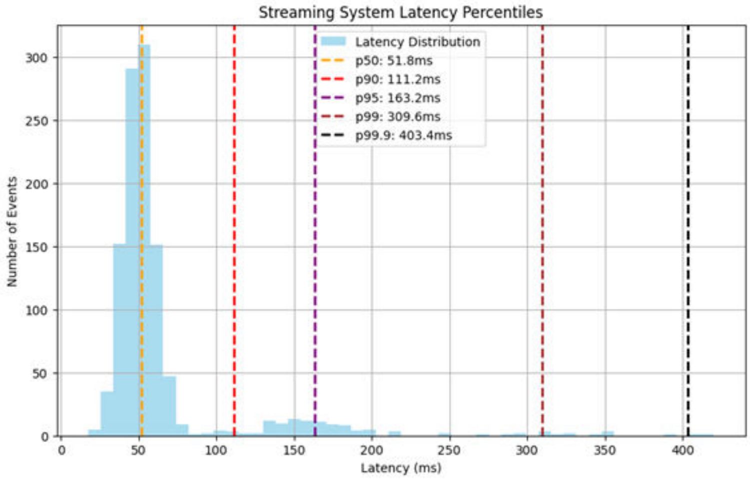


Figure 3: Streaming System Latency Percentiles

IV. CONCLUSION

Integrating autonomous, multi-agent AI systems with real-time streaming data systems (such as Kafka and Kinesis) will transform supply chain forecasting by providing speed and accuracy in demand forecasting, risk detection, procurement, and logistics. The three enablers of this modern supply chain intelligence will include real-time, multi-source data sensing for situational awareness; multi-agent orchestration of decentralized processes at scale; and an integrated layer of execution with existing systems to enable autonomous decision-making. Learning from feedback and reinforcement is foundational for sustained performance, and governance can deliver transparency, security, and compliance. Monitoring AI systems in real-time must be critical to ensure operational health and facilitate achievement of business goals; alarms for anomalies must facilitate a timely intervention. Enhanced AI explainability and human integration will not only increase user trust but will produce fully autonomous supply chains. Edge computing and decentralized data platforms will fully operationalize real-time decisions—an improved ethical framework will operationalize compliance inside AI workflows. Standardizing open datasets and open standards will accelerate scale and innovation. Businesses will now have the ability to enable effective autonomy within the evolving supply chain allowing for the increase of efficient and risk management.

REFERENCES

1. “MVP Development Trends for 2025”, Yashaswini S. P., April 29, 2025, https://think201.com/blog/mvp-development-trends/.
2. “How to Use MVP Development to Mitigate Risk: A Complete Guide”, May 16, 2023, https://www.iotric.com/blog/mvp-development-to-mitigate-risk/.
3. “How AI is Transforming Supplier Risk Management in Retail & CPG”, Michael Tracy, Jul 26, 2024, https://www.wns.com/perspectives/articles/how-ai-is-transforming-supplier-risk-management-in-retail-cpg.
4. “5 Common Data Pipeline Failures and How Rakuten SixthSense Solves Them”, January 22, 2025, https://sixthsense.rakuten.com/blog/5-Common-Data-Pipeline-Failures-and-How-Rakuten-SixthSense-Solves-Them.
5. “Data Ingestion: 7 Challenges and 4 Best Practices”, Michael Segner, https://www.montecarlodata.com/blog-data-ingestion/.
6. “What is Data Transformation? Types, Process and Techniques Explained”, Jeffrey Richman, June 26, 2025, https://estuary.dev/blog/what-is-data-transformation/.
7. “How to improve data quality: 8 steps and best practices”, Will Harris, February 19, 2025, https://www.metaplane.dev/blog/how-to-improve-data-quality.
8. “Artificial intelligence in supply chain management: A systematic literature review of empirical studies and research directions”, Giovanna Culot, Matteo Podrecca, Guido Nassimbeni, November 2024, https://doi.org/10.1016/j.compind.2024.104132.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

9. "AI-Driven Intelligent Choice Architecture: Empowering Leaders to Orchestrate the Future Supply Chain", Ajith Thomas, https://www.tcs.com/insights/global-studies/blogs/agentic-ai-smarter-faster-supply-chain.
10. "AGENTIC AI SUSTAINABILITY ASSESSMENT FOR SUPPLY CHAIN DOCUMENT INSIGHTS", Diego Gosmar, Anna Chiara Pallota, Giovanni Zenezini, November 11, 2025, https://arxiv.org/pdf/2511.07097.
11. "Demand Planning Best Practices for Inventory Accuracy", November 26, 2024, https://www.impactanalytics.co/blog/best-practices-in-demand-planning-and-forecasting.
12. "Top Metrics to Track When Measuring Agentic AI Performance", https://www.aiprime.global/blog/top-metrics-to-track-when-measuring-agentic-ai-performance.
13. "Why High Accuracy Doesn't Guarantee Reliable AI Agents", Conor Bronsdon, https://galileo.ai/blog/ai-agent-reliability-metrics.
14. "Unveiling Latency: Exploring Percentiles and Mean for Optimal Performance", Arunkumar HG, July 6, 2023, https://www.linkedin.com/pulse/unveiling-latency-exploring-percentiles-mean-optimal-performance-hg/.