ISSN 2964-0490

Resilient Data Analytics Pipelines for Fault-Tolerant Smart Manufacturing Systems

[1]Kenechukwu Favour Anagwu, [2]Okechukwu Chiedu Ezeanyim
[1]Department of Production Technology, Nnamdi Azikiwe University, P.M.B. 5025 Awka, Anambra State – Nigeria; [1,2]Industrial and Production Engineering Department, Nnamdi Azikiwe University, P.M.B. 5025 Awka, Anambra State - Nigeria.
Email : *oi.ezeanyim@unizik.edu.ng*

This review examines resilient data analytics pipelines as critical infrastructure for fault-tolerant smart manufacturing systems. It addresses the need for reliable, low-latency, and integrity-preserving data flows in industrial environments where sensor failures, network disruptions, data corruption, platform faults, and cyber-physical disturbances can compromise real-time analytics and autonomous decision-making. The study synthesizes recent literature on Industrial Internet of Things-enabled manufacturing, distributed stream processing, edge-fog-cloud computing, fault-tolerant architectures, and data governance. It analyses pipeline layers covering data sources, edge preprocessing, fault-tolerant ingestion, stream analytics, distributed storage, security, governance, and decision-support applications. Core resilience mechanisms examined include redundancy, replication, monitoring, checkpointing, rollback recovery, graceful degradation, secure aggregation, audit logging, and adaptive recovery. The review also evaluates technologies such as Apache Kafka, Apache Flink, Apache Spark, Storm, HDFS, Cassandra, MongoDB, cloud-agnostic platforms, microservices, and digital twins. Findings show that resilient analytics pipelines support predictive maintenance, real-time process monitoring, quality assurance, supply-chain optimization, and autonomous manufacturing by preserving data continuity and analytical reliability during faults. However, major challenges remain in balancing scalability with performance, maintaining data integrity across heterogeneous edge-fog-cloud layers, achieving low-latency recovery, integrating legacy systems, securing distributed data flows, and establishing standardized design frameworks. The review identifies future directions in AI-driven fault detection, digital-twin-based resilience testing, scalable distributed architectures, autonomous data management, and benchmarking protocols. Resilient pipelines are therefore essential operational assets for sustaining reliable, secure, and fault-tolerant smart manufacturing intelligence.

Keywords: Resilient data analytics pipeline; Smart manufacturing; Fault tolerance; Industrial Internet of Things; Edge-fog-cloud computing.

Corresponding Author:
Okechukwu Chiedu Ezeanyim
Industrial and Production Engineering Department,
Nnamdi Azikiwe University, P.M.B. 5025 Awka, Anambra State - Nigeria
oi.ezeanyim@unizik.edu.ng

1. Introduction

Background and Context

Smart manufacturing systems depend on continuous data flows from sensors, Industrial Internet of Things (IIoT) devices, and enterprise platforms to support real-time analytics, process optimization, and autonomous decision-making (Khattach et al., 2025; Syafrudin et al., 2018). The IIoT serves as an underlying platform for various smart manufacturers seeking effective and powerful data management structures (AlSuwaidan, 2020). These data streams require robust processing architectures capable of ingesting, transforming, and delivering actionable insights with minimal latency (Khattach et al., 2025; Çakır et al., 2022). Technologies such as Apache Kafka provide fault-tolerant, high-throughput data ingestion (Isah et al., 2019; Syafrudin et al., 2018), while frameworks like Apache Spark and Apache Flink enable scalable batch and stream processing (Igbokwe and Nwamekwe, 2025; Nasiri et al., 2019). The

effectiveness of manufacturing analytics depends on the reliability of the data pipelines that connect sensors to decision systems (Peres et al., 2018; Syafrudin et al., 2018). Industrial environments generate high-volume, heterogeneous data that demand distributed processing with strong delivery guarantees (Isah et al., 2019; Nwamekwe et al., 2025b). As manufacturing systems grow in complexity, the need for resilient data infrastructure becomes a foundational requirement for operational continuity and intelligent automation (Okpala et al., 2025; Çakır et al., 2022).

Problem Statement

Data analytics pipelines in manufacturing must handle high-velocity, heterogeneous data streams under uncertain and failure-prone conditions (Isah et al., 2019; Mehmood & Anees, 2020). Industrial settings are susceptible to sensor failures, network outages, data corruption, and system faults, all of which compromise data integrity and disrupt analytics processes (Javed et al., 2020; Dongen & Poel, 2021). Traditional pipeline architectures lack built-in fault tolerance and often fail under adverse conditions (Mehmood & Anees, 2020; Chidiebube et al., 2025). Stream processing frameworks address some of these concerns through mechanisms such as checkpointing, replication, and message delivery guarantees (Dongen & Poel, 2021; Geldenhuys et al., 2022). For instance, Flink offers distributed snapshots for state recovery (Dongen & Poel, 2021; Nwamekwe et al., 2025d), while Kafka provides replicated partitions ensuring data durability (Isah et al., 2019; Davoudian & Liu, 2020). Designing resilient pipelines requires integrating distributed systems, data management strategies, and fault-tolerant algorithms into unified frameworks (Nwamekwe et al., 2026; Marosi et al., 2022). There remains a gap in comprehensive approaches that address scalability, data integrity, and real-time fault recovery simultaneously within manufacturing contexts (Mehmood & Anees, 2020; Javed et al., 2020).

Objective of the Review

This review synthesizes current knowledge on resilient data analytics pipelines for fault-tolerant smart manufacturing systems. It examines architectural components, including ingestion layers, stream processing engines, and storage systems that form the backbone of modern data pipelines (Khattach et al., 2025; Marosi et al., 2022). It evaluates fault-tolerance strategies such as checkpointing, replication, and adaptive recovery mechanisms employed across distributed stream processing frameworks (Dongen & Poel, 2021; Geldenhuys et al., 2022). The review also considers application domains where these pipelines support predictive maintenance, quality control, and real-time monitoring in manufacturing (Ezeanyim et al., 2025a; Syafrudin et al., 2018; Peres et al., 2018). Challenges related to scalability, data integrity, and real-time implementation are identified alongside opportunities for future research (Isah et al., 2019; Mehmood & Anees, 2020; Assunção et al., 2018).

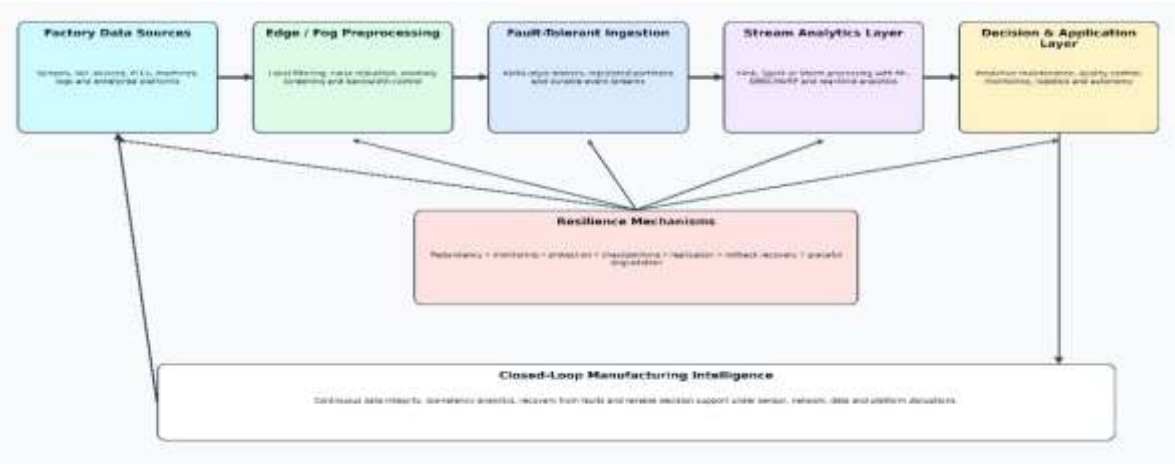

Figure 1: Conceptual Framework for Resilient Data Analytics Pipelines in Smart Manufacturing Systems

Figure 1 presents the end-to-end conceptual framework of a resilient manufacturing analytics pipeline. Data originating from sensors, IIoT devices, PLCs, and enterprise platforms passes through edge preprocessing, fault-tolerant ingestion, and stream analytics layers before supporting operational decisions. Resilience mechanisms, including redundancy, monitoring, replication, checkpointing, and recovery, operate across all stages to ensure continuous data integrity, reliable analytics, and closed-loop manufacturing intelligence.

2. Results

Conceptual Foundations of Resilient Data Analytics

1. Definition of Resilience in Data Systems

Resilience in data systems refers to the ability of a system to maintain an acceptable level of service in the face of various faults and challenges to normal operation (Khattach et al., 2025). Khattach et al. define resilience as encompassing behavioural stability when facing changes such as disruptions, attacks, or accidental faults, and note that a fault occurring in an individual component should not lead to the system becoming unable to match its behavioural requirements (Khattach et al., 2025). In practical terms, resilience involves four core mechanisms: redundancy, monitoring, protection, and recovery (Khattach et al., 2025). Redundancy masks present errors and faults, monitoring detects faults or attacks, protection shields against the occurrence of harm, and recovery steers the system back to a well-defined functional state (Khattach et al., 2025). Prokhorenko and Babar further elaborate that resilience in distributed systems requires both design-time architecture definition and run-time system adaptation, where failure recovery mechanisms include graceful degradation techniques to keep a system at least partially operational while recovery steps are being taken (Syafrudin et al., 2018). In the context of data analytics pipelines, resilience means the pipeline continues to ingest, process, and deliver data despite component failures, network disruptions, or data corruption, and adapts its behaviour to restore full functionality (Khattach et al., 2025; Syafrudin et al., 2018).

2. Characteristics of Smart Manufacturing Data

Data generated in smart manufacturing environments exhibits distinct characteristics that impose specific demands on analytics pipelines. Manufacturing data is high-frequency and continuous, originating from sensors, Industrial Internet of Things (IIoT) devices, and enterprise platforms that produce real-time streams requiring low-latency processing (AlSuwaidan, 2020; Çakır et al., 2022). Çakır et al. characterize IoT-generated sensor data from manufacturing processes as real-time, large in volume, and unstructured in type (Çakır et al., 2022). This data is distributed across multiple systems, spanning edge devices, fog nodes, and cloud platforms, which necessitates distributed processing architectures (Isah et al., 2019; Nasiri et al., 2019). Isah emphasizes that the massive data generated in IIoT environments comes from diverse sources including sensors, machine learning modules, performance management systems, and business intelligence platforms (Isah et al., 2019). The data is also heterogeneous in format and structure, combining structured numerical readings from accelerometers and temperature sensors with semi-structured logs and unstructured text or image data (AlSuwaidan, 2020; Peres et al., 2018). Isah et al. note that streaming data sources are not only common but depreciate rapidly if not processed quickly, and the ever-increasing volume and irregular nature of data rates pose new challenges to processing systems (Emeka et al., 2025a). These characteristics demand pipelines capable of handling high-volume, high-velocity, and heterogeneous data streams under uncertain conditions (AlSuwaidan, 2020; Chidiebube et al., 2025a).

3. Types of Failures in Data Pipelines

Data analytics pipelines in manufacturing environments face multiple categories of failures. Sensor and device failures represent a primary concern, as IoT devices deployed in harsh industrial settings are prone to malfunction, producing erroneous readings or ceasing to transmit data entirely (Khattach et al., 2025; Mehmood & Anees, 2020). Khattach et al. discuss sensor data quality analyses that attempt to quantify the quality of received data, noting this mechanism is highly application-specific (Khattach et al. (2025). Network disruptions constitute another common failure type, where node reachability faults, whether permanent or temporary, are expressed in terms of traffic-related degradation and measured through metrics such as frequency of failures, average outage duration, or packet loss rate (Syafrudin et al., 2018). Data corruption and loss occur when transmitted data is altered during transit or storage, compromising the integrity of downstream analytics (Javed et al., 2020; AlSuwaidan, 2020). Dongen and Poel identify several fault categories in stream processing systems, including master failures, worker failures, application failures, and task failures, each producing different impacts on system behaviour such as outages, downtime, data loss, and duplicate processing (Javed et al., 2020). Software and hardware faults at the platform level, including failures in processing engines, storage systems, and orchestration components, add further complexity (Javed et al., 2020; Syafrudin et al., 2018). Understanding these failure types is essential for designing targeted fault-tolerance strategies within analytics pipelines.

4. Fault-Tolerance Principles

Fault tolerance involves designing systems that continue to operate despite the occurrence of failures. The foundational principles include redundancy, replication, and graceful degradation (Khattach et al., 2025; Dongen & Poel, 2021; Syafrudin et al., 2018). Redundancy, the most widely applied principle, takes multiple forms: hardware redundancy through physical replication of components, software redundancy through replicated execution, information redundancy through error-correcting codes, and time redundancy through re-execution of operations (Dongen & Poel, 2021; Assunção et al., 2018). Replication in stream processing frameworks manifests as replicated partitions in systems like Apache Kafka, which ensure data durability, and distributed snapshots in Apache Flink, which enable state recovery after failures (AlSuwaidan, 2020; Javed et al., 2020). Graceful degradation allows a system to continue operating at a reduced level of service rather than failing completely when components are lost (Khattach et al., 2025; Syafrudin et al., 2018). Prokhorenko and Babar describe graceful degradation as techniques that keep a system at least partially operational while recovery steps are being taken (Syafrudin et al., 2018). Khattach et al. categorize resilience mechanisms into redundancy for masking faults, monitoring for detecting faults, protection for preventing harm, and recovery for restoring functional state (Chidiebube et al., 2025b). Checkpointing and rollback recovery represent additional fault-tolerance mechanisms, where the system periodically saves its state and rolls back to the last consistent checkpoint upon failure detection (Javed et al., 2020; Geldenhuys et al., 2022). These principles form the building blocks for constructing resilient data analytics pipelines.

5. Performance Metrics for Resilient Pipelines

Evaluating the resilience of data analytics pipelines requires well-defined performance metrics. System availability, commonly stated by the number of nines (e.g., 99.99%), depends on three properties: rate of failures, resiliency to failures, and recovery speed (Emeka et al., 2025). Dongen and Poel measure fault recovery performance across several dimensions, including whether there is an outage, the duration of downtime, recovery time, data loss, duplicate processing, accuracy, and the cost and behaviour of different message delivery guarantees (Javed et al., 2020). Recovery time is a critical metric, as it determines how quickly a pipeline returns to normal operation after a disruption. Their experiments show that Kafka Streams achieves the shortest downtime, while Flink offers end-to-end exactly-once semantics at low cost, and

Spark frameworks recover quickly due to their task-based scheduling approach (Javed et al., 2020). Data integrity metrics assess whether processed results remain correct and complete despite failures, with exactly-once processing semantics serving as the strongest guarantee (Javed et al., 2020; AlSuwaidan, 2020). Throughput under failure conditions measures the system's ability to sustain processing rates during and after faults, which is essential for real-time manufacturing analytics (AlSuwaidan, 2020; Nwamekwe et al., 2025c). Syafrudin et al. add that quantitative network metrics such as frequency of failures, average outage duration, and packet loss rate determine a system's overall survivability (Syafrudin et al., 2018). Together, these metrics provide a comprehensive framework for assessing and comparing the resilience of data analytics pipelines in smart manufacturing systems.

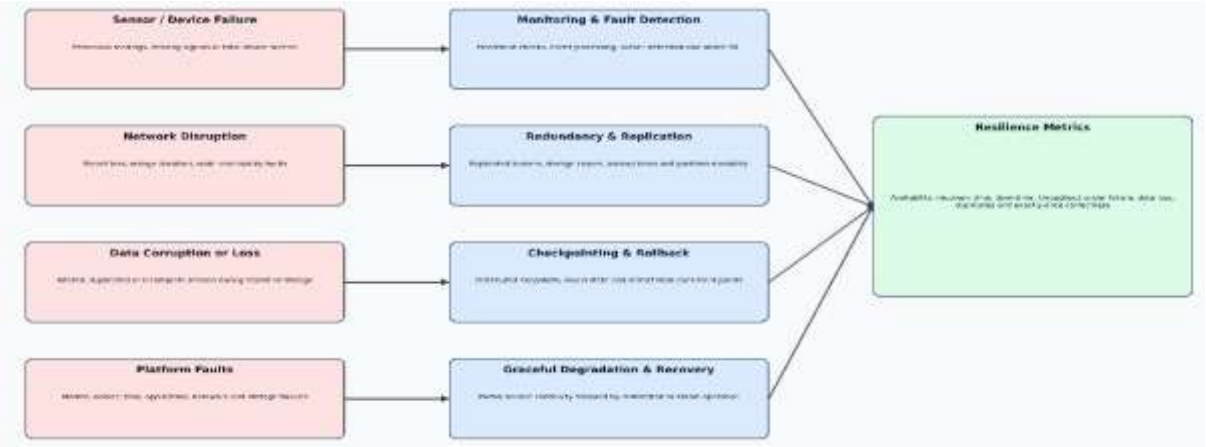

Figure 2: Failure Modes and Resilience Mechanisms in Data Analytics Pipelines

Figure 2 maps major pipeline failure modes to corresponding resilience strategies and performance outcomes. Sensor failures, network disruptions, data corruption, and platform faults are mitigated through monitoring, fault detection, redundancy, replication, checkpointing, rollback recovery, and graceful degradation. The integrated resilience framework improves availability, minimizes downtime and data loss, sustains throughput during disruptions, and supports accurate, fault-tolerant manufacturing analytics.

Architecture of Resilient Data Analytics Pipelines

1. Data Ingestion and Edge Processing Layer

The ingestion layer forms the entry point of any data analytics pipeline, accepting streams of data from sensors, IoT devices, and other sources into the processing system (Isah et al., 2019). Isah et al. describe a generic data stream processing architecture where the ingestion layer ensures scalable, resilient, and fault-tolerant data distribution across the pipeline from multiple input data streams, decoupling the input sources from the rest of the system (Isah et al., 2019). In smart manufacturing, this layer collects high-frequency sensor readings from equipment on the factory floor and transmits them to downstream components for analysis (Khattach et al., 2025). Khattach et al. present an architecture where IoT data streams are ingested via Apache Kafka, a distributed event streaming platform that acts as the ingestion layer, providing high-throughput and fault-tolerant message brokering (Khattach et al., 2025). Edge processing adds a local computation step before data reaches centralized systems. Cañizo et al. describe how data is first gathered from industrial machines through programmable logic controllers installed on each machine, persisted in a local database, and then forwarded to the cloud (Cañizo et al., 2019). This local preprocessing at the edge reduces bandwidth consumption, filters noise, and enables preliminary anomaly detection closer to the data source (Srirama, 2024; Cardellini et al., 2022). Srirama discusses how fog computing enables proximal computational and storage devices to perform sensor data analytics, providing preliminary insights from

data streams and protecting cloud-based storage against massive volumes and high data velocity (Srirama, 2024). The combination of edge preprocessing with a robust ingestion layer creates a first line of defence against data quality issues and network disruptions in manufacturing environments (Igbokwe et al., 2024; Cañizo et al., 2019).

2. Stream Processing and Analytics Layer

The stream processing layer sits at the core of the pipeline, analysing data in real time as it flows through the system. Isah et al. describe this layer as responsible for preprocessing and analysing data in one or more steps, forming the computational backbone of any distributed stream processing system (Isah et al., 2019). Modern stream processing engines such as Apache Flink, Apache Spark Streaming, and Apache Storm provide the distributed infrastructure needed to handle high-velocity manufacturing data (Isah et al., 2019; Dongen & Poel, 2021). Dongen and Poel compare four state-of-the-art frameworks and find that each offers distinct trade-offs: Spark frameworks recover quickly from faults due to their task-based scheduling, Kafka Streams achieves the shortest downtime, and Flink offers end-to-end exactly-once semantics at low cost (Dongen & Poel, 2021). Khattach et al. present an architecture where Spark Streaming processes data in micro-batches, with a modular machine learning pipeline handling automated data preprocessing, training, and evaluation (Khattach et al., 2025). Handling missing or corrupted data is a critical function of this layer. Syafrudin et al. (2018) propose a hybrid prediction model combining DBSCAN-based outlier detection with Random Forest classification to remove outlier sensor data and provide fault detection during the manufacturing process (Igbokwe et al., 2025). Cardellini et al. note that data stream processing applications are typically long-running and experience varying workloads over time, requiring runtime adaptation strategies to maintain consistent service levels (Cardellini et al., 2022). The stream processing layer must therefore integrate both analytical functions and data quality mechanisms to ensure reliable outputs for downstream decision-making (Isah et al., 2019; Khattach et al., 2025).

3. Storage and Data Management Layer

Distributed storage systems ensure data availability and durability across the pipeline. Isah et al. identify the storage layer as responsible for storing, indexing, and managing data and generated knowledge within the processing architecture (Isah et al., 2019). The Hadoop Distributed File System (HDFS) remains a foundational technology, providing high-throughput access to application data across thousands of machines in a fault-tolerant manner through data replication (Khalid & Yousaf, 2021). Khalid and Yousaf explain that HDFS and similar systems employ data replication and redundant storage of in-process metadata to recover from faults during data processing (Khalid & Yousaf, 2021). Dubuc et al. describe multiple storage policies for preventing data loss, including duplicating task-sensitive data on separate hardware to avoid cascading failures and using fault-tolerant file systems to mitigate hardware or system failure Dubuc et al., 2021). NoSQL databases such as Apache Cassandra and MongoDB serve specific roles in manufacturing pipelines. Syafrudin et al. (2018) use MongoDB to store sensor data from the manufacturing process, accommodating the unstructured nature of IoT-generated data (Çakır et al., 2022). Oza et al. describe Cassandra as offering a flexible, scalable NoSQL data store for processed pipeline data, supporting the creation of resilient real-time pipelines when combined with Kafka and Spark (Oza et al., 2024). Alsuwaidan emphasizes that effective data management in the Industrial Internet of Things requires structures that address data sources, machine learning integration, performance management, and business intelligence (AlSuwaidan, 2020). The storage layer must balance write throughput for real-time ingestion with read performance for analytics queries, all while maintaining replication for fault tolerance (Okpala et al., 2024; Dubuc et al., 2021).

4. Fault Detection and Recovery Mechanisms

Fault detection and recovery mechanisms form the resilience backbone of data analytics pipelines. Khalid and Yousaf state that fault detection is the starting point of any fault-tolerant mechanism, enabling faults to be detected as soon as they appear within the system, with most big data frameworks relying on heartbeat detection approaches (Khalid & Yousaf, 2021). Power and Kotonya propose a microservices-based framework that implements fault tolerance through two complementary services: one using complex event processing for real-time fault detection, and another using online machine learning to detect fault patterns and pre-emptively mitigate faults before activation (Power & Kotonya, 2018). Checkpointing represents a primary recovery mechanism in stream processing systems. Jayasekara et al. explain that state-of-the-art distributed stream processing systems such as Apache Flink and Storm use checkpointing to provide fault tolerance for stateful applications, periodically saving system state for rollback upon failure (Jayasekara et al., 2020). Geldenhuys et al. present Khaos, an approach that dynamically optimizes checkpoint configurations at runtime by borrowing from chaos engineering principles: establishing steady-state conditions, conducting failure experiments, and using the knowledge to minimize quality of service violations (Geldenhuys et al., 2022). Dongen and Poel provide empirical evidence showing that different frameworks exhibit different recovery behaviours, with Spark frameworks recovering without application restarts in most cases, while Flink requires longer recovery times but maintains stronger correctness guarantees (Dongen & Poel, 2021). Cheng et al. introduce approximate fault tolerance through AF-Stream, which adaptively issues backups while ensuring bounded errors, offering a practical trade-off between performance and accuracy (Cheng et al., 2019). These mechanisms work together to maintain pipeline continuity in the face of the diverse failure modes encountered in manufacturing environments (Khalid & Yousaf, 2021; Dongen & Poel, 2021; Jayasekara et al., 2020).

5. Integration with Cloud and Distributed Systems

Cloud platforms provide the scalability and redundancy needed to support large-scale manufacturing analytics. Marosi et al. introduce a scalable, cloud-agnostic, and fault-tolerant data analytics platform built from open-source reusable building blocks, serving as an architecture blueprint for processing and analysing various feeds of structured and unstructured input data from IoT-based use cases (Marosi et al., 2022). Javed et al. propose IoTEF, a federated edge-cloud architecture for fault-tolerant IoT applications, where processing occurs closer to data sources while sharing results across heterogeneous clusters (Javed et al., 2020). This edge-cloud continuum simplifies data exchanges between IoT systems, the cloud, and the edge for managing low latency, minimal bandwidth, and fault-tolerant applications (Javed et al., 2020). Vergilio et al. present MC-BDP, a reference architecture for big data stream processing in a containerized, multi-cloud environment, identifying fault tolerance and scalability as key issues and mitigating vendor lock-in through technology agnosticism (Vergilio et al., 2022). Assunção et al. survey resource elasticity features of cloud computing in stream processing, noting that elasticity allows applications to scale out or in according to fluctuating demands, though achieving elastic systems that make efficient resource management decisions based on current load remains challenging (Igbokwe et al., 2024a). Distributed architectures enhance fault tolerance by eliminating single points of failure and enabling workload distribution across multiple nodes (Khalid & Yousaf, 2021; Marosi et al., 2022). The integration of cloud resources with edge and fog computing creates a layered processing architecture well suited to the latency and throughput requirements of smart manufacturing systems (Javed et al., 2020; Srirama, 2024).

6. Security and Data Governance Layer

Security mechanisms protect data integrity and prevent unauthorized access throughout the analytics pipeline. Asaithambi et al. identify security as a core component of their microservice-oriented big data

ISSN 2964-0490

architecture, listing four essential components of the security broker layer for managing sensitive data, including authentication, authorization, encryption, and audit logging (Asaithambi et al., 2020). Khan et al. present FESDAO, a fog-enabled secure data analytics scheme that incorporates secure aggregation, authentication, fault tolerance, and resilience against insider threats, achieving privacy during data aggregation through a modified cryptographic scheme (Khan et al., 2025). Data governance encompasses policies and practices for maintaining data quality, lineage, and compliance across the pipeline. Asaithambi et al. describe data lineage as capturing data visibility across all stages of the pipeline, where every action is traced back to its source, helping in troubleshooting and recovery of middle-stage data by running only the affected parts of the pipeline (Igbokwe et al., 2025a). Dubuc et al. recommend duplicating task-sensitive data on separate hardware and using fault-tolerant file systems to prevent data loss, while also noting the role of Redundant Arrays of Inexpensive Disks within the architecture for storage failure recovery (Dubuc et al., 2021). In manufacturing contexts, security and governance are especially important because compromised data integrity leads directly to incorrect process decisions and potential safety hazards (Khan et al., 2025; Asaithambi et al., 2020). The security layer must operate across all pipeline stages, from ingestion through storage and analytics, without introducing latency that would compromise real-time processing requirements (Okpala et al., 2025a; Dubuc et al., 2021).

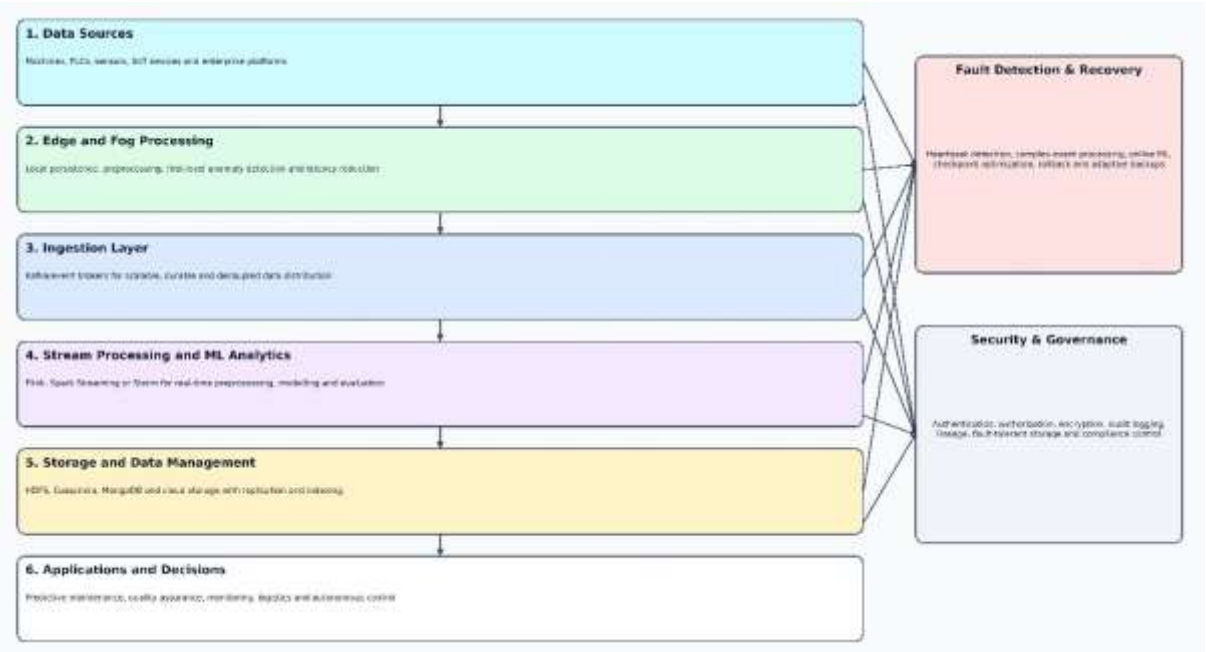

Figure 3: Multi-Layer Architecture of Resilient Data Analytics Pipelines with Fault Detection and Recovery Mechanisms

Figure 3 illustrates the layered architecture of resilient manufacturing analytics systems. Data progresses through source, edge processing, ingestion, stream analytics, storage, and application layers. Fault detection and recovery functions operate horizontally across all layers, while security and governance provide authentication, encryption, auditability, and compliance controls. The architecture demonstrates how resilience, scalability, and data integrity are embedded throughout the analytical workflow.

Applications in Smart Manufacturing Systems

1. Predictive Maintenance and Fault Detection

Predictive maintenance represents one of the most established applications of resilient data analytics pipelines in smart manufacturing. The core idea is to use real-time sensor data to anticipate equipment failures before they occur, reducing unplanned downtime and maintenance costs (Kang et al., 2021;

Rehman et al., 2019). Rehman et al. note that the concept of adopting big data analytics for intelligent predictive maintenance is gaining traction, though new avenues need exploration to fully realize real-time prediction systems (Rehman et al., 2019). Angelopoulos et al. describe how machine learning solutions for fault detection and prediction in Industry 4.0 rely on cloud, fog, and edge architectures to acquire manufacturing data for training algorithms, where fault detection involves data collection, feature extraction, and fault classification (Angelopoulos et al., 2019). The reliability of these predictive models depends directly on the quality and continuity of the data feeding them. Kang et al. emphasize that cyber-manufacturing is expected to enhance predictive maintenance through data-driven decision-making, but information quality challenges arise when a single dataset serves multiple computation tasks such as fault diagnosis, maintenance scheduling, and quality prediction simultaneously (Kang et al., 2021). Peres et al. present the IDARTS framework, which combines distributed data acquisition, machine learning, and run-time reasoning to assist in predictive maintenance and quality control, reducing the impact of events on production (Peres et al., 2018). Syafrudin et al. (2018) demonstrate a practical implementation where a hybrid prediction model combining DBSCAN-based outlier detection and Random Forest classification achieves better fault prediction accuracy than other models when applied to sensor data from an automotive manufacturing assembly line in Korea (Çakır et al., 2022). Resilient pipelines ensure that sensor data reaches these models without corruption or loss, even during network disruptions or component failures, making the fault detection process dependable under real operating conditions (Kang et al., 2021; Okeagu et al., 2024).

2. Real-Time Process Monitoring

Continuous monitoring of manufacturing processes supports early detection of deviations and enables timely corrective actions. Cañizo et al. present a large-scale platform for cyber-physical system real-time monitoring based on big data technologies, validated on industrial press machines in a real work environment, where the implementation improved overall equipment effectiveness (Cañizo et al., 2019). The platform gathers data from programmable logic controllers installed on each machine, persists it in a local database, and forwards it to cloud-based analytics (Cañizo et al., 2019). Peres et al. describe how the IDARTS framework integrates a cyber-physical system at the edge with cloud computing to provide real-time supervision for manufacturing environments, aligned with the industry 4.0 trend (Peres et al., 2018). Çakır et al. propose a real-time monitoring system for automotive manufacturing that uses IoT-based sensors collecting temperature, humidity, accelerometer, and gyroscope data, processed through Apache Storm as a real-time processing engine (Çakır et al., 2022). Nwamekwe et al. (2026a) present an analytics environment architecture for industrial cyber-physical systems that supports real-time analytics through an attribute-driven design approach, gathering requirements from smart production planning and manufacturing process monitoring scenarios (Igbokwe et al., 2025b). Their architecture includes a data-driven event component that monitors submitted data in real time and returns predictions for fault detection events (Nwamekwe et al., 2020). The resilience of the underlying data pipeline is essential for these monitoring systems, as any interruption in data flow leads to blind spots in process visibility. Cardellini et al. note that data stream processing applications are long-running and experience varying workloads over time, requiring runtime adaptation strategies to maintain consistent service levels during monitoring operations (Cardellini et al., 2022).

3. Quality Control and Assurance

Reliable data pipelines improve quality analytics and defect detection by ensuring that measurement data reaches analytical models without degradation. Huang et al. describe how AI-driven digital twins serve as core elements in modern manufacturing for quality control, where the digital process twin learns and

interprets correlations between manufacturing processes and material, process, and environmental parameters from heterogeneous data (Huang et al., 2021). Yang et al. discuss multidimensional real-time data analytics for quality control, enabling manufacturers to build models between key quality metrics and process parameters for proactive and precise quality management (Yang et al., 2020). Caiazzo et al. propose a five-layer IoT-based monitoring architecture that combines control charts, autoencoders, LSTM networks, and fuzzy inference systems to detect product defects and recognize their causalities during solar thermal panel production (Caiazzo et al., 2022). Their system provides human operators with information about anomalous events and their risk levels, enabling targeted intervention (Caiazzo et al., 2022). Suvarna et al. describe how data-driven machine learning approaches offer a generic and computationally faster approach for process monitoring, quality control, and effective system integration in manufacturing environments (Suvarna et al., 2020). Angelopoulos et al. further note that machine learning capabilities for timely processing of abundant data are critical for safeguarding quality in IIoT-enabled interconnected manufacturing environments (Angelopoulos et al., 2019). The integrity of the data pipeline directly determines the accuracy of quality predictions. When data is corrupted or lost during transmission, defect detection models produce unreliable outputs, leading to either missed defects or false alarms that disrupt production flow (Huang et al., 2021; Caiazzo et al., 2022).

4. Supply Chain and Logistics Management

Resilient data systems ensure continuity of data-driven decision-making across supply chain and logistics operations. Rehman et al. identify logistics and supply chain management as a typical industrial analytics application, where the appropriate use of analytics helps with condition monitoring, supply chain optimization, fleet management, and strategic supplier management (Rehman et al., 2019). Panetto et al. describe how supply chain and operations analytics applications include logistics and supply chain control with real-time data, inventory control using sensing data, and dynamic resource allocation in Industry 4.0 customized assembly systems (Panetto et al., 2019). Sahal et al. propose a blockchain-empowered digital twins collaboration framework for smart logistics, where digital twins use real-time operational data analytics and distributed consensus algorithms to predict potential risks within distributed manufacturing systems (Sahal et al., 2021). Their framework addresses logistics data generated from sensors attached to containers, fleets, warehouses, and robots that capture real-time data about logistic items and report on environmental changes (Sahal et al., 2021). Kang et al. note that disruptive events such as the COVID-19 pandemic accelerate the adoption of data-driven decision-making in manufacturing, requiring that production, supply chain, quality engineering, and reliability engineering decisions be made automatically (Kang et al., 2021). Vital-Soto and Olivares-Aguila emphasize that manufacturing systems are frequently exposed to disturbances and risks affecting everyday operations, making it imperative to analyse and design proactive and reactive strategies for supply chain disruptions (Vital-Soto & Olivares-Aguila, 2023). When data pipelines fail during supply chain operations, the resulting information gaps lead to delayed shipments, inventory imbalances, and missed demand signals, all of which carry direct financial consequences (Rehman et al., 2019; Panetto et al., 2019).

5. Autonomous Manufacturing Systems

Fault-tolerant pipelines support autonomous decision-making systems that operate with minimal human intervention. Kang et al. describe how cyber-manufacturing, founded upon advanced communication, computation, and control infrastructure, is expected to enhance intelligent production planning, flexible and autonomous manufacturing processes, and human-machine integration (Kang et al., 2021). Zeadally et al. discuss self-adaptation techniques in cyber-physical systems, where an important feature of current and future systems is the ability to adapt autonomously, yet designing and implementing self-adaptive

mechanisms remains challenging (Zeadally et al., 2019). Parri et al. describe a hardware and software framework supporting operation and maintenance of software-controlled systems that enhances resilience through a model-driven engineering process, where a reflective architecture developed around digital twins enables representation and control of system configuration items with support for runtime self-assessment and dynamic adaptation (Parri et al., 2020). Their framework empowers the system with self-recovery and self-adaptation properties demonstrated on a real cyber-physical system (Parri et al., 2020). Andronie et al. note that in cyber-physical production systems, smart connected devices team up automatically to optimize manufacturing processes, manage disturbances, and adjust to variable conditions (Andronie et al., 2021). Chae et al. survey industrial cyber-physical systems and describe how AI augmentation enables monitoring techniques across defect detection, fault detection, fault prediction, and quality prediction, all of which require continuous and reliable data streams Chae et al., 2023). The autonomy of these systems depends on uninterrupted access to accurate data. A pipeline failure during autonomous operation leads to decisions based on stale or incomplete information, which in safety-critical manufacturing contexts carries risks of equipment damage, product defects, or worker safety hazards (Zeadally et al., 2019; Parri et al., 2020; Chae et al., 2023).

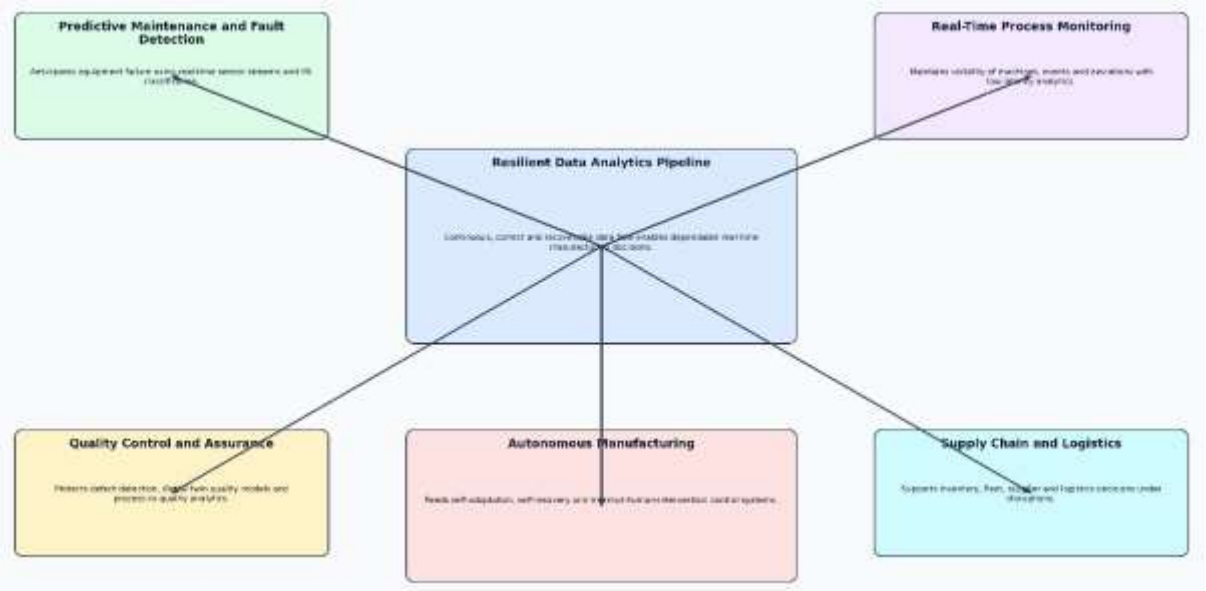

Figure 4: Application Domains of Resilient Data Analytics Pipelines in Smart Manufacturing Systems

Figure 4 demonstrates how resilient analytics pipelines support critical manufacturing applications. Reliable data flows enable predictive maintenance, real-time process monitoring, quality assurance, supply chain optimization, and autonomous manufacturing operations. The central pipeline serves as the foundation for continuous decision support, ensuring that operational intelligence remains accurate and available despite failures, disruptions, or fluctuating manufacturing conditions.

Key Challenges and Research Gaps

1. Scalability and Performance Trade-offs

Balancing resilience mechanisms with processing performance remains a persistent challenge in data analytics pipelines for smart manufacturing. Fault tolerance features such as checkpointing, replication, and state backup introduce computational overhead that directly affects throughput and latency (Khattach et al., 2025; Syafrudin et al., 2018). Dongen and Poel (2021) demonstrate through empirical testing that enabling exactly-once semantics in stream processing frameworks comes at a measurable cost, with different frameworks exhibiting different trade-offs between recovery guarantees and processing speed (Khattach et al., 2025). Geldenhuys et al. (2022) observe that static fault tolerance configurations often fail

to meet quality of service constraints with low overhead, because the statistical probability of partial failures and the variability of workloads make a single configuration suboptimal across operating conditions (Syafrudin et al., 2018). Assunção et al. (2018) note that resource elasticity, the ability to scale out or in according to fluctuating demands, has been extensively investigated for enterprise applications, but stream processing poses specific challenges in making efficient resource management decisions based on current load (AlSuwaidan, 2020). Cardellini et al. (2022) add that data stream processing applications experience varying workloads over time, and maintaining consistent service levels requires runtime adaptation strategies that themselves consume resources (Çakır et al., 2022). Kang et al. (2021) point out that the complexity of information technology infrastructure in cyber-manufacturing leads to fundamental challenges, where a single dataset often serves multiple computation tasks such as fault diagnosis, maintenance scheduling, and quality prediction simultaneously, creating competing demands on pipeline resources (Isah et al., 2019). Rehman et al. (2019) identify that big data processing in IIoT is challenging due to limited computational, networking, and storage resources at the IoT device end, making scalability a core concern (Ezeanyim et al., 2025). The research gap lies in developing adaptive mechanisms that dynamically adjust the level of fault tolerance based on current system load and failure probability, rather than relying on fixed configurations (Vitalis et al., 2024; AlSuwaidan, 2020).

2. Data Integrity and Consistency

Ensuring consistent data across distributed systems presents a complex challenge for resilient manufacturing pipelines. Distributed stream processing systems must guarantee that data is processed correctly despite failures, and the strength of this guarantee varies across frameworks (Onyeka et al., 2024; Peres et al., 2018). Dongen and Poel compare message delivery guarantees across four frameworks and find significant differences: Flink offers end-to-end exactly-once semantics at low cost, while other frameworks provide at-least-once guarantees that result in duplicate processing during recovery (Khattach et al., 2025). Isah et al. note that the ever-increasing volume and irregular nature of data rates pose new challenges to data stream processing systems, particularly regarding accurate ingestion and integration of data streams from various sources and locations (Peres et al., 2018). Khalid and Yousaf explain that most big data frameworks rely on heartbeat detection approaches for fault detection, and data replication combined with redundant storage of in-process metadata enables recovery from faults during data processing (Nwamekwe et al., 2025e). Cheng et al. introduce the concept of approximate fault tolerance through AF-Stream, which adaptively issues backups while ensuring bounded errors, offering a practical trade-off between performance and accuracy (Mehmood & Anees, 2020). Their approach allows users to specify bounds on both state divergence and the loss of non-backup streaming items, issuing backups only when bounds are reached (Mehmood & Anees, 2020). Syafrudin et al. demonstrate that data quality issues in manufacturing sensor streams require dedicated outlier detection mechanisms, as corrupted or anomalous readings directly affect downstream analytics accuracy (Javed et al., 2020). The research community still lacks unified approaches for maintaining end-to-end data integrity across heterogeneous pipeline components spanning edge, fog, and cloud layers, where each layer introduces its own consistency challenges (Peres et al., 2018; Nwamekwe et al., 2025).

3. Real-Time Processing Constraints

Maintaining low latency under failure conditions is a difficult requirement for manufacturing analytics pipelines. Yang et al. identify that cloud-based big data analytics and decision-making often fail to meet the requirements of latency-sensitive applications on shop floors, as data communication between edge devices and the cloud, along with data collection, cleaning, and synchronization, consumes significant time (Dongen & Poel, 2021). Dongen and Poel measure recovery times across stream processing frameworks

and find that downtime during failures varies from seconds to minutes depending on the framework and fault type, with master failures causing longer outages than worker failures (Khattach et al., 2025). Jayasekara et al. demonstrate that current systems use nominal checkpoint intervals that do not account for salient aspects of the checkpoint process or system scale, leading to inefficient operation, and they derive an optimal checkpoint interval dependent on checkpoint cost and failure rate (Assunção et al., 2018). Srirama discusses how fog computing enables proximal computational and storage devices to perform sensor data analytics, providing preliminary insights from data streams and protecting cloud-based storage against massive volumes and high data velocity (Geldenhuys et al., 2022). Cañizo et al. present a practical implementation where data is first gathered from industrial machines through programmable logic controllers, persisted in a local database, and then forwarded to the cloud, reducing latency for time-critical monitoring (Nwamekwe et al., 2025a). Nwamekwe and Igbokwe, (2024) describe a data-driven event component that monitors submitted data in real time and returns predictions for fault detection events, requiring consistent low-latency performance (Davoudian & Liu, 2020). The gap between theoretical latency guarantees and practical performance during failure recovery remains significant, particularly for safety-critical manufacturing applications where millisecond-level response times are required (Khattach et al., 2025; Dongen & Poel, 2021; Assunção et al., 2018).

4. Integration with Legacy Systems

Existing manufacturing systems often lack the architectural features needed to support resilient data analytics pipelines. Yang et al. observe that existing manufacturing systems lack sufficient reconfigurability, openness, and evolvability to deal with shop-floor disturbances and market changes (Dongen & Poel, 2021). Panetto et al. describe how complex and heterogeneous enterprise systems, built by different stakeholders at different times, need an environment that allows integration of systems forming a System-of-Systems, and the changing environment requires models that adapt over time (Marosi et al., 2022). Kang et al. note that the complexity of information technology infrastructure leads to fundamental challenges in cyber-manufacturing, ranging from information-poor datasets to a lack of reproducibility of analytical studies (Nwamekwe & Nwabunwanne, 2025). Rehman et al. identify that although existing literature still lacks the concept of automated data pipelines in IIoT systems, big data analytics processes are executed as a sequence of operations during data engineering, preparation, and analytics, requiring a holistic approach to execute and administer these processes across all layers of concentric computing systems (Nasiri et al., 2019). Alsuwaidan discusses how the transformation of the traditional manufacturing paradigm toward smart manufacturing requires effective data management structures, but legacy systems in industries such as oil and gas were not designed for the data volumes and velocities generated by modern IIoT deployments (Isah et al., 2019). Soni et al. note that industrial data platforms need optimization for deployment in environments with limited connectivity and extreme conditions, and research should focus on creating resilient architectures capable of maintaining functionality under such constraints (Khattach et al., 2025). Bridging the gap between legacy operational technology and modern data analytics infrastructure remains an open problem, as many factories operate equipment with decades-old communication protocols and proprietary data formats (Marosi et al., 2022; Onyeka & Emeka, 2025).

5. Security and Privacy Concerns

Distributed systems increase exposure to cyber threats, and securing data analytics pipelines across multiple layers presents significant challenges. Leng et al. explain that the traditional centralized IIoT framework is vulnerable to cyber-attacks and single-point failure, making it unsuitable for achieving resilient manufacturing (Cañizo et al., 2019). They note that IIoT contains confidential data and private information, and many security issues arise through vulnerabilities in the infrastructure (Cañizo et al., 2019). Khan et al.

present FESDAO, a fog-enabled secure data analytics scheme that incorporates secure aggregation, authentication, fault tolerance, and resilience against insider threats, addressing the need for privacy during data aggregation through a modified cryptographic scheme (Srirama, 2024). Asaithambi et al. identify four essential components of the security broker layer for managing sensitive data in big data architectures: authentication, authorization, encryption, and audit logging (Cardellini et al., 2022). Vijayakumaran et al. present a next-generation cyber security architecture for IIoT environments that detects and thwarts cybersecurity threats through an automated cyber-defence authentication mechanism generating cryptographically encrypted identity tokens verified by a virtual gateway system (Dongen & Poel, 2021). Khalid & Yousaf underscore that IoT devices play a critical role in Industry 4.0 data collection and transmission but are not inherently equipped to run strong encryption mechanisms to secure the data they transmit over wired or wireless channels (Khalid & Yousaf, 2021). The challenge intensifies as pipelines span edge, fog, and cloud layers, each with different security postures and attack surfaces, and securing the entire data flow without introducing latency that compromises real-time processing requirements remains an open research problem (Cañizo et al., 2019); Srirama, 2024; Dongen & Poel, 2021).

6. Lack of Standardized Frameworks

No unified approach exists for designing and implementing resilient data analytics pipelines in smart manufacturing. Isah et al. identify the difficulty in selecting the right stream processing framework for different use cases as a key challenge in developing streaming analytics infrastructure (Peres et al., 2018). Vergilio et al. present MC-BDP as a reference architecture for big data stream processing in a containerized, multi-cloud environment, identifying fault tolerance and scalability as key issues and mitigating vendor lock-in through technology agnosticism, but acknowledge that this represents one approach among many (Dubuc et al., 2021). Marosi et al. introduce a scalable, cloud-agnostic, and fault-tolerant data analytics platform built from open-source reusable building blocks, serving as an architecture blueprint, yet note that it represents a baseline for further new reference architectures rather than a definitive standard (Oza et al., 2024). Power and Kotonya observe three main approaches for designing software architectures for big data systems: adopting a reference architecture, following an architectural design methodology, and using a reference model, but find no consensus on which approach best serves manufacturing contexts (Power & Kotonya, 2018). Panetto et al. call for an infrastructure that supports loose coupling and evolving systems of systems, noting that models and systems need to be modular and support modification and self-adaptation (Marosi et al., 2022). Rehman et al. identify the need for an end-to-end industrial analytics pipeline that handles big data from various data sources in parallel and finds correlated knowledge patterns across entire IIoT systems (Nasiri et al., 2019). The absence of standardized frameworks forces each manufacturing organization to design custom solutions, leading to fragmented implementations that are difficult to benchmark, compare, or replicate across different industrial settings (Peres et al., 2018; Okpala et al., 2024a; Oza et al., 2024)

ISSN 2964-0490

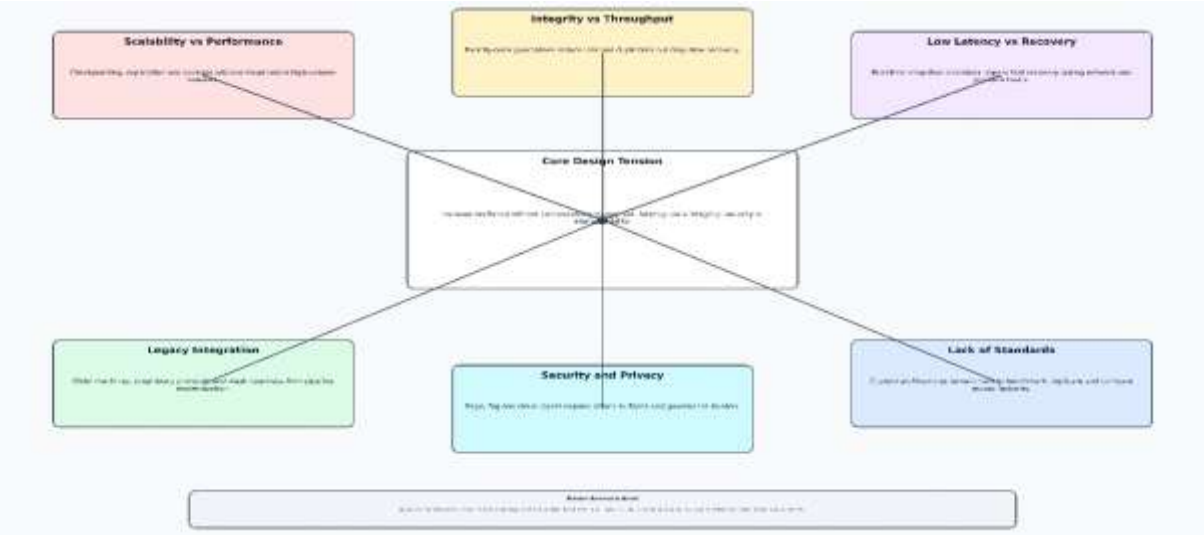

Figure 5: Challenges and Trade-offs in Designing Resilient Data Analytics Pipelines

Figure 5 summarizes the principal design challenges affecting resilient manufacturing analytics infrastructures. Key trade-offs exist between scalability and performance, integrity and throughput, latency and recovery, security and usability, and modernization and legacy system integration. The figure highlights the absence of standardized frameworks and emphasizes the need for adaptive, scalable, and fault-tolerant architectures capable of balancing resilience requirements with operational efficiency.

Future Directions and Conclusion

1. AI-Driven Fault Detection and Recovery

Future research should advance resilient data analytics pipelines from rule-based fault handling toward intelligent, AI-driven fault detection and recovery. Smart manufacturing pipelines operate under dynamic failure conditions involving sensor malfunction, packet loss, corrupted streams, processing-node failure, and inconsistent workload patterns. Conventional recovery mechanisms such as replication, checkpointing, and rollback remain essential, but they are often reactive and configuration-dependent. Their effectiveness may reduce when failure frequency, data velocity, and workload intensity change during real-time production.

AI-driven recovery systems should therefore combine anomaly detection, predictive fault modelling, and adaptive orchestration. Machine learning models can learn normal pipeline behaviour from throughput, latency, packet loss, recovery time, duplicate-processing rate, checkpoint cost, and resource-utilization patterns. When deviations occur, the system can identify whether the fault originates from the sensor layer, ingestion broker, processing engine, storage layer, or network path. This would allow faster and more precise recovery than generic restart or rollback procedures.

Future systems should also use online learning and reinforcement learning to adjust recovery actions during operation. For example, checkpoint intervals, replication levels, backup frequency, and resource allocation can be modified according to current failure probability and service-level requirements. This direction directly addresses the manuscript's identified challenge of balancing resilience with performance. The goal should not be maximum fault tolerance at all times, but context-aware resilience that preserves data integrity, minimizes latency, and avoids unnecessary computational overhead.

2. Integration with Digital Twins

Digital twins represent a major future direction for resilient data analytics pipelines in smart manufacturing. A digital twin can replicate the behaviour of the physical production system and its supporting data

infrastructure, allowing failures to be simulated before they affect real operations. This capability is important because manufacturing pipelines are not isolated software systems; they are connected to machines, sensors, controllers, operators, quality systems, and supply-chain platforms.

Future research should develop digital-twin environments that model both the manufacturing process and the data pipeline that supports it. Such twins should simulate sensor degradation, communication delays, broker failure, stream-processing overload, storage inconsistency, and cyber-physical disturbances. This would allow researchers and practitioners to test whether a pipeline can maintain acceptable throughput, recovery time, data completeness, and decision accuracy under different fault scenarios.

Digital twins can also support resilience optimization. Before deploying a new pipeline configuration, the digital twin can evaluate the effect of checkpoint frequency, edge preprocessing, fog-node placement, cloud scaling, and data replication on real-time performance. This would reduce the risk of deploying architectures that appear robust theoretically but fail under shop-floor constraints. In this sense, digital twins should become validation environments for fault-tolerant pipeline design, not only visualization tools for manufacturing systems.

3. Scalable Distributed Architectures

Future resilient pipelines must be designed as scalable distributed architectures that operate across the edge, fog, cloud, and enterprise layers. The manuscript shows that smart manufacturing data is high-volume, high-velocity, heterogeneous, and geographically distributed. A centralized pipeline cannot always satisfy the latency, availability, and fault-tolerance needs of such environments. Edge and fog computing should therefore play a stronger role in future pipeline designs.

A key direction is the development of layered architectures in which time-critical analytics occur close to the machine, while large-scale model training, historical analysis, and cross-site optimization occur in the cloud. This structure can reduce bandwidth pressure, improve response time, and preserve operational visibility when cloud connectivity is degraded. However, distributed architectures also introduce new challenges, including synchronization, consistency, version control, and secure data movement across layers.

Future studies should therefore focus on architectures that are cloud-agnostic, vendor-neutral, modular, and interoperable with existing industrial systems. Manufacturing firms often operate legacy machines, proprietary protocols, and heterogeneous data formats. A resilient pipeline must therefore support loose coupling, standardized interfaces, and gradual integration rather than assuming a fully modernized factory environment. Scalable architecture should be measured not only by throughput, but also by recoverability, portability, maintainability, and ability to support real-time decision-making under partial failure.

4. Autonomous Data Management Systems

The next generation of resilient analytics pipelines should move toward autonomous data management. In the current literature, many pipelines still depend on manually configured fault-tolerance settings, static thresholds, and human intervention during abnormal conditions. This limits their suitability for autonomous manufacturing systems, where decisions must be made continuously and with minimal delay.

Autonomous data management systems should be capable of self-monitoring, self-diagnosis, self-healing, and self-optimization. They should detect missing data, corrupted records, outliers, duplicate messages, delayed streams, and inconsistent states without waiting for manual inspection. Once detected, the system should automatically apply the most appropriate corrective action, such as data imputation, stream

rerouting, checkpoint restoration, node replacement, model recalibration, or escalation to human operators when risk exceeds acceptable limits.

This direction is especially important for quality control, predictive maintenance, and autonomous production planning. These applications depend on accurate and continuous data flow. A failure in the analytics pipeline can produce false alarms, missed defects, delayed maintenance decisions, or unsafe autonomous actions. Future research should therefore integrate data governance, lineage tracking, audit logging, and explainable recovery decisions into autonomous pipeline management. Resilience must be transparent, traceable, and defensible, especially in manufacturing environments where data-driven decisions affect product quality, equipment safety, and operational continuity.

5. Standardization, Benchmarking, and Evaluation

A major future requirement is the development of standardized frameworks for designing, testing, and benchmarking resilient data analytics pipelines in smart manufacturing. The reviewed literature shows strong progress in stream processing, edge-cloud architectures, checkpointing, replication, fault recovery, and real-time monitoring. However, the field remains fragmented. Different studies use different architectures, fault assumptions, metrics, workloads, and validation environments. This makes direct comparison difficult.

Future research should establish common evaluation protocols for resilient manufacturing pipelines. These protocols should include standard fault scenarios such as sensor dropout, corrupted data, broker failure, worker-node crash, master-node failure, cloud disconnection, network latency spikes, and storage inconsistency. They should also define core metrics, including recovery time, downtime, data loss, duplicate-processing rate, throughput under failure, latency under recovery, accuracy degradation, and resilience overhead.

Benchmarking should also reflect manufacturing reality. Pipeline resilience should be evaluated using industrial workloads, continuous sensor streams, heterogeneous data sources, and safety-critical analytics tasks. Synthetic tests are useful, but they should be complemented by realistic case studies in predictive maintenance, process monitoring, quality control, logistics, and autonomous manufacturing. A standardized evaluation framework would help researchers compare solutions more rigorously and help manufacturers select architectures that fit their operational risk profile.

3. Conclusion

This review has established that resilient data analytics pipelines are foundational to fault-tolerant smart manufacturing systems. Modern manufacturing depends on continuous data streams from sensors, IIoT devices, edge platforms, cloud systems, and enterprise applications. These data streams support predictive maintenance, real-time monitoring, quality assurance, supply-chain coordination, and autonomous decision-making. However, their value depends on the reliability of the pipeline that ingests, processes, stores, secures, and delivers the data.

The review shows that resilience is achieved through the coordinated use of redundancy, monitoring, protection, recovery, checkpointing, replication, distributed storage, edge-cloud collaboration, security controls, and governance mechanisms. Technologies such as Kafka, Flink, Spark, Storm, HDFS, Cassandra, MongoDB, fog computing, cloud-agnostic platforms, and microservice-based architectures provide important building blocks. Yet no single technology fully resolves the combined requirements of scalability, data integrity, low latency, fault recovery, interoperability, and security. The major unresolved challenge is the trade-off between resilience and performance. Stronger recovery guarantees often increase computational overhead, while lightweight configurations may expose the pipeline to data loss, duplicate

processing, or delayed recovery. This tension becomes more critical in smart manufacturing, where delayed or inaccurate analytics can affect equipment reliability, product quality, process stability, and worker safety. The future of the field should therefore focus on adaptive, intelligent, and standardized pipeline architectures. AI-driven fault detection, digital-twin-based resilience testing, scalable edge-fog-cloud deployment, autonomous data management, and benchmark-driven evaluation will define the next stage of research. These directions will move resilient pipelines from passive fault recovery toward proactive, self-optimizing infrastructure.

In conclusion, resilient data analytics pipelines are not merely technical support systems for smart manufacturing. They are core operational assets that determine whether manufacturing intelligence remains reliable under real-world uncertainty. Their successful development will enable smart factories to maintain continuity, preserve data integrity, support autonomous decisions, and sustain high-performance production even when faults occur.

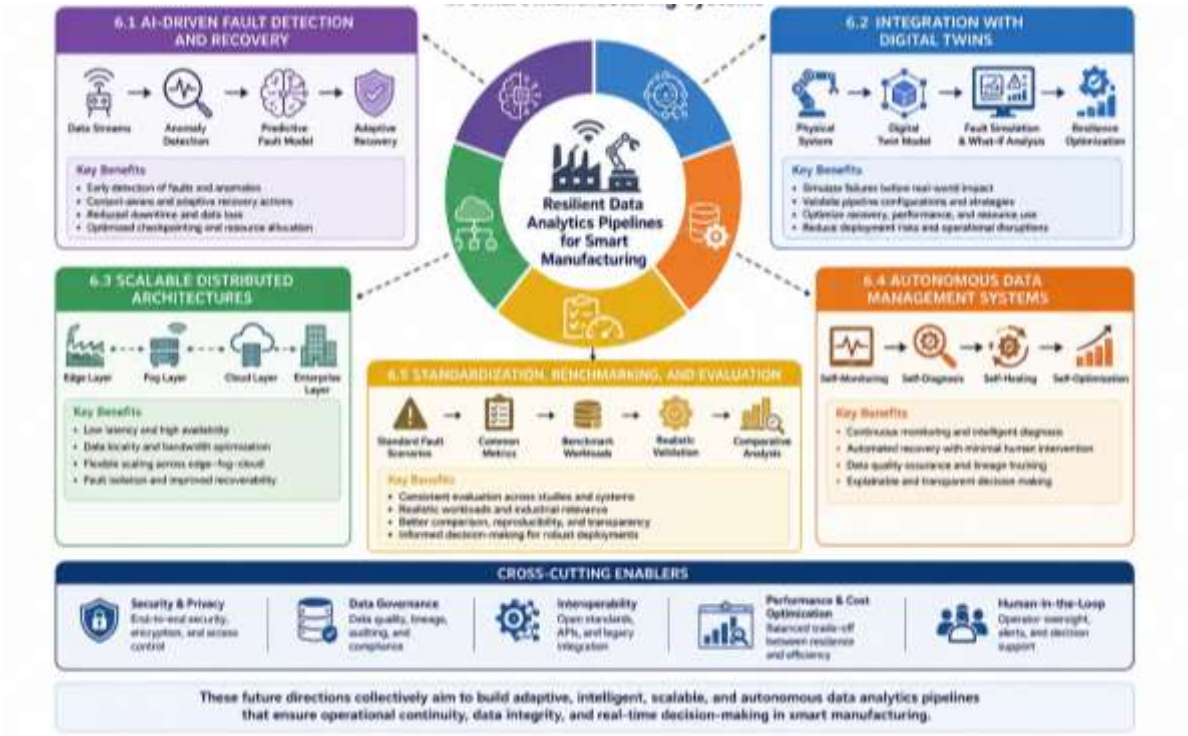

Figure 6: Future Roadmap for Resilient Data Analytics Pipelines in Smart Manufacturing Systems

4. References

AlSuwaidan, L. (2020). The role of data management in the Industrial Internet of Things. Concurrency and Computation Practice and Experience, 33(23). https://doi.org/10.1002/cpe.6031

Andronie, M., Lăzăroiu, G., Iatagan, M., Uţă, C., Ștefănescu, R., & Cocoşatu, M. (2021). Artificial Intelligence-Based Decision-Making Algorithms, Internet of Things Sensing Networks, and Deep Learning-Assisted Smart Process Management in Cyber-Physical Production Systems. Electronics, 10(20), 2497. https://doi.org/10.3390/electronics10202497

Angelopoulos, A., Michailidis, E., Νομικός, N., Trakadas, P., Hatziefremidis, A., Voliotis, S., ... & Zahariadis, T. (2019). Tackling Faults in the Industry 4.0 Era—A Survey of Machine-Learning Solutions and Key Aspects. Sensors, 20(1), 109. https://doi.org/10.3390/s20010109

Asaithambi, S., Venkatraman, R., & Venkatraman, S. (2020). MOBDA: Microservice-Oriented Big Data Architecture for Smart City Transport Systems. Big Data and Cognitive Computing, 4(3), 17. https://doi.org/10.3390/bdcc4030017

Assunção, M., Veith, A., & Buyya, R. (2018). Distributed data stream processing and edge computing: A survey on resource elasticity and future directions. Journal of Network and Computer Applications, 103, 1-17. https://doi.org/10.1016/j.jnca.2017.12.001

Caiazzo, B., Murino, T., Petrillo, A., Piccirillo, G., & Santini, S. (2022). An IoT-based and cloud-assisted AI-driven monitoring platform for smart manufacturing: design architecture and experimental validation. Journal of Manufacturing Technology Management, 34(4), 507-534. https://doi.org/10.1108/jmtm-02-2022-0092

Çakır, A., Akın, Ö., Deniz, H., & Yılmaz, A. (2022). Enabling real time big data solutions for manufacturing at scale. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00672-6

Cañizo, M., Conde, Á., Charramendieta, S., Miñón, R., Cid-Fuentes, R., & Onieva, E. (2019). Implementation of a Large-Scale Platform for Cyber-Physical System Real-Time Monitoring. Ieee Access, 7, 52455-52466. https://doi.org/10.1109/access.2019.2911979

Cardellini, V., Presti, F., Nardelli, M., & Russo, G. (2022). Runtime Adaptation of Data Stream Processing Systems: The State of the Art. Acm Computing Surveys, 54(11s), 1-36. https://doi.org/10.1145/3514496

Chae, J., Lee, S., Jang, J., Hong, S., & Park, K. (2023). A Survey and Perspective on Industrial Cyber-Physical Systems (ICPS): From ICPS to AI-Augmented ICPS. Ieee Transactions on Industrial Cyber-Physical Systems, 1, 257-272. https://doi.org/10.1109/ticps.2023.3323600

Cheng, Z., Huang, Q., & Lee, P. (2019). On the performance and convergence of distributed stream processing via approximate fault tolerance. The VLDB Journal, 28(5), 821-846. https://doi.org/10.1007/s00778-019-00565-w

Chidiebube, I. N., Nwamekwe, C. O., Chukwuemeka, G. H., and Wilfred, M. (2025). Optimization Of Overall Equipment Effectiveness Factors in a Food Manufacturing Small and Medium Enterprise. Journal of Research in Engineering and Applied Sciences, 10(1), 836-845.

Chidiebube, I.N., Onyeka, N.C., Sunday, A.P., et al. (2025a) 'A comparative analysis of machine learning models for inventory demand forecasting in a food manufacturing SME', Indonesian Journal of Innovation Science and Knowledge, 2(3), pp. 35-48.

Chidiebube, I.N., Uzochukwu, M.G., Nwamekwe, C.O., et al. (2025b) 'Evaluating machine learning models for optimizing overall equipment effectiveness in food manufacturing SMEs', Jurnal Inovasi Teknologi Dan Edukasi Teknik, 5(2). https://hal.science/hal-05149408v1/file/igbokwe-nkemakonam-chidiebube-layout-jitet.pdf

Davoudian, A. and Liu, M. (2020). Big Data Systems. Acm Computing Surveys, 53(5), 1-39. https://doi.org/10.1145/3408314

Dongen, G. and Poel, D. (2021). A Performance Analysis of Fault Recovery in Stream Processing Frameworks. Ieee Access, 9, 93745-93763. https://doi.org/10.1109/access.2021.3093208

Dubuc, T., Stahl, F., & Roesch, E. (2021). Mapping the Big Data Landscape: Technologies, Platforms and Paradigms for Real-Time Analytics of Data Streams. Ieee Access, 9, 15351-15374. https://doi.org/10.1109/access.2020.3046132

Emeka, U. C., Chikwendu, O. C., & Onyeka, N. C. (2025). Human-Centric Design Integration in Industry 5.0: A Framework for Resilient Smart Manufacturing. INTERNATIONAL JOURNAL, 3(4).

Emeka, U. C., Okpala, C., and Nwamekwe, C. O. (2025a). Circular Economy Principles'implementation in Electronics Manufacturing: Waste Reduction Strategies in Chemical Management. International journal of industrial and production engineering, 3(2), 29-42.

Ezeanyim, O. C., Ewuzie, N. V., Aguh, P. S., Nwabueze, C. V., and Nwamekwe, C. O. (2025). Effective Maintenance of Industrial 5-Stage Compressor: A Machine Learning Approach. Gazi University Journal of Science Part A: Engineering and Innovation, 12(1), 96-118. https://dergipark.org.tr/en/pub/gujsa/issue/90827/1646993

Ezeanyim, O.C., Nwabunwanne, E.C., Igbokwe, N.C. and Nwamekwe, C.O. (2025a) 'Patient flow and service efficiency in public hospitals: data-driven approaches, strategies, challenges, and future directions', Journal Health of Indonesian, 3(02), pp. 104–124. https://doi.org/10.58471/health.v3i02.228

Geldenhuys, M., Pfister, B., Scheinert, D., Thamsen, L., & Kao, O. (2022). Khaos: Dynamically Optimizing Checkpointing for Dependable Distributed Stream Processing., 30, 553-561. https://doi.org/10.15439/2022f225

Huang, Z., Shen, Y., Li, J., Fey, M., & Brecher, C. (2021). A Survey on AI-Driven Digital Twins in Industry 4.0: Smart Manufacturing and Advanced Robotics. Sensors, 21(19), 6340. https://doi.org/10.3390/s21196340

Igbokwe, N. C., and Nwamekwe, C. O. (2025). Application of Machine Learning in Predicting Emergency Obstetric Cases in Sub-Saharan Africa: An Early Appraisal. International Journal of Industrial Engineering, Technology and Operations Management, 3(1), 13-22.

Igbokwe, N. C., Christiana, C., Nweke, C. O. N., and Onyeka, C. (2025). Data-Driven Solutions for Shuttle Bus Travel Time Prediction: Machine Learning Model Evaluation at Nnamdi Azikiwe University. African Journal of Computing, Data Science and Informatics (AJCDSI), 1(1), 31-55.

Igbokwe, N. C., Nwamekwe, C. O., Ono, C. G., Nwabunwanne, E. C., & Aguh, P. S. (2024). The role of digital twins in optimizing renewable energy utilization and energy efficiency in manufacturing. Siber International Journal of Digital Business, 1(4), 93-111.

Igbokwe, N. C., Okeagu, F. N., Onyeka, N. C., Onwuliri, J. B., and Godfrey, O. C. (2024a). Machine Learning-Driven Maintenance Cost Optimization: Insights from a Local Industrial Compressor Case Study. Jurnal Inovasi Teknologi dan Edukasi Teknik, 4(11), 2.

Igbokwe, N.C., Emmanuel, U.N. and Nwamekwe, C.O. (2025a) 'Advances in post-harvest fish processing: an appraisal of traditional and modern smoking techniques for improved quality and efficiency', Jurnal Integrasi Dan Harmoni Inovatif Ilmu-Ilmu Sosial', 5 (9), pp. 1-13. https://philarchive.org/rec/IGBAIP

Igbokwe, N.C., Nwamekwe, C.O. and Aguh, P.S. (2025b) 'Predictive modeling of manufacturing defects using machine learning: A comparative performance study in a manufacturing SME', African Journal of Advances in Engineering and Technology (AJAET), 1(02), pp. 93-115.

Isah, H., Abughofa, T., Mahfuz, S., Ajerla, D., Zulkernine, F., & Khan, S. (2019). A Survey of Distributed Data Stream Processing Frameworks. Ieee Access, 7, 154300-154316. https://doi.org/10.1109/access.2019.2946884

Javed, A., Robert, J., Heljanko, K., & Främling, K. (2020). IoTEF: A Federated Edge-Cloud Architecture for Fault-Tolerant IoT Applications. Journal of Grid Computing, 18(1), 57-80. https://doi.org/10.1007/s10723-019-09498-8

Jayasekara, S., Harwood, A., & Karunasekera, S. (2020). A utilization model for optimization of checkpoint intervals in distributed stream processing systems. Future Generation Computer Systems, 110, 68-79. https://doi.org/10.1016/j.future.2020.04.019

Kang, S., Jin, R., Deng, X., & Kenett, R. (2021). Challenges of modelling and analysis in cybermanufacturing: a review from a machine learning and computation perspective. Journal of Intelligent Manufacturing, 34(2), 415-428. https://doi.org/10.1007/s10845-021-01817-9

Khalid, M. and Yousaf, M. (2021). A Comparative Analysis of Big Data Frameworks: An Adoption Perspective. Applied Sciences, 11(22), 11033. https://doi.org/10.3390/app112211033

Khan, H., Jabeen, F., Khan, A., Waqar, M., & Kim, A. (2025). IoT-Enabled Fog-Based Secure Aggregation in Smart Grids Supporting Data Analytics. Sensors, 25(19), 6240. https://doi.org/10.3390/s25196240

Khattach, O., Moussaoui, O., & Hassine, M. (2025). End-to-End Architecture for Real-Time IoT Analytics and Predictive Maintenance Using Stream Processing and ML Pipelines. Sensors, 25(9), 2945. https://doi.org/10.3390/s25092945

Marosi, A., Emődi, M., Farkas, A., Lovas, R., Beregi, R., Pedone, G., ... & Gáspár, P. (2022). Toward Reference Architectures: A Cloud-Agnostic Data Analytics Platform Empowering Autonomous Systems. Ieee Access, 10, 60658-60673. https://doi.org/10.1109/access.2022.3180365

Mehmood, E. and Anees, T. (2020). Challenges and Solutions for Processing Real-Time Big Data Stream: A Systematic Literature Review. Ieee Access, 8, 119123-119143. https://doi.org/10.1109/access.2020.3005268

Nasiri, H., Nasehi, S., & Goudarzi, M. (2019). Evaluation of distributed stream processing frameworks for IoT applications in Smart Cities. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0215-2

Nwamekwe, C. O., and Igbokwe, N. C. (2024). Supply Chain Risk Management: Leveraging AI for Risk Identification, Mitigation, and Resilience Planning. International Journal of Industrial Engineering, Technology and Operations Management, 2(2), 41–51. https://doi.org/10.62157/ijietom.v2i2.38

Nwamekwe, C. O., and Nwabunwanne, E. C. (2025). Immersive Digital Twin Integration in the Metaverse for Supply Chain Resilience and Disruption Management. Journal of Engineering Research and Applied Science, 14(1), 95-105.

Nwamekwe, C. O., Chidiebube, I. N., Godfrey, O. C., Celestine, N. E., and Sunday, A. P. (2025). Resilience and Risk Management in Social Robot Systems: An Industrial Engineering Perspective. Culture education and technology research (Cetera), 2(2), 1-12.

Nwamekwe, C. O., Chidiebube, I. N., Godfrey, O. C., Celestine, N. E., and Aguh, P. S. (2025a). Human-Robot Collaboration in Industrial Engineering: Enhancing Productivity and Safety. Journal of Industrial Engineering and Management Research, 6(5), 1-20.

Nwamekwe, C. O., Chinwuko, C. E. and Mgbemena, C. E. (2020). Development and Implementation of a Computerised Production Planning and Control System. UNIZIK Journal of Engineering and Applied Sciences, 17(1), 168-187. https://journals.unizik.edu.ng/ujeas/article/view/1771

Nwamekwe, C. O., Edokpia, R. O., & Eboigbe, C. I. (2026). Integration of Machine Learning into Lean Six Sigma: A Systematic Review for Enhancing Predictive Analytics in the Pharmaceutical Industry. Siber Journal of Advanced Multidisciplinary, 3(4), 133-151.

Nwamekwe, C. O., Edokpia, R. O., and Igbinosa, E. C. (2025b). Exploring the Role of Artificial Intelligence in Enhancing Lean Manufacturing and Six Sigma for Smart Factories. International Journal of Industrial Engineering, Technology and Operations Management, 3(1), 1-12.

Nwamekwe, C. O., Ewuzie, N.V., Igbokwe, N. C., Nwabunwanne, E. C., and Ono, C. G. (2025c). Digital Twin-Driven Lean Manufacturing: Optimizing Value Stream Flow. Letters in Information Technology Education (LITE), 8 (1), pp.1-13. https://hal.science/hal-05127340/

Nwamekwe, C. O., Nwabunwanne, E. C., Okeagu, F. N., and Ono, C. G. (2025d). Lean Manufacturing Principles in the Design and Production of Social Robots. International Journal of Industrial Engineering, Technology and Operations Management, 3(1), 23-34.

Nwamekwe, C. O., Okpala, C. C., and Nwabunwanne, E. C. (2025e). Design Principles and Challenges in Achieving Zero-Energy Manufacturing Facilities. Journal of Engineering Research and Applied Science, 14(1), 1-21.

Nwamekwe, C. O., Uchenna, P. C., Onyedika, S. C. (2026a). Leveraging Emerging Technologies to Enhance Business Processes in Blue Economy Sectors: A Case Study of Anambra State's Industrial Landscape.

International Journal of Technology, Health and Sustainability, 2(2), pp. 559-572. https://ijths.com/wp-content/uploads/IJTHS-0202024.pdf

Okeagu, F., Nwamekwe, C., and Nnamani, B. (2024). Challenges and Solutions of Industrial Development in Anambra State, Nigeria. Iconic Research and Engineering Journals, 7(11), 467-472. https://www.irejournals.com/formatedpaper/1705825.pdf

Okpala C. C., Chukwudi Emeka Udu, and Charles Onyeka Nwamekwe. (2025). Sustainable HVAC Project Management: Strategies for Green Building Certification. International Journal of Industrial and Production Engineering, 3(2), 14-28. https://journals.unizik.edu.ng/ijipe/article/view/5595.

Okpala, C. C., Ezeanyim, O. C., and Nwamekwe, C. O. (2024). The Implementation of Kaizen Principles in Manufacturing Processes: A Pathway to Continuous Improvement. International Journal of Engineering Inventions, 13(7), 116-124. https://www.ijeijournal.com/papers/Vol13-Issue7/1307116124.pdf

Okpala, C. C., Udu, C. E., and Nwamekwe, C. O. (2025a). Artificial Intelligence-Driven Total Productive Maintenance: The Future of Maintenance in Smart Factories. International Journal of Engineering Research and Development (IJERD), (21)1, 68-74. https://www.ijerd.com/paper/vol21-issue1/21016874.pdf

Okpala, C., Onyeka, C. and Igbokwe, N.C. (2024a) 'The implementation of Internet of Things in the manufacturing industry: An appraisal', International Journal of Engineering Research and Development, 20(7), pp. 510-516.

Onyeka, N. C., and Emeka, N. (2025). Circular Economy and Zero-Energy Factories: A Synergistic Approach to Sustainable Manufacturing. Journal of Research in Engineering and Applied Sciences, 10(1), 829-835.

Onyeka, N. C., Vitalis, E. N., Chidiebube, I. N., U-Dominic, C. M., and Chibuzo, N. (2024). Adoption of Smart Factories in Nigeria: Problems, Obstacles, Remedies and Opportunities. International journal of industrial and production engineering, 2(2), 68-81. https://journals.unizik.edu.ng/ijipe/article/view/4167

Oza, J., Patil, A., Maniyath, C., More, R., Kambli, G., & Maity, A. (2024). Harnessing Insights from Streams: Unlocking Real-Time Data Flow with Docker and Cassandra in the Apache Ecosystem. https://doi.org/10.36227/techrxiv.170475337.78884732/v1

Panetto, H., Iung, B., Ivanov, D., Weichhart, G., & Wang, X. (2019). Challenges for the cyber-physical manufacturing enterprises of the future. Annual Reviews in Control, 47, 200-213. https://doi.org/10.1016/j.arcontrol.2019.02.002

Parri, J., Patara, F., Sampietro, S., & Vicario, E. (2020). A framework for Model-Driven Engineering of resilient software-controlled systems. Computing, 103(4), 589-612. https://doi.org/10.1007/s00607-020-00841-6

Peres, R., Rocha, A., Leitão, P., & Barata, J. (2018). IDARTS – Towards intelligent data analysis and real-time supervision for industry 4.0. Computers in Industry, 101, 138-146. https://doi.org/10.1016/j.compind.2018.07.004

Power, A. and Kotonya, G. (2018). A Microservices Architecture for Reactive and Proactive Fault Tolerance in IoT Systems., 588-599. https://doi.org/10.1109/wowmom.2018.8449789

Rehman, M., Yaqoob, I., Salah, K., Imran, M., Jayaraman, P., & Perera, C. (2019). The role of big data analytics in industrial Internet of Things. Future Generation Computer Systems, 99, 247-259. https://doi.org/10.1016/j.future.2019.04.020

Sahal, R., Alsamhi, S., Brown, K., O'Shea, D., McCarthy, C., & Guizani, M. (2021). Blockchain-Empowered Digital Twins Collaboration: Smart Transportation Use Case. Machines, 9(9), 193. https://doi.org/10.3390/machines9090193

Srirama, S. (2024). Distributed edge analytics in edge-fog-cloud continuum. Internet Technology Letters, 8(3). https://doi.org/10.1002/itl2.562

Suvarna, M., Büth, L., Hejny, J., Mennenga, M., Li, J., Ng, Y., … & Wang, X. (2020). Smart Manufacturing for Smart Cities—Overview, Insights, and Future Directions. Advanced Intelligent Systems, 2(10). https://doi.org/10.1002/aisy.202000043

Syafrudin, M., Alfian, G., Fitriyani, N., & Rhee, J. (2018). Performance Analysis of IoT-Based Sensor, Big Data Processing, and Machine Learning Model for Real-Time Monitoring System in Automotive Manufacturing. Sensors, 18(9), 2946. https://doi.org/10.3390/s18092946

Vergilio, T., Kor, A., & Mullier, D. (2022). A Unified Vendor-Agnostic Solution for Big Data Stream Processing in a Multi-Cloud Environment. https://doi.org/10.21203/rs.3.rs-1253161/v1

Vitalis, E. N., Nwamekwe, C. O., Chidiebube, I. N., Chibuzo, N., Nwabunwanne, E. C., and Ono, C. G. (2024). Application Of Machine-Learning-Based Hybrid Algorithm for Production Forecast in Textile Company. Jurnal Inovasi Teknologi dan Edukasi Teknik, 4(12), 1-9.

Vital-Soto, A. and Olivares-Aguila, J. (2023). Manufacturing Systems for Unexpected Events: An Exploratory Review for Operational and Disruption Risks. Ieee Access, 11, 96297-96316. https://doi.org/10.1109/access.2023.3311362

Yang, C., Lan, S., Wang, L., Shen, W., & Huang, G. (2020). Big Data Driven Edge-Cloud Collaboration Architecture for Cloud Manufacturing: A Software Defined Perspective. Ieee Access, 8, 45938-45950. https://doi.org/10.1109/access.2020.2977846

Zeadally, S., Sanislav, T., & Moiş, G. (2019). Self-Adaptation Techniques in Cyber-Physical Systems (CPSs). Ieee Access, 7, 171126-171139. https://doi.org/10.1109/access.2019.2956124