Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

Designing Real-Time Data Pipelines for Smart Healthcare Factories

Suchitra Venkatesan

Independent Researcher
California, USA
suchitrav.93@gmail.com

Abstract:
Smart healthcare factories—facilities that combine pharmaceutical or medical-device manufacturing with pervasive sensing, automation, and analytics—generate continuous, high-velocity data streams from process sensors, quality-inspection systems, environmental monitors, and enterprise resource-planning platforms. Coordinating these heterogeneous streams into a coherent, low-latency information fabric is a non-trivial systems-engineering challenge that directly affects product quality, regulatory compliance, and patient safety. This paper presents an architectural framework for real-time data pipelines in smart healthcare factories. It examines the domain-specific requirements imposed by Good Manufacturing Practice (GMP) regulations, discusses candidate streaming technologies and their trade-offs, proposes a layered reference architecture, and identifies open research problems. The paper draws on publicly available standards (IEC 62443, ISA-95, ICH Q10, FDA 21 CFR Part 11) and peer-reviewed literature to ground each design decision.

Index Terms: data pipeline, stream processing, smart factory, healthcare manufacturing, Industrial IoT, real-time analytics, GMP compliance, edge computing, data governance.

I. INTRODUCTION

The convergence of Industrial Internet of Things (IIoT), edge computing, and advanced analytics is reshaping pharmaceutical and medical-device manufacturing. Regulatory agencies and standards bodies have articulated a vision of Pharma 4.0—an extension of the broader Industry 4.0 paradigm—in which continuous manufacturing processes are governed by real-time data feedback loops [1]. Achieving this vision requires robust, auditable, and low-latency data pipelines that can ingest signals from heterogeneous field devices, transform and contextualize them, and deliver actionable insights to control systems and human operators within process-relevant time horizons.

Unlike discrete manufacturing, healthcare factories operate under strict regulatory oversight. Every data record is potentially subject to regulatory inspection; any loss or corruption of process data can trigger batch rejection, warning letters, or consent decrees. This regulatory backdrop creates requirements that are absent from mainstream stream-processing literature: immutable audit trails, validated software lifecycle management, electronic signatures, and data-integrity assurance throughout the pipeline.

This paper makes the following contributions:

- A characterization of the domain-specific requirements for real-time data pipelines in healthcare manufacturing.
- A layered reference architecture that maps IIoT data flows to ISA-95 enterprise levels.
- A comparative analysis of streaming middleware technologies evaluated against healthcare factory constraints.
- A discussion of edge–fog–cloud workload partitioning strategies.

- An identification of open research challenges at the intersection of real-time streaming and regulated manufacturing.

II. BACKGROUND AND RELATED WORK

A. Smart Manufacturing and Pharma 4.0

The ISA-95 standard (IEC 62264) defines a five-level functional hierarchy for enterprise-to-control integration [2]. Level 0 encompasses physical processes; Levels 1–2 cover control and supervisory systems; Level 3 addresses manufacturing operations management (MOM/MES); and Level 4 encompasses enterprise resource planning (ERP). Data pipelines in smart healthcare factories must span all five levels while preserving semantic integrity across the boundaries between operational technology (OT) and information technology (IT).

The Pharma 4.0 initiative, articulated by the International Society for Pharmaceutical Engineering (ISPE), extends Industry 4.0 principles to regulated environments by adding a sixth dimension: data integrity and quality culture [3]. Key enablers include real-time release testing (RTRT), continuous process verification (CPV), and digital batch records.

B. Stream Processing Foundations

Stream processing systems execute computations over unbounded data sequences with strict latency constraints. The foundational models—DataFlow [4] and the Kappa architecture [5]—treat all computation as transformations over event streams, thereby unifying batch and real-time processing. Modern frameworks such as Apache Flink [6] provide exactly-once semantics, event-time processing, and stateful operators, which are prerequisites for reliable statistical process-control (SPC) computations in manufacturing.

The CAP theorem [7] constrains distributed stream-processing systems to favor either consistency or availability during network partitions. Healthcare factory pipelines must carefully evaluate this trade-off: process-control applications demand high availability and low latency, while audit-log systems demand strong consistency and durability.

C. IIoT Communication Protocols

Three protocol families dominate factory floor communication: (i) fieldbus and real-time Ethernet protocols (PROFINET, EtherNet/IP) at Levels 0–1; (ii) OPC Unified Architecture (OPC-UA), which provides a platform-independent, secure information model at Levels 1–3 [8]; and (iii) MQTT and AMQP, which provide lightweight publish-subscribe semantics suitable for cloud ingestion at Levels 3–4. Integrating these heterogeneous protocols without data loss or latency amplification is a central challenge for pipeline designers.

D. Regulatory Data-Integrity Frameworks

The FDA's Data Integrity and Compliance with Drug CGMP guidance (2018) and the EU GMP Annex 11 establish the ALCOA+ principles—data must be Attributable, Legible, Contemporaneous, Original, Accurate, and additionally Complete, Consistent, Enduring, and Available [9]. These principles translate directly into pipeline design requirements: timestamping at source, immutable storage, cryptographic provenance, and redundant replication.

III. DOMAIN-SPECIFIC REQUIREMENTS

A. Latency and Throughput

Healthcare factory data pipelines must simultaneously satisfy heterogeneous latency targets. Closed-loop control applications (e.g., continuous granulation feedback) require end-to-end latencies in the sub-millisecond to single-digit millisecond range and are typically handled within the control network itself. Condition monitoring and SPC applications tolerate latencies of tens to hundreds of milliseconds. Batch-

record aggregation and regulatory reporting can tolerate latencies of seconds to minutes. Throughput requirements vary from a few hundred samples per second for a simple sensor network to millions of events per second for high-resolution spectroscopic or vision-inspection systems.

B. Data Integrity and Auditability

Every event entering the pipeline must carry a verifiable timestamp, a source identifier, and a chain of custody that cannot be repudiated. Pipeline stages that perform transformations must record their inputs, outputs, and processing parameters in an immutable audit log. This log must be tamper-evident—accomplished through cryptographic hashing (SHA-256 or stronger) or blockchain anchoring—and must be retained for a period consistent with the applicable regulatory dossier, typically ten or more years for pharmaceutical products [10].

C. Validated State and Change Control

Under EU GMP Annex 11 and GAMP 5, any software system that affects product quality must be validated. For a data pipeline, this implies that every component—broker, stream-processing topology, data-store schema—must have a defined validation status and a formal change-control process. Continuous delivery practices common in cloud engineering must be adapted to ensure that no pipeline change reaches the production environment without documented impact assessment and regression testing.

D. Cybersecurity

The IEC 62443 series defines security levels (SL 0–4) for industrial automation and control systems, and the zone-conduit model for segmenting networks by security criticality [11]. A healthcare factory pipeline that bridges OT and IT networks necessarily traverses multiple security zones. All inter-zone communication must pass through validated data diodes or firewalls, and all data in transit must be encrypted (TLS 1.3 minimum). Authentication must rely on X.509 certificates with hardware-backed key storage wherever process latency permits.

E. Interoperability and Semantic Consistency

Sensor readings from different vendors often share identical tag names but carry different units, calibration offsets, or engineering ranges. Without a shared semantic layer, downstream analytics produce silently incorrect results. The industry is converging on OPC-UA companion specifications (e.g., the OPC-UA for Machinery, OPC-UA for Batch) and ISO/IEC 21823 for IoT interoperability as mechanisms for encoding semantics in the data stream itself, rather than relying on out-of-band documentation.

IV. REFERENCE ARCHITECTURE

Fig. 1 (conceptually described below) depicts the proposed five-layer reference architecture. Each layer is loosely coupled to its neighbors via well-defined message schemas and back-pressure protocols, enabling independent scaling and replacement of individual components.

A. Layer 1: Edge Acquisition

The acquisition layer resides at ISA-95 Levels 0–1. PLCs, distributed control systems (DCS), and smart sensors produce time-stamped process values at rates governed by the underlying control cycle (typically 1–100 ms). OPC-UA servers expose these values as a browsable address space; OPC-UA PubSub or lightweight MQTT clients publish changed values to an on-premises broker. Signal conditioning—noise filtering, range checking, unit normalization—is performed at this layer to minimize upstream bandwidth. An industrial-grade buffering mechanism (e.g., OSIsoft PI, InfluxDB Edge) provides a short-horizon store-and-forward capability to tolerate transient network interruptions without data loss.

B. Layer 2: Transport and Brokering

The transport layer decouples producers from consumers through a durable, partitioned log. Apache Kafka and Apache Pulsar are the primary candidates at this tier; both provide ordered, replayable, and durably stored event streams with sub-ten-millisecond end-to-end latency under typical factory workloads. Kafka's compaction feature enables materialized views of the latest sensor state, while Pulsar's tiered storage simplifies long-term regulatory archiving. Partition keys should be chosen to collocate events from the same equipment unit, preserving ordering for SPC calculations without sacrificing parallelism.

Topic naming should follow a hierarchical convention aligned with ISA-88 physical-model terminology (enterprise/site/area/unit/phase/parameter) to facilitate semantic routing and access-control policies. Schema enforcement via Apache Avro or Protobuf, governed by a schema registry, prevents schema drift from silently corrupting downstream consumers.

C. Layer 3: Stream Processing

The processing layer implements continuous transformations, enrichments, and analytics over the event streams. Apache Flink is the recommended framework for this tier due to its native support for event-time semantics, exactly-once checkpointing, and stateful window operators—all essential for accurate SPC charting (Shewhart, CUSUM, EWMA) over out-of-order sensor data. Processing topologies are modelled as directed acyclic graphs (DAGs) whose nodes correspond to operators (filter, join, aggregate, enrich) and whose edges carry bounded-latency SLAs.

Key processing workloads at this layer include: (i) univariate and multivariate SPC; (ii) golden-batch comparison for continuous process verification; (iii) real-time anomaly detection using sliding-window statistics or pre-trained ML models served via ONNX; (iv) digital batch record assembly; and (v) alarm rationalization and de-bouncing to reduce operator alarm fatigue.

D. Layer 4: Contextualization and Data Mesh

Raw process values acquire meaning only when contextualized against the batch record, the equipment state, and the material genealogy. The contextualization layer enriches events from the stream with reference data fetched from the MES, LIMS, and ERP systems. A unified asset model—expressed as a knowledge graph conforming to the Asset Administration Shell (AAS) specification defined by Industry 4.0 consortia—provides a machine-readable description of each equipment unit, including its calibration history, maintenance status, and associated product specifications. The data mesh pattern [12], adapted to the regulated context, assigns each manufacturing domain (granulation, compression, coating, packaging) ownership of its own data product, including its pipeline topology, schema, and SLA contract.

E. Layer 5: Serving and Archival

The serving layer provides query interfaces for operational dashboards, predictive-maintenance models, and regulatory reporting. A time-series database (e.g., InfluxDB, TimescaleDB, or QuestDB) serves real-time trends and SPC charts with sub-second query latency. A columnar analytical store (e.g., Apache Parquet on object storage, queried via Apache Spark or DuckDB) serves batch analytics and CPV studies. A separate immutable audit-log store—written to via append-only APIs and cryptographically chained—serves as the system of record for regulatory inspection. All stores must implement role-based access control (RBAC) tied to an enterprise identity provider and must log all access events to the audit trail.

V. TECHNOLOGY COMPARISON

Table I presents a comparative evaluation of candidate streaming technologies across dimensions relevant to healthcare factory workloads.

TABLE I: *Streaming Technology Comparison for Healthcare Factory Workloads*

Technology	Latency	Throughput	Ordering	Replay	GMP-Audit Fit
Apache Kafka	< 10 ms	Very High	Per-partition	Yes (log)	High
Apache Pulsar	< 10 ms	Very High	Per-topic	Yes (tiered)	High
MQTT Broker	< 5 ms	Medium	QoS-based	Limited	Medium
OPC-UA PubSub	< 2 ms	Medium	Message seq.	No	Medium
Apache Flink	< 1 ms	Very High	Event-time	Via source	High
Spark Structured Streaming	~100 ms	High	Micro-batch	Via source	Medium

Kafka and Pulsar are the strongest candidates for the transport tier, primarily because of their durable log semantics and replay capability, which directly support the ALCOA+ requirement for original, enduring data. MQTT and OPC-UA PubSub are better suited to the edge acquisition tier where bandwidth and compute are constrained. Flink surpasses Spark Structured Streaming for latency-sensitive SPC workloads, but Spark remains viable for batch-mode CPV studies where its richer ecosystem of statistical libraries is advantageous.

VI. REGULATORY CONTROLS MAPPING

Table II maps the primary regulatory requirements that apply to healthcare factory data pipelines to concrete pipeline design controls.

TABLE II: *Regulatory Requirements Mapped to Pipeline Design Controls*

Regulation / Standard	Requirement	Pipeline Control
FDA 21 CFR Part 11	Electronic records & signatures	Immutable audit log, PKI signing
EU GMP Annex 11	Validated computerized systems	IQ/OQ/PQ for pipeline components
ICH Q10 PQS	Process performance monitoring	Statistical process control stream
ISA-95 / IEC 62264	Enterprise-control integration	Layered data model (L0–L4)
IEC 62443	OT/IT cybersecurity	Zone-conduit model, TLS 1.3
GAMP 5	Risk-based software validation	Change-control pipeline versioning

A critical observation is that regulatory compliance cannot be retrofitted into a pipeline designed without it. Controls such as electronic record immutability, time-synchronization to UTC (traceability requires NTP/PTP synchronization to Stratum 1 sources), and validated change management must be designed into the pipeline from inception. Particular attention is warranted for FDA 21 CFR Part 11, whose requirements for audit trails and electronic signatures apply to any computerized system that creates, modifies, maintains, archives, retrieves, or transmits electronic records that are required under FDA regulations [13].

VII. EDGE–FOG–CLOUD WORKLOAD PARTITIONING

A naive architecture that forwards all raw sensor data to a central cloud incurs prohibitive bandwidth costs, introduces latency incompatible with closed-loop control, and creates a single point of failure that violates GMP availability requirements. A hierarchical edge–fog–cloud continuum, summarized in Table III, allocates workloads to the tier best suited to their latency, compute, and data-retention requirements.

TABLE III: *Workload Allocation Across the Edge–Fog–Cloud Continuum*

Tier	Typical Hardware	Functions	Target Latency
Edge (L0/L1)	PLCs, embedded gateways	Signal filtering, protocol translation, SPC triggers	< 5 ms
Fog (L2/L3)	Industrial PCs, MEC nodes	Batch aggregation, ML inference, local historian	5–100 ms
Cloud (L4)	Data-center / PaaS	Long-term analytics, regulatory archive, ML training	> 1 s

A. Edge Tier

Edge nodes implement the acquisition and first-stage processing described in Layers 1–2 of the reference architecture. Deterministic real-time operating systems (RTOS) or time-sensitive networking (TSN) switches are required where sub-millisecond jitter is specified. Edge ML inference using quantized models (INT8 or FP16) on neural processing units (NPUs) or FPGAs is viable for high-speed defect detection on packaging lines. Local store-and-forward buffers must be sized to accommodate the maximum anticipated network outage duration without data loss; a typical design target is 72 hours of buffering at maximum sensor rate.

B. Fog Tier

Fog nodes—industrial PCs or multi-access edge compute (MEC) servers co-located with the manufacturing area—host the stream-processing engine (Layer 3), the local MES interface, and the on-premises segment of the time-series historian. This tier is the primary execution environment for SPC, alarm management, and real-time batch record assembly. Fog nodes are typically deployed in redundant active-passive pairs, with automatic failover triggered by the orchestration layer (Kubernetes with hardware-aware scheduling is suitable here).

C. Cloud Tier

The cloud tier hosts the data mesh governance plane, long-term regulatory archives, ML model training pipelines, and cross-site analytics. In regulated industries, the choice between public cloud and private on-premises data centers must be evaluated against data residency requirements, supplier qualification obligations (cloud providers become computerized system suppliers under EU GMP Annex 11 and must be qualified accordingly), and business continuity considerations. Hybrid cloud designs, in which regulatory archives remain on-premises while analytics workloads run in public cloud, are increasingly

common and technically well-supported by managed Kafka services (Confluent Cloud, Amazon MSK, Azure Event Hubs with Kafka compatibility).

VIII. DATA GOVERNANCE AND QUALITY ASSURANCE

A. Schema Registry and Versioning

A central schema registry (e.g., Confluent Schema Registry or AWS Glue Schema Registry) is a mandatory component of the governance architecture. All producers must register their schemas before writing to a topic; all consumers must validate incoming messages against the registered schema. Schema evolution must follow compatibility rules (backward, forward, or full) that are agreed upon and documented as part of the change-control process. Schema changes that are not backward-compatible require a new topic and a migration plan, analogous to a software change order under GAMP 5.

B. Data Lineage

Regulators expect manufacturers to demonstrate that a given analytical result or batch-record entry can be traced back to calibrated instruments through a documented chain of transformations. Apache Atlas, OpenLineage, or DataHub can capture lineage metadata at the operator level within the stream-processing topology. These tools emit lineage events when a Flink job starts, updates, or terminates, creating a machine-readable record that can be presented to inspectors as part of a computerized system validation package.

C. Observability

Pipeline observability encompasses metrics (throughput, lag, error rate), traces (end-to-end event latency), and logs (operational events, audit records). A unified observability stack—Prometheus for metrics, OpenTelemetry for traces, Loki or Elasticsearch for logs—provides the visibility required to detect data-quality degradation before it affects batch outcomes. SLA-based alerting should be configured to notify QA personnel when pipeline lag exceeds the threshold at which SPC charts may miss out-of-control conditions.

IX. CYBERSECURITY CONSIDERATIONS

The convergence of OT and IT networks in smart healthcare factories substantially increases the attack surface relative to traditional air-gapped control environments. Several high-profile ransomware incidents against pharmaceutical manufacturers in recent years have demonstrated that a pipeline breach can halt production and compromise patient safety.

The IEC 62443-3-3 standard defines system-level security requirements for industrial automation systems and specifies that data flows across zone boundaries must be restricted by authenticated and encrypted conduits [11]. Applied to the data pipeline, this means:

- All inter-zone message flows must traverse a dedicated firewall or data diode; no direct socket connections between OT and IT networks are permitted.
- Kafka/Pulsar brokers must require mutual TLS authentication; topic-level ACLs must enforce least-privilege access for each producer and consumer.
- Secrets (TLS private keys, API tokens) must be managed by a secrets manager (HashiCorp Vault or a cloud-native equivalent) with automatic rotation and hardware security module (HSM) backing.
- All privileged administrative actions on pipeline components must require multi-factor authentication and must be logged to the tamper-evident audit trail.
- A software bill of materials (SBOM) must be maintained for all pipeline components to enable rapid response to disclosed vulnerabilities.

X. OPEN RESEARCH CHALLENGES

Despite significant advances in stream processing and IIoT platforms, several challenges remain insufficiently addressed for healthcare factory contexts.

A. Federated Learning over Streaming Data

Multi-site pharmaceutical manufacturers wish to train predictive-quality models on data from all sites without centralizing sensitive batch records. Federated learning addresses privacy, but its adaptation to continuously streaming, non-stationary factory data—where concept drift is common and data distributions differ across sites—remains an active research problem.

B. Formal Verification of Pipeline Correctness

Regulatory inspectors increasingly ask manufacturers to demonstrate that a validated pipeline cannot, under any sequence of inputs, produce an incorrect batch record. Formal methods (TLA+, Alloy, model checking) have been applied to distributed system specifications, but their application to the full stack of a heterogeneous IIoT pipeline—spanning RTOS firmware, OPC-UA address spaces, Kafka topic partitioning, and Flink operator state—remains largely unexplored in the literature.

C. Real-Time Digital Twin Synchronization

A digital twin of a pharmaceutical process must reflect the physical state of the process within a bounded synchronization latency. Achieving sub-second twin synchronization while maintaining exactly-once update semantics in the presence of network partitions is an open systems problem, particularly for continuous manufacturing processes where the physical state evolves faster than current consensus protocols can reliably commit.

D. Adaptive Pipeline Autoscaling Under Regulatory Constraints

Cloud-native pipelines can auto scale consumers horizontally in response to load; however, in a validated environment, adding a new consumer instance constitutes a configuration change that may trigger a revalidation obligation. The tension between operational agility (autoscaling) and regulatory stability (validated, fixed configuration) is unresolved and represents a significant barrier to cloud adoption in regulated manufacturing.

XI. CONCLUSION

Real-time data pipelines are foundational infrastructure for smart healthcare factories. The domain imposes requirements—immutable audit trails, validated software lifecycle management, OT/IT security segmentation, and strict latency targets—that significantly constrain the technology choices available relative to a general-purpose streaming architecture. This paper has presented a layered reference architecture grounded in established standards (ISA-95, IEC 62443, ICH Q10, FDA 21 CFR Part 11) and has systematically mapped regulatory requirements to pipeline design controls.

The framework demonstrates that a compliant, high-performance data pipeline is technically achievable using mature open-source components (Apache Kafka, Apache Flink, OPC-UA) combined with disciplined governance practices (schema registry, lineage tracking, validated change control). However, several research challenges—federated learning over streaming data, formal pipeline verification, digital twin synchronization, and regulatory-compatible autoscaling—remain open and warrant focused attention from both the academic and industrial communities.

As regulatory frameworks continue to evolve toward outcome-based, data-driven oversight, the manufacturers who invest in robust, auditable data pipelines today will be best positioned to demonstrate process understanding, accelerate real-time release testing, and ultimately improve patient outcomes.

REFERENCES:

[1] International Society for Pharmaceutical Engineering (ISPE), "ISPE Industry 4.0 Glossary and Technical Guidance," ISPE, Tampa, FL, 2022.
[2] International Electrotechnical Commission, "IEC 62264-1: Enterprise-Control System Integration – Part 1: Models and Terminology," IEC, Geneva, Switzerland, 2013.
[3] S. Schlindwein and R. Rehrl, "Pharma 4.0: An Industry 4.0 Framework for the Pharmaceutical Industry," Drug Discovery Today, vol. 24, no. 4, pp. 1061–1064, 2019.
[4] G. Akidau et al., "The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing," Proc. VLDB Endowment, vol. 8, no. 12, pp. 1792–1803, 2015.
[5] J. Kreps, "Questioning the Lambda Architecture," O'Reilly Radar, Jul. 2014. [Online]. Available: https://www.oreilly.com/radar/questioning-the-lambda-architecture/
[6] P. Carbone et al., "Apache Flink: Stream and Batch Processing in a Single Engine," IEEE Data Eng. Bull., vol. 38, no. 4, pp. 28–38, 2015.
[7] E. Brewer, "CAP Twelve Years Later: How the 'Rules' Have Changed," IEEE Computer, vol. 45, no. 2, pp. 23–29, 2012.
[8] S. Hannelius et al., "OPC Unified Architecture – Service-Oriented Architecture for Industrial Applications," in Proc. 6th IEEE Int. Conf. Industrial Informatics, Daejeon, Korea, 2008, pp. 68–73.
[9] U.S. Food and Drug Administration, "Data Integrity and Compliance With Drug CGMP," Guidance for Industry, FDA, Silver Spring, MD, Dec. 2018.
[10] European Medicines Agency, "EMA Guideline on Electronic Systems and Electronic Data in Clinical Trials (Annex 11)," EMA, Amsterdam, Netherlands, 2022.
[11] International Electrotechnical Commission, "IEC 62443-3-3: Industrial Communication Networks – Security for Industrial Automation and Control Systems – Part 3-3: System Security Requirements and Security Levels," IEC, Geneva, Switzerland, 2013.
[12] Z. Dehghani, "Data Mesh: Delivering Data-Driven Value at Scale," O'Reilly Media, Sebastopol, CA, 2022.
[13] U.S. Food and Drug Administration, "21 CFR Part 11: Electronic Records; Electronic Signatures," Code of Federal Regulations, FDA, Silver Spring, MD, 1997 (as amended).