CONFLUENT

Accelerating Cancer Research Through Real-Time Data Interoperability Across NIH with the Confluent Platform

Executive Summary

The National Cancer Institute (NCI), through its Office of Data Sharing (ODS), has launched an ambitious effort to enhance the interoperability, accessibility, and reuse of cancer-related research data across the NIH enterprise. Current data systems are siloed, asynchronous, and limited in their ability to support real-time discovery, precision medicine, and AI/ML research. Confluent Federal proposes the deployment of the Confluent Platform, built on Apache Kafka®, to create a unified, real-time data streaming layer that allows institutes, programs, and researchers to share data seamlessly across heterogeneous systems while maintaining compliance and security.

1. Introduction: The Challenge of Interoperability in Cancer Research

NIH's cancer research initiatives—spanning clinical, genomic, imaging, and patient-reported data—are distributed across platforms such as:

- The Cancer Genome Atlas (TCGA)
- SEER Program
- The Cancer Imaging Archive (TCIA)
- ClinicalTrials.gov
- NCI Cancer Research Data Commons (CRDC)

These systems often rely on static data dumps, incompatible formats, or custom APIs that hinder timely integration. Without a real-time, scalable, and standards-based architecture, researchers face delays in data availability, duplication of effort, and barriers to collaborative discovery.

2. Solution Overview: Confluent as a Real-Time Interoperability Backbone

Confluent provides a cloud-native, enterprise-ready data streaming platform that enables real-time data movement and processing across systems, data types, and research domains. Key features include:

- Event-Driven Architecture: Producers (e.g., sequencing labs, clinical trial systems) publish data to Kafka topics; consumers (e.g., researchers, analytics platforms) subscribe as needed.
- Built-In Schema Registry: Ensures data consistency and governance across evolving datasets.
- Prebuilt Connectors: 120+ integrations with PostgreSQL, S3, HL7 FHIR, Snowflake, Databricks, and more.
- Stream Processing: Enables transformation, filtering, and enrichment of data in motion.

3. Use Cases and Examples

3.1. Genomic Data Sharing Across Institutes

A cancer center sequences tumor genomes and streams annotated variant data to Kafka topics. NHGRI and other research entities can subscribe in real time to ingest and analyze these datasets with ML pipelines, reducing delays from weeks to minutes.

3.2. Real-Time Clinical Trial Monitoring

Adverse event data from NCI's Cancer Therapy Evaluation Program (CTEP) can be streamed in real time to safety monitoring systems. Enables adaptive trial design, earlier risk detection, and more responsive IRB reviews.

3.3. Imaging and AI/ML Model Training

Large-scale imaging data from TCIA can be streamed to federated compute environments using Kafka and Confluent's tiered storage. Researchers train AI models on distributed data without relocating sensitive PHI.

3.4. Public Health Surveillance Integration

SEER data integrated with real-time hospital feeds (via HL7 or FHIR) allows population-scale analytics for cancer epidemiology. CDC, NCI, and academic centers can build dashboards to detect cancer incidence trends in near real-time.

4. Governance, Compliance, and Security

Confluent meets federal compliance standards:

- FISMA Moderate & FedRAMP-compatible deployments
- RBAC and ACLs for fine-grained access control
- Audit logging and encryption at rest/in transit
- Integration with NIH Identity and Access Management (IAM) tools

Confluent's centralized governance framework supports versioning, lineage tracking, and schema evolution—critical for multi-agency collaboration and reproducibility of research.

5. Scalability and Deployment Options

Confluent supports:

- Hybrid deployments across on-premise and cloud (NIH STRIDES Initiative-compatible)
- Tiered storage for handling large datasets like high-resolution pathology slides
- Elastic scalability, from pilot projects to enterprise-wide implementation

6. Strategic Alignment with NIH Data Science Goals

Confluent supports:

- FAIR Principles (Findable, Accessible, Interoperable, Reusable)
- NIH Strategic Plan for Data Science (SPDS)
- Bridge2AI, AIM-AHEAD, and DataWorks!

By creating a streaming data mesh, Confluent enables NIH to break down silos, reuse critical datasets, and engage in distributed, real-time cancer research collaboration.

7. Conclusion and Recommendations

Confluent is uniquely positioned to serve as the interoperability backbone for NIH cancer research initiatives. We recommend a phased deployment beginning with high-impact use cases (e.g., genomic sharing and clinical trial monitoring) followed by broader integration across NIH institutes. This platform will drive real-time collaboration, enable advanced analytics, and reduce time-to-insight for groundbreaking cancer research.

CONFLUENT

Appendices – Available upon request

A. Architecture Diagram

A sample architecture diagram showing Confluent integrated across NIH systems.

B. Sample Deployment Plans

Phase 1: Pilot with genomic streaming and safety data from NCI clinical trials

Phase 2: Expand to include imaging and public health analytics (SEER)

Phase 3: Integrate across all NIH institutes with hybrid cloud deployment via STRIDES

C. Security and Compliance Summary

- Data encryption in transit and at rest

- Role-based access controls (RBAC) and audit logs

- Supports HIPAA, FISMA Moderate, and FedRAMP-compliant architectures

- Integration with existing NIH IAM and logging tools

John Poulos
Jpoulos@confluent.io
301-704-8410