IBM | CLOUDERA

SOLUTION BRIEF

DATA-IN-MOTION PHILOSOPHY

A Blueprint for Enterprise-wide Streaming Data Architecture

Real Life Data-in-Motion

Below describes how the data-in-motion capabilities described in this solution brief have been successfully applied by a Cloudera customer, end-to-end.

A global medical device manufacturer successfully modernized their messaging architecture to support a new line of implantable medical devices that generate more data, more often, and at a higher resolution than previous products.

- **Flow management**—Due to the private nature of medical data, the data flow was complex, requiring in motion and at rest encryption. NiFi's no code user interface enabled the initiative to be 100% business driven, only engaging the technology teams as needed.
- **Streams messaging**—Messaging volume jumped from quarterly reporting of device status to real-time health monitoring. Kafka enabled the business to scale that volume across multiple on-premises and cloud environments.
- **Stream processing and analytics**—The company had to transition from batch to real-time data processing. Flink handles both models along with complex event processing that is planned for the near future. The company, therefore, only needs to adopt and support one type of stream processing and analytics engine.

Don't Let Your End-to-End Streaming Architecture Fall Short

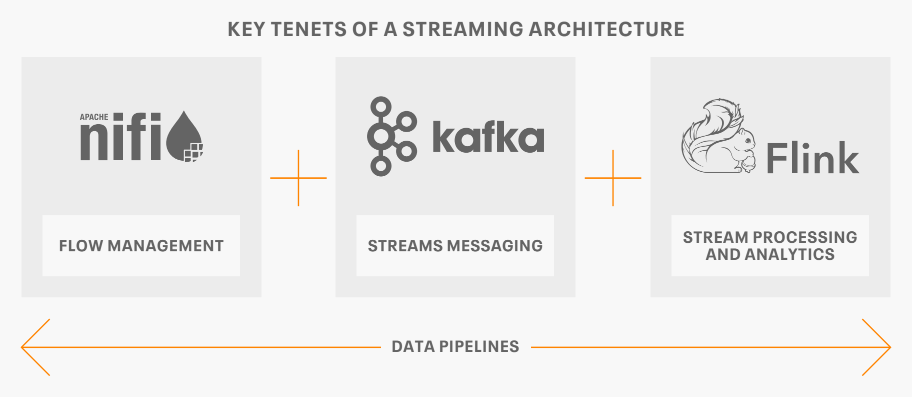
It's not enough to have the best messaging solution at the heart of your end-to-end streaming architecture. We've learned that as a result of supporting our customers through their data journeys. Flow management, along with stream processing and analytics, are two additional tenets that need to be unified with streams messaging capabilities. These three tenets, if properly integrated will ensure a sustainable, scalable, and adaptable end-to-end streaming architecture. Like a three legged stool, one weak tenet can make the entire structure fall short.

This solution brief describes Cloudera's data-in-motion philosophy and is meant as a blueprint to help business and technology decision makers evaluate and simplify their approach to streaming data across their enterprise.

Streaming Architecture in Context

Below we have listed the three tenets that together provide a unified end-to-end streaming architecture.

- **Flow management**, broadly speaking, refers to the collection, distribution, and transformation of data across multiple points of producers and consumers
- **Streams messaging** is the provisioning and distribution of messages between producers and consumers
- **Stream processing and analytics** is how you generate real-time analytical insights from the data streaming between producers and consumers



KEY TENETS OF A STREAMING ARCHITECTURE

Cloudera's data-in-motion philosophy is rooted in the complementary powers that are brought to the table by Apache NiFi for flow management, Apache Kafka for streams messaging, and Apache Flink for stream processing and analytics.

The Flow Management Customer Journey
Our flow management journey with our customers goes back many years.

What started out as a better way to ingest data into a date lake evolved to address scalable streaming event architecture requirements and then to let customers tap into and run analytics on those streams.

We list the stages below that highlights how best-in-class compute engines became our data-in-motion vision.

1. **NiFi** first deployed to enable codeless data lake ingestion. Later it is applied to federated clusters to enable full enterprise data movement.
2. **Kafka** is integrated with NiFi as the primary messaging substrate to support scalable event architectures.
3. **Flink** introduced a third generation stream processing and analytics engine to enable real-time processing as easily as batch.
4. Integration with **Cloudera Data Platform** creates a flexible platform as a service (PaaS) through a shared data experience.

Data-in-Motion Philosophy

Cloudera believes that a holistic end-to-end data pipeline must leverage best-in-class compute engines for their respective tenets. Additionally, high level platform abstractions should handle the complexities of connecting and managing engines in the background, so that users can focus on business logic. Cloudera Data Platform (CDP) realizes that two part philosophy as follows:

1. Cloudera DataFlow (CDF) is the data-in-motion platform that supports the entire streaming data journey and integrates all three tenets from:
   - Data capture and flow management at the edge; to
   - Provisioning that data directly to/from your Kafka messaging backbone; and/or
   - Stream processing and analytics
2. Shared Data Experience (SDX) provides a common set of integrated services, including unified security and governance across data centers and cloud environments (see The Shared Data Experience on page 5).

The Limits of 'Messaging as the Center of Everything'

A scalable messaging solution at the heart of your end-to-end architecture is important but it's not enough to handle advanced real-time use cases or the technical and operational data movement requirements of expanding enterprises.

In response to real-time business demands, technology teams have been pivoting from large monolithic database architectures to event driven applications and microservices design. In addition to analytics on data at rest, more and more decisioning is being derived directly from real-time data streams.

Kafka has emerged as the single central backbone of streaming architectures for large organizations and for good reason. It addresses the fundamental challenges of scalability and is highly optimized for both ad-hoc and sustained exchange of messages. However, it does not address the challenges related to ingesting real-time data from various sources or generating real-time insights on the data as it is being streamed. Those are best addressed by a combination of flow management and stream processing and analytics capabilities in your streaming architecture.

Key Considerations for Flow Management

There are three important aspects to consider when evaluating your flow management needs. They are: extensible tooling, ease of use, and data provenance. Apache NiFi is a real-time integrated data logistics and simple event processing platform that inherently addresses all three.

Extensible Tooling

The lifecycle of data, from where it is produced to points of consumption, can span vast geographic borders and security boundaries. For example, sensor data may need to be routed sequentially or in parallel to regional business centers, commercial partners, and global headquarters for different reasons. Some of that data may be sensitive and need to be sequestered or filtered out. Organizational dynamics offer another level of complexity as data producers and consumers are added, removed, modified, or redesigned ad-hoc.

The tooling needed to address varied and unique data flow scenarios needs to be flexible and extensible. NiFI does this by acting on behalf of data sources and targets that are at vastly different points in the enterprise and different levels of maturity. It is data source agnostic and can handle disparate and distributed sources of differing formats, schemas, protocols, speeds, and sizes.

Machine sensors, geolocation devices, click streams, files, social feeds, log files, videos, and more all fall under the NiFi purview. Simple event processing like combining streams and enriching or filtering data are also handled by NiFi.

The Streams Messaging Customer Journey
Interviews with top Cloudera Kafka customers made it clear that the lack of visibility across Kafka clusters was a huge problem. We refer to this as "Kafka Blindness".

In response, Cloudera developed a unified toolset (see Full Visibility of Streams Messaging on this page) to solve the unique challenges of teams across the enterprise, including:

- **Kafka platform operations teams** now have visibility at a cluster and broker level, the effects of the broker on the infrastructure, and vice versa.
- **Kafka DevOps and application teams** can fully visualize how data flows between producers, brokers, topics, and consumers and monitor the key performance metrics of each.
- **Governance and security teams** have full transparency around chain of custody, audit, metadata, access control, and data lineage.

For a detailed overview of how Cloudera cured "Kafka Blindness", see the white paper, "Manage, Monitor and Replicate Apache Kafka Across the Enterprise and Cloud."

Data Provenance

The data integration complexities described above also make it extraordinarily hard for organizations to understand the origin and attribution of data as it moves throughout the enterprise. This concept is generally referred to as data provenance and is a critical top-of-mind issue for CDOs and CISOs. At any point in time, they need to count on their teams being able to explain exactly how any data point was affected by any system.

Data provenance is inherent to NiFI because it naturally generates data lineage information in everything that it does at a very fine grained level that records changes before and after an event.

Since NiFi controls the data flow between producers and consumers, enterprise-wide end-to-end data lineage is captured. Beyond data governance and security value, NiFi expands operational awareness of which and how systems communicate and the latency in between. That is very powerful.

Ease of Use

Subject matter experts who lead business solutions and understand end-to-end data flow may not know how to write good code. For this reason, NiFi makes the integration work described above approachable through a no-code, drag-and-drop interface. The experience is interactive because the orchestration flows they build translate into real functions that affect real data in real-time.

Your team can quickly build powerful data flows because the subject matter experts have very real and tactile feedback as to what does and doesn't work.

Additionally, with 300+ pre-built processors, you have the capability to seamlessly connect any data source to any target with simplicity and ease.

Full Visibility of Streams Messaging

The streams messaging tenet requires high scalability and stability and the best-in-class compute engine in that space is Apache Kafka. Kafka has emerged as the single central backbone of streaming architectures for large organizations, igniting data-in-motion innovations across financial services, telecommunications, manufacturing, and numerous other industries.

Cloudera is dedicated to the Kafka community and continues to be actively involved with deep engineering relationships that have led to critical innovations and product improvements.

As part of the CDF data-in-motion platform, Cloudera provides an entire ecosystem of components that support and enhance high performing Kafka environments, some of which are described below.

- **Cloudera Streams Messaging Manager (SMM)** is a single monitoring/management dashboard that provides end-to-end visibility into how data moves across Kafka clusters between producers, brokers, topics, and consumers.
- As part of a unified toolset, directly incorporated into SMM is the **Streams Replication Manager (SRM)**. SRM is an enterprise-grade cross-cluster Kafka topic replication solution that enables your teams to ensure business continuity and high availability for your streaming architecture.
- **Schema Registry** enables your teams to safely mitigate interruptions that occur due to schema mismatches. It manages, shares, and supports the evolution of all producer and consumer schemas across the Kafka landscape.
- **Cruise Control** is a Kafka load balancing component used in large Kafka installations. While automatically balancing partitions based on user defined goals, Cruise Control also detects and actively addresses anomalies.

The Stream Processing and Analytics Customer Journey

At Stage 3 of The Data Flow Customer Journey (see The Flow Management Customer Journey on page 2), customers had an easy way to move data across the enterprise with NiFi and became integrated with Kafka as the scalable event/streams messaging architecture. Now they need to tap into those streams to extract real-time insights.

While Cloudera supported Spark Structured Streaming and Kafka Streams, they didn't address the challenges of late or missing data, complex event processing, or ensuring resilience, high availability, and no data loss.

Recently, Apache Flink emerged as the 3rd generation stream processing and analytics engine that addresses those complex analytics needs and has since been introduced into Cloudera's data-in-motion portfolio.

Learn about Flink's capabilities and the technical and operational factors that are crucial in selecting a stream processing and analytics engine, see analytics engine with this white paper, "Choose the Right Stream Processing Engine for Your Data Needs."

Stream Processing and Analytics With Loads of Control and Options

The third tenet of capabilities is stream processing and analytics. While the first two tenets provide a powerful way to move, provision, and replicate streaming data with full visibility and provenance, that data also needs to be processed in real-time to get actionable intelligence for business decisions.

Our customers, through their data-in-motion journey, realized that they need a best-in-class stream processing and analytics engine that covers the full range of data pipeline requirements.That is why Apache Flink has been incorporated into CDF (see The Stream Processing and Analytics Customer Journey on page 4).

Apache Flink is a distributed processing engine and a scalable data analytics framework that can process millions of data points or complex events very easily and deliver predictive insights in real-time. It is best-in-class because it gives you loads of technological and operational control to address some of the more sophisticated analytic use cases.

Flink has a streaming-first (over batch) approach to processing high-volume streams of data at high-scale, while supporting key features such as stateful streaming, exactly-once delivery, built-in fault tolerance/resilience, and advanced windowing techniques. Flink can process real-time data as it is generated as well as store data in storage filesystems, public cloud object stores, or other durable repositories.

From an ease of use perspective, Cloudera SQL Stream Builder enables developers, analysts and data scientists to write streaming applications using industry-standard SQL. It provides an interactive experience so the development process is quick, easy, and productive. It provides an advanced materialized view engine to interface with applications, tooling, and services via REST so the results are reported through alerts, visualization dashboard, and other analytics applications in real-time.

Below are a few reasons why some of the biggest brand name companies have already invested in large deployments of Flink for their real-time stream processing and analytics needs.

- Flexibility across microservices, batch, and streaming use cases is required
- High throughput is necessary
- Low latency is crucial
- Advanced state capabilities like complex event processing are needed
- Operational efficiency is as important as technical capabilities
- Ease of use and scalability is important to encourage adoption across the enterprise
- Developers don't need to understand the Java and Scala programing languages and complexities of watermarks

Flink is an integral component of Cloudera's data-in-motion philosophy and as such it will fully integrate into your organization's security framework, operational processes, and support structure.

The Shared Data Experience (SDX)

A key part of Cloudera's data-in-motion philosophy is to provide high level platform abstraction that shields users from the complexities of connecting and managing the best-in-class compute engines described in this brief.

Cloudera SDX is the foundational element of that principle and a key differentiator from other platform providers. SDX lets you set data security, governance, and control policies once and consistently enforce them across all components and anywhere—in the data center or in hybrid and multi-clouds.

It boosts deployment choice and flexibility with the following:

- **Apache Ranger** delivers enterprise-scale security with centralized, granular, consistent access control. It monitors and protects assets across your end-to-end streaming platform through a single pane of glass.
- **Apache Atlas** provides enterprise-grade auditing, lineage, and governance capabilities. It is a key component to providing enterprise visibility and understanding.
- **Apache Knox** is a gateway based SSO that simplifies security controls with seamless, secure user access to cluster data and permissions to execute jobs while maintaining compliance with enterprise security policies.

Unification Through A High Level of Abstraction

Cloudera's philosophy is that best-in-class compute engines are required to adequately address the unique challenges of the end-to-end management needs of streaming data.

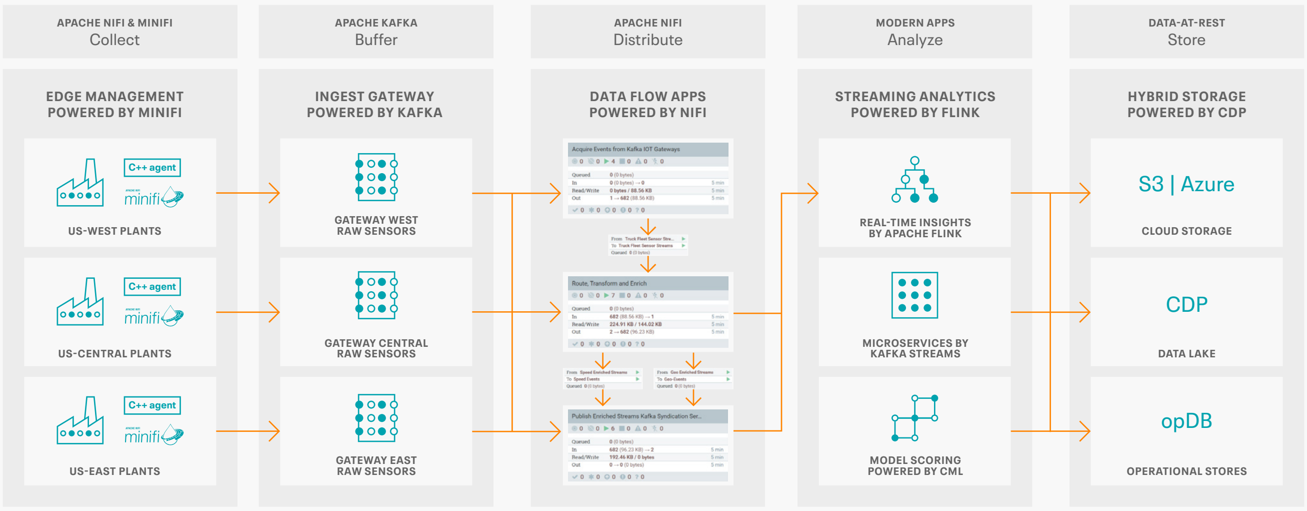
To turn that philosophy into reality, we provide a unified platform to handle the complexities of connecting, managing, and integrating those engines through a high level of abstraction.

This means that your teams can focus on the true business logic that goes into building an end-to-end data pipeline because Cloudera seamlessly renders that logic across the respective engines. This shields the user from that complexity. We've already described a few examples of how we do this:

- A no code user interface that enables subject matter experts to translate data flow diagrams into real functions that affect real data in real-time
- A single monitoring/management dashboard that provides end-to-end visibility and replication solutions to ensure business continuity and high availability across large streams messaging environments
- Expressive and flexible APIs enable developers to build sophisticated stream processing and analytics applications easily
- Tight integration with a common set of services that offer unified security and governance across your enterprise's data center and cloud environments

The reference architecture diagram below shows how NiFi, Kafka, and Flink work together within the larger context of an end-to-end data pipeline.

A DATA-IN-MOTION REFERENCE ARCHITECTURE


APACHE NIFI & MINIFI
Collect
APACHE KAFKA
Buffer
APACHE NIFI
Distribute
MODERN APPS
Analyze
DATA-AT-REST
Store
EDGE MANAGEMENT POWERED BY MINIFI
C++ agent
minifi
US-WEST PLANTS
US-CENTRAL PLANTS
US-EAST PLANTS
INGEST GATEWAY POWERED BY KAFKA
GATEWAY WEST RAW SENSORS
GATEWAY CENTRAL RAW SENSORS
GATEWAY EAST RAW SENSORS
DATA FLOW APPS POWERED BY NIFI
Acquire Events from Kafka IOT Gateways
From Truck Fleet Sensor Stre...
To Truck Fleet Sensor Streams
Route, Transform and Enrich
From Speed Enriched Streams
To Speed Events
From Geo Enriched Streams
To Geo-Events
Publish Enriched Streams Kafka Syndication Ser...
STREAMING ANALYTICS POWERED BY FLINK
REAL-TIME INSIGHTS BY APACHE FLINK
MICROSERVICES BY KAFKA STREAMS
MODEL SCORING POWERED BY CML
HYBRID STORAGE POWERED BY CDP
S3 | Azure
CLOUD STORAGE
CDP
DATA LAKE
opDB
OPERATIONAL STORES

Vision

The Cloudera data-in-motion vision is to provide one unified set of services that simplify the development of end-to-end data pipelines and integrate with your organization's security framework, operational processes, support structure, and to scale up and down in line with business demand.

We deliver on that vision by providing an integrated platform of best-in-class data streaming compute engines with a high level of abstraction so that you and your teams can focus on the true business logic of building streaming data pipelines.

About IBM

IBM provides a complete set of AI-enabled solutions that allow organizations to collect data of any type, source and structure to make it simple and accessible across multiple vendors, deployments and workloads. Our enterprise grade, secure solutions support hybrid multi-cloud environments and automation through embedded AI capabilities.

Find out more here:
ibm.com/analytics/partners/cloudera

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more about Cloudera Data Platform at cloudera.com/cdp

Learn more about Cloudera DataFlow at cloudera.com/cdf

Learn more at cloudera.com

Cloudera, Inc. 5470 Great America Pkwy, Santa Clara, CA 95054 USA cloudera.com