CLOUDERA

WHITE PAPER

CLOUDERA DELIVERS THE BEST KAFKA ECOSYSTEM TODAY

Serving Hundreds of Customers Globally

DATA IN MOTION

Why Cloudera's Kafka Offering is the Best in the Industry

Global businesses continually transform and adapt to changing economic conditions and consumer expectations. Within each organization, both the business and technology communities share the responsibility of delivering, in near real-time, the innovative products and services that their customers, employees, and regulators expect. Apache Kafka is the key architectural component to a wide range of streaming data initiatives that enable enterprises to deliver on those responsibilities.

As a result of providing excellent support to hundreds of customers on their advanced and large scale deployments of tens of thousands of Kafka brokers, we've learned that it is not enough to just have the best messaging solution at the heart of your end-to-end streaming architecture. This is because data management challenges exist all the way from data ingestion to preparation to processing data in real-time to gain predictive insights. Flow management, along with stream processing and analytics, are two additional tenets that need to be unified with streams messaging capabilities to implement a complete end-to-end streaming architecture.

While there are different vendors in the market that claim support for Kafka, Cloudera stands out distinctively from the pack with its holistic and comprehensive platform offering. Cloudera's commitment to the open source community and its penchant for listening to the voice of the customer helps it deliver advanced innovations in the Kafka ecosystem of components. Also, as a trailblazer in the Enterprise Data Cloud market, Cloudera has also been delivering on extending the Kafka ecosystem from on-premises deployments to public and private cloud environments.

This paper describes how all the three data-in-motion tenets are unified through a common data experience across on-premises, hybrid, and multi-cloud environments. Through the integration of Cloudera DataFlow (CDF) with the Cloudera Data Platform (CDP) that leverages best-in-class, open source-based engineering, Cloudera delivers the best Kafka ecosystem today that ensures a sustainable, scalable, and adaptable end-to-end streaming architecture.

Table of Contents

It Takes a Complete Streaming Platform 4

The Complete Kafka Offering 5
Kafka Streams 5
Kafka Connect 6
Kafka Cruise Control 6
Schema Registry 6
Streams Messaging Manager 7
Streams Replication Manager 7
SQL Stream Builder 8

Security and Governance Are First-Class Citizens 9
Shared Data Experience (SDX) 9
Apache Ranger 10
Apache Knox 10
Apache Atlas 10

CDP Enables Multi-Cloud Support 11

The Complete Edge-to-Cloud Streaming Data Platform 12
Data-in-Motion Philosophy 13

A Steel Manufacturing Success Story 14
Challenges 14
Solutions 14
Results 14

The Reasons Why Cloudera is Superior in the Kafka Space 15

The Three Tenets of a Unified End-to-End Streaming Architecture
There are 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.

Cloudera DataFlow (CDF) is a comprehensive edge-to-cloud real-time streaming data platform that collects, curates, and analyzes data so customers gain key insights for immediate actionable intelligence. It meets the challenges faced with data-in-motion, such as real-time stream processing, streaming analytics, data provenance, and data ingestion from IoT devices and other sources.

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
- Provisioning that data directly to/from your Kafka messaging backbone
- Stream processing and analytics

Read our solution brief, "Data-In-Motion Philosophy: A Blueprint for Enterprise-wide Streaming Data Architecture" to understand how it all comes together.

It Takes a Complete Streaming Platform

Business and technology teams are driven to improve the flexibility, speed, efficiency, accuracy, and security of capturing, provisioning, distributing, and analyzing data that is streaming across their enterprise and with external parties. Accordingly, organizations have been pivoting from large monolithic database platforms to event driven streaming architectures and microservices design.

Apache Kafka has emerged as the single central backbone of event-based architectures because it addresses the fundamental challenges of scalability and is highly optimized for both ad-hoc and sustained exchange of messages. However, Kafka doesn't address the challenges of how the data is ingested from multiple sources into your enterprise or cloud. It also doesn't address the challenges of how these real-time data streams are analyzed with extremely low latency to produce meaningful and actionable insights for key decision makers.

Cloudera delivers the best Kafka ecosystem today by integrating our data-in-motion platform, Cloudera DataFlow (CDF) with Cloudera Data Platform (CDP), the world's first enterprise data cloud.

The diagram in Figure 1 illustrates how CDF supports the entire streaming data journey from data capture and flow management at the edge (1) to provisioning that data directly to/from your Kafka messaging backbone (2) and/or stream processing and analytics (3). It is tightly integrated with CDP's Shared Data Experience (SDX)—a common set of services that offer unified security, governance, lineage, and control (4) across your enterprise's data center and cloud environments (5).

Figure 1

Stream Processing and Analytics: Choose the Right Tool for the Job
Real-time streaming applications are confronted with both simple and complex sets of challenges and there are a number of ways in which to address them.

The stream processing and analytics engine that is best for you depends on your organization's use cases, team makeup, and various technology, operational, and organizational factors.

To understand the best fit for purpose across the stream processing and analytics engines, including Kafka Streams, Spark Structured Streaming, Storm with Trident, and Flink, read this white paper, "Choose the Right Stream Processing Engine for Your Data Needs."

In short, CDF provides flow management and stream processing capabilities that IT teams need while the data engineering and platform teams can deploy, manage, monitor and replicate Kafka clusters with full end-to-end visibility. CDP enables safe and consistent deployments and migrations of such streaming capabilities across hybrid, private, or multi-cloud environments.

The rest of this paper describes how Cloudera provides the best Kafka ecosystem of components in the industry today.

The Complete Kafka Offering

As described earlier, Kafka is the key architectural component to a wide range of streaming data initiatives that enable enterprises to keep up with customer demand, provide better services, and proactively manage risk. Although there may be an assortment of streaming data approaches within the same organization, Kafka is the enterprise-wide common denominator because it provides:

- **High throughput and low latency**—Kafka supports millions of messages per second, with latencies as low as a few milliseconds.
- **Scalability**—A Kafka cluster can be elastically and transparently expanded without downtime.
- **Durability and reliability**—Messages are persisted on disk and replicated across clusters to prevent data loss.
- **Fault tolerance**—The platform is immune to machine failure in the Kafka cluster.
- **High concurrency**— Ability to simultaneously handle thousands of diverse clients, writing to and reading from Kafka.

Cloudera has hundreds of happy customers getting excellent support on their advanced Kafka deployments that process billions of messages per second. That is because we provide the most comprehensive Kafka platform with all necessary ecosystem components and some productivity-boosting innovations for a complete Kafka implementation.

Only Cloudera provides simple deployment and robust troubleshooting and monitoring of Kafka, as well as shared compliance-ready security, governance, lineage, and control in one simple application across multiple on-premises, hybrid, private, public, or multi-cloud environments.

Embedded in that ecosystem is a unified set of tooling to connect data sources, manage and reuse schemas, optimize clusters, and enable high availability and disaster recovery replication use cases. We will discuss some of the key Kafka ecosystem components in the following sections.

Kafka Streams

Kafka Streams is the built-in stream processing library of the Apache Kafka project and provides real-time stream processing and analytics with high throughput and very low latency. It is a good fit if you are developing solely within a Kafka to Kafka pipeline, you don't need or want another cluster for stream processing and analytics in the future, and operational and resilience requirements are simple or handled elsewhere.

Kafka Streams enables you to perform common stream processing functions like filtering, joins, aggregations, and enrichments on the data stream. Good use cases include building lightweight microservices, straight forward ETL jobs, and simple stream analytics apps. For more sophisticated use cases consider Apache Flink (see Data-in-Motion Philosophy on page 13).

Because Kafka Streams is an integral part of Cloudera's Kafka ecosystem, you have the additional capability to build microservices apps that address complex security, governance and audit requirements (see Security and Governance Are First-Class Citizens on page 9).

We've Got Your Back with Our Expertise
With more experience across more production customers for more use cases, Cloudera Professional Services and Training (PS&T) is the leader in Kafka end-to-end services and support. For example:

- Cloudera PS&T helped one of the largest truck manufacturers in North America build a vehicle telematics pipeline using a streaming platform built with Kafka. Sensor data is continuously sent from over 150,000 trucks in North America every 2 minutes. This data is processed in real-time and provides vehicle details—from speed and idling, to fuel use, low tire pressure, and more.
- One of the largest energy companies in the world engaged Cloudera PS&T to build a mission critical commodity trading platform using Kafka. Combined with other complementary Cloudera technologies, this platform processes extremely high throughput and low latency messages to generate trader alerts based on proprietary machine learning algorithms on streaming data.

Cloudera PS&T is considered a trusted advisor by all of our customers throughout the globe across a range of industries.

+1k

Successful customer engagements

+300

Professional services consultants and architects globally

+150

Kafka experts

Kafka Connect

Kafka also includes a connectivity framework called Kafka Connect and, like Kafka Streams, it is a good fit if you are developing solely within a Kafka to Kafka pipeline because it was engineered to simplify reading and writing to and from Kafka only. It is best used for simple use cases because Kafka Connect has limited data transformation capabilities. For use cases that require a higher sophistication of data pipelines or if you need a no-code user interface with a wide range of pre-built processors, consider Apache NiFi (see Data-in-Motion Philosophy on page 13).

Any security and governance restrictions and complexities of a standalone Kafka Connect are mitigated when it is implemented as part of CDP (see Shared Data Experience (SDX) on page 9).

Kafka Cruise Control

Kafka Cruise Control enables you to manage and load balance large Kafka installations. It is the solution for platform teams that need first class management services that address hard problems such as frequent hardware/virtual machine failures, cluster expansion/reduction, and load skew among brokers. It solves these challenges by balancing clusters intelligently and with automated anomaly detection and remediation.

While it automatically balances partitions based on user defined goals, Kafka Cruise Control also detects and actively addresses anomalies. For example, if there is a broker failure, Kafka Cruise Control will fix the cluster by removing the failed brokers. In the case of disk failure, all the offline replicas will be moved to healthy brokers.

Schema Registry

Schema Registry is an important component of the Cloudera Kafka ecosystem because it 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. You can also avoid having to attach a schema to every piece of data.

As part of CDF's streams messaging capabilities, Schema Registry provides a shared repository of schemas that allows applications to flexibly interact with each other across the Kafka landscape by using the same schemas from end-to-end. This is particularly useful for managing data flows with schema-based routing. For example, parsing a syslog event to extract the event type, and then based on that type, route it to a downstream Kafka topic.

The screenshot in Figure 2 below shows how you would use the Schema Registry UI to create schema groups, schema metadata, and add schema versions.

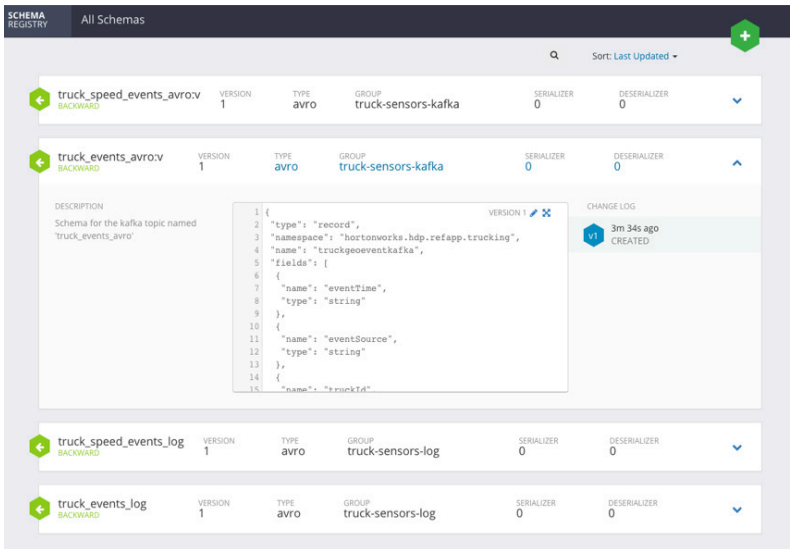

Figure 2

Get More out of Streams Replication Manager

Replication is often associated with disaster recovery and high availability use cases. But Streams Replication Manager enables additional business critical use cases, some of which are described below.

- **Aggregation for analytics:** Aggregate data from multiple streaming pipelines and across multiple data centers to run batch analytics jobs that provide a holistic view across the enterprise.
- **Data deployment after analytics:** This is the opposite of the aggregation use case in which the data generated by the analytics application in one cluster (say the aggregate cluster) is broadcast to multiple clusters across data centers for end user consumption.
- **Isolation:** Due to performance or security reasons, data needs to be replicated between different environments to isolate access. In many deployments, the ingestion cluster is isolated from the consumption clusters.
- **Geo proximity:** In geographically distributed access patterns where low latency is required, replication is used to move data closer to the access location.
- **Cloud migration:** As more enterprises have an on-premises and cloud presence, Kafka replication can be used to migrate data to the public or private cloud and back.
- **Legal and compliance:** Much like the isolation use case, a policy driven replication is used to limit what data is accessible in a cluster to meet legal and compliance requirements.

For more about innovation in replication, read the white paper, "Manage, Monitor and Replicate Apache Kafka Across the Enterprise."

Streams Messaging Manager

Probably the most striking component of the Cloudera Kafka ecosystem is Cloudera Streams Messaging Manager (SMM) because it provides so much power across so many teams. 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. It is a complete Kafka toolset that addresses the unique needs of DevOps, application development, platform operations, governance, and security teams.

As an example, the image in Figure 3 below shows interactive visualizations that enable you to fully understand how data flows across Kafka clusters.

Figure 3

An example why SMM is important to the Cloudera Kafka ecosystem is that you're able to optimize Kafka environments based on key performance insights gathered from various brokers and topics. With the tight integration with the Schema Registry (see page 6) schemas can be managed from the same user interface.

SMM is a differentiating innovation designed by our engineering teams to enable hundreds of Kafka customers across the globe to gain complete visibility into their Kafka clusters and cure themselves of their Kafka blindness. Read more about this in the white paper titled, "Manage, Monitor and Replicate Apache Kafka Across the Enterprise."

Streams Replication Manager

While we are focused on delivering key management and monitoring capabilities for important personas across the IT landscape, we are also deliberate about ensuring business continuity and high availability for your streaming architecture.

Streams Replication Manager (SRM) is an enterprise-grade replication solution that enables fault tolerant, scalable and robust cross-cluster Kafka topic replication and enables a number or business critical replication use cases such as high availability, disaster recovery, cloud migrations, geo proximity, and many others (see Get More out of Streams Replication Manager on this page).

SRM is built on the innovations that Cloudera has brought to MirrorMaker, the original Kafka open source messaging replication tool. Cloudera addresses some of the severe shortcomings of the original by unveiling MirrorMaker2, which infuses the concepts of clusters, global configuration, and global management APIs.

Improved Customer Experience in the Telecommunications Industry

The Ooredoo Group is an international communications company serving consumers and businesses in 10 countries across the Middle East, North Africa and Southeast Asia.

Ooredoo Kuwait needed a comprehensive platform to scrutinize customer network traffic at scale as well as analyze usage metrics and communication channels in order to provide a better service and take the right actions. They deployed Cloudera's streaming data platform to tackle a variety of critical use cases, including stream processing, log aggregation, large-scale messaging and customer insights.

Results included:

- Improved overall customer experience through strategic use of data analysis
- Reduced infrastructure management costs and TCO
- Enablement of real-time actions to improve business outcomes

Read the complete customer success story, "Ooredoo Kuwait: leveraging big data to improve the customer experience in real-time."

The image in Figure 4 below shows how SRM provides replication monitoring, details, and metrics at cluster and topic levels including the status, throughput, and replication latency for all topics being replicated.

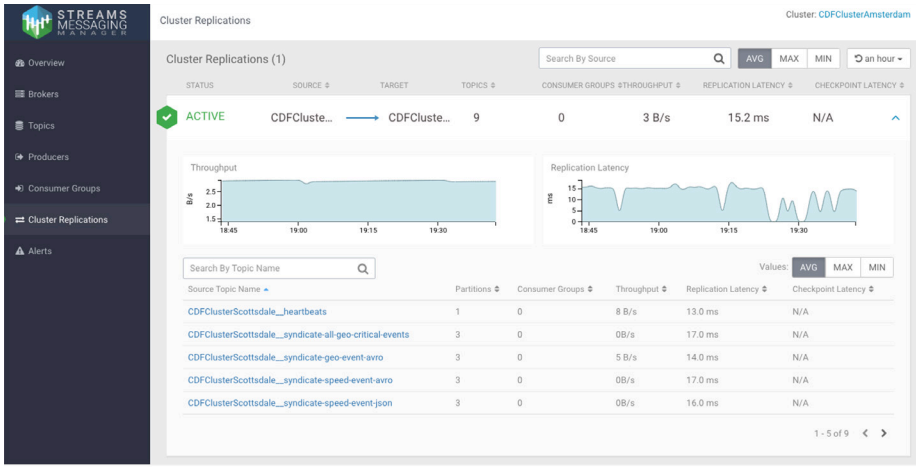

Figure 4

SQL Stream Builder

Cloudera SQL Stream Builder enables developers, data analysts, and data scientists to write streaming applications using just SQL. It provides an interactive experience, so the development process is quick, easy, and productive. It offers syntax checking, error reporting, schema detection, query creation, sampling results, and creating outputs with its powerful interface and APIs.

SQL Stream Builder continuously runs SQL via Apache Flink and in Apache Kafka. Developers don't need to understand the Java and Scala programming languages or complexities like watermarks. The SQL job inherits and leverages the robust nature of Apache Flink and can be restarted and retain state (for the fast restart and upgrade), and have massive scalability and robust run-time framework.

Learn more about SQL Stream Builder.

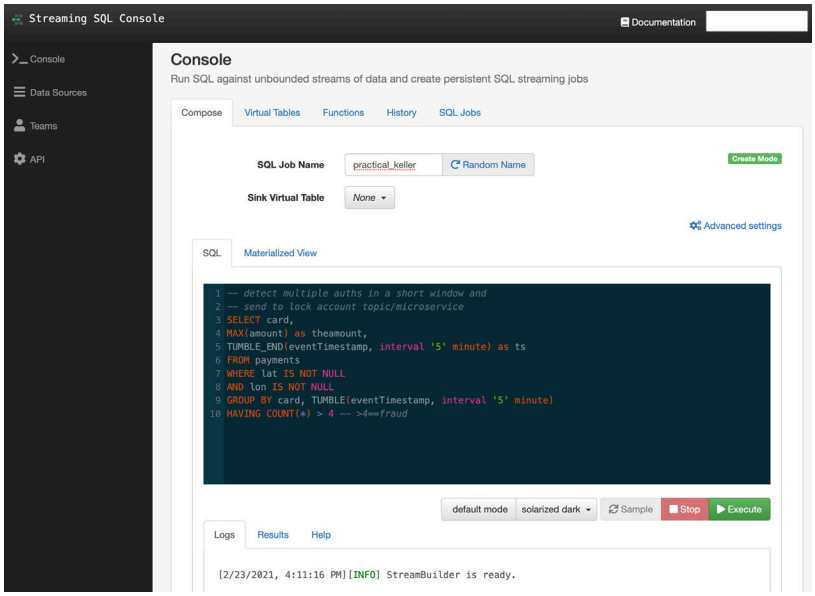

Figure 5

Security and Governance Are First-Class Citizens

The previous pages described the strong capabilities of each key component that makes up the Kafka ecosystem within the CDP framework. Below, we describe why and how security and governance are first-class citizens in our platform.

Shared Data Experience (SDX)

CDP's Shared Data Experience (SDX) is the key differentiator from other platform providers. It is what enables the seamless integration of all parts of our Kafka ecosystem and safe, efficient, and consistent experience of deployments and migrations across all data environments: on-premises, hybrid, private, or multi-cloud (see Figure 6 below).

This is because data security, control policies, governance, and lineage are set once and automatically enforced on every data platform and across all components of your streaming architecture. Below, we briefly describe some of SDX's key components.

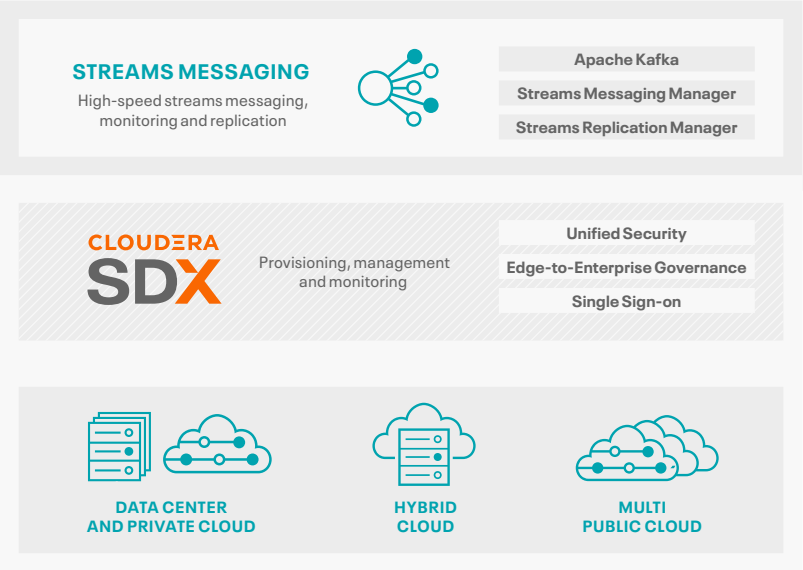

Figure 6

Apache Ranger

A single pane of glass for enterprise-wide security with centralized, granular, and consistent access control is provided by Apache Ranger. It addresses the four main pillars of security that are needed to support sensitive and regulated data use cases: identity, access, data protection, and visibility.

A sample of capabilities includes data encryption, dynamic row filtering, dynamic column masking, attribute-based and fine-grained access control, and the ability to enforce logical and physical separation of administrative duties on all infrastructures.

In another example of seamless integration across CDP, gateway-based SSO access is provided through Apache Knox (below) while Apache Atlas (below) enables full end-to-end data classification and audit in order to perform analytics on regulated data and redact sensitive data when needed.

Apache Knox

Apache Knox is a gateway based SSO that simplifies security controls with seamless, secure user access to cluster data and the permissions needed to execute jobs while maintaining compliance with enterprise security policies. It provides a single entry point for all user interfaces across the Kafka ecosystem, alleviating the need to remember which node uses what services, for example. From a security administration point of view, Knox reduces the number of ports that need to be opened.

Knox is one of the reasons why all teams across the enterprise are able to safely and easily access their applications regardless of the data environment it is located in.

Apache Atlas

Earlier, we described how Atlas enhances Ranger security and data protection functions with full end-to-end data classification and audit capabilities. However, Atlas' enterprise-grade auditing, lineage, and governance capabilities are critical across the entire Kafka ecosystem. For instance, with Apache Atlas, you have access to the metadata and metrics about every Kafka topic and can produce complete data lineage and audit trails, even across multiple Kafka hops.

The example in Figure 7 below shows the seamless integration between Atlas and Streams Messaging Manager, which we introduced earlier (see page 7). A user can drill down from an edge sensor consumer (1) and launch a data lineage diagram (2) to directly see related flows across Kafka topics (3) as an efficient way to troubleshoot problems.

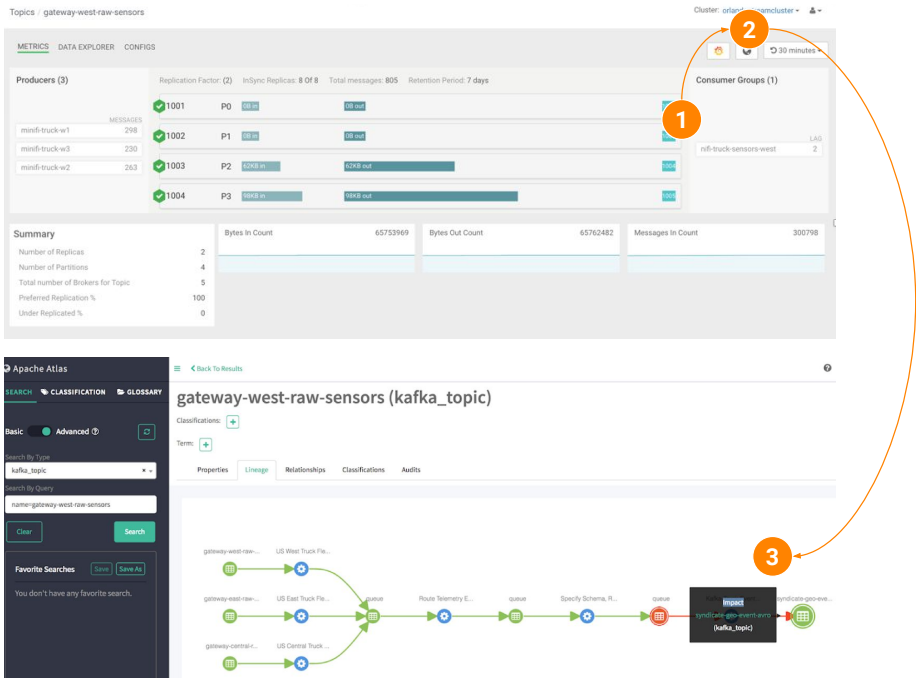

Figure 7

CDP Enables Multi-Cloud Support

Enterprises struggle to take their streaming data to the cloud because they often need to retain their on-premises footprint, for reasons like data sensitivity, data gravity, or security. In this scenario, they need to adopt a hybrid cloud architecture. For enterprises facing such challenges within such complex environments, CDP is an excellent vendor-agnostic platform to embrace multi-cloud or hybrid cloud strategies.

Cloudera took the best of CDF's streaming abilities into the CDP world so that the same holistic enterprise-wide on-premises streaming experience can be extended to the cloud as well. For example, the capabilities described earlier with Kafka and Streams Messaging Manager can be quickly provisioned into a public cloud in just a matter of minutes and can continue to take advantage of CDP's unified data security, governance, lineage, and control. Similarly, the other tenets of CDF, like Flow Management and Stream Processing & Analytics, are also made available on CDP, giving you complete control of how you deploy your streaming architecture across all environments.

This model allows development teams to leverage the same tools and platform across multiple environments and avoid the struggle of managing data across multiple tools. This also enables DevOps teams to easily spin up clusters of the streaming component of choice based on the specific use case they are handling. With CDP's SDX offering the seamless security and governance experience across all components and across all environments, security and governance personnel can also feel assured about how data is managed across such diverse environments.

The Complete Edge-to-Cloud Streaming Data Platform

This paper has focused primarily on the streams messaging aspects of the Kafka ecosystem with regard to how to secure, monitor, balance, and replicate large scale Kafka environments across on-premises, hybrid, private, and public cloud environments. The Data-in-Motion reference architecture diagram in Figure 8 below, puts this all in perspective.

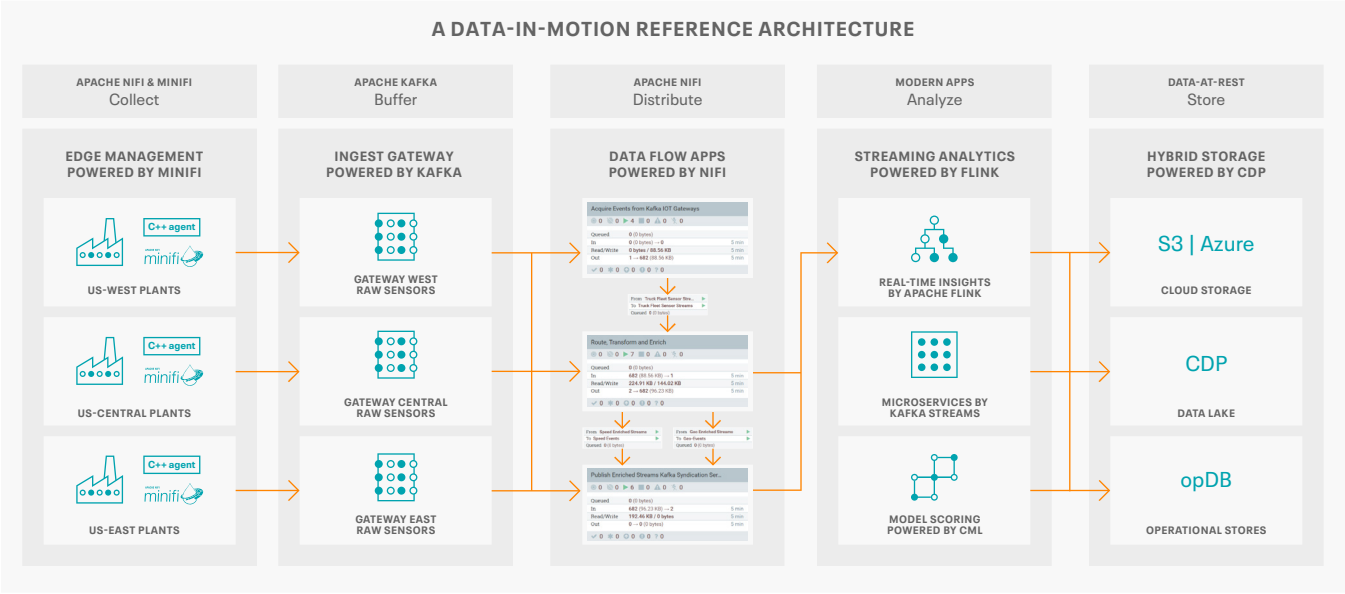

Figure 8

As a result of supporting our customers through their data journeys, we've learned that it is not enough to have the best messaging solution at the heart of your end-to-end streaming architecture. As represented in the diagram above, 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 and is the basis to our data-in-motion philosophy.

Real Life Data-in-Motion
Cloudera's data-in-motion philosophy is illustrated below in real life terms.

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.

Data-in-Motion Philosophy

Cloudera's philosophy is that best-in-class compute engines are required to adequately address the unique challenges of data flow management, streams messaging, and stream processing and analytics. As described in the table below, we deliver on that vision by supporting best-in-class data streaming compute engines while providing a high level of abstraction so that you and your teams can focus on the true business logic of building streaming data pipelines.

THE THREE TENETS OF A UNIFIED END-TO-END STREAMING ARCHITECTURE

Tenet	Compute Engine	Why It Is Best-in-Class	Importance of CDP
Flow management, broadly speaking, refers to the collection, distribution, and transformation of data across multiple points of producers and consumers.	APACHE nifi	Apache NiFi is a real-time integrated data logistics and simple event processing platform. It is best-in-class because it inherently addresses the three important aspects of flow management: extensible tooling, ease of use, and data provenance.	As with the entire Kafka ecosystem, CDP provides a unified platform to handle the complexities of connecting, managing, and integrating these best-in-class 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.
Streams messaging is the provisioning and distribution of messages between producers and consumers.	kafka	Kafka has emerged as the single central backbone of streaming architectures for large organizations because it addresses the fundamental challenges of scalability and is highly optimized for both ad-hoc and sustained exchange of messages.	
Stream processing and analytics is how you generate real-time analytical insights from the data streaming between producers and consumers.	Flink	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.	

For additional insight into above, read our solution brief, "Data-In-Motion Philosophy: A Blueprint for Enterprise-wide Streaming Data Architecture".

Positive Impact of a Complete Streaming Platform

The success story at the right describes how one of the world's largest steel mining businesses worked with Cloudera experts to significantly increase steel production by expanding their current streams messaging Kafka ecosystem with MiNiFi edge management, NiFi flow management, and CDP data lake for high capacity data storage.

The business impact includes:

+6.5%

Increase in productivity

100K

Tons of new steel

.7 SEC

For data to go through the solution

A Steel Manufacturing Success Story

Severstal is one of the world's leading vertically integrated steel and steel-related mining companies, with major assets in Russia. Severstal is Russia's prime high-quality supplier of flats, longs and steel pipes for the construction, automotive, machinery, and oil & gas industries.

Challenges

The company's strategy's key elements are creating an excellent customer experience and achieving leadership in expenses. It could be producing more steel from the same amount of raw material or for the same period.

To do this, the company has developed the largest data lake in Russia's industrial sector (6PB of capacity) in a push to be able to store and work with big data. Besides, Severstal needed an end-to-end solution to help collect, manage, and analyze data. Several challenges were considered in choosing an appropriate solution. Every minute, a single industrial assembly can generate several millions of data points. Moreover, Severstal's production system consists of different mills and mines across Russia, and every facility consists of plenty of these assemblies. Another requirement was handling real-time data feeds. Because steel production contains many high-speed processes, it is crucial to receive, process, and send control action to a facility within seconds.

Solutions

Severstal migrated to a Cloudera CDH data lake, as it offered a complete end-to-end solution to support the company's objectives. The architecture also includes streaming data - with Kafka, NiFi, MiNiFi, and machine learning (ML) models. Professional Services has been instrumental in supporting these efforts.

Severstal applies computer vision to control industrial safety, quality of steel products and raw materials. For example, Severstal implemented a solution that detects defects of the steel surface. It includes ten cameras set on the production line sending over five million photos of steel surface each day to a CV-model. The model processes received photos and predicts whether the photo contains defects or not, and sends the result to a web-application. This solution helps to reduce waste of production time and helps to provide the clients with high-quality steel.

Using NiFi and MiNiFi, Severstal started collecting millions of messages per minute from IoT devices and sensor data from the machinery producing the steel. The data lake stores data collected from transmitters on industrial equipment (IoT), process management information system servers, and MES-systems. This vast amount of data makes it possible to use advanced analytic techniques for operations optimization.

Severstal uses historical data stored in CDH, ranging from several months to five years, to train the ML models. Moreover, using data from Kafka models can retrain and work on a real-time basis. In total, it takes between 0.7-1.5 seconds for data to go through the solution. For example, this approach has been used for implementing a solution that automatically controls the speed of the continuous pickling line.

Results

With its data lake built on Cloudera, Severstal can significantly improve manufacturing processes and productivity. The stored data supports development of advanced digital solutions, which provide cost reduction, supplies of steel products with high quality and increasing production volume.

Within the continuous pickling line use case alone, performance has increased by more than 6.5%, which provides more than 100 thousand tonnes of additional metal processing per year.

Severstal is planning to move to CDP at some point, and this should open up additional opportunities to gain further flexibility and improve processes.

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 at cloudera.com

Connect with Cloudera
About Cloudera:
cloudera.com/more/about.html

Join the Cloudera Community:
community.cloudera.com

Read about our customers' successes:
cloudera.com/more/customers.html

The Reasons Why Cloudera is Superior in the Kafka Space

In this paper, we described how Cloudera delivers the best Kafka ecosystem by not only covering the streams messaging aspects of securing, monitoring, balancing, and replicating large scale Kafka environments but also by incorporating best-in-class flow management and stream processing and analytics engines to ensure a sustainable, scalable and adaptable end-to-end streaming architecture.

The key reasons why Cloudera delivers the best Kafka ecosystem in the industry:

- **Commitment to the open source community**—Cloudera is dedicated to the Kafka ecosystem and continues to be actively involved with the Kafka open source community through deep engineering relationships with other Kafka committers. This relationship has led to critical innovations and product improvements, many of which have been described here.
- **Respect for the voice of the customer**—Cloudera is constantly listening to feedback from our customers on what they are asking from us and our products. This is evident in the innovations and product enhancements we have been delivering over the years. Case in point is Streams Messaging Manager, which was primarily created as a response to what our Kafka customers most needed—to cure their Kafka blindness.
- **Kafka innovations**—Kafka is just one of the open source projects we are committed to across our platform. But, Kafka is a very key part of our overall streaming offering and so we have been delivering disruptive innovations in this space for our customers. Other than SMM, Streams Replication Manager is another example of best-in-class engineering and innovation. This is based on Mirrormaker 2, which is a much needed innovation that our engineers delivered for the open source community.
- **Security and governance are table stakes**—Cloudera's SDX is a true differentiator for us when compared to other vendor products. The promise of a unified security and governance layer across all components and environments is what our customers truly want. Understanding end-to-end lineage of your streaming data from the edge to the cloud across your ingestion, messaging, and stream processing components is made possible with SDX. This is super critical for companies that are struggling with compliance and regulations.
- **Multi-cloud and hybrid cloud support**—CDP is the world's first enterprise data cloud and thus we are able to help our customers support streaming architectures that must retain an on-premises footprint but also need to leverage the cost efficiencies of public cloud providers. Cloudera's streaming platform components can be quickly provisioned into your private or public cloud while leveraging the unified data security, governance, lineage, and control provided through SDX.
- **Global customer support**—Cloudera has hundreds of customers running sophisticated Kafka deployments across tens of thousands of broker nodes. We have enabled our customers to implement true end-to-end streaming architectures across multiple industry verticals. Beyond software support, our global team of knowledgeable professional services and training staff make it a breeze for our customers to implement their streaming architectures.
- **One platform**—Cloudera has a true edge-to-cloud streaming data platform like none other in the industry. Instead of adopting Kafka alone as a point solution, we address the data management challenges of the enterprise across all aspects of the data-in-motion journey with one integrated platform.

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

Learn more about Cloudera DataFlow at cloudera.com/cdf

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