CLOUDERA

Extend Your On-premises Streaming to the Public Cloud

Go hybrid. Make it easy.

Table of Contents

Streaming Analytics is Critical to Business Success

Global digitization has resulted in a vast array of new products and services with such high levels of convenience that it fuels a continuous loop of greater expectations for immediacy of data insights. This means that business opportunities that directly impact revenue or boost operational efficiency need to be addressed in near real-time.

Great value comes from comparing "right now" with "the recent past" and "a full history of events." That is what enables enterprises to analyze past behaviors to predict events, gain insights faster, and immediately address opportunities or potential risks.

Seamless integration and analysis of real-time events from the edge to the enterprise, in conjunction with stored and operational data across on-premises and public cloud data environments, is critical to business success. That is the approach that enables organizations across a variety of industries to make the most of their data, including:

- **Manufacturers** that keep humming with predictive maintenance
- **Retailers** that restock inventory or replenish proactively
- **Utilities** that prevent power outages
- **Telecoms** that deliver continuous Quality of Service
- **Financial Services** companies that reduce cyber threats

"Many leading enterprises realize that real-time analytics – the analytics of the present – is an incredible competitive advantage because they can act now to serve fickle customers, fix operational problems, power internet-of-things (IoT) apps, and respond decisively to competitors."

Sridharan, S. "The Best of Times as a Data and Analytics Leader." Forrester, The Insights Beat, Oct. 11, 2019.

Streaming Data Challenges

Your streaming analytics implementation is only as good as your ability to acquire and analyze the real-time data you have captured.

Enterprises are continuously inundated with streaming data that hinders their ability to capture, process, and harness such data in real time due to challenges such as:

- **Data volume**—The sheer volume of data streams across multiple disparate data sources and targets is a significant challenge for traditional enterprise tools to capture, ingest, and process. Furthermore, it is critical for enterprises to manage such data streams in real-time so as to not let the data become stale.
- **Data velocity**—Organizations often find it difficult to turn large volumes of continuous high-velocity event flows into actionable insights in real-time. Traditional batch ETL processes cannot process and analyze billions of events per second and serve the data in real-time to stream processing engines.
- **Data variety**—IoT implementations and other streaming initiatives bring forth a wide range of devices, applications, data structures, and formats to deal with. Traditional systems struggle to manage the dozens of data formats across multiple vendor products as well as with the types of data, being either structured, unstructured, or semi-structured.
- **Data security and governance**—This topic is top of mind for executives, board members, regulators, and customers alike, with data governance rising to the top spot of Chief Audit Executive's concerns, replacing cybersecurity preparedness.[1]

Siloed systems that, while perhaps providing best-in-class point solutions, don't integrate well nor provide holistic data governance and security. This opens organizations up to security risks from cyber intrusion and internal threats, regulatory penalties for breaking compliance rules, and reputational risk.

Data Inundation[2]

41.5B

More than 41.5 billion IoT devices will be active by 2025.

79ZB

Billions of IoT devices will create 79ZB of data by 2025, causing organizations to reevaluate their data governance, retention, and usage policies.

40%

40% of manufacturers will use field asset IoT data to intelligently diagnose issues and resolve autonomously, improving unplanned downtime by 25%.

Cloudera DataFlow Provides the Complete Streaming Platform

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

With regard to addressing challenges with streaming data, the diagram to the right illustrates how CDF supports the entire set of streaming data functions. From data capture and flow management at the edge (1) to provisioning that data directly to/from your messaging backbone (2) and/or stream processing and analytics (3).

CDF addresses the key challenges of high volume, high velocity, and extreme variety of streaming data. With an edge-to-cloud comprehensive set of streaming capabilities, CDF is a natural fit to any enterprise embarking on a streaming analytics or IoT implementation.

To address data security and governance risks, CDF 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 hybrid on-premises and public cloud environments (5).

Holistic Streaming Platform of Integrated Components

Cloudera DataFlow meets all the management, security, and governance challenges that enterprises face with streaming data and analytics. It helps them deliver a better customer experience, boost operational efficiency, and stay ahead of the competition across all strategic digital initiatives:

1. **Edge & Flow Management** capabilities enable enterprises to orchestrate the acquisition, enrichment, transformation, and routing of large volumes of any type of data from edge to cloud in a secure and governed manner. Powered by Apache MiNiFi and NiFi, the no-code approach boosts developer productivity and time-to-market.

2. **Streams Messaging** enables enterprises to ingest, buffer, and scale massive volumes of real-time data to serve on-premises and cloud applications. Primarily powered by Apache Kafka, the Streams Messaging capabilities enable real-time data access to a wide-range of applications such as analytic engines, data lakes, time series databases, etc.

3. **Stream Processing and Analytics** employs the latest generation of low-latency stream processing and analytics engines that addresses the requirements of real-time insights and predictive analytics. Headlined by Apache Flink, it helps to democratize streaming analytics across the enterprise to deliver better business outcomes, faster.

4. **Shared Data Experience (SDX)** is the key differentiator from other platform providers. It is what enables the seamless integration of all parts of your streaming ecosystem. It provides a safe, efficient, and consistent experience of deployments and migrations across all data environments: on-premises, hybrid, or multi-cloud.

5. CDP is **the world's first enterprise data cloud** and supports running a variety of analytic workloads seamlessly across on-premises, multi-cloud, and hybrid cloud environments. With CDP's Data Hub framework, the CDF streaming components can be quickly provisioned as a cluster into a public cloud environment within minutes. This cluster can take advantage of CDP's SDX capabilities for seamless security and governance.

Growth of Hybrid and Multi-cloud Architectures

Enterprises have been forced to deal with an exponential increase in data and have done so by adopting cloud as an easy option for storage and compute. This is because of the economic benefits and flexibility that comes from easily spinning up infrastructure as a service (IaaS) on the cloud.

While many have enjoyed the fruits of cloud success, that experience has often been due to the simplicity of utilizing one cloud. But over time, things have gotten more complex as more data infrastructures got spun up in multiple locations within an enterprise.

Those with the responsibility of securing and managing data across the enterprise struggle with the fact that data resides in multiple places: on-premises, in their private cloud, and public cloud environments. IT organizations adopted hybrid architectures accepting the fact that data has to reside on-premises and in the cloud simultaneously. This made sense from a data storage perspective, but digital transformation initiatives became harder to implement due to the lack of effective tooling to bridge both worlds in a holistic way.

Over time, companies had more cloud vendors to choose from as a result of price wars between IaaS providers. But, migrating data from one cloud to another was also not easy. Moreover, as companies merged with or acquired other companies, a new data management challenge arose. The parent entity's IT organization now needed to manage data silos spread across multiple cloud vendors (multi-cloud) while being saddled with incompatible data management tools.

For the foreseeable future, hybrid and multi-cloud architectures are here to stay. In order for enterprises to adapt and leverage such architectures, they must rely on data management vendors to provide the proper tooling in a cloud-agnostic or cloud-neutral manner.

The Future is Hybrid – for Most[3]

58%

58% of enterprises moving toward a hybrid environment.

42%

42% of enterprises considering or developing a hybrid IT strategy.

34%

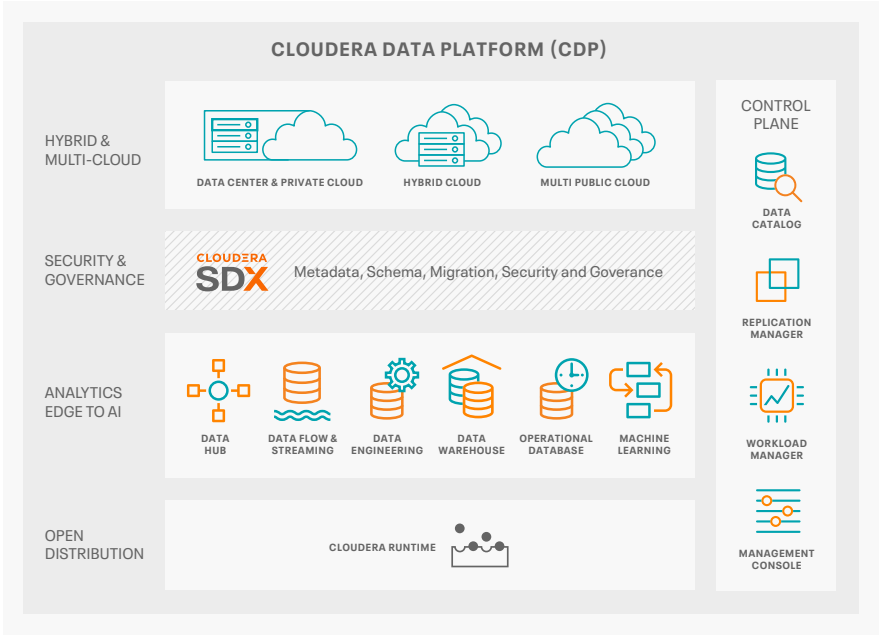
34% of enterprises actively executing a hybrid IT strategy.

Leverage the Unified Experience

Optimized for hybrid and multi-cloud, CDP seamlessly delivers the same data management and analytic capabilities across on-premises and public clouds.

Cloudera Data Platform (CDP) is a data cloud built for the enterprise. With CDP, businesses manage and secure the end-to-end data lifecycle —collecting, enriching, analyzing, experimenting, and predicting with their data—to drive actionable insights and data-driven decision making. Optimized for hybrid cloud, CDP delivers the same data management and analytic capabilities seamlessly across private and public clouds. CDP separates data management from infrastructure strategy, enabling companies to move data and apps from one environment to another without rewriting applications and retraining personnel.

As illustrated to the right, CDP can handle any type of workload (e.g.: data streaming, engineering, warehousing, time series, or machine learning) regardless of location.



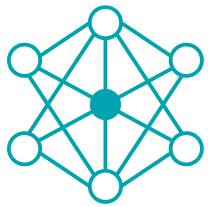
The Advantages of CDP—The World's First Enterprise Data Cloud

- **Hybrid & multi-cloud**—Operates across all major public clouds and the private cloud with a public cloud experience everywhere.
- **Multi-function analytics**—Integrates data management and analytic experiences across the data lifecycle for data anywhere.
- **Secure & governed**—Delivers security, compliance, migration, and metadata management across all environments.
- **Open platform**—Open source, open integrations, extensible and open to multiple data stores and compute architectures.

Challenges of Streaming Platforms in a Hybrid World

Most enterprises have divvied up their data platforms across environments that include on-premises and public cloud data centers.

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 but also need to leverage the flexibility, economic and performance benefits of the cloud. But, more importantly, they struggle with finding a streaming data platform that can span across a hybrid environment seamlessly. Without one, those enterprises face challenges such as:

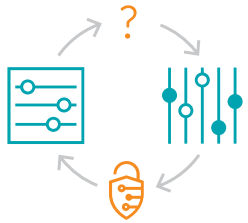


Complexity

Cloud configuration and infrastructure setups are complex and time consuming. While many cloud vendors provide some tools for developers to manage streaming data, few provide a single, simple approach for easy configuration, compute, and storage to extend their on-premises streaming platform to the cloud.

Incompatibility

Data streaming tools and apps are either built to work on-premises or in the cloud but not both. Organizations are therefore challenged to create holistic business solutions with data tools that are disjointed and incompatible, which results in precious time getting lost in stitching them together.



Risking Governance and Security

The fragmented and completely disconnected nature of hybrid environments makes it more challenging to consistently enforce strong data security and governance on data streams across diverse environments. The result is an overall lack of comprehensive authentication and policy control.

Extend Data Streaming to the Public Cloud with CDP Data Hub

Cloudera Data Hub is a cloud-native service powered by a suite of integrated open source technologies that delivers the widest range of analytical workloads with comprehensive security, governance, and control.

CDP addresses the challenges of streaming data across hybrid environments. As a key service, CDP Data Hub enables you to extend the same on-premises streaming experience of Cloudera DataFlow to the cloud.

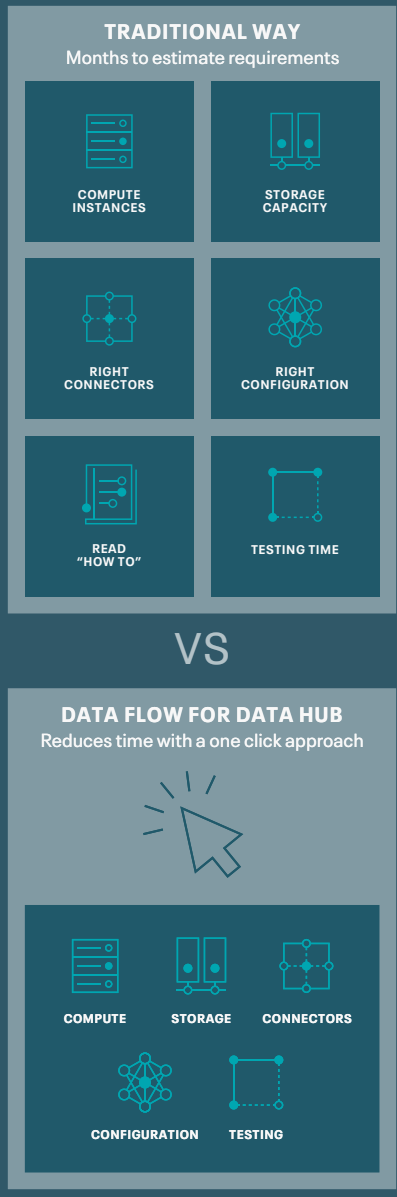
With a "one-click" approach, the CDP Data Hub interface accelerates deployment of flow management, streams messaging and streaming analytics clusters to the cloud in just minutes, without the heavy burden of generating infrastructure requirements to develop and configure them.

Cloudera Data Hub lets you spin such streaming component clusters by using cluster definitions. Cluster definitions are like templates that let you choose from a list of pre-defined cluster sizes aligned with a public cloud option—all in one step. Your team will spend more time designing data flows and less time setting up cloud infrastructure.

Given that CDP extends CDF's streaming capabilities to the cloud, there are no concerns about disconnected or incompatible streaming platforms across environments. Enterprises can now use the same comprehensive capabilities of CDF across both on-premises and their public cloud deployments. This enables easy portability and exchange of data flow assets across such environments very easily.

Finally, with Cloudera SDX, you alleviate data security and governance concerns because control policies are set once and consistently enforced across all components and across all environments to provide a unified authentication process for all users and end-to-end data governance for all the data streaming through the platform.

A New Approach



What is DataFlow for Data Hub?

DataFlow for Data Hub makes hybrid use cases possible by extending on-premises flow management, streams messaging, and stream processing and analytics capabilities to the public cloud.

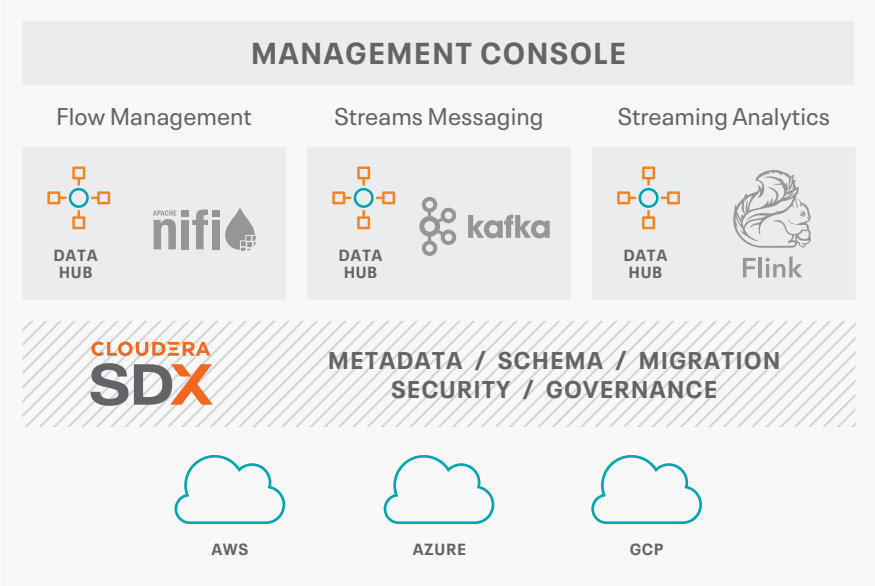
Flow Management for Data Hub spins up NiFi and NiFi Registry into your public cloud cluster.

While NiFi orchestrates data collection, distribution and transformation, NiFi Registry keeps those data flows versioned and synchronized. This mitigates functional gaps that often occur when migrating on-premises workloads to the cloud or integrating use cases across hybrid environments.

Streams Messaging for Data Hub extends your on-premises Apache Kafka investment by spinning up a second Kafka cluster into the public cloud along with Schema Registry and Streams Messaging Manager.

On-premises only solutions are often not feasible due to the sheer volume and processing requirements of streaming analytics use cases. With bi-directional replication across hybrid environments, apps can consume cloud data or burst to the cloud as needed and without interruption.

Streaming Analytics for Data Hub spins up Apache Flink and its related components into the public cloud, bringing stream processing of real-time data into the hybrid world. With Data Hub enabling NiFi and Kafka to be on the cloud, Flink can easily process the data from either of those streaming components in real-time by being close to them on the cloud as well.



Benefits of DataFlow on Data Hub

CDF provides a variety of benefits for enterprises, enabling you to:

Reduce data integration development time with a no-code approach to building complex data pipelines with minimal effort. Flow Management for Data Hub offers a simple visual UI for building sophisticated data ingestion, transformation, and enrichment requirements across a variety of streaming data sources and targets. This enables you to quickly ingest data from devices, enterprises applications, partner systems, and cloud applications that generate real-time streaming data.

Accelerate streams messaging across your cloud clusters and ensure that all ingested data is buffered in a transient state from where each enterprise application consumes only the data that it needs. Streams Messaging for Data Hub's advanced messaging and processing capabilities enable enterprises to effectively scale data streams.

Democratize real-time insights with streaming analytics to enable quick and actionable intelligence for your business decisions. Streaming Analytics for Data Hub helps businesses to democratize streaming analytics across the firm and improves detection and response to critical events that deliver better business outcomes.

Manage and secure your data from edge to cloud inclusive of high volume data collection from any streaming source. The cohesiveness of the different DataFlow clusters allows you to setup a very distributed data collection architecture that can span from the edge to the cloud. The tight integration with SDX gives DataFlow the unique advantage of seamless security and governance across all your data-in-motion and data-at-rest locations.

Spin up DataFlow clusters in mere minutes using the different cluster definitions that are available within Data Hub. This allows you to launch NiFi, Kafka, or Flink clusters very quickly in a public cloud environment for a specific workload or for other long-running tasks too. Easily deploy your on-premises flows or other streaming assets to the cloud with this model.

> "Cloudera helped our organization get to the next level by providing us with a streaming data platform, which provides us with real-time data."
>
> Martijn Groen, IT lead of the Data Lake, Rabobank

Adopt the New Approach Today

Take the next step towards modernizing your enterprise data ecosystem by bridging your on-premises data streaming architecture to the cloud and building a foundation for the next generation of business opportunities.

Why Cloudera

- Cloudera delivers powerful self-service analytics across hybrid and multi-cloud environments.
- CDP provides end-to-end data lifecycle integration, including data streaming, data engineering, data warehouse, and machine learning to help businesses improve productivity and continue their transformation journey of being data-driven organizations.
- Cloudera's Shared Data Experience (SDX) ensures consistent data security, governance, and control across the data lifecycle and all environments while mitigating risk and costs.
- By setting the foundation of 100% open source, Cloudera helps businesses accelerate innovation, prevent vendor lock-in, and deliver future-proof software.
- Cloudera DataFlow is the industry's most comprehensive streaming data platform for your edge-to-cloud use cases. It has only become more advanced with the availability of these capabilities within CDP.

Learn More

Cloudera DataFlow on CDP

Watch the video

Contact an expert

Sources

[1] Gartner.com, "Gartner Says Data and Cyber-Related Risks Remain Top Worries for Audit Executives". November 7, 2019.

[2] Carrie MacGillivray and David Reinsel, "Worldwide Global DataSphere IoT Device and Data Forecast, 2019–2023," ***IDC***, May 2019, Market Forecast Doc #US45066919, https://www.idc.com/getdoc.jsp?containerId=US45066919; Carrie MacGillivray et al. "IDC FutureScape: Worldwide IoT 2020 Predictions," ***IDC***, Oct 2019, IDC FutureScape Doc #US45591819, https://www.idc.com/getdoc.jsp?containerId=US45591819.

[3] 451 Research, "Voice of the Enterprise— Cloud, Hosting & Managed Services, Vendor Evaluations", 2019.

CLOUDERA