First Look

Data in Motion: Simplify Your Approach to Streaming Data with IBM and Cloudera DataFlow (CDF)

Date: July 2022 **Author:** Tony Palmer, Principal Validation Analyst

Hybrid Cloud Challenges:[1]

47% The percentage of organizations that believe *providing better/more differentiated customer experiences* is **one of their most important** digital transformation objectives.

29% The percentage of organizations that consider *increasing the capability to handle data where it is generated* to be **an important priority** in support of new and/or ongoing data initiatives.[2]

A growing number of businesses are implementing data streaming to drive value. 2022 saw an 83% increase in the number of organizations telling ESG that collecting and analyzing data from a multitude of internet of things (IoT) devices would be the most important initiatives for their organization. Traditionally, this was complex and required many tools. Nearly a third of organizations (31%) report that their organizations have a problematic shortage of skills in data analytics, making it even more of a challenge.
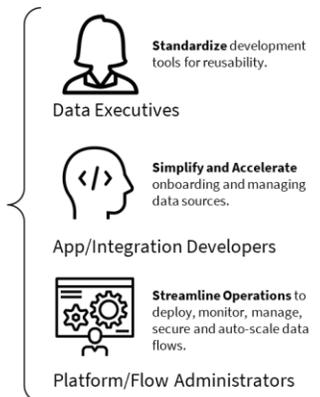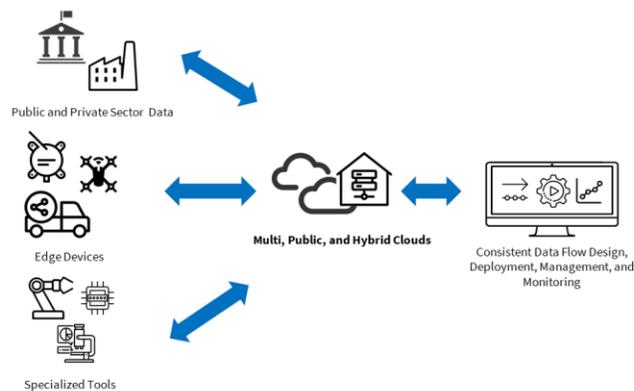
The volume, variety, and complexity of data continue to increase and data is increasingly distributed across heterogenous environments. To simplify this evolving data pipeline, flow management, stream processing, and analytics need to be unified. It is critical for businesses to understand their data and be able to act on it even before it reaches its destination. Real-time analytics provides insights organizations act on to make better decisions when it matters most. The more quickly and reliably data is processed and analyzed, the more likely organizations are to be able to optimize business outcomes.

Cloudera Data in Motion with IBM

IBM and Cloudera have been partnering since 2019. Their current solutions include on-premises and public cloud offerings with data management, data science, security**,** and governance to build and offer an enterprise-class data, AI, and analytics solution. Organizations can choose flexible hybrid cloud options, across multiple public clouds and on-premises deployments for Cloudera with IBM providing a single point of support through their OEM agreement with Cloudera. Joint IBM and Cloudera customers can benefit from easy-to-use, self-service tooling and access industry and technology experts globally to address a variety of use cases and speed production at scale. Cloudera DataFlow is designed as a scalable, real-time streaming data platform built for simpler collection, curation, and analysis of data to provide customers with key insights for immediate actionable intelligence. IBM and Cloudera DataFlow help organizations manage data from edge to cloud with both no-code and code-based approaches to developing sophisticated streaming applications. Additional benefits include: simplification of streaming in any-to-any environments; highly scalable data ingestion, transformation, and

---

[1] Source: ESG Research Report, *2022 Technology Spending Intentions Survey,* November, 2021. Unless otherwise stated, all ESG research references and charts in this first look have been taken from this research report.
[2] Source: ESG Complete Survey Results, *2022 Technology Spending Intentions Survey*, November 2021.

management; accelerated onboarding for data and AI, application development, and administration teams; and edge-to-cloud streaming data across on-premises, public cloud, and hybrid cloud environments.

Cloudera DataFlow (CDF) Use Cases

- **Customer Experience Optimization—** Ingest, transform, and combine customer data from multiple sources into a single 360-degree view of the customer and use the insight to optimize the next best actions.
- **Cybersecurity Optimization** — Optimize log analytics solutions with CDF by simplifying log ingestion from the edge, and leverage streaming analytics insights for fraud detection and network threat analysis.
- **Operational Efficiency —** Streamline operations with predictive and preventive maintenance, asset tracking, fleet monitoring, and quality control/improvement processes to optimize business decisions and gain competitive advantage.
- **Enterprise Data Management and Movement** — Manage massive volumes of high-velocity data to and from legacy systems, ETL tools, and other data stores while optimizing resource utilization by moving data between data centers or between on-premises and cloud infrastructures.


Cloudera Flow Management (CFM)

Cloudera Flow Management provides a centralized data movement tool using no-code flow design to manage end-to-end data flows, from many sources to many destinations with built-in security and governance. CFM supports over 400 pre-built processors and is designed to make it easy to build custom processors. Apache NiFi Registry is leveraged to provide a comprehensive flow development lifecycle. In addition, users can easily drill down for a closer look at individual processors. This means that CFM can deliver highly scalable data movement, transformation, and management capabilities to the enterprise.
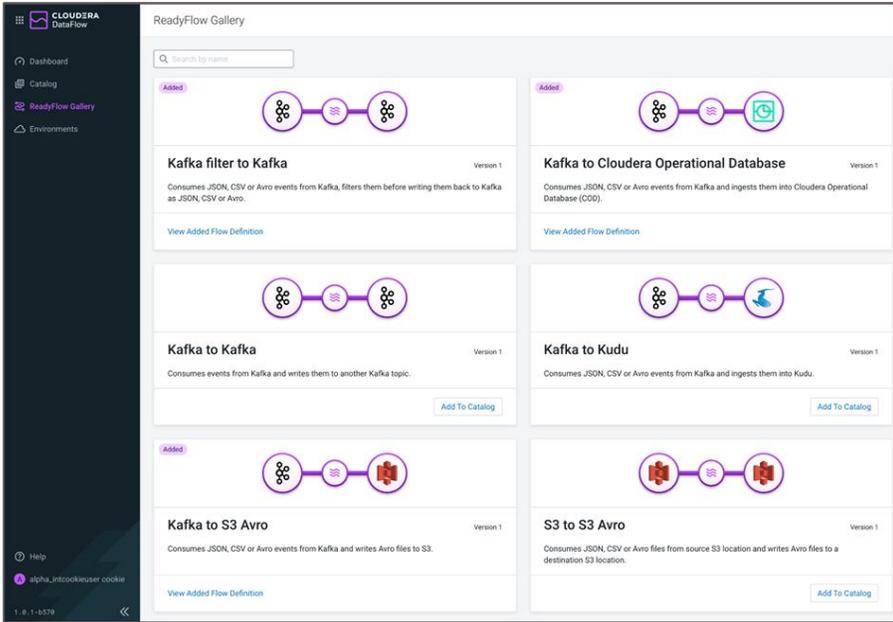
The CDF ReadyFlow Gallery is designed to enable deployment of data flows using a continuously updated library of pre-built flow templates for the most common data movement use cases that allow for fast customization and deployment. Instead of spending time on building data flows in NiFi, practitioners can focus on deploying flows and defining the right KPIs for easy monitoring.

Cloudera Stream Processing

Next, ESG looked at SQL Stream Builder, a service in Cloudera DataFlow Stream Processing. SQL Stream Builder enables developers, analysts, and data scientists to write streaming applications using industry-standard SQL. It also offers an interactive experience designed to make the development process faster, simplifying access to continuous streaming data within Kafka and Flink, syncing data with a variety of destinations like Db2, improving productivity.
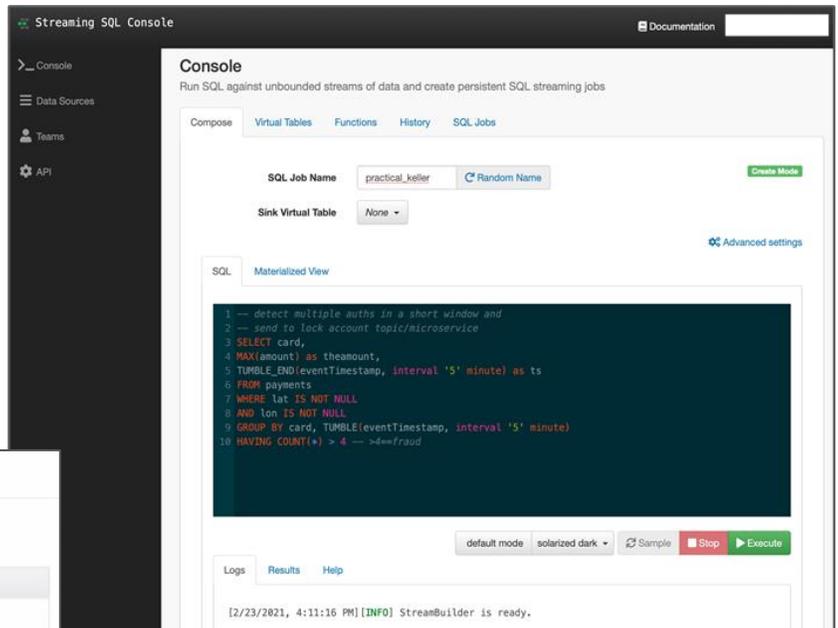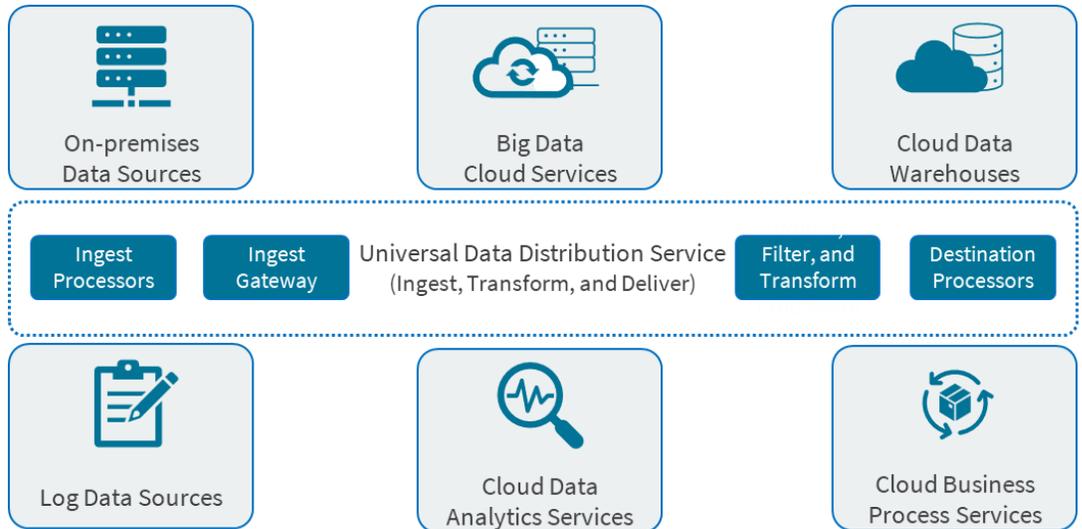

Cloudera makes it easier for organizations to process data streams at scale and to deliver real-time analytical insights about processed data with streaming applications.

As seen here, CDF makes connecting to data providers fast and easy as well, since adding connectors takes just a couple of clicks.
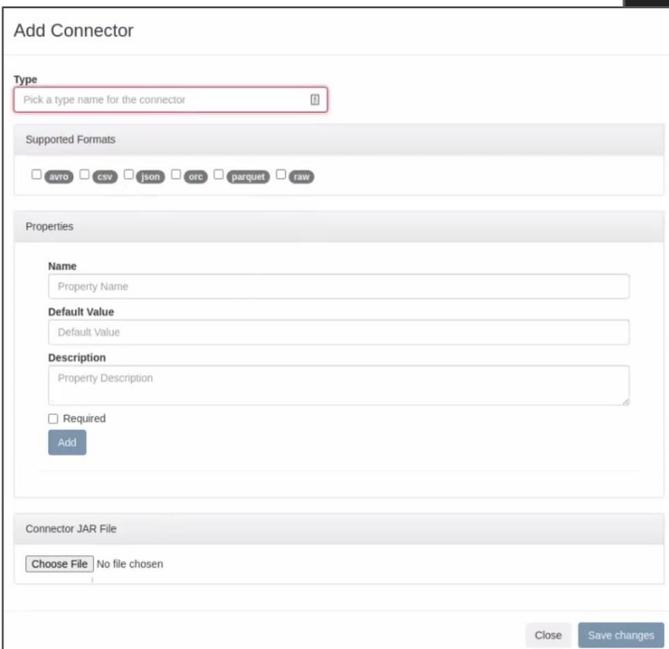
Universal Data Distribution Service

Finally, ESG looked at how IBM and Cloudera make it possible to connect to any data source anywhere, process the data, and deliver it to any destination, making it extremely well-suited to use as part of a data fabric.

On-premises Data Sources

Big Data Cloud Services

Cloud Data Warehouses

Ingest Processors | Ingest Gateway | Universal Data Distribution Service (Ingest, Transform, and Deliver) | Filter, and Transform | Destination Processors

Log Data Sources

Cloud Data Analytics Services

Cloud Business Process Services

First Impressions

Business demands for harnessing the power of data and AI to drive better outcomes are higher than ever. ESG observed an 83% increase in businesses' interest in collecting and analyzing IoT data in 2022 and yet, about one-third of respondents report a data analytics skill shortage as one of the impediments in deriving value from streaming data. Responding to these challenges, Cloudera and IBM are simplifying the businesses' approach to streaming data by bringing together users and stakeholders with a diverse set of no-code and code-based tools in a multi-cloud unified environment.

ESG validated that the integration of Cloudera DataFlow (CDF) with the Cloudera Data Platform (CDP) leverages 100% open source-based engineering and ensures a sustainable, scalable, and adaptable end-to-end streaming architecture.

Cloudera unifies data flow management, streaming messaging, and streaming analytics in data in motion and integrates with data at rest. It takes advantage of the end-to-end lifecycle to predict and optimize outcomes in diverse use cases such as elevator maintenance using sensors, detecting and scoring fraud in credit card and similar transactions, or cybersecurity defense using security information and event management (SIEM).

Organizations can use Cloudera Dataflow, together with IBM, as part of their data analytics and AI toolsets to improve customer experience, increase operational efficiency, and drive innovations while mitigating security and compliance risks. The IBM-Cloudera OEM agreement is in place so that joint customers can enjoy a one-stop-shop at IBM to procure, license, and deploy Cloudera solutions, while exploiting IBM's scale and deep expertise in data and AI.