HARNESSING DATA IN MOTION

How Cloudera DataFlow supports government data movement and processing

As the volume, velocity, and variety of data swells across federal, state, and local government, many agencies struggle to acquire maximum value from an ever-expanding number of sources. For all government agencies, modernizing data movement—ingestion, routing, management, and delivery—in particular, represents a complex and costly challenge.

A simple, flexible, open-ended solution is required to streamline access to both structured and unstructured data from sources from the edge and across the enterprise so agencies can use their legacy and modernized systems to take advantage of the new insights available.

5000x

The Kentucky Transportation Cabinet reduced weather related IoT data ingestion and availability for analysis from 2 days to 35 seconds.

The Data-Flow Challenge

In the effort to leverage data for their missions, some agencies find themselves bogged down by manual data processing tasks, spending time and money writing scripts to move data from its source to the various points of analysis. Furthermore, making data available to a variety of applications adds complexity as each may have specific requirements. It's an inefficient use of resources to have skilled data scientists and talented IT experts mired in such labor-intensive efforts.

Agencies that have tapped the power of the cloud often end up constrained when it comes to data movement, trapped by vendor lock when they seek to move data from one public cloud provider to another. Even agencies with a strong core capability around analytics find it difficult to lift and shift data from disparate sources, across legacy systems and into various cloud services. To support rapid, meaningful analysis, they need an enabling technology that makes it easier to flow data across systems without requiring major overhauls. Most would prefer to stay with systems and processes that are familiar to their users and that support their needs.

Cloudera DataFlow

Cloudera DataFlow (CDF), a scalable, real-time streaming data platform, ingests, encrypts, curates, and analyzes data for key insights and immediate actionable intelligence. For agencies that rely on a vast array of data feeding downstream applications to drive missions, CDF offers a way forward by addressing the government's key challenges. For data-rich agencies that draw information from a wide range of feeds, CDF offers real-time processing of data streaming at high volume.
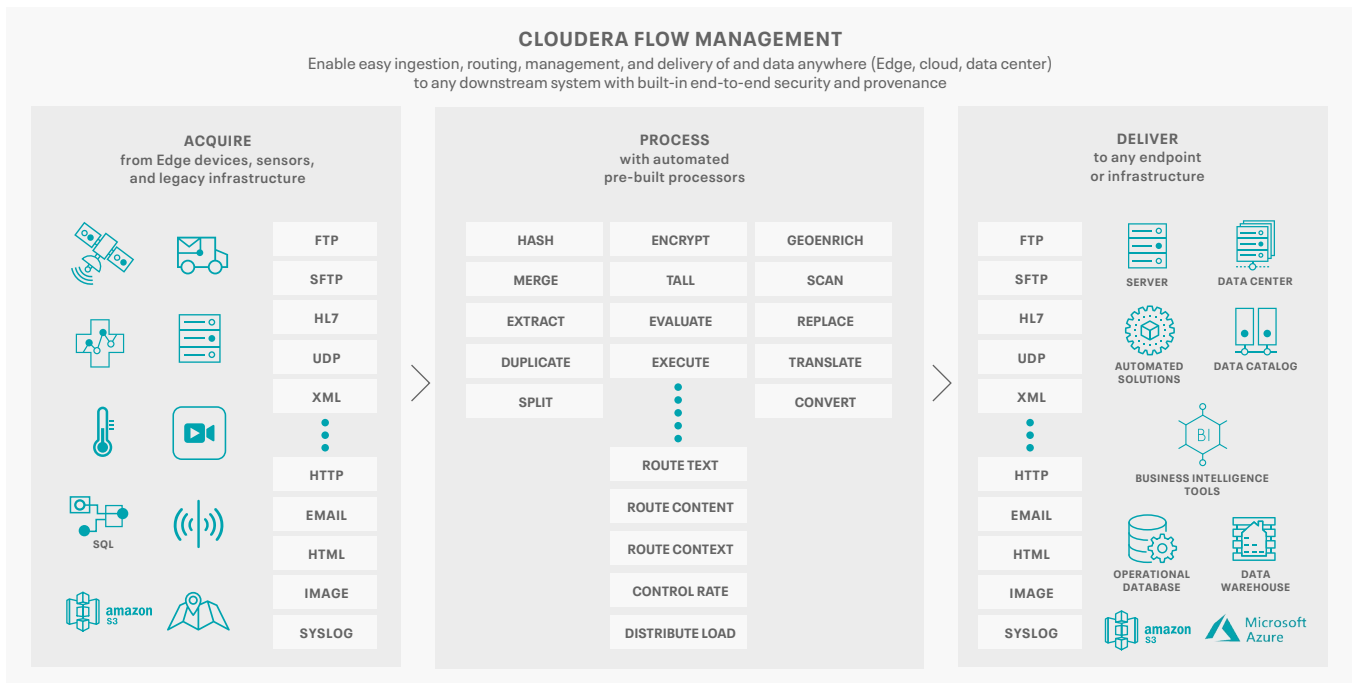
330M

Citizens worth of data
processed by the
U.S. Census Bureau

"The 2020 census is a massive undertaking to ensure
that everyone living in the United States is counted.
The Enterprise Data Lake and Cloudera DataFlow
enables us to efficiently process the vast amounts of
data we anticipate generating in 2020 via massively
parallel processing and taking advantage of cloud-
based infrastructure."

Kevin Smith, Chief Information Officer, U.S. Census Bureau

At the same time, CDF addresses a critical government need around accountability. The
system encrypts data in motion and tracks and retains provenance of streaming data, ensuring
agencies can meet their regulatory requirements around transparency, accountability and
auditability.

Government agencies require a seamless process to take in data and get it to its destination.
CDF manages and monitors edge applications and streaming sources, enabling easy ingestion,
routing, and management of any type of data from thousands of internet-of-things devices and
other sources. Audio, video, Lidar, temperature sensors—whatever the source, CDF can ingest
and enrich that data and deliver it to any downstream system for further processing and
comprehensive analysis.

CLOUDERA FLOW MANAGEMENT

Enable easy ingestion, routing, management, and delivery of and data anywhere (Edge, cloud, data center)
to any downstream system with built-in end-to-end security and provenance

For agencies seeking to lighten the workload on valuable IT and data science personnel,
CDF eliminates custom scripting by offering of over 300 pre-built processes. And it is readily
scalable, capable of processing trillions of events per second and moving petabytes of data
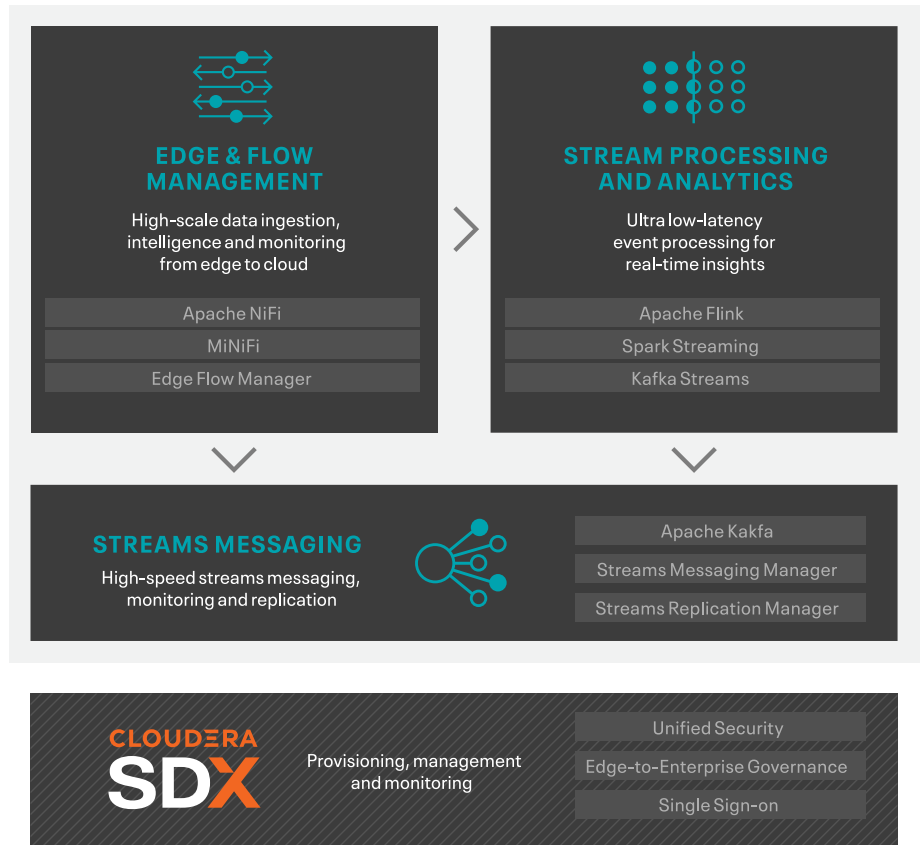quickly.

CLOUDERA

Cloudera offers the US Government:

- **Connected Platform:** Rigorously tested to ensure integration, security, quality and ease-of-use. Includes an enterprise-grade management console.

- **U.S. Government Support:** Fully cleared U.S. citizens working in data centers on U.S. soil with extensive knowledge on government best practices, available 24x7 with experts on every component of our stack.

- **Professional Services:** Hands-on installation, integration, optimization, upgrade, patching and troubleshooting.

- **Training:** Expert training on each component of Cloudera, offering agencies experience and certification. Available on-site in set periods of time, IT staff can train at their own pace on-demand online.

An Open Approach

A data-in-motion solution should deliver maximum flexibility that includes the ability to work across legacy and modernized platforms and adapt to future technological enhancements and innovations.

CDF is built on Apache NiFi, an open source infrastructure originally developed by the National Security Agency to automate the management, encryption, and processing of large data streams in real time. This core component of CDF has been tested in a range of government use cases, including critical applications in the Department of Defense, as well as in civilian, state, and local governmental entities.

EDGE & FLOW MANAGEMENT
High-scale data ingestion, intelligence and monitoring from edge to cloud

Apache NiFi
MiNiFi
Edge Flow Manager

STREAM PROCESSING AND ANALYTICS
Ultra low-latency event processing for real-time insights

Apache Flink
Spark Streaming
Kafka Streams

STREAMS MESSAGING
High-speed streams messaging, monitoring and replication

Apache Kakfa
Streams Messaging Manager
Streams Replication Manager

CLOUDERA SDX
Provisioning, management and monitoring

Unified Security
Edge-to-Enterprise Governance
Single Sign-on

Hundreds of government agencies spanning

40+

countries rely on Cloudera

Entirely open source, the CDF ingestion platform is secure enough for government users thanks to rigorous testing, regular improvements, and support provided by Cloudera. It fast-tracks the integration of data by ingesting any type of data from any end point, processing it in real time, and delivering it to any downstream system.

CDF enables high-volume data users to optimize resources by moving data between data centers or between on-premises infrastructure and clouds. Agencies can reduce their log ingestion costs and avoid unnecessary Splunk costs: By filtering data through CDF, only truly necessary data is sent through into Splunk for further exploration, analysis, and visualization.

A Bridge to the Future

While CDF is part of the more comprehensive Cloudera Data Platform, it also works as a stand-alone solution for agencies that want to continue using their already-familiar analytic stores, applications, and storage solutions. And for organizations with an eye on modernization, CDF can channel data in parallel into both legacy and modern infrastructures, enabling government users to gain maximum value from existing systems while allowing themselves a path toward future capabilities.

The net result: For federal, state, and local agencies, as well as other data-intensive government entities, CDF delivers key insights with streaming analytics. It accelerates ROI on big data by analyzing streaming data for patterns, comparing the outputs with machine learning models, and delivering actionable intelligence.

Government needs a simplified approach to building complex data pipelines. CDF's visual interface makes it easy for users to construct sophisticated data flows for data ingestion, transformation, and enrichment from a variety of streaming sources.

Learn how Cloudera DataFlow helps agencies manage data from edge to cloud at cloudera.com.

A Bridge to the Future