CLOUDERA DATAFLOW FOR DATA HUB

KEY FEATURES AND CAPABILITIES

Flow Management for Data Hub

- Based on the latest open-source version of Apache NiFi
- Build data streams quickly by leveraging pre-created NiFi access policies and controller services
- Robust and reliable through automated replacement of failed NiFi nodes

Streams Messaging for Data Hub

- Based on the latest open-source version of Apache Kafka
- Streams Messaging Manager for management and monitoring of your Kafka clusters
- Schema Registry for central schema management
- Robust and reliable through automated replacement of failed brokers

Streaming Analytics for Data Hub

- Based on the latest open-source version of Apache Flink
- Designed for low-latency, stateful stream processing
- Execute complex event processing and other windowing functions

Cloudera SDX

- Unified security and governance across all DataFlow clusters on Data Hub with SDX
- All clusters are integrated with CDP's central user management system to allow seamless single-sign-on
- Central permission management for Kafka topics and NiFi policies through integrated Apache Ranger
- Pre-configured Apache Atlas integration to ensure data lineage is automatically captured for all data stream

Cloudera DataFlow for Data Hub comprises of three key services - **Flow Management for Data Hub, Streams Messaging for Data Hub, and Streaming Analytics for Data Hub**. Together, they offer a comprehensive streaming data platform for the public cloud. They address the key data management challenges with streaming and IoT data for all types of enterprises. They can modernize data streaming services by delivering real-time business use cases, provide actionable intelligence and ensure comprehensive security, governance, and control.

The Pitfalls of Data Streaming in Hybrid Cloud

Hybrid cloud is a bridge between on-premises and public cloud infrastructure management for analytics and data streaming applications. Getting hybrid cloud models to work for real-time use cases can be challenging for IT administrators and developers.

KEY CHALLENGES
COMPLEXITY
NOT COMPATIBLE
RISKING COMPLIANCE AND SECURITY

Data Streaming with Hybrid Cloud - Three Challenges

Complexity of Cloud Configuration and Infrastructure Setups

Many cloud vendors provide a plethora of tools for developers to stream data, but few can offer a simple way to configure and size compute and storage resources. To design a data streaming solution for the hybrid cloud, IT administrators have to estimate the requirements for storage capacity, choose the right compute instances, understand the network demands, and the entire stack's configuration before deployment is possible. These efforts require research, testing and validation, making the whole process very time-consuming.

Different Streaming Applications Cannot Work Across Hybrid Environments

The integration of different data streams and other applications across hybrid cloud is difficult because some are built for on-premises and some for the public cloud. The inconsistent tooling results in disjointed and incompatible products leading to extra development time to stitch the hybrid environments together.

Risking Compliance and Security

The fragmented and completely disconnected nature of the hybrid environments makes it more challenging to consistently enforce strong data security and governance across Flow Management, Streams Messaging, and Streaming Analytics solutions. This inconsistency opens the door for potential cyber threats and unnecessary penalties for violating data compliance.

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

The New Approach

Cloudera DataFlow for Data Hub is the industry's only modern, flexible, and scalable data streaming solution built purposely for hybrid use cases. It offers a single platform to process real-time streaming data using Apache NiFi, Apache Kafka, and Apache Flink. You get the same advantages of the DataFlow platform on the public cloud, as you enjoyed on-premises.

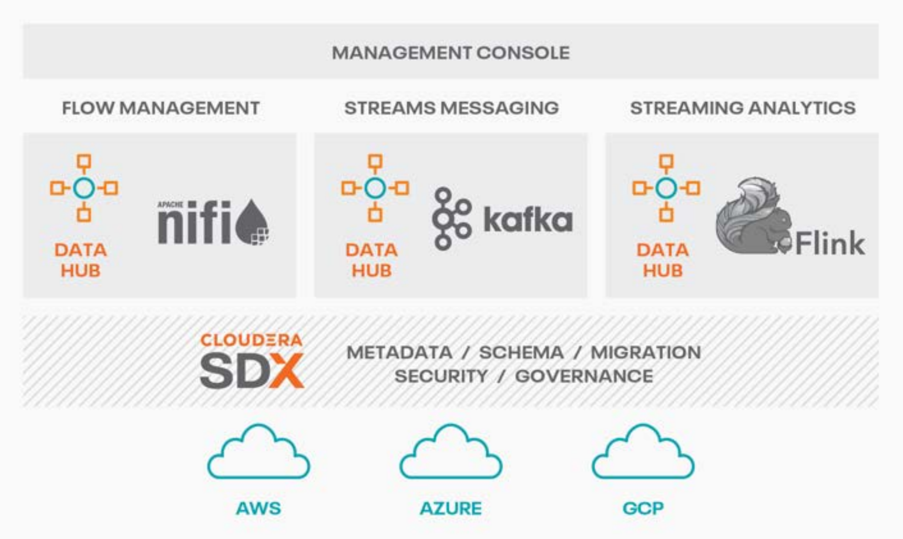
CDP DataHub is a "one-click" approach to create Flow Management, Streams Messaging, and Streaming analytics clusters in the cloud without the heavy burden of estimating requirements to develop and configure them. Enterprises can manage secured data streams by leveraging pre-built "Cluster Definitions" in CDP Data Hub to orchestrate, manage, govern, and execute real-time streaming applications and analytics.

Flow Management for Data Hub can spin-up a cluster of NiFi and NiFi Registry into your public cloud environment. With this, you get an extensive library of over 300+ pre-built processors that can process a wide variety of data formats, including structured and unstructured data. NiFi data flows can easily extend on-premises data streams to other data and analytics applications in the cloud. 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 Kafka clusters in the public cloud with Schema Registry and Streams Messaging Manager. Get complete visibility into your cloud Kafka clusters using Streams Messaging Manager. This allows administrators and operations teams to visualize the data streams across Kafka clusters over hybrid environments.

Streaming Analytics for Data Hub spins up Apache Flink and its related components to the public cloud, bringing stream processing of real-time data into hybrid cloud environments. With CDP Data Hub enabling NiFi and Kafka in the cloud, Flink can easily process streaming data from either of those clusters, enabling faster development of real-time analytics in the cloud.

With Cloudera SDX, data security, governance, and control policies are set once and consistently enforced across Flow Management, Streams Messaging, and Streaming Analytics. These clusters are secured by default and reduce the possibility of a cyber-attack and enforce data governance to ensure data policy is followed.


MANAGEMENT CONSOLE
FLOW MANAGEMENT
STREAMS MESSAGING
STREAMING ANALYTICS
DATA HUB
APACHE nifi
DATA HUB
kafka
DATA HUB
Flink
CLOUDERA SDX
METADATA / SCHEMA / MIGRATION
SECURITY / GOVERNANCE
AWS
AZURE
GCP