CLOUDERA DATAFLOW (CDF)

Why Cloudera DataFlow?

- **100% open source** - Invest in your architecture and scale with confidence knowing that there is no vendor lock-in
- **300+ pre-built processors** - Only product to offer such comprehensive connectivity and a no-code user experience from edge-to-cloud
- **Built-in data provenance** - Only platform in the market to offer out-of-the-box end-to-end data lineage tracking and provenance across MiNiFi, NiFi, Kafka, Flink, etc
- **Choice of multiple stream processing engines** - Supports Apache Flink, Kafka Streams and Spark Structured Streaming for real-time insights and predictive analytics
- **Hundreds of Kafka customers** - Cloudera has hundreds of happy customers getting excellent support on their advanced Kafka deployments
- **Edge IoT use cases** - Collect, process and manage IoT data from thousands of endpoints from the edge to the cloud with ease
- **Multi-cloud/Hybrid cloud strategy** - With the power of CDP, easily chooses a flexible deployment model for your streaming architecture that spans across edge, on-premises and multiple cloud environments.

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. Built on 100% open source technology, CDF supports secure and governed data ingestion, data transformation and content routing. CDF helps you deliver a better customer experience, boost your operational efficiency and stay ahead of the competition across all your strategic digital initiatives.

The three tenets that make up CDF deliver the complete toolset you need to manage, secure and govern data from the edge to the cloud - Edge and Flow Management, Streams Messaging and Stream Processing & Analytics. With the tight integration with Cloudera Data Platform's (CDP) Shared Data Experience (SDX), you get unified security and governance across the stack.

"Cloudera helped our organization get to the next level by providing us with a streaming data platform, which provides us with real-time data. Rabobank is aiming for a self-service environment for our data, and we want our customers to be able to access the data at a click of the button in a secure and controlled manner. Within a financial institution it's very important to be in control of your data, and Cloudera is helping us to support that."

Martijn Groen, IT lead of the Data Lake, Rabobank

Edge & Flow Management

Getting the data from the edge is a challenge that many enterprises struggle with. Collecting and processing data from thousands of endpoints, devices and sensors at the edge is super critical for a lot of real-time scenarios including IoT use cases like predictive maintenance, connected cars, utilities monitoring etc. The next step in that process would be to assimilate all that data and ingest it into the enterprise or the cloud for immediate refinement and processing.

The edge management capability is delivered with a combination of edge agents and an edge management hub. One can manage, control and monitor edge agents to collect data from edge devices and push intelligence back to the edge. This allows you to develop, deploy, run and monitor edge flow apps on thousands of edge devices. Apache MiNiFi is a light-weight edge agent that implements the core features of Apache NiFi, focusing on data collection & processing at the edge. Edge Flow Manager (EFM) is an agent management hub that supports a graphical flow-based programming model to develop, deploy & monitor edge flows on thousands of MiNiFi agents.

CDF's Flow Management is powered by Apache NiFi, a no-code data ingestion and management solution. Apache NiFi is a very mature open source solution meant for large scale, high velocity enterprise data ingestion use cases. Primarily meant for real-time streaming sources such as clickstreams, social streams, log data etc., Apache NiFi can handle all types of data across any type of data source. With 300+ processors, NiFi is the most comprehensive toolset in the market for data ingestion and management. NiFi Registry, which augments NiFi, enables DevOps teams with versioning, deployment and development lifecycle of flow applications.

Streams Messaging

The second tenet in the CDF platform is to ensure that all the ingested data streams can be buffered in a transient state from where other applications can consume the data for their needs. This allows the enterprise to scale effectively when the data streams start growing to the tune of petabytes of data from thousands of origination points. Streams Messaging is the ability to buffer such data streams in a publish-subscribe type of model to achieve IoT-scale.

CDF's Streams Messaging capabilities are powered by Apache Kafka. It provides advanced messaging and processing capabilities. Kafka is fully integrated with Cloudera Manager (CM) for cluster management and monitoring, Apache Ranger for role-based authorization and a Schema Registry service that provides governance and schema management capabilities. Extended capabilities around Kafka include support for Kafka Streams for light-weight stream processing, Kafka Connect for native connectors into Kafka and Cruise Control for optimizing and scaling your Kafka clusters.

Beyond this, CDF's Streams Management capabilities include management and monitoring of your Kafka clusters with an intuitive and easy-to-use interface called Streams Messaging Manager. Also, to ensure business continuity and offer disaster recovery for your Kafka clusters, Streams Replication Manager, based on the latest Mirrormaker 2, is tightly integrated into the platform.

Stream Processing & Analytics

The third tenet in the CDF platform is the ability to process the incoming data streams in real-time and with low latency to offer actionable intelligence in the form of predictive and prescriptive insights. This step is critical in fulfilling the Data-in-Motion lifecycle for an enterprise since there is zero value in ingesting all those real-time streams if no real-time value is derived from them and acted upon to create a positive impact for your business.

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

While Cloudera supports multiple stream processing engines such as Kafka Streams and Spark Structured Streaming, the de facto engine of choice for key enterprise use cases today is Apache Flink. It has gained immense popularity across multiple industries for a range of use cases across fraud detection, predictive maintenance, etc. Apache Flink delivers predictive insights from real-time stateful processing of IoT-scale data streams and complex events. CDF's support for Flink also includes support for data sources/sinks like Kafka, HDFS, HBase, Kudu etc. Support for DataStream and ProcessFunction APIs are also included to improve the developer experience. Integration with Apache Atlas allows for true data governance and lineage tracking from a source at the edge all the way to the point where insights are generated about the data. SQL and Table API support is also available to query data directly from Kafka or Kudu via plain SQL.

Shared Data Experience (SDX)

Beyond all the integration between the three tenets of CDF, the most important element that makes CDF a true platform is Cloudera Data Platform's SDX. A powerful data fabric for complete security, governance and control across infrastructures, providing ultimate deployment choice and flexibility. Since all the components of CDF integrate tightly with SDX, you get a unified experience for security (with Apache Ranger), governance (with Apache Atlas) and data lineage from edge-to-cloud.

Embracing the Cloudera Data Platform (CDP)

As we see enterprises struggle to take their streaming data to the cloud but still want to retain their on-premises footprint, they start adopting a hybrid cloud architecture. To ease the challenges of these enterprises in such complex environments, CDP comes across as a fantastic platform to embrace a multi-cloud or hybrid cloud strategy. Naturally, Cloudera took the best of CDF into the CDP world as well to enable enterprises to have the same amazing streaming experience they have on-premises, on the cloud as well.

Flow Management and Streams Messaging cluster templates are available on CDP Data Hub to enable you to instantly provision pre-defined NiFi or Kafka clusters into your favorite public cloud in just a matter of minutes. Both the templates take advantage of CDP's SDX and support extensive security, central user management, single sign-on, central permission management and comprehensive data lineage tracking for all data streams.

Similarly, all the tenets of CDF are also made available on CDP Private Cloud as well, giving you complete control of how you deploy your streaming architecture.

Modernize your data lifecycle

Take the next step towards modernizing your data streams by bridging your on-premises to the cloud and build a foundation for the next generation data streaming platform with Cloudera DataFlow.