Cloudera DataFlow

*Accelerating big data collection and data flow management*

Cloudera DataFlow (CDF), powered by Apache NiFi, is an integrated platform that solves the challenges of collecting and transporting data from a multitude of sources. HDF is designed to provide simple, fast data acquisition, security-rich data transport, prioritized data flow and clear traceability of data from the edge of your network to the core data center. It uses a combination of an intuitive visual interface, a high-fidelity access and authorization mechanism, and an always-on chain of custody—data provenance—framework.

**Figure 1: Cloudera DataFlow**

CDF was designed to meet the challenges of collecting data from a wide range of data sources securely, efficiently and over a geographically dispersed and possibly fragmented network.
CDF helps enterprises:

<table>
<thead>
<tr>
<th>Make use of operational efficiency</th>
<th>Make better business decisions</th>
<th>Increase data security</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerate big data return on investment (ROI) through simplified data collection and an intuitive data flow management interface.</td>
<td>Make better business decisions with highly granular data-sharing policies.</td>
<td>Support data security from source to storage with an implementation process designed for ease of use.</td>
</tr>
<tr>
<td>Reduce the cost and complexity of managing and maintaining data flows.</td>
<td>Automate data flow routing, management and troubleshooting without coding.</td>
<td>Improve compliance and reduce risk through highly granular data access, data access, sharing and usage policies.</td>
</tr>
<tr>
<td>Trace and verify the value of data sources for future investments.</td>
<td>Enable on-time, immediate decision making by using real-time bidirectional data flows.</td>
<td>Create a security-rich data flow environment that can run the same security and encryption on small-scale Java virtual machine (JVM)-capable data sources and enterprise-class datacenters.</td>
</tr>
<tr>
<td>Adapt to new data sources through an extremely scalable, extensible platform.</td>
<td>Increase business agility with prioritized data collection policies.</td>
<td>Increase business agility with prioritized data collection policies.</td>
</tr>
<tr>
<td>Accelerate ROI through a single data-source agnostic collection platform.</td>
<td>Reduce cost and complexity with an intuitive, real-time visual user interface.</td>
<td>Implement data security from source to storage.</td>
</tr>
<tr>
<td>React in real time with bidirectional data flows and prioritized data feeds.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Common applications of Cloudera DataFlow

CDF accelerates time to insight by enabling off-the-shelf, flow-based programming for big data infrastructure in a security-rich environment. It’s also designed to help simplify the current complexity of secure data acquisition, ingestion and real-time analysis of distributed, disparate data sources.

Example 1: Accelerated data collection and operational effectiveness

Current big data collection and ingest tools are purpose-built and over-engineered because they weren’t created with universally applicable, operationally efficient design principles in mind. This issue creates a complex architecture of disparate acquisition, messaging and often customized transformation tools that make operations time-consuming and expensive.

Streamlined big data ingestion

CDF accelerates big data pipeline ingest through a single integrated and extensible visual interface. This process results in faster return ROI of big data projects and increased operational effectiveness.

Example 2: Increased security and powerful chain of custody

The tools used for transporting electronic data today aren’t designed for future security requirements. It’s difficult for current tools to share discrete bits of data, much less do so dynamically.

Increased security and provenance with Cloudera DataFlow

CDF provides end-to-end data provenance. Beyond the ability to meet compliance regulations, data provenance provides a method for tracing data from its point of origin, from virtually any point in the data flow, to determine which data sources are most used and most valuable.
Example 3: The Internet of Things

CDF is a scalable platform for the acquisition and ingestion of the Internet of Things (IoT) or, even more broadly, the Internet of Anything (IoA).

Adaptive to resource constraints

There are many challenges in enabling an ever-connected yet physically dispersed IoT. Data sources may be remote, physical footprints may be limited, power and bandwidth are likely to be both variable and constrained. Much of the data being produced is data in motion. Unlocking the business value from this data is crucial.

CDF supports the prioritization of data within a data flow. Bidirectional data flows adapt to fluctuations in data volume, network connectivity, and source and endpoint capacity. This process means that, should there be resource constraints, the data source can be instructed to automatically promote the most important information to be sent first. It will hold less important data for future windows of transmission. It could even possibly not be sent at all. Additionally, with a fine-grained command and control interface, data queues can be slowed or accelerated to balance the demands of the situation at hand with the current availability and cost of resources.

Secure data collection

CDF addresses the security needs of the IoT with a security-rich, reliable and integrated big data collection platform designed with simplicity in mind. The security features of CDF include end-to-end data provenance: a chain of custody for data. This feature enables the IoT systems to verify origins of the data flow, troubleshoot problems from point of origin through destination and determine which data sources are most frequently used and most valuable.

With the ability to seamlessly adapt to resource constraints in real time, help ensure secure data collection and prioritized data transfer, CDF is an ideal platform for the IoT.

Conclusion

CDF employs an intuitive visual interface, a high-fidelity access and authorization mechanism, and data provenance to help ease the collection and transport of data from multiple sources. Backed by the power of Apache NiFi, CDF is designed to offer simple, fast data acquisition, security-rich data transport, prioritized data flow and clear traceability of data from the edge of your network to the core data center.

Why IBM?

IBM is also a major research organization, holding the record for most patents generated by a business, as of 2017, for 24 consecutive years.

For more information

To learn more about Cloudera DataFlow, please contact your IBM representative or IBM Business Partner, or visit www.ibm.com/analytics/partners/cloudera.