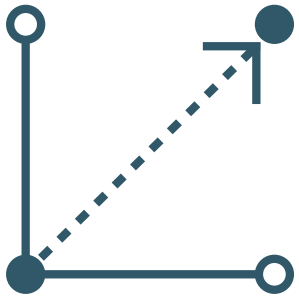
CLOUDERA

Scaling NiFi for the Enterprise with Cloudera

Table of Contents

Introduction

If you have been managing your Apache NiFi environment internally and are considering a transition to a managed service model, this guide is tailored to provide you with valuable insights and guidance you need for decision-making.



NiFi offers organizations the flexibility and control to build complex data flow architectures. However, self-supporting NiFi deployments can be demanding, requiring significant expertise, time, and resources to ensure optimal performance, scalability, and security.

Moving from self-supporting open source to Cloudera's support can bring numerous benefits to your organization. By entrusting the management and maintenance of your NiFi environment to a specialized managed service, you can alleviate the operational burden and focus on your core business objectives.

In this guide, we will explore the advantages of leveraging managed services, including round-the-clock support, expert guidance, proactive monitoring, and rapid issue resolution. You will gain a deeper understanding of how a managed service model can streamline your operations, enhance performance, and improve the overall reliability of your data flows.

Throughout this guide, we will address common concerns and challenges organizations face that may indicate the need for a switch. We will explore topics such as data security and compliance, adjusting to new support workflows, and optimizing cost-effectiveness. In addition, we will discuss what exactly a managed service provides through a series of customer use cases.

By the end of this ebook, you will have a comprehensive understanding of what to expect during the switch from self-supporting NiFi open source to CDP. Armed with this knowledge, you will be better equipped to navigate the transition process and leverage the benefits of managed services to drive operational efficiency, scalability, and reliability in your NiFi environment.

The Origins of NiFi

As you probably already know, Apache NiFi is an open-source data integration tool that provides an intuitive web-based interface for building, managing, and monitoring data flows, making it easy for businesses to automate the flow of data between different systems.

NiFi was initially developed by the United States National Security Agency (NSA) and was later released as an open-source project in 2014. Since then, it has become a popular tool for data integration, processing, and delivery, with a growing community of users and contributors — you very likely may be one of them.

One of the key features of NiFi is its ability to handle real-time data flows in various formats. It also provides a wide range of processors for manipulating data, such as filtering, splitting, merging, and transforming data. These processors can be easily configured using the graphical/visuald interface, making it easy for users to build complex data processing workflows without writing any code.

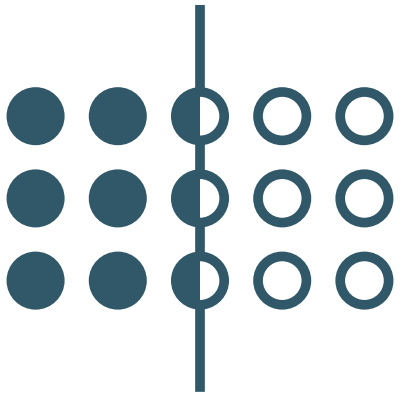
NiFi is based on a flow-based programming model, where data flows through a series of processors, each of which performs a specific action on the data. This approach makes it easy for users to visualize the flow of data and quickly identify any issues or bottlenecks in the data processing pipeline. NiFi also provides features for monitoring data flows, such as real-time statistics on data flow rates and the ability to set alerts for certain conditions, such as data flow errors.

What Does the NiFi Open Source Self-Support Model Look Like?

NiFi open source requires a combination of issue tracking, community support, and combing through existing documentation. If your team has the time, resources, and technical expertise, you may be able to analyze the NiFi source code to gain a deeper understanding of how it works. This can help you diagnose issues and develop custom solutions or modifications.

To solve issues, users can search the NiFi issue tracking system, JIRA, to find information, bug reports, and feature requests. You can search for specific problems you encounter and see if there are any workarounds or planned fixes available. Additionally, you can report new issues if you come across any.

Successful NiFi use optimizes performance, efficiency, and meets or exceeds security requirements. All of this takes time, resource coordination, and know-how to do effectively.



What Are the Challenges of Self-Supporting NiFi Open Source?

While NiFi is a powerful tool to collect, process, and distribute data — with the ability to adapt to a user's needs and quickly build innovative solutions to their challenges — using NiFi requires substantial operational overhead. Some of these overhead tasks include:

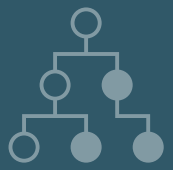
- Managing software (Downloading, installing, provisioning, etc.)
- Ongoing maintenance (Ensuring configuration of deployments to optimize workloads and maintain security)
- Data (Managing schemas)

The more of these tasks an organization takes on, managing them among increasing use cases becomes more difficult.

And any time a user has to deal with structured data, a data flow should always make reference to schemas. This prevents downstream systems from receiving data with unexpected schemas. Schemas evolve of course, and as they do sources and sinks must be kept in sync to prevent errors or data quality issues. NiFi only provides rudimentary schema management for Avro schemas without additional features like schema versioning, schema evolution or other schema formats.

With all of this in mind, using a schema management system becomes mandatory.

NiFi is a powerful tool to collect, process, and distribute data — with the ability to adapt to a user's needs and quickly build innovative solutions to their challenges — using NiFi requires substantial operational overhead.



NiFi is a feature-rich data integration platform, which can make it complex to understand and troubleshoot issues. NiFi's extensive capabilities, including data routing, transformation, and integration with various systems, require a good understanding of its architecture and components. Navigating through the different processors, controller services, and configurations can be overwhelming for newcomers.

As an open-source project, NiFi also has a learning curve associated with it. The concepts and terminology used in NiFi may be unfamiliar to users who are new to the platform. Understanding the flow-based programming model, NiFi's user interface, and the configuration options can take time and effort. Users need to invest in learning the platform to effectively troubleshoot and resolve issues.

While NiFi provides extensive documentation, it may not cover every specific use case or troubleshooting scenario. Some issues or challenges that users encounter may not have clear solutions documented. In such cases, users may need to rely on other sources or experimentation to find resolutions.

Highly skilled and ambitious developers often adopt NiFi to quickly hack together their innovation projects, but NiFi is a tool that requires technical expertise in data integration and related technologies. Troubleshooting complex issues may require understanding various systems, protocols, and data formats. Users without prior experience or knowledge in these areas may find it challenging to diagnose and resolve problems effectively.

And a DIY approach is likely to be sub-optimal in terms of performance, efficiency and security. And when there are problems, DIY means working through forums looking for answers. All in all, this means that open source NiFi is a viable long-term solution only for those development teams that have lots of experience and have the flexibility to spend time researching or experimenting to find answers to their challenges.

How CDP Helps Scale NiFi Efforts

Cloudera not only provides support from a team loaded with NiFi knowledge, but offers a packaged product that reduces the operational/administrative overhead of DIY.

This becomes especially important as you scale beyond the innovation stage and start to put more data pipelines into production and need to manage more NiFi clusters.

CDP is the natural evolution of Apache NiFi.

Cloudera offers a cloud-optimized data service with self-service pipeline development experience built to power universal data distribution efficiently at massive scale.

Benefits

By switching from self-supporting NiFi open source to CDP, you can focus on building robust and highly efficient data pipelines. Cloudera's offering is optimally configured for cloud or on-prem deployments greatly reduces administrative burden, while simplifying security and governance.

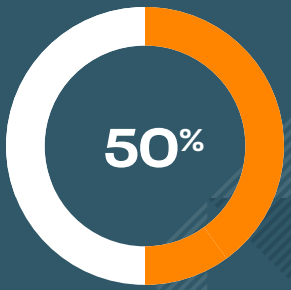
Furthermore, with Cloudera support, you can enjoy quicker resolution of incidents, improved quality of service, and reduced downtime. According to Gartner, the average cost per minute in a major breakdown is $5,600, however, Cloudera customers can experience a 35% drop in time to resolution, reducing outage time. The improved TCO and risk exposure with CDP makes it an obvious business decision. CDP brings the innovation of open source to you with the added benefit of smoother operations.

The time savings associated with this service are significant. CDP can save businesses 60-85% on cloud infrastructure spend. Administrators enjoy 60-80% time savings. And pipeline developers can slash 20-40% of their time, all according to The Forrester Wave Report.[1]

What to Expect When Switching from Self-Supporting NiFi Open Source to CDP

Cloudera has worked with customers across a range of industries to manage their NiFi workloads.

With the added support and security of a managed service, these organizations are seeing major results, like saving 50% on resource utilization.



Organizations are seeing major results, like saving 50% on resource utilization

CASE STUDY | Transportation

A European Rail Service Faced the Pain of Self-Support

A rail service in Europe must collect and distribute data about a train system of more than 2,000 trains.

NiFi was effective in quickly building pipelines, but the team needed to ensure security and development efficiency at massive scale. Previously, the massive volumes of data needed to ensure safe and smooth operations were only available through a relational database with partial data sets due to capacity and limited user access. Managing the data was very labor intensive and expensive. The company was particularly concerned with security and safety, so they decided to switch from self-support to a managed data service model.

They chose CDP because of its functionality, scalability, and lower total cost of ownership relative to other options. Since migrating to CDP, data-driven decisions are nearly instant.

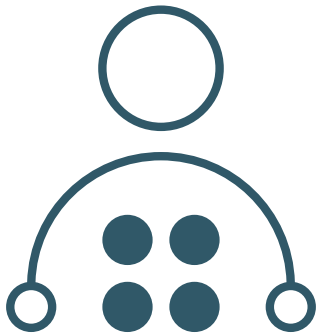
With Cloudera technology this information determines the exact number of trains that should drive on each line at a certain time or the exact number of passengers per square meter in each wagon.

CASE STUDY | Health Insurance

A Large Health Insurance Provider

A large health insurance provider was ingesting data from many different sources, both streaming and in batch, to support data apps with high SLAs.

The organization has highly variable workloads, especially during enrollment periods when bursts of data volumes demand elastic scalability. In less than two months, the team was able to move more than 100 data pipelines to Cloudera, saving almost 50% in resource utilization from optimized cloud configuration and auto scaling features.



CASE STUDY | Data and Analytics

A Retail Data Science, Insights, and Media Company

A retail data science, insights, and media company faced several challenges self-supporting NiFi open source and turned to CDP.

With Cloudera, they found operational reliability and performance, rapid scaling, visibility and fine grain control of resources, user management for strong security, and development flexibility.

Cloudera enables operational reliability and performance. Data pipelines are therefore able to meet stringent SLAs for uptime and performance. Furthermore, with visibility and fine grained control of data distribution and resource utilization, the company is able to accurately attribute costs.

Ultimately, data is delivered with minimal latency and burden on resources, including administrative overhead because Cloudera is always optimally configured for performance and easy to manage from a central control pane. They are able to quickly add or reduce capacity quickly because cluster deployment scaling is no longer a manual task.

CASE STUDY | Manufacturing

An American Multinational Food Manufacturing Company

An American multinational food manufacturing company was using open-source Apache NiFi. They faced increasing issues with their poorly developed flows on unsupported Apache NiFi and were looking to improve performance, resource utilization, and actionability of their data pipelines.

They needed a solution to help with event-driven ETL, waiting for files to land to trigger ETL processing on the file and time-driven ETL orchestration, where ETL jobs would run at specific times during the day/week. They also needed to create daily sales reports across cereal, snack, and plant-based products sold in stores globally.

The solution was to replace Apache NiFi with CDP. The business outcome was faster onboarding of new data sources and avoiding the risk of running an unsupported NiFi version on IaaS AWS, which could have resulted in significant expenses. With Cloudera, they gained operational stability of data pipelines in production, reduced security risks, and optimized configurations for resource utilization.

CASE STUDY | Health Insurance

An Israeli Health Maintenance Organization

This company is one of the four health maintenance organizations currently active in Israel. They were using the open-source version of NiFi for over four years, integrating it with Couchbase, SQL, Kafka, and Elastic.

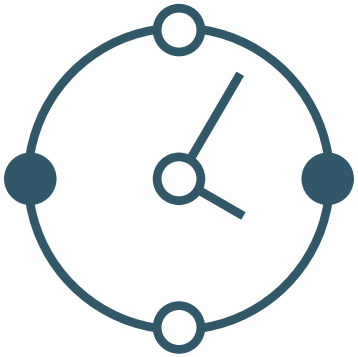
Their challenges with open-source NIFi were mainly security, like securing the cluster with Kerberos and ongoing support. The BI department's business challenges were to move the data while maintaining health compliance, such as FHIR. Ultimately, they required NiFi to read and write to Kafka.

They became a CDP customer in 2023, enjoying the business outcome of rapid platform learning — a situation where no more learning is necessary after two use cases. They also benefited from full security in an easy-to-manage single environment.

Conclusion

Self-supporting NiFi open source not only takes time and resources, but a DIY approach is likely to be sub-optimal in terms of performance, efficiency, and security.

Cloudera customers who switched from NiFi open source to CDP enjoyed better functionality, scalability, and lower total cost of ownership relative to other options. Time savings for all teams across the board were significant. Customers also received real-time alerts and notifications, faster onboarding of new data sources, and increased speed in pipeline deployment. All this without diving into source code, without combing JIRA for bug reports, and without requiring technical expertise in data integration and related technologies. With CDP, customers can harness the full power of Apache NiFi for data integration, processing, and automation needs.



Learn More

Ready to learn more about the specific products that CDP offers? See our product pages for more detail.

- Cloudera Flow Management for on-prem deployment
- Cloudera Data Flow in the public cloud
- Cloudera Edge Management to integrate Apache NiFi into the ecosystem where management of distributed edge agents is necessary.

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 | US: +1 888 789 1488 | Outside the US: +1 650 362 0488

Sources

[1] The Forrester Wave: Notebook-Based Predictive Analytics and Machine Learning, Q3 2020," Cloudera Blog, accessed June 21, 2023.

Cloudera, Inc. 5470 Great America Pkwy, Santa Clara, CA 95054 USA cloudera.com

CLOUDERA