

**CLOUDBERA**

# **Data Distribution Architecture to Drive Innovation Using Cloudera Data Flow on AWS**

Simplify Your Streaming Data  
on the AWS Cloud



---

# Table of Contents

<b>What If You Could Simplify the Delivery of Data?</b>	3
<b>Complexity Expands Exponentially</b>	6
<b>Analytics and Data are Fundamental to Business</b>	8
<b>The Solution is Universal Data Distribution</b>	9
Support Active and Passive, Real-time, and Batch	9
Get There Faster, Cheaper, and Safer	10
<b>Implement a Universal Data Distribution Solution</b>	11
Get Started Quickly	11
Optimize Through DataFlow Deployments and Functions	12
<b>Boost Developer Productivity</b>	13
<b>Case Study: Optimize Financial Data Feeds</b>	15
<b>Case Study: Enhance Security Information and Event Management (SIEM)</b>	16
<b>Go to the Next Level</b>	18

---

# What If You Could Simplify the Delivery of Data?

The exponential growth of data that is generated each year shows no signs of abating and is only matched by the significant increase in the number of systems. These systems are needed to ingest, process, and store that data via a complex network of middleware and routers that traverse various hybrid cloud and on-prem environments that each have their own particular set of features and nuances.

That is how things are today. Your job to transform your organization from the relatively straightforward world of on-prem computing, with a small number of storage platforms, to a dynamic and complex world where data is constantly generated, distributed, and stored anywhere and everywhere can seem unrealizable. The point-to-point solutions that enabled you to deliver data before are now too complex to consider.

Let's take a common example where mobile telemetry data is generated by many devices. It then needs to be captured and delivered to the cloud for

future analysis and storage while simultaneously being routed to a real-time analytics pipeline where machine learning algorithms are executed. Furthermore, the results of those algorithmic processes must then be joined with data from other applications. In that scenario, there are a plethora of systems to integrate that essentially all use the same data but in different ways. In building a custom solution for each destination, an engineer cannot ensure that data can be rapidly onboarded and transformed to where it is best utilized for business and/or analytics purposes.



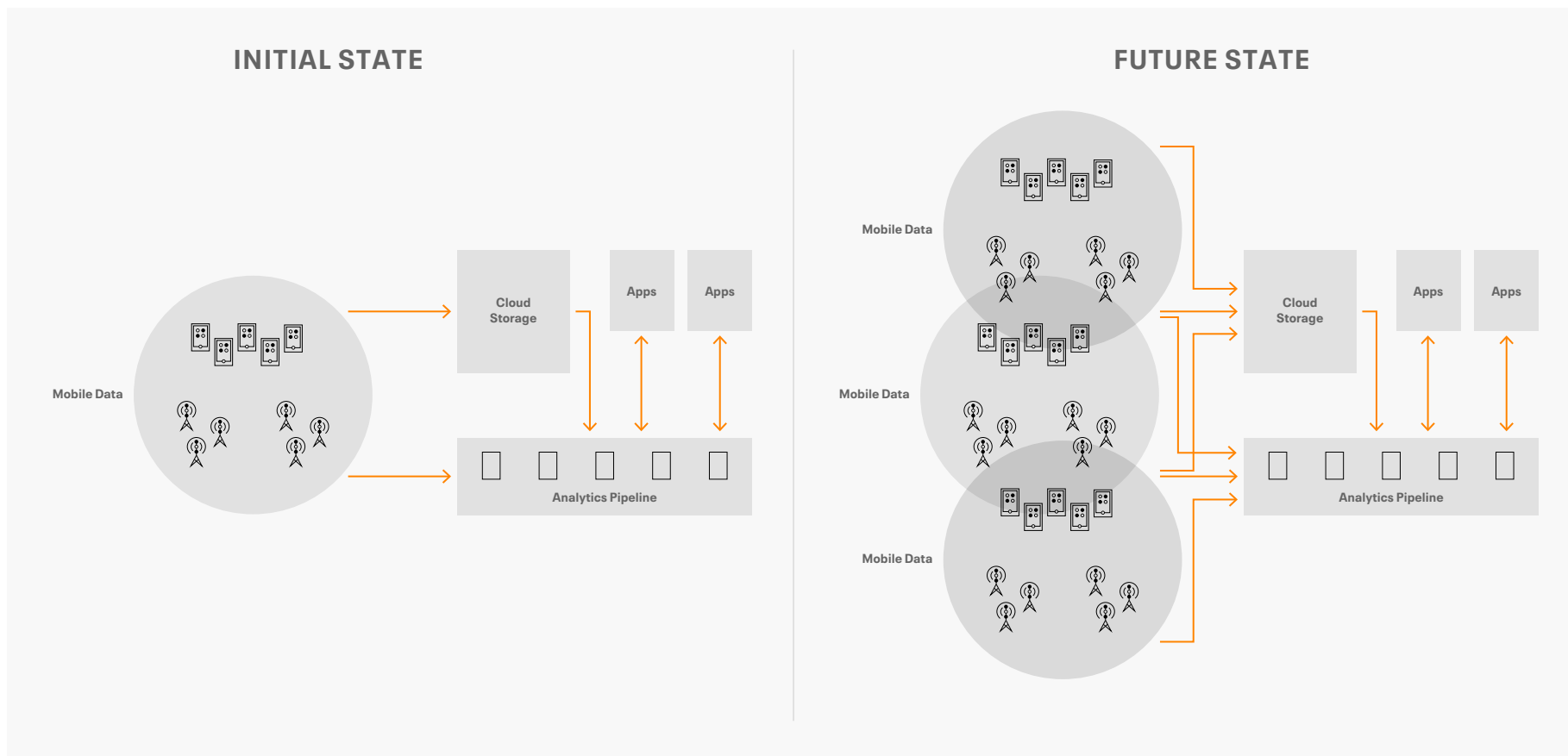


Figure 1: An example of the exponential growth of data in relation to mobile devices.

Managing business demand and complexity, while making sure that your team of data engineers and analysts are as productive as possible, is a challenge for every organization.

What if they, through a no-code point-and-click tool, could design the appropriate number of flows that move and enrich data from one location to the best location, such as to the Amazon Web Services (AWS) Cloud, quickly, easily, and in the most efficient way possible?

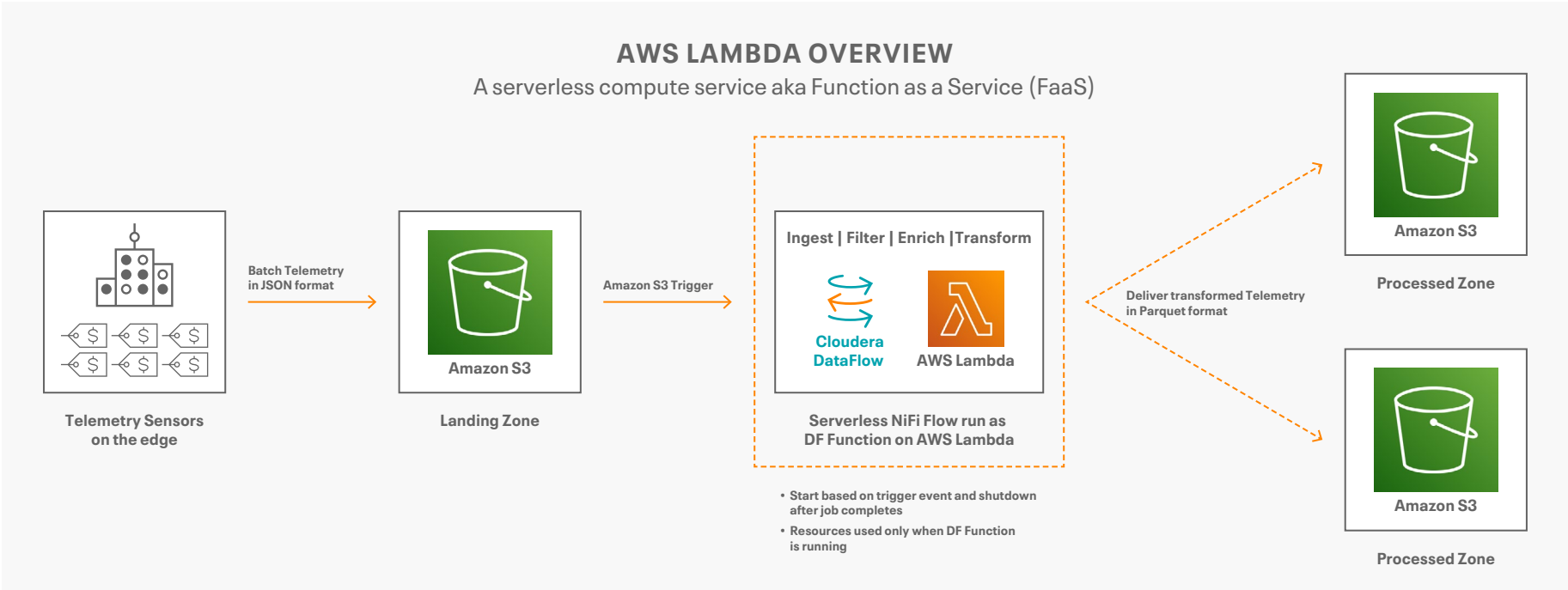


Figure 2: An example of a recommended approach to telemetry data architecture using Cloudera DataFlow for Public Cloud with AWS Lambda.

---

# Complexity Expands Exponentially

The complexity of data systems and infrastructure is increasing, especially when considering a move to the public cloud. Historically, there were a limited number of data platforms and so delivering data from one system to another could follow a traditional point-to-point model. Typically a central database contained all the relevant data and through a bespoke set of SQL scripts, data could be shaped and shipped to a few destination platforms. This model was relatively straightforward and not technically complicated since you were dealing with structured data and simple feed layouts.

Inevitably, limitations and inefficiencies with this model started to appear. For example, the development of new feeds often involved copying old ones and thus leaving the team with redundant code in numerous locations. The management and maintenance of these changes became costly and time-consuming, mostly due to non-standardization.

Today, delivering data is no longer solutioned by simply copying and writing a new SQL feed because there is often the need to map from one schema to another with a different data format.

Additionally, new layers of data distribution functions have made things immensely complex. For example, think about how difficult it can be to take a real-time JSON message, transform and store it in a relational database, write a copy to Amazon S3 (for future analysis), and all the while, enrich and stream it via Kafka to a real-time on-prem analytics service.

Business is moving along faster than it has in the past, and the changes that are needed to incorporate new data paradigms can no longer take months, but instead need to be done in hours or days.

## Hybrid: Don't Forget On-Prem Connectivity

Before the advent of cloud services, most organizations had their own data centers and could pretty easily manage their infrastructure. Networks were tuned for high performance between servers and they could ensure that the connections between geographically diverse data centers had the necessary bandwidth to communicate effectively.

Compared to today, that was relatively simple. Organizations now utilize their cloud provider, such as AWS, to guarantee stability and resiliency but even with this transition, many services are still run on-prem.

Imagine integrating thousands of nightly feeds between on-prem to the cloud.

Challenging? It can be.

Impossible? No.

Is there a solution? Yes. Read on.

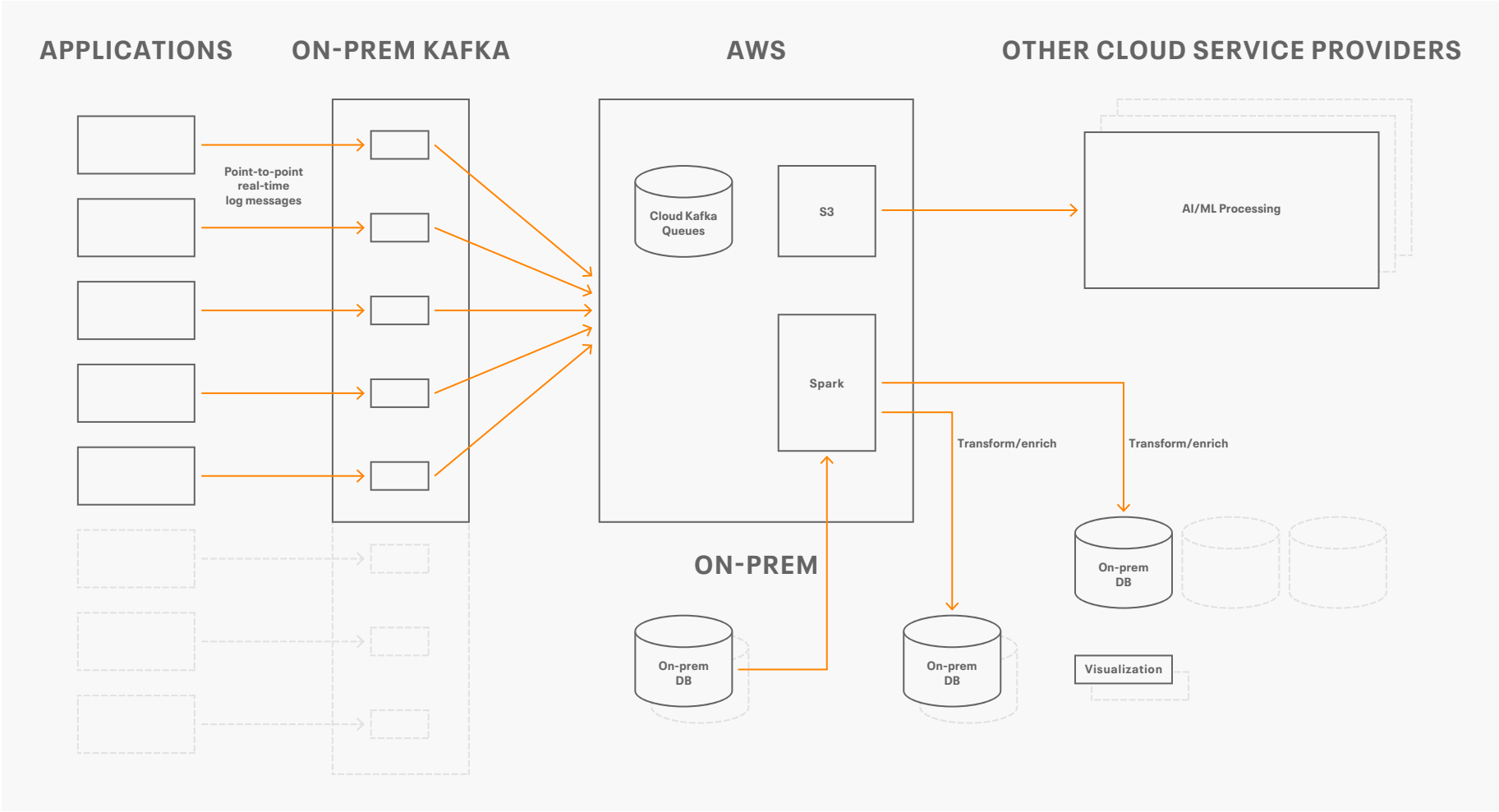


Figure 3: An example of the growing complexity of point-to-point solutions. Gray dotted shapes indicate expanding complexity.

---

# Analytics and Data are Fundamental to Business

Data and analytics are the energy upon which corporations run and both are fundamental to every aspect of business execution. In order to make data and analytics work together, businesses need to have a clear understanding of what data they have, where it came from, and how it can be used to inform decisions. Additionally, businesses need the right tools and technologies to collect, transform, deliver, and analyze data.

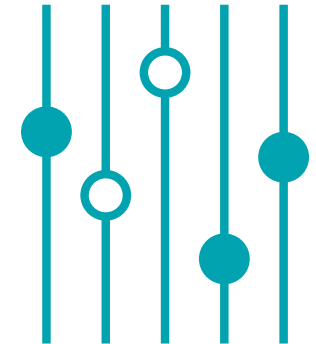
From using historical data that analyzes the past and teases out new operational efficiencies to using real-time information for new product recommendations and to detect in-progress fraud, the variety, complexity, source, and structure of data has grown over a very short period of time.

The highest value is achieved by delivering information to where it is best utilized. Efficient data distribution is critical to maintaining a responsive, modern, and data-driven business. The very speed by which you and your team enable new data feeds is what will help your organization become digitally sophisticated and consequently, make the best use of the most updated data.

For instance, considering fraud analytics use cases, what if you could process and make available a new JSON message that holds a piece of information that would reduce your credit card fraud rate by 3%?

The cost of not incorporating that information into your fraud analytics is now quantifiable because, without it, you are losing 3% of your sales to fraud every day. Being able to bring that piece of information into your fraud detection models and use it for decision-making is a huge opportunity and now takes on a new urgency. Tools like Cloudera DataFlow for the Public Cloud (CDF-PC), enable you to rapidly bring that particular piece of data to where it is most needed. But this cannot be done through a traditional point-to-point model.

Cloudera is an AWS Data and Analytics Competency Partner that provides a fast, easy, and secure platform to help customers use data to solve demanding business challenges. Organizations can accelerate their journey of multi-function analytics to the cloud with the [Cloudera Data Platform on AWS Marketplace](#).





---

# The Solution is Universal Data Distribution

Large data repository hub-and-spoke and point-to-point data distribution are examples of how organizations have not treated data collection and distribution as a first-class problem.

This is especially true with hybrid cloud, such as the combination of AWS cloud and on-prem environments because building connectors to all the various systems and infrastructures is difficult. As the number of systems increases the complexity and time of managing change becomes a significant obstacle to growth along with numerous other technical challenges.

Simplifying your streaming data, connecting to any data source, anywhere, processing, and delivering it to any destination is a necessity in business. A smart way to address this need is to establish what is called universal data distribution. This paradigm supports connectivity to hundreds of different applications and systems while allowing a developer or business user to graphically design new dataflows in a no-code UI as they are needed. With this, your organization will be able to take control of all its data pipelines in a way that allows for rapid deployment and easy management.

## Support Active and Passive, Real-time, and Batch

There is a clear path to universal data distribution and integration if you can solve the root problems. Cloudera, collaborating with AWS and working directly with their customers across many industries to integrate data across disparate systems, has found that there are actually only a finite number of challenges and that they are similar. Those challenges are:

- **Unifying different formats and schemas:** For example transferring JSON messages to a SQL database.
- **Bridging different protocols:** Moving data between different protocols such as Kafka messages into a document database or a cloud application such as Salesforce.
- **Getting the right fit for batch and stream:** It is more efficient to move large objects in a batch and smaller objects in a stream.

# 450+

---

## Connect Them All

CDF-PC in the AWS cloud comes with an ecosystem of over 450 data connectors to integrate and distribute data universally, without worrying about the complexity and specifics of the architectures. It takes the hard part out of data integration and distribution, especially with the challenges that come with a complex multi-cloud and on-prem environment.

- **Filtering data:** Removing unwanted records for any number of use cases such as data cleansing or records that are not needed at the destination application.
- **Enriching and sanitizing data:** Augment and enrich data with additional datasets or fix specific fields before they are transferred to the destination application.

Moving towards a universal data distribution service remediates these issues and rather than having numerous tools and applications to manage data flow throughout the organization, you only need one.

### Get There Faster, Cheaper, and Safer

A key feature of a data distribution platform is the breadth of available connectors and how easy it is to extend them to custom use cases. A comprehensive universal data distribution solution has these capabilities and it supports real-time and batch data alike, making connections between systems that have different characteristics possible.

Once you're able to connect any data source, you then address the data processing capabilities needed to route data based on content, filter out records, convert between data formats, and enrich data, all of which are critical functions to a data distribution layer.

Cloudera DataFlow for the Public Cloud (CDF-PC) on AWS is a single solution that allows you to distribute data from any source, to any destination, through a no-code visual designer. There is no longer the need to utilize vendor specific platforms and one-off solutions to move data throughout the organization. CDF-PC, with an ecosystem of 450+ connectors on the AWS cloud, gives you the ability to seamlessly integrate systems whether they are on-prem, in the cloud, or hosted vendor applications.

However, none of this is practicable without considering data protection. This single solution also comes with enterprise grade security and data governance that is fully integrated to facilitate greater security and data transparency. For example with CDF-PC on AWS, Cloudera customers gain an enormous degree of observability through built-in monitoring. With that, robust data provenance is established because data is tracked and monitored from the beginning to the end of the flow.

With CDF-PC and an AWS cloud-first strategy, data distribution is treated as a top priority and will put you in a position to take back control of your data pipelines from disparate systems. CDF-PC helps your organization operate at the speed of business by delivering data where it is needed and faster.

A large, stylized graphic of the number '90%' in a light teal color. The '9' and '0' are composed of horizontal lines, and the percentage sign is a solid circle with a horizontal line through it.

### Faster response time for cybersecurity threats

It starts with the ability to connect to any data source that was created in the AWS cloud or on-prem system.

The outcomes of implementing a universal data distribution model with CDF-PC on AWS include faster response times and lower costs. See Use Case: Enhance Security Information and Event Management (SIEM) on page 16 for more information.

# Implement a Universal Data Distribution Solution

Powered by AWS, Cloudera DataFlow for the Public Cloud (CDF-PC) is an Apache NiFi based cloud service, which enables developers to connect to any data source anywhere with any structure, process it, and deliver it to any destination using a low-code authoring experience. In short, it's the foundation to enable universal data distribution in an organization.

## Get Started Quickly

As part of CDF-PC, ReadyFlows make it easy for developers to get quickly started by providing templates for the most common use cases. For example, moving data between Confluent Cloud and Snowflake can be accomplished with just a few clicks. The image of the ReadyFlow interface below shows how you would browse a gallery of common use cases from which to deploy your flows in minutes.

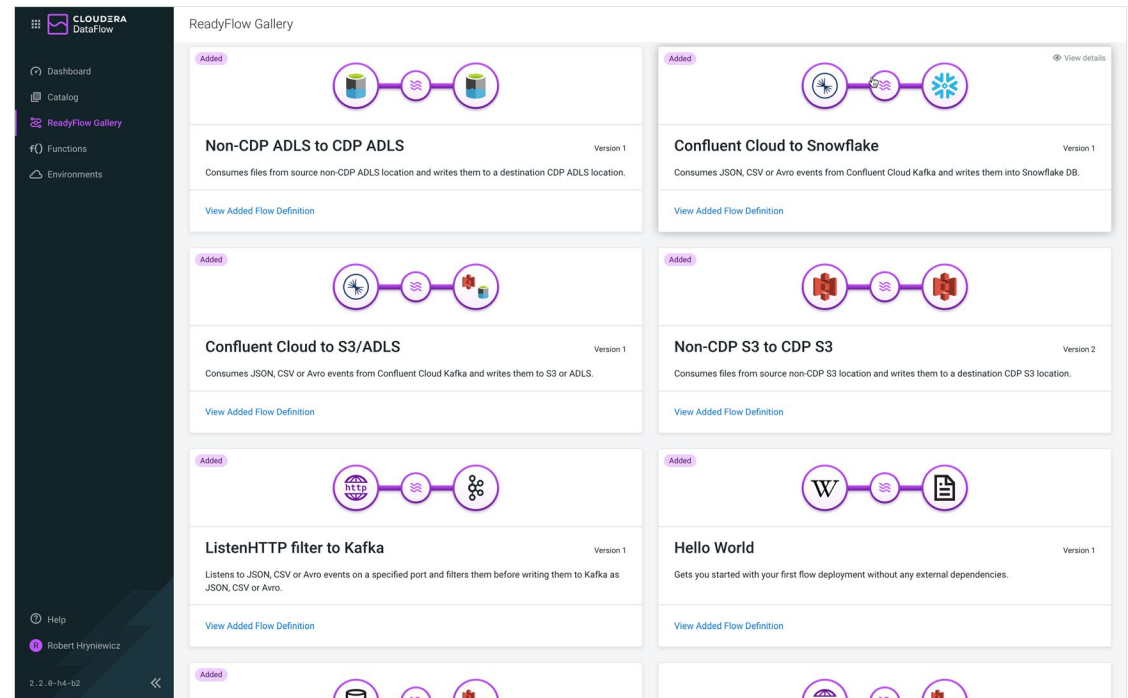


Figure 4: An example of ReadyFlow pre-built flow templates from which to quickly customize and deploy new data flows.

## Optimize Through DataFlow Deployments and Functions

Once you have developed your flows, you need to decide how they will be productized and that depends on the use case, service level agreements, and other requirements. Cloudera DataFlow Deployments and DataFlow Functions provide the options you need to support fundamentally different runtimes so that you can design pipelines that best fit your needs. The table and the diagram below provide such a comparison.

With DFF, users have the choice of deploying NiFi flows not only as long-running auto scaling Kubernetes clusters on Amazon Elastic Kubernetes Service (Amazon EKS) but also as short-lived functions on AWS Lambda.

### CLUDERA DATAFLOW FOR THE PUBLIC CLOUD (CDP-PC)

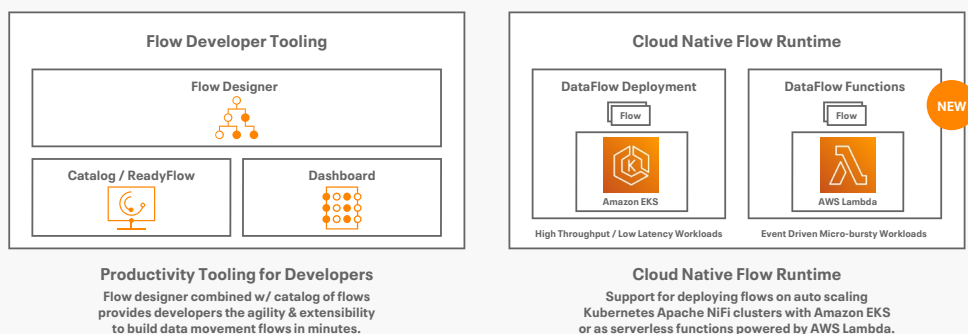


Figure 5: DataFlow functions in Cloudera DataFlow service.

RUNTIME OPTIONS IN THE PUBLIC CLOUD		
<b>Feature</b>	DataFlow Deployments	DataFlow Functions
<b>Cloud Runtime</b>	NiFi Clusters using Kubernetes/Containers	NiFi flows running on cloud providers' serverless compute services (AWS Lambda)
<b>Use Case</b>	Use cases that need low latency for high throughput workloads requiring always running NiFi flows	Event driven, micro-bursty use cases with no sub-second latency requirement where NiFi flows do not need to run continuously
<b>Benefits</b>	Auto-scaling Kubernetes clusters for long running workflows with centralized monitoring	Efficient, cost optimized, scalable way to run NiFi flows serverless allowing developers to focus on business logic

## Build a Serverless Trigger-Based Data Movement Pipeline Using Apache NiFi, DataFlow Functions, and AWS Lambda

Learn about how DataFlow Functions, combined with the serverless compute services provided by AWS Lambda, enables developers to implement a wide spectrum of use cases using the low-code NiFi flow designer UI, and deploy the flows as short-lived serverless functions.

---

# Boost Developer Productivity

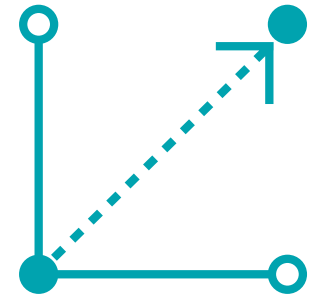
The capabilities so far described in this ebook tend to benefit flow administrators the most because they are free to choose and apply the best options from which to run their production data flows in the cloud. It is important to note that Cloudera has taken a big picture approach, by streamlining the process from development to production for both flow administrators and flow developers alike, especially when moving to the cloud.


CDF-PC on AWS addresses the challenges and needs of flow developers with a self service solution. DataFlow Designer reinforces NiFi's most popular features through an effective and enjoyable user experience. With DataFlow Designer, developers can:

- **Quickly edit processor configurations** without losing focus on the big picture. This is done through a context-sensitive side panel that instantly displays relevant configuration information as you navigate through your flow components.
- **Directly upload JDBC Drivers, Python scripts, and other files** through the designer UI. Developers are now much more self-sufficient in building their own flows, not needing to wait on administrators to open up SSH access to each NiFi instance.


- **Immediately build data flows** without waiting for resources to be created. Developers are able to drag and drop processors to the canvas, create parameters and services, and apply configuration changes. When ready, they then initiate test sessions that provision and terminate resources on-demand. Not only are developers self-sufficient, the economics are better because resources are used only as needed.

The collaboration between flow developers and flow administrators is as follows: flow developers draft flows, build them out, and test them with FlowDesigner and then publish to the central DataFlow catalog. At that point, flow administrators can deploy them in their AWS cloud and benefit from the aforementioned features like auto-scaling, one-button NiFi version upgrades, centralized monitoring through KPIs, and automation through a powerful CLI.




Flow Design / pse-aw-env / czorzin-flow-design / Canvas
Active Test Session

- Dashboard
- Catalog
- ReadyFlow Gallery
- Flow Design
- Functions
- Environments
- Help
- N Nijjwol Lamsal



**ListS3**

ListS3 1.18.0.2.3.7.0-100


IN	0 (0.00 B)
READ/WRITE	0.00 B / 0.00 B
OUT	0 (0.00 B)
TASKS	29 / 00:00:11.184

5 min

NAME **success**

QUEUED 0 (0.00 B)

↓



Funnel

↓

QUEUED 0 (0.00 B)

↓

**czorzin-flow-design**

Process Group

[More Details](#)

Settings

Process Group Name

czorzin-flow-design

FlowFile Concurrency

Unbounded

Outbound Policy

Stream When Available

Default FlowFile Expiration

0 sec

Default Back Pressure Object Threshold

10000

Figure 6: The context-sensitive side panel of FlowDesigner enables developers to quickly edit processor configurations without losing focus on the big picture.

# Optimize Financial Data Feeds

## Challenge

This Cloudera customer, a large asset management firm, offers investment solutions to a broad range of clients. With a diverse client base comes a number of challenges, including:

- Data processing and support for a large variety of market analytic strategies
- Exorbitant cost control of historical market data feeds from numerous 3rd party data providers
- Implementation of agile and scalable development and deployment processes

## Solution

- Cloudera DataFlow (CDF-PC) on AWS
- Cloudera Data Warehouse (CDW)
- Amazon S3

## Outcome

This asset management firm reduced operational overhead and data subscription costs by using CDF-PC to process lower cost real-time daily feeds captured in Amazon S3 and transform them into historical views in CDW. CDW is then used to query the aggregated data and provide the analysis each of their clients need. Additionally, CDF-PC enables an agile and scalable deployment process so that any of their internal bespoke applications can access and use historical or real-time data views directly from Amazon S3.

## Tame All Your Streaming Data Pipelines

Watch this webinar to learn how to streamline development and operation of your data flows with Cloudera DataFlow on AWS. A simple demo use case is provided as an introduction to broader capabilities.



# Enhance Security Information and Event Management (SIEM)

## Challenge

This multinational oil and gas corporation succeeded in building a manufacturing data lake that maximizes an on-prem and AWS public cloud hybrid environment to power real-time analytics while providing a consolidated view of its operations.

Among the many obstacles they had to overcome related to the development, deployment, and management of global dataflows, particularly with regard to cybersecurity. Their mission to optimize log analytics for security information and event management (SIEM) included these challenges:

- Ingest data from multiple clouds and on-prem sources, including PCs (100K+), Linux servers (30K+), and global network routers
- Process and distribute multiple data formats to applications hosted across cloud and on-prem environments
- Lower the cost of log analytics so that budgets could be repurposed to high value initiatives
- Increase the speed of threat detection

## Solution

To address this particular challenge and gain additional flexibility, the oil and gas corporation established a universal data distribution model powered by Cloudera DataFlow (CDF-PC) on AWS. With CDF-PC, developers connect to any data source anywhere with any structure, process it, and deliver to any destination using a low-code authoring experience.

In this particular case, Apache NiFi is used to capture logs from their cloud environment because it proved to be an effective choice for collecting data out of Windows operating systems, ingesting log data from over 100K PC's around the world in real-time. NiFi is then used to parse, process, and distribute specific types of log data to the respective applications regardless of where they reside. Additionally, autoscaling and stabilization was facilitated through NiFi data collection flows.

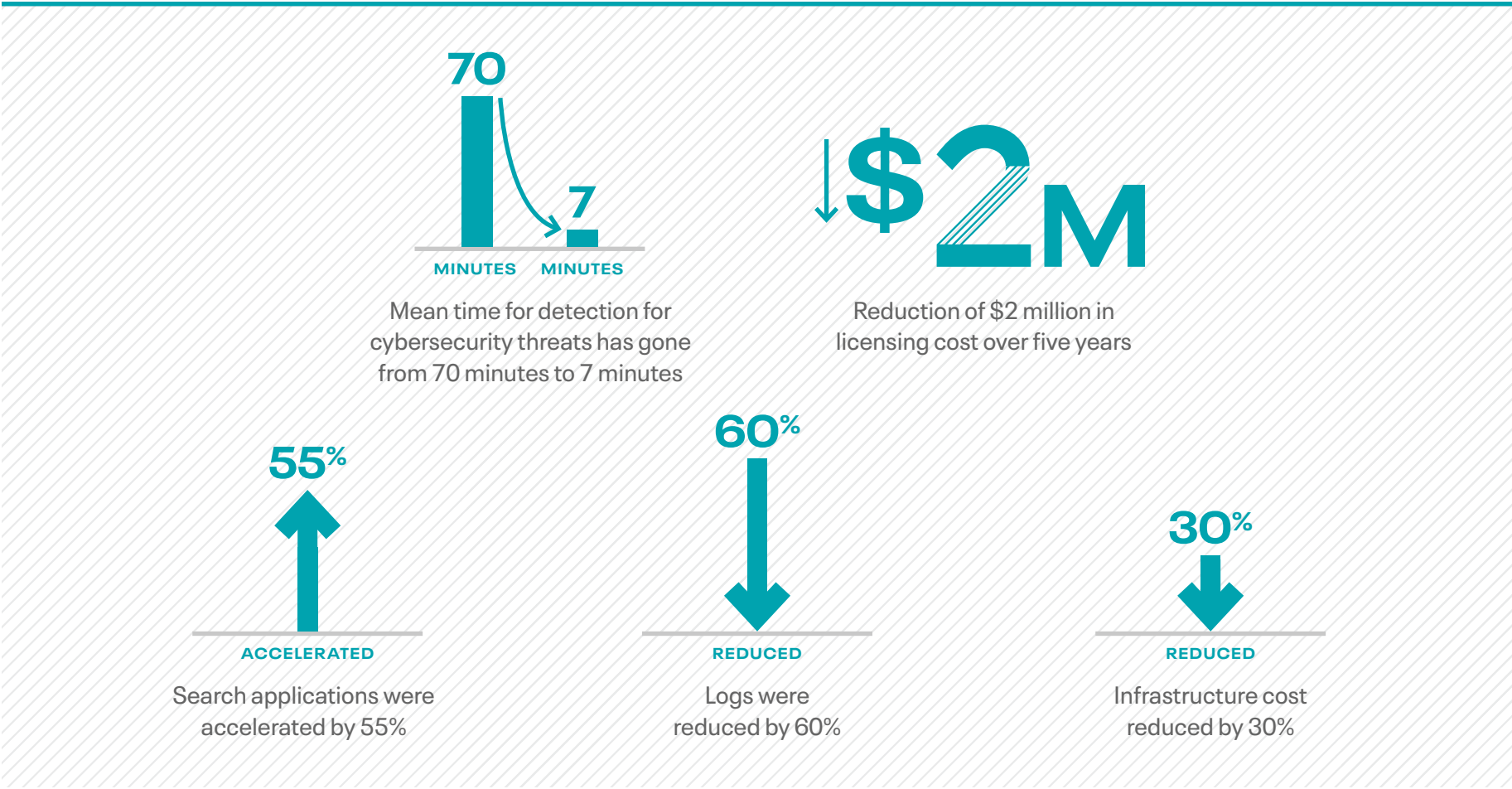
## Serverless on the Public Cloud

Check out the following video for a full end-to-end demo: [Serverless NiFi Flows with DataFlow Functions and AWS Lambda.](#)



**Outcome**

Powered by AWS, the CDF-PC universal data distribution model was implemented as an integral part of the hybrid manufacturing data lake and continues to be critical to keeping operational overhead and infrastructure cost low. Additional impacts of this initiative include:



---

# Go to the Next Level

Getting your universal data distribution right is fundamental to your efforts to transform your organizations from the relatively simple world of on-prem computing to a constantly dynamic and complex world, where data is constantly generated, distributed, and stored everywhere. Once your data is moving you can start to address your needs for real-time analytics and stream processing.

With Cloudera DataFlow for the Public Cloud on AWS, the aspirations to help your organization to become digitally sophisticated is no longer a futile pipe dream, but an entirely possible event.

---

# Learn More

To learn more about implementing your own IoT use cases, ingesting data into your data lakes and lakehouses, or delivering data to various cloud services, take our [interactive product tour](#).

Learn more about Cloudera and AWS [here](#).

## 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](https://cloudera.com) | US: +1 888 789 1488 | Outside the US: +1 650 362 0488

---

Cloudera, Inc. 5470 Great America Pkwy, Santa Clara, CA 95054 USA [cloudera.com](https://cloudera.com)

© 2023 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice. 5695-001 April 6, 2023

[Privacy Policy](#) | [Terms of Service](#)

**CLOUdera**

