CLOUDERA

Cloudera DataFlow (CDF)

Dinesh Chandrasekhar

Product Marketing, Data-in-Motion

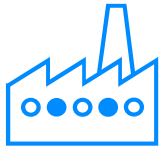
@AppInt4All

TODAY'S NEEDS FOR DATA STREAMING

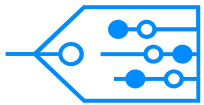
Gain Competitive Advantage

"Many leading enterprises realize that real-time analytics — the analytics of the present — is an incredible competitive advantage because they can act now to ***serve fickle customers, fix operational problems, power internet-of-things (IoT) apps, and respond decisively to competitors.***"

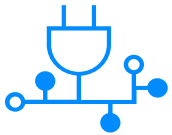
Forrester



Supply chain impacts manufacturing



Predict customer buying pattern



Utilities prevent power outage

Telecoms deliver continuous QoS

Reduce cyber threats

GROWTH OF DATA STREAMING EXPONENTIAL

Data overload with more connected devices

41.5B

More than 41.5 billion connected IoT devices are expected to be active by 2025. Majority of that will be in industrial and automotive sectors.

Source: Worldwide Global DataSphere IoT Device and Data Forecast, 2019-2023 (IDC #US45066919, May 2019)

79 ZB

By 2025, there will be 79ZB of data created by billions of IoT devices, causing organizations to re-evaluate their data governance, retention, and usage policies.

Source: IDC FutureScape: Worldwide IoT 2020 Predictions

40%

By 2024, 40% of manufacturers will use field asset IoT data to intelligently diagnose issues and resolve autonomously, improving unplanned downtime by 25%.

Source: IDC FutureScape: Worldwide IoT 2020 Predictions

CONNECTED DEVICES ARE EVERYWHERE

EDGE

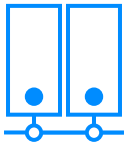
Miles driven

Wear-out tires

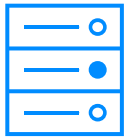
Wearing of the doors

Engine wear

DATA CENTER



Data capacity



Compute speed

Rising temperature

Capture data from all these sources and scale with a data streaming platform in a hybrid architecture

DATA MANAGEMENT CHALLENGES WITH DATA-IN-MOTION

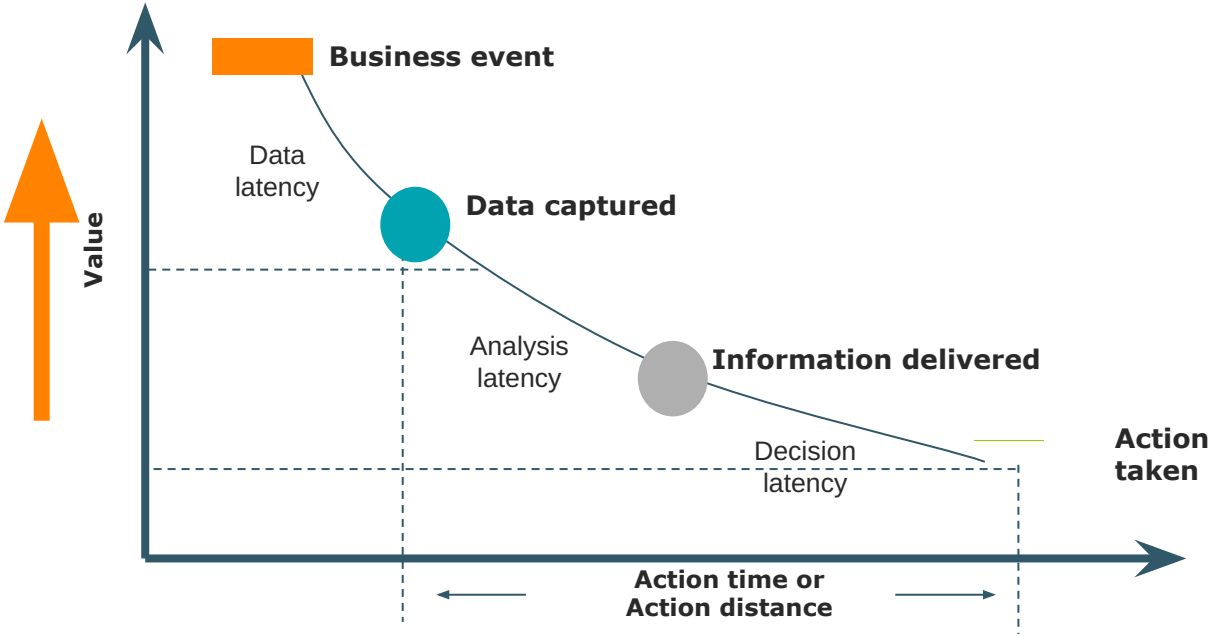
REAL-TIME ANALYTICS

Source: The ASEAN Appetite for Data-in-Motion, AOPG Insights, 2019

Your IoT / Streaming Analytics implementation is only as good as your ability to acquire and analyze the real-time data you have captured.

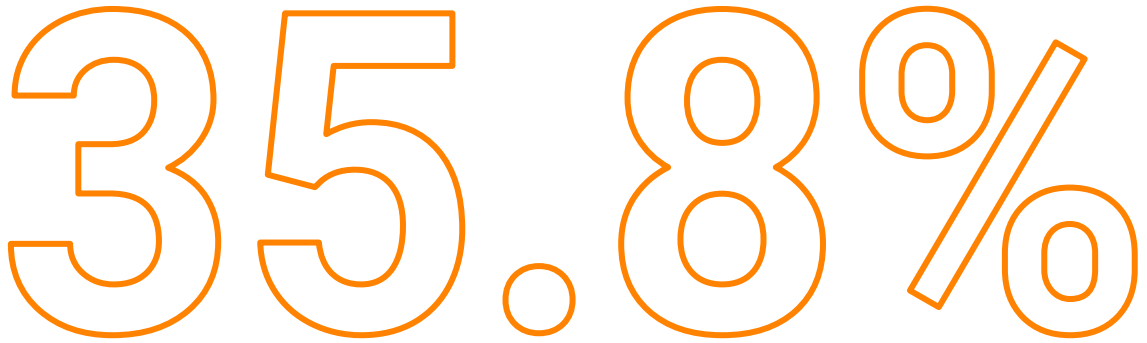
Data decays!

What does “real-time” mean to you?



CURRENT STREAMING ANALYTICS INVESTMENT

Only 35.8% of companies are invested in streaming analytics. 60.2% percent of them are still stuck on traditional historical analytics.



Source: A study conducted by Forrester Consulting on behalf of Cloudera, October 2019
Base: 157 manager-level and above decision makers with responsibility for streaming analytics

STREAMING PLATFORM VALUE

76% of companies say that streaming platforms are important for their current needs, but 89% say that they are important for their future needs as well

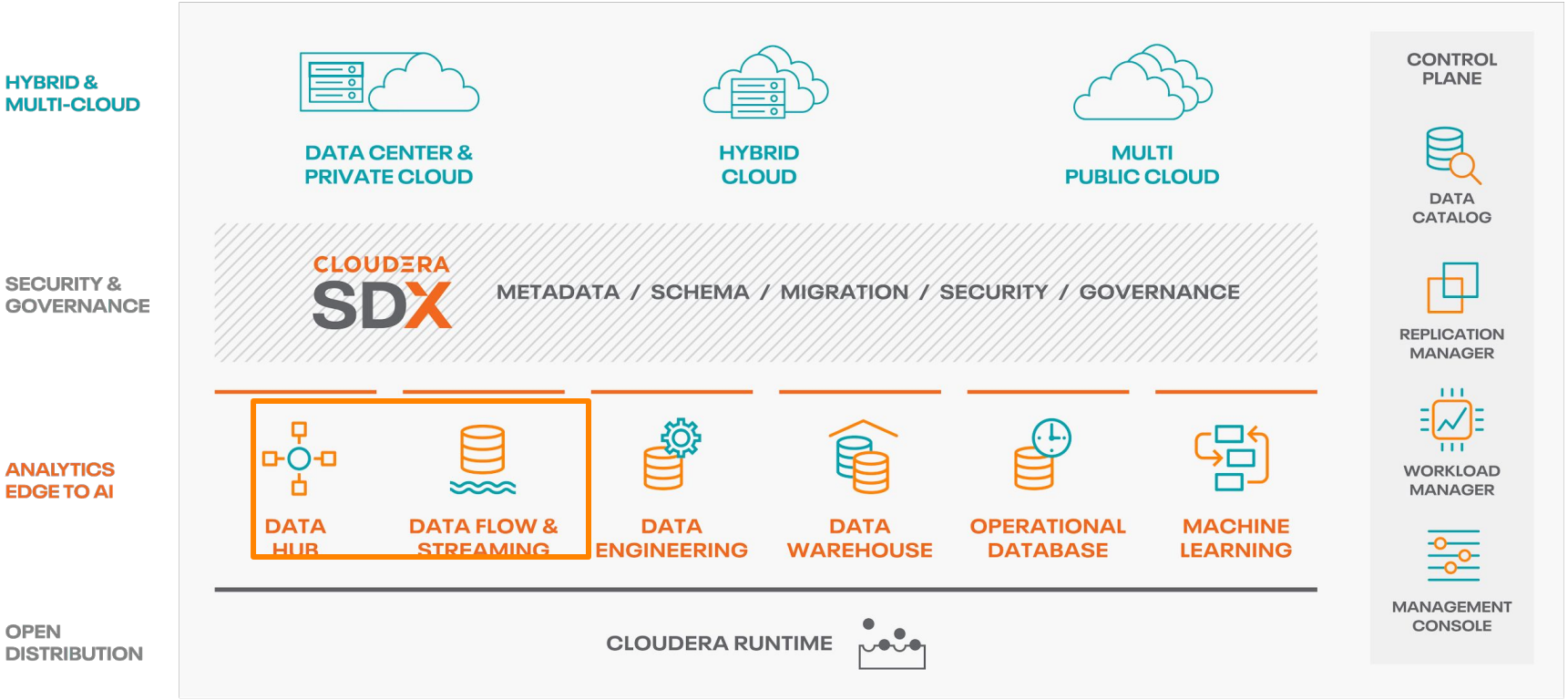
Source: A study conducted by Forrester Consulting on behalf of Cloudera, October 2019
Base: 157 manager-level and above decision makers with responsibility for streaming analytics

WHAT IS CLOUDERA DATAFLOW (CDF)?

Cloudera DataFlow (CDF) is a scalable, real-time streaming data platform that collects, curates, and analyzes data so customers gain key insights for immediate actionable intelligence.

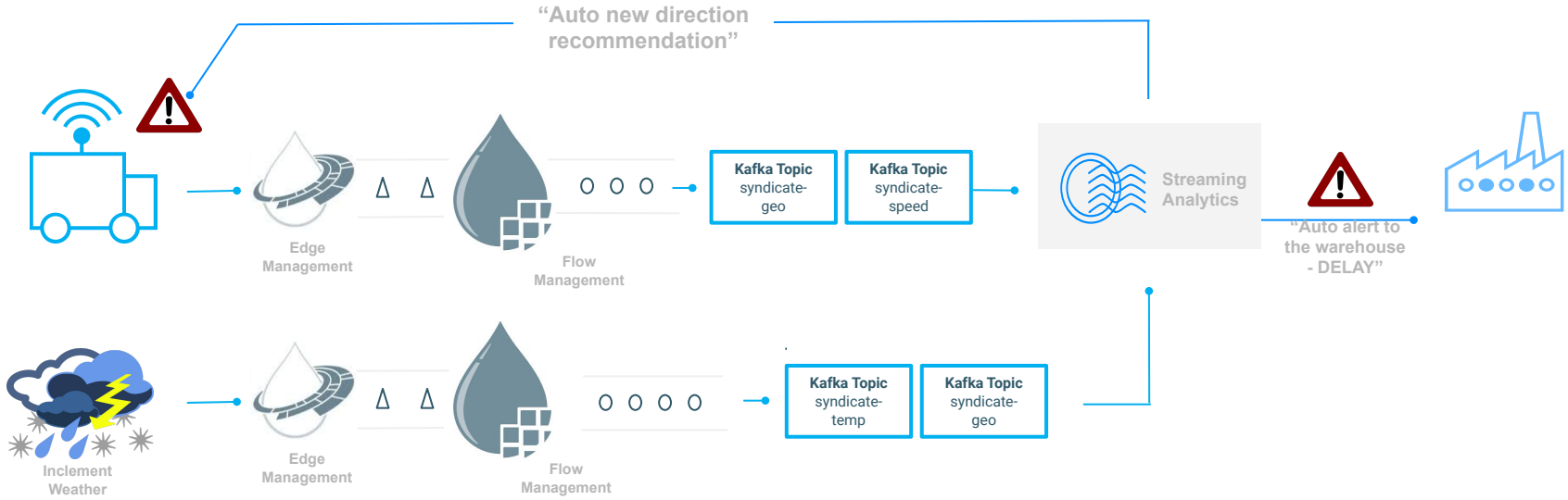
CLOUDERA DATAFLOW DATA-IN-MOTION PLATFORM

CLOUDERA DATA PLATFORM

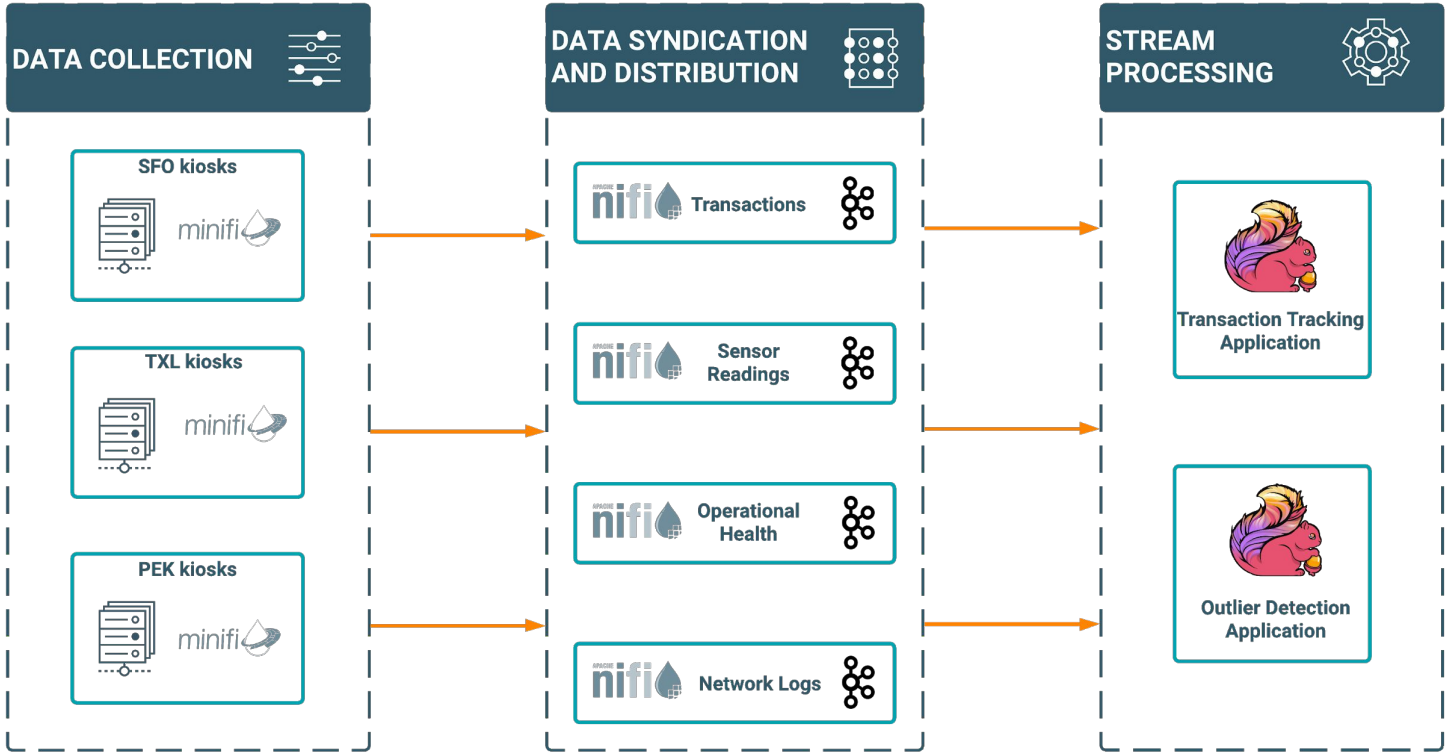


WHY CLOUDERA DATAFLOW?

Real-time insights drive business value



Data-in-Motion Components in Context



CLOUDERA DATAFLOW CUSTOMERS

Healthcare

Johns Hopkins University

Clearsense

vizient

Geisinger Health System
Redefining Boundaries

Yale New Haven Health

Rush University Medical Center

Public Sector

United States Census Bureau

Joint Improvised Threat Defeat Organization

Telecom

ooredoo

Globe

Telefónica

Magenta

Deutsche Telekom

vodafone

Technology

Micron

Symantec Endpoint Protection

Banking, Finance and Insurance

PROGRESSIVE

Rabobank

ZURICH

Bombay Stock Exchange

BANK BRI
Melayani Dengan Setulus Hati

ACI

Santander

ATB Financial

Manufacturing and Transportation

Trimble

NISSAN

NAVISTAR

Severstal

Travel & Leisure

HILTON WORLDWIDE

Utilities/O&G

centrica

Int'l Gov

دبي الذكية
SMART DUBAI

Cloudera DataFlow Use Cases

Data Movement

Optimize resource utilization by moving data between data centers or between on-premises and cloud infrastructures

e.g. intercontinental data exchange

Logging Modernization

Optimize log analytics solutions by with CDF in simplifying log ingestion from the edge, reducing costs and gaining key analytics

e.g. Splunk / Logstash offload

Streaming analytics insights

Make key business decisions by analyzing streaming data for complex patterns, gaining actionable intelligence etc.

e.g. Fraud detection, Network threat analysis, app monitoring, Clickstream analysis

360° view of customer

Ingest, transform and combine customer data from multiple sources into a single data view / lake

e.g. Real-time customer offers, Loan approvals

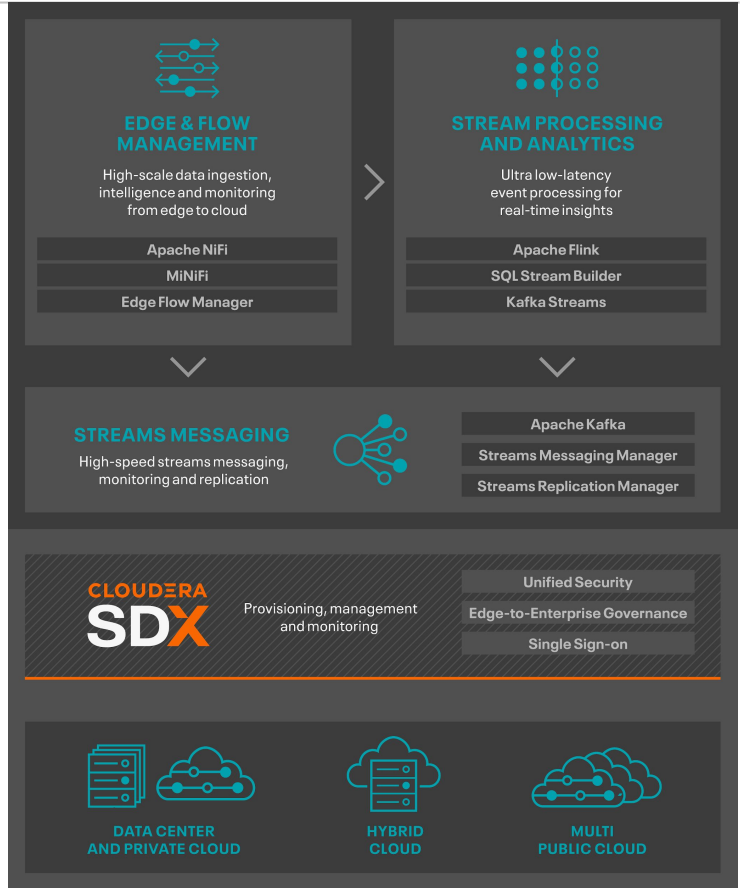
IoT & Edge use cases

e.g. Predictive Maintenance, Asset Tracking / Monitoring, Patient Monitoring, Quality Processes, Fleet Management, Connected Cars and more

Enterprise data management

Managing massive volumes of high-velocity data to/from legacy systems, ETL tools and other data stores

e.g. Flume offload, ETL replacement, payment data processing, integration with Oracle

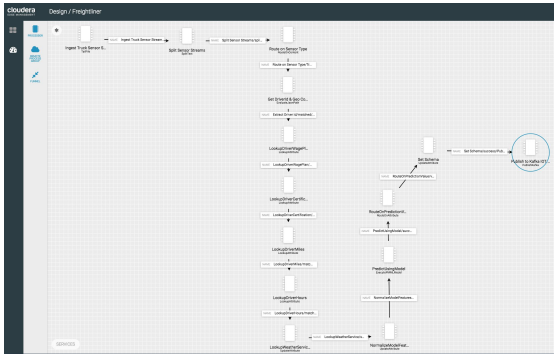


Edge & Flow Management

Cloudera Edge Management

Manage, control, and monitor the edge for all your streaming and IoT initiatives

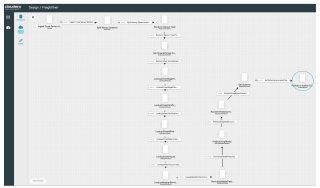
Edge Flow Manager



APACHE NIFI minifi

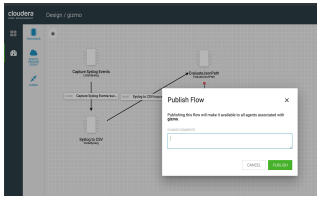
A lightweight edge agent that implements the core features of Apache NiFi, focusing on data collection and processing at the edge

Flow Authorship



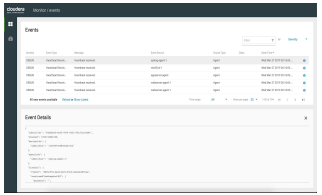
- Small footprint agent with MiNiFi
- Java and C++ agents
- Rich edge processors (edge collection & processing)
- End to end lineage and security

Flow Deployment

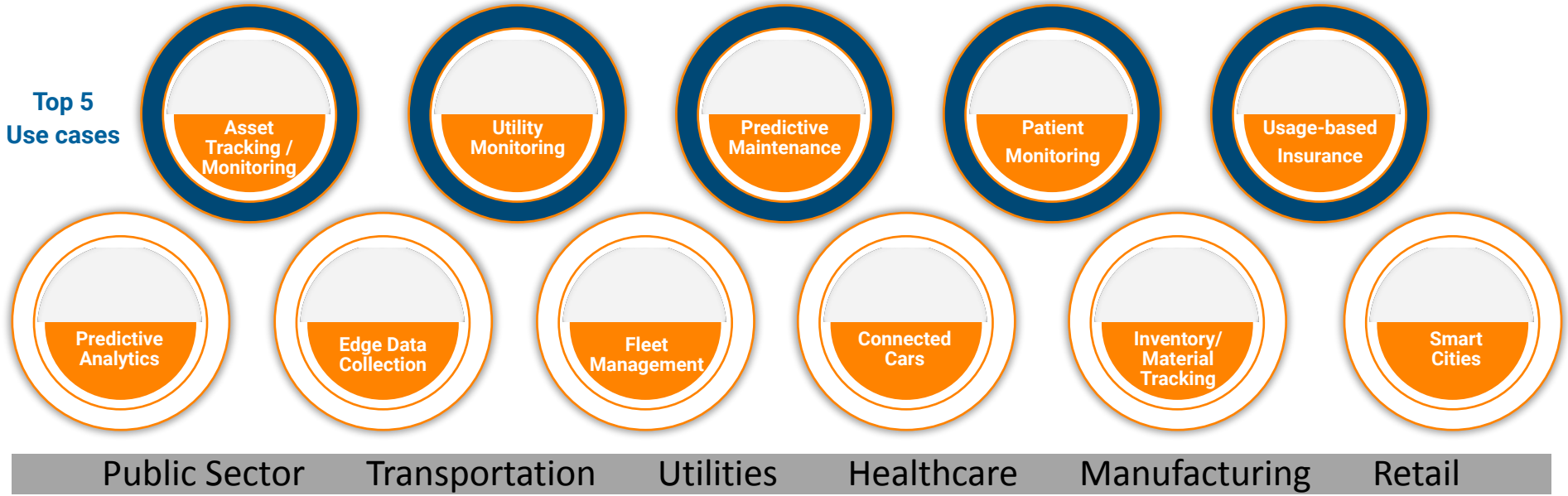


- Central Command and Control
- Design and deploy to thousands of agents
- Edge Applications lifecycle management
- Multitenancy with Agent classes
- Native integration with other CDF services

Flow Monitoring



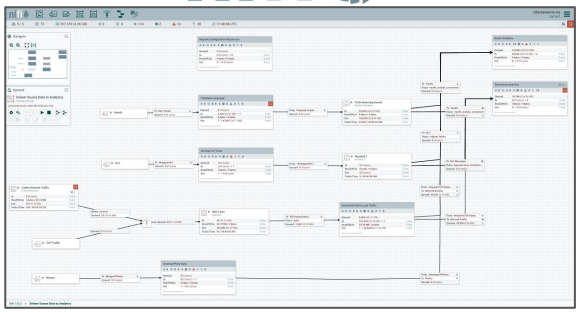
COMMON IoT USE CASES BY INDUSTRY



- IoT is a $1.13T market opportunity in 2021.
- Americas - $329B IoT spending. Manufacturing and Transportation are top industries, accounting for 26% of total spending.
- APAC - $500B IoT spending. Manufacturing, Utilities and Transportation are top industries.
- EMEA - $264B IoT spending. Manufacturing is top industry, powered by Industry 4.0 initiatives.
- Worldwide IoT Analytics and Information Management Market = $573M

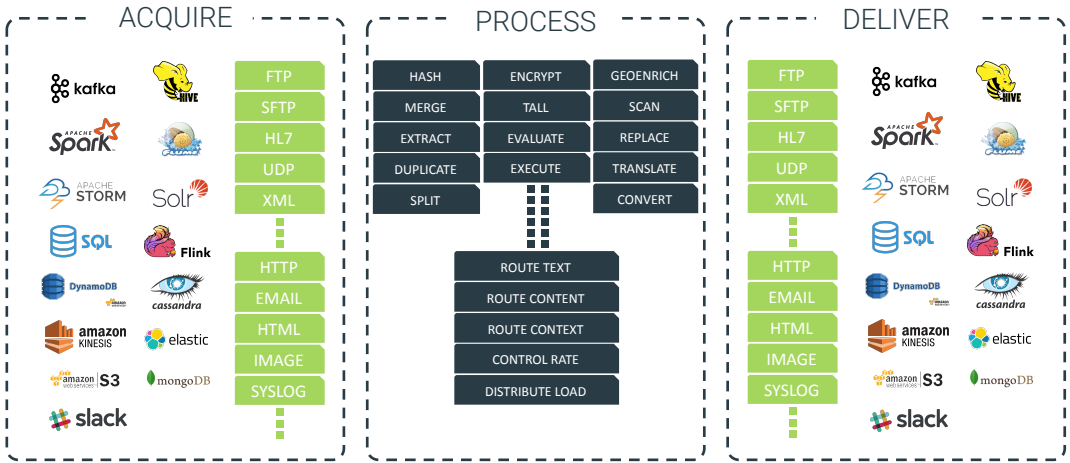
Cloudera Flow Management

Ingest and manage data from edge-to-cloud using a no-code interface



APACHE NIFI registry

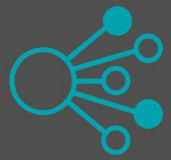
Advanced tooling to industrialize flow development (Flow Development Life Cycle)



- Over 300 pre-built processors
- Easy to build your own processors
- Parse, enrich & apply schema
- Filter, Split, Merge & Route
- Throttle & Backpressure
- Guaranteed delivery
- Full data provenance
- Eco-system integration

STREAMS MESSAGING

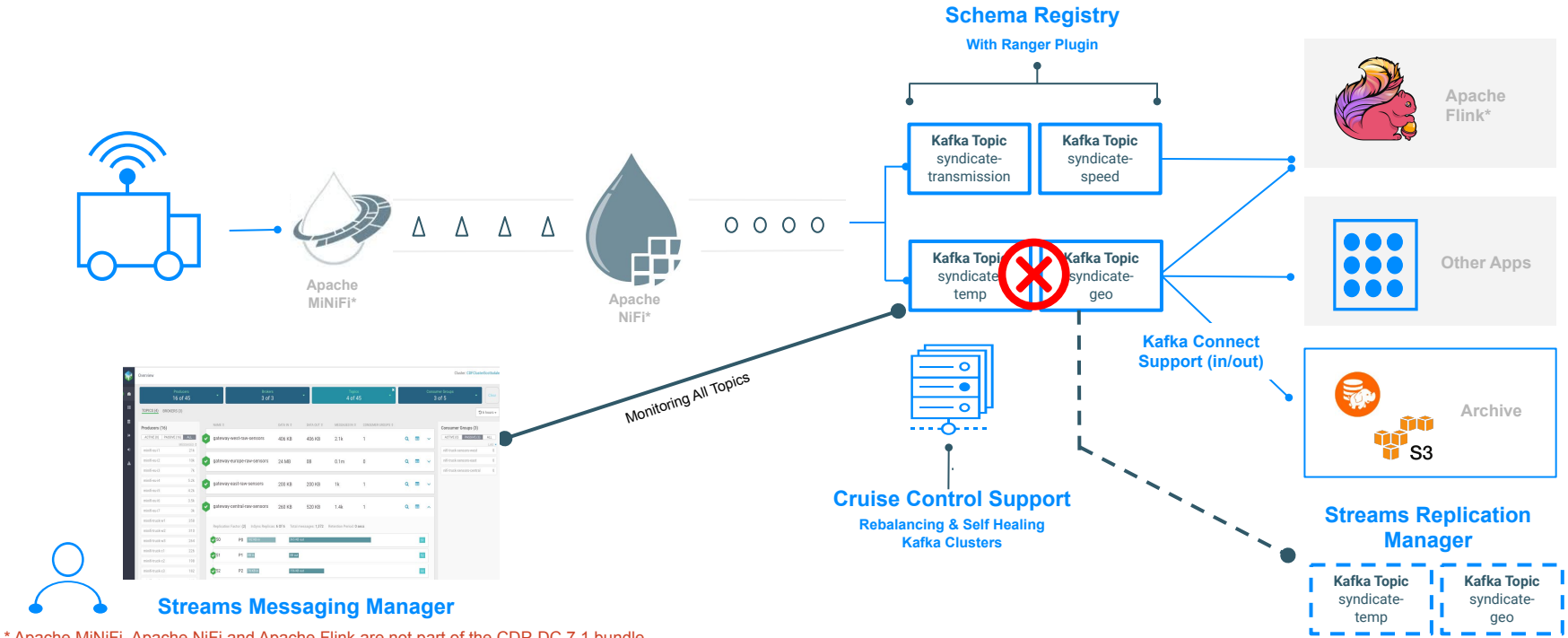
High-speed streams messaging, monitoring and replication



- Apache Kafka
- Streams Messaging Manager
- Streams Replication Manager

Streams Messaging

Comprehensive Streams Messaging (now available in CDP)



* Apache MiNiFi, Apache NiFi and Apache Flink are not part of the CDP-DC 7.1 bundle

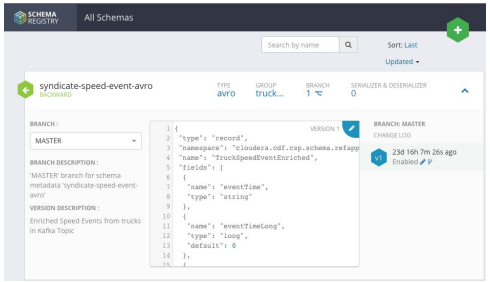
Comprehensive Streams Messaging

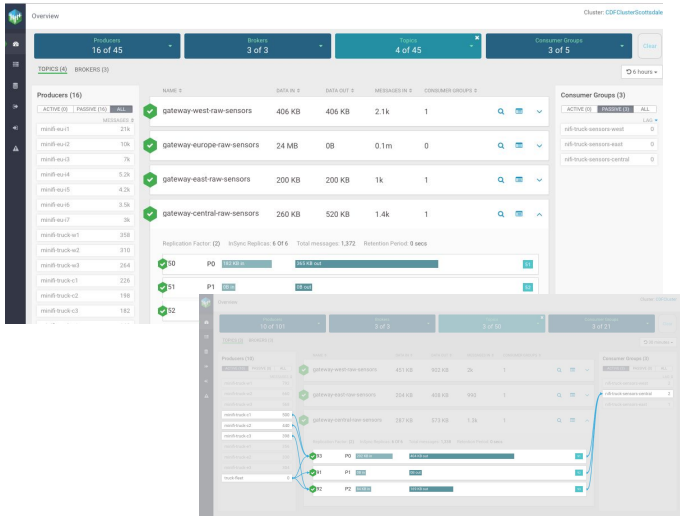
Extend streams messaging services for Schema Mgmt, Replication & Monitoring

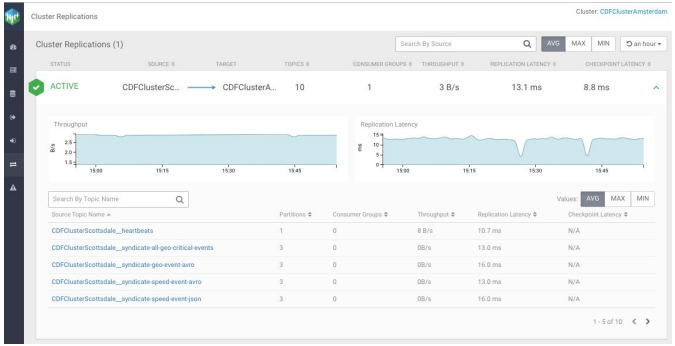
Schema Registry
Kafka Schema Governance

Streams Messaging Manager
Management & Monitoring Service
for all of your Kafka clusters

Streams Replication Manager
Kafka Replication Service powered by
MirrorMaker2





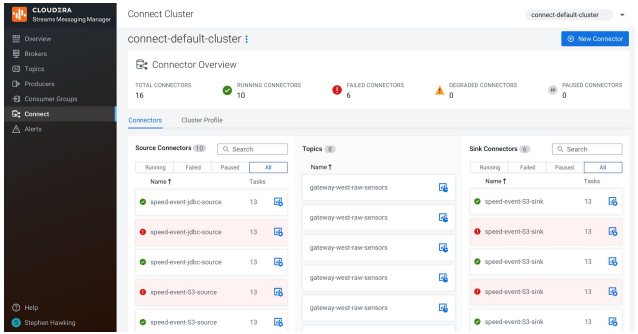


Comprehensive Streams Messaging - New capabilities

New Kafka Services Delivered with CDP PvC Base

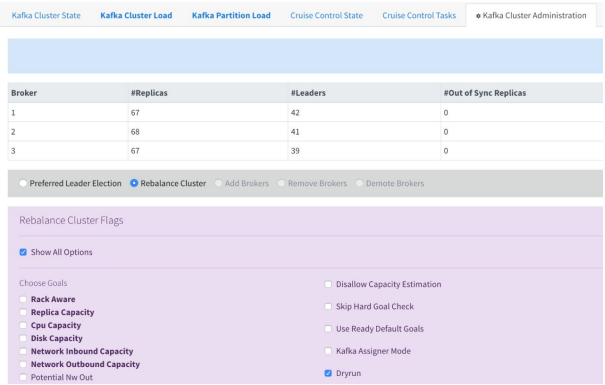
Kafka Connect Support

Simple Data Movement
In/Out of Kafka



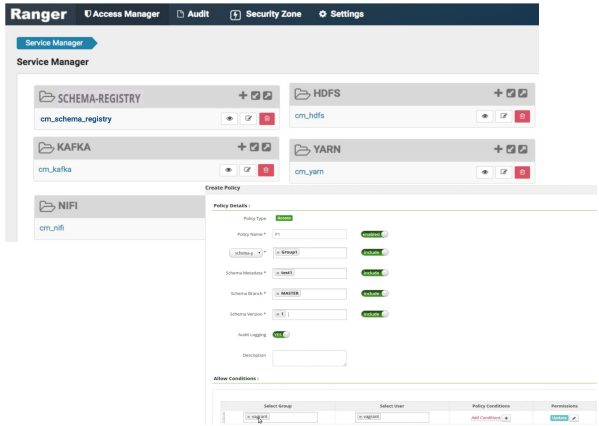
Cruise Control Support

Intelligent Rebalancing
& Self-Healing of your
Kafka Clusters



Schema Registry Ranger Plugin

Improved ACL and Audit for
Kafka and Schema Registry



Stream Processing & Analytics

Next Generation Stream Processing & Analytics

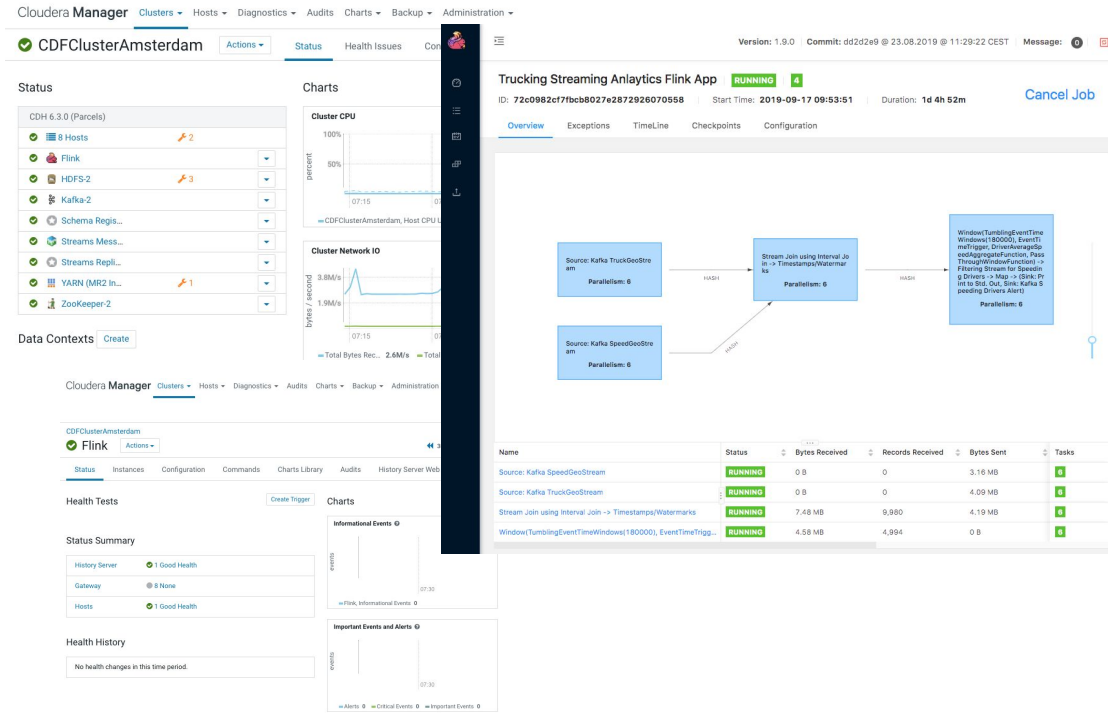
Low latency stateful stream processing

Low Latency

Event Processing

Real-Time Insights

- Advanced features - late arriving data, checkpointing, event time processing, Exactly Once Processing
- We support Apache Flink along with other stream processing engines like Kafka Streams and Spark Structured Streaming.



Shared Data Experience (SDX)

SDX: CONSISTENT SECURITY AND GOVERNANCE

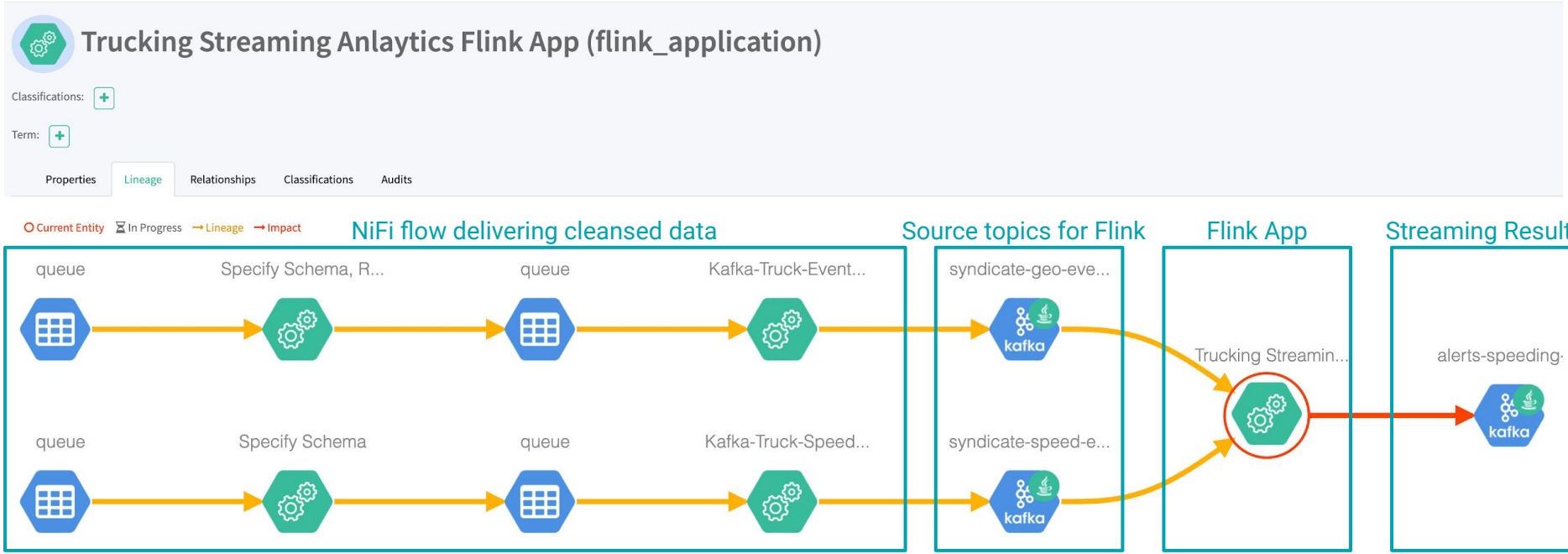
Built for multi-functional analytics anywhere

- **Security:** dynamic fine-grained access control applied consistently across all clouds and data center. Includes full stack encryption and key management
- **Governance:** enterprise-grade auditing, lineage, and governance capabilities applied across the platform with rich extensibility for partner integrations
- **Metadata:** establish information assets for increased usability, trust and value leveraging all metadata (structural, operational, business and social)
- **Catalog:** single pane of glass for administration and use of data assets spanning all analytics and deployments
- **Intelligence:** insight into how data, metadata and analytics are being used in the platform, leading to recommendations and automations for optimization

BASED ON COMMUNITY OPEN SOURCE COMPONENTS

Ensuring consistent security and governance

Streaming End-To-End Lineage



CLOUDERA DATAFLOW DATA-IN-MOTION PLATFORM

Customer Successes

Clearsense®

Improving Healthcare with SMART data

CHALLENGE

Combine multi-format data streams, with hundreds of sources, into one platform

- Needed a platform that could combine multi-format data streaming
- Data scarcity & latency problems
- Machine learning & data science

IMPACT

Lack of medical expertise around patient care, post surgery

- Patient Code Blue status
- Possible cardiac arrest 4–6 hours post surgery

SOLUTION

Cloud-based systems architected to deliver SMART data, using HDP and CDF

- First to deliver SMART real-time streaming data
- Clearsense's Inception™ product enables fast decisions for clinicians
- Customers have access to all data sources with HDP & CDF

RESULT

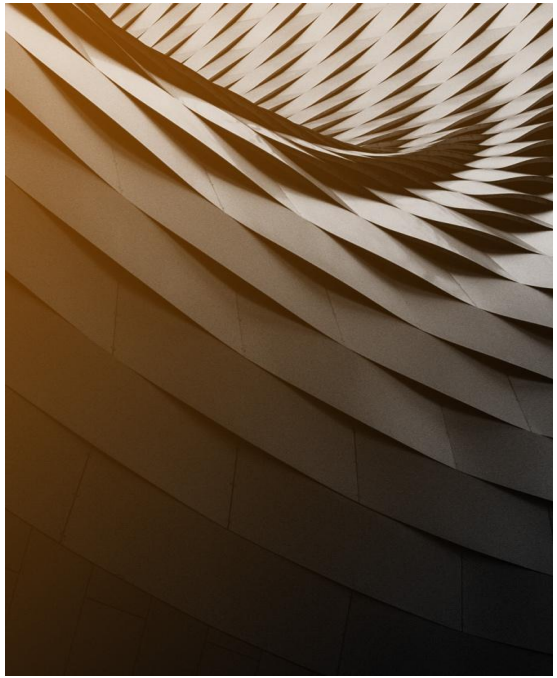
Mission-critical data and relevant insight for 2,000 rural providers

- Mission critical data is now available for doctors to make critical decisions
- Cost efficiencies led to access for 2,000 rural providers
- Real-time data helps prevent "Code Blue"

Photo by rawpixel on Unsplash

OOREDOO KUWAIT + CLOUDERA

"REAL-TIME DATA ENHANCES CUSTOMER VALUE MANAGEMENT"



CHALLENGE

Increased CDR data volume and need to store them longer term due to compliance.

Need better customer segmentation to provide better services.

SOLUTION

Tackled various critical use cases using Cloudera's platform and Streams Messaging within a single environment

OUTCOMES

Improved customer experience by empowering customer service and marketing teams with accurate insight

Drive accurate policies across the firm.

SEVERSTAL

100K

Tonnes of additional metal processing per year

CHALLENGE

Millions of data points generated by factories and refineries required operational modernization

SOLUTION

Using NiFi and MiNiFi to collect millions of messages per minute from IoT devices and sensor data from the machinery producing the steel

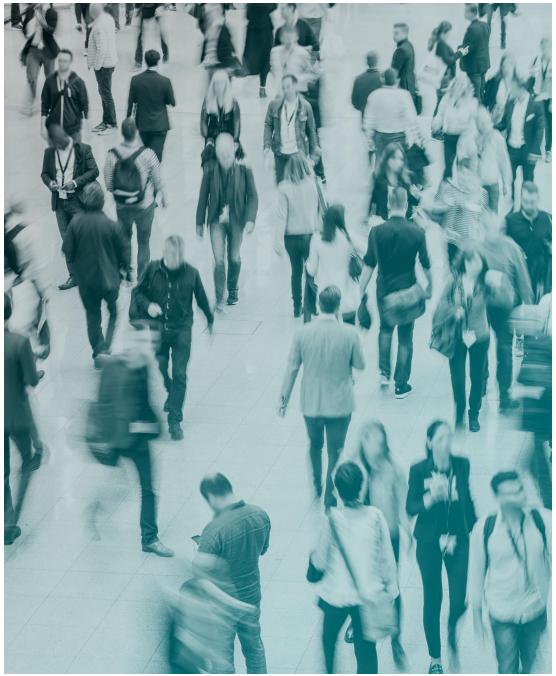
OUTCOMES

AI has automated speed and adjustments of the assembly line - increasing speed by more than 5%

U.S. Census

330M

Households served



CHALLENGE

To share data across teams and the government with complete lineage, security and governance

SOLUTION

CDF to ingest data and provide real-time analytics. HDP serves as the data lake and repository for the massive amount of data collected

OUTCOMES

Improved insights that help the federal government in emergency services, education and healthcare

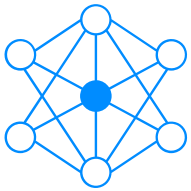
CDF in the hybrid world

ENTERPRISES WANT A HYBRID APPROACH BUT STRUGGLE WITH CLOUD INTEGRATION

Data Streaming
On-Premises

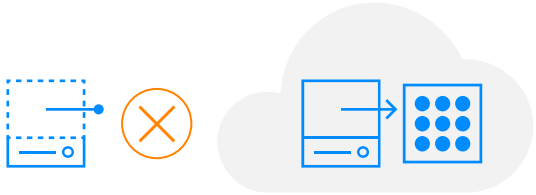
Data Streaming
in the Cloud

KEY CHALLENGES



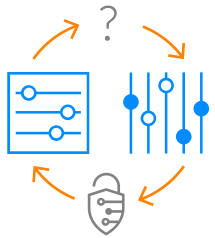
COMPLEXITY

Cloud configuration and infrastructure setups are complex



NOT COMPATIBLE

Different data streaming tools, applications cannot work together across on-premises and cloud

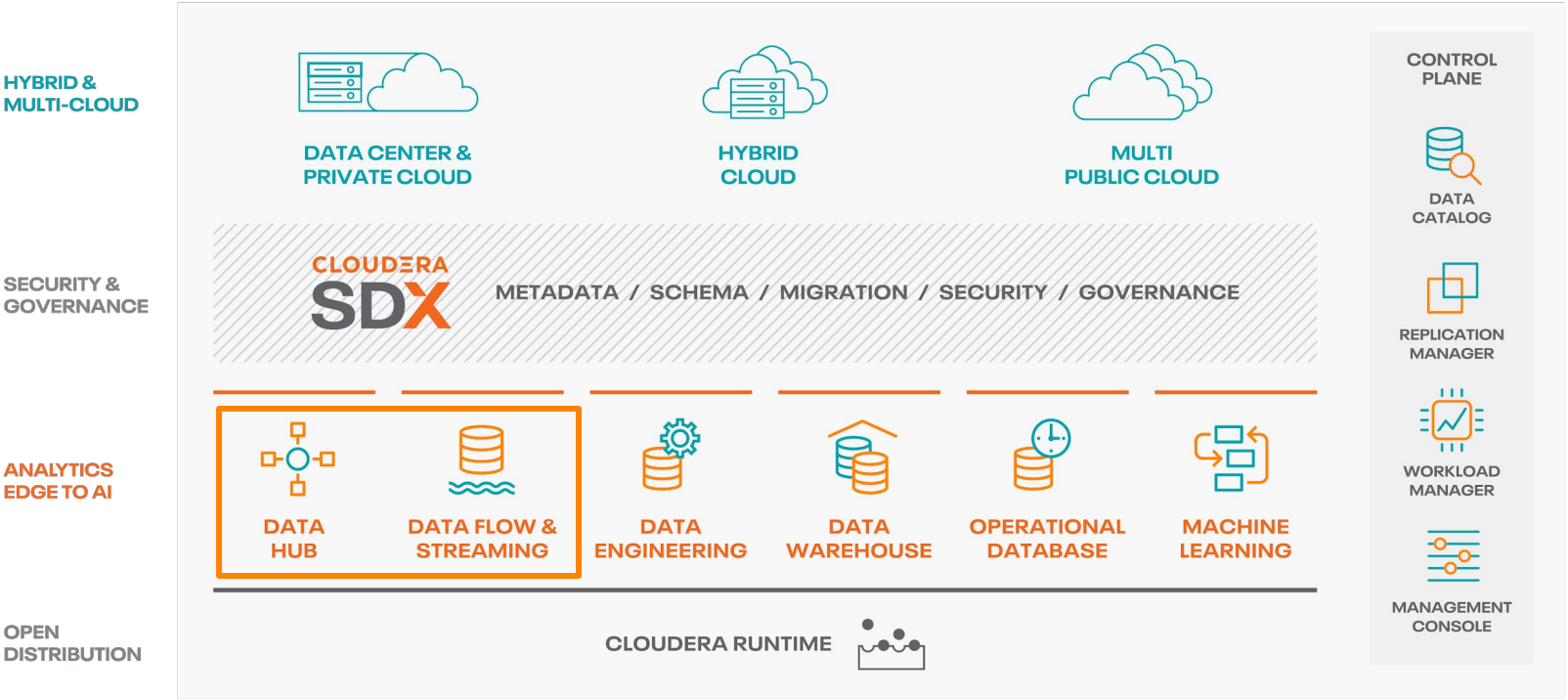


RISKING COMPLIANCE AND SECURITY

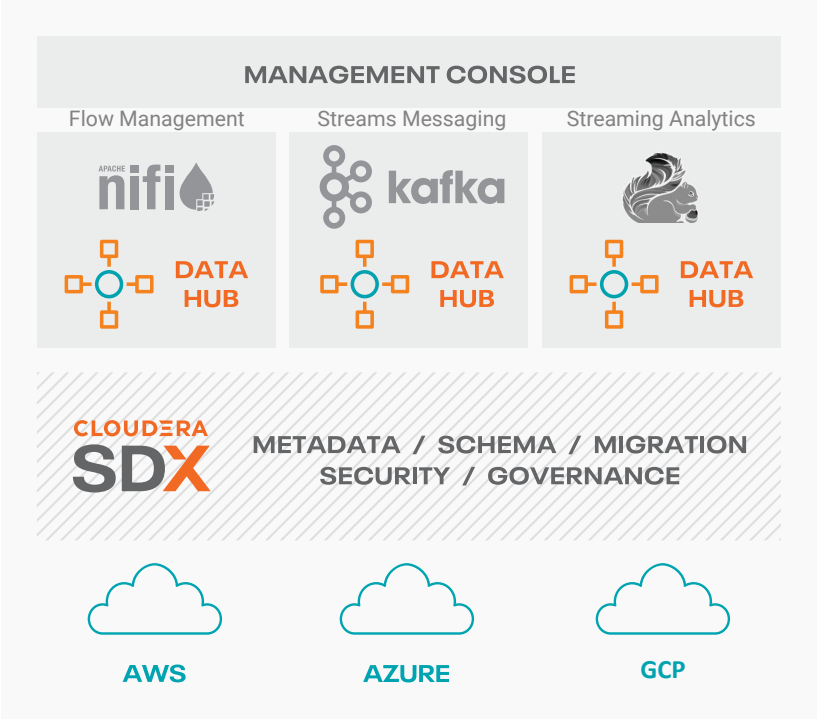
Data Governance is an afterthought, even with sensitive data

Lack of authentication and policy control

CLOUDERA DATA PLATFORM



CDF in CDP Public Cloud Data Hub



Web service hosted and managed by Cloudera

Hosted in the customer's cloud environment, but managed by the CDP Management Console

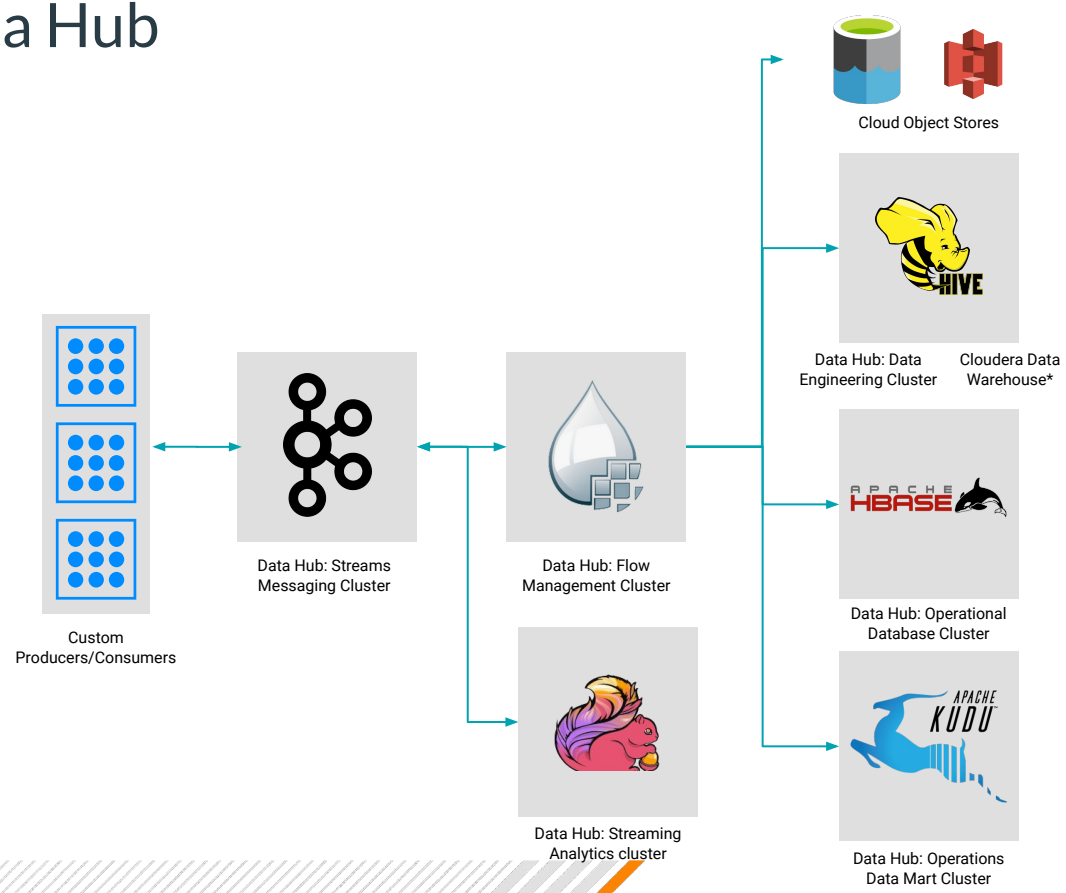
Shared Data Experience (SDX) technologies form a secure and governed data lake backed by object storage (S3, ADLS, GCS)

CDP services are optimized for the elastic compute & 'always-on' storage services provided by the customer's chosen cloud provider

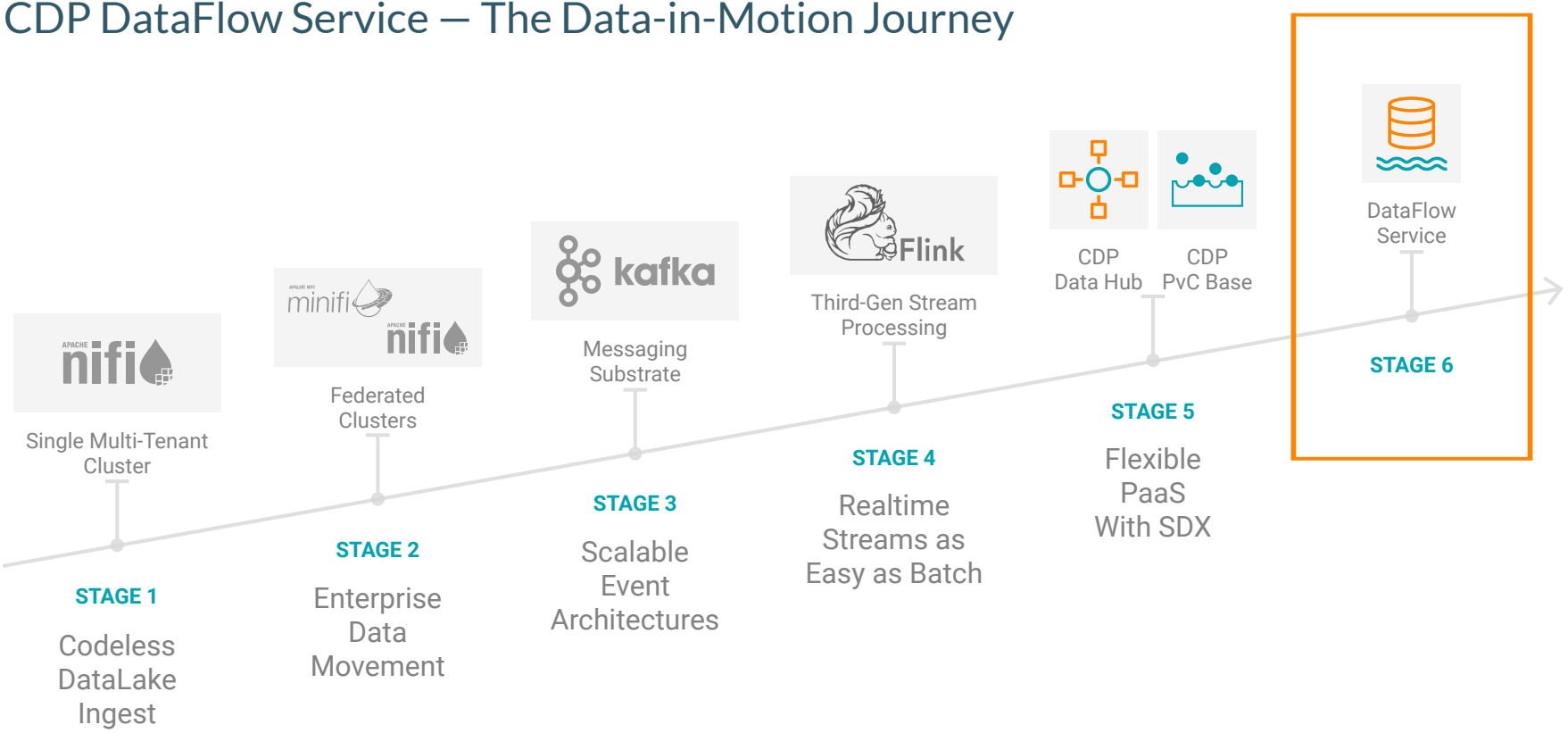
CDF Use Cases in CDP Data Hub

Cloud Native Streaming / CDP Ingest

- Enables cloud native streaming applications
- Use NiFi or custom producers/consumers to interact with Kafka
- Use NiFi to power CDP Ingest
  - Hive
  - HBase
  - Kudu
  - Cloud object stores

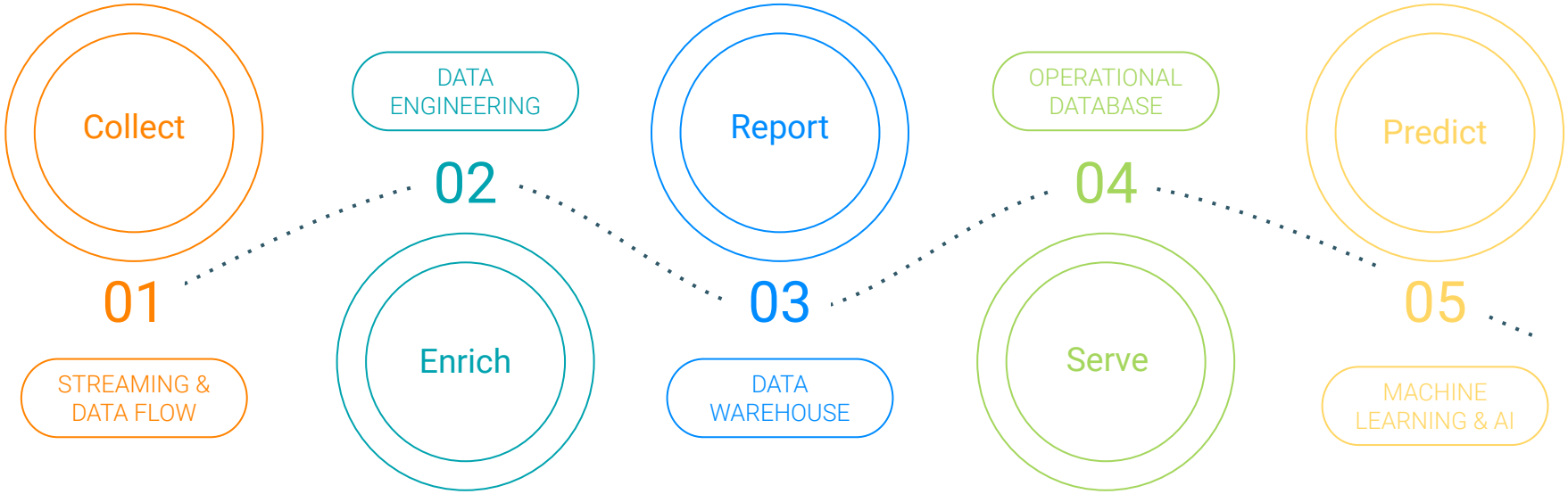


CDP DataFlow Service — The Data-in-Motion Journey



CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

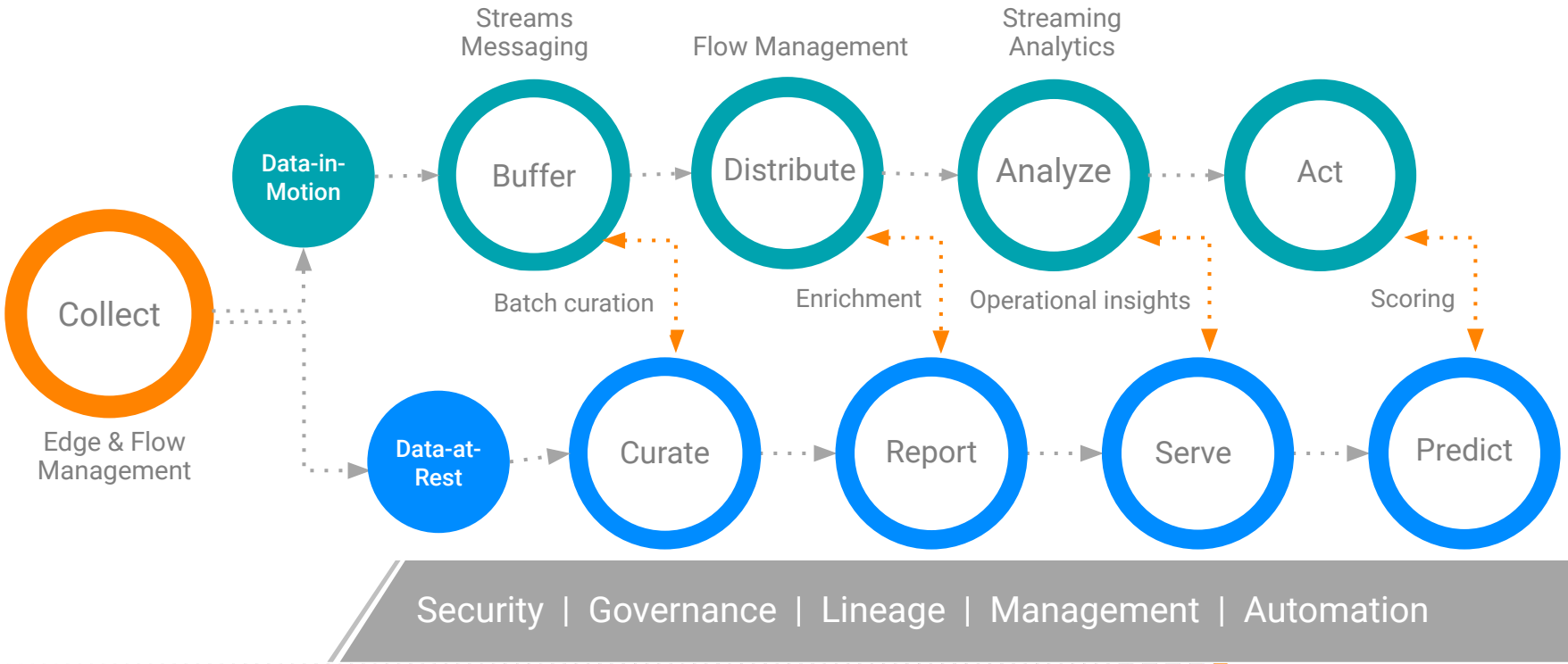
Manage and secure the data lifecycle in any cloud or datacenter



CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

Complete and Connected Data Lifecycle



Cloudera is a Strong Performer in Forrester Wave for Streaming Analytics Q3 2019

Cloudera brings data management discipline to streaming analytics

"Cloudera DataFlow (CDF) is more than streaming analytics. It brings to streaming analytics such features as data provenance and other data management capabilities usually found only in batch-oriented big data and data warehouse platforms."
"Cloudera offers strengths in streaming data platforms, management, security, development, extensibility, and deployment."

[Download the full report]

Cloudera DataFlow in the news

Cloudera is a Strong Performer in the Forrester Streaming Analytics Wave, Q3 2019

"Cloudera DataFlow (CDF) is more than streaming analytics. It brings to streaming analytics such features as data provenance and other data management capabilities usually found only in batch-oriented big data and data warehouse platforms."

"Cloudera offers strengths in streaming data platforms, management, security, development, extensibility, and deployment."

Cloudera DataFlow recognized as an IoT Data Management Solution across cloud, on-premises and systems-based services

- Now Tech: IoT Data Management Solutions, Q2 2019

Cloudera DataFlow recognized as an enterprise-focused Edge Intelligence Fabric

- A decoder ring for Edge Computing, Forrester, Dec 2019

Apache Flink Powers Cloudera's New Streaming Analytics Product - datanami, Jan 15, 2020

KEY DIFFERENTIATORS

Stream to Cloud – Extend the same on-premises streaming capabilities to the cloud with full support for multi-cloud and hybrid cloud models

340+ pre-built processors – Only product to offer such comprehensive connectivity to a wide range of data sources from edge to cloud

Democratize access to real-time data – Enable data analysts and other personas to quickly build streaming applications with just SQL

Enterprise-Grade Security & Governance – Deploy your streaming applications with confidence and trust with Cloudera SDX offering unified security and governance across the entire platform

Comprehensive streaming platform – Only vendor to offer a open and comprehensive streaming platform for real-time data ingestion and processing to produce prescriptive and predictive analytics

Download these assets today


CLOUDERA

WHITE PAPER

CLOUDERA DELIVERS THE BEST KAFKA ECOSYSTEM TODAY

Serving Hundreds of Customers Globally

CLOUDERA

WHITE PAPER

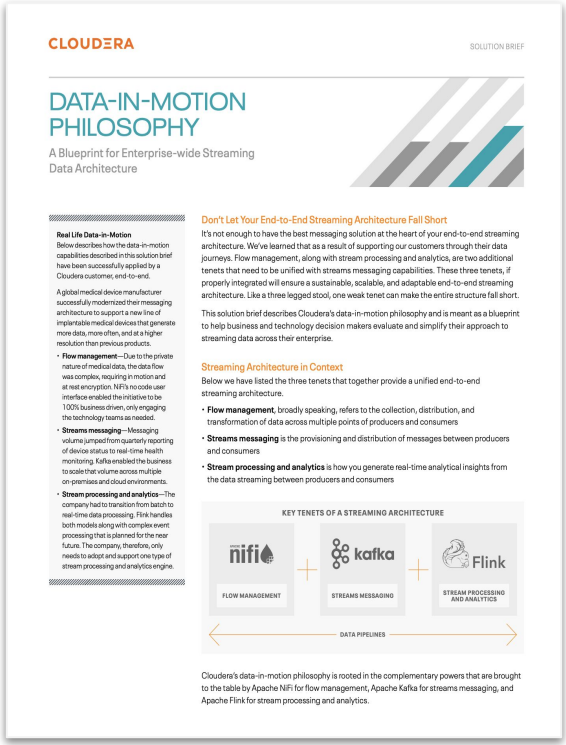
The Complete Edge-to-Cloud Streaming Data Platform

This paper has focused primarily on the streams messaging aspects of the Kafka ecosystem with regard to how to secure, monitor, balance, and replicate large scale Kafka environments across on-premises, hybrid, private, and public cloud environments. The Data-in-Motion reference architecture diagram in Figure 7 below, puts this all in perspective.

A DATA-IN-MOTION REFERENCE ARCHITECTURE

Figure 7

As a result of supporting our customers through their data journeys, we've learned that it is not enough to have the best messaging solution at the heart of your end-to-end streaming architecture. As represented in the diagram above, flow management, along with stream processing and analytics, are two additional tenets that need to be unified with streams messaging capabilities. These three tenets, if properly integrated, will ensure a sustainable, scalable and adaptable end-to-end streaming architecture and is the basis to our data-in-motion philosophy.


CLOUDERA

SOLUTION BRIEF

DATA-IN-MOTION PHILOSOPHY

A Blueprint for Enterprise-wide Streaming Data Architecture

Real Life Data-in-Motion

Below describes how the data-in-motion capabilities described in this solution brief have been successfully applied by a Cloudera customer, end-to-end.

A global medical device manufacturer successfully modernized their messaging architecture to support a new line of implantable medical devices that generate more data, more often, and at a higher resolution than previous products.

- **Flow management**—Due to the private nature of medical data, the data flow was complex, requiring in motion and at rest encryption. NiFi's no code user interface enabled the initiative to be 100% business driven, only engaging the technology teams as needed.
- **Streams messaging**—Messaging volume jumped from quarterly reporting of device status to real-time health monitoring. Kafka enabled the business to scale that volume across multiple on-premises and cloud environments.
- **Stream processing and analytics**—The company had to transition from batch to real-time data processing. Flink handles both models along with complex event processing that is planned for the near future. The company, therefore, only needs to adopt and support one type of stream processing and analytics engine.

Don't Let Your End-to-End Streaming Architecture Fall Short

It's not enough to have the best messaging solution at the heart of your end-to-end streaming architecture. We've learned that as a result of supporting our customers through their data journeys. Flow management, along with stream processing and analytics, are two additional tenets that need to be unified with streams messaging capabilities. These three tenets, if properly integrated will ensure a sustainable, scalable, and adaptable end-to-end streaming architecture. Like a three legged stool, one weak tenet can make the entire structure fall short.

This solution brief describes Cloudera's data-in-motion philosophy and is meant as a blueprint to help business and technology decision makers evaluate and simplify their approach to streaming data across their enterprise.

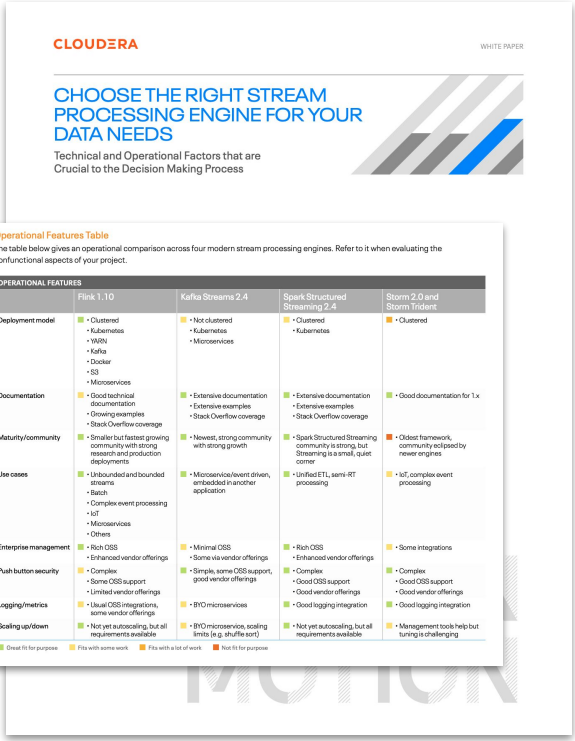
Streaming Architecture in Context

Below we have listed the three tenets that together provide a unified end-to-end streaming architecture.

- **Flow management**, broadly speaking, refers to the collection, distribution, and transformation of data across multiple points of producers and consumers
- **Streams messaging** is the provisioning and distribution of messages between producers and consumers
- **Stream processing and analytics** is how you generate real-time analytical insights from the data streaming between producers and consumers

KEY TENETS OF A STREAMING ARCHITECTURE

Cloudera's data-in-motion philosophy is rooted in the complementary powers that are brought to the table by Apache NiFi for flow management, Apache Kafka for streams messaging, and Apache Flink for stream processing and analytics.


CLOUDERA

WHITE PAPER

CHOOSE THE RIGHT STREAM PROCESSING ENGINE FOR YOUR DATA NEEDS

Technical and Operational Factors that are Crucial to the Decision Making Process

Operational Features Table

The table below gives an operational comparison across four modern stream processing engines. Refer to it when evaluating the nonfunctional aspects of your project.

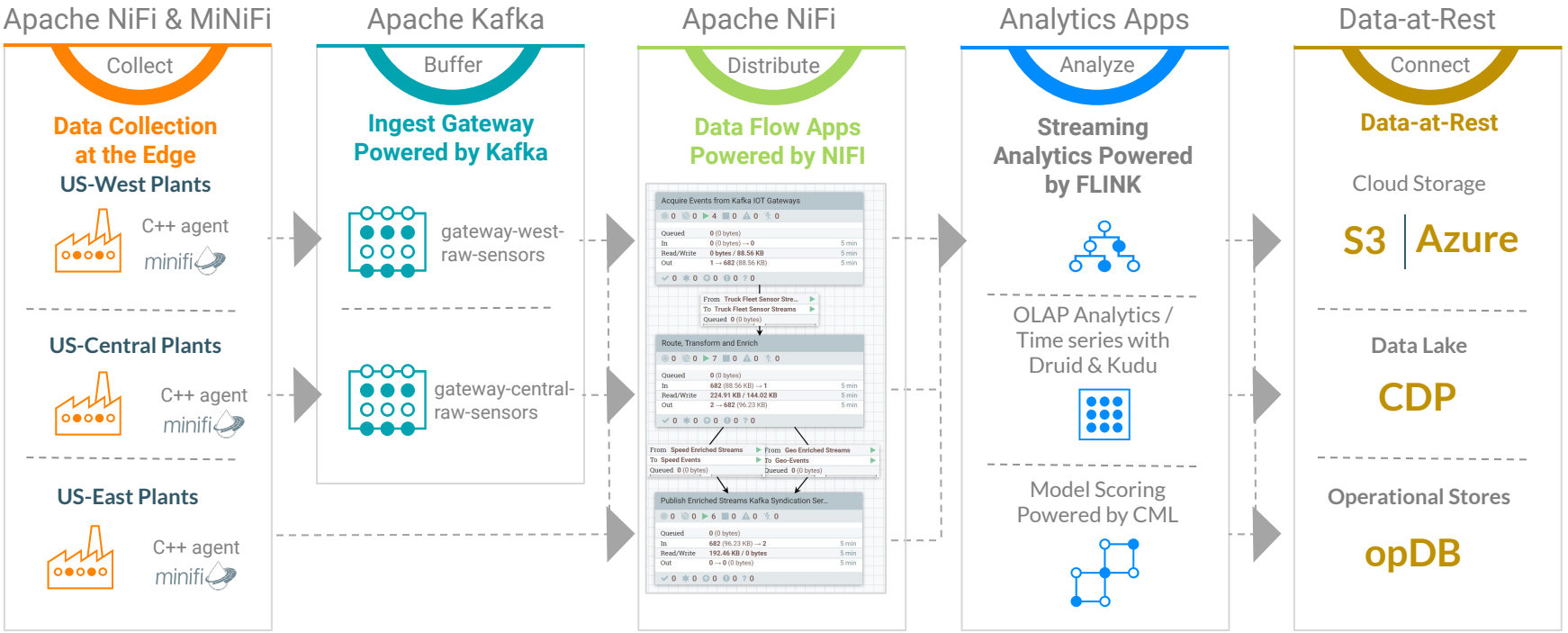
OPERATIONAL FEATURES	Flink 1.10	Kafka Streams 2.4	Spark Structured Streaming 2.4	Storm 2.0 and Storm Trident
Deployment model	• Clustered • Kubernetes • YARN • Kafka • Docker • SS • Microservices	• Not clustered • Kubernetes • Microservices	• Clustered • Kubernetes	• Clustered
Documentation	• Good technical documentation • Growing examples • Stack Overflow coverage	• Extensive documentation • Extensive examples • Stack Overflow coverage	• Extensive documentation • Extensive examples • Stack Overflow coverage	• Good documentation for 1.x
Maturity/community	• Smaller but fastest growing community with strong research and production deployments	• Newest, strong community with strong growth	• Spark Structured Streaming community is strong, but Streaming is a small, quiet corner	• Oldest framework, community eclipsed by newer engines
Use cases	• Unbounded and bounded streams • Batch • Complex event processing • IoT • Microservices • Others	• Microservice/event driven, embedded in another application	• Unified ETL, semi-RT processing	• IoT, complex event processing
Enterprise management	• Rich OSS • Enhanced vendor offerings	• Minimal OSS • Some via vendor offerings	• Rich OSS • Enhanced vendor offerings	• Some integrations
Push button security	• Complex • Some OSS support • Limited vendor offerings	• Simple, some OSS support, good vendor offerings	• Complex • Good OSS support • Good vendor offerings	• Complex • Good OSS support • Good vendor offerings
Logging/metrics	• Usual OSS integrations, some vendor offerings	• BYO microservices	• Good logging integration	• Good logging integration
Scaling up/down	• Not yet autoscaling, but all requirements available	• BYO microservices, scaling limits (e.g. shuffle sort)	• Not yet autoscaling, but all requirements available	• Management tools help but tuning is challenging

Great fit for purpose · Fits with some work · Fits with a lot of work · Not fit for purpose

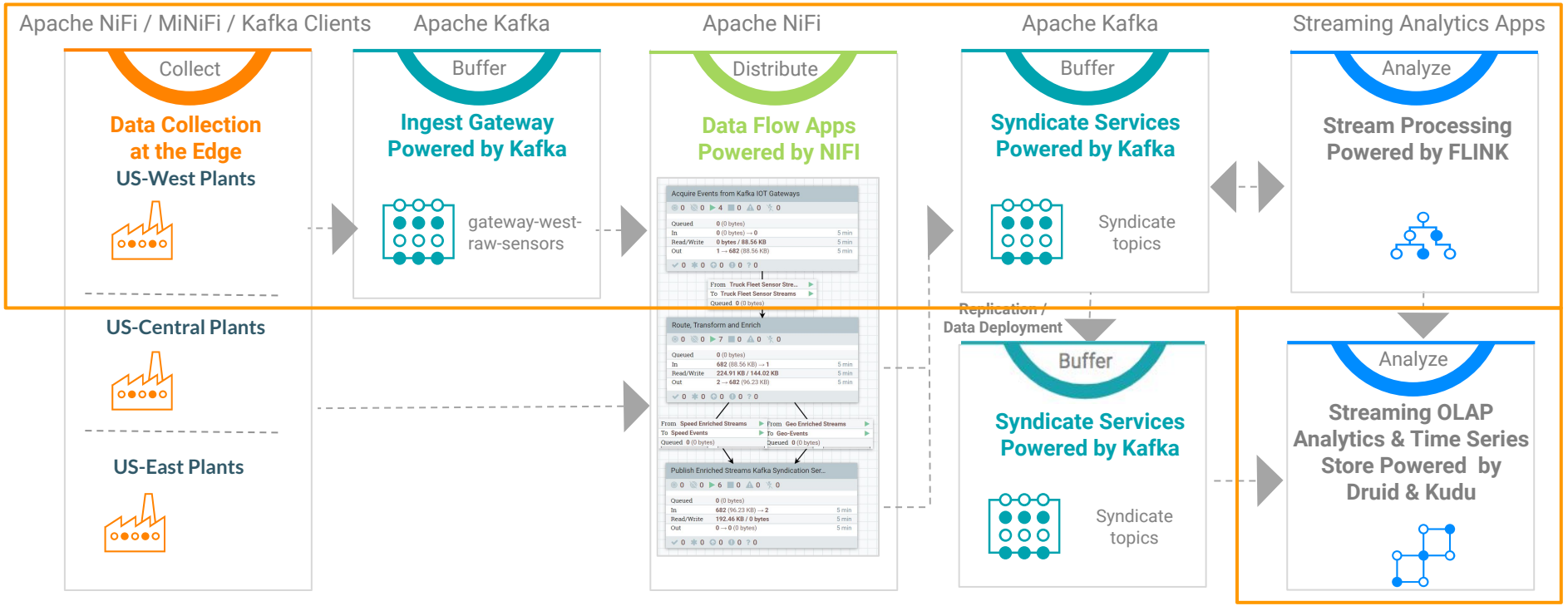
Summary

Additional Slides

A DATA-IN-MOTION REFERENCE ARCHITECTURE



A DATA-IN-MOTION REFERENCE ARCHITECTURE



REFERENCE ARCHITECTURE FOR FRAUD DETECTION