CLOUDERA

Introduction to Cloudera Data Flow (CDF)

Data In Motion Field Engineering

Agenda

Introduction

Flow management

Edge Management

Stream processing

Enterprise Services

Cloud Services

Conclusions

Introduction

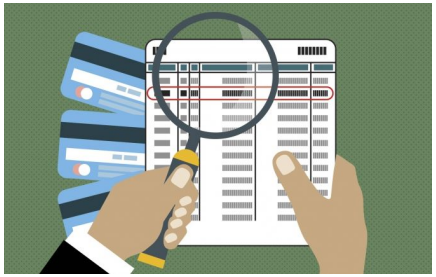
Low Latency Streaming Analytics Use cases

Telco Network Monitoring

Content Recommendations

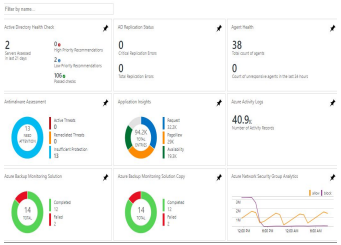
Search Optimization

Clickstream Analysis

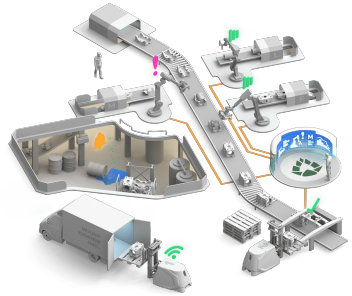


Fraud Detection

Gaming Analytics



Application Monitoring



Industrial IoT

Business Value Realized with Cloudera Data Flow

1

Faster insights

By ingesting and processing events in real-time. Whether it's part of an application, or a standalone pipeline, insights are available at low latency

2

Better customer experience

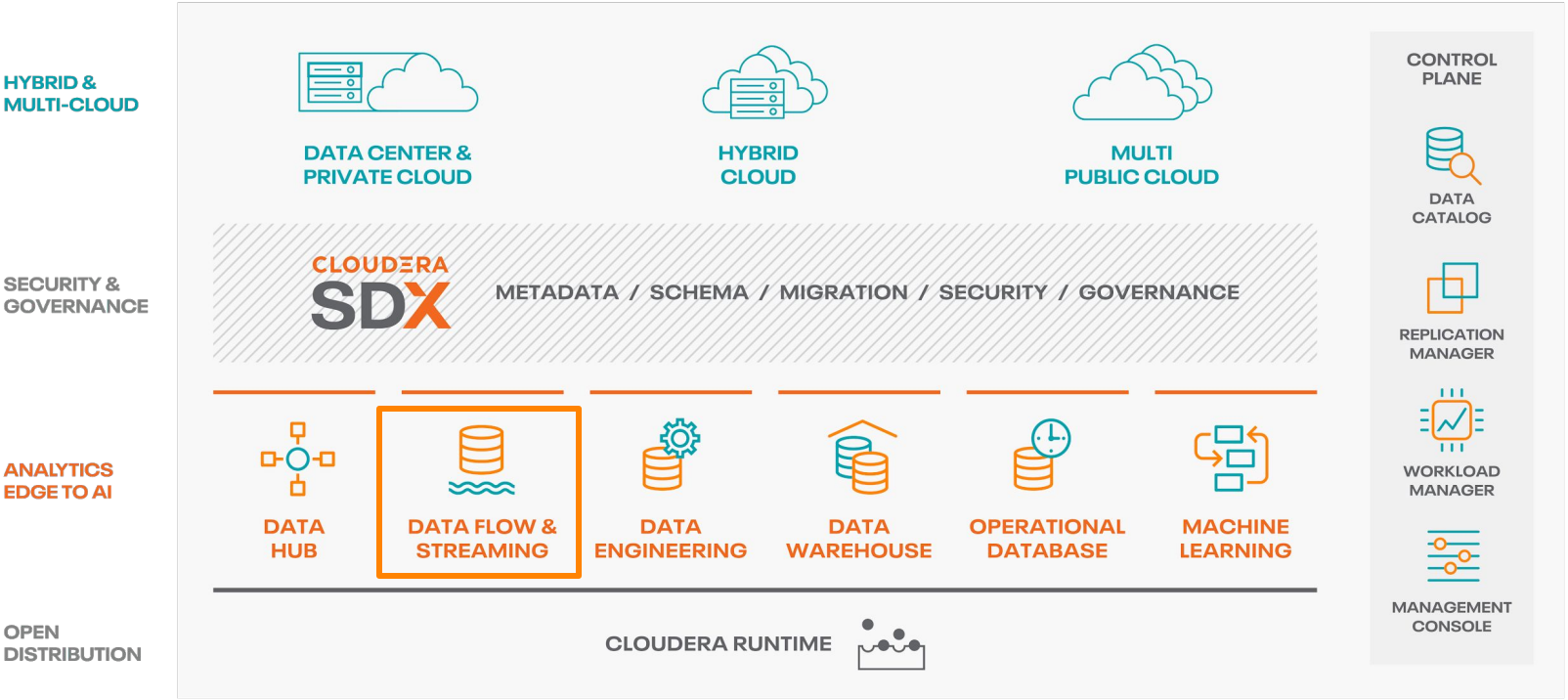
Reacting and answering the customer in real-time is a game changer in a world where we are used to smartphones, continuous connectivity and social networks.

3

Increased agility at lower cost

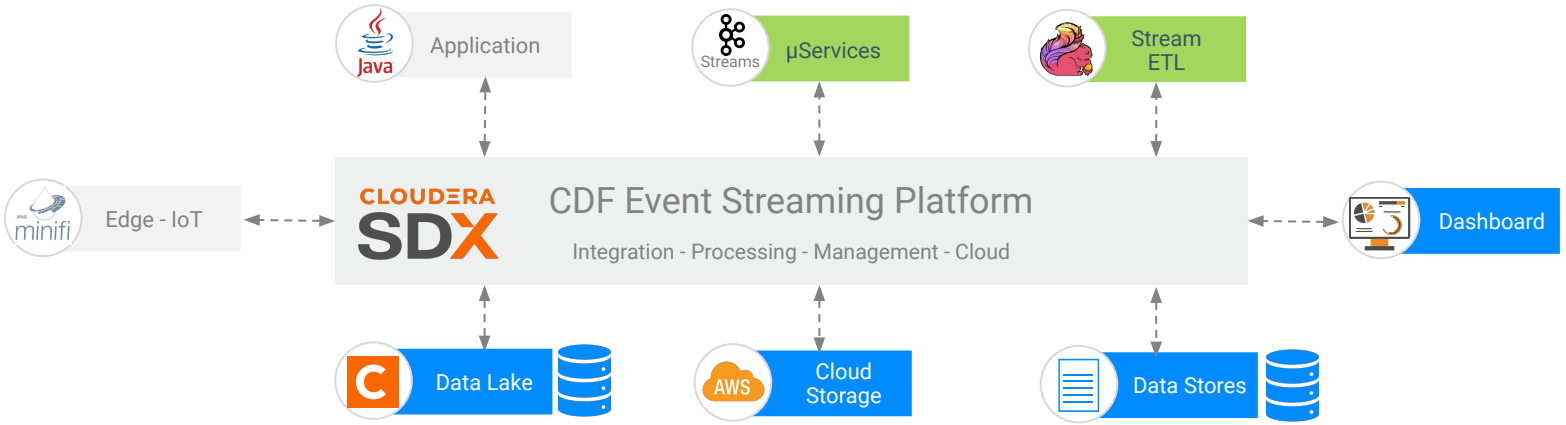
By putting events at the center of the data architecture, it becomes a living experimentation lab. new applications can be developed easily at lower cost

Cloudera Data Platform

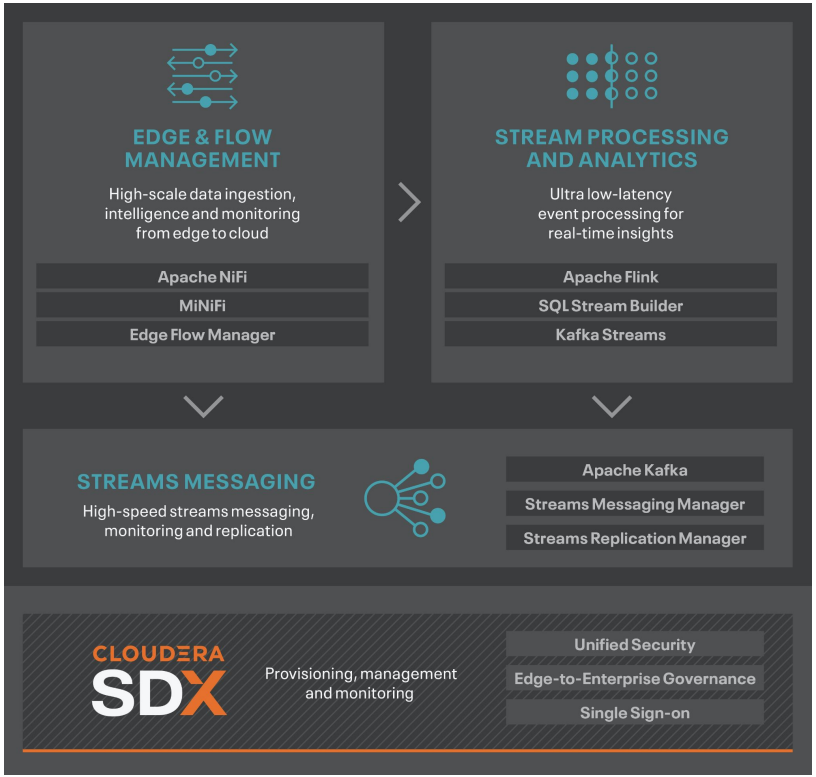


Event Driven Organization with CDF

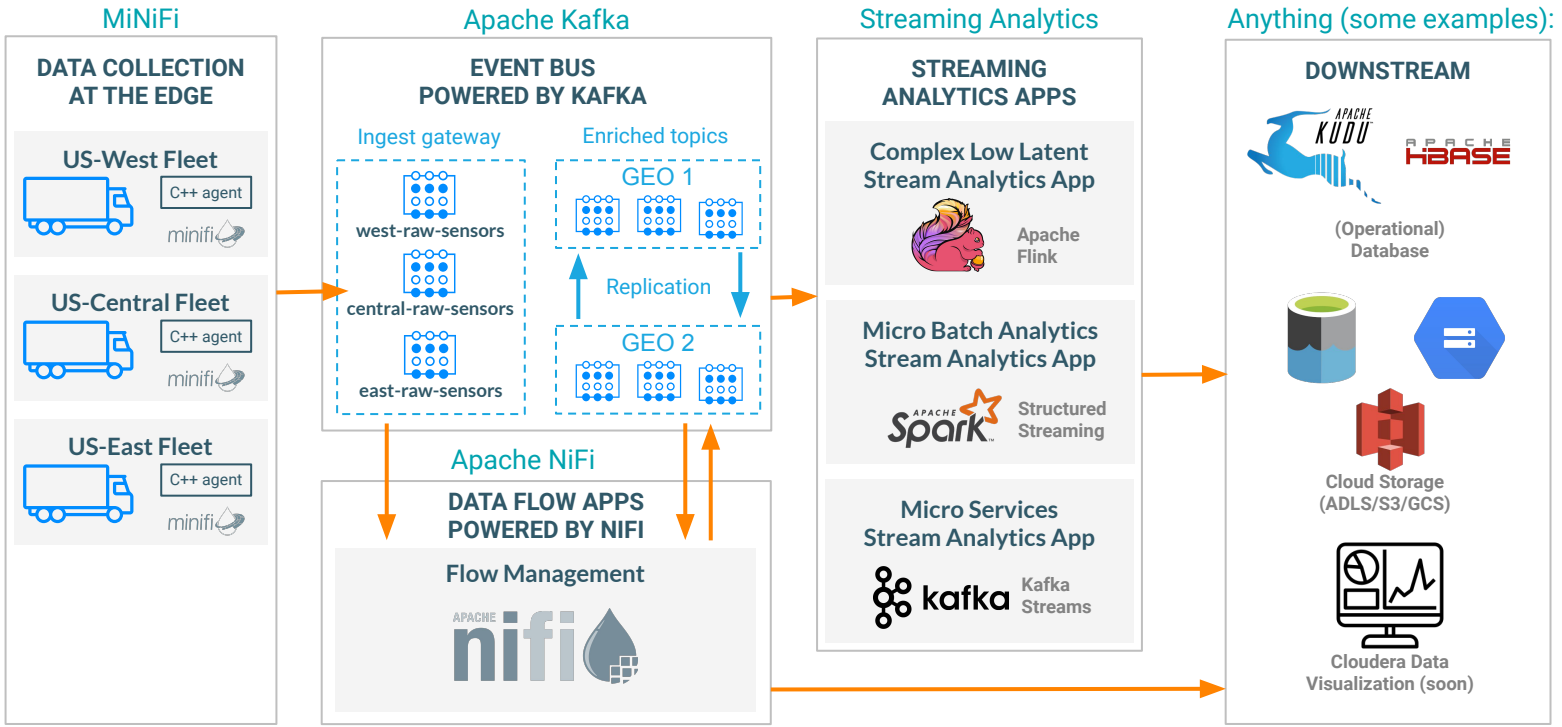
Modernize your data and applications



Cloudera DataFlow Data-in-motion Platform



Cloudera DataFlow reference architecture

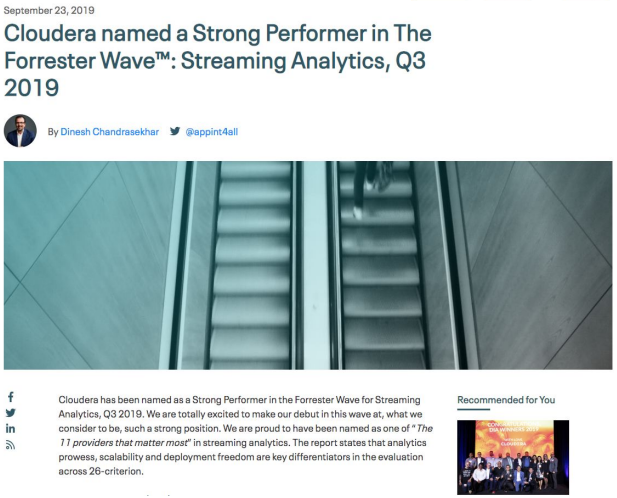


Strong Performer in Forrester Wave for Streaming Analytics

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

Report

Blog



Flow Management

First Mile Problem

Lack of Agile Data Collection Tools Has Hindered the Start of the Data Lifecycle Journey

Data Source Diversity and Scale

Platform/Data Ingestion Team

"BU has asked for tooling to ingest data from 80 different sources and observe 1 million events per hour. Our current tools don't support these diverse data sources or extreme scale requirements."

Agility

Data/App Developer

"Onboarding new data sources and updating ingestion services is extremely painful and slow. Too much code to write and maintain to provide enterprise ingestion services (failures, scheduling, re-tries)"

Compliance Requirements

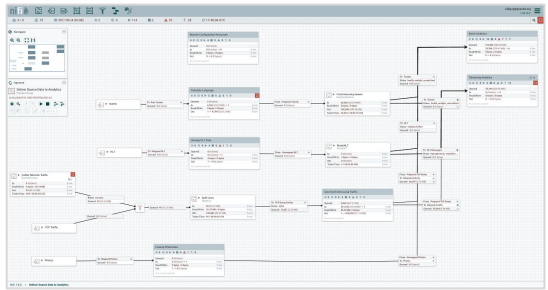
CISO / VP of Platform

"Our homegrown ingestion tool has been around for 20+ years. The ingestion tool doesn't meet new requirements for Sarbanes Oxley, CCPA and GDPR."

Cloudera Flow and Edge Management

Move data from anywhere to anywhere with ease

minifi APACHE nifi



APACHE NIFI registry

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

ACQUIRE

kafka, HIVE, Spark, Storm, Solr, SQL, Flink, DynamoDB, cassandra, amazon KINESIS, elastic, S3, mongoDB, slack

- FTP
- SFTP
- HL7
- UDP
- XML
- HTTP
- EMAIL
- HTML
- IMAGE
- SYSLOG

PROCESS

- HASH
- MERGE
- EXTRACT
- DUPLICATE
- SPLIT
- ENCRYPT
- TALL
- ROUTE TEXT
- ROUTE CONTENT
- ROUTE CONTEXT
- ROUTE RATE
- DISTRIBUTE LOAD
- EVALUATE
- EXECUTE
- GEOENRICH
- SCAN
- REPLACE
- TRANSLATE
- CONVERT

DELIVER

FTP, SFTP, HL7, UDP, XML, HTTP, EMAIL, HTML, IMAGE, SYSLOG

kafka, HIVE, Spark, Storm, Solr, SQL, Flink, DynamoDB, cassandra, amazon KINESIS, elastic, S3, mongoDB, slack

- Over 300 Prebuilt Processors
- Easy to build your own
- Parse, Enrich & Apply Schema
- Filter, Split, Merger & Route
- Throttle & Backpressure

- Guaranteed Delivery
- Full data provenance from acquisition to delivery
- Diverse, Non-Traditional Sources
- Eco-system integration

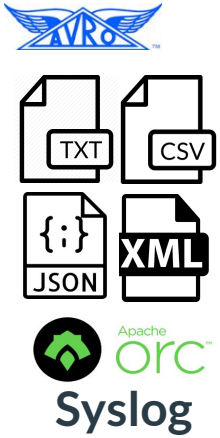
An overview of NiFi capabilities

Data Ingest

SFTP
HL7
UDP
HTTP
WS
Syslog
MQTT

Data Transformation

Hash
Merge
Compress
Duplicate
Split
Encrypt

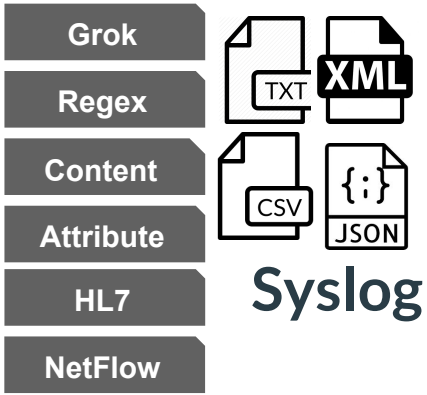


Data Enrichment

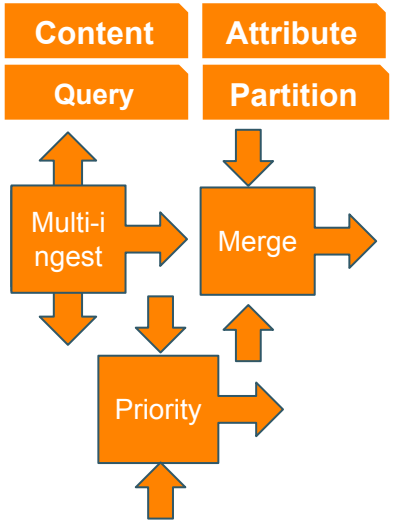
GeoIP
Enrich IP
REST
XML
Mapcach

An overview of NiFi capabilities

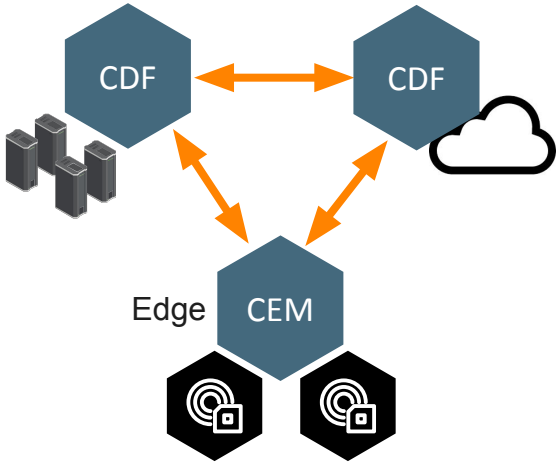
Parsing



Routing

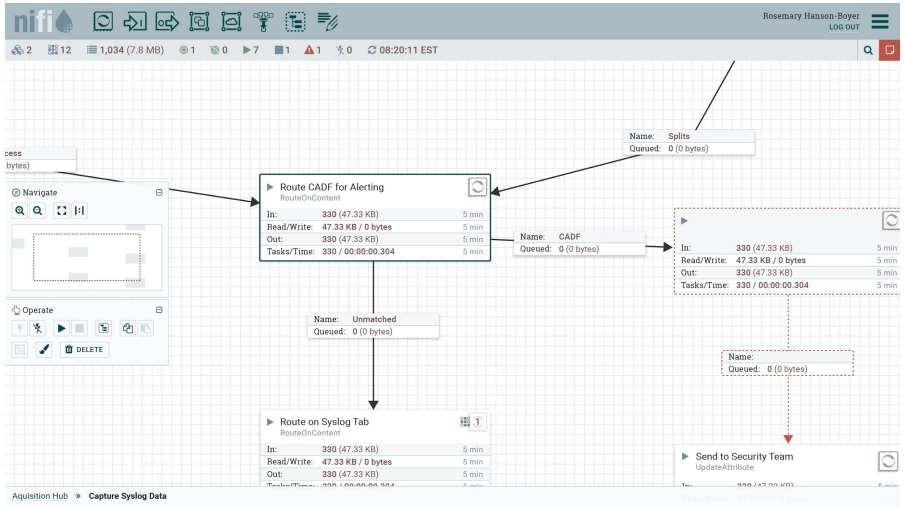


Data Movement



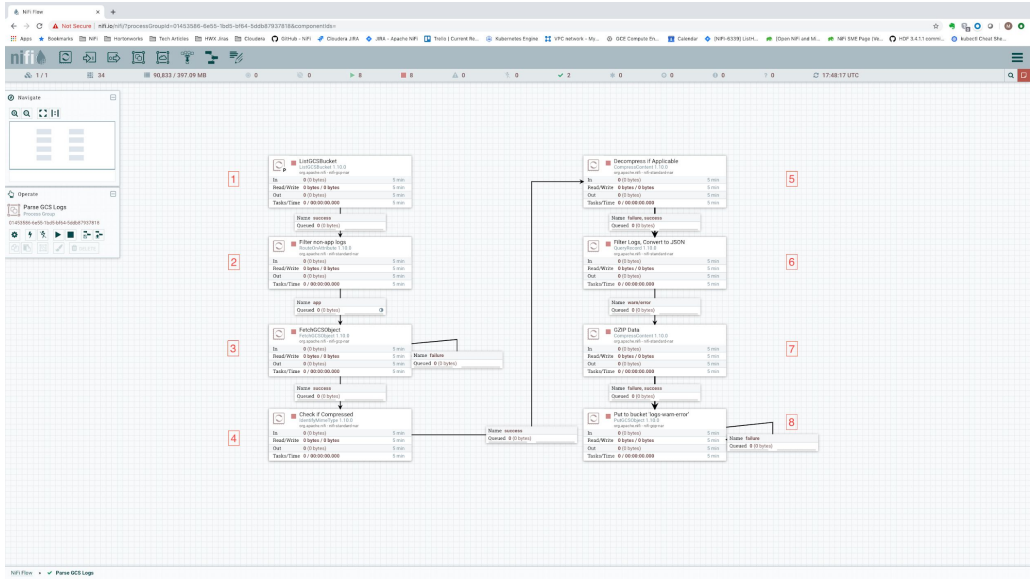
Apache NiFi High Level Capabilities

- Scale horizontal and vertically
  - Scale your data flow to millions event/s
  - Ingest TB to PB of data per day
- Adapt to your flow requirements
  - Back pressure & Dynamic prioritization
  - Loss tolerant vs guaranteed delivery
  - Low latency vs high throughput
- Secure
  - SSL, HTTPS, SFTP, etc.
  - Governance and data provenance
- Extensible
  - Build your own processors and Controller services (providers)
  - Integrate with external systems (Security, Monitoring, Governance, etc)



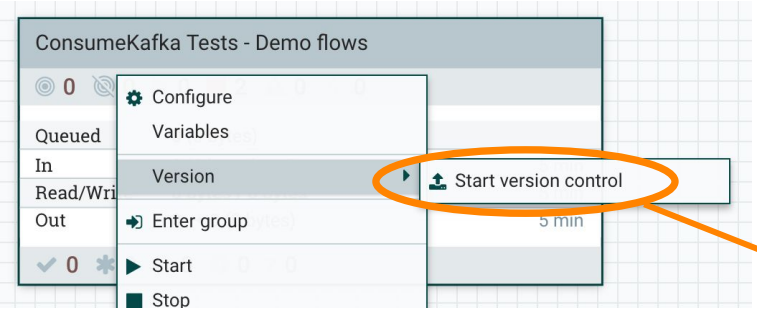
Processing one billion events per second with NiFi

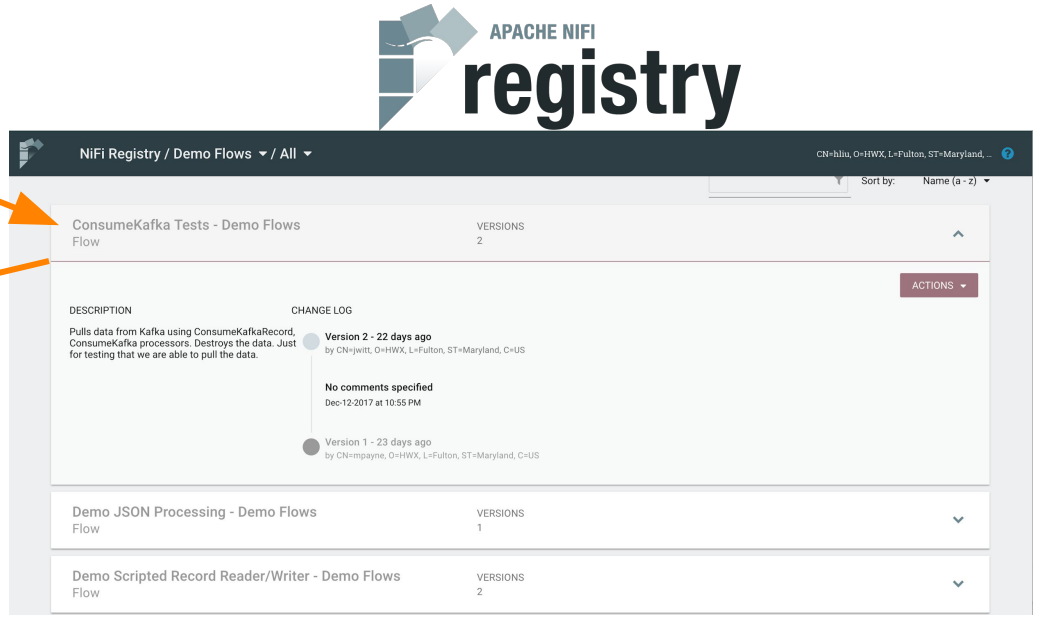
https://blog.cloudera.com/benchmarking-nifi-performance-and-scalability/

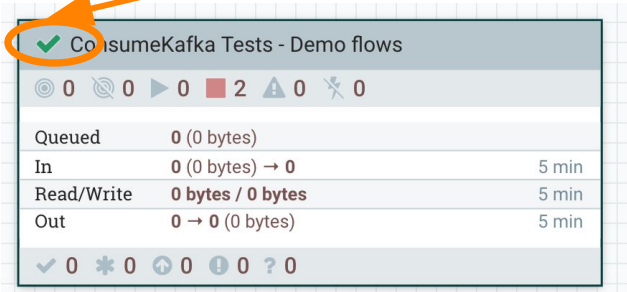


Nodes	Data rate/sec	Events/sec	Data rate/day	Events/day
1	192.5 MB	946,000	16.6 TB	81.7 Billion
5	881 MB	4.97 Million	76 TB	429.4 Billion
25	5.8 GB	26 Million	501 TB	2.25 Trillion
100	22 GB	90 Million	1.9 PB	7.8 Trillion
150	32.6 GB	141.3 Million	2.75 PB	12.2 Trillion

NiFi Flow Registry

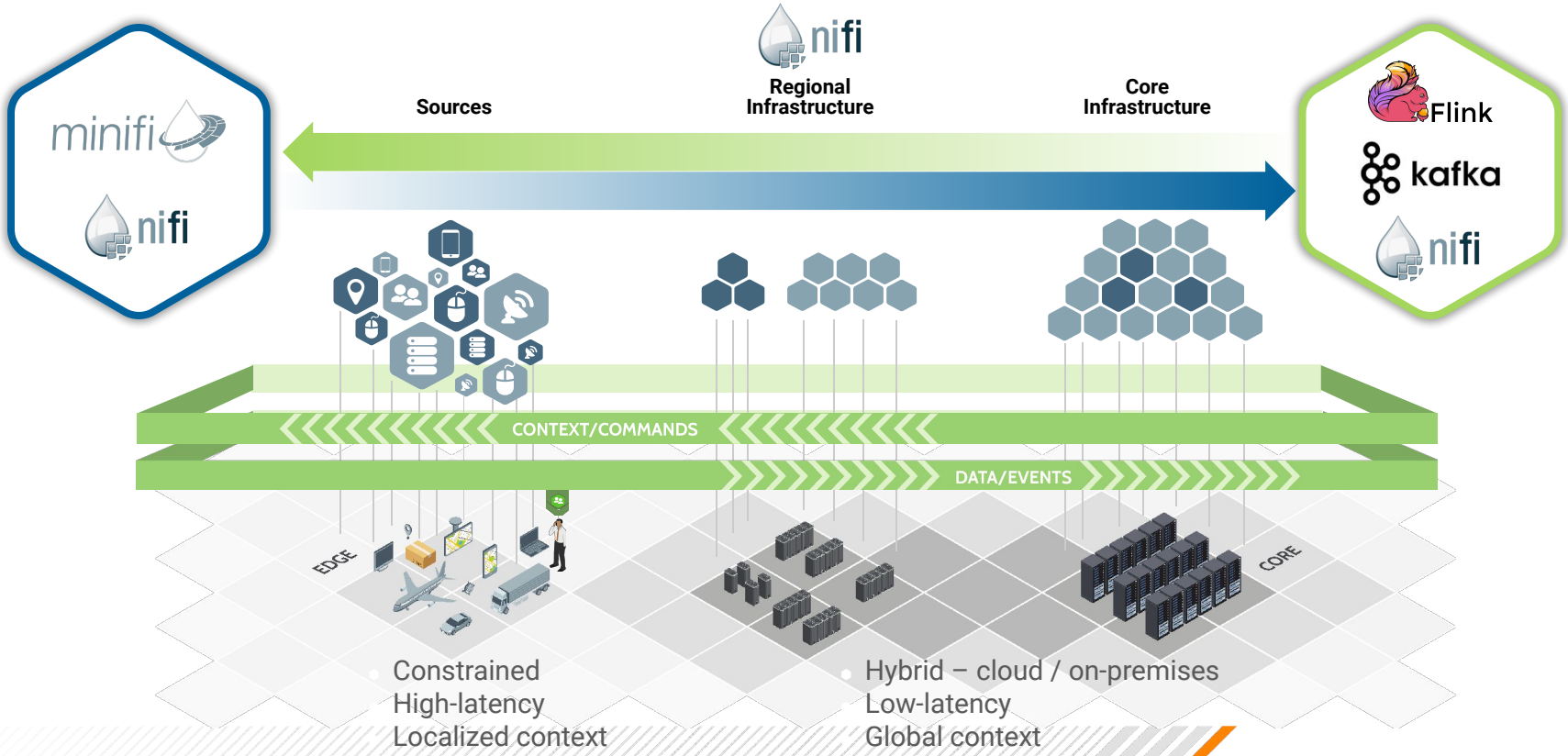






Edge Management

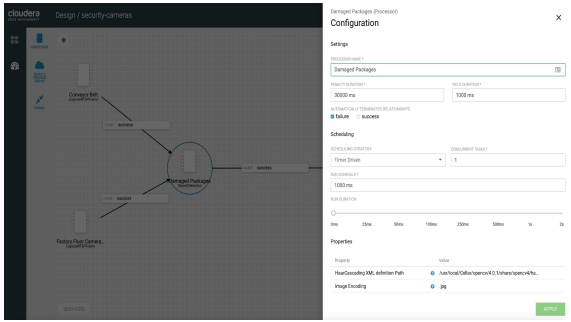
End to End data management



Cloudera EDGE Management

Edge device data collection and processing with easy to use central command and control

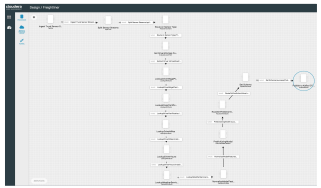
Edge Flow Manager



APACHE NIFI minifi

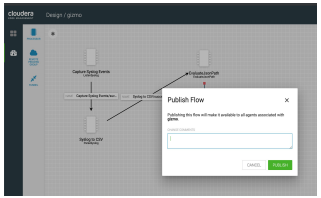
A lightweight edge agent that implements the core features of Apache NiFi, focusing on data collection and processing at the edge

Flow Authorship



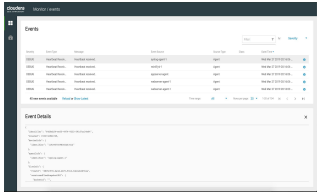
- Small footprint agent with MiNiFi
- Java and C++ agents
- Rich edge processors (edge collection & processing)
- End to end lineage and security

Flow Deployment



- Central Command and Control (C2)
- Design and deploy to thousands of agents
- Edge Applications lifecycle management
- Multitenancy with Agent classes
- Native integration with other CDF services

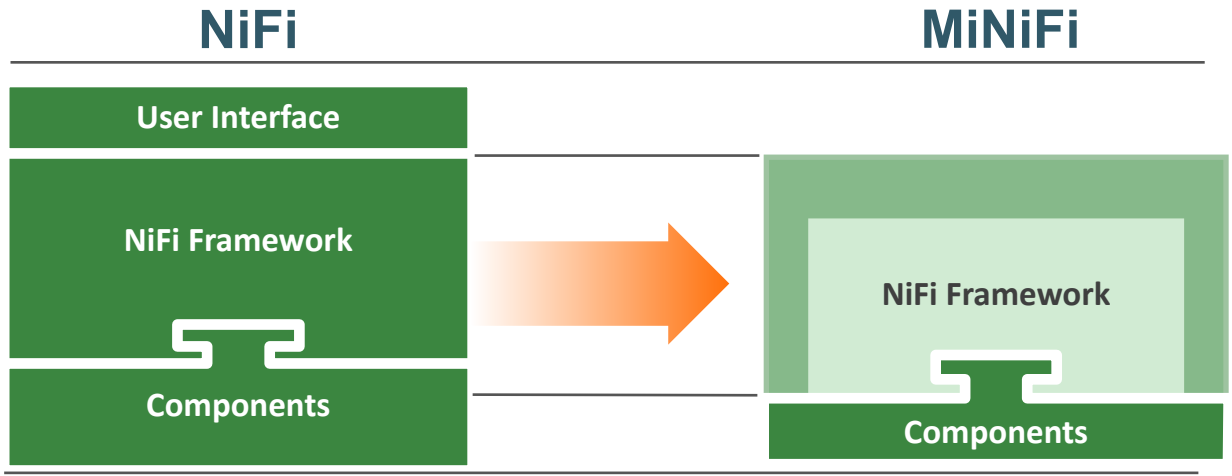
Flow Monitoring



MiNiFi agents

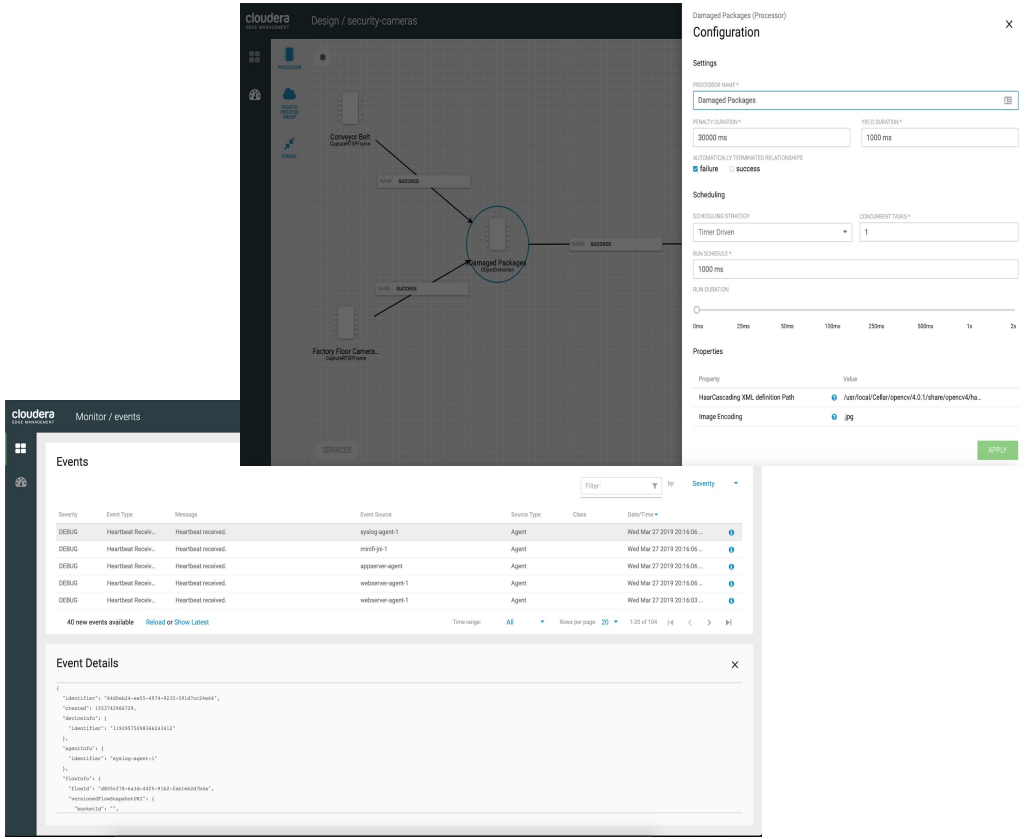
- C++ and Java agents
- Security & Data provenance
- Guaranteed delivery
- Data buffering, Backpressure
- Prioritized queuing, Flow QoS
- Loss tolerance
- TensorFlow support of Edge AI
- Supports IoT ecosystem


```
Security Properties:
    keystore: /tmp/ssl/localhost-ks.jks
    keystore type: JKS
    keystore password: localtest
    key password: localtest
    truststore: /tmp/ssl/localhost-ts.jks
    truststore type: JKS
    truststore password: localtest
    ssl protocol: TLS
    Sensitive Props:
        key:
        algorithm: PBEWITHMD5AND256BITAES-CBC-OPENSSL
        provider: BC

Processors:
    - name: TailAppLog
      class: org.apache.nifi.processors.standard.TailFile
      max concurrent tasks: 1
      scheduling strategy: TIMER_DRIVEN
      scheduling period: 10 sec
      penalization period: 30 sec
      yield period: 1 sec
      run duration nanos: 0
      auto-terminated relationships list:
      Properties:
```



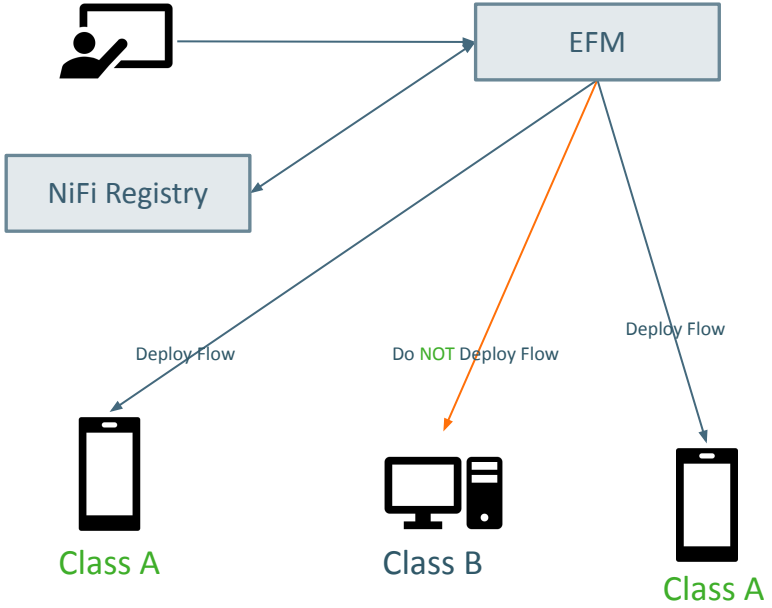
Edge Flow Manager (EFM)

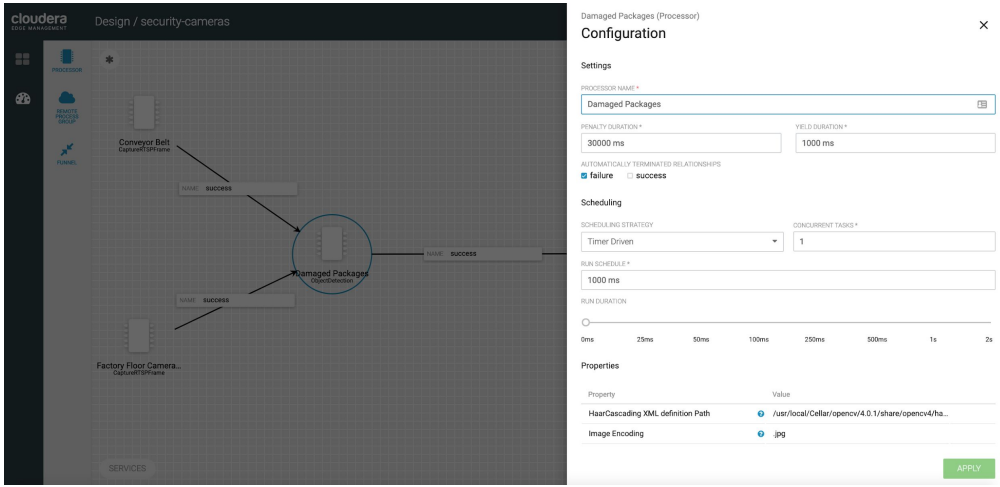
- Manage thousands of deployed agents
- NiFi-like user interface to develop and deploy flow files to the edge
- Application lifecycle management
- Update and deploy ML model files to the edge agents
- Monitor thousands of edge agents
- Integration with NiFi Registry



EFM Command & Control

Flow Authorship & deployment

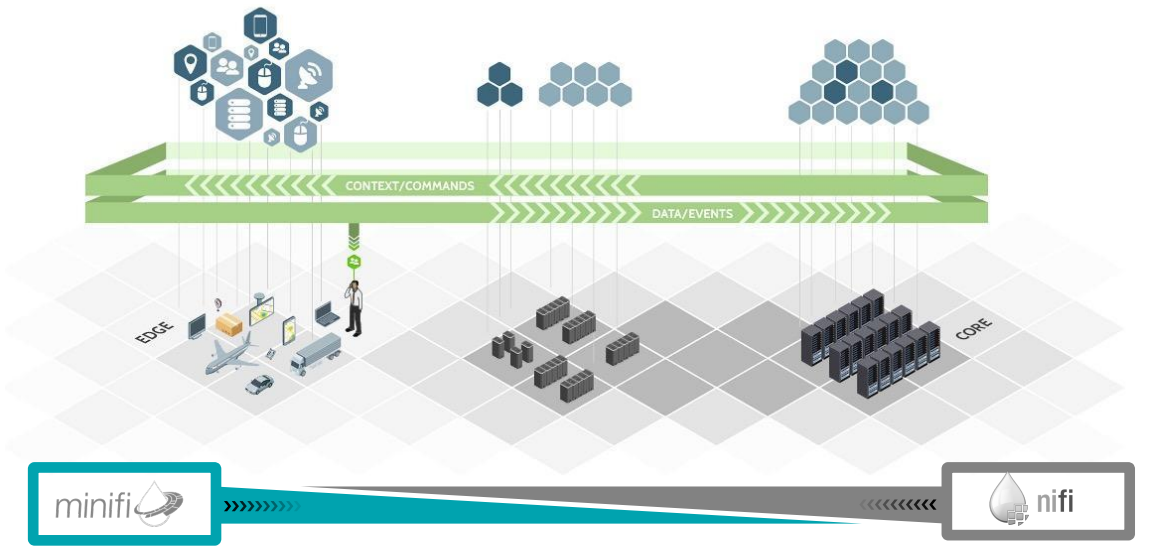




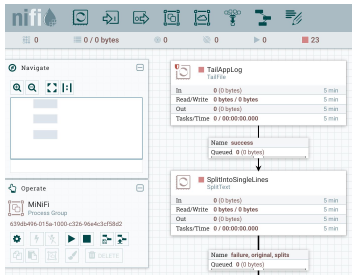
- User designs flow in EFM UI
- Completed flow is saved to NiFi Registry
- Class A flow is pushed to only class A devices

How Does MiNiFi Interact With NiFi?

- MiNiFi
  - Receive flows
  - Collect data
  - Send for processing
- NiFi
  - Design flows
  - Aggregate data from many sources
  - Perform routing / analysis / SEP



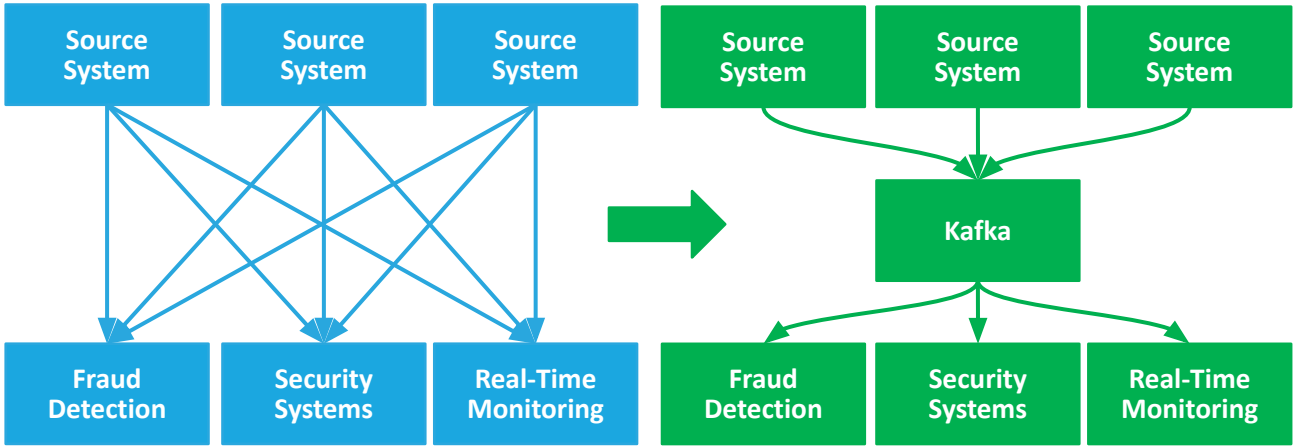

```
Security Properties:
    keystore: /tmp/ssl/localhost-ks.jks
    keystore type: JKS
    keystore password: localtest
    key password: localtest
    truststore: /tmp/ssl/localhost-ts.jks
    truststore type: JKS
    truststore password: localtest
    ssl protocol: TLS
    Sensitive Props:
        key:
        algorithm: PBEWITHMD5AND256BITAES-CBC-OPENSSL
        provider: BC

Processors:
    - name: TailAppLog
      class: org.apache.nifi.processors.standard.TailFile
      max concurrent tasks: 1
      scheduling strategy: TIMER_DRIVEN
      scheduling period: 10 sec
      penalization period: 30 sec
      yield period: 1 sec
      run duration nanos: 0
      auto-terminated relationships list:
      Properties:
```



Stream Messaging

Apache Kafka

- Highly reliable distributed messaging system
- Decouple applications, enables many-to-many patterns
- Publish-Subscribe semantics
- Horizontal scalability
- Efficient implementation to operate at speed with big data volumes
- Organized by topic to support several use cases



Apache Kafka Tooling

Kafka Services for Schema Management, Replication and Monitoring

Schema Registry

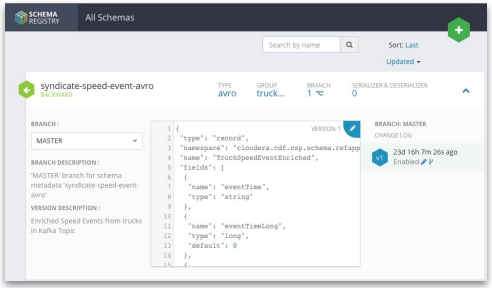
New Kafka Schema Governance

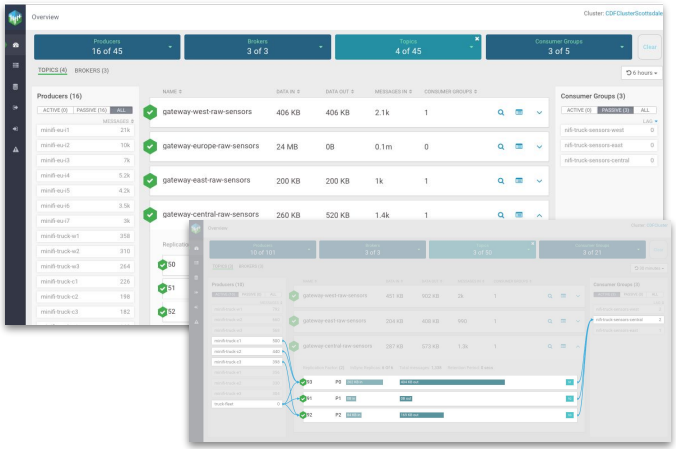
Streams Messaging Manager (SMM)

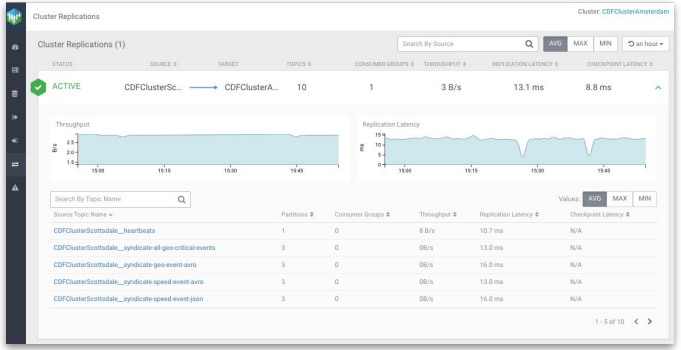
Kafka Monitoring / Management Service

Streams Replication Manager (SRM)

New Kafka Replication Engine powered by MirrorMaker2





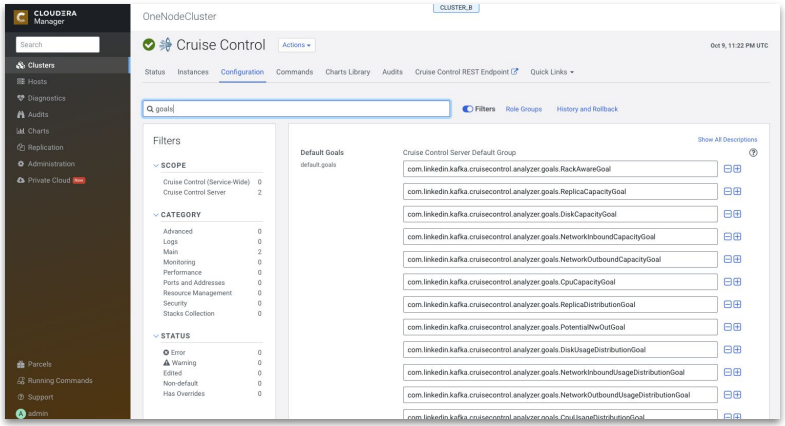


Apache Kafka Tooling

Kafka Services for Schema Management, Replication and Monitoring

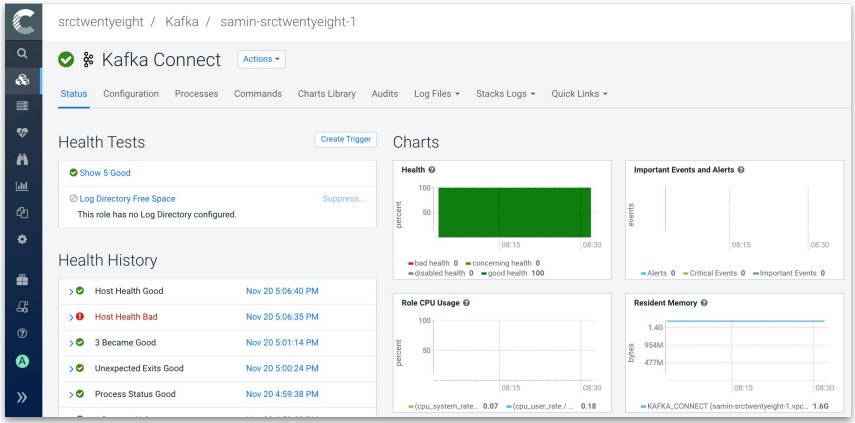
Cruise Control

Intelligent Kafka Cluster
Rebalancing & Self Healing



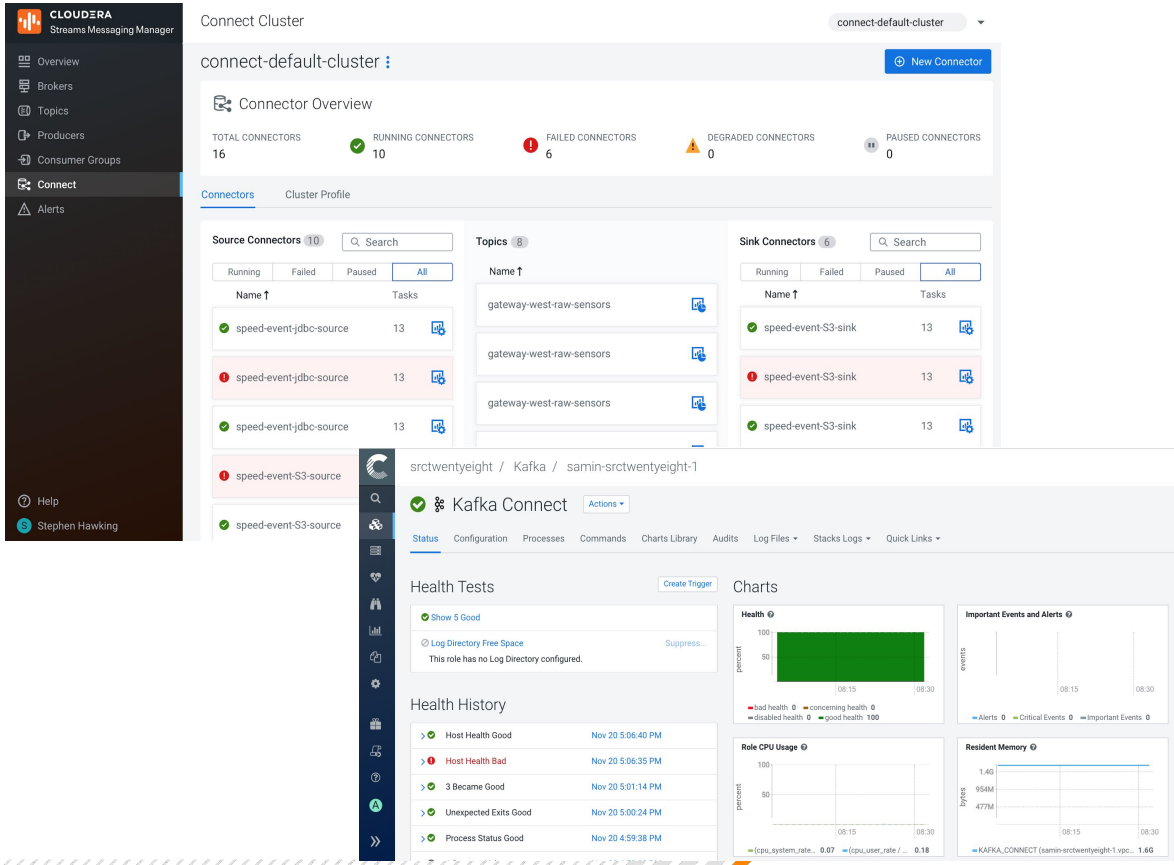
Kafka Connect

Simple Data Movement
for Kafka



Kafka Connect Fully Integrated

- Support for Kafka Connect Runtime.
- Provision & manage Kafka Connect cluster via Cloudera Manager
- Deploy, monitor and management Kafka Connect connectors in SMM
- Supported set of sink and source connectors: HDFS, S3, File based
- Next supported connectors: JMS, MQ, MQTT, JDBC
- Custom/community connectors, can be integrated to SMM

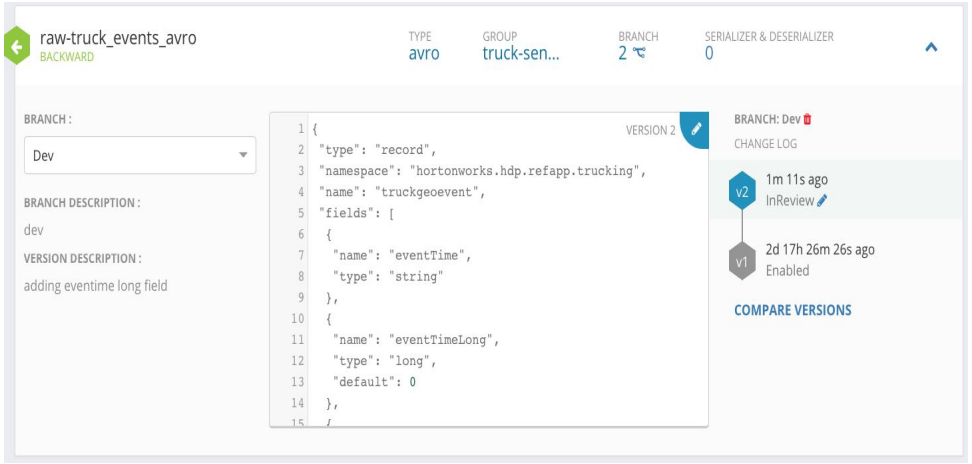


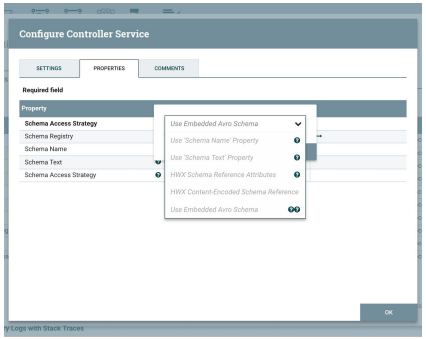
NiFi vs. Kafka Connect for Data Ingestion

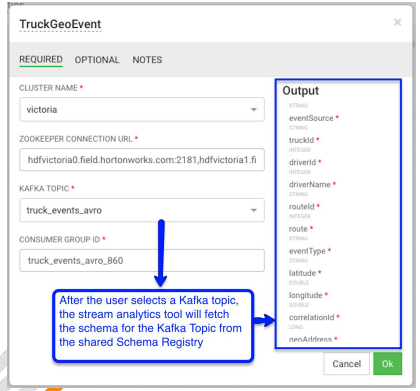
Feature	NiFi	Kafka Connect
Data movement	Move data from any source to any destination. Kafka is one of many the supported systems. ✓	Kafka Connect can only move data to and from Kafka. Several use case don't need data to go through Kafka.
Richness of connectors	NiFi offers more extensive set of processors to read/write from other data source ✓	Some connectors from Confluent. Others are from vendors or community
Data Pipeline	Usually data is not ready to be consumed by apps. NiFi can filter, transform, enrich data before even publishing it for consumption or processing. ✓	An integral component of an ETL pipeline with Kafka and a Kafka streams. This means that we need to move and store raw data in Kafka before preparing it (expensive duplication). Transformation are not factored nor reusable.
Edge collection	MiNiFi extends NiFi capabilities to the edge ✓	Push for MQTT proxy for IoT
Data pipeline design	UI based design, Flow/Variable Registries ✓	Configuration and script files
Universality	Not all data sources are compatible with pub/sub model. NiFi is polyglot (Request/Response, Polling, etc) ✓	Mainly publish/subscribe
Simple data movement with ecosystem containing only Kafka	NiFi provides processors more most sources and sinks in enterprise and not just Kafka	Kafka Connect is all about moving data to and from Kafka ✓
Data Ordering guarantees	Ordering of events from source to destination is not guaranteed	Ordering of events in Kafka partition is preserved when writing to downstream systems (HDFS, S3, ADLS) ✓

Schema Registry

- Centralized registry to provide reusable schema
- Version management to define relationship between schemas
- Format validation to enable generic format conversion
- Avoid attaching schema to every piece of data
- Integration with the stack (NiFi, Kafka, SAM)







Streams Messaging Manager (SMM)

- Single Monitoring Dashboard for all your Kafka Clusters across 4 entities
- Message viewer
- Topic administration: CRUD
- Notification on available metrics
- REST as a First Class Citizen
- Designed for the Enterprise
  - Support for Secure Kafka cluster
  - Rich Access Control Policies (ACLS)

Kafka Replication Use Cases

Disaster Recovery

In an event of a partial or complete datacenter disaster, providing failover/failback to a secondary cluster in a different region / DC

Geo-Locality

Active-active geo-localized deployments allows users to access a near-by data center to optimize their architecture for low latency and high performance.

Data Movement / Deployment

Use Kafka to synchronize data between on-prem applications and cloud deployments

Centralized Analytics

Aggregate data from multiple Kafka clusters into one location for organization-wide analytics

Workload Isolation

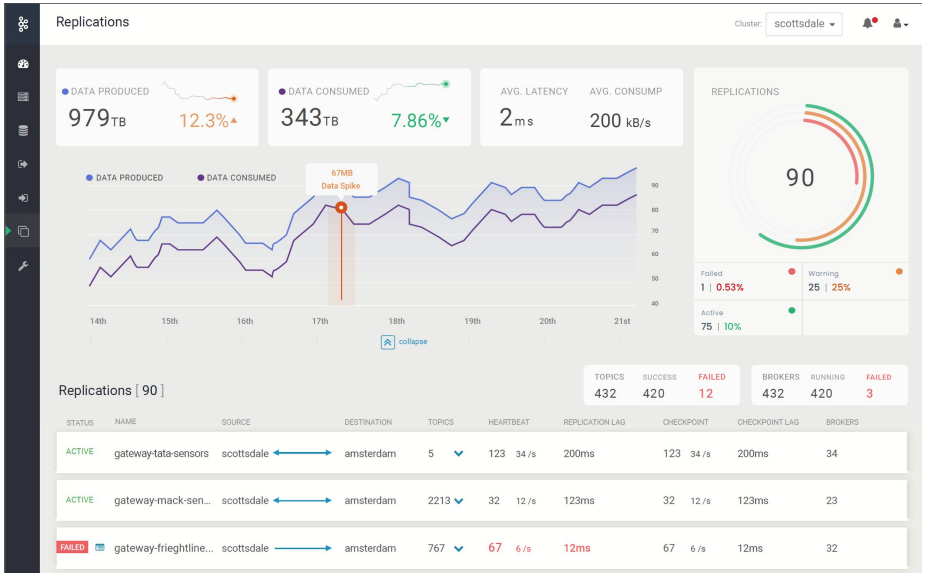
Creation of different envs for SDLC: Dev, Test, Prod. Clusters for specific use case cases (ETL, ingestion, analytics, etc)

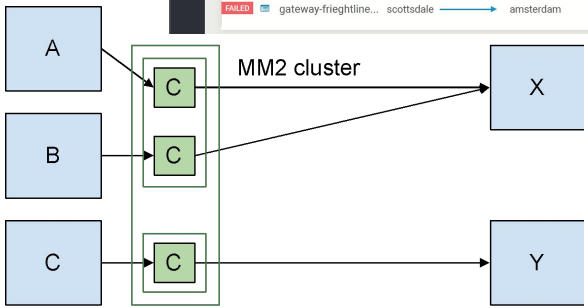
Legal / Compliance

Different data storage and security policies require clusters to be created in region but data still needs to be shared.

Streams Replication Manager (SRM)

- Event Replication engine for Kafka
- Supports active-active, multi-cluster, cross DC replication scenarios
- Leverage Kafka Connect for scalability and HA
- Replicate data and configurations (ACL, partitioning, new topics, etc)
- Offset translation & smart client for simplified failover and failback
- Integrate replication monitoring with SMM





Kafka Cruise Control Service

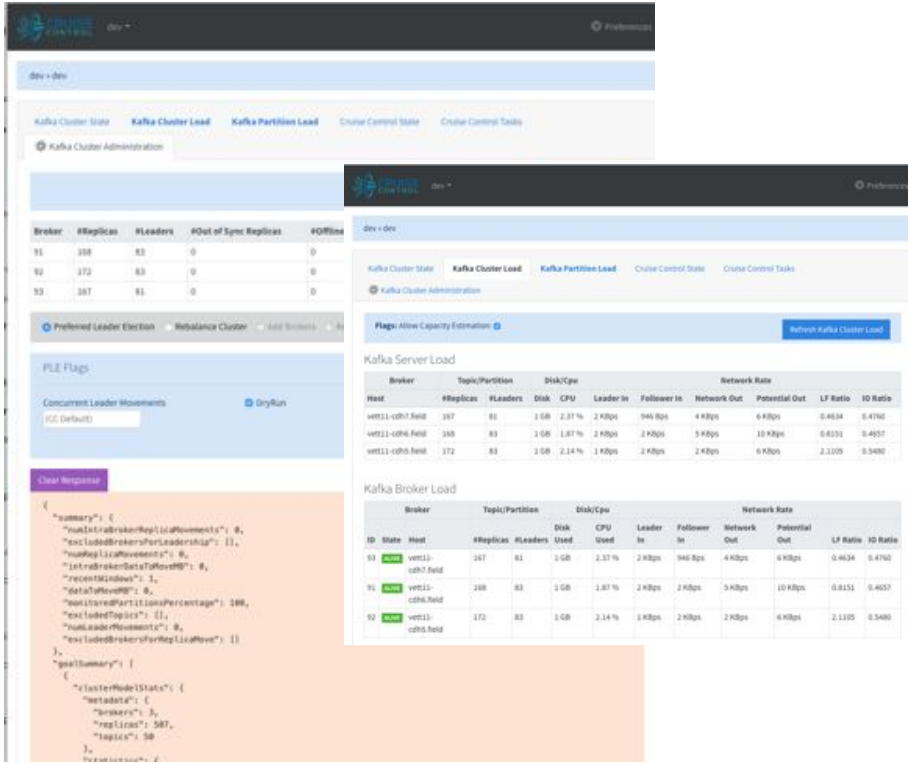
Managing Kafka Clusters at Large Scale

Problem Statement / Requirements

- Platform teams need first class management services that **address hard problems such as frequent hardware/vm failures, cluster expansion/reduction, and load skew among brokers.**
- These class of problems require the need to **balance the cluster intelligently with automated anomaly detection and remediation.**

Solution / Benefit

- Support for Cruise Control in CSP, a open source project created by LinkedIn.
- Includes admin operations such add/decommission/demote brokers, rebalance the cluster.
- Generation of rebalancing proposals based on goals with automated anomaly detection and remediation
- Foundation for providing first class Kafka Cloud Workload in Cloudera Data Hub

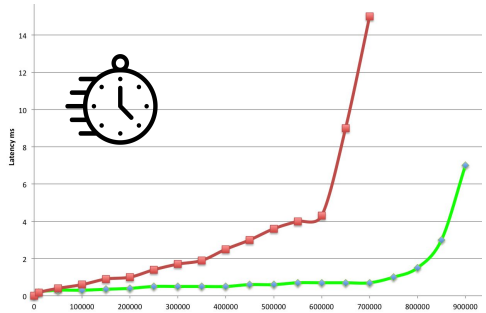


Stream Processing and Analytics

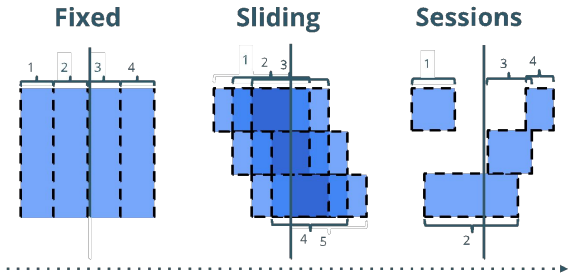
Choosing a streaming engine

Things to consider

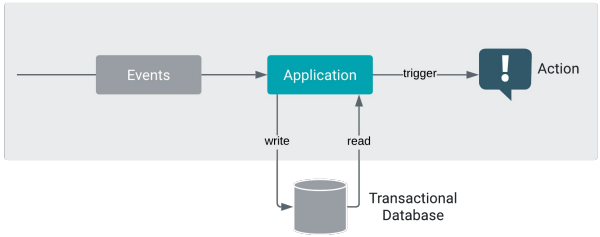
Latency



Windowing



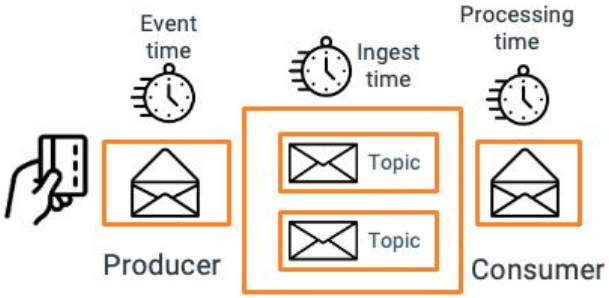
State Management



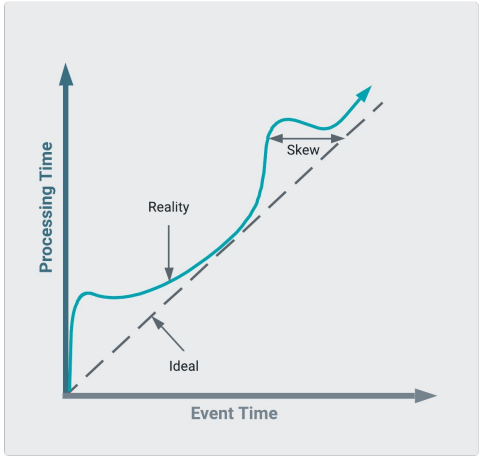
Choosing a streaming engine

Things to consider

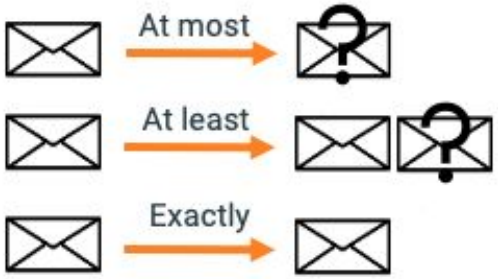
Event time



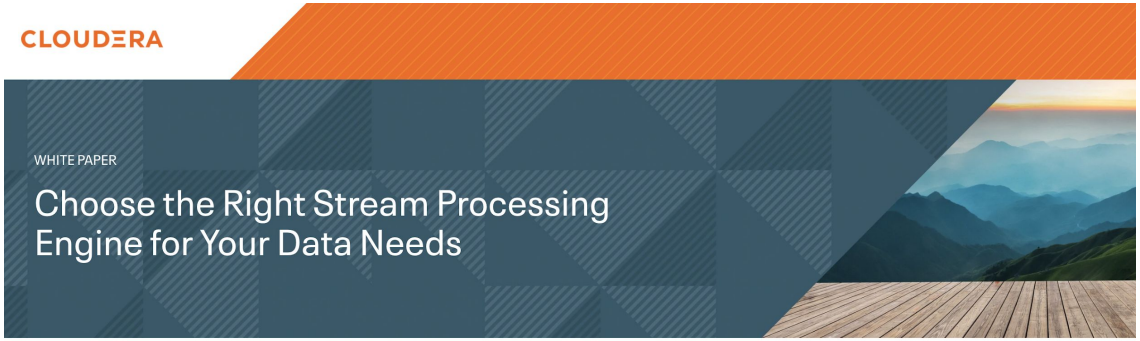
Watermarking



Processing Semantics



Choose The Right Stream - Download the Whitepaper



bit.ly/choose-the-right-stream

Streaming Engines in a nutshell

Already using Spark?

Need highest throughput?
Want unified batch/stream?
Don't need low latency?
Don't need advanced time/state features?

Like Kafka?

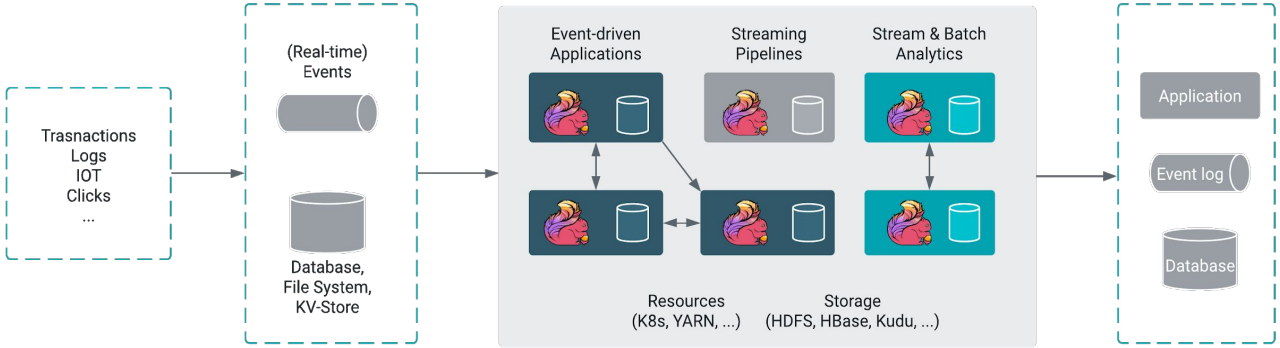
Want low latency?
Want just Kafka?
Only need microservices?
Don't need advanced Time/State features OOTB?

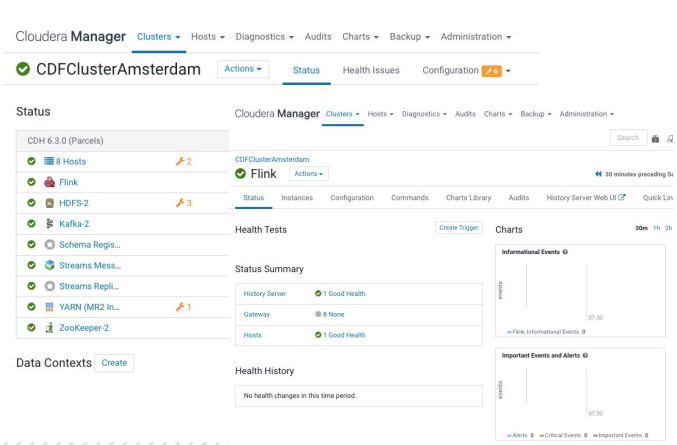
Like Flink?

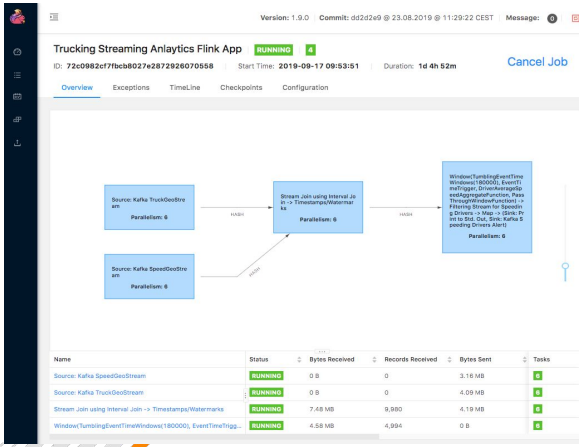
Want flexibility?
Want low latency?
Want leading-edge technology?

Apache Flink

- A distributed processing engine for stateful computations
- Manages 10s TBs of state
- In-memory processing at scale
- Flexible and expressive APIs
- Guaranteed correctness & Exactly-once state consistency
- Event-time semantics
- Flexible deployment & large ecosystem (K8S, YARN, S3, HDFS..)
- Fully integrated with CM



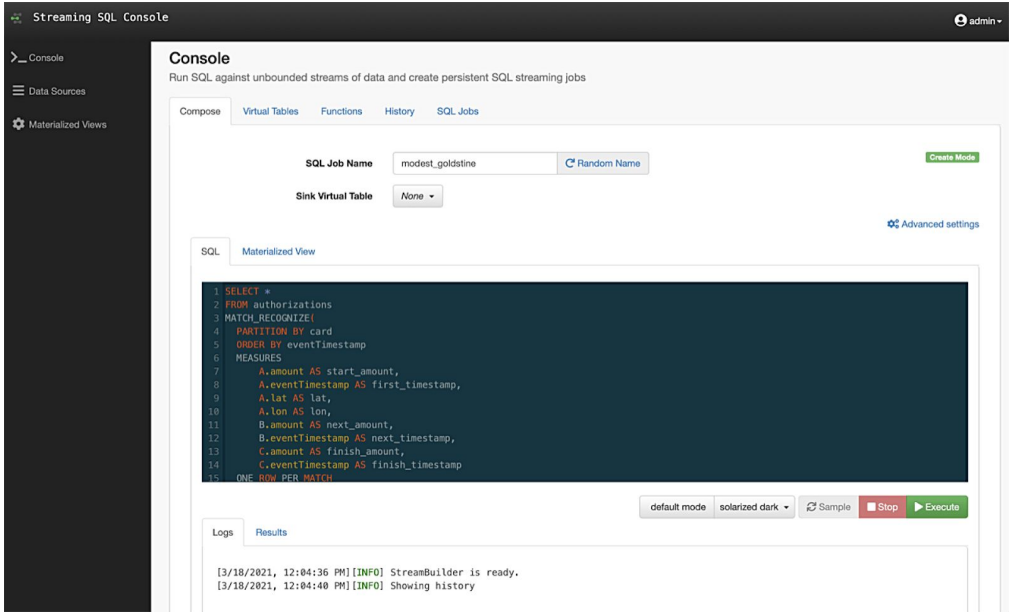




Flink SQL: Complex event processing made easy

Agile Streaming App Development using SQL

- SQL Console for writing and scheduling SQL Streaming jobs
- Unified APIs for streaming data and data at rest
- Run the same query on batch and streaming data
- ANSI SQL: No stream-specific syntax or semantics!
- Many common stream analytics use cases supported



Cloudera SQL Stream Builder

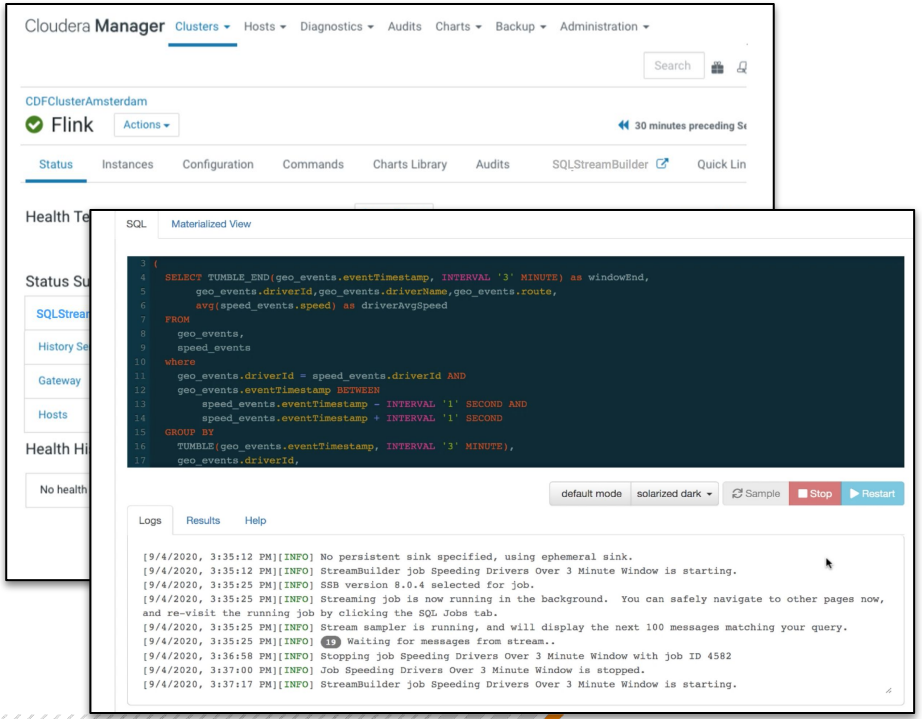
Making Streaming Analytics accessible to everyone with SQL

Application Developer

- Develop & test SQL queries with a powerful UI
- Expose streaming data to applications through materialized views
- Single button “Push to production” turns SQL queries into Flink application

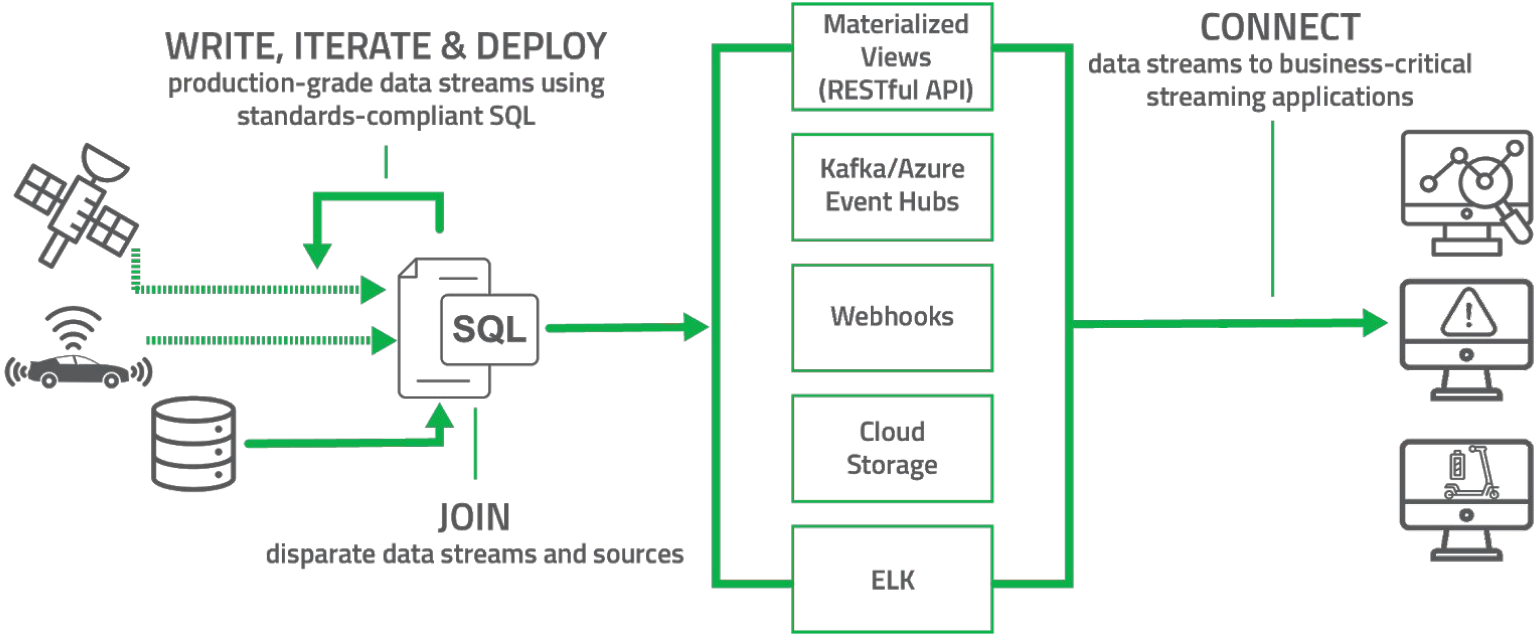
Business Analyst,

- Explore Streaming Data using SQL without learning new skills
- Build new real-time business reporting applications



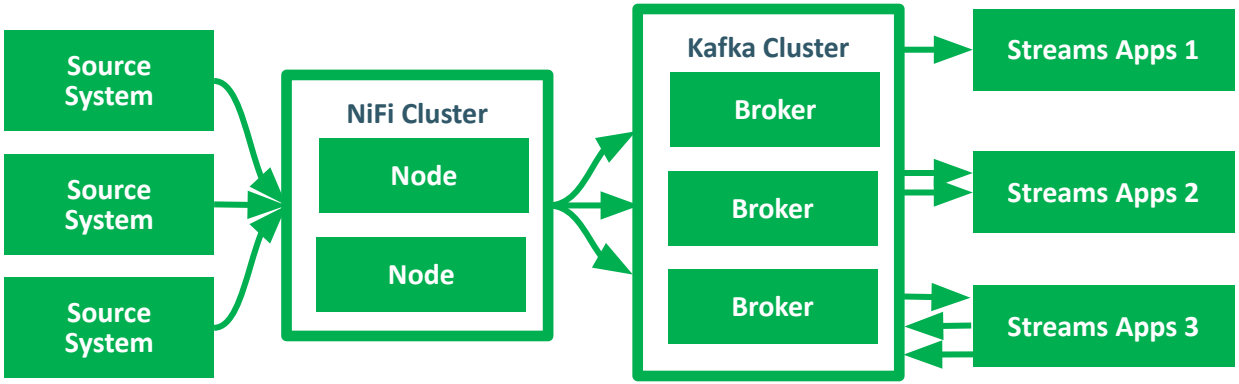
Streaming SQL

Democratizing access to streams of data via structured query language



Kafka Streams

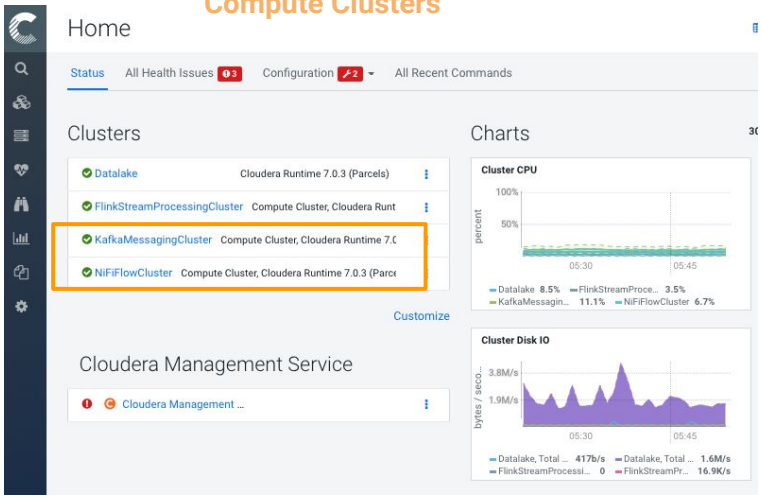
- A client library for applications and microservices on top of Kafka
- Exists in Java and Scala
- Elastic, highly scalable, fault-tolerant
- Supports At-Least-Once and Exactly-once semantics inside Kafka
- Deploy to containers, VMs, bare metal, cloud

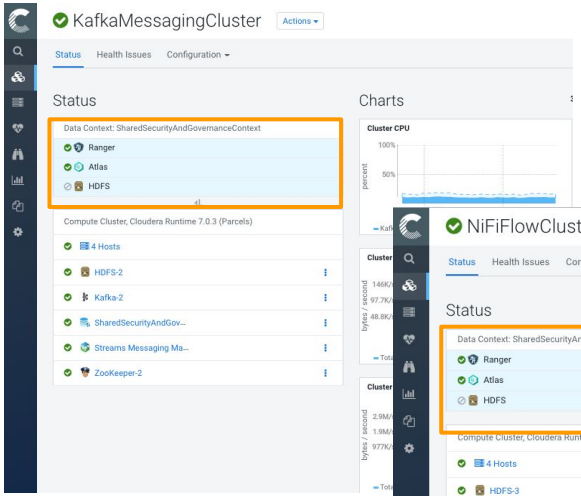


Shared Data Experience

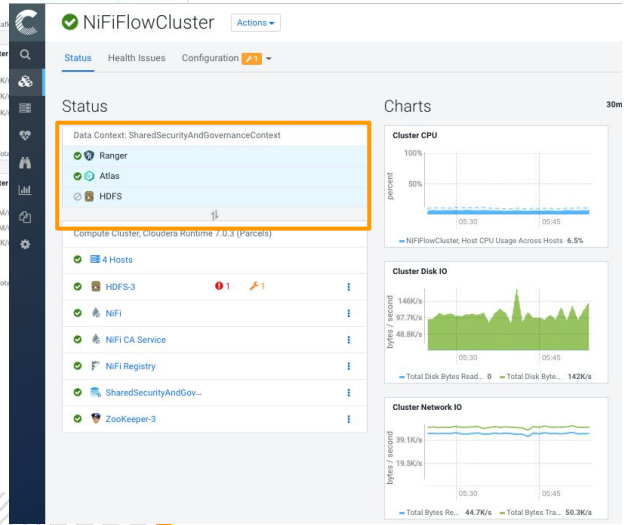
Cloudera Manager support, Virtual Cluster Support

Support for Kafka and NiFi Compute Clusters



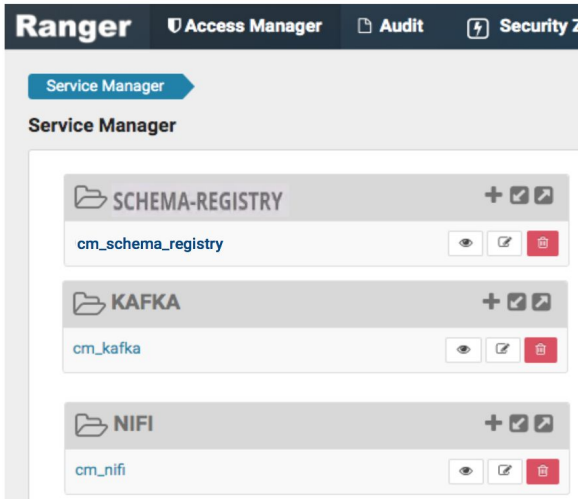


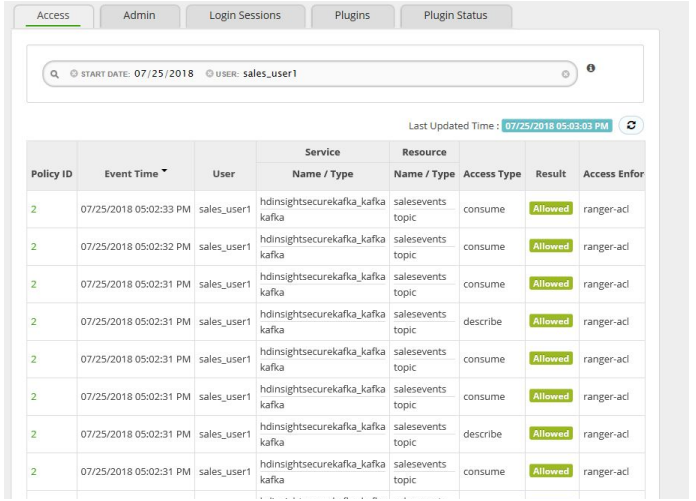
Shared Security Context from DataLake consisting of Ranger and Atlas

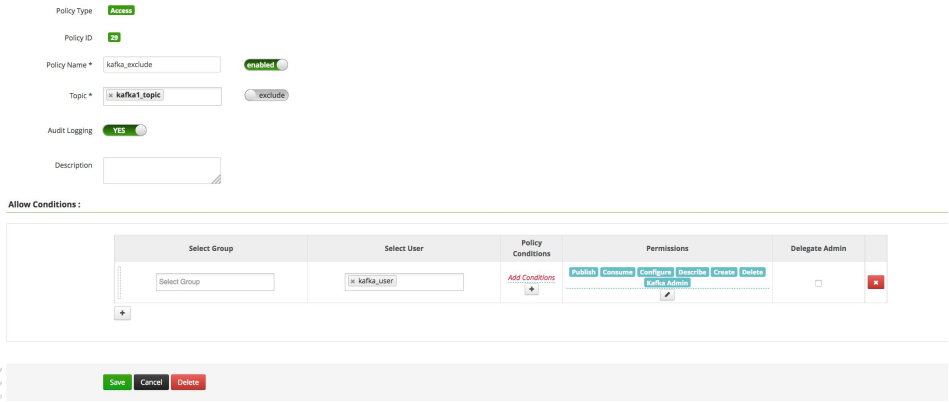


Authorization & audit with Apache Ranger

- Single pane of glass security: NiFi, Kafka, SMM, SR, etc
- Fine-grained access control
- Role & Attribute Based ACL
- Integration with LDAP/AD
- Centralized audit

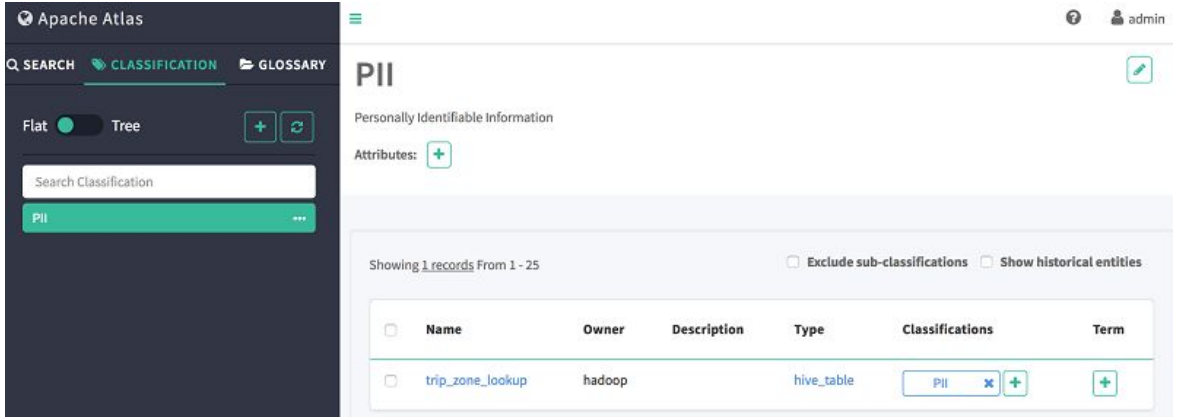


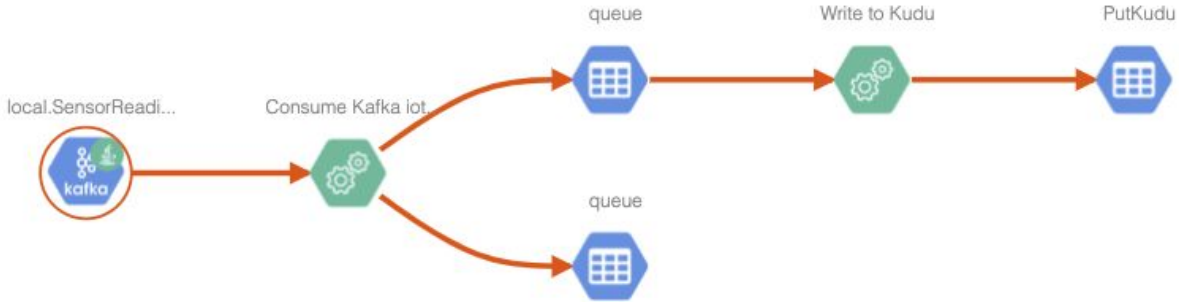




Governance & Lineage with Apache Atlas

- End-To-End lineage
- Integration with NiFi, Spark and Kafka for automatic lineage capture
- Business taxonomy
- Integration with Ranger for Attribute Based Access
- Multi-facets search engine





CDP DataHub Cloud Services

How Does CDP DataHub Bring Value to End Users?

Platform Team

Data/App Developer

Data Steward

"Sorry. It will take me 2 weeks to provision a fully secured and governed Kafka and NiFi Cluster?"

Hm..you can spin up clusters for me easily. But what is the point if i can't get data in & out and move it around easily?

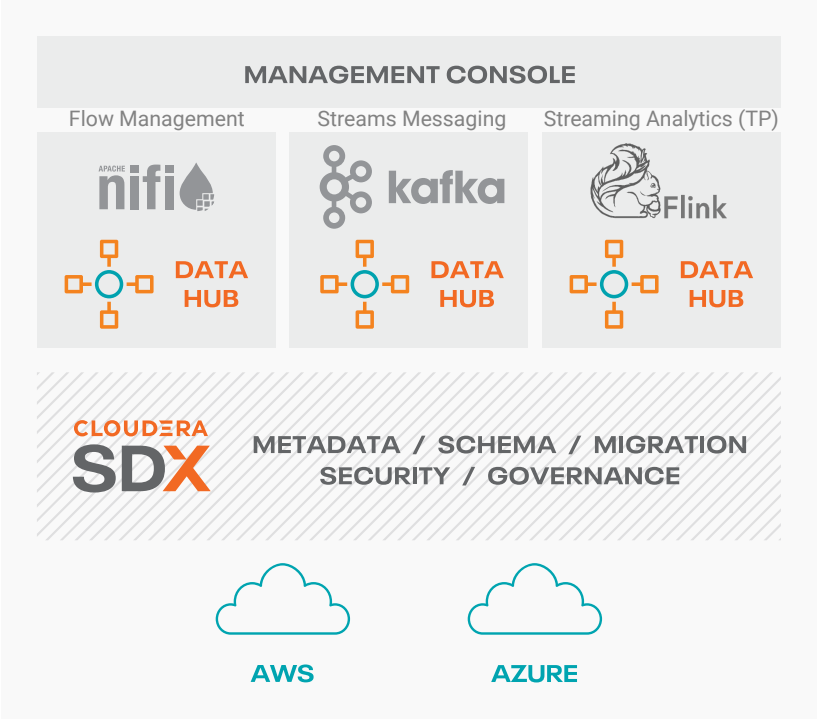
I have no way to track where data is coming from and where its going and security policies are spread everywhere

Sure. I'll get a cluster created for you in the next 30 mins. Do you want on Azure or AWS?

"Wow. I love this code-less approach of NiFi to get data in and out of Kafka, DW, DE Cluster, or any cloud source"

"Lineage and Provenance of my Data Flows. Security policies for NiFi, Kafka, Hive, Spark and Flink are in a single place with Audit"

CDF on CDP Public Cloud Data Hub



CDP Management Console runs as a web service hosted and managed by Cloudera

Data Hub clusters running Kafka, NiFi and Flink 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

Flow Management Cluster Definition (CDP 7.2.8)

Apache NiFi in CDP Data Hub

Latest NiFi innovation

Apache NiFi 1.11.4

Apache NiFi Registry 0.5

Build flows immediately

Pre-created Ranger policies allowing immediate access to NiFi

Support for Kerberos Principal/Password eliminating need for managing keytabs

Pre-configured with TLS settings through RestrictedSSLContextProvider

CDP Platform integration

Managed FreeIPA server to fully secure clusters by default

- Kerberos authentication
- SSL/TLS wire encryption

Integrated with Apache Ranger for managing granular authorization policies on NiFi resources

Protected against cloud infrastructure failures - VM reprovisioning

Pre-configured Atlas lineage reporting

Flow Management in Data Hub

Apache NiFi 1.11 Highlights

Parameter Support across environments

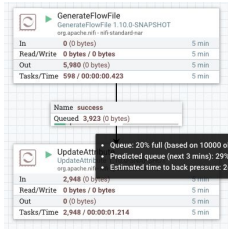
Allows parameterization of all processor properties

Support for sensitive parameters

Allows easy flow promotion through NiFi registry from Dev to Prod

Predictive Monitoring

Queue Length and time to Backpressure are now predicted



https://www.youtube.com/watch?v=Tt8TSlHu7PE

General Improvements

Support for "public" (accessible to remote site to site clients) ports for any processor

Support for flow versioning across NiFi versions

New Retry FlowFile processor for improved error handling

Encrypted content and flow file repositories (preview)

Fore more information, check out: https://medium.com/@abdelkrim.hadjidj , https://www.datainmotion.dev

Streams Messaging Cluster Definition (CDP 7.2.8)

Apache Kafka in CDP Data Hub

Latest Kafka innovation

Apache Kafka 2.5

LDAP Callback Handler enabling SASL/PLAIN authentication

Powerful Kafka Tools

Cloudera Streams Messaging Manager (SMM) for Kafka monitoring and operations

Cloudera Schema Registry to centrally manage your schemas for Kafka and NiFi

CDP Platform integration

Managed FreeIPA server to fully secure clusters by default

- Kerberos authentication
- SSL/TLS wire encryption

Integrated with Apache Ranger for managing granular authorization policies on topics

Protected against cloud infrastructure failures - VM reprovisioning

Streaming Analytics Cluster Definition (CDP 7.2.8)

Apache Flink in CDP Data Hub

 Latest Flink innovation

Apache Flink 1.10

 Cloud optimized

Store application data like checkpoints / savepoints on cloud storage

Scale worker nodes up and down

Heavy / Light Duty templates for different workloads

 CDP Platform integration

Managed FreeIPA server to fully secure clusters by default

- Kerberos authentication
- SSL/TLS wire encryption

Data access secured by Apache Ranger for managing granular authorization policies data sources

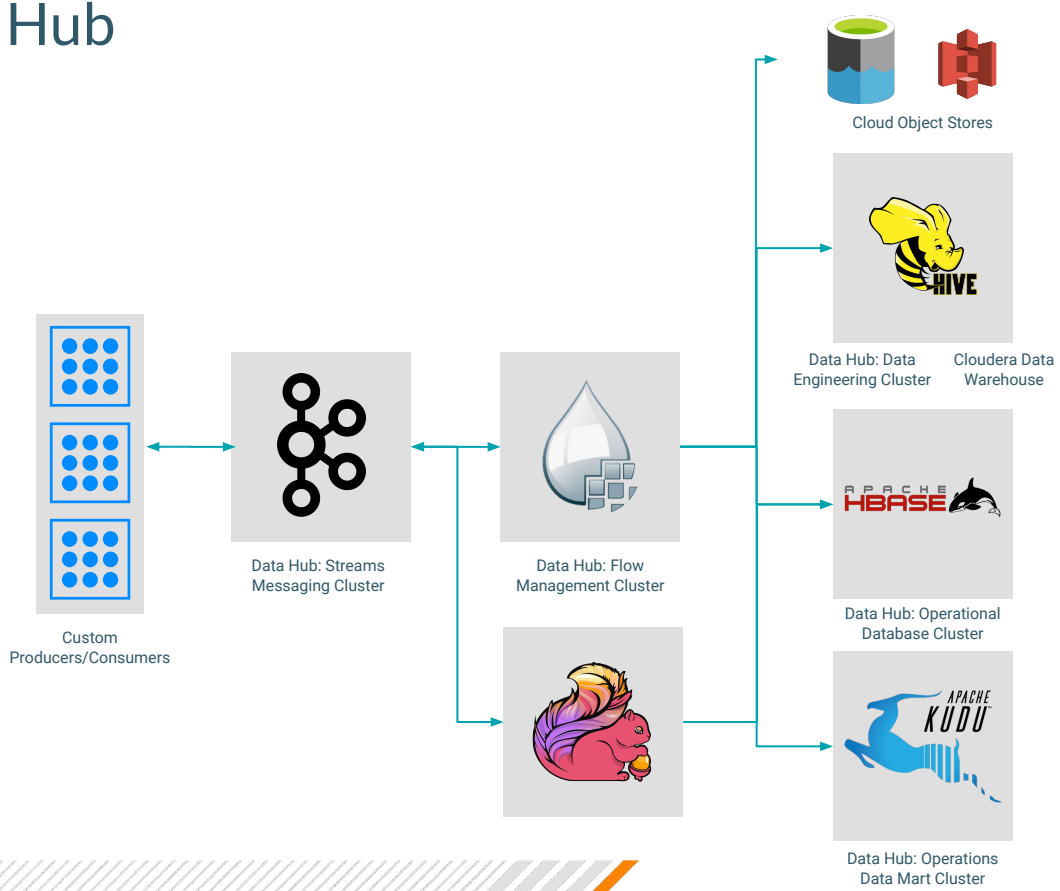
Protected against cloud infrastructure failures - VM reprovisioning

Pre-configured Atlas lineage integration

CDF Use Cases in CDP Data Hub

Cloud Native Streaming / CDP Ingest

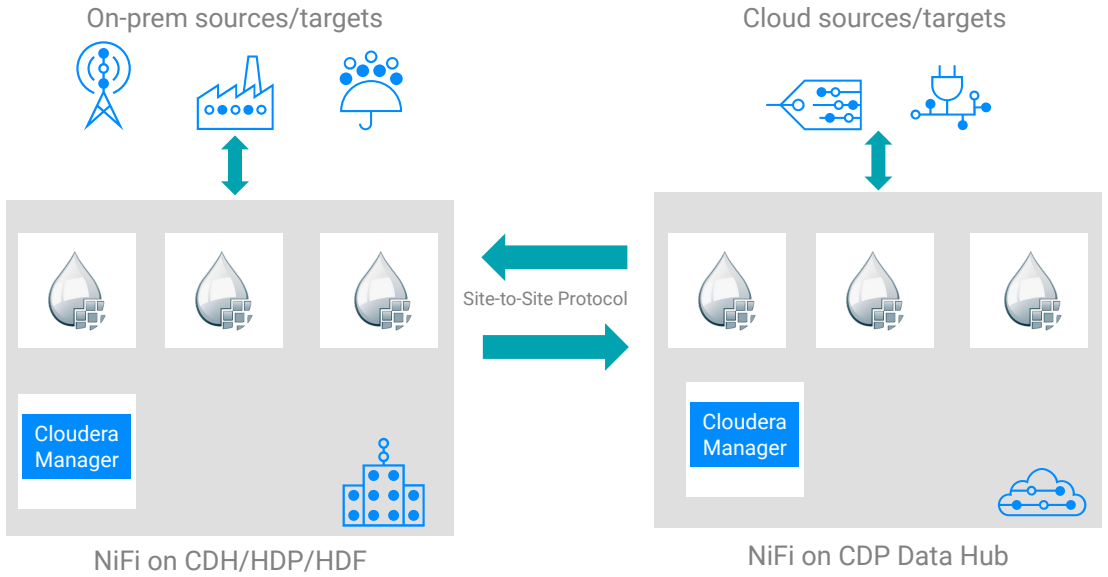
- Enables cloud native streaming applications
- Use NiFi or custom producers/consumers to interact with Kafka
- Use NiFi to power CDP Ingest
  - Hive
  - HBase
  - Kudu
  - Cloud object stores
- Use Flink for Streaming Analytics
  - Stateful and scalable applications
  - Complex Event Processing
  - SQL Streaming



Kafka and NiFi Use Cases in CDP Data Hub

Hybrid NiFi deployments / Cloud migrations

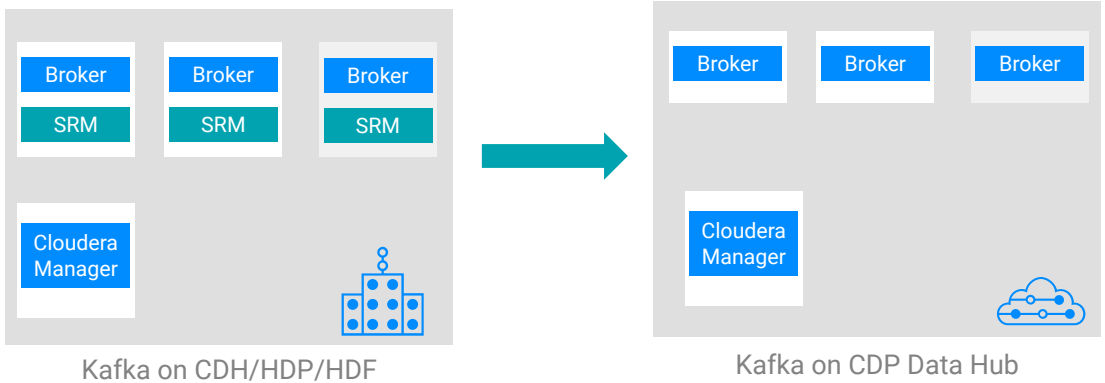
- Set up DataFlows that span on-prem and public cloud environments
- Merge on-prem sources with cloud sources
- Migrate on-prem data sets and make them available for cloud applications



Kafka and NiFi Use Cases in CDP Data Hub

Hybrid Kafka deployments / Cloud migrations

- Set up Kafka in Data Hub as DR cluster
- Migrate existing on-prem Kafka clusters to the cloud
- Leverage Cloudera Streams Replication Manager on on-prem cluster



Kafka on CDH/HDP/HDF

Kafka on CDP Data Hub

EAP*

Cloud Native Streaming Pipelines

* **E**arly **A**ccess **P**rogram

Simplifying the User Experience

EAP

EAP

Cloudera DataFlow Service

Phased Approach for DataFlow Service Delivery

Phase 1 - Flows as a Service

- Shift users' mindsets from thinking about NiFi clusters to Flows
- Focus on Flow Deployment and Monitoring
- Define and assign SLAs to your flows
- Automatic infrastructure sizing and scaling based on workload characteristics and SLAs
- Central monitoring console for all your flows across environments

Phase 2 - Topic as a Service (Kafka) & Stream Apps (e.g: Flink)

- Shift users' mindsets from thinking about Kafka clusters to Topics and end to end data pipelines
- Users provide inputs on the profile, usage characteristics & schema of the topic. DataFlow Service creates the topic on either an existing or new Kafka cluster using K8S operator & scales the clusters to meet the.
- Provide common SDLC experience While Using the Right Compute Engine (e.g: Flink, NiFi) for the Use Case

EAP

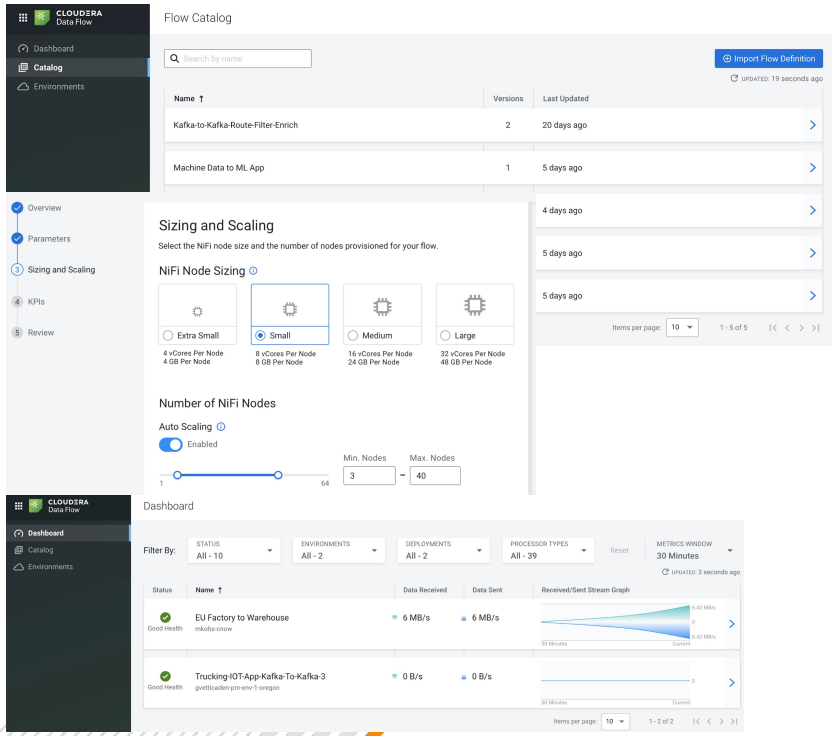
Flows as a Service

Cluster-Less Data Flow deployment and management

Shifting Users from Thinking about NiFi clusters to Data Flows

Key Capabilities

- Central *DataFlow Catalog* to discover and reuse existing Flow Definitions
- Provide *Flow Parameters, Configure Auto-Scaling, Define KPIs* in Deployment Wizard
- Monitor and Manage Flow Deployments in a central dashboard



Roadmap

Topics as a Service

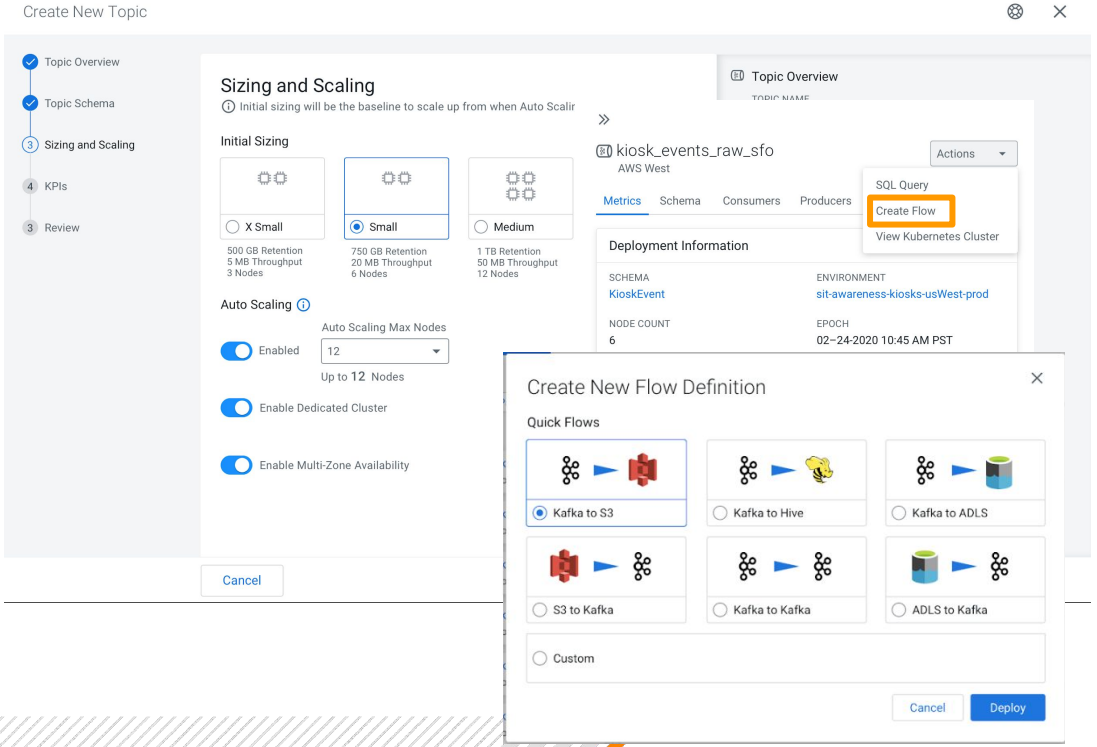
Roadmap

Cluster-Less Topic Creation and Code-less Producers/Consumers

Shifting Users from Thinking about Kafka Clusters to Topics

Key Capabilities

- Users provide **topic profile info including usage characteristics, schema and SLAs**
- DataFlow Service **creates the topic on either an existing or new Kafka cluster using K8S operator** & scales the clusters to meet the SLA/KPIS configured.
- Kafka topic actions include creating **data movement or streaming analytics apps**



Roadmap

Streaming Apps as a Service

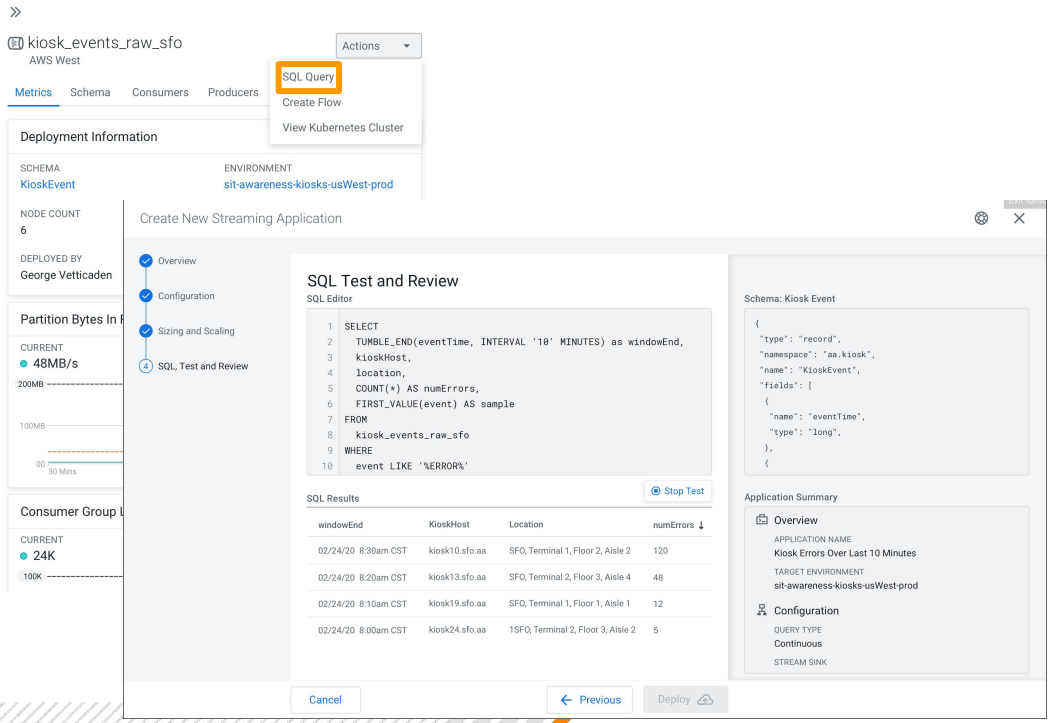
Roadmap

Agile Streaming Analytics App Development using SQL

Shifting Users from Thinking about Flink Clusters to Streaming Apps

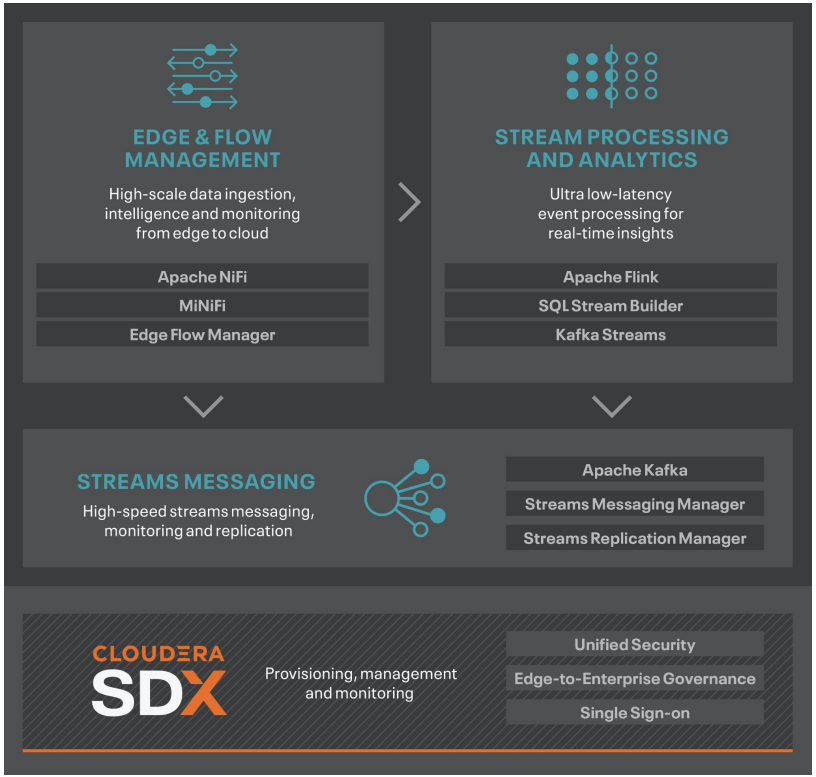
Key Capabilities

- Support for **Flink StreamingSQL**
- Select a Kafka Topic and **execute streaming SQL that runs as a continuous query**
- Configure streaming SQL to **write streaming results into Kafka Endpoint**
- **SQL Editor** in Data Flow Service to write, test and deploy streaming SQL jobs



Conclusions

Cloudera DataFlow Data-in-motion Platform



What can I use today?

	CDH / HDP / HDF	CDP Private Cloud Base	CDP Public Cloud
NiFi / NiFi Registry	Yes	Yes	Yes
MiNiFi / EFM	Yes	Yes	
Kafka / KStreams	Yes	Yes	Yes
Schema Registry	Yes	Yes	Yes
SMM	Yes	Yes	Yes
SRM	Yes	Yes	Yes
KConnect		Yes	
Flink		Yes	Yes
SQL Stream Builder		Yes (7.1.6+)	
Ranger / Atlas	HDP/HDF only	Yes	Yes

Major organizations leverage CDF

Few examples from Dataworks Summit

Flow & Streaming use cases

https://dataworkssummit.com/san-jose-2018/session/using-nifi-to-simplify-data-flow-streaming-use-cases-mastercard/

Connected Plants & Industrial IoT

https://dataworkssummit.com/berlin-2018/session/apache-nifi-best-practices-and-lessons-learnt/

Distributed financial service with CDF

https://dataworkssummit.com/berlin-2018/session/synchronicity-of-a-distributed-financial-system/

Real-time Freight Visibility

https://dataworkssummit.com/san-jose-2018/session/real-time-freight-visibility-how-tmw-systems-uses-hdf-to-create-sub-second-transportation-visibility/

Major organizations leverage CDF

Few examples from Dataworks Summit

Internet of Fleet Management Things

https://dataworkssummit.com/washington-dc-2019/session/iofmt-internet-of-fleet-management-things/

Event Standardization Service

https://dataworkssummit.com/san-jose-2018/session/using-spark-streaming-and-nifi-for-the-next-generation-of-etl-in-the-enterprise/

IoT and Time Series data

https://dataworkssummit.com/san-jose-2018/session/exxonmobils-journey-to-unleash-time-series-data-with-open-source-technology/

High Speed Cybersecurity

https://dataworkssummit.com/washington-dc-2019/session/building-the-high-speed-cyber-security-data-pipeline-using-apache-nifi/