cloudera

INTRODUCTION TO CDF

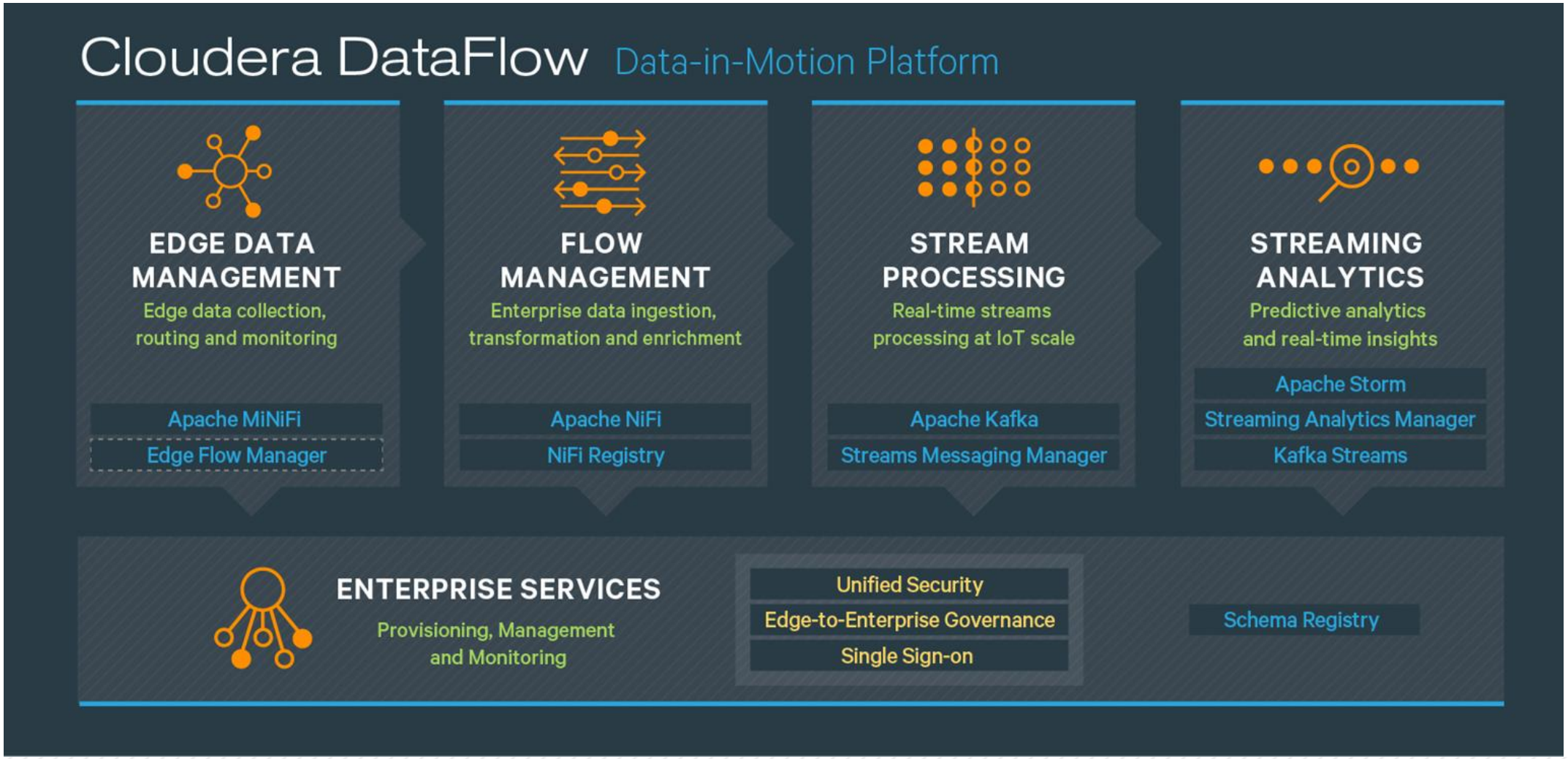
Pedro Algaba – Solution Engineer

AGENDA

- Cloudera Data Flow platform
- Flow & Edge management
- Technical details
- Stream Processing

CLOUDERA DATA FLOW

Cloudera dataflow Platform



HDF 3.3 Platform

Orange available in **HDP+HDF** scenario

FLOW MANAGEMENT

Data acquisition and delivery
Simple transformation and data routing
Simple event processing
End to end provenance
Edge intelligence & bi-directional communication

Java Agent

C++ Agent

STREAM PROCESSING

Scalable data broker for streaming apps
Scale Out Streaming Computation Engine

STREAM ANALYTICS

Pattern Matching
Prescriptive & Predictive Stream Analytics
Complex Event Processing
Continuous Insight

ENTERPRISE SERVICES

Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance, Multi-tenancy

NiFi Flow Registry

Hortonworks Schema Registry

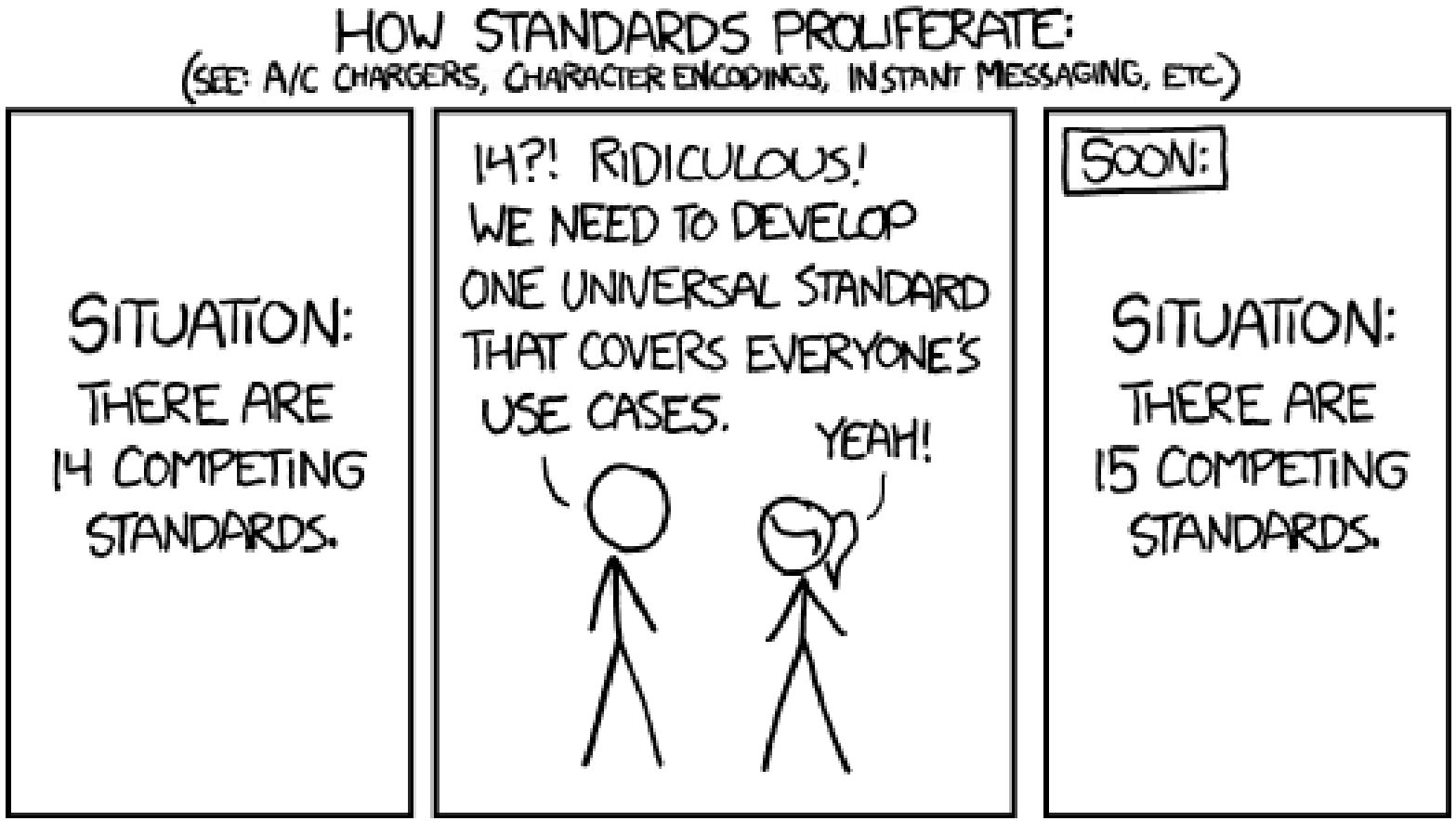
Flow & Edge Management

Today’s Data Landscape

- Teradata
- RDBMS
- Hive
- HDFS
- S3
- ADLS
- Google PubSub
- HBase
- Druid
- Memcache
- Solr
- Kafka
- Cassandra
- ElasticSearch
- AWS Redshift
- Couchbase
- JMS Queues
- MQTT
- REST
- SFTP
- MongoDB
- AeroSpike
- Redis
- Splunk
- Syslog
- InfluxDB

What tools do you use?

- Sqoop
- Flume
- ETL tools
- Script
- Script_V2
- Script_incremental_tets
- …



Challenges with Current Frameworks

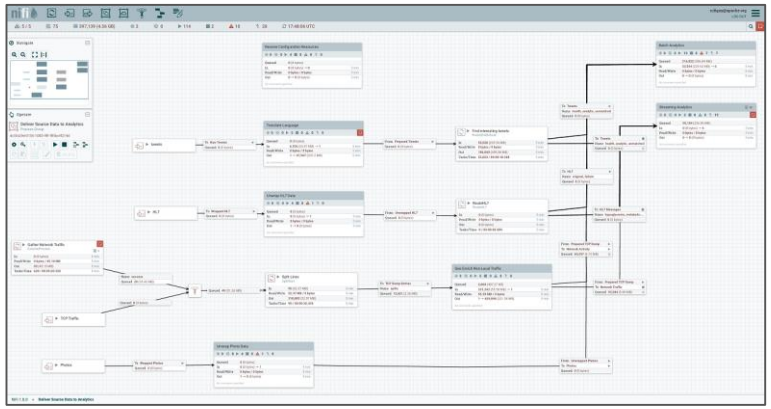
- Most frameworks on the market are focused on transformations (i.e. ETL vs EL)
  - **Traditional**: Informatica, Pentaho, Talend, SSIS
  - **Big Data**: MR, Tez, Pig, Hive, Spark, Storm, Flink
- EL only requires simple processing (e.g. translation)
  - i.e. it is a map-only MR job with no shuffle; a Storm topology with one bolt
- Sqoop, DistCp, and MirrorMaker are specialized to a specific route
- Kafka Connect requires Kafka as an intermediary

=> NiFi is a general purpose EtL engine

Flow Management

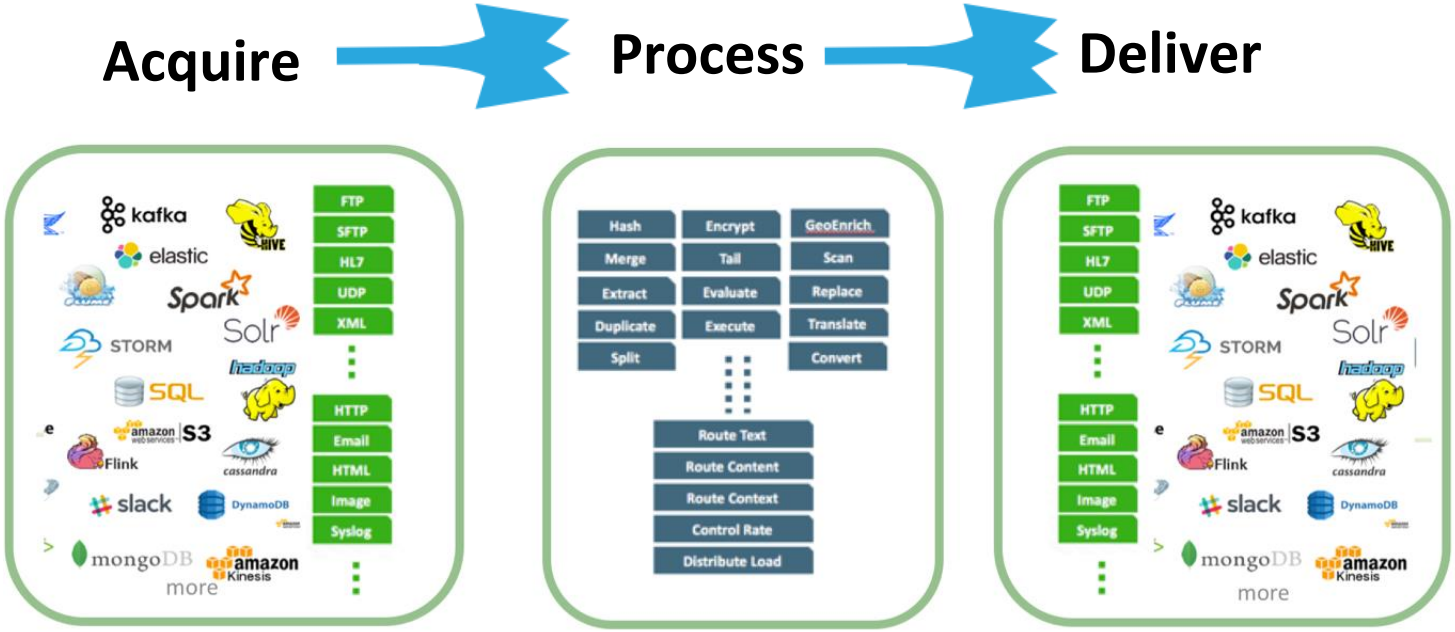
Enable easy ingestion, routing, management and delivery of any data anywhere (Edge, cloud, data center) to any downstream system with built in end-to-end security and provenance

APACHE nifi APACHE NIFI minifi



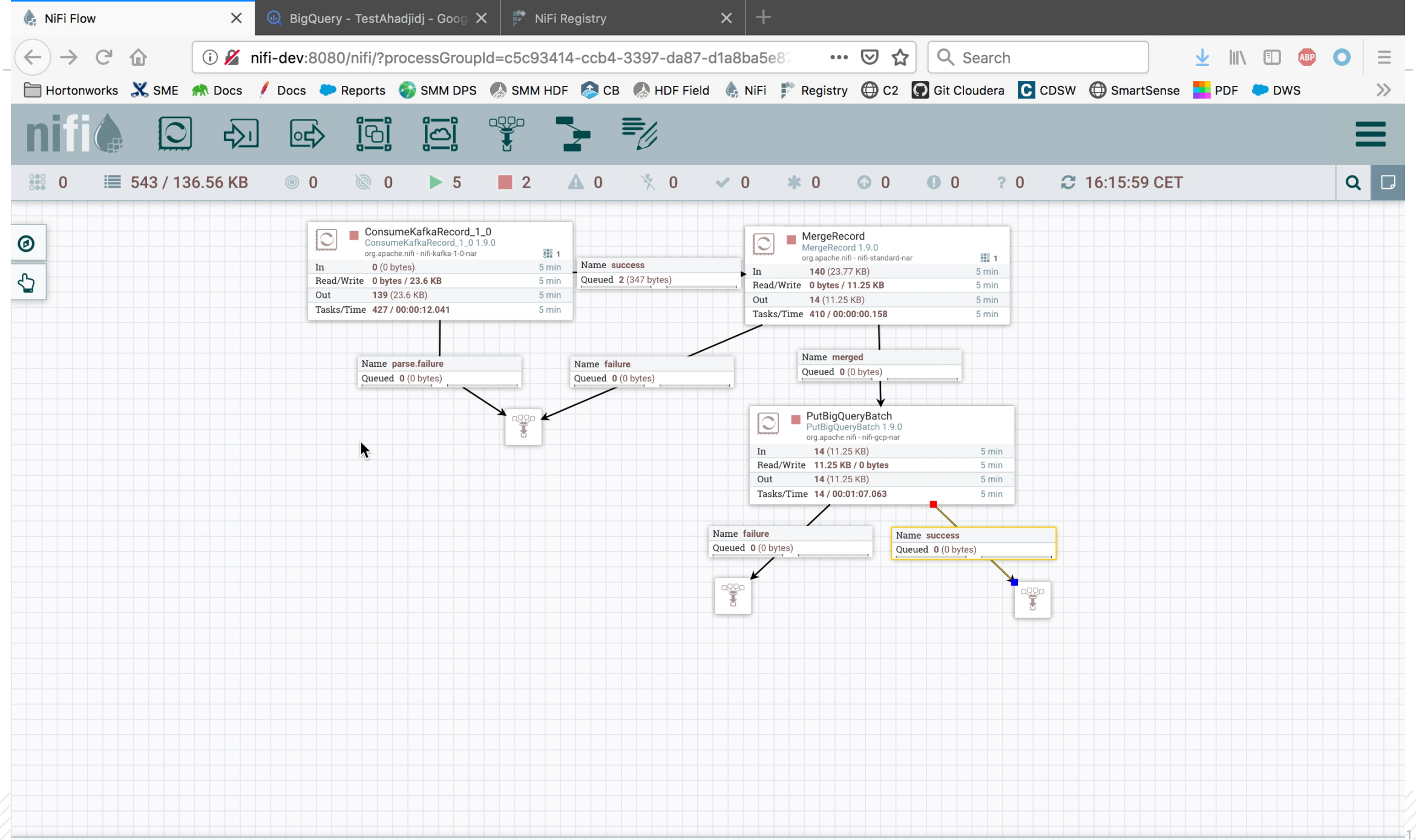
APACHE NIFI registry

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



- Over 280+ 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


NiFi Flow
BigQuery - TestAhadjidj - Goog
NiFi Registry
nifi-dev:8080/nifi/?processGroupId=c5c93414-ccb4-3397-da87-d1a8ba5e87
Search
Hortonworks
SME
Docs
Docs
Reports
SMM DPS
SMM HDF
CB
HDF Field
NiFi
Registry
C2
Git Cloudera
CDSW
SmartSense
PDF
DWS
0
543 / 136.56 KB
0
0
5
2
0
0
0
0
0
0
16:15:59 CET
ConsumeKafkaRecord_1_0
ConsumeKafkaRecord_1_0 1.9.0
org.apache.nifi - nifi-kafka-1-0-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 23.6 KB 5 min
Out 139 (23.6 KB) 5 min
Tasks/Time 427 / 00:00:12.041 5 min
Name success
Queued 2 (347 bytes)
MergeRecord
MergeRecord 1.9.0
org.apache.nifi - nifi-standard-nar
In 140 (23.77 KB) 5 min
Read/Write 0 bytes / 11.25 KB 5 min
Out 14 (11.25 KB) 5 min
Tasks/Time 410 / 00:00:00.158 5 min
Name parse.failure
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)
Name merged
Queued 0 (0 bytes)
PutBigQueryBatch
PutBigQueryBatch 1.9.0
org.apache.nifi - nifi-gcp-nar
In 14 (11.25 KB) 5 min
Read/Write 11.25 KB / 0 bytes 5 min
Out 14 (11.25 KB) 5 min
Tasks/Time 14 / 00:01:07.063 5 min
Name failure
Queued 0 (0 bytes)
Name success
Queued 0 (0 bytes)
NiFi Flow » Cloud Demo

An overview of NiFi capabilities

Data Ingest

SFTP
HL7
UDP
HTTP
WS
Syslog
MQTT

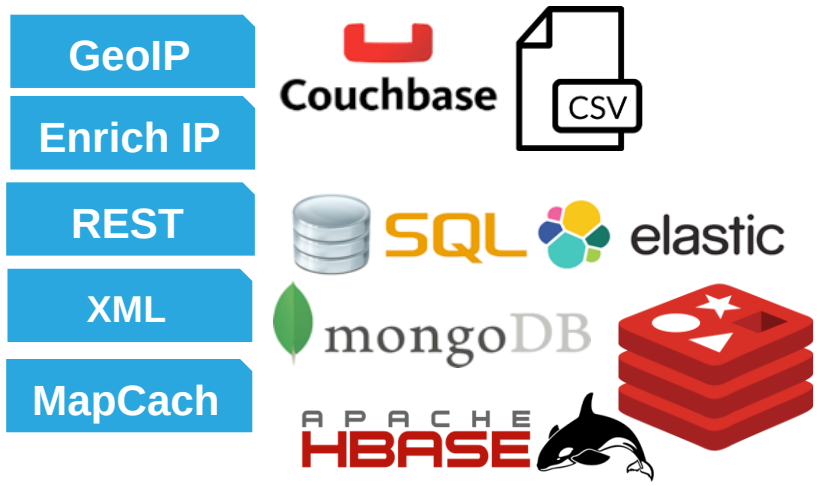
Data Transformation

Hash
Merge
Compress
Duplicate
Split
Encrypt

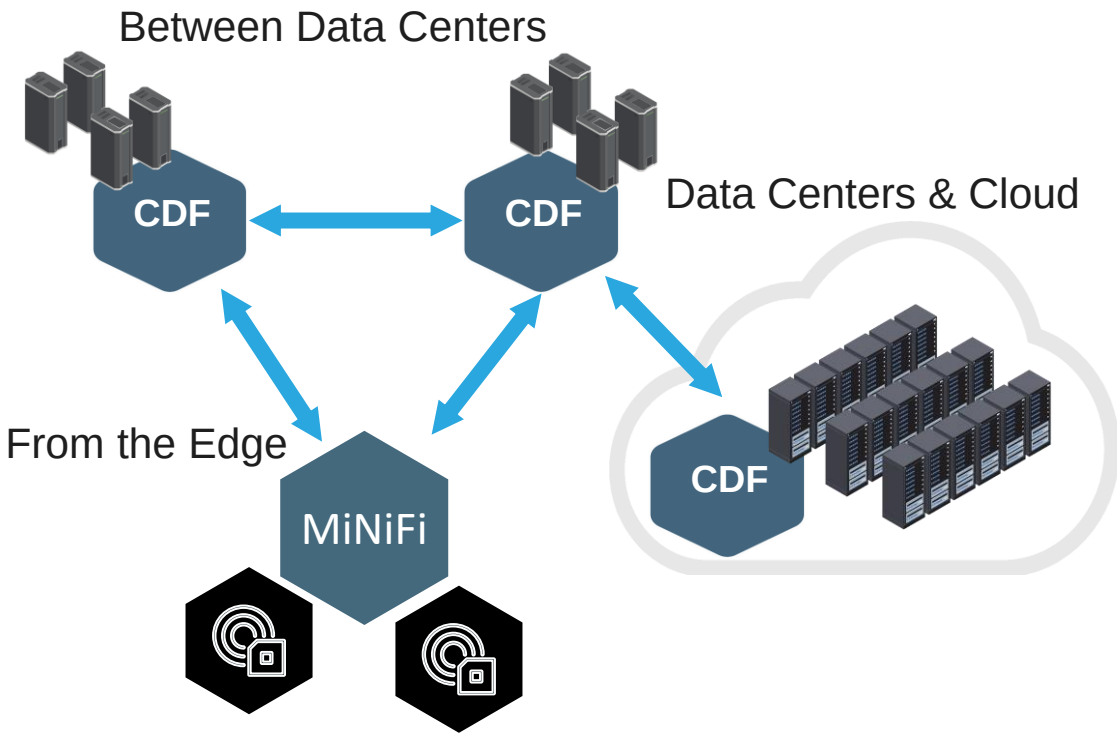
Syslog

An overview of NiFi capabilities

Data Enrichment



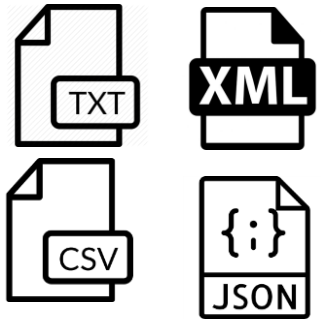
Data Movement



An overview of NiFi capabilities

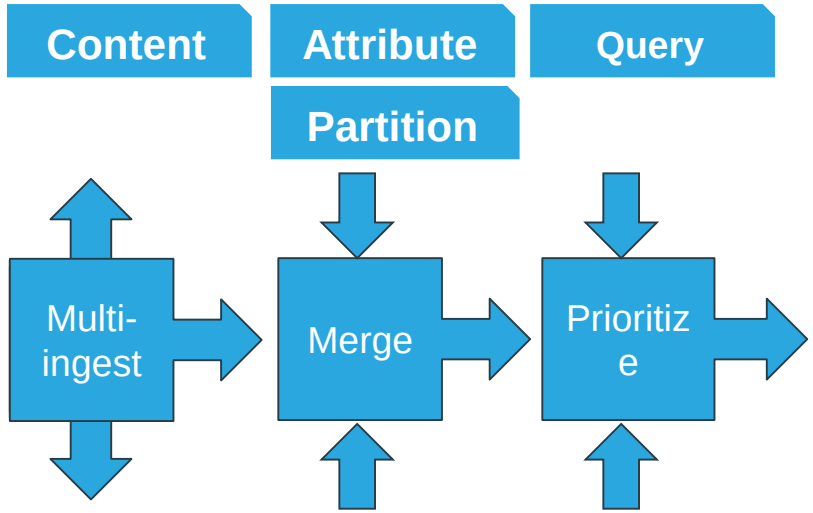
Parsing

- Grok
- Regex
- Content
- Attribute
- HL7
- NetFlow



Syslog

Routing



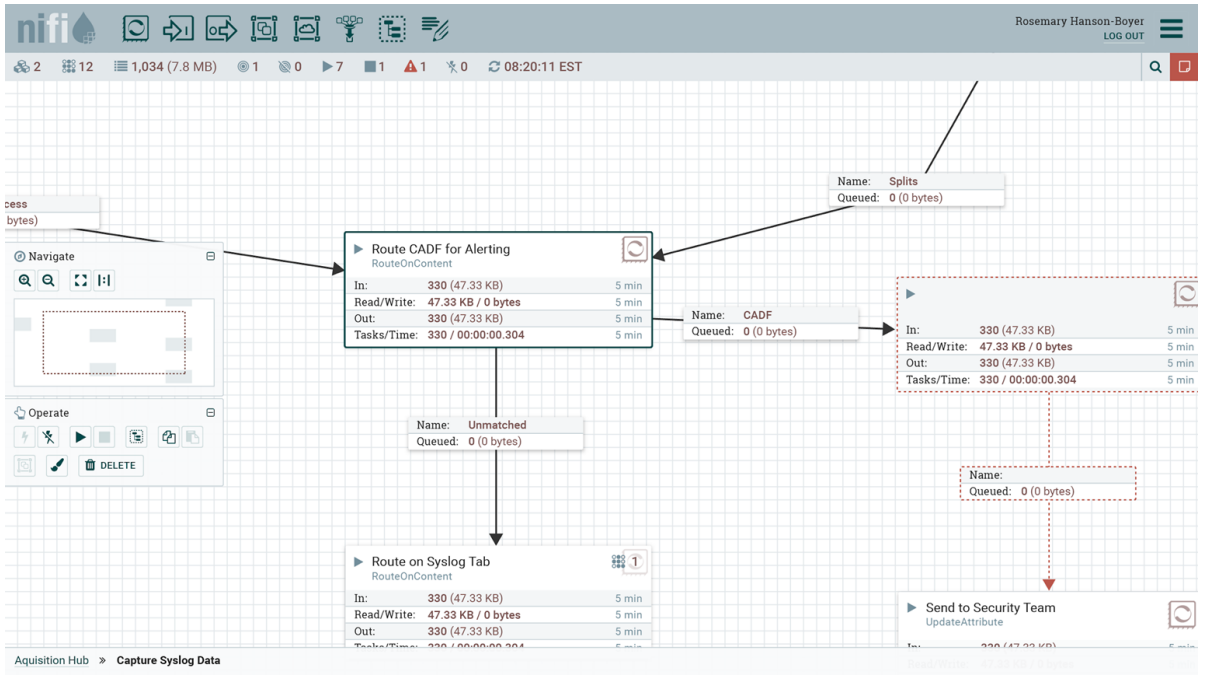
Why should you care about NiFi?

- **Accelerates development time** with its UI and more than 260 OOTB processors
- **Reduces costs** by offloading expensive techs such as CFT, Talend
- **Industrializes data movements** with central monitoring, flow/variable registries
- **Reduces risks** with Security and lineage (Ranger, Atlas, NiFi provenance)
- **Improves visibility** with data ingest centralization and monitoring
- **Breaks silos** with large ecosystem integration
- **Improves application performance** with horizontal and vertical scalability
- **Enables new use case** such as log optimization, IoT, streaming, etc

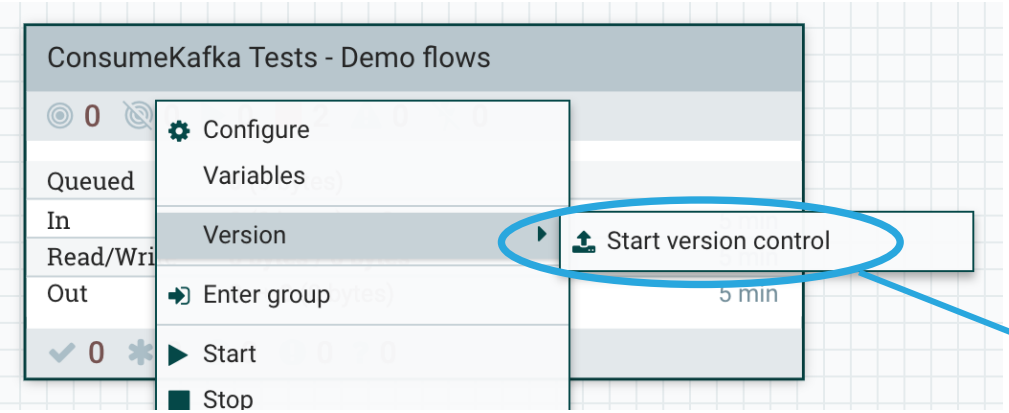
TECHNICAL DETAILS

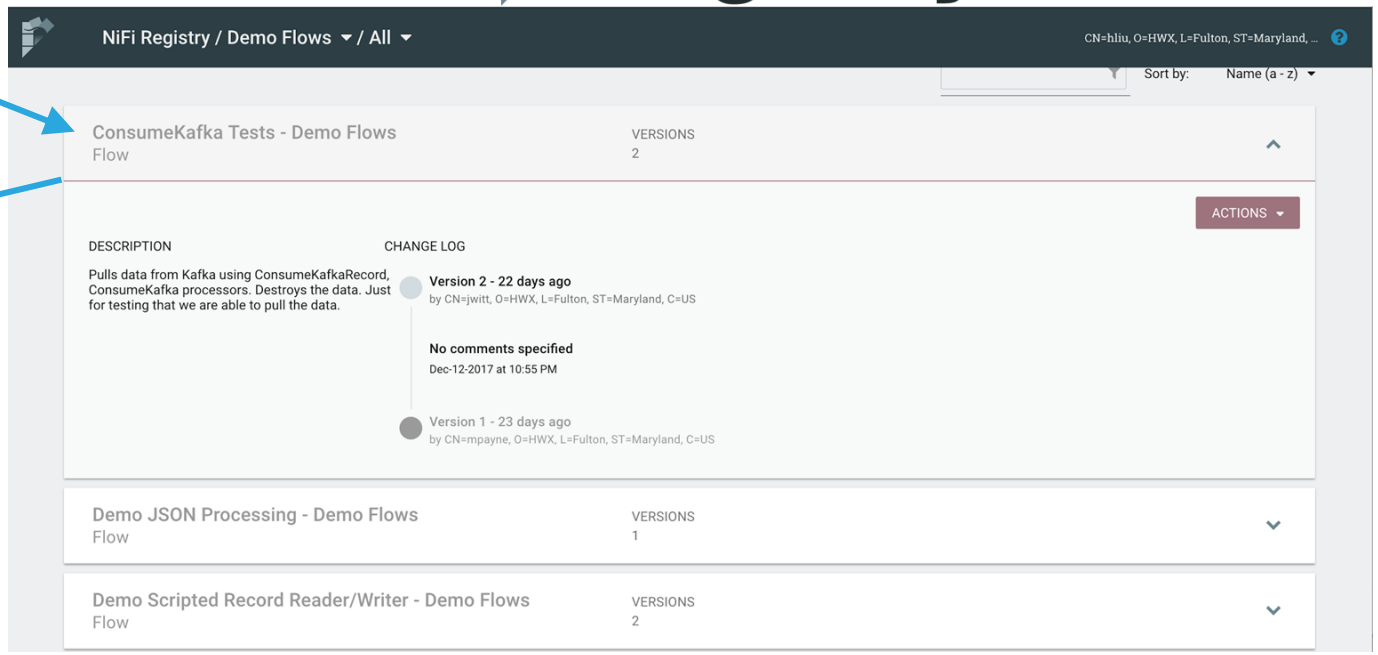
Apache NiFi High Level Capabilities

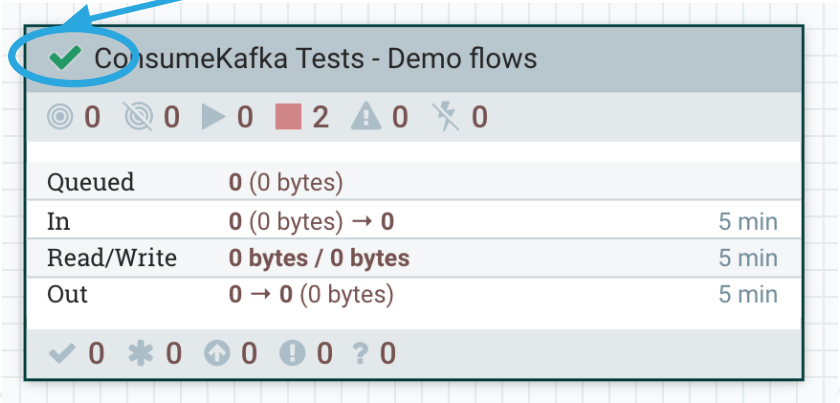
- Strong Out Of The Box
  - Flow can be modified at runtime
  - Back pressure
  - Loss tolerant vs guaranteed delivery
  - Low latency vs high throughput
  - Dynamic prioritization
- Secure
  - SSL, HTTPS, SFTP, etc.
  - End to end data provenance
- Designed for extension
  - Provider are a first class citizen
  - Build your own processors and Controller services
  - Integrate with external systems (Security, Monitoring, Governance, etc)

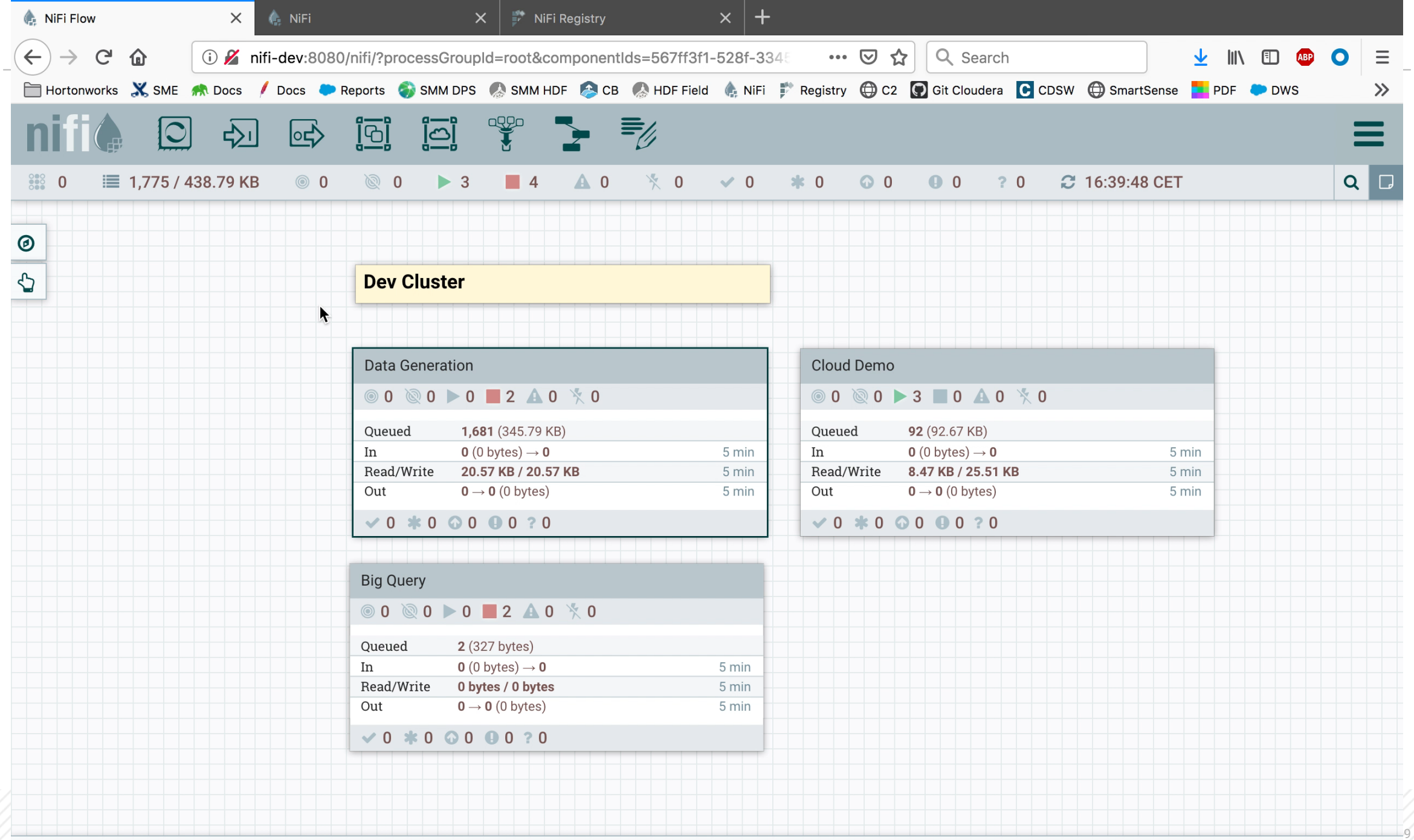


NiFi Flow Registry

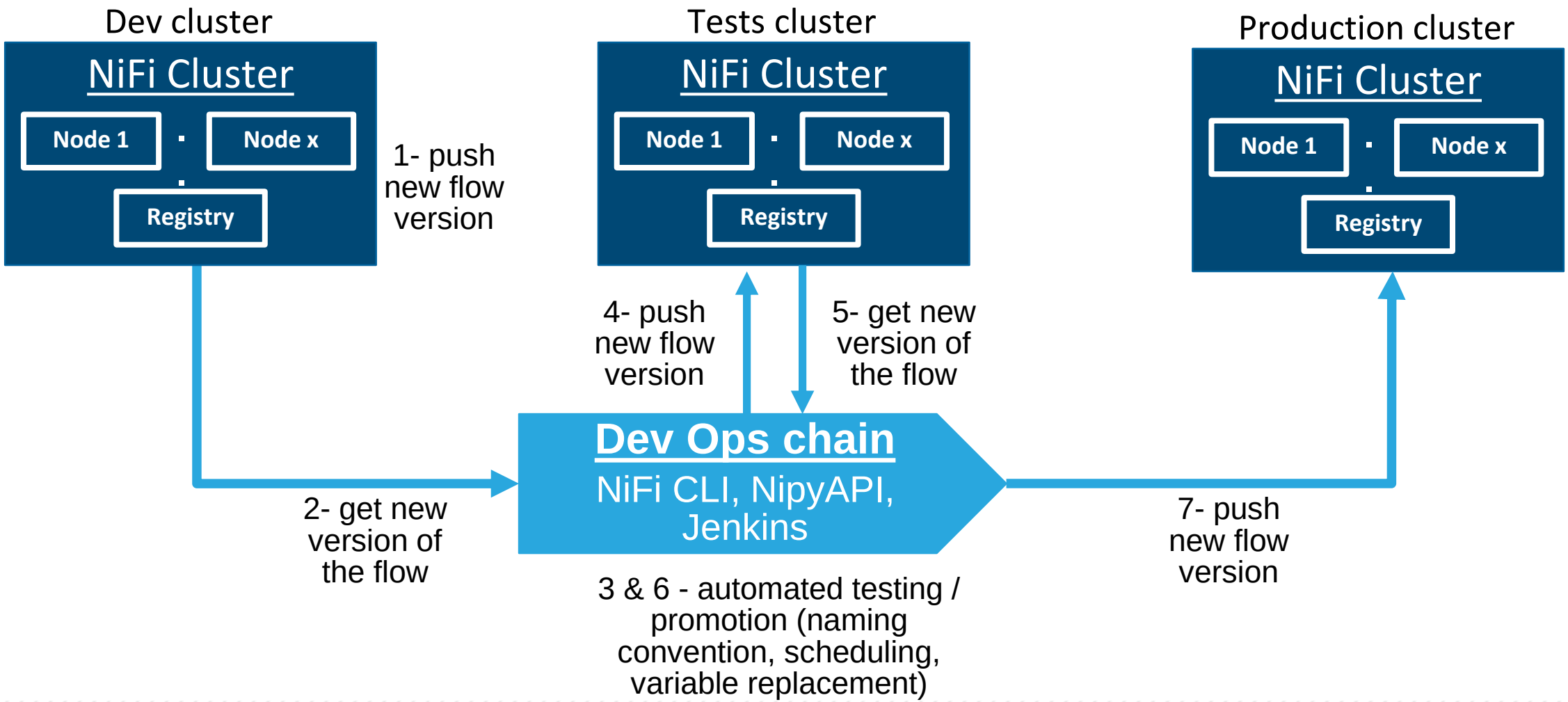






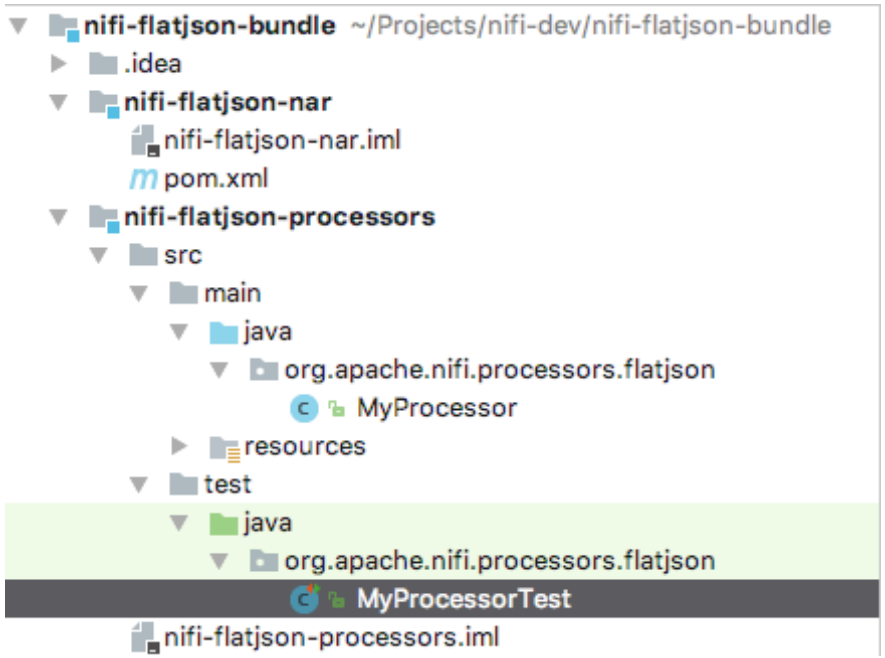

NiFi Flow
NiFi
NiFi Registry
nifi-dev:8080/nifi/?processGroupId=root&componentIds=567ff3f1-528f-334
Search
Hortonworks
SME
Docs
Docs
Reports
SMM DPS
SMM HDF
CB
HDF Field
NiFi
Registry
C2
Git Cloudera
CDSW
SmartSense
PDF
DWS
nifi
0
1,775 / 438.79 KB
0
0
3
4
0
0
0
0
0
0
16:39:48 CET
Dev Cluster
Data Generation
0 0 0 2 0 0
Queued 1,681 (345.79 KB)
In 0 (0 bytes) → 0 5 min
Read/Write 20.57 KB / 20.57 KB 5 min
Out 0 → 0 (0 bytes) 5 min
0 0 0 0 0
Cloud Demo
0 0 3 0 0 0
Queued 92 (92.67 KB)
In 0 (0 bytes) → 0 5 min
Read/Write 8.47 KB / 25.51 KB 5 min
Out 0 → 0 (0 bytes) 5 min
0 0 0 0 0
Big Query
0 0 0 2 0 0
Queued 2 (327 bytes)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 → 0 (0 bytes) 5 min
0 0 0 0 0

FDLC: Flow Development LifeCycle



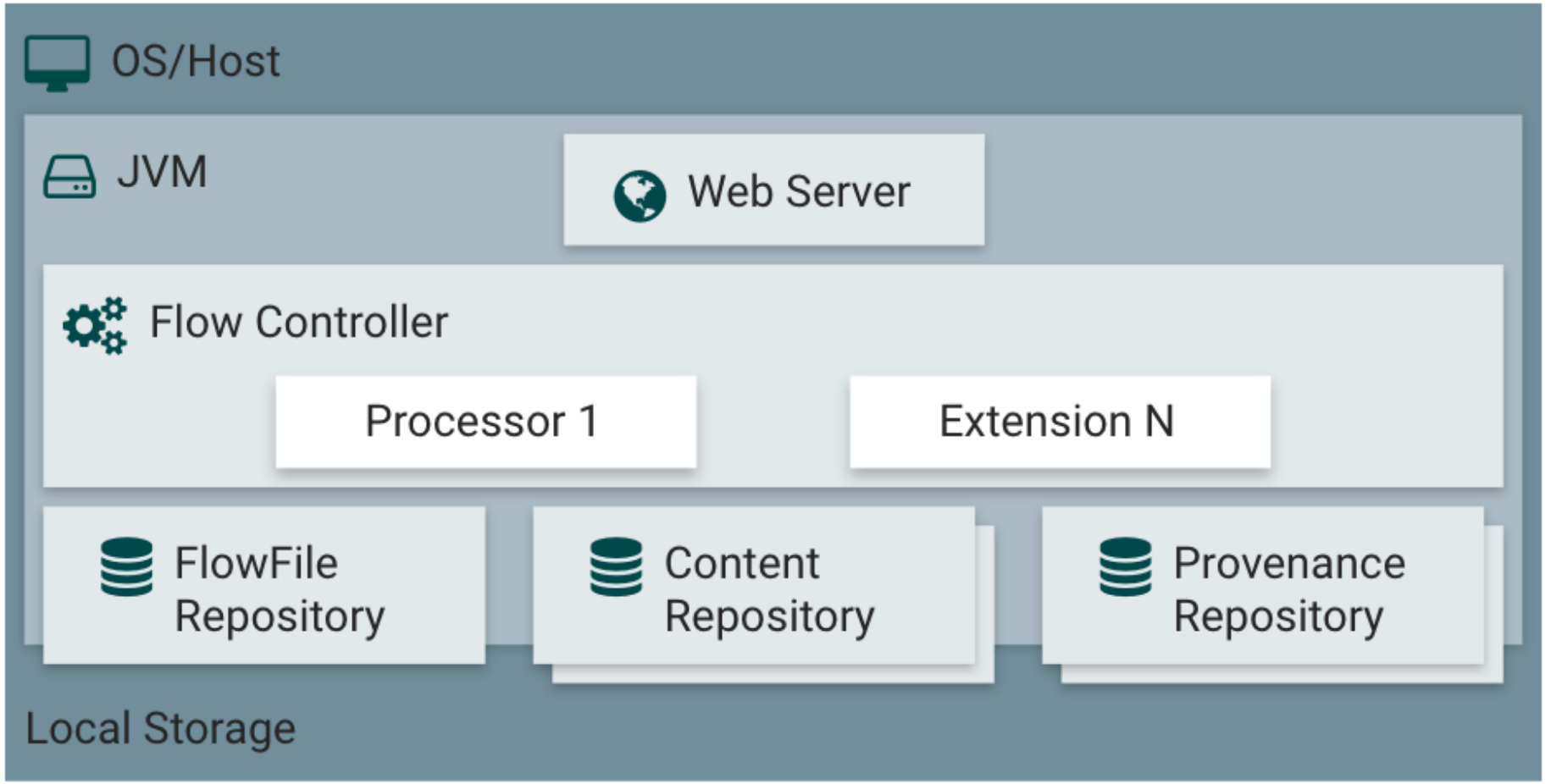
NiFi - Extensibility

- Built from the ground up with extensions in mind
- Service-loader pattern for…
  - Processors
  - Controller Services
  - Reporting Tasks
  - Prioritizers
- Extensions packaged as NiFi Archives (NARs)
  - Deploy NiFi lib directory and restart
  - Same model as standard components

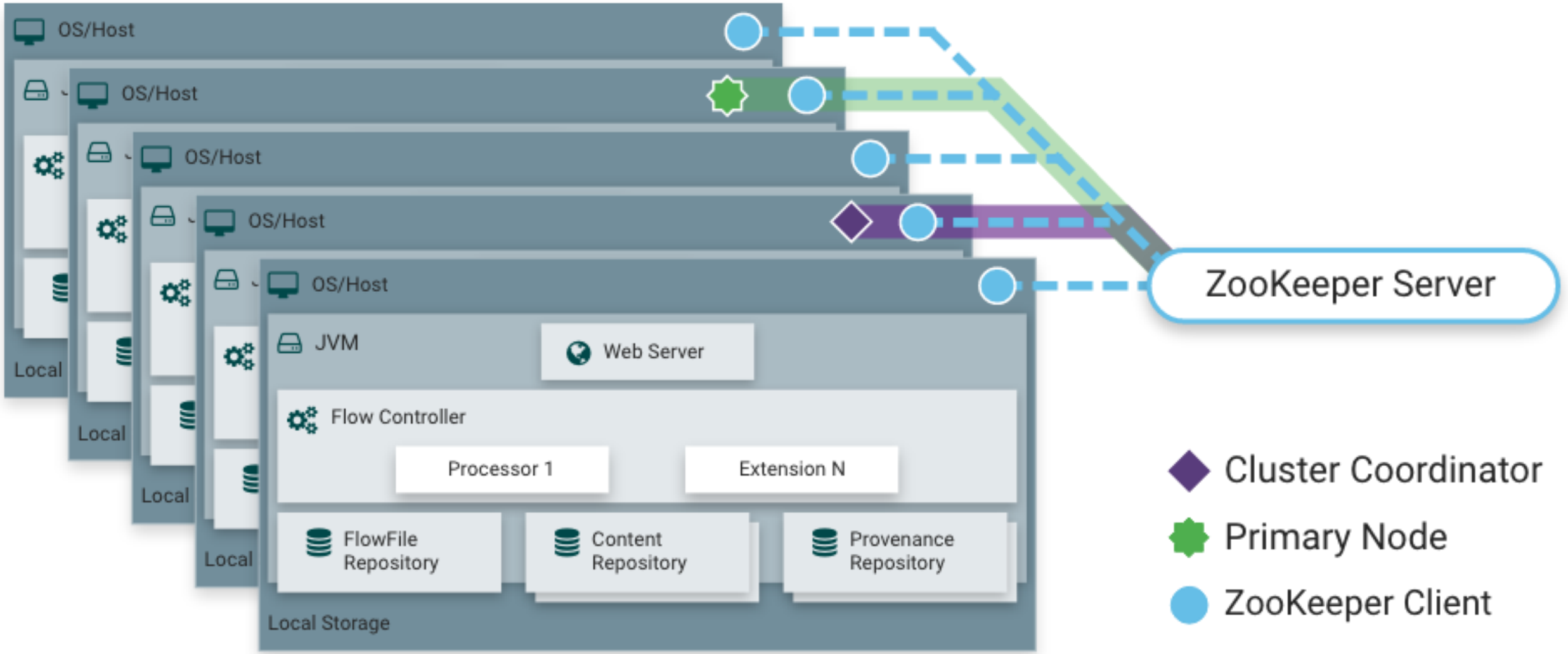



```
<parent>
    <groupId>org.apache.nifi</groupId>
    <artifactId>nifi-flatjson-bundle</artifactId>
    <version>1.0-SNAPSHOT</version>
</parent>

<artifactId>nifi-flatjson-processors</artifactId>
<packaging>jar</packaging>

<dependencies>
    <dependency>
        <groupId>com.github.wnameless</groupId>
        <artifactId>json-flattener</artifactId>
        <version>0.4.0</version>
    </dependency>
    <dependency>
        <groupId>org.apache.nifi</groupId>
        <artifactId>nifi-api</artifactId>
    </dependency>
    <dependency>
        <groupId>org.apache.nifi</groupId>
        <artifactId>nifi-utils</artifactId>
    </dependency>
```

NiFi - Architecture

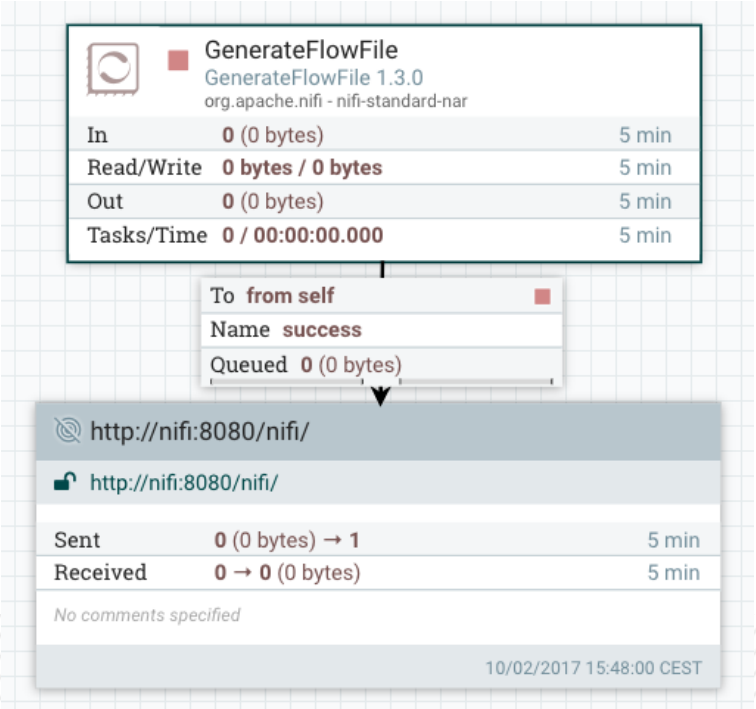


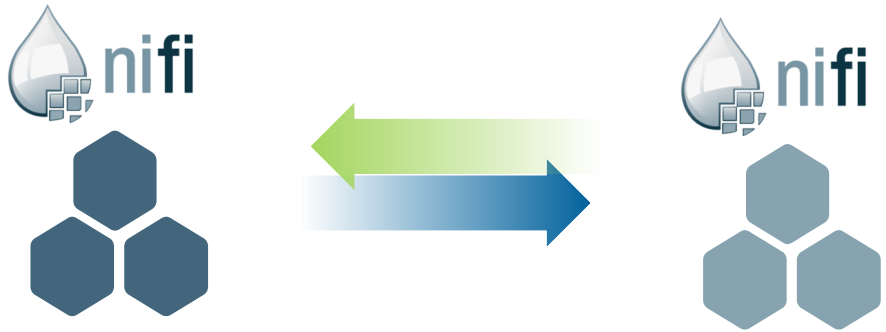
Scalable and distributed architecture

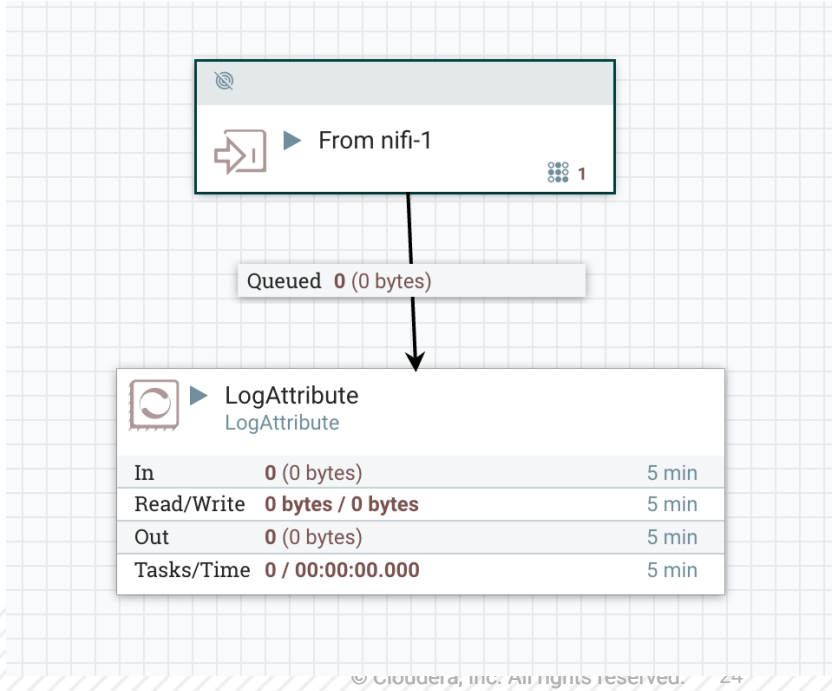


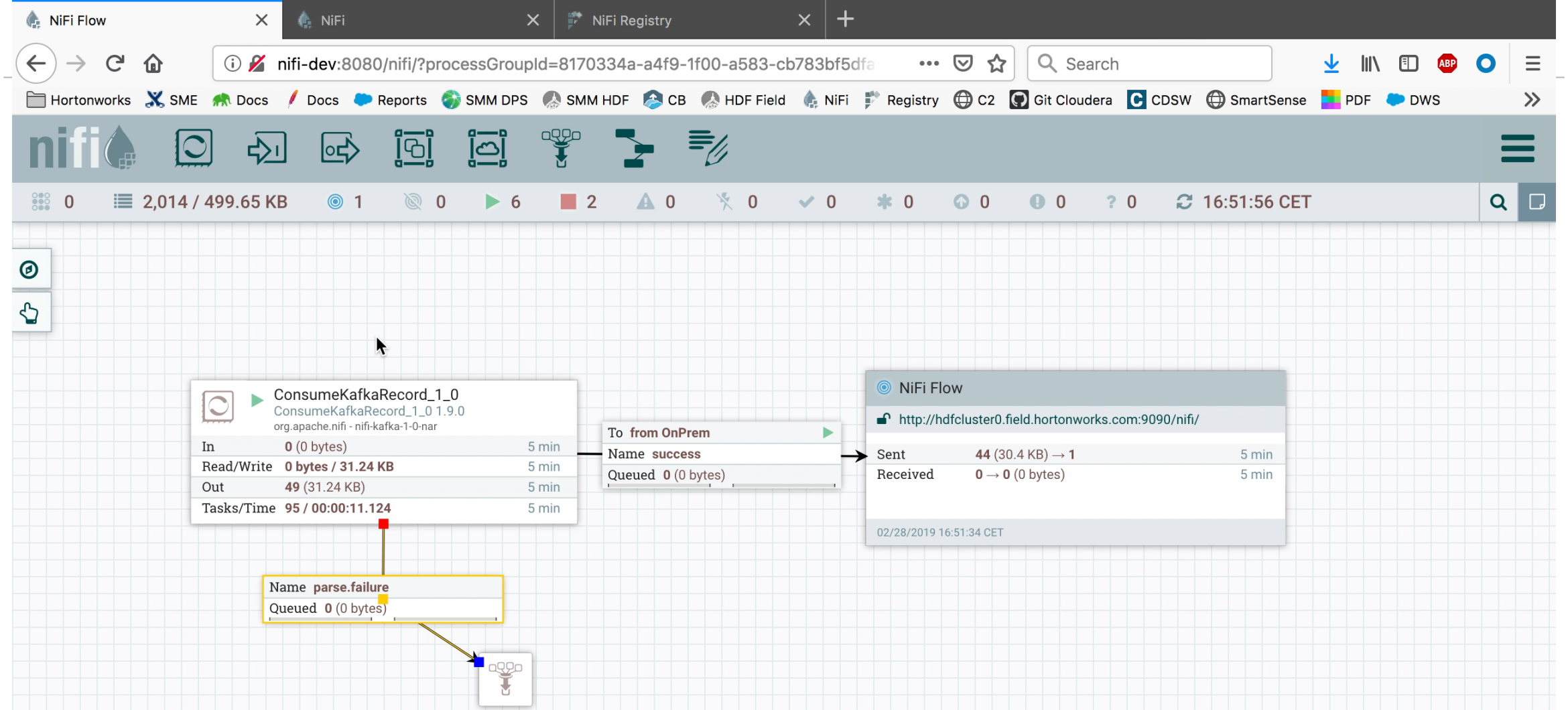
Hub and Spoke Architecture with Site-to-Site (S2S)

- Send data from a NiFi instance/cluster to one or multiple NiFi instances/clusters
- Preferred communications protocol when NiFi on both ends
- 2-way secure protocol, push & pull, high availability and load balancing

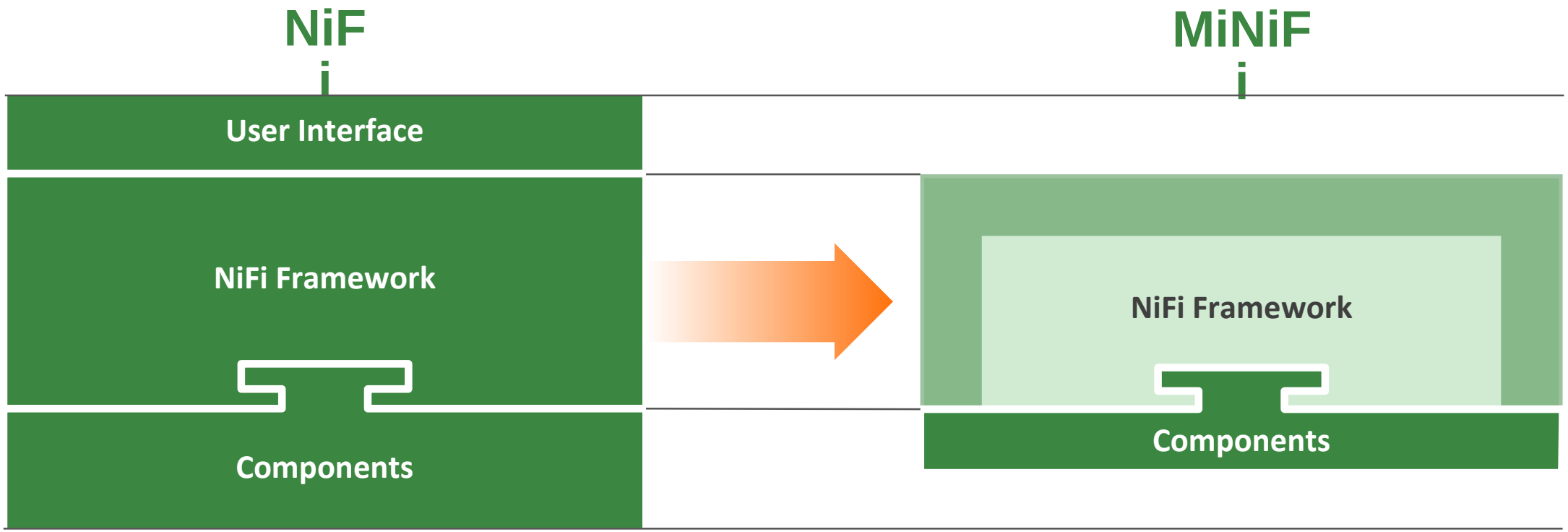






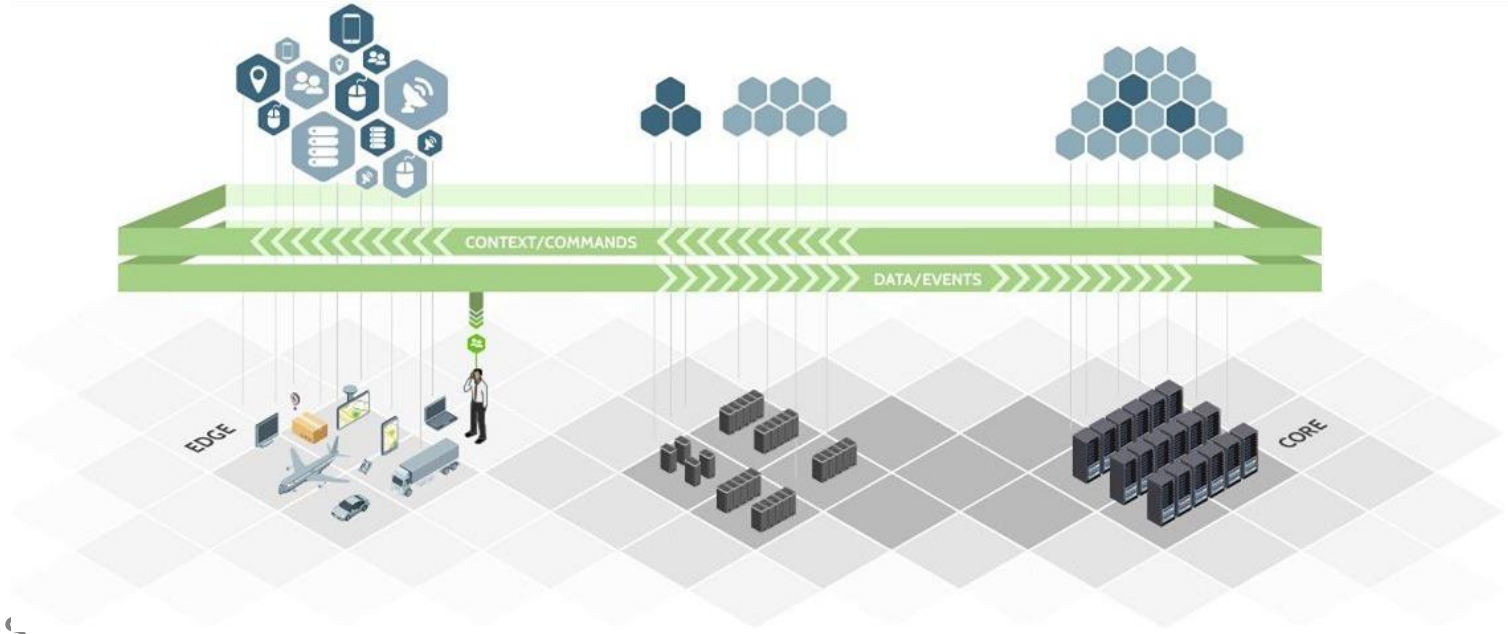

NiFi Flow
NiFi
NiFi Registry
nifi-dev:8080/nifi/?processGroupId=8170334a-a4f9-1f00-a583-cb783bf5dfa
Search
Hortonworks
SME
Docs
Docs
Reports
SMM DPS
SMM HDF
CB
HDF Field
NiFi
Registry
C2
Git Cloudera
CDSW
SmartSense
PDF
DWS
nifi
0
2,014 / 499.65 KB
1
0
6
2
0
0
0
0
0
0
16:51:56 CET
ConsumeKafkaRecord_1_0
ConsumeKafkaRecord_1_0 1.9.0
org.apache.nifi - nifi-kafka-1-0-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 31.24 KB 5 min
Out 49 (31.24 KB) 5 min
Tasks/Time 95 / 00:00:11.124 5 min
To from OnPrem
Name success
Queued 0 (0 bytes)
NiFi Flow
http://hdfcluster0.field.hortonworks.com:9090/nifi/
Sent 44 (30.4 KB) → 1 5 min
Received 0 → 0 (0 bytes) 5 min
02/28/2019 16:51:34 CET
Name parse.failure
Queued 0 (0 bytes)

NiFi vs. MiNiFi Java Prcessor, Smaller Footprint ~40 MB

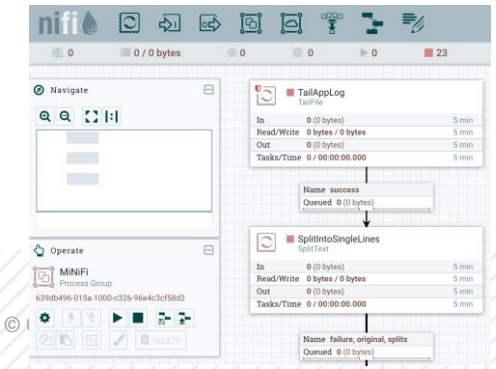


How Does MiNiFi Interact With NiFi?

- 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 Processing

Stream Processing

Apache Kafka

"Pub & Sub"

Massively Scalable Publish-Subscribe Message Queue

Apache Storm

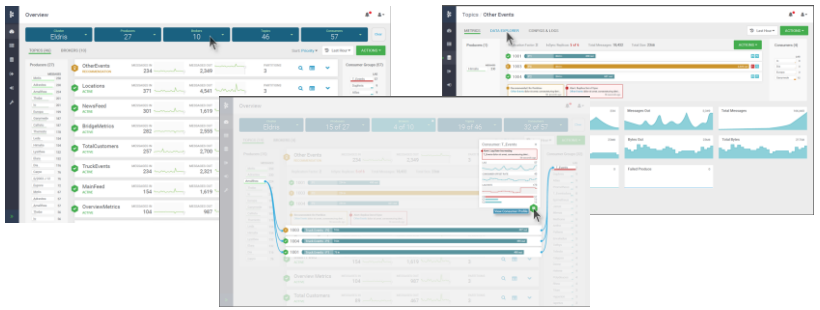
"Process"

Real-time, distributed data and task processing engine

Streams Messaging Manager

DPS

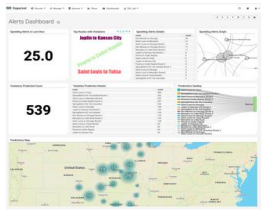
"Manage & Monitor"



- Cure Kafka blindness and help the different streaming personas be more productive
- End-to-end integration with Ambari, Grafana, Ranger & Atlas
- Comprehensive REST service for open integration

Streaming Analytics Manger

"Build & Analyze"



- Create streaming applications easily
- Visualize enrichment, matching, aggregations, notifications, etc

Our committer focus areas…

- Time to insights
- Security and governance
- DevOps and SDLC Support
- Management & Monitoring
- Stability and performance

5 Building blocks for robust and scalable real time solutions

A real-time integrated data logistics and simple event processing platform

Publish subscribe high-throughput, low-latency broker

API for building real-time applications and microservices on top of Kafka

Kafka's Omnipresence Has Led to the Onset of "Kafka Blindness"

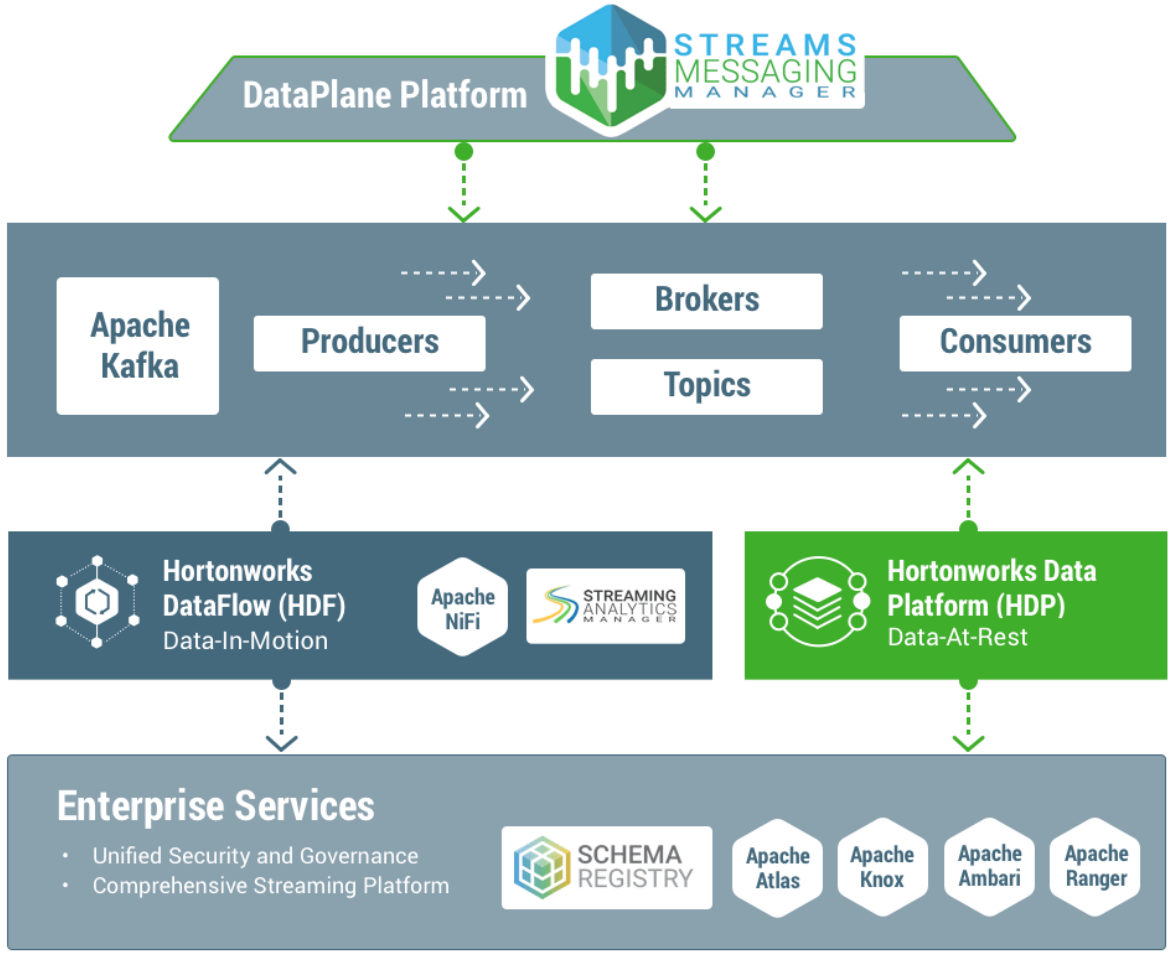
- What is **"Kafka Blindness"**?
  - Customers who use Kafka today struggle with monitoring / "seeing"/troubleshooting what is happening in their clusters

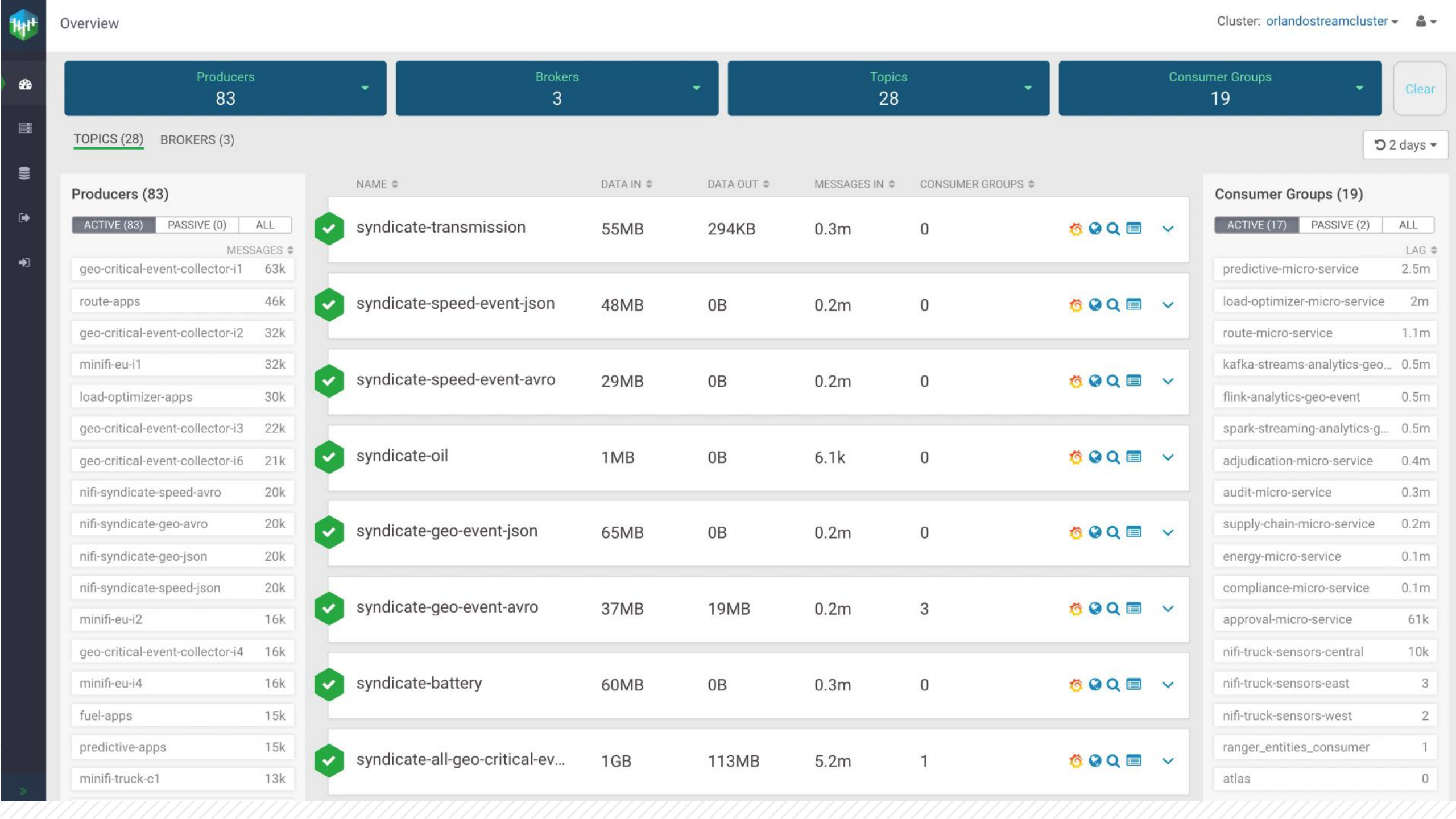
- Who is **Affected**?
  - Platform Operation Teams
  - Developers / DevOps Teams
  - Security / Governance Teams

- What are the **Symptoms**?
  - Difficulty seeing who is producing and consuming data
  - Difficulty understanding the flow of data from producers -> topics → consumers
  - Difficulty troubleshooting/monitoring.

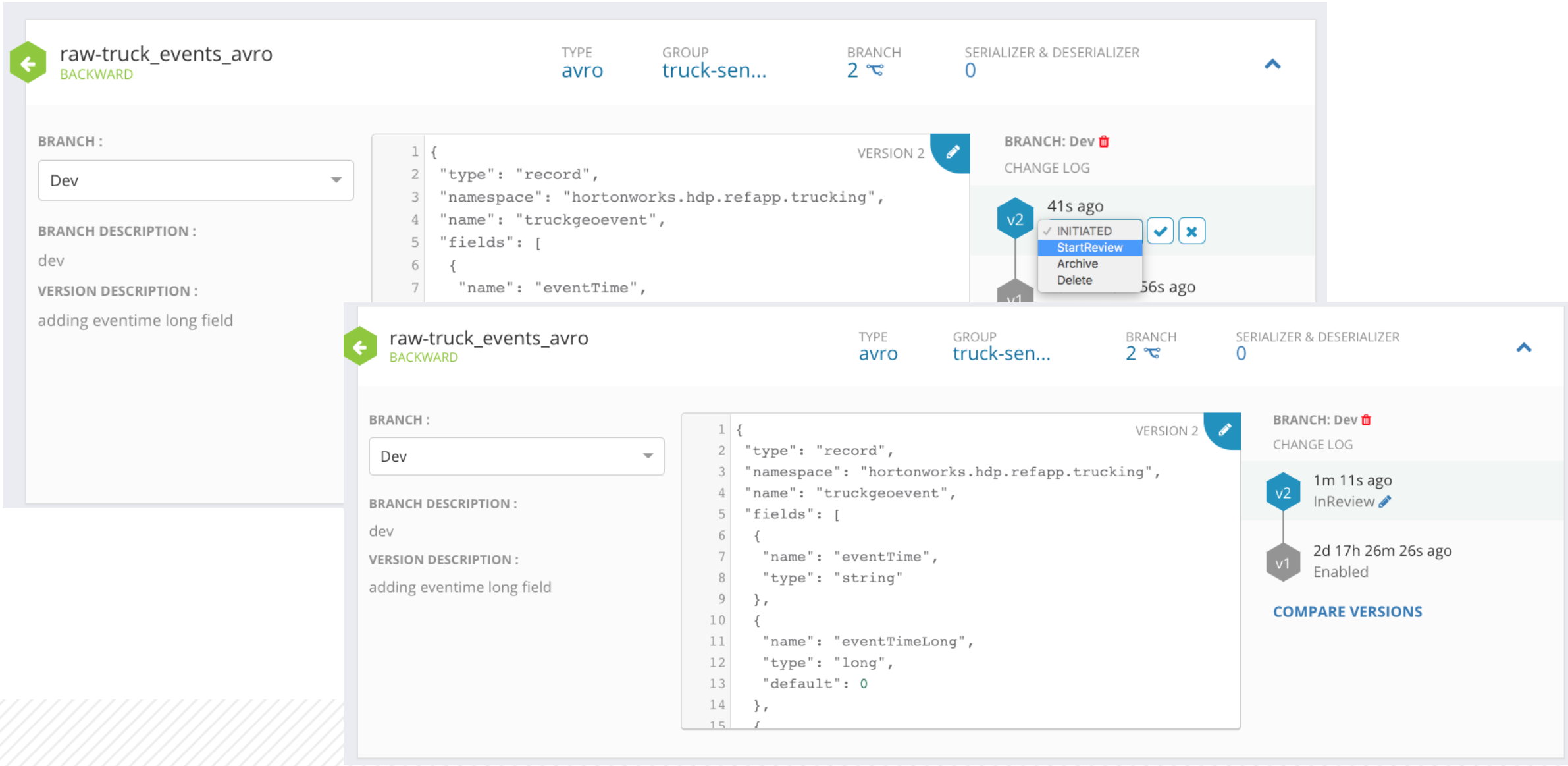
Hortonworks Streams Messaging Manager (SMM)

- New Open Source project led by Hortonworks to Cure the “Kafka Blindness”
- Single Monitoring Dashboard for all your Kafka Clusters across 4 entities
  - Broker
  - Topic
  - Producer
  - Consumer
- Designed for the Enterprise
  - Support for Secure/Kerborized Kafka cluster
  - Rich Access Control Policies (ACLS)
  - Supports multiple HDP and/or HDF Kafka Clusters
- REST as a First Class Citizen
- Delivered as a DataPlane Service

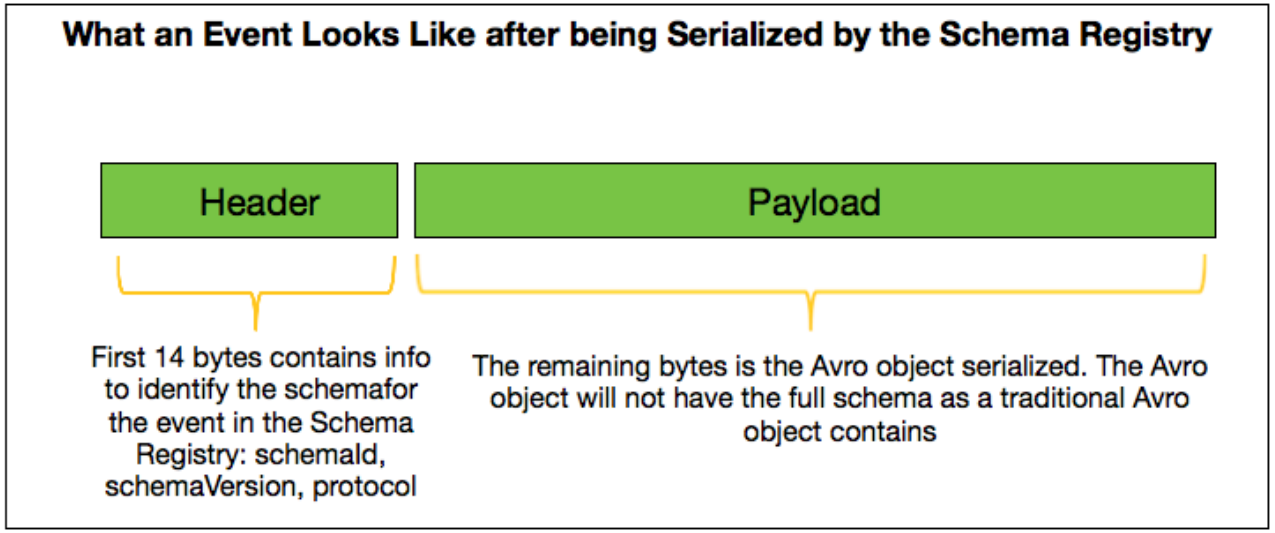


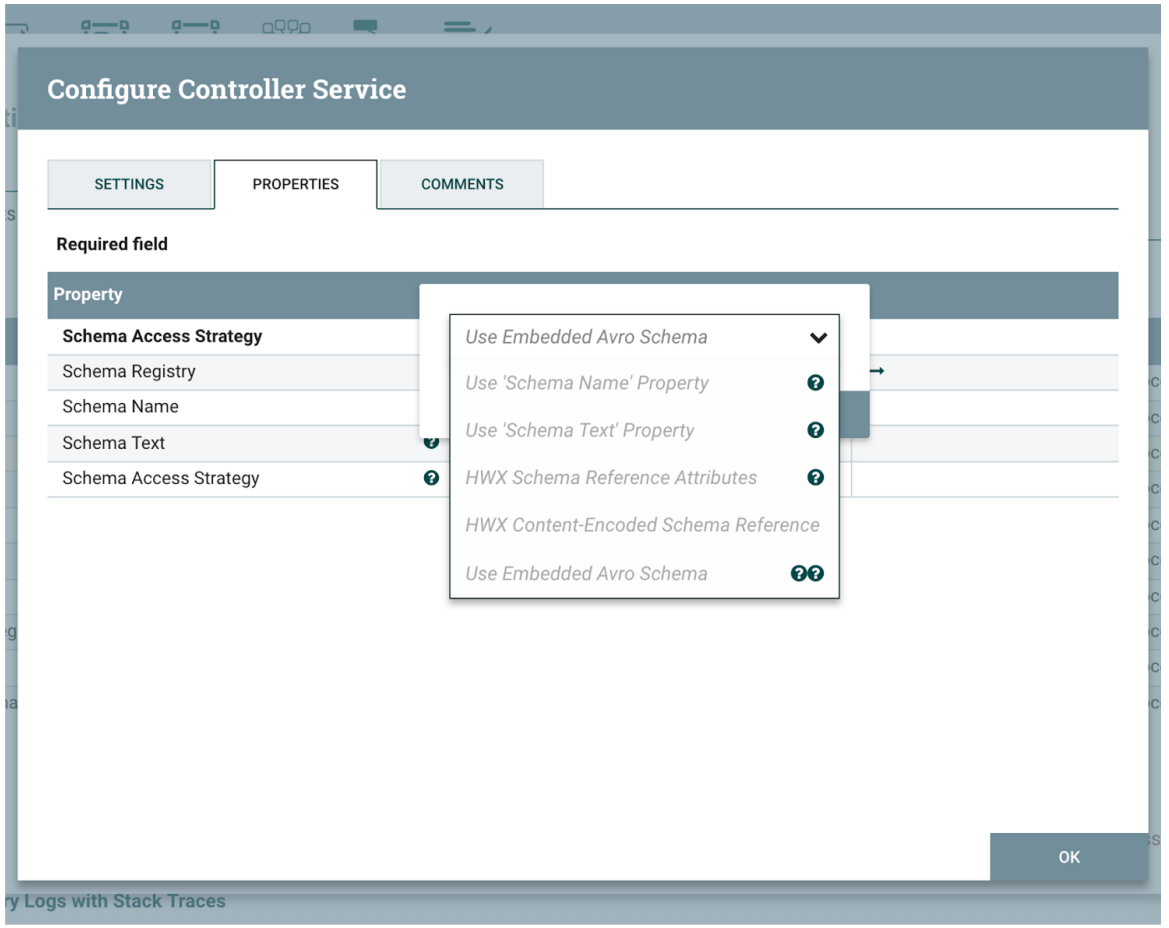

Overview
Cluster: orlandostreamcluster
Producers
83
Brokers
3
Topics
28
Consumer Groups
19
Clear
TOPICS (28)
BROKERS (3)
2 days
Producers (83)
ACTIVE (83)
PASSIVE (0)
ALL
MESSAGES
geo-critical-event-collector-i1 63k
route-apps 46k
geo-critical-event-collector-i2 32k
minifi-eu-i1 32k
load-optimizer-apps 30k
geo-critical-event-collector-i3 22k
geo-critical-event-collector-i6 21k
nifi-syndicate-speed-avro 20k
nifi-syndicate-geo-avro 20k
nifi-syndicate-geo-json 20k
nifi-syndicate-speed-json 20k
minifi-eu-i2 16k
geo-critical-event-collector-i4 16k
minifi-eu-i4 16k
fuel-apps 15k
predictive-apps 15k
minifi-truck-c1 13k
NAME
DATA IN
DATA OUT
MESSAGES IN
CONSUMER GROUPS
syndicate-transmission 55MB 294KB 0.3m 0
syndicate-speed-event-json 48MB 0B 0.2m 0
syndicate-speed-event-avro 29MB 0B 0.2m 0
syndicate-oil 1MB 0B 6.1k 0
syndicate-geo-event-json 65MB 0B 0.2m 0
syndicate-geo-event-avro 37MB 19MB 0.2m 3
syndicate-battery 60MB 0B 0.3m 0
syndicate-all-geo-critical-ev... 1GB 113MB 5.2m 1
Consumer Groups (19)
ACTIVE (17)
PASSIVE (2)
ALL
LAG
predictive-micro-service 2.5m
load-optimizer-micro-service 2m
route-micro-service 1.1m
kafka-streams-analytics-geo... 0.5m
flink-analytics-geo-event 0.5m
spark-streaming-analytics-g... 0.5m
adjudication-micro-service 0.4m
audit-micro-service 0.3m
supply-chain-micro-service 0.2m
energy-micro-service 0.1m
compliance-micro-service 0.1m
approval-micro-service 61k
nifi-truck-sensors-central 10k
nifi-truck-sensors-east 3
nifi-truck-sensors-west 2
ranger_entities_consumer 1
atlas 0

Schema Registry



Schema registry is integrated with NiFi, Kafka and SAM





THANK YOU

cloudera