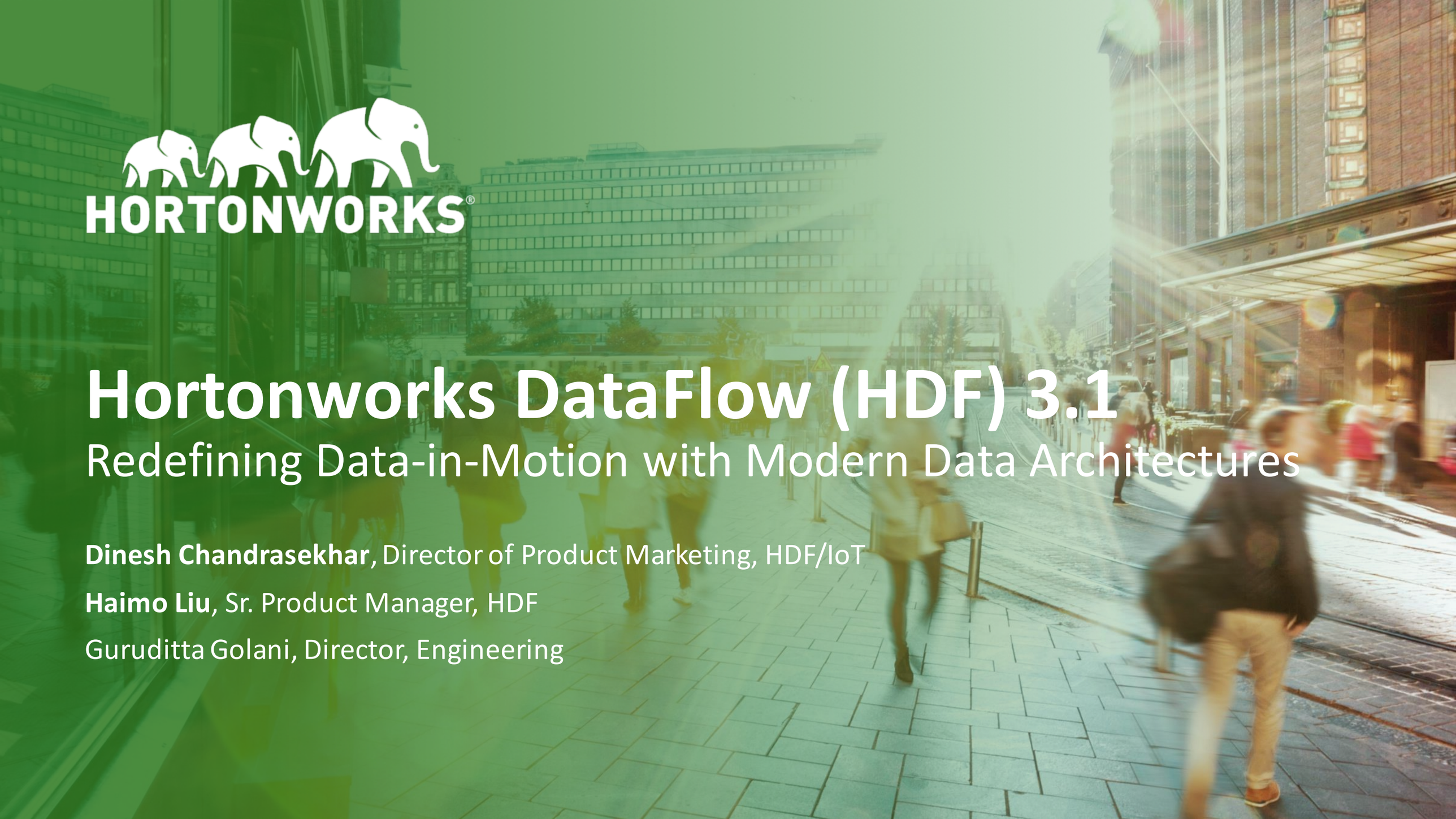


Hortonworks DataFlow (HDF) 3.1

Redefining Data-in-Motion with Modern Data Architectures

Dinesh Chandrasekhar, Director of Product Marketing, HDF/IoT

Haimo Liu, Sr. Product Manager, HDF

Guruditta Golani, Director, Engineering

The New Way of Business Is Fueled By Connected Data

DEVELOPMENT

- Connected Customers, Vehicles, Devices
- Socially crowd-sourced requirements
- Digital design and analysis
- Digital prototypes and tests (simulations)

MANUFACTURING

- Connected Factories, Sensors, Devices
- Human-robotic interaction
- 3D-printing on demand

DISTRIBUTION

- Connected Trucks, Inventory
- Location, traffic, weather-aware distribution
- Real-time inventory visibility
- Dynamic rerouting

MARKETING/SALES

- Connected Customers, Devices
- Omni- channel demand sensing
- Real-Time Recommendations

SERVICE

- Connected Assets
- Remote service monitoring & delivery
- Predictive maintenance
- OTA Updates

HORTONWORKS

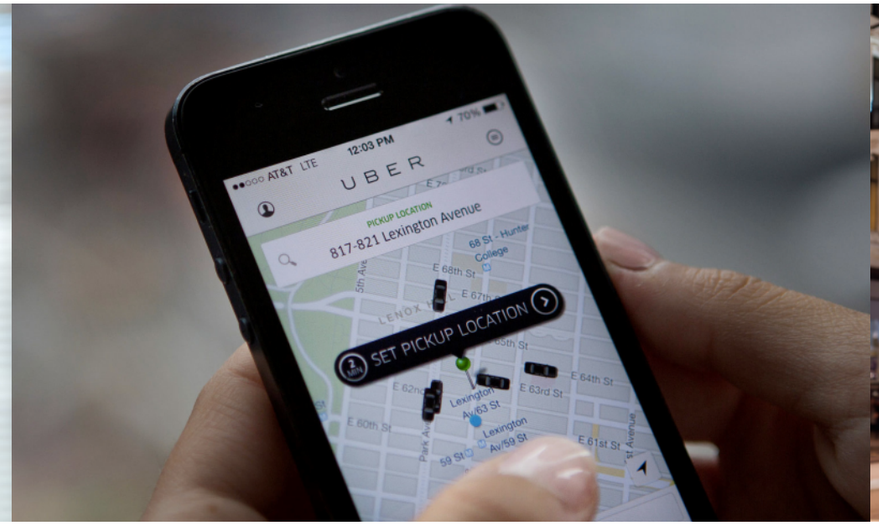
Technology Trends: Shifting the Data Paradigm

INTERNET OF THINGS	ARTIFICIAL INTELLIGENCE	CLOUD COMPUTING	STREAMING DATA
			
Industrial Internet *Connected Business* *Consumer Devices*	*Smart Devices* *Autonomy* *Prescriptive Analytics*	*SaaS/PaaS Applications* *Ephemeral Use Cases* *Operational Efficiency* *Collaboration*	*Real-time Applications* *Targeted Retail* *Recommendations* *Industrial Applications*

HORTONWORKS

…And Changing the Consumption Trends

Search
Recommendations
Reviews

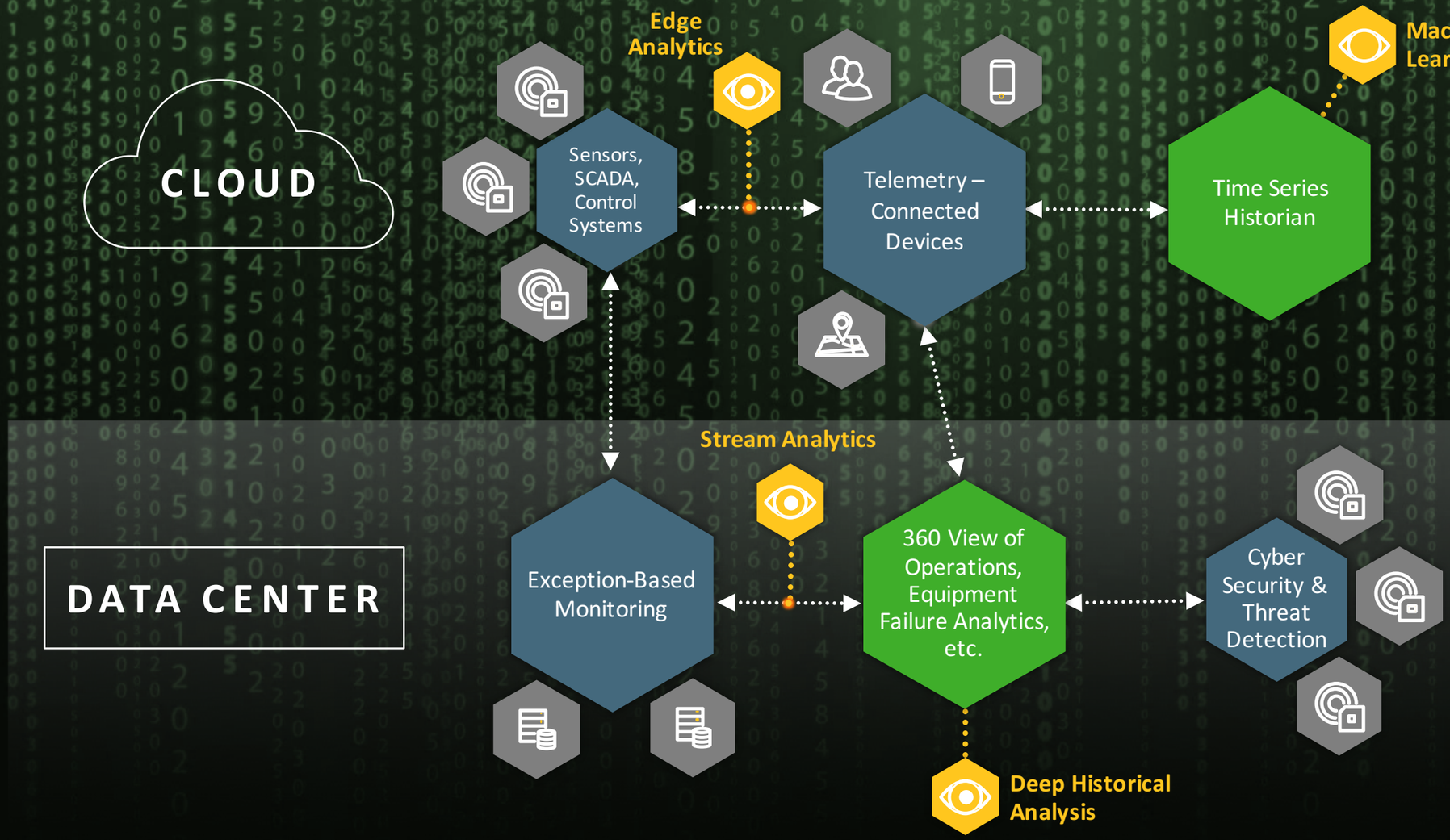


Summon my service
Frictionless commerce
Instant feedback

Immersive
Ambient
Physical meets digital

HORTONWORKS

Modern Data Architecture



HORTONWORKS

Challenges in adapting to the Modern Data Architecture

CHALLENGES

1. **Data** challenges have become more complex

sprawl | volume | speed | security | governance

2. **Systems** challenges remain

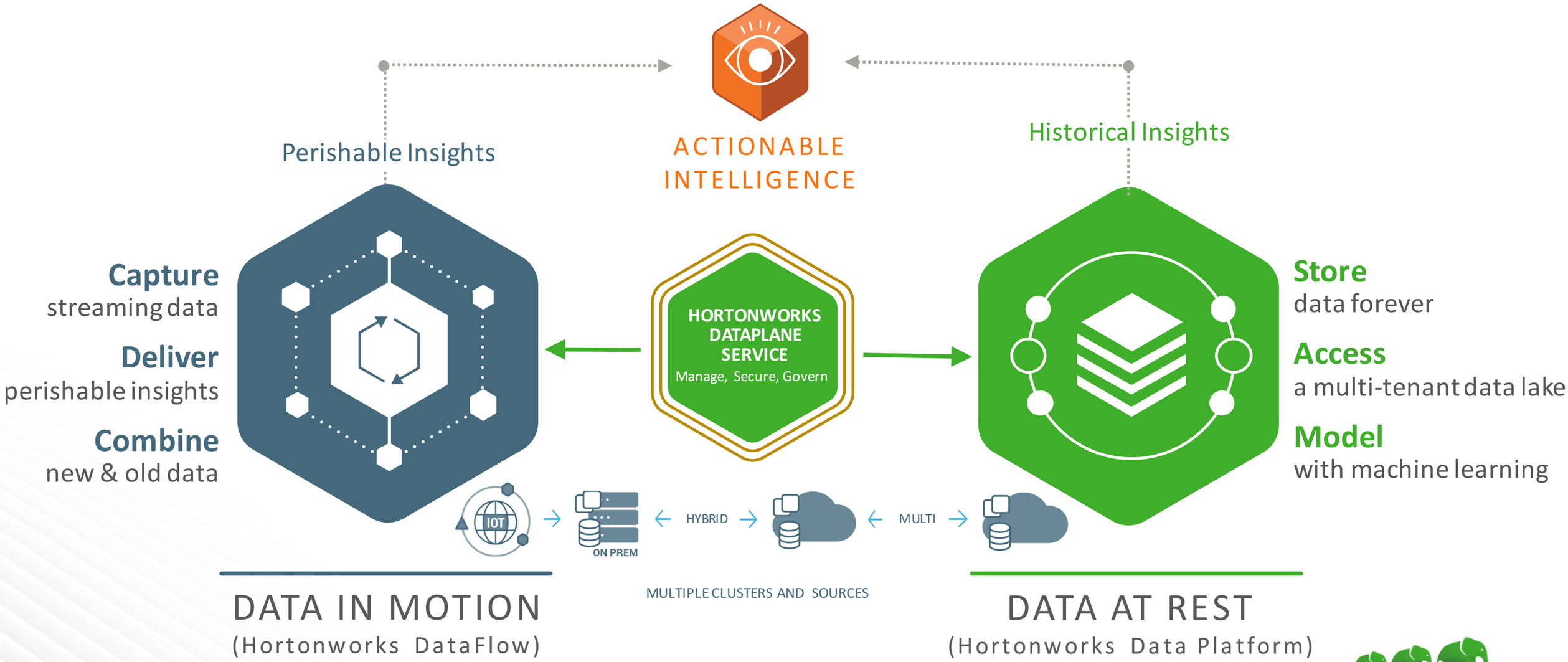
integration | management | reliability | security | costs

3. **Ecosystem** challenges grow

standards | latency | diversity | costs | SLAs

HORTONWORKS

A Connected Data Strategy Solves for All Data



HDF Data-In-Motion Platform

Flow Management

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

nifi

minifi C++ Agent Java Agent

Stream Processing

Scalable data broker for streaming apps
Scale out streaming computation engine

kafka

STORM

Stream Analytics

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

STREAMING ANALYTICS MANAGER

Enterprise Services

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

SCHEMA REGISTRY

APACHE NIFI registry

Ambari

Apache Ranger

Flow Management with Apache NiFi

HDF - Flow Management powered by Apache NiFi

- Ingestion: connectors to read/write data from/to several data sources
- Transformation:
  - Format conversion
  - Compression/decompression, Merge, Split, encryption, etc
- Data enrichment
  - Attribute, content, rules, etc
- Routing
  - Priority, dynamic/static, based on content or metadata, etc
- Parsing

220+ Processors for Deeper Ecosystem Integration

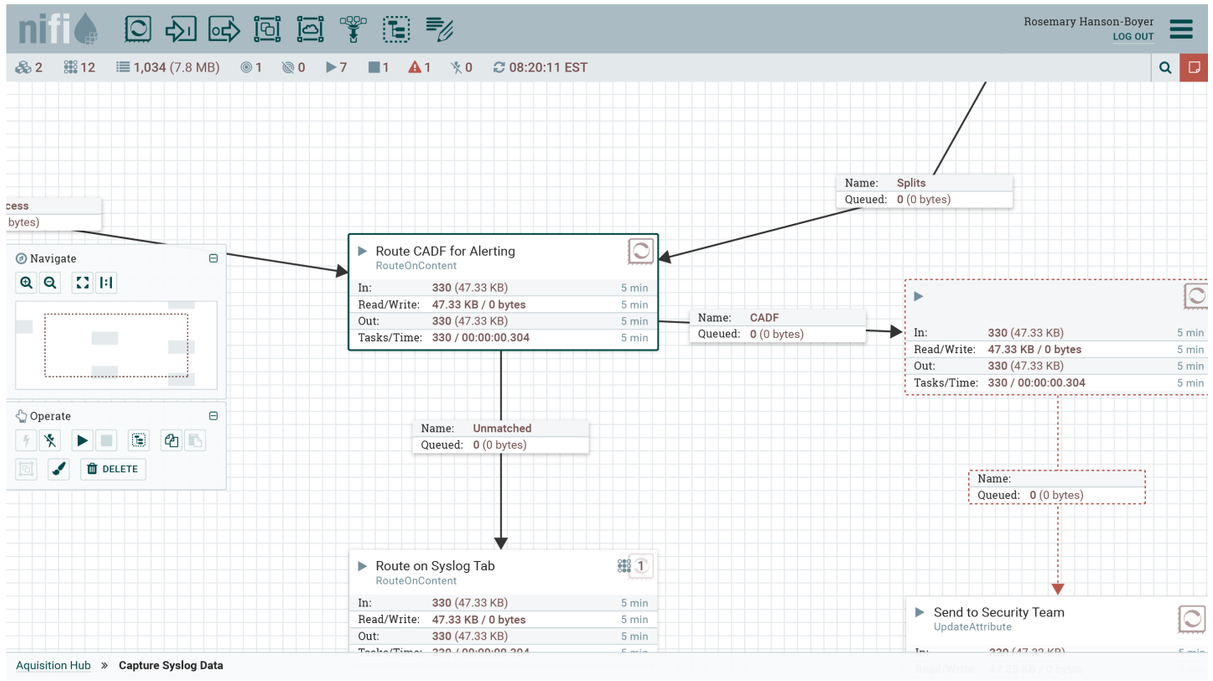
- FTP
- SFTP
- HL7
- UDP
- XML
- HTTP
- WebSocket
- Email
- HTML
- Image
- Syslog
- AMQP

- Hash
- Encrypt
- GeoEnrich
- Merge
- Tail
- Scan
- Extract
- Evaluate
- Replace
- Duplicate
- Execute
- Translate
- Split
- Fetch
- Convert
- Route Text
- Distribute Load
- Route Content
- Generate Table Fetch
- Route Context
- Jolt Transform JSON
- Control Rate
- Prioritized Delivery

HORTONWORKS

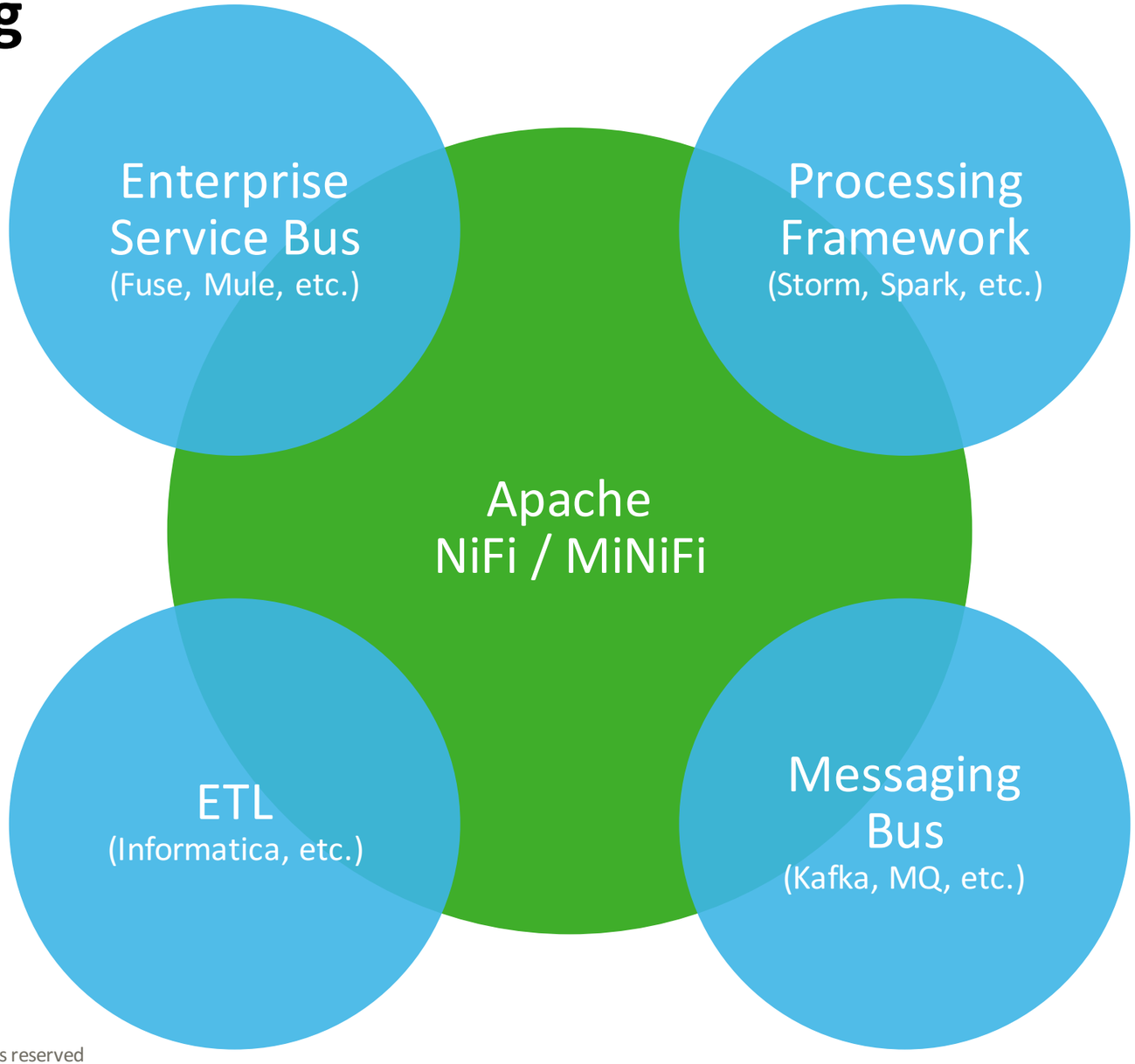
Apache NiFi High Level Capabilities

- Web-based user interface
  - Design, control, feedback & monitoring
- Highly configurable
  - Loss tolerant vs guaranteed delivery
  - Low latency vs high throughput
  - Dynamic prioritization
  - Flow can be modified at runtime
  - Back pressure
- Data provenance
  - Track dataflow from beginning to end
- Designed for extension
  - Build your own processors
- Secure
  - SSL, SSH, HTTPS, etc.



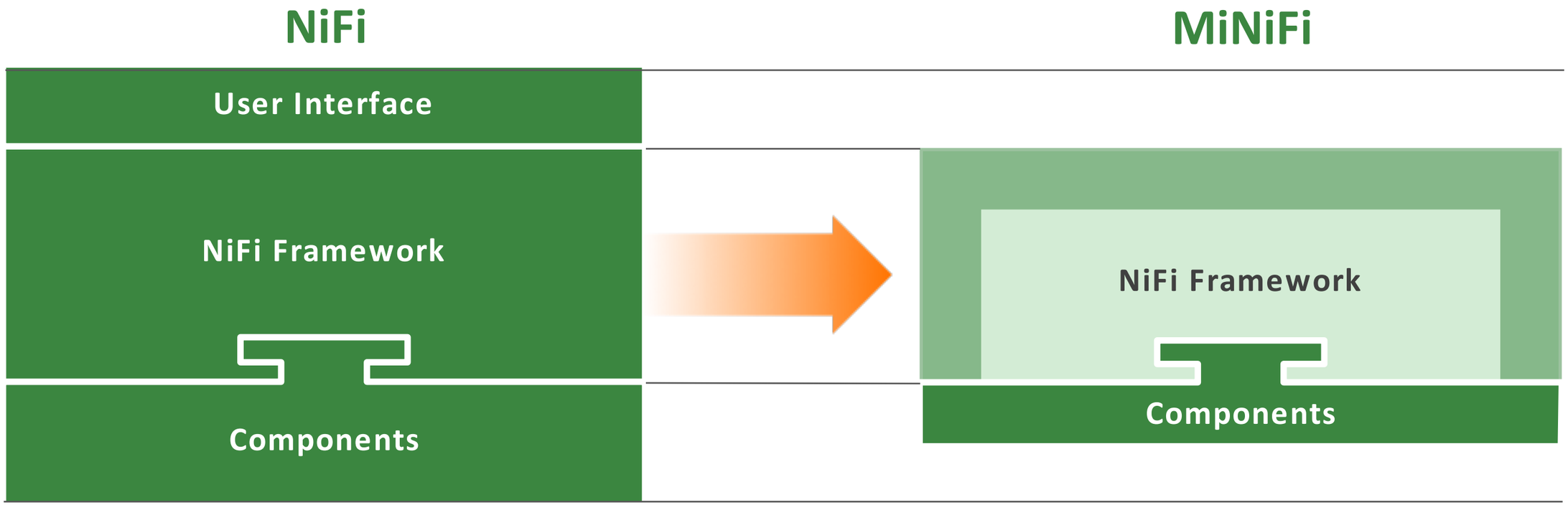
HORTONWORKS

NiFi Positioning



HORTONWORKS

NiFi vs. MiNiFi - Smaller Footprint ~40 MB



HORTONWORKS

Edge Intelligence with Apache MiNiFi

Key Features

- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance

- Recovery / recording a rolling log of fine-grained history
- Designed for extension

Different from Apache NiFi

- **Design and Deploy**
- **Warm re-deploys**

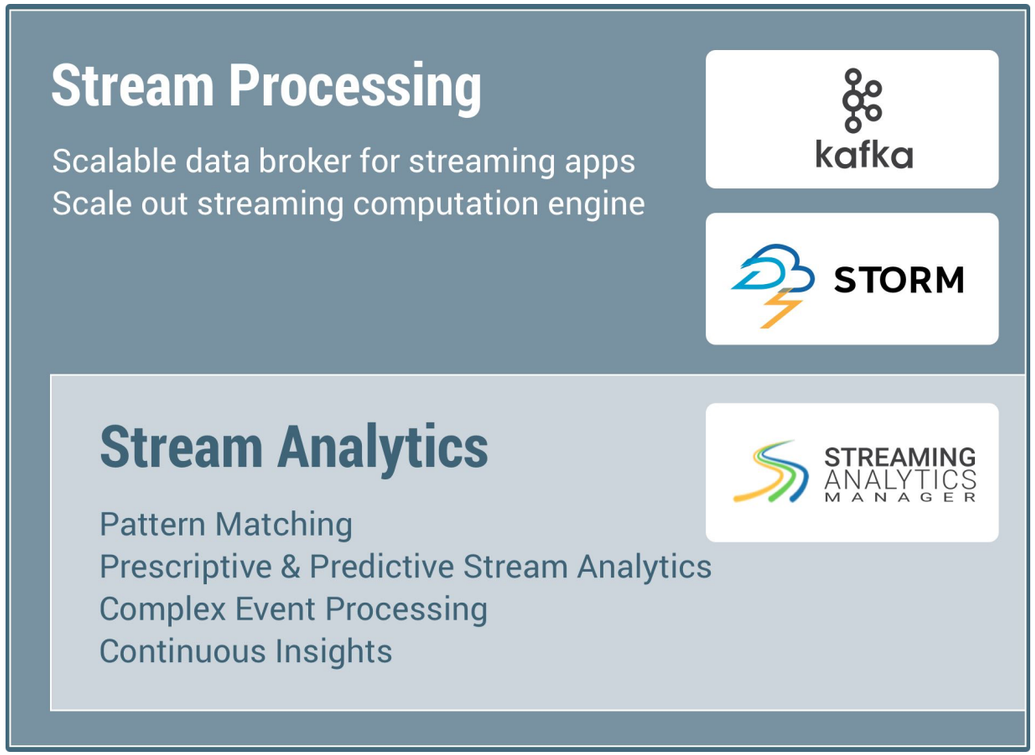
HORTONWORKS

Stream Processing

HORTONWORKS

HDF Stream Processing – Streaming Analytics Manager (SAM)

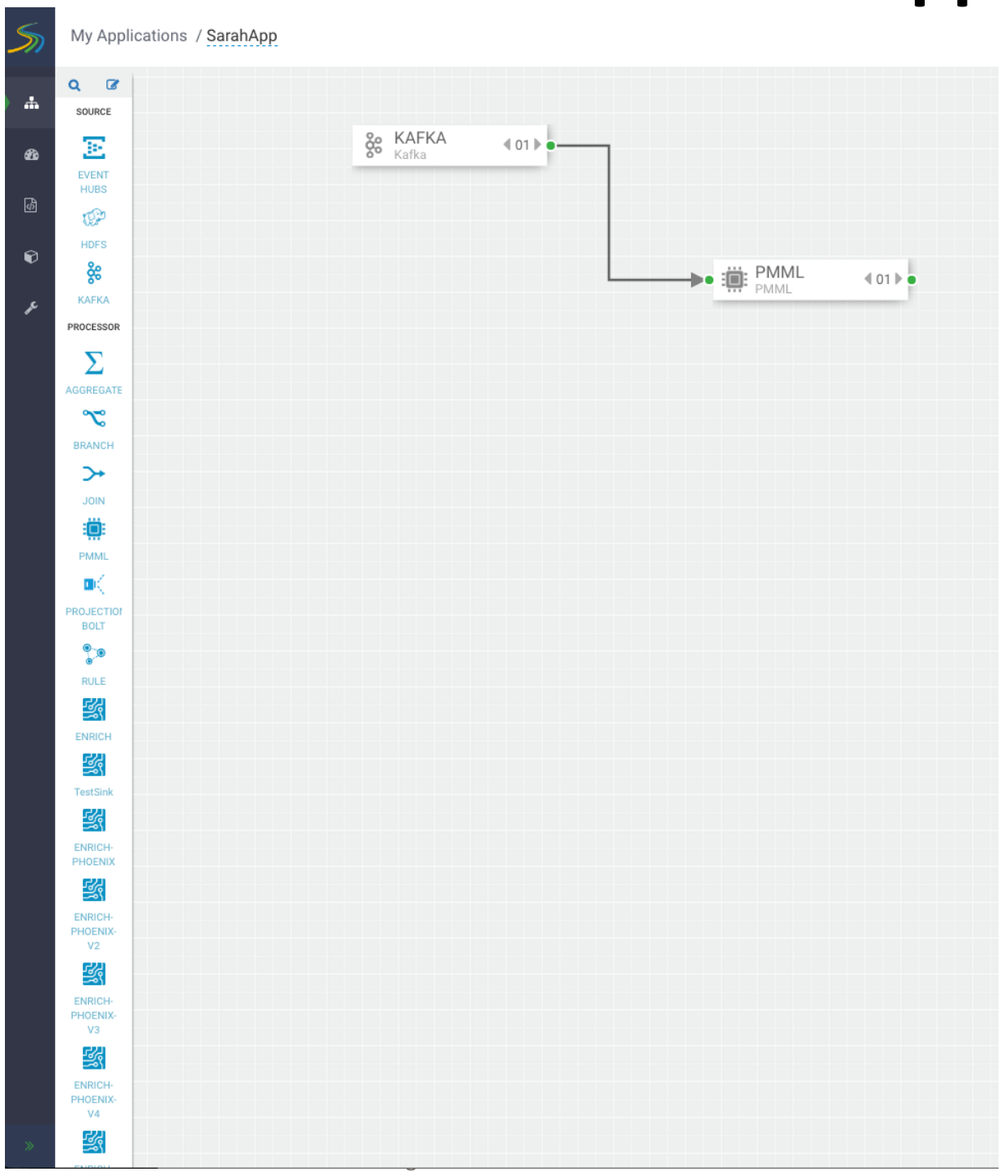
- A brand product module in the HDF stack to design, develop, deploy and manage streaming analytics app with drag-and-drop ease
  - Build streaming analytics applications that do event correlation, context enrichment , complex pattern matching, analytical aggregations and creation of alerts/notifications when insights are discovered.
  - Supports multiple streaming substrates (e.g: Storm, Spark Streaming, Flink)
  - Extensibility is a first class citizen (add custom sinks, processors, spouts, etc..)



Who Uses SAM?

HORTONWORKS

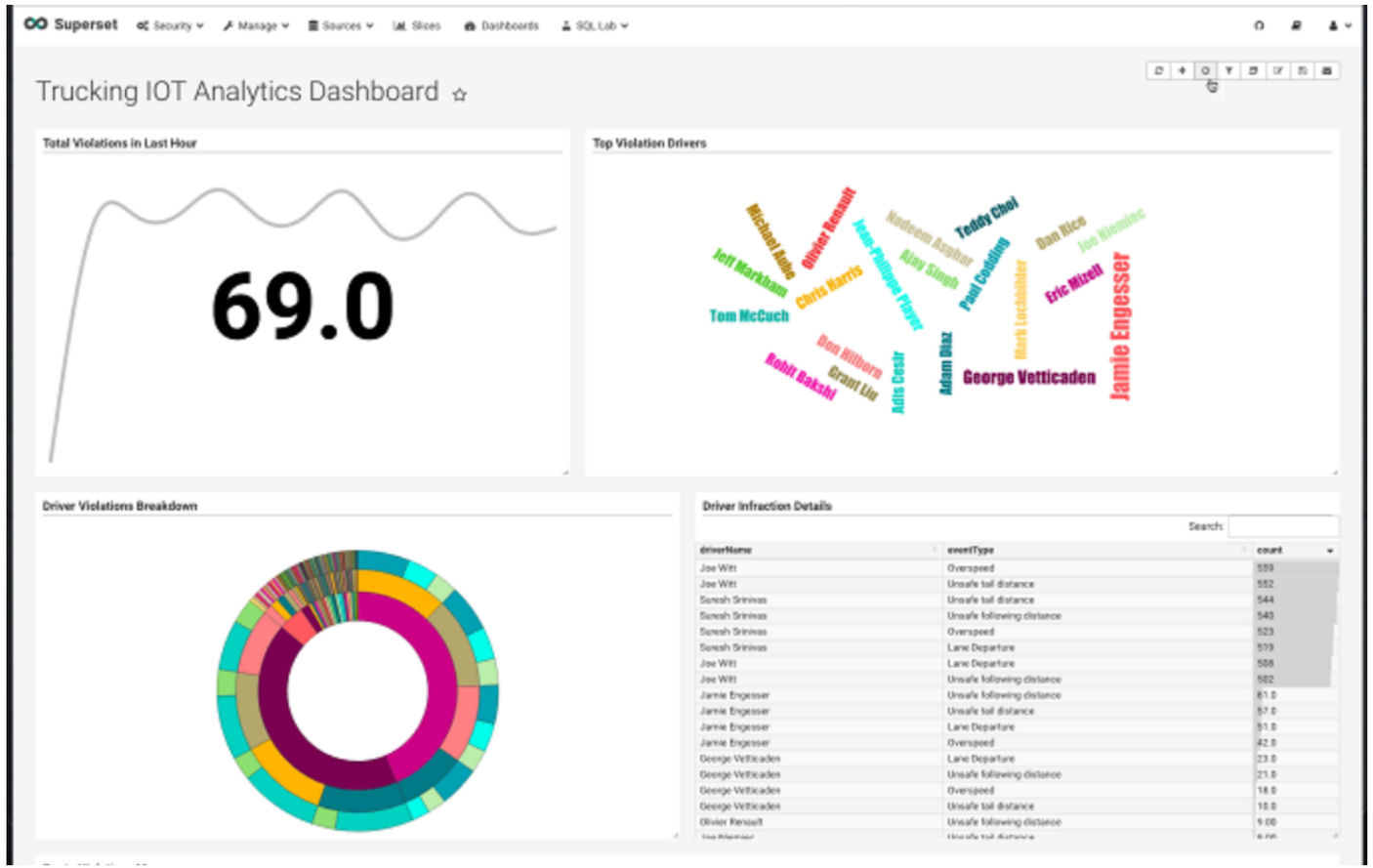
Stream Builder Module for App Developers



- Builder components, shown on the canvas palette, are the building blocks used by the app developer to build streaming apps.
- Drag and drop to build a working streaming application without writing a single line of code.
- 4 Types of Components: Sources, Processors, Sinks and Custom

HORTONWORKS

Stream Insight Module for Business Analysts



- A tool to create time-series and real-time analytics dashboards, charts and graphs
- 30+ visualization charts out of the box with customization capability
- Druid is the Analytics Engine that powers the Stream Insight Module.

HORTONWORKS

4 Building blocks for robust and scalable real time solutions

Publish subscribe

high-throughput, low-latency broker

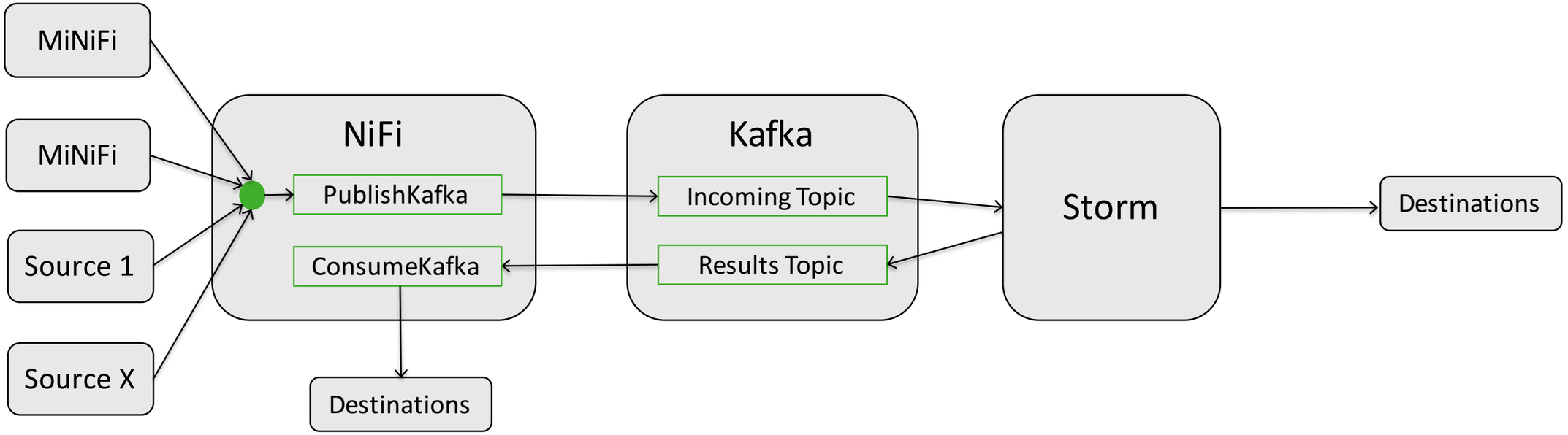
reliable real-time data processing engine able to process a million tuples per second per node

A tool used to design, develop, deploy and manage streaming analytics applications using a drag drop paradigm

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

HORTONWORKS

Better Together



- MiNiFi – Collection, filtering, and prioritization at the edge
- NiFi - Central data flow management, routing, enriching, and transformation
- Kafka - Central messaging bus for subscription by downstream consumers
- Storm - Streaming analytics focused on complex event processing


Enterprise Services

HDF Enterprise Services

- Schema Registry
- Apache NiFi Registry
- Ambari
- Apache Ranger
- Apache Knox
- Apache Atlas
- SmartSense

HORTONWORKS

Use cases

HDF Use Cases

Data Movement

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

Optimize Log Collection & Analysis

Optimize log analytics solutions such as Splunk by using HDF as a single platform to collect and deliver multiple data sources and using HDP for lower cost storage options

Feed Data to Streaming Analytics:

Accelerate big data ROI by streaming data into analytics systems such as Apache Storm or Apache Spark Streaming

Ingest Logs for Cyber Security:

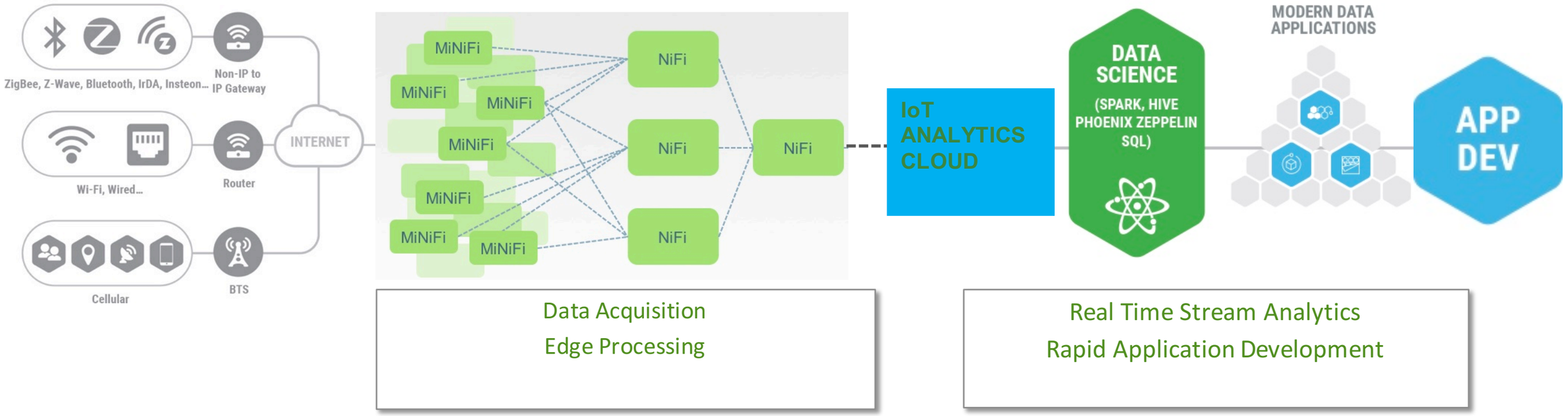
Integrated and secure log collection for real-time data analytics and threat detection

Capture IoT Data

Transport disparate and often remote IoT data in real time, despite any limitations in device footprint, power or connectivity—avoiding data loss

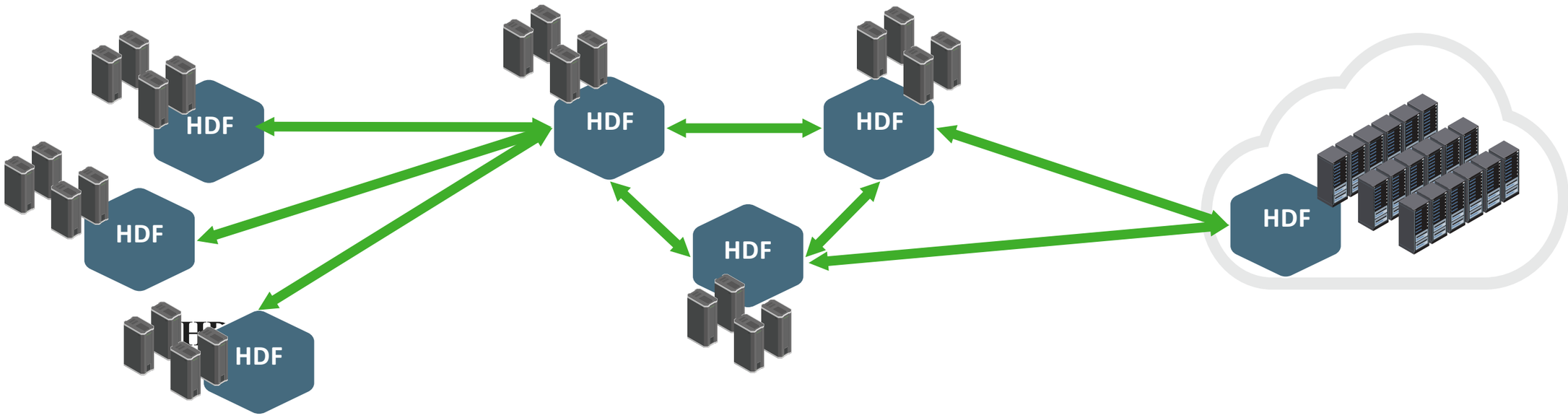
HORTONWORKS

Data Ingestion/Stream Processing

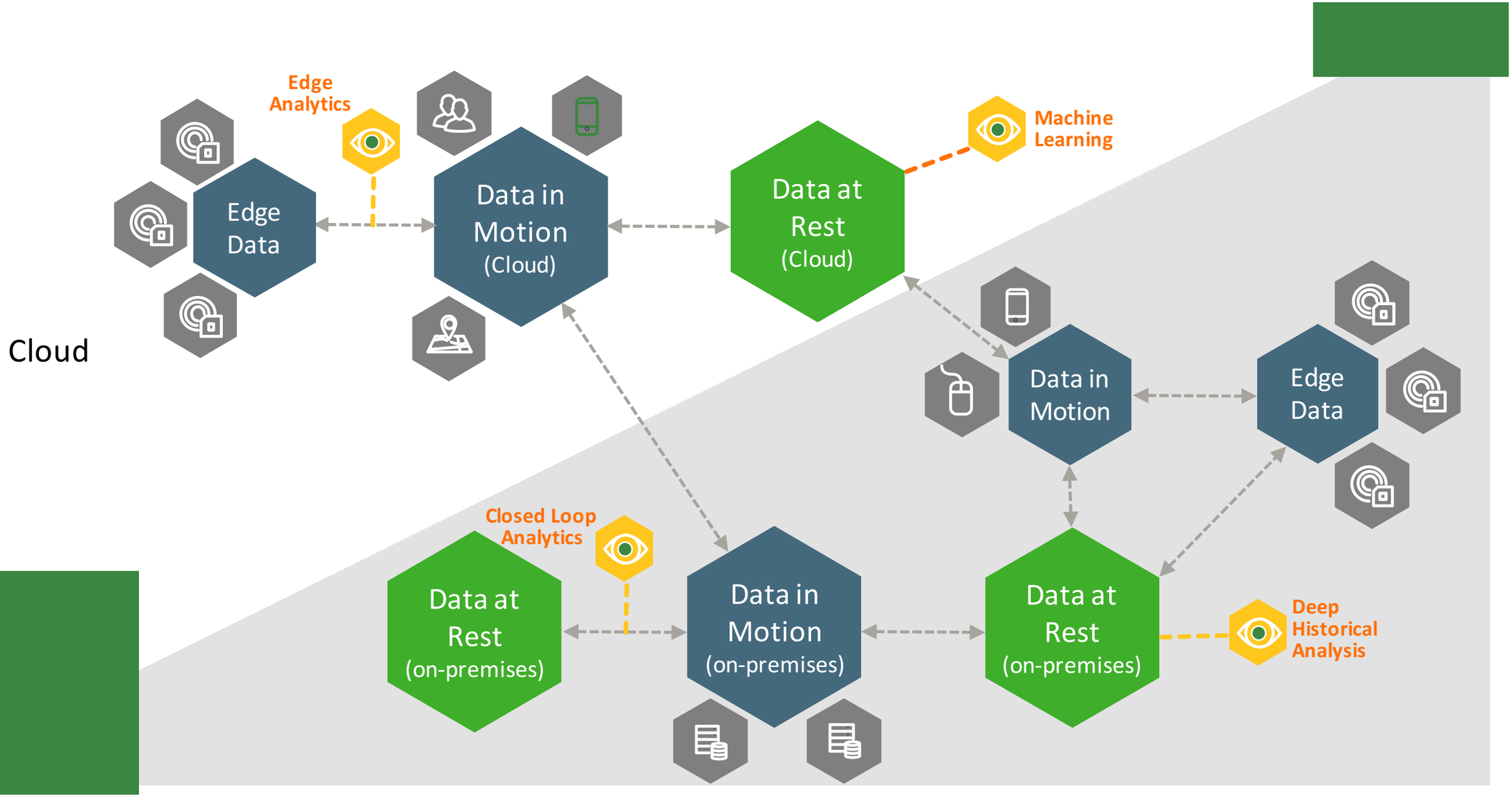


Enterprise Data Movement

- Seamlessly fuse dataflows between data centers
  - Data center to data center,
  - Remote location to data center,
  - Data center to cloud



Architectural transformation



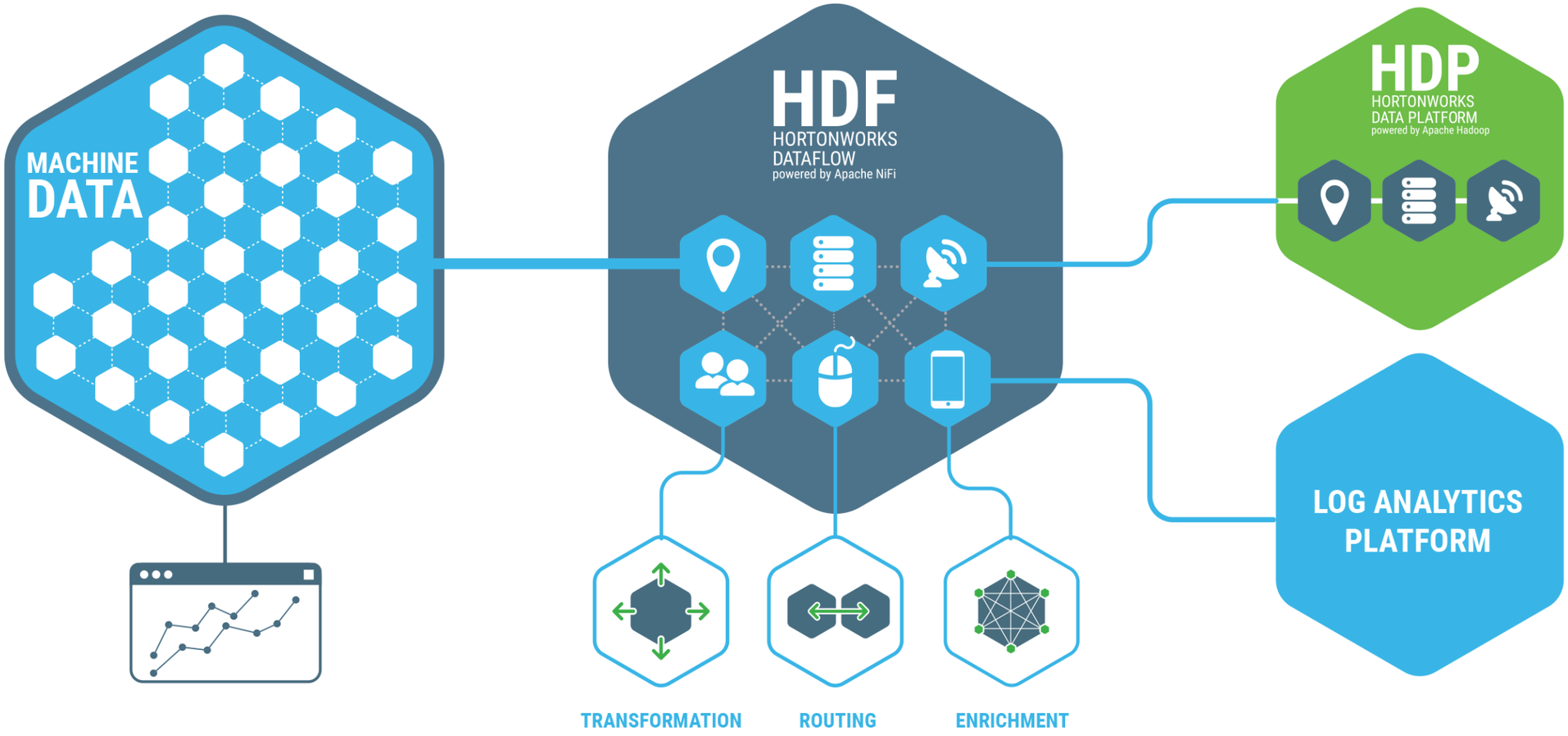
HORTONWORKS

Data Ingestion: Optimize Log Analytics with Content Based Routing

Edge analytics for cost-effective and efficient movement of machine data

Intelligent, content based routing, transformation and enrichment

Send data to alternative systems based on value, content, priority



HORTONWORKS

What’s new in HDF 3.1?

HDF 3.1 New and Enhanced Features

	Core Enhancements	Cross-Product Integration	Ease of Use
Flow Management	• Apache NiFi Registry (new) • Flow migration and version control • MiNiFi C++, Java, Andriod/IOS libraries GA • Containerized deployment (Docker)	• NiFi-Atlas, -SmartSense, and -Knox integration (HDF on HDP scenario only) • NiFi-Ranger: Group based policy support for NiFi resources	• Improved Ambari experience: Automate adding NiFi nodes to existing cluster
Stream Processing	• Kafka 1.0 Support • Schema Registry • Schema Version Lifecycle Mgmt. • SAM extensibility improvements	• Ambari and Ranger support for Kafka 1.0	• New SAM operations module • SAM ”Test Mode”

HORTONWORKS

HDF 3.1 Data-In-Motion Platform

Flow Management

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

nifi
NiFi 1.5

minifi
C++ Agent
NEW
Java Agent

Stream Processing

Scalable data broker for streaming apps
Scale out streaming computation engine

kafka
Kafka 1.0

STORM

Stream Analytics

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

STREAMING ANALYTICS MANAGER

Enterprise Services

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

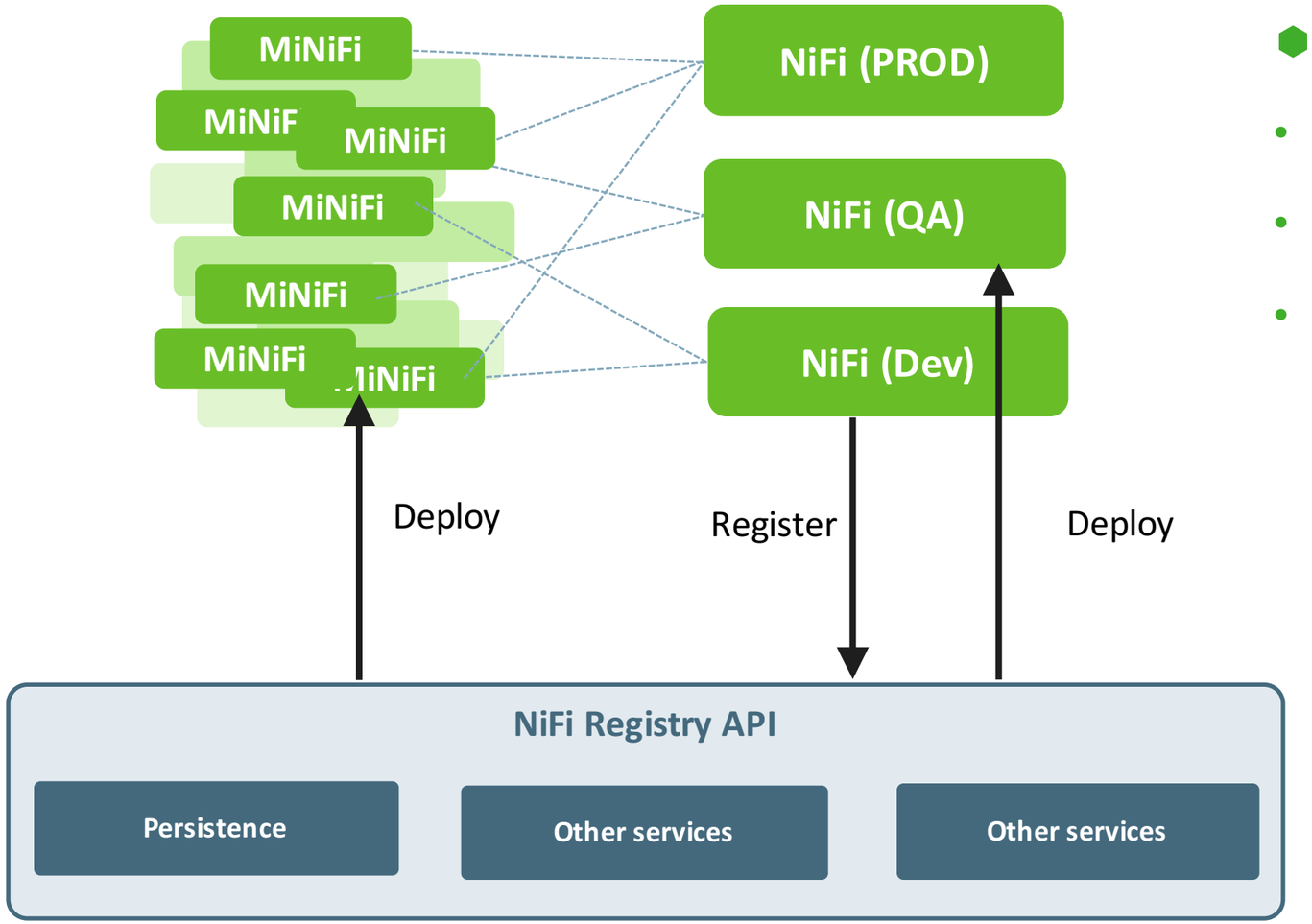
SCHEMA REGISTRY

APACHE NIFI registry
NEW

Ambari

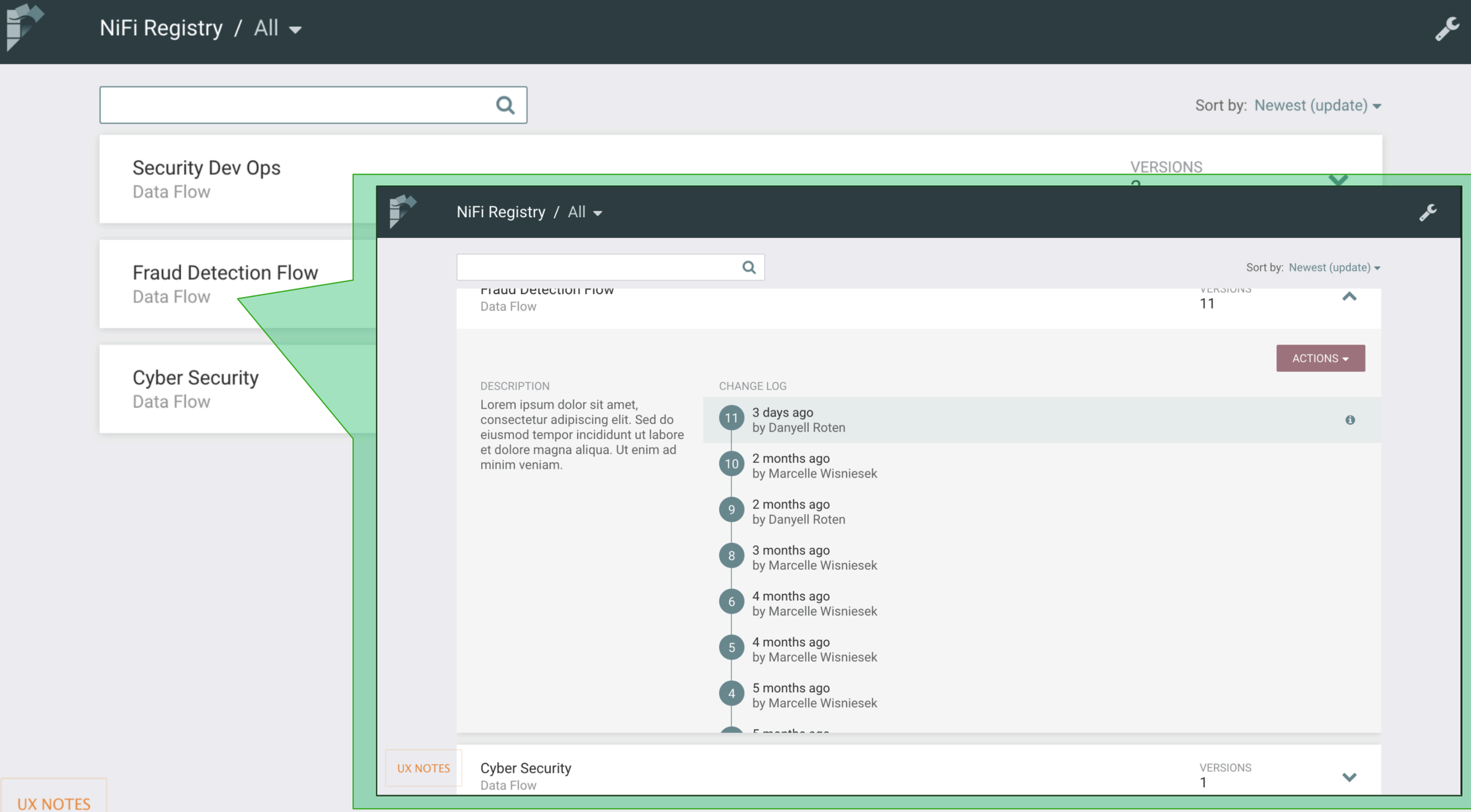
Apache Ranger

Increased Developer Productivity
Apache NiFi Registry



- **NiFi Registry**
- Repository of versioned flows
- Portability
- Design and deploy mechanism

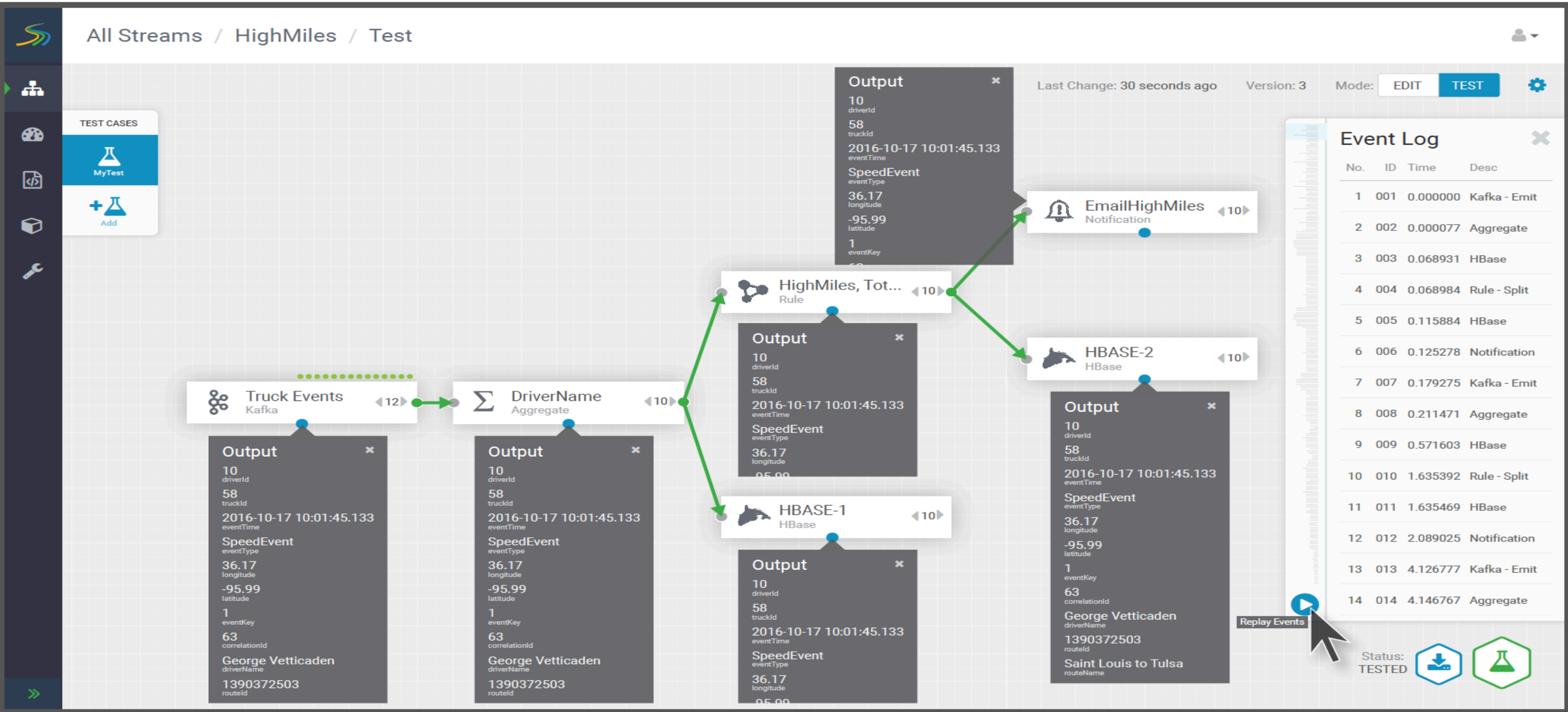
HORTONWORKS


NiFi Registry / All
Sort by: Newest (update)
Security Dev Ops
Data Flow
VERSIONS
Fraud Detection Flow
Data Flow
Cyber Security
Data Flow
NiFi Registry / All
Sort by: Newest (update)
Data Flow
VERSIONS
11
ACTIONS
DESCRIPTION
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam.
CHANGE LOG
11 3 days ago by Danyell Roten
10 2 months ago by Marcelle Wisniesek
9 2 months ago by Danyell Roten
8 3 months ago by Marcelle Wisniesek
6 4 months ago by Marcelle Wisniesek
5 4 months ago by Marcelle Wisniesek
4 5 months ago by Marcelle Wisniesek
UX NOTES
Cyber Security
Data Flow
VERSIONS
1
UX NOTES

Improved Operational Efficiency
SAM Improvements

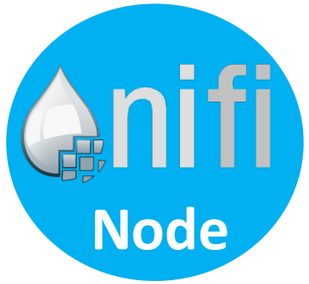
HORTONWORKS

Improved Operational Efficiency
SAM “Test Mode”

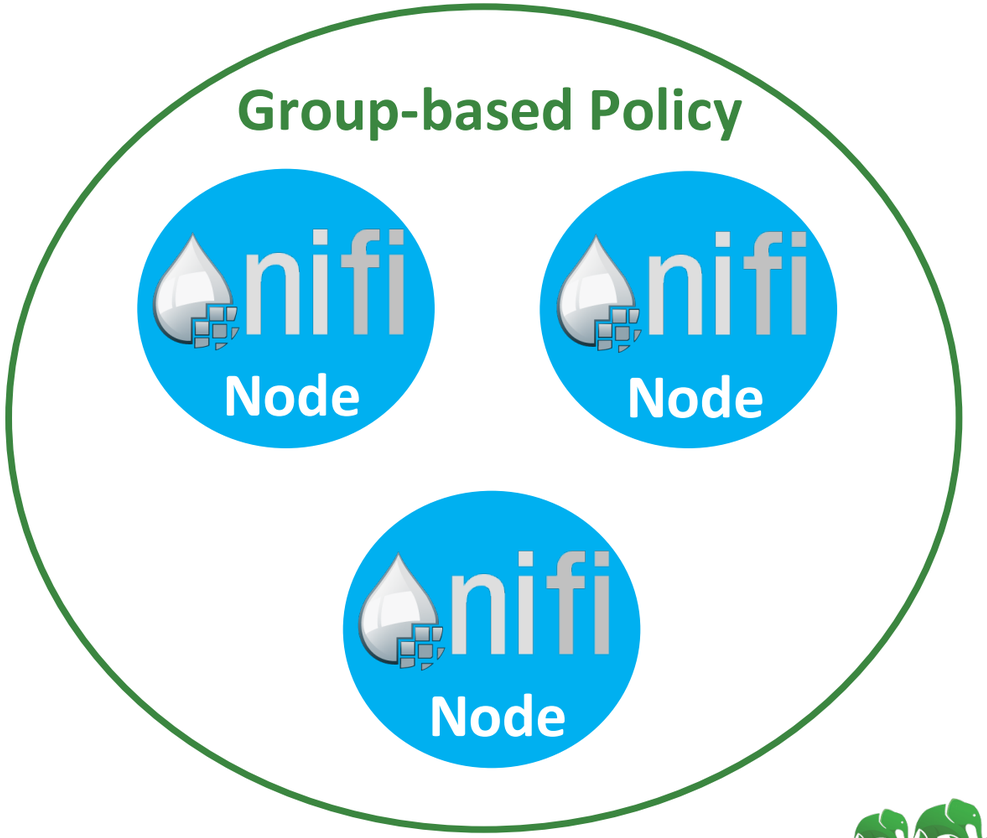


Improved Operational Efficiency
Ambari / Ranger Improvements

Ambari

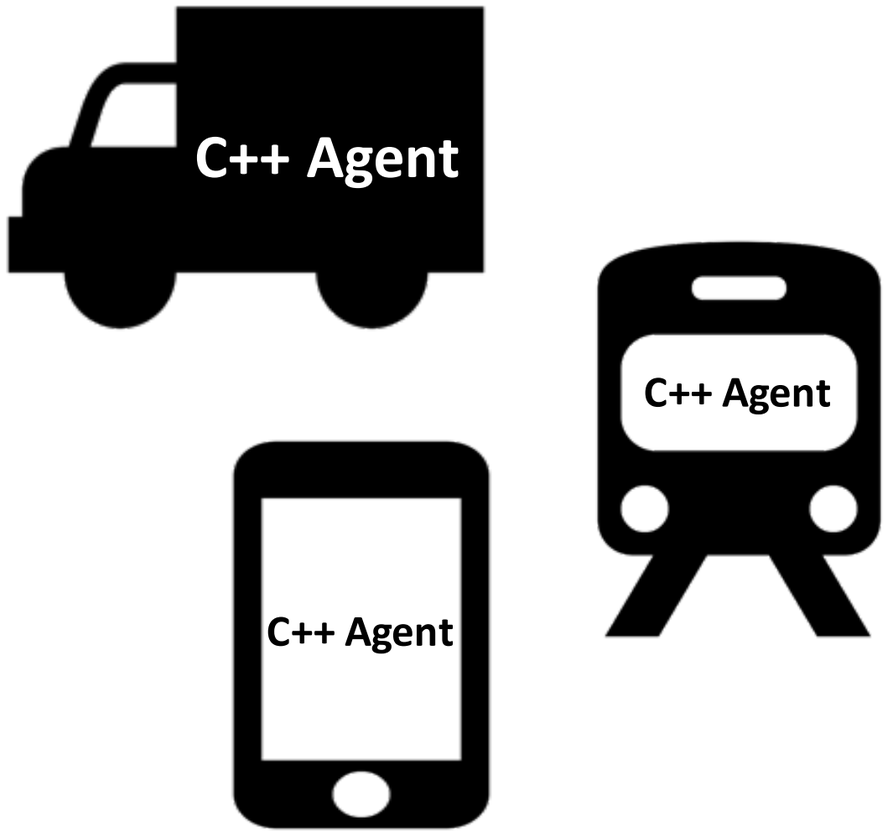


Apache Ranger



HORTONWORKS

Improved Operational Efficiency
MiNiFi C++ Agent



There are many configuration options for MiNiFi C++, all dependent on the use case, they may help with:

- Minimizing memory footprint
- Lowering CPU consumption
- Reducing size on disk

HORTONWORKS

Integrated Provisioning and Security
Kafka 1.0 Support

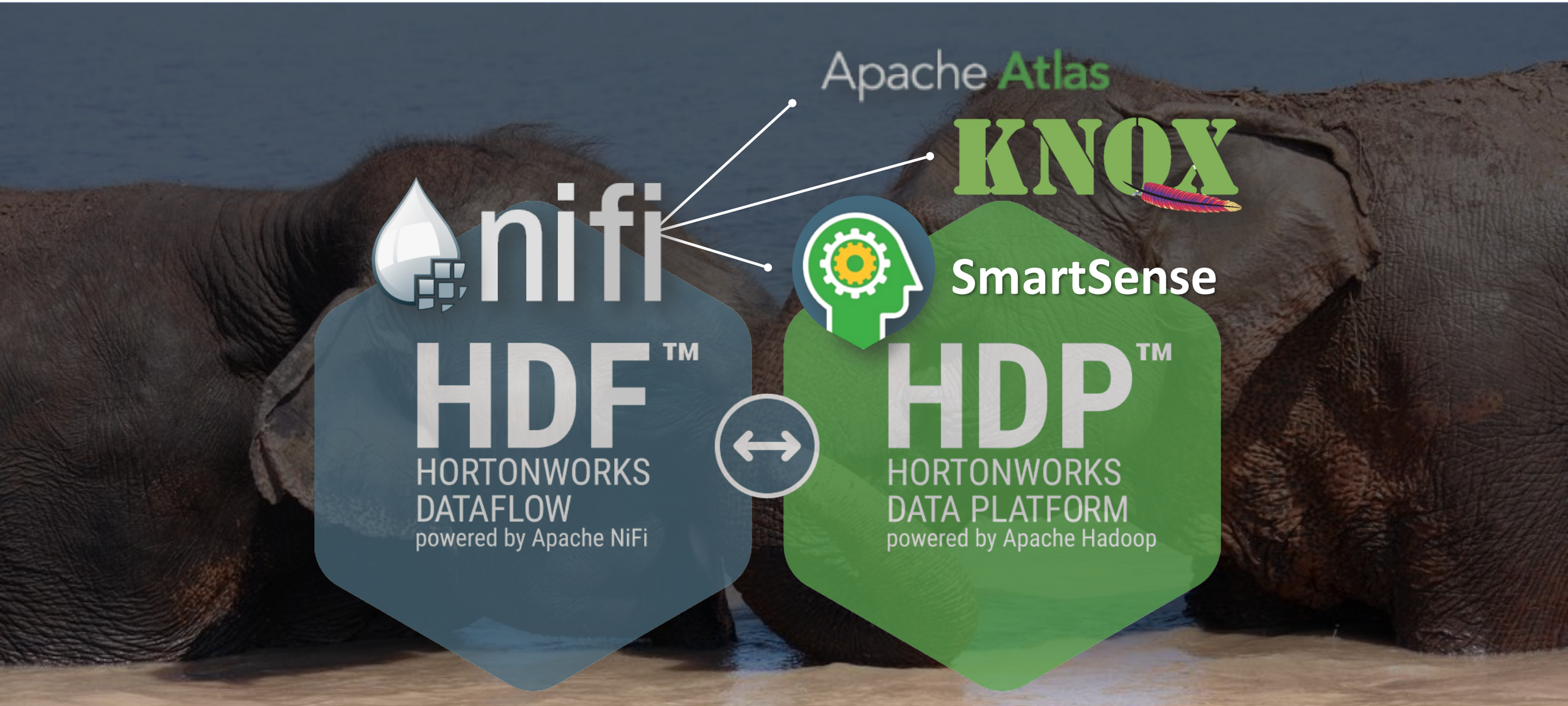
Users can now install, configure, manage, upgrade, monitor, and secure Kafka 1.0 clusters with **Ambari**.

To enhance data governance and lineage, users can now manage access control policies using resource or tag-based security in **Ranger** for Kafka 1.0 clusters.

New processors in **NiFi and Streaming Analytics Manager** support Kafka 1.0 features including message headers and transactions.

HORTONWORKS

When HDF is co-located with HDP…
Integrations with Atlas, Knox and SmartSense



I'm excited. What can I do next?

- 7-part Blog series by PMs with videos

Part 1: Announcing GA of HDF 3.1
Part 2: Introducing NiFi Registry in HDF 3.1
Part 3: Apache Kafka is now supported in HDF 3.1
Part 4: Unit Testing and Continuous Integration/Delivery of Streaming Analytics Apps using SAM's New Test Mode and SAM REST
Part 5: NiFi and Atlas Integration
Part 6: SAM's Stream Operations: Managing, Monitoring, and Debugging Streaming Apps Made Easier!
Part 7: Powerful New Extensibility in SAM: Building, Registering and using custom Kinesis Sources and S3 Sinks in SAM Streaming Analytics App

- Attend next two deep-dive webinars
  http://hortonworks.com/webinars

HDF 3.1 SERIES – PART 2: A TECHNICAL DEEP-DIVE ON NEW STREAMING FEATURES	March 1st 2018
HDF 3.1 SERIES – PART 3: TECHNICAL DEEP-DIVE ON NEW FLOW MANAGEMENT FEATURES	March 8th 2018

- Download the HDF 3.1 Sandbox today
  https://hortonworks.com/downloads/#dataflow

HORTONWORKS

Thank you

HORTONWORKS®