HDF™
HORTONWORKS DATAFLOW
powered by Apache NiFi

Hortonworks DataFlow Workshop
October 13th,2017

Microsoft Azure

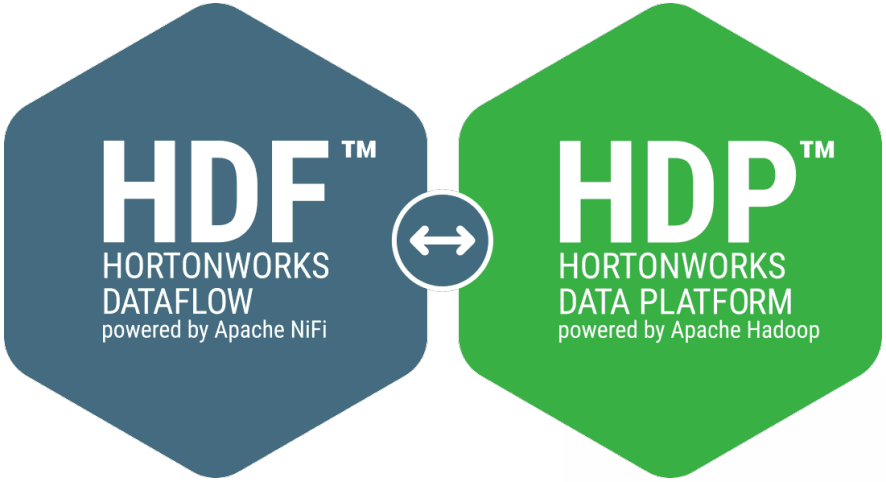
Microsoft Canada
2000 McGill College Avenue Suite 550
Montreal, QC H3A 3H3

© Hortonworks Inc. 2011 – 2015. All Rights Reserved

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Agenda

- 8:30 - 9:00: Registration, Breakfast & Networking
- 9:00 - 9:30: Welcome, Kick-off & Introduction
- 9:30 - 10:30: HDF and NiFi Overview
- 10:30 - 10:45: Break
- 10:45 - 12:30: NiFi / Dataflow Lab 1 & Lab 2
- 12:30 - 1:15: Lunch
- 1:15 - 2:15: HDF 3 Components Review, MiniFi
- 2:15 - 2:30: Break
- 2:30 – 2:45: MiNiFi – Lecture
- 2:45 - 3:20: Lab 3
- 3:20 - 3:30: Wrap-up



HORTONWORKS®
POWERING THE FUTURE OF DATA™

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

About Hortonworks

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Hortonworks Connected Data Platforms and Solutions

Hortonworks Solutions

Enterprise Data Warehouse Optimization

Cyber Security and Threat Management

Internet of Things and Streaming Analytics

Hortonworks Connection

Subscription Services

SmartSense™

Premier Operational Support

Educational Services

Professional Services

Community Connection

Data Center

Hortonworks Data Suite

HDP HDF

Cloud

Hortonworks Data Cloud

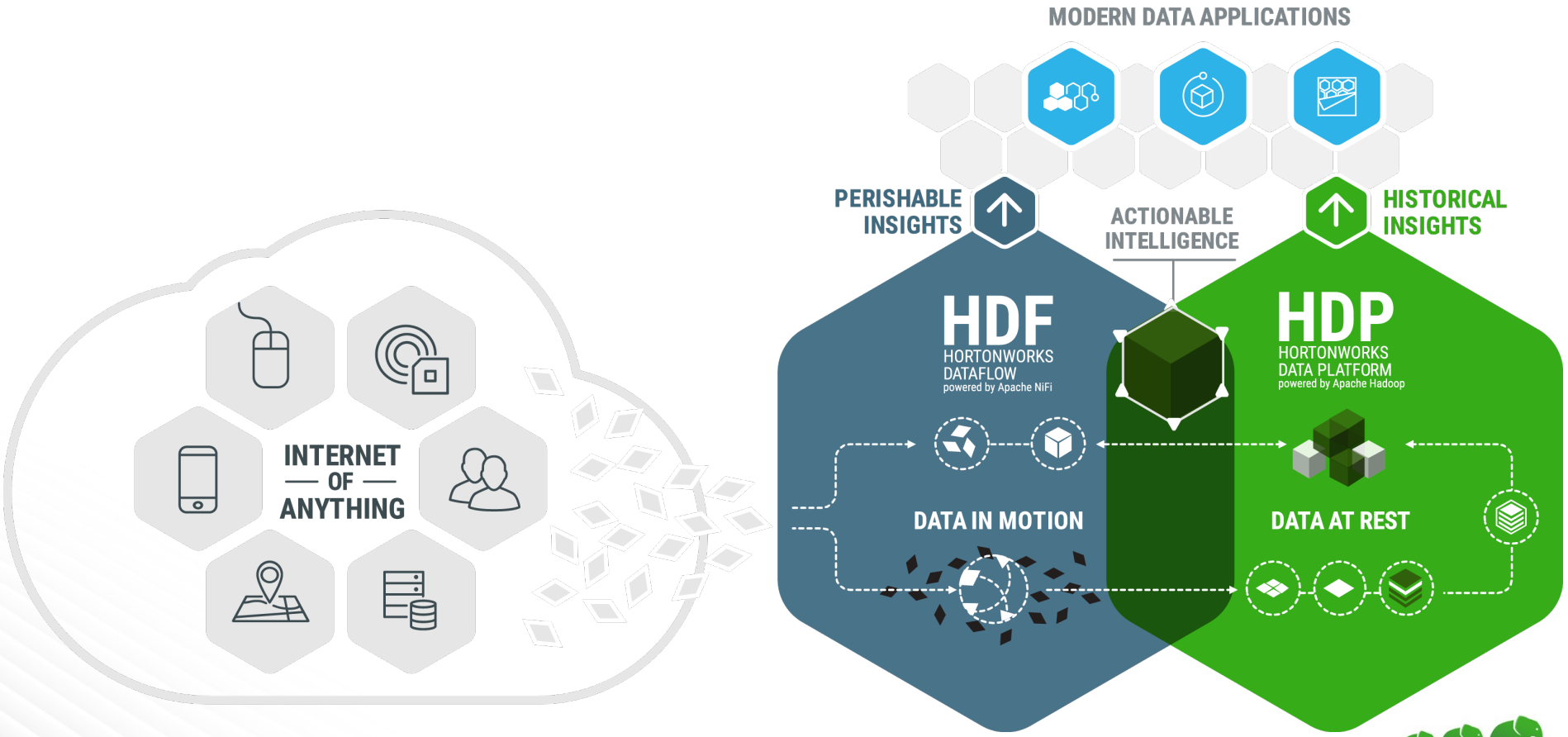
AWS HDInsight

Data Services

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

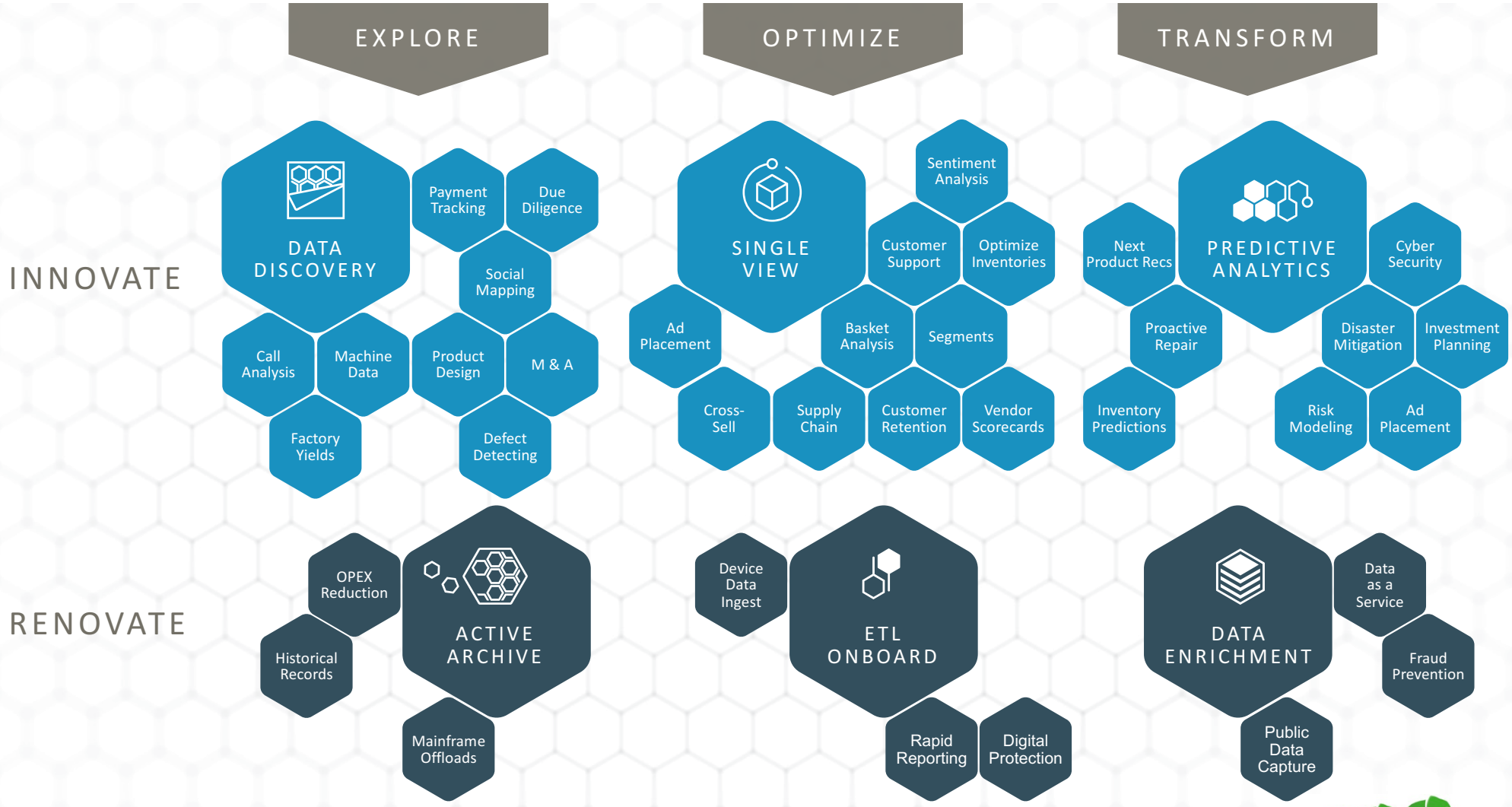
HORTONWORKS

The Connected Data Platform



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

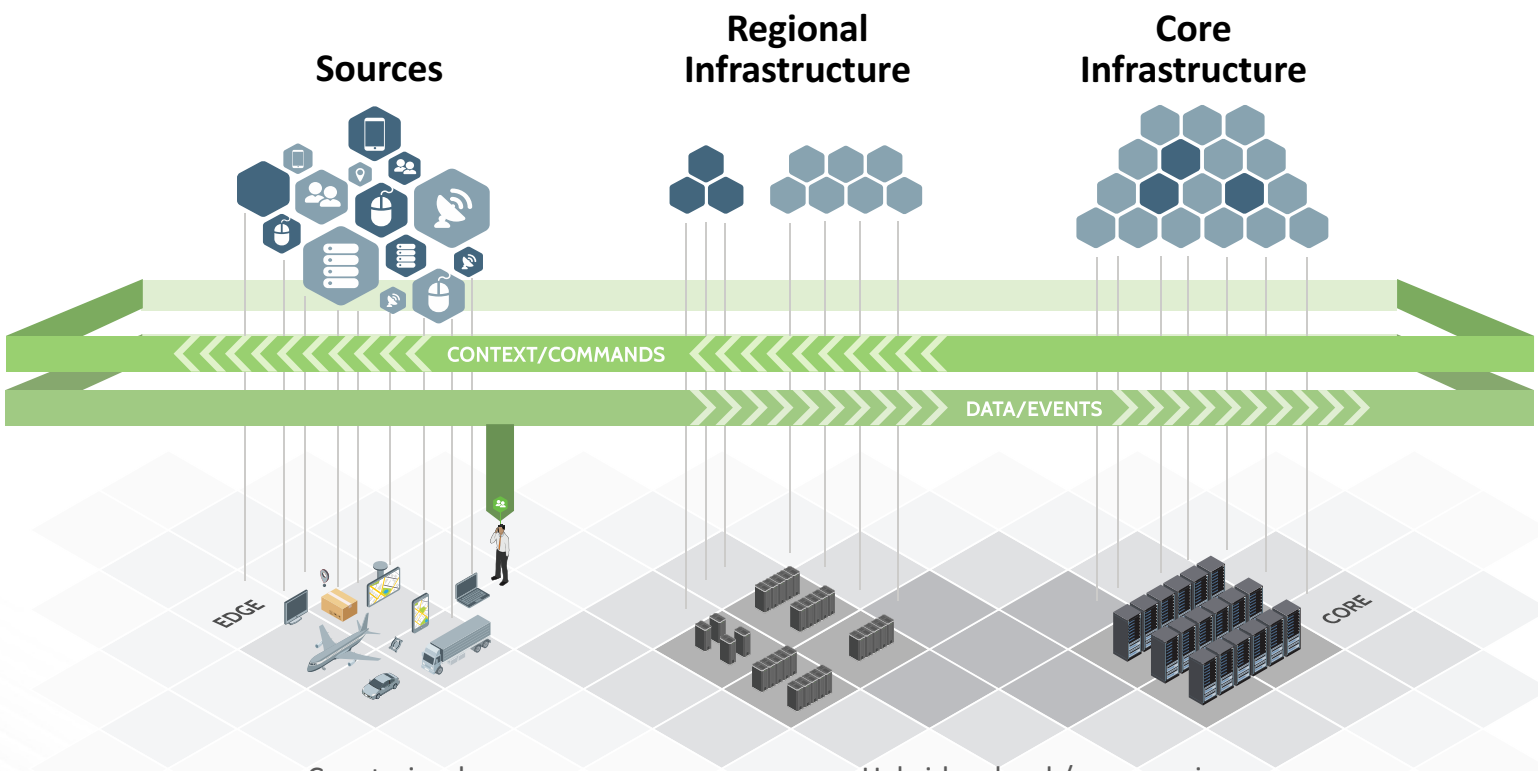


© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Hortonworks Data Flow (HDF)

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Hortonworks DataFlow Manages Data in Motion



Constrained
High-latency
Localized context

Hybrid – cloud / on-premises
Low-latency
Global context

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HDF Use Cases

Continuous Data Ingestion

- Dataflow development, deployment and management from **months to minutes.**
- Ingest Data from **any data source.**
- Make **real-time** data ingest flow changes.
- Dynamic back pressure support and **resilience**.
- First Mile problem – **Acquisition of data** at the edge.

Data Movement

- **Enterprise Data Movement.**
- **Move and synch data** between from on-prem and cloud environments.
- Acquire and Deliver Data **anywhere**, over any transport.

Stream Processing

- **Complex Event Processing**.
- **Streaming Analytics:** Real-time descriptive, predictive, and prescriptive analytics on perishable insights.
- Real-time **Actionable Intelligence**.

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

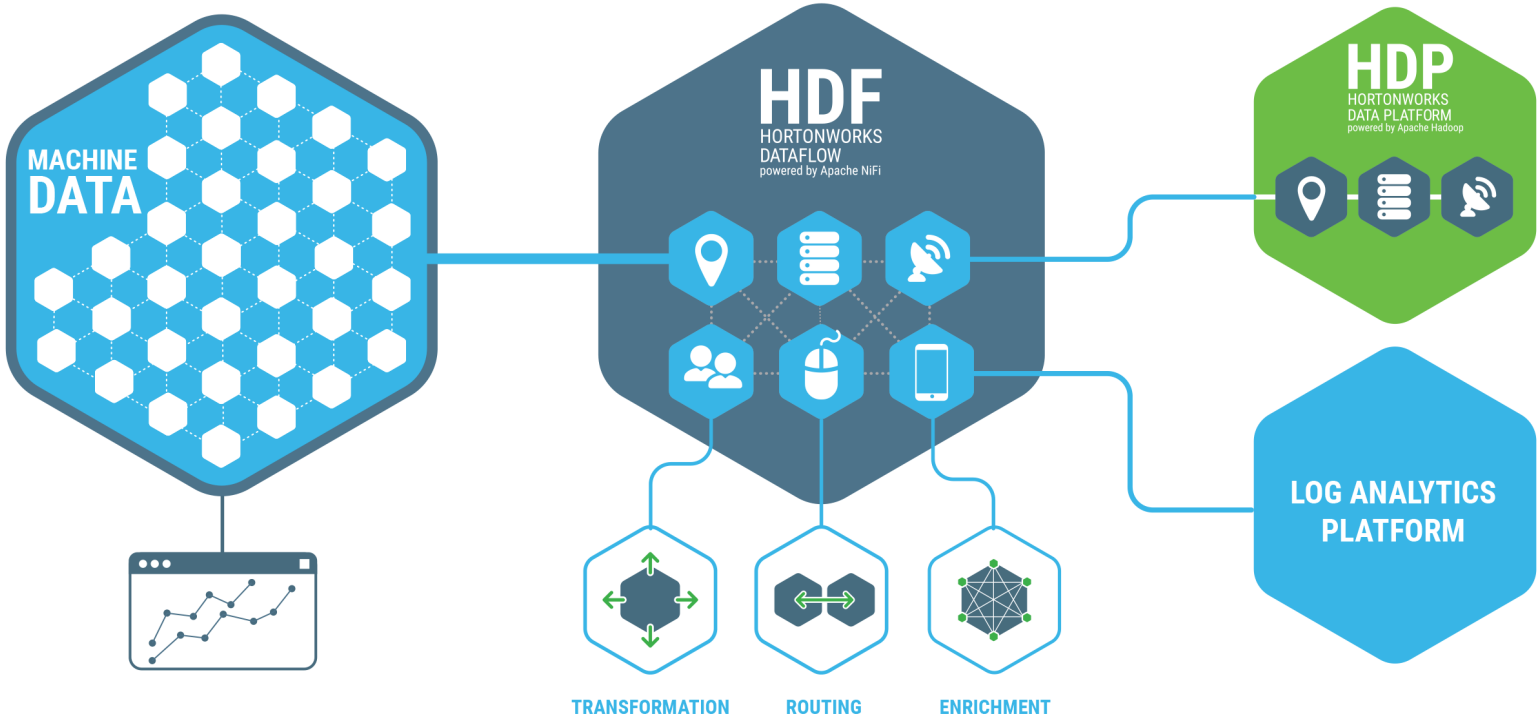
HORTONWORKS

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

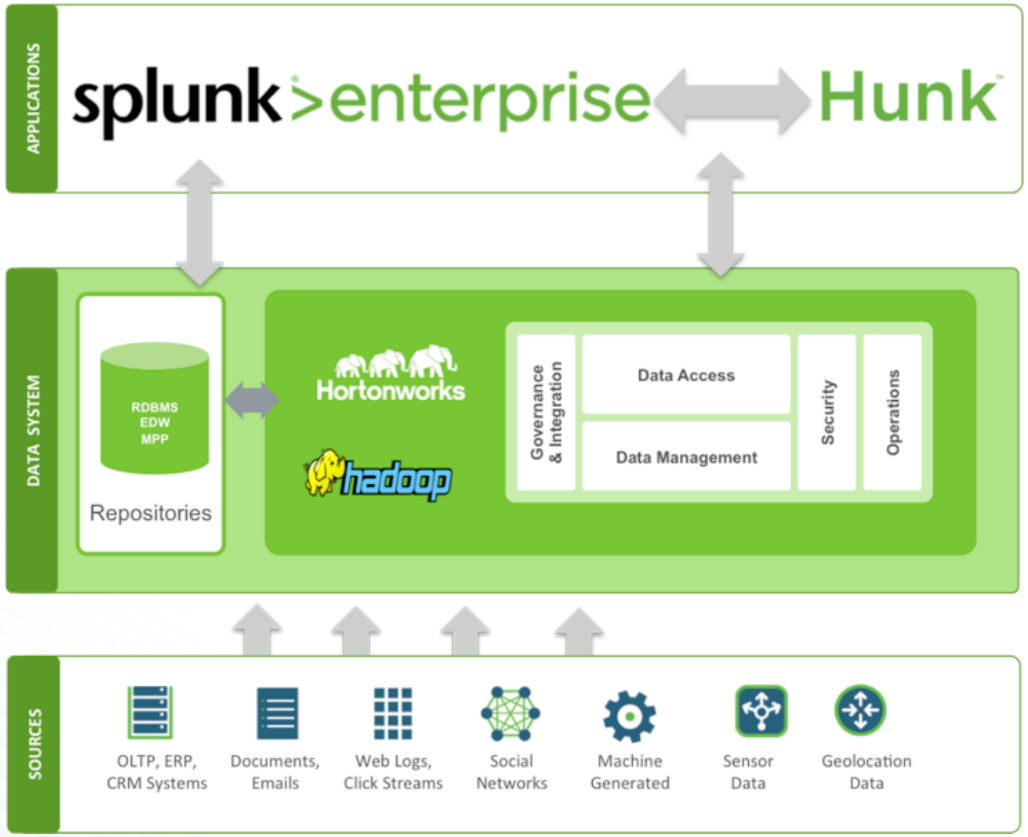
Send data to alternative systems based on value, content, priority



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

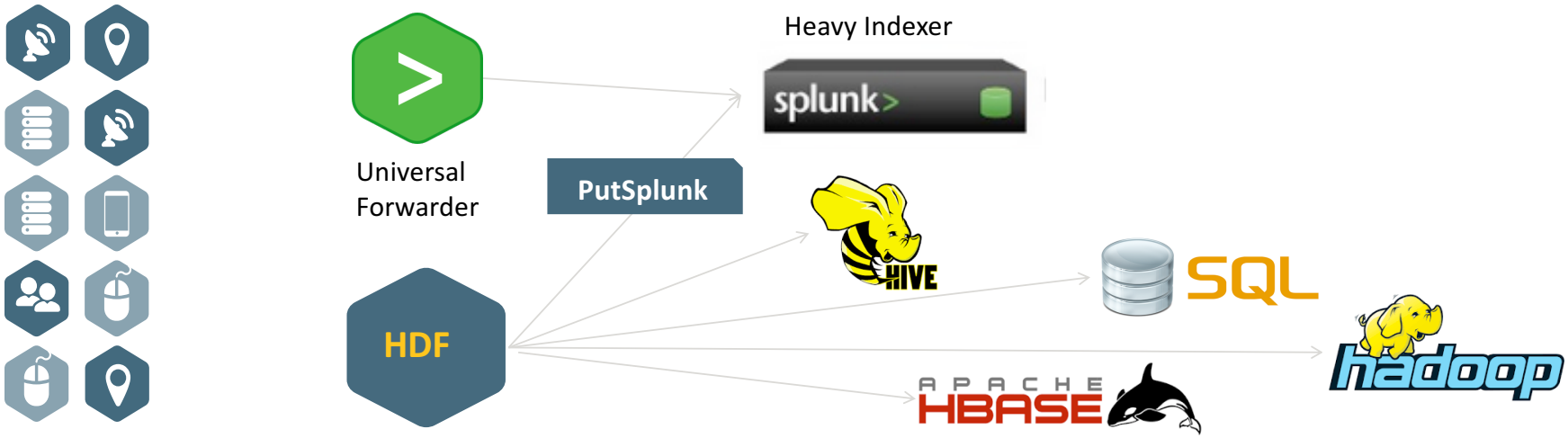
Splunk, Part of the Modern Data Architecture



- Bi-directional data integration between Splunk & HDP
- Collect data from across the organization, deliver it to Hadoop for refining data and batch analytics
- Output of Hadoop jobs can be imported into Splunk Enterprise for rapid analysis and visualization
- Archiving from Splunk Enterprise to Hadoop

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Augment Existing Splunk Deployment with HDF to Distribute Data



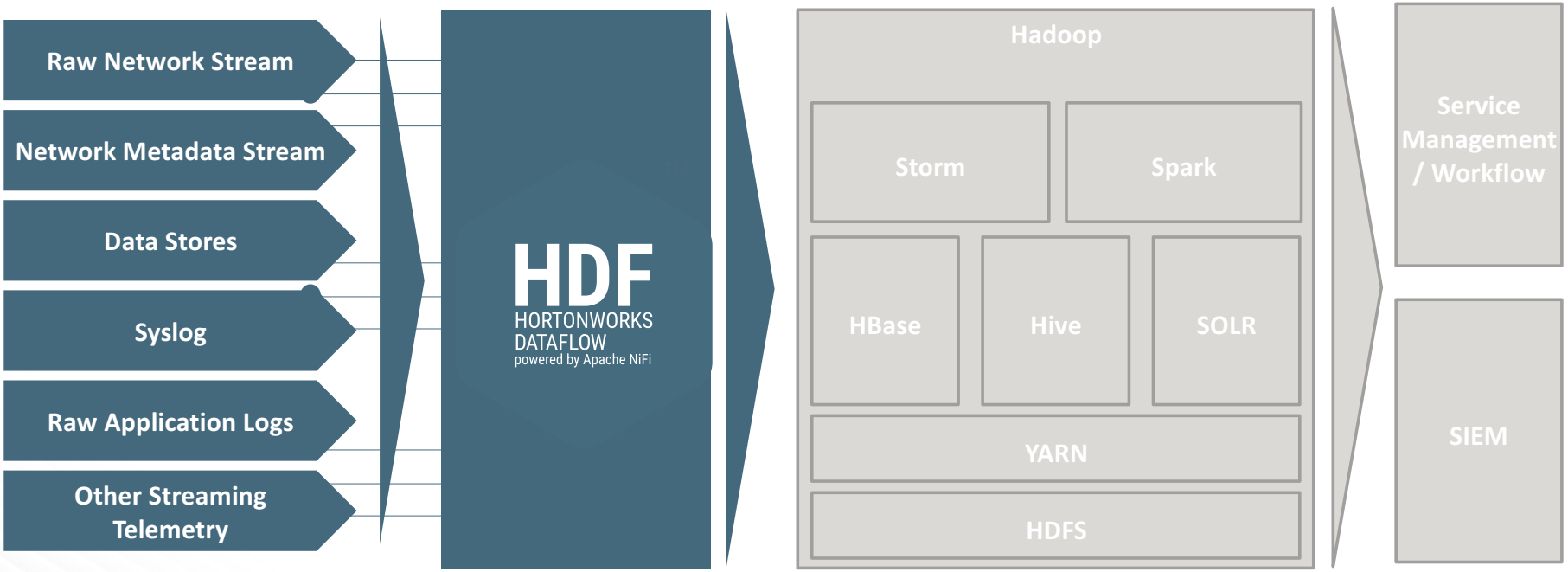
HDF Enables

- Single integrated system to simultaneously send data to Splunk and alternative systems
- Intelligence at the edge to route, enrich, prioritize data

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

Cybersecurity Log Ingestion



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HDF Value proposition

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Problems Today: Timely Access to Data and Decisions

Big data's biggest problem: It's too hard to get the data in

While big data has been turned into more of a marketing term than a technology, it still has enormous untapped potential. But, one big issue has to get solved first.

By Jason Hiner | April 10, 2016 -- 22:00 GMT (15:00 PDT) | Topic: Big Data Analytics

"HDF helps us to

so that my team

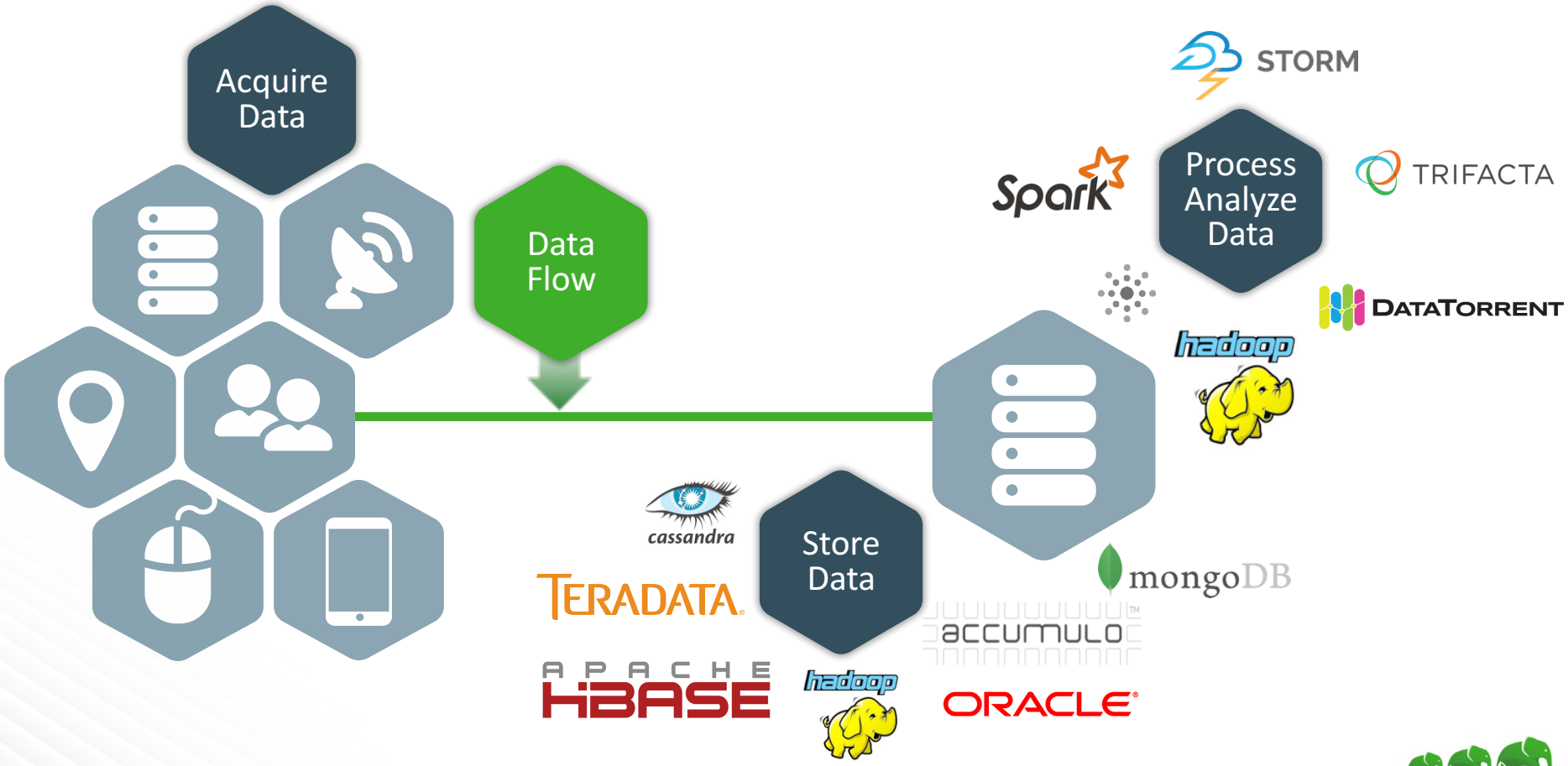
Royal Mail

http://diginomica.com/2016/04/22/royal-mail-starts-to-deliver-on-hortonworks-data-in-motion-promise

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Big Data to Nowhere !?

Simplistic View of Dataflows: Easy, Definitive

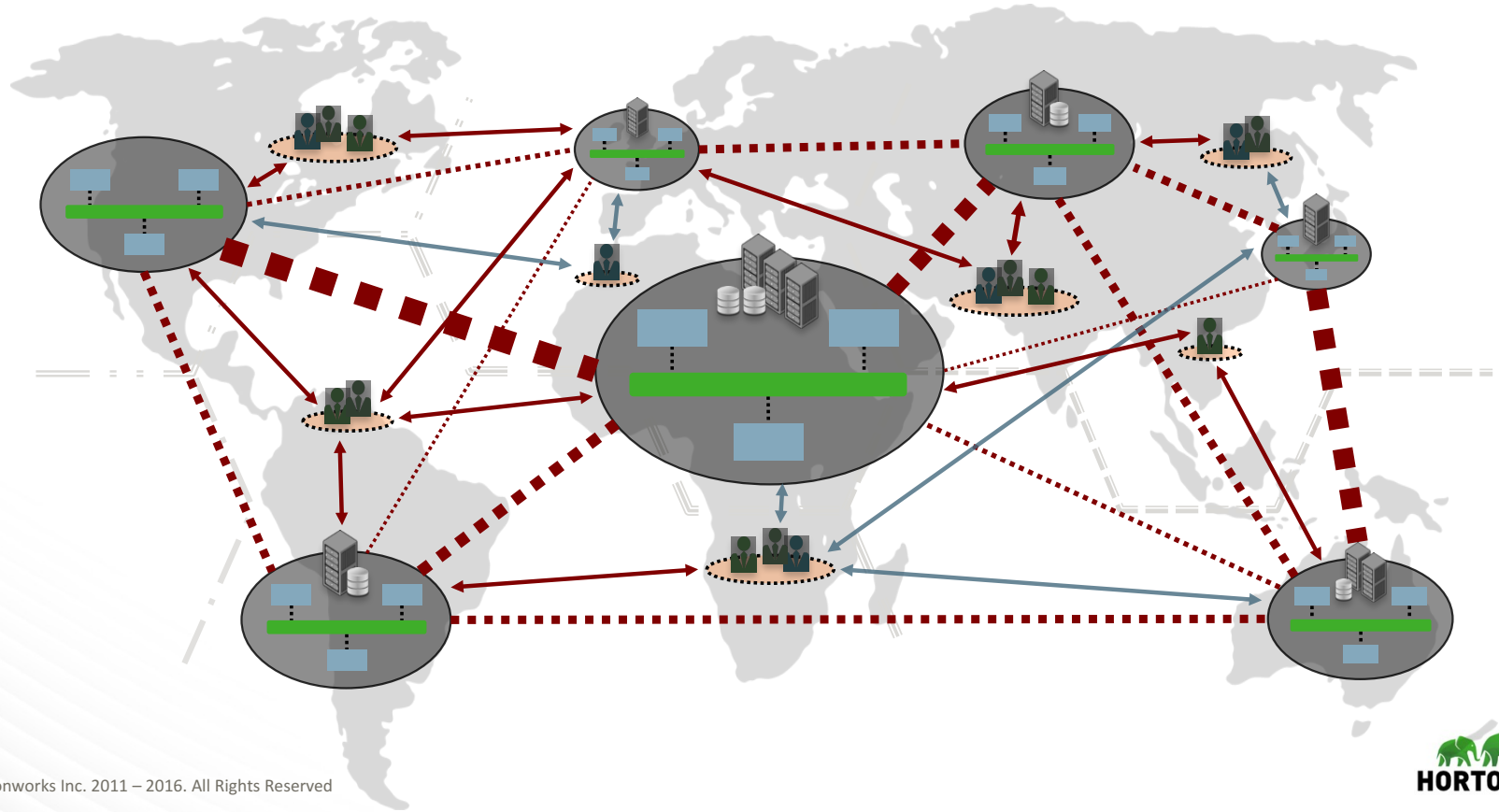


© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

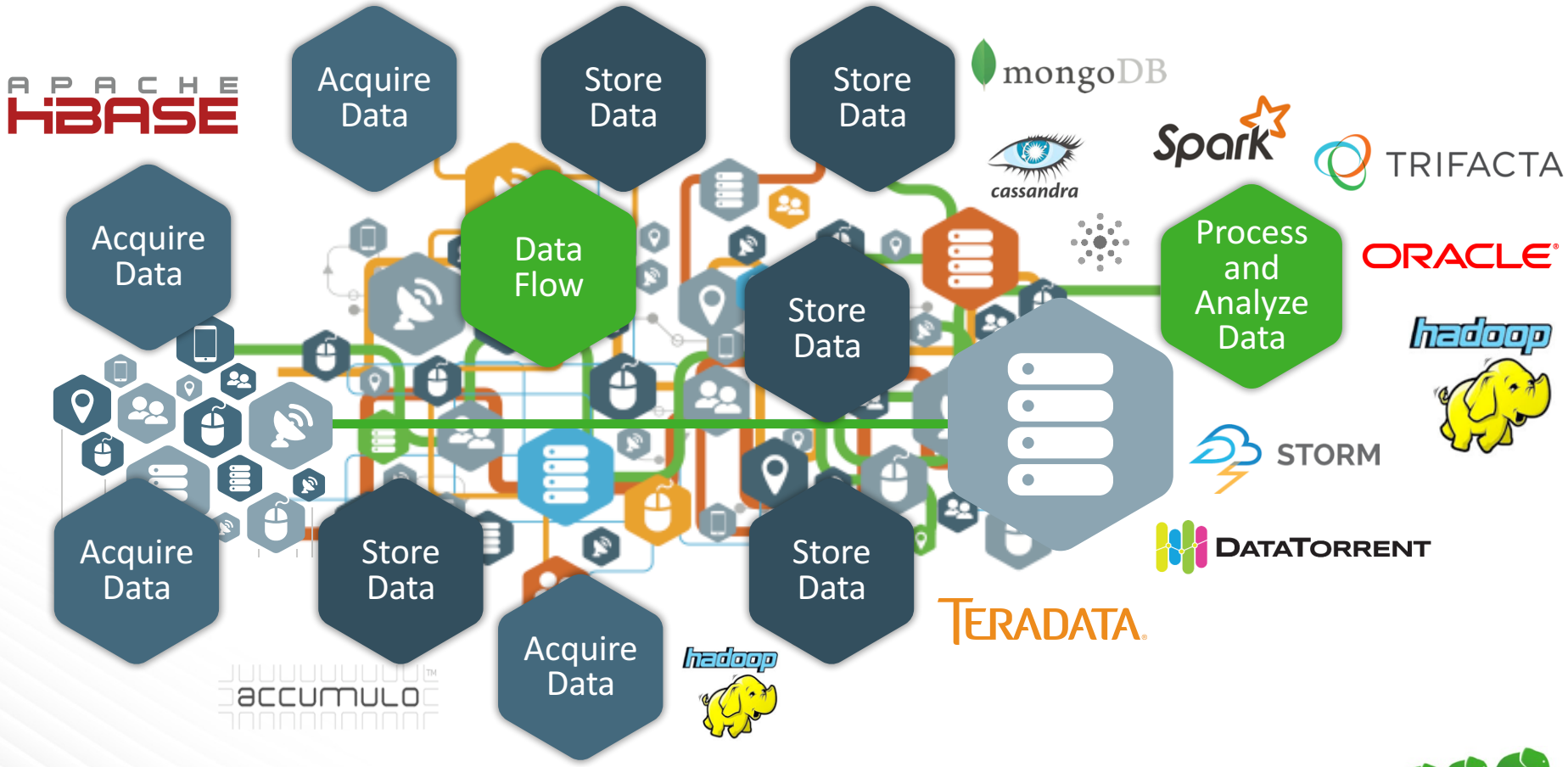
Realistic View of Managing Flow of Enterprise Data

Global interactions with customers, partners, systems and things spanning different volume, velocity, bandwidth, and latency needs



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Realistic View of Dataflows: Complex, Convoluted



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

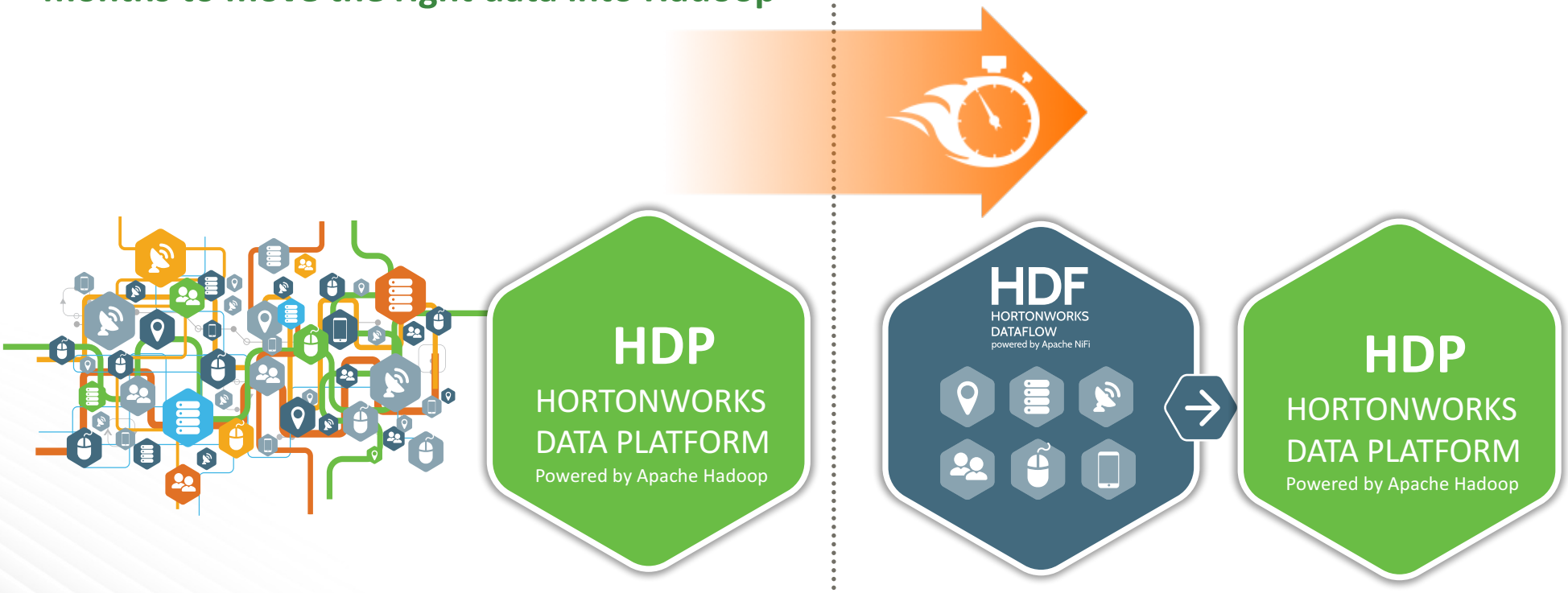
© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Yet we had a solid roadmap for a data lake !

HDF Makes Big Data Ingest Easy

Complicated, messy, and takes weeks to months to move the right data into Hadoop

Streamlined, Efficient, Easy



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

HDF Addresses Challenges

- Logs
- Files
- Feeds
- Sensors

Collect: Bring Together

Aggregate all data from sensors, geo-location devices, machines, logs, files, database and feeds via a highly secure lightweight agent

- Deliver
- Secure
- Govern
- Audit

Conduct: Mediate the Data Flow

Mediate point-to-point and bi-directional data flows, delivering data reliably to real-time applications and storage platforms such as HDP

- Parse
- Filter
- Transform
- Fork
- Clone

Curate: Gain Insights

Parse, filter, join, transform, fork, and clone data in motion to empower analytics and perishable insights

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

Data in Motion Needs Dataflow Management and Stream Processing

1. **Acquire** truck telemetry events from different Geo Cloud Instances and from the edge (telemetry devices on certain trucks)
2. **Move and Filter** Data from Customer X Cloud Instances to on-premise instance
3. **Intelligent Routing** of data. The routing and filtering rules will be often changed at run-time.
4. **Deliver** the data data to various downstream systems.
5. **Parse** the device data to standardized format that downstream sysem can understand
6. **Enrich** the data with contextual information HR, Payroll, Truck Maintenance records, weather
7. **Join multiple sensor** streams via attributes in real-time
   1. E.g: Join separate streams like **geo-location stream** of a truck with the **driver speed stream**.
8. Detect **complex patterns in real-time:**
   1. E.g: **Over a 3 minute period, if a dirver has multiple infractions across 2 different infraction types**, then alert (Braking, Speeding)
   2. E.g**: Over a 5 minute period, if the average speed of a driver is > 80 in known dangerous routes**, then alert.

Flow Management

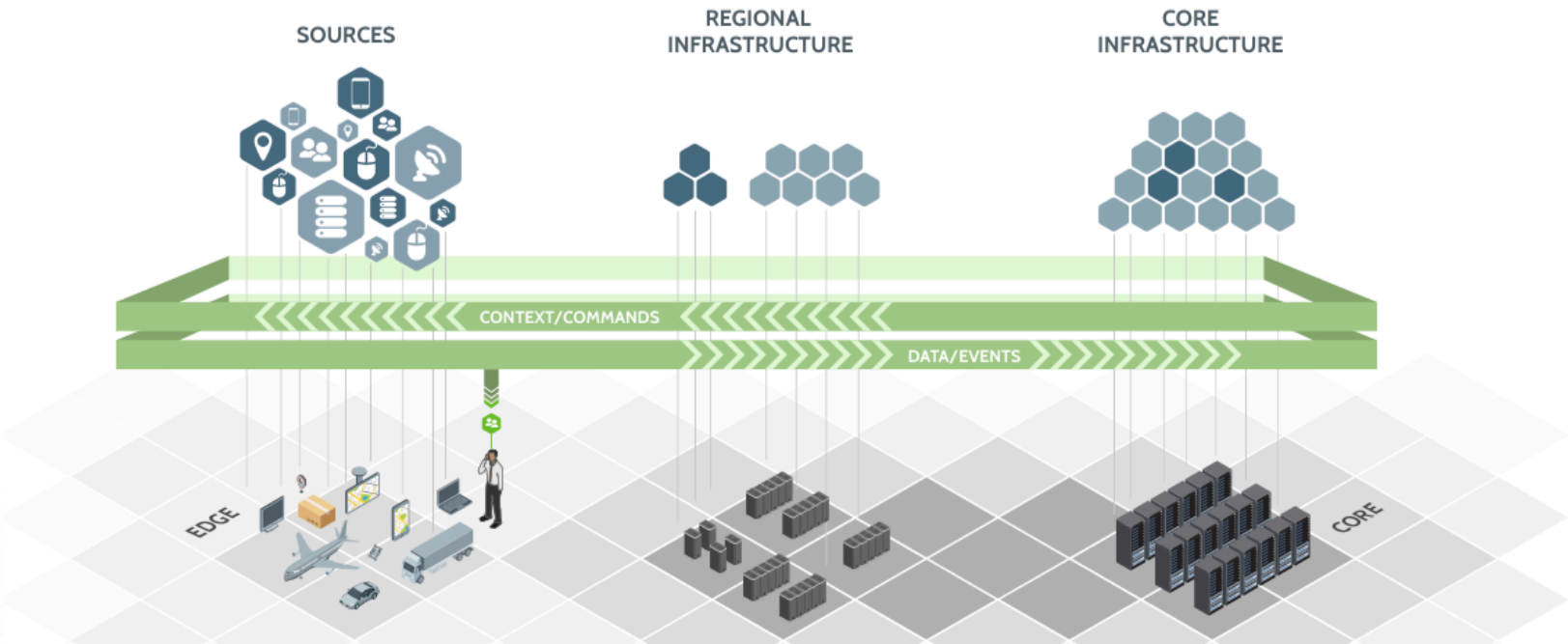
Stream Processing

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

End-to-End Data Ingestion : IoT

Resolves real world connectivity and transmission issues often overlooked by assuming connectivity is always perfect



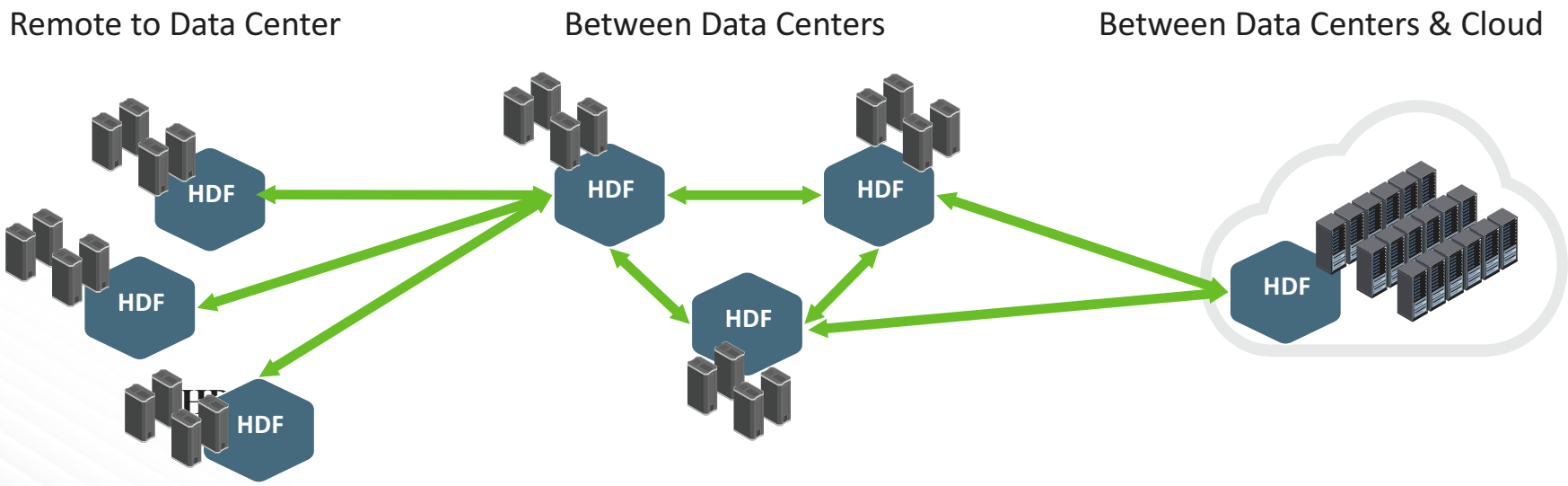
Constrained
High-Latency
Localized Context

Hybrid – Cloud/On-Premise
Low-Latency
Global Context

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

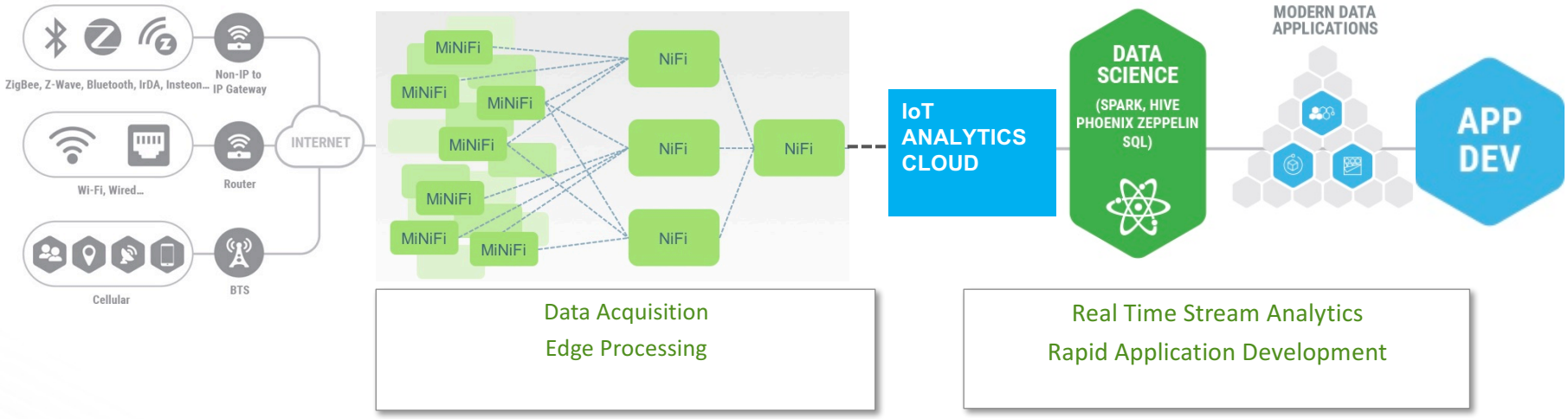
Enterprise Data Movement and Hybrid Cloud

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



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Stream Processing

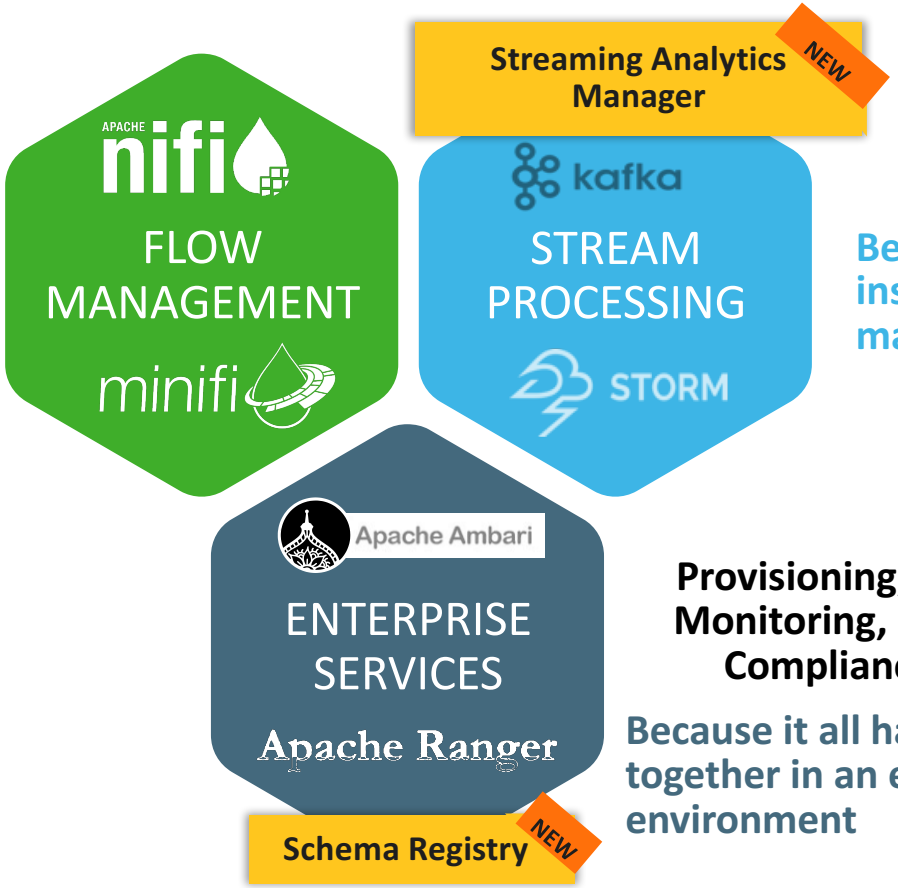


© Hortonworks Inc. 2011 – 2016. All Rights Reserved
© Hortonworks Inc. 2011

Hortonworks Data Flow Component Overview

Easy, Secure, Reliable Way to Get the Data You Need

From the edge to anywhere with intelligence



Immediate and Continuous Insights

Because acting on perishable insights in real time maximizes value

Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance

Because it all has to work together in an enterprise environment

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

Data Flow Management with Apache NiFi

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

A little bit of history

NATIONAL SECURITY AGENCY
CENTRAL SECURITY SERVICE
FORT GEORGE G. MEADE, MARYLAND 20755-6000

Declassified

NSA PRESS RELEASE
25 November 2014
For further information contact:
NSA Public and Media Affairs, 301-688-6524

NSA Releases First in Series of Software Products to Open Source Community

New technology automates high-volume data flows

The National Security Agency announced today the public release of its new technology that automates data flows among multiple computer networks, even when data formats and protocols differ. The tool, called "Niagarafiles (Nifi)," could benefit the U.S. private sector in various ways. For example, commercial enterprises could use it to quickly control, manage, and analyze the flow of information from geographically dispersed sites – creating comprehensive situational awareness.

Developed by the NSA over the last 8 years.

"NSA's innovators work on some of the most challenging national security problems imaginable,"

"Commercial enterprises could use it to quickly control, manage, and analyze the flow of information from geographically dispersed sites – creating comprehensive situational awareness"

-- Linda L. Burger,
Director of the NSA

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

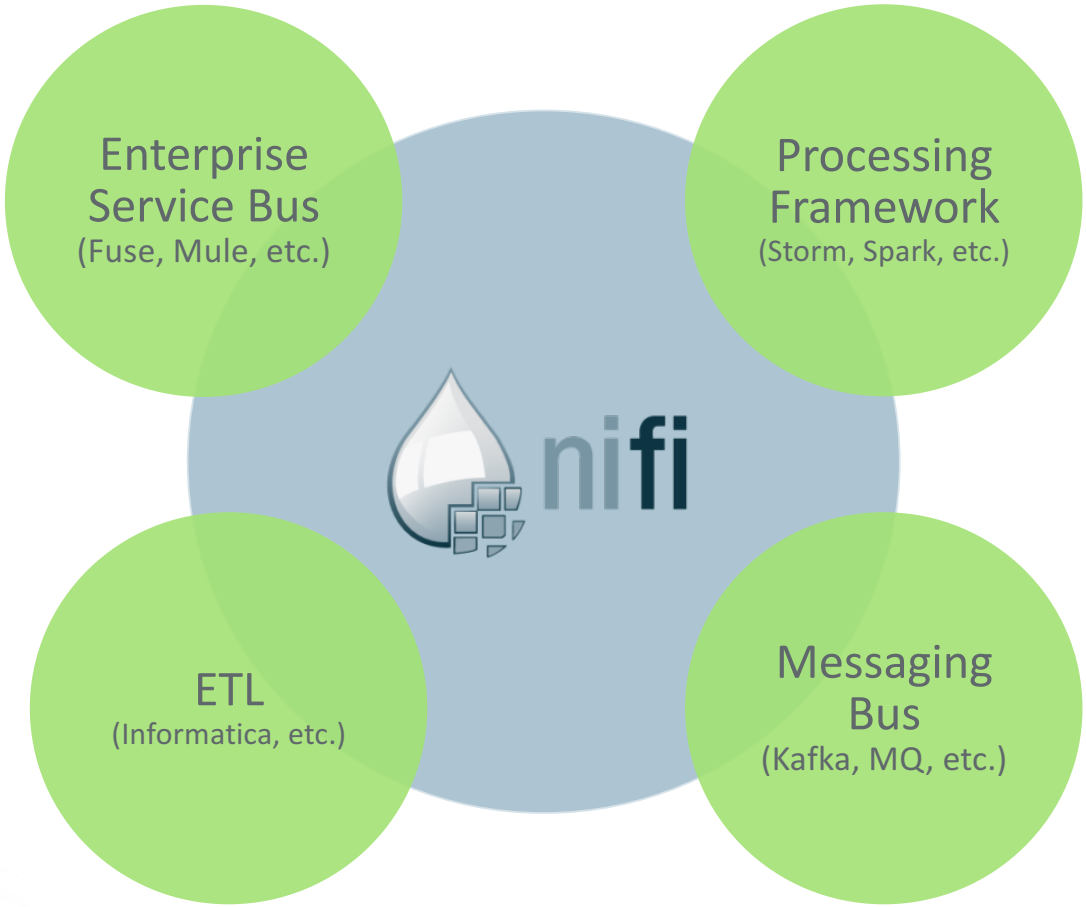
A little bit of history

- The NiFi project was created by the US National Security Agency and originally named **NiagaraFiles**.
- In 2014, the NSA released it as open-source software, via its technology transfer program.
- Development of Apache NiFi continued at Onyara Inc.
- Onyara acquired by Hortonworks in August 2015.

HORTONWORKS

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

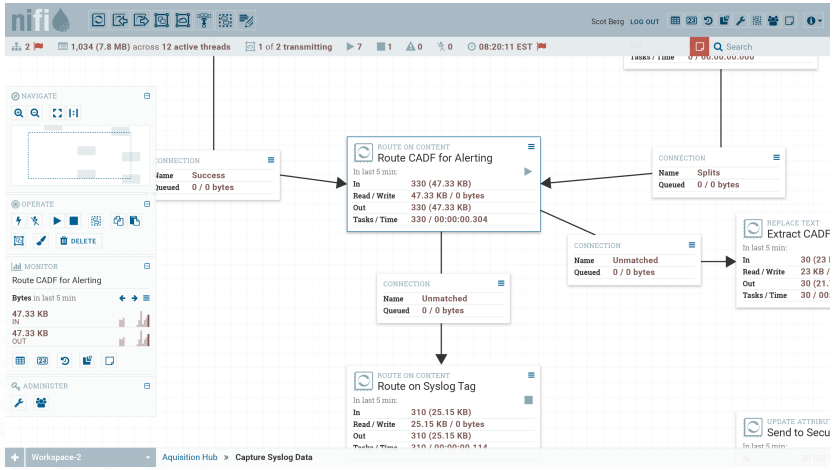
Apache NiFi Positioning



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

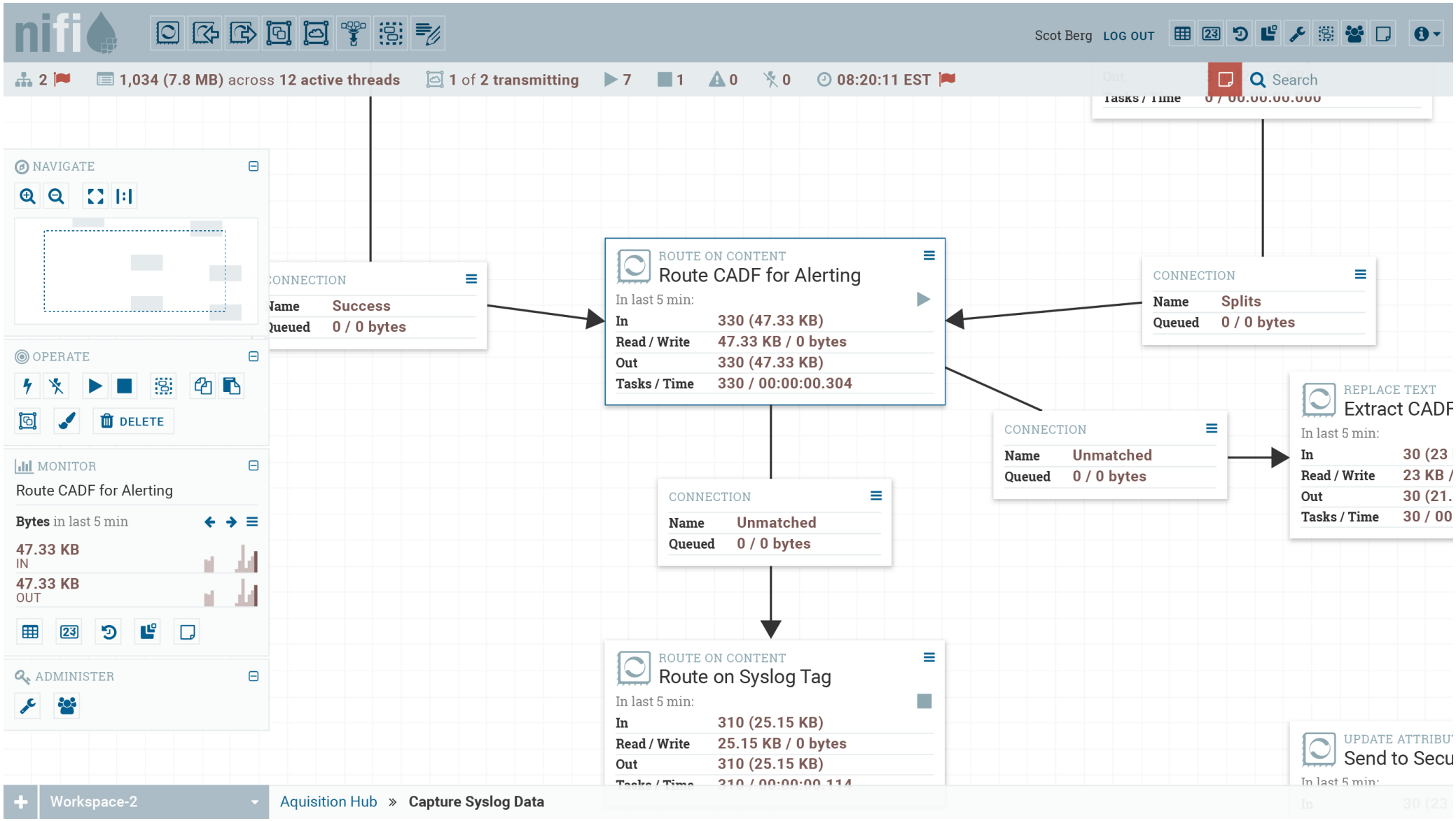
Apache NiFi

Key Features

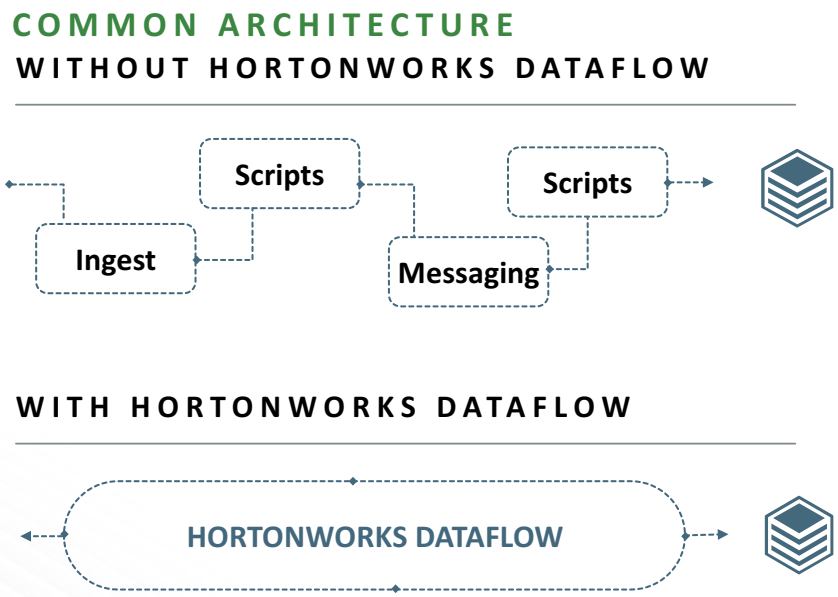


- **Guaranteed delivery**
- **Data buffering**
  - Backpressure
  - Pressure release
- **Prioritized queuing**
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- **Data provenance**
- Supports push and pull models
- Recovery/recording a rolling log of fine-grained history
- **Visual command and control**
- Flow templates
- Pluggable/multi-role security
- **Designed for extension**
- Clustering

© Hortonworks Inc. 2011 – 2016. All Rights Reserved


nifi
Scot Berg
LOG OUT
2
1,034 (7.8 MB) across 12 active threads
1 of 2 transmitting
7
1
0
0
08:20:11 EST
Search
NAVIGATE
OPERATE
DELETE
MONITOR
Route CADF for Alerting
Bytes in last 5 min
47.33 KB
IN
47.33 KB
OUT
ADMINISTER
CONNECTION
Name Success
Queued 0 / 0 bytes
ROUTE ON CONTENT
Route CADF for Alerting
In last 5 min:
In 330 (47.33 KB)
Read / Write 47.33 KB / 0 bytes
Out 330 (47.33 KB)
Tasks / Time 330 / 00:00:00.304
CONNECTION
Name Splits
Queued 0 / 0 bytes
CONNECTION
Name Unmatched
Queued 0 / 0 bytes
REPLACE TEXT
Extract CADF
In last 5 min:
In 30 (23
Read / Write 23 KB /
Out 30 (21.
Tasks / Time 30 / 00
CONNECTION
Name Unmatched
Queued 0 / 0 bytes
ROUTE ON CONTENT
Route on Syslog Tag
In last 5 min:
In 310 (25.15 KB)
Read / Write 25.15 KB / 0 bytes
Out 310 (25.15 KB)
UPDATE ATTRIBU
Send to Secu
Workspace-2
Aquisition Hub » Capture Syslog Data

Integrated Processes and Control for Data Flows and Streaming Analytics



Optimize the Architecture

Reduce cost an complexity with the most efficient data collection technologies

Assure Efficient Operations

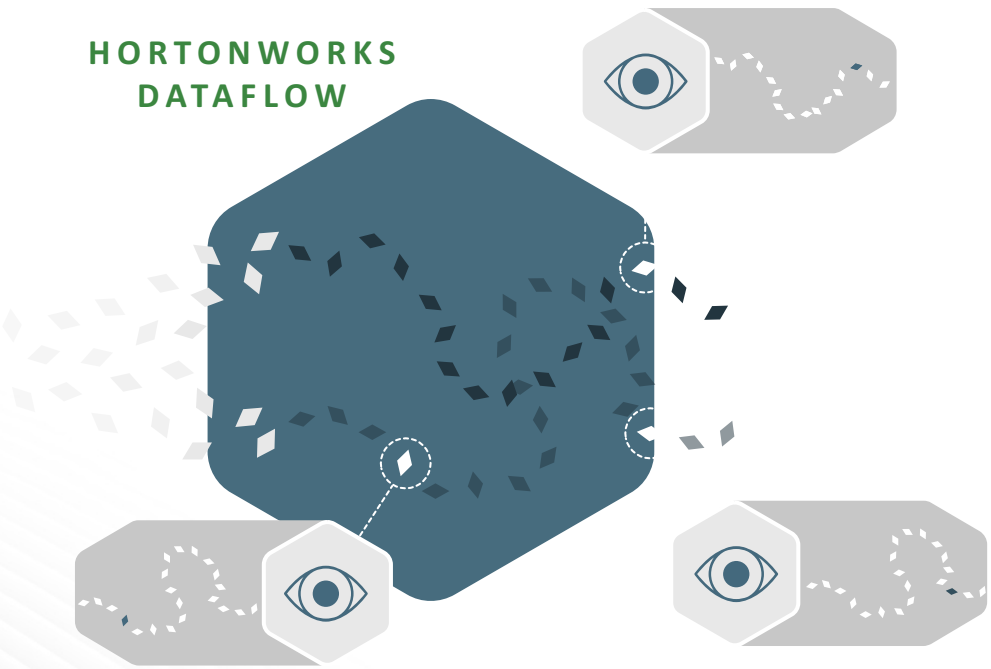
Via real-time control of data inputs, outputs, transportation and transformations

Rely on a Common Foundation

Eliminating dependence on multiple customized systems

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Secure Flows with Chain of Custody and Provenance



End-to-End Security

Apply security rules to encrypt, decrypt, filter and replace data from the point of collection at the jagged edge to its final destination

Granular Control and Sharing

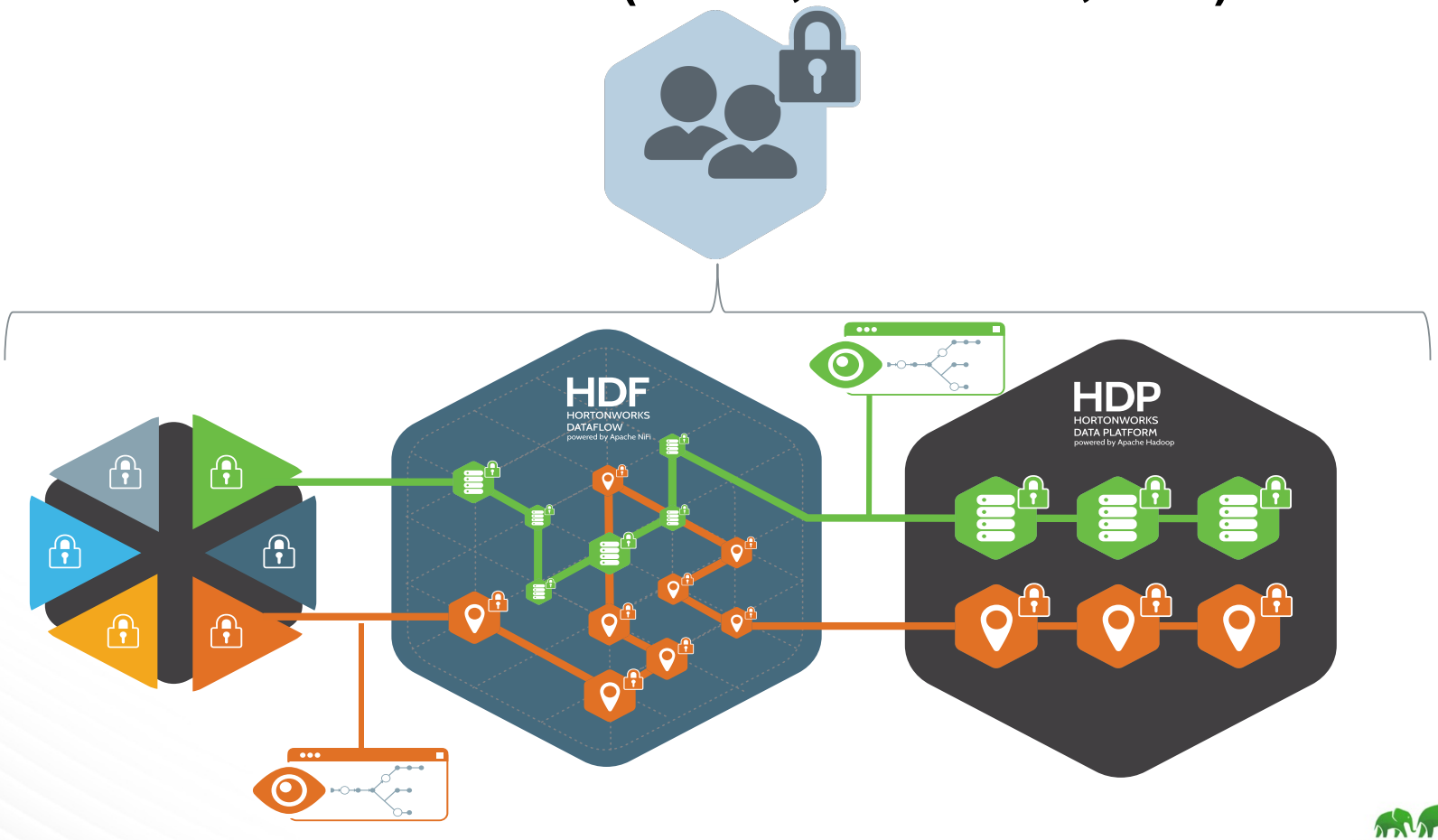
Move beyond role-based access and dynamically share an entire dataflow

Real-Time Traceability

Rich metadata and contextual detail helps troubleshoot security issues and informs timely decisions

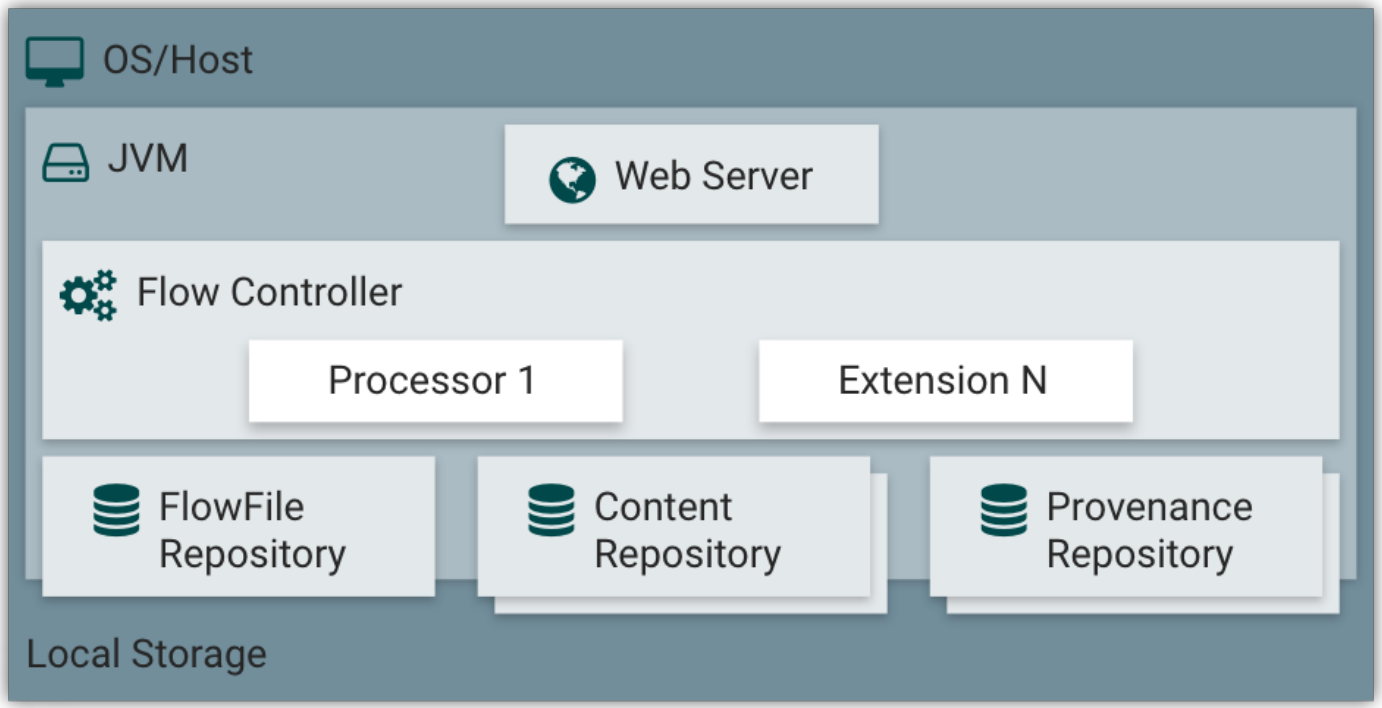
© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Centralized Authentication (LDAP, Kerberos, etc.)



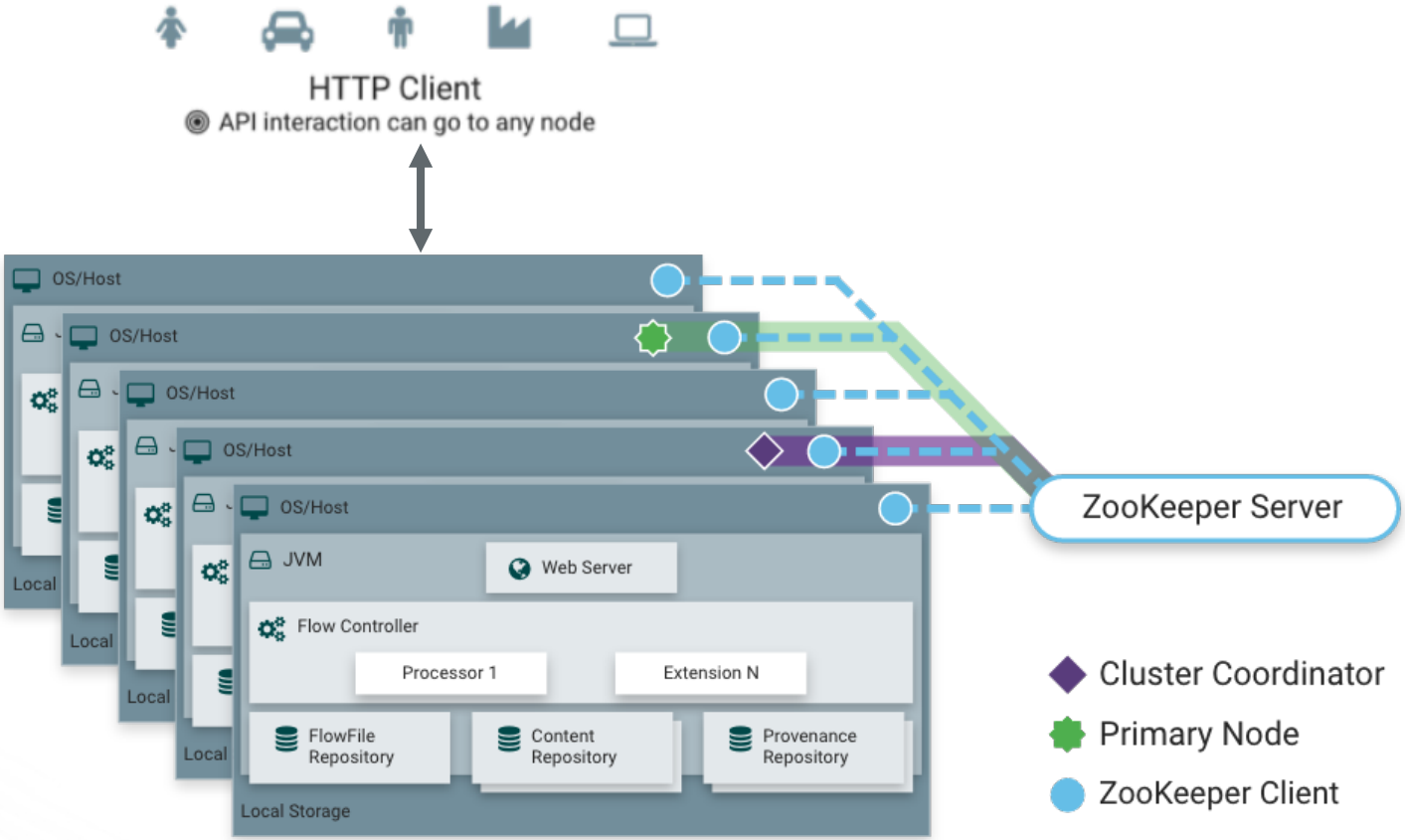
© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Nifi Architecture : Primary components



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Nifi Architecture : Zero-Master Clustering



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Connecting Data Between Ecosystems Without Coding: 200+ Processors

- 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	

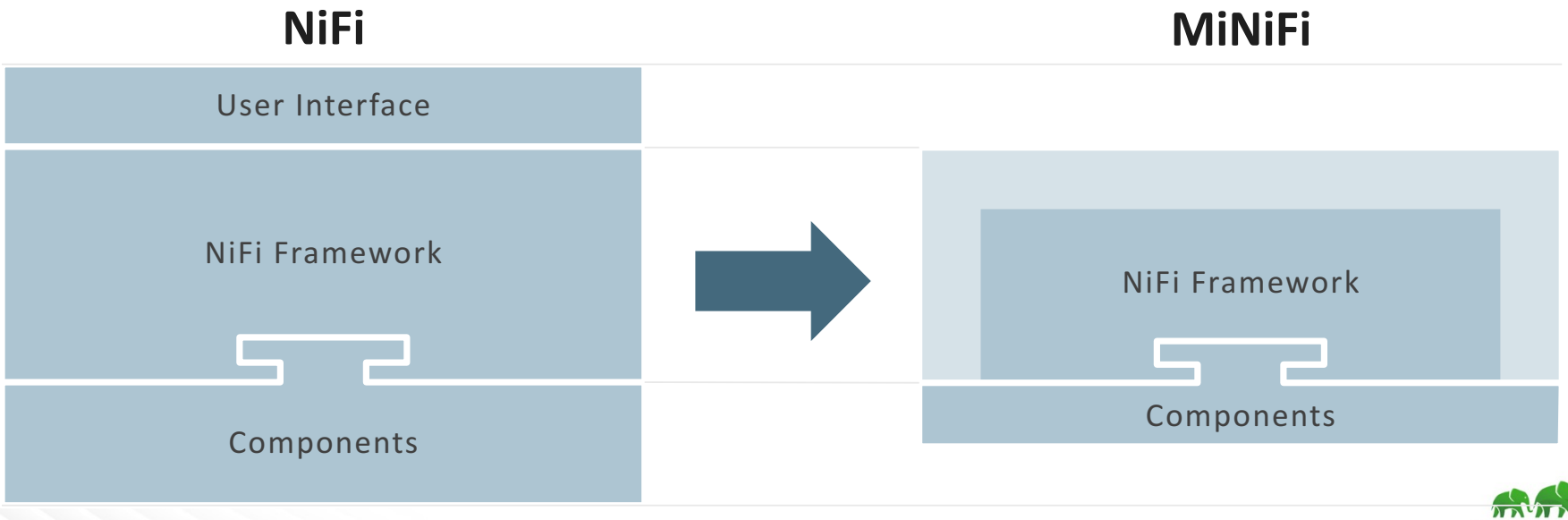
All Apache project logos are trademarks of the ASF and the respective projects.

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

MiNiFi: Edge Intelligence

- Lightweight NiFi version at the edge, where data begins and provide bidirectional communication
- NiFi lives in the data center. Give it an enterprise server or a cluster of them.
- MiNiFi lives close to where data is born and is a guest on that device or system



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

NiFi Lab Demo

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Use Case

As social media analytics company, the first product we are working on is a Meetup Analytics dashboard.

The first step in that project consist in building an ingestion pipeline to automatically fetch all relevant data regarding the meetups happening right now and store the results on a file system.

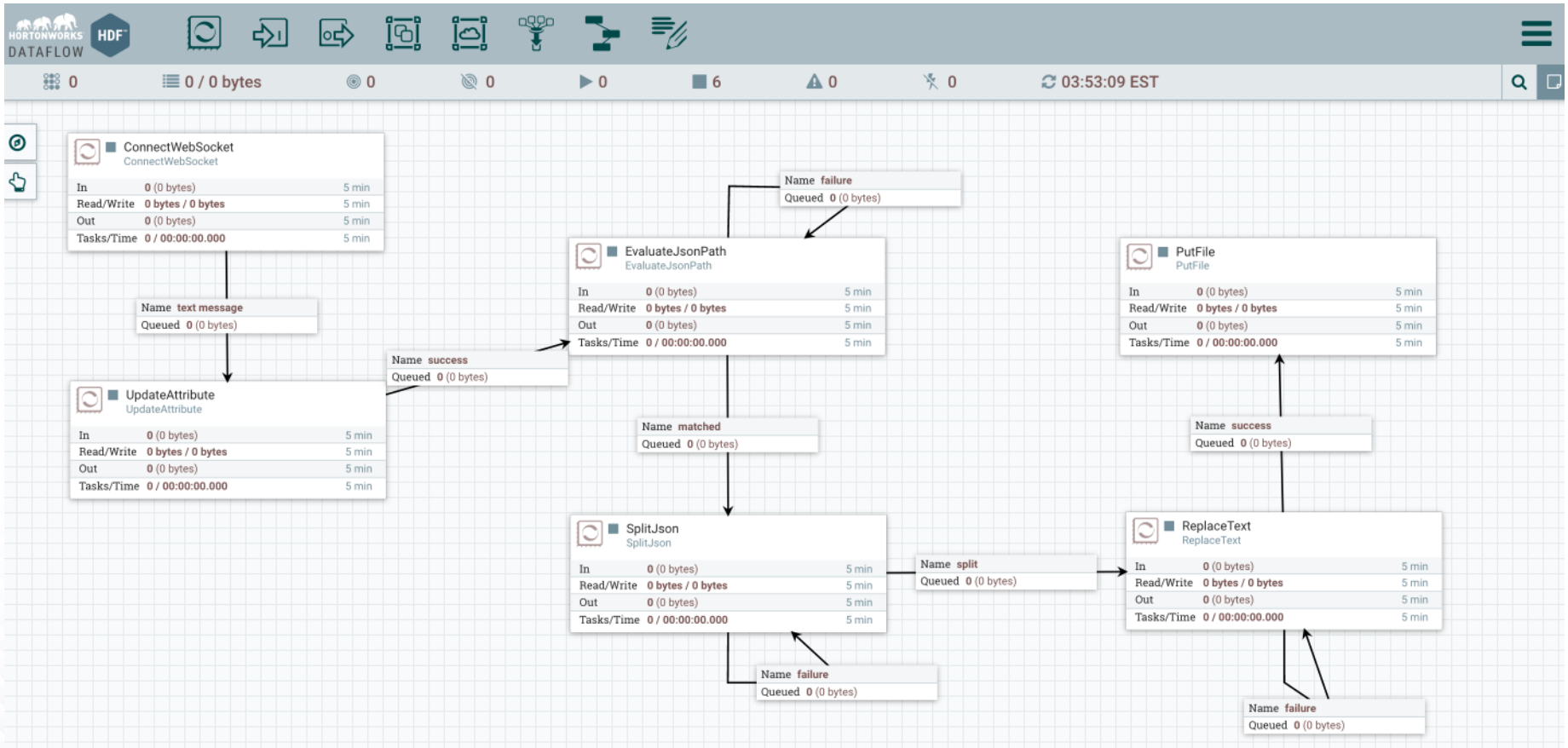
Goals:

- Consume Meetup RSVP stream
- Extract the JSON elements we are interested in
- Split the JSON into smaller fragments
- Write the JSON files to disk

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

Ingestion pipeline



© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Where to find Lab exercises and presentations ?

https://github.com/oascofare/hdflab

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

Good Luck !

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

Hardware Considerations for HDF

General Hardware

A key design point of NiFi is to use typical enterprise class application servers.

Hardware failure:

- Nodes are typically configured with RAID 5/10 to handle hardware failures through replication, redundancy.
- More nodes means less impact from failure.
- More nodes provide increased throughput.

Machine Class

A NiFi cluster consists of a single class of machine

Balanced NiFi Node:

- 8 CPU cores per node minimum.
- 6 Hard disks per node minimum, Spinning or SSD base on throughput requirements.
- 8 GB of RAM per node minimum.
- Designed for availability.
- Typical enterprise class application server.
- Resilience built into the server itself (RAID)
- Cost reduced where possible to strike proper price/performance ratio owing to volume.

Networking

Network decisions play a role due to the clustered nature of data processing.

In-rack backplane/Top-of-rack Switch:

- Keeps traffic local and reduces load on expensive aggregate switches.

Dual NIC Recommended:

- Depends on NTW requirements.

10G Recommended:

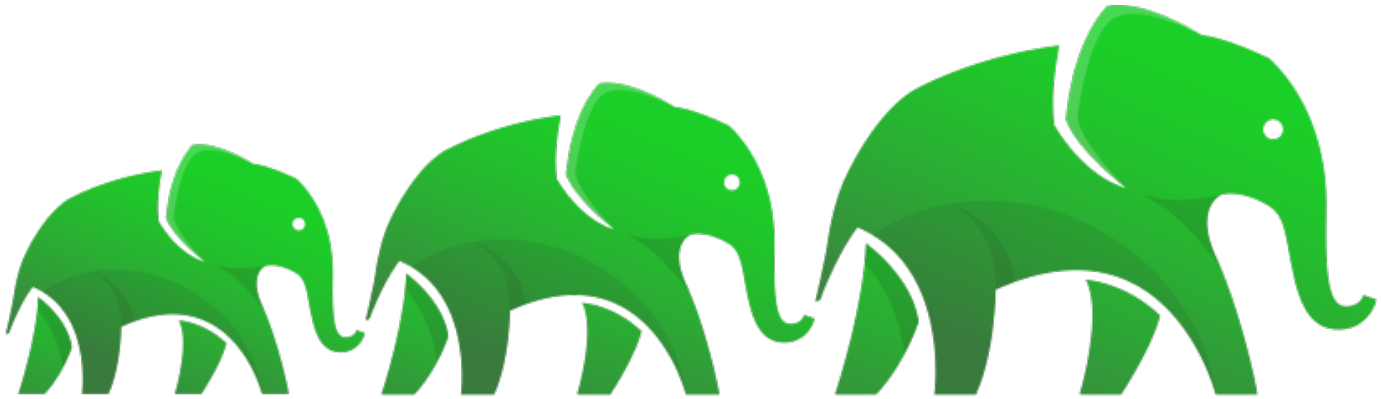
- Immediate cost vs Future-proofing.
- Investment in 10G upfront will survive next 2-3 server hardware upgrades.
- In-rack/top-of-rack switches allow Cat6 copper or Twinax to reduce 10G costs.

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

NiFi Sizing Guidelines

Throughput Target	Number of NiFi nodes	CPU Cores/node	Number of disks/node, size of each disk (RAID 5/10)	RAM/node	Ideal Networking Setup
50 MB/s, 1000 events/s	3	8	6+, 1TB	16+ GB	1-10 Gigabit bonded NICs
100 MB/s, 10,000 events/s	5	16+	6+, 2TB	16+ GB	1-10 Gigabit bonded NICs
200 MB/s, 100,000 events/s	7	24+	12+, 4TB	16+ GB	10 Gigabit bonded NICs
400 MB/s, 100,000+ events/s	9	24+	12+, 8TB	16+ GB	10 Gigabit bonded NICs

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

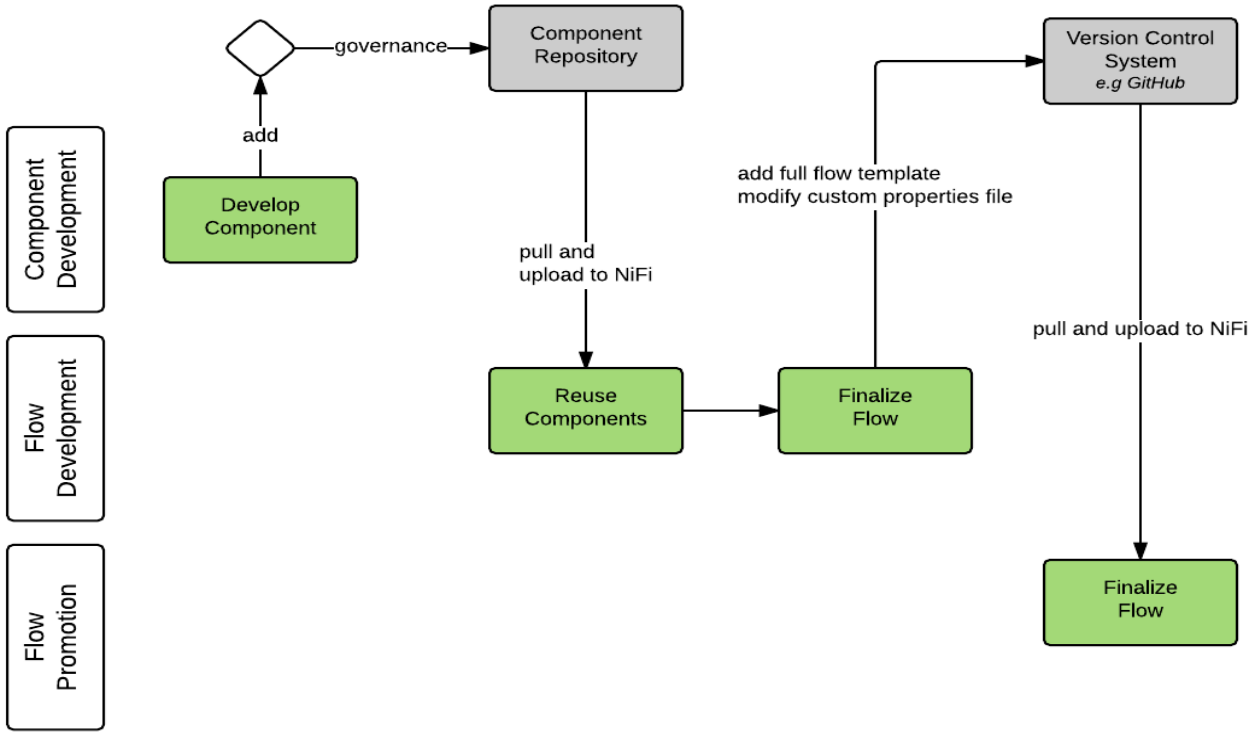


HORTONWORKS®

POWERING THE FUTURE OF DATA™

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

SDLC with Apache NiFi



- Reusable components are added as templates to a central repository.
- Development groups pull reusable components and upload to their NiFi environment to build new flows.
- Other environment-specific config values should similarly use Expression Language references.
- Developers finalize their flows and submit the template of the flow to version control, eg Git (and also submit custom property files).

© Hortonworks Inc. 2011 – 2016. All Rights Reserved