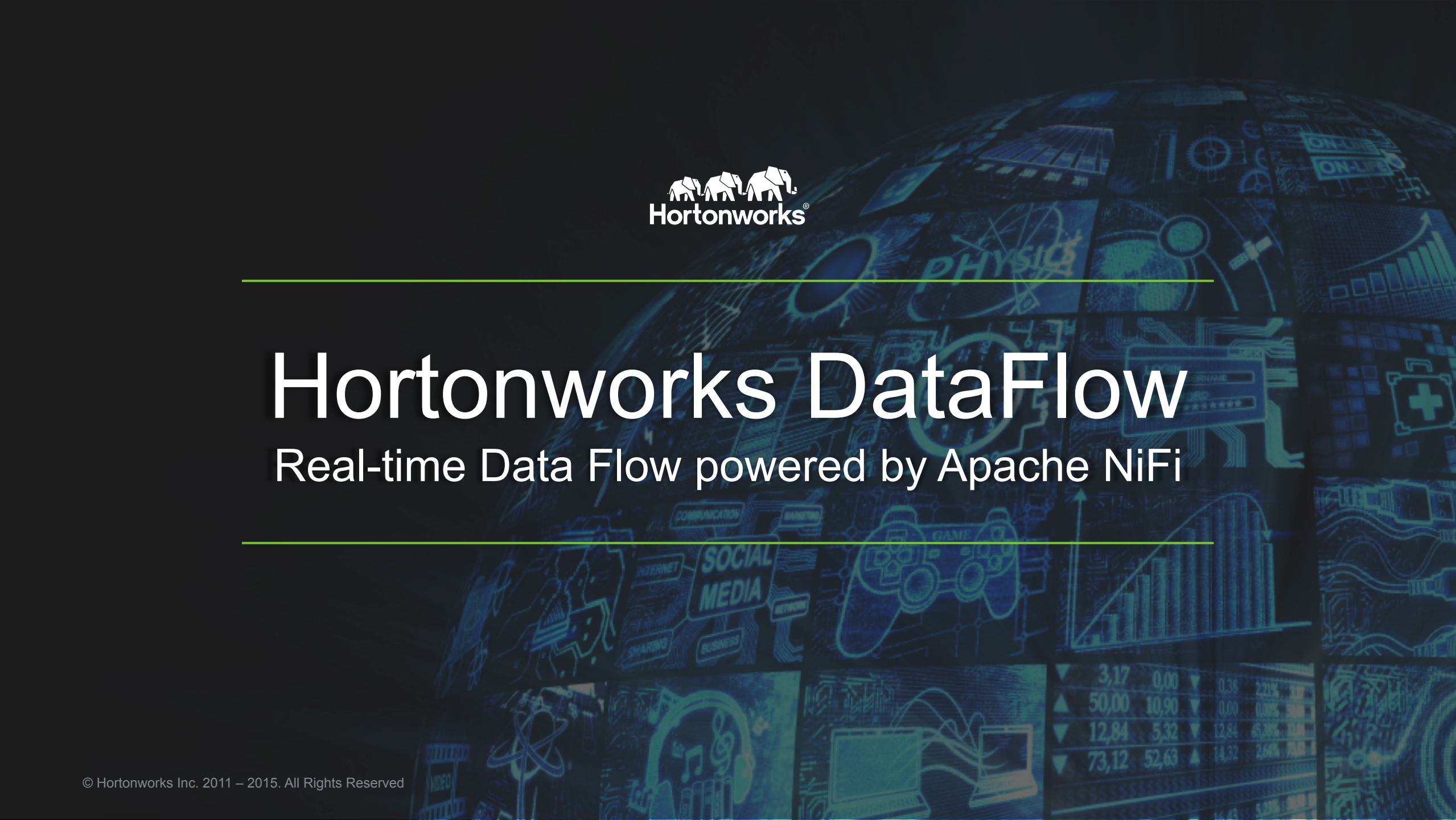
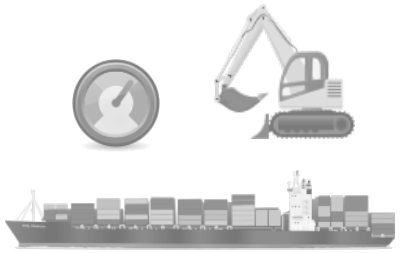


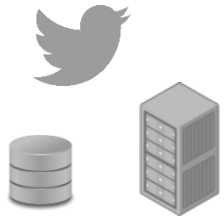
Hortonworks DataFlow

Real-time Data Flow powered by Apache NiFi

Simplistic View of Enterprise Data Flow



Acquire Data



Process and Analyze Data

Spark

hadoop

STORM

TRIFACTA

DATATORRENT

The Data Flow Thing

Store Data

mongoDB

cassandra

APACHE HBASE

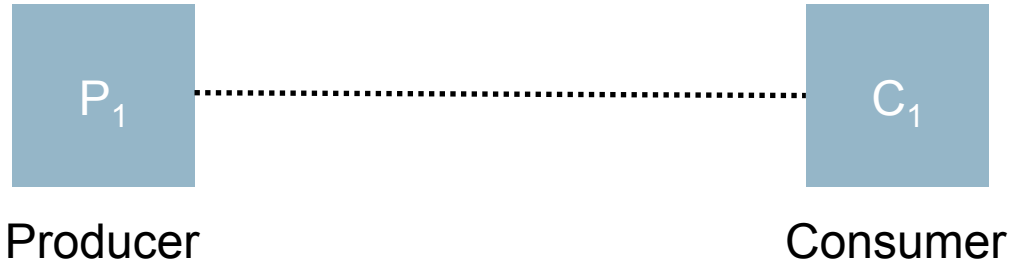
TERADATA

ORACLE

accumulo

Hortonworks

Realty of Point to Point Connections



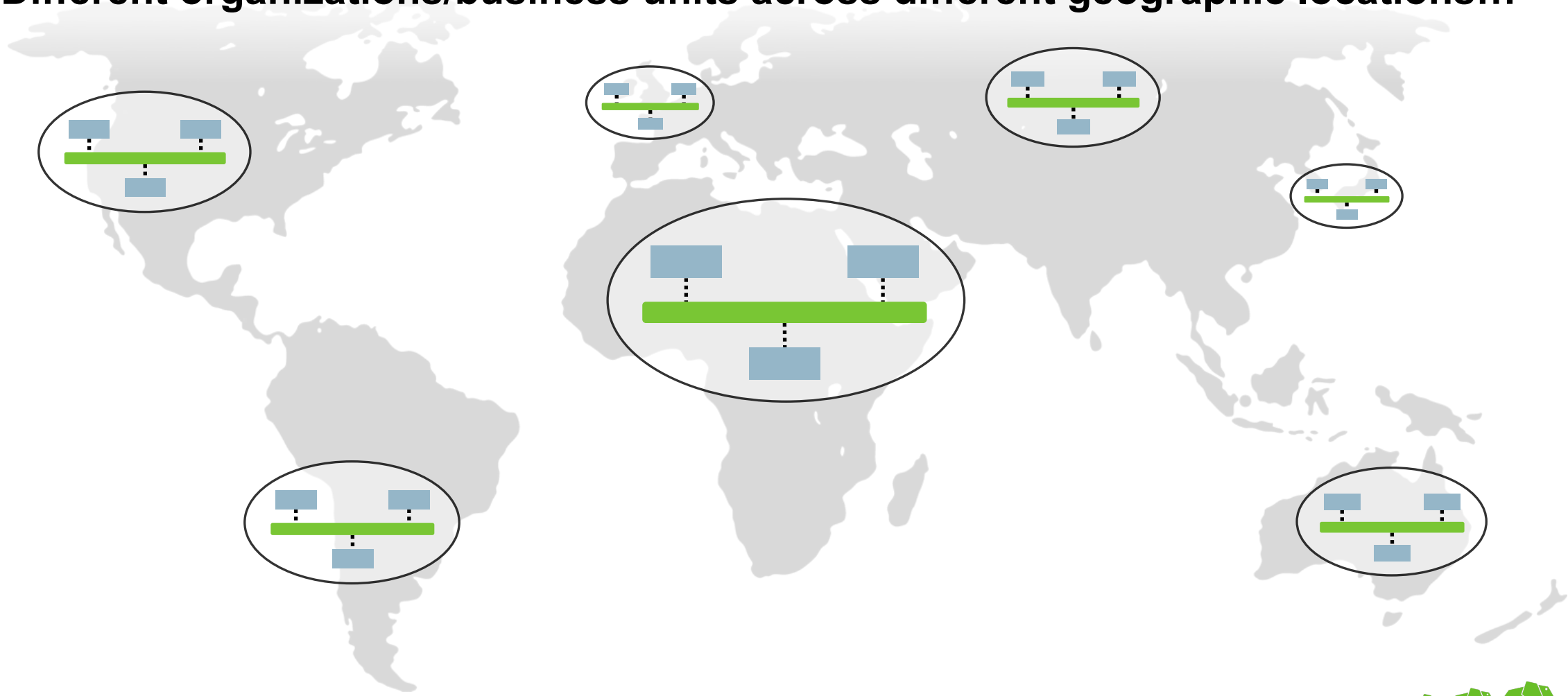
For every connection, these must agree:

1. Protocol
2. Format
3. Schema
4. Priority
5. Size of event
6. Frequency of event
7. Authorization access
8. Relevance
9. Security

Hortonworks

Realistic View of Enterprise Data Flow

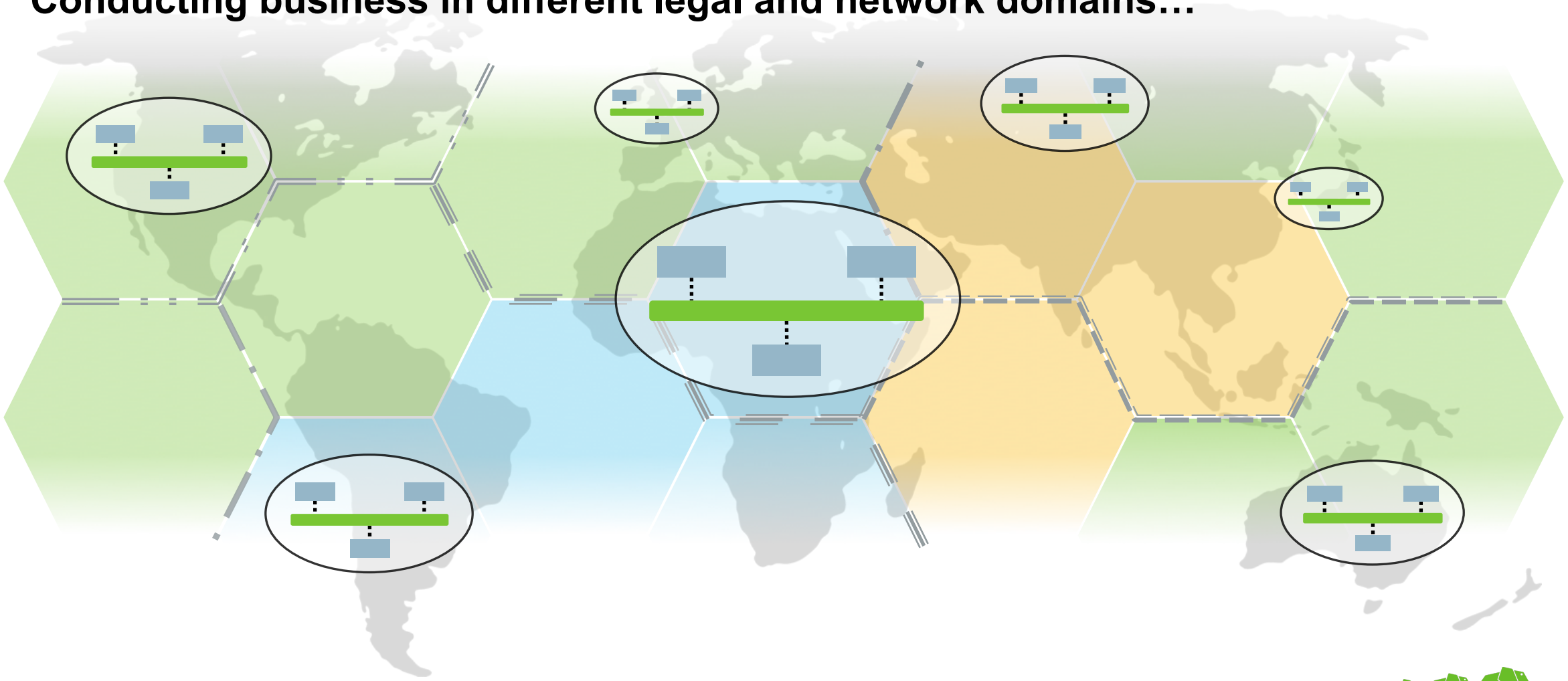
Different organizations/business units across different geographic locations…



Hortonworks

Realistic View of Enterprise Data Flow

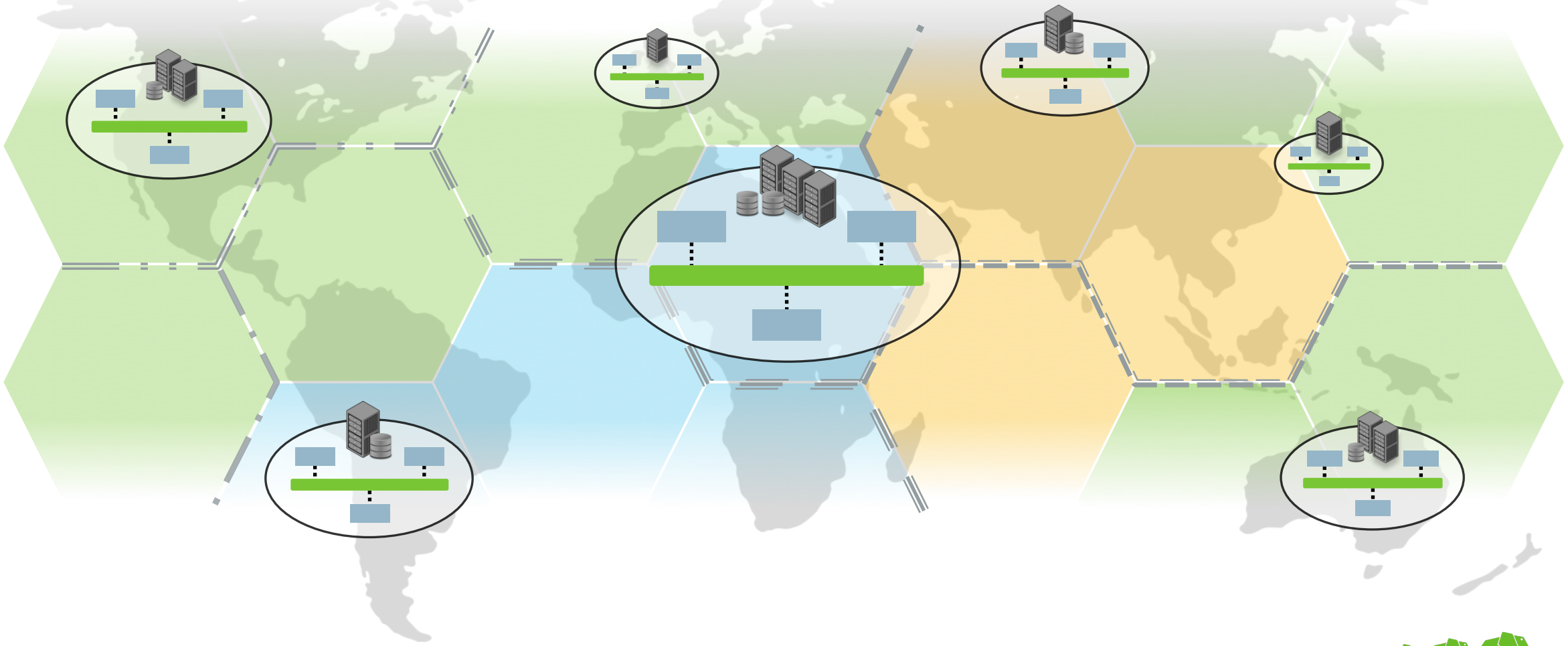
Conducting business in different legal and network domains…



Hortonworks

Realistic View of Enterprise Data Flow

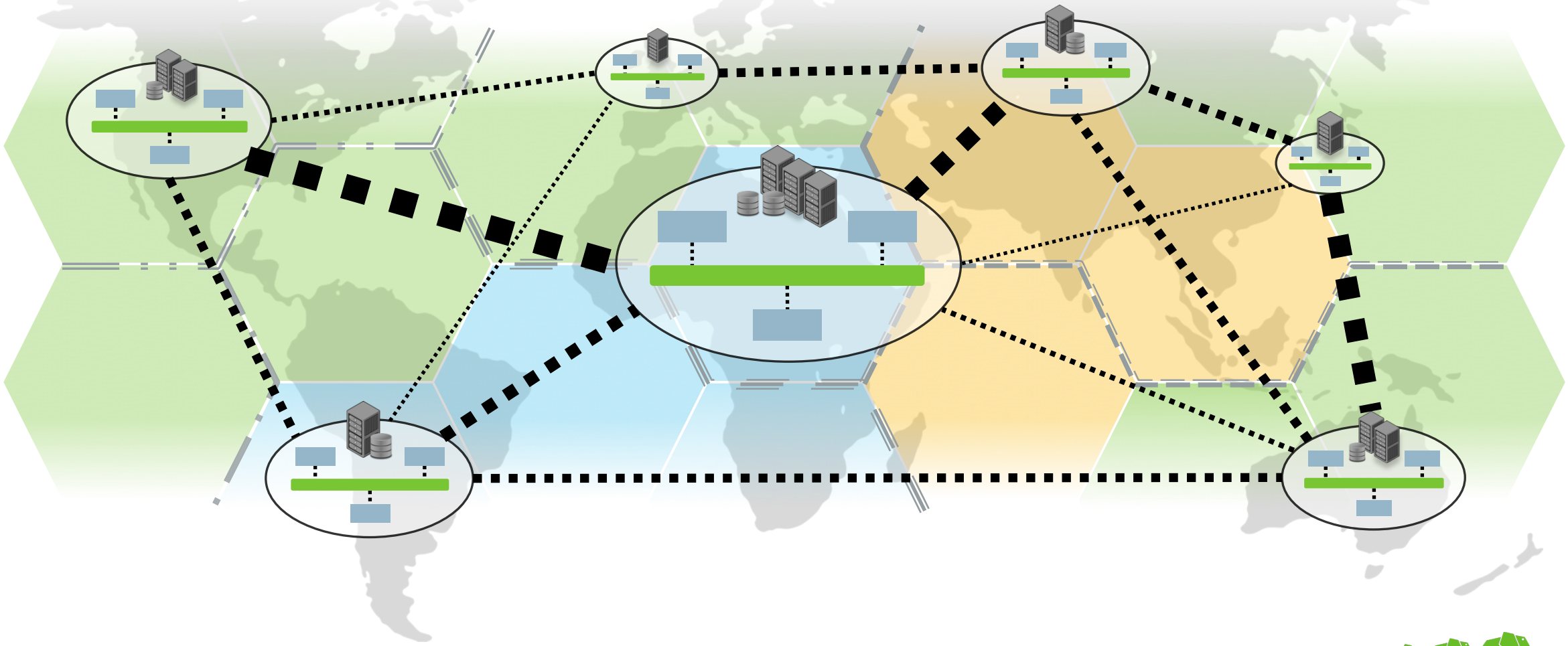
Operating on very different infrastructure (power, space, cooling)…



Hortonworks

Realistic View of Enterprise Data Flow

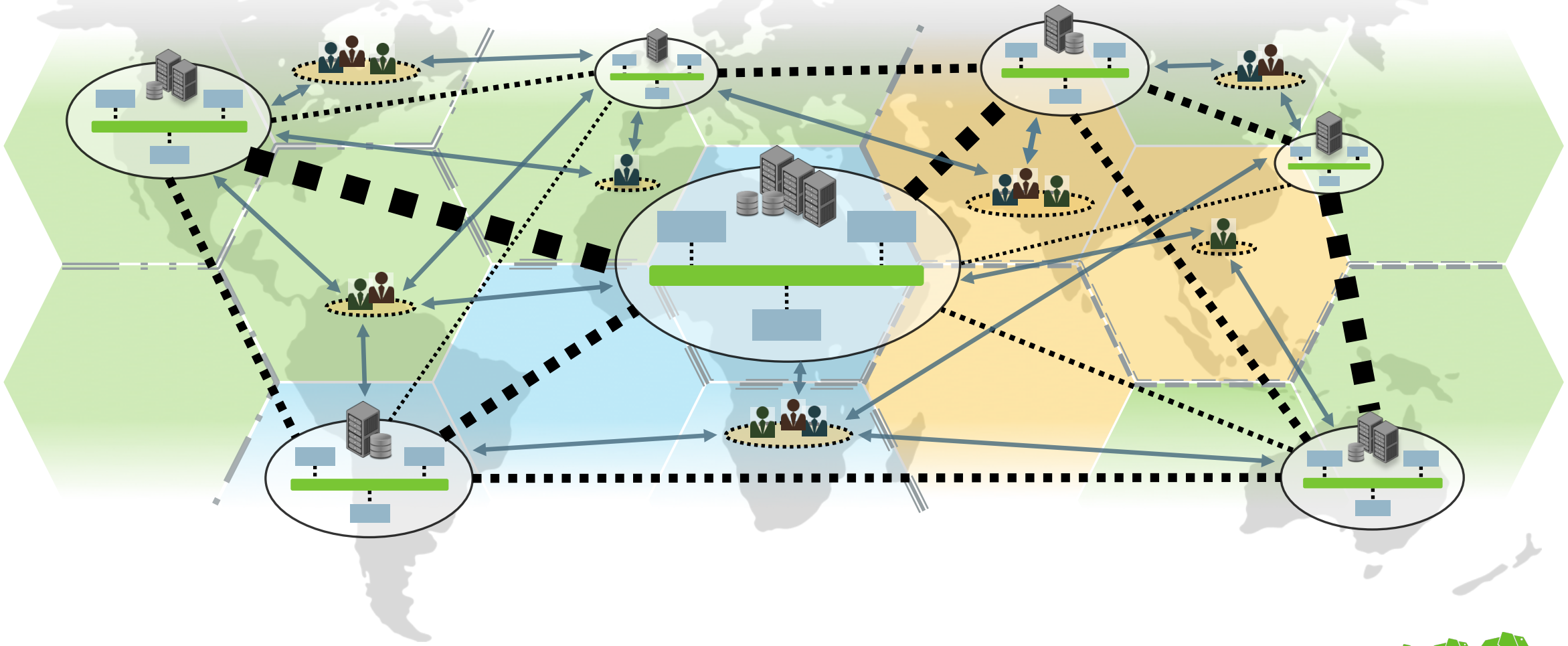
Capable of different volume, velocity, bandwidth, and latency…



Hortonworks

Realistic View of Enterprise Data Flow

Interacting with different business partners and customers



Hortonworks

The Need for Data Provenance

For Operators

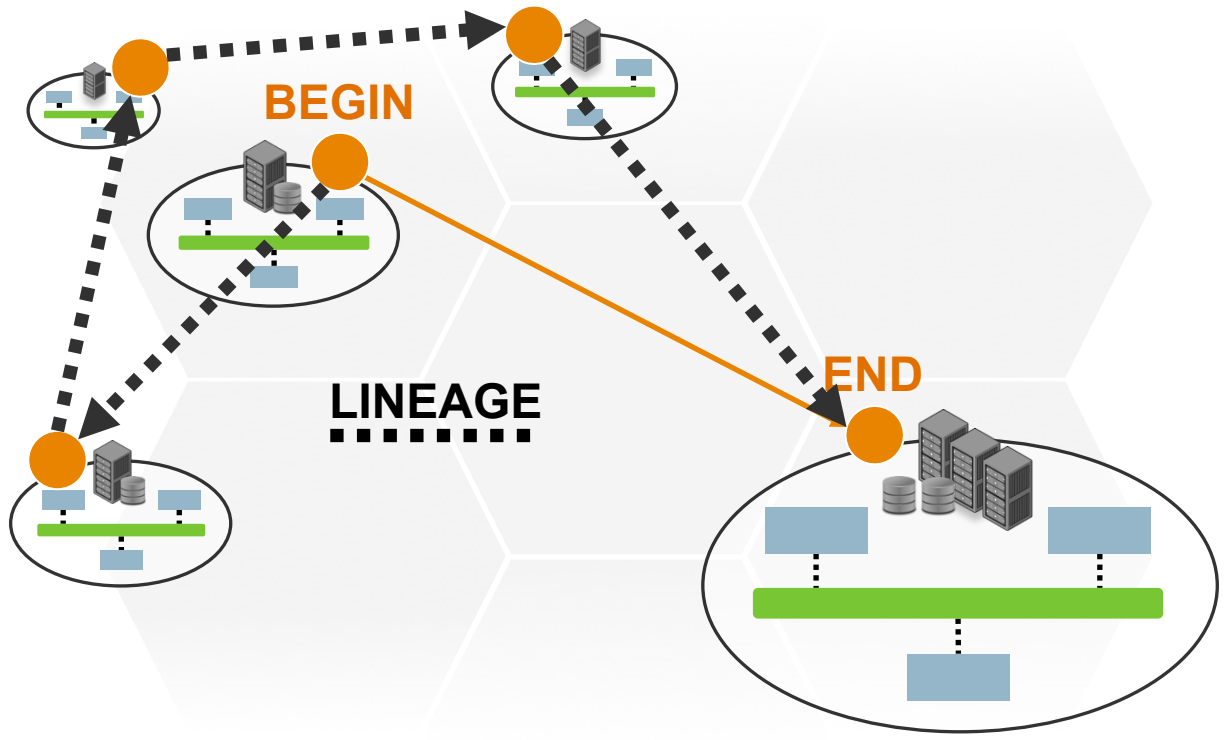
- Traceability, lineage
- Recovery and replay

For Compliance

- Audit trail
- Remediation

For Business

- Value sources
- Value IT investment

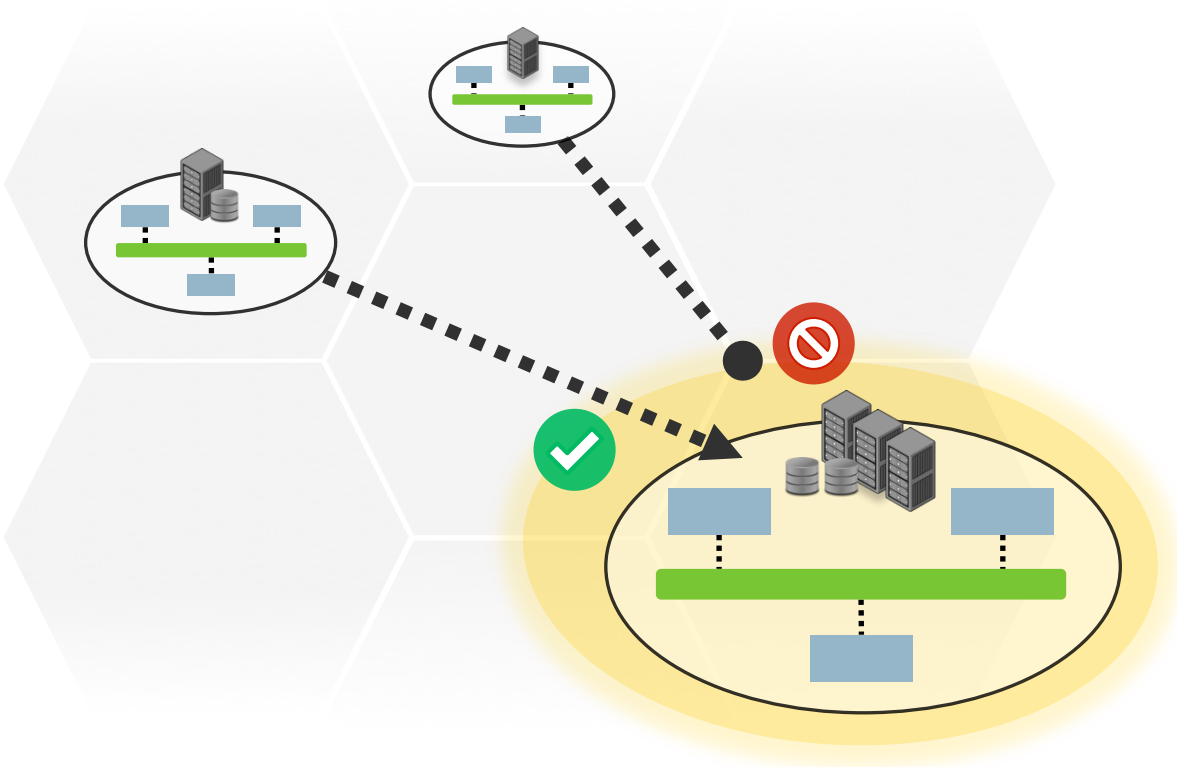


Hortonworks

The Need for Fine-grained Security and Compliance

It's not enough to say you have encrypted communications

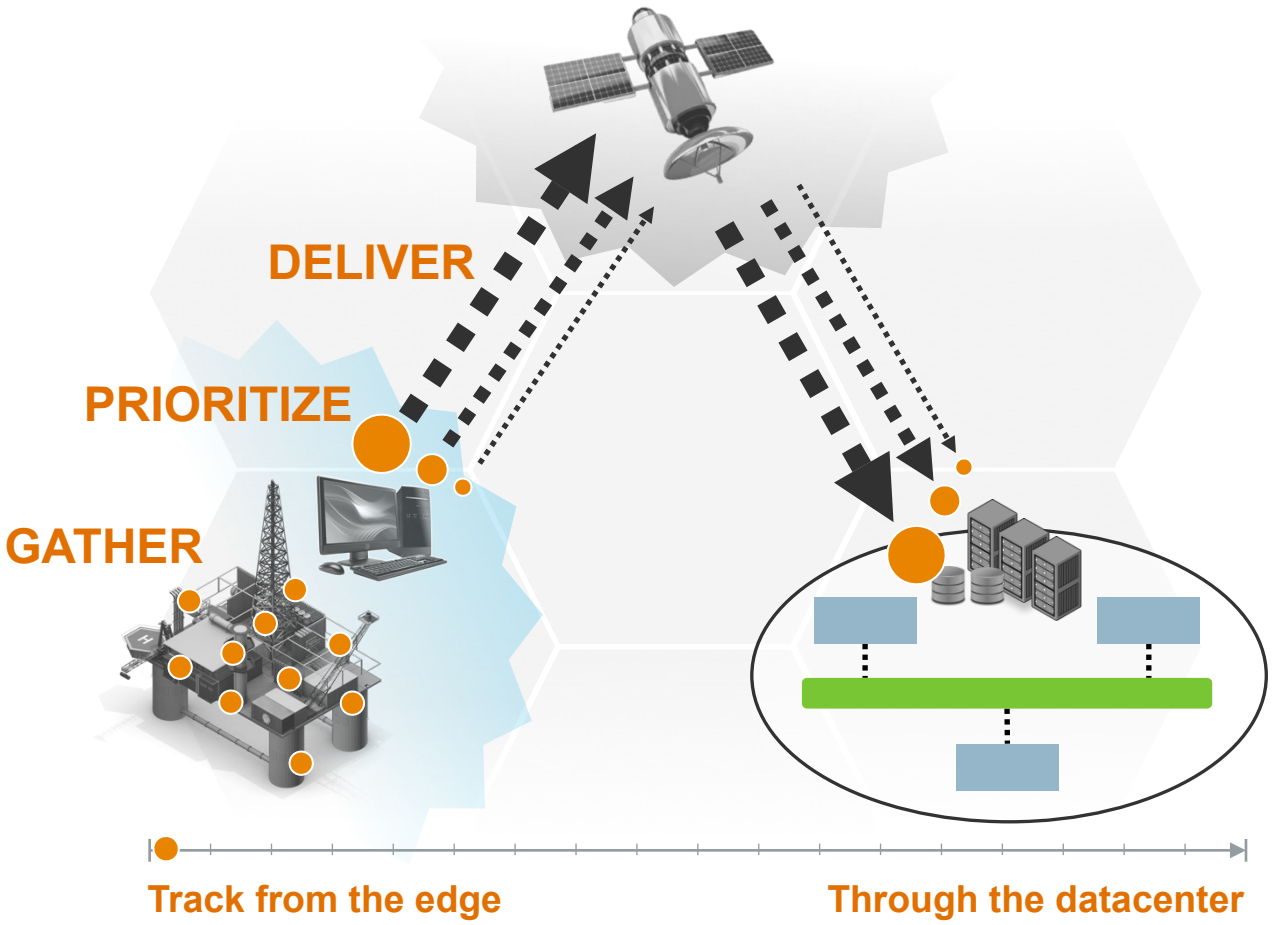
- Enterprise authorization services –entitlements change often
- People and systems with different roles require difference access levels
- Tagged/classified data



Hortonworks

Challenges at the Jagged Edge

- Small footprint
- Low power
- Expensive bandwidth
- High latency
- Access to data exceeds bandwidth (if you're doing it right)
- Needs recoverability
- Distributed Agent Model
- Needs to be secured for both the data plane and control plane

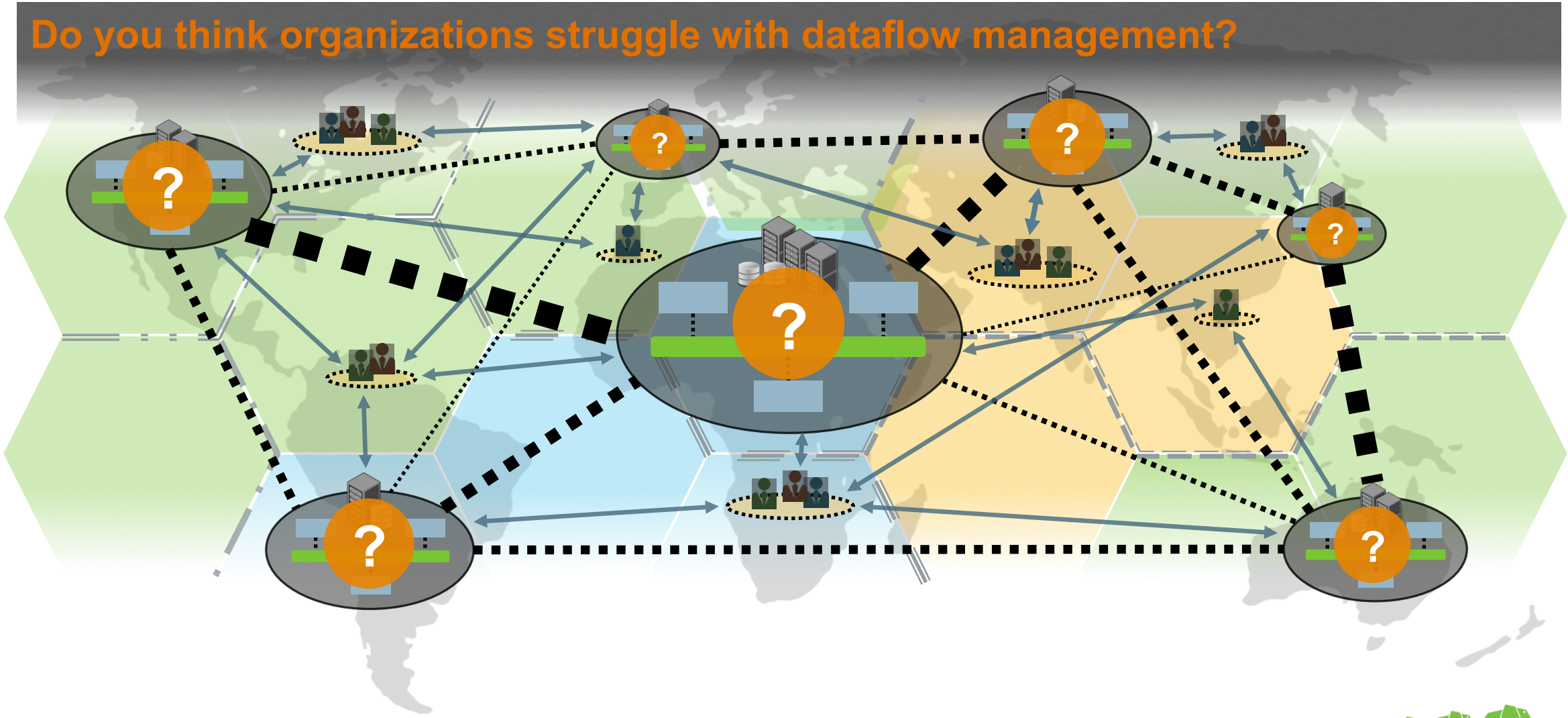


Hortonworks

Real-time Data Flow

It’s not just how quickly you move data – it’s about how quickly you can change behavior and seize new opportunities

Realistic View of Enterprise Data Flow



Hortonworks®

HDF Powered by Apache NiFi Addresses Modern Data Flow Challenges

- Logs
- Files
- Feeds
- Sensors

Collect: Bring Together

Aggregate all IoAT data from sensors, geo-location devices, machines, logs, files, 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®

NiFi Developed by the National Security Agency

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®

Proven Technology, through Onyara Acquisition

Hortonworks with Onyara's technology delivers a state-of-the-art solution for the Internet of Anything

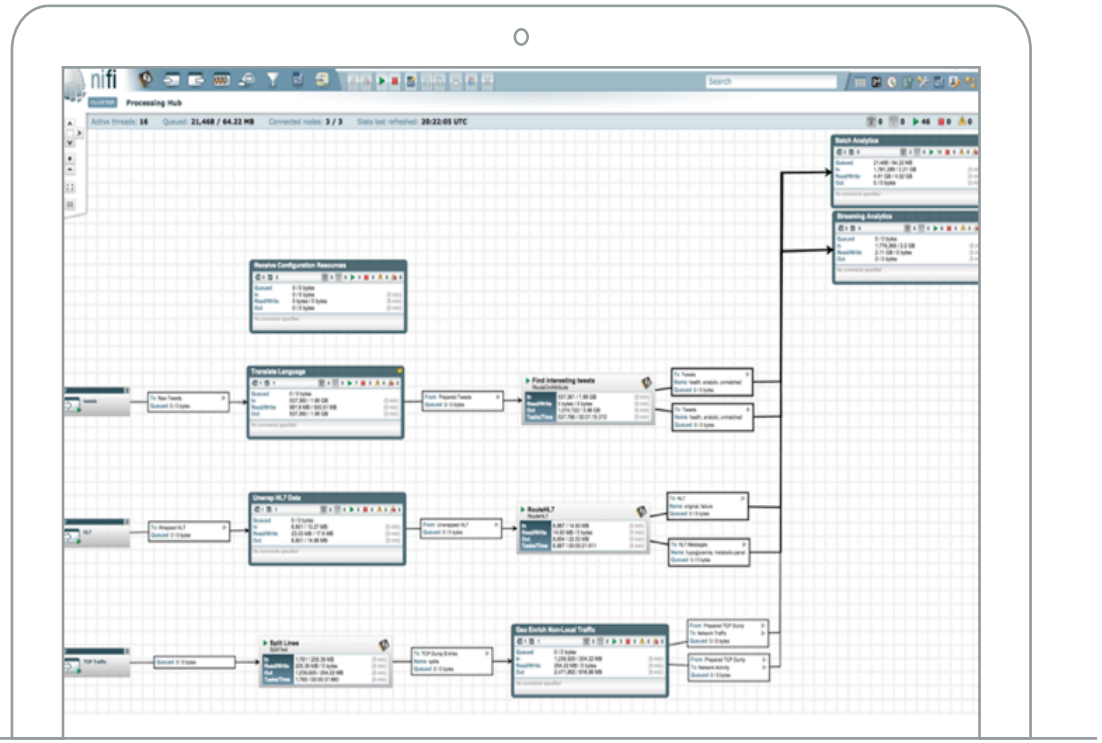
- Easier, reliable and secure data collection
- From anywhere and anything

Availability

- Hortonworks DataFlow subscription available now
- Hortonworks Data Platform subscription since 2012

Hortonworks®

Designed In Response to Real World Demands



nifi

Powered by Apache NiFi

Visual User Interface

Drag and drop for efficient, agile operations

Immediate Feedback

Start, stop, tune, replay dataflows in real-time

Adaptive to Volume and Bandwidth

Any data, big or small

Provenance Metadata

Governance, compliance & data evaluation

Secure Data Acquisition & Transport

Fine grained encryption for controlled data sharing

Hortonworks®

Apache NiFi – Key Features

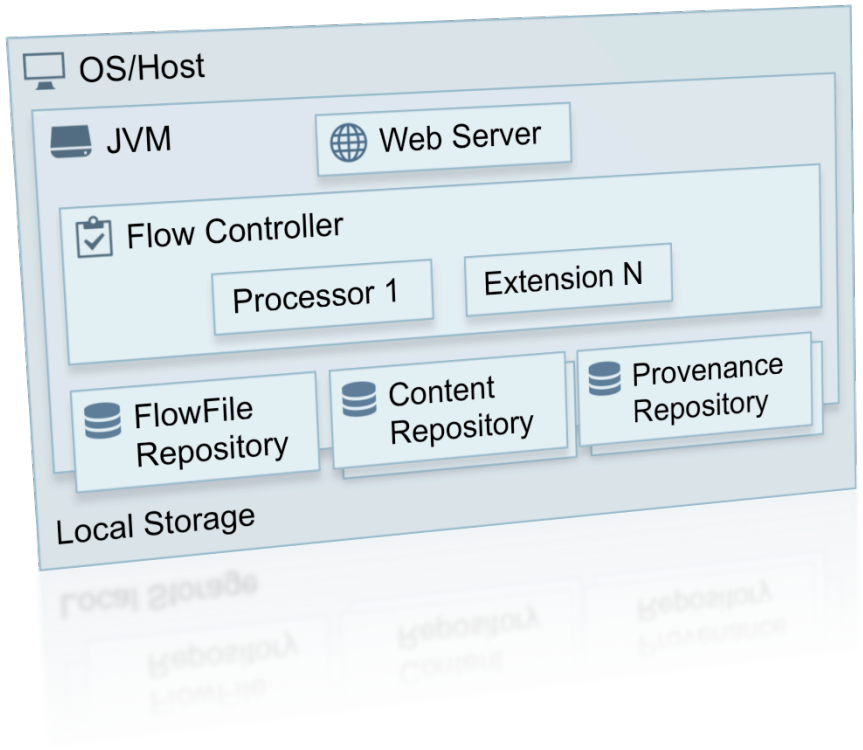
- **Guaranteed delivery**
- **Data buffering**
  - Backpressure
  - Pressure release
- **Prioritized queuing**
- **Flow specific QoS**
  - Latency vs. throughput
  - Loss tolerance
- **Data provenance**
- **Recovery/recording a rolling log of fine-grained history**
- **Visual command and control**
-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

Hortonworks

Key Concepts

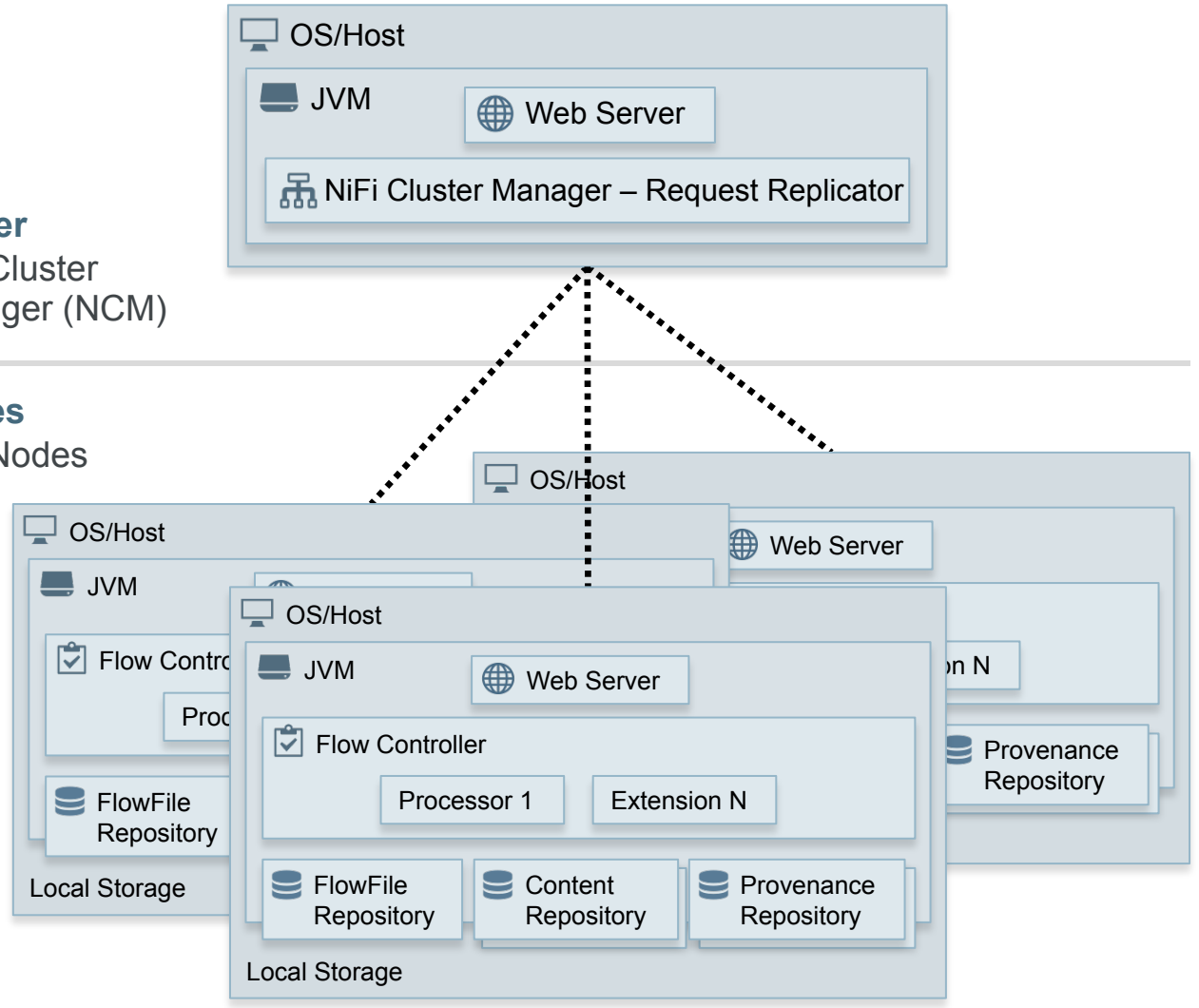
NiFi Term	FBP Term	Description
FlowFile	Information Packet	Each object moving through the system.
FlowFile Processor	Black Box	Performs the work, doing some combination of data acquisition, routing, transformation, or mediation between systems.
Connection	Bounded Buffer	The linkage between processors, acting as queues and allowing various processes to interact at differing rates.
Flow Controller	Scheduler	Maintains the knowledge of how processes are connected, and manages the threads and allocations thereof which all processes use.
Process Group	Subnet	A set of processes and their connections, which can receive and send data via ports. A process group allows creation of entirely new component simply by composition of its components.

Architecture – Supports Single and Clustered Implementation



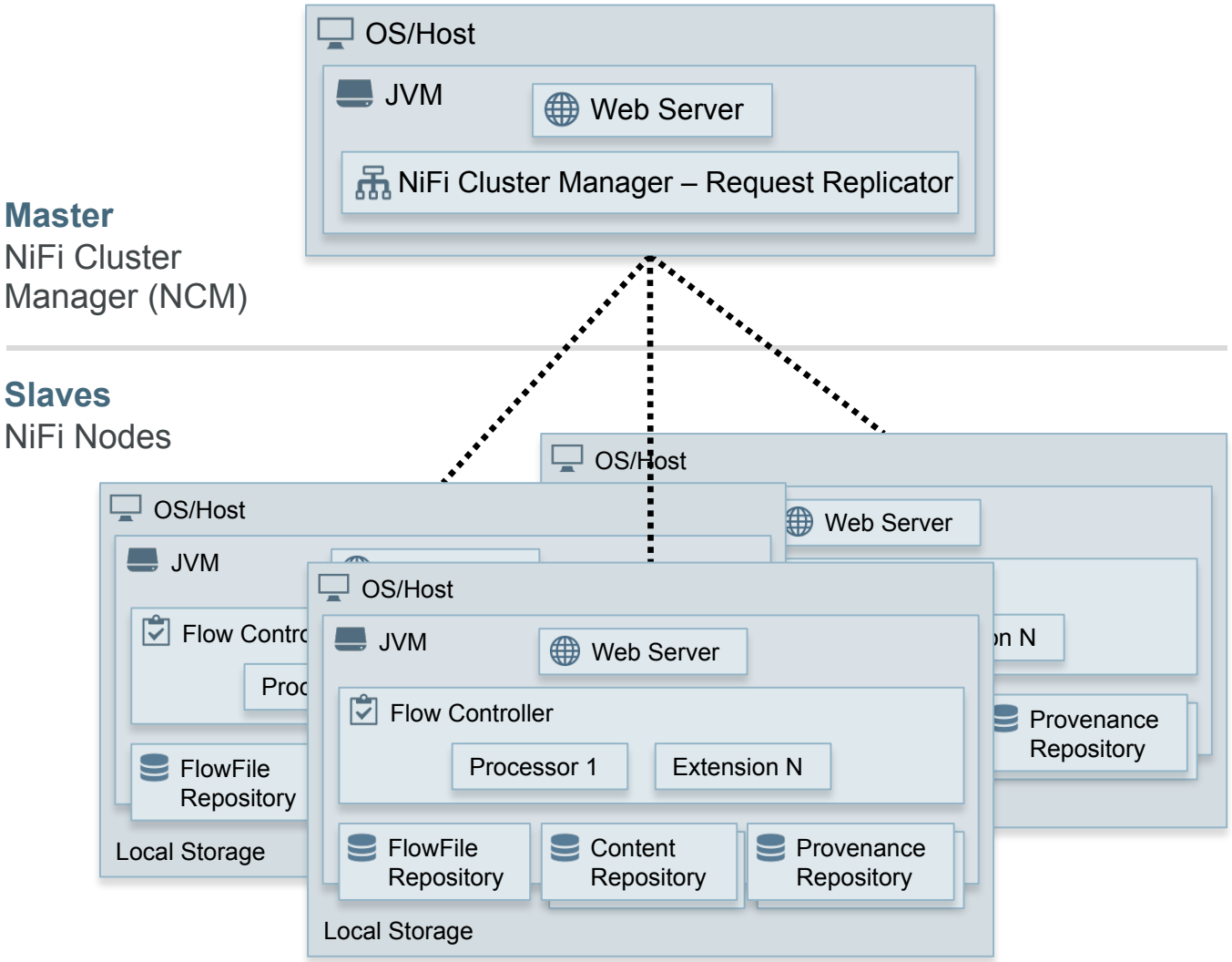
Master
NiFi Cluster Manager (NCM)

Slaves
NiFi Nodes



Hortonworks

Cluster Architecture



- **Master**
  - Tracks of which Nodes are in the cluster, their status, and to replicate requests to modify or observe the flow
- **Slaves**
  - Do the flow processing
  - Manage Local Repositories
    - FlowFile Repository – Metadata of FlowFile currently active
    - Content Repository – Contents of FlowFile
    - Provenance Repository – Archive of provenance events

Hortonworks

Security

Administration Central management and consistent security	• NiFi Cluster Manager
Authentication Authenticate users and systems	• 2-Way SSL support out of the box; additional types coming
Authorization Provision access to data	• Pluggable authorization designed to fit any Identity and Access Management (IAM) scheme • File-based authority provider out of the box • Multi-role
Audit Maintain a record of data access	• Detailed logging of all user actions • Detailed logging of key system behaviors • Data Provenance enables unparalleled tracking from the edge through the Lake
Data Protection Protect data at rest and in motion	• Support a variety of SSL/encrypted protocols • Tag and utilize tags on data for fine grained access controls • Encrypt/decrypt content using pre-shared key mechanisms

Administrator	Configure system threads, user accounts, and flow audit history
Data Flow Manager	Manipulate the dataflow
Read Only	View the dataflow only

NiFi	Configure system threads, user accounts, and flow audit history
Proxy	Manipulate the dataflow

Provenance	Query the provenance repository and download content

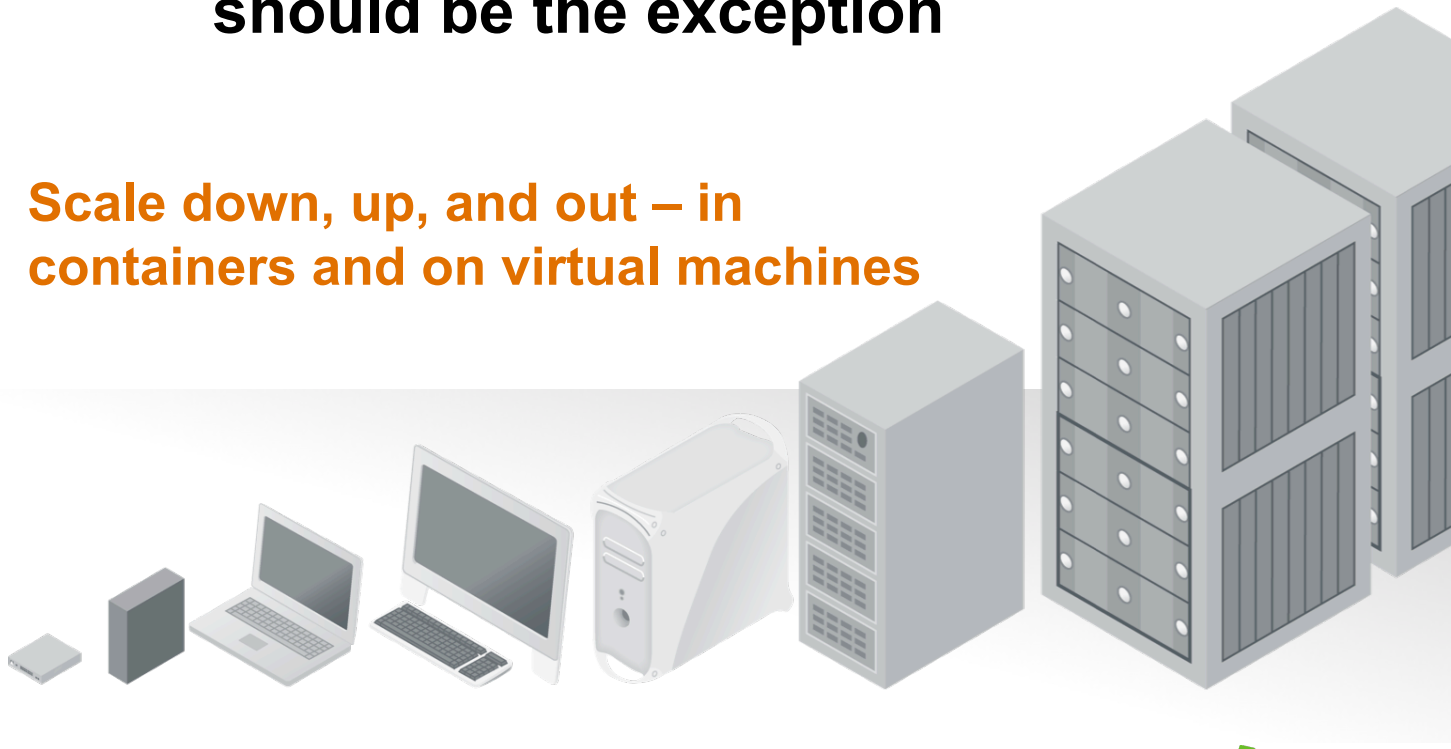
Hortonworks

Operations

- **Dynamic flow changes**
  - Push new business rules via REST API (closed loop)
  - Pull updates periodically from web services
- **Site-to-site**
  - Stay at the 'flow level' not suddenly doing file transfer protocols
- **Reporting tasks (push)**
- **Statistics / status (pull)**

- **Extensible**
- **Optimized user experience – log hunts should be the exception**

Scale down, up, and out – in containers and on virtual machines



Hortonworks

Typical HDF Sizing Scenarios: Sustained Throughput

For Sustained Throughput of 50MB/sec and thousands of events per second	For Sustained Throughput of 100MB/sec and tens of thousands of events per second	For Sustained Throughput of 200MB/sec and hundreds of thousands of events per second	For Sustained Throughput of 400-500MB/sec and hundreds of thousands of events per second
• 1-2 nodes • 8+ cores per node (more is better) • 6+ disks per node (SSD or Spinning) • 2 GB of mem per node • 1GB bonded NICs ideally	• 3-4 nodes • 8+ cores per node (more is better) • 6+ disks per node (SSD or Spinning) • 2 GB of mem per node • 1GB bonded NICs ideally	• 5-7 nodes • 24+ cores per node (effective cpus) • 12+ disks per node (SSD or spinning) • 4GB of mem per node • 10GB bonded NICs	• 7-10 nodes • 24+ cores per node (effective cpus) • 12+ disks per node (SSD or spinning) • 6GB of mem per node • 10GB bonded NICs

Hortonworks

FAQs

- **Can I use NiFi for single site deployment**
  - Yes. Data movement between machines and services in a site face many of the same challenges as data movement across sites

- **How does NiFi compare with Apache Flume**
  - NiFi supersedes Flume capabilities providing visual interface, dynamic flow changes, information provenance and advanced transformation capabilities

- **Will Flume be supported by Hortonworks**
  - Currently there are no plans to remove Flume from HDP
  - Existing Flume users are encouraged to evaluate NiFi for their usage case

Hortonworks

FAQs

- **Are NiFi and Kafka overlapping in functionality**
  - No. NiFi and Kafka are complimentary
  - NiFi focus is data acquisition, data transformation, data routing and data delivery with data provenance and replay capabilities. Kafka is a pub-sub message bus
  - NiFi can publish and consume data from Kafka
  - NiFi supports data delivery to applications using it's supported protocols, format and schema. No application change is required

- **Can I score and act on incoming events using NiFi without leveraging Storm/Spark-Streaming**
  - Yes. NiFi is a simple event processing system. It is ideal for scoring and triggering actions on events that can be analyzed using the event's data and metadata.
  - Storm/Spark-Streaming are complex event processing systems. Usage of Storm/Spark-Streaming is recommended when analysis is required over a window of events.

Hortonworks

FAQs

- **Does NiFi overlap with ETL tools**
  - No. ETL focus is bulk extract, bulk data joins, batch data quality from traditional data sources. ETL tools operate in the hadoop cluster.
  - NiFi focus is streaming and batch data ingest from new (machine data) and traditional data sources. NiFi operates outside of the hadoop cluster and is often deployed at remote sites.

Hortonworks

Labs

https://gist.github.com/abajwa-hw/e884bcc423cfbadf4a1d
Lab 1: https://github.com/abajwa-hw/ambari-nifi-service
Lab 2: https://github.com/abajwa-hw/nifi-network-processor

Hortonworks

Thank you