Objectives of the HDF Masterclass

- Understand what HDF is capable of and the use cases that it can address
- Provide you with information on the key features of the product
- Enable you to get hands on with the software building out your first data flows
- Answer any questions you have

Hortonworks®

Agenda - Morning

- Format of the masterclass
- How to access the software
- Introduction to Hortonworks Data Flow
- User Interface Orientation

Break

- *Creating & modifying a basic data flow*
- *Organising data flows and best practice*
- *Data Provenance walk through*
- Creating a routing data flow

LUNCH

Hortonworks®

Agenda - Afternoon

- Graphs and stats of dataflows
- *Modifying data flows live*
- Backpressure and pressure release
- *Linking multiple ni-fi clusters*
- Custom processor options
- *Adding a custom processor*
- Production infrastructure for HDF
- HDF vs Storm/Kafka/Flume/Spark

Break

- Roadmap
- What's next?
- Questions

Hortonworks®

The Growth of the Flow of Data

Internet of Anything

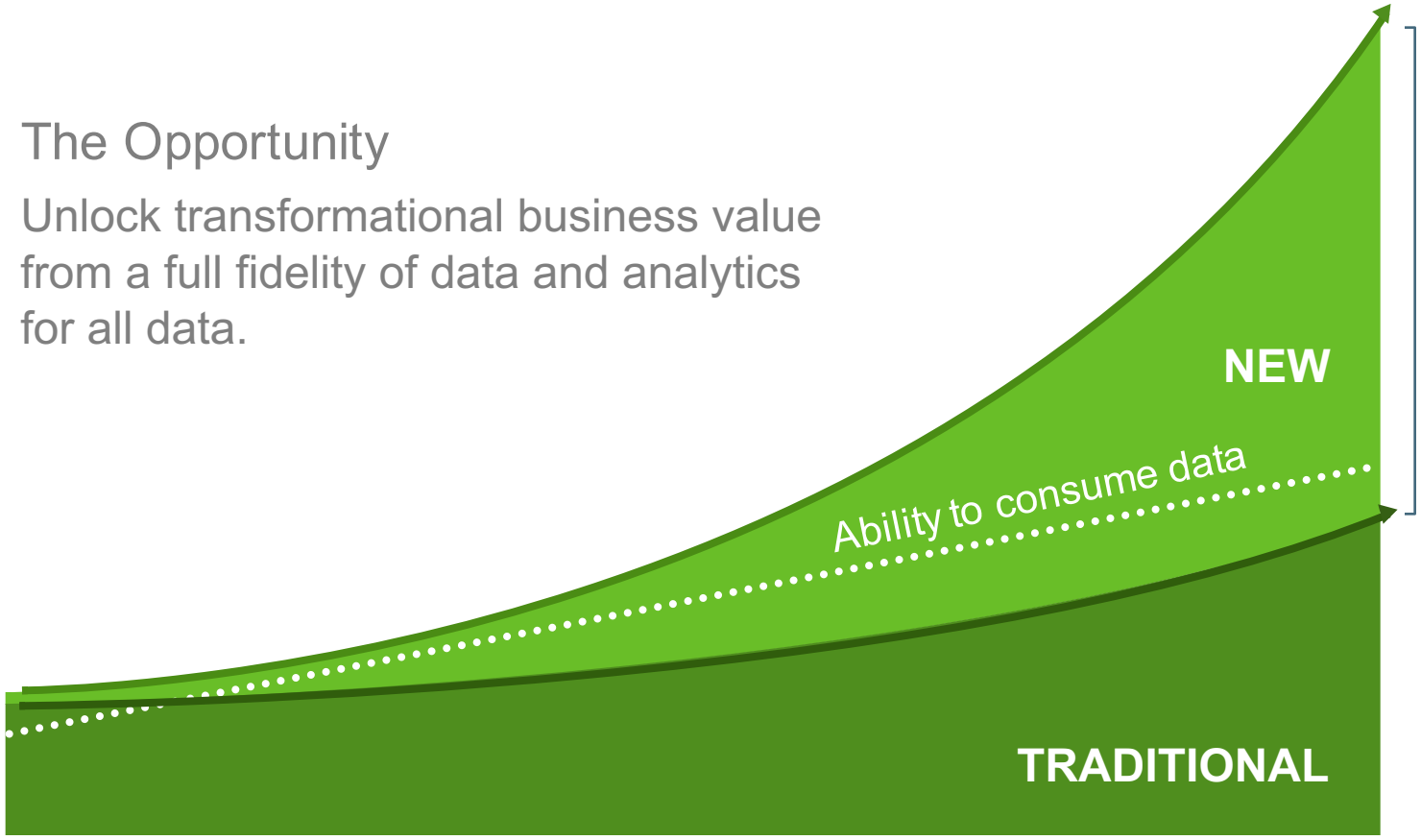
- Sensors and machines
- Geolocation
- Server logs
- Clickstream
- Web & social
- Files & emails

Traditional Data Sources

- ERP, CRM, SCM

The Opportunity

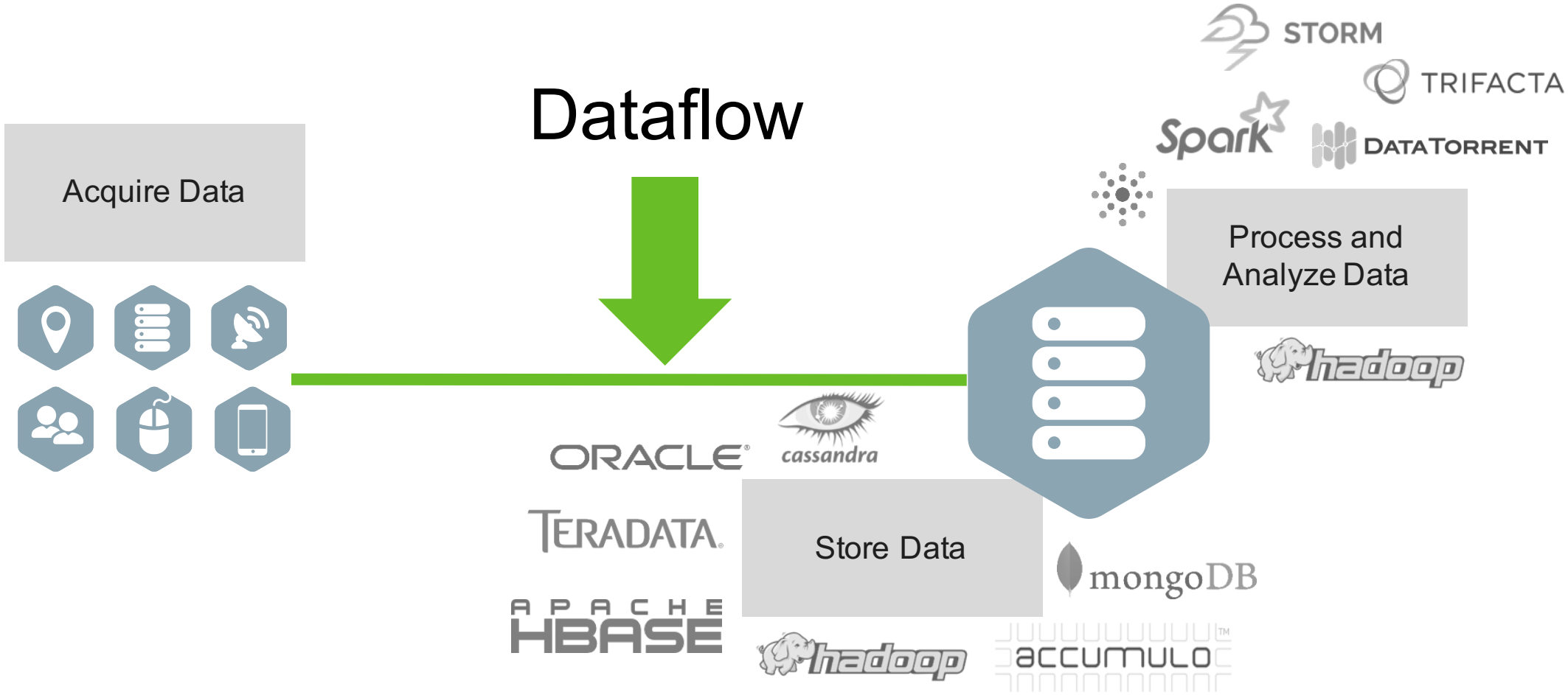
Unlock transformational business value from a full fidelity of data and analytics for all data.



Much of the new data exists in-flight between systems and devices as part of the Internet of Anything

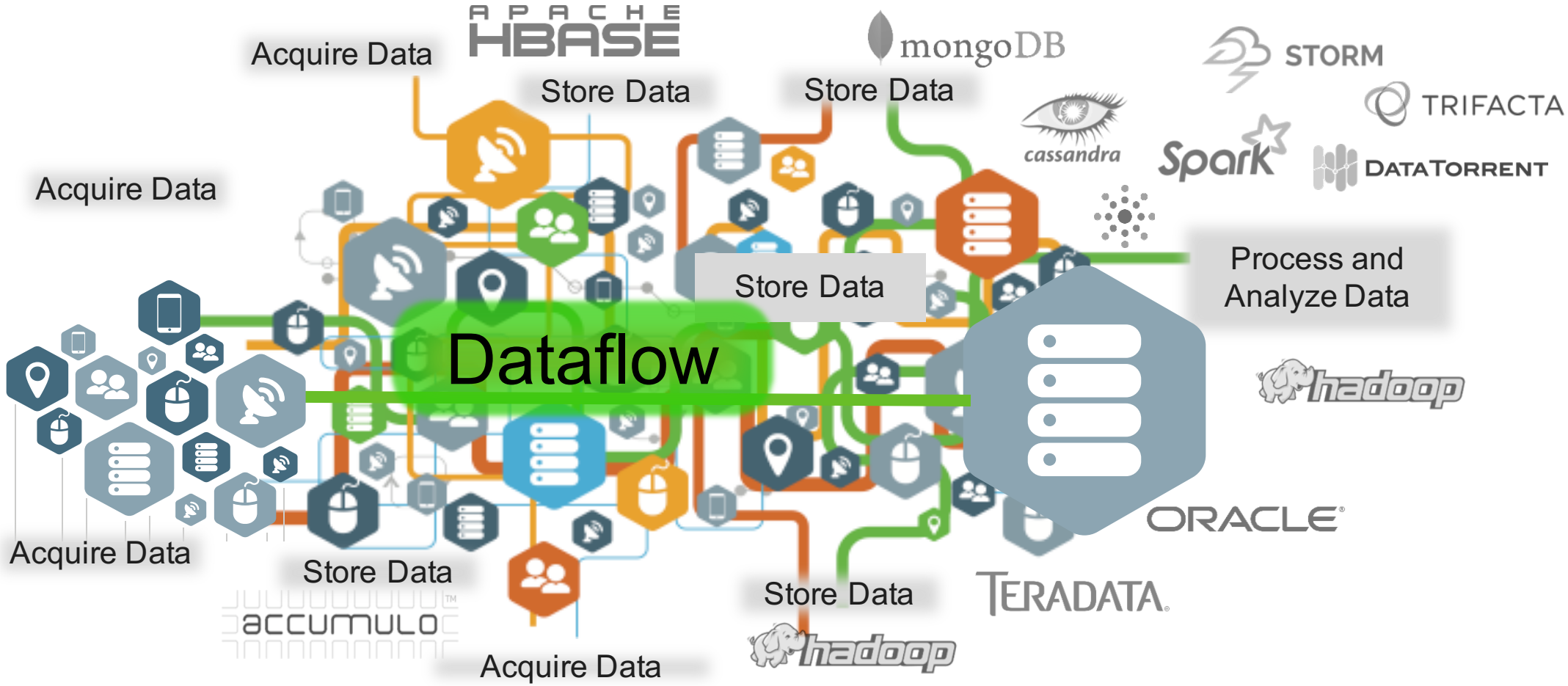
Hortonworks®

Simplistic View of DataFlows: Easy, Definitive



Hortonworks®

Realistic View of Dataflows: Complex, Convoluted



Hortonworks®

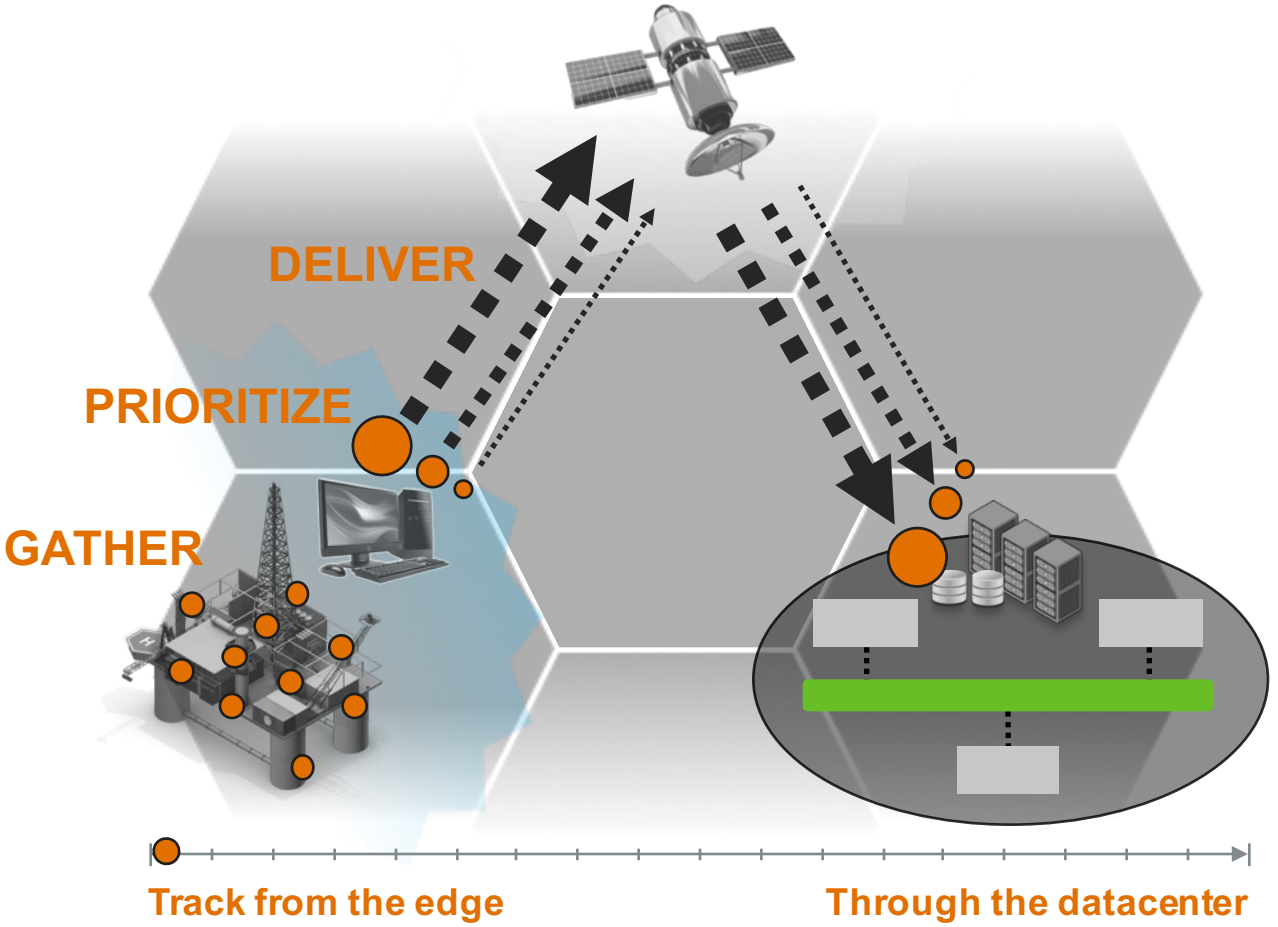
Challenges of the Internet of Anything

- Businesses need trusted, real-time insights from full-fidelity data
- Applications need access to both data in motion and data at rest
- IT needs to capture multi-directional IoAT data flows from point to point
- Challenge: the IoAT perimeter is jagged and outside the data center

Hortonworks®

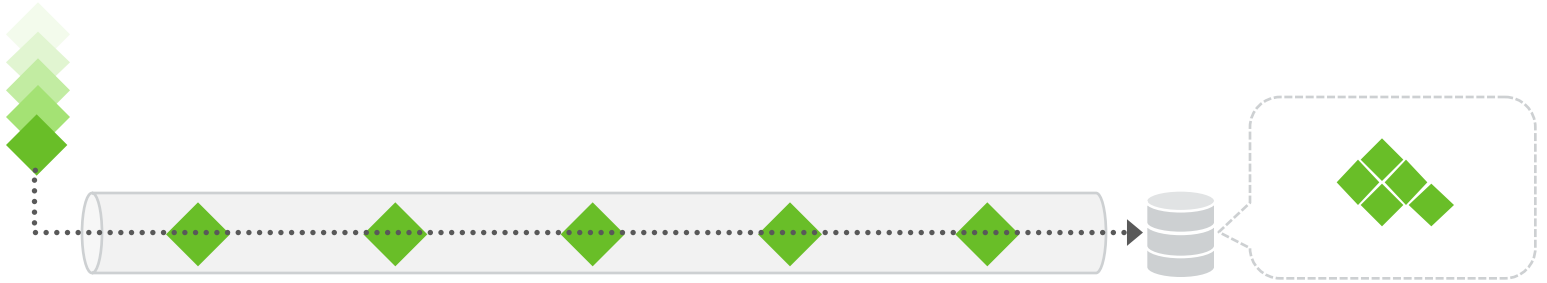
Challenges of IoAT Edges

- Data flow edges require small footprint and operate on low power
- Bandwidth is limited and high latency can be the norm
- Access to data often exceeds the bandwidth to transmit it
- Agents need recoverability capacity
- Must be secured across both the data plane and control plane

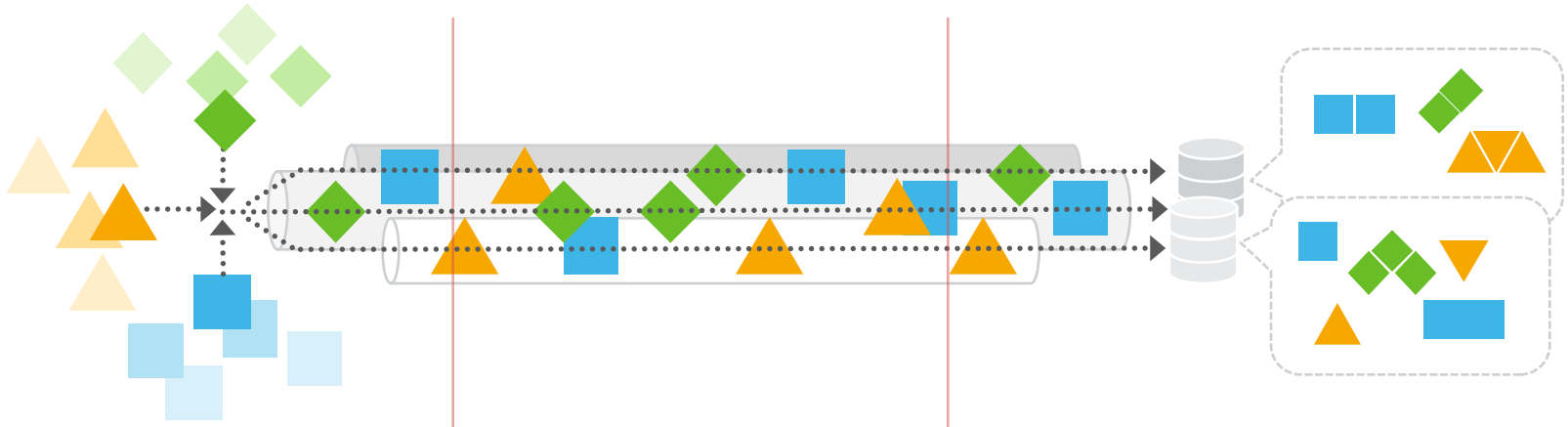


Challenges of Current Data Flow Architecture

Traditional data movement SW has been built for the world of standardized data and one way flows.



Tools built for newer types of data tend to be custom, difficult to manage, and architecturally disjoint.



This creates business problems of cost, complexity, incomplete data

Overly complicated collection systems and processes.

IoAT data flows are not optimized, or prioritized

Difficult to gain insights from the data-in-motion. .

Hortonworks®

Why Hortonworks DataFlow

Because even the best data scientists and most powerful platforms need the right data to analyze

Hortonworks®

Typical Answer to Challenges

Add new systems to handle the protocol differences

Add new systems to convert the data

Add new systems to reorder the data

Add new systems to filter the unauthorized data

Add new system to slow down or speed up data

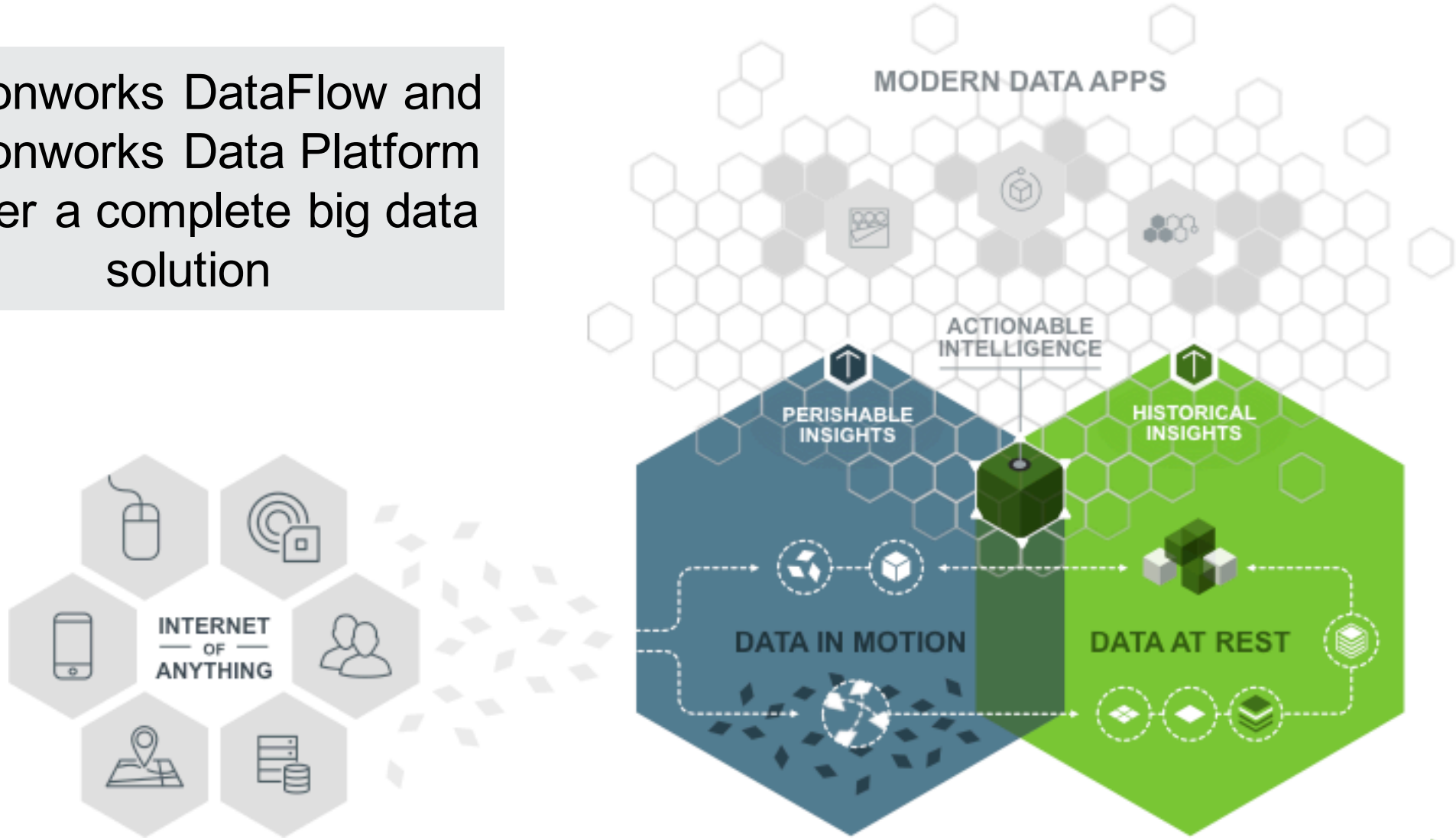
Add new topics to represent 'stages of the flow'

Added Complexity!

Hortonworks®

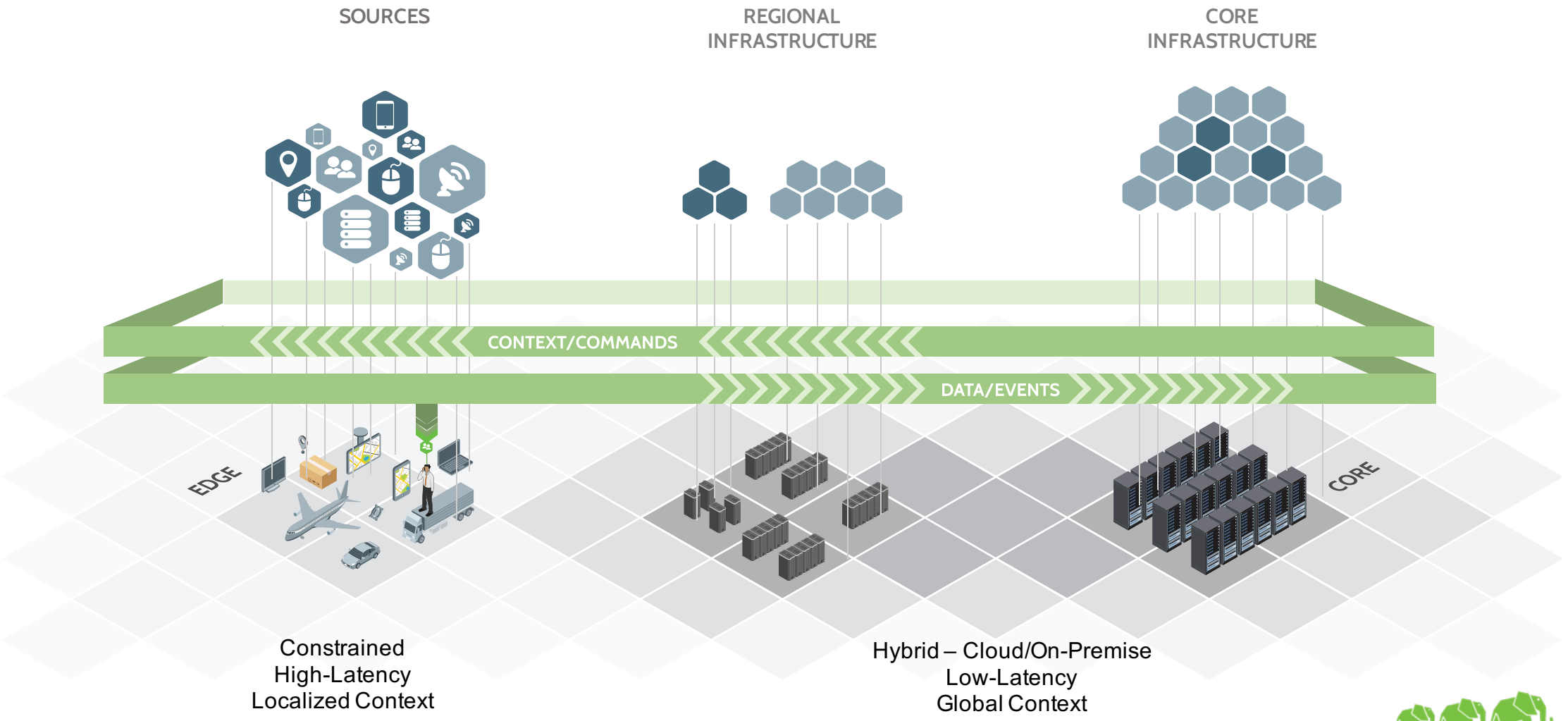
Part of Hortonworks Connected Data Platforms

Hortonworks DataFlow and Hortonworks Data Platform deliver a complete big data solution



Hortonworks®

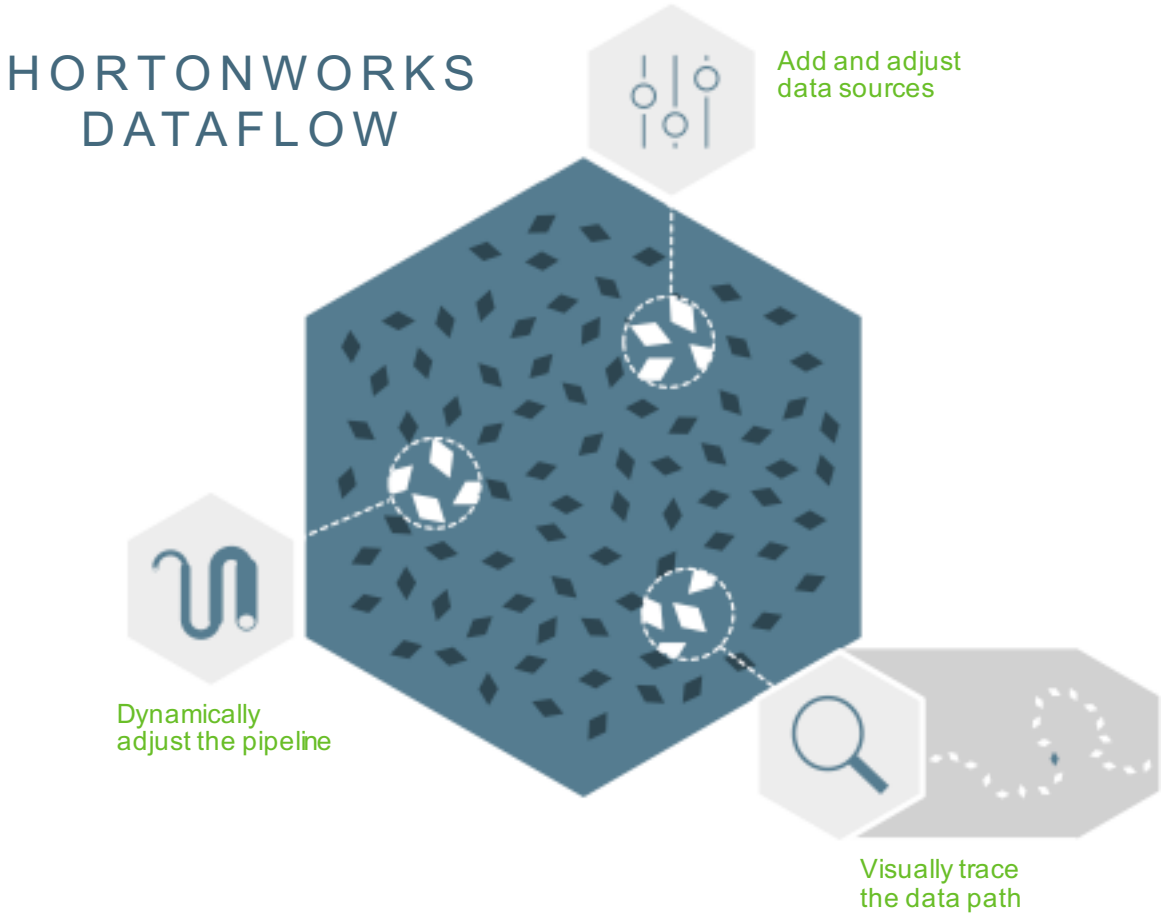
HDF Manages Dataflow



About Hortonworks DataFlow

- Immediate operational visibility and control
- Accelerates big data ROI
- Secure collection from the Internet of Any Thing
- Scalability from enterprise data centers to sensors
- Enables real time decisions from streaming data

Real-Time, Visual, Interactive Control of Data Flows



Add and Adjust Data Sources

to maximize the opportunity that you capture from perishable insights

Visually Trace the Data Path

to manage the what, who, where and how around data in motion

Dynamically Adjust the Pipeline

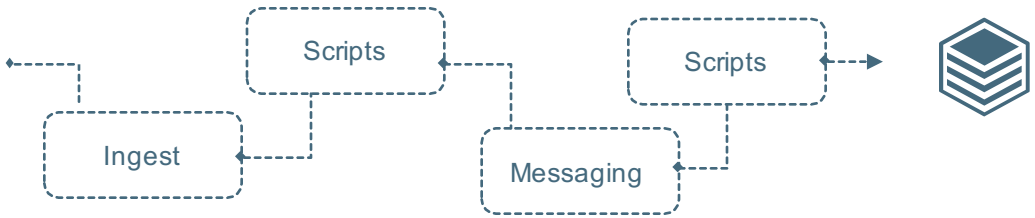
to match the dataflow with your bandwidth

Hortonworks®

Integrated Processes and Control

COMMON ARCHITECTURE

WITHOUT HORTONWORKS DATAFLOW



WITH HORTONWORKS DATAFLOW

Optimize Your Architecture

Reduce cost and complexity with the most efficient data collection technologies

Assure Efficient Operations

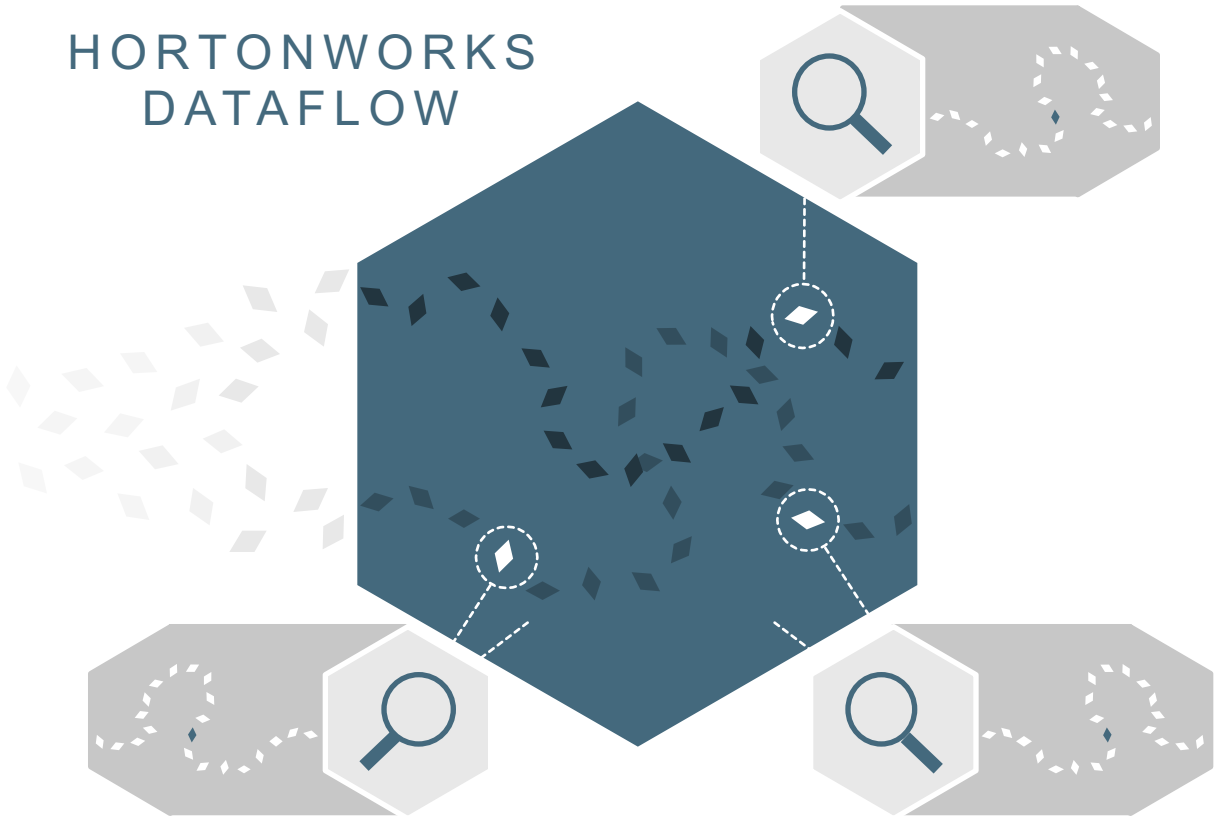
Via real-time control of data inputs, outputs, transportation and transformations

Rely on a Common Foundation

Eliminating dependence on multiple customized systems

Hortonworks®

Secure, Track, and Troubleshoot Dataflows



End-to-end Security

Apply security rules from the point of collection at the jagged edge to the data's final destination

Granular Control and Sharing

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

Real-time Traceability and Event-Level Provenance

Rich event-level metadata and contextual detail help troubleshoot security issues and inform timely decisions

Hortonworks®

Adapt to a Broad Range of Data Flow Demands

Automated

Bi-directional communication between source and destination adapts data flows automatically, according to current priorities

On-demand

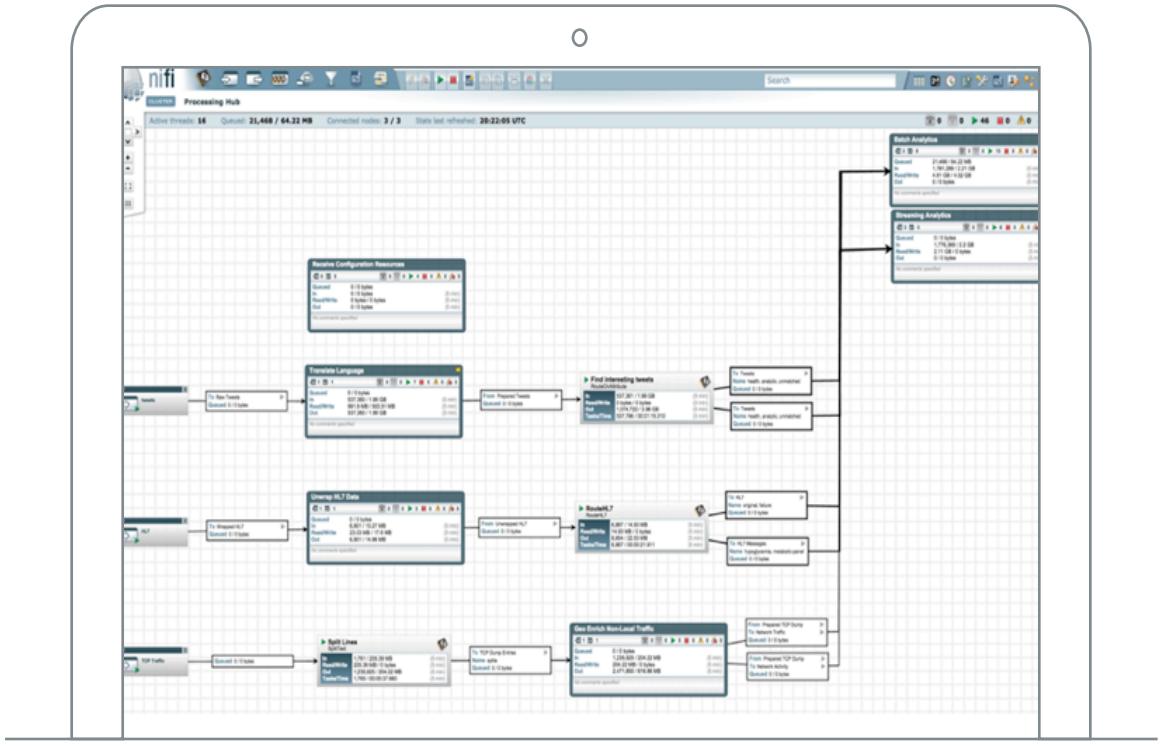
Operational control to adapt to changing conditions and requirements

Scalable

By incorporating data from any device—small machine sensors to enterprise data centers—HDF connects you to the broadest set of disparate data sources

Hortonworks®

Operational Effectiveness Proven at Scale



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

Event Level Data Provenance

Governance, compliance & data evaluation

Secure Data Acquisition & Transport

Fine grained encryption for controlled data sharing and selective data democratization

Hortonworks®

NiFi Developed by the National Security Agency

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

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,"

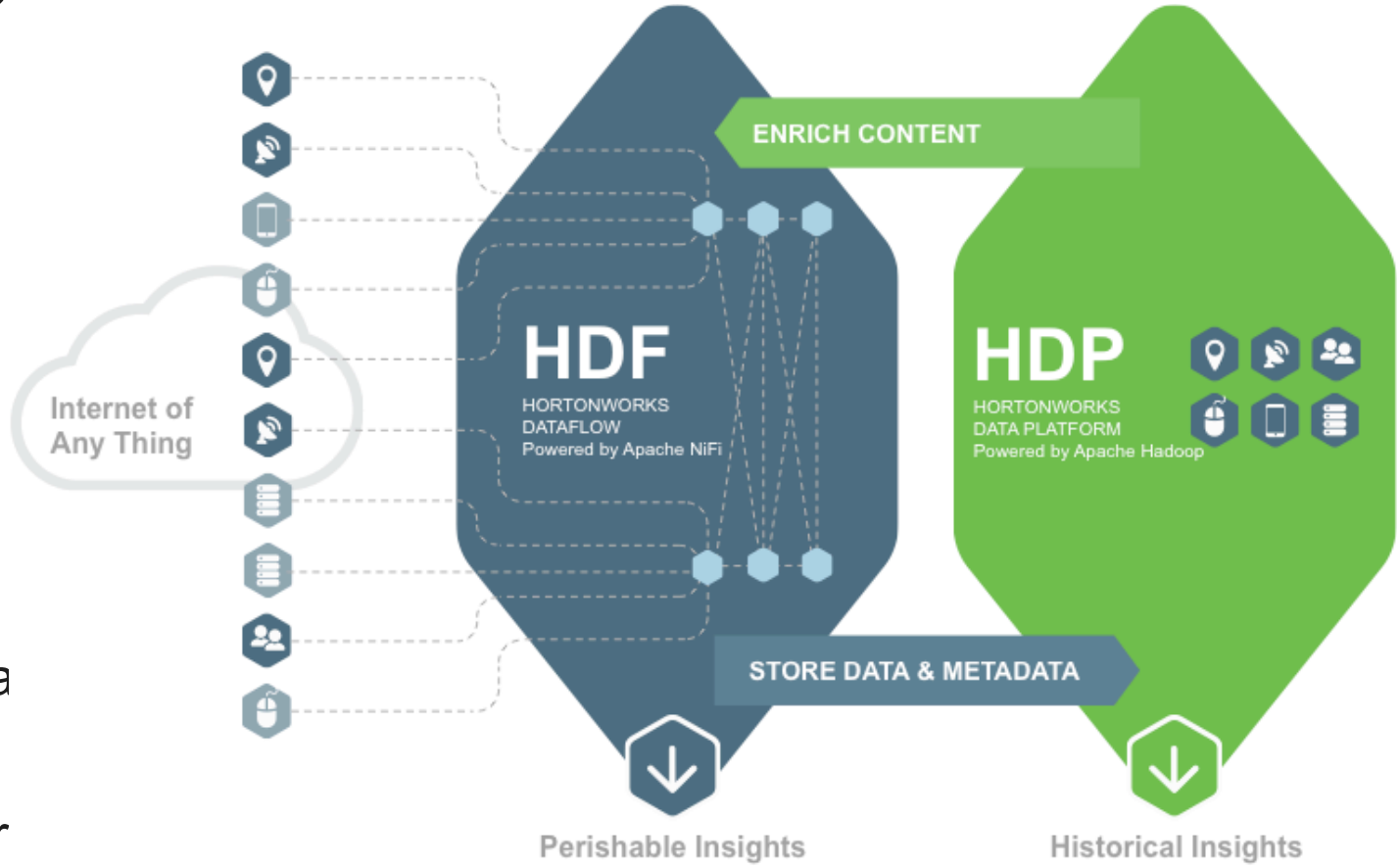
"Commercial enterprises could use it to quickly control, manage, and analyze the flow of information from geographically dispersed sites – creating comprehensive situational awareness"

-- Linda L. Burger, Director of the NSA

Hortonworks®

HDF and HDP Deliver a Complete Big Data Solution

- HDF dynamically connects HDP to data at the edge
- HDF secures and encrypts the movement of data into HDP
- HDF includes mature IoAT data protocols that improve device extensibility
- HDF supports easily adjustable bi-direction IoAT dataflows
- HDF offers traceability of IoAT data with lineage and audit trails
- HDF brings a real-time, visual user interface to manipulate live dataflows



Hortonworks®

Transform your Business with Hortonworks DataFlow

- Faster ROI from accelerated big data pipeline ingest and operational simplicity
- Secure, highly discrete data sharing and unprecedented chain-of-custody
- Prioritized, secure and reliable data collection from high output, low bandwidth environments

Hortonworks®

HDF Use Cases

Optimize Splunk:

Reduce costs by pre-filtering data so that only relevant content is forwarded into Splunk

Ingest Logs for Cyber Security:

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

Feed Data to Streaming Analytics:

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

Move Data Internally:

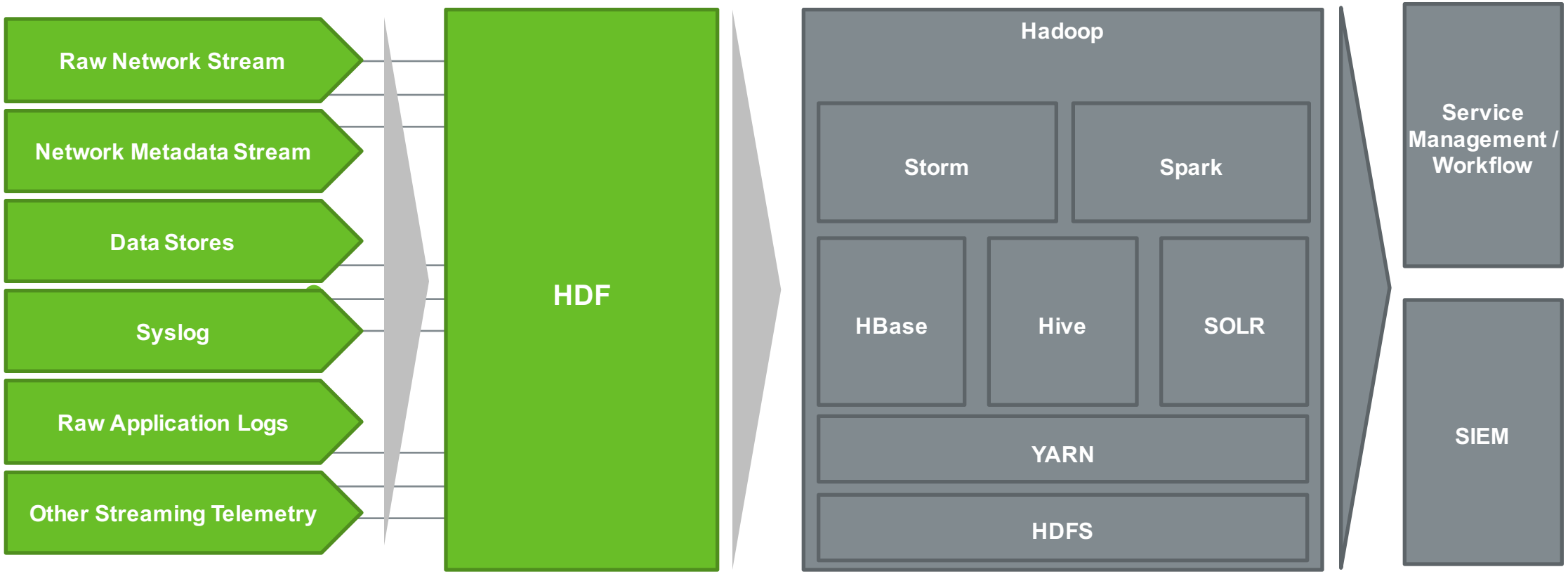
Optimize resource utilization by moving data between data centers or between on-premises infrastructure and cloud infrastructure

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®

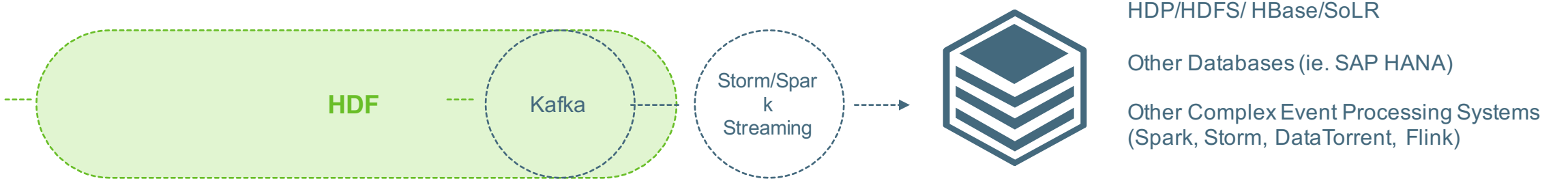
Cybersecurity Log Ingestion



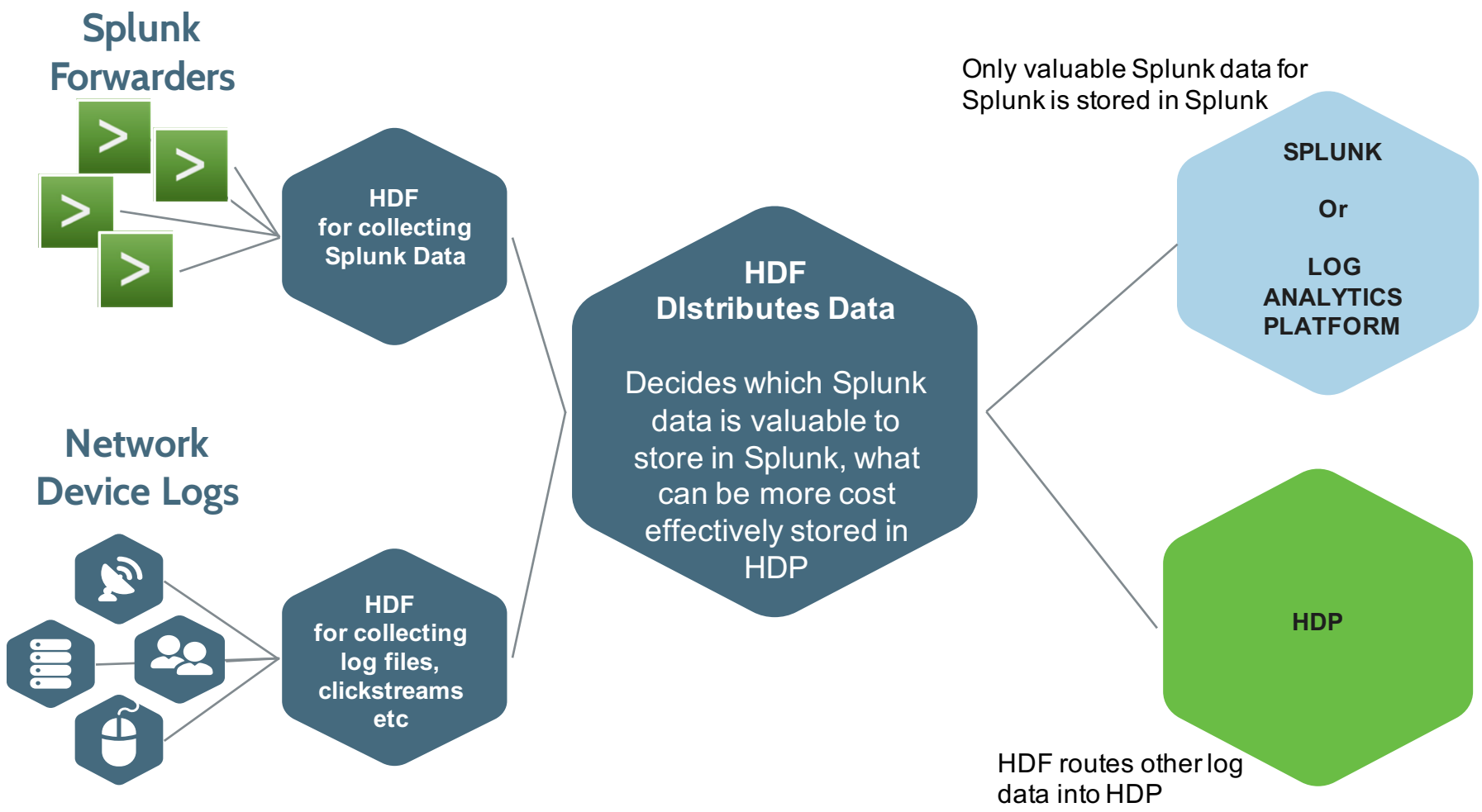
Feed Streaming Analytics

WITHOUT HORTONWORKS DATAFLOW

WITH HORTONWORKS DATAFLOW



Log (i.e Splunk) Optimization



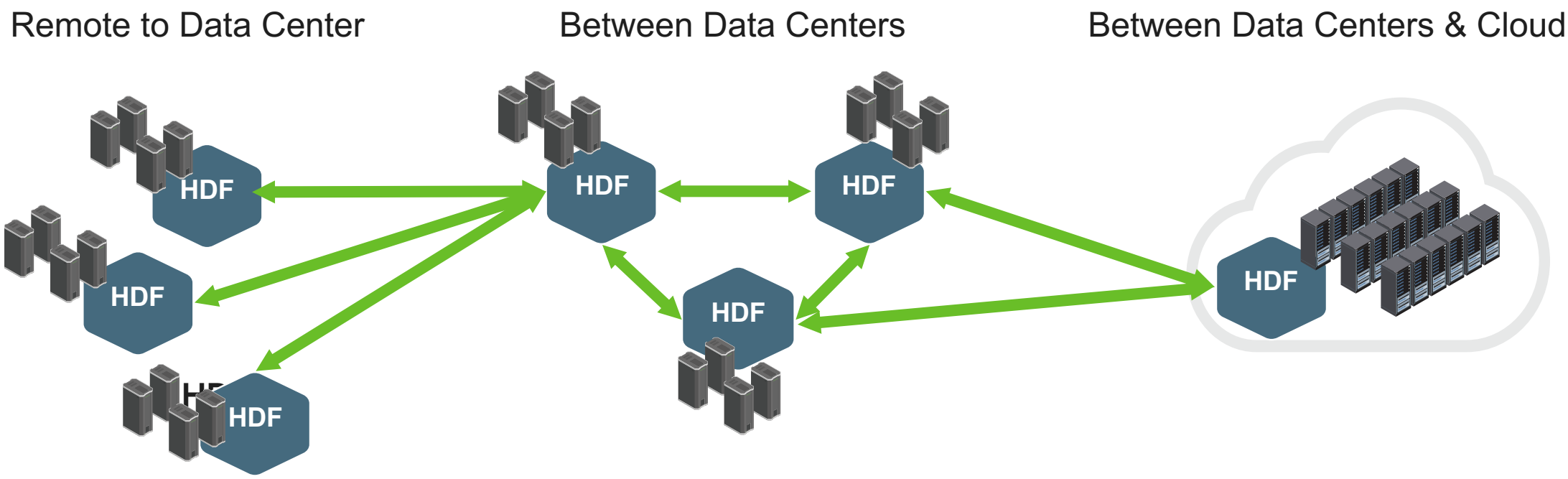
Benefits of HDF

- Re-Use Current Log Forwarders
- Secure data transfer via Site to Site
- Any Size or Speed of Log Event
- Bi-Directional Flow of Data
- Prioritize data at the Edge
- Data Provenance or Audit Trail
- Enrichment and Transformation at the Edge

Hortonworks®

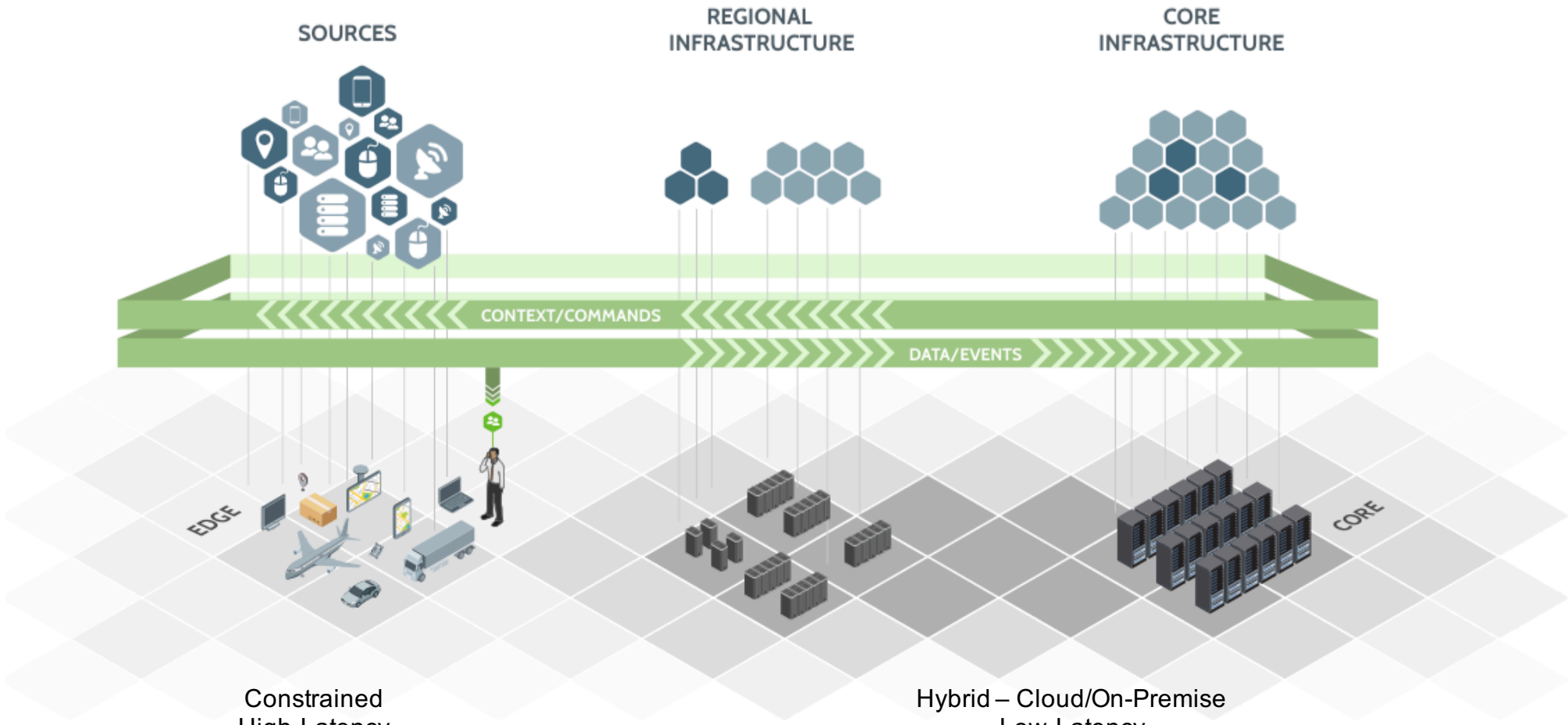
Enterprise Data Movement

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



Hortonworks®

IoT Data Collection and Transport



Constrained
High-Latency
Localized Context

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

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

Hortonworks®

Industry Specific Data Ingest Use Cases

1. CDR Ingest for Telcos
2. Mainframe data into HDFS
3. 340B Prescription Program
4. Fraud Loss Prevention- Gift Cards and merchandise
5. Real time inventory for e-commerce sites and for supply chain/replenishment
6. Social Media ingest
7. Connected Car
8. Gateway for pushing data from a site to multiple vendors that want to use
9. IoT- Condition Monitoring- Control Valve, Steam Traps
10. Facilities monitoring- proactive, predictive and condition monitoring for planned downtime
11. Real time movement of transport vehicles, planes, trains for route optimization
12. Fuel Optimization and Credit Card for airlines
13. Smart City, lighting, parking, cameras

Contact your industry marketing representative for support

What we setup for you

- Single HDF instance

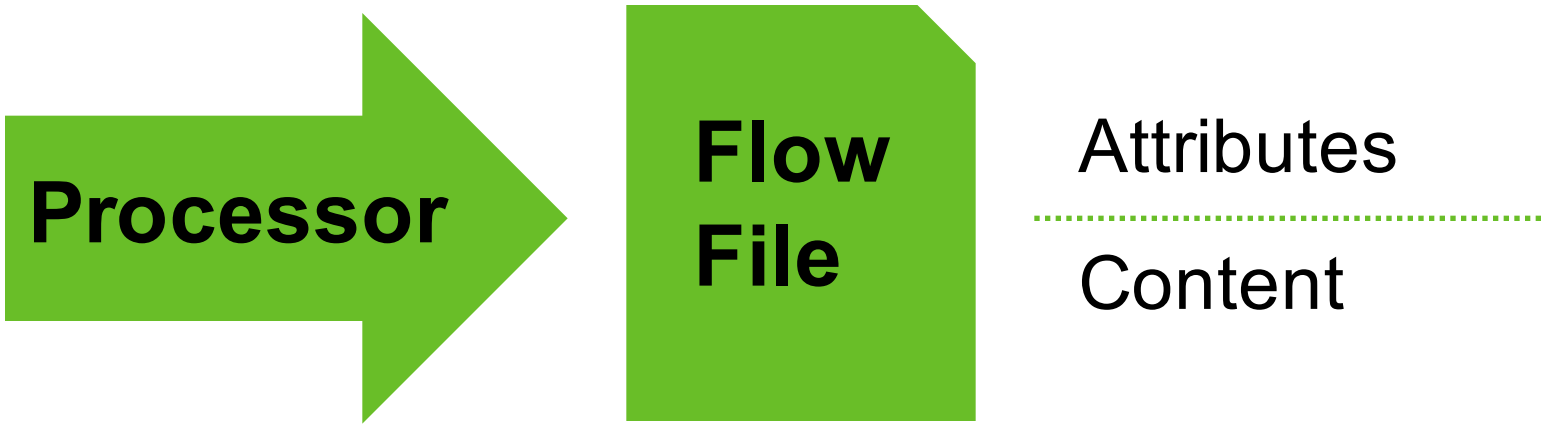
- Single NiFi instance
- HDP 1 nodes cluster
  - HDFS
  - Hive
  - Solr
- A database server with some sample data.

Hortonworks®

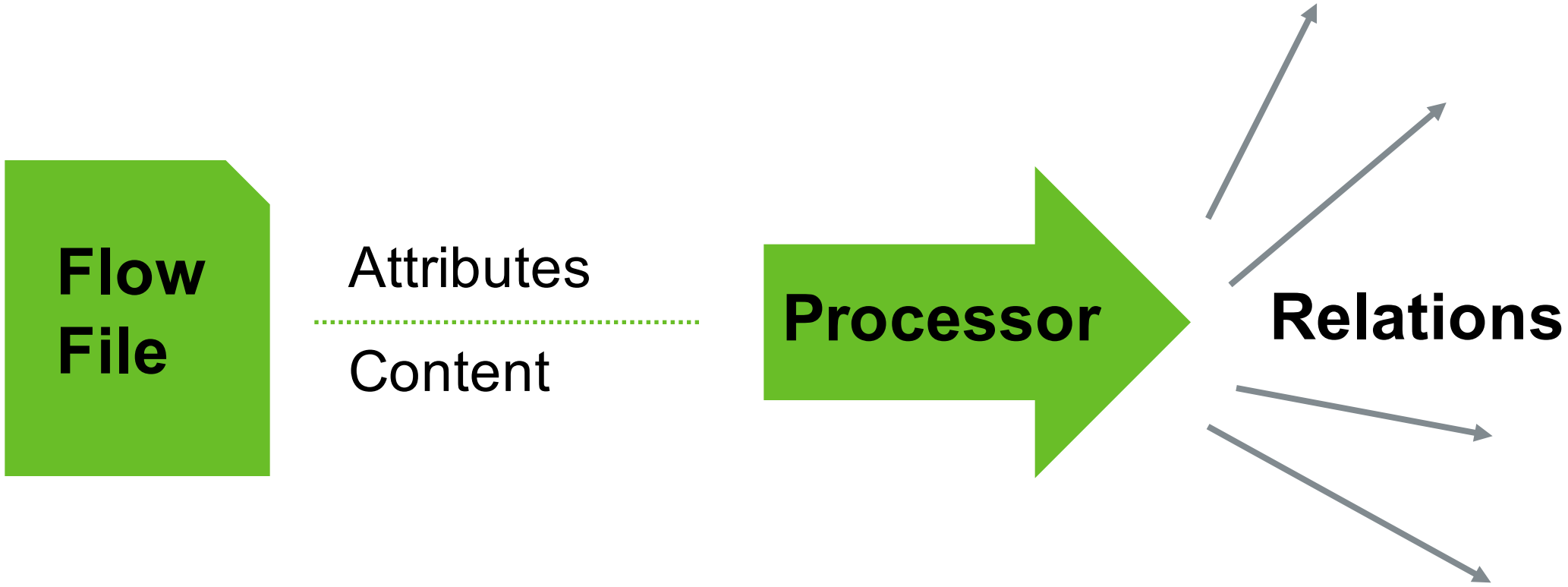
How it all works

HDF Masterclass

NiFi concepts



NiFi concepts



Hands on: A basic NiFi flow

HDF Masterclass

Hands on: Building a basic flow

- **Explore the processors**
- **Create some random data**
- **Change the destination filename**
- **Compress it**
- **Write it out somewhere**

Hortonworks®

Hands on: Building a basic flow

- **Explore the processors**
- **Create some random data** (GenerateFlowFile)
- **Change the destination filename** (UpdateAttribute)
- **Compress it** (CompressFlowFile)
- **Write it out somewhere** (PutFile)

Hands on: Adding to the flow

- **Ingesting some log files**
- **Add some log streaming sources**
- **Combine multiple flows**

Hortonworks®

Hands on: Adding to the flow

- **Ingesting some log files** (TailFile)
- **Add some log streaming sources** (ListenSyslog)
- **Combine multiple flows** (Funnels, Connections)

Hortonworks®

Provenance and Lineage

HDF Masterclass

HDF Data Provenance Improves Business Outcomes

IT and Cloud Operators

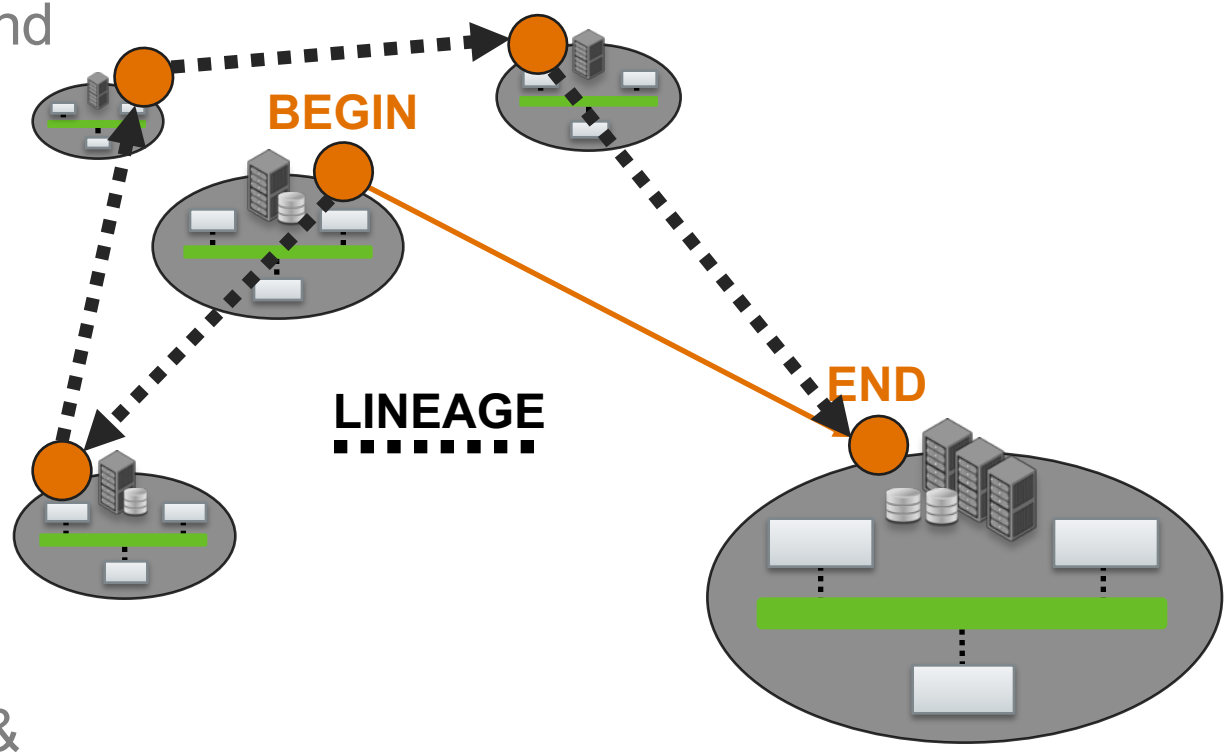
- Visual graph to immediately understand traceability, lineage
- Quickly determine chain of custody
- Enable recovery and replay

Compliance Regulations

- Provide an audit trail
- Remediation capabilities

Business

- Deterministically value data sources & IT investment based on actual usage
- Tune investments for maximum impact



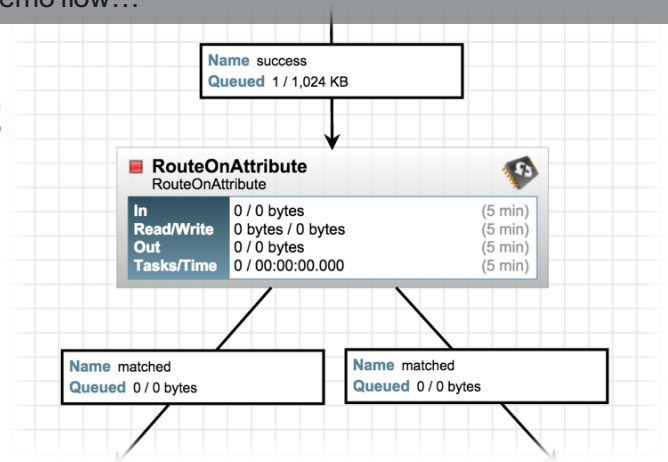
Hortonworks®

NiFi Architecture – Repositories - Pass by reference

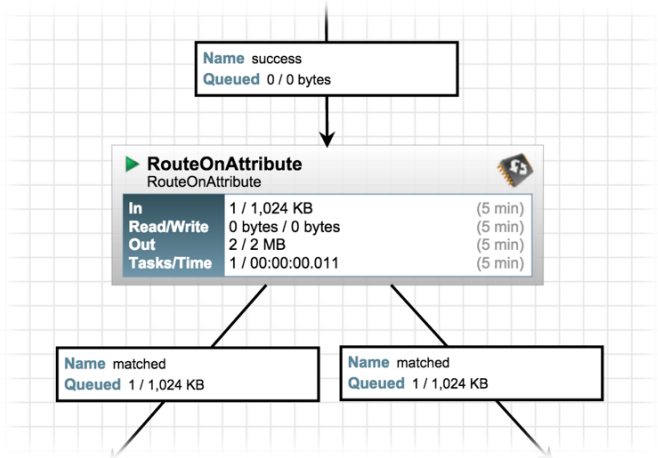
Excerpt of demo flow…

What's happening inside the repositories…

BEFORE



AFTER



FlowFile	Content	Provenance
$F_1 \rightarrow C_1$	C_1	$P_1 \rightarrow F_1$
$F_1 \rightarrow C_1$ $F_2 \rightarrow C_1$	C_1	$P_1 \rightarrow F_1$ – Create $P_2 \rightarrow F_1$ – Route $P_3 \rightarrow F_2$ – Clone (F_1)

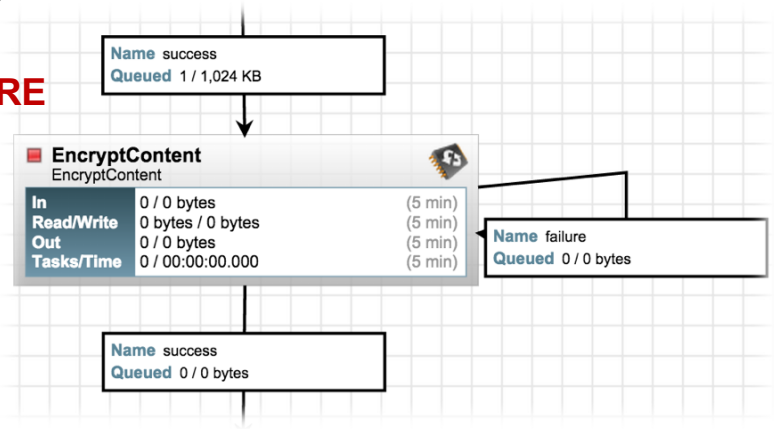
Hortonworks®

NiFi Architecture – Repositories – Copy on Write

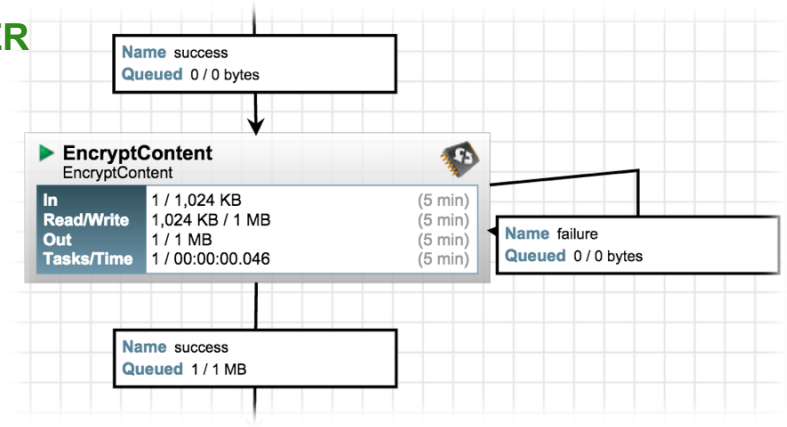
Excerpt of demo flow…

What's happening inside the repositories…

BEFORE



AFTER



FlowFile	Content	Provenance
$F_1 \rightarrow C_1$	C_1	$P1 \rightarrow F1$ - CREATE
~~$F_1 \rightarrow C_1$~~ $F_{1.1} \rightarrow C_2$	C_1 (plaintext) C_2 (encrypted)	$P_1 \rightarrow F_1$ - CREATE $P_2 \rightarrow F_{1.1}$ - MODIFY

Hortonworks®

Organisation and Groups

HDF Masterclass

Hands on: Process Groups and Organisation

- **Creating process groups**
- **Input and output ports**
- **Templates and re-using flows**

Hortonworks®

Hortonworks®

External Sources and Destinations

HDF Masterclass

Hands on: External sources

- **Grabbing data from a web based source**
- **The List, Fetch pattern: bringing in SFTP data**
- **Interpreting the data to pull out attributes**
- **Making decisions about what to do next**
- **Sending data to different endpoints**

Hortonworks®

Modifying flows

HDF Masterclass

Hands on: Modifying data flows on the fly

- **Adding to an additional flow**
- **Changing the scheduling of a processor**
- **Updating queue settings**

Hortonworks®

Hints: Modifying flows

- **Stop the processor to reconfigure**
- **Transactions and changes**
- Optimistic locking
- **Changing a queue, or connection:**
- Stop either side
- Sometime drain the queue (DANGER!)

Remote clusters

HDF Masterclass

Hands on: Linking multiple clusters

- **Setting up a collection agent**
- **Remote port setup**
- **Settings up remote links**

Hortonworks®

Custom Processors

HDF Masterclass

Custom Processors

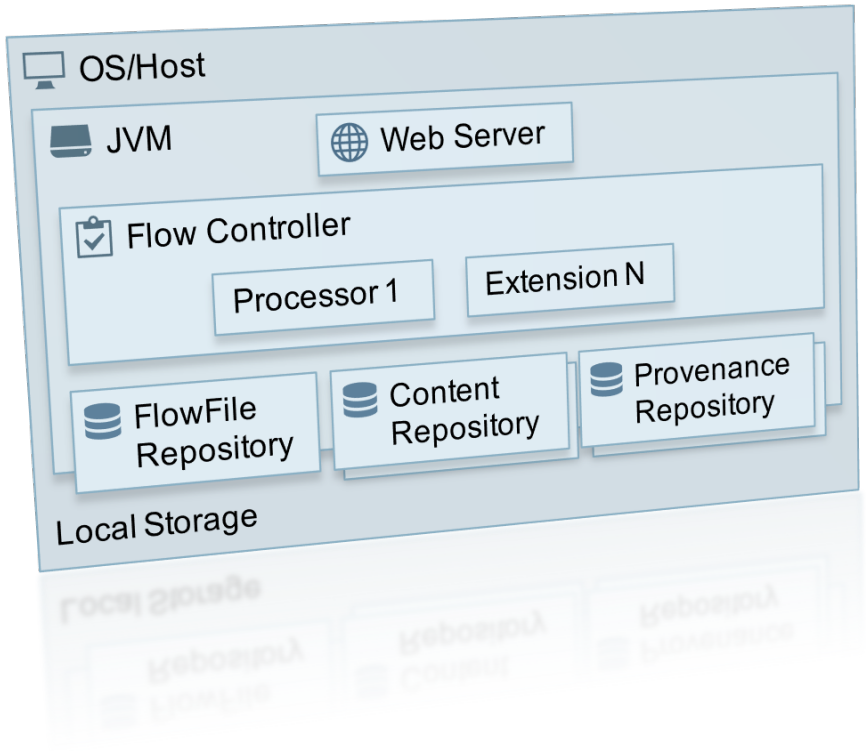
- **Ingress**
- **Egress**
- **Routing**
- One to One
- One to Many
- Split
- **Update attributes**
- **Enrich and modify content**

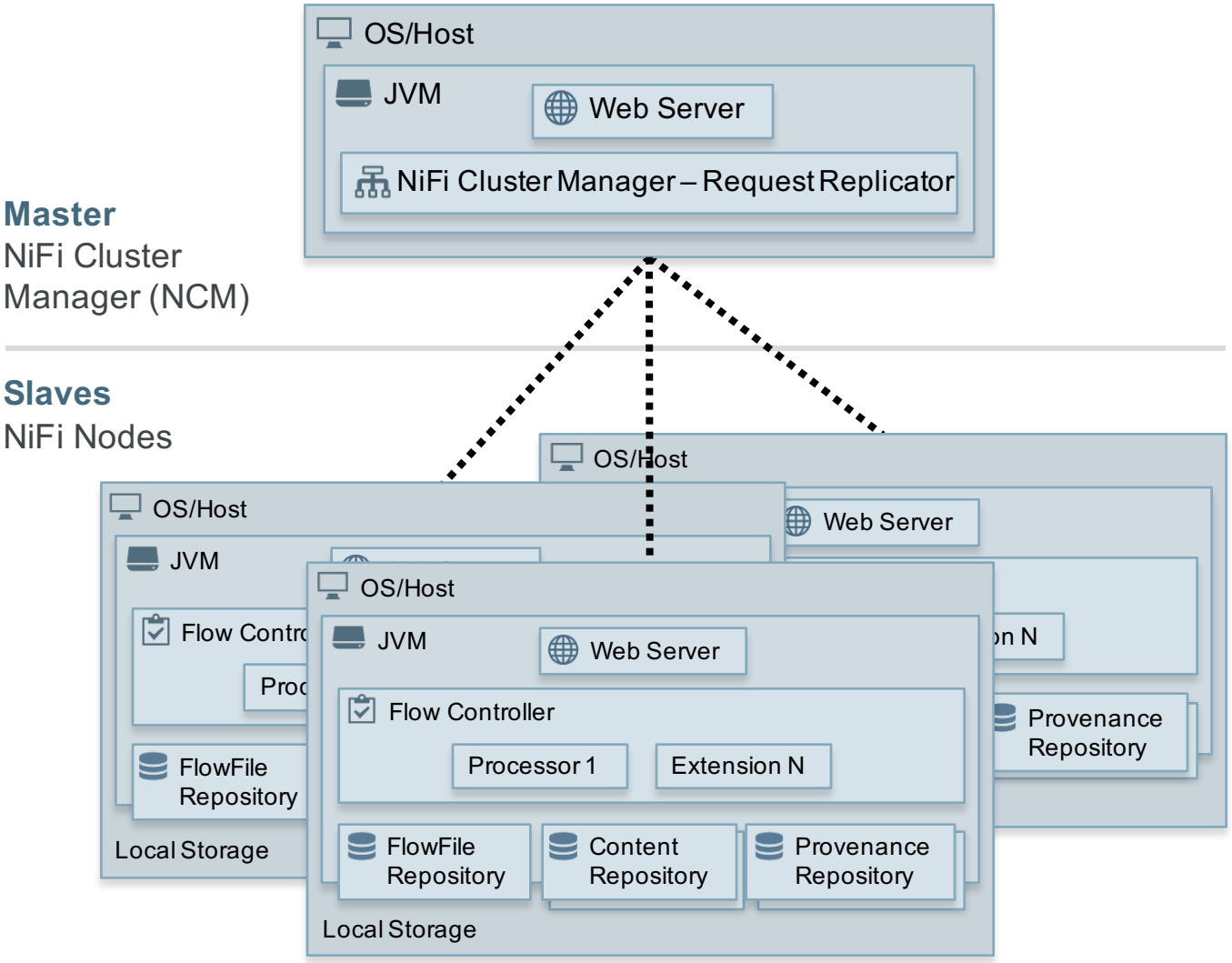
Hortonworks®

Hortonworks®

NiFi in Production

Architecture





Hortonworks®

Production concerns

Security: SSL (two-way)

Authentication: LDAPProvider

Authorization: AuthorityProvider

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®