Dataflow with Apache NiFi

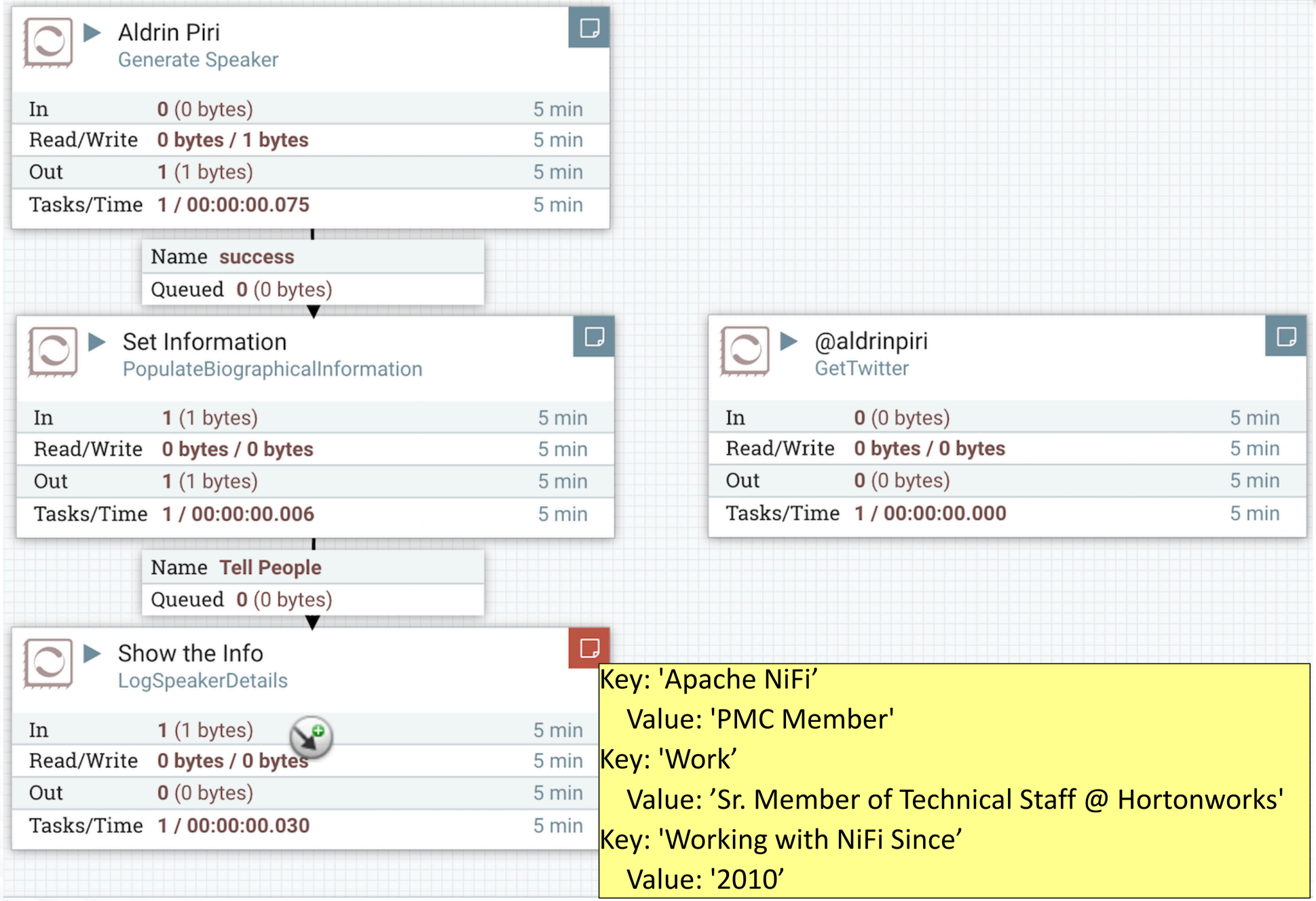
Aldrin Piri - @aldrinpiri

Apache NiFi Crash Course

DataWorks Summit 2017 – Munich

6 April 2017

HORTONWORKS®
POWERING THE FUTURE OF DATA™



Agenda

What is dataflow and what are the challenges?

Apache NiFi

Architecture

Live Demo

Community

Agenda

- What is dataflow and what are the challenges?
- Apache NiFi
- Architecture
- Live Demo
- Community

HORTONWORKS

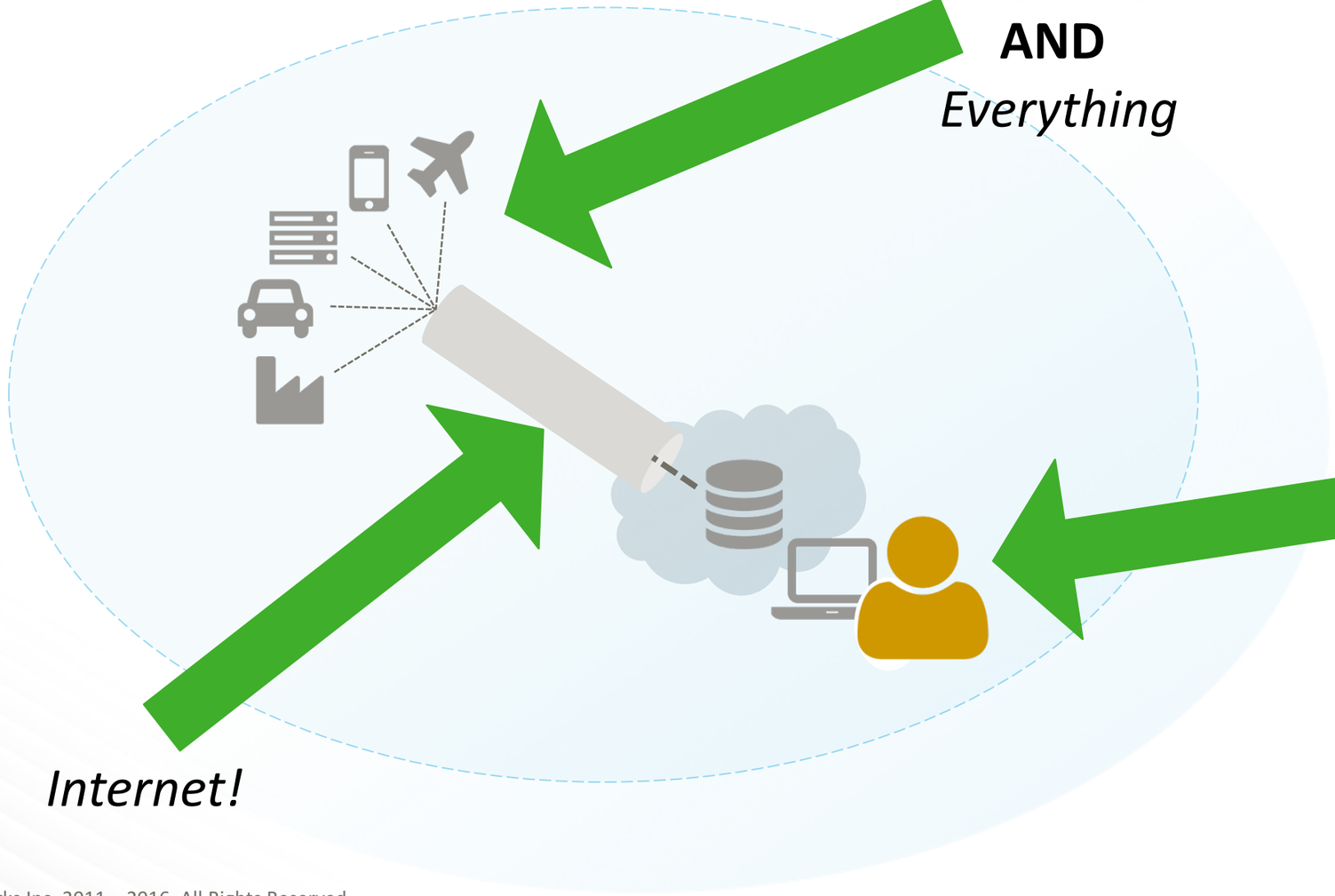
Let's Connect A to B

Producers A.K.A Things

Anything

AND

Everything



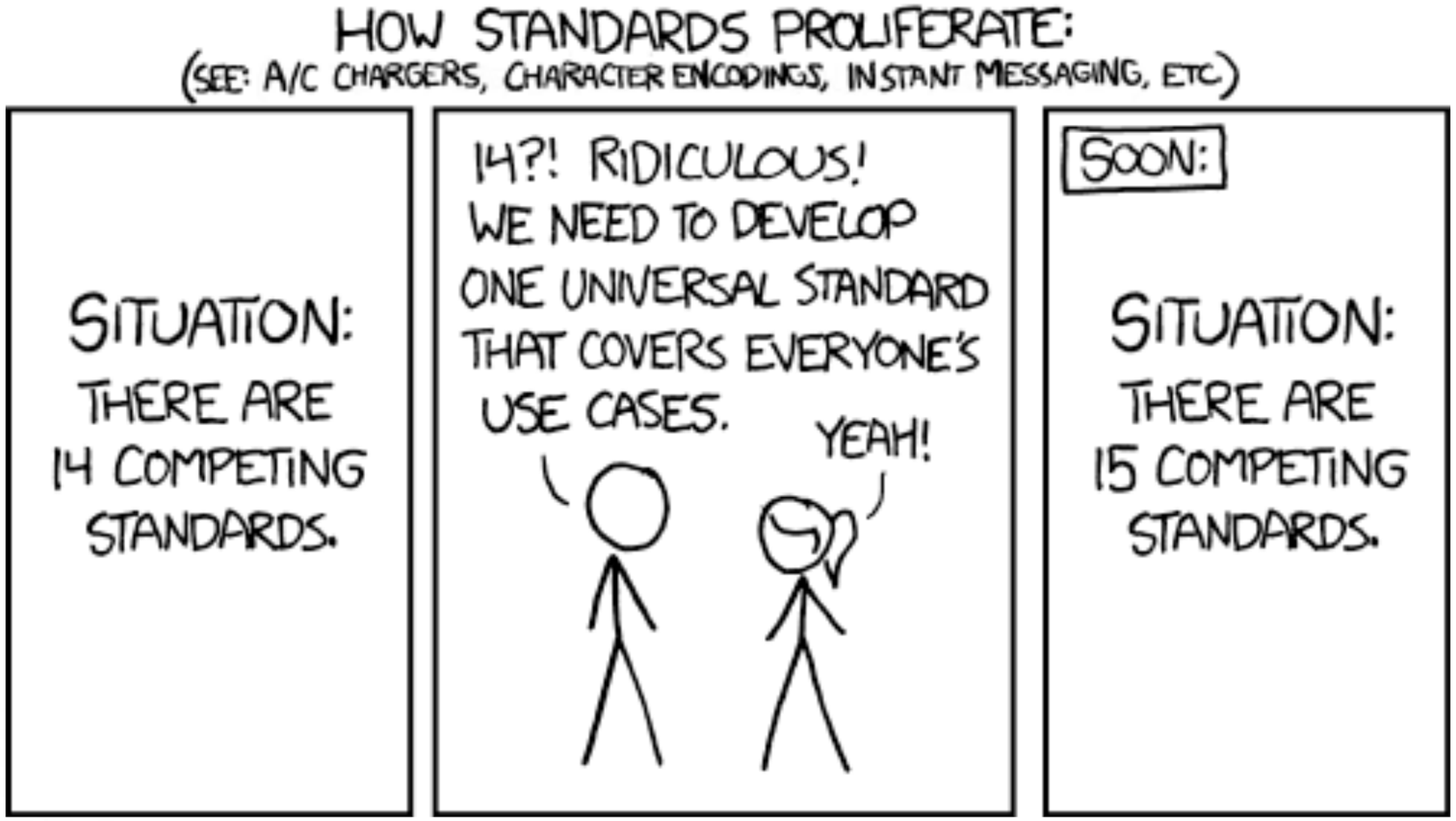
Consumers

- *User*
- *Storage*
- *System*
- ***…More** Things*

Internet!

HORTONWORKS

Moving data *effectively* is hard



Standards: http://xkcd.com/927/

HORTONWORKS

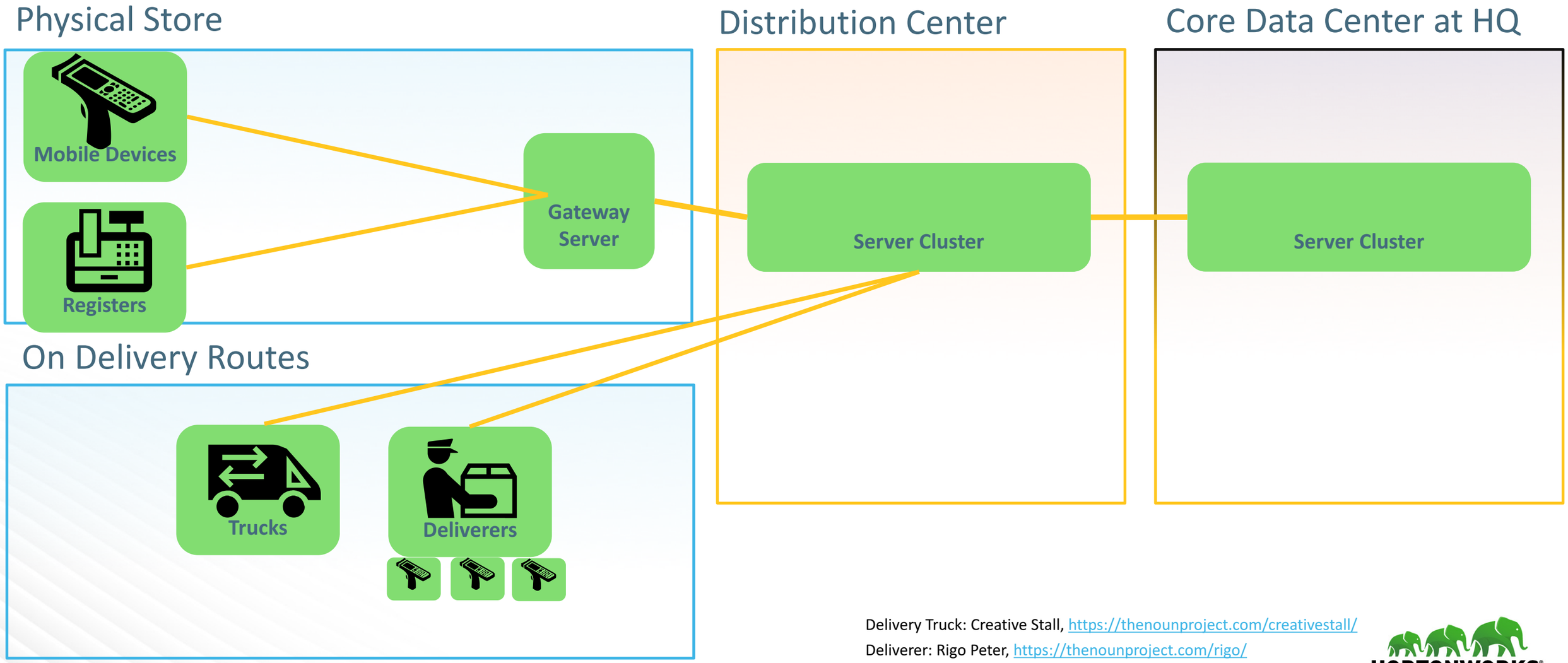
Why is moving data *effectively* hard?

- Standards
- Formats
- “Exactly Once” Delivery
- Protocols
- Veracity of Information
- Validity of Information
- Ensuring Security
- Overcoming Security
- Compliance
- Schemas
- Consumers Change
- Credential Management
- “*That* [person|team|group]”
- Network
- “Exactly Once” Delivery

HORTONWORKS

Let's Connect Lots of As to Bs to As to Cs to Bs to Δs to Cs to φs

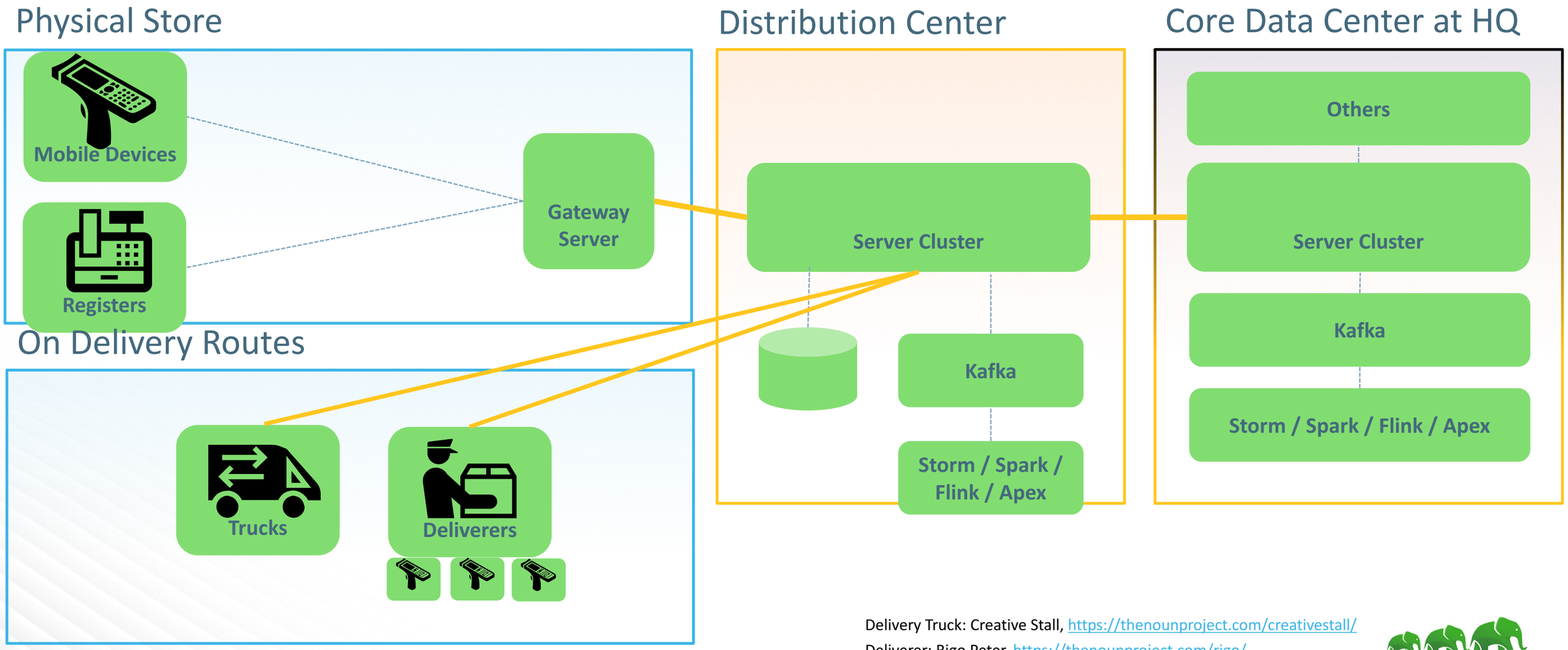
Let's consider the needs of a courier service



Delivery Truck: Creative Stall, https://thenounproject.com/creativestall/
Deliverer: Rigo Peter, https://thenounproject.com/rigo/
Cash Register: Sergey Patutin, https://thenounproject.com/bdesign.by/
Hand Scanner: Eric Pearson, https://thenounproject.com/epearson001/

Great! I am collecting all this data! Let's use it!

Finding our needles in the haystack



Delivery Truck: Creative Stall, https://thenounproject.com/creativestall/
Deliverer: Rigo Peter, https://thenounproject.com/rigo/
Cash Register: Sergey Patutin, https://thenounproject.com/bdesign.by/
Hand Scanner: Eric Pearson, https://thenounproject.com/epearson001/

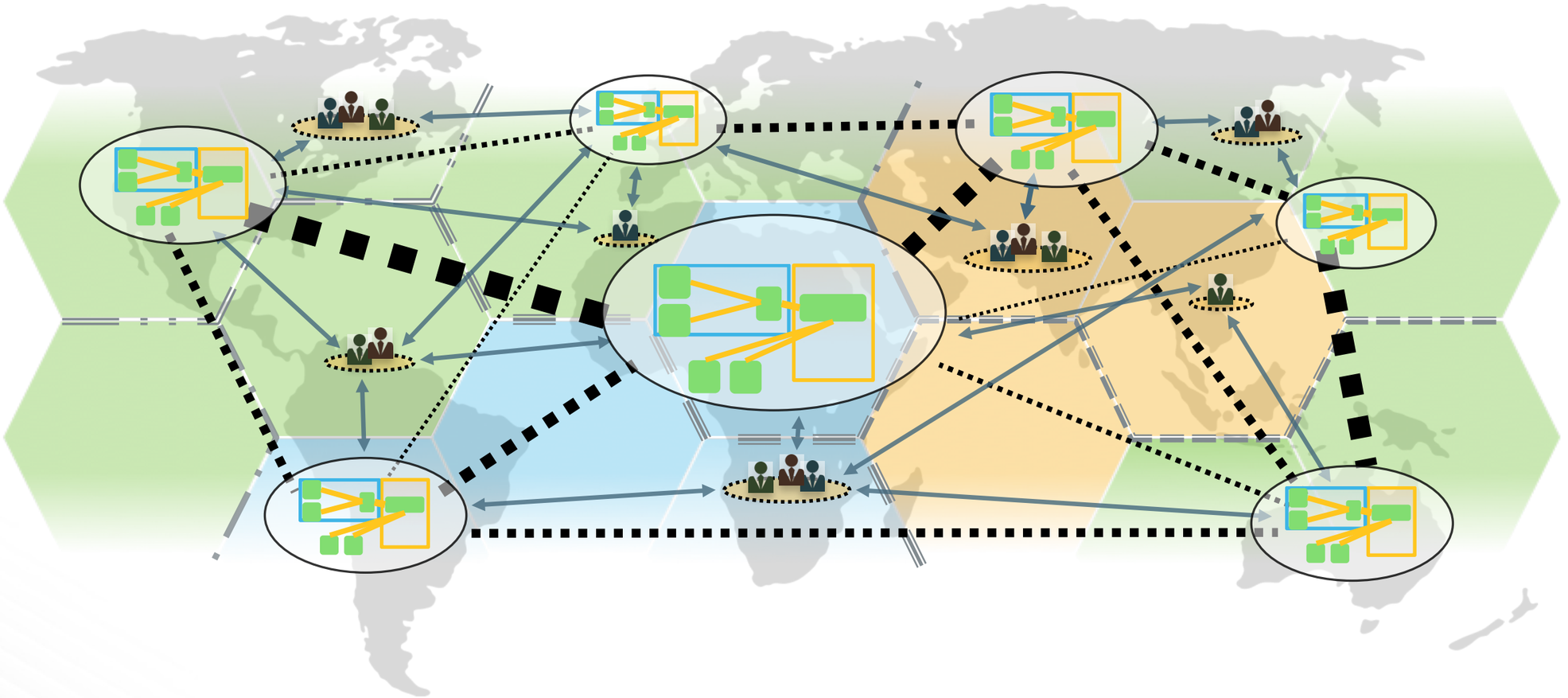
Why is moving data *effectively* hard when scoped *internally*?

- Standards
- Formats
- “Exactly Once” Delivery
- Protocols
- Veracity of Information
- Validity of Information
- Ensuring Security
- Overcoming Security
- Compliance
- Schemas
- Consumers Change
- Credential Management
- “*That* [person|team|group]”
- Network
- “Exactly Once” Delivery

HORTONWORKS

Let’s Connect Lots of As to Bs to As to Cs to Bs to Δs to Cs to φs

Oh, that courier service is global



HORTONWORKS

Why is moving data *effectively* hard when scoped *globally*?

- Standards
- Formats
- “Exactly Once” Delivery
- Protocols
- Veracity of Information
- Validity of Information
- Ensuring Security
- Overcoming Security
- Compliance
- Schemas
- Consumers Change
- Credential Management
- “*That* [person|team|group]”
- Network
- “Exactly Once” Delivery

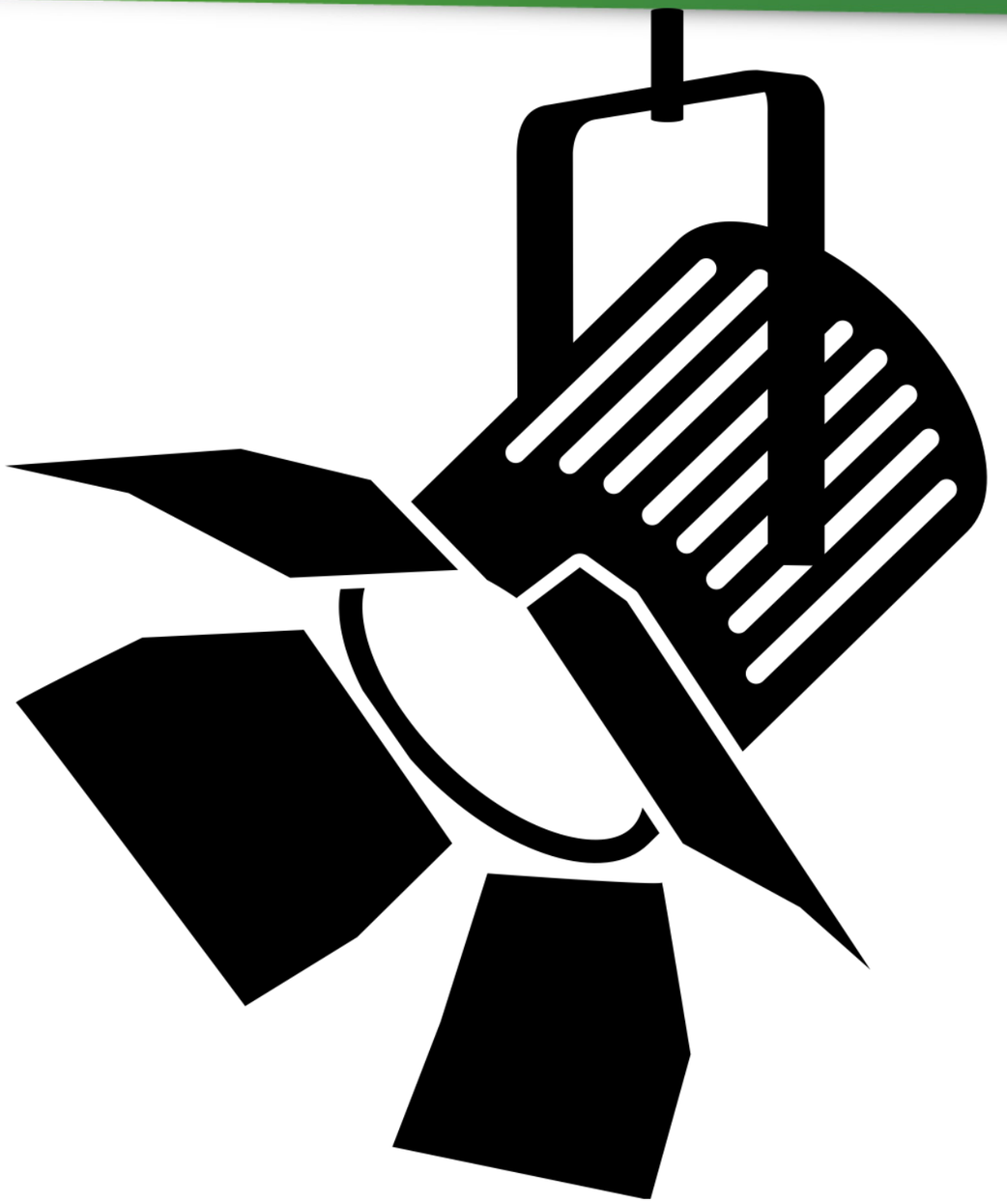
HORTONWORKS

The Unassuming Line: A Case Study

We've seen a few lines show up in the wild thus far

Internet!

Inter- & Intra- connections in our global courier enterprise



Spotlight: Arthur Lacôte, https://thenounproject.com/turo/

Dataflow Line Anatomy 101

Let's dissect what this line typically represents

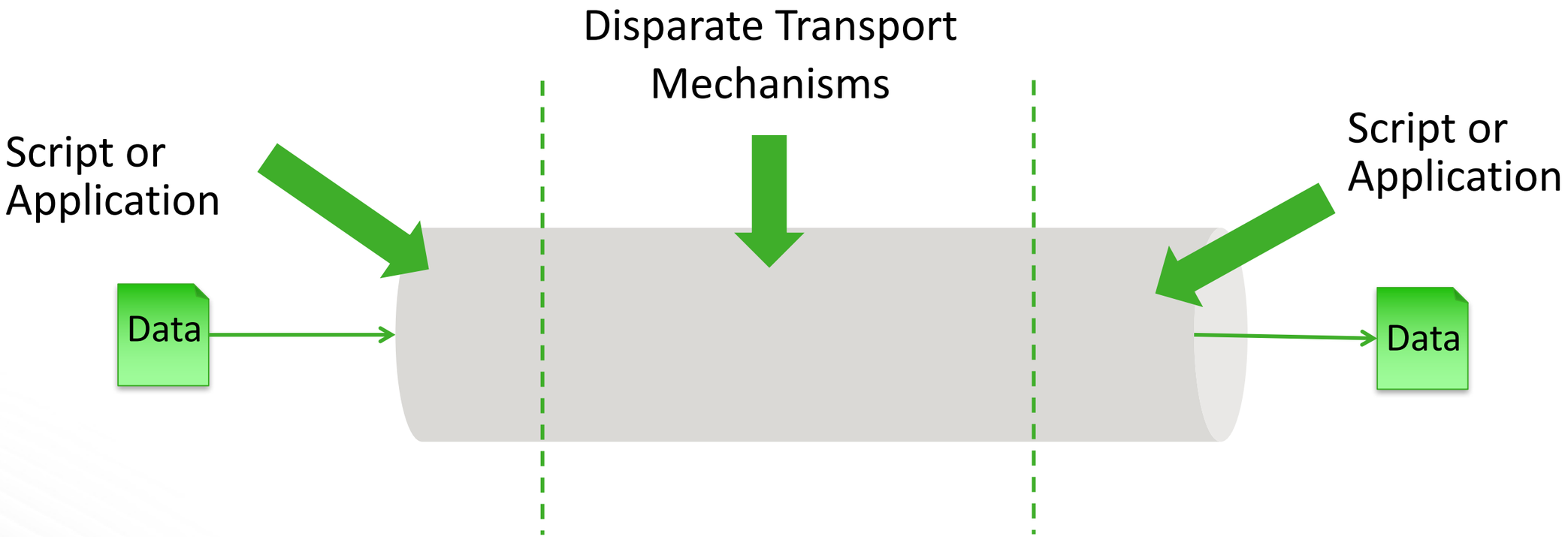


Fig 1. Lineus Worldwidewebus. Common Name: Internet!

HORTONWORKS

Dataflow Line Anatomy 201

Sometimes that transport is just more lines

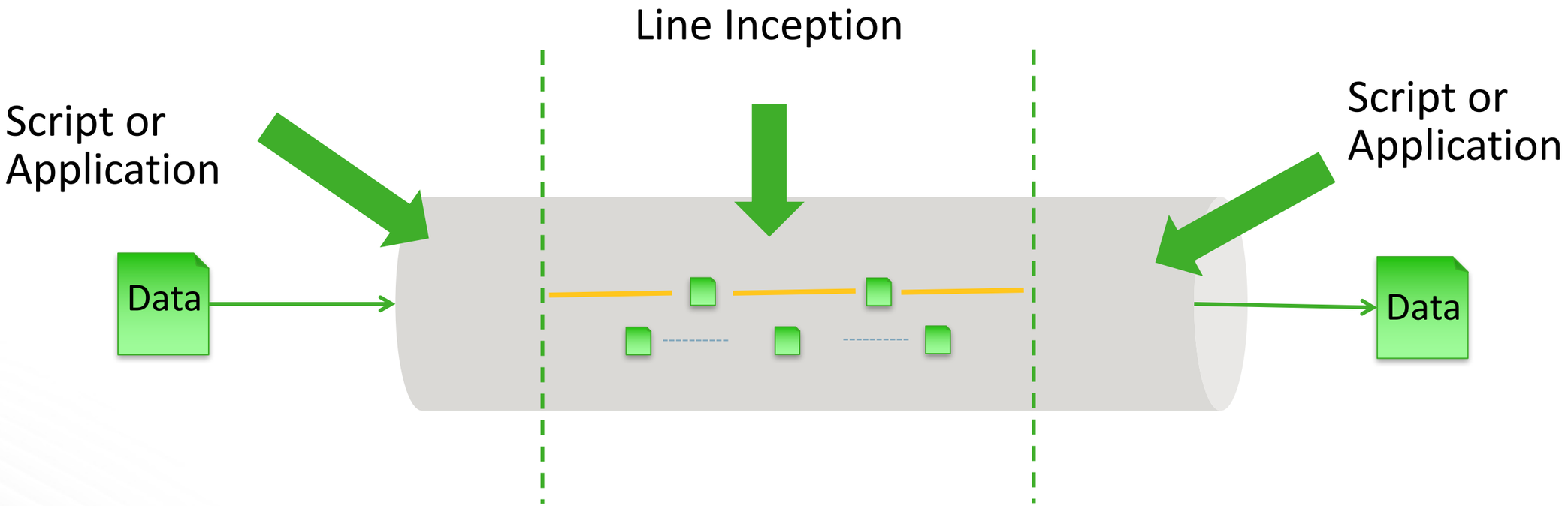


Fig 1. Lineus Worldwidewebus. Common Name: Internet!

HORTONWORKS

Dataflow Line Anatomy 301

But those lines could also have components…

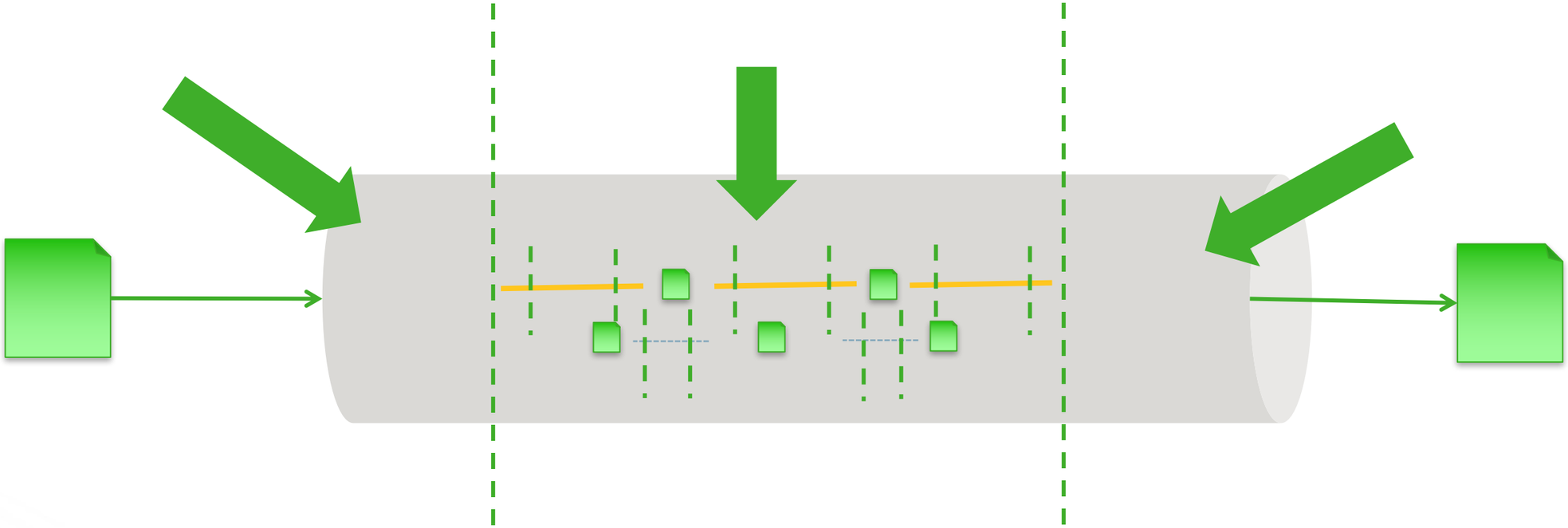


Fig 1. Lineus Worldwidewebus. Common Name: Internet!

Agenda

What is dataflow and what are the challenges?

Apache NiFi

Architecture

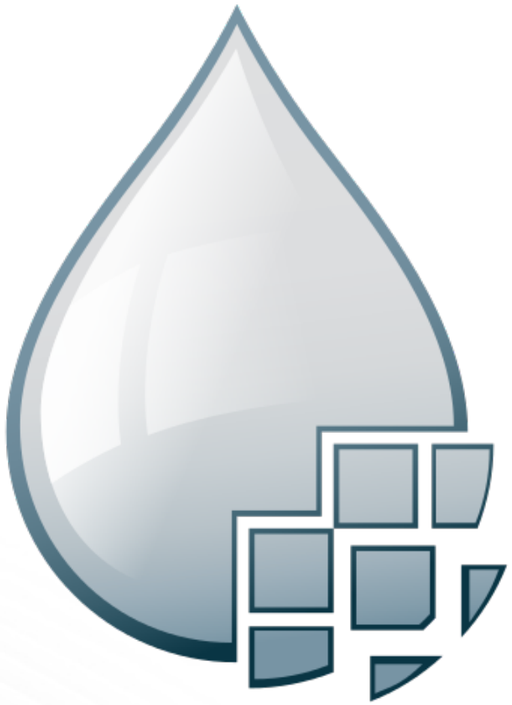
Live Demo

Community

HORTONWORKS

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

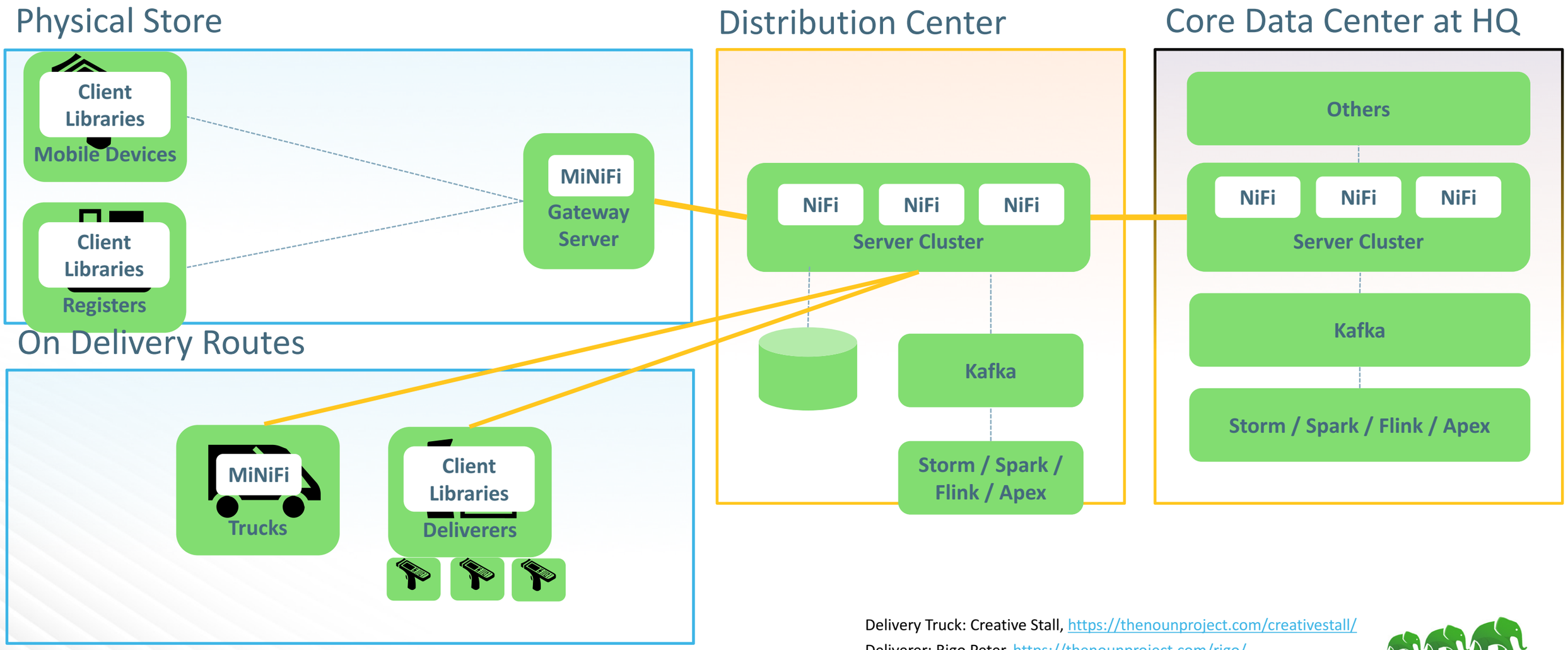
Apache NiFi Subproject: MiNiFi

- Let me get the key parts of NiFi close to where data begins and provide bidrectional communication

- NiFi lives in the data center. Give it an enterprise server or a cluster of them.
- MiNiFi lives as close to where data is born and is a guest on that device or system

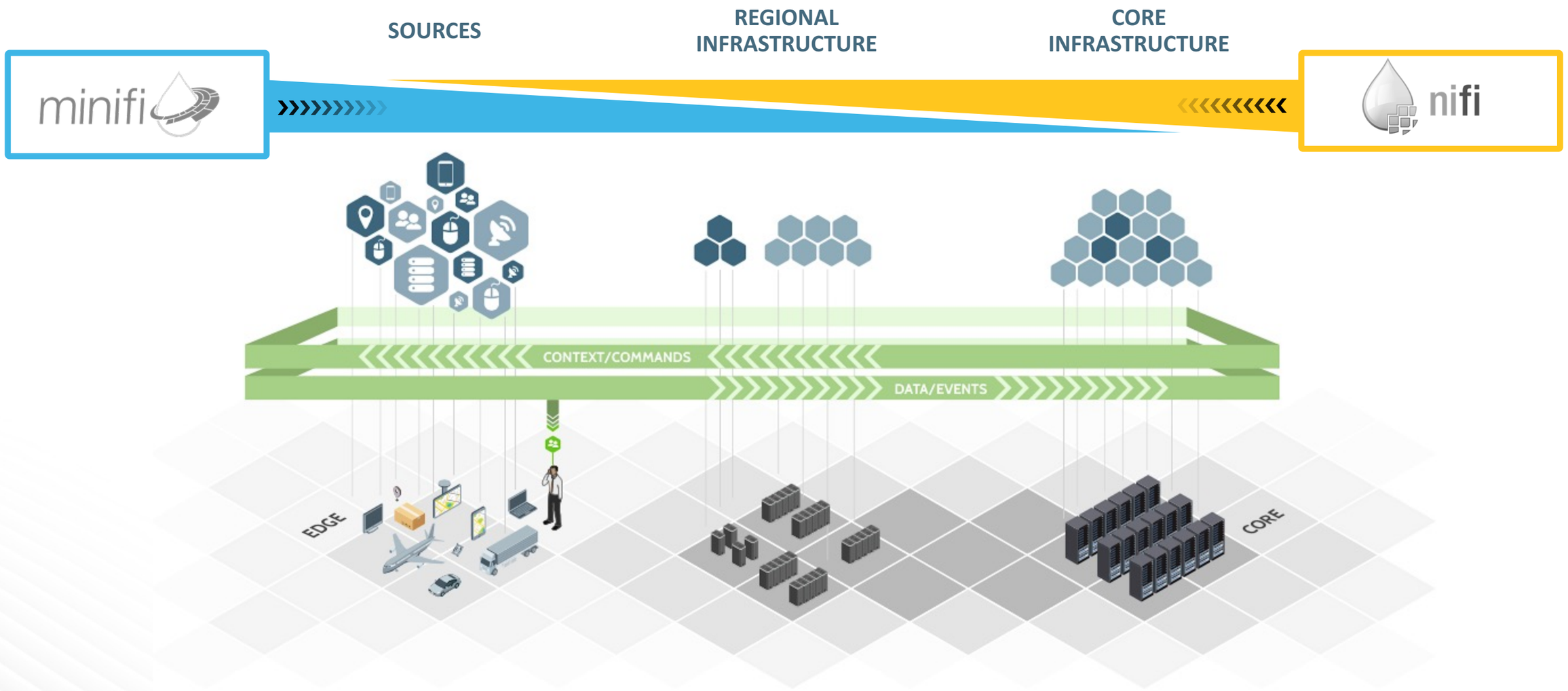
HORTONWORKS

Let’s revisit our courier service from the perspective of NiFi



Delivery Truck: Creative Stall, https://thenounproject.com/creativestall/
Deliverer: Rigo Peter, https://thenounproject.com/rigo/
Cash Register: Sergey Patutin, https://thenounproject.com/bdesign.by/
Hand Scanner: Eric Pearson, https://thenounproject.com/epearson001/

Apache NiFi Managed Dataflow



HORTONWORKS

NiFi is based on Flow Based Programming (FBP)

FBP Term	NiFi Term	Description
Information Packet	FlowFile	Each object moving through the system.
Black Box	FlowFile Processor	Performs the work, doing some combination of data routing, transformation, or mediation between systems.
Bounded Buffer	Connection	The linkage between processors, acting as queues and allowing various processes to interact at differing rates.
Scheduler	Flow Controller	Maintains the knowledge of how processes are connected, and manages the threads and allocations thereof which all processes use.
Subnet	Process Group	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.

HORTONWORKS

FlowFiles & Data Agnosticism

- NiFi is data agnostic!
- But, NiFi was designed understanding that users can care about specifics and provides tooling to interact with specific formats, protocols, etc.

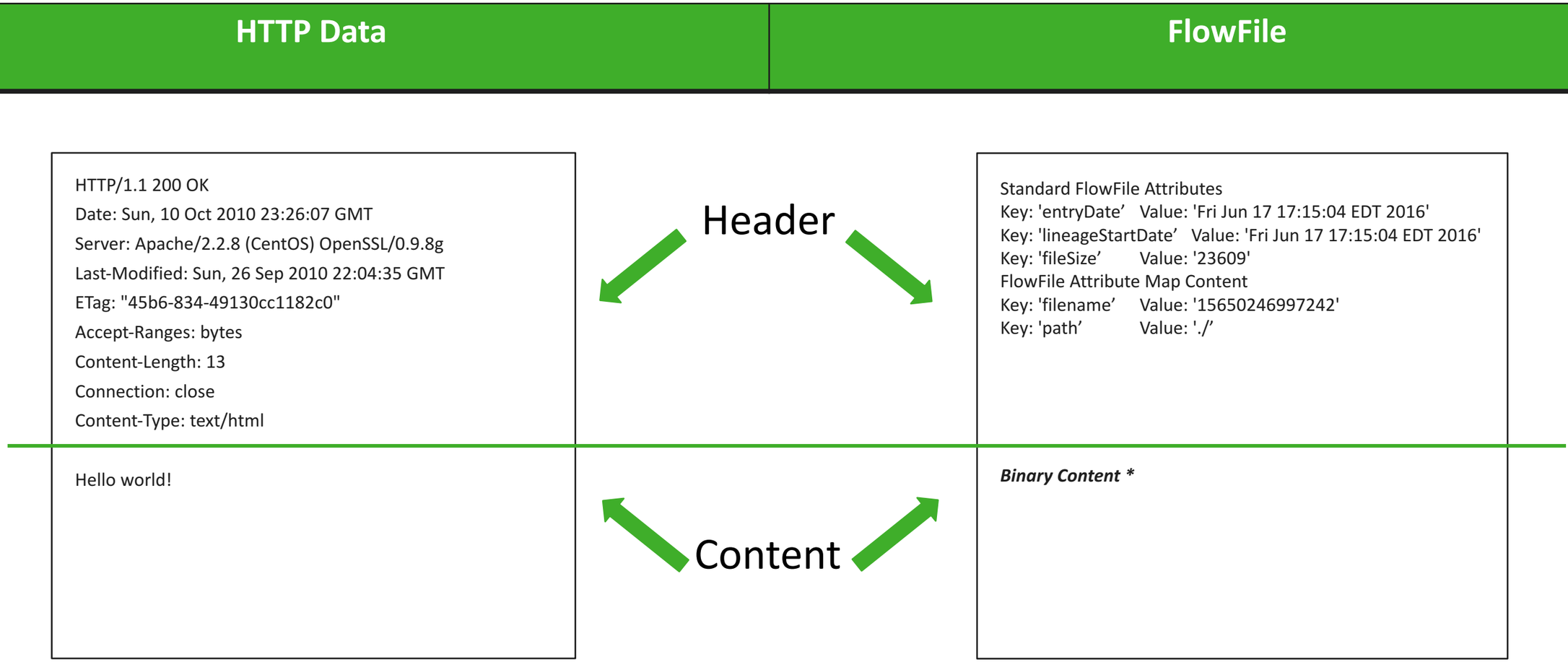
Robustness principle

> "Be conservative in what you do,
> be liberal in what you accept from others

ISO 8601 - http://xkcd.com/1179/

HORTONWORKS

FlowFiles are like HTTP data



Agenda

What is dataflow and what are the challenges?

Apache NiFi

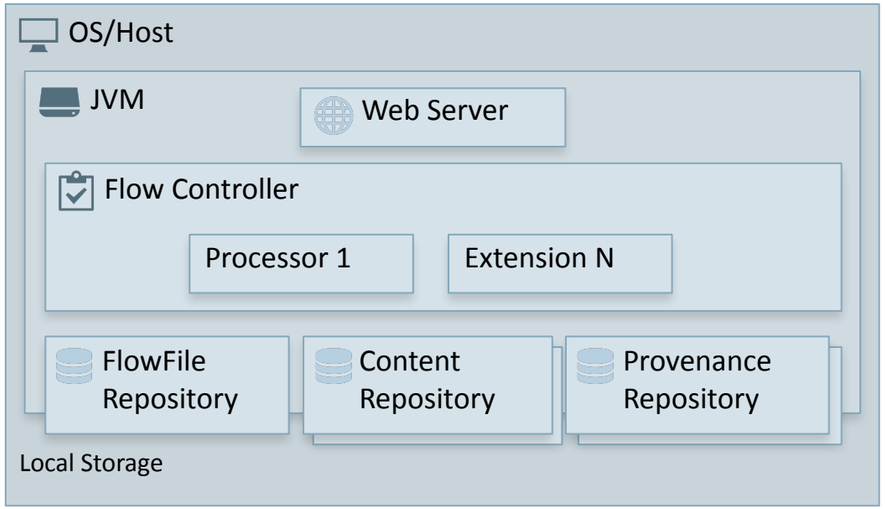
Architecture

Live Demo

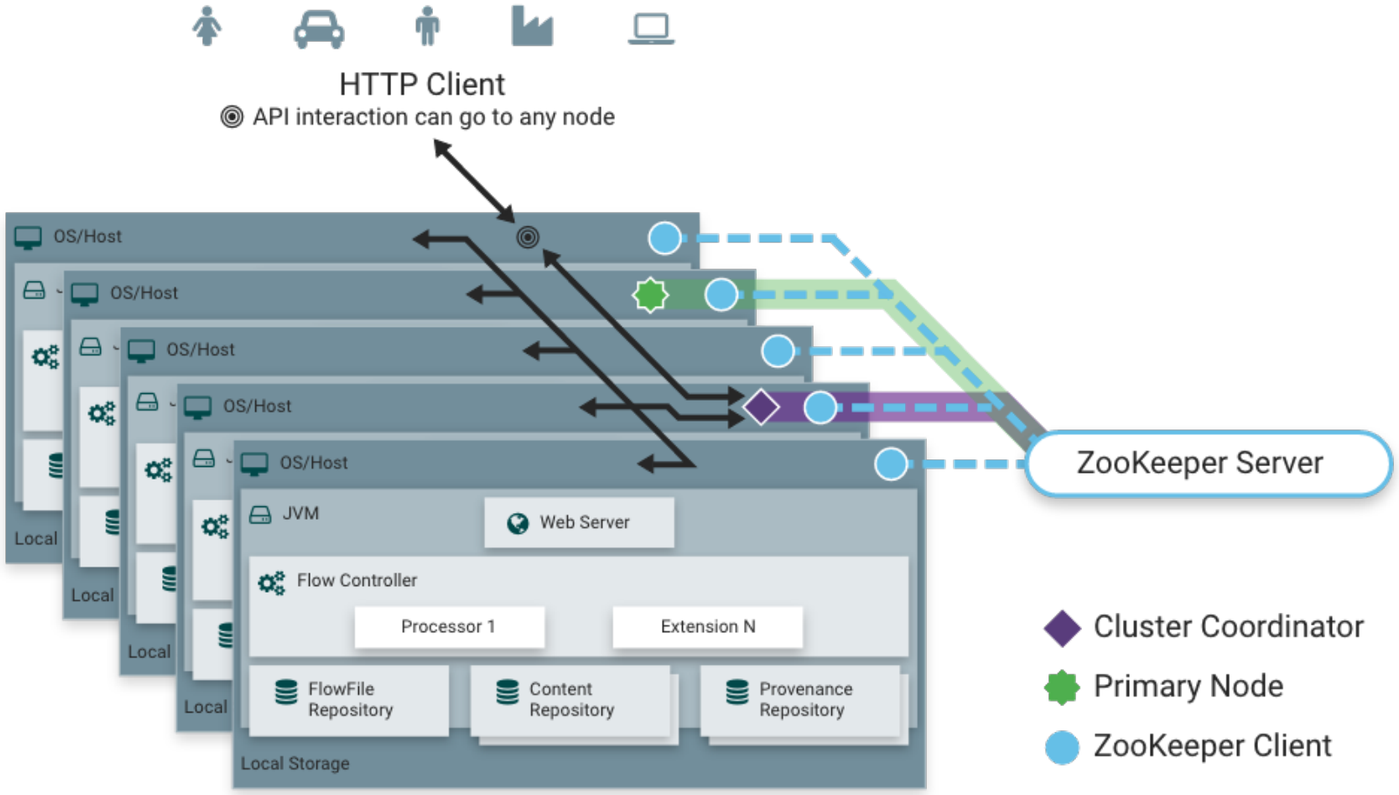
Community

HORTONWORKS

Architecture



Standalone



Cluster

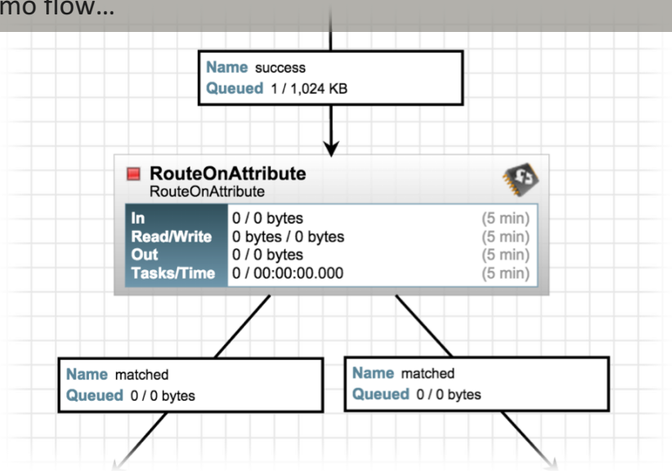
HORTONWORKS

NiFi Architecture – Repositories - Pass by reference

Excerpt of demo flow…

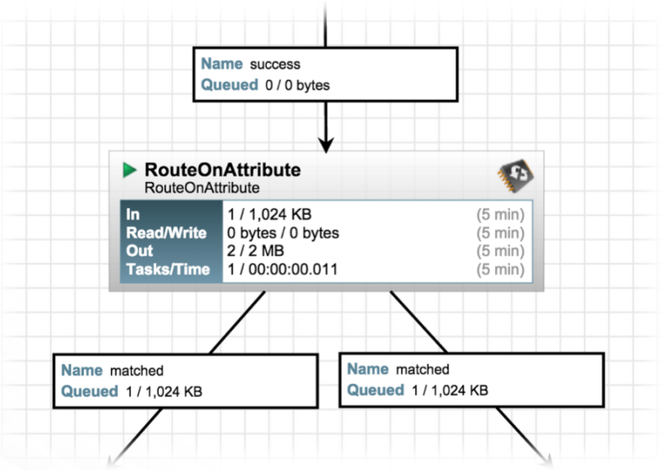
What's happening inside the repositories…

BEFORE



FlowFile	Content	Provenance
$F_1 \rightarrow C_1$	C_1	$P_1 \rightarrow F_1$

AFTER

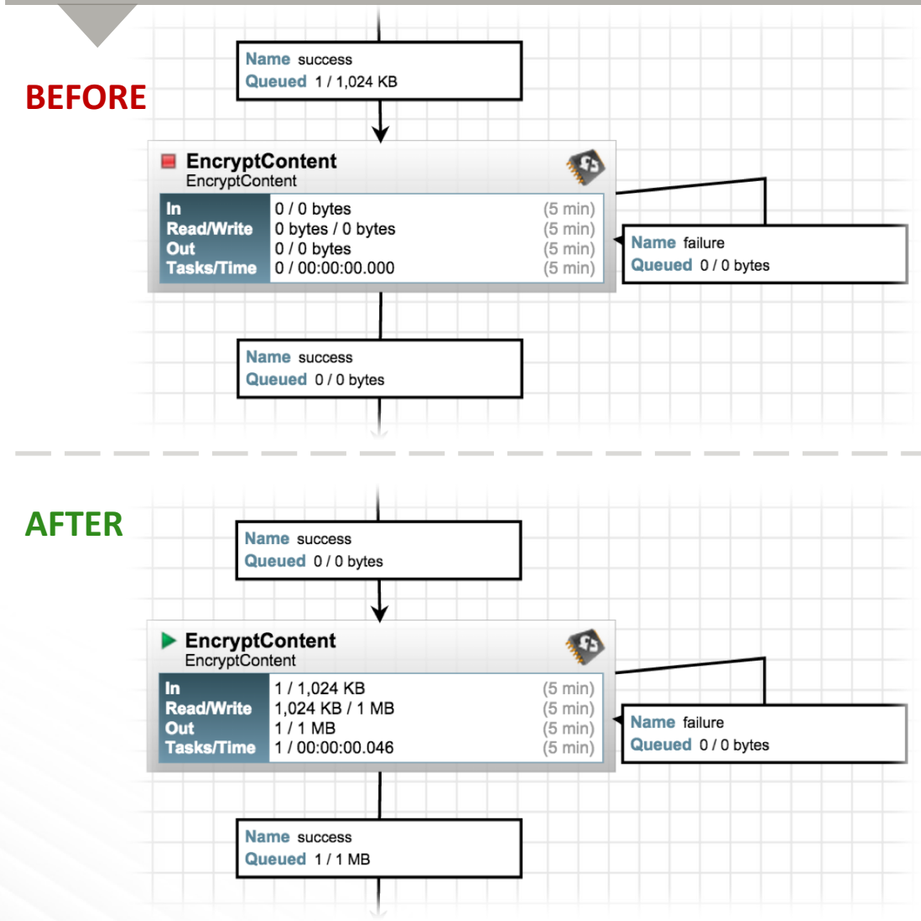


FlowFile	Content	Provenance
$F_1 \rightarrow C_1$	C_1	$P_1 \rightarrow F_1$ – Create
$F_2 \rightarrow C_1$		$P_2 \rightarrow F_1$ – Route
		$P_3 \rightarrow F_2$ – Clone (F_1)

HORTONWORKS

NiFi Architecture – Repositories – Copy on Write

Excerpt of demo flow…



What's happening inside the repositories…

FlowFile	Content	Provenance
$F_1 \rightarrow C_1$	C_1	P1→ 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

Agenda

What is dataflow and what are the challenges?

Apache NiFi

Architecture

Demo

Community

HORTONWORKS

Learn, Share at Birds of a Feather

IOT, STREAMING & DATA FLOW

Thursday, April 6

5:50 pm, Room 5

HORTONWORKS

Why NiFi?

- Moving data is multifaceted in its challenges and these are present in different contexts at varying scopes
  - Think of our courier example and organizations like it: inter vs intra, domestically, internationally

- Provide common tooling and extensions that are commonly needed but be flexible for extension
  - Leverage existing libraries and expansive Java ecosystem for functionality
  - Allow organizations to integrate with their existing infrastructure

- Empower folks managing your infrastructure to make changes and reason about issues that are occurring
  - Data Provenance to show context and data's journey
  - User Interface/Experience a key component

HORTONWORKS

Learn more and join us!

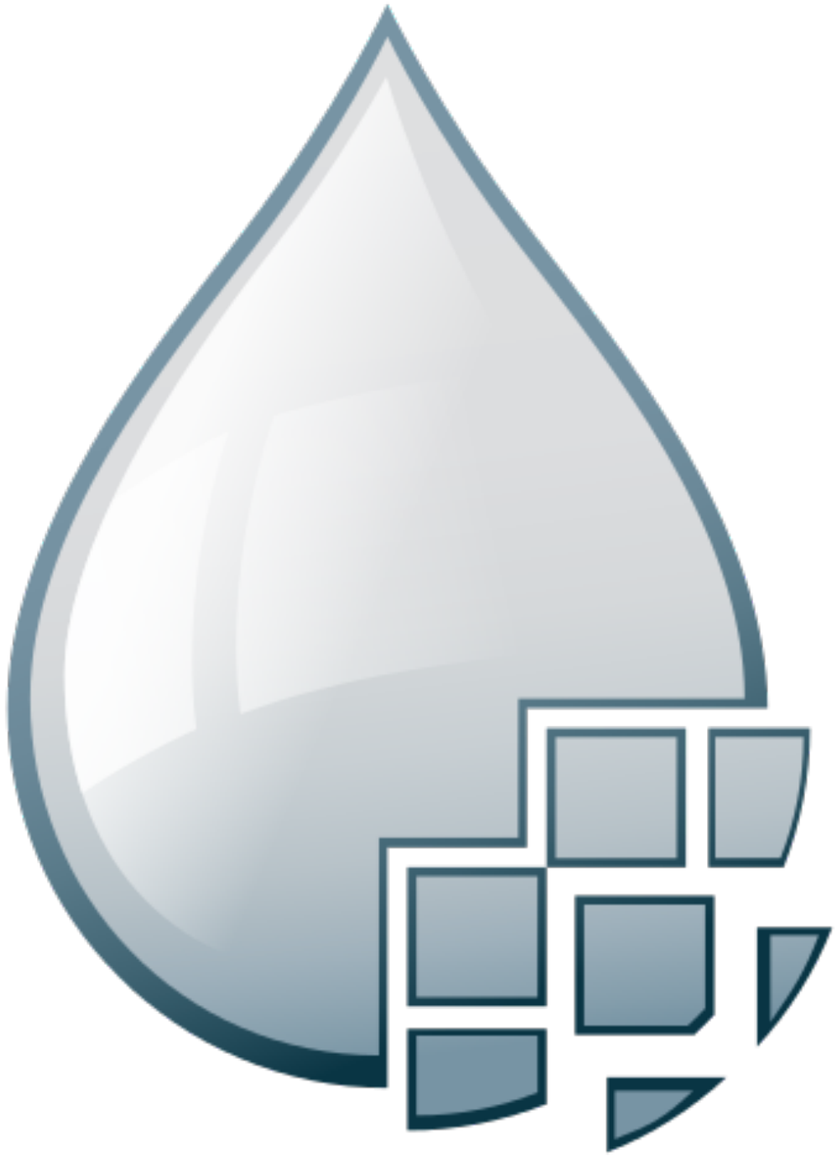
Apache NiFi site
http://nifi.apache.org

Subproject MiNiFi site
http://nifi.apache.org/minifi/

Subscribe to and collaborate at
dev@nifi.apache.org
users@nifi.apache.org

Submit Ideas or Issues
https://issues.apache.org/jira/browse/NIFI

Follow us on Twitter
@apachenifi



HORTONWORKS

Our Lab for Today

- We will be exploring some examples to work through creating a dataflow with Apache NiFi

- Use Case: An urban planning board is evaluating the need for a new highway, dependent on current traffic patterns, particularly as other roadwork initiatives are under way. Integrating live data poses a problem because traffic analysis has traditionally been done using historical, aggregated traffic counts. To improve traffic analysis, the city planner wants to leverage real-time data to get a deeper understanding of traffic patterns. NiFi was selected for for this real-time data integration.

- Labs are available at http://tinyurl.com/nificrashcourse

Thank You

HORTONWORKS®
POWERING THE FUTURE OF DATA™