

# Dataflow with Apache NiFi/MiNiFi

Andy LoPresto - @yolohey

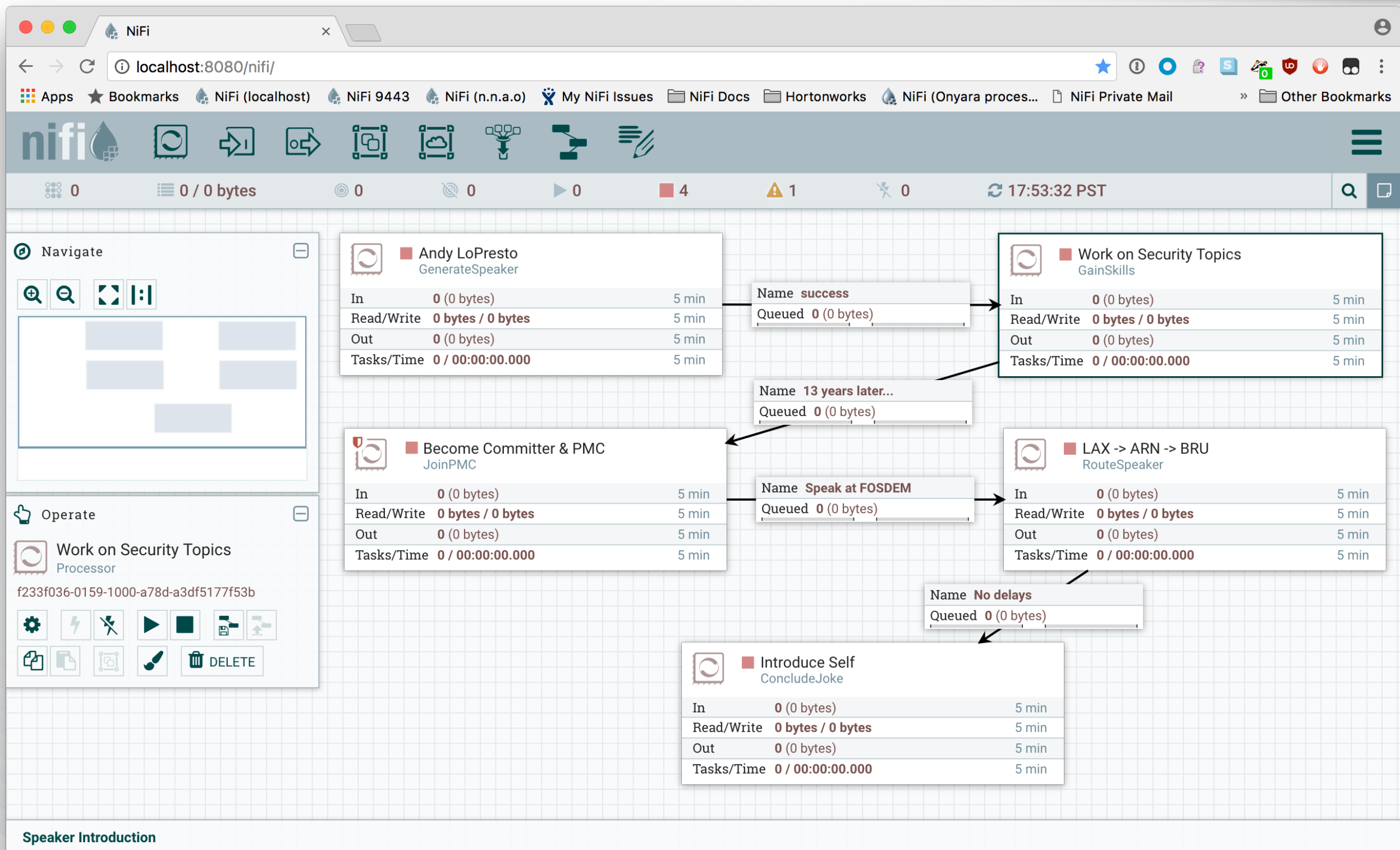
Intelligently Collecting Data at the Edge

Apache NiFi PMC

FOSDEM '17 - Brussels

04 Feb 2017





# Agenda

What is dataflow and what are the challenges?

Apache NiFi

Apache MiNiFi

Architecture

Community

# Agenda



What is dataflow and what are the challenges?

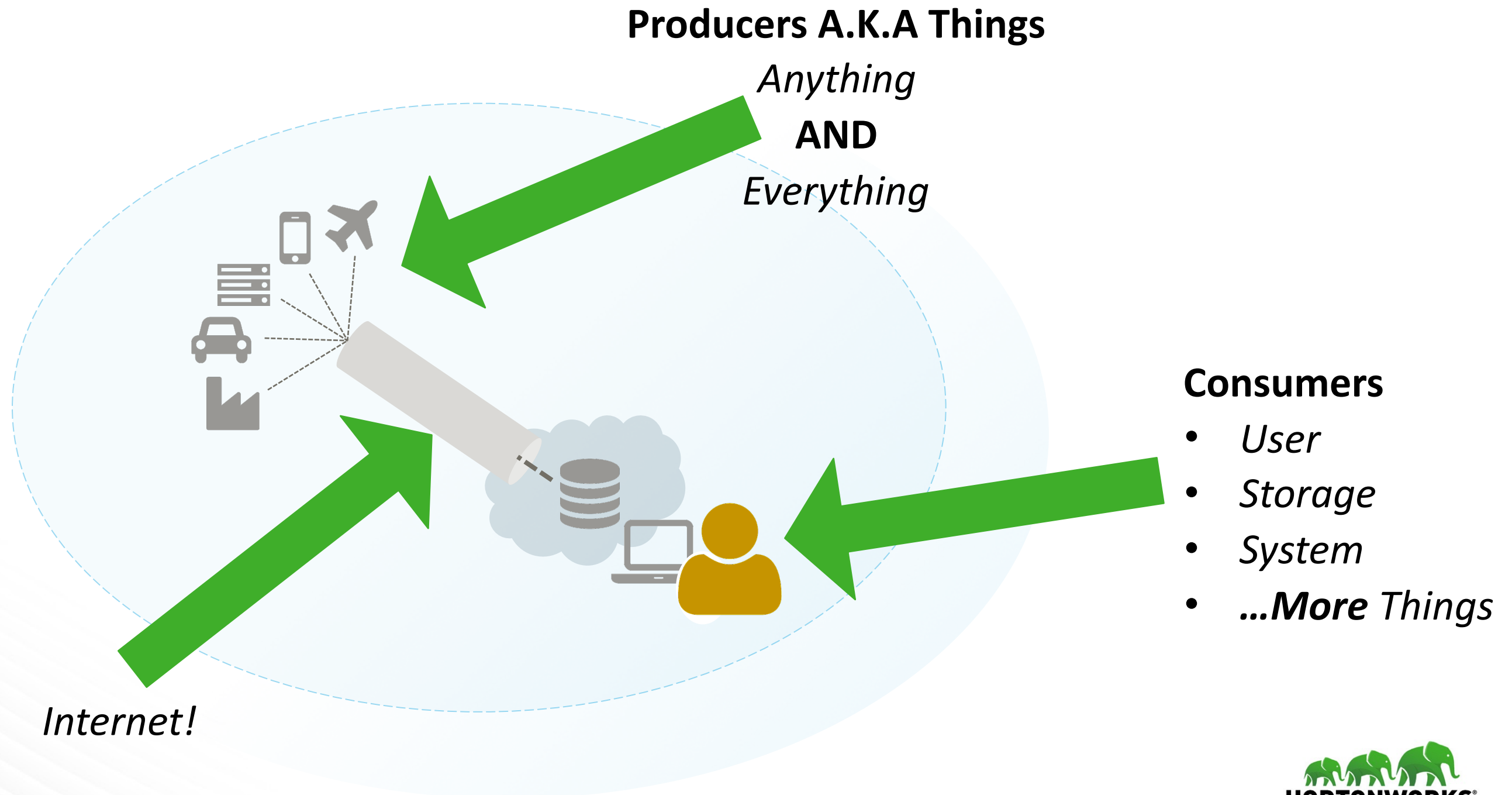
Apache NiFi

Apache MiNiFi

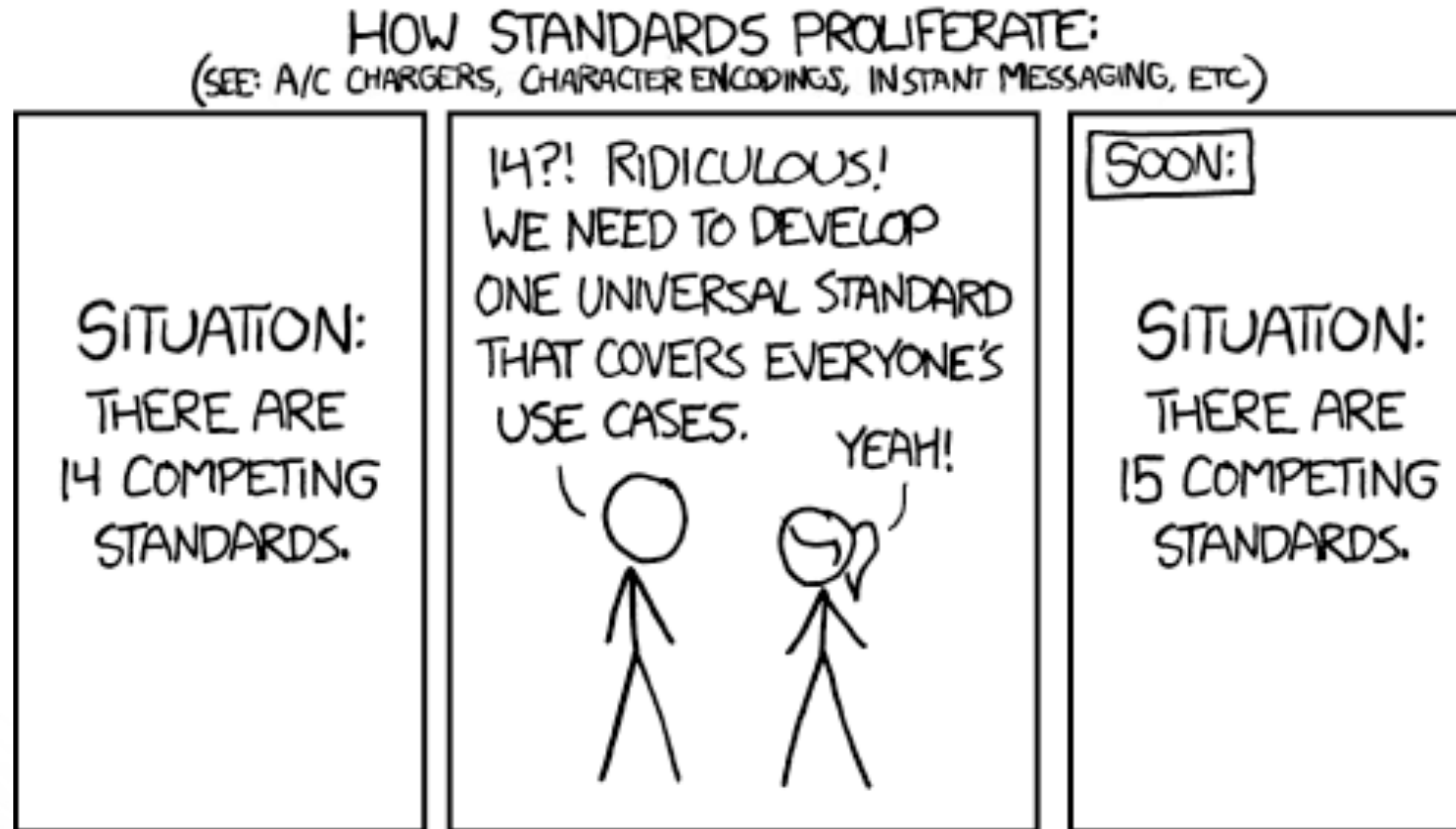
Architecture

Community

# Let's Connect A to B



# Moving data *effectively* is hard



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

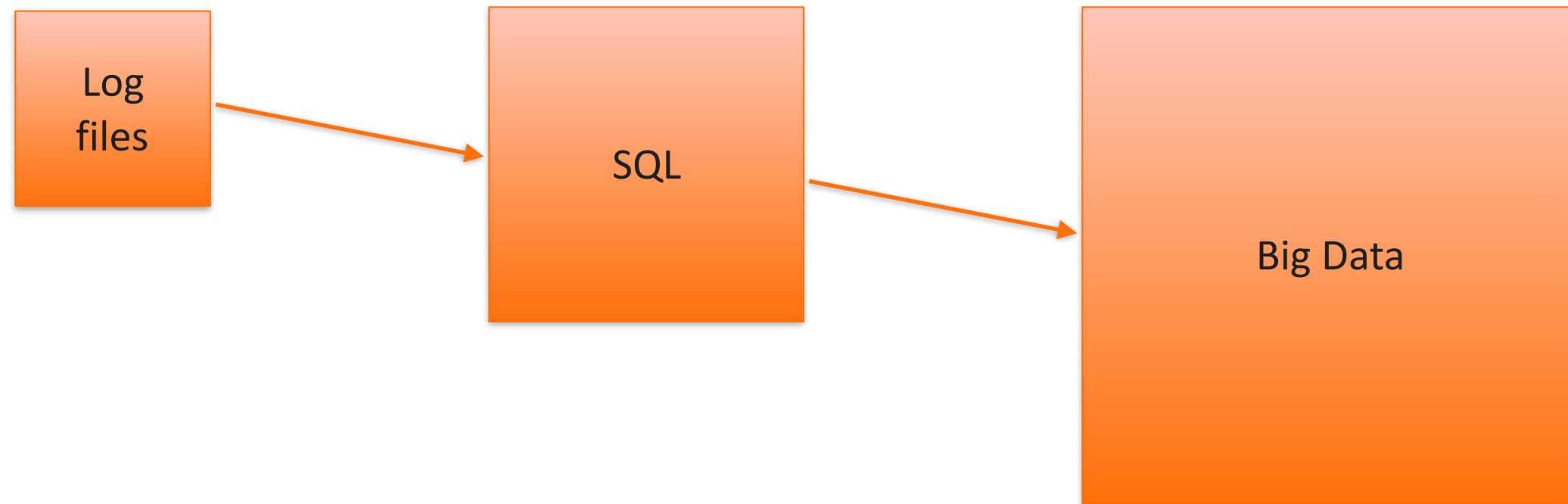


# Why is moving data *effectively* hard?

- ◆ 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

# Connecting A to B to C

Easy enough with Bash scripts, Ruby/Python/Groovy, etc.

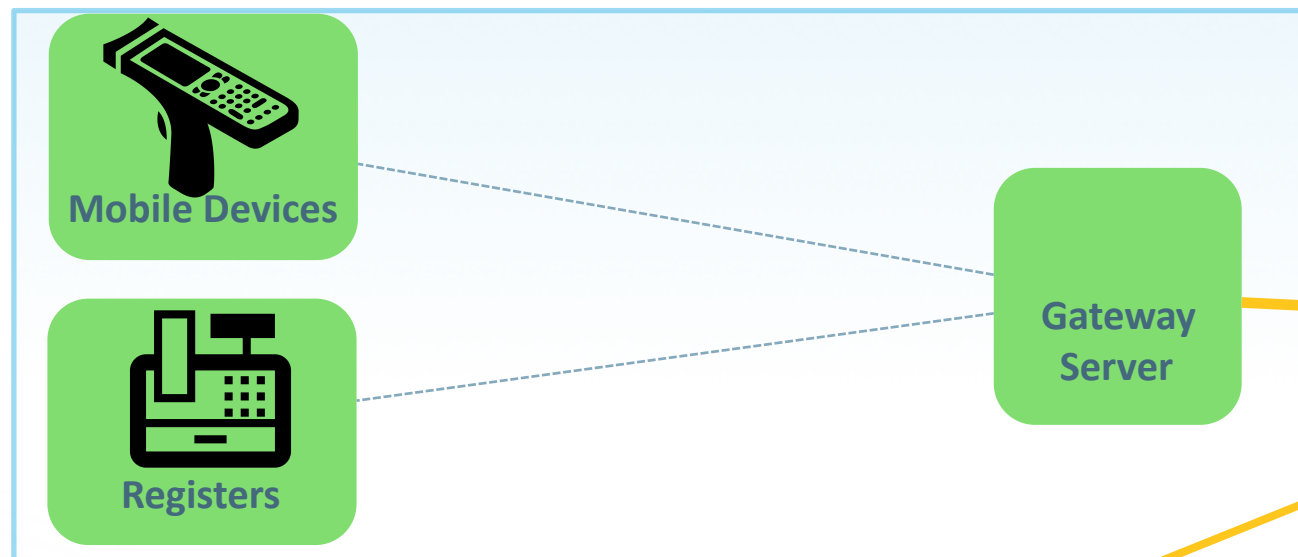




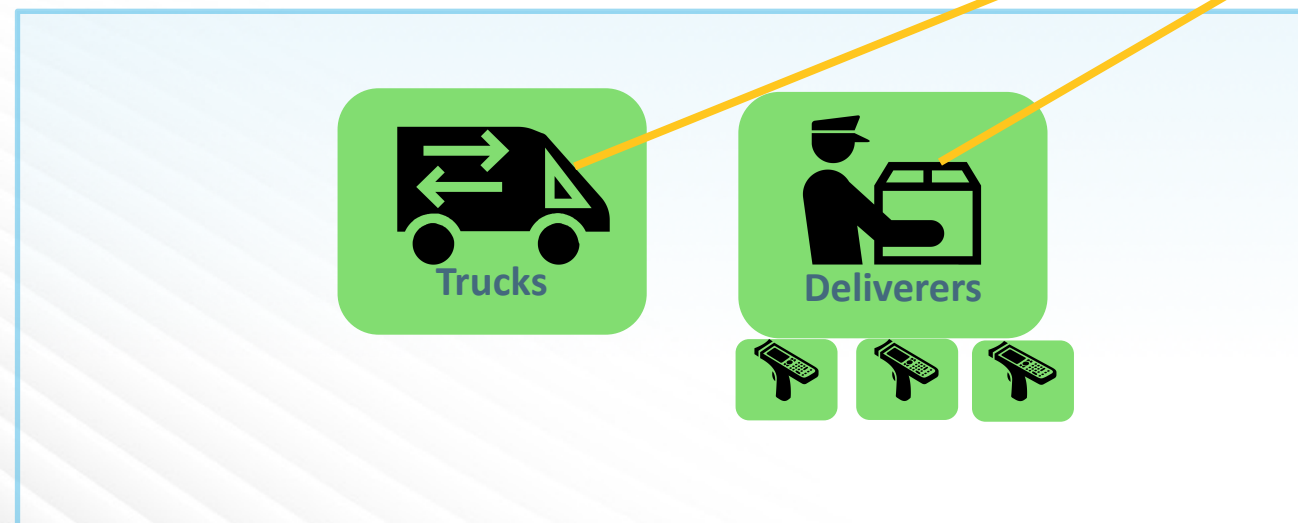
# Let's Connect Lots of As to Bs to As to Cs to Bs to Δs to Cs to φs

Let's consider the needs of a courier service

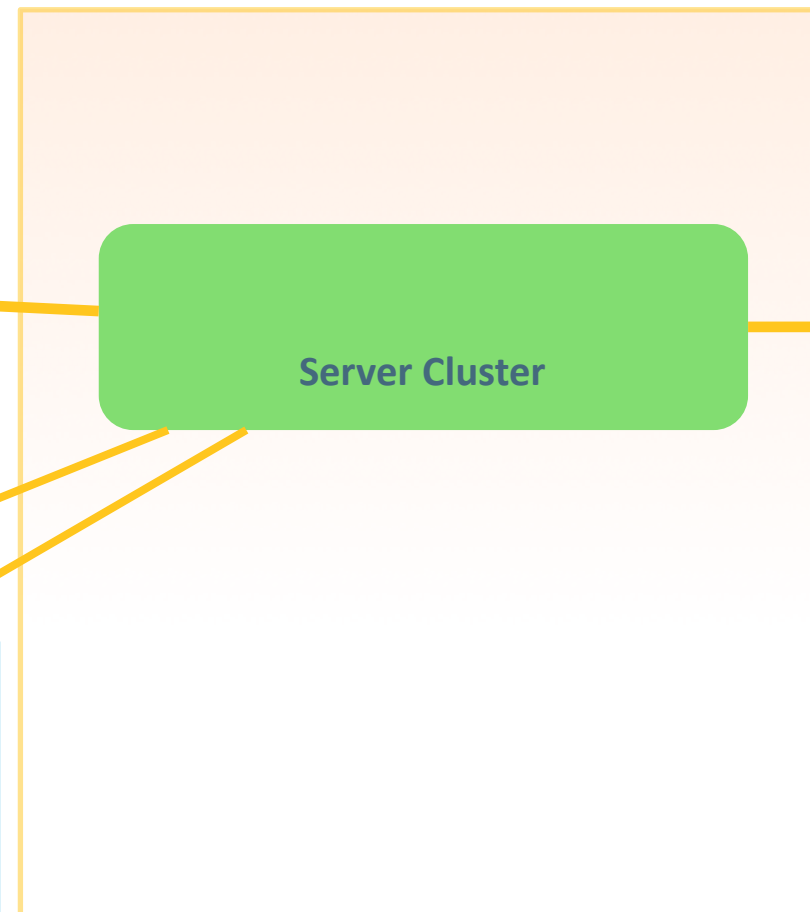
Physical Store



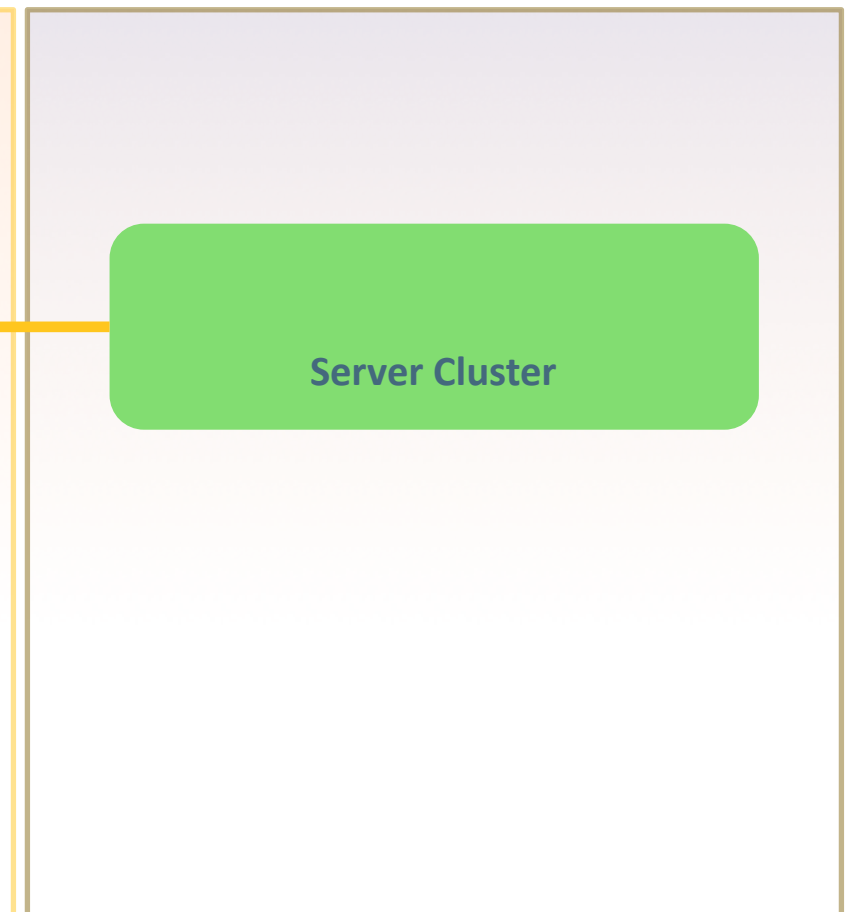
On Delivery Routes



Distribution Center



Core Data Center at HQ



Delivery Truck: Creative Stall, <https://thenounproject.com/creativestall/>

Deliverer: Rigo Peter, <https://thenounproject.com/rigo/>

Cash Register: Sergey Patutin, <https://thenounproject.com/bdesign.by/>

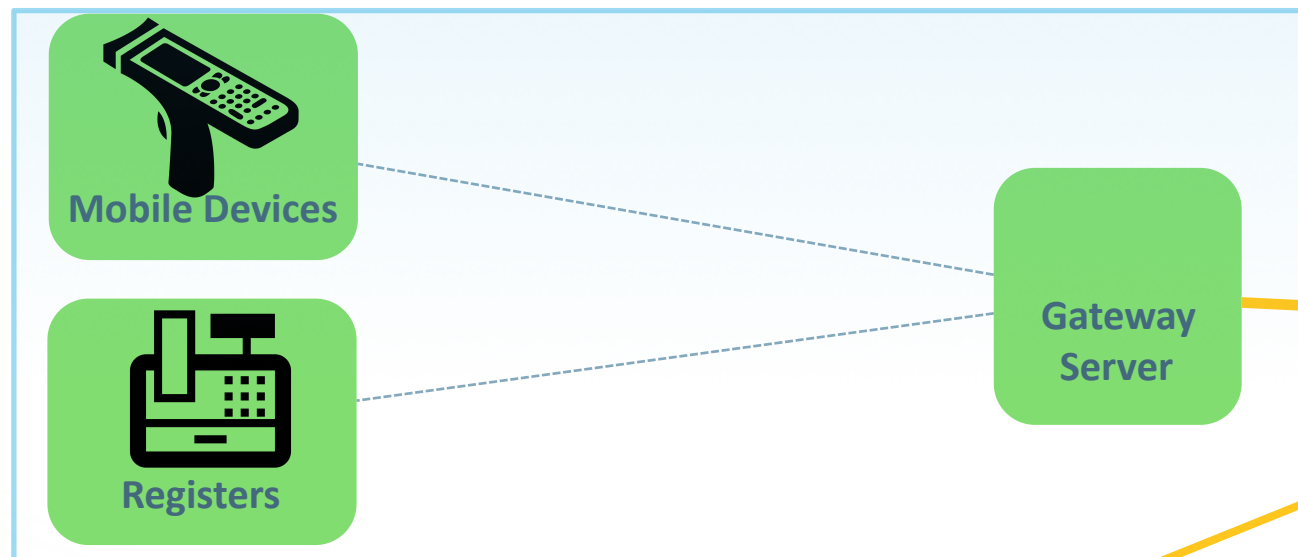
Hand Scanner: Eric Pearson, <https://thenounproject.com/epearson001/>



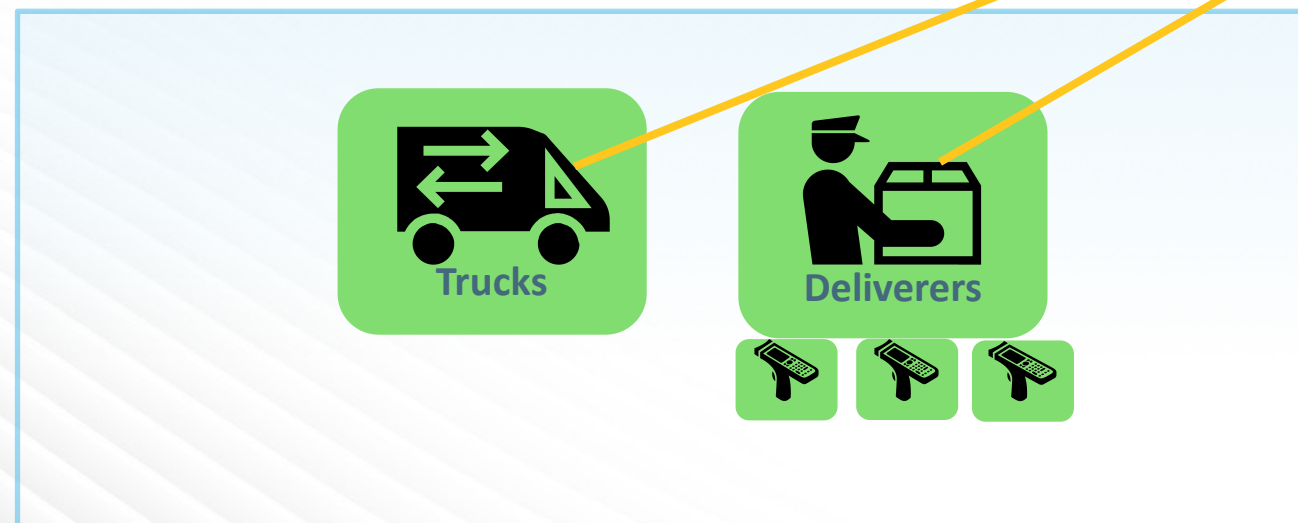
# Great! I am collecting all this data! Let's use it!

## Finding our needles in the haystack

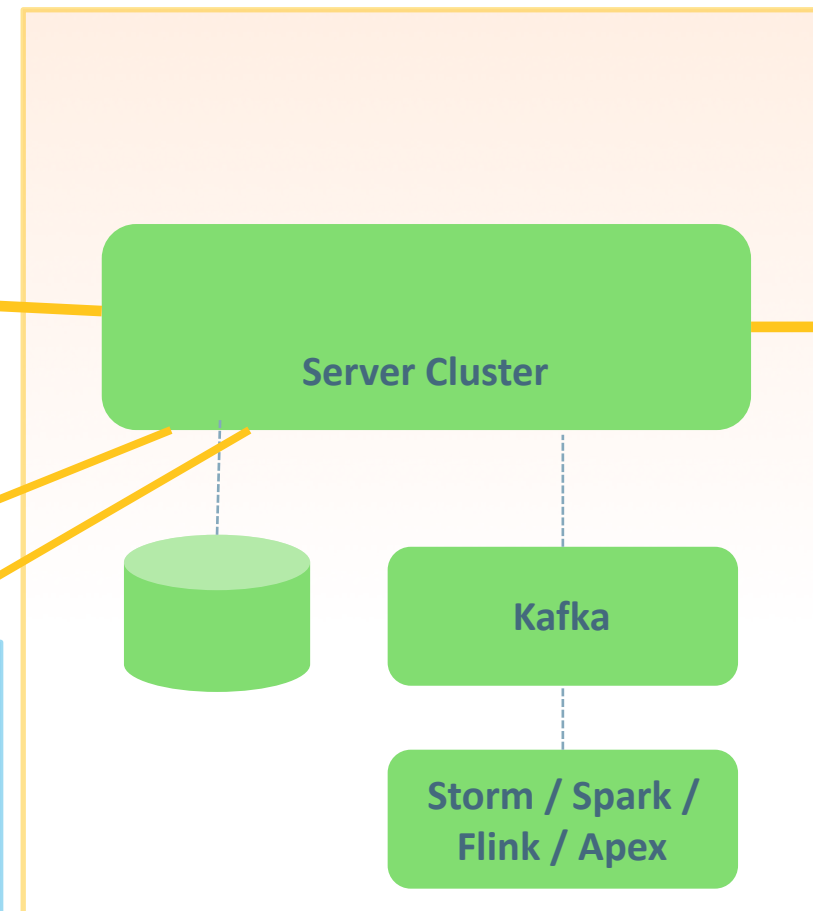
### Physical Store



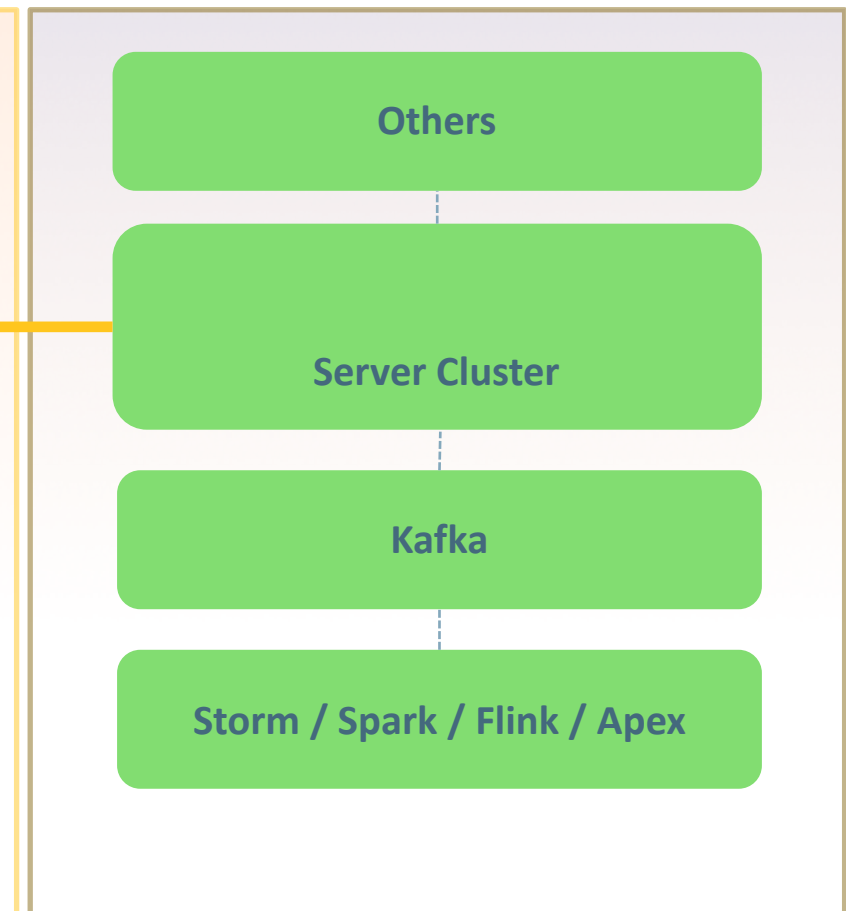
### On Delivery Routes



### Distribution Center

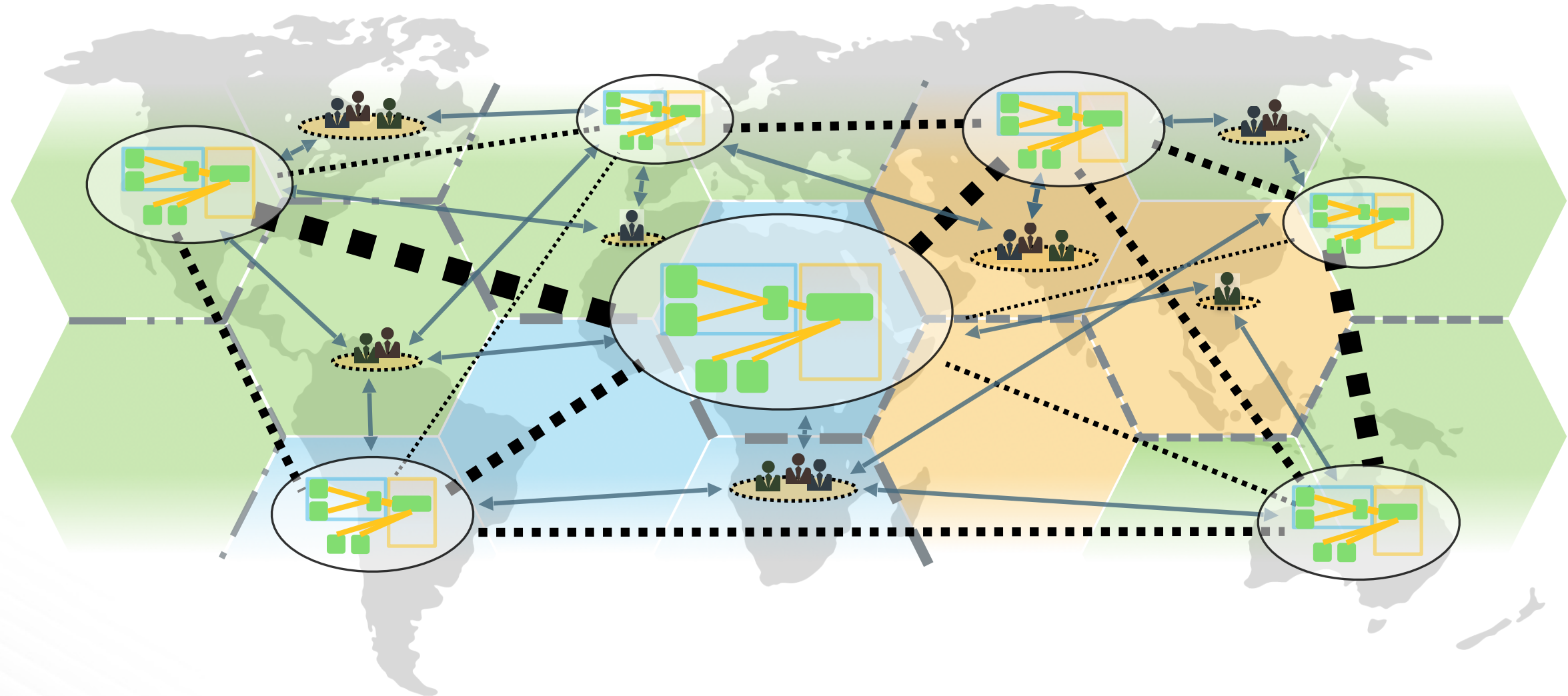


### Core Data Center at HQ



# Let's Connect Lots of As to Bs to As to Cs to Bs to $\Delta$ s to Cs to $\varphi$ s

Raise your hand if you want to maintain Python scripts for the rest of your life



# Agenda



What is dataflow and what are the challenges?

Apache NiFi

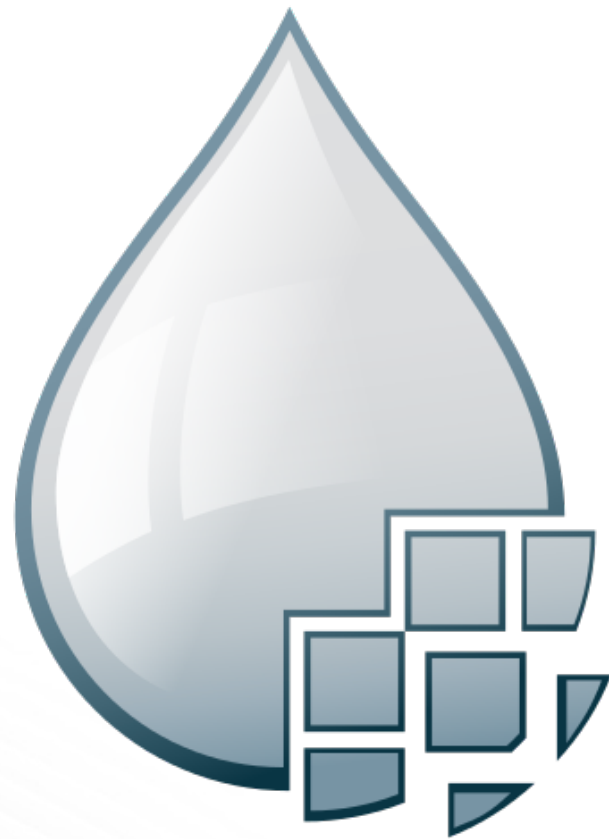
Apache MiNiFi

Architecture

Community

# 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-tenant security
- Designed for extension
- Clustering

# 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.



## 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

### PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

**2013-02-27**


THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013. II. 27. 27½-13 2013.158904109

MMXIII-II-XXVII MMXIII <sup>LVII</sup>/<sub>CCCLXV</sub> 1330300800

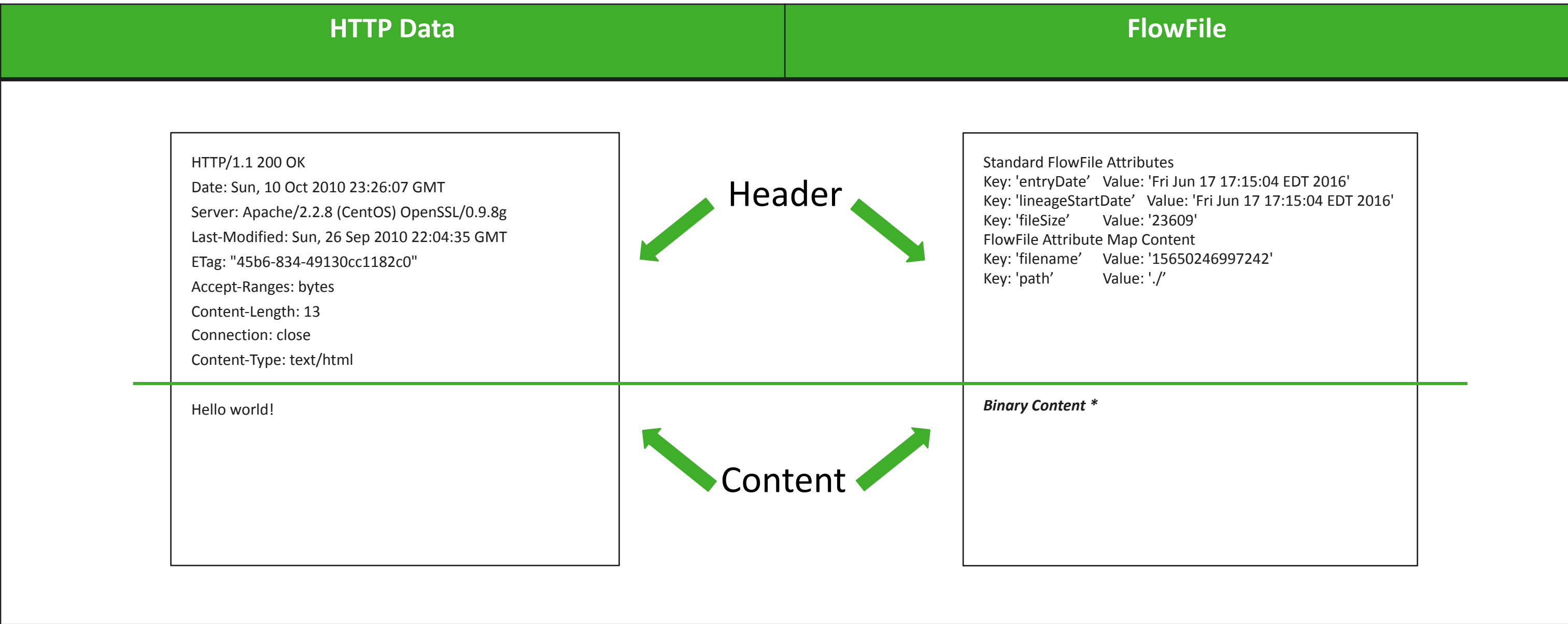
$((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$  2013 

10/11011/1101 02/27/20/13  $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 \end{matrix}$

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

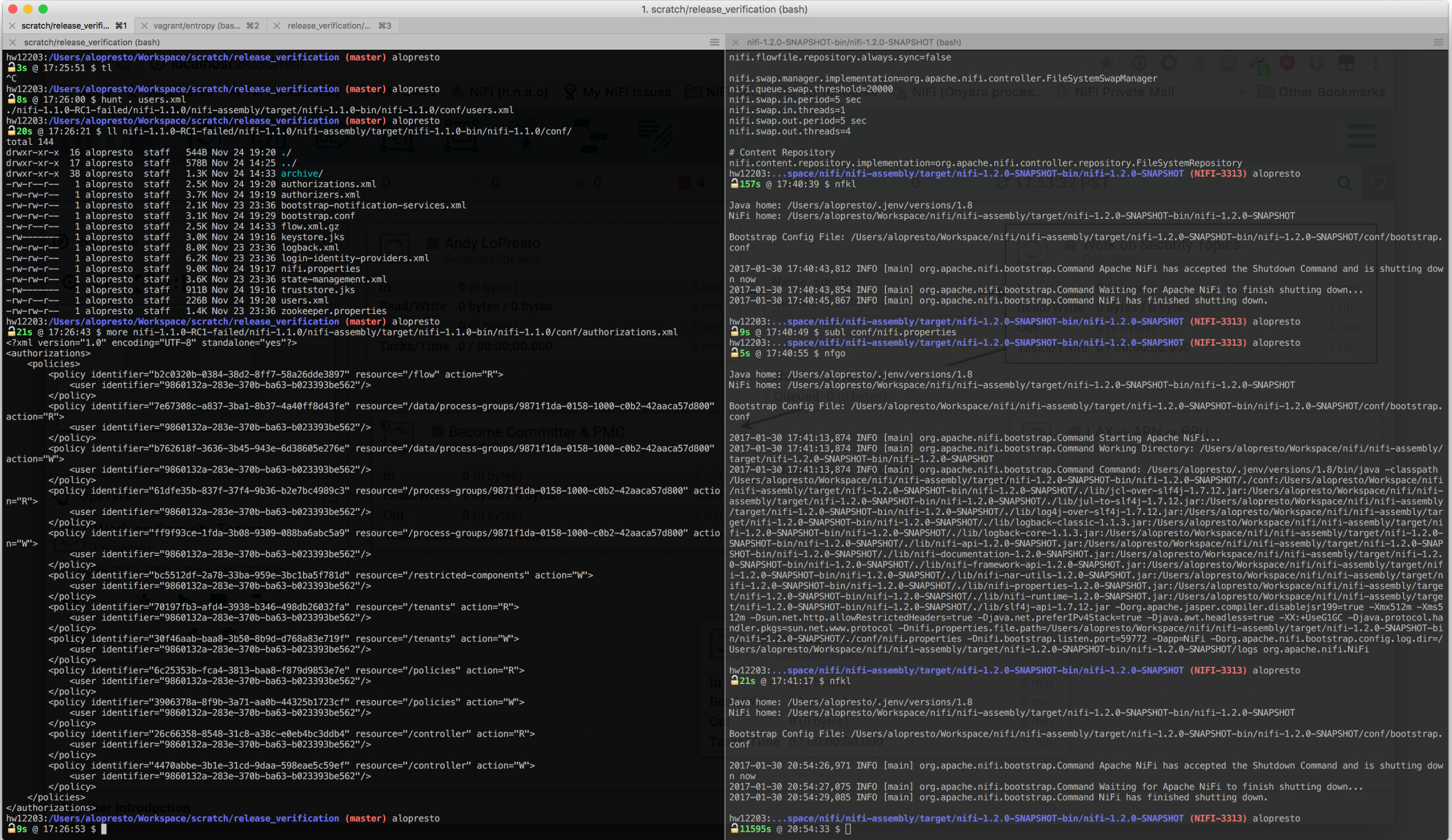


# FlowFiles are like HTTP data



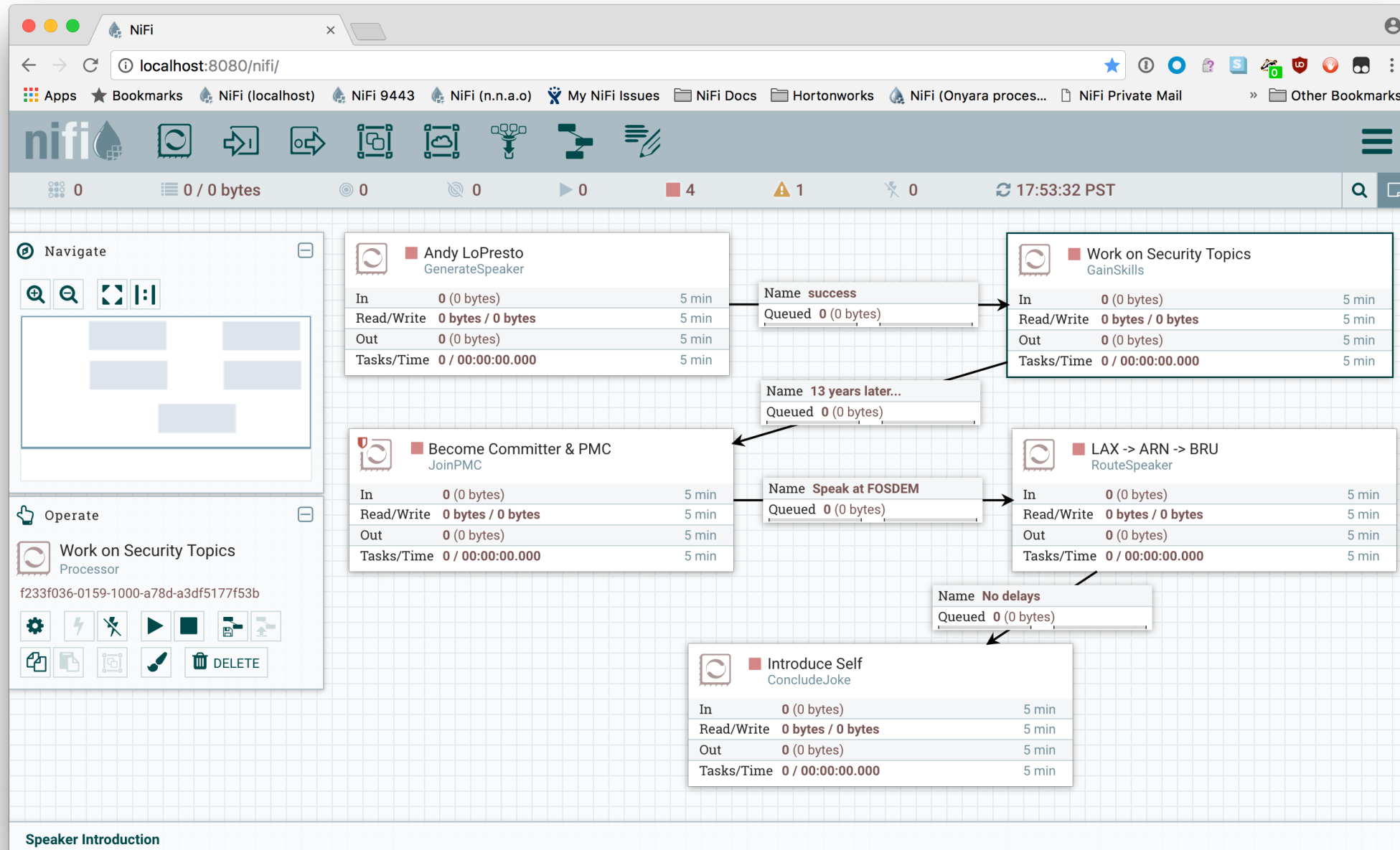
# User Interface

Less of this...

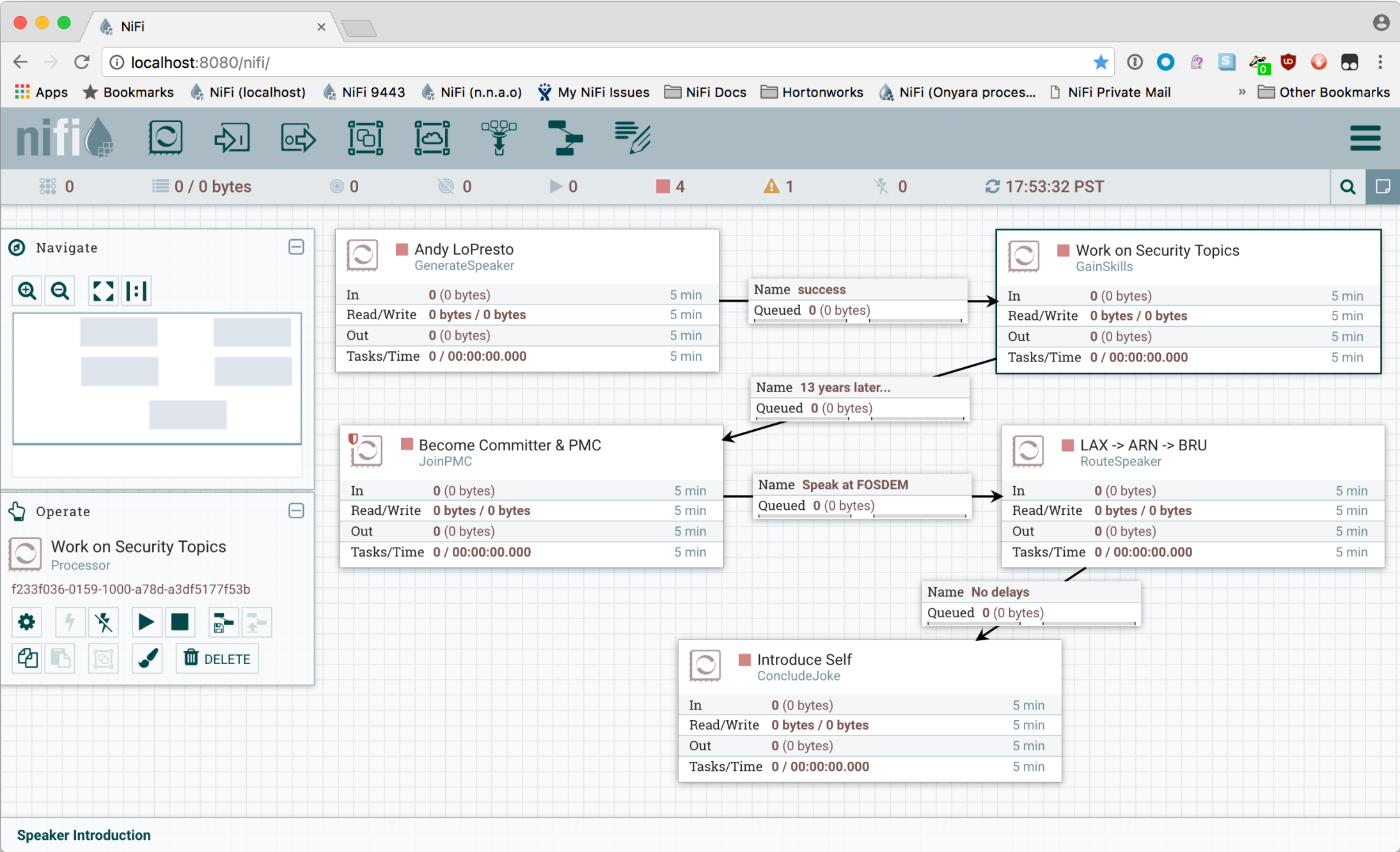


# User Interface

Less of this... .. more of this



# User Interface





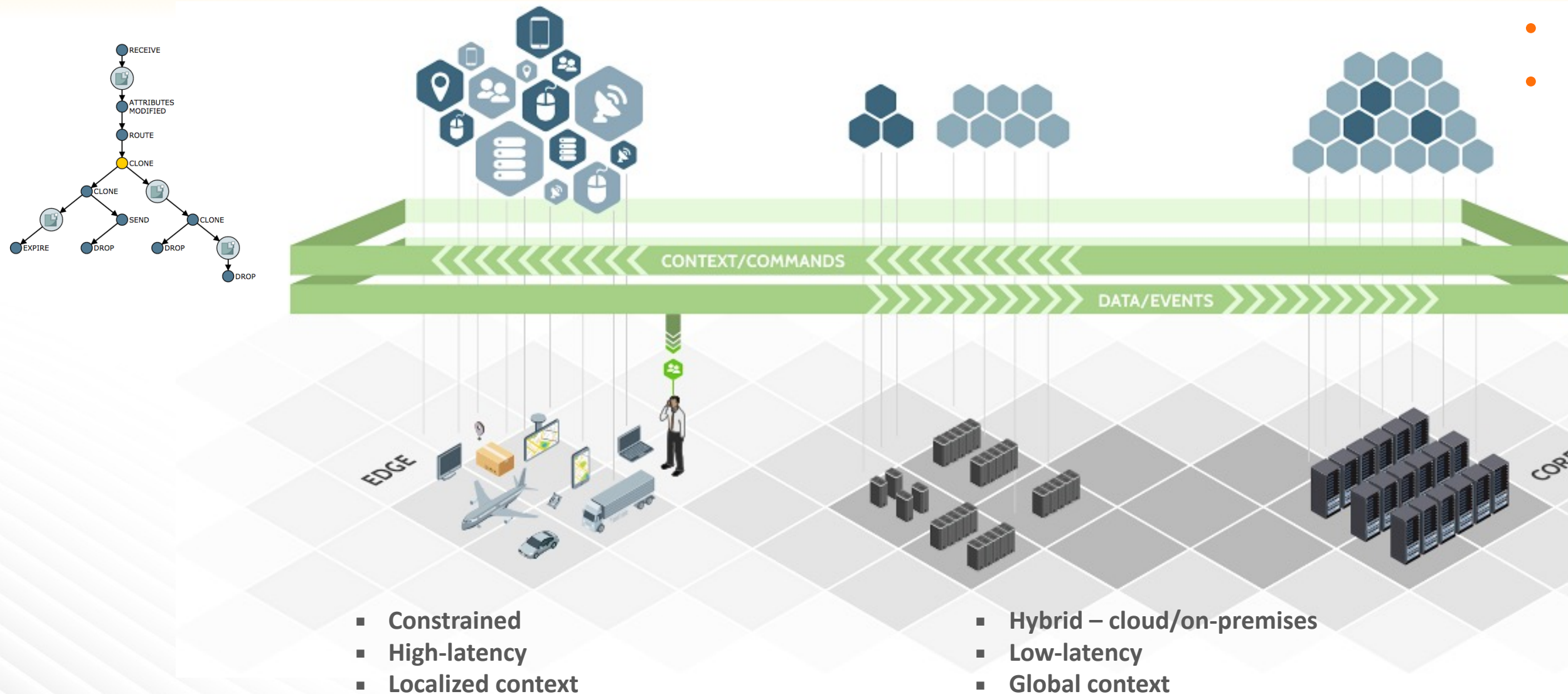
# Data Provenance

Origin – attribution  
Replay – recovery

Evolution of topologies  
Long retention

## Types of Lineage

- Event
- Configuration



# Deeper Ecosystem Integration: 180+ 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.



# Agenda

What is dataflow and what are the challenges?

Apache NiFi



Apache MiNiFi

Architecture

Community



## Apache NiFi Subproject: MiNiFi

- Let me get the key parts of NiFi close to 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 as close to where data is born and is a guest on that device or system
  - IoT
  - Connected car
  - Legacy hardware
  - S2S client libs



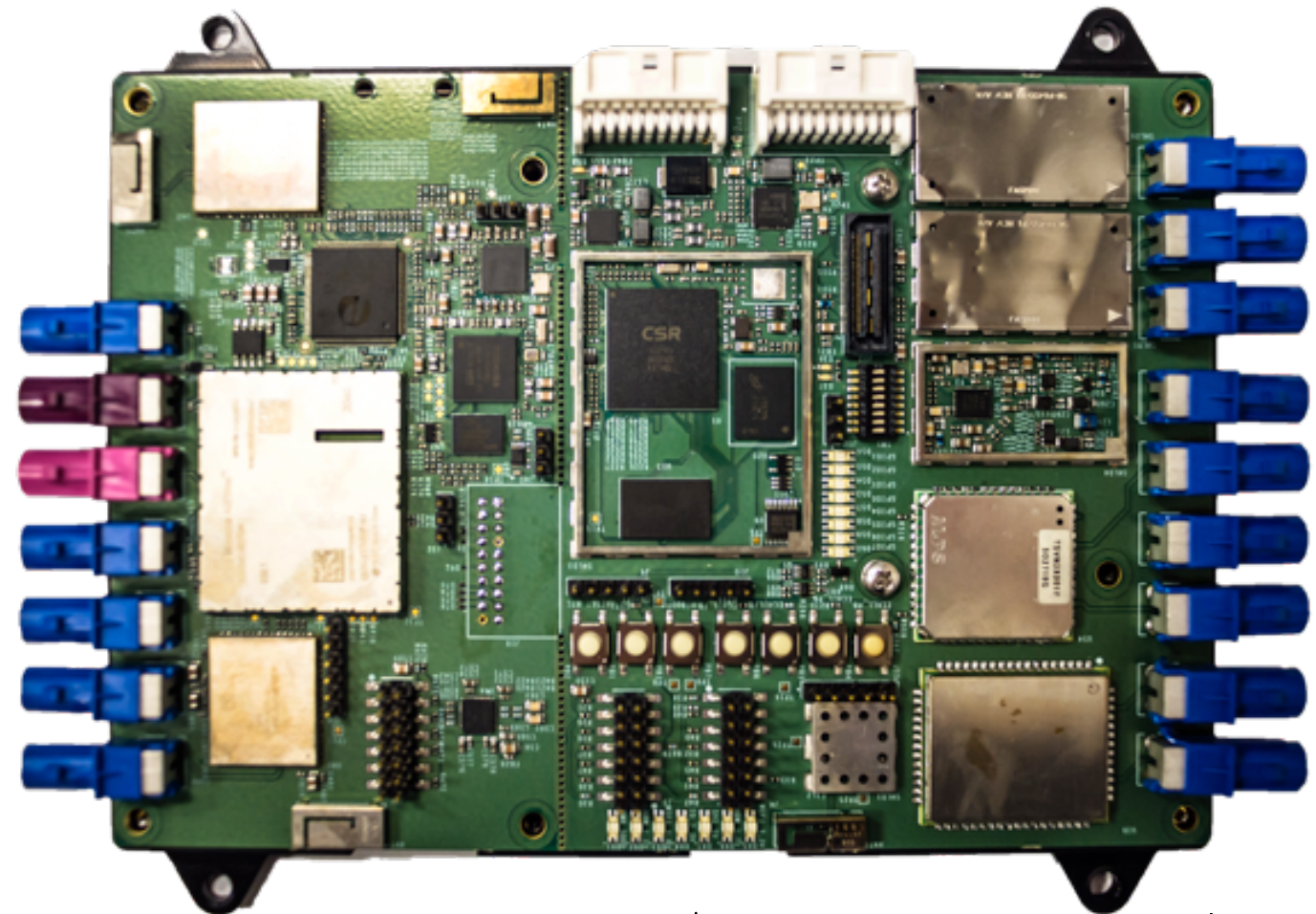
# Why build MiNiFi?

- ◆ NiFi is big
  - ◆ 1.1.0 release is 726 MB compressed
  - ◆ Can be modified to run in restricted environments, but requires manual surgery
  - ◆ Provides UI, provenance query, etc.
  - ◆ Runs on dedicated machines/clusters — “owns the box”
- ◆ MiNiFi lives at the edge
  - ◆ No UI
  - ◆ 0.1.0 Java binary is 45 MB, C++ binary is 746 KB
  - ◆ “Good guest”

# What does MiNiFi provide?

- ◆ Data tagging/provenance
- ◆ Governance from edge (geopolitical restrictions)
- ◆ Security (encryption, certificate-based authentication)
- ◆ Low latency (immediate reactions & decision-making)

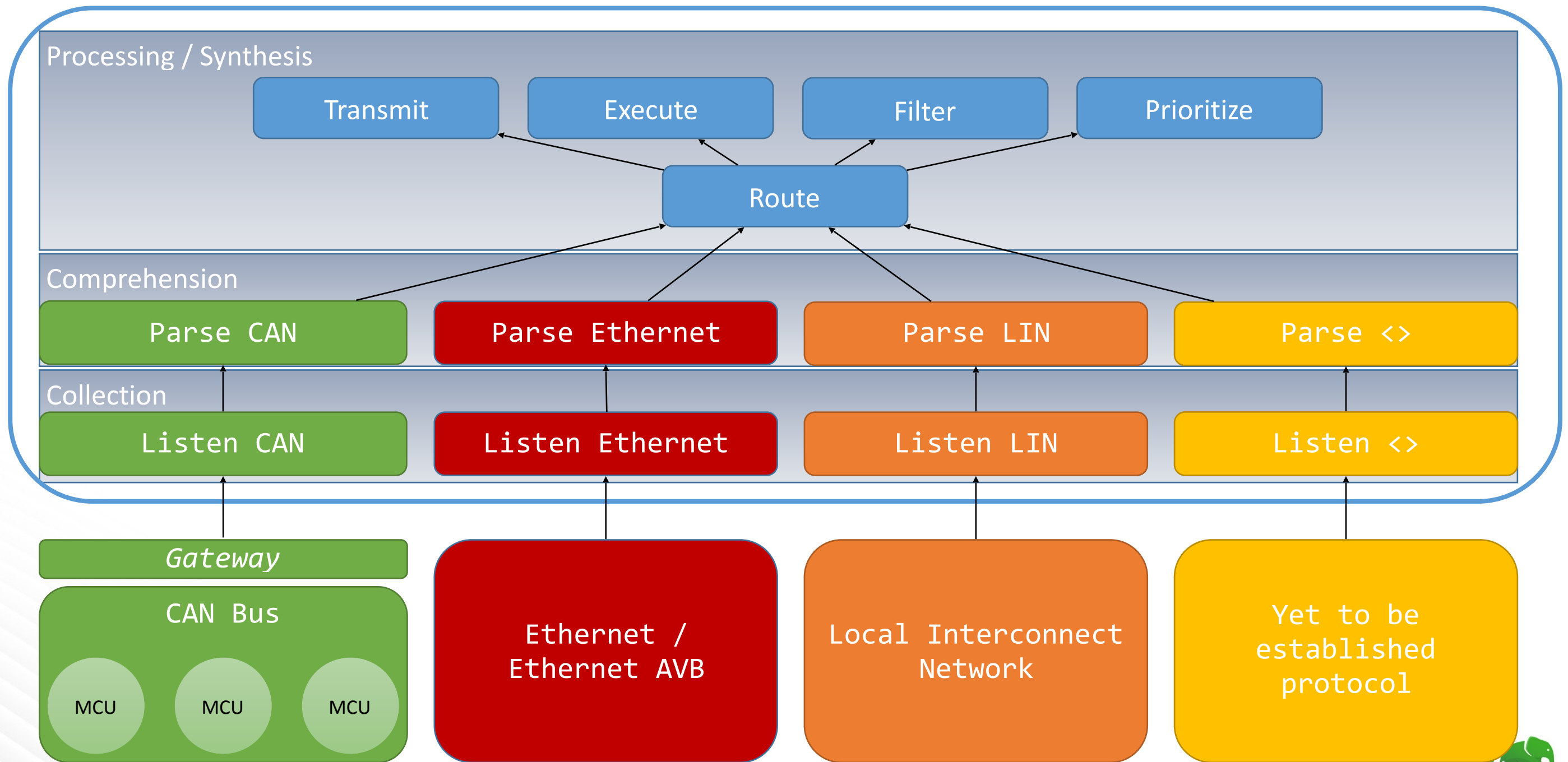
## Connected Car Reference Platform Box



Connectivity Card

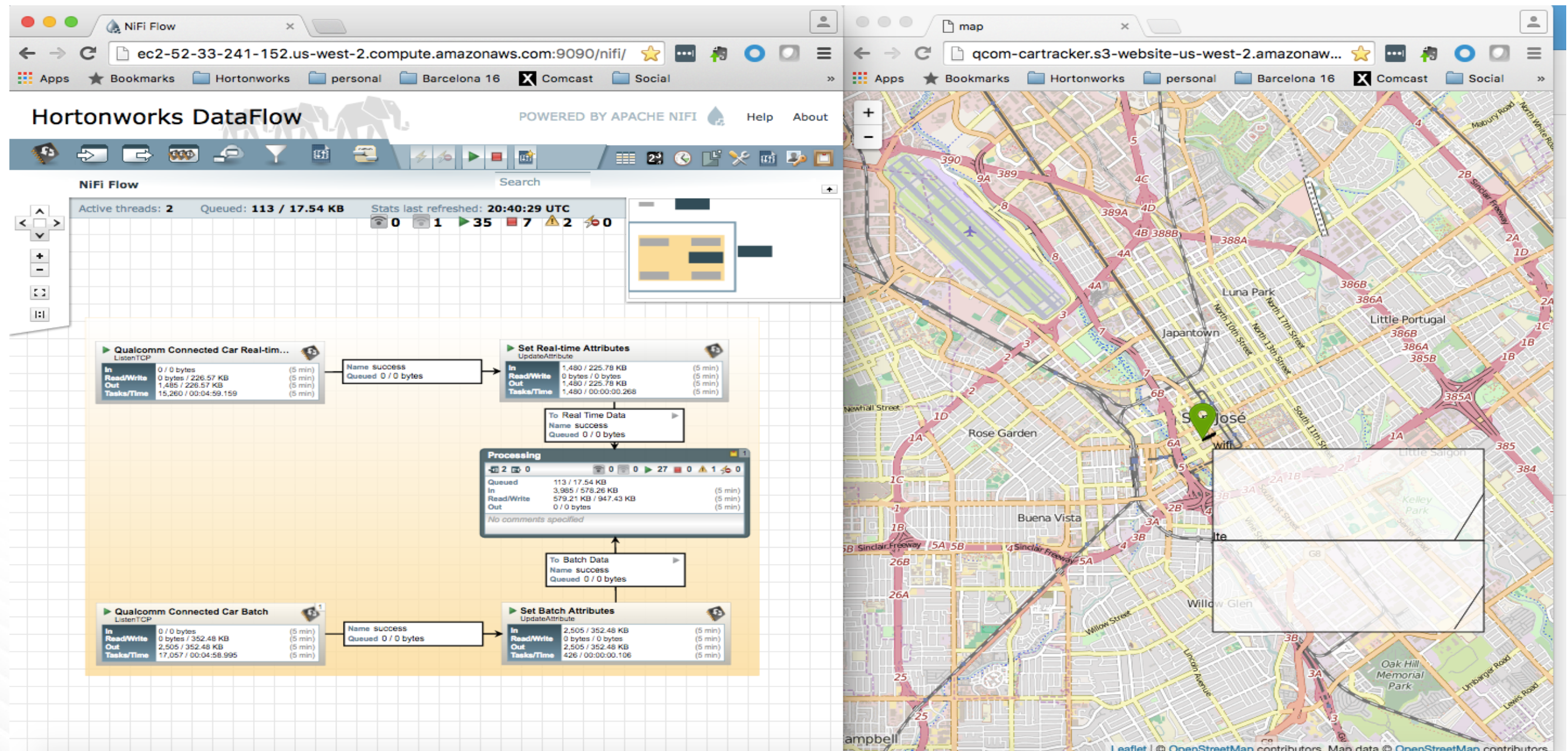
Tuner + DSRC Card

# MiNiFi on a Connected Car





# MiNiFi on a Connected Car



# MiNiFi Feature Proposals

- ◆ Flow Versioning
  - ◆ Develop flows for class of MiNiFi instances
- ◆ Command & Control (C2) API
  - ◆ FileChangeIngestor
  - ◆ RestAPIIngestor
  - ◆ PullHTTPIngestor

# Agenda

What is dataflow and what are the challenges?

Apache NiFi

Apache MiNiFi

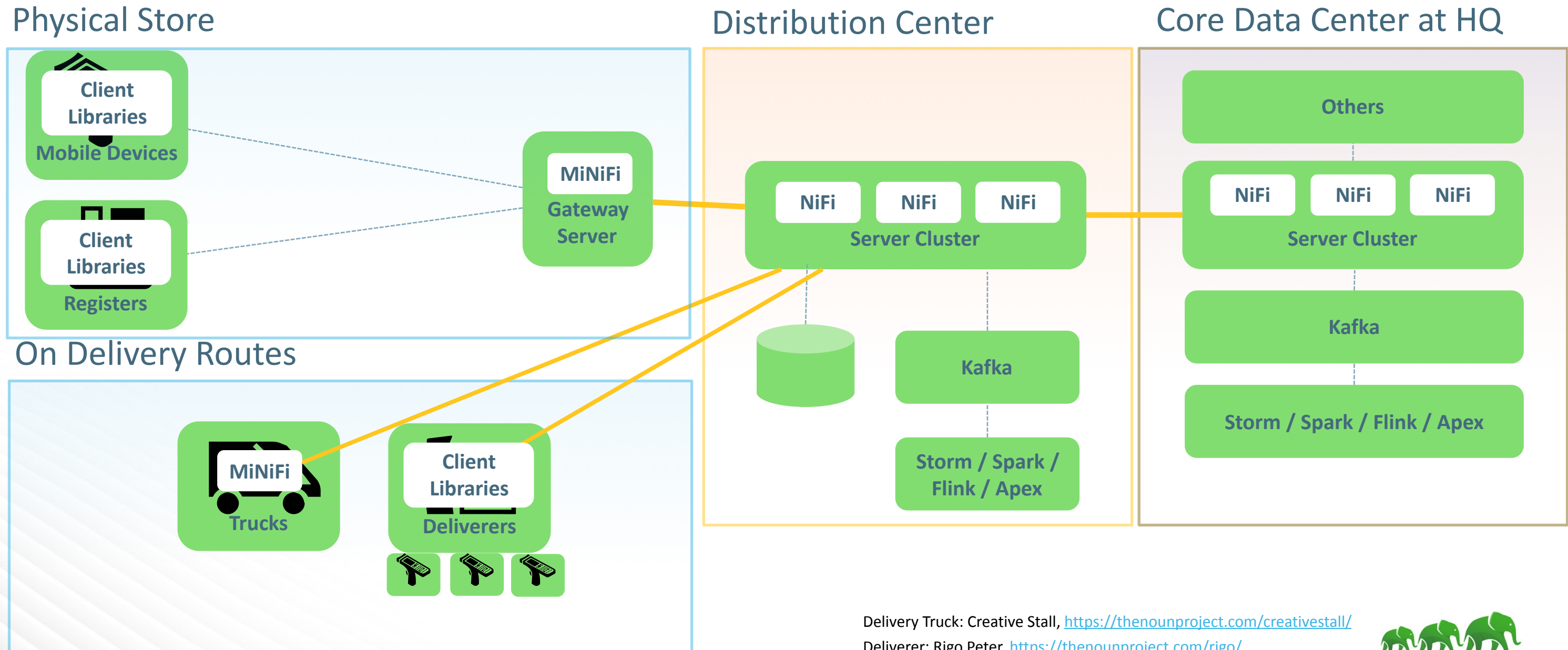


Architecture

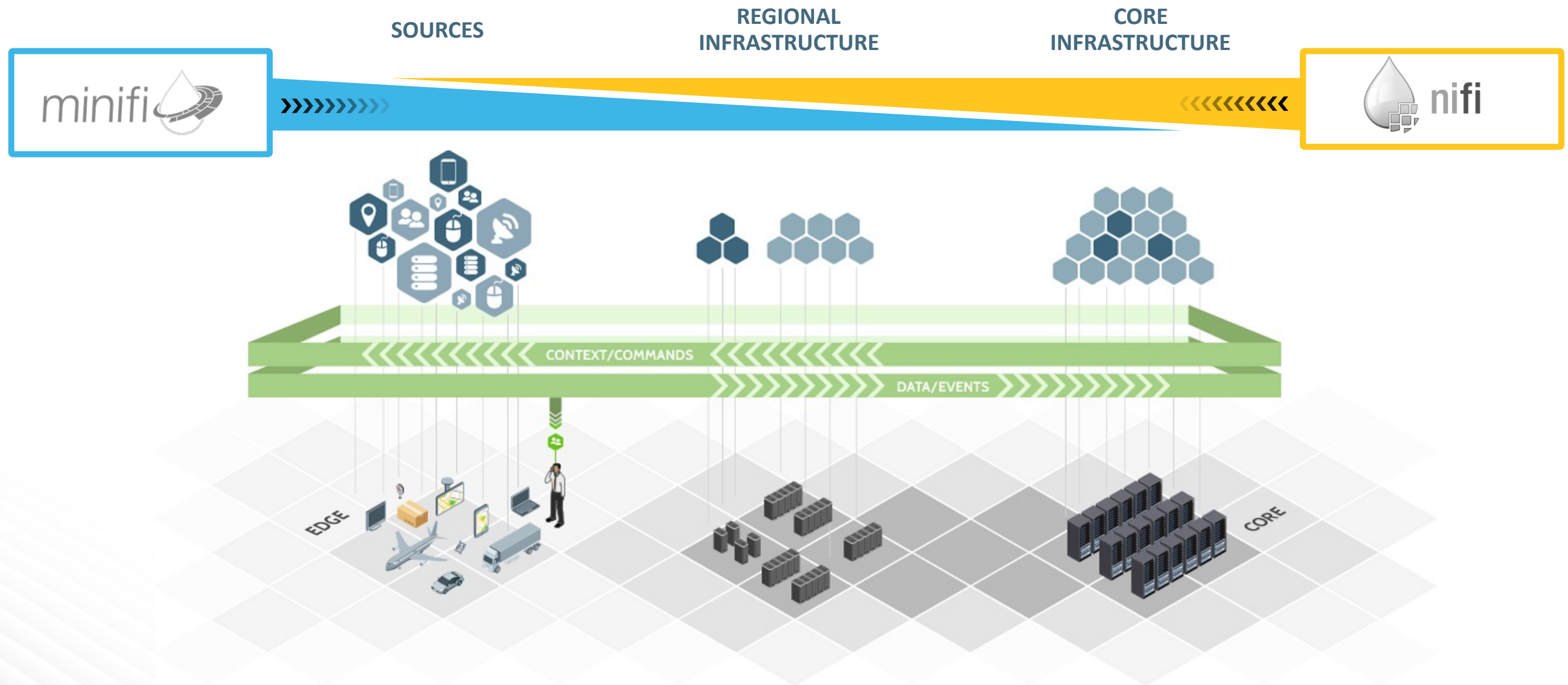
Community



# Let's revisit our courier service from the perspective of NiFi



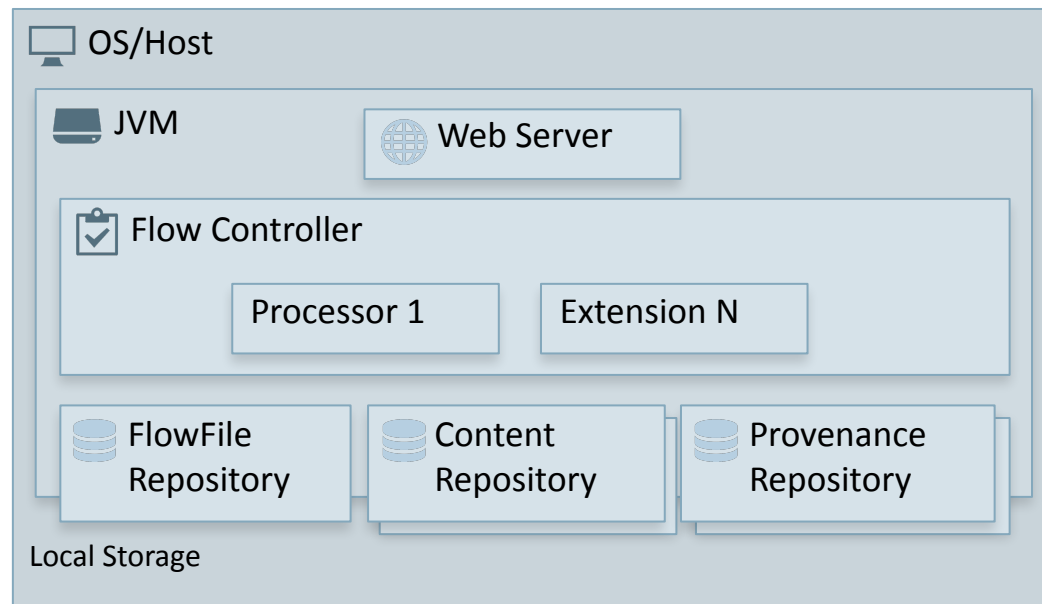
# Apache NiFi Managed Dataflow



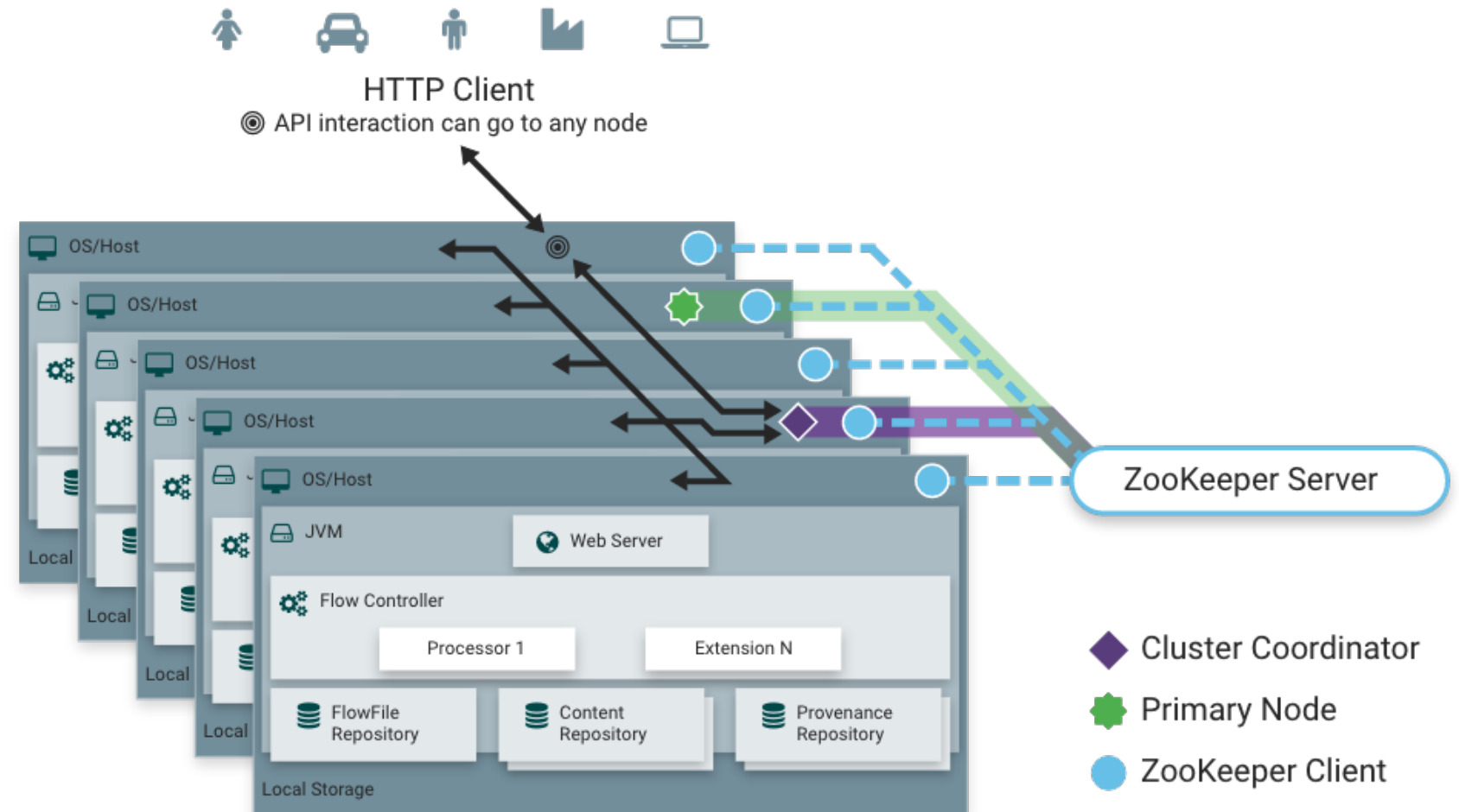
## Extension / Integration Points

NiFi Term	Description
Flow File Processor	Push/Pull behavior. Custom UI
Reporting Task	Used to push data from NiFi to some external service (metrics, provenance, etc.)
Controller Service	Used to enable reusable components / shared services throughout the flow
REST API	Allows clients to connect to pull information, change behavior, etc.

# Architecture



Standalone

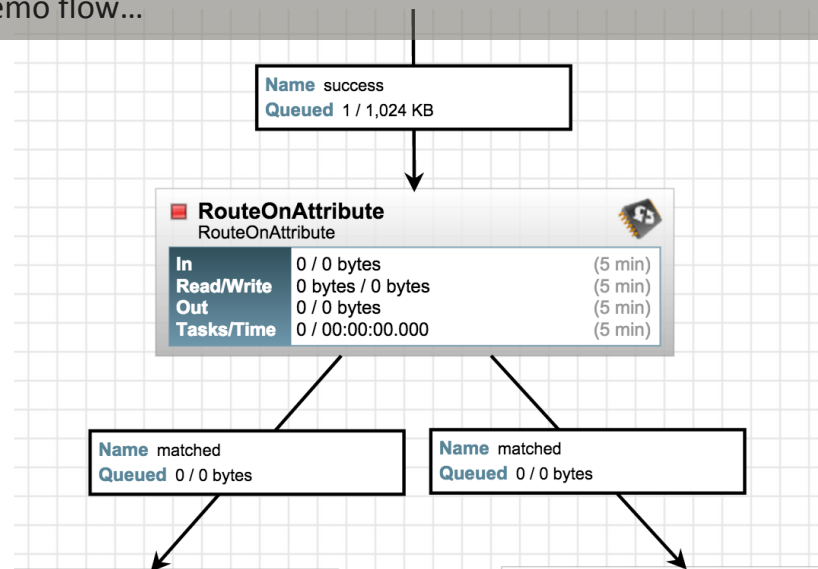


Cluster

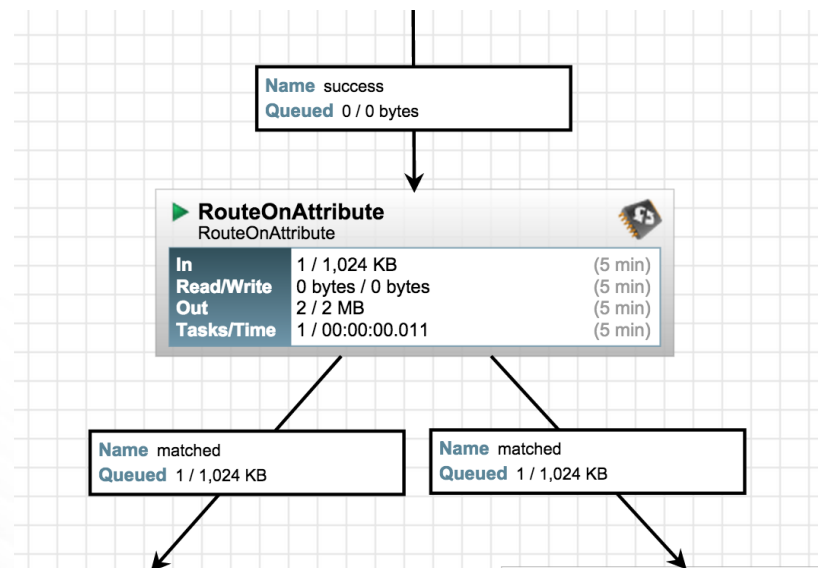
# NiFi Architecture – Repositories - Pass by reference

Excerpt of demo flow...

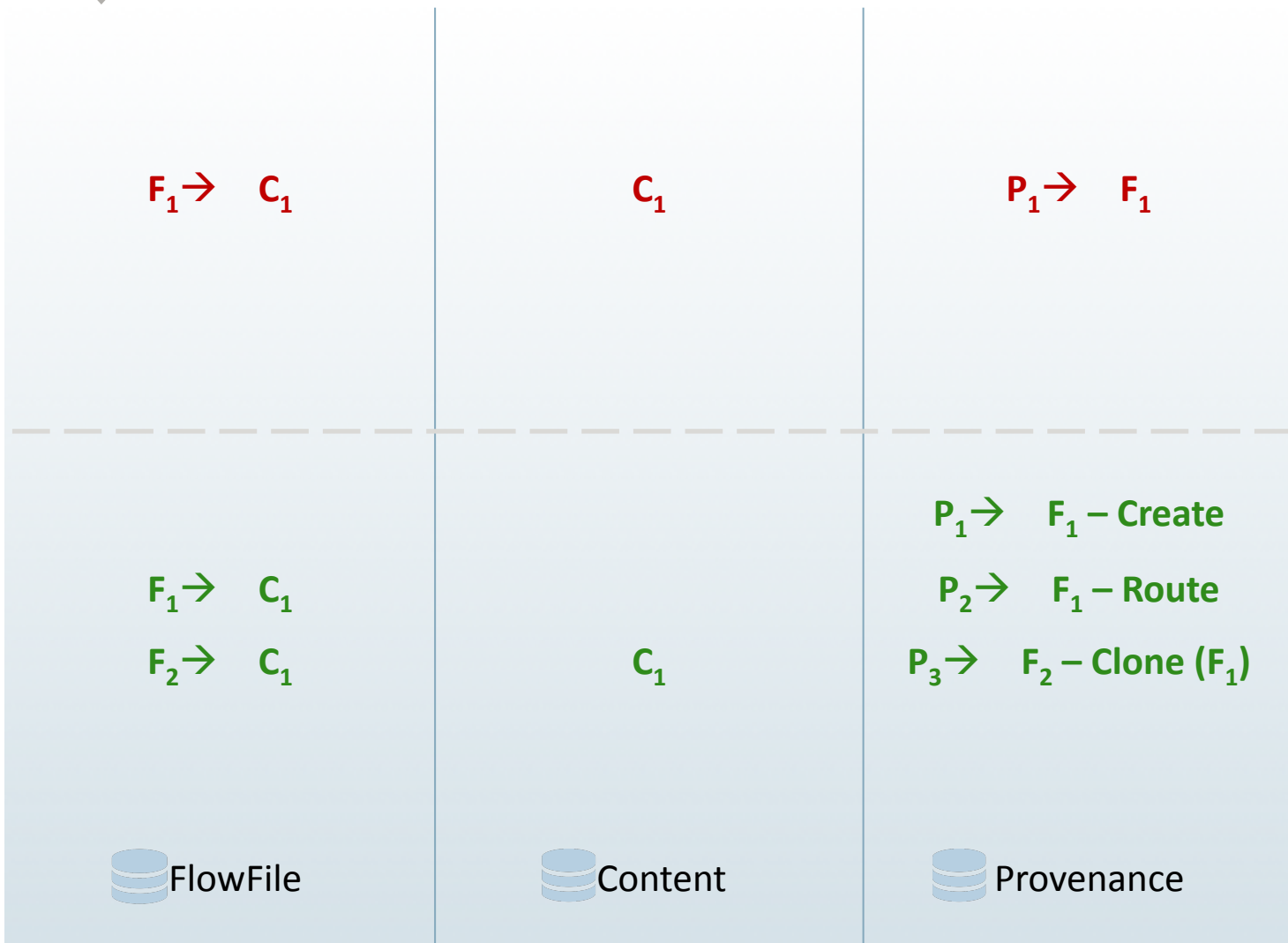
**BEFORE**



**AFTER**



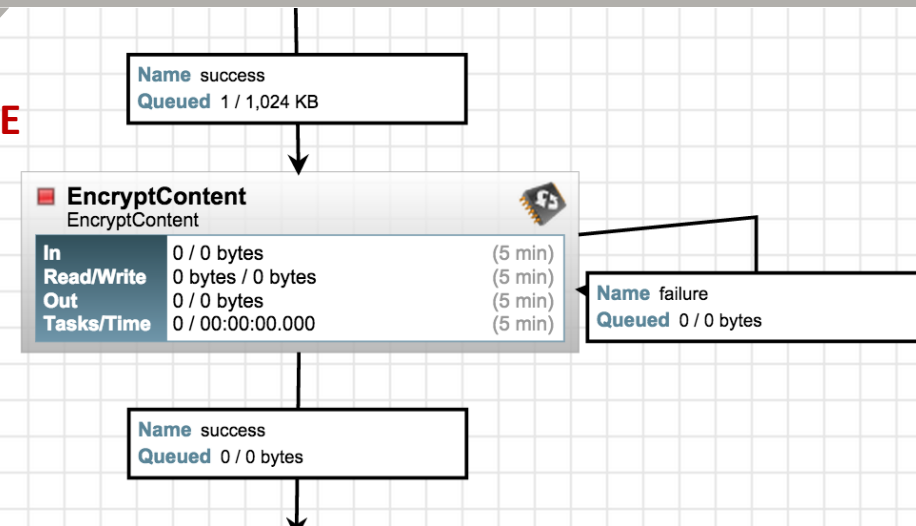
What's happening inside the repositories...



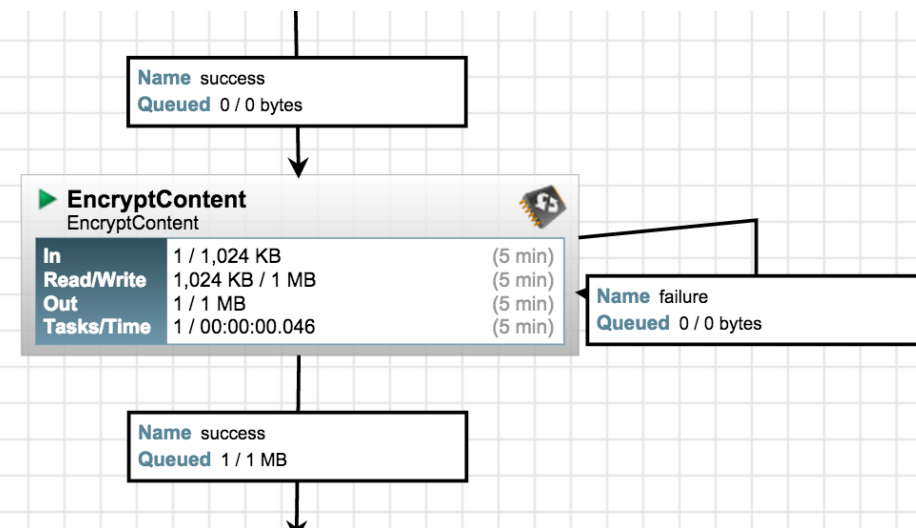
# NiFi Architecture – Repositories – Copy on Write

Excerpt of demo flow...

BEFORE



AFTER



What's happening inside the repositories...

$F_1 \rightarrow C_1$

$C_1$

$P_1 \rightarrow F_1 - \text{CREATE}$

$F_1 \rightarrow C_1$   
 $F_{1.1} \rightarrow C_2$

$C_1$  (plaintext)  
 $C_2$  (encrypted)

$P_1 \rightarrow F_1 - \text{CREATE}$   
 $P_2 \rightarrow F_{1.1} - \text{MODIFY}$

FlowFile

Content

Provenance



# Agenda

What is dataflow and what are the challenges?

Apache NiFi

Apache MiNiFi

Architecture




Community



# 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 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

# Healthy Community

 **apache / nifi**  
mirrored from [git://git.apache.org/nifi.git](https://git.apache.org/nifi.git)

Unwatch ▾101

★ Unstar446

🍴 Fork416

<> Code

🔗 Pull requests96

📁 Projects0

📶 Pulse

📊 Graphs

Mirror of Apache NiFi

📄 3,176 commits

🌿 33 branches

🏷 40 releases

👤 106 contributors

## Unresolved: By Priority

Priority	Issues	Percentage
🛑 Blocker	6	■ 1%
🔴 Critical	19	■ 2%
🔴 Major	495	■ 45%
🟢 Minor	516	■ 47%
🟢 Trivial	71	■ 6%

[View Issues](#)

## Status Summary

Status	Issues	Percentage
Open	1008	■ 30%
In Progress	20	■ 1%
Reopened	10	
Resolved	2234	■ 66%
Closed	63	■ 2%
Patch Available	69	■ 2%

[View Issues](#)



# Learn more and join us

## Apache NiFi site

<https://nifi.apache.org>

## Subproject MiNiFi site

<https://nifi.apache.org/minifi/>

## Subscribe to and collaborate at

[dev@nifi.apache.org](mailto:dev@nifi.apache.org)

[users@nifi.apache.org](mailto:users@nifi.apache.org)

## Submit Ideas or Issues

<https://issues.apache.org/jira/browse/NIFI>

## Follow us on Twitter

[@apachenifi](https://twitter.com/apachenifi)



# Thank You

I'm sticking around for discussions/questions

 @yolohey / @apachenifi

[alopresto@apache.org](mailto:alopresto@apache.org)

PGP: 70EC B3E5 98A6 5A3F D3C4 BACE 3C6E F65B **2F7D EF69**