# Apache NiFi and MiNiFi: Edge to Core

**Andy LoPresto - @yolopey**

Apache NiFi PMC

DataWorks Summit 2017 - Sydney

19 Sep 2017

HORTONWORKS®
POWERING THE FUTURE OF DATA™

# Agenda

*What is dataflow and what are the challenges?*

*Apache NiFi*

IoT Challenges

Apache MiNiFi

Exploration

Community




HORTONWORKS

# Agenda

*What is dataflow and what are the challenges?*

*Apache NiFi*

IoT Challenges

Apache MiNiFi

Exploration

Community

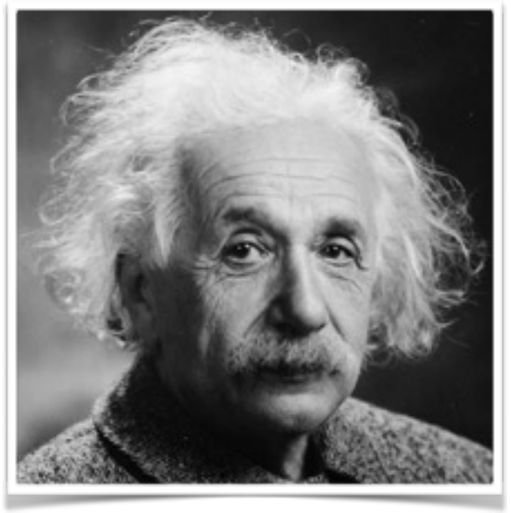



HORTONWORKS

# Gauging Audience Familiarity With NiFi

*"What's a NeeFee?"*

No experience with dataflow
No experience with NiFi



*"I can pick this up pretty quickly"*

Some experience with dataflow
Some experience with NiFi

*"I refactored the Ambari integration endpoint to allow for mutual authentication TLS during my coffee break"*

Forgotten more about NiFi than most of us will ever know



HORTONWORKS

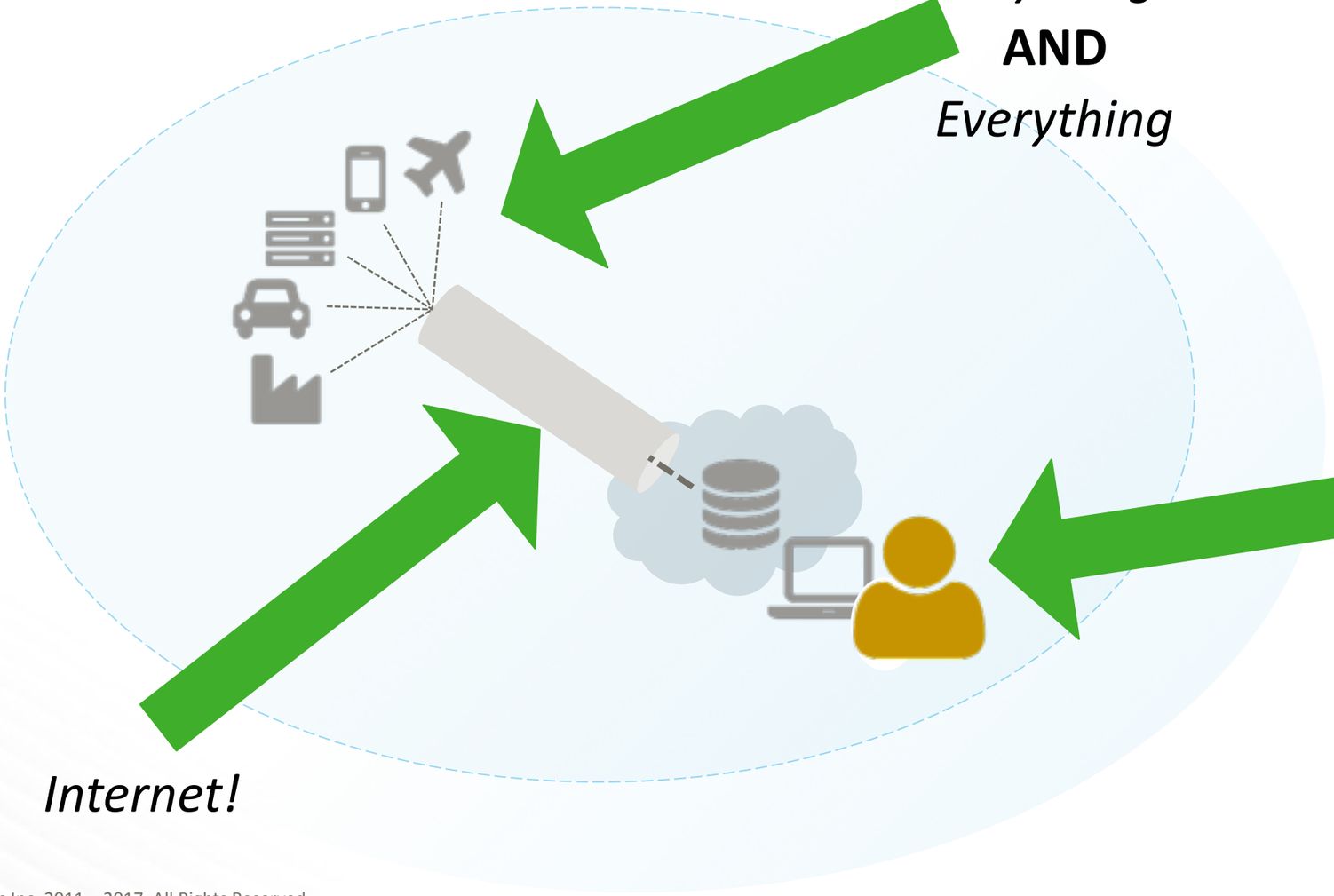
# Let's Connect A to B

**Producers A.K.A Things**

*Anything*

**AND**

*Everything*



**Consumers**

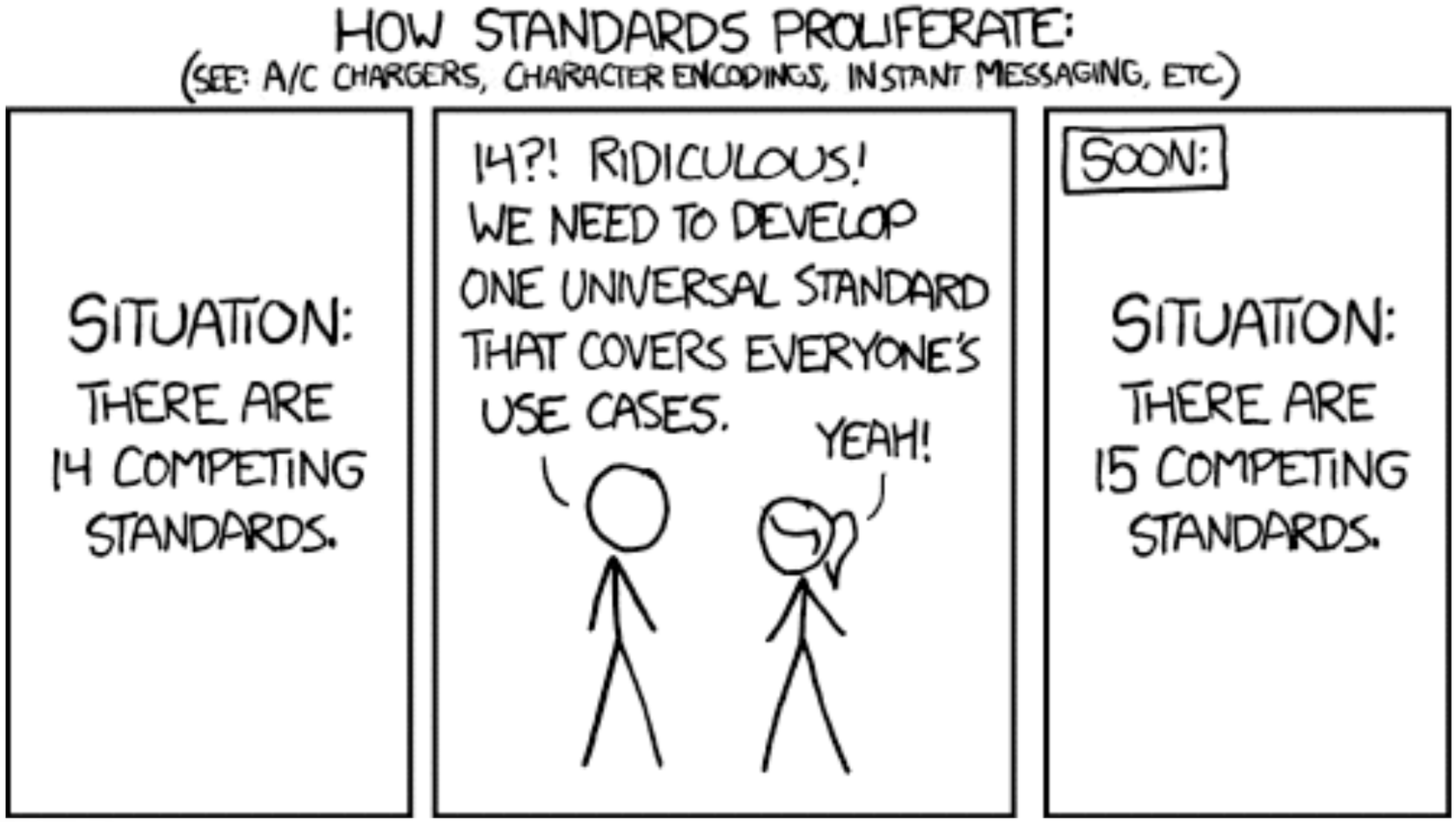
- *User*
- *Storage*
- *System*
- ***…More** Things*

*Internet!*




HORTONWORKS

# Moving data *effectively* is hard




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




HORTONWORKS

# Why is moving data *effectively* hard?

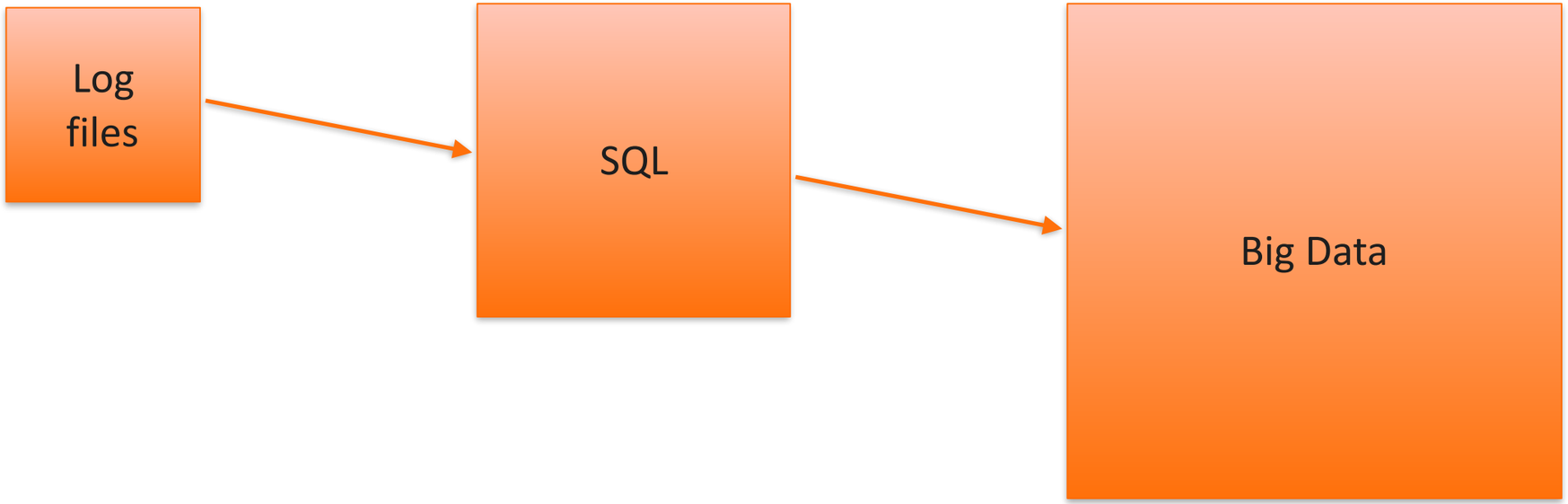
- Standards
- Formats
- “Exactly Once” Delivery
- Protocols
- Veracity of Information
- Validity of Information
- Ensuring Security
- Overcoming Security
- Compliance
- Schemas
- Consumers Change
- Credential Management
- “*That* [person|team|group]”
- Network*
- “Exactly Once” Delivery




HORTONWORKS

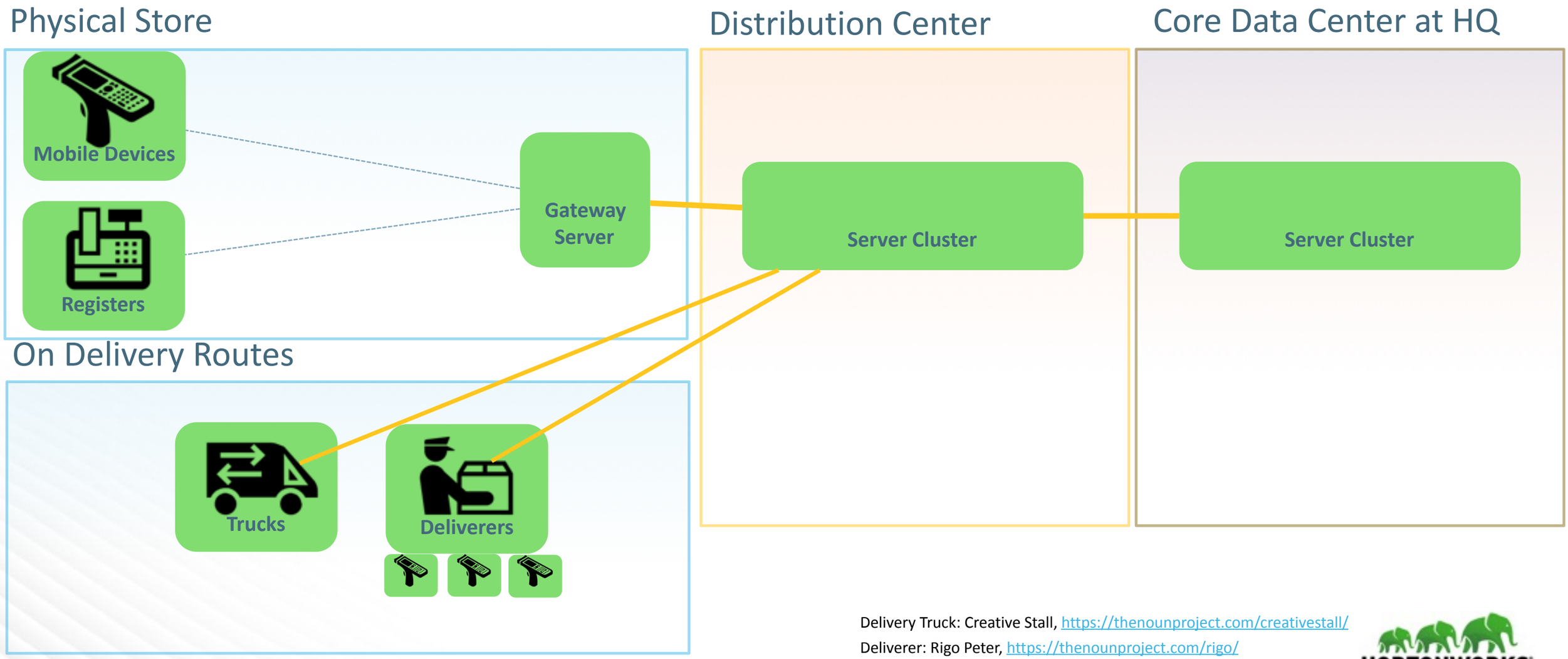
# Connecting A to B to C

## 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

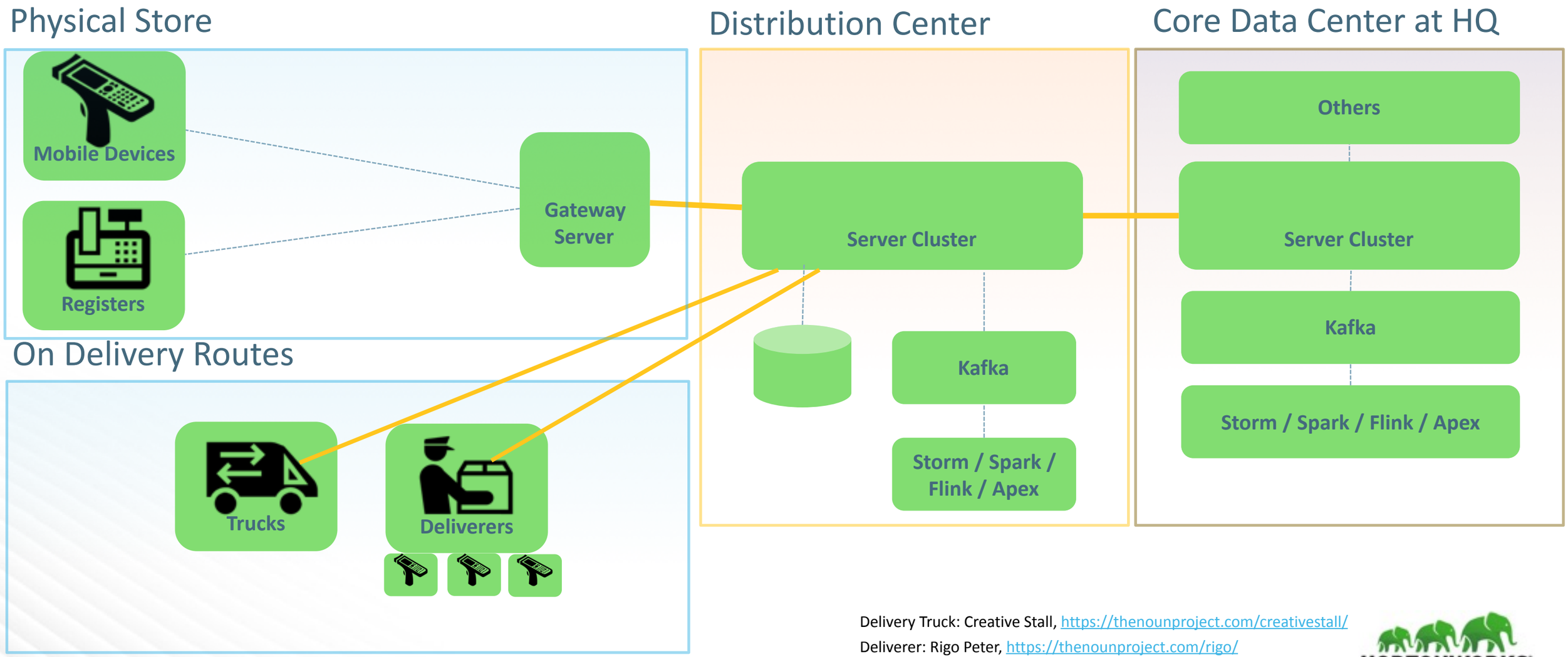
## Let’s consider the needs of a courier service




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!

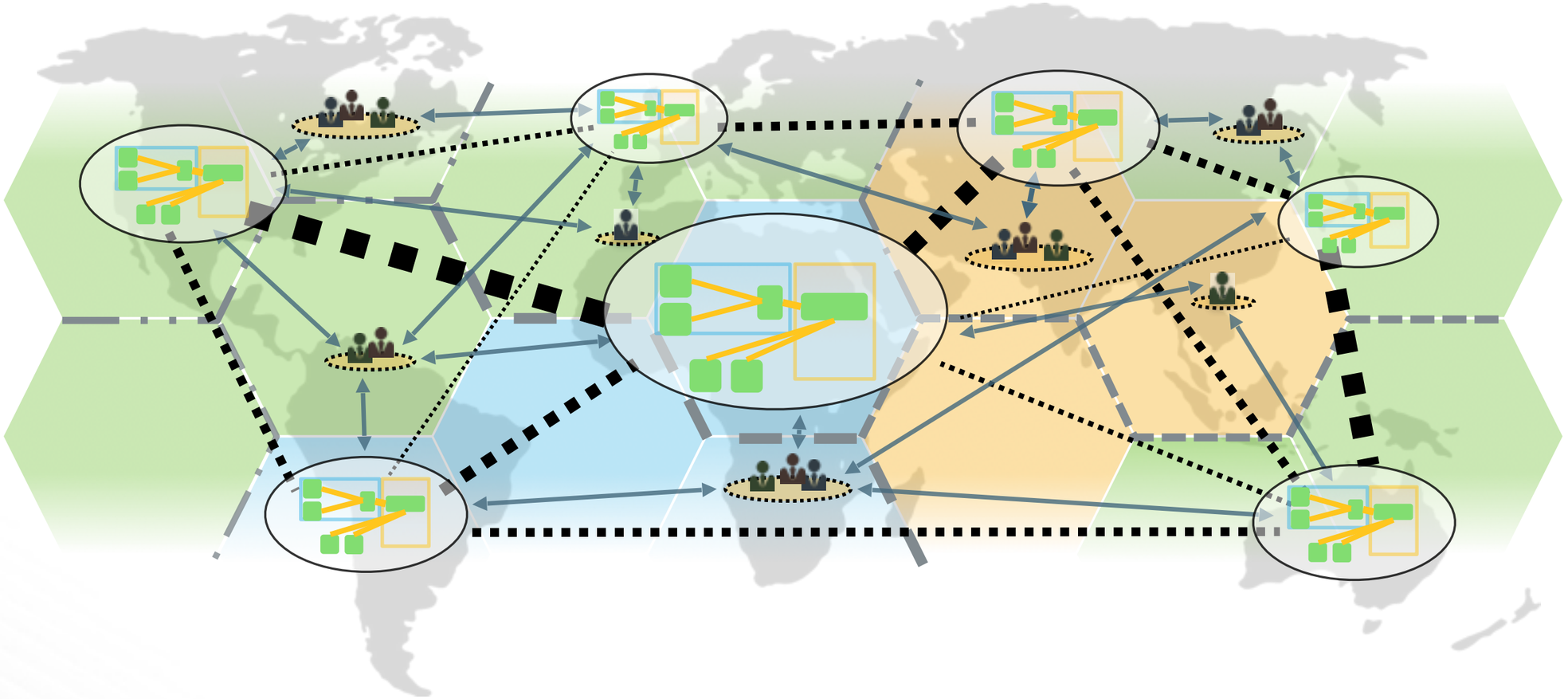
## Finding our needles in the haystack




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/

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

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






HORTONWORKS

# Agenda

*What is dataflow and what are the challenges?*

*Apache NiFi*

IoT Challenges

Apache MiNiFi

Exploration

Community




HORTONWORKS

# NiFi is based on Flow Based Programming (FBP)

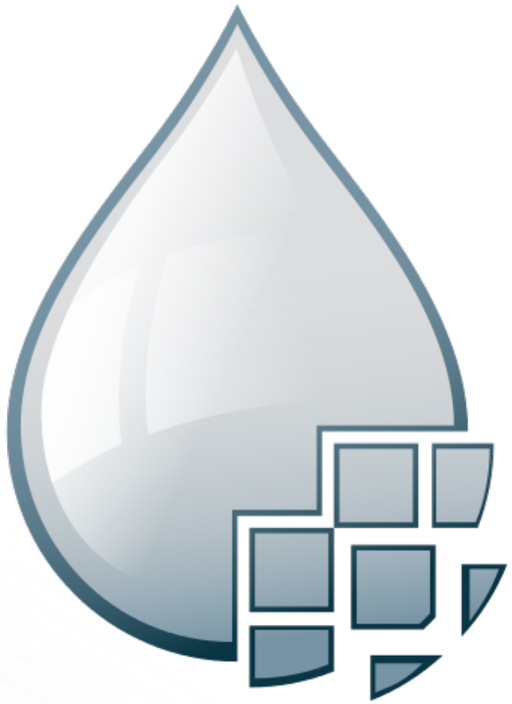
| FBP Term | NiFi Term | Description |
|---|---|---|
| Information Packet | FlowFile | Each object moving through the system. |
| Black Box | FlowFile Processor | Performs the work, doing some combination of data routing, transformation, or mediation between systems. |
| Bounded Buffer | Connection | 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

# Apache NiFi

## Key Features

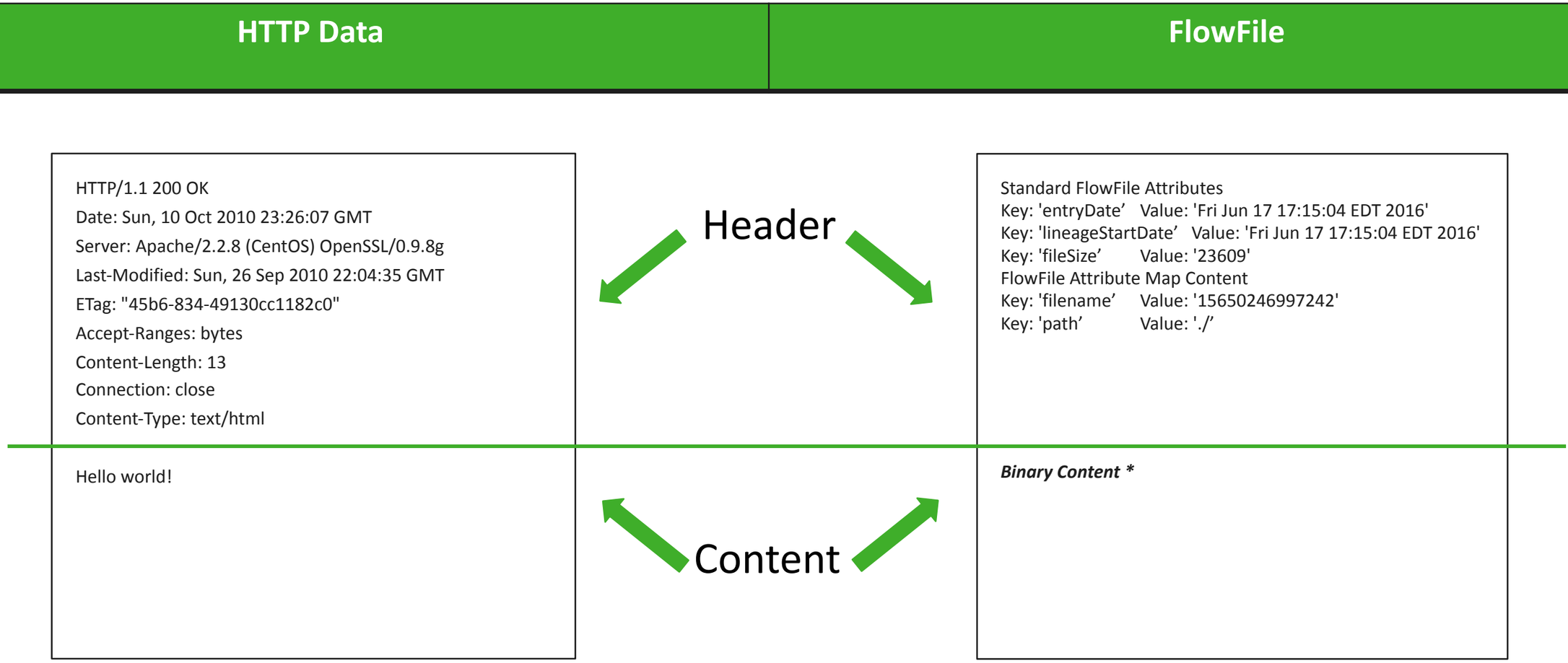


- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models
- Recovery/recording a rolling log of fine-grained history
- Visual command and control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



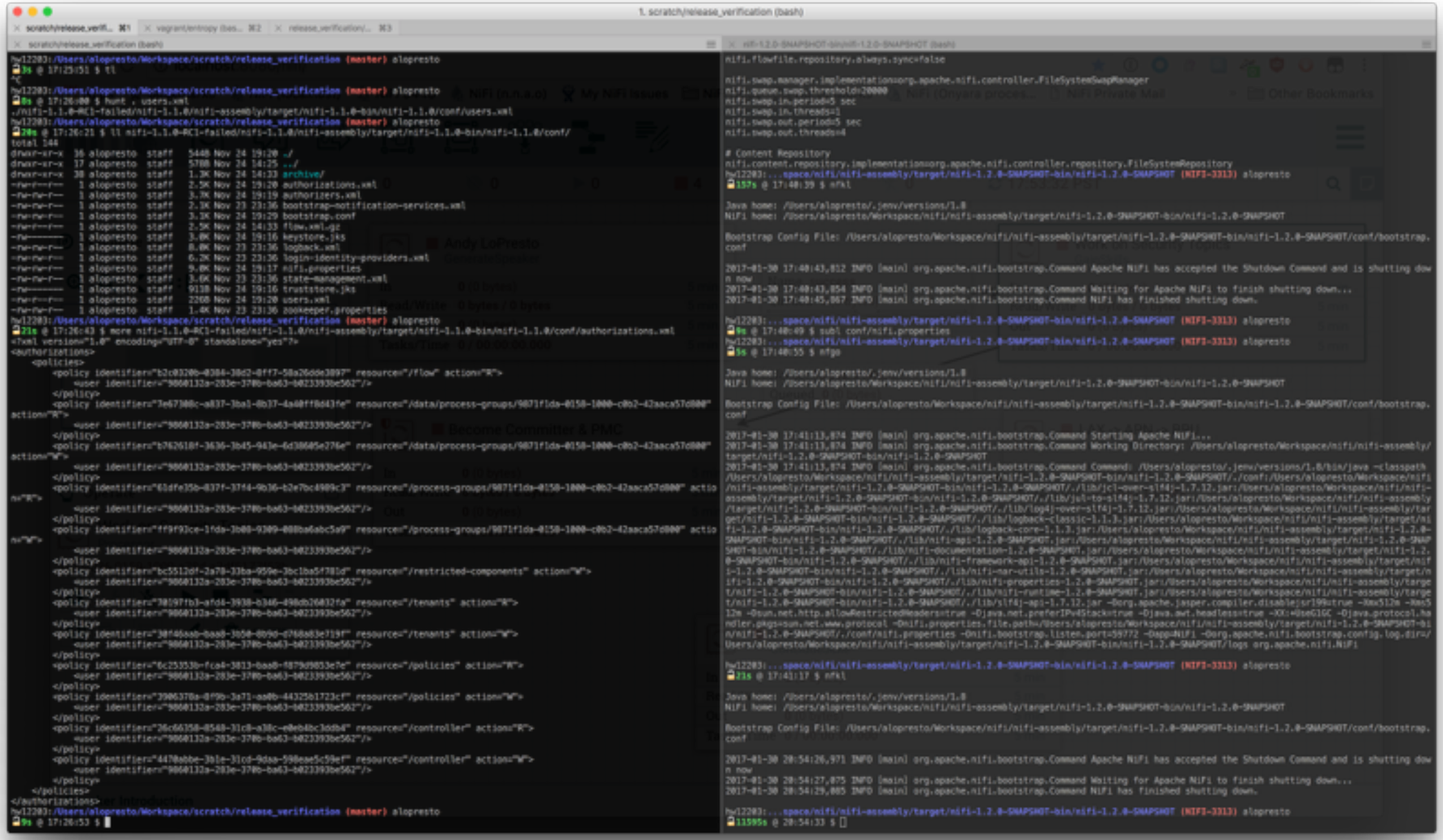

HORTONWORKS

# FlowFiles are like HTTP data



# User Interface

## Less of this… … more of this

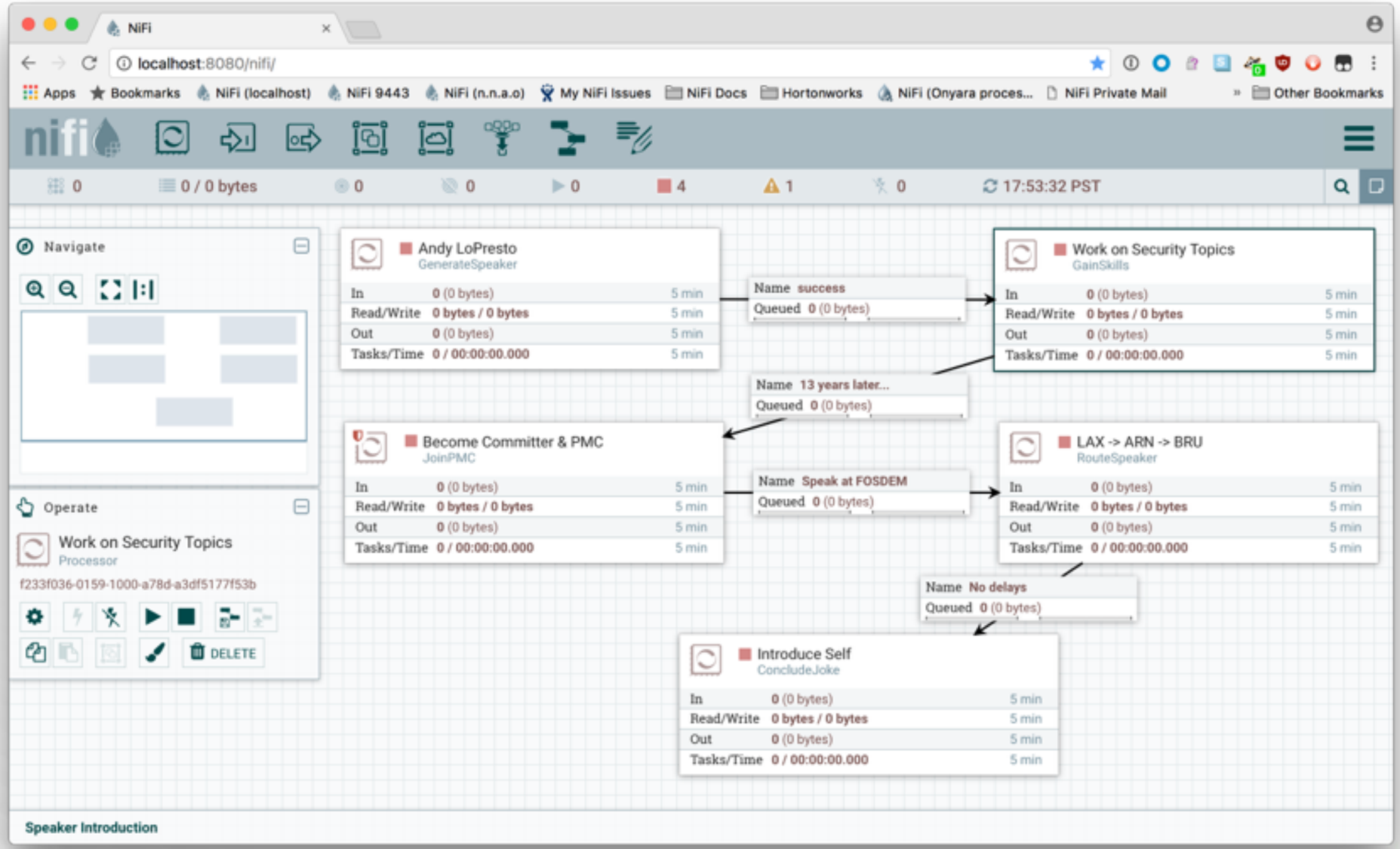






HORTONWORKS

# User Interface







HORTONWORKS

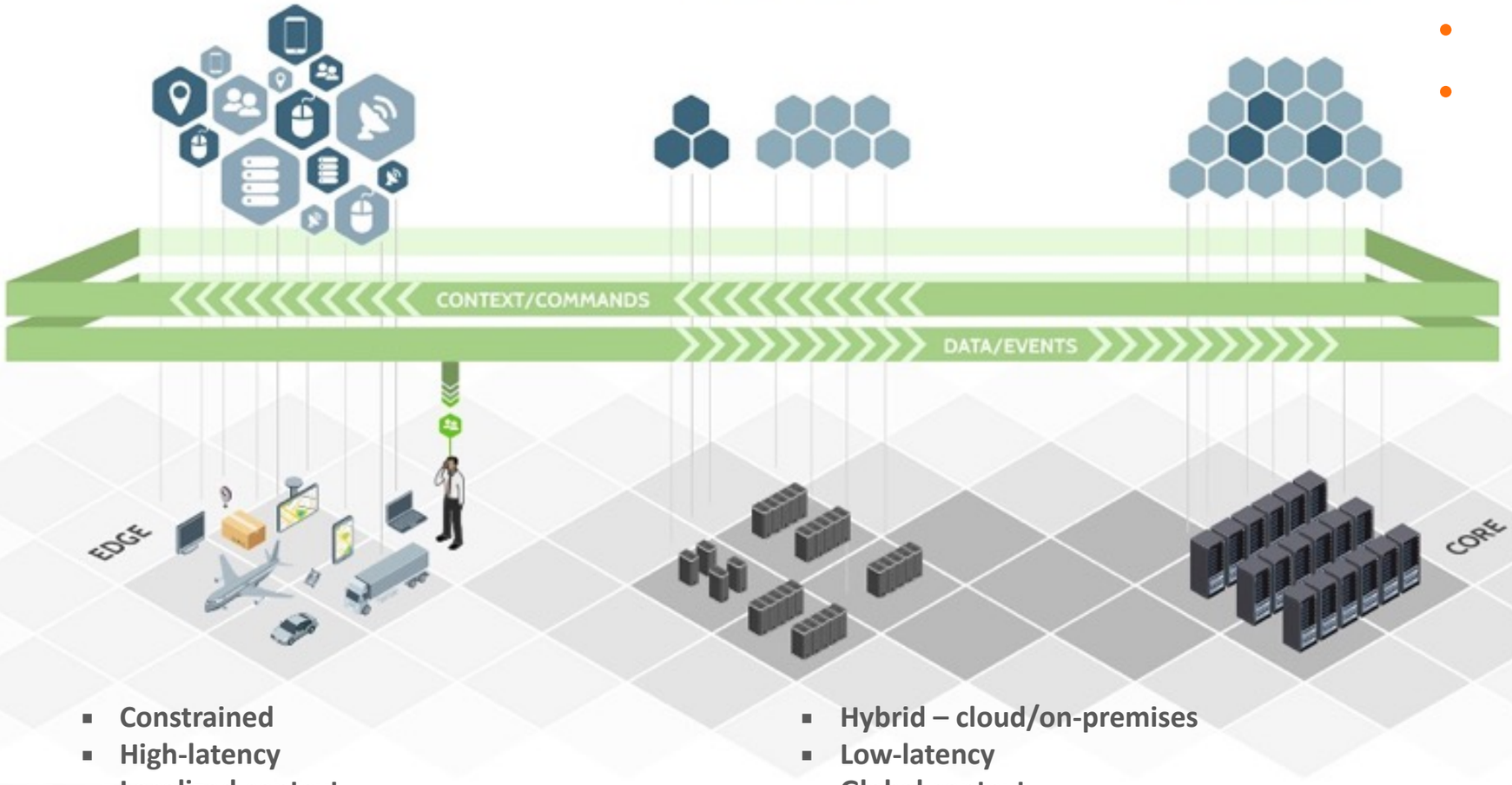
# Data Provenance

**Origin – attribution**
**Replay – recovery**

**Evolution of topologies**
**Long retention**

## Types of Lineage

- Event
- Configuration


- Constrained
- High-latency
- Localized context

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




HORTONWORKS

# Deeper Ecosystem Integration: 220+ 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

# Edge Challenges

- Limited computing capability
- Limited power/network
- Restricted software library/platform availability
- No UI
- Physically inaccessible
- Not frequently updated
- Competing standards/protocols
- Scalability
- Privacy & Security



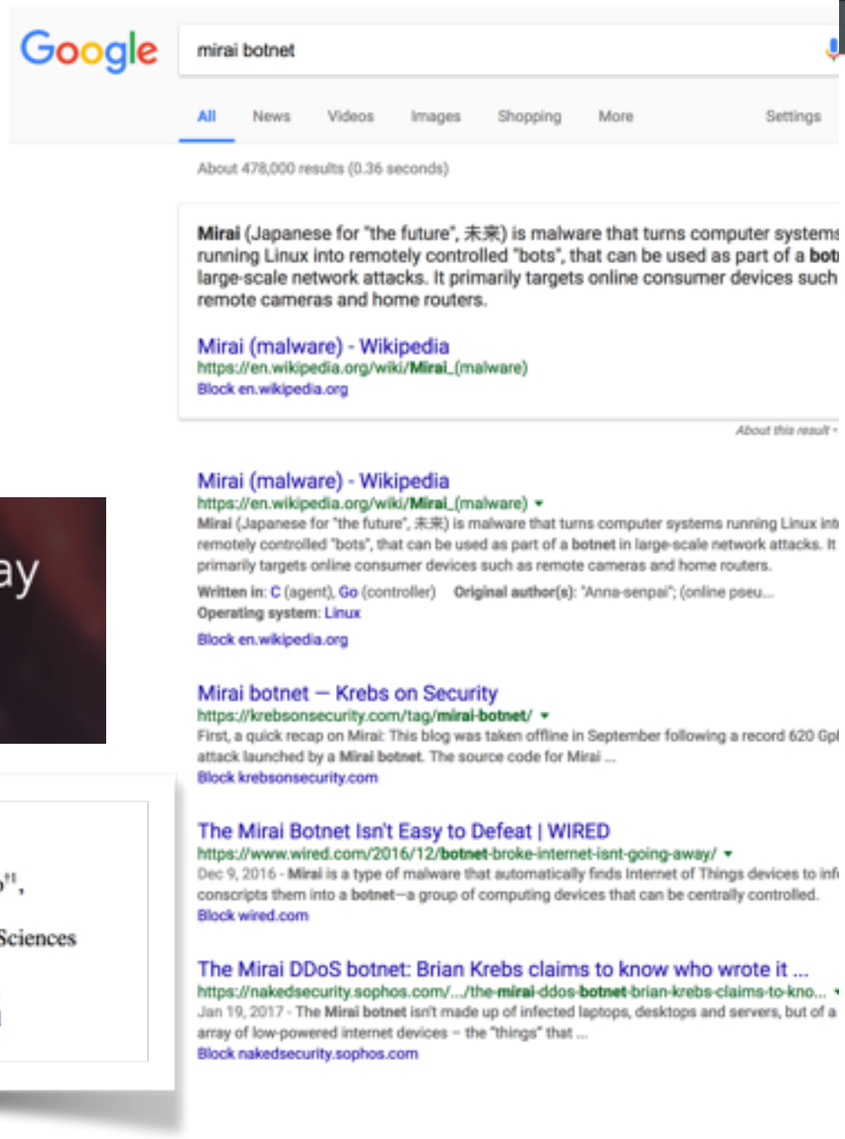

HORTONWORKS

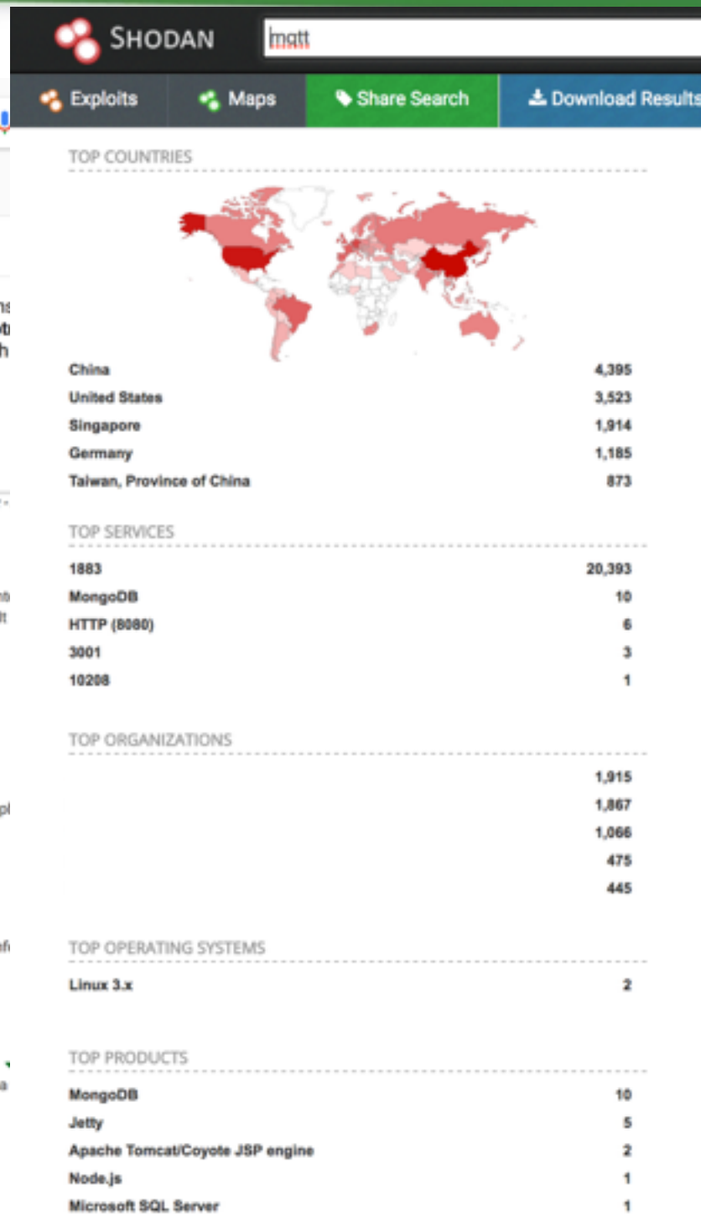
# Recent Examples

- When the Mirai attack has its own Wikipedia page, that's not good


**IoTPOT: Analysing the Rise of IoT Compromises**

Yin Minn Pa Pa[†1], Shogo Suzuki[†1], Katsunari Yoshioka[†1], Tsutomu Matsumoto[†1],
Takahiro Kasama[†2], Christian Rossow[†3]
[†1]Graduate School of Environment and Information Sciences/Institute of Advanced Sciences
[†1] Yokohama National University, Japan
[†2]National Institute of Information and Communications Technology, Japan
[†3]Institute of Advanced Sciences, Yokohama National University, Japan and
[†3]Cluster of Excellence, MMCI, Saarland University, Germany







HORTONWORKS

# NiFi Solves Everything*

- Runs on JVM
- Provides UI for flow design & monitoring
- Security built-in
  - TLS, authn/authz, encrypted data
- Handles practically any format/protocol

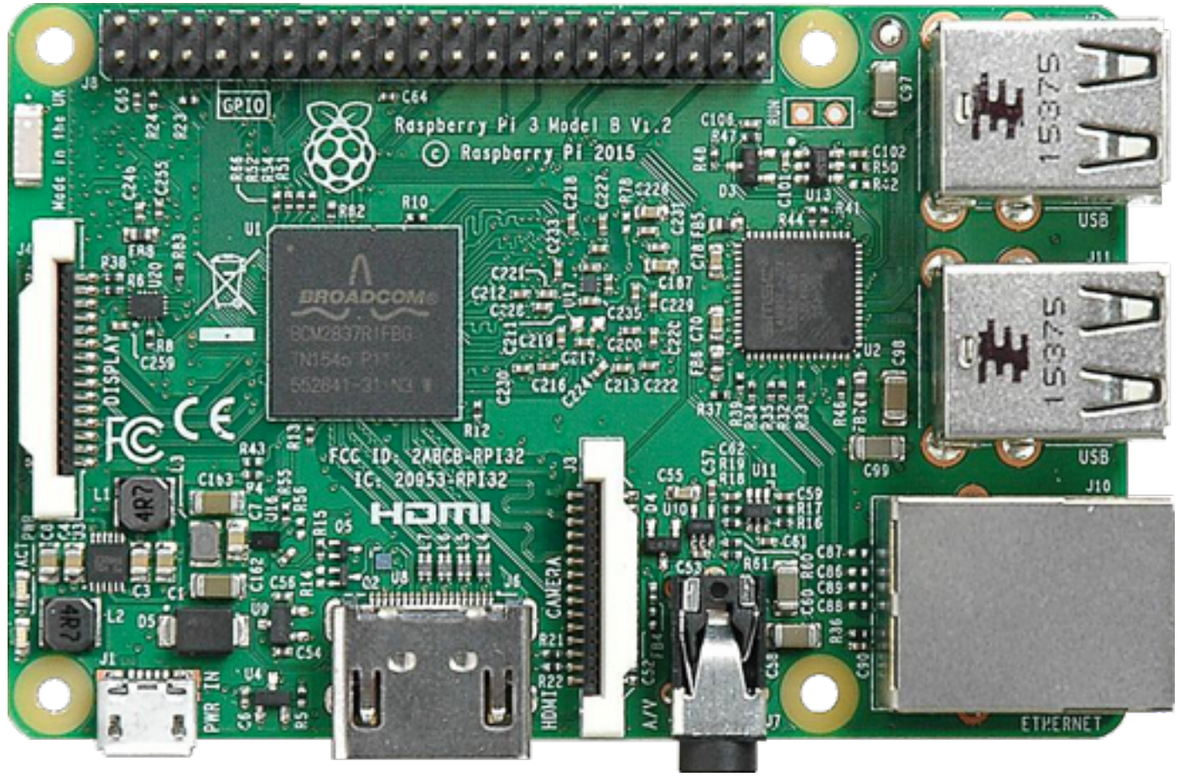



HORTONWORKS

# NiFi for IoT

- NiFi supports AMQP, MQTT, UDP, TCP, HTTP(S), CEF, JMS, (S)FTP, *AWSIoT*
- With a little pruning, NiFi can run on a Raspberry Pi

```
.
├── bootstrap
├── jcl-over-slf4j-1.7.12.jar
├── jul-to-slf4j-1.7.12.jar
├── log4j-over-slf4j-1.7.12.jar
├── logback-classic-1.1.3.jar
├── logback-core-1.1.3.jar
├── nifi-api-0.6.1.jar
├── nifi-documentation-0.6.1.jar
├── nifi-framework-nar-0.6.1.nar
├── nifi-html-nar-0.6.1.nar
├── nifi-http-context-map-nar-0.6.1.nar
├── nifi-jetty-bundle-0.6.1.nar
├── nifi-kerberos-iaa-providers-nar-0.6.1.nar
├── nifi-ldap-iaa-providers-nar-0.6.1.nar
├── nifi-nar-utils-0.6.1.jar
├── nifi-properties-0.6.1.jar
├── nifi-provenance-repository-nar-0.6.1.nar
├── nifi-runtime-0.6.1.jar
├── nifi-scripting-nar-0.6.1.nar
├── nifi-ssl-context-service-nar-0.6.1.nar
├── nifi-standard-nar-0.6.1.nar
├── nifi-standard-services-api-nar-0.6.1.nar
├── nifi-update-attribute-nar-0.6.1.nar
└── slf4j-api-1.7.12.jar
```



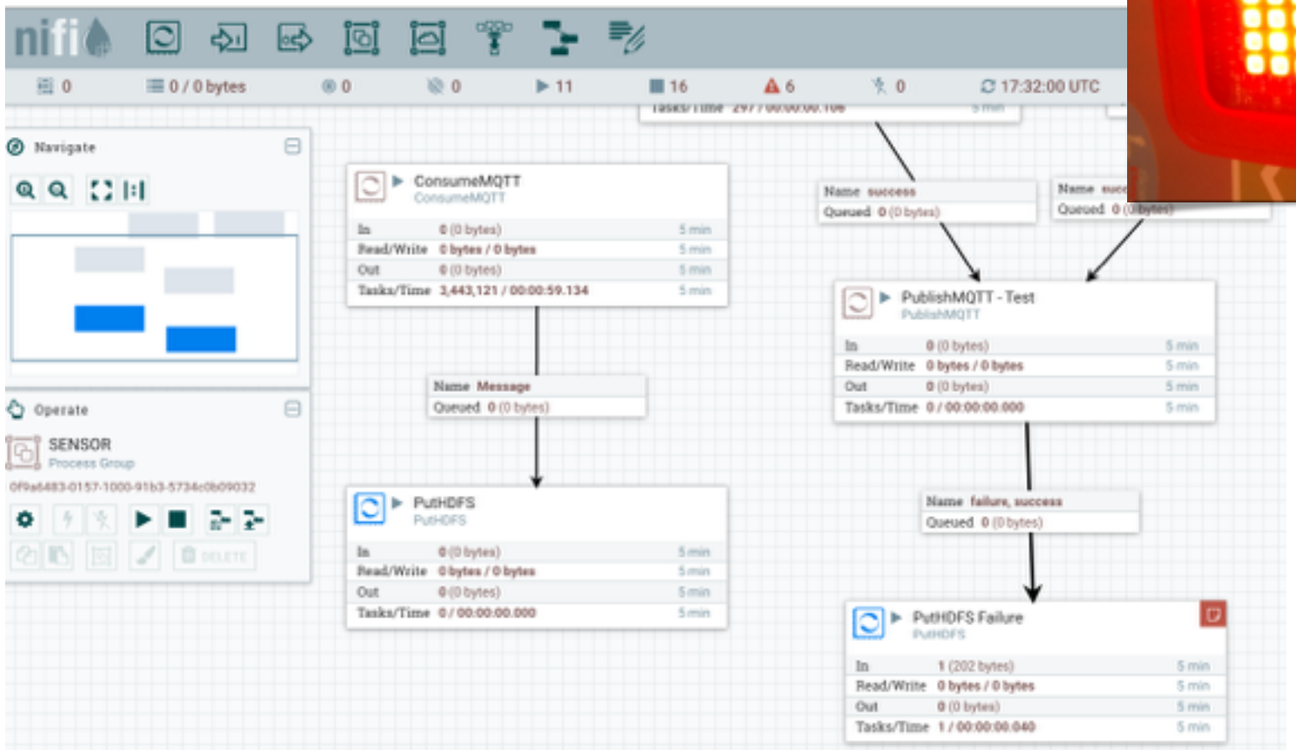




HORTONWORKS

# Example — Sensor Readings via RP3B

- Tim Spann
- Sense Hat sensor attachment
  - Temp, humidity, pressure
  - 8x8 LED display
- Python Flask server reading sensor and pushing to MQTT
- NiFi consuming MQTT



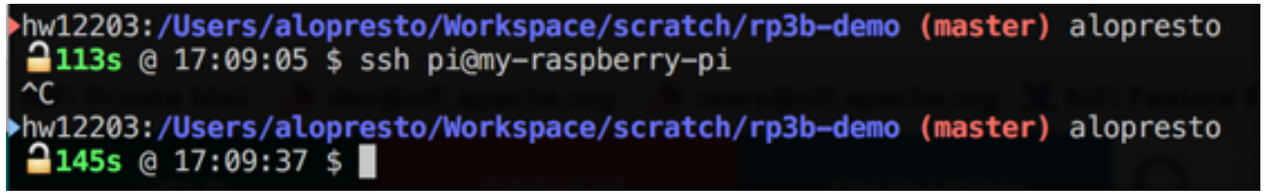

https://community.hortonworks.com/articles/55839/reading-sensor-data-from-remote-sensors-on-raspber.html




HORTONWORKS

# So Why Do We Need A Different Solution?

- NiFi is designed to “own the box”
- NiFi 0.7.x started up in about 10-15 minutes on RP3 (593 MB)
- NiFi 1.x started up in about 30 minutes on RP3 (760 MB)
  - 33 new processors
  - Rewrite for multi tenant authorization
  - Complete UI overhaul







HORTONWORKS

# Apache NiFi Subproject: MiNiFi

- 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

# Why build MiNiFi?

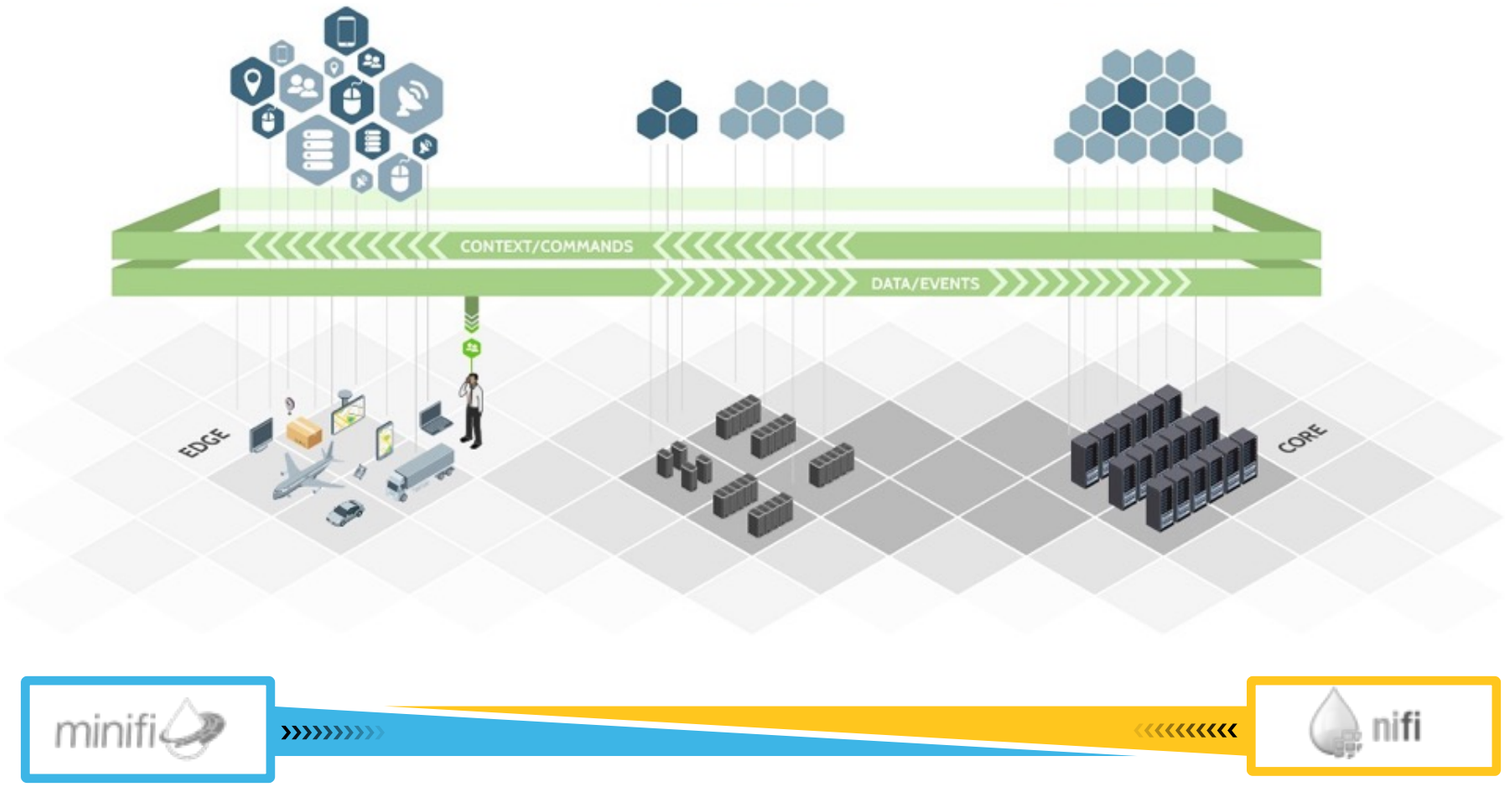
- NiFi is big
  - 1.3.0 release is 933 MB compressed
  - Can be modified to run in restricted environments, but requires manual surgery
  - Provides UI, provenance query, etc.
  - Runs on dedicated machines/clusters — “owns the box”
- MiNiFi lives at the edge
  - No UI
  - 0.1.0 Java binary is 45 MB, C++ binary is 746 KB
  - “Good guest”




HORTONWORKS

# How Does MiNiFi Interact With NiFi?

- NiFi
  - Design flows
  - Aggregate data from many sources
  - Perform routing/analysis/ SEP
- MiNiFi
  - Receive flows
  - Collect data
  - Send for processing



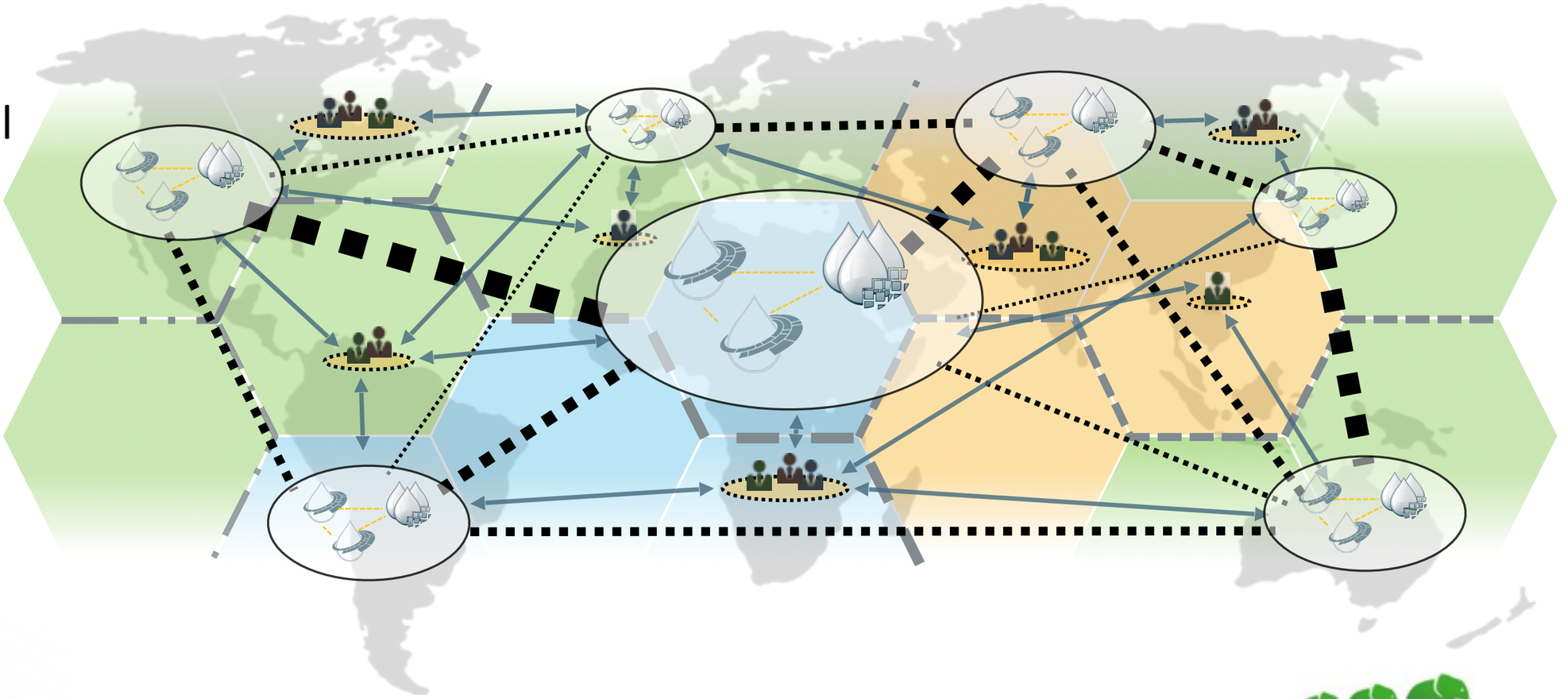




HORTONWORKS

# Let’s Add Dimensionality

- We’ve been imagining EDGE to CORE as a bi-directional linear system
- Let’s expand that to the real world

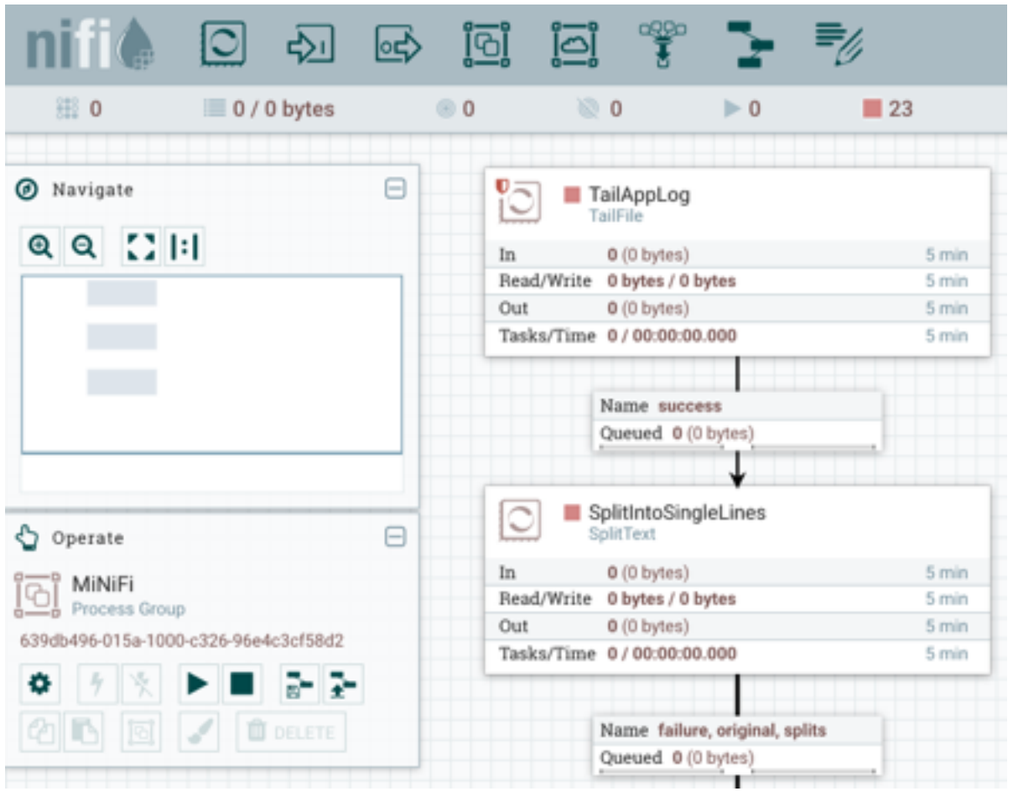


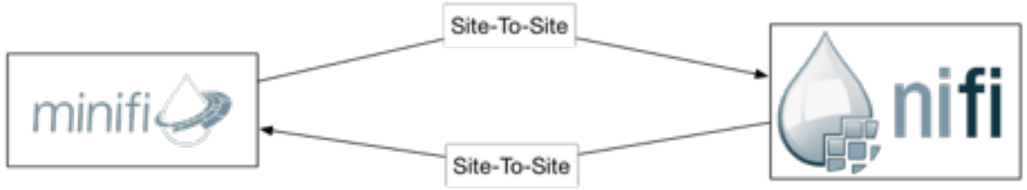



HORTONWORKS

# Flavors of MiNiFi

- MiNiFi Java (v0.2.0)
  - Modified version of NiFi
    - No UI
    - YAML configuration
  - Reduced processor count
    - 110 by default, more available with additional NARs
- MiNiFi C++ (v0.2.0)
  - Written from scratch
  - 10 processors by default
  - Bi-directional site-to-site & provenance data



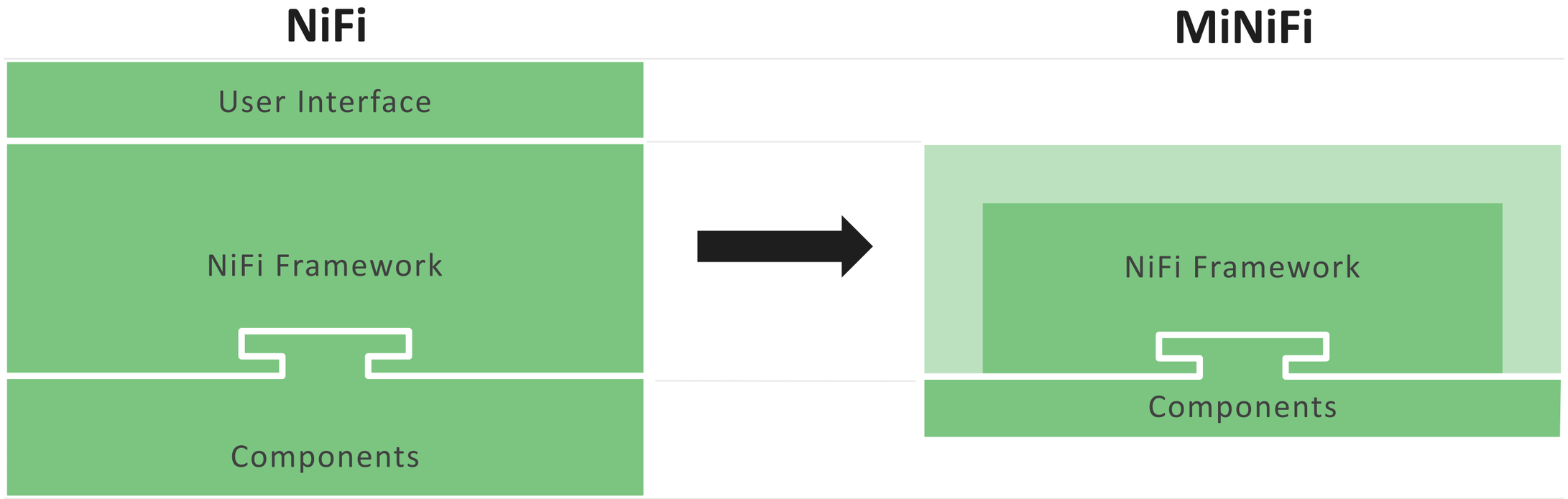

```
Security Properties:
    keystore: /tmp/ssl/localhost-ks.jks
    keystore type: JKS
    keystore password: localtest
    key password: localtest
    truststore: /tmp/ssl/localhost-ts.jks
    truststore type: JKS
    truststore password: localtest
    ssl protocol: TLS
    Sensitive Props:
        key:
        algorithm: PBEWITHMD5AND256BITAES-CBC-OPENSSL
        provider: BC

Processors:
    - name: TailAppLog
      class: org.apache.nifi.processors.standard.TailFile
      max concurrent tasks: 1
      scheduling strategy: TIMER_DRIVEN
      scheduling period: 10 sec
      penalization period: 30 sec
      yield period: 1 sec
      run duration nanos: 0
      auto-terminated relationships list:
      Properties:
          File to Tail: logs/minifi-app.log
          Rolling Filename Pattern: minifi-app*
          Initial Start Position: Beginning of File
    - name: SplitIntoSingleLines
      class: org.apache.nifi.processors.standard.SplitText
      max concurrent tasks: 1
      scheduling strategy: TIMER_DRIVEN
      scheduling period: 0 sec
      penalization period: 30 sec
      yield period: 1 sec
      run duration nanos: 0
      auto-terminated relationships list:
          - failure
          - original
      Properties:
          Line Split Count: 1
          Header Line Count: 0
          Remove Trailing Newlines: true
```




HORTONWORKS

# NiFi vs MiNiFi Java Processes

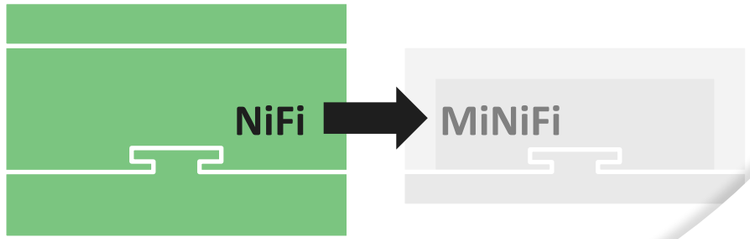


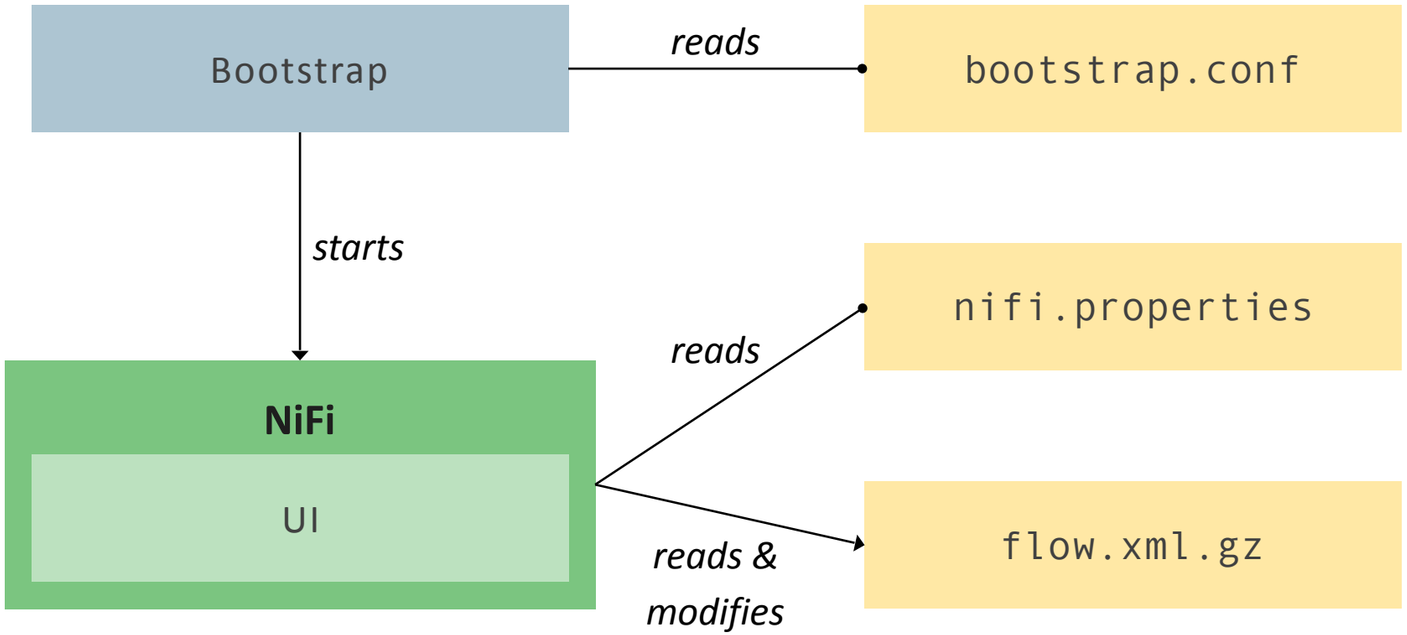




HORTONWORKS

# NiFi Java Processes



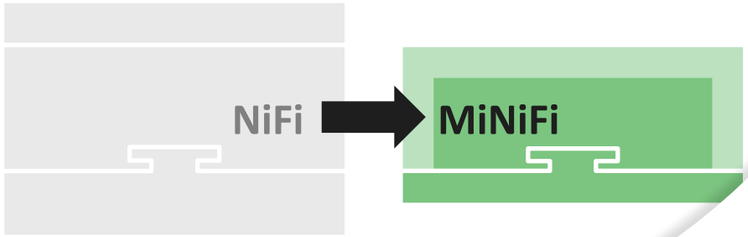



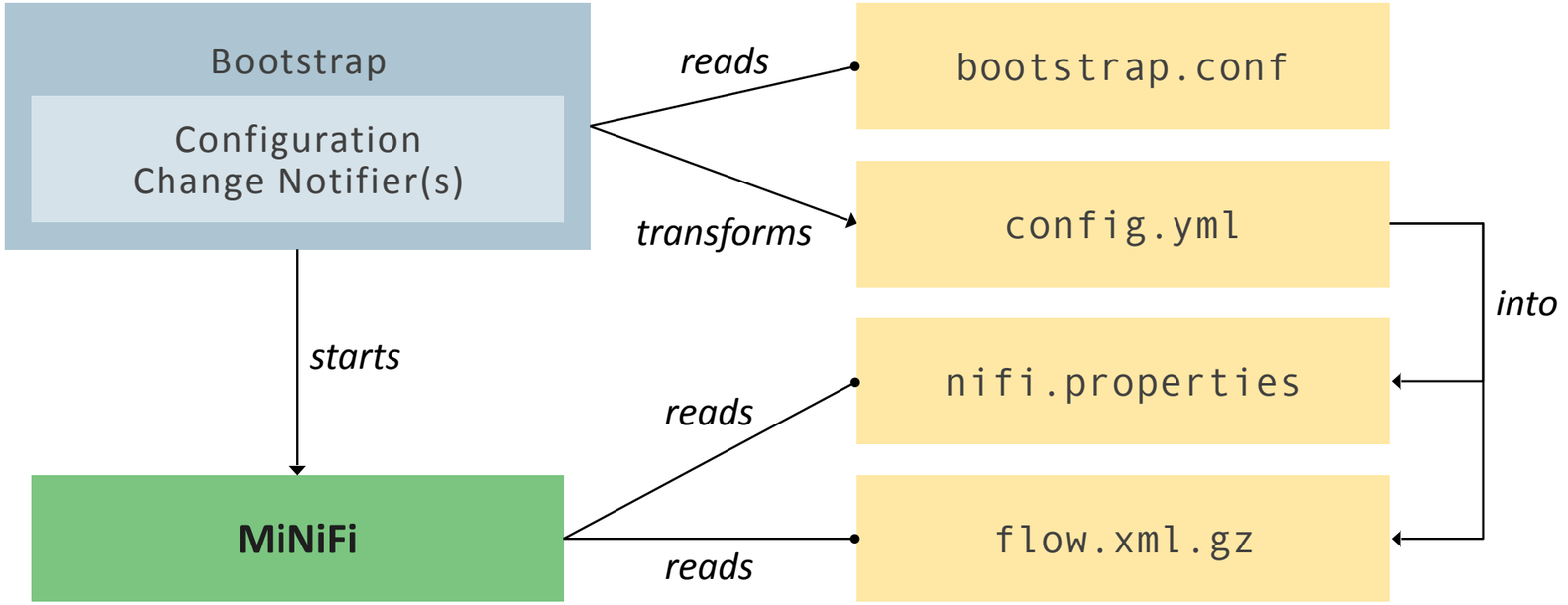




HORTONWORKS

# MiNiFi Java Processes

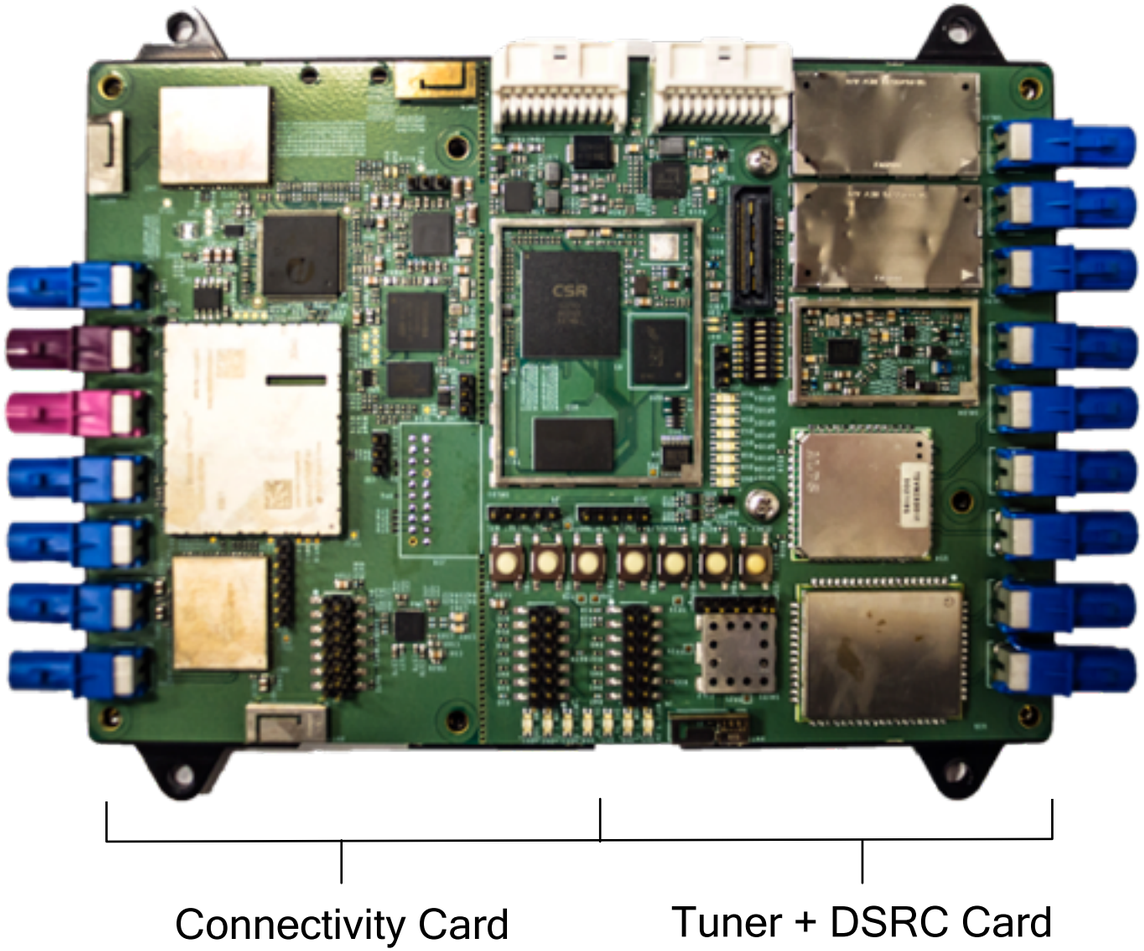


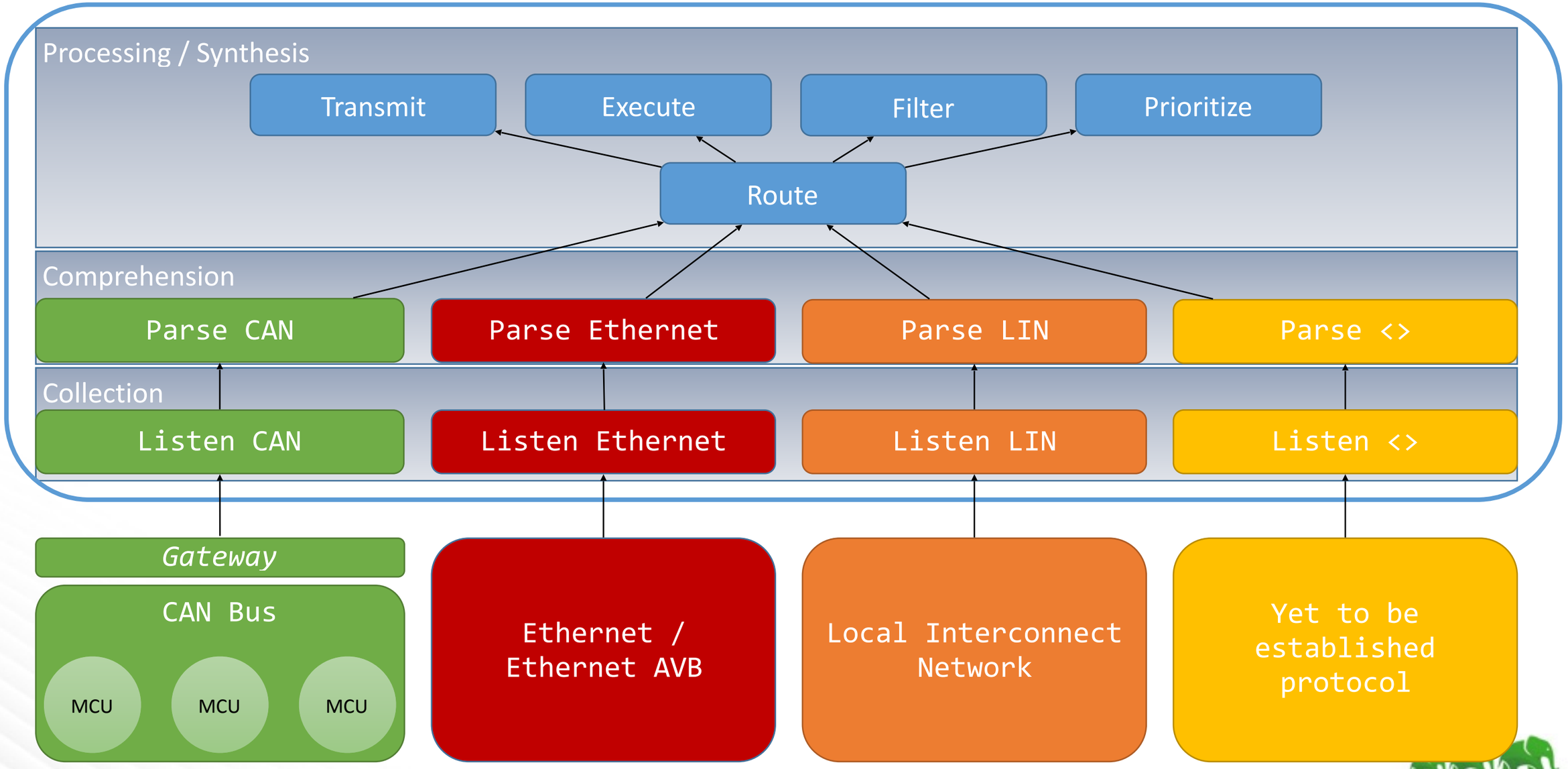
# What does MiNiFi provide?

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

## Connected Car Reference Platform Box



# MiNiFi on a Connected Car

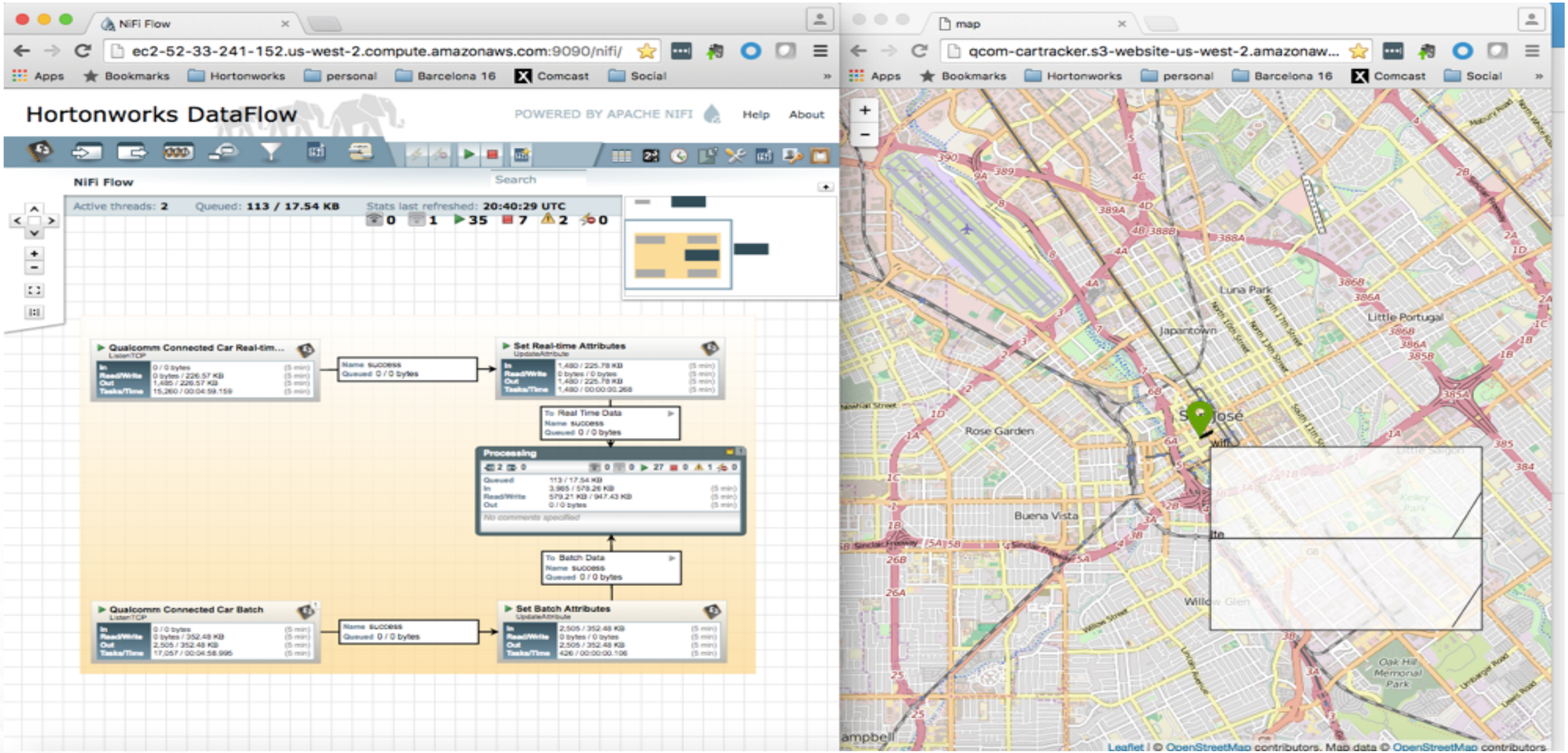






HORTONWORKS

# MiNiFi on a Connected Car



# MiNiFi Exfil

- Site-to-Site
  - NiFi protocol
  - Two implementations
    - Raw socket
    - HTTP(S) *(Java only)*
  - Secured with mutual authentication TLS
- HTTP(S), (S)FTP, JMS, Syslog, File, Email, Process *(Java only)*




HORTONWORKS

# Advanced Topics

- New features in Apache NiFi 1.2.0 & 1.3.0
- New features in Apache MiNiFi Java 0.2.0 & C++ 0.2.0
- New subproject Apache NiFi Registry




HORTONWORKS

# New in 1.2.0/1.3.0

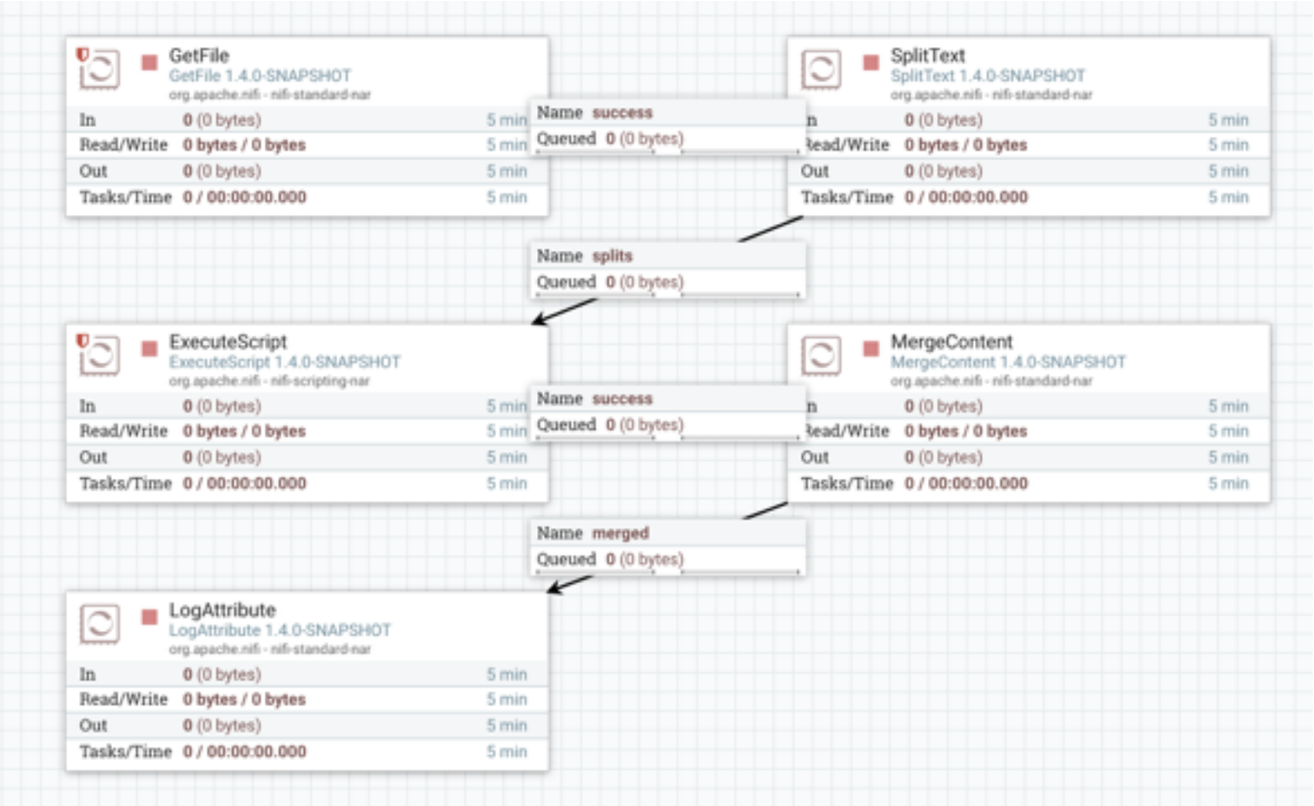
- Record Parsing
- Encrypted Provenance Repository




HORTONWORKS

# Record Parsing

- Previously, data had to be divided into individual flowfiles to perform work
- CSV output with 50k lines would need to be split, operated on, re-merged
- 1 + 50k + 50k + 1 flowfiles = 100k flowfiles



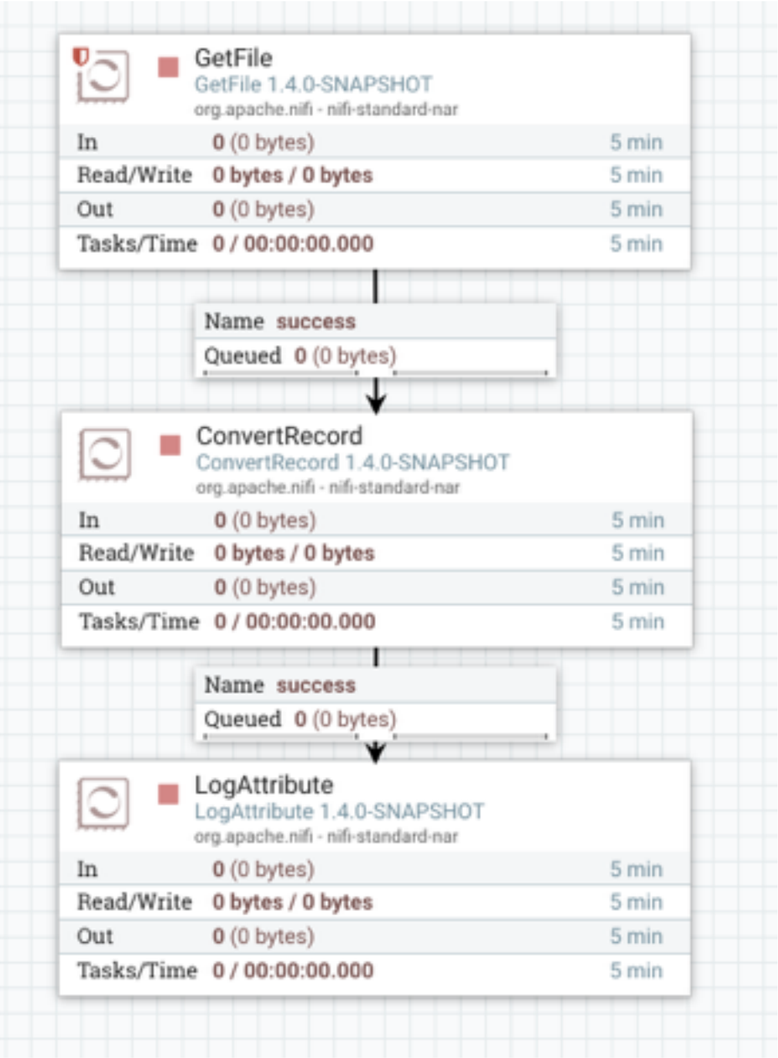




HORTONWORKS

# Record Parsing

- Now flowfile content can contain many “record” elements
- Read and write with **Reader* and **Writer* Controller Services
- Perform lookups, routing, conversion, SQL queries, validation, and more…
- 1 + 1 flowfiles = 2 flowfiles



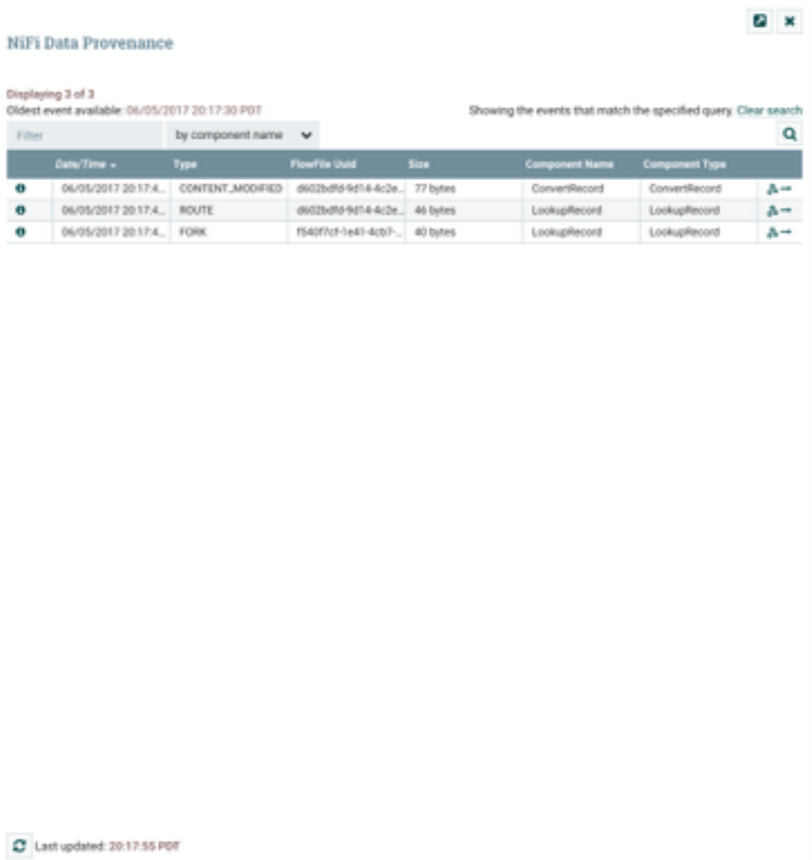




HORTONWORKS

# Encrypted Provenance Repository

- Every provenance event record is encrypted with AES G/CM before being persisted to disk
- Decrypted on deserialization for retrieval/query
- Random access via offset seek
- Handles key migration & rotation







HORTONWORKS

# MiNiFi Java 0.2.0

- Upgrading of core component dependencies to NiFi 1.2.0
- Initial command and control server capabilities
- Increased support for NiFi features in configuration YAML inclusive of:
- Support for HTTP Site to Site Proxy Properties
- Controller Services
- Binding site to site to a specific network interface




HORTONWORKS

# MiNiFi C++ 0.2.0

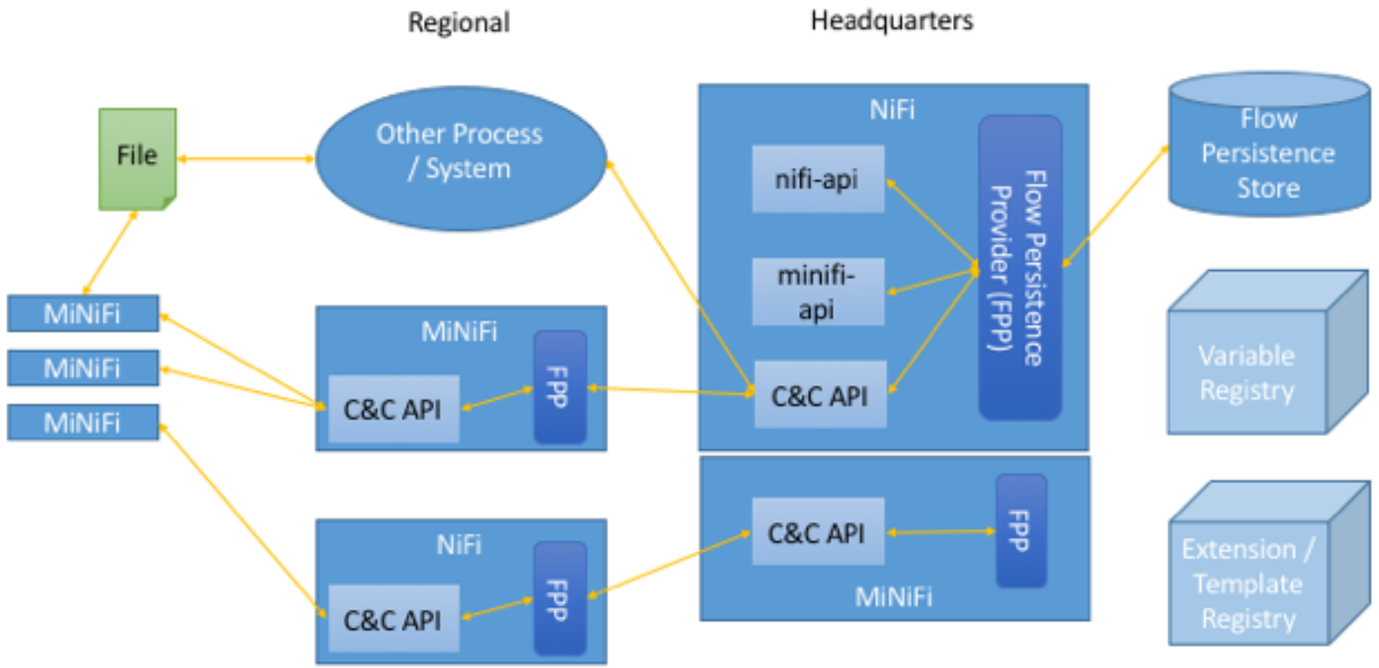
- Incorporation of Catch testing framework and Google linting for code quality and enhanced test coverage
- Providing support for reporting tasks and an initial implementation of Site to Site Provenance reporting
- New Processors inclusive of PutFile, LIstenHTTP




HORTONWORKS

# MiNiFi Feature Proposals

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







HORTONWORKS

# Apache NiFi Registry

- "…**complementary application** that provides a **central location** for storage and management of **shared resources** across **one or more instances** of NiFi and/or MiNiFi."





HORTONWORKS

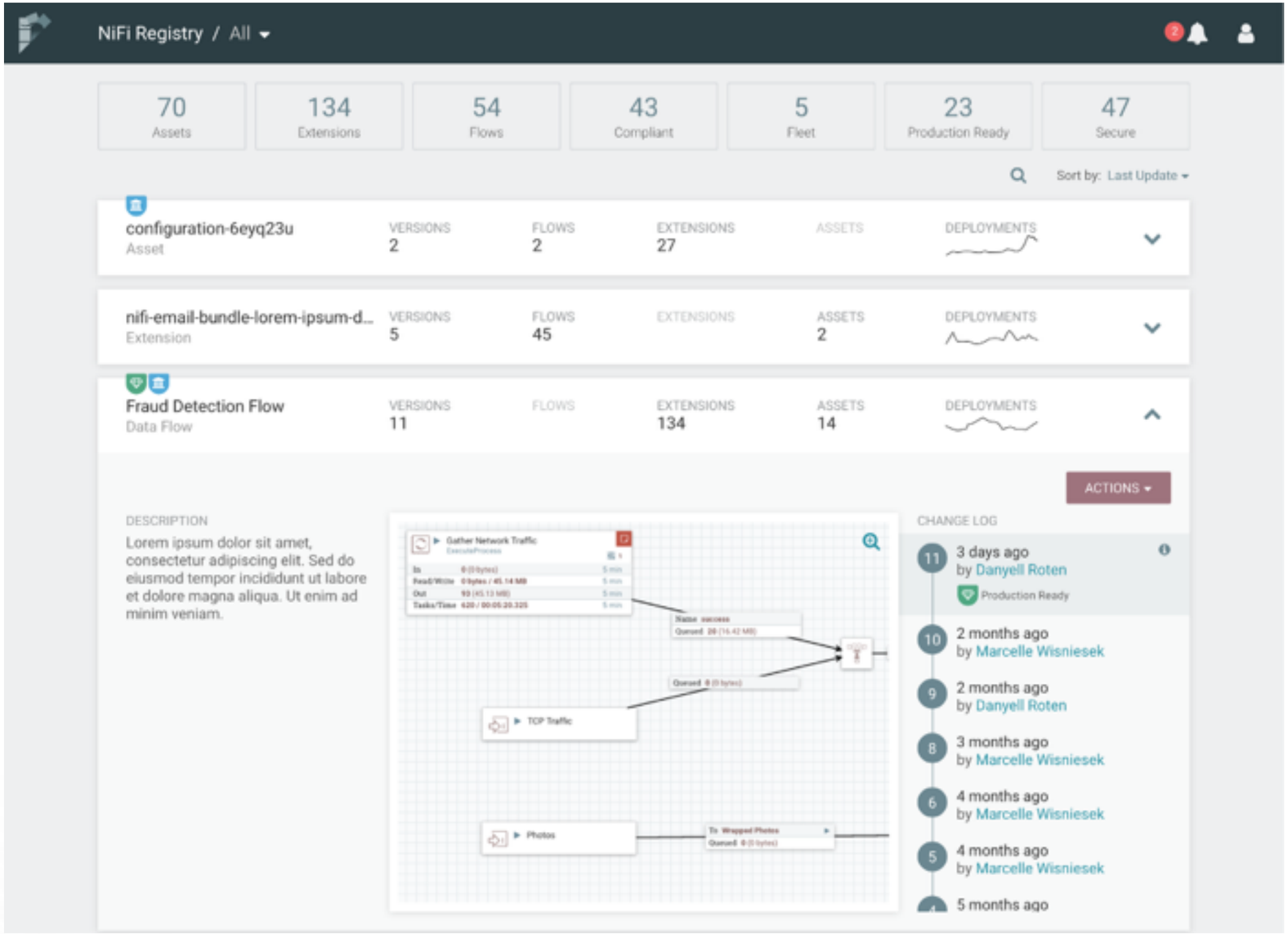
# Apache NiFi Registry - Flow Registry

- Flow registry stores & manages versioned flow definitions
- Integrated with NiFi to allow save/retrieve/ upgrade operations from canvas
- Admin of users, groups, and policies




HORTONWORKS

# Apache NiFi Registry - Flow Registry

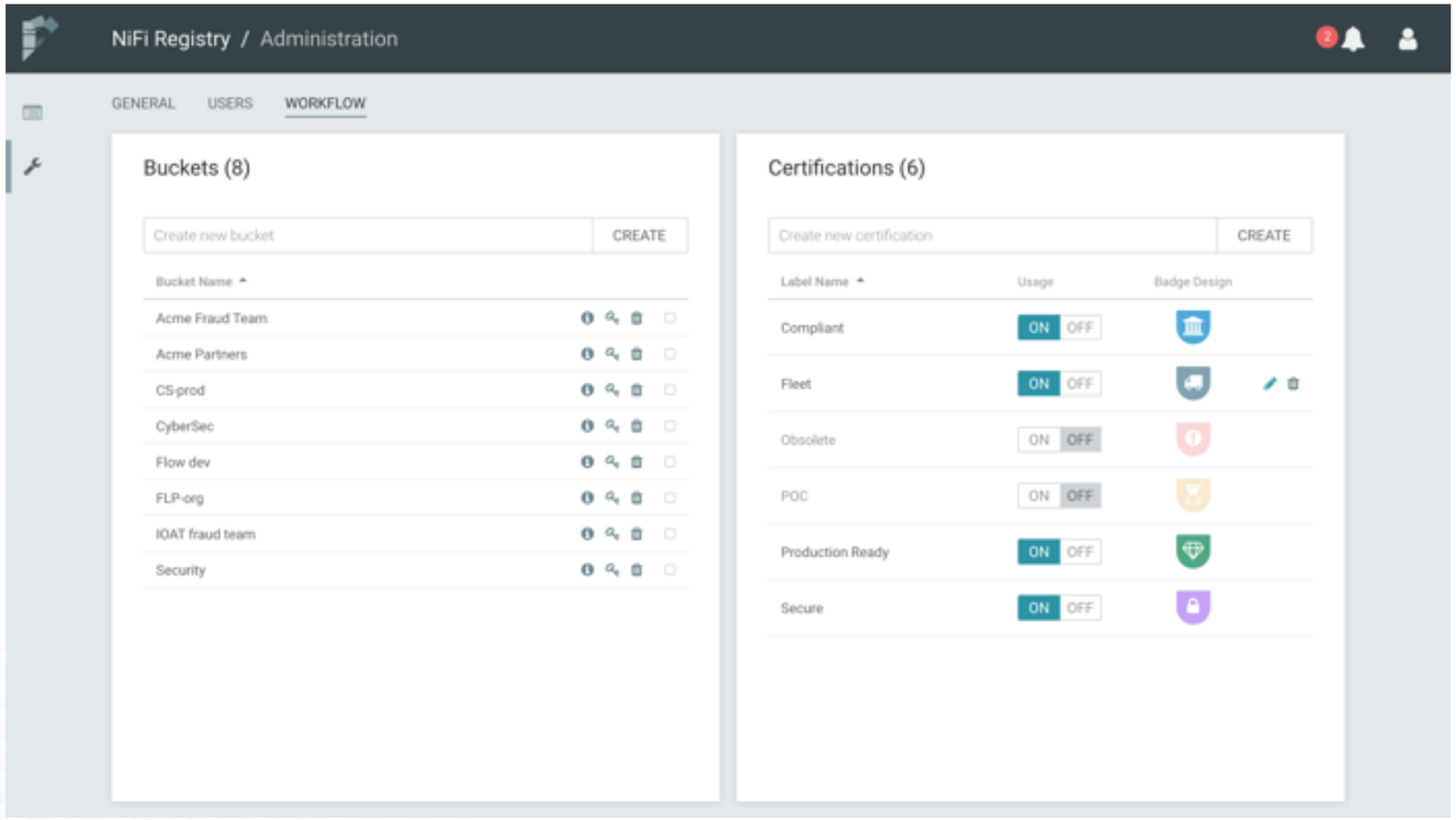


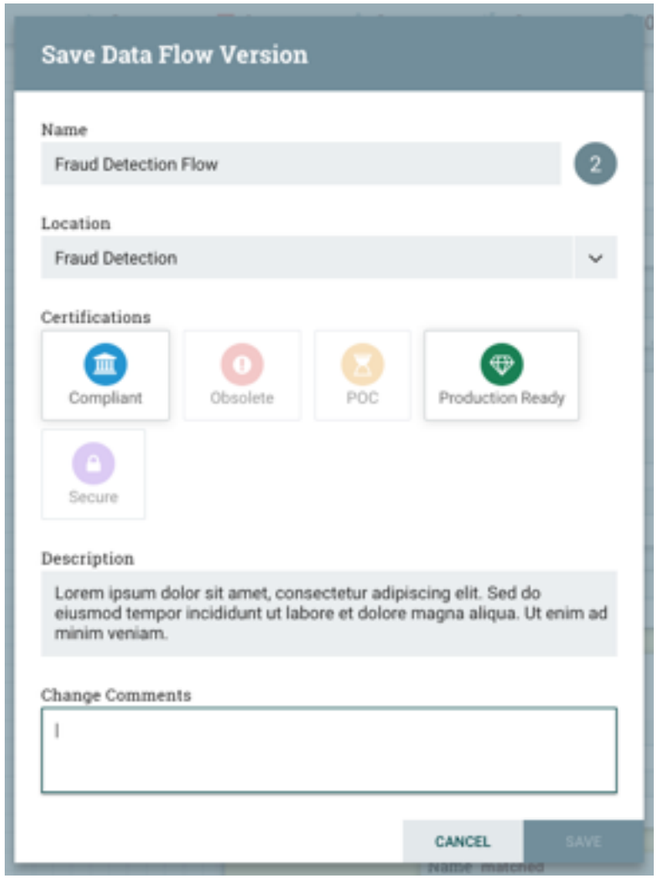




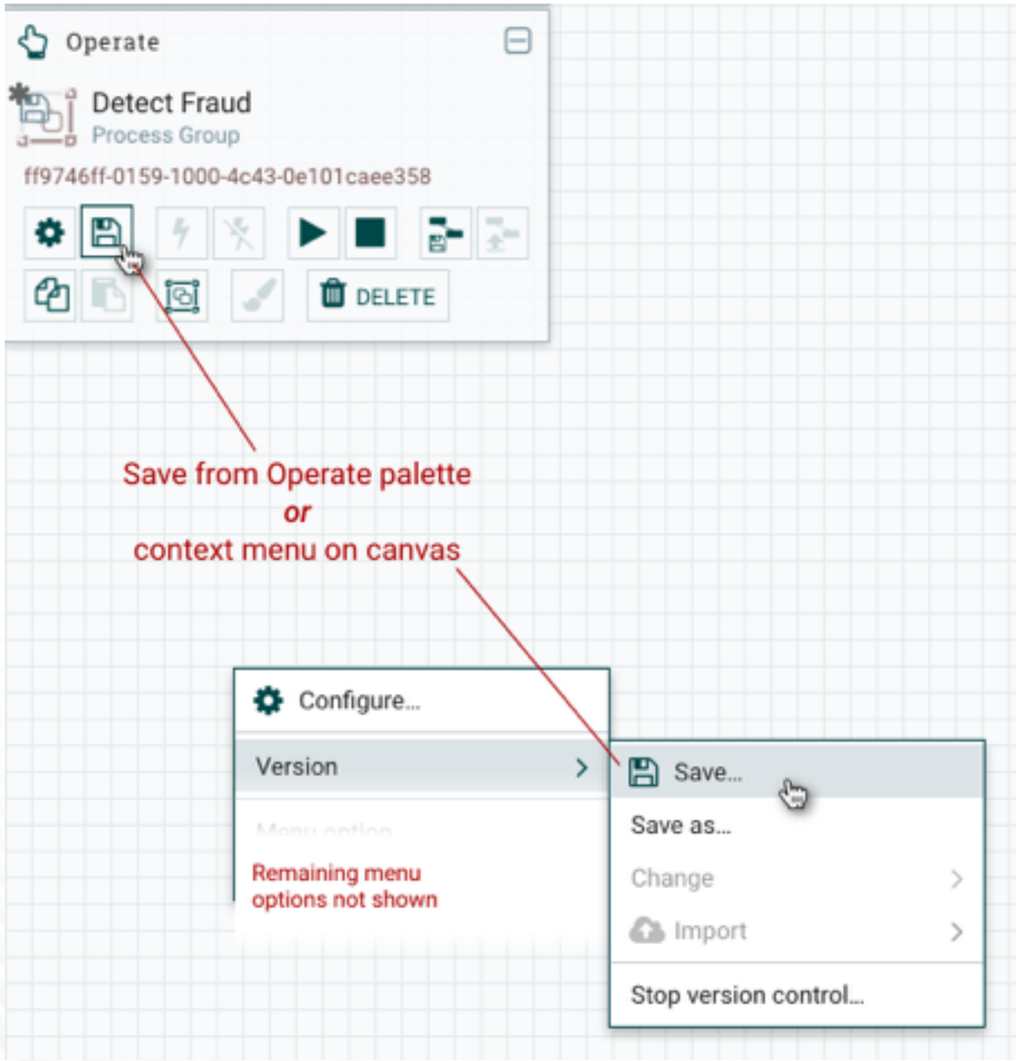
HORTONWORKS

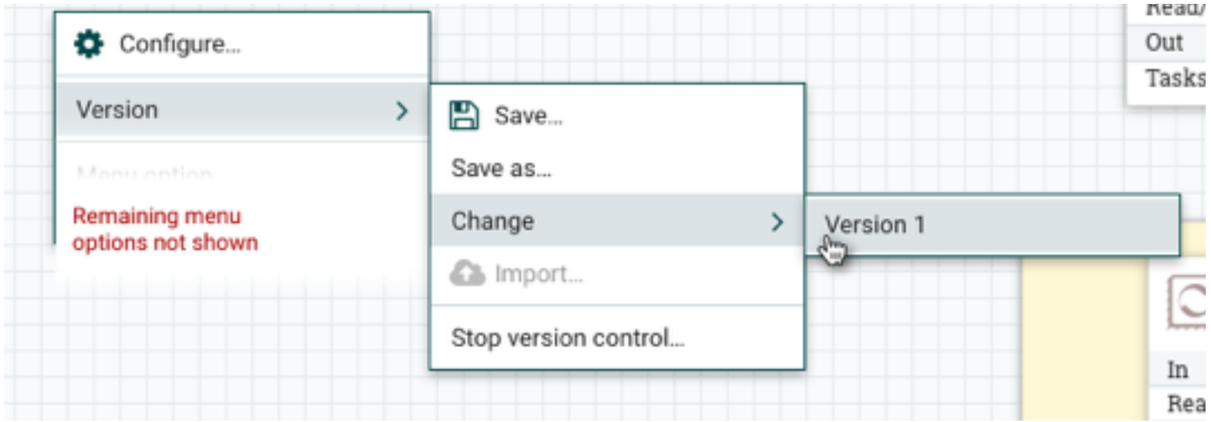
# Apache NiFi Registry - Flow Registry



# Apache NiFi Registry - Flow Registry







HORTONWORKS

# Why NiFi & MiNiFi?

- Moving data is multifaceted in its challenges and these are present in different contexts at varying scopes
  - 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

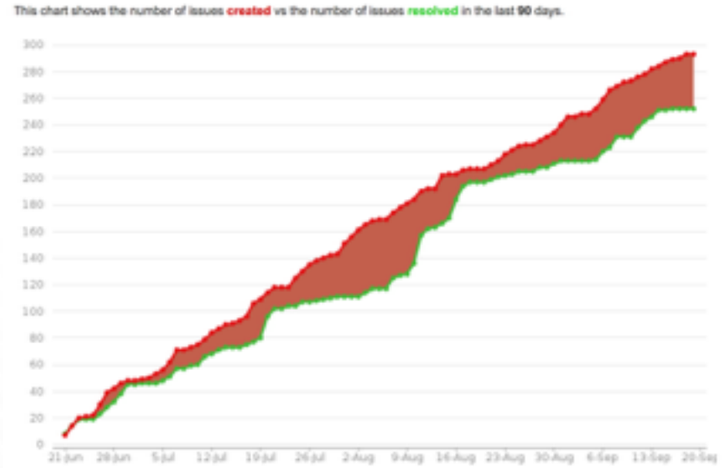
apache / nifi
mirrored from git://git.apache.org/nifi.git

Unwatch 133 | Unstar 656 | Fork 629

Code | Pull requests 104 | Projects 0 | Insights

Mirror of Apache NiFi

3,841 commits | 43 branches | 54 releases | 138 contributors | Apache-2.0




Status Summary

| Status | Issues | Percentage |
|---|---|---|
| Open | 1190 | 27% |
| In Progress | 31 | 1% |
| Reopened | 12 | |
| Resolved | 2982 | 68% |
| Closed | 72 | 2% |
| Patch Available | 77 | 2% |




HORTONWORKS

# Learn more and join us

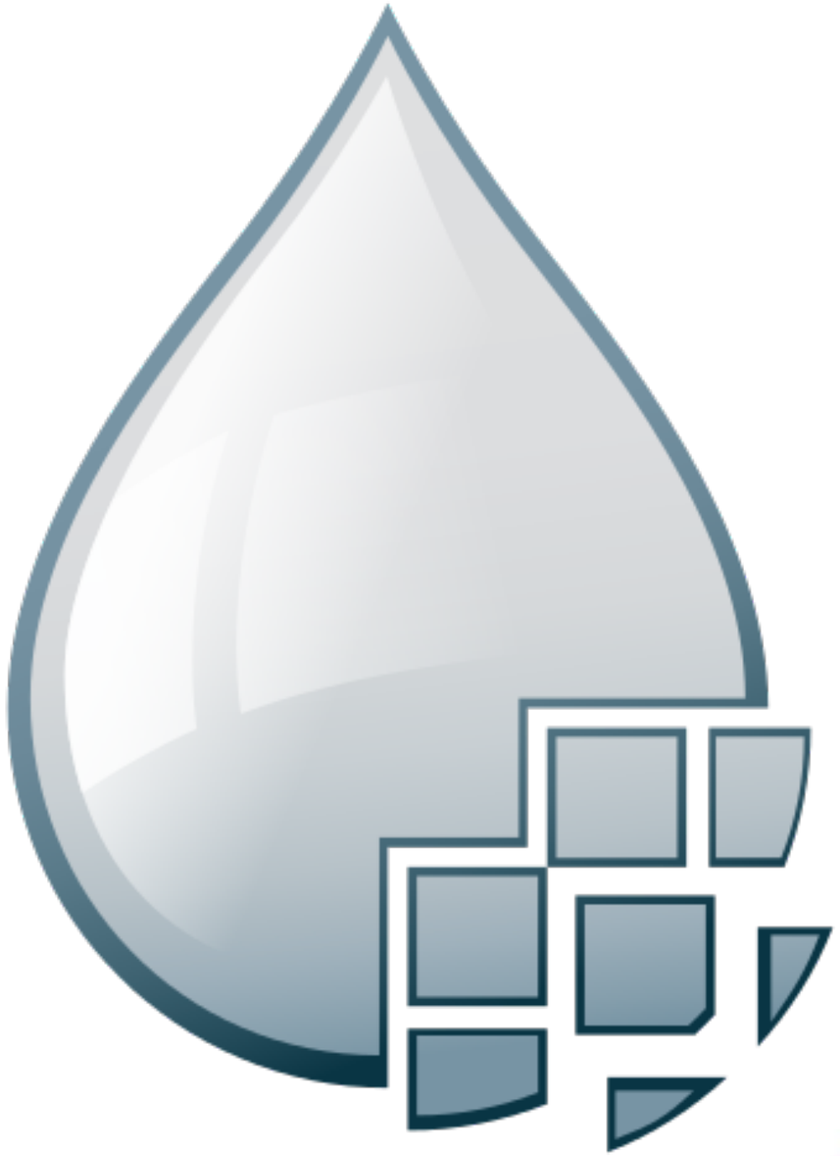
**Apache NiFi site**
https://nifi.apache.org

**Subproject MiNiFi site**
https://nifi.apache.org/minifi/

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

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

**Follow us on Twitter**
@apachenifi






HORTONWORKS

# Learn and share at Birds of a Feather

# IOT, STREAMING & DATA FLOW

Thursday September 21

6:00 pm, C4.6





HORTONWORKS

# Thank You

I’m sticking around for discussions/questions

 @yolopey / @apachenifi

alopresto@apache.org

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




HORTONWORKS