Special Crash Course WiFi

SSID: CrashCourse
PW: DataWorks19

1. Download NiFi 1.8.0
    1. Browse to https://nifi.apache.org/download.html
    2. `wget http://ftp.riken.jp/net/apache/nifi/1.8.0/nifi-1.8.0-bin.tar.gz`
2. Untar file
    1. `tar -xvzf nifi-1.8.0-bin.tar.gz`
3. Change into `$NIFI_HOME` dir
    1. `cd nifi-1.8.0`
4. Start NiFi
    1. `../bin/nifi.sh start`
    2. `tail -f logs/nifi-app.log`

Special Crash Course WiFi

SSID: CrashCourse
PW: DataWorks19

Apache NiFi Crash Course

Andy LoPresto | @yolopey

Sr. Member of Technical Staff at Hortonworks, Apache NiFi PMC & Committer

06 February 2019 Dataworks Summit Melbourne

Acknowledgement of Country

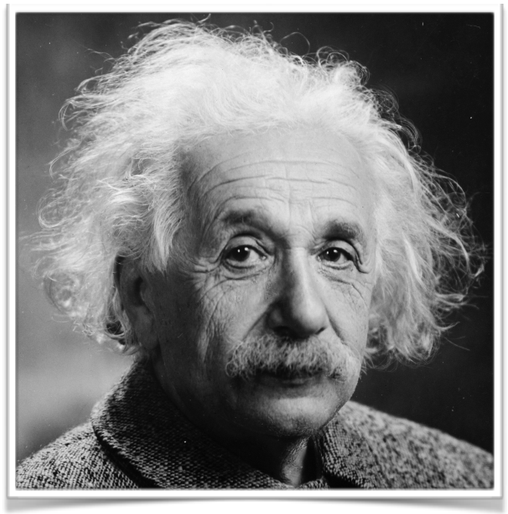
I acknowledge the Traditional Owners of the land on which we are meeting. I pay my respects to their Elders, past and present, and the Aboriginal Elders of other communities who may be here today.

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

Agenda

- Introduction
- What is dataflow?
- What is NiFi?
- What's next?

- All slides provided online, so no need to transcribe

HORTONWORKS

Goals for today

- Not to demonstrate what I know about NiFi
- Teach you something new
- Encourage exploration on topics that interest you
- Empower you to build flows on your own

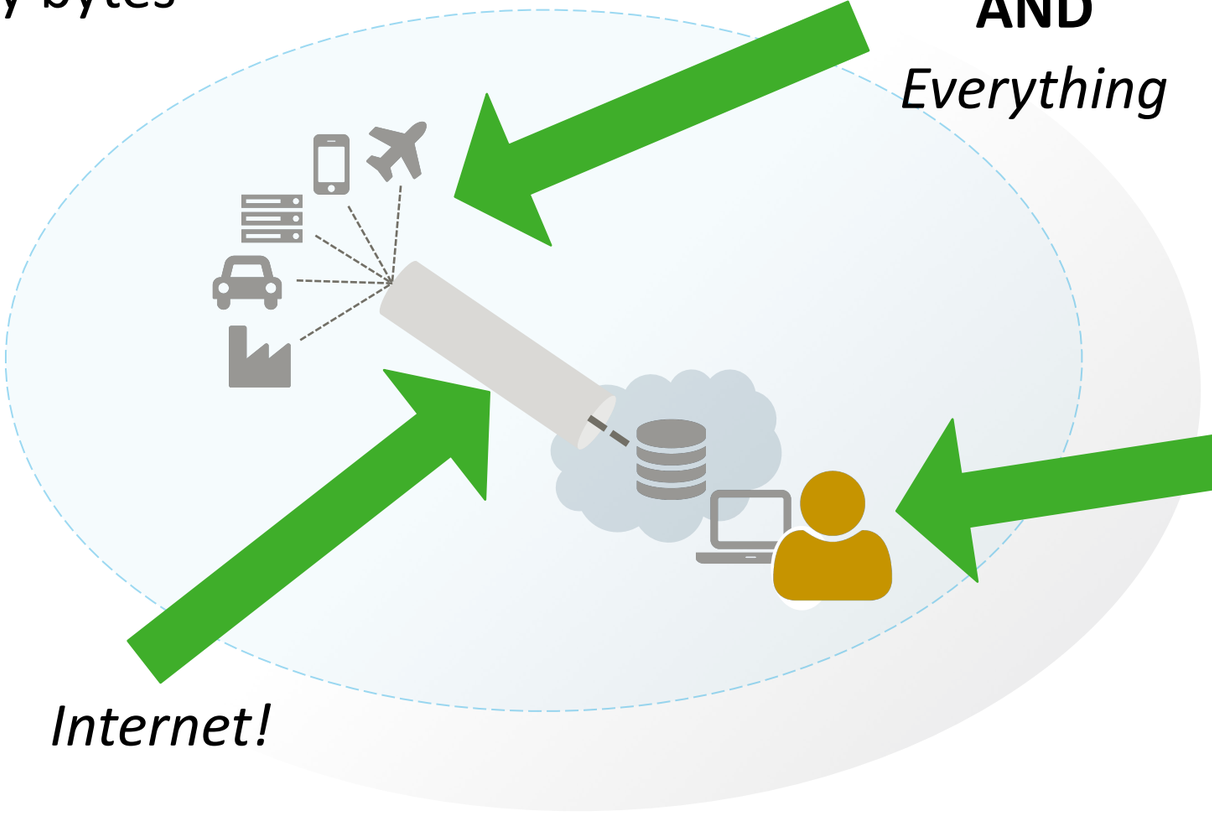
HORTONWORKS®

What is dataflow?

HORTONWORKS®

What is dataflow?

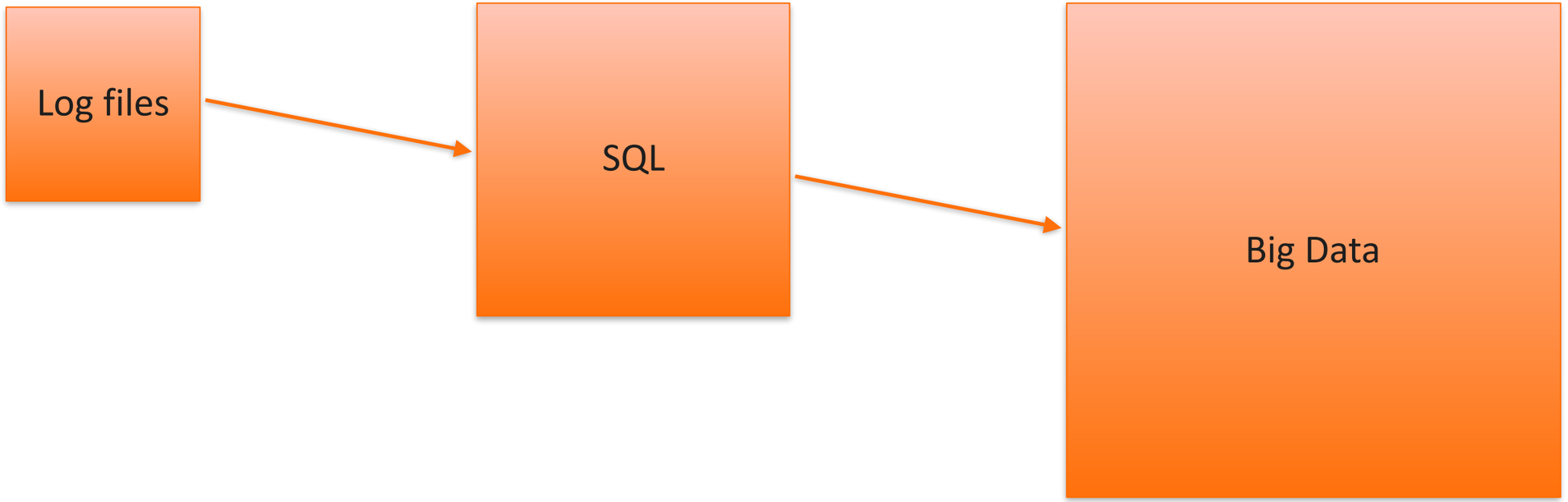
- Moving some content from A to B
- Content could be any bytes
  - Logs
  - HTTP
  - XML
  - CSV
  - Images
  - Video
  - Telemetry



HORTONWORKS

Connecting Data Points Is Easy

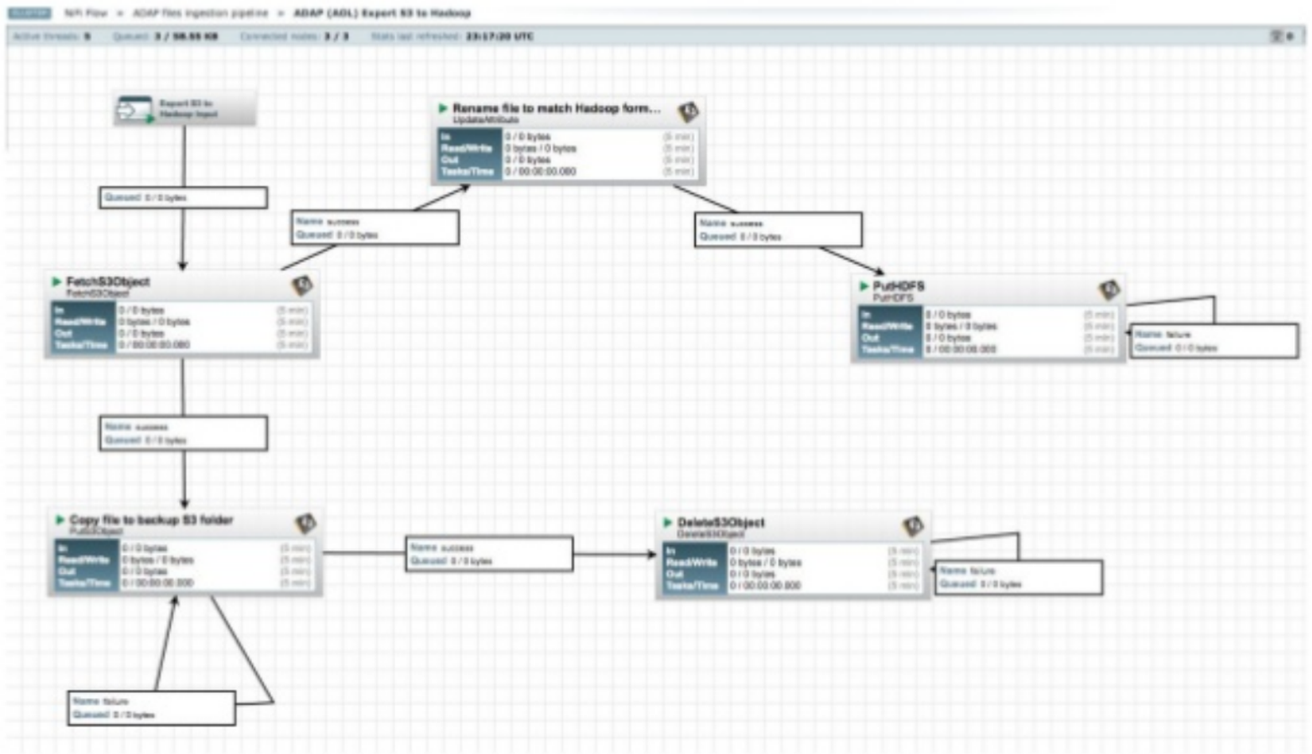
- Simple enough to write a process
  - Bash/Ruby/Python
  - SQL proc
  - etc.



HORTONWORKS

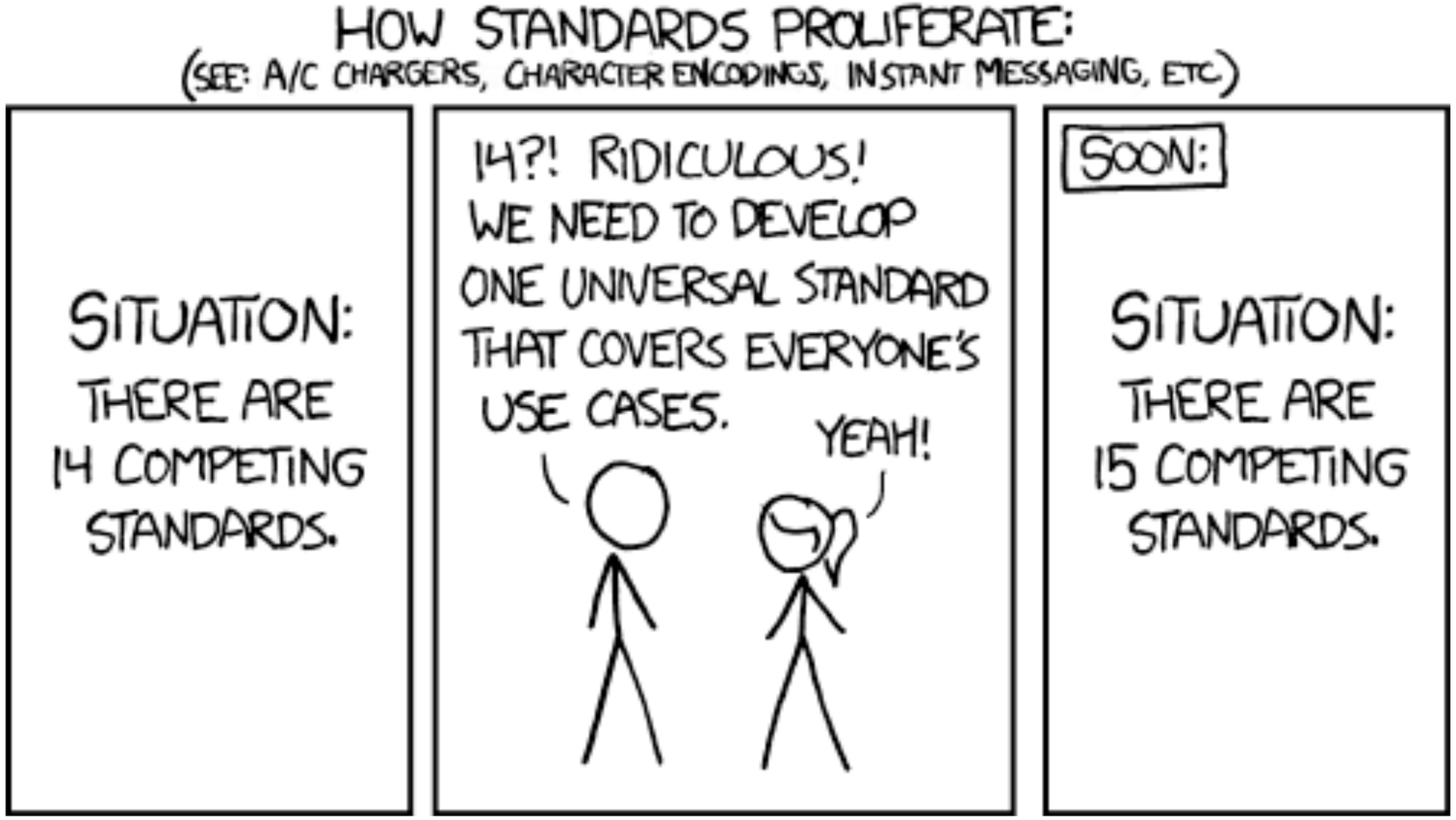
Big Data Is About Scale…

- …and this doesn't scale
- Example use case:
  - AOL Data Processing
  - AWS -> HDFS
  - 20 TB ingested/day
  - Lev Brailovskiy, "Data Ingestion and Distribution with Apache NiFi", Slide 27, 02/2017
  - https://www.slideshare.net/LevBrailovskiy/data-ingestion-and-distribution-with-apache-nifi



HORTONWORKS

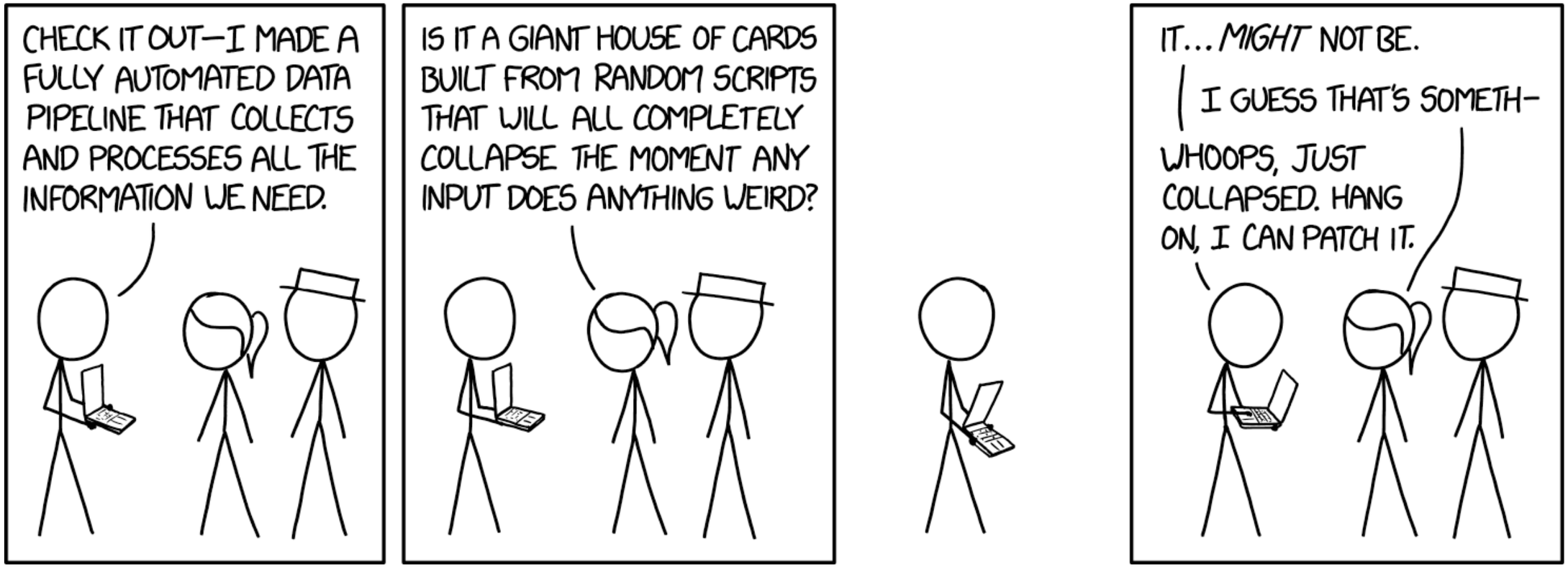
Moving data *effectively* is hard



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

HORTONWORKS

Moving data *effectively* is *really* hard



"Data Pipeline" https://xkcd.com/2054/

HORTONWORKS

Dataflow Challenges In 3 Categories

Data

- Standards
- **Formats**
- Protocols
- Veracity
- Validity
- Schemas
- Partitioning/ Bundling

Infrastructure

- “Exactly Once” Delivery
- Ensuring Security
- **Overcoming** Security
- Credential Management
- Network

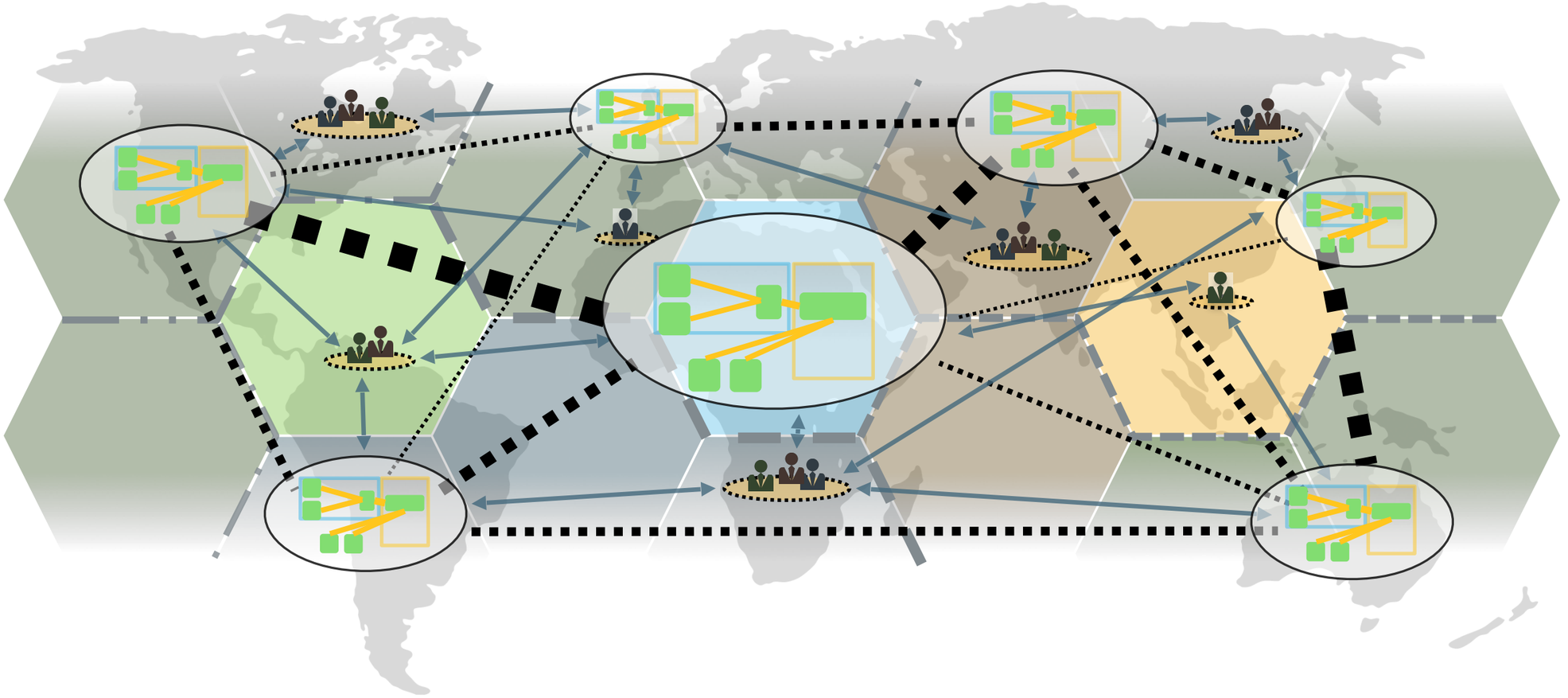
People

- Compliance
- “**That** [person|team|group]”
- **Consumers Change**
- **Requirements Change**
- “Exactly Once” Delivery

HORTONWORKS

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

What is Apache NiFi?

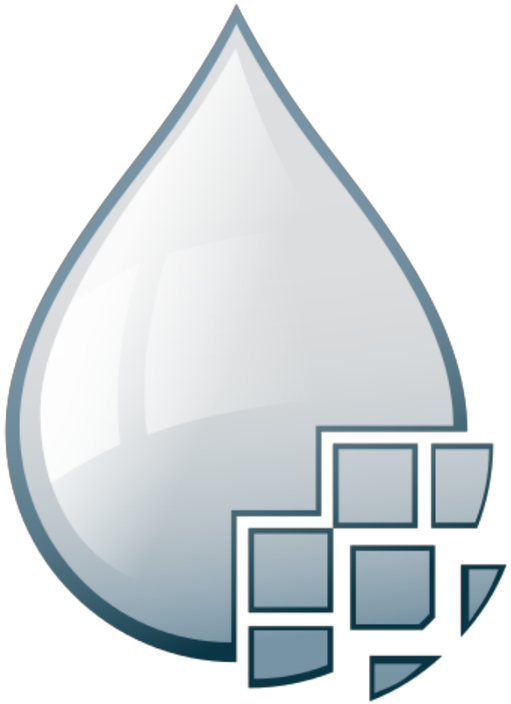
NiFi is based on Flow Based Programming (FBP)

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

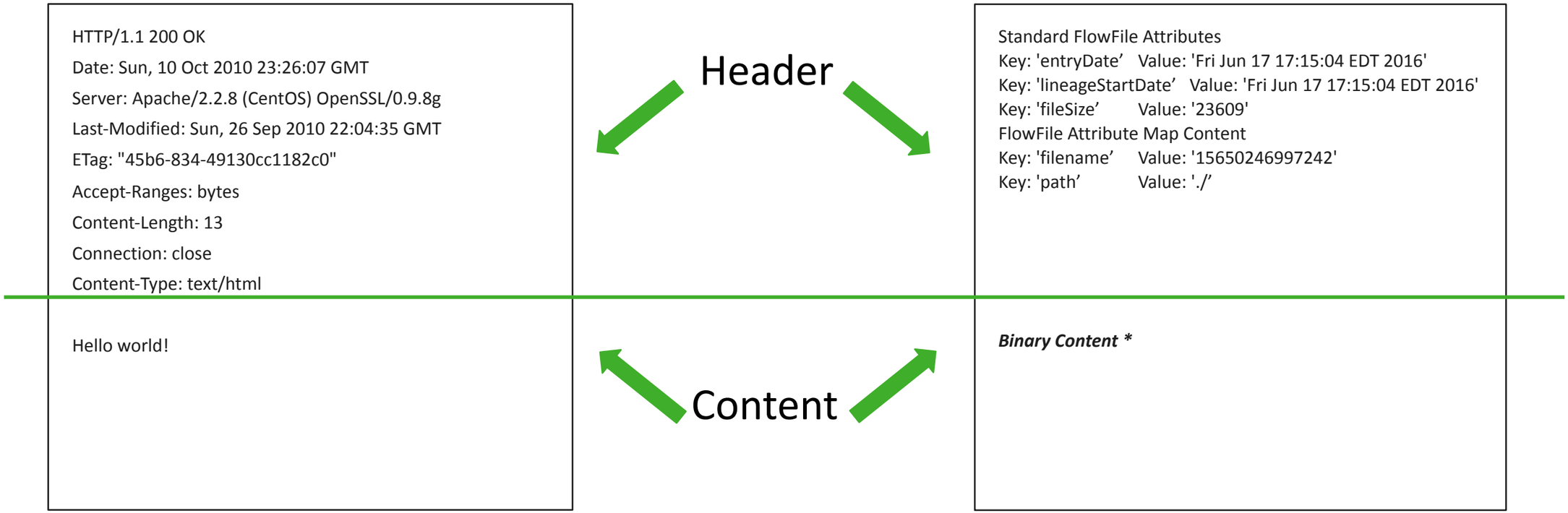
Key Features



- 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

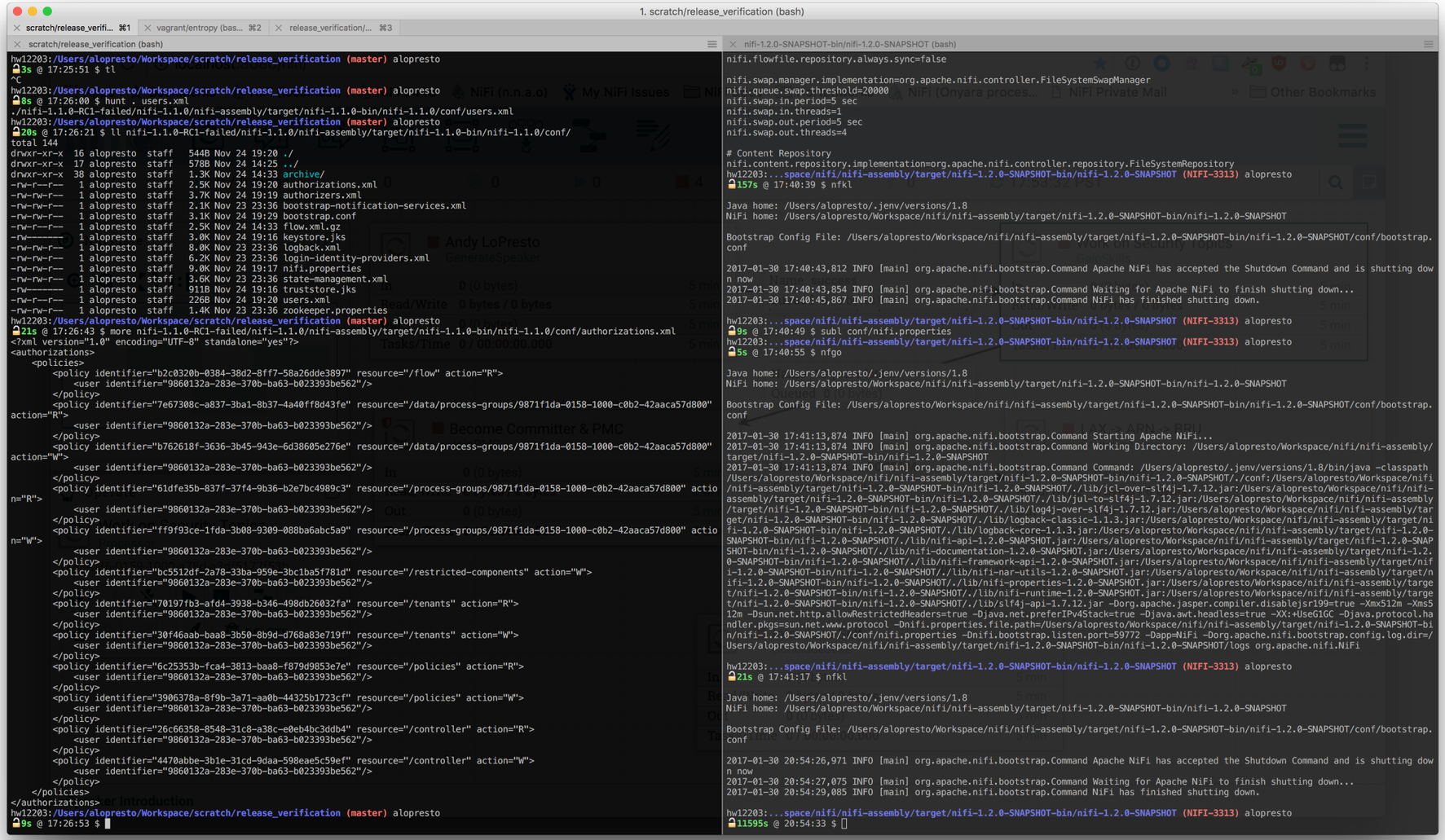
HORTONWORKS

Flowfiles Are Like HTTP Data



User Interface

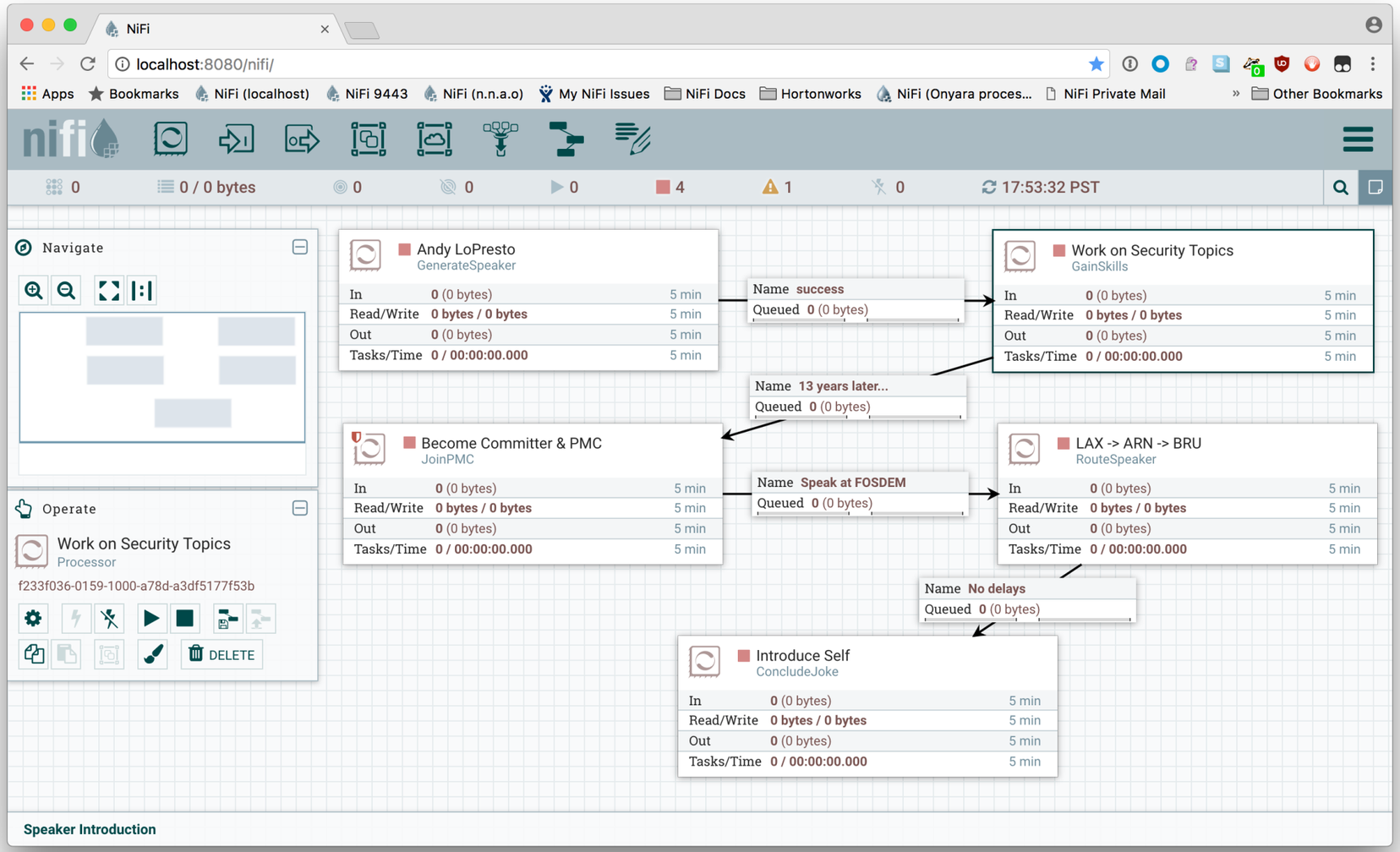
Less of this…



HORTONWORKS

User Interface

Less of this..... more of this



HORTONWORKS

Deeper Ecosystem Integration: 286+ Processors, 61 Controller Services

- 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

- Parse Records
- Convert Records
- Route Text
- Distribute Load
- Route Content
- Generate Table Fetch
- Route Context
- Jolt Transform JSON
- Control Rate
- Prioritized Delivery

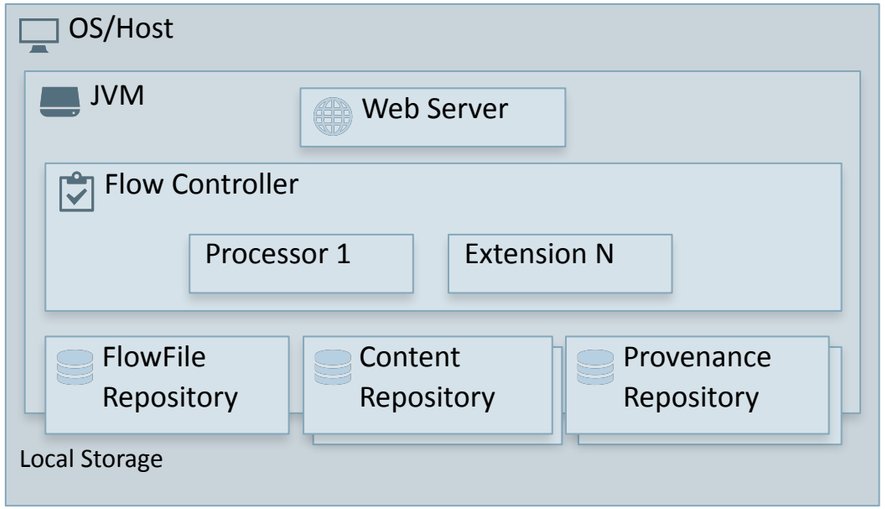
HORTONWORKS

Extension / Integration Points

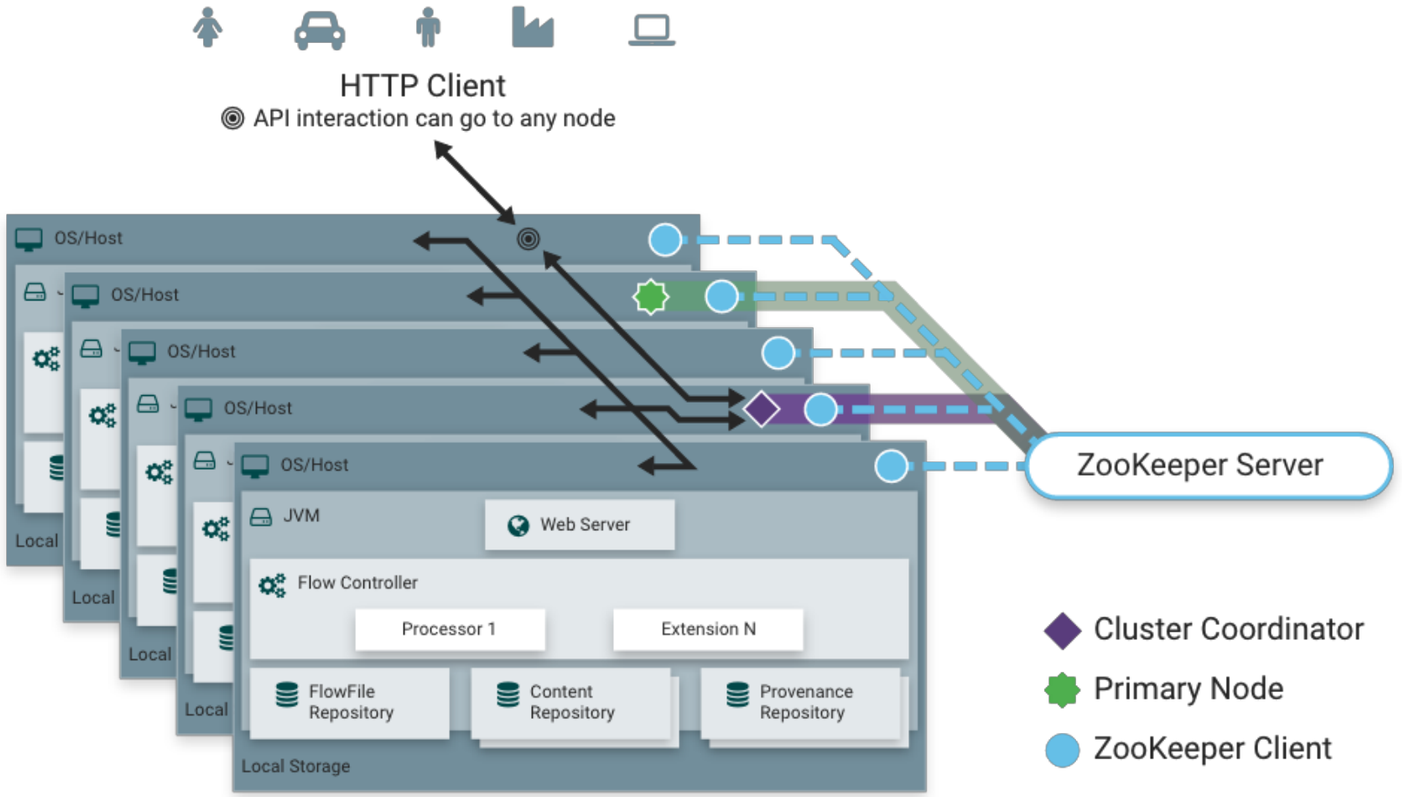
NiFi Term	Description
Flow File Processor	Push/Pull behavior. Custom UI
Reporting Task	Used to push data from NiFi to some external service (metrics, provenance, etc.)
Controller Service	Used to enable reusable components / shared services throughout the flow
REST API	Allows clients to connect to pull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

HORTONWORKS

Architecture



Standalone



Cluster

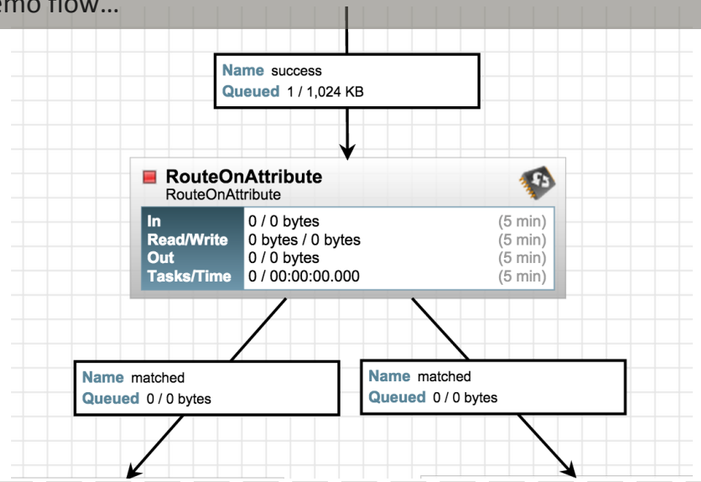
HORTONWORKS

NiFi Architecture – Repositories - Pass by reference

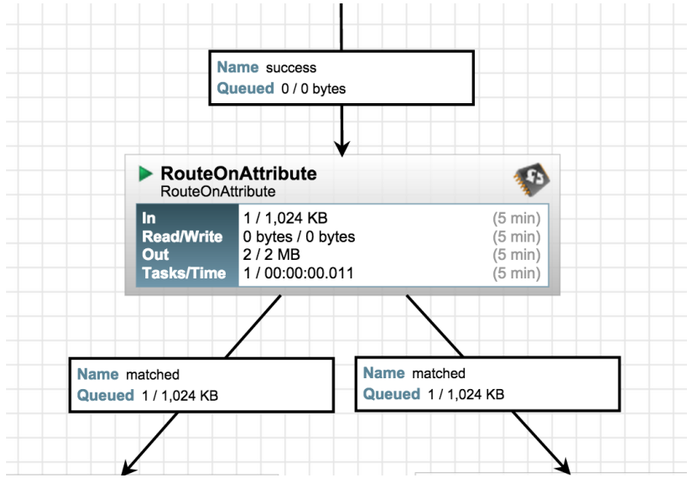
Excerpt of demo flow…

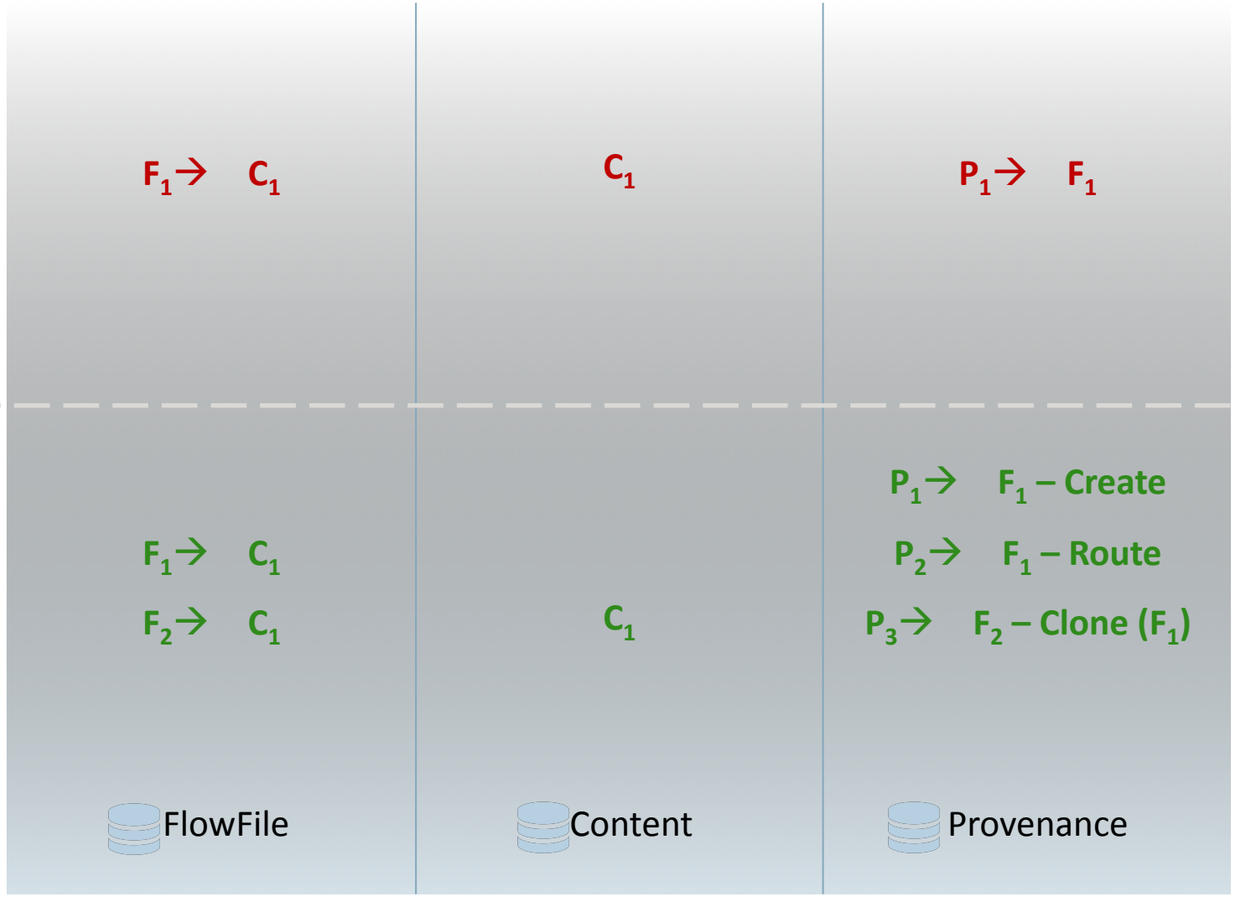
What's happening inside the repositories…

BEFORE



AFTER

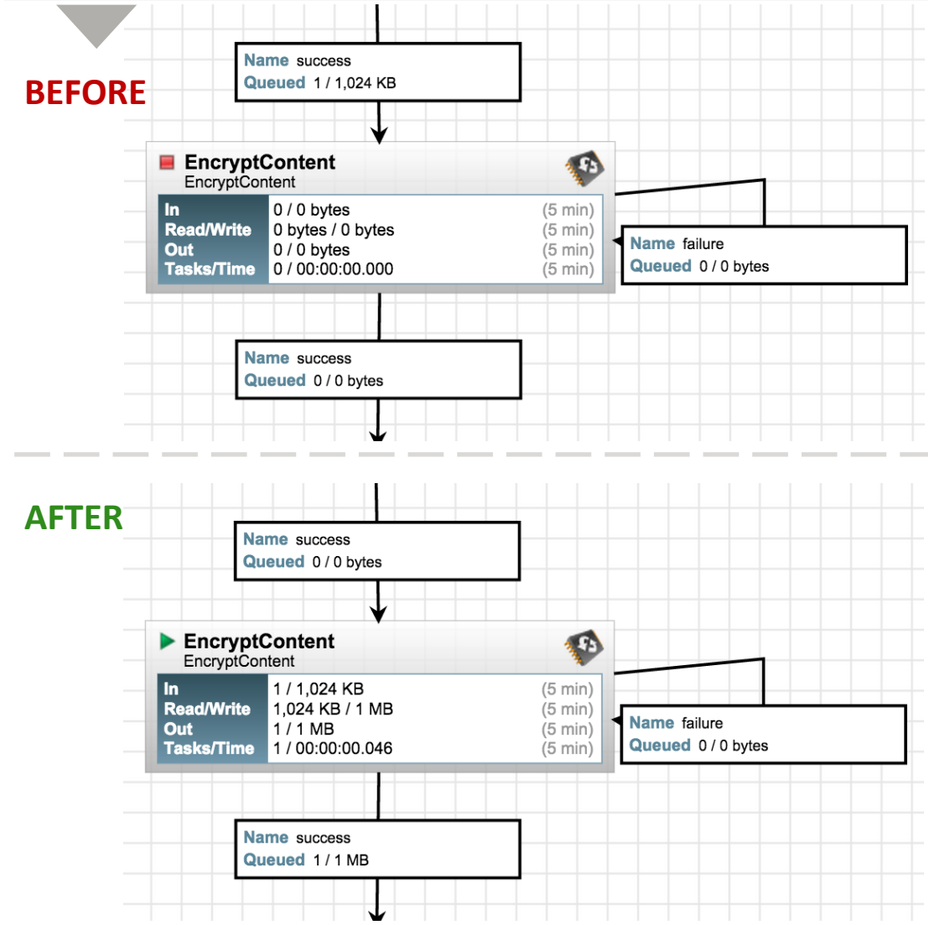




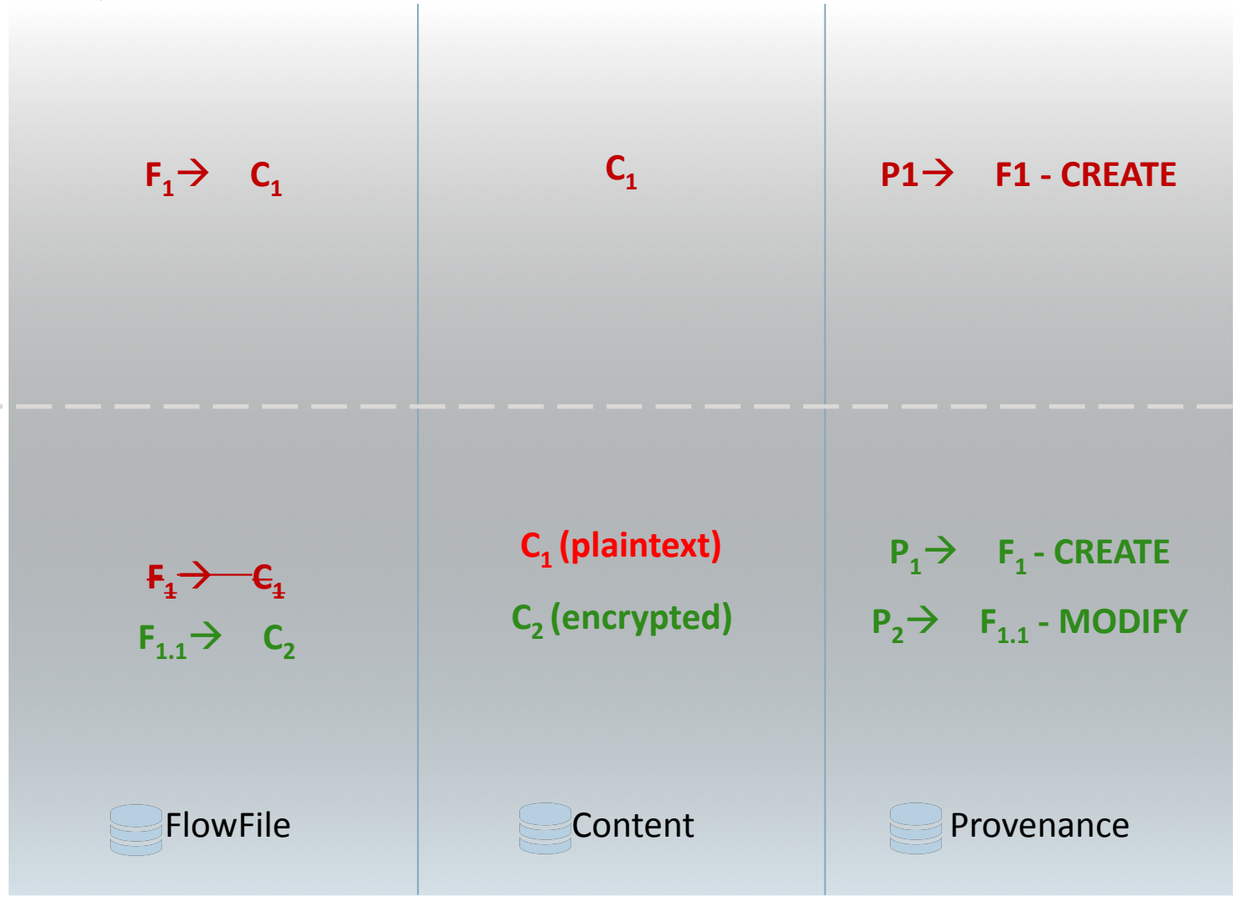
HORTONWORKS

NiFi Architecture – Repositories – Copy on Write

Excerpt of demo flow…



What's happening inside the repositories…



HORTONWORKS

People want to know about their data

Colin the Chicken | Portlandia | IFC

HORTONWORKS

People want to know about their data

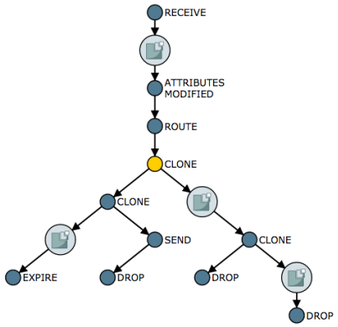
Colin the Chicken | Portlandia | IFC

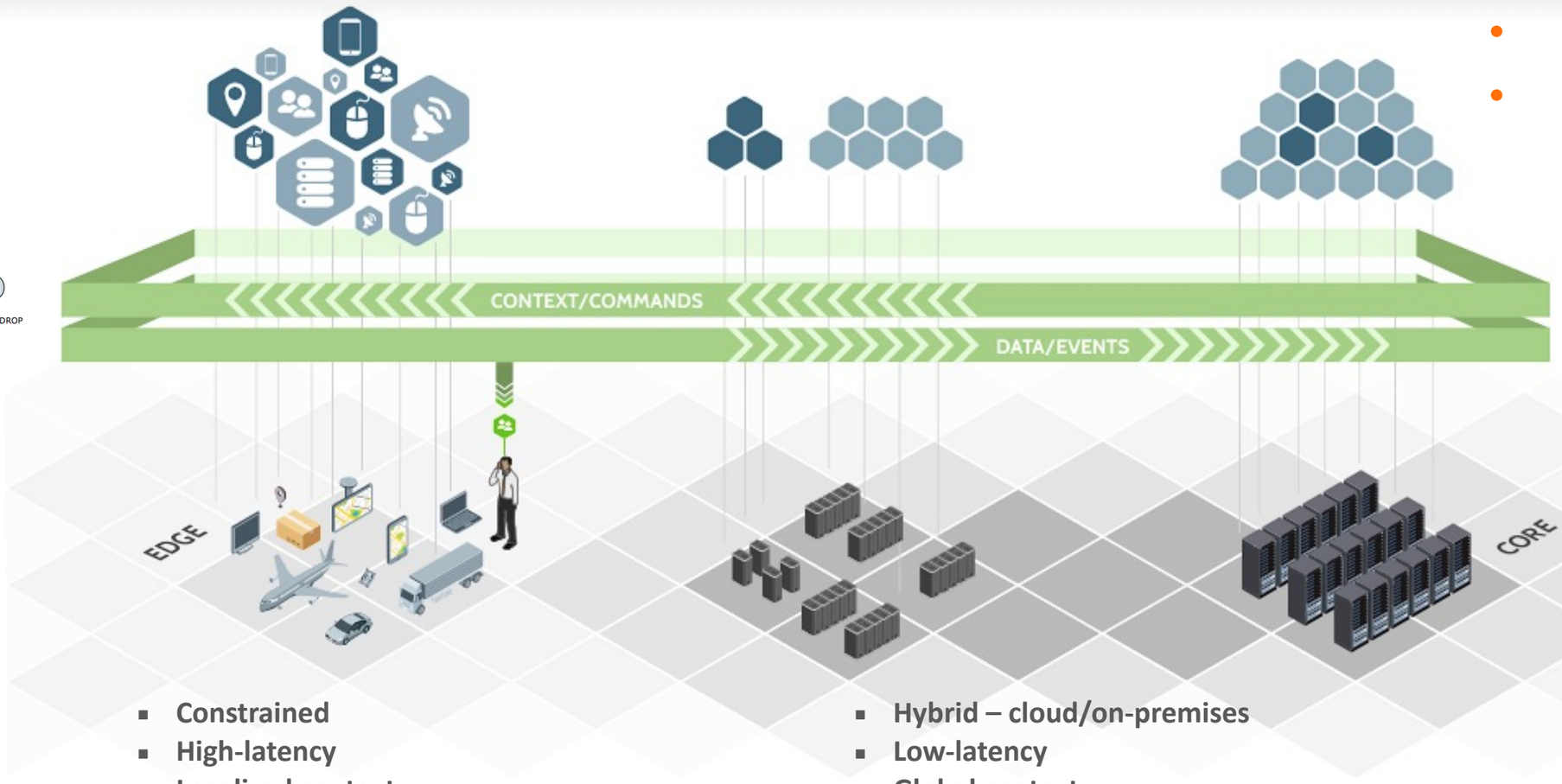
HORTONWORKS

Data Provenance

Origin – attribution
Replay – recovery

Evolution of topologies
Long retention





- **Constrained**
- **High-latency**
- **Localized context**

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

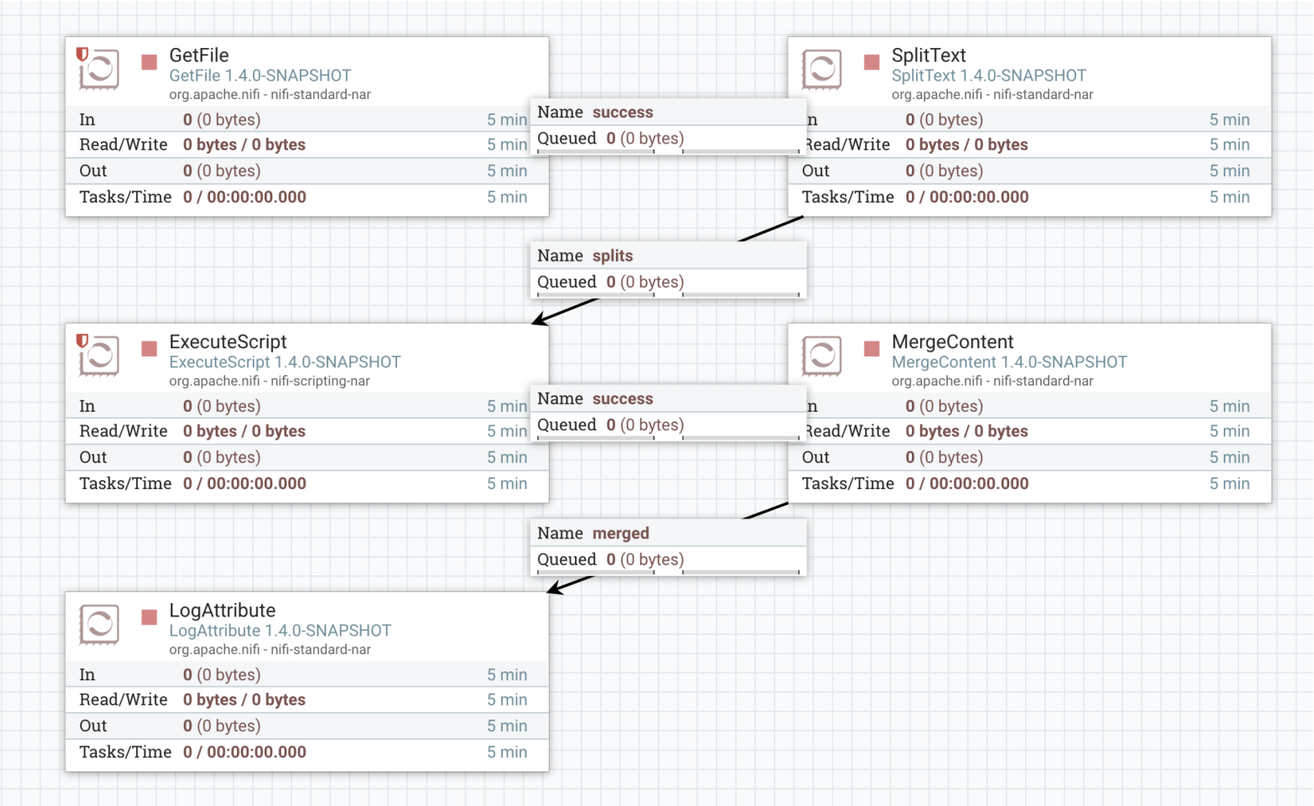
Types of Lineage

- Event
- Configuration

HORTONWORKS

Record Parsing

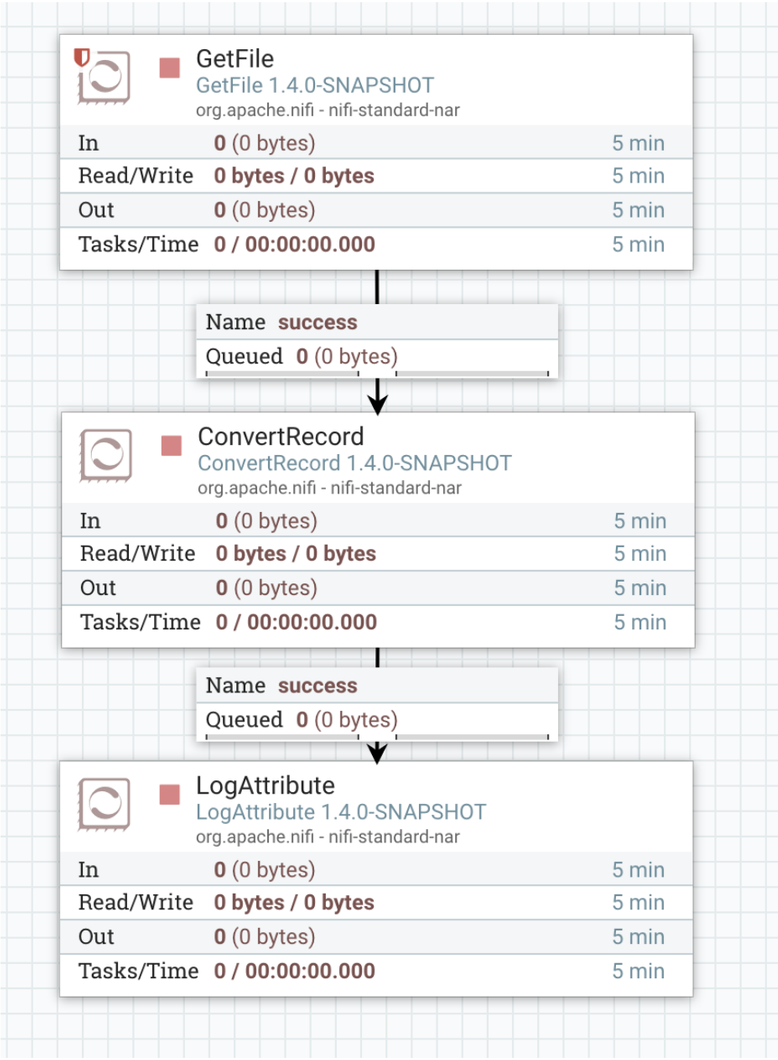
- Previously, data had to be divided into individual flowfiles to perform work
- CSV output with 50k lines would need to be split, operated on, re-merged
- 1 + 50k + 50k + 1 flowfiles = 100k flowfiles



HORTONWORKS

Record Parsing

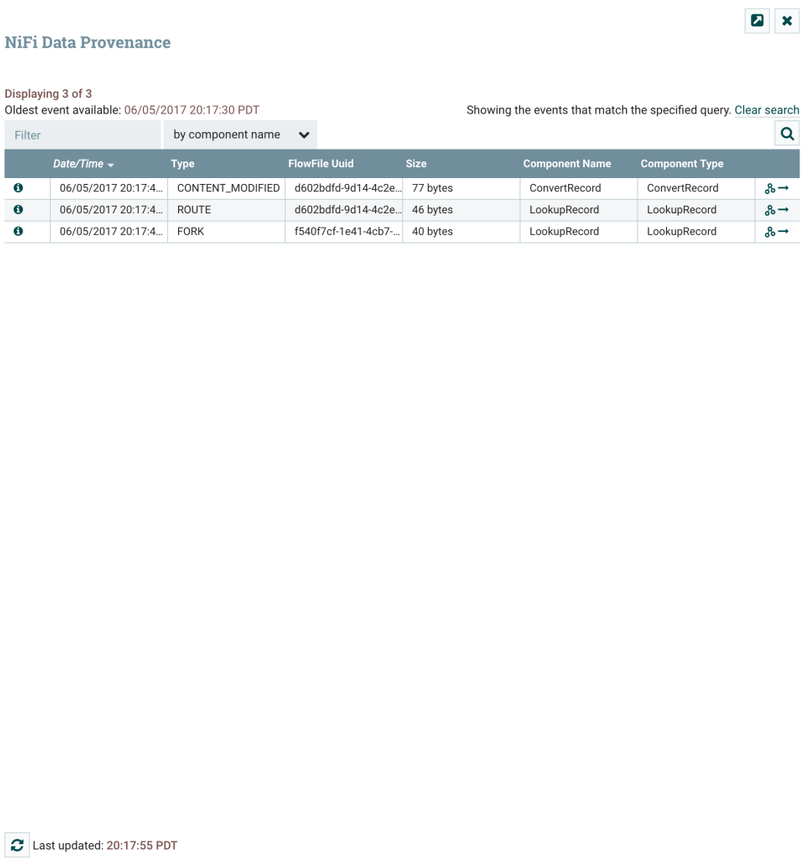
- Now flowfile content can contain many “record” elements
- Read and write with **Reader* and **Writer* Controller Services
- Perform lookups, routing, conversion, SQL queries, validation, and more…
- 1 + 1 flowfiles = 2 flowfiles

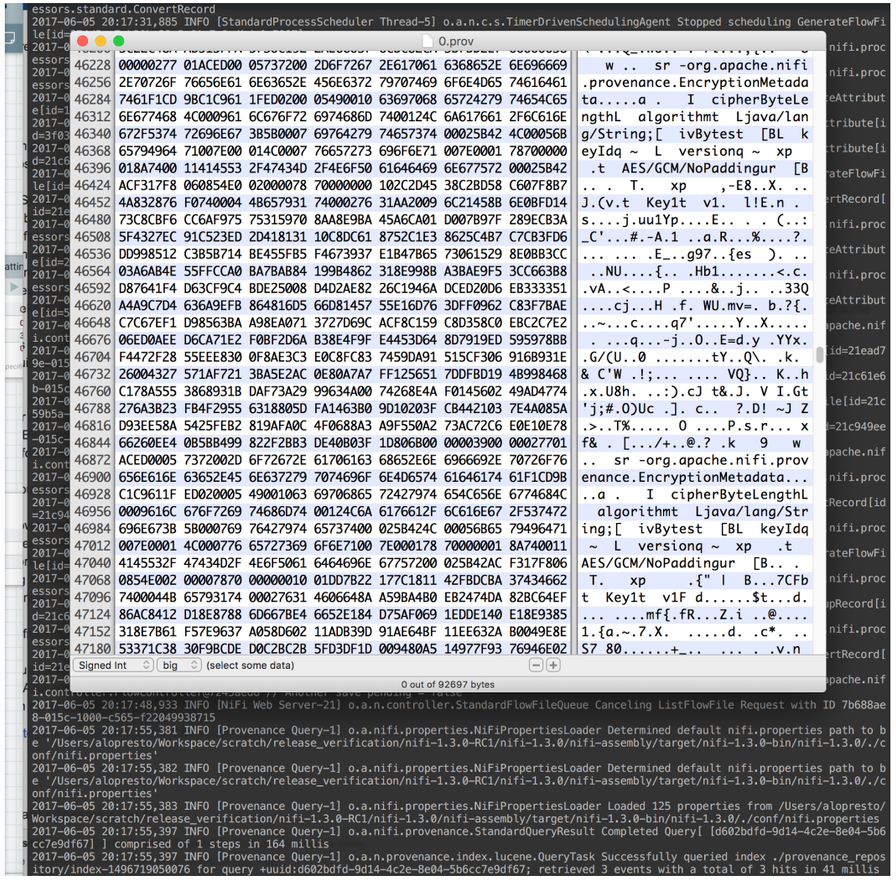


HORTONWORKS

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





What’s Next?

HORTONWORKS

New Announcements

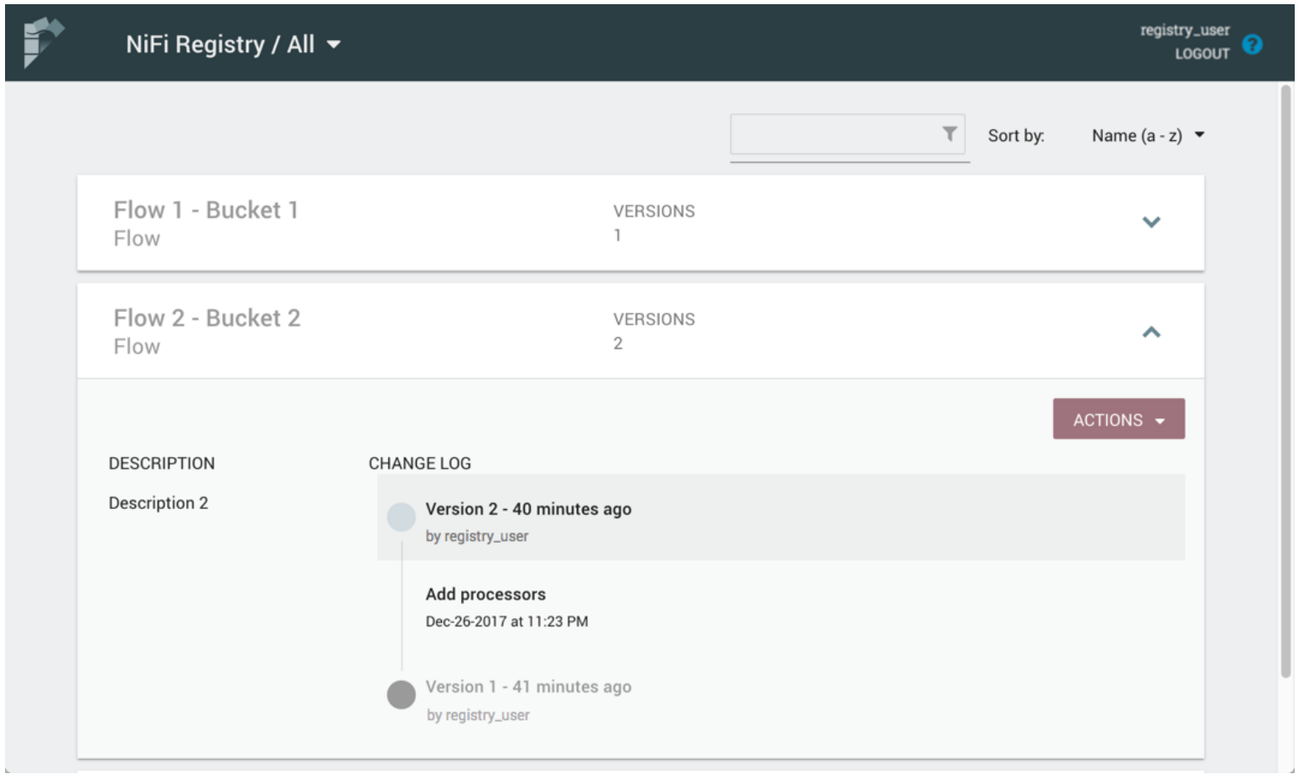
- NiFi 1.8.0 — 26 Oct 2018 (212+ Jiras)
  - Jetty, DB improvements
  - Auto load-balancing queues
  - TLS Toolkit w/ external CA
  - Record processor improvements
- MiNiFi C++ 0.5.0 — 6 June 2018
- MiNiFi Java 0.5.0 — 7 July 2018
- NiFi Registry 0.3.0 — 25 Sept 2018

HORTONWORKS

NiFi Registry for Dataflows

Introducing Apache NiFi Registry 0.3.0

- Previously, flows were exported via XML templates
  - Didn’t contain sensitive values
  - Couldn’t be updated in-place
  - No tracking system
- NiFi Registry brings asset management as first-class citizen to NiFi
- Flows can be versioned
- Flows can be promoted between environments



HORTONWORKS

Community Health

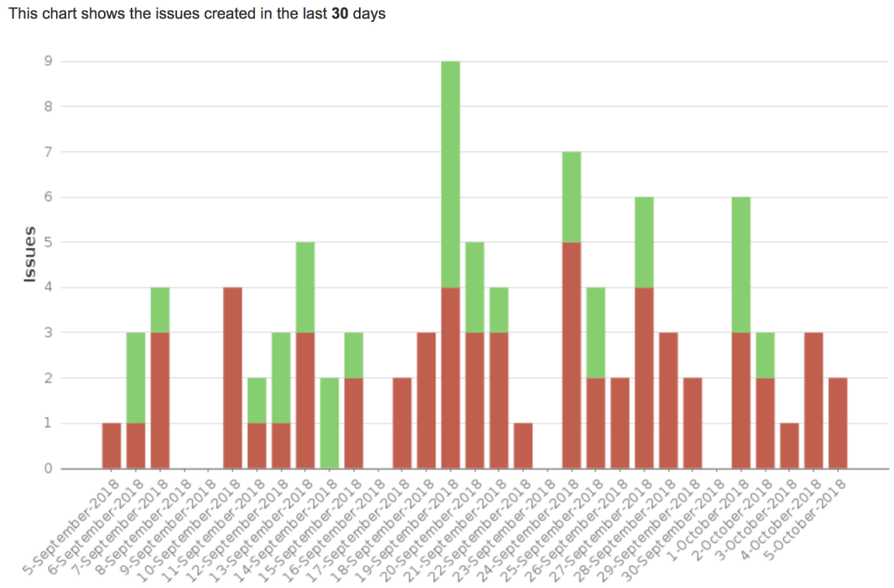
apache / nifi

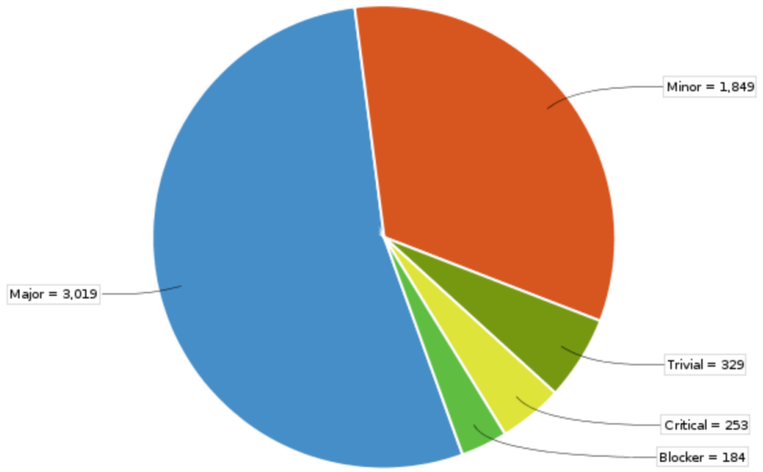
Unwatch 154 | Unstar 1,103 | Fork 1,041

Code | Pull requests 145 | Projects 0 | Insights

Mirror of Apache NiFi

4,720 commits | 52 branches | 63 releases | 200 contributors | Apache-2.0





HORTONWORKS

Learn more and join us

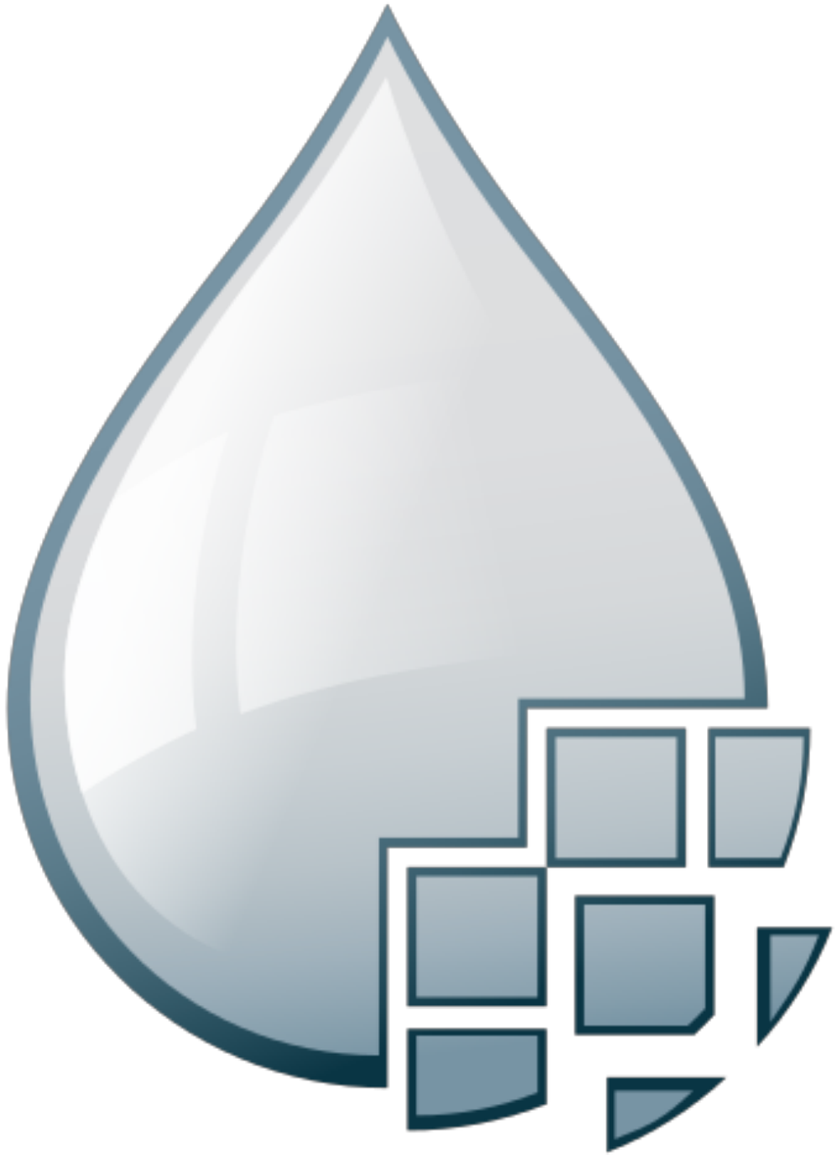
Apache NiFi site
https://nifi.apache.org

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

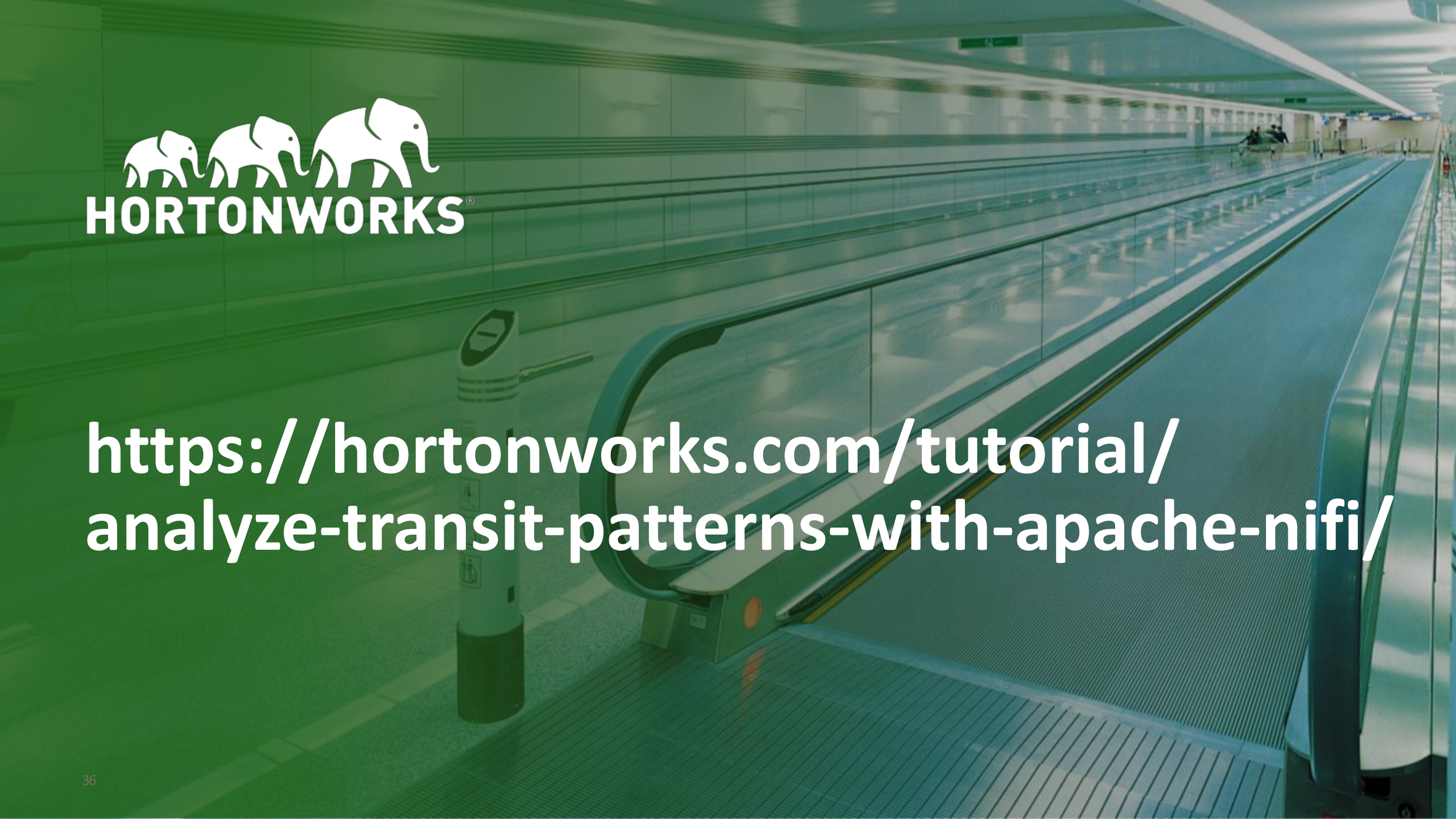


HORTONWORKS

More NiFi Today…

Title	Time	Room
The First Mile – Edge and IoT Data Collection with Apache NiFi and MiNiFi	1100 - 1140	Room 103
Apache NiFi Crash Course	1400 - 1600	Room 109
Dataflow Management From Edge to Core with Apache NiFi	1650 - 1730	Room 112
Using Spark Streaming and NiFi for the Next Generation of ETL in the Enterprise	1650 - 1730	Room 103

HORTONWORKS



https://hortonworks.com/tutorial/analyze-transit-patterns-with-apache-nifi/

Thank you

alopresto@hortonworks.com | alopresto@apache.org | @yolopey

github.com/alopresto/slides