# Hands-on with Apache NiFi and MiNiFi

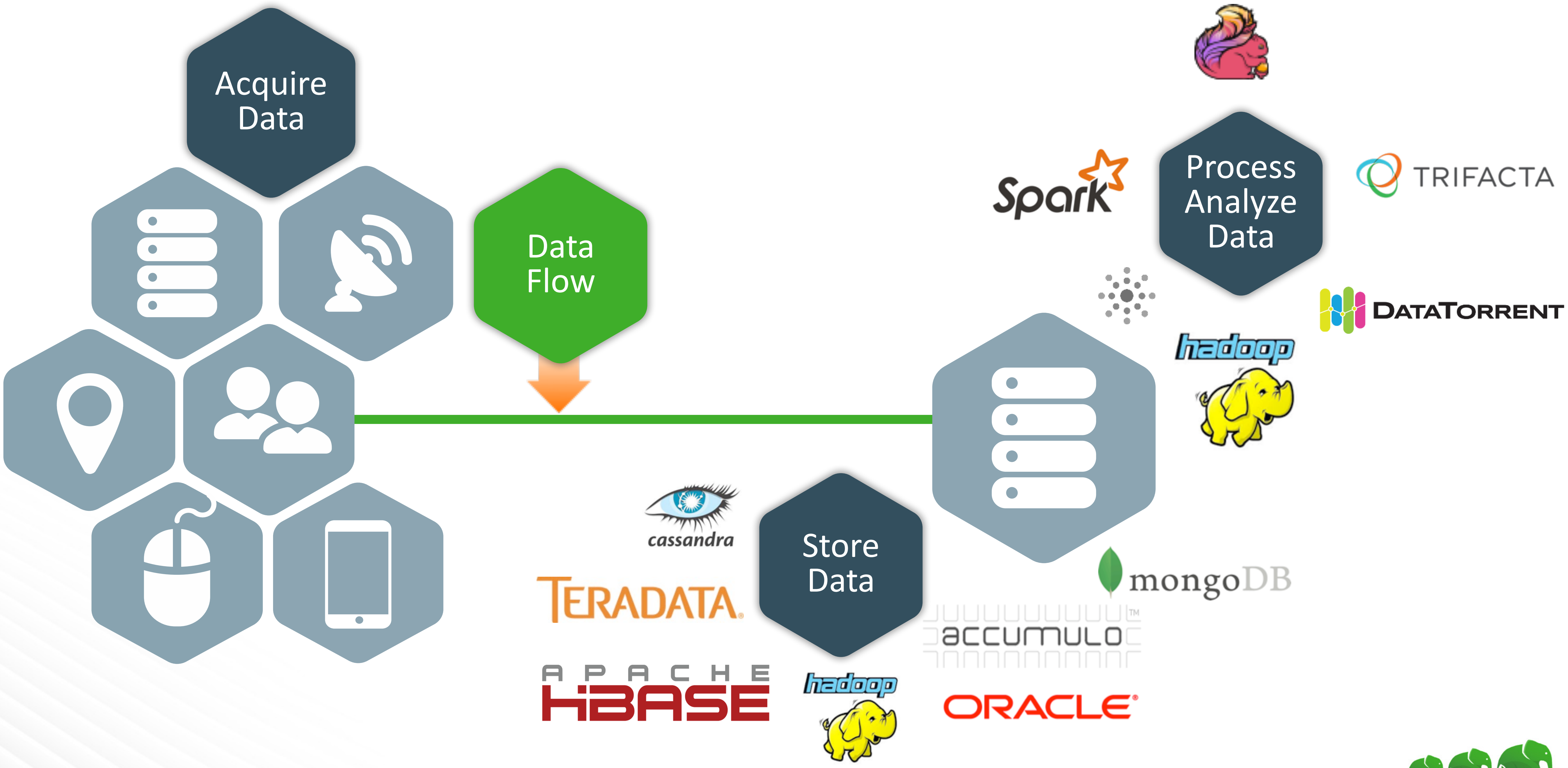
Andrew Psaltis - @itmdata

Berlin Buzzwords 2017

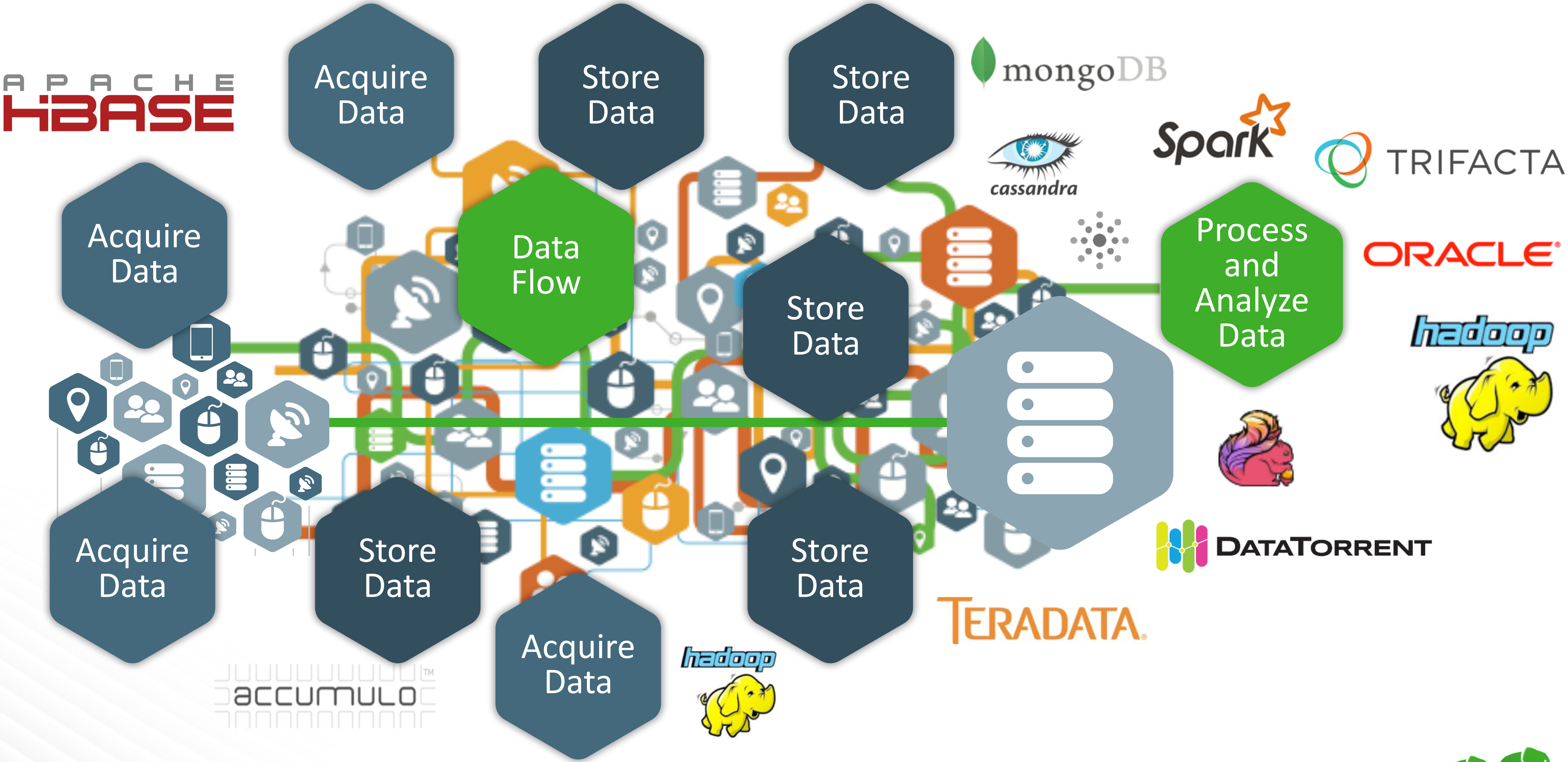
HORTONWORKS®
POWERING THE FUTURE OF DATA™

# Dataflow and the associated problems

# Simplistic View of Dataflows: Easy, Definitive

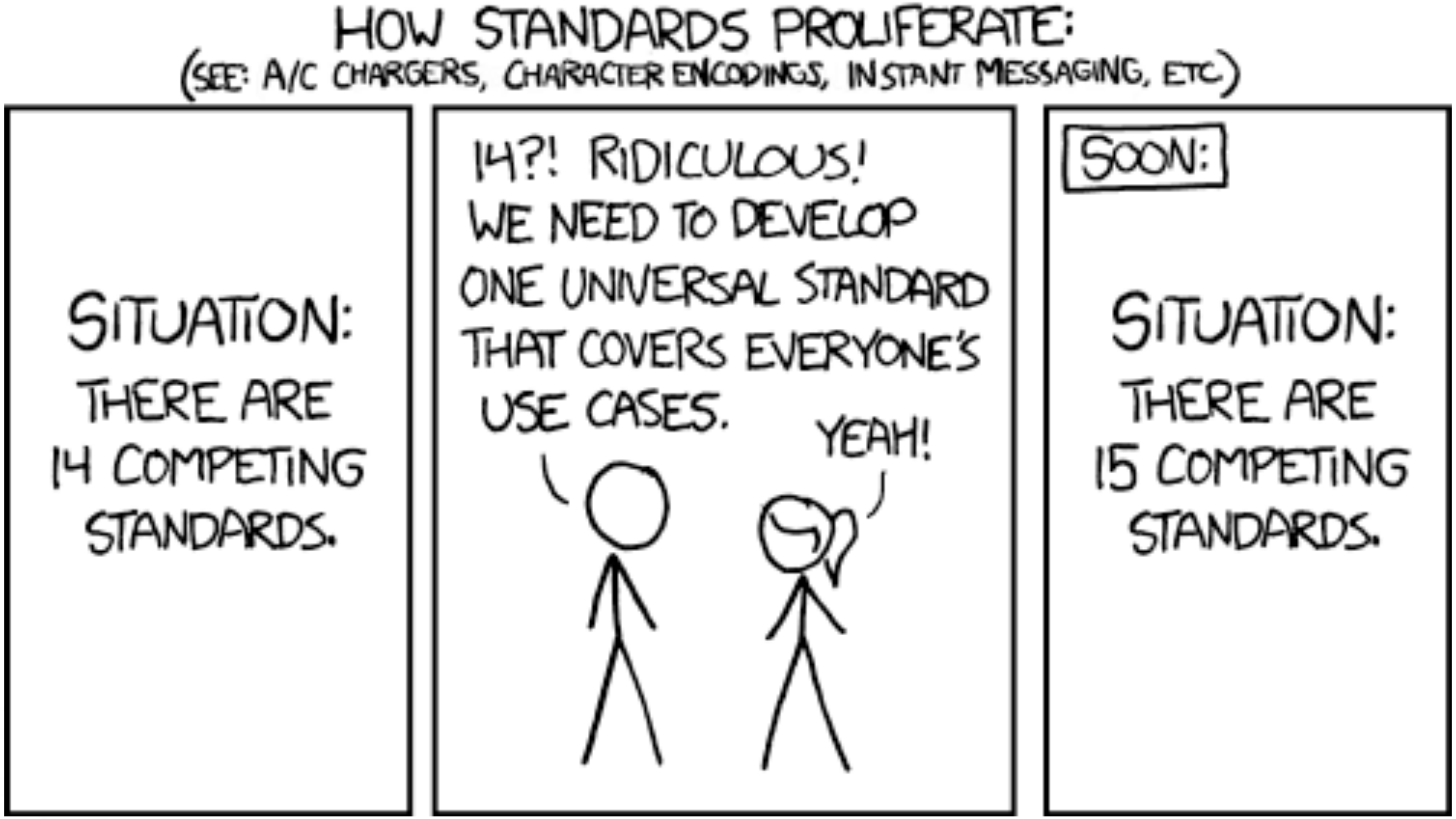


# Realistic View of Dataflows: Complex, Convoluted



HORTONWORKS

# Moving data *effectively* is hard


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

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

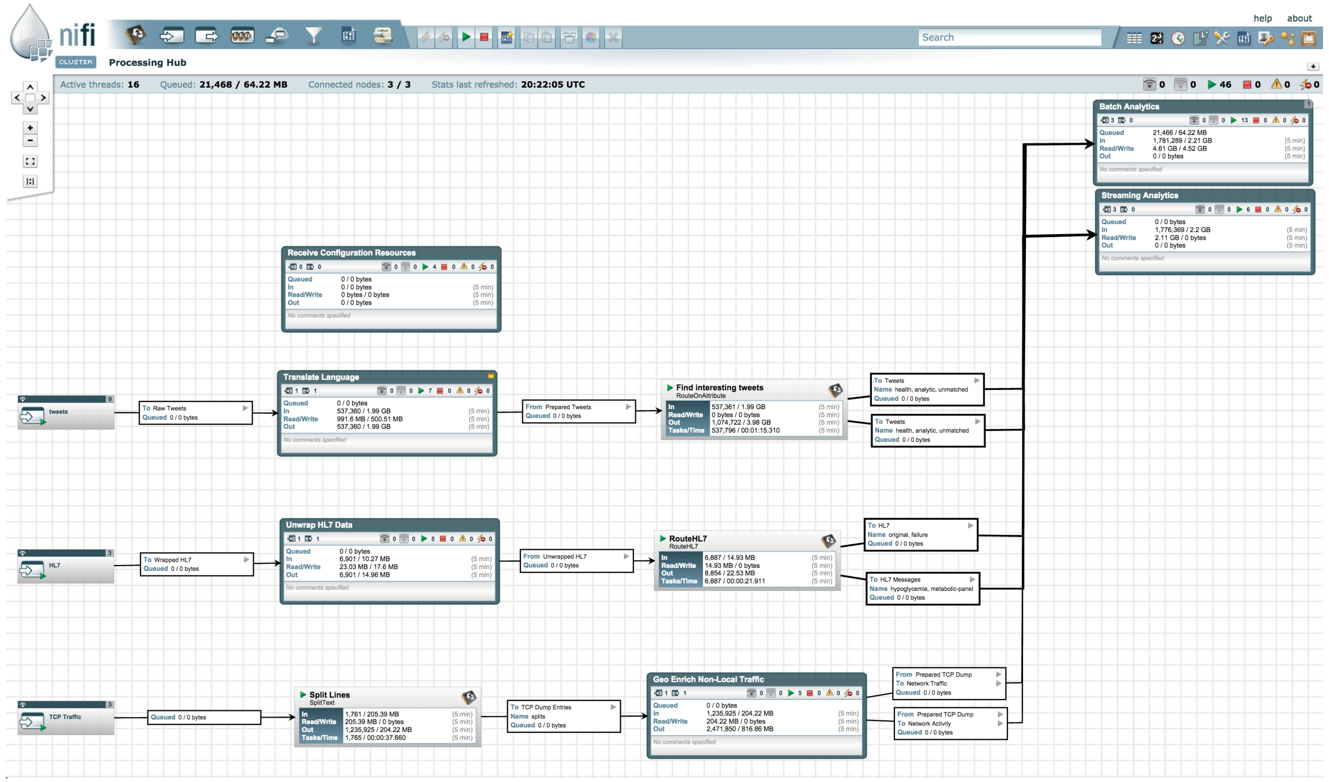
# Apache NiFi

Hortonworks

# Apache NiFi

## Dataflow

- Web-based User Interface for creating, monitoring, & controlling data flows
- Directed graphs of data routing and transformation
- Highly configurable - modify data flow at runtime, dynamically prioritize data
- Easily extensible through development of custom components
- Data Provenance tracks data through entire system



Active threads: 16 Queued: 21,468 / 64.22 MB Connected nodes: 3 / 3 Stats last refreshed: 20:22:05 UTC

0 0 46 0 0 0

**Receive Configuration Resources**
0 0 0 0 4 0 0 0
Queued 0 / 0 bytes
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
No comments specified

To Raw Tweets
Queued 0 / 0 bytes

**Translate Language**
1 1 0 0 7 0 0 0
Queued 0 / 0 bytes
In 537,360 / 1.99 GB (5 min)
Read/Write 991.6 MB / 500.51 MB (5 min)
Out 537,360 / 1.99 GB (5 min)
No comments specified

From Prepared Tweets
Queued 0 / 0 bytes

**Find interesting tweets**
RouteOnAttribute
In 537,361 / 1.99 GB (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 1,074,722 / 3.98 GB (5 min)
Tasks/Time 537,796 / 00:01:15.310 (5 min)

To Tweets
Name health, analytic, unmatched
Queued 0 / 0 bytes

To Tweets
Name health, analytic, unmatched
Queued 0 / 0 bytes

To Wrapped HL7
Queued 0 / 0 bytes

**Unwrap HL7 Data**
1 1 0 0 5 0 0 0
Queued 0 / 0 bytes
In 6,901 / 10.27 MB (5 min)
Read/Write 23.03 MB / 17.6 MB (5 min)
Out 6,901 / 14.96 MB (5 min)
No comments specified

From Unwrapped HL7
Queued 0 / 0 bytes

**RouteHL7**
RouteHL7
In 6,887 / 14.93 MB (5 min)
Read/Write 14.93 MB / 0 bytes (5 min)
Out 8,854 / 22.53 MB (5 min)
Tasks/Time 6,887 / 00:00:21.911 (5 min)

To HL7
Name original, failure
Queued 0 / 0 bytes

To HL7 Messages
Name hypoglycemia, metabolic-panel
Queued 0 / 0 bytes

Queued 0 / 0 bytes

**Split Lines**
SplitText
In 1,761 / 205.39 MB (5 min)
Read/Write 205.39 MB / 0 bytes (5 min)
Out 1,235,925 / 204.22 MB (5 min)
Tasks/Time 1,765 / 00:00:37.660 (5 min)

To TCP Dump Entries
Name splits
Queued 0 / 0 bytes

**Geo Enrich Non-Local Traffic**
1 1 0 0 5 0 0 0
Queued 0 / 0 bytes
In 1,235,925 / 204.22 MB (5 min)
Read/Write 204.22 MB / 0 bytes (5 min)
Out 2,471,850 / 816.86 MB (5 min)
No comments specified

From Prepared TCP Dump
To Network Traffic
Queued 0 / 0 bytes

From Prepared TCP Dump
To Network Activity
Queued 0 / 0 bytes

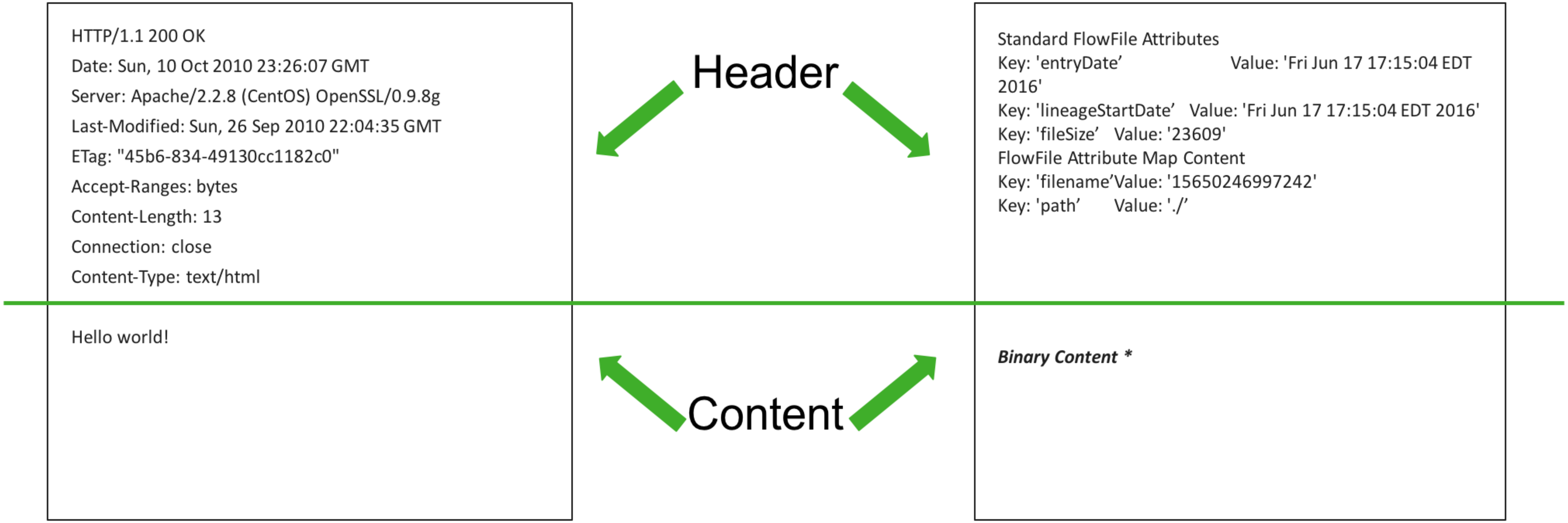
**Batch Analytics**
3 0 0 0 13 0 0 0
Queued 21,466 / 64.22 MB
In 1,781,289 / 2.21 GB (5 min)
Read/Write 4.61 GB / 4.52 GB (5 min)
Out 0 / 0 bytes (5 min)
No comments specified

**Streaming Analytics**
3 0 0 0 6 0 0 0
Queued 0 / 0 bytes
In 1,776,369 / 2.2 GB (5 min)
Read/Write 2.11 GB / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
No comments specified

*[1] https://nifi.apache.org/*

# NiFi - Terminology

| HTTP Data | FlowFile |
|---|---|



HTTP/1.1 200 OK
Date: Sun, 10 Oct 2010 23:26:07 GMT
Server: Apache/2.2.8 (CentOS) OpenSSL/0.9.8g
Last-Modified: Sun, 26 Sep 2010 22:04:35 GMT
ETag: "45b6-834-49130cc1182c0"
Accept-Ranges: bytes
Content-Length: 13
Connection: close
Content-Type: text/html

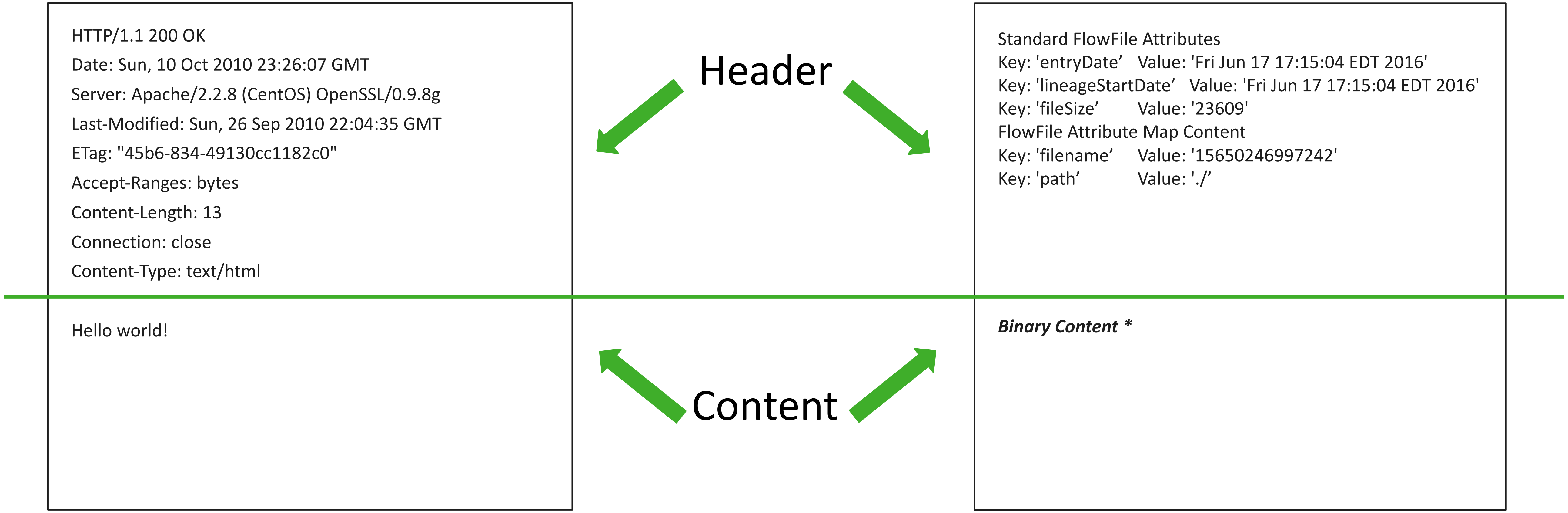
Hello world!

Standard FlowFile Attributes
Key: 'entryDate’ Value: 'Fri Jun 17 17:15:04 EDT 2016'
Key: 'lineageStartDate’ Value: 'Fri Jun 17 17:15:04 EDT 2016'
Key: 'fileSize’ Value: '23609'
FlowFile Attribute Map Content
Key: 'filename’Value: '15650246997242'
Key: 'path’ Value: './’

***Binary Content ****

# FlowFiles are like HTTP data

| HTTP Data | FlowFile |
|---|---|



HTTP/1.1 200 OK
Date: Sun, 10 Oct 2010 23:26:07 GMT
Server: Apache/2.2.8 (CentOS) OpenSSL/0.9.8g
Last-Modified: Sun, 26 Sep 2010 22:04:35 GMT
ETag: "45b6-834-49130cc1182c0"
Accept-Ranges: bytes
Content-Length: 13
Connection: close
Content-Type: text/html

Hello world!

Standard FlowFile Attributes
Key: 'entryDate’ Value: 'Fri Jun 17 17:15:04 EDT 2016'
Key: 'lineageStartDate’ Value: 'Fri Jun 17 17:15:04 EDT 2016'
Key: 'fileSize’ Value: '23609'
FlowFile Attribute Map Content
Key: 'filename’ Value: '15650246997242'
Key: 'path’ Value: './’

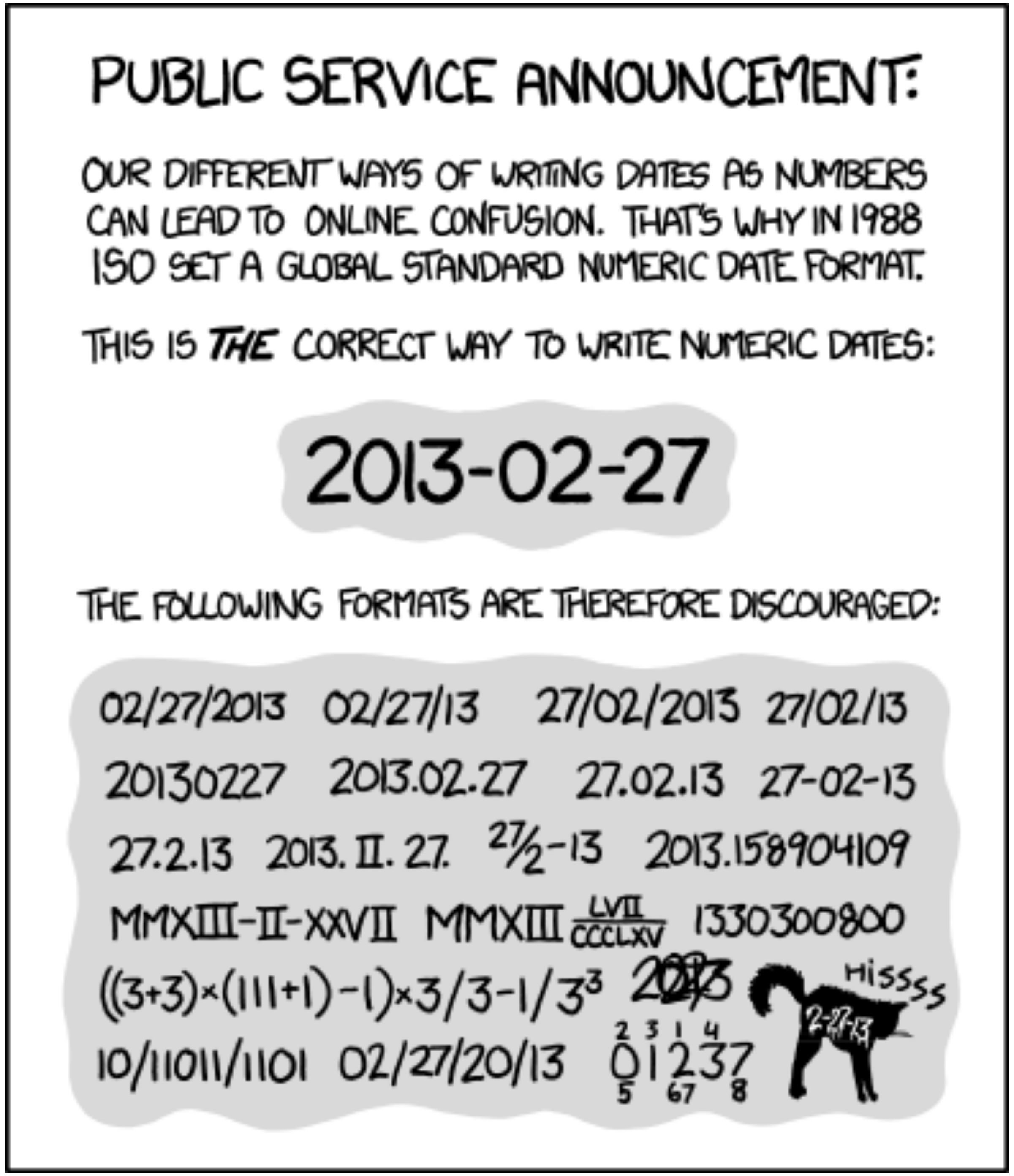
***Binary Content ****

# 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. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
Hisssss
2-27-13
10/11011/1101 02/27/20/13 01237
2 3 1 4
5 67 8

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

# NiFi - Terminology

- FlowFile
  - Unit of data moving through the system
  - Content + Attributes (key/value pairs)

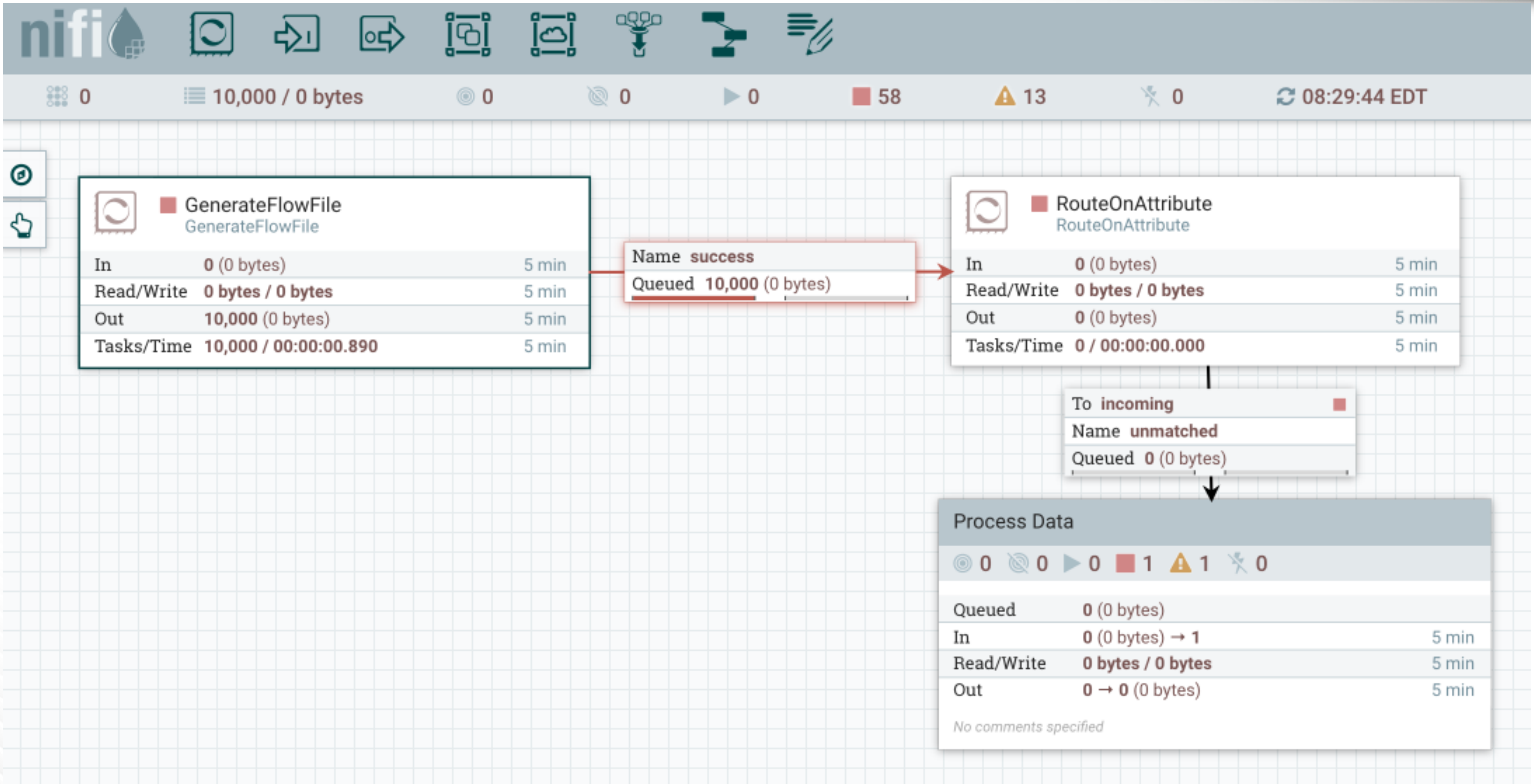
- Processor
  - Performs the work, can access FlowFiles

- Connection
  - Links between processors
  - Queues that can be dynamically prioritized

- Process Group
  - Set of processors and their connections
  - Receive data via input ports, send data via output ports

# 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. |



0 | 10,000 / 0 bytes | 0 | 0 | ▶ 0 | 🟥 58 | ⚠ 13 | 0 | 08:29:44 EDT

🟥 GeneratedFlowFile
GenerateFlowFile

| | | |
|---|---|---|
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 10,000 (0 bytes) | 5 min |
| Tasks/Time | 10,000 / 00:00:00.890 | 5 min |

Name success
Queued 10,000 (0 bytes)

🟥 RouteOnAttribute
RouteOnAttribute

| | | |
|---|---|---|
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

To incoming 🟥
Name unmatched
Queued 0 (0 bytes)

Process Data

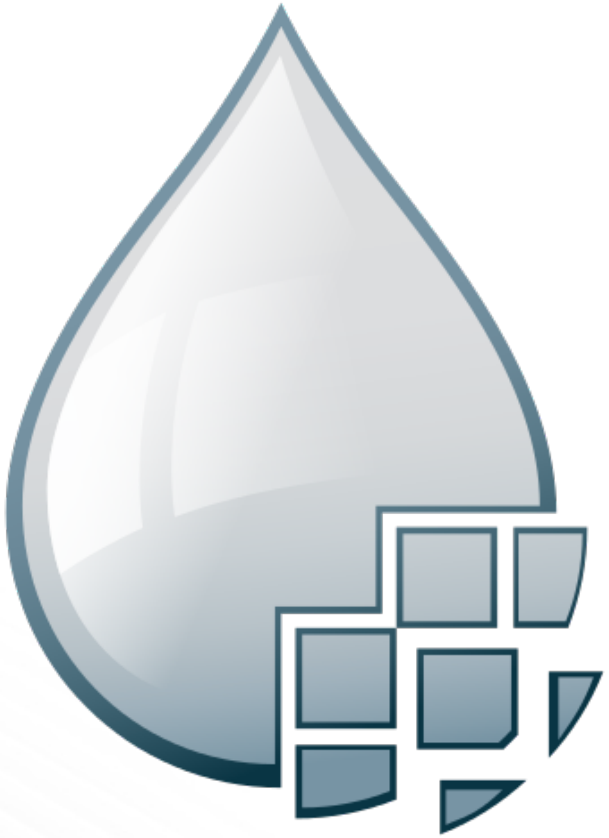
0 | 0 | ▶ 0 | 🟥 1 | ⚠ 1 | 0

| | | |
|---|---|---|
| Queued | 0 (0 bytes) | |
| In | 0 (0 bytes) → 1 | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 → 0 (0 bytes) | 5 min |

*No comments specified*

# Apache NiFi

## Key Features and Principles



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

# The need for data provenance

**For Operators**

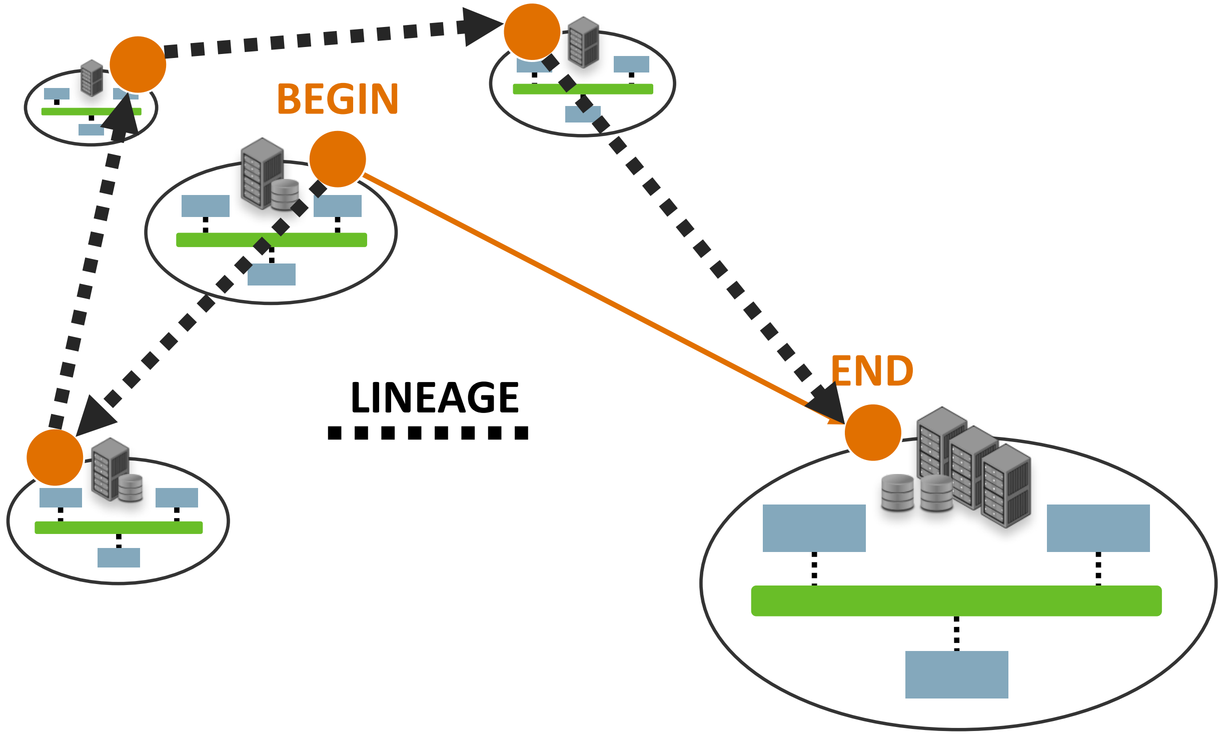
- Traceability, lineage
- Recovery and replay

**For Compliance**

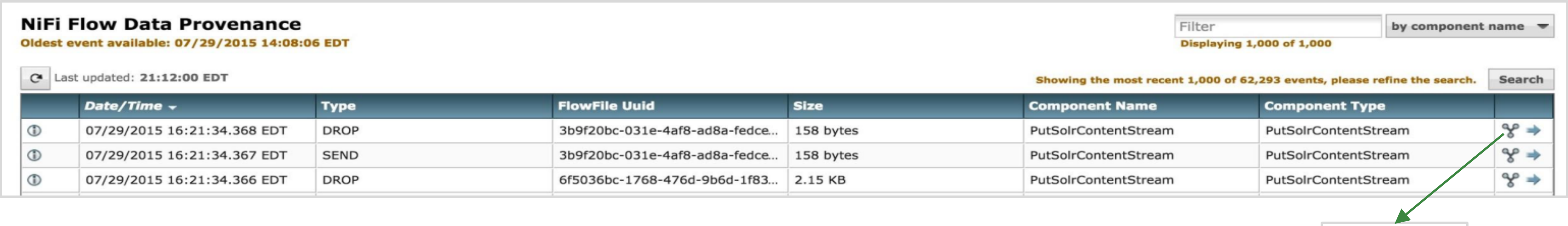
- Audit trail
- Remediation

**For Business / Mission**

- Value sources
- Value IT investment



# Data Provenance (Not just Lineage)



**NiFi Flow Data Provenance**

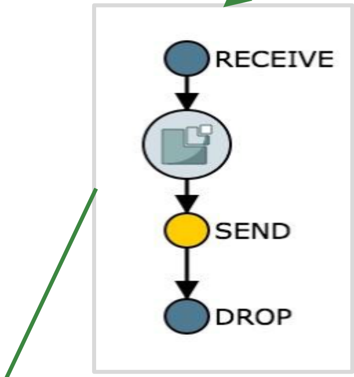
Oldest event available: 07/29/2015 14:08:06 EDT

Last updated: **21:12:00 EDT**

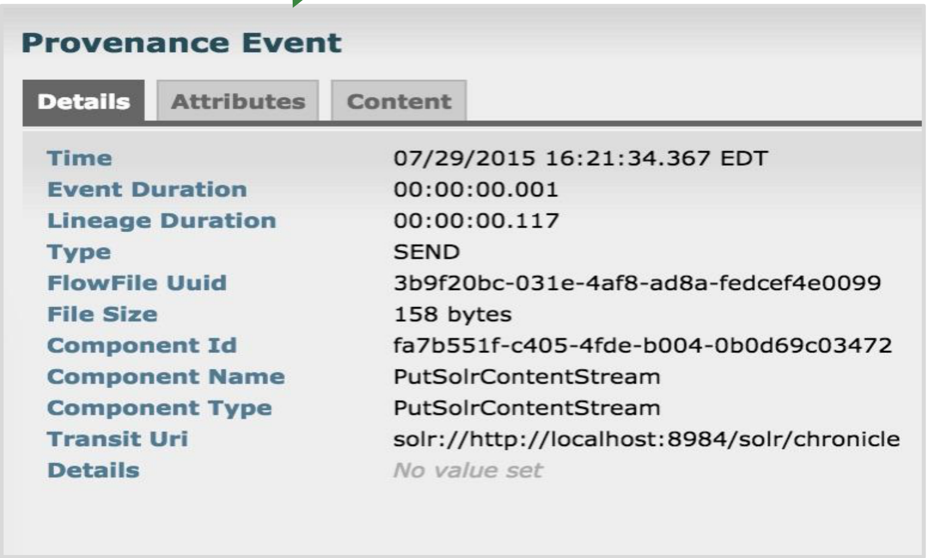
Displaying 1,000 of 1,000

Showing the most recent 1,000 of 62,293 events, please refine the search.

| | Date/Time | Type | FlowFile Uuid | Size | Component Name | Component Type | |
|---|---|---|---|---|---|---|---|
| ⓘ | 07/29/2015 16:21:34.368 EDT | DROP | 3b9f20bc-031e-4af8-ad8a-fedce... | 158 bytes | PutSolrContentStream | PutSolrContentStream | → |
| ⓘ | 07/29/2015 16:21:34.367 EDT | SEND | 3b9f20bc-031e-4af8-ad8a-fedce... | 158 bytes | PutSolrContentStream | PutSolrContentStream | → |
| ⓘ | 07/29/2015 16:21:34.366 EDT | DROP | 6f5036bc-1768-476d-9b6d-1f83... | 2.15 KB | PutSolrContentStream | PutSolrContentStream | → |



- View attributes and content at given points in time (before and after each processor) !!!
- Records, indexes, and makes events available for display



**Provenance Event**

| | |
|---|---|
| Time | 07/29/2015 16:21:34.367 EDT |
| Event Duration | 00:00:00.001 |
| Lineage Duration | 00:00:00.117 |
| Type | SEND |
| FlowFile Uuid | 3b9f20bc-031e-4af8-ad8a-fedcef4e0099 |
| File Size | 158 bytes |
| Component Id | fa7b551f-c405-4fde-b004-0b0d69c03472 |
| Component Name | PutSolrContentStream |
| Component Type | PutSolrContentStream |
| Transit Uri | solr://http://localhost:8984/solr/chronicle |
| Details | *No value set* |

# Provenance

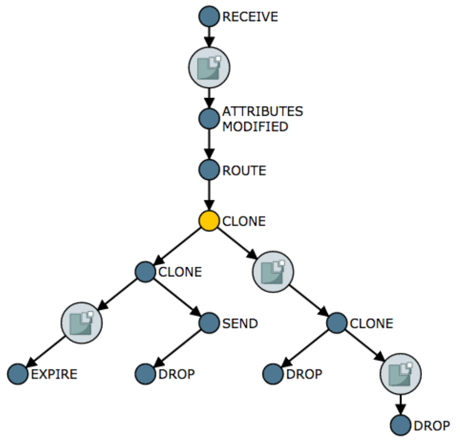
SOURCES

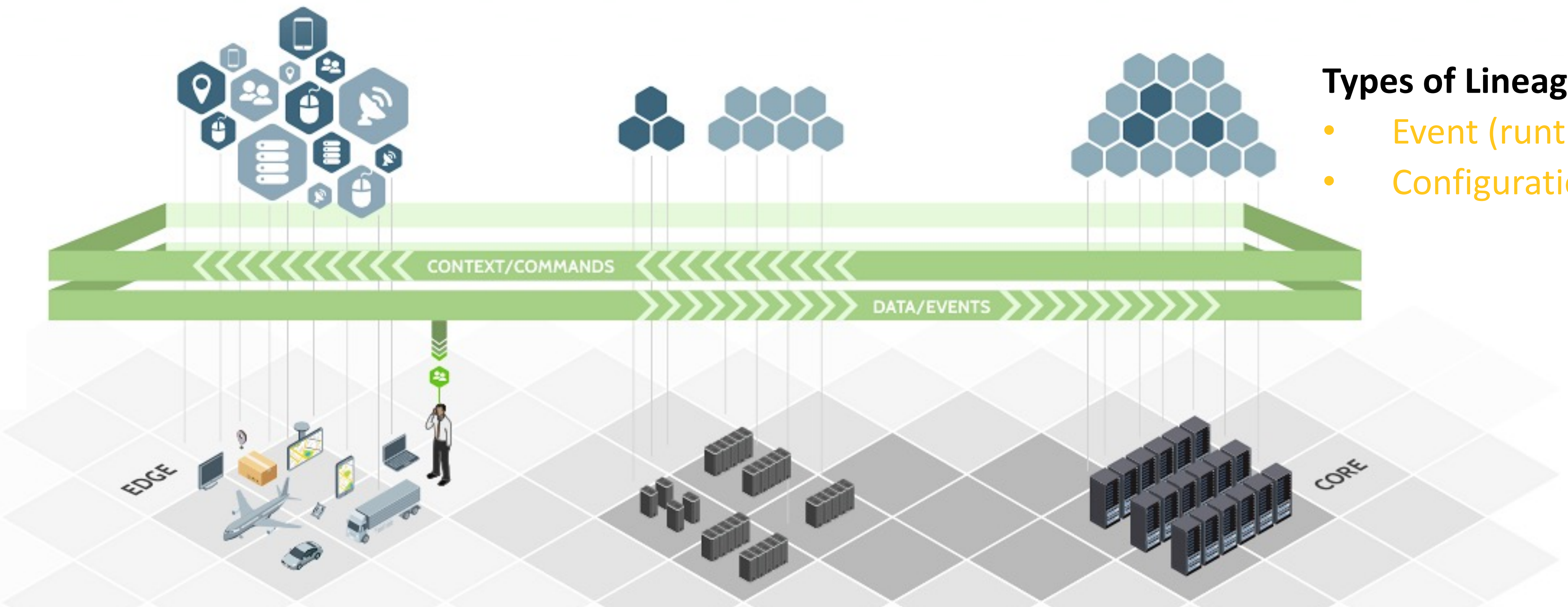
Origin – attribution
Replay – recovery

REGIONAL
INFRASTRUCTURE

CORE
INFRASTRUCTURE

Evolution of topologies
Long retention





**Types of Lineage**

- Event (runtime)
- Configuration (design time)

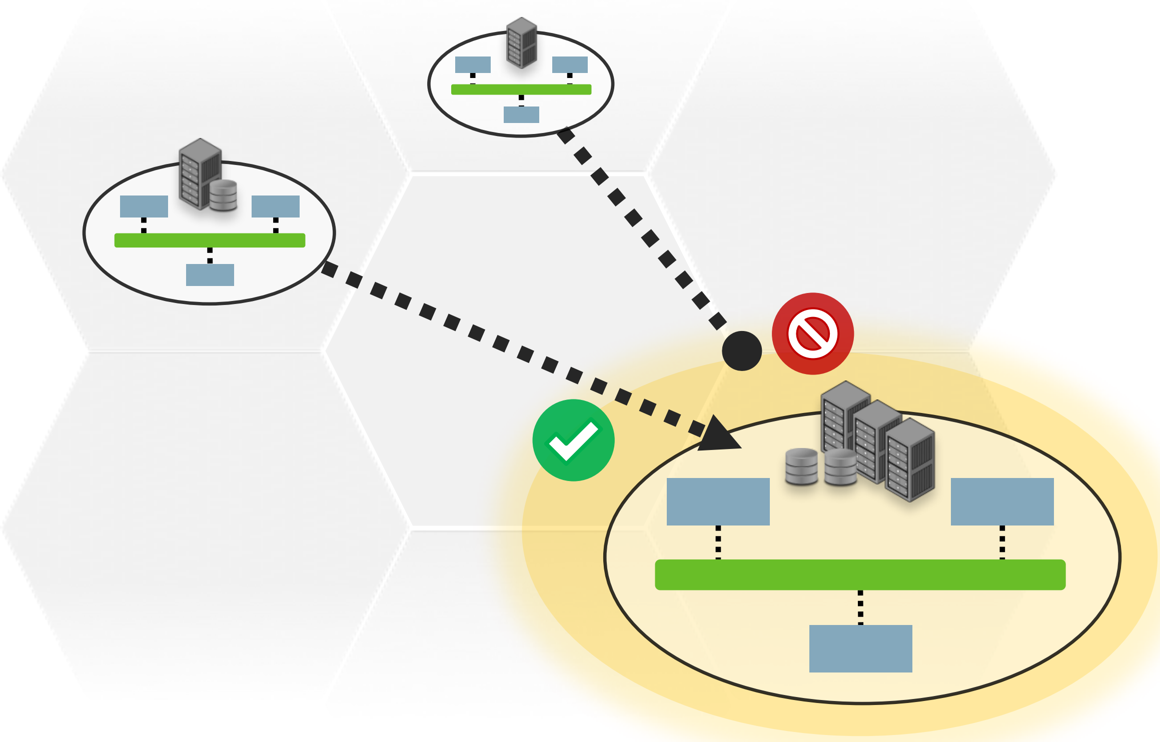
- Constrained
- High-latency
- Localized context

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

# The need for fine-grained security and compliance

**It's not enough to say you have encrypted communications**

- Enterprise authorization services –entitlements change often
- People and systems with different roles require difference access levels
- Tagged/classified data



# Security

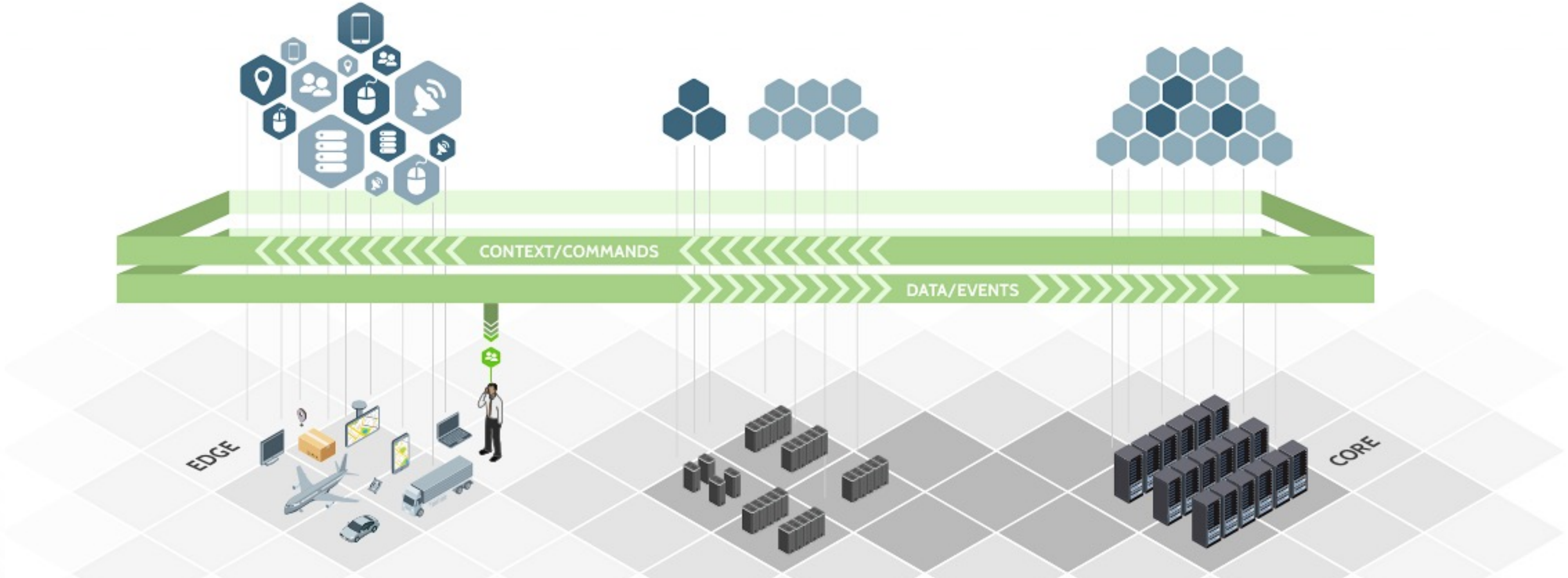
SOURCES

REGIONAL INFRASTRUCTURE

CORE INFRASTRUCTURE

Sign, encrypt, static
(data and control)

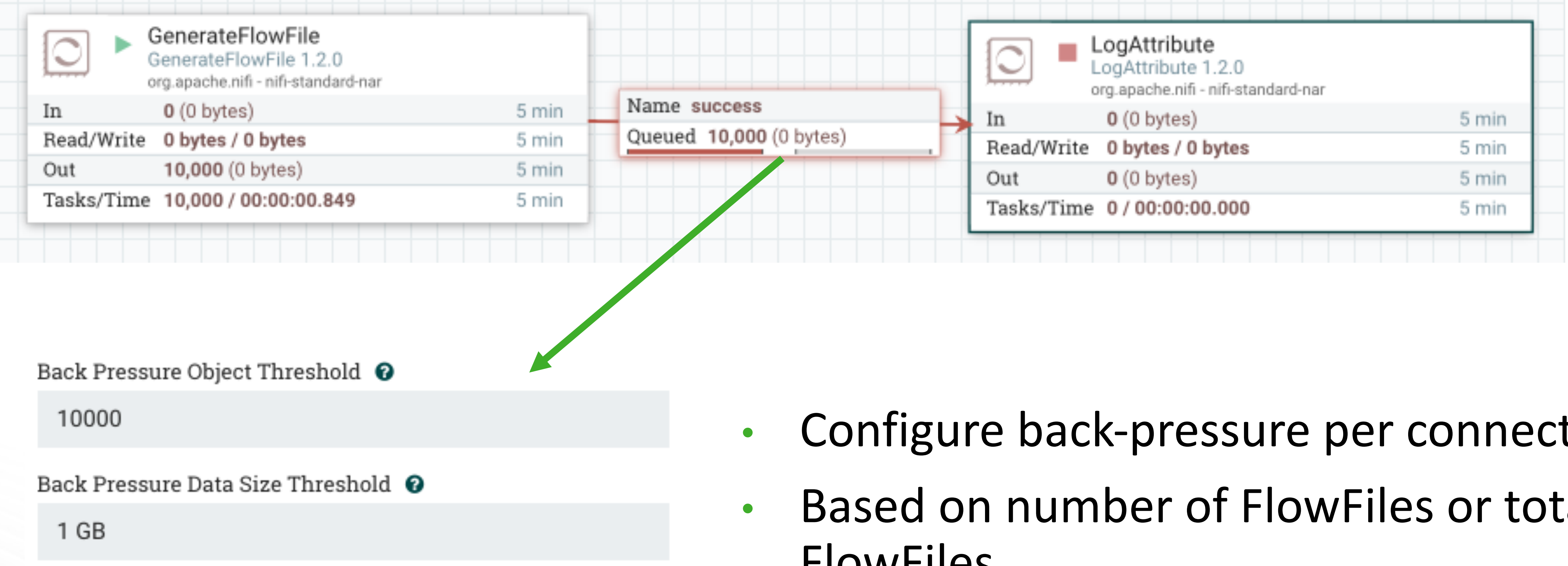
TLS, obfuscation, dynamic entitlements
Kerberos, PKI, AD/DS, etc.



- Constrained
- High-latency
- Localized context

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

# Back Pressure



GenerateFlowFile
GenerateFlowFile 1.2.0
org.apache.nifi - nifi-standard-nar

| | | |
|---|---|---|
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 10,000 (0 bytes) | 5 min |
| Tasks/Time | 10,000 / 00:00:00.849 | 5 min |

Name success
Queued 10,000 (0 bytes)

LogAttribute
LogAttribute 1.2.0
org.apache.nifi - nifi-standard-nar

| | | |
|---|---|---|
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

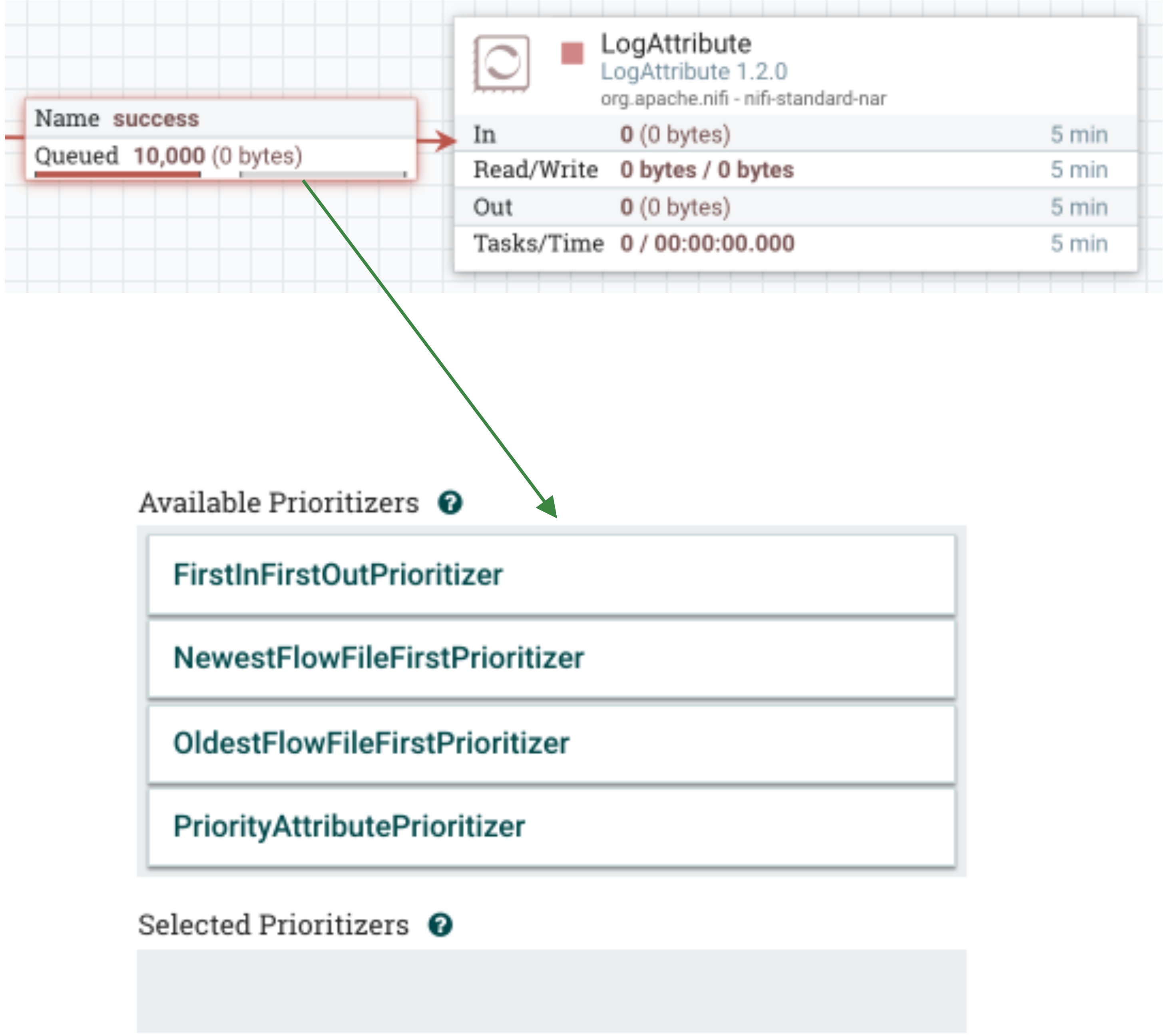
Back Pressure Object Threshold ?
10000

Back Pressure Data Size Threshold ?
1 GB

- Configure back-pressure per connection
- Based on number of FlowFiles or total size of FlowFiles
- Upstream processor no longer scheduled to run until below threshold

# Prioritization

- Configure a prioritizer per connection
- Determine what is important for your data – time based, arrival order, importance of a data set
- Funnel many connections down to a single connection to prioritize across data sets
- Develop your own prioritizer if needed



Name success

Queued 10,000 (0 bytes)

LogAttribute

LogAttribute 1.2.0

org.apache.nifi - nifi-standard-nar

| | | |
|---|---|---|
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

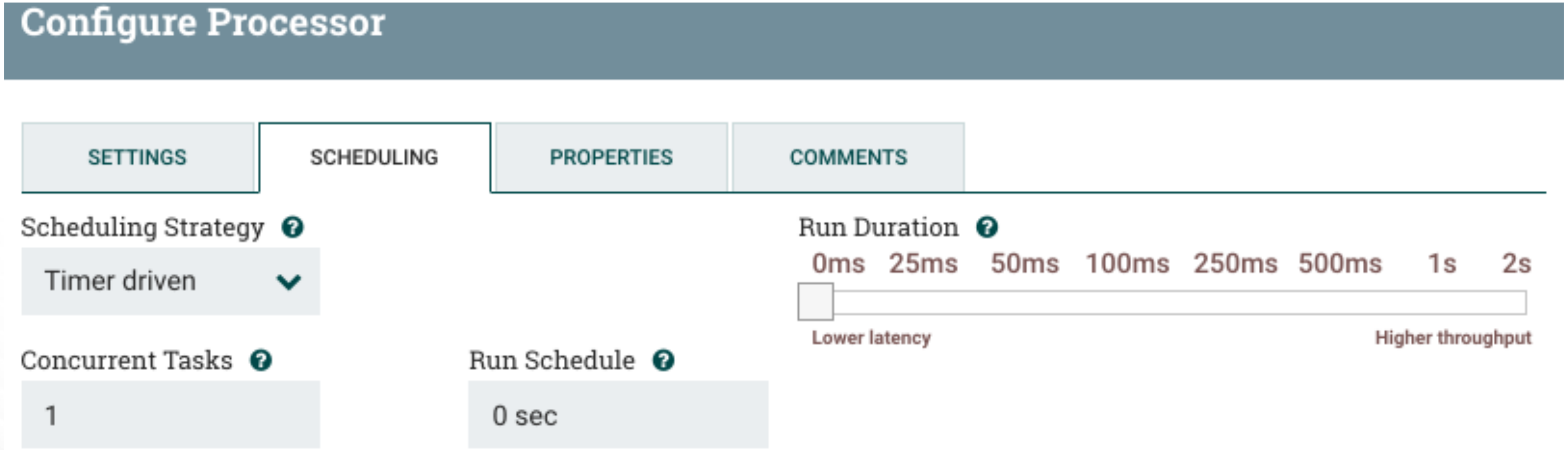
Available Prioritizers

- FirstInFirstOutPrioritizer
- NewestFlowFileFirstPrioritizer
- OldestFlowFileFirstPrioritizer
- PriorityAttributePrioritizer

Selected Prioritizers

# Latency vs. Throughput

- Choose between lower latency, or higher throughput on each processor
- Higher throughput allows framework to batch together all operations for the selected amount of time for improved performance
- Processor developer determines whether to support this by using @SupportsBatching annotation



Configure Processor

Scheduling Strategy: Timer driven

Concurrent Tasks: 1

Run Schedule: 0 sec

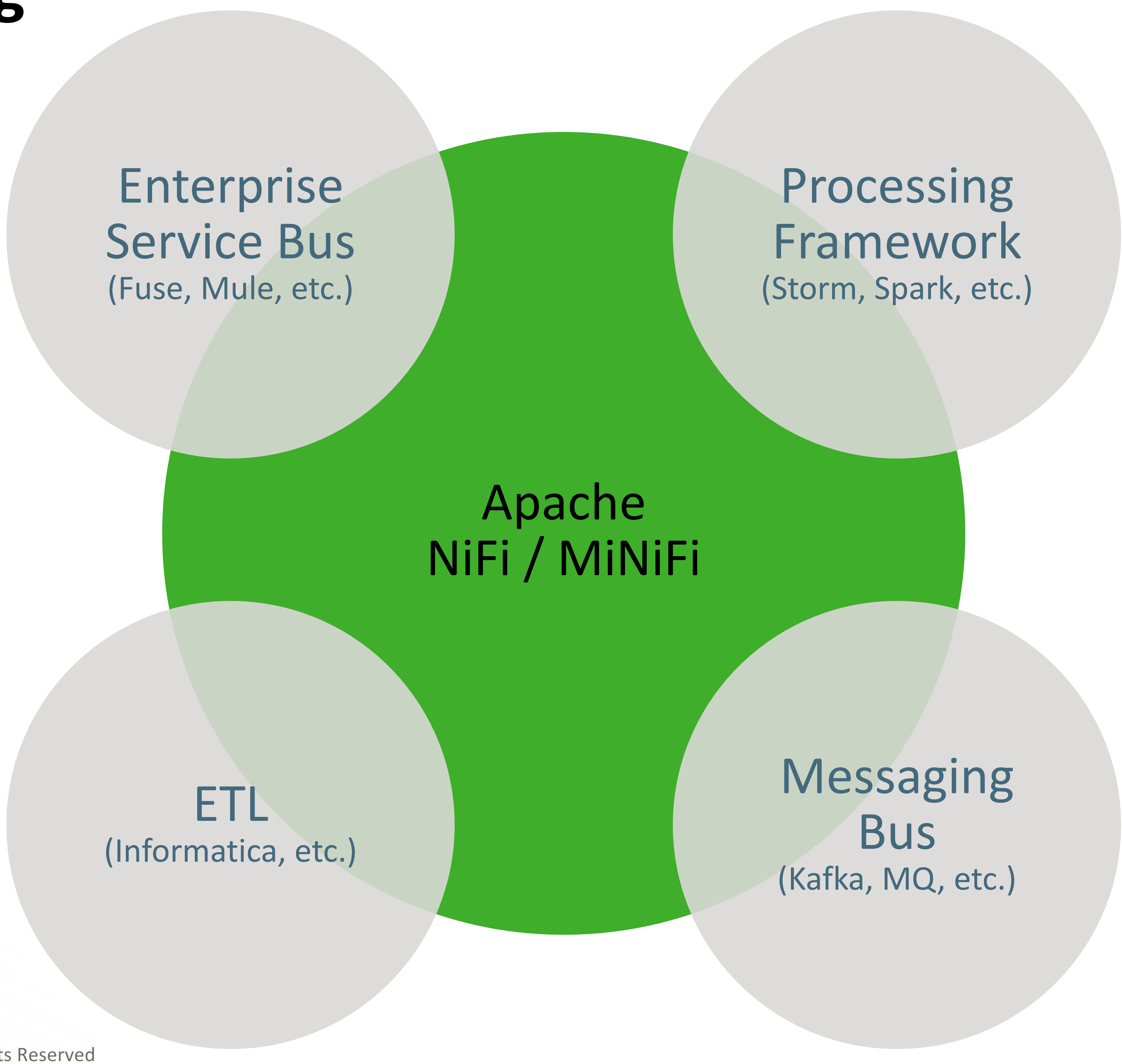
Run Duration: 0ms 25ms 50ms 100ms 250ms 500ms 1s 2s

# 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.. |

# NiFi Positioning

# NiFi Positioning



Enterprise Service Bus
(Fuse, Mule, etc.)

Processing Framework
(Storm, Spark, etc.)

Apache NiFi / MiNiFi

ETL
(Informatica, etc.)

Messaging Bus
(Kafka, MQ, etc.)

# Apache NiFi / Processing Frameworks

## NiFi

### Simple event processing

- **Primarily feed data into processing frameworks, can process data, with a focus on simple event processing**
- **Operate on a single piece of data, or in correlation with an enrichment dataset (enrichment, parsing, splitting, and transformations)**
- **Can scale out, but scale up better to take full advantage of hardware resources, run concurrent processing tasks/threads (processing terabytes of data per day on a single node)**

⚠ Not another distributed processing framework, but to feed data into those

## Processing Frameworks (Storm, Spark, etc.)

### Complex and distributed processing

- **Complex processing from multiple streams (JOIN operations)**
- **Analyzing data across time windows (rolling window aggregation, standard deviation, etc.)**
- **Scale out to thousands of nodes if needed**

⚠ Not designed to collect data or manage data flow

# Apache NiFi / Messaging Bus Services

## NiFi

### Provide dataflow solution

- Centralized management, from edge to core
- Great traceability, event level data provenance starting when data is born
- Interactive command and control – real time operational visibility
- Dataflow management, including prioritization, back pressure, and edge intelligence
- Visual representation of global dataflow

⚠ Not a messaging bus, flow maintenance needed when you have frequent consumer side updates

## Messaging Bus (Kafka, JMS, etc.)

### Provide messaging bus service

- Low latency
- Great data durability
- Decentralized management (producers & consumers)
- Low broker maintenance for dynamic consumer side updates

⚠ Not designed to solve dataflow problems (prioritization, edge intelligence, etc.)

⚠ Traceability limited to in/out of topics, no lineage

⚠ Lack of global view of components/connectivities

# Apache NiFi / Integration, or ingestion, Frameworks

## NiFi

### End user facing dataflow management tool

- Out of the box solution for dataflow management
- Interactive command and control in the core, design and deploy on the edge
- Flexible failure handling at each point of the flow
- Visual representation of global dataflow and connectivities
- Native cross data center communication
- Data provenance for traceability

⚠ Not a library to be embedded in other applications

## Integration framework (Spring Integration, Camel, etc), ingestion framework (Flume, etc)

### Developer facing integration tool with a focus on data ingestion

- A set of tools to orchestrate workflow
- A fixed design and deploy pattern
- Leverage messaging bus across disconnected networks

⚠ Developer facing, custom coding needed to optimize

⚠ Pre-built failure handling, lack of flexibility

⚠ No holistic view of global dataflow

⚠ No built-in data traceability

# Apache NiFi / ETL Tools

## NiFi

### NOT schema dependent

- Dataflow management for both structured and unstructured data, powered by separation of metadata and payload
- Schema is not required, but you can have schema
- Minimum modeling effort, just enough to manage dataflows
- Do the plumbing job, maximize developers' brainpower for creative work

⚠ Not designed to do heavy lifting transformation work for DB tables (JOIN datasets, etc.). You can create custom processors to do that, but long way to go to catch up with existing ETL tools from user experience perspective (GUI for data wrangling, cleansing, etc.)

## ETL (Informatica, etc.)

### Schema dependent

- Tailored for Databases/WH
- ETL operations based on schema/data modeling
- Highly efficient, optimized performance

⚠ Must pre-prepare your data, time consuming to build data modeling, and maintain schemas

⚠ Not geared towards handling unstructured data, PDF, Audio, Video, etc.

⚠ Not designed to solve dataflow problems

# Apache MiNiFi

# Apache MiNiFi

**“Let me get the key parts of NiFi close to where data begins and provide bidirectional data transfer"**

- 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

# Realities of computing outside the comforts of the data center

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

# Apache ~~NiFi~~ MiNiFi

## Key Features

- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Recovery/recording a rolling log of fine-grained history
- Designed for extension

- Design and Deploy
- Warm re-deploys

# MiNiFi: Precedent from NiFi

## A quick look at NiFi Site to Site

- Provides the semantics between two NiFi components across network boundaries
  - A custom protocol for inter-NiFi communication
  - Secure, Extensible, Load Balanced & Scalable Delivery to Cluster
- Extracted out to a client library which powers integration into popular frameworks like Apache Spark, Apache Storm, Apache Flink, and Apache Apex
- Attributes and the FlowFile format maintained

*https://nifi.apache.org/docs/nifi-docs/html/user-guide.html#site-to-site*

# MiNiFi: Precedent from NiFi

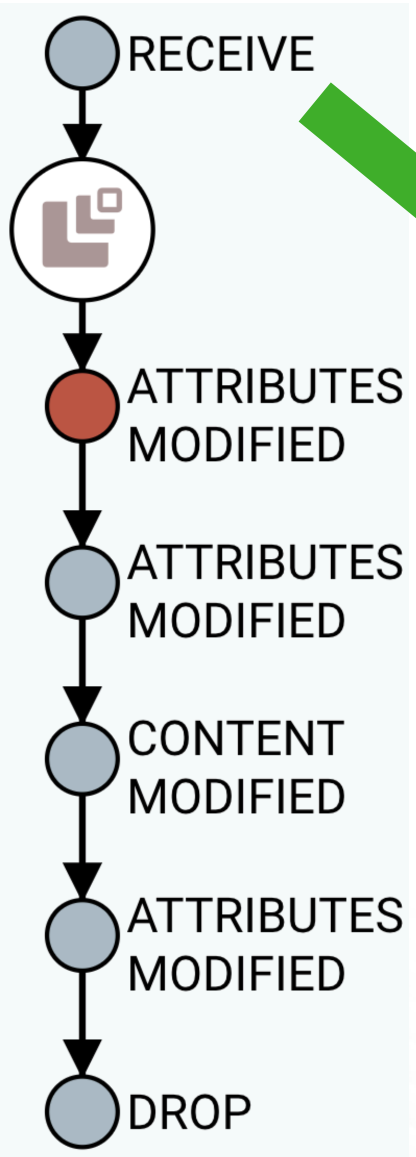
## A deeper dive into provenance

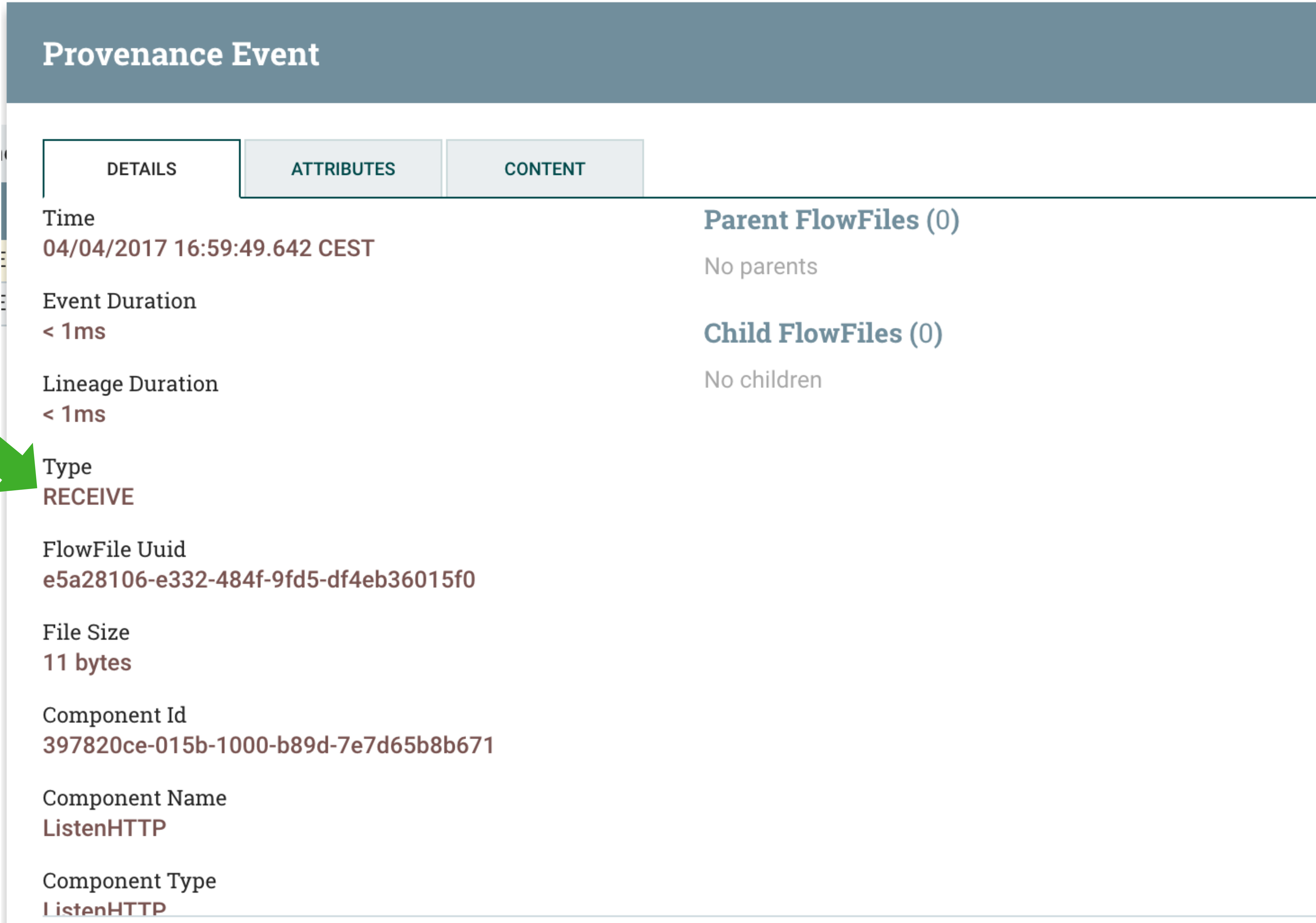
- Fine-grained, event level access of interactions with FlowFiles
  - CREATE, RECEIVE, FETCH, SEND, DOWNLOAD, DROP, EXPIRE, FORK, JOIN …
- Captures the associated attributes/metadata at the time of the event
- A map of a FlowFile’s journey and how they relate to other FlowFiles in a system
  - MiNiFi enables us to get more and further illuminate the map of data processing

http://nifi.apache.org/docs/nifi-docs/html/user-guide.html#data-provenance

# MiNiFi: Precedent from NiFi

## RECEIVE event





### Provenance Event

Time
04/04/2017 16:59:49.642 CEST

Event Duration
< 1ms

Lineage Duration
< 1ms

Type
RECEIVE

FlowFile Uuid
e5a28106-e332-484f-9fd5-df4eb36015f0

File Size
11 bytes

Component Id
397820ce-015b-1000-b89d-7e7d65b8b671

Component Name
ListenHTTP

Component Type
ListenHTTP

Parent FlowFiles (0)

No parents

Child FlowFiles (0)

No children

# Apache MiNiFi

## Departures from NiFi in getting the right fit

- The feedback loop is longer and not guaranteed
  - Removal of Web Server and UI

- Declarative configuration
  - Lends itself well to CM processes
  - Extensible interface to support varying formats
    - Currently provided in YAML

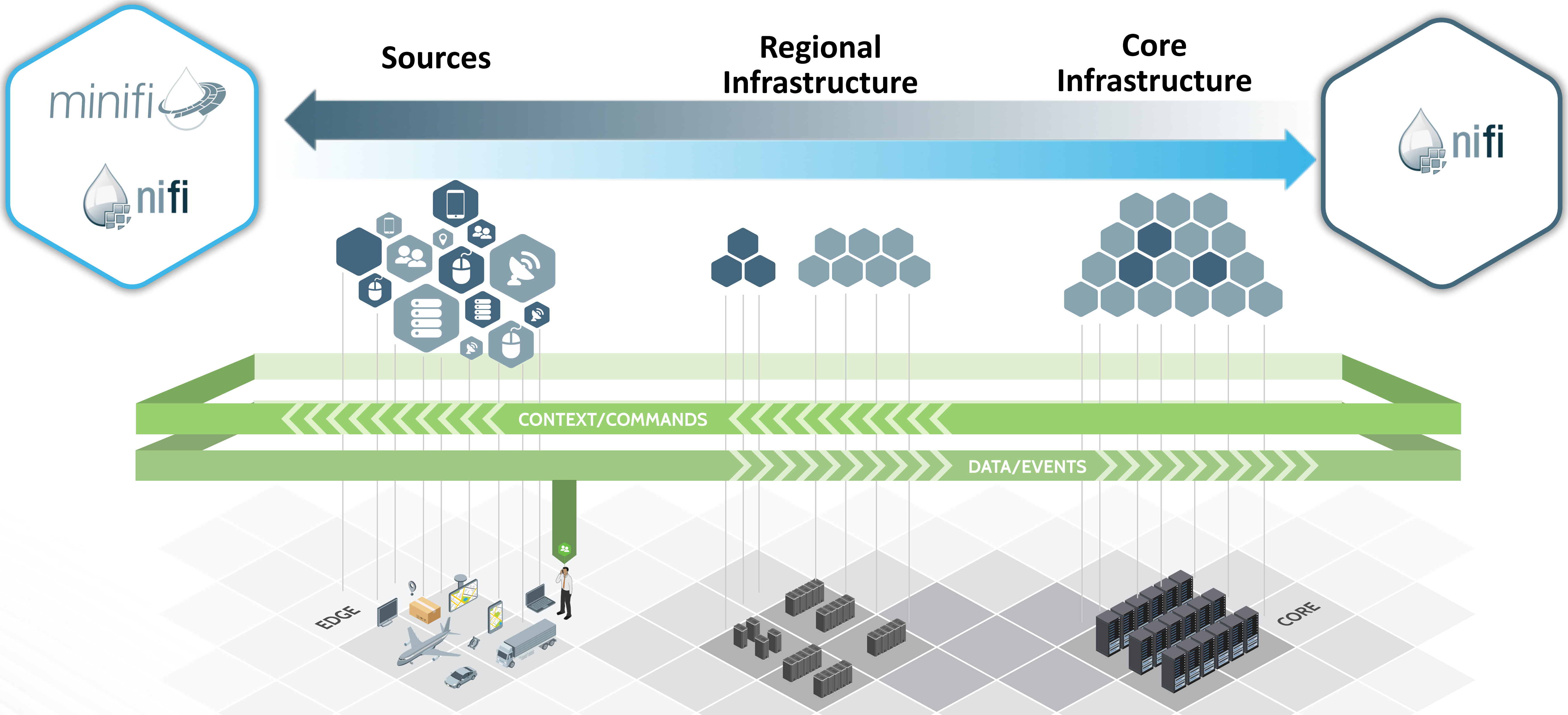
- Reduced set of bundled components

# Apache MiNiFi:  Scoping

## Provide all the key principles of NiFi in varying, smaller footprints

- ***Go small***:  Java – *Write once, run anywhere**
  - Feature parity and reuse of core NiFi libraries
- ***Go smaller***:  C++ – *Write once**, run anywhere*
- ***Go smallest***: *Write n-many times, run anywhere*
  Language libraries to support tagging, FlowFile format, Site to Site protocol, and provenance generation without a processing framework
  - Mobile:  Android & iOS
  - Language SDKs

# Harnessing Data in Motion



- Constrained
- High-latency
- Localized context

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

# Demo

# 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
https://issues.apache.org/jira/browse/MINIFI

**Follow on Twitter**
@apachenifi

# Thank You

HORTONWORKS®
POWERING THE FUTURE OF DATA™