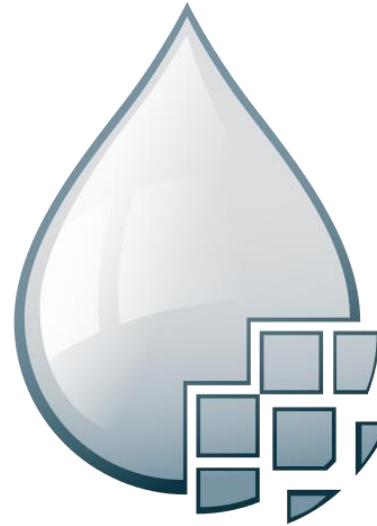


Apache NiFi



Better Analytics Demand Better Dataflow

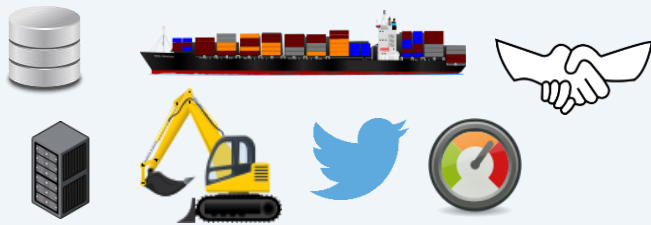
Presented by: Joe Witt

Apache NiFi PPMC Member



Apache NiFi's job: Enterprise Dataflow Management

Automate the flow of data from any source



...to systems which extract meaning and insight



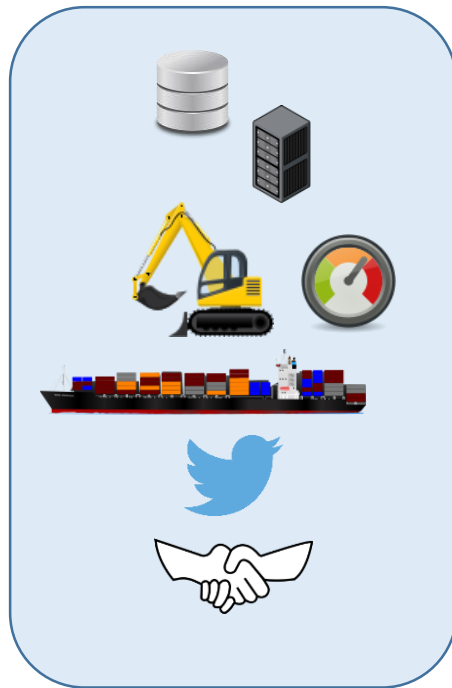
...and to those that store and make it available for users

Analytics need data with the following characteristics:

Quality	Correct, complete, reliable
Relevance	Right size, rate, format, schema, content, lightweight analysis
Timeliness	All data has a half-life. Not all data is created equal.
Secure	Confidential, unaltered
Compliant	Authorized, traceable
Recoverable	Errors happen. Iterate until it's right.

Enterprise Dataflow: “What could possibly go wrong?”

Acquire



Analyze



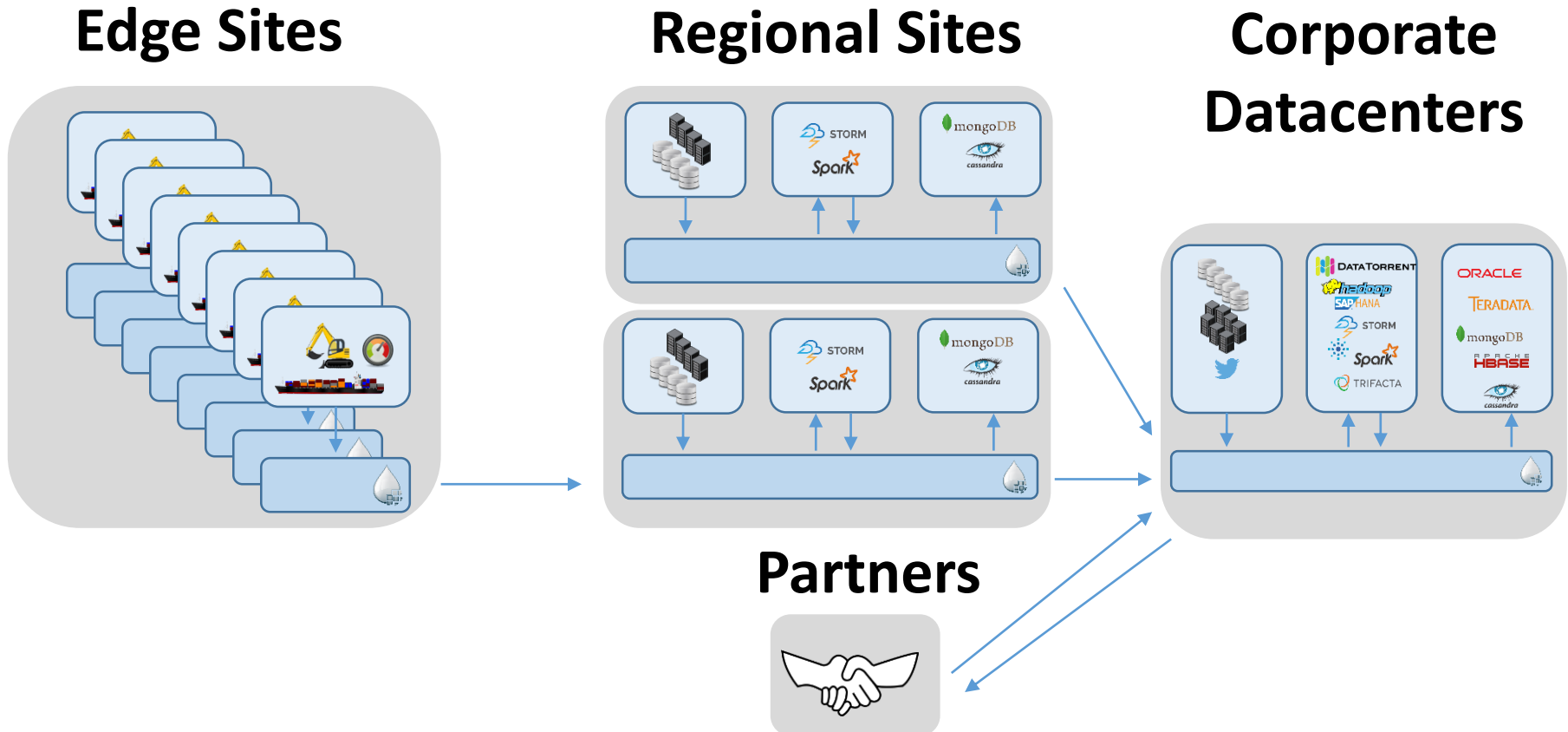
Store



Dataflow – Route, Transform, Mediate

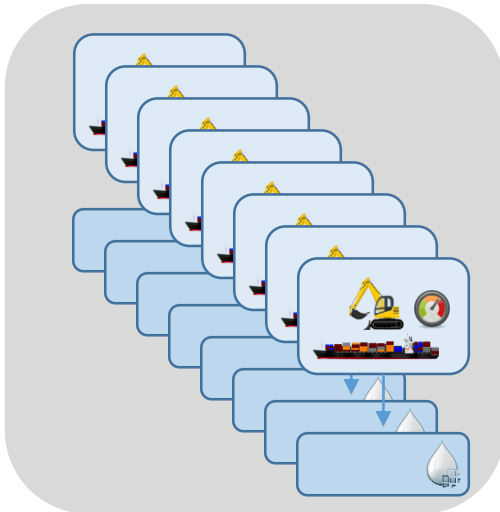


Dataflow across the enterprise



Challenges at the edge

Edge Sites



- **Devices may**
 - **Have low power**
 - **Use legacy protocols and formats**
 - **Use emerging protocols and formats**
- **Communications may be**
 - **Unstable**
 - **High latency / Low Throughput**
 - **Expensive**
- **Data acquired may be**
 - **Erroneous**
 - **Devoid of value or ‘noisy’**
 - **Time sensitive or tolerant**
 - **Of differing priority**
 - **Sensitive**

Challenges at the core

Data may need transformation

- Enrichment
- Format/schema conversion
- Splitting or Aggregation

Systems may be

- Down, degraded, returning to service
- Rate or throughput sensitive
- Authorized for a subset of data

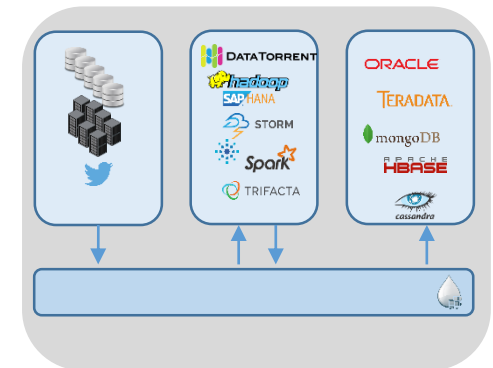
Scaling and reliability

- Controlled data loss only
- Up (node efficient) & Out (global volume)

Governance

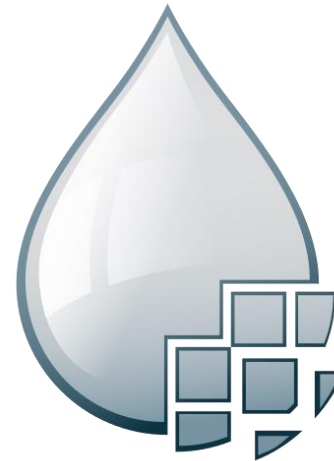
- Keeping track of all the information flows
- Ability to understand and manage the flows
- Ability to detect and recover from mistakes

Corporate Datacenters



Apache NiFi Foundational Concepts

- 1 The basic building blocks
- 2 Real-time Command and Control
- 3 The Power of Provenance



Flow File

- UUID
- Name
- Size
- Entry Time


HEADER


Attributes Map
[[Key | Value]]

CONTENT

- **Types**
 - Events
 - Objects
 - Files
 - Messages
 - Media
- **Formats**
 - JSON
 - Avro
 - Text
 - Mp4
 - Proprietary
- **Sizes**
 - Bytes to GBs

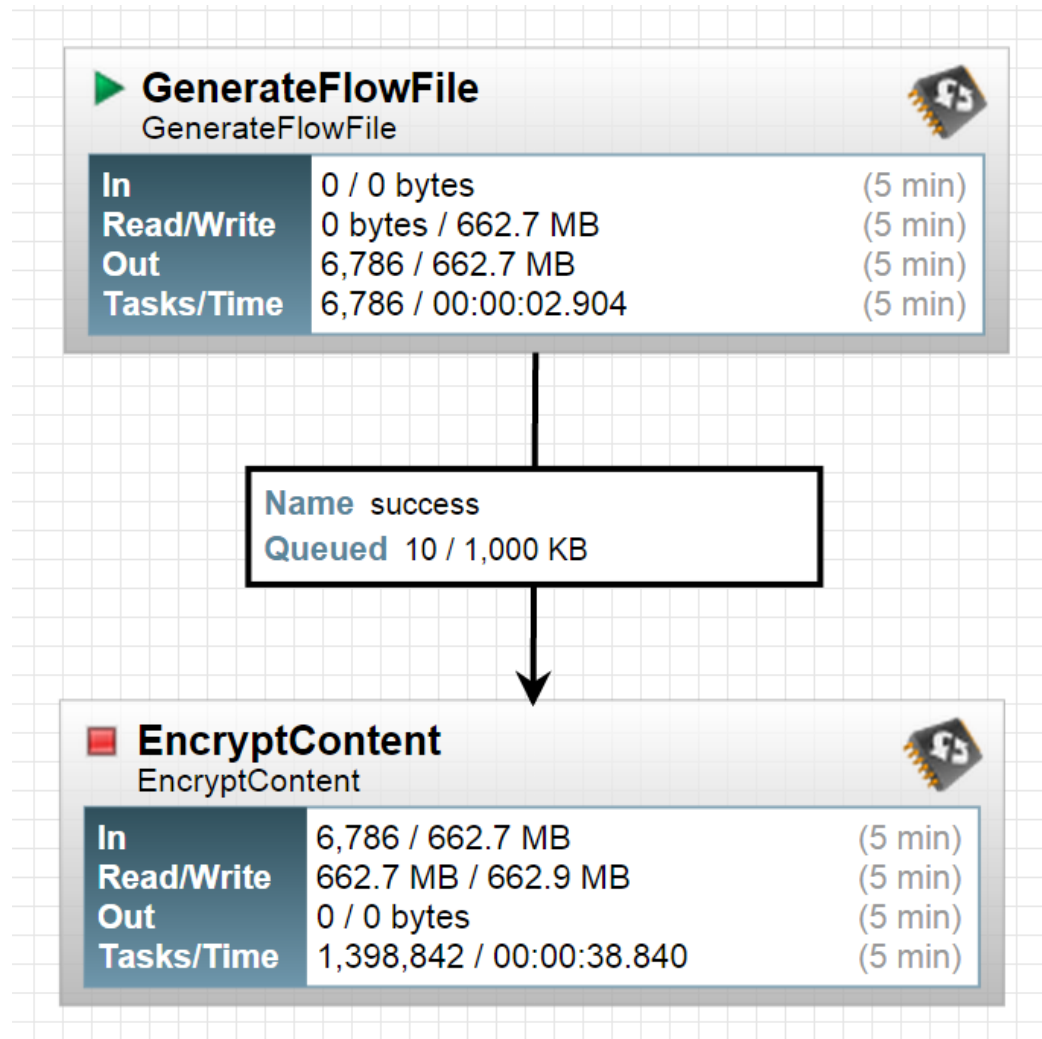
Flow File Processor

**EncryptContent**
EncryptContent

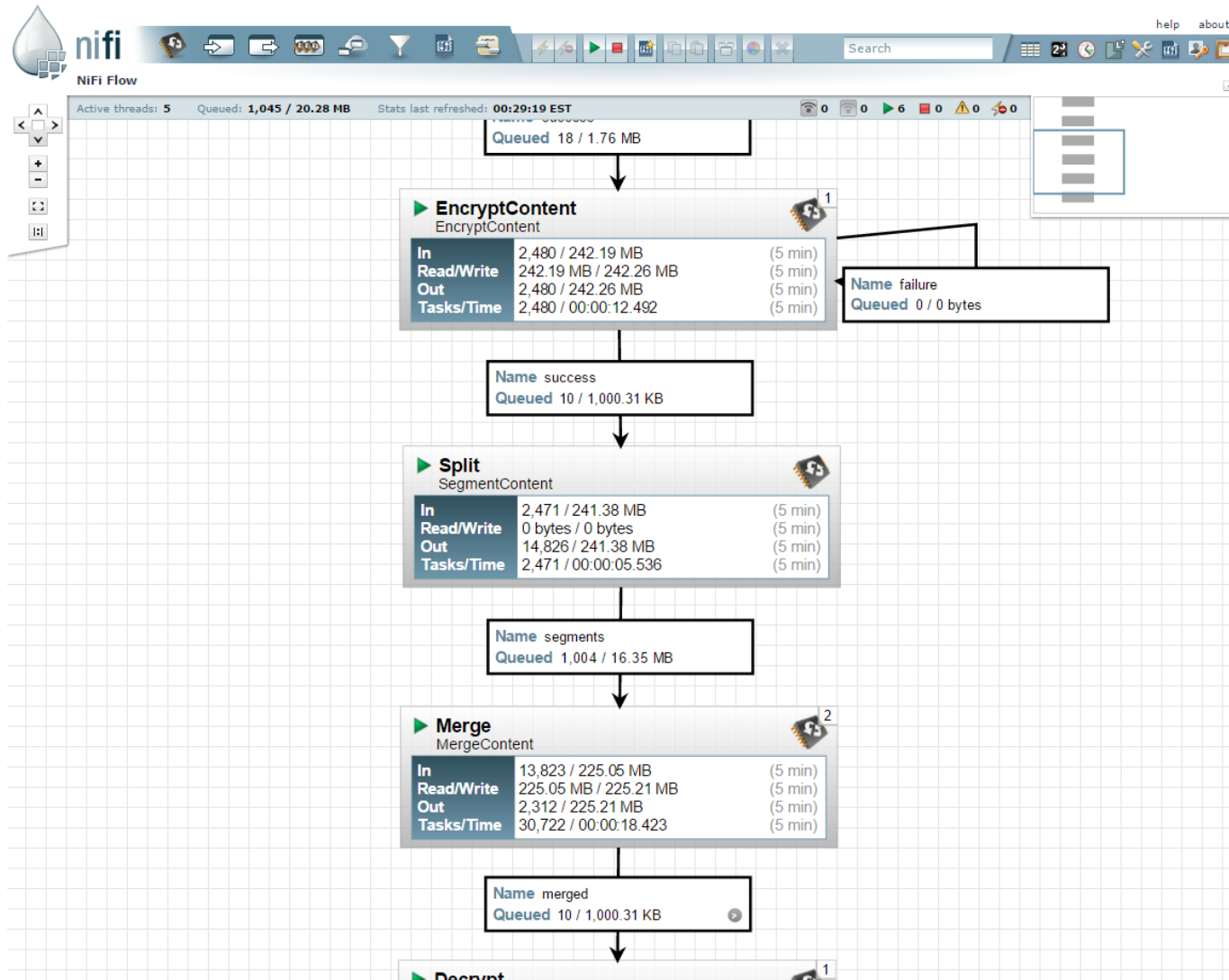
**1**

In	13,157 / 1.25 GB	(5 min)
Read/Write	1.25 GB / 1.26 GB	(5 min)
Out	0 / 0 bytes	(5 min)
Tasks/Time	15,613 / 00:00:36.886	(5 min)

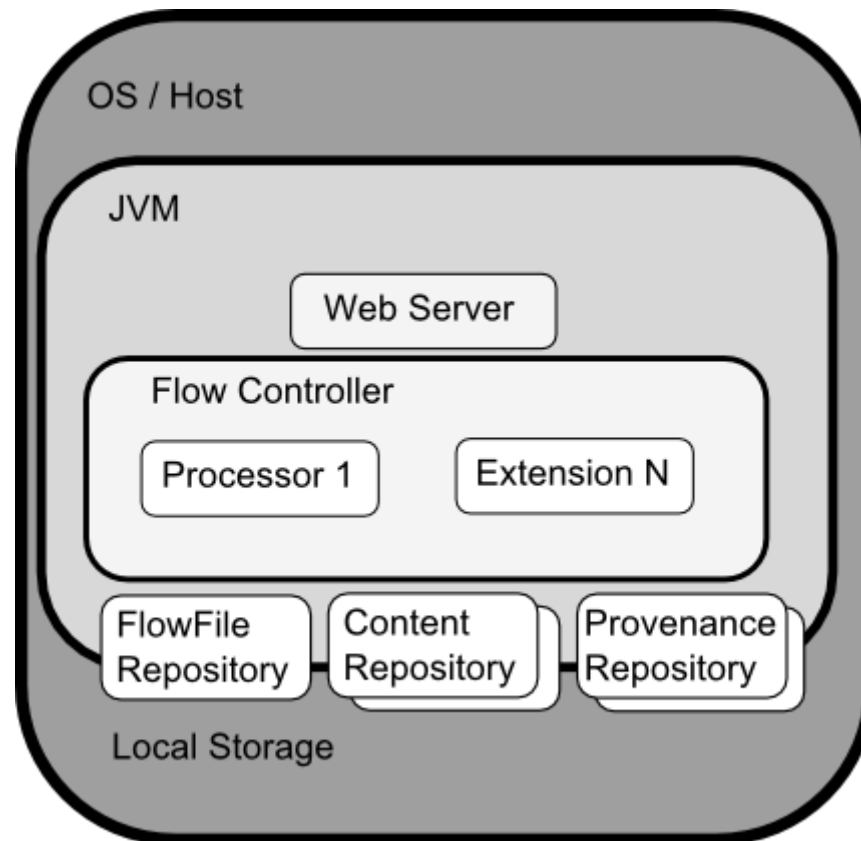
Connections



Flow Controller

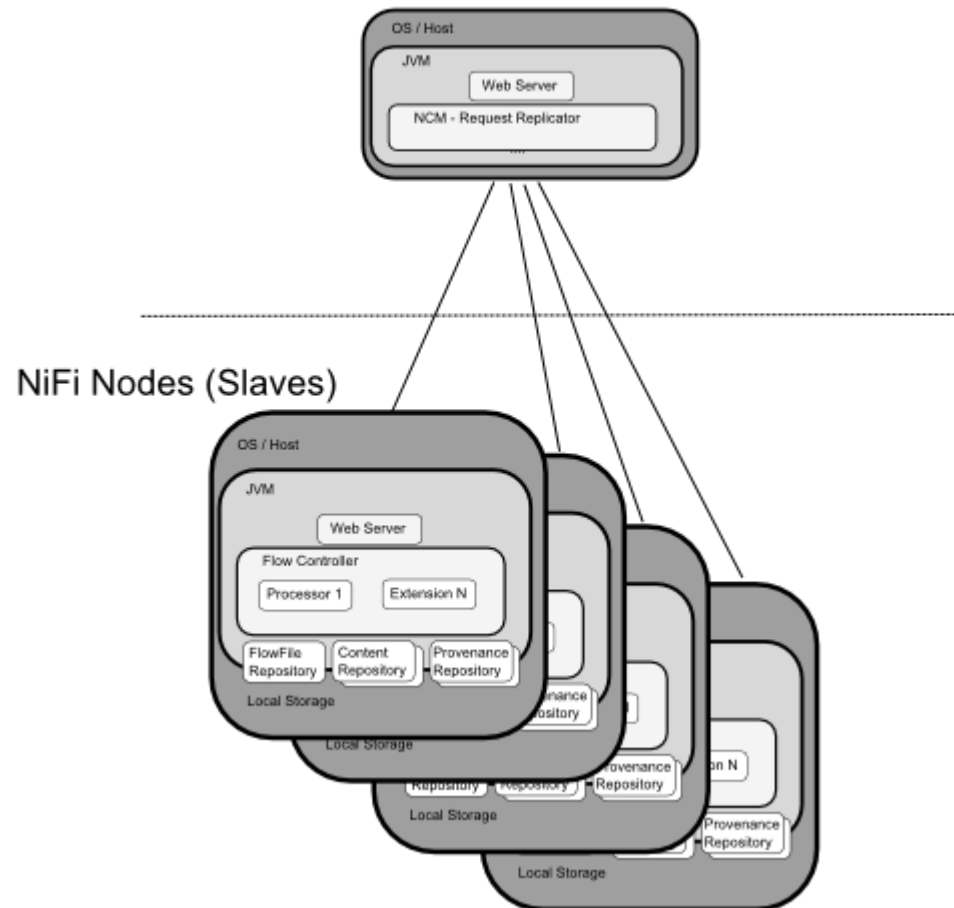


NiFi Architecture



NiFi Clustering Model

NiFi Cluster Manager (Master)



2 Real-time command and control

Tighten the feedback loop

- **Changes have consequences (good or bad)**
- **And you see them as they occur**

Continuous Improvement

- **Compare real-time vs. historical statistics**
- **View data provenance**
- **View Content at any stage**

Intuitive user experience

- **Visual programming**
- **Logical flow graph**

3

The Power of Provenance aka “Dude, where’s my data?”

Latency Optimization

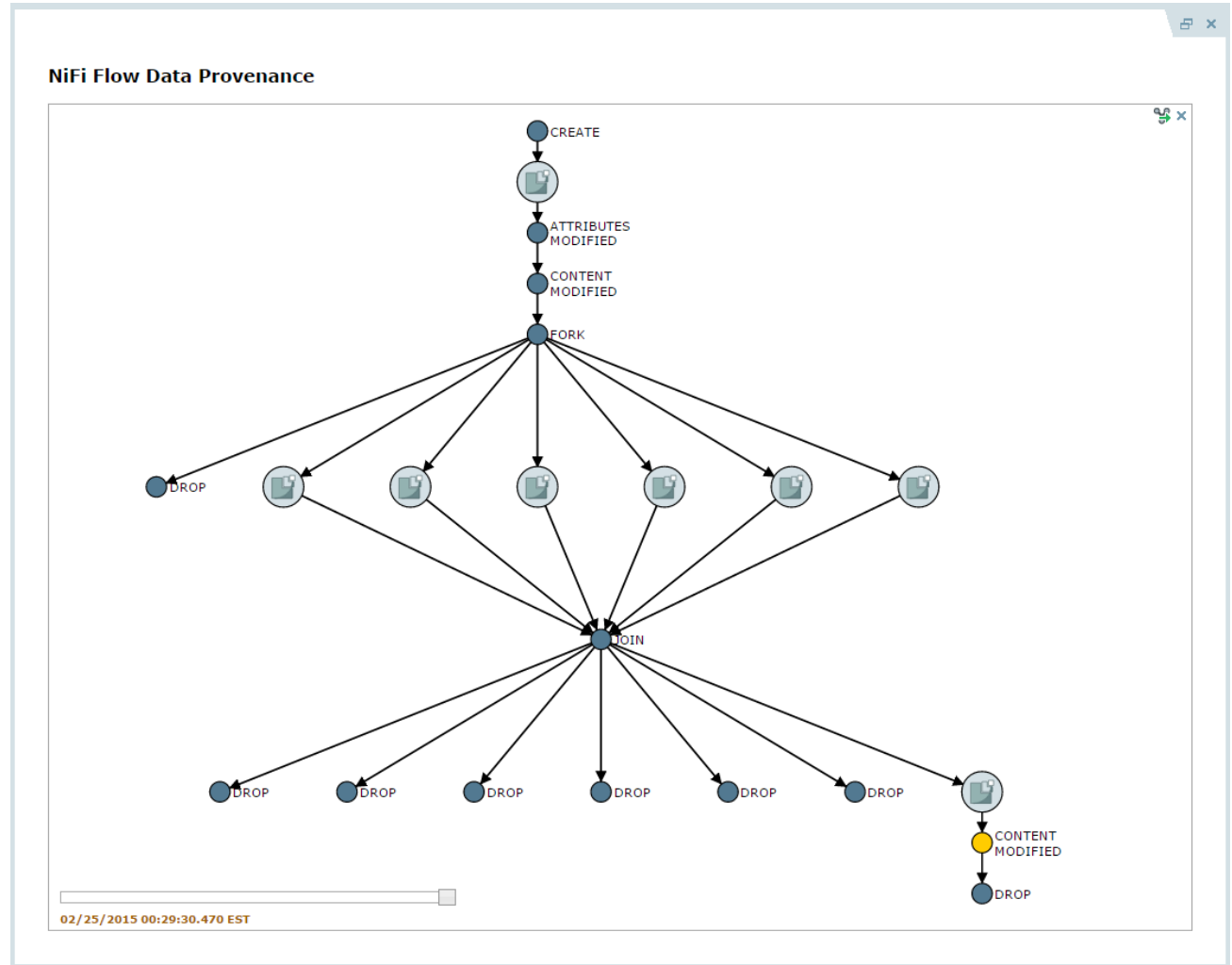
- Intra process
- Inter process
- End-to-end

Compliance

- Prove handling
- Assess impact

Understanding

- Step through time
- View content
- View Context



Status and direction for NiFi

Existing Strengths

Efficient use of each node

- 100s of MB/s per node
- 100Ks transactions/s per node

Simple / Effective scaling model

Runtime Command and Control

Data Provenance

Roadmap Highlights

Distributed durability of data

- Maybe Kafka backed queues

High Availability Cluster Manager

Live / Rolling Upgrades

Provenance Query Language / Reporting

A complete user experience enabled by provenance

Learn more about Apache NiFi

Apache NiFi (incubating) site

<http://nifi.incubator.apache.org>

Subscribe to and collaborate at

dev@nifi.incubator.apache.org

Submit Ideas or Issues

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

@ApacheNifi

