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

Producers A.K.A Things

Anything
AND
Everything

Consumers
• User
• Storage
• System
• ...More Things

Internet!
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 $\Delta s$ to Cs to $\varphi s$

Let’s consider the needs of a courier service

Physical Store

- Mobile Devices
- Registers

Gateway Server

On Delivery Routes

- Trucks
- Deliverers

Distribution Center

Server Cluster

Core Data Center at HQ

Server Cluster

Delivery Truck: Creative Stall, https://thenounproject.com/creativestall/
Deliverer: Rigo Peter, https://thenounproject.com/rigo/
Cash Register: Sergey Patutin, https://thenounproject.com/bdesign.by/
Great! I am collecting all this data! Let’s use it!

Finding our needles in the haystack

Physical Store
- Mobile Devices
- Registers

On Delivery Routes
- Trucks
- Deliverers

Distribution Center
- Gateway Server
- Server Cluster
- Kafka
- Storm / Spark / Flink / Apex

Core Data Center at HQ
- Server Cluster
- Kafka
- Storm / Spark / Flink / Apex
- Others

Delivery Truck: Creative Stall, https://thenounproject.com/creativestall/
Deliverer: Rigo Peter, https://thenounproject.com/rigo/
Cash Register: Sergey Patutin, https://thenounproject.com/bdesign.by/
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
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)

<table>
<thead>
<tr>
<th>FBP Term</th>
<th>NiFi Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Packet</td>
<td>FlowFile</td>
<td>Each object moving through the system.</td>
</tr>
<tr>
<td>Black Box</td>
<td>FlowFile Processor</td>
<td>Performs the work, doing some combination of data routing, transformation, or mediation between systems.</td>
</tr>
<tr>
<td>Bounded Buffer</td>
<td>Connection</td>
<td>The linkage between processors, acting as queues and allowing various processes to interact at differing rates.</td>
</tr>
<tr>
<td>Scheduler</td>
<td>Flow Controller</td>
<td>Maintains the knowledge of how processes are connected, and manages the threads and allocations thereof which all processes use.</td>
</tr>
<tr>
<td>Subnet</td>
<td>Process Group</td>
<td>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.</td>
</tr>
</tbody>
</table>
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”

FlowFiles are like HTTP data

<table>
<thead>
<tr>
<th>HTTP Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTTP/1.1 200 OK</td>
</tr>
<tr>
<td>Date: Sun, 10 Oct 2010 23:26:07 GMT</td>
</tr>
<tr>
<td>Server: Apache/2.2.8 (CentOS) OpenSSL/0.9.8g</td>
</tr>
<tr>
<td>Last-Modified: Sun, 26 Sep 2010 22:04:35 GMT</td>
</tr>
<tr>
<td>ETag: &quot;45b6-834-49130cc1182c0&quot;</td>
</tr>
<tr>
<td>Accept-Ranges: bytes</td>
</tr>
<tr>
<td>Content-Length: 13</td>
</tr>
<tr>
<td>Connection: close</td>
</tr>
<tr>
<td>Content-Type: text/html</td>
</tr>
<tr>
<td>Hello world!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FlowFile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard FlowFile Attributes</td>
</tr>
<tr>
<td>Key: ‘entryDate’ Value: ‘Fri Jun 17 17:15:04 EDT 2016’</td>
</tr>
<tr>
<td>Key: 'lineageStartDate' Value: 'Fri Jun 17 17:15:04 EDT 2016'</td>
</tr>
<tr>
<td>Key: 'fileSize' Value: '23609'</td>
</tr>
<tr>
<td>FlowFile Attribute Map Content</td>
</tr>
<tr>
<td>Key: ‘filename’ Value: ‘15650246997242’</td>
</tr>
<tr>
<td>Key: ‘path’ Value: ‘./’</td>
</tr>
</tbody>
</table>

Header

Content

Binary Content *
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

- Constrained
- High-latency
- Localized context

- Hybrid – cloud/on-premises
- Low-latency
- Global context
Deeper Ecosystem Integration: 180+ Processors

- 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)
MiNiFi on a Connected Car

Processing / Synthesis
- Transmit
- Execute
- Filter
- Prioritize

Comprehension
- Parse CAN
- Parse Ethernet
- Parse LIN
- Parse <>

Collection
- Listen CAN
- Listen Ethernet
- Listen LIN
- Listen <>

Gateway
- CAN Bus
- Ethernet / Ethernet AVB
- Local Interconnect Network
- Yet to be established protocol

Yet to be established protocol
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

**Physical Store**

- Client Libraries
- Mobile Devices
- Registers

**Distribution Center**

- MiNiFi Gateway Server
- NiFi
- NiFi
- NiFi
- Server Cluster
- Kafka
- Storm / Spark / Flink / Apex

**Core Data Center at HQ**

- Others
- Kafka
- Storm / Spark / Flink / Apex

**On Delivery Routes**

- MiNiFi Trucks
- Deliverers

DELIVERY TRUCK: Creative Stall, [https://thenounproject.com/creativestall/](https://thenounproject.com/creativestall/)
DELIVERER: Rigo Peter, [https://thenounproject.com/rigo/](https://thenounproject.com/rigo/)
CASH REGISTER: Sergey Patutin, [https://thenounproject.com/bdesign.by/](https://thenounproject.com/bdesign.by/)
HAND SCANNER: Eric Pearson, [https://thenounproject.com/epearson001/](https://thenounproject.com/epearson001/)
Apache NiFi Managed Dataflow
# Extension / Integration Points

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

OS/Host
- JVM
- Flow Controller
  - Processor 1
  - Extension N
- FlowFile Repository
- Content Repository
- Provenance Repository

Web Server

Local Storage

Standalone

Cluster

HTTP Client

API interaction can go to any node

ZooKeeper Server

Cluster Coordinator
Primary Node
ZooKeeper Client
NiFi Architecture – Repositories - Pass by reference

Excerpt of demo flow…

What’s happening inside the repositories…

BEFORE

AFTER

FlowFile

Content

Provenance
NiFi Architecture – Repositories – Copy on Write

Excerpt of demo flow...

BEFORE

| Name success | Queued 1/1,024 KB
| EncryptContent | EncryptContent |
| Read/Write | Out | Tasker/Time |
| 0 / 0 bytes | 0 bytes / 0 bytes | 0 / 0 bytes |

WHAT'S HAPPENING INSIDE THE REPOSITORIES...

<table>
<thead>
<tr>
<th>F_1 \rightarrow C_1</th>
<th>C_1</th>
<th>P_1 \rightarrow F_1 - CREATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>F_1.1 \rightarrow C_2</td>
<td>C_1 (plaintext)</td>
<td>P_1 \rightarrow F_1 - CREATE</td>
</tr>
<tr>
<td>P_2 \rightarrow F_1.1 - MODIFY</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AFTER

| Name success | Queued 0/0 bytes
| EncryptContent | EncryptContent |
| Read/Write | Out | Tasker/Time |
| 1/1,024 KB / 1 MB | 1/1 MB | 0/0 bytes |

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

Mirror of Apache NiFi

Unresolved: By Priority

<table>
<thead>
<tr>
<th>Priority</th>
<th>Issues</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocker</td>
<td>6</td>
<td>1%</td>
</tr>
<tr>
<td>Critical</td>
<td>19</td>
<td>2%</td>
</tr>
<tr>
<td>Major</td>
<td>495</td>
<td>45%</td>
</tr>
<tr>
<td>Minor</td>
<td>516</td>
<td>47%</td>
</tr>
<tr>
<td>Trivial</td>
<td>71</td>
<td>6%</td>
</tr>
</tbody>
</table>

Status Summary

<table>
<thead>
<tr>
<th>Status</th>
<th>Issues</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>1008</td>
<td>30%</td>
</tr>
<tr>
<td>In Progress</td>
<td>20</td>
<td>1%</td>
</tr>
<tr>
<td>Reopened</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Resolved</td>
<td>2234</td>
<td>66%</td>
</tr>
<tr>
<td>Closed</td>
<td>63</td>
<td>2%</td>
</tr>
<tr>
<td>Patch Available</td>
<td>69</td>
<td>2%</td>
</tr>
</tbody>
</table>

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
users@nifi.apache.org

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

Follow us on Twitter
@apachenifi
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