Apache NiFi

Better Analytics Demand Better Dataflow

Presented by: Joe Witt
Apache NiFi PPMC Member
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:

<table>
<thead>
<tr>
<th>Quality</th>
<th>Correct, complete, reliable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Right size, rate, format, schema, content, lightweight analysis</td>
</tr>
<tr>
<td>Timeliness</td>
<td>All data has a half-life. Not all data is created equal.</td>
</tr>
<tr>
<td>Secure</td>
<td>Confidential, unaltered</td>
</tr>
<tr>
<td>Compliant</td>
<td>Authorized, traceable</td>
</tr>
<tr>
<td>Recoverable</td>
<td>Errors happen. Iterate until it’s right.</td>
</tr>
</tbody>
</table>
Enterprise Dataflow: “What could possibly go wrong?”

**Acquire**

- Database
- Hadoop
- Twitter
- Handshake

**Analyze**

- DataTorrent
- Hadoop
- SAP HANA
- Storm
- Spark
- Trifacta

**Store**

- Oracle
- Teradata
- MongoDB
- Cassandra
- Handshake

Dataflow – Route, Transform, Mediate
Dataflow across the enterprise

Edge Sites

Regional Sites

Corporate Datacenters

Partners
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
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

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>In</td>
<td>13,157 / 1.25 GB</td>
<td>(5 min)</td>
</tr>
<tr>
<td>Read/Write</td>
<td>1.25 GB / 1.26 GB</td>
<td>(5 min)</td>
</tr>
<tr>
<td>Out</td>
<td>0 / 0 bytes</td>
<td>(5 min)</td>
</tr>
<tr>
<td>Tasks/Time</td>
<td>15,613 / 00:00:36.886</td>
<td>(5 min)</td>
</tr>
</tbody>
</table>
Flow Controller
NiFi Clustering Model

NiFi Cluster Manager (Master)

NiFi Nodes (Slaves)
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
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