Managing Data in Motion with the Connected Data Architecture

Dmitry Baev
Director, Solutions Engineering
Disclaimer

This document may contain product features and technology directions that are under development, may be under development in the future or may ultimately not be developed.

Project capabilities are based on information that is publicly available within the Apache Software Foundation project websites ("Apache"). Progress of the project capabilities can be tracked from inception to release through Apache, however, technical feasibility, market demand, user feedback and the overarching Apache Software Foundation community development process can all effect timing and final delivery.

This document’s description of these features and technology directions does not represent a contractual commitment, promise or obligation from Hortonworks to deliver these features in any generally available product.

Product features and technology directions are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.

Since this document contains an outline of general product development plans, customers should not rely upon it when making purchasing decisions.
Open source is the norm, and Apache is the center of gravity.
Connected Data Defines the New Way of Business

- Systems of Insight
- Personalized Experiences
- Balanced Supply Chains
- New Products & Services
- Operational Efficiencies
Harnessing Data in Motion

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

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

- **Core Infrastructure**
Hortonworks DataFlow: Data in Motion Platform

Platform to build dataflow management and streaming analytics solutions that collect, curate, analyze and act on data in motion across the data center and cloud.
Hortonworks DataFlow for Data in Motion
Powered by Apache NiFi, Kafka, and Storm

Easy, Secure, Reliable Way to Get the Data You Need
From the edge to anywhere with intelligence

Immediate and Continuous Insights
Because acting on perishable insights in real time maximizes value

Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance
Because it all has to work together in an enterprise environment
Hortonworks DataFlow Manages Data in Motion

Sources

Regional Infrastructure

Core Infrastructure

Constrained
High-latency
Localized context

Hybrid – cloud / on-premises
Low-latency
Global context
Simplistic View of Dataflows: Easy, Definitive
Realistic View of Dataflows: Complex, Convoluted
Requirements for Data in Motion

- Prioritization
- Connectivity
- Scalability
- Extensibility
- Real-Time
- Adaptability
- Security
- Provenance
- Perishable Insights
# Apache NiFi: Designed for 8 challenges of global enterprise dataflow

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visual Command and Control</strong></td>
<td>• For agile and immediate creation, configuration, control of dataflows</td>
</tr>
<tr>
<td><strong>Data Lineage (Provenance)</strong></td>
<td>• Ensures trust of your data</td>
</tr>
<tr>
<td><strong>Data Prioritization</strong></td>
<td>• Because not all data is of equal importance</td>
</tr>
<tr>
<td><strong>Data Buffering/Back-Pressure</strong></td>
<td>• Since not all senders/receivers/connections work perfectly all the time</td>
</tr>
<tr>
<td><strong>Control Latency vs Throughput</strong></td>
<td>• Adapt to different situations with different requirements</td>
</tr>
<tr>
<td><strong>Secure Control Plane/Data Plane</strong></td>
<td>• Security of data, and data access</td>
</tr>
<tr>
<td><strong>Scale out Clustering</strong></td>
<td>• Scalability</td>
</tr>
<tr>
<td><strong>Extensibility</strong></td>
<td>• Ecosystem flexibility and growth</td>
</tr>
</tbody>
</table>
Hortonworks DataFlow, powered by Apache NiFi
Connecting Data Between Ecosystems Without Coding: 180+ Processors

- FTP
- SFTP
- HL7
- UDP
- XML

- HTTP
- Email
- HTML
- Image
- Syslog
- AMQP
- MQTT

- Kafka
- Hadoop
- Spark
- Storm
- Elastic
- Hive
- Couchbase
- S3
- Cassandra
- Flink
- Minifi
- Splunk
- MongoDB

- Hash
- Encrypt
- GeoEnrich
- Merge
- Tail
- Scan
- Extract
- Evaluate
- Replace
- Duplicate
- Execute
- Translate
- Split
- Convert
- Route Text
- Distribute Load
- Generate Table Fetch
- Route Content
- Jolt Transform JSON
- Route Context
- Prioritized Delivery
- Control Rate
Edge Intelligence with Apache 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

**Different from Apache NiFi**

- Design and Deploy
- Warm re-deploys
Hortonworks DataFlow Manages Data in Motion

<table>
<thead>
<tr>
<th>EDGE/ SOURCES</th>
<th>REGIONAL INFRASTRUCTURE</th>
<th>CORE INFRASTRUCTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Data Acquisition</td>
<td>- Data Acquisition &amp; Movement</td>
<td>- Detect Complex Patterns in Real-Time</td>
</tr>
<tr>
<td>- Edge Intelligence</td>
<td>- Intelligent Routing &amp; Filtering</td>
<td>- Descriptive, Prescriptive &amp; Predictive Analytics</td>
</tr>
<tr>
<td>- Bi-directional Comm.</td>
<td>- Parse, Enrich &amp; Data Delivery</td>
<td>- Join/Split Streams</td>
</tr>
</tbody>
</table>

- Data Acquisition
- Edge Intelligence
- Bi-directional Comm.

- Java Agent
- C++ Agent

- Data Acquisition & Movement
- Intelligent Routing & Filtering
- Parse, Enrich & Data Delivery

- Detect Complex Patterns in Real-Time
- Descriptive, Prescriptive & Predictive Analytics
- Join/Split Streams

- Streaming Analytics Manager

**Context/Commands**
- Constrained
- High-latency
- Localized context

**Data/Events**
- Hybrid – cloud/on-premises
- Low-latency
- Global context
Stream Processing using Streaming Analytics Manager
Data in Motion Needs DataFlow Management and Stream Processing

Data in Motion Using DataFlow Management
- Data Ingestion
- Edge Intelligence
- First Mile Problem
- Physical Data Movement
- Simple event processing such as Route, Filter, Enrich, Transform, etc.

Data in Motion Using DataFlow Management and Stream Processing
- Flow Management to deliver data for Stream Processing
- Stream Processing for complex pattern matching on unbounded streams of data.
Hortonworks DataFlow: Data-in-Motion Platform

Flow management

Data Acquisition

Edge Processing

Flow management + Stream Processing

Real Time Stream Analytics

Rapid Application Development

Enterprise Services

Ambari

Ranger

Other services

Streaming Analytics Manager

DATA SCIENCE

(SPARK, HIVE, PHOENIX, ZEPPELIN, SQL)

MODERN DATA APPLICATIONS

APP DEV
**Example:** Company X provides alerting services when users’ resting heart rate higher than a threshold.
Data in Motion Needs Dataflow Management and Stream Processing

- **Acquire** data from various Wearable Device’s Cloud Instances
- **Move** Data from Customer Cloud Instances to on-premise instance
- **Intelligent Routing & Filtering** of data. The routing and filtering rules will be often changed at run-time.
- **Deliver** the data to various downstream systems. New downstream apps will be added over time
- **Parse** the device data to standardized format that downstream system can understand
- **Enrich** the data with contextual information including patient/customer info (age, gender, etc..)
- **Recognize** the Pattern when the resting heart rate exceeds a certain threshold (the insight), and then create an alert/notification.
- **Run a Outlier detection model** on streaming heart rate that comes in. If the score is above a certain threshold, alert on the heart rate.
Thank You
Managing Data in Motion with the Connected Data Architecture

Capture streaming data
Deliver perishable insights
Combine new & old data

Perishable Insights

ACTIONABLE INTELLIGENCE

Historical Insights

Store data forever
Access a multi-tenant data lake
Model with artificial intelligence

DATA IN MOTION
(Hortonworks DataFlow)

DATA AT REST
(Hortonworks Data Platform)
Hortonworks DataFlow: Data in Motion Platform

HDF provides the **flow management**, **stream processing**, and **enterprise services** needed to collect, curate, analyze and act on data-in-motion across the data center and cloud.

**Flow Management**
- Data acquisition and delivery
- Simple transformation and data routing
- Simple event processing
- End-to-end provenance
- Edge intelligence & bi-directional communication

**Stream Processing**
- Scalable data broker for streaming apps
- Scale out complex transformation

**Stream AnalyZcs**
- Pa8 ern Matching
- Prescrip6ve & Predic6ve Stream Analy6cs
- Complex Event Processing
- Continuous Insight

**Enterprise Services**
- Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance, Mul: -tenancy

- Apache Ambari
- Apache Ranger
- Hortonworks Schema Registry
Real-Time, Visual Control of Data Flows between Datacenter and the Cloud

Add and Adjust Data Sources to maximize the opportunity that you capture from perishable insights

Visually Trace the Data Path to manage the what, who, where and how around data in motion

Dynamically Adjust the Pipeline to match the dataflow with your bandwidth
Integrated Processes and Control for Data Flows and Streaming Analytics

**COMMON ARCHITECTURE**

**WITHOUT HORTONWORKS DATAFLOW**

- Ingest
- Messaging
- Scripts
- Scripts

**WITH HORTONWORKS DATAFLOW**

- HORTONWORKS DATAFLOW

**Optimize the Architecture**
Reduce cost and complexity with the most efficient data collection technologies

**Assure Efficient Operations**
Via real-time control of data inputs, outputs, transportation and transformations

**Rely on a Common Foundation**
Eliminating dependence on multiple customized systems
Secure Flows with Chain of Custody and Provenance

**End-to-End Security**
Apply security rules to encrypt, decrypt, filter and replace data from the point of collection at the jagged edge to its final destination

**Granular Control and Sharing**
Move beyond role-based access and dynamically share an entire dataflow

**Real-Time Traceability**
Rich metadata and contextual detail helps troubleshoot security issues and informs timely decisions
Streaming Data Flows with Adaptive Control

**Automated**
Bi-directional communication between source and destination adapts data flows automatically, according to current priorities.

**On-Demand**
Operational control to adapt to changing conditions and requirements.

**Scalable**
By incorporating data from any device—small machine sensors to enterprise data centers—HDF connects you to the broadest set of disparate data sources.
Data Acquisition & Delivery: First Mile Problem - Architecture

Consumer Devices
- MiNiFi
- Client Libraries
- DVR

Regional Data Center
- NiFi
- NiFi
- NiFi
- Server Cluster

Core Data Center
- NiFi
- NiFi
- NiFi
- Server Cluster
End to End Data in Motion Architecture

**Tier 1: Regional Data Center (~ 1000)**
- Router
- Switch
- Gateway
- Storage
- Client Libraries
- DVR
- Home Security

**Tier 2: Distribution Data Center (~ 100)**
- Gateway
- Storage
- Networking device
- Server Cluster
- Kafka

**Tier 3: National Data Center (3 - 6)**
- Storm/Spark/Flink/Apex
- Kafka
- Splunk

- Streaming Analytics Manager

---

© Hortonworks Inc. 2011 – 2016. All Rights Reserved