HORTONWORKS

SOLUTION SHEET

Hortonworks DataFlow (HDF) — End-to-End Data Flow Management and Streaming Analytics Platform

CREATE STREAMING ANALYTICS APPLICATIONS IN MINUTES WITHOUT WRITING CODE

The increasing growth of data, especially data-in-motion, presents enterprises with the challenges of managing streaming data and getting actionable intelligence. Hortonworks DataFlow (HDF) provides the only end-to-end streaming data platform with flow management, stream processing, and enterprise services which collect, curate, analyze and act on data in the data center and cloud. Complementary to the Hortonworks Data Platform (HDP®), HDF is powered by key open sourced projects including Apache® NiFi, Apache MiniFi, Apache Kafka®, Apache Storm™, and Druid.

- **Easy, Flexible, Secure Way to Get the Data You Need**
  The biggest challenge to getting data insights to work for your organization is getting the data in the first place; ingestion, cleansing, and preparing the data for analysis. This is complicated by data-in-motion, which could operate under varying conditions such as velocity and bandwidth over a geographically dispersed and fragmented network. HDF is designed to meet these data collection challenges securely and efficiently while giving real-time operational visibility, control, and management of the data flow. No more digging through log files.

Immediate and Continuous Insights
How do you analyze data-in-motion when it has not landed in a database yet? HDF's Streaming Analytics Manager (SAM) feature allows organizations to create analytics applications in minutes to capture perishable insights in real-time without writing a single line of code. Streaming Analytics Manager is a tool used to design, develop, deploy and manage streaming analytics applications using a drag-drop visual paradigm. A developer can build complex streaming analytics applications without having to know the complexities of the underlying streaming engine.

Enterprise Grade Corporate Governance, Security and Operations
Streaming data needs to meet the same enterprise corporate governance and security standards for operations as other traditional data types. HDF provides a visual tool for comprehensive provisioning, management, monitoring, security, auditing, compliance, and governance that's integrated with the rest of your Hadoop environment. With a central schema repository, IT DevOps can easily manage and govern the schemas needed for data flow across the enterprise for faster analytics application development.

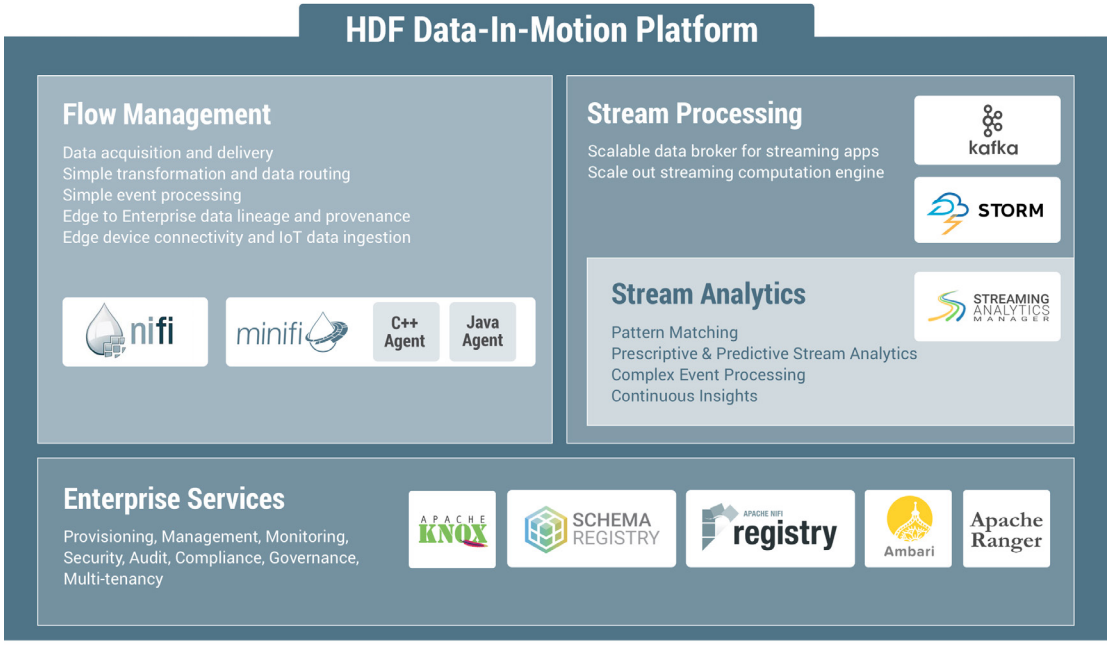

Figure 1: Hortonworks DataFlow (HDF) 3.2

FLOW MANAGEMENT

HDF provides an interactive data flow management platform powered by Apache NiFi/MiniFi for easy ingestion, routing, management, and delivery of any data anywhere (edge, cloud, data center) to any downstream system with intelligence. Our Flow Management is data source agnostic. Included security and encryption features protect data from source to storage over geographically dispersed communication links on a small scale, JVM-capable data sources, as well as enterprise-class data centers. With Flow Management, you get the following benefits:

Extremely Easy Data Collection

- Integrated with over 260+ data processors
- Source agnostic data collection
- Bi-directional command and control
- IoT device connectivity and data ingestion
- Support HDP 3.0 services such as Apache Hive 3 and HDFS 2

Runtime Adaptability

- Real-time visual control of data flows to add or adjust data sources and pipeline
- Ability to add contextual data to streaming data for immediate impact
- Adapt to system resource constraints in real-time with prioritized data transfer
- Intuitive visual interface

Always-On Data Provenance Audit Trails

- Data traceability and lineage to visually verify where data came from, how it was used, who viewed it, whether it was sent, copied, transformed or received
- Metadata supports data sharing compliance requirements and data flow troubleshooting and optimization

APACHE nifi

APACHE NIFI minifi

STREAM PROCESSING

HDF streaming analytics integrates with multiple processing engines such as Kafka and Storm. With the newly introduced integrated Streaming Analytics Manager, immediate and continuous insights using aggregations over windows, pattern matching, predictive and prescriptive analytics can be done. Analytics applications can also be built and deployed in minutes without writing any code. A robust SDK ensures that developers can also create custom analytics features.

New processors in NiFi and Streaming Analytics Manager support Kafka 1.1 features including message headers and transactions. From an operational standpoint, users can now install, configure, manage, upgrade, monitor, and secure Kafka 1.1 clusters with Apache Ambari™.

Streaming Analytics Manager (SAM)

SAM is built with application developers, business analysts as well as devops operators in mind. With SAM, get the following benefits:

Build Easily

- Data streams automatically connected through schema registry
- Drag and drop visual paradigm to build analytics applications
- Drop down analytics functions such as filtering, routing, rules engine and alerting
- Choose services from Service Pool to create development environment
- Experiment with creation of SAM apps using mock data and create unit tests for SAM apps using the new SAM “Test mode”

Operate Efficiently

- Easily test, debug, troubleshoot, and monitor the deployed applications
- Prebuilt monitoring dashboards of application system metrics
- Create and manage service pools for developer to easily create any dev environments
- Manage schema registry for easy schema attach to data streams

Analyze Quickly

- Analytics engine powered by Druid, an open source data store designed for OLAP queries on event data
- Rich visual dashboard powered by Apache Superset, with over 30 visualization charts right out of the box
- Easy deployment which is processing engine agnostic

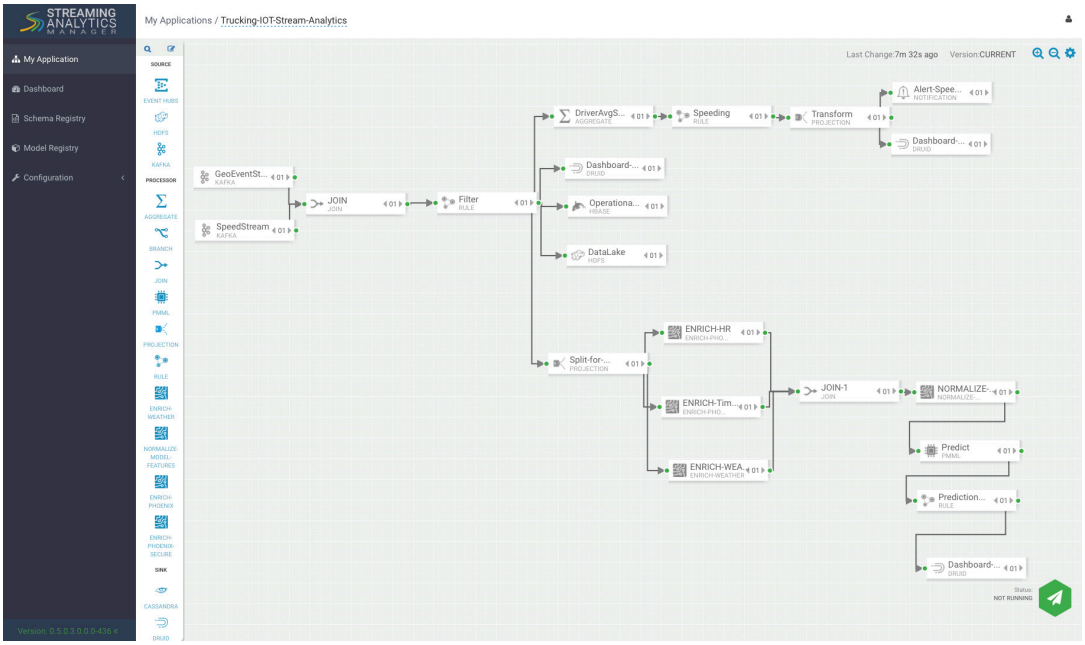

Figure 2: Streaming Analytics Manager

ENTERPRISE SERVICES

The Flow Management and Stream Processing services that power the HDF data-in-motion platform are complemented with enterprise services for provisioning, management, monitoring, security, audit, compliance and governance. These enterprise services use familiar components including Apache Ambari for operational cluster management and Apache Ranger for security across both HDP and HDF. These services enable IT to manage the entire HDF cluster efficiently and comprehensively.

Hortonworks Schema Registry

The Hortonworks Schema Registry provides a simple way to validate schema, enable format conversion, and enable the data producer and data consumer to evolve at different rates.

Schema Registry improves end-to-end data governance and operational efficiency by providing a centralized registry, supporting version management and enabling schema validation. With Schema Registry, you get the following benefits:

Centralized Registry

- Eliminates the need to attach schema to every piece of data
- Allows apps to flexibly interact with each other to save or retrieve schemas for the data they need to access
- Fully integrated with the flow management component of HDF, including NiFi
- Allows schemas created using NiFi to be easily managed and reused by the entire platform

Version Management

- Supports schema evolution so that a consumer and producer can understand different schema versions but still read all the information shared between them

Schema Validation

- Enables generic format conversion and generic routing within NiFi
- Facilitates schema validation to ensure data quality schema validation by enabling generic format conversion and generic routing to ensure data quality

Apache NiFi Registry

Apache NiFi Registry, a new Apache sub-project now included within HDF Enterprise Services, facilitates the development, management and portability of data flows. Core to its functionality is the ability to abstract data flow schemas and programs to enable users to track and monitor data flow changes at a more granular level. Data flow schemas are stored in a shared repository that allows for easy sharing on a global basis as well as versioning of schemas.

Through this, the export and import of data flows allow easy porting and enables smooth migration of data flows from one environment to another. The functionality significantly improves the storage, control, and management of versioned flows, further shortening the software development life cycle and accelerating application deployment to achieve faster time to value.

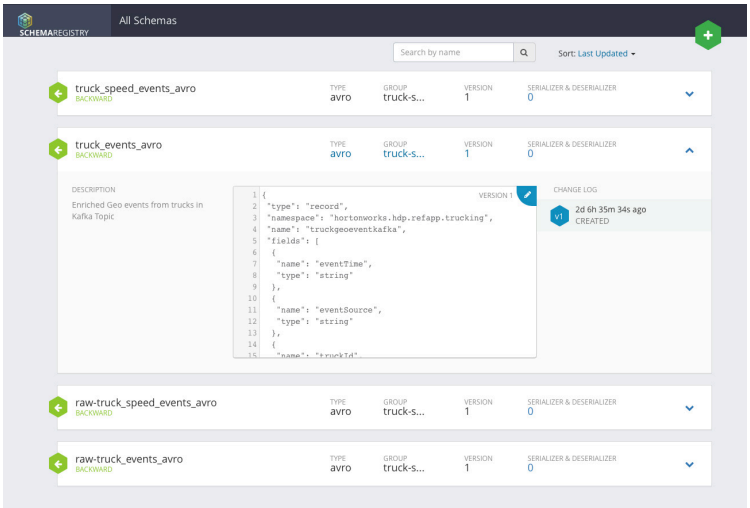

Figure 3: Schema Registry

KEY HDF USE CASES

- **Data movement**—Use HDF to move data within a data center, between data centers and between cloud and on-premises with intelligent movement.
- **Continuous data ingest**—Acquire data from the edge and ingest data from any data source with the ability to make flow changes in real-time.
- **Streaming ETL**—Ability to process and prepare streaming data for analysis with full data lineage of all extraction, transformation, and loads.
- **Streaming Analytics**—Ability to capture perishable and continuous insights with analytics modeling and enabling real-time actionable responses.
- **IoT**—Securely connect with edge devices, ingest data from the edge and stream it through the enterprise for gaining real-time insights.

CONCLUSION

Capture Continuous Insights from Data-in-Motion
Enterprises worldwide are undergoing digital transformations in their businesses. The growth of data, especially data-in-motion, and getting their actionable insights are challenges to be overcome. Hortonworks DataFlow meets these challenges with an end-to-end data flow management and streaming analytics platform. With an easy to use management UI and a visual analytics builder, developers and analysts can work closely together to build analytics applications quickly, and place into action the continuous insights from data-in-motion in no time at all.

About Hortonworks

Hortonworks is an industry leading innovator that creates, distributes and supports enterprise-ready open and Connected Data Platforms and Modern Data Applications that deliver actionable intelligence from all data: data-in-motion and data-at-rest. Hortonworks is focused on driving innovation in open source communities such as Apache Hadoop, Apache NiFi and Apache Spark. Along with its 2,100+ partners, Hortonworks provides the expertise, training and services that allow customers to unlock transformational value for their organizations across any line of business.

Contact

For further information,
visit hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

HORTONWORKS