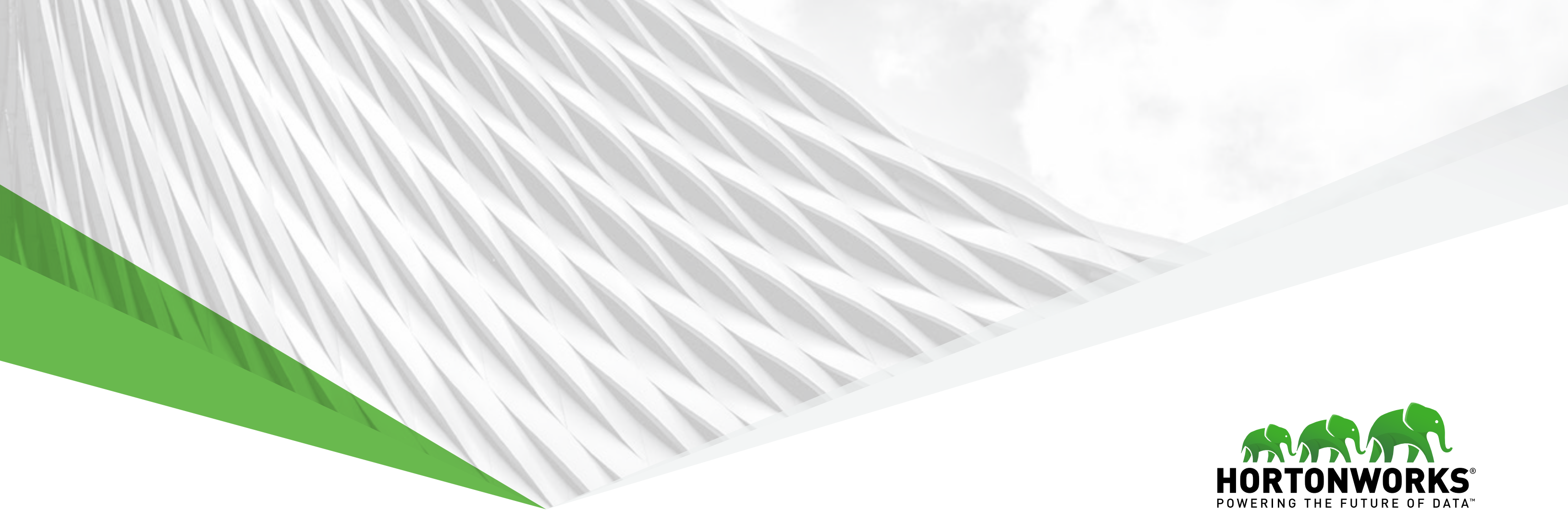


HORTONWORKS
POWERING THE FUTURE OF DATA™

HORTONWORKS DATAFLOW

Accelerating Big Data Collection and DataFlow Management

A HORTONWORKS WHITE PAPER
FEBRUARY 2016

Contents

What is Hortonworks DataFlow?

Hortonworks DataFlow (HDF), powered by Apache™ NiFi, is the first integrated platform that solves the complexity and challenges of collecting and transporting data from a multitude of sources be they big or small, fast or slow, always connected or intermittently available.

Hortonworks DataFlow is a single combined platform for data acquisition, simple event processing, transport and delivery, designed to accommodate the highly diverse and complicated dataflows generated by a world of connected people, systems and things.

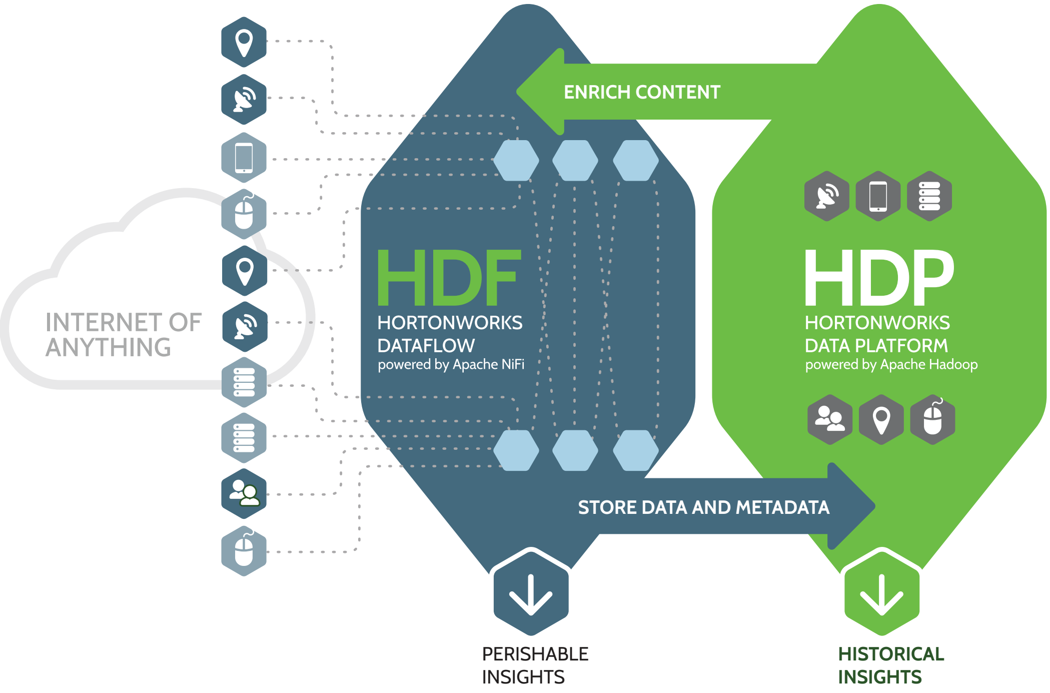
An ideal solution for the Internet of Any Thing (IoAT), HDF enables simple, fast data acquisition, secure data transport, prioritized data flow and clear traceability of data from the very edge of your network all the way to the core data center. Through a combination of an intuitive visual interface, a high fidelity access and authorization mechanism and an “always on” chain of custody (data provenance) framework, HDF is the perfect complement to HDP to bring together historical and perishable insights for your business. A single integrated platform for data acquisition, simple event processing, transport and delivery mechanism from source to storage.

A single integrated platform for data acquisition, simple event processing, transport and delivery mechanism from source to storage.

Hortonworks DataFlow is based on Apache NiFi, technology originally created by the NSA (National Security Agency) in 2006 to address the challenge of automating the flow of data between systems of all types– the very same problem that enterprises are encountering today. After eight years of development and use vat scale, the NSA Technology Transfer Program released NiFi to the Apache Software Foundation in the fall of 2014.

Hortonworks DataFlow enables the real time collection and processing of perishable insights.

Hortonworks DataPlatform can be used to enrich content and support changes to real-time dataflows.



Hortonworks DataFlow is designed to securely collect and transport data from highly diverse data sources be they big or small, fast or slow, always connected or intermittently available.

Figure 1: Hortonworks DataFlow

Benefits of Hortonworks DataFlow

DataFlow was designed inherently to meet the practical challenges of collecting data from a wide range of disparate data sources; securely, efficiently and over a geographically disperse and possibly fragmented network. Because the NSA encountered many of the issues enterprises are facing now, this technology has been field-proven with built-in capabilities for security, scalability, integration, reliability and extensibility and has a proven track record for operational usability and deployment.

HORTONWORKS DATAFLOW ENABLES ENTERPRISES TO

Leverage Operational Efficiency

- Accelerate big data ROI via simplified data collection and a visually intuitive dataflow management interface
- Significantly reduce cost and complexity of managing, maintaining and evolving dataflows
- Trace and verify value of data sources for future investments
- Quickly adapt to new data sources through an extremely scalable extensible platform

Make Better Business Decisions

- Make better business decisions with highly granular data sharing policies
- Focus on innovation by automating dataflow routing, management and trouble-shooting without the need for coding
- Enable on-time, immediate decision making by leveraging real time data bi-directional dataflows
- Increase business agility with prioritized data collection policies

Increase Data Security

- Support unprecedented yet simple to implement data security from source to storage
- Improve compliance and reduce risk through highly granular data access, data sharing and data usage policies
- Create a secure dataflow ecosystem with the ability to run the same security and encryption on smale scale JVM capable data sources as well as enterprise class datacenters

Accelerate big data ROI through a single data-source agnostic collection platform

Reduce cost and complexity through an intuitive, real-time visual user interface

Unprecedented yet simple to implement data security from source to storage

Better business decisions with highly granular data sharing policies

React in real time by leveraging bi-directional data flows and prioritized data feeds

Adapt to new data sources through an extremely scalable, extensible platform

Features of Hortonworks DataFlow

DATA COLLECTION

Integrated collection from dynamic, disparate and distributed sources of differing formats, schemas, protocols, speeds and sizes such as machines, geo location devices, click streams, files, social feeds, log files and videos

REAL TIME DECISIONS

Real-time evaluation of perishable insights at the edge as being pertinent or not, and executing upon consequent decisions to send, drop or locally store data as needed

OPERATIONAL EFFICIENCY

Fast, effective drag and drop interface for creation, management, tuning and troubleshooting of dataflows, enabling coding free creation and adjustments of dataflows in five minutes or less

SECURITY AND PROVENANCE

Secure end-to-end routing from source to destination, with discrete user authorization and detailed, real time visual chain of custody and metadata (data provenance)

BI-DIRECTIONAL DATAFLOW

Reliably prioritize and transport data in real time leveraging bi-directional dataflows to dynamically adapt to fluctuations in data volume, network connectivity and source and endpoint capacity

COMMAND AND CONTROL

Immediate ability to create, change, tune, view, start, stop, trace, parse, filter, join, merge, transform, fork, clone or replay dataflows through a visual user interface with real time operational visibility and feedback

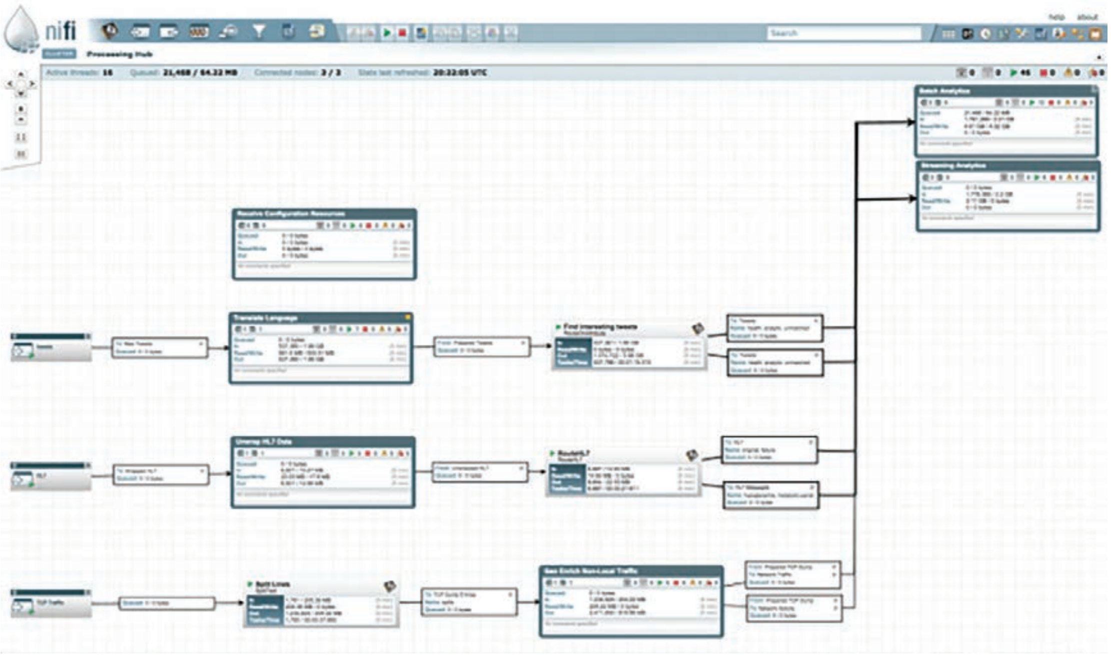


Figure 2: Apache NiFi Real-Time Visual User Interface

Common Applications of Hortonworks DataFlow

Hortonworks DataFlow accelerates time to insight by securely enabling off-the shelf, flow based programming for big data infrastructure and simplifying the current complexity of secure data acquisition, ingestion and real time analysis of distributed, disparate data sources.

An ideal framework for collection of data and management of dataflows, the most popular uses of Hortonworks DataFlow are for: simplified, streamlined big data ingest, increased security for collection and sharing of data with high fidelity chain of custody metadata and as the underlying infrastructure for the internet of things.

CASE 1: ACCELERATED DATA COLLECTION AND OPERATIONAL EFFECTIVENESS

Streamlined Big Data Ingestion

Hortonworks DataFlow accelerates big dwata pipeline ingest through a single integrated and easily extensible visual interface for acquiring, and ingesting data from different, disparate, distributed data sources in real time. The simplified and integrated creation, control and analyses of data flows results in faster ROI of big data projects and increased operational effectiveness.

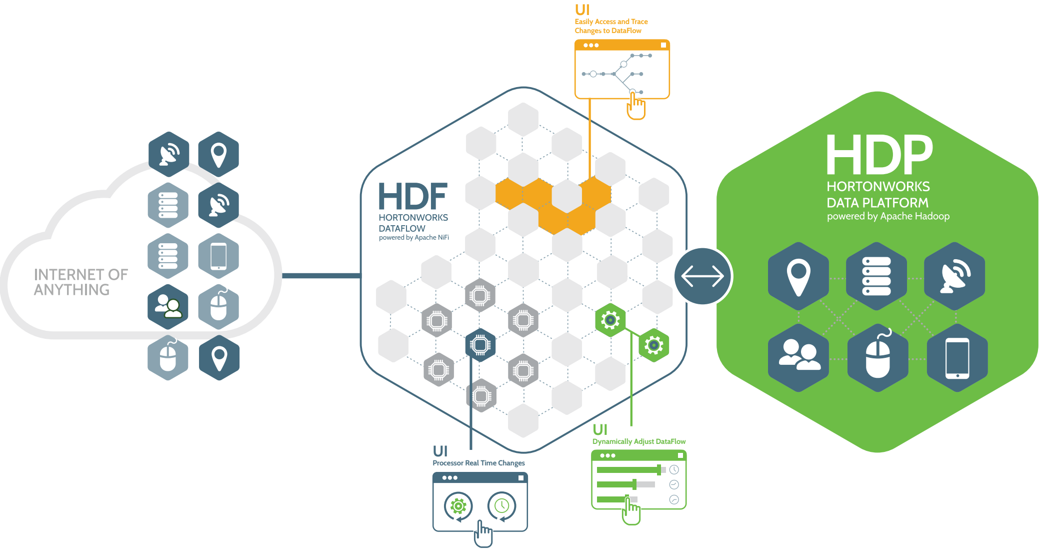

INTERNET OF ANYTHING
HDF
HORTONWORKS DATAFLOW
powered by Apache NiFi
UI
Easily Access and Trace Changes to DataFlow
UI
Processor Real Time Changes
UI
Dynamically Adjust DataFlow
HDP
HORTONWORKS DATA PLATFORM
powered by Apache Hadoop

Figure 2: An integrated, data source agnostic collection platform

WHAT IS A COMMAND AND CONTROL INTERFACE?

A command and control interface provides the ability to manipulate the dataflow in real time such that current contextual data can be fed back to the system to immediately change its output. This is in contrast to a design and deploy approach which involves statically programming a dataflow system before the data is flowed through it, and then returning to a static programming phase to make adjustments and restarting the dataflow again. An analogy to this would be the difference between 3D printing which requires pre-planning before execution, and molding clay which provides immediate feedback and adjustment of the end product in real time.

To learn more about Hortonworks DataFlow go to http://hortonworks.com/hdf

Why aren't current data collection systems ideal?

Current big data collection and ingest tools are purpose-built and over-engineered simply because they were not originally designed with universally applicable, operationally efficient design principles in mind. This creates a complex architecture of disparate acquisition, messaging and often customized transformation tools that make big data ingest complex, time consuming and expensive from both a deployment and a maintenance perspective. Further, the time lag associated with the command line and coding dependent tools fetters access to data and prevents the on-time, operational decision making required of today's business environment.

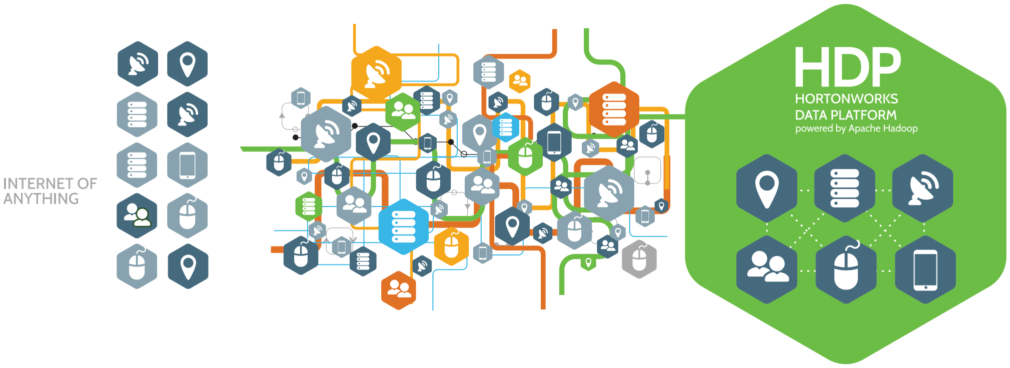


Figure 3: Current big data ingest solutions are complex and operationally inefficient

CASE 2: INCREASED SECURITY AND UNPRECEDENTED CHAIN OF CUSTODY

Increased security and provenance with Hortonworks DataFlow

Data security is growing ever more important in the world of ever connected devices and the need to adhere to compliance and data security regulations is currently difficult, complex and costly. Verification of data access and usage is difficult, time consuming and often involves a manual process of piecing together different systems and reports to verify where data is sourced from, how it is used, who has used it and how often.

Current tools utilized for transporting electronic data today are not designed for the expected security requirements of tomorrow. It is difficult, if not almost impossible for current tools to share discrete bits of data, much less do so dynamically—a problem that had to be addressed in the environment of Apache NiFI as a dataflow platform used in governmental agencies.

Hortonworks Dataflow addresses the security and data provenance needs in an electronic world of distributed real time big data flow management. Hortonworks DataFlow augments existing systems with a secure, reliable, simplified and integrated big data ingestion platform that ensures data security from all sources – be they centrally located, high volume data centers or remotely distributed data sources over geographically dispersed communication links. As part of its security features, HDF inherently provides end-to-end data provenance—a chain-of-custody for data. Beyond the ability to meet compliance regulations, provenance provides a method for tracing data from its point of origin, from any point in the dataflow, in order to determine which data sources are most used and most valuable.

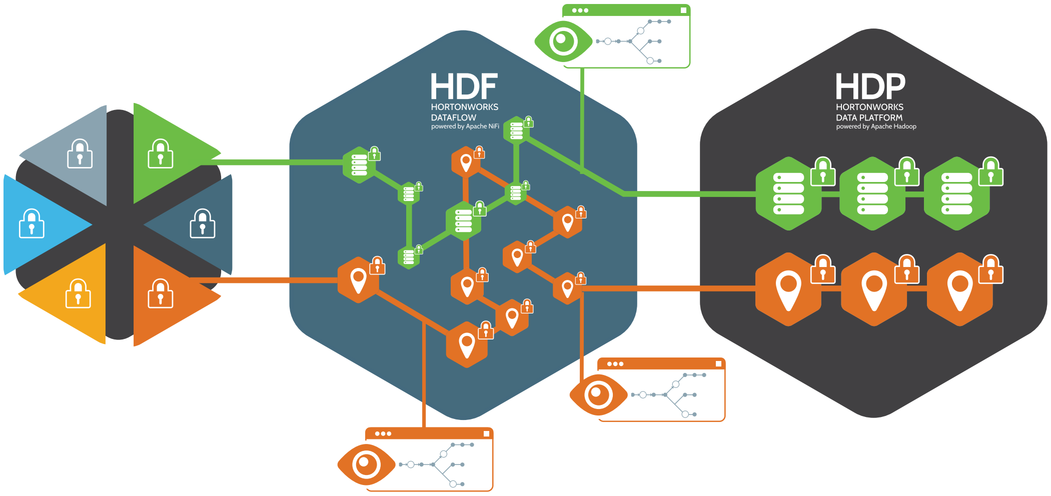


Figure 4: Secure from source to storage with high fidelity data provenance

WHAT IS DATA PROVENANCE?

Provenance is defined as the place of origin or earliest known history of some thing. In the context of a dataflow, data provenance is the ability to trace the path of a piece of data within a dataflow from its place of creation, through to its final destination. Within Hortonworks DataFlow data provenance provides the ability to visually verify where data came from, how it was used, who viewed it, whether it was sent, copied, transformed or received. Any system or person who came in contact with a specific piece of data is captured in its completeness in terms of time, date, action, precedents and dependents for a complete picture of the chain of custody for that precise piece of data within a dataflow. This provenance metadata information is used to support data sharing compliance requirements as well as for dataflow troubleshooting and optimization.

Data democratization with unprecedented security

Hortonworks DataFlow opens new doors of business insights by enabling secure access without hindering analysis, as very specific data can be shared or not shared. For example, Mary could be given access to discrete pieces of data tagged with the term “finance” within a dataflow, while Jan could be given access to the same dataflow but with access only to data tagged with “2015” and “revenue”. The removes the disadvantages of role based data access which can inadvertently create security risks, while still enabling democratization of data for comprehensive analysis and decision making.

Hortonworks Dataflow, with its inherent ability to support fine grained provenance data and metadata throughout the collection, transport and ingest process provides comprehensive and detailed information needed for audit and remediation unmatched by any existing data ingest system in place today.

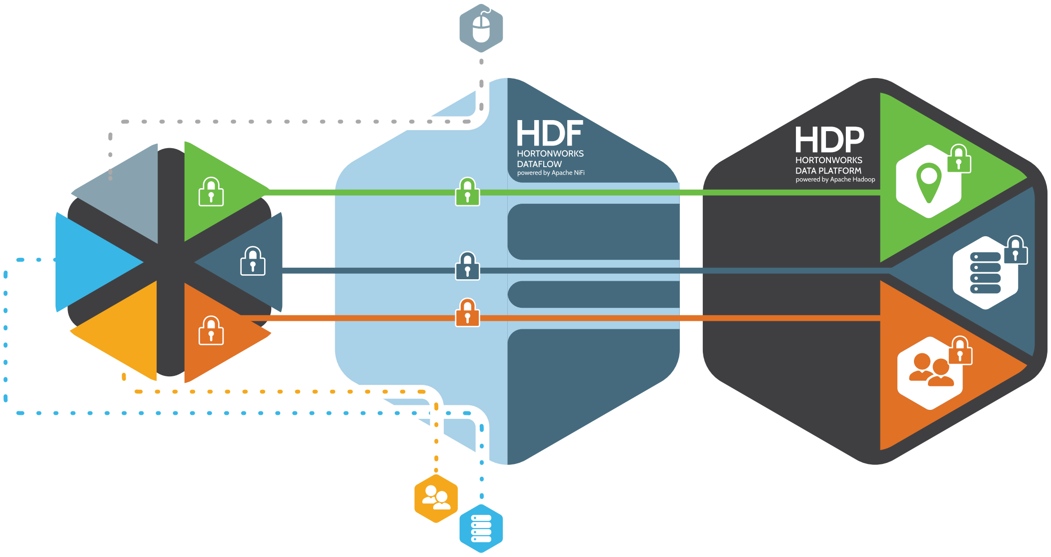


Figure 5: Fine-grained data access and control

CASE 3: THE INTERNET OF ANY THING (IOAT)

The Internet of Any Thing with Hortonworks DataFlow

Designed in the field, where resources are scarce (power, connectivity, bandwidth), Hortonworks DataFlow is a scalable, proven platform for the acquisition and ingestion of the Internet of Things (IoT), or even more broadly, the Internet of Any Thing (IoAT).

Adaptive to Resource Constraints

There are many challenges in enabling an ever connected yet physically dispersed Internet of Things. Data sources may often be remote, physical footprints may be limited, power and bandwidth are likely to be both variable and constrained. Unreliable connectivity disrupts communication and causes data loss while the lack of security on most of the world's deployed sensors puts businesses and safety at risk.

At the same time, devices are producing more data than ever before. Much of the data being produced is data-in-motion and unlocking the business value from this data is crucial to business transformations of the modern economy.

Yet business transformation relies on accurate, secure access to data from the source through to storage. Hortonworks DataFlow was designed with all these real-world constraints in mind: power limitations, connectivity fluctuations, data security and traceability, data source diversity and geographical distribution, altogether, for accurate, time-sensitive decision making.

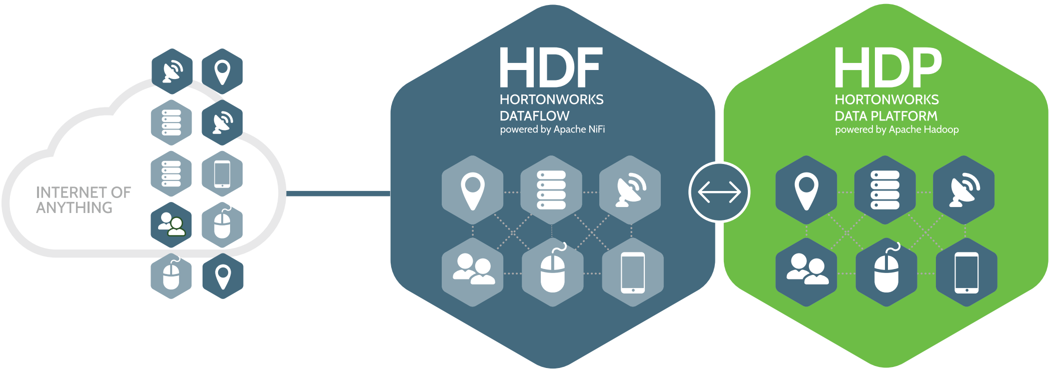


Figure 6: A proven platform for the Internet of Things

Secure Data Collection

Hortonworks Dataflow addresses the security needs of IoT with a secure, reliable, simplified and integrated big data collection platform that ensures data security from distributed data sources over geographically dispersed communication links. The security features of HDF includes end-to-end data provenance — a chain-of-custody for data. This enables IoT systems to verify origins of the dataflow, trouble-shoot problems from point of origin through destination and the ability to determine which data sources are most frequently used and most valuable.

Hortonworks DataFlow is able to run security and encryption on small scale, JVM-capable data sources as well as enterprise class datacenters. This enables the Internet of Things with a reliable, secure, common data collection and transport platform with a real-time feedback loop to continually and immediately improve algorithms and analysis for accurate, informed on-time decision making.

Prioritized Data Transfer and Bi-directional Feedback Loop

Because connectivity and available bandwidth may fluctuate, and the volume of data being produced by the source may exceed that which can be accepted by the destination, Hortonworks DataFlow supports the prioritization of data within a dataflow. This means that should there be resource constraints, the data source can be instructed to automatically promote the more important pieces of information to be sent first, while holding less important data for future windows of transmission opportunity, or possibly not sent at all. For example, should an outage from a remote device occur, it is critical to send the "most important" data from that device first as the outage is repaired and communication is re-established, Once this critical "most important" data has been sent, it can then be followed by the backlog of lower priority data that is vital to historical analysis, but less critical to immediate decision making.

Hortonworks DataFlow enables the decision to be made at the edge of whether to send, drop or locally store data, as needed, and as conditions change. Additionally, with a fine grained command and control interface, data queues can be slowed down, or accelerated to balance the demands of the situation at hand with the current availability and cost of resources.

With the ability to seamlessly adapt to resource constraints in real time, ensure secure data collection and prioritized data transfer, Hortonworks DataFlow is a proven platform ideal for the Internet of Things.

Why Hortonworks for Apache Hadoop?

Founded in 2011 by 24 engineers from the original Yahoo! Apache Hadoop development and operations team, Hortonworks has amassed more Apache Hadoop experience under one roof than any other organization. Our team members are active participants and leaders in Apache Hadoop developing, designing, building and testing the core of the Apache Hadoop platform. We have years of experience in Apache Hadoop operations and are best suited to support your mission-critical Apache Hadoop project.

For an independent analysis of Hortonworks Data Platform and its leadership among Apache Hadoop vendors, you can download the Forrester Wave™: Big Data Apache Hadoop Solutions, Q1 2014 report from Forrester Research.

About Hortonworks

Hortonworks is a leading innovator at creating, distributing and supporting enterprise-ready open data platforms. Our mission is to manage the world's data. We have a single-minded focus on driving innovation in open source communities such as Apache Hadoop, NiFi, and Spark. Our open Connected Data Platforms power Modern Data Applications that deliver actionable intelligence from all data: data-in-motion and data-at-rest. Along with our 1600+ partners, we provide the expertise, training and services that allows our customers to unlock the transformational value of data across any line of business. We are Powering the Future of Data™.

Contact

For further information visit
For more information,
visit www.hortonworks.com.

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

© 2011-2016 Hortonworks Inc. All Rights Reserved.
Privacy Policy | Terms of Service