IBM Power Systems
Solution Brief

Hortonworks DataFlow and IBM Power Systems

Accelerating data science and real-time analytics at scale

Highlights

- Boost data scientist productivity with collaboration tools and rapid model training
- Enable quick and easy access to machine learning, deep learning and artificial intelligence applications
- Capitalize on data in motion at the edge of the network with streaming analytics and dataflow management solutions from Hortonworks
- Support your next-generation data science and analytics solutions with the performance, flexibility and openness of IBM Power Systems

Until now, organizations looking to get the most business value possible from big data have focused primarily on deploying cost-efficient storage and performing deep analytics across diverse data sources. However, the rules of the game are changing: the growing volume of Internet of Things (IoT) devices is creating a new class of streaming data, while artificial intelligence (AI) is making it possible to capitalize on that data like never before.

To keep up in this changing world, businesses must be able to catch insights not only inside the data lake but also at the edge of the network and take action immediately. They can accomplish this by accelerating development of machine learning models with data science, leveraging Hadoop storage and accelerated computing, and then applying those models in real-time within streaming data flows to drive immediate actions based on model predictions.

Data science and real-time edge analytics can support use cases across a variety of industries, including:

- Logistics: Monitor trucking fleets in real time to mitigate driving infractions
- Retail: Analyze and visualize social media data about particular products to support real-time promotions
- Energy and utilities: Monitor transmission lines with drones and smart meters to predict and prevent failures
- Finance: Protect credit card customers from fraud
- Insurance and healthcare: Provide personalized policies and treatments

IBM® Power Systems™ is the ideal server platform to support the latest data science and analytics solutions. This includes solutions that make data scientists more productive and effective (IBM Data Science Experience), solutions that make deep learning, machine learning and AI more accessible (IBM PowerAI), and solutions that enable real-time analytics of data in motion (Hortonworks DataFlow).

Boost data scientist productivity with IBM Data Science Experience

Data scientists are a highly valuable asset in any organization, but it's often difficult to find enough qualified data science professionals to hire. In fact, Gartner predicted there will be a shortfall of 100,000 data scientists by 2020.[1] In addition, data scientists must pull together skills and experience from across computer science, statistics, and specific business or industry domains. No one individual is likely to be an expert in all three of these areas, so it's important to give your data scientists tools that can accentuate their strengths, while also shoring up their weaknesses.

Developed based on input from hundreds of real data scientists, IBM Data Science Experience (DSX) helps you get the most productivity possible from your data science team. DSX gives data scientists the ability to select the tools and capabilities that best meet their needs, including the most popular open source tools, as well as unique IBM value-added functionality.

DSX also provides a social environment, allowing data scientists to collaborate with one another to solve data challenges, while sharing their expertise. When data scientists can find out what works and what doesn't, without having to do the trial and error themselves, they'll be empowered to work faster and get more done.

By deploying DSX on Power Systems, you can also take advantage of accelerated business insights, thanks to Power Systems' performance advantages. For instance, organizations that run DSX on Power can complete model training in half the time required by comparable x86 systems.[2] Since these models will be based on the most current data and trends, data scientists can feel confident acting on the insights they turn out.

Accelerate deep learning, machine learning, and AI with IBM PowerAI

By harnessing the power of deep learning—a subset of machine learning based on training patterns that allow neural networks to make sense of the data they encounter—developers and data scientists can begin taking advantage of business data to draw value in ways never thought possible before.

IBM PowerAI is a software distribution built specifically to help organizations use deep learning, machine learning, and AI to their full potential. Exclusively available on Power Systems, the Best Server for Enterprise AI, it's even easier for organizations to start unlocking the value of ML/DL applications. PowerAI is integrated as part of IBM DSX on Power Systems.

PowerAI brings together all the most popular deep learning frameworks and their supporting libraries in a single easy-to-deploy platform. The platform is fully optimized and supported by IBM out of the box, meaning that it gives organizations everything they need to start drawing value from ML/DL quickly. PowerAI also helps data scientists significantly cut down the amount of time they waste performing data preparation tasks or experimenting with settings to optimize performance. This leaves them more time for the actual high-value work of ML/DL.

Built from the ground-up for enterprise AI, POWER9 is the only processor with state-of-the-art I/O subsystem technology, including next-generation NVIDIA NVLink, PCIe Gen4, and OpenCAPI. These interfaces give POWER9 a memory bandwidth superhighway that allows you to run ML/DL applications with unmatched performance and scalability, handling even very large data sets with ease. As a result, you can identify business insights faster and drive greater value through a variety of industry use cases, including understanding and responding to customers better and discovering new business opportunities.

Harness real-time streaming analytics and data flow management with Hortonworks DataFlow

Hortonworks DataFlow (HDF) is the only end-to-end streaming data platform on the market today. Built with 100 percent open source components, HDF empowers organizations to collect, curate, analyze, secure, govern and act upon data at the edge of the network in real time. Three main elements make up the HDF platform, as shown in Figure 1: flow management, stream processing, and enterprise services.

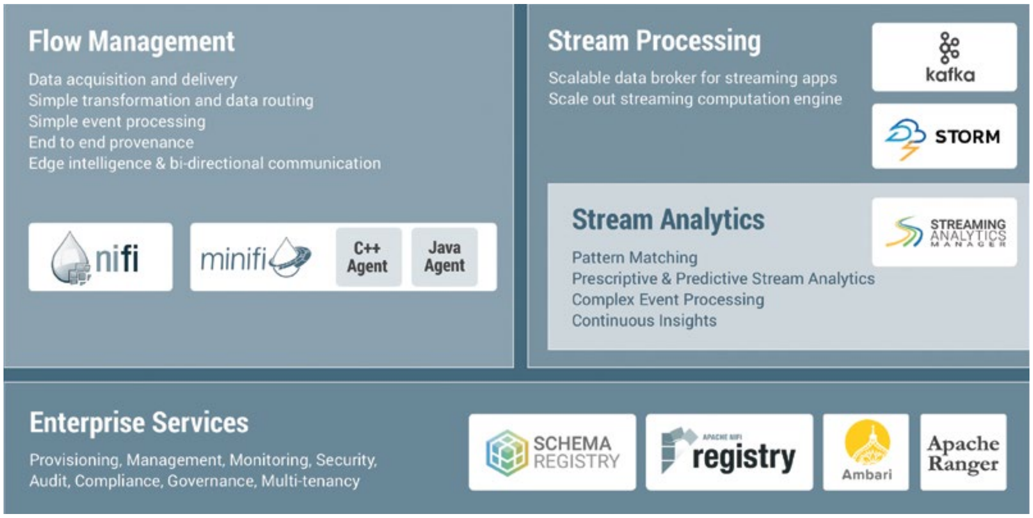
Flow Management
Data acquisition and delivery
Simple transformation and data routing
Simple event processing
End to end provenance
Edge intelligence & bi-directional communication
nifi
minifi
C++ Agent
Java Agent
Stream Processing
Scalable data broker for streaming apps
Scale out streaming computation engine
kafka
STORM
Stream Analytics
Pattern Matching
Prescriptive & Predictive Stream Analytics
Complex Event Processing
Continuous Insights
STREAMING ANALYTICS MANAGER
Enterprise Services
Provisioning, Management, Monitoring, Security, Audit, Compliance, Governance, Multi-tenancy
SCHEMA REGISTRY
APACHE NIFI registry
Ambari
Apache Ranger

Figure 1: Hortonworks DataFlow Data-In-Motion Platform. Some HDF components shown may not be included in a particular version of HDF on Power.

Flow management

HDF allows you to quickly and securely ingest, route, manage and deliver data from the edge to cloud or on-prem data centers. HDF flow management also provides an intuitive visual interface, allowing users to control data flows in real time, and built-in encryption to keep data secure, from source to storage.

Stream processing

HDF's Streaming Analytics Manager (SAM) makes it easier to build new streaming applications without coding, using a simple drag-and-drop interface. This empowers users to build and manage new streaming apps quickly and consistently.

HDF's stream processing and analytics capabilities expand into pattern matching, predictive stream analytics, complex event processing and continuous insights of data in motion.

Enterprise services

HDF offers enterprise services such as provisioning, management, monitoring, security, audit, compliance and governance, to ensure your streaming analytics platform works well within an enterprise environment.

Also, your operations teams can use the visual management interface to do their day-to-day tasks seamlessly.

HDF on Power

Thanks to its high performance, flexibility and openness, Power Systems makes a great choice to support HDF for a variety of important big data use cases, including physical data movement, continuous data ingestion and streaming analytics.

Bringing it all together with Hortonworks Data Platform

Once streaming data is processed and acted upon to capture any perishable insights, a version of the data, often filtered or aggregated, must be stored in a secure, reliable enterprise data lake so it can extend the data set for future modelling and deep insights.   Hortonworks Data Platform (HDP) on IBM Power Systems is the perfect combination of openness, reliability and performance for this data at rest.

HDP is the leading 100 percent open source Apache Hadoop and Spark distribution. IBM Power Systems is a member of the openPOWER family of servers, which benefit from a community of hundreds of organizations innovating around the Power processor. Together, IBM and Hortonworks have more than three times the code committers to Apache open source projects than the next closest contributor.   This commitment to open software and hardware development ensures that innovation is delivered rapidly, while avoiding vendor lock-in.

In the world of connected data, it is even more important to ensure the security and governance of data is maintained as it flows through the pipeline from inception to disposition. IBM and Hortonworks have lead the development of enterprise features in the open source community in this area, with projects like Apache Ranger, Atlas and Knox. This ensures that security policies and data lineage can be easily preserved across the entire ecosystem of the connected data platform to maintain compliance, including General Data Protection Regulation (GDPR) readiness.

Organizations that can harness the insights from streaming data to better serve and protect their clients and business goals will ensure they can continue to compete and prosper. Building a data science practice to take advantages of AI techniques such as deep learning and applying dataflow management and analytics at the edge of the network are critical steps in this journey.

About the IBM/Hortonworks partnership

IBM and Hortonworks are working to bring together the best of DSX with the best of HDP, creating an enterprise-class data science platform that's secure, scalable, and supported by the latest tools for data scientists. Integrating HDF on Power for data in motion with HDP on Power for data at rest ensures that users are getting insights from new data in real time, while also having a secure enterprise data lake they can use to store data for future analytics workloads.

Together, IBM and Hortonworks are helping data scientists and developers across industries innovate and build models faster, putting them in a better position to draw insights and business value from their data. Since we support an open approach, our work drives big data innovation for the entire Hadoop community, not just for our own clients.

For more information

To learn more, contact your IBM representative or IBM Business Partner, or visit **ibm.biz**/hortonworksOnPower.

For more information about Hortonworks®, please visit their website at www.hortonworks.com.

HORTONWORKS®

IBM®

IBM Corporation
IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
February 2018

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Clients are responsible for ensuring their own compliance with various laws and regulations, including the European Union General Data Protection Regulation. Clients are solely responsible for obtaining advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulations that may affect the clients' business and any actions the clients may need to take to comply with such laws and regulations. The products, services, and other capabilities described herein are not suitable for all client situations and may have restricted availability. IBM does not provide legal, accounting or auditing advice or represent or warrant that its services or products will ensure that clients are in compliance with any law or regulation.

Learn more about IBM's own GDPR readiness journey and our GDPR capabilities and offerings to support your compliance journey here.

[1] Gartner, *Defining and Differentiating the Role of the Data Scientist*, (https://blogs.gartner.com/doug-laney/defining-and-differentiating-the-role-of-the-data-scientist)

[2] Test results based on running a machine learning workload based on k-means clustering algorithm on data sets size ranging from 1GB to 15 GB. Test system details: Power Systems S822 LC HPC – 20 Cores, 512 GB RAM and SSD, Power Systems S822LC Big Data – 20 Cores, 512 GB, HDDs, Intel Server with Broadwell E5 2640 v4 – 20 cores, 512 GB and SSD, Intel Server with Broadwell E5 2699 v4 – 44 cores, 512 GB, HDD.

Please Recycle