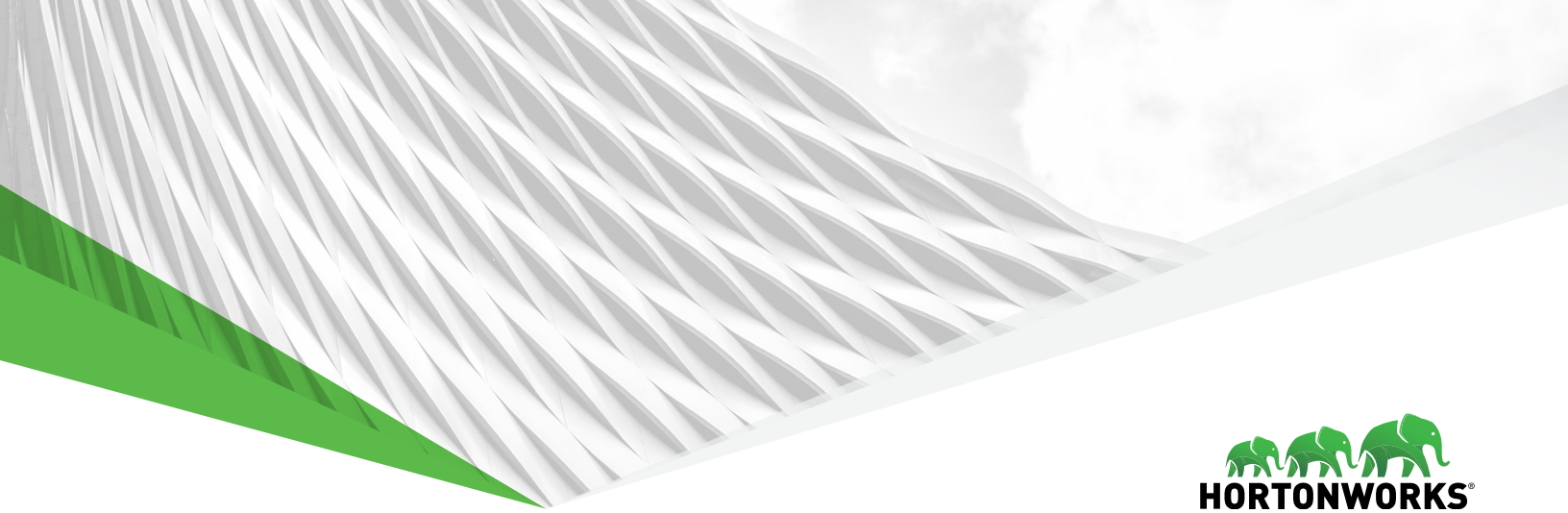

The Power of One: IBM + Hortonworks

Overcome Enterprise Data Challenges with One Solution

A HORTONWORKS WHITE PAPER

MAY 2018

IBM® and Hortonworks® have partnered to give enterprises easy access to the capabilities, scalability and economy of Apache™ Hadoop®, plus additional governance and security features as well as tools for data federation, advanced query and management of their data. The result is an analytic solution that's open-source, enterprise-ready, and future-proof.

The Challenge: Driving Analytics Benefit from Today's Data

Enterprise-wide access to and analysis of data is essential. The challenge is to accommodate new technologies such as cloud, artificial intelligence (AI) and the Internet of Things (IoT) that are driving tremendous volume, velocity and variety of data. The biggest challenge is the growth in data variety, fueled by growing semi- and unstructured data from social media, streaming audio/video, log data, images, clickstreams and more. This results in cascade effect of exponential increase in data volume at unprecedented velocity. Evidently, as study shows that 85% of today's data is not used, the scope of the challenge becomes clear.

Organizations need their data scientists, business analysts and developers to draw from all relevant data sources to stay competitive, whether by gaining a holistic view of customer behavior or innovating internal operations and processes. The challenges of new data diversity and the growing demands for insight have spurred the adoption of data lakes as a flexible and powerful platform for data management and analytic insight.

One of the most popular data lake technologies is Apache Hadoop, a highly scalable open-source platform designed to process very large data sets across hundreds to thousands of parallel computing nodes. It provides a cost-effective storage solution for data ingestion with no initial format requirements. As an open-source platform, it is built on code contributions from a community of some of the world's best developers. Tapping into such a broad group of collaborators drives continuous innovation and provides community-driven support.

However, with the emergence of new business models and innovations, Hadoop is at the verge of reinventing itself to be the next-generation data platform that powers technology advancement. With a newer, more powerful version release of Hadoop in 2018, the partnership of IBM and Hortonworks has extended its capabilities, driving new levels of data science and machine learning, and further building Hadoop into an enterprise-class data platform with advanced analytic capabilities. The result is a single solution that combines Hortonworks Data Platform (HDP®), Hortonworks DataFlow (HDF) and IBM technology.

HORTONWORKS®

- #1 pure open-source Hadoop distribution
- 1300+ customers and 2100+ ecosystem partners
- Expertise from the original architects, developers and operators of Hadoop (formerly at Yahoo)

IBM®

- #1 SQL engine for complex, analytical workloads
- #1 data science platform (according to Gartner)
- Leadership in on-premises and hybrid cloud solutions
- OpenPOWER performance leadership
- Flexible, software-defined storage

A Combined Solution from IBM and Hortonworks

The IBM and Hortonworks solution enables a data-lake-based Hadoop infrastructure with improved exploration, discovery, testing and advanced data query. It also offers massive scalability, security and governance along with the ability to federate both data-at-rest and data-in-motion across the entire organization. Users can easily query both relational databases and Hadoop, on-premises or in the cloud. They benefit from self-service data access as well as the ability to do ad-hoc and real-time queries. Ultimately, the IBM and Hortonworks solution is built to better support machine learning and data science at enterprise scale.

Hortonworks Data Platform and Hortonworks DataFlow

Hortonworks Data Platform (HDP) is the industry's only truly secure, enterprise-ready open source Apache Hadoop distribution based on a centralized resource allocation architecture (YARN). YARN maximizes data ingestion by enabling enterprises to analyze data to support advanced use cases and coordinates cluster-wide services for operations, data governance and security. HDP addresses the complete needs of data-at-rest and data-in-motion, powering real-time customer applications and delivering real-time analytics for your data scientists, analysts and developers, either on-premises or in the cloud.

With Hortonworks Data Platform, users can:

- Deploy, integrate and work with unprecedented volumes of semi-structured, unstructured and structured data.
- Use an open source platform to avoid vendor lock-in.
- Minimize the expense and effort required to connect their IT infrastructure with Hadoop .
- Save time and money by leveraging their current IT infrastructure .
- Ensure security is consistently administered across their data lake.

Hortonworks DataFlow (HDF) provides simple, fast data acquisition, security-rich data transport, prioritized data flow and clear data traceability. This end-to-end platform collects, curates, analyzes and acts on data-in-motion in real time, on-premise or in the cloud. It was designed to handle all types of data delivery from data sources to complex processing systems such as Hadoop, and integrate with other data technologies.

HDF helps users to:

- Maximize the value of data-in-motion from sources such as IoT.
- Aggregate all types of data delivered from complex processing systems such as Spark, Storm, Google Cloud DataFlow, as well as Hadoop and other data storage systems.
- Manage streaming real-time analytics with a drag-and-drop visual interface that includes Data Flow Management Systems, Stream Processing and Enterprise Service.
- Integrate Apache NiFi/MiNiFi, Apache Kafka, Apache Storm and Druid.

By combining HDF with HDP, enterprise users can more easily integrate many data types from various sources and store them in one place. This level of integration means they can run the same industry-leading, open-source platform both on-premises and in the cloud, and ensure that security measures are consistently administered across different data access engines.

Together HDP and HDF offer a comprehensive platform to accommodate the volume, variety and velocity of data at enterprise scale. With the addition of IBM technology, an IBM–Hortonworks ecosystem is created for the extended usage of data, enabling improved query, advanced analytics, and predictive insight that drive the speed of business required to stay competitive today.

The IBM and Hortonworks Ecosystem

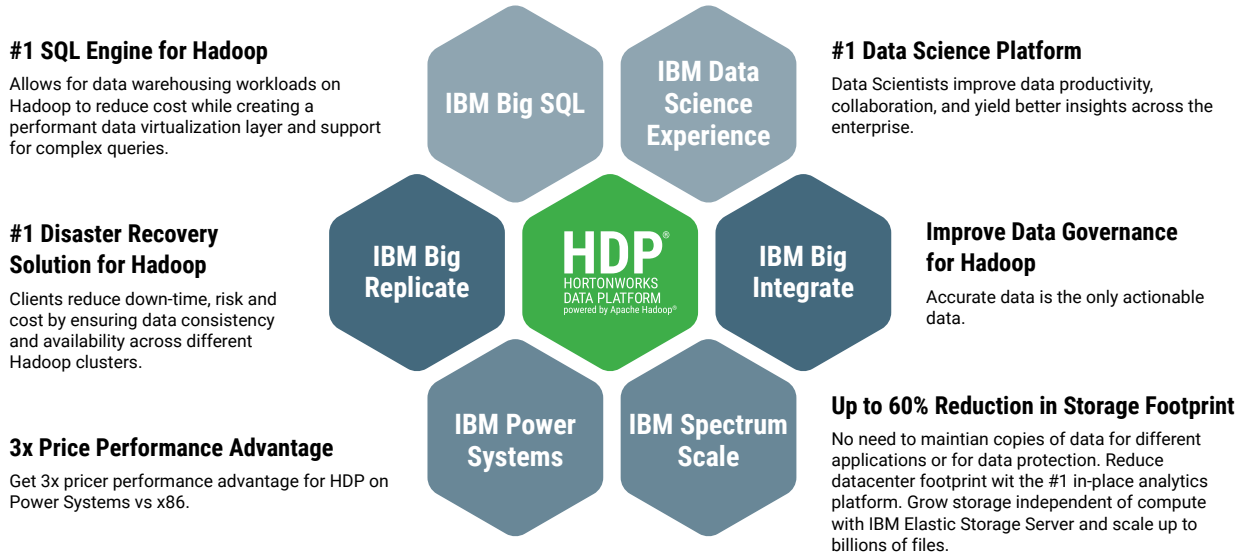

Figure 1: IBM and Hortonworks value proposition

By combining HDP, HDF and IBM, enterprises can gain the following benefits:

- **An enterprise-capable Hadoop distribution**
  Provides massive scalability, security and governance.
- **IoT data ingestion**
  Enables running analytics on the edge and making educated decisions before sending to data center.
- **Enterprise data movement and hybrid cloud**
  Enable data movement: remote to data center; between data centers; and between data center and cloud. Seamlessly fuses data flows between data centers.
- **Stream processing**
  Delivers insights across multiple data streams.
- **Unified queries that support timely and responsive analytic insight**
  Deliver insights necessary to act on time-sensitive matters that affect the entire enterprise.
- **Tools to collect, aggregate, federate and query virtually any data**
  Provide data insights throughout the organization, in real time, from both Hadoop and relational databases, on-premises and in the cloud.

The success of joint solution adoption is well exemplified at a leading casual gaming company in the U.S. Their use case involves combining structured customer data from disparate data systems with semi-structured data of gaming activity logs stored in Hadoop. The profound impact of an integrated view of data allows the company to focus on answering business-critical questions, without being sidetracked by repetitive data engineering that consumes valuable time and IT resources. Furthermore, the speed and quality of data insights have improved tremendously because analytics are done on enriched sets of data with 3x faster data movement to Hadoop than the previous solution. The resulting analytics improves the attractiveness of the gaming company's algorithms, and presents cross- and up-selling opportunities that directly impact the top- and bottom-line for the company.

Enterprises that choose the IBM + Hortonworks solution can leverage the following IBM data management tools to extend and enrich their data insights:

IBM Db2® Big SQL®

A hybrid, highly scalable enterprise-grade SQL engine for Apache Hadoop that delivers easy data querying across the enterprise. It concurrently exploits Hive, HBase and Spark using a single database connection — or even a single query.

With Db2 Big SQL, organizations can use a single database connection or query to connect to disparate sources such as HDFS, RDMS, NoSQL databases, object stores and WebHDFS. Enjoy low latency, support for ad-hoc and complex queries, high performance, robust security, SQL compatibility and federation capabilities to get the most from your data warehouse and SQL on Hadoop.

IBM Data Science Experience

Data science is an interdisciplinary field that combines machine learning, statistics, advanced analysis, and programming. IBM Data Science Experience provides a set of critical tools and a collaborative environment through which analysts and developers can create new analytic models quickly and easily. For example, IBM Machine Learning, found in the Data Science Experience, can halve the time it takes to build and deploy analytic models for application development, according to IBM testing.

IBM Big Replicate

Active-active data replication for Hadoop across supported environments, distributions, and hybrid deployments. It replicates big data from lab to production, from production to disaster recovery sites, or from ground to cloud object stores governed by the most demanding business and regulatory requirements.

IBM BigIntegrate

Superior connectivity, fast transformation and reliable, easy-to-use data delivery features that execute on the data nodes of a Hadoop cluster. This in-memory data integration solution provides superior connectivity, data profiling capabilities, metadata management, and integration with IBM Streams.

IBM Power® Systems

Cloud-ready servers built for the most demanding, data-intensive computing on earth. Unleash insight from your data pipeline — from managing mission-critical data, to managing your operational data stores and data lakes, to delivering the best server for cognitive computing.

IBM Spectrum Scale™

Advanced storage management of unstructured data for cloud, big data, analytics, objects and more. Supports both new-era big data and traditional applications with security, reliability and high performance. IBM Spectrum Scale is a high-performance solution for managing data at scale with the distinctive ability to perform archive and analytics in place.

For more information

To learn more about the IBM and Hortonworks solution, please contact your Hortonworks representative, or visit the **Hortonworks** website.

About Hortonworks

Hortonworks is a leading provider of enterprise-grade, global data management platforms, services and solutions that deliver actionable intelligence from any type of data for over half of the Fortune 100. Hortonworks is committed to driving innovation in open source communities, providing unique value to enterprise customers. Along with its partners, Hortonworks provides technology, expertise and support so that enterprise customers can adopt a modern data architecture. For more information, visit hortonworks.com.

Contact

For further information,
visit **hortonworks.com**

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

HORTONWORKS

05/2018