Hortonworks Data Platform on IBM Power Systems

Secure, enterprise-ready open source Apache Hadoop distribution for the leading open server for big data analytics and artificial intelligence

Highlights

- **Maximum productivity**—Double your data scientist productivity[1] with faster insights or more supported users.
- **Maximum efficiency**—Faster ingest, unmatched scalability and up to 60 percent smaller storage footprint[2] with IBM® Elastic Storage™ Server.
- **Ready for cognitive**—Allows seamless integration of machine and deep learning using IBM's PowerAI deep learning platform.
- **Committed to client success and open innovation**—Complete, enterprise-ready solution built on open hardware and software technology that is fully tested and offered with industry-leading support and expertise.

Hortonworks Data Platform (HDP) on IBM Power Systems™ delivers a superior solution for the connected enterprise data platform. With industry-leading performance and IT efficiency combined with the best of open technology innovation to accelerate big data analytics and artificial intelligence (AI), organizations can unlock and scale data-driven insights for the business like never before.

Hortonworks Data Platform

An industry-leading, secure and enterprise-ready open source Apache Hadoop distribution, HDP addresses a range of data-at-rest use cases, powering real-time customer applications and delivering robust analytics to accelerate decision-making and innovation.

HDP uses the Hadoop Distributed File System (HDFS) or HDFS API-compatible alternative storage systems for scalable, fault-tolerant big data storage and Hadoop's centralized Yet Another Resource Negotiator (YARN) architecture for resource and workload management. YARN enables a range of data processing engines including SQL, real-time streaming and batch processing, among others, to interact simultaneously with shared datasets, avoiding unnecessary and costly data silos and unlocking an entirely new approach to analytics. HDP now supports GPU-enabled YARN, which allows machine learning, deep learning, or artificial intelligence applications to be processed on servers with GPUs.

An open and flexible data platform, HDP includes a comprehensive set of capabilities including data access, governance and integration, along with security and operations management. HDP's open source community development model allows your organization to take advantage of rapid innovation and deep integration across the enterprise.

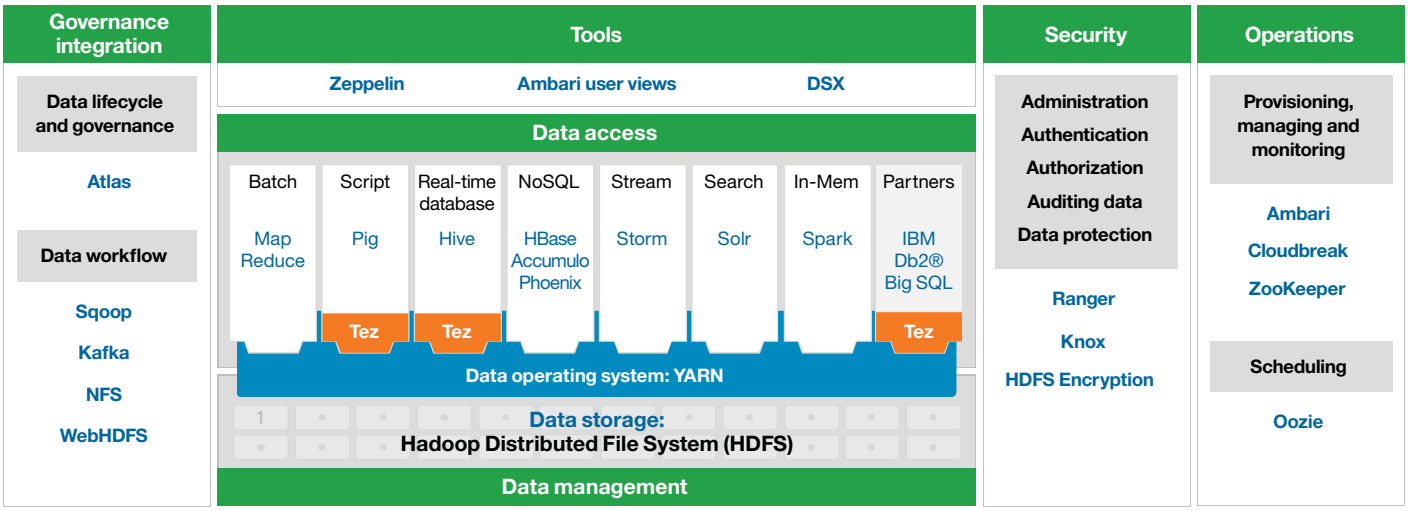

Figure 1: Hortonworks Data Platform

Hortonworks Data Platform on Power Systems—better together

Power Systems with IBM POWER9® processors and differentiated hardware acceleration technology are designed to deliver breakthrough performance for big data analytics workloads. The POWER9 processor delivers industry-leading performance for big data analytics and AI applications running on HDP, with multi-threading designed for fast execution of analytics (four threads per core), multi-level cache for continuous data load and fast response (including an L4 cache) and a large, high-bandwidth memory workspace to maximize throughput for data-intensive applications.

The POWER9 processor's leading thread density, large cache and memory bandwidth, and superior I/O capabilities are a great match for in-memory Apache Spark and AI workloads, including SQL, streaming, graph, and machine or deep learning applications.

IBM Power Systems OpenPOWER LC server family

Designed for flexibility and seamless integration to existing clusters and clouds, the IBM OpenPOWER LC server family offers the data-crushing POWER9 processor in a range of purpose-built system configurations, from compute-dense to storage-rich. The LC family's innovative design in partnership with the OpenPOWER Foundation offers hardware accelerator-offload for compute, storage and networking workloads—for incredible speed-ups to analytics and massive efficiencies in data movement.

OpenPOWER brings the leading processor together with the best of our partners and end users across the ecosystem—from High Performance Computing installations, enterprise IT and hyperscale data centers and to system designers worldwide.

The HDP on IBM Power Systems reference architecture suggests options for a solution that's sized to your specific needs, built with a combination of the high-performance, storage-rich Power Systems LC922 and the lightweight, yet compute-powerful Power Systems LC921. See Figure 2 for a summary of the reference architecture. For complete details, refer to the IBM reference architecture documentation.[3]

Superior performance for Apache Hadoop, Spark and data science workloads

HDP on Power Systems delivers more data faster, enabling valuable analytics for better and quicker decision making. In IBM performance testing for a typical Spark workload, HDP on Power Systems versus x86-based solutions demonstrated 59 percent better performance[4] for the same infrastructure cost.

	System Management Node	Master Node	Edge Node	Worker Node		
Cluster Type	All	All	All	**Balanced**	**Performance**	**Storage Dense**
Server Model	1U LC921	1U LC921	1U LC921	2U LC922	2U LC922	2U LC922
# Servers (Min/Default/Max)	1 / 1 / 1	3 / 3 / Any	1 / 1 / Any	4 / 8 / Any	4 / 8 / Any	4 / 8 / Any
Sockets	2	2	2	2	2	2
Cores (total)	16	40	40	44	44	44
Memory	32GB	256GB	256GB	256GB	512GB	128GB
Storage—HDD (front) Storage—SSD (front)	2x 4TB HDD	4x 4TB HDD	4x 4TB HDD	12x 4TB HDD	8x 4TB HDD + 4x 3.8TB SSD	12x 10TB HDD
Storage - HDD (rear for OS)				2x 1.2TB HDD	2x 1.2TB HDD	2x 1.2TB HDD
Storage Controller	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)
Network* — 1GbE	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)
Cables* - 1 GbE	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)
Network** - 10 GbE	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)	2x 2-port Intel (4 ports)	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)
Cables** - 10 GbE	2 cables (DACs)	2 cables (DACs)	4 cables (DACs)	2 cables (DACs)	2 cables (DACs)	2 cables (DACs)
Operating System	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9

* The 1 GbE network infrastructure hosts the following logical networks: campus, management, provisioning and service networks.

** The 10Gbe network infrastructure hosts the data network.

Figure 2: HDP on IBM Power Systems reference configuration

For machine learning model development, IBM Power Systems supported twice the number of data scientists with better response times compared to equivalent x86-based solutions.[1]

Ready for cognitive with PowerAI and Data Science Experience

HDP on IBM Power Systems clients can seamlessly integrate with IBM's fully optimized and supported PowerAI platform for deep learning. Optimized for blazing performance on the Power Systems AC922 with NVIDIA NVLink Technology and Tesla P100 GPUs, PowerAI includes the most popular deep learning frameworks, precompiled and easily deployed for maximum-throughput, scalable deep learning on your connected data.

For organizations looking for a complete data science productivity workbench, HDP and PowerAI integrate with the IBM Data Science Experience (DSX). DSX runs optimized on IBM Power Systems LC922 and AC922 servers to deliver more accurate and faster AI models.

IBM Spectrum Scale and Elastic Storage Server

An integrated storage system running IBM Spectrum Scale software on IBM Power Systems, IBM Elastic Storage Server can serve as the underlying storage for HDP. IBM Spectrum Scale is a software-defined storage system based on a parallel file system architecture that provides File (NFS, SMB, POSIX) and Object (S3, Swift) access and supports HDFS APIs. Support for HDFS APIs enables in-place analytics on enterprise storage instead of copying data from enterprise storage to analytics silos. In-place analytics not only eliminates duplication of data but also avoids the problems of running analytics on stale data. Support for POSIX access enables super-fast ingest. In addition, Spectrum Scale provides shared storage to HDP, which allows for de-coupling of compute and storage to enable optimized configurations.

Elastic Storage Server provides a consolidated storage solution that can store a wide variety of data types and a range of applications with standard access methods, creating a dynamic, shared data ocean that scales capacity and performance with demand. A shared data lake (a storage repository that holds a vast amount of raw data in its native format until it is needed) allows for the same data to be shared across different application domains and global locations while helping reduce the need for data movement and copies. This helps save significant costs for storage, floor space and administration.

Commitment to client success

HDP on IBM Power Systems and Elastic Storage Server is fully tested, validated and supported by Hortonworks and IBM who provide deep industry expertise and dedicated commitment to client success. IBM and Hortonworks are leading supporters of the open source Apache Hadoop and Spark communities, driving innovation and advancements in the big data ecosystem.

IBM and Hortonworks partner to offer a wide range of complementary solutions to support a data lake based on HDP and Power Systems. For a list of example solutions, visit: **ibm.com**/developerworks/library/l-isv-solution-hortonworks

For more information

To learn more, contact your IBM representative or visit: **ibm.com**/power

IBM Corporation
IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
August 2018

1 Based on IBM internal testing of the core computational step to form 5 clusters using a 350694 x 301 float64 data set (1 GB) running the K-means algorithm.

2 60 percent reduction in storage is due to HDFS making two additional copies for data protection and availability while IBM ESS only requires 30 percent overhead. Additional copies can be avoided with ESS by sharing the same data over NAS and Object protocols with other enterprise applications.

3 Hortonworks Data Platform 3 on IBM POWER9 with internal storage: **ibm.com**/common/ssi/cgi-bin/ssialias?htmlfid=06017906USEN&

Hortonworks Data Platform 3 on IBM POWER9 with IBM Elastic Storage Server: **ibm.com**/common/ssi/cgi-bin/ssialias?htmlfid=07017907USEN&

4 Results are based on IBM internal measurements running four concurrent streams of 99 TPC-DS like queries against a 3TB dataset.

Please Recycle

POS03160-USEN-01