Hortonworks Data Platform (HDF®) 3.0 with IBM—faster, smarter, hybrid data

Open source framework for distributed storage and processing of big data

Key challenges faced by enterprises today in their digital transformation journey:

- Flexibility and agility to build and deploy data-intensive apps quickly: developers want faster time to deployment, in minutes rather than hours or weeks
- Infinite scalability to support billions of files and thousands of nodes
- Support for deep learning apps and workloads, including TensorFlow and Caffe
- Enterprise security and governance requirements for data lakes
- Real-time enterprise data warehouse (EDW) database to optimize queries for historical and real-time data, on premises and in the cloud
- Independent software vendor (ISV) strategy to integrate third-party apps, including IBM® Data Science Experience (DSX)

Companies looking at new, innovative technologies to retool business processes and better serve their customers often encounter serious hurdles along the way. These hurdles include building and deploying new applications quickly, accessing the data timely and scaling to support the massive volumes of data generated.

In addition, the rise of machine learning and deep learning apps have brought processing challenges—with new use cases in manufacturing, insurance and other industries—along with increasing pressure at an affordable price. Nevertheless, it’s now crucial for businesses to seize these opportunities for growth and efficiency, and to take full advantage of the availability of cheaper data storage, distributed processing and more powerful computers.

Hortonworks Data Platform 3.0 with IBM

Hortonworks Data Platform (HDP) 3.0 with IBM delivers significant new features, including the ability to launch apps in a matter of minutes and address new use cases for high-performance deep learning and machine learning apps. In addition, this new version of HDP enables enterprises to gain value from their data faster, smarter, in a hybrid environment.

Key benefits and value

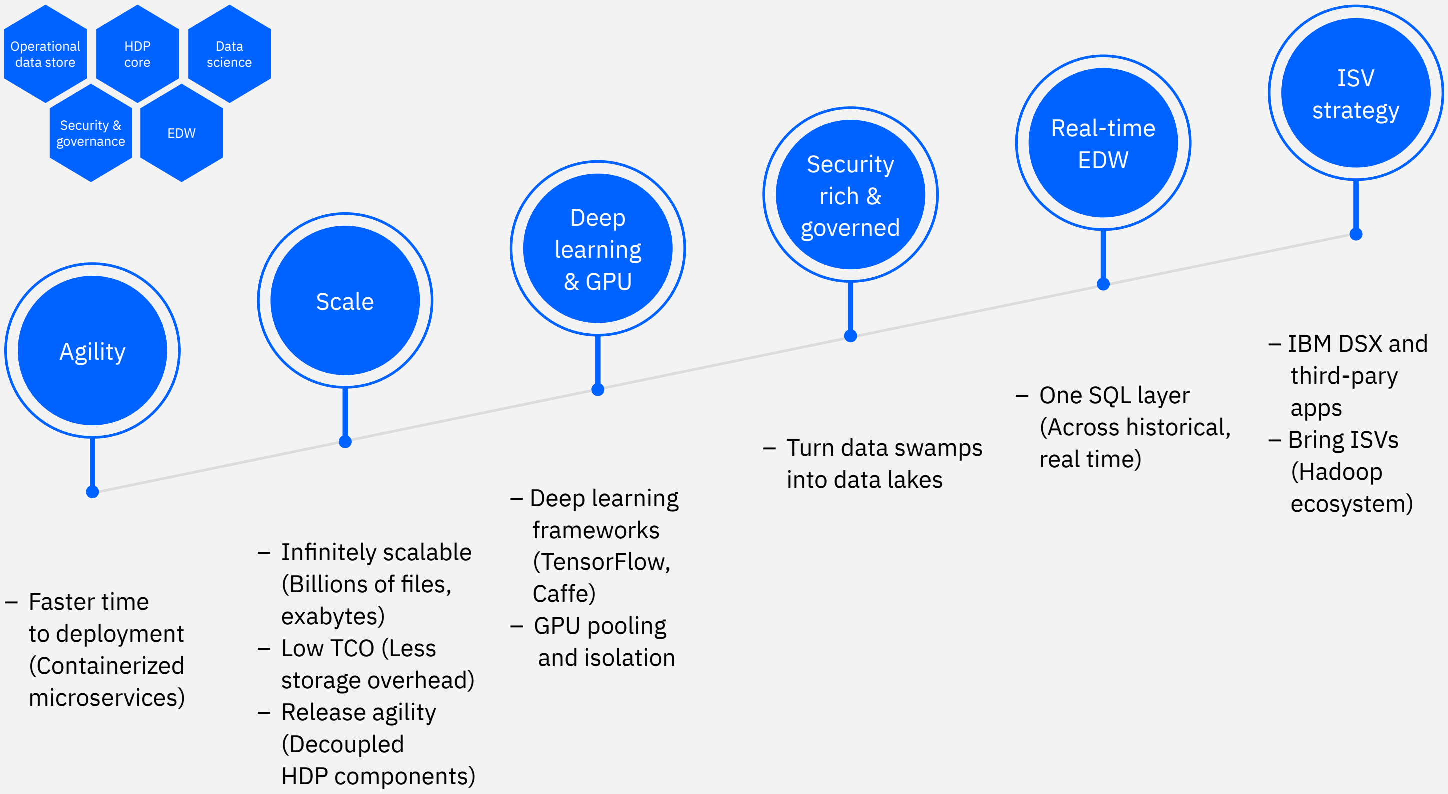
Faster
Agile app deployment from containerization allows customers to build data-intensive apps quickly, rolling them out in minutes and decreasing time to market. HDP 3.0 is your “cleanest” path to containerization.

Smarter
Deep learning app support through graphics processing units (GPUs) to make intelligent decisions, empowering customers to run GPU-loving workloads, such as machine learning and deep learning apps. GPU support can accelerate the time to insight from months to days.

Hybrid
Modern hybrid data architecture includes cloud storage support to store endless amounts of data in its native format, including Microsoft Azure Data Lake Store (ADLS), Windows Azure Storage Blob (WASB), Amazon Simple Storage Service (S3), Google Cloud Platform (GCP) tech preview. Data-at-rest and data-in-motion support on premises and in the cloud.

Bigger
Scalable with NameNode federation, allowing customers to scale to thousands of nodes and a billion files, while providing higher availability with multiple name nodes, at significantly lower total cost of ownership (TCO) with erasure coding.

Market drivers



Agility

- Faster time to deployment (Containerized microservices)

Scale

- Infinitely scalable (Billions of files, exabytes)
- Low TCO (Less storage overhead)
- Release agility (Decoupled HDP components)

Deep learning & GPU

- Deep learning frameworks (TensorFlow, Caffe)
- GPU pooling and isolation

Security rich & governed

- Turn data swamps into data lakes

Real-time EDW

- One SQL layer (Across historical, real time)

ISV strategy

- IBM DSX and third-pary apps
- Bring ISVs (Hadoop ecosystem)

Real-time database
Integration of Apache Hive offers the only unified solution to provide interactive query at scale—whether the data sits on premises or in the cloud.

Data science
Performance improvements around Apache Spark and Apache Hive integration, Spark connectors to other features, including cloud and containerized TensorFlow tech preview combined with GPU pooling.

Security rich
Comprehensive security and governance from Apache Ranger and Apache Atlas provide the ability to track the lineage of data from its origin to the data lake, allowing auditors to follow data through the entire enterprise.

Key capabilities

Containerization provides YARN support for Docker containers, allowing third-pary applications to run on Hadoop, for example, containerized applications, enabling:

- Faster time to deployment by enabling third-party apps
- The ability to run multiple versions of an application, enabling users to rapidly create features by developing and testing new versions of services without disrupting old ones
- Improved resource utilization and increased task throughput for containers, yielding faster time to market for services

GPU pooling and isolation helps ensure a first-class resource type in Hadoop, helping make it easier for customers to run machine learning and deep learning workloads.

- Compute-intensive analytics require not only a large compute pool, but also a fast and expensive processing pool with GPUs in tandem.
- Customers can share cluster-wide GPU resources without having to dedicate a GPU node to a single tenant or workload.
- GPU isolation dedicates a GPU to an application so that no other application has access to that GPU.

Erasure coding offers a data protection method that until now has mostly been found in object stores.

- Hadoop 3.0 will no longer default to storing three full copies of each piece of data across its clusters. Only 1.5 copies are needed for data recovery.
- Erasure coding boosts storage efficiency by 50 percent, allowing more efficient data replication.
- **NameNode federation** protects data in case of failures or disaster recovery.
- Erasure coding allows scaling up to thousands of nodes and billions of files.
- It supports multiple standby NameNodes. If one goes down, the cluster can continue to operate.

What makes HDP 3.0 unique?

Faster
Fastest time to deployment for data-intensive containers apps

Smarter
High-performance compute for deep learning and machine learning apps though GPU support

Hybrid
Cloud storage support to store endless amounts of data in its native format, including Amazon S3, ADLS, WASB

Scalable
Scalability with NameNode federation, allowing customers to scale to a billion files and thousands of nodes

Real-time database
SQL data access enhancements and improved query performance to eliminate the gap between low latency and high throughput queries

Security rich
Greater regulatory compliance, including General Data Protection Regulation (GDPR), through full chain of data custody, as well as fine-grained auditing of events

For more information

To learn more about HDP, visit the **IBM and Cloudera webpage** or **contact an IBM data management expert**.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
December 2019

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.

It is the user’s responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

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.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed, misappropriated or misused or can result in damage to or misuse of your systems, including for use in attacks on others. No IT system or product should be considered completely secure and no single product, service or security measure can be completely effective in preventing improper use or access. IBM systems, products and services are designed to be part of a lawful, comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM DOES NOT WARRANT THAT ANY SYSTEMS, PRODUCTS OR SERVICES ARE IMMUNE FROM, OR WILL MAKE YOUR ENTERPRISE IMMUNE FROM, THE MALICIOUS OR ILLEGAL CONDUCT OF ANY PARTY.