IBM

Hortonworks Data Platform

An open-architecture platform to manage data in motion and at rest

- Addresses a range of data-at-rest use cases

- Powers real-time customer applications

- Delivers robust analytics

Every business is now a data business. Data is your organization's future and its most valuable asset. The Hortonworks Data Platform (HDP) is a security-rich, enterprise-ready, open source Apache Hadoop distribution based on a centralized architecture (YARN). HDP addresses the needs of data at rest, powers real-time customer applications, and delivers robust analytics that help accelerate decision making and innovation.

The Hortonworks difference

HDP helps enterprises transform their businesses by unlocking the full potential of big data with the following benefits:

Open	Central	Interoperable	Enterprise ready
HDP is composed of numerous Apache Software Foundation (ASF) projects that enable enterprises to deploy, integrate and work with unprecedented volumes of structured and unstructured data. ASF's approach is to deliver enterprise-grade software that fosters innovation and prevents vendor lock-in.	YARN is the architectural center of open-enterprise Hadoop. It allocates resources among diverse applications that process data. YARN coordinates cluster-wide services for operations, data governance and security. YARN also maximizes data ingestion by enabling enterprises to analyze data to support diverse use cases. This process empowers Hadoop operators to confidently extend their big data assets to the largest possible audience in their organizations.	Its 100 percent open-source architecture enables HDP to be interoperable with a broad range of data center and business intelligence applications. HDP's interoperability helps minimize the expense and effort required to connect customers' IT infrastructures with HDP's data and processing capabilities. With HDP, customers can preserve their investment in existing IT architecture as they adopt Hadoop.	HDP is built for enterprises. Open-enterprise Hadoop provides consistent operations, with centralized management and monitoring of clusters through a single pane of glass. With HDP, security and governance is built into the platform. This feature helps provide a security-rich environment that's consistently administered across data access engines.

The Hortonworks Data Platform

HDP offers a security-rich, enterprise-ready open-source Hadoop distribution based on a centralized architecture. HDP addresses a range of data-at-rest use cases, powers real-time customer applications and delivers robust analytics that accelerate decision making and innovation.

Data management

The foundational components of HDP are Apache Hadoop YARN and the Hadoop Distributed File System (HDFS). While HDFS provides the scalable, fault-tolerant, cost-efficient storage for a big data lake, YARN provides the centralized architecture that enables organizations to process multiple workloads simultaneously. YARN also provides the resource management and pluggable architecture for enabling a wide variety of data access methods.

Data access

With YARN at its architectural center, HDP provides a range of processing engines that allow users to simultaneously interact with data in multiple ways. YARN enables a range of access methods to coexist in the same cluster against shared data sets. This feature avoids unnecessary and costly data silos. HDP enables multiple data processing engines that range from interactive structured query language (SQL) and real-time streaming to data science and batch processing to use data stored in a single platform.
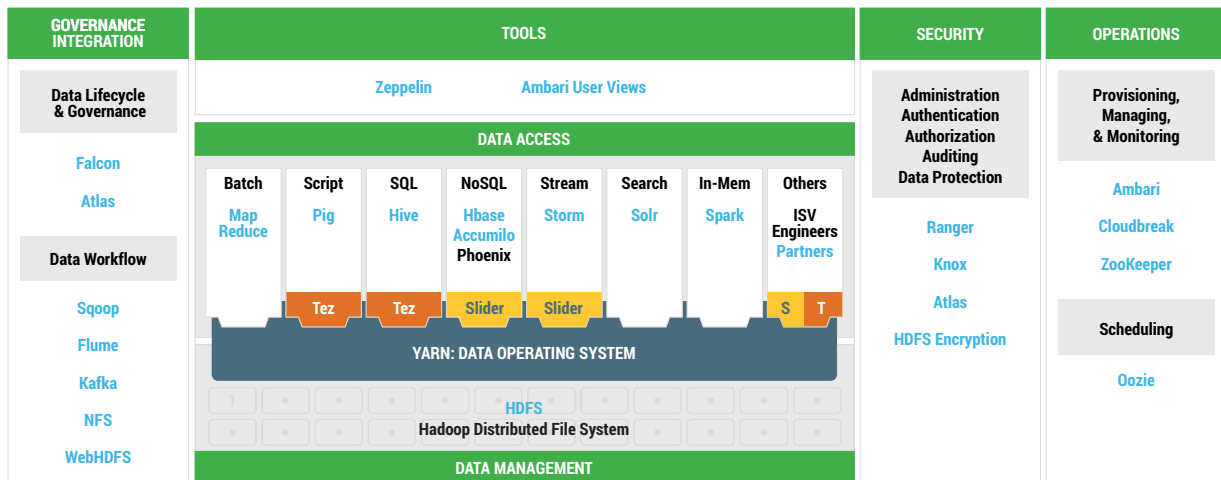

Figure 1: Next-generation Hadoop security

Security and governance

As organizations pursue Hadoop initiatives to capture new opportunities for data-driven insights, data governance and security requirements can pose a key challenge. In response to this challenge, the Data Governance Initiative (DGI), a consortium of cross-industry leaders, was created to address the need for an open-source governance solution to manage data classification, lineage, security and data lifecycle management.

Apache Atlas, created as part of DGI, empowers organizations to apply consistent data classification across the data ecosystem. Apache Ranger provides centralized security administration for Hadoop. By integrating Atlas with Ranger, Hortonworks empowers enterprises to institute dynamic access policies at runtime that proactively help prevent violations from occurring.

This integration enables enterprises to implement dynamic classification-based security policies. Ranger's centralized platform empowers data administrators to define security policy based on Atlas metadata tags or attributes. They can then apply this policy in real time to the entire hierarchy of data assets, including databases, tables and columns.

Security

A Hadoop-powered data lake can provide a robust foundation for a new generation of analytics and insight. It's important, however, to secure the data before launching or expanding a Hadoop initiative. By ensuring that data protection and governance are built into their big data environments, enterprises can use the full value of advanced analytics without exposing their businesses to new risks.

Governance

As organizations pursue Hadoop initiatives to capture new opportunities for data-driven insight, data governance requirements can pose a key challenge. The management of information to identify its value and enable effective control, security and compliance for customer and enterprise data is a core requirement for both traditional and big data architectures.

Operations

HDP Operations is designed to enable IT organizations to bring Hadoop online quickly by taking the guesswork out of the manual processes and replacing them with automated, preconfigured best practices, guided configurations and full operation control. HDP operations help simplify operation of distributed multiuser, multitenant and multidata access engines and manage HDP clusters at scale through an integrated web user interface or single pane of glass.

HDP uses Apache Ambari, an open-source management platform for provisioning, managing, monitoring and securing Hadoop clusters. Ambari removes the manual and often error-prone tasks associated with operating Hadoop. It also provides the necessary integration points to fit seamlessly into the enterprise.
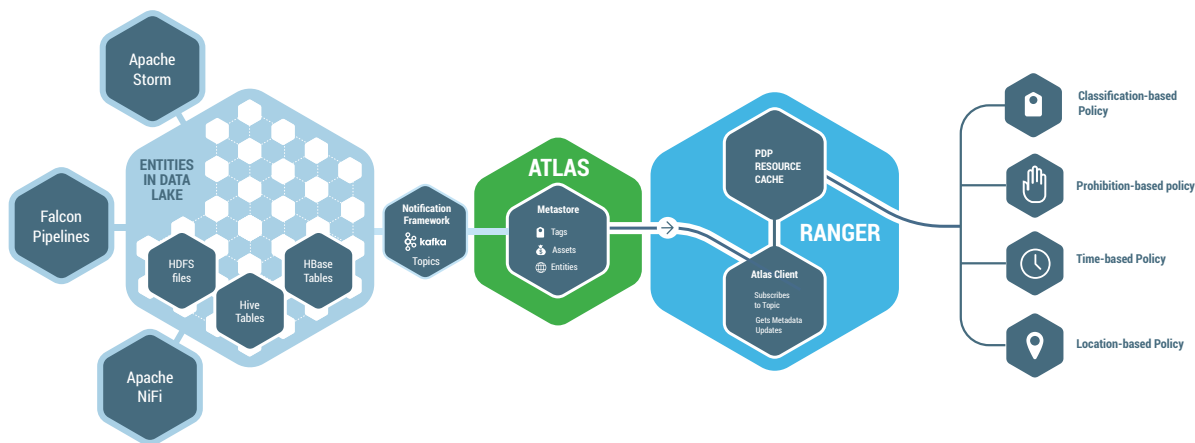

Figure 2: Next-generation Hadoop security

Deployment options

HDP offers a range of infrastructure choices to deploy an open and flexible data platform. Users have the flexibility to combine the infrastructure options that best suit their unique use cases.

On premises

Several organizations that have invested in data center infrastructure and managed services and are now considering Hadoop capabilities will find on-premise implementation to be a viable option. HDP is designed to be easily deployed on premises to integrate with existing data centers.

Cloud

HDP can be deployed in the cloud as part of Microsoft Azure HDInsight. Azure HDInsight is a managed service offering on the Microsoft Azure cloud, powered by HDP. This deployment option enables organizations to scale from terabytes to petabytes of data on demand by spinning up any number of nodes at any time. With HDInsight, enterprises can also connect their on-premises Hadoop clusters to the cloud.

Hybrid cloud and Cloudbreak

Cloudbreak is a solution for provisioning Hadoop clusters on a cloud infrastructure. As part of HDP, and powered by Apache Ambari, Cloudbreak helps enterprises simplify the provisioning of clusters in the cloud and optimize the use of cloud resources with elastic scaling. It's designed for customers that have an on-premises Hadoop deployment and want to set up clusters in the cloud with greater ease. With Cloudbreak, customers can choose their cloud provider of choice and let Cloudbreak configure the cluster in the cloud.

HDP for teams

Successful deployment of Hadoop in any organization depends on using existing skill sets and resources to adopt the big data architecture. HDP provides valuable tools and capabilities for every role on your big data team.

The data scientist

Apache Spark, part of HDP, plays an important role when it comes to data science. Data scientists commonly use machine learning, a set of techniques and algorithms that can learn from data. These algorithms are often iterative, and Spark's ability to cache the data in memory greatly accelerates the iterative data processing, making it an ideal processing engine for implementing such algorithms.

The business analyst

HDP provides business analysts with fast access to vast amounts of data through SQL on Hadoop interfaces provided by Apache Hive, Spark SQL and Apache Phoenix. With these interfaces, business analysts can use their favorite business intelligence and business analytics tools to create reports, visualizations, dashboards and scorecards to make more effective insight-driven decisions.

The developer

HDP provides a rich set of tools for application developers, such as Spark and Apache Zeppelin. HDP also features a set of native application program interfaces (APIs) that helps ease development: WebHDFS provides a REST interface to write, manipulate and delete HDFS files, while WebHCAT provides a critical point of integration to access metadata and schema for Hadoop data.

Hadoop operators

Using Ambari, Hadoop operators get the following benefits to streamline Hadoop operations:

- More flexible upgrades: Ambari enables a faster way to upgrade clusters by automating both maintenance and feature releases while the cluster is down.
- Simplified security operations: Service configurations for Ranger provide a continuation of the new user experience. Optional storage of Kerberos credentials and customizable security settings help simplify administration and provide a security-rich environment.
- Improved troubleshooting: Ambari provides a customizable metric widget graph display and the ability to export metrics to identify and respond to problems quickly.

Industries

HDP enables enterprises across a range of industries to add value to their core functions:

Financial services	Telecommunications
• Manage default risk • Improve customer cross-sell • Detect money laundering	• Analyze call detail records (CDRs) • Proactively service transmission infrastructure • Rationalize infrastructure investments • Develop new products and services
Retail	**Oil and gas**
• Build a 360° view of their customers • Localize and personalize consumer experiences • Manage supply chains effectively • Understand changes in brand sentiment through sentiment analysis • Optimize websites, campaigns and store layouts	• Monitor upstream production in remote locations • Slow decline curves • Proactively repair valuable equipment • Report on compliance with environmental health and safety regulations

Conclusion

Data is a fundamental tool to every business in every industry. It determines how you develop new products and services, operate efficiently and more. Hortonworks DataFlow (HDF) manages data in motion by helping to securely acquire and transport data to the Hortonworks Data Platform (HDP). HDP manages data at rest for virtually all types of data, with enterprise-grade governance, security and operations, helping to ensure your business remains competitive.

Why IBM?

IBM is an American multinational technology and consulting corporation, with its headquarters in Armonk, New York. IBM manufactures and markets computer hardware, middleware and software, and offers hosting and consulting services in areas ranging from mainframe computers to nanotechnology. IBM is also a major research organization, holding the record for most patents generated by a business, as of 2017, for 24 consecutive years.

For more information

To learn more about Hortonworks Data Platform, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/hadoop

Please Recycle