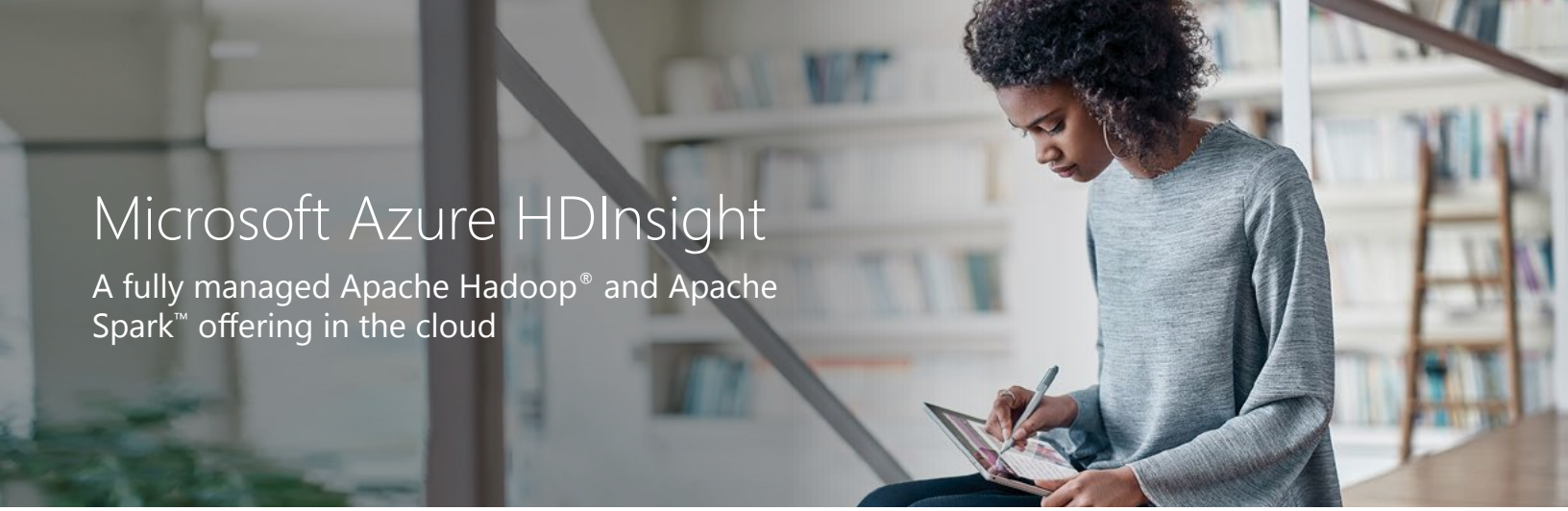


# Microsoft Azure HDInsight

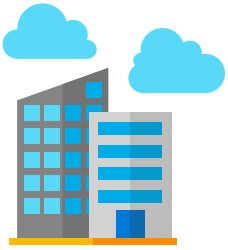
## A fully managed Apache Hadoop® and Apache Spark™ offering in the cloud

Microsoft and Hortonworks are working together to create Azure HDInsight—a fully managed Apache Hadoop and Apache Spark cloud service that has been hardened for the enterprise and made simpler for your users. As a managed Hadoop-as-a-service offering, HDInsight was designed to make Apache Hadoop and Apache Spark simple to use, with lower manageability costs and higher developer productivity. Customers have seen a 63% lower total cost of ownership (TCO) and 66% higher IT staff efficiencies by deploying HDInsight over on-premises Hadoop deployments.[1] It is also the only managed cloud Hadoop offering using Hortonworks, the leading contributor to Apache Hadoop, helping to ensure that Microsoft is supremely qualified to support your deployment through expert technical assistance and the ability to fix and commit code back to open source.

### Your enterprise can benefit from an Apache Hadoop deployment that is:

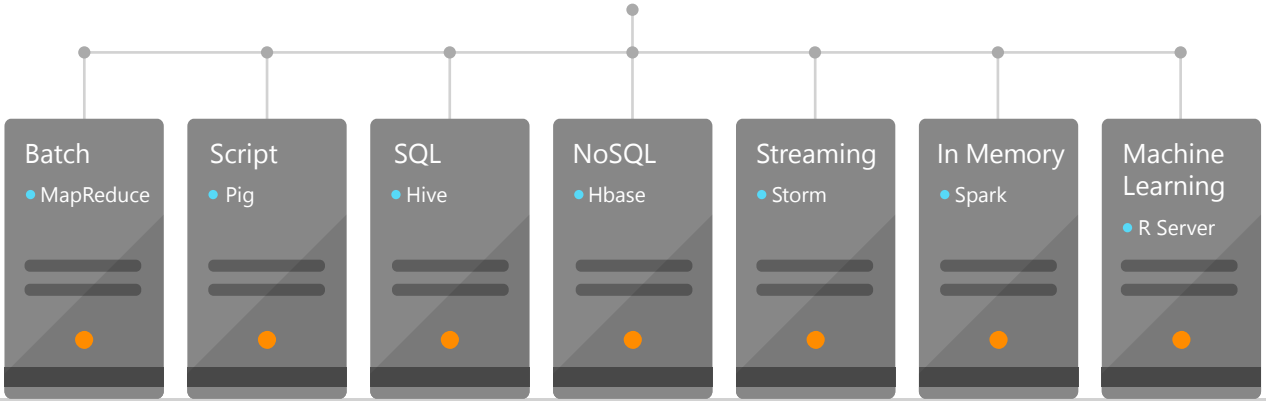
**Enterprise-ready**

**Easy for everybody**



**Hybrid over on-premises and cloud**

## Azure HDInsight


HORTONWORKS

Microsoft

# Customer challenges

### Large volumes

The volume of data today is exploding—expanding tenfold every five years. Much of this new data is driven by devices from the more than 1.2 billion people who are connected to the Internet worldwide, with an average of 4.3 connected devices per person by 2020.[2]

### Unstructured data

Today's data doesn't fit neatly into relational databases because up to 85% of new data is unstructured: text files, images, videos, sensor data, and weblogs. Customers need to manage all this data alongside their relational data in databases and data warehouses.[3]

### High velocity

Companies are using real-time data to change, build, and optimize their businesses as well as to sell, transact, and engage in dynamic, event-driven processes like market trading or Internet of Things (IoT) scenarios.

### Lack of skills

While Hadoop has become popular among organizations unlocking insights from data of all size, shape, and speed, Gartner has found the top challenge of Hadoop adoption to be obtaining the skills and capabilities needed to be productive.[4]

### Manageability

As organizations scale out the amount of data they are capturing, manageability of on-premises infrastructure becomes a challenge, requiring more IT staff to deploy and administer the solution.

Beth Israel Deaconess Medical Center

"It's part of our audit requirements that we keep data for seven years, and some information has to be retained for as long as thirty years. With HDInsight, we can store more data and query it as needed."

**–Don Wood, Beth Israel Deaconess Medical Center**

ASCRIBE

"With a solution based on SQL Server and the Azure HDInsight service, we can capture data written in plain English and use it to improve services. This will reinvent the way we work with medical records in the future."

**–Paul Henderson, Ascribe**

McKESSON

"Because we're on an elastic cloud with Azure, we don't have to worry about setting up infrastructure or whether we can sustain growth with the current capacity in our data centers."

**–Sujatha Bayyapureddy, McKesson**

# Why Azure HDInsight?

With SQL Server as the most widely deployed relational database management system, Microsoft understands what it takes to run your business at mission-critical scale. Azure HDInsight is a fully managed Apache Hadoop and Apache Spark cloud service that has been hardened by Microsoft for the enterprise and made simpler for your users.

# Solution advantages

## Enterprise-ready

### Scale elastically on demand

Azure HDInsight is an Apache Hadoop distribution powered by the cloud. This means that it handles virtually any amount of data, scaling from terabytes to petabytes on demand. Spin up any number of nodes at any time. Integration with Azure Data Lake Store provides a hyperscale repository that won't require application or code changes as the data increases.

### High availability

Azure HDInsight is architected for full redundancy and high availability. Both data and head nodes are replicated, allowing Microsoft to provide the highest availability guarantee in the industry with a 99.9% service level agreement, ensuring continuity and protection against catastrophic events. Azure also offers 24/7 enterprise support and cluster monitoring.

### Backed by industry leader

Azure HDInsight is the only managed cloud Hadoop offering using Hortonworks, the leading contributor to Apache Hadoop, helping to ensure that Microsoft is supremely qualified to support your deployment through expert technical assistance and the ability to fix and commit code back to open source.

### The most Apache projects and integration with Cortana Intelligence Suite

As a managed Hadoop service, Azure HDInsight has the most Apache projects in the cloud, including core Hadoop (HDFS, YARN, MapReduce, Hive, Tez, Pig, Sqoop, Oozie, Mahout, and Zookeeper) and advanced workloads (Spark, Storm, HBase, and R Server). Azure HDInsight is also part of the Cortana Intelligence Suite, providing an end-to-end solution, from information management to machine learning, dashboards, and cognitive services.

### Available in more datacenters

Azure is generally available in 24 regions (for 140 countries) around the world, and has announced plans for eight additional regions. This makes Azure HDInsight available in more data-center regions than any other cloud provider of Hadoop.

## Easy for everybody

### Easy for administrators

Deploy Hadoop in the cloud without time-consuming installation, setup, and administration—you can launch your first cluster in minutes. You'll always be on the latest version of Hadoop without doing version upgrades, OS patching, and security updates. Customers have seen 66% higher IT staff efficiencies by deploying HDInsight instead of on-premises deployments.[5]

### Easy for developers

HDInsight has focused on making developers productive with the deepest set of tooling available for Hadoop today. This includes extensive Visual Studio integration, allowing you to author, debug, troubleshoot, and submit Hive scripts, YARN logs, or Storm topologies. Integration with IntelliSense in Visual Studio provides a modern development experience, including keyword completion, syntax highlighting, goto definition/reference, and auto-format. To make development on Spark easier, we introduced IntelliJ Spark Tooling, which gives developers native authoring support for Scala and Java, local testing, remote debugging, and the ability to submit Spark applications to the Azure cloud.

### Easy for data scientists

HDInsight provides out-of-the-box integration with Jupyter (iPython), the most popular open source notebook for data scientists. Unlike other managed Spark offerings that might require you to install your own notebooks, we worked with the Jupyter OSS community to enhance the kernel to allow Spark execution through a REST endpoint. As a result, Jupyter notebooks are now accessible within HDInsight out-of-the-box.

### Easy for business analysts

Business analysts can interact with Spark through their favorite business intelligence tool, including Power BI, Tableau, SAP Lumira, and QlikView. New in Power BI is a streaming connector that integrates with Spark, allowing you to publish real-time events from Spark Streaming directly to Power BI.

## Hybrid across on-premises and cloud

Because Azure HDInsight uses Hortonworks Data Platform, it makes hybrid deployments easy over on-premises and cloud. By addressing both environments with a common platform, Microsoft is guiding users on a path that delivers the advantages of the public cloud while respecting the complexities of today's on-premises large data initiatives. When customers want to move code or projects from on-premises to the cloud, it's done with a few clicks and within a few minutes without buying hardware or hiring specialized operations teams typically associated with big data infrastructure.

## Validated

Forrester recently recognized Microsoft Azure as a leader in its Big Data Hadoop Cloud Solutions.[6] Forrester notes that leaders have the most comprehensive, scalable, and integrated platforms.[7] Microsoft specifically was called out for having a cloud-first strategy that is paying off.[8]

IDC interviewed Azure HDInsight customers and found that these organizations on average will achieve five-year discounted benefits worth $255,000 per terabyte of data in their Hadoop environment and an average five-year return on investment (ROI) of 418%.[9] In addition, IDC's comparison of the costs of using Azure HDInsight and Hadoop deployed on-premises showed that HDInsight has a 63% lower average TCO on a per-terabyte basis.[10]

## Hortonworks and Microsoft can help your enterprise optimize insights from data

Hortonworks and Microsoft have come together to transform data within your organization into intelligent action, to help you unlock new business insights, increase efficiency, and reduce costs.

Learn more about Azure HDInsight at **hortonworks.com/HDInsight** or email **HDI@hortonworks.com** to schedule a consultation.


1. Matthew Marden and Carl Olofson, *The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight*, International Data Corporation (IDC), 2015.
2. http://www.mobileworldlive.com/featured-content/home-banner/connected-devices-to-hit-4-3-per-person-by-2020-report/
3. https://enterprisetechnologyconsultant.wordpress.com/2013/02/17/semi-structured-or-unstructured-data-metadata/
4. Merv Adrian and Nick Heudecker, *Survey Analysis: Hadoop Adoption Drivers and Challenges*, Gartner, 2015.
5. Marden and Olofson
6. Mike Gualtieri and Noel Yuhanna, *The Forrester Wave™: Big Data Hadoop Cloud Solutions*, Q2 2016, Forrester, 2016.
7. Ibid.
8. Ibid.
9. Marden and Olofson
10. Ibid.