



Hadoop

What it is and why it matters

Hadoop is an open-source software framework for storing and processing big data in a distributed fashion on large clusters of commodity hardware. Essentially, it accomplishes two tasks: massive data storage and faster processing.

For starters, let's take a quick look at some of those terms and what they mean.

- **Open-source software.** Open source software differs from commercial software due to the broad and open network of developers that create and manage the programs. Traditionally, it's free to download, use and contribute to, though more and more commercial versions of Hadoop are becoming available.
- **Framework.** In this case, it means everything you need to develop and run your software applications is provided – programs, tool sets, connections, etc.
- **Distributed.** Data is divided and stored across multiple computers, and computations can be run in parallel across multiple connected machines.
- **Massive storage.** The Hadoop framework can store huge amounts of data by breaking the data into blocks and storing it on clusters of lower-cost commodity

hardware.

- **Faster processing.** How? Hadoop processes large amounts of data in parallel across clusters of tightly connected low-cost computers for quick results.

With the ability to economically store and process any kind of data (not just numerical or structured data), organizations of all sizes are taking cues from the corporate web giants that have used Hadoop to their advantage (Google, Yahoo, Etsy, eBay, Twitter, etc.), and they're asking "What can Hadoop do for me?"

How did Hadoop get here?

As the World Wide Web grew at a dizzying pace in the late 1990s and early 2000s, search engines and indexes were created to help people find relevant information amid all of that text-based content. During the early years, search results were returned by humans. It's true! But as the number of web pages grew from dozens to millions, automation was required. Web crawlers were created, many as university-led research projects, and search engine startups took off (Yahoo, AltaVista, etc.).

One such project was Nutch – an open-source web search engine – and the brainchild of Doug Cutting and Mike Cafarella. Their goal was to invent a way to return web search results faster by distributing data and calculations across different computers so multiple tasks could be accomplished simultaneously. Also during this time, another search engine project called Google was in progress. It was based on the same concept – storing and processing data in a distributed, automated way so that more relevant web search results could be returned faster.

In 2006, Cutting joined Yahoo and took with him the Nutch project as well as ideas based on Google's early work with automating distributed data storage and processing. The Nutch project was divided. The web crawler portion remained as Nutch. The distributed computing and processing portion became Hadoop (named after Cutting's son's toy elephant). In 2008, Yahoo released Hadoop as an open-source project, and, today Hadoop's framework and family of technologies are managed and maintained by the non-profit Apache Software Foundation (ASF), a global community of software developers and contributors.

Why is Hadoop important?

Since its inception, Hadoop has become one of the most talked about technologies. Why? One of the top reasons (and why it was invented) is its ability to handle huge amounts of data – any kind of data – quickly. With volumes and varieties of data growing each day, especially from social media and automated sensors, that's a key consideration for most organizations. Other reasons include:

- **Low cost.** The open-source framework is free and uses commodity hardware to store large quantities of data.
- **Computing power.** Its distributed computing model can quickly process very large volumes of data. The more computing nodes you use, the more processing power you have.
- **Scalability.** You can easily grow your system simply by adding more nodes. Little administration is required.
- **Storage flexibility.** Unlike traditional relational databases, you don't have to preprocess data before storing it. And that includes unstructured data like text, images and videos. You can store as much data as you want and decide how to use it later.
- **Inherent data protection and self-healing capabilities.** Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. And it automatically stores multiple copies of all data.

➤ Video Timeout

Why should organizations care about Hadoop? Ryan Schmiedl, Senior Director at SAS, explains.



What's in Hadoop?

Hadoop components have funny names, which is sort of understandable knowing that “Hadoop” was the name of a yellow toy elephant owned by the son of one of its inventors. Here's a quick rundown on names you may hear. Currently three core components are included with your basic download from the Apache Software Foundation.

- **HDFS** – the Java-based distributed file system that can store all kinds of data without prior organization.
- **MapReduce** – a software programming model for processing large sets of data in parallel.
- **YARN** – a resource management framework for scheduling and handling resource requests from distributed applications.

Other components that have achieved top-level Apache project status and are available include:

- **Pig** – a platform for manipulating data stored in HDFS. It consists of a compiler for MapReduce programs and a high-level language called Pig Latin. It provides a way to perform data extractions, transformations and loading, and basic analysis without having to write MapReduce programs.
- **Hive** – a data warehousing and SQL-like query language that

presents data in the form of tables. Hive programming is similar to database programming. (It was initially developed by Facebook.)

- **HBase** – a nonrelational, distributed database that runs on top of Hadoop. HBase tables can serve as input and output for MapReduce jobs.
- **Zookeeper** – an application that coordinates distributed processes.
- **Ambari** – a web interface for managing, configuring and testing Hadoop services and components.
- **Flume** – software that collects, aggregates and moves large amounts of streaming data into HDFS.
- **Sqoop** – a connection and transfer mechanism that moves data between Hadoop and relational databases.
- **Oozie** – a Hadoop job scheduler.

In addition, commercial software distributions of Hadoop are growing. Two of the most prominent (Cloudera and Hortonworks) are startups formed by the framework's inventors. And there are plenty of others entering the Hadoop sphere. With distributions from software vendors, you pay for their version of the framework and receive additional software components, tools, training, documentation and other services.

How does data get into Hadoop?

There are numerous ways to get data into Hadoop. Here are just a few:

- You can load files to the file system using simple Java commands, and HDFS takes care of making multiple copies of data blocks and distributing those blocks

over multiple nodes in Hadoop.

- If you have a large number of files, a shell script that will run multiple “put” commands in parallel will speed up the process. You don’t have to write MapReduce code.
- Create a cron job to scan a directory for new files and “put” them in HDFS as they show up. This is useful for things like downloading email at regular intervals.
- Mount HDFS as a file system and simply copy or write files there.
- Use Sqoop to import structured data from a relational database to HDFS, Hive and HBase. It can also extract data from Hadoop and export it to relational databases and data warehouses.
- Use Flume to continuously load data from logs into Hadoop.
- Use third-party vendor connectors (like [SAS/ACCESS®](#)).

Then what happens?

Going beyond its original goal of searching millions (or billions) of web pages and returning relevant results, many organizations are looking to Hadoop as their next big data platform. Here are some of the more popular uses for the framework today.

1. **Low-cost storage and active data archive.** The modest cost of commodity hardware makes Hadoop useful for storing and combining big data such as transactional, social media, sensor, machine, scientific, click streams, etc. The low-cost storage lets you keep information that is not currently critical but could become useful later for business analytics.
2. **Staging area for a data warehouse and analytics store.** One of the most prevalent uses is to stage large amounts of raw data for loading into an enterprise data warehouse (EDW) or an analytical store for activities such as advanced analytics, query and reporting, etc. Organizations are looking at Hadoop to handle new types of data (e.g., unstructured), as well as to offload some historical data from their EDWs.
3. **Data lake.** Hadoop is often used to store large amounts of data without the constraints introduced by schemas commonly found in the SQL-based world. It is used as a low-cost compute-cycle platform that supports processing ETL and data quality jobs in parallel using hand-coded or commercial data management technologies. Refined results can then be passed to other systems (e.g., EDWs, analytic marts) as needed.
4. **Sandbox for discovery and analysis.** Because Hadoop was designed to deal

with volumes of data in a variety of shapes and forms, it can enable analytics. Big data analytics on Hadoop can help run your organization more efficiently, uncover new opportunities and derive next-level competitive advantage. The sandbox setup provides a quick and perfect opportunity to innovate with minimal investment.

Certainly Hadoop provides an economical platform for storing and processing large and diverse data. The next logical step is to transform and manage the diverse data and use analytics to quickly identify undiscovered insights.

What challenges may be encountered?

First of all, MapReduce is not a good match for all problems. It's good for simple requests for information and problems that can be broken up into independent units. But it is inefficient for iterative and interactive analytic tasks. MapReduce is file-intensive. Because the nodes don't intercommunicate except through sorts and shuffles, iterative algorithms require multiple map-shuffle/sort-reduce phases to complete. This creates multiple files between MapReduce phases and is very inefficient for advanced analytic computing.

Second, there's a talent gap. Because it is a relatively new technology, it is difficult to find entry-level programmers who have sufficient Java skills to be productive with MapReduce. This talent gap is one reason distribution providers are racing to put relational (SQL) technology on top of Hadoop. It is much easier to find programmers with SQL skills than MapReduce skills. And, Hadoop administration seems part art and part science, requiring low-level knowledge of operating systems, hardware and Hadoop kernel settings.

Another challenge centers around the fragmented data security issues in Hadoop, though new tools and technologies are surfacing. The [Kerberos](#) authentication protocol is a great step forward for making Hadoop environments secure. And, Hadoop does not have easy-to-use, full-feature tools for data management, data cleansing, governance and metadata. Especially lacking are tools for data quality and standardization.

Big Data, Hadoop and SAS

SAS support for big data implementations, including Hadoop, centers on a singular

goal – helping you know more, faster, so you can make better decisions. Regardless of how you use the technology, every project should go through an iterative and continuous improvement cycle. And that includes data preparation and management, data visualization and exploration, model development, model deployment and monitoring.

SAS capabilities span this entire analytics (data-to-decision) life cycle. From data aggregation to powerful analytics – you can derive insights and quickly turn your big Hadoop data into bigger opportunities.

Because SAS is focused on analytics, not storage, we offer a flexible approach to choosing hardware and database vendors. We work with you to deploy the right mix of technologies, including the ability to deploy Hadoop with other data warehouse technologies.

And as always, remember that the success of any project is determined by the value it brings. So metrics built around revenue generation, margins, risk reduction and process improvements will help small pilot projects gain wider acceptance and garner more interest from other departments. Many organizations are looking at how they can implement a project or two in Hadoop, with plans to add more in the future.

Who does SAS partner with, and who are the players?

Cloudera is the most widely known and used commercial distribution of Hadoop, followed by Hortonworks, Pivotal, IBM, MapR and a growing number of other providers. At SAS, Cloudera and Hortonworks are the primary distributions used for development and testing of SAS software, and the ones we've found our customers are most interested in.



Big Data Insights

Get more insights on big data including articles, research and other hot topics.

Learn more about Hadoop

- [Fast and Furious: Big Data Analytics Meets Hadoop](#) (white paper)
- [Bringing the Power of SAS to Hadoop](#) (white paper)
- [Eight Considerations for Using Big Data with Hadoop](#) (TDWI Checklist Report)
- [Eyes Wide Open: Open Source Analytics Software](#) (IIA Research Brief)
- [Three ways to use a Hadoop data platform without throwing out your data warehouse](#) (blog)

Fun Fact:



"Hadoop" was the name of a yellow toy elephant owned by the son of one of its inventors.

Hadoop Solutions From SAS

Access and Manage Hadoop Data



SAS® Data Loader for Hadoop

Manage big data on your own terms – and avoid burdening IT – with self-service data integration.



SAS/ACCESS® Interface to Hadoop

Get out-of-the-box connectivity between SAS and Hadoop, via Hive.



SAS/ACCESS® Interface to Impala

Gain low-latency response times and work faster with this out-of-the-box solution connecting SAS and Impala.



SAS® Data Management

Ensure better, more reliable data integrated from any source.



SAS® Federation Server

Centralize and streamline business views of your data without moving it.



Base SAS®

Use a flexible programming language for powerful data access, transformation and reporting.

Explore and Visualize



SAS® Visual Analytics

Visually explore all data, discover new patterns and publish reports to the web and mobile devices.

For the Data Scientist



SAS® In-Memory Statistics

Find insights in big data with a single environment that moves you quickly through each phase of the analytical life cycle.

Analyze and Model



SAS® Visual Statistics

Create and modify predictive models faster than ever using a visual interface and in-memory processing.



SAS® High-Performance Data Mining

Quickly analyze big data to derive more accurate insights and make timely business decisions.



SAS® High-Performance Text Mining

Quickly discover categories and themes in huge volumes of unstructured data.



SAS® High-Performance Statistics

Apply advanced statistics to huge volumes of data using all available processing power.



SAS® High-Performance Optimization

Model and solve very large, cumbersome optimization problems with greater efficiency and effectiveness.

Deploy and Integrate



SAS® Scoring Accelerator

Automate data scoring processes within the database to improve model performance and get faster results.



SAS® Event Stream Processing Engine

Gain immediate analytic insights from real-time data streaming into your organization.

Want more insights?



Data Management

Get more insights on data management – articles, research, videos and more.




Analytics

Connect with the latest insights on analytics through related articles and research.



Marketing

Explore insights from marketing movers and shakers on a variety of timely topics.

Ready to learn more about our
Hadoop & big data solutions? 

[Privacy Statement](#) | [Terms of Use](#) | © SAS Institute Inc. All Rights Reserved