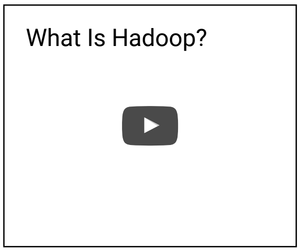
Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a

Hadoop makes it possible to run applications on systems with thousands of commodity hardware nodes, and to handle thousands of terabytes of data. Its distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating in case of a node failure. This approach lowers the risk of catastrophic system failure and unexpected data loss, even if a significant number of nodes become inoperative. Consequently, Hadoop quickly emerged as a foundation for big data processing tasks, such as scientific analytics, business and sales planning, and processing enormous volumes of sensor data, including from internet of things sensors.

Hadoop was created by computer scientists Doug Cutting and Mike Cafarella in 2006 to support distribution for the Nutch search engine. It was inspired by Google's MapReduce, a software framework in which an application is broken down into numerous small parts. Any of these parts, which are also called fragments or blocks, can be run on any node in the cluster. After years of development within the open source community, Hadoop 1.0 became publically available in November 2012 as part of the Apache project sponsored by the Apache Software Foundation.

Since its initial release, Hadoop has been continuously developed and updated. The second iteration of Hadoop (Hadoop 2) improved resource management and scheduling. It features a high-availability file-system option and support for Microsoft Windows and other components to expand the framework's versatility for data processing and analytics.



What is Hadoop?

Organizations can deploy Hadoop components and supporting software packages in their local data center. However, most big data projects depend on short-term use of substantial computing resources. This type of usage is best-suited to highly scalable public cloud services, such as Amazon Web Services (AWS), Google Cloud Platform and Microsoft Azure. Public cloud providers often support Hadoop components through basic services, such as AWS Elastic Compute Cloud and Simple Storage Service instances. However, there are also services tailored specifically for Hadoop-type tasks, such as AWS Elastic MapReduce, Google Cloud Dataproc and Microsoft Azure HDInsight.

Hadoop modules and projects

As a software framework, Hadoop is composed of numerous functional modules. At a minimum, Hadoop uses Hadoop Common as a kernel to provide the framework's essential libraries. Other components include Hadoop Distributed File System (HDFS), which is capable of storing data across thousands of commodity servers to achieve high bandwidth between nodes; Hadoop Yet Another Resource Negotiator (YARN), which provides resource management and scheduling for user applications; and Hadoop MapReduce, which provides the programming model used to tackle large distributed data processing -- mapping data and reducing it to a result.

Hadoop also supports a range of related projects that can complement and extend Hadoop's basic capabilities. Complementary software packages include:

- **Apache Flume**. A tool used to collect, aggregate and move huge amounts of streaming data into HDFS.
- **Apache HBase**. An open source, nonrelational, distributed database;
- **Apache Hive**. A data warehouse that provides data summarization, query and analysis;
- **Cloudera Impala**. A massively parallel processing database for Hadoop, originally created by the software company Cloudera, but now released as open source software;
- **Apache Oozie**. A server-based workflow scheduling system to manage Hadoop jobs;
- **Apache Phoenix**. An open source, massively parallel processing, relational database engine for Hadoop that is based on Apache HBase;
- **Apache Pig**. A high-level platform for creating programs that run on Hadoop;
- **Apache Sqoop**. A tool to transfer bulk data between Hadoop and structured data stores, such as relational databases;
- **Apache Spark**. A fast engine for big data processing capable of streaming and supporting SQL, machine learning and graph processing;
- **Apache Storm**. An open source data processing system; and
- **Apache ZooKeeper**. An open source configuration, synchronization and naming registry service for large distributed systems.

Marga ks:

How has Hadoop affected the big data initiatives in your organization?

Join the Discussion

This was last updated in September 2016

Next Steps

Learn how a low-cost, high-performance computing framework like Hadoop can help an organization's employees improve the way they manage massive amounts of data.

To address specific use cases, Hadoop vendors bundle Hadoop distributions with different levels of support.

To help you decide which vendor distribution subscription best fits your needs, take a look at our in-depth product descriptions of the six leading subscriptions: Amazon Web Services Elastic MapReduce, Cloudera CDH, Hortonworks HDP, IBM BigInsights, Microsoft Azure HDInsight and MapR.

Continue Reading About Hadoop

- Learn how Hadoop can help you manage big data
- Explore different Hadoop distributions for big data projects
- Use Spark on AWS to optimize big data workloads
- IT pros see benefits from Hadoop workflow automation