What is Hadoop and What is it Used For? | Google Cloud

Hadoop history

Hadoop has its origins in the early era of the World Wide Web. As the Web grew to millions and then billions of pages, the task of searching and returning search results became one of the most prominent challenges. Startups like Google, Yahoo, and AltaVista began building frameworks to automate search results. One project called Nutch was built by computer scientists Doug Cutting and Mike Cafarella based on Google's early work on MapReduce (more on that later) and Google File System. Nutch was eventually moved to the Apache open source software foundation and was split between Nutch and Hadoop. Yahoo, where Cutting began working in 2006, open sourced Hadoop in 2008.

While Hadoop is sometimes referred to as an acronym for High Availability Distributed Object Oriented Platform, it was originally named after Cutting's son's toy elephant.

Hadoop defined

Hadoop is an open source framework based on Java that manages the storage and processing of large amounts of data for applications. Hadoop uses distributed storage and parallel processing to handle big data and analytics jobs, breaking workloads down into smaller workloads that can be run at the same time.

Four modules comprise the primary Hadoop framework and work collectively to form the Hadoop ecosystem:

Hadoop Distributed File System (HDFS): As the primary component of the Hadoop ecosystem, HDFS is a distributed file system in which individual Hadoop nodes operate on data that resides in their local storage. This removes network latency, providing high-throughput access to application data. In addition, administrators don't need to define schemas up front.

Yet Another Resource Negotiator (YARN): YARN is a resource-management platform responsible for managing compute resources in clusters and using them to schedule users' applications. It performs scheduling and resource allocation across the Hadoop system.

MapReduce: MapReduce is a programming model for large-scale data processing. In the MapReduce model, subsets of larger datasets and instructions for processing the subsets are dispatched to multiple different nodes, where each subset is

processed by a node in parallel with other processing jobs. After processing the results, individual subsets are combined into a smaller, more manageable dataset.

Hadoop Common: Hadoop Common includes the libraries and utilities used and shared by other Hadoop modules.

Beyond HDFS, YARN, and MapReduce, the entire Hadoop open source ecosystem continues to grow and includes many tools and applications to help collect, store, process, analyze, and manage big data. These include Apache Pig, Apache Hive, Apache HBase, Apache Spark, Presto, and Apache Zeppelin.

How does Hadoop work?

Hadoop allows for the distribution of datasets across a cluster of commodity hardware. Processing is performed in parallel on multiple servers simultaneously.

Software clients input data into Hadoop. HDFS handles metadata and the distributed file system. MapReduce then processes and converts the data. Finally, YARN divides the jobs across the computing cluster.

All Hadoop modules are designed with a fundamental assumption that hardware failures of individual machines or racks of machines are common and should be automatically handled in software by the framework.

What are the benefits of Hadoop?

Scalability

Hadoop is important as one of the primary tools to store and process huge amounts of data quickly. It does this by using a distributed computing model which enables the fast processing of data that can be rapidly scaled by adding computing nodes.

Low cost

As an open source framework that can run on commodity hardware and has a large ecosystem of tools, Hadoop is a low-cost option for the storage and management of big data.

Flexibility

Hadoop allows for flexibility in data storage as data does not require preprocessing before storing it which means that an organization can store as much data as they like and then utilize it later.

Resilience

As a distributed computing model, Hadoop allows for fault tolerance and system resilience, meaning if one of the hardware nodes fail, jobs are redirected to other nodes. Data stored on one Hadoop cluster is replicated across other nodes within the system to fortify against the possibility of hardware or software failure.

What are the challenges of Hadoop?

MapReduce complexity and limitations

As a file-intensive system, MapReduce can be a difficult tool to utilize for complex jobs, such as interactive analytical tasks. MapReduce functions also need to be written in Java and can require a steep learning curve. The MapReduce ecosystem is quite large, with many components for different functions that can make it difficult to determine what tools to use.

Security

Data sensitivity and protection can be issues as Hadoop handles such large datasets. An ecosystem of tools for authentication, encryption, auditing, and provisioning has emerged to help developers secure data in Hadoop.

Governance and management

Hadoop does not have many robust tools for data management and governance, nor for data quality and standardization.

Talent gap

Like many areas of programming, Hadoop has an acknowledged talent gap. Finding developers with the combined requisite skills in Java to program MapReduce, operating systems, and hardware can be difficult. In addition, MapReduce has a steep learning curve, making it hard to get new programmers up to speed on its best practices and ecosystem.

Why is Hadoop important?

Research firm IDC estimated that 62.4 zettabytes of data were created or replicated in 2020, driven by the Internet of Things, social media, edge computing, and data created in the cloud. The firm forecasted that data growth from 2020 to 2025 was expected at 23% per year. While not all that data is saved (it is either deleted after consumption or overwritten), the data needs of the world continue to grow.

Hadoop tools

Hadoop has a large ecosystem of open source tools that can augment and extend the capabilities of the core module. Some of the main software tools used with Hadoop include:

Apache Hive: A data warehouse that allows programmers to work with data in HDFS using a query language called HiveQL, which is similar to SQL

Apache HBase: An open source non-relational distributed database often paired with Hadoop

Apache Pig: A tool used as an abstraction layer over MapReduce to analyze large sets of data and enables functions like filter, sort, load, and join

Apache Impala: Open source, massively parallel processing SQL query engine often used with Hadoop

Apache Sqoop: A command-line interface application for efficiently transferring bulk data between relational databases and Hadoop

Apache ZooKeeper: An open source server that enables reliable distributed coordination in Hadoop; a service for, "maintaining configuration information, naming, providing distributed synchronization, and providing group services"

Apache Oozie: A workflow scheduler for Hadoop jobs

What is Apache Hadoop used for?

Here are some common uses cases for Apache Hadoop:

Analytics and big data

A wide variety of companies and organizations use Hadoop for research, production data processing, and analytics that require processing terabytes or petabytes of big data, storing diverse datasets, and data parallel processing.

Data storage and archiving

As Hadoop enables mass storage on commodity hardware, it is useful as a low-cost storage option for all kinds of data, such as transactions, click streams, or sensor and machine data.

Data lakes

Since Hadoop can help store data without preprocessing, it can be used to complement to data lakes, where large amounts of unrefined data are stored.

Marketing analytics

Marketing departments often use Hadoop to store and analyze customer relationship management (CRM) data.

Risk management

Banks, insurance companies, and other financial services companies use Hadoop to build risk analysis and management models.

AI and machine learning

Hadoop ecosystems help with the processing of data and model training operations for machine learning applications.

Related products and services

Companies often choose to run Hadoop clusters on public, or hybrid cloud resources versus on-premises hardware to gain flexibility, availability, and cost control. Many cloud solution providers offer fully managed services for Hadoop. With this kind of prepackaged service for cloud-first Hadoop, operations that used to take hours or days can be completed in seconds or minutes, with companies paying only for the resources used.

On Google Cloud, Dataproc is a fast, easy-to-use, and fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, integrated, most cost-effective way. It fully integrates with other Google Cloud services that meet critical security, governance, and support needs, allowing you to gain a complete and powerful platform for data processing, analytics, and machine learning.

Big data analytics tools from Google Cloud—such as Dataproc, BigQuery, Vertex AI Workbench, and Dataflow—can enable you to build context-rich applications, build new analytics solutions, and turn data into actionable insights.