 Apache Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Learn more » (http://hadoop.apache.org/docs/stable)	Download » (/releases.html)

Getting started » (http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html)

Latest news

Apache Hadoop (/)	2019 May 7

News (/news.html)	2019 May 7

Ozone 0.4.0-alpha is released (/news/2019-05-07-ozone-0.4.0-alpha.html)	2019 May 7

Ozone 0.4.0 alpha version supports kerberos and transparent data encryption. This is first secure Ozone release. It is compatible with apache Spark, Hive and Yarn.

For more information check the ozone site (https://hadoop.apache.org/ozone/release/0.4.0-alpha/).

Release 3.1.2 available (/release/3.1.2.html)	2019 Feb 6

This is the second stable release of Apache Hadoop 3.1 line. It contains 325 bug fixes, improvements and enhancements since 3.1.1.

Users are encouraged to read the overview of major changes (/docs/r3.1.2/index.html) since 3.0.x. For details of 325 bug fixes, improvements, and other enhancements since the previous 3.1.1 release, please check release notes (http://hadoop.apache.org/docs/r3.1.2/hadoop-project-dist/hadoop-common/release/3.1.2/RELEASENOTES.3.1.2.html) and changelog (http://hadoop.apache.org/docs/r3.1.2/hadoop-project-dist/hadoop-common/release/3.1.2/CHANGES.3.1.2.html)

Release 3.2.0 available (/release/3.2.0.html)	2019 Jan 16

This is the first stable release of Apache Hadoop 3.2 line. It contains 1092 bug fixes, improvements and enhancements since 3.1.0.

Users are encouraged to read the overview of major changes (/docs/r3.2.0/index.html) since 3.1.0. For details of 1092 bug fixes, improvements, and other enhancements since the previous 3.1.0 release, please check release notes (http://hadoop.apache.org/docs/r3.2.0/hadoop-project-dist/hadoop-common/release/3.2.0/RELEASENOTES.3.2.0.html) and changelog (http://hadoop.apache.org/docs/r3.2.0/hadoop-project-dist/hadoop-common/release/3.2.0/CHANGELOG.3.2.0.html) detail the changes since 3.1.0.

Release archive → (/release.html)

News archive → (/news.html)

Modules

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.
- **Hadoop Ozone (https://hadoop.apache.org/ozone/)**: An object store for Hadoop.
- **Hadoop Submarine (https://hadoop.apache.org/submarine/)**: A machine learning engine for Hadoop.

Who Uses Hadoop?

A wide variety of companies and organizations use Hadoop for both research and production. Users are encouraged to add themselves to the Hadoop PoweredBy wiki page (https://wiki.apache.org/hadoop/PoweredBy).

Related projects

Other Hadoop-related projects at Apache include:

- **Ambari™** (https://ambari.apache.org): A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™** (https://avro.apache.org): A data serialization system.
- **Cassandra™** (https://cassandra.apache.org/): A scalable multi-master database with no single points of failure.
- **Chukwa™** (https://chukwa.apache.org/): A data collection system for managing large distributed systems.
- **HBase™** (https://hbase.apache.org/): A scalable, distributed database that supports structured data storage for large tables.
- **Hive™** (https://hive.apache.org/): A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™** (https://mahout.apache.org/): A Scalable machine learning and data mining library.
- **Pig™** (https://pig.apache.org): A high-level data-flow language and execution framework for parallel computation.
- **Spark™** (https://spark.apache.org): A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez™** (https://tez.apache.org): A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
- **ZooKeeper™** (https://zookeeper.apache.org): A high-performance coordination service for distributed applications.

Privacy policy (/privacy_policy.html)