Welcome to Apache™ Hadoop®!

Table of contents

1. What Is 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.

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.

Other Hadoop-related projects at Apache include:

- **Ambari™**: 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™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A Scalable machine learning and data mining library.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.
- **Spark™**: 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™**: 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™**: A high-performance coordination service for distributed applications.

2. Getting Started

To get started, begin here:

1. Learn about Hadoop by reading the documentation.
2. Download Hadoop from the release page.
3. Discuss Hadoop on the mailing list.

3. Download Hadoop

Please head to the releases page to download a release of Apache Hadoop.

4. 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.

5. News

5.1. 18 November, 2014: release 2.6.0 available

Apache Hadoop 2.6.0 contains a number of significant enhancements such as:

- Hadoop Common
  - Key management server (beta)
  - Credential provider (beta)
- Hadoop HDFS
  - Heterogeneous Storage Tiers - Phase 2
    - Application APIs for heterogeneous storage
    - SSD storage tier
    - Memory as a storage tier (beta)
  - Support for Archival Storage
  - Transparent data at rest encryption (beta)

- Operating secure DataNode without requiring root access
- Hot swap drive: support add/remove data node volumes without restarting data node (beta)
- AES support for faster wire encryption

- Hadoop YARN
  - Support for long running services in YARN
    - Service Registry for applications
  - Support for rolling upgrades
    - Work-preserving restarts of ResourceManager
    - Container-preserving restart of NodeManager
  - Support node labels during scheduling
  - Support for time-based resource reservations in Capacity Scheduler (beta)
  - Global, shared cache for application artifacts (beta)
  - Support running of applications natively in Docker containers (alpha)

Full information about this milestone release is available at Hadoop Releases.

5.2. 19 November, 2014: release 2.5.2 available

Full information about this milestone release is available at Hadoop Releases.

5.3. 12 September, 2014: release 2.5.1 available

Full information about this milestone release is available at Hadoop Releases.

5.4. 11 August, 2014: release 2.5.0 available

Full information about this milestone release is available at Hadoop Releases.

5.5. 30 June, 2014: release 2.4.1 available

Full information about this milestone release is available at Hadoop Releases.

5.6. 27 June, 2014: release 0.23.11 available

Full information about this milestone release is available at Hadoop Releases.

5.7. 07 April, 2014: release 2.4.0 available

Full information about this milestone release is available at Hadoop Releases.

5.8. 20 February, 2014: release 2.3.0 available

Full information about this milestone release is available at Hadoop Releases.

5.9. 11 December, 2013: release 0.23.10 available

Full information about this milestone release is available at Hadoop Releases.

5.10. 15 October, 2013: release 2.2.0 available

Apache Hadoop 2.x reaches GA milestone! Full information about this milestone release is available at Hadoop Releases.

5.11. 25 August, 2013: release 2.1.0-beta available

Apache Hadoop 2.x reaches beta milestone! Full information about this milestone release is available at Hadoop Releases.

5.12. 27 December, 2011: release 1.0.0 available

Hadoop reaches 1.0.0! Full information about this milestone release is available at Hadoop Releases.

5.13. March 2011 - Apache Hadoop takes top prize at Media Guardian Innovation Awards

Described by the judging panel as a "Swiss army knife of the 21st century", Apache Hadoop picked up the *innovator of the year* award for having the potential to change the face of media innovations.

See The Guardian web site

5.14. January 2011 - ZooKeeper Graduates

Hadoop's ZooKeeper subproject has graduated to become a top-level Apache project.

Apache ZooKeeper can now be found at http://zookeeper.apache.org/

5.15. September 2010 - Hive and Pig Graduate

Hadoop's Hive and Pig subprojects have graduated to become top-level Apache projects.

Apache Hive can now be found at http://hive.apache.org/

Pig can now be found at http://pig.apache.org/

5.16. May 2010 - Avro and HBase Graduate

Hadoop's Avro and HBase subprojects have graduated to become top-level Apache projects.

Apache Avro can now be found at http://avro.apache.org/

Apache HBase can now be found at http://hbase.apache.org/

5.17. July 2009 - New Hadoop Subprojects

Hadoop is getting bigger!

- Hadoop Core is renamed Hadoop Common.
- MapReduce and the Hadoop Distributed File System (HDFS) are now separate subprojects.
- Avro and Chukwa are new Hadoop subprojects.

See the summary descriptions for all subprojects above. Visit the individual sites for more detailed information.

5.18. March 2009 - ApacheCon EU

In case you missed it.... ApacheCon Europe 2009

5.19. November 2008 - ApacheCon US

In case you missed it.... ApacheCon US 2008

5.20. July 2008 - Hadoop Wins Terabyte Sort Benchmark

Hadoop Wins Terabyte Sort Benchmark: One of Yahoo's Hadoop clusters sorted 1 terabyte of data in 209 seconds, which beat the previous record of 297 seconds in the annual general purpose (Daytona) terabyte sort benchmark. This is the first time that either a Java or an open source program has won.