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 a simple programming model. 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-avaiability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-availabile service on top of a cluster of computers, each of which may be prone to failures.

The project includes these subprojects:

- **Hadoop Common**: The common utilities that support the other Hadoop subprojects.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce**: A software framework for distributed processing of large data sets on compute clusters.

Other Hadoop-related projects at Apache include:

- **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.
- **ZooKeeper™**: A high-performance coordination service for distributed applications.

2 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.

3 News

3.1 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

3.2 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/

3.3 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/

3.4 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/

3.5 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.

3.6 March 2009 - ApacheCon EU

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

3.7 November 2008 - ApacheCon US

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

3.8 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.