

# Home



This HADOOP2 space was migrated from old Hadoop wiki. Please check <https://cwiki.apache.org/confluence/display/HADOOP> for the current information.

## Apache Hadoop

**Apache Hadoop** is a framework for running applications on large cluster built of commodity hardware. The Hadoop framework transparently provides applications both reliability and data motion. Hadoop implements a computational paradigm named **Map/Reduce**, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system (**HDFS**) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both **MapReduce** and the Hadoop Distributed File System are designed so that node failures are automatically handled by the framework.

- **Apache Hadoop**
  - **General Information**
    - **Related-Projects**
  - **User Documentation**
    - **Setting up a Hadoop Cluster**
    - **Tutorials**
    - **MapReduce**
  - **Contributed parts of the Hadoop codebase**
  - **Developer Documentation**
  - **Related Resources**

## General Information

- **Official Apache Hadoop Website**: download, bug-tracking, mailing-lists, etc.
- **Overview** of Apache Hadoop
- **FAQ** Frequently Asked Questions.
- **What Hadoop is not**
- **Distributions and Commercial Support** for Hadoop (RPMs, Debs, AMIs, etc)
- **Presentations, books, articles** and **papers** about Hadoop
- **PoweredBy**, a growing list of sites and applications powered by Apache Hadoop
- **Support**
  - **Getting help from the hadoop community**.
  - **People and companies for hire**.
- **Hadoop Community Events and Conferences**
- **HadoopUserGroups** (HUGs)

## Related-Projects

- **HBase**, a Bigtable-like structured storage system for Hadoop HDFS
- **Apache Pig** is a high-level data-flow language and execution framework for parallel computation. It is built on top of Hadoop Core.
- **Hive** a data warehouse infrastructure which allows sql-like adhoc querying of data (in any format) stored in Hadoop
- **ZooKeeper** is a high-performance coordination service for distributed applications.
- **Hama**, a Google's Pregel-like distributed computing framework based on BSP (Bulk Synchronous Parallel) computing techniques for massive scientific computations.
- **Mahout**, scalable Machine Learning algorithms using Hadoop
- **Hadoop Compatible FileSystems (HCFS)**
- **Apache Gora**, open source framework provides an in-memory data model and persistence for big data. Gora supports persisting to column stores, key value stores, document stores and RDBMSs, and analyzing the data with extensive Apache Hadoop **MapReduce** support.

## User Documentation

- **Available Java Runtime Environments for Hadoop**
- **Important Concepts**
- **GettingStartedWithHadoop** (lots of details and explanation)
- **QuickStart** (for those who just want it to work *now*)
- **Command Line Options** for the Hadoop shell scripts.
- **Hadoop Code Overview**
- **Troubleshooting** What do when things go wrong

## Setting up a Hadoop Cluster

- **Starting a Single-Node Hadoop Cluster**
- **HowToConfigure** Hadoop software
- **WebApps** for monitoring your system
- **Configure NameNode High-Availability**
- **How to get metrics into ganglia**

- [Tips for managing a large cluster](#)
- [Disk Setup: some suggestions](#)
- [Topology Scripts / Rack Awareness](#)
- [Build and Install Hadoop 2.2 or newer on Windows](#)
- Virtual Clusters including Amazon AWS
  - [Virtual Hadoop](#) - the theory
  - How to set up a [Virtual Cluster](#)
  - Running Hadoop on [AmazonEC2](#)
  - Running Hadoop with AmazonS3

## Tutorials

- [Running Hadoop On Ubuntu Linux \(Single-Node\\_Cluster\)](#) Tutorial by Michael Noll on installing, configuring and running Hadoop on a single Ubuntu Linux machine.
- [Running Hadoop On Ubuntu Linux \(Multi-Node Cluster\)](#) Tutorial by Michael Noll on how to setup a multi-node Hadoop cluster.
- [Cloudera basic training](#)
- [Hadoop Windows/Eclipse Tutorial](#): How to develop Hadoop with Eclipse on Windows.
- [Yahoo! Hadoop Tutorial](#): Hadoop setup, HDFS, and [MapReduce](#)
- [Running Hadoop on Mac OSX \(Multi-Node Cluster\)](#) Tutorial on how to setup a multi-node Hadoop cluster on Macintosh OSX (Lion).

## MapReduce

The [MapReduce](#) algorithm is the foundational algorithm of Hadoop, and is critical to understand.

- [HadoopMapReduce](#)
- [HadoopMapRedClasses](#)
- [HowManyMapsAndReduces](#)
- [TaskExecutionEnvironment](#)
- [HowToDebugMapReducePrograms](#)
- Examples
  - [WordCount](#)
  - [Python Word Count](#)
  - [C/C++ Word Count](#)
  - [Grep](#)
  - [Sort](#)
  - [RandomWriter](#)
  - [How to read from and write to HDFS](#)
- Benchmarks
  - [Hardware benchmarks](#)
  - [Data processing benchmarks](#)

## Contributed parts of the Hadoop codebase

These are independent modules that are in the Hadoop codebase but not tightly integrated with the main project -yet.

- [HadoopStreaming](#) (Useful for using Hadoop with other programming languages)
- [DistributedLucene](#), a Proposal for a distributed Lucene index in Hadoop
- [MountableHDFS](#), Fuse-DFS & other Tools to mount HDFS as a standard filesystem on Linux (and some other Unix OSs)
- [HDFS-APIs](#) in Perl, Python, PHP and other languages.
- [Chukwa](#) a data collection, storage, and analysis framework
- [The Apache Hadoop Plugin for Eclipse](#) (An Eclipse plug-in that simplifies the creation and deployment of [MapReduce](#) programs with an HDFS Administrative feature)
- [HDFS-RAID](#) Erasure Coding in HDFS

## Developer Documentation

- [Roadmap](#), listing release plans.
- [HowToContribute](#)
- [HowToUseInjectionFramework](#)
- [HowToUseSystemTestFramework](#)
- [HowToSetupYourDevelopmentEnvironment](#)
- [HowToUseConcurrencyAnalysisTools](#)
- [GithubIntegration](#)
- [HowToUseJCarder](#)
- [HowToCodeReview](#)
- [Jira](#) usage guidelines
- [HowToCommit](#)
- [HowToRelease](#)
- [HudsonBuildServer](#)
- [HowToSetupUbuntuBuildMachine](#)
- [DevelopmentHints](#)
- [ProjectSuggestions](#)

- [Building/Testing under IntelliJ IDEA](#)
- [Git And Hadoop](#)
- [ProjectSplit](#)

## Related Resources

- [Nutch Hadoop Tutorial](#) (Useful for understanding Hadoop in an application context)
- [IBM MapReduce Tools for Eclipse](#) - Out of date. Use the Eclipse Plugin in the [MapReduce](#)/Contrib instead
- Hadoop IRC channel is #hadoop at irc.freenode.net.
- [Using Spring and Hadoop](#) (Discussion of possibilities to use Hadoop and Dependency Injection with Spring)
- [Univa Grid Engine Integration](#) A blog post about the integration of Hadoop with the Grid Engine successor Univa Grid Engine
- [Hadoop Grid Engine Integration](#) Open Grid Scheduler/Grid Engine Hadoop integration setup instructions.
- [Hadoop Tutorial Series](#) Learning progressively important core Hadoop concepts with hands-on experiments using the Cloudera Virtual Machine
- [Pydoop](#) A Python MapReduce and HDFS API for Hadoop ([tutorial](#)).
- [Dumbo](#) Dumbo is a project that allows you to easily write and run Hadoop programs in Python.
- [Hadoop distributed file system](#) New Hadoop Connector Enables Ultra-Fast Transfer of Data between Hadoop and Aster Data's MPP Data Warehouse.
- [Hadoop + CUDA](#)
- [Hadoop on ARM cluster](#) A study that compares Hadoop [MapReduce](#) applications' energy consumption and performance between ARM cluster and general X86\_64 cluster
- [HDFS Architecture Documentation](#) An overview of the HDFS architecture, intended for contributors.

---

[CategoryHomepage](#)