

Cloudera's Distribution Including Apache Hadoop (CDH)



Cloudera's Distribution Including Apache Hadoop (CDH) is the #1 distribution of Hadoop in commercial and non-commercial environments.

Available for free download at www.cloudera.com/downloads, CDH delivers a streamlined path for putting Apache Hadoop to work solving business problems in production. Ideal for enterprises seeking a stable, tested Hadoop solution without proprietary vendor lock-in, CDH is the bridge between the insights of organizations using Hadoop in production and the continuous stream of innovations from the Apache community.

Fully documented and ready to go; only CDH consists of 100% open source Apache Hadoop plus the comprehensive set of open source software components needed to use Hadoop, which can be activated on an as-needed basis. CDH is thoroughly tested and certified to integrate with the widest range of operating systems and hardware, databases and data warehouses, and business intelligence and ETL systems, enabling users to protect existing technology investments by taking advantage of Hadoop while leveraging their existing tools and resources.

Simplicity and Certainty

Due to the explosive growth in data volumes and complexity, companies of all sizes are adopting the open source Apache Hadoop platform to store, process and derive value from all their data. Hadoop has many advantages, particularly in its flexibility and economics; but just like any open source technology, implementation into a production environment presents a unique set of challenges and questions that must be addressed. How can you guarantee system stability? Which components of the open source platform do you need and how can you ensure that they will integrate correctly? How well is the technology documented? Who do you call for support? The uncertainty caused by questions like these drives some companies to invest in proprietary vendor solutions – despite the higher price tag and risk of vendor lock-in and product bloat.

Cloudera created its distribution of Hadoop for this very purpose – to remove the uncertainty and barriers that sometimes dissuade enterprises from deploying open source Hadoop in production.

CDH is an enterprise-ready distribution of Apache Hadoop.

STREAMLINED PATH TO SUCCESS WITH HADOOP

- Contains all of the components needed for use in production, which can be implemented on an as-needed basis
- Fully tested and documented
- The integration work is done for you – works with the widest range of operating systems, hardware, and data analysis tools

STABLE AND RELIABLE HADOOP WITHOUT VENDOR LOCK-IN

- 100% pure Apache Hadoop, hardened for production environments
- Thoroughly tested by Cloudera's expert engineers
- Proven at scale across tens of thousands of nodes in enterprise environments

COMMUNITY DRIVEN

- Incorporates only open source components – no forks or proprietary underpinnings
- Access to continuous stream of innovations from the community
- Completely free and open source

“Cloudera's distribution and support have been instrumental in helping Groupon deliver on our goal to be a technology leader.”

GROUPON

What CDH Delivers

A complete, packaged set of all components needed to use Apache Hadoop in production

CDH contains everything you need for a successful implementation, and you can deploy the different components as you need them. All of the integration work is done for you, and the entire solution is fully documented. By taking the guesswork out of building out your Hadoop deployment, CDH gives you a streamlined path to success in solving real business problems.

100% pure Apache Hadoop, hardened for production environments

With CDH, Cloudera maintains a vigorous commitment to open source. From the core file system to the UI framework, CDH contains absolutely no code forks or proprietary underpinnings and is free to download. The entire distribution is QA tested by Cloudera and proven at scale across tens of thousands of nodes and varying workloads in enterprise environments. CDH is stable, reliable Apache Hadoop without vendor lock-in, which is why it's the #1 distribution deployed in commercial and non-commercial environments.

The Cloudera Ready certification program

CDH is tested and certified to integrate with a wide range of hardware, operating systems and data analysis tools – lowering risk, easing deployment and protecting your investments in your existing data analysis and business intelligence solutions.

CDHv3 Components and Distribution Details

Component	Description	Version
Apache Hadoop	Reliable, scalable distributed storage and computing	v0.20.2 + 923.194
Apache Flume	Distributed service for collecting and aggregating log and event data	v0.9.4 + 25.40
Apache HBase	Hadoop database for random, real-time read/write access	v0.90.4+49.137
Apache Hive	SQL-like language and metadata repository	v0.7.1 + 42.36
Apache Mahout	Library of machine learning algorithms for Apache Hadoop	v0.5 + 9.3
Apache Oozie	Server-based workflow engine for Hadoop activities	v2.3.2 + 27.12
Apache Pig	High-level language for expressing data analysis programs	v0.8.1 + 28.26
Apache Sqoop	Data transport engine for integrating Hadoop with RDBMS	v1.3.0 + 5.68
Apache Whirr	Library for running Hadoop in the cloud	v0.5.0 + 4.8
Apache Zookeeper	Highly-reliable distributed coordination service	v3.3.4 + 19
Fuse-DFS	Module within Hadoop for mounting HDFS as a traditional file system	v0.20.2 + 923.194
Hue	Browser-based desktop interface for interacting with Hadoop	v1.2.0.0 + 114.4

Supported Operating System Versions

CentOS	CentOS 5, CentOS 6
Debian	Squeeze, Lenny
Oracle	Oracle Linux 5.6 with Unbreakable Enterprise Kernel
Red Hat	RHEL 5, RHEL 6
SUSE	SLES 11 (service pack 1 or later)
Ubuntu	Lucid, Maverick

Supported Build Infrastructure and Cloud Platforms

Build Infrastructure	Apache Maven
Cloud Platforms	Rackspace, Amazon EC2, Softlayer