

Dell | Cloudera Apache Hadoop Solution, accelerated by Intel

Leverage enterprise-optimized hardware, software and services with the Dell | Cloudera Apache Hadoop Solution

Streamlined to solve business problems

When you move forward with the Dell™ | Cloudera® Apache™ Hadoop® Solution, you have the confidence that comes with Cloudera's distribution including Apache® Hadoop® (CDH), one of the world's leading distributions of Hadoop in commercial and non-commercial environments.

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, CDH consists of 100-percent 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 a wide range of operating systems and hardware, databases and data warehouses, and business intelligence and extract, transform and load (ETL) systems. This broad compatibility helps your organization take advantage of Hadoop while leveraging your existing tools and resources.

CDH is incorporated in Cloudera Enterprise, which also includes Cloudera Search, Impala and Spark. Cloudera Enterprise streamlines your IT environment with one tool for data processing and interactive analysis, allowing for faster access to critical business insights with interactive analytics and the ability to scale as needed. Cloudera Enterprise is thoroughly tested and validated by Dell to support a wide range of big data use cases—and to give organizations complete confidence in their Hadoop environments.

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 of their data. There are many advantages to Hadoop, particularly in the areas of flexibility and economics. But just like any open source technology, implementation into a production environment presents a unique set of challenges and questions.

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.

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.

Everything you need

The Dell | Cloudera Apache Hadoop Solution, accelerated by Intel, gives you everything you need to tackle your big-data challenges—including software, hardware, services, reference architecture and deployment guide. Begin with PowerEdge™ R730xd series servers and Dell Networking—components proven in cloud and big-data deployment—then select Cloudera Enterprise or CDH distributions of Apache Hadoop.

Dell Services makes getting started easy. Options include, but are not limited to, custom solution design, hardware and software deployment, ongoing support and training. This gives you the assurance that your solution is backed by expert hardware and software support that can be tailored to your specific needs.



Streamlined path to success with Hadoop

- Gain all the components needed for use in production; implement components on an as-needed basis.
- Remove uncertainty with a fully supported distribution.
- Protect existing investments and the integration work is done for you.

Stable and reliable Hadoop without vendor lock-in

- Leverage 100-percent pure Apache Hadoop, hardened for production environments.
- Increase your confidence with a distribution thoroughly tested by expert engineers.
- Scale across tens of thousands of nodes in enterprise environments.

Community driven

- Incorporate only open source components—no forks or proprietary underpinnings.
- Gain access to a continuous stream of innovations from the Hadoop community.
- Leverage a proven distribution that is free to download.

What the Dell | Cloudera Apache Hadoop Solution delivers

A complete, packaged set of all components needed to use Apache Hadoop in production	The Dell Reference Architecture, with CDH 5.7, contains everything needed for a successful implementation. It takes the guesswork out of building a Hadoop stack, with support, hardware, networking, CDH 5.7 software and services to enable a streamlined path to success in solving real business challenges.
100-percent pure Apache Hadoop, hardened for production environments	With CDH, Dell, together with Cloudera, maintains a vigorous commitment to upholding the open source standard. From the core file system to the user interface, CDH contains absolutely no code forks or proprietary underpinnings and is free to download. The entire distribution is quality tested by Cloudera and proven at scale across tens of thousands of nodes and varying workloads in enterprise environments. CDH is stable, reliable Hadoop without vendor lock-in.
The Cloudera Ready certification framework	CDH has been 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.

Core components

CDH	At the core of Cloudera Enterprise is CDH, which combines Apache Hadoop with a number of other open source projects to create a single, massively scalable system where you can unite storage with an array of powerful processing and analytic frameworks.
Automated Cluster Management – Cloudera Manager	Cloudera Enterprise includes Cloudera Manager to help you easily deploy, manage, monitor and diagnose issues with your cluster. Cloudera Manager is critical for operating clusters at scale
Cloudera Support	Get the industry's best technical support for Hadoop. With Cloudera Support, you'll experience more uptime, faster issue resolution, better performance to support your mission-critical applications, and faster delivery of the platform features you care about.

Advanced components

Online NoSQL – HBase	HBase is a distributed key-value store that helps you build real-time applications on massive tables (billions of rows, millions of columns) with fast, random access.
Analytic SQL – Impala	Impala is the industry's leading massively-parallel processing (MPP) SQL engine built for Hadoop.
Search – Cloudera Search	Cloudera Search, based on Apache Solr, lets your users query and browse data in Hadoop just as they would search Google or their favorite ecommerce site.
In-Memory Machine Learning and Stream Processing – Apache Spark	Apache Spark delivers fast, in-memory analytics and real-time stream processing for Hadoop.
Data Management – Cloudera Navigator	Cloudera Navigator provides critical enterprise data audit, lineage and data discovery capabilities that enterprises require.

A deeper dive

For a closer look at the solution components, see these resources in the Dell TechCenter:

- [Dell | Cloudera Apache Hadoop Solution Overview](#)
- [Dell | Cloudera Apache Hadoop Solution Reference Architecture Guide – Version 5.7](#)

Highlights of this release include:

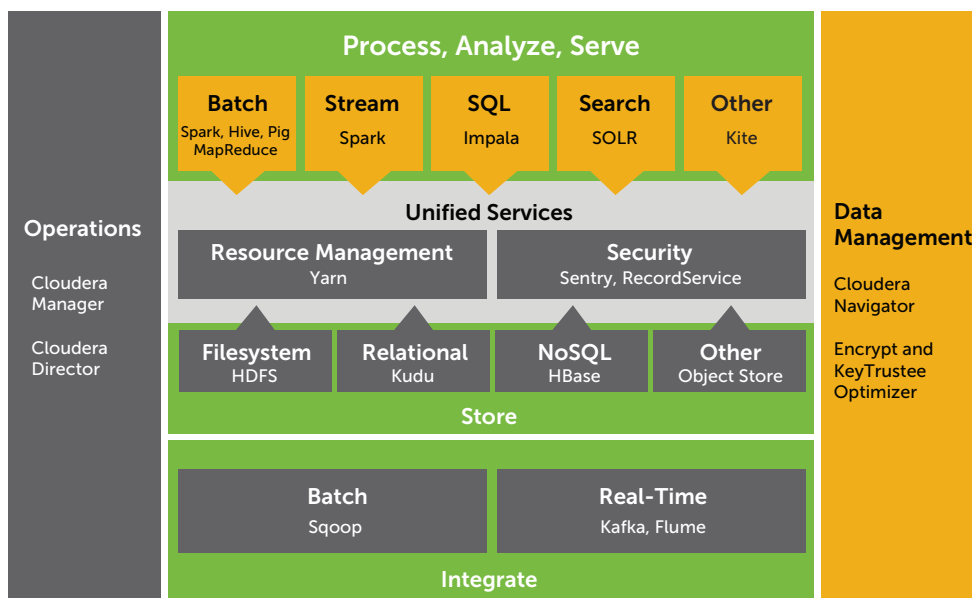
For a closer look at the solution components, see these resources in the Dell TechCenter:

- Production-ready Hive-on-Spark functionality
- Accelerate the use of Apache Spark as a data processing standard
- Two-X performance gains for Apache Impala (incubating)
- Easier cluster configuration and utilization reporting
- End-to-end encryption for Apache Spark data



CDH 5.7 distribution details	
Component versions	
Cloudera Manager	v5.7
Cloudera Navigator	v2.6
Apache Avro	v1.7.6 +88
Apache Crunch	v0.11.0 +46
DataFu	v1.1.0 +16
Apache Flume	v1.5.0 +122
Apache Hadoop	v2.6.0 +572
Apache HBase	v1.2
HBase-Solr	v1.5 +50
Apache Hive	v1.1.0 +141
Hue	v3.7.0 +1220
Cloudera Impala (incubating)	v2.5
Apache Kafka	v0.9
Kite SDK	v1.0.0 +22
Llama	v1.0.0 +0
Apache Mahout	v0.9 +21
Apache Oozie	v4.1.0 +142
Parquet	v1.5.0 +93
Parquet-format	v2.1.0 +9
Apache Pig	v0.12.0 +56
Cloudera Search	v1.0.0 +0
Apache Sentry (incubating)	v1.4.0 +165
Apache Solr	v4.10.3 +251
Apache Spark	v1.6
Apache Sqoop	v1.4.5 +98
Apache Sqoop2	v1.99.5 +32
Apache Whirr	v0.9.0 +14
Apache ZooKeeper	v3.4.5 +88
Supported operating system versions	
Red Hat	Red Hat Enterprise Linux Server 6.7
CentOS	CentOS 6.7
Ubuntu	Lucid, Precise
SUSE	SUSE 11 SP1
Debian	7.8
JDK	7_80
JDK	8_60
Supported build infrastructure and cloud platforms	
Build Infrastructure	Apache Maven
Cloud Platforms	Rackspace, Amazon AWS, Amazon EC2, Softlayer, Microsoft Azure, OpenStack

CDH components	
Apache Hadoop	Reliable, scalable distributed storage and computing
Apache DataFu	Library with user defined functions for data mining and statistics tasks
Apache Flume	Distributed service for collecting and aggregating log and event data
Apache HBase	Hadoop database for random, real-time read/write access
Apache Hive	Data access and processing language
Apache Hue	High-level language that provides a web interface for analyzing data
Apache Kafka	A high throughput distributed messaging system, is publish-subscribe messaging rethought as a distributed commit log.
Apache Mahout™	Tool that provides machine learning capabilities
Apache MRv2	MapReduce software framework for processing large data sets
Apache Oozie	Engine that provides a general workflow capability for coordinating complex sequences of production jobs
Apache Pig	SQL-like language and metadata repository
Apache Spark	Fast and general data processing engine that supports cyclic flow and in-memory computing
Apache Sqoop	Transfers bulk data between Apache Hadoop and structured datastores
Apache Whirr	Library for running Hadoop in the cloud
Apache Zookeeper	Resource manager for Hadoop - the next generation of MapReduce framework
Apache YARN	Highly reliable distributed coordination service



To learn more, visit Dell.com/Hadoop | Dell.com/BigData

For questions, please contact us at: Hadoop@Dell.com

© 2016 Dell Inc. All rights reserved. Dell, the DELL logo, the DELL badge and PowerEdge are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell disclaims proprietary interest in the marks and names of others. This document is for informational purposes only. Dell reserves the right to make changes without further notice to the products herein. The content provided is as-is and without expressed or implied warranties of any kind.

