Distributions and Commercial Support

Products that include Apache Hadoop or derivative works and Commercial Support

The following companies provide products that include Apache Hadoop, a derivative work thereof, commercial support, and/or tools and utilities related to Hadoop.

Please see Defining Hadoop to see the Apache Hadoop's project's copyright, naming, trademark and compatibility policies.

This listing is provided as a reference only. No endorsements are given or implied.

The sole products that can be called a release of Apache Hadoop come from apache.org. Some companies release or sell products that include the official Apache Hadoop release files, and/or their own and other useful tools. Other companies or organizations release products that include artifacts build from modified or extended versions of the Apache Hadoop source tree. Such derivative works are not supported by the Apache Team: all support issues must be directed to the suppliers themselves.

The Apache Software Foundation strongly encourages users of Hadoop —in any form— to get involved in the Apache-hosted mailing lists. Even though you may only get support through the supplier of any derivative work of Apache Hadoop, by participating in the Hadoop user and developer lists, you can become an active part of the Hadoop community. Your needs may be addressed in future versions of the code, and you will be able to get in touch with many other users of the technology.

The Hadoop developers would like you to be aware that filing JIRA issues is not a way to get support to get your Hadoop installation up and running. Those bug reports are not bugs, and will be closed as invalid. Either use the hadoop-user mailing list, the organisations providing support listed below, or delve into the Hadoop source itself.

Entries are listed alphabetically by company name.

- Amazon Web Services
  - Amazon offers a version of Apache Hadoop on their EC2 infrastructure, sold as Amazon Elastic MapReduce.
- Apache Bigtop
  - Apache Bigtop is a project for the development of packaging and tests of the Apache Hadoop ecosystem. This includes testing at various levels (packaging, platform, runtime, upgrade, etc...) developed by a community with a focus on the system as a whole, rather than individual projects.
  - Apache Bigtop doesn't provide binary artifacts of its releases, it is source only project.
- Cascading - Cascading is a popular feature-rich API for defining and executing complex and fault tolerant data processing workflows on a Apache Hadoop cluster. Cascading 2.0 is Apache-licensed.
- Cloudera
  - Cloudera distributes a platform of open-source projects called Cloudera's Distribution including Apache Hadoop or CDH. In addition, Cloudera offers its enterprise customers a family of product and services that complement the open-source Apache Hadoop platform. These include comprehensive training sessions, architectural services and technical support for Hadoop clusters in development or in production. We serve a wide range of customers including retail, government, financial service, healthcare, life sciences, digital media, advertising, networking and telephony enterprises.
- Cloudspace
  - Cloudspace is a web technology consulting company, since 1996. Cloudspace uses Apache Hadoop to scale client and internal projects on Amazon's EC2 and bare metal architectures.
- Datameer
  - Datameer Analytics Solution (DAS) is a Hadoop-based solution for big data analytics that includes data source integration, storage, an analytics engine and visualization.
  - DAS Log File Aggregator is a plug-in to DAS that makes it easy to import large numbers of log files stored on disparate servers.
- Data Mine Lab
  - Data Mine Lab is a London-based consultancy developing solutions based on Apache Hadoop, Apache Mahout, Apache HBase and Amazon Web Services. Data Mine Lab uses combination of cloud computing, MapReduce, columnar databases and open source Business Intelligence tools to develop solutions that add value to their customers' businesses and the data they collect.
- Datasalt
  - Datasalt is an Apache Hadoop consulting company which has released two open-source products on top of Hadoop (Pangool, an easier low-level API for Hadoop, and Splout SQL, a low-latency SQL serving engine on top of Hadoop). Datasalt provides commercial support, public / private training and custom Hadoop development.
- DataStax
  - DataStax provides a product which fully integrates Apache Hadoop with Apache Cassandra and Apache Solr in its DataStax Enterprise platform. DataStax Enterprise is completely free to use for development environments with no restrictions. In addition, DataStax supplies OpsCenter for visual management and monitoring, along with expert support, training, and consulting services for Apache Hadoop, Apache Cassandra, and Apache Solr.
- DataTorrent
  - DataTorrent provides a Hadoop 2.0 real-time stream-processing platform, designed for today's Big Data needs. DataTorrent is certified on Apache Hadoop, and all leading distributions.
  - Enables enterprises to ingest, process, analyze, and act on data in motion. The DataTorrent platform includes built-in fault tolerance and auto-scaling, DataTorrent can process billions of events/second.
  - Applications that run on DataTorrent are comprised of Java code objects called "Operators". Malhar is a an open-source free library of over 400 commonly-used operators, UI widgets, and application templates that are open sourced under the Apache 2.0 license.
- Debian
  - A Debian package of Apache Hadoop is available. Please see the Debian Wiki on Hadoop.
- Emblocsoft
  - Delivers enterprise Hadoop edition based on Apache Hadoop, or EDH, to meet the demand of enterprise data processing:
    - Big Data data visualization and advanced analytics
    - Real time stream processing

- Machine learning at scale
- Enterprise integration
  - Asia Pacific time zone support for China, Hong Kong, Macau, Taiwan, and other countries in Asia including Singapore, Korea, Malaysia, Thailand, Japan, India, Australia and New Zealand.
- Hortonworks
  - Major contributors to Apache Hadoop and dedicated to working with the community to make Apache Hadoop more robust and easier to install, manage, use, integrate and extend.
  - Provides Hortonworks Data Platform Powered by Apache Hadoop, which is a 100% open source big-data platform based upon Apache Hadoop. HDP-2.2 is built on Apache Hadoop 2.6.
  - Provider of expert technical support, training and partner-enablement services for both end-user organizations and technology vendors.
- HStreaming
  - HStreaming offers real-time stream processing and continuous advanced analytics built into Hadoop. Scales to 1000+ cluster nodes and processes millions of events per second.
  - Single Hadoop platform for both stream and batch processing jobs with common code base including Apache Pig. Compatible with all major Hadoop distributions.
  - Available as free community edition, enterprise edition, and cloud service.
- IBM
  - IBM InfoSphere BigInsights brings the power of Apache Hadoop to the enterprise. BigInsights Enterprise Edition builds on Apache Hadoop with capabilities to withstand the demands of an enterprise including:
    - Advanced Text Analytics
    - Performance Optimizations
    - Workload Management & Scheduling
    - Professional-Grade Visualization & Developer Tooling
    - Enterprise Integration & Security
  - The result is a more developer and user-friendly solution for complex, large scale analytics.
  - Learn Hadoop using InfoSphere BigInsights in the IBM Cloud (www.bigdatauniversity.com).
  - Platform Computing provides an Enterprise Class MapReduce solution for Big Data Analytics with high scalability and fault tolerance. Platform MapReduce provides unique scheduling capabilities and its architecture is based on almost two decades of distributed computing research and development. Based on the same low-latency distributed architecture deployed in the leading financial institutions on Wall Street, the solution meets the needs of the most demanding enterprise customers.
- Impetus
  - Impetus' LADAP system is built for large enterprises and Websites to effectively derive intelligence out of raw data from discrete sources. With LADAP, an in-depth analysis can be undertaken on data from many different sources including social networks, to find out the patterns and structures within it. More info about LADAP @Impetus
  - The HCM (Hadoop Cluster Management) tool is a solution that automates the cluster setup and management activities, thus reducing the overall time, cost and effort required for setting up and managing Hadoop clusters. More info about HCM @Impetus
  - With a strong focus, established thought leadership and open source contributions in the area of Big Data analytics and consulting services, Impetus uses its Global Delivery Model to help technology businesses and enterprises evaluate and implement solutions tailored to their specific context, without being biased towards a particular solution. More info about BigData @Impetus
- Jaspersoft – Embedded Open Source Business Intelligence
  - Build reports, dashboards and analytics directly from Hadoop and other Big Data stores, without having to move the data to another database.
  - Blend your Hadoop data with other data sources using Data Virtualization or traditional ETL capabilities.
  - Embed Hadoop visualizations & reports inside your app or use the insights to optimize your business.
  - Deploy Jaspersoft either on-premises or by-the-hour in the Cloud
- Karmasphere
  - Distributes Karmasphere Studio for Hadoop, which allows cross-version development and management of Apache Hadoop jobs in a familiar integrated development environment.
- Apache Mahout
  - Another Apache project using Hadoop to build scalable machine learning algorithms like canopy clustering, k-means and many more.
- MapR Technologies
  - MapR sells a high performance map-reduce framework based on Apache Hadoop that includes many of the standard eco-system components. A replacement file system and significant amount of re-engineering of the MapReduce components allows significantly higher performance, higher level of fault tolerance, distributed incremental backups, read-write access to the cluster file store via NFS and other features (the HDFS team would dispute some of these assertions).
- Nutch - Apache Nutch: flexible web search engine software
- NGDATA
  - Makes available Lily Open Source that further builds upon Apache Hadoop, Apache HBase and Apache Solr. Lily wraps these leading edge Apache Software Foundation technologies into an easy-to-use, all in one solution bringing Big Data storage, indexing and search to the enterprise.
  - Distributes Lily Enterprise which delivers a unique combination of interactive Big Data management, machine learning technologies and consumer intelligence applications in one integrated solution to allow better, and more dynamic, consumer insights.
- Pentaho – Open Source Business Intelligence
  - Pentaho provides a complete, end-to-end open-source BI alternative to proprietary offerings like Oracle, SAP and IBM.
  - Offers an easy-to-use, graphical ETL tool that is integrated with Apache Hadoop for managing data and coordinating Hadoop related tasks in the broader context of your ETL and Business Intelligence workflow.
  - Provides Reporting and Analysis capabilities against big data in Hadoop.
  - Learn more at http://www.pentaho.com/hadoop.
- Pervasive Software
  - Provides Pervasive DataRush, a parallel dataflow framework which improves performance of Apache Hadoop and MapReduce jobs by exploiting fine-grained parallelism on multicore servers. (contact)
- Pivotal
  - Pivotal HD delivers the foundation for a Business Data Lake architecture, providing the world's most advanced real-time analytics and most extensive set of advanced analytical toolsets for data scientists, IT and business analysts, and developers.
  - Pivotal HD includes the capabilities of Apache Hadoop in a fully-supported, enterprise-ready distribution combined with a rich proven, parallel SQL query processing engine from Pivotal HAWQ and in-memory, real-time analytics from Pivotal GemFire XD.
  - Pivotal HD powers the industry's only closed loop real-time analytics platform for OLAP and OLTP with HDFS as the common data storage layer. By combining the day-to-day events of your business with your analytics systems, you can act prescriptively to events in real-time.

- Sematext International
  - Provides consulting services around Apache Hadoop and Apache HBase, along with large-scale search using Apache Lucene, Apache Solr, and Elastic Search.
  - Runs the popular search-hadoop.com search service.
- Syncsort
  - Syncsort provides high-performance software to collect, process and distribute data. Syncsort runs natively in Hadoop. It does not generate any Java, Pig or HiveQL; doesn't need to compile, and doesn't have any tuning or code maintenance requirements.
  - Syncsort Hadoop Solutions are used for:
    - Extracting, loading and transforming data with Hadoop (Hadoop ETL)
    - Offloading expensive and inefficient workloads from data warehouses or mainframes into Hadoop
    - Accelerate MapReduce processes such as HiveQL & Pig
  - With Syncsort you can:
    - Develop sophisticated data flows in Windows without writing any code and deploy natively in Hadoop
    - Access and move data from/to virtually any data source to/from Hadoop, including best-in-class mainframe data access and translation capabilities
    - Deploy, maintain, monitor and secure the Hadoop environment with tight integration with Hadoop job tracker, Cloudera Manager, Ambari and standard security protocols such as LDAP and Kerberos
    - Syncsort supports and integrates with all major Hadoop distros including Cloudera, Hortonworks, and MapR
    - Also available at the Amazon AWS marketplace for Amazon EMR
    - Run on-premises or in the cloud
- Talend - The Open Source Integration Company
  - Talend Platform for Big Data includes support and management tools for all the major Apache Hadoop distributions.
  - Talend Open Studio for Big Data is an Apache License Eclipse IDE, that provides a set of graphical components for HDFS, HBase, Pig, Sqoop and Hive. A configuration of a big data connection is represented graphically and the underlying code is automatically generated and then can be deployed as a service, executable or stand-alone job.
- Think Big Analytics - Flexible Data Solution Services
  - Think Big Analytics offers expert consulting services specializing in Apache Hadoop, MapReduce and related data processing architectures.
  - Offers 1 week Brainstorm Workshops and 6 week Deployment Iterations delivered collaboratively with our customers.
- Tresata - Big Data Analytics Platform for the Financial Services Industry
  - Big Data as a Service Platform built for Data used in the Financial Services Industry
  - Financial Industry's first software platform architected from the ground up on Hadoop
  - Analyze structured (transactions, balances) & unstructured (online, mobile, voice, social) data
  - Data storage, processing, analytics and visualization all done on Hadoop
- VMware - Initiate Open Source project and product to enable easily and efficiently deploy and use Hadoop on virtual infrastructure.
  - HVE - Hadoop Virtual Extensions to make Hadoop virtualization-aware
  - Serengeti - enable the rapid deployment of an Apache Hadoop cluster
- WANdisco is a committed member & sponsor of the Apache Software community and has active committers on several projects including Apache Hadoop.
  - WANdisco provides:
  - Non-Stop Hadoop: WANdisco's patented replication technology turns the NameNode into an active-active shared-nothing cluster that delivers the first and only continuous availability solution for globally distributed Hadoop deployments. Clients at every location have LAN-speed read and write access to same data at every location, and failover and recovery are immediate and automatic in and across data centers.
  - Non-Stop Hadoop for Hortonworks: WANdisco's patented replication technology applied to Hortonworks Data Platform (HDP) to deliver the first and only continuous availability solution for globally distributed HDP deployments.
  - Non-Stop Hadoop for Cloudera: This applies WANdisco's patented replication technology to deliver the first and only continuous availability solution for globally distributed CDH deployments and integrates seamlessly with Cloudera Manager.