CITO Research

Advancing the craft of technology leadership

# Choosing a Provider from the Hadoop Ecosystem

Sponsored by MapR Technologies

## Contents



## Introduction: The Hadoop Opportunity

Enterprises are faced with new requirements for data. We now have big data that is different from the structured, cleansed corporate data repositories of the past. Before, we had to plan out structured queries. In the Hadoop world, we don't have to sort data according to a predetermined schema when we collect it. We can store data as it arrives and decide what to do with it later. Today, there are different ways to analyze data collected in Hadoop—but which one is the best way forward?

*There are different ways to analyze data collected in Hadoop—but which one is the best way forward?*

The key questions facing companies are simple and direct:

- How can we use Hadoop as a tool in our analytics arsenal to gain significant benefits?
- How do we know if our Hadoop investment and related analytics generate the maximum return?
- How can we create a Hadoop infrastructure for big-data analysis that meets our current and future needs?
- How can we ensure our Hadoop infrastructure is flexible enough for all our use cases and provides support for mission critical applications?

CITO Research has written this paper for companies that are considering Hadoop as a way to gain value from big data. This paper is not an exhaustive guide to Hadoop and all of its capabilities and components. Instead, it offers a framework that you can use to evaluate product-strategy differences among the three primary distributors of Hadoop.

Companies considering Hadoop must answer questions about its usability and functionality and whether Hadoop can play nicely with existing enterprise systems, analytics tools, and databases. They may also have concerns when choosing a Hadoop distributor. Will such a choice create vendor lock-in? Will future changes to the Hadoop open source get folded into the distribution that they choose?

In this paper, CITO Research will:

- Explain the place of Hadoop in the context of the contemporary enterprise
- Present a framework for comparing the commercial Hadoop alternatives

Based on this information, companies can analyze each alternative with respect to their detailed requirements.



*Hadoop is quickly becoming an enterprise IT platform, but presents some challenges that are best solved before committing to a distribution.*

## What Is Hadoop?

Hadoop is an open-source software framework, available as a free license from the Apache Software Foundation. It supports data-intensive, distributed applications on large clusters of low-cost hardware. Hadoop was inspired by a Google white paper on its MapReduce programming model, which allows large-scale computations to run in parallel across large server clusters. While at Yahoo!, Doug Cutting implemented the MapReduce concept as an open-source framework, composed in Java and running on Linux. Hadoop became very popular with Internet companies such as Yahoo!, Amazon, Facebook, and others as they struggled to handle the enormous amount of data generated by the intensive use of the Internet.

Hadoop analyzes massive datasets. It helps companies move their algorithms and computing power to the data, rather than the other way around. Today, companies must analyze an ocean of data created by their internal systems and by other systems that track customer behavior and other activities. Hadoop is quickly becoming an enterprise IT platform, but presents challenges best solved before committing to a distribution.

### How Hadoop Expands Data Analysis Capabilities

Hadoop  handles the masses of unstructured data now produced by electronic devices, social media posts, and the like. It expands the analyses businesses can conduct. While relational databases use a structured query—essentially a pre-designed question that contains expectations about where data will be located and how it will be categorized—Hadoop is much more fluid, expanding the range of analytical techniques to include facial recognition, sentiment analysis, and more.



**Understanding What Hadoop Means to Your Business**

Answering the following questions will close the gap between what Hadoop can do generally and what it can do specifically for your business. This will provide the basis for choosing the right Hadoop implementation and creating an infrastructure.

- What data will be analyzed?
- What questions do you want to answer?
- What methods of analysis do you plan to use?
- What data do you want to bring into Hadoop?
- What is the scope of applications to be supported and what are their requirements (disaster recovery, availability, and so on)?
- What use cases do you anticipate?
- How will your use of Hadoop expand across users, applications, and so on?
- What existing tools and infrastructure do you want to integrate with Hadoop?
- Hadoop's core distribution does not provide management capabilities. Will your administrators need management tools or will administrators be hired or trained for managing Hadoop?

## Hadoop in Context

Because it is an open-source framework, Hadoop is often compared to the open-source operating system Linux. But there are critical differences. Because Unix was already a well-understood technology, Linux was easily commoditized, and the "ast mile completed by commercial distributions was to provide a layer of management over Linux implementations, and, in the case of Red Hat, enterprise-level support. Also, because Linux offered a low-cost way to run existing Unix programs on much cheaper hardware, there was a readymade market for its adoption and a built-in incentive for major hardware manufacturers such as IBM and HP to fund and support its continued development.

Hadoop, on the other hand, is at the beginning of its lifecycle. It was created to crunch massive web crawls into indexes to support searches and to enable real-time analysis of streaming data sets, but not necessarily to support application development, or to provide a robust, scalable distributed file system. The ethic of the open-source community is to create a good enough piece of code and optimize it later.

*Hadoop's open source community began as an interpretation of a white paper, and many features remained to be filled in.*

Hadoop and related projects have become hotbeds of computer science research and activity. Its open source community began as an interpretation of a white paper, and many features remained to be filled in. These include value-added services that ease implementation, integration, and data protection that are offered through commercial distributions. A debate exists between distribution vendors about the importance of adhering to Hadoop as a "pure" open-source project, since it has some design flaws that impact many enterprise use cases.

## The Hadoop Ecosystem

Like many technologies, Hadoop has benefits and drawbacks. Here, we'll quickly outline the benefits and drawbacks of Hadoop and identify key efforts of commercial vendors to resolve the drawbacks. Hadoop's ecosystem of related commercial providers is expanding the capabilities and ease of use of the system, but their approaches differ significantly.

### How Hadoop Works

The Hadoop Distributed File System (HDFS) affordably stores a large amount of data on local disk across a distributed cluster of computers. MapReduce is Hadoop's functional algorithm that decides how to distribute data and optimally control the processing across the cluster. The name MapReduce is quite literal. In the Map function, a master node splits an algorithm into subtasks on worker nodes, each of which processes the algorithm on the portion of relevant data it contains. Next, the Reduce function reassembles, disambiguates, and de-duplicates the data, delivering a readable response.

The success of a distributed Hadoop processing job hangs on the performance of the NameNode—the part of the master node that identifies the location of each block in relation to the master file being computed. There are some scalability issues associated with the NameNode—it can accommodate from 100 to 200 million files, depending on the memory capacity of the node. If the NameNode fails, it may lose track of the blocks and must reestablish communication with each individual block, a repair process than can take eight hours or more in a typical cluster.

HDFS was written in Java and is a write once storage layer. Updates to closed files are conducted via an append process. The batch updates of HDFS are a major limitation. There is no support for continuous updates to a file. Moreover, HDFS relies on the underlying Linux file system to store the HDFS contents.



## A Framework for Considering Hadoop Distributions

Commercial distributions for Hadoop assemble the various enhancement projects from the Apache repository and present them in a unified product so businesses don't have to embark on a science project of assembling each of these elements into a functional whole.

Vendors of open-source distributions often vary what they offer in a distribution. Sometimes, the vendor offers open source software and support, consulting, and education services. It is quite common for vendors to also offer extra, proprietary software that is only provided to paying customers.

Our proposed Hadoop framework evaluates the layers of innovation offered by commercial vendors in the Hadoop community:

- **Core distribution:** All vendors use the Apache Hadoop core and package it for enterprise use.
- **Management capabilities:** Some vendors provide an additional layer of management software that helps administrators configure, monitor, and tune Hadoop.
- **Enterprise reliability and integration:** A third class of vendors offers a more robust package, including a management layer augmented with connectors to existing enterprise systems and engineered to provide the same high level of availability, scalability, and reliability as other enterprise systems.

The following table shows vendors and features for the Hadoop framework.

*Table 1. A Framework for Considering Hadoop Distributions*

| Layer | Predominant Vendor | Features |
|---|---|---|
| Core distribution | Cloudera, MapR Technologies, Hortonworks | Subscription service model; support services |
| Management capabilities | Cloudera, MapR Technologies | Tuning, configuration, monitoring |
| Enterprise reliability and integration | MapR Technologies | Enterprise-quality engineering, built-in connectors to existing systems, data protection, snapshots, real-time data streaming |



## Core Distribution/Services Focus – Hortonworks

*Hortonworks' objective is to stay as close to the original open-source trunk of Hadoop as possible, limiting enhancements to its product to those offered publicly by the Hadoop community.*

Some distributions predicate their value proposition on their proximity to the original Hadoop trunk, the core distribution of the software committed to a source-code management system. Hortonworks, founded by Yahoo! engineers who had worked on the original implementation of Hadoop, offers a services-only model for Hadoop. Its objective is to stay as close to the original open-source trunk of Hadoop as possible, limiting any enhancements to its product to those offered publicly by the Hadoop community.

Hortonworks charges for support, but not for the distribution itself. Their rationale is twofold. First, Hortonworks believes that the Hadoop community, to which Hortonworks is a contributor, will devise the optimum solution. (It's worth noting that in practice the community has not changed Hadoop much in the past two years.) Hortonworks also believes that it's important to avoid vendor lock-in by committing to a forked version of Hadoop.

Hortonworks is focused on improving the usability of the Hadoop platform as it is written and presented in the Apache repository by offering a free distribution and subscription support. This approach works for parties with a vested interest in Hadoop remaining open-source but also reliable.

The services model seems to be geared toward developers committed to using vanilla Hadoop as their data management platform. However, the Apache distribution is not optimized to support snapshots (point-in-time images of the entire file system or sub-trees thereof, often used for error recovery), Network File System (NFS) integration, or real-time data streaming into the cluster. Such enhancements require re-architecting the underlying infrastructure.

## Management Capabilities – Cloudera

*Cloudera's focus on the management layer, to the exclusion of all other aspects of Hadoop that prevent it from being truly enterprise-ready seems a limited proposition.*

Distributions with added management capabilities, typified by Cloudera, seek to improve on one area of Hadoop while leaving the rest to the open-source community. Cloudera adds capabilities to Hadoop based on small variations in the trunk.

Cloudera offers a subscription service that packages Hadoop and supplies support and management software that helps administrators configure, monitor, and tune Hadoop. The company also offers training and consulting services that fill the gap between what the community can provide and what enterprises need to integrate Hadoop as part of their data management strategy.



Cloudera's focus on the management layer, to the exclusion of all other aspects of Hadoop that prevent it from being truly enterprise-ready, seems a limited proposition. Cloudera counts on the open-source community—including contributors from Cloudera, Hortonworks, LinkedIn, Hauwei, IBM, Facebook and Yahoo!—to innovate faster than vendors that change the distribution in a significant way.

Cloudera's leadership also asserts that the strong Apache governance model can resolve many of the conflicts in Hadoop's development path that may arise. However, recent history raises some questions since the flaws in Hadoop's architecture have not been addressed over the past 7 years.

## Enterprise Reliability and Integration – MapR Technologies

*MapR is designed to give Hadoop the same high-availability, scalability, and reliability as other enterprise systems.*

The enterprise reliability and integration category, typified by MapR Technologies, is built so that Hadoop connects to as many processing sources and consuming applications as possible. It also gives Hadoop the same high availability, scalability, and reliability as other enterprise systems. Hadoop lets organizations process data at scale, but requires other platforms to interpret and extract value from that data. The founders of MapR asserted that Hadoop was too important to leave it in an underdeveloped state. While acknowledging the power of open-source development, MapR operates under the belief that a market-driven entity is more likely to support market needs, faster.

While maintaining the core distribution, MapR has also conducted proprietary development in some critical areas where the open-source community has not solved Hadoop's flaws.

**Improving HDFS for high performance and high availability.** HDFS does not support enterprise-grade performance, and MapR sought to change that in a number of ways.

- MapR replaced HDFS so that it would not be reliant on Java or on the underlying Linux file system (see Figure 1).
- MapR's version of HDFS allows dynamic reading and writing, whereas the original HDFS is an append-only file system that can only be written once.
- MapR de-centralized and distributed the NameNode, increasing its capacity from 100 million files to 1 trillion. Because each node contains a copy of the NameNode, all nodes can participate in failure recovery, instead of requiring each node to call back to a central instance.

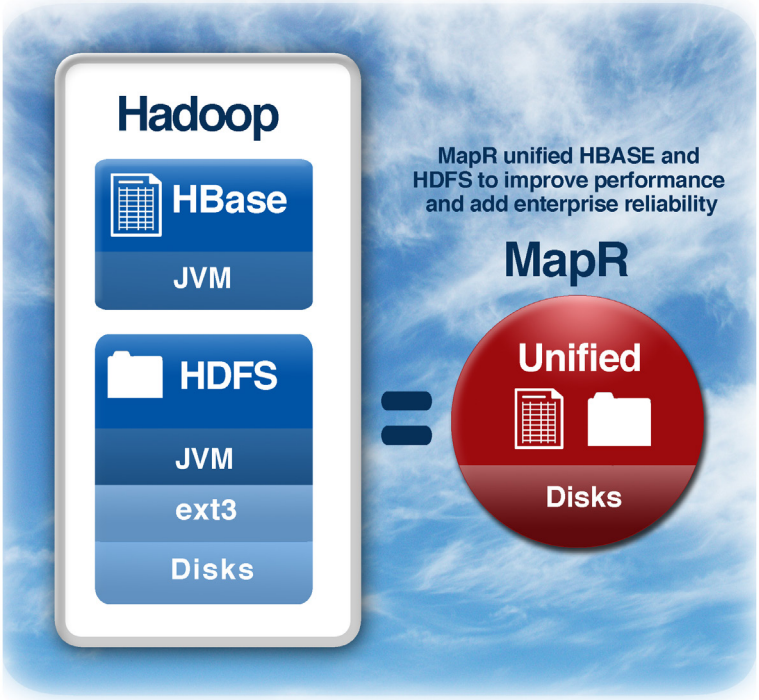


*Figure 1. MapR rewrote HDFS for enterprise usage*

*MapR offers data protection through the use of snapshots that offer point-in-time recovery of files and tables to protect against user or application errors.*

**Integrating with existing environments.** MapR provides links between Hadoop clusters and other common environments in the enterprise, including NFS and relational databases, so that Hadoop gains a full read/write storage system that can support multiple and full random readers and writers.

**Protecting data.** MapR offers data protection through the use of snapshots that offer point-in-time recovery of files and tables to protect against user or application errors. Files, as well as HBase tables, can be read directly from snapshots and recovered, avoiding costly downtime.

**Offering a management suite.** MapR offers a purpose-built management suite that eliminates the need to manually administer data availability, cluster health, or recovery.

In essence, MapR tamed the wild elephant of Hadoop for the enterprise.

Figure 2 summarizes the three distribution types and the vendors discussed here. Note that all vendors provide the core distribution.

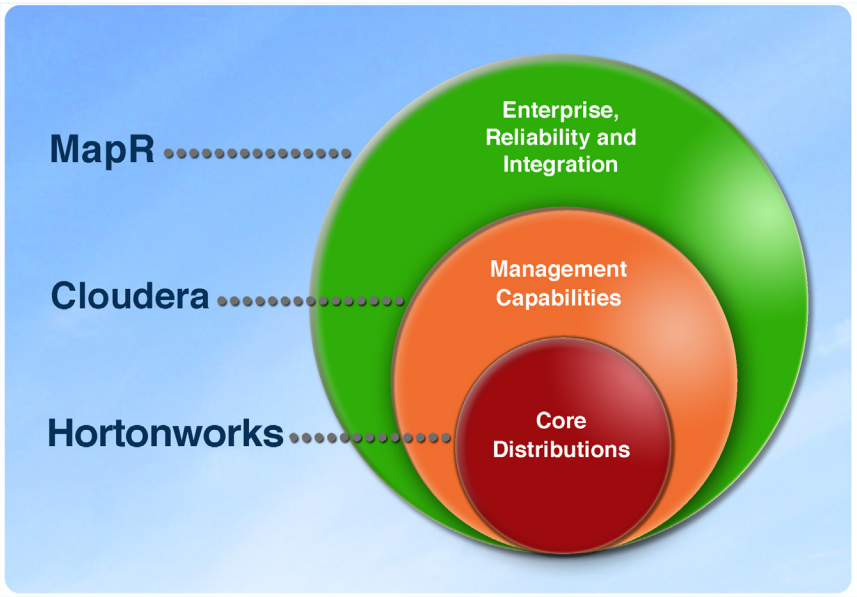


*Figure 2. Capabilities of Hadoop Distributions*

*An organization that has an investment in analytics tools that use SQL needs Open Database Connectivity (ODBC) interfaces.*

## Choosing a Distribution

Enterprises can benefit from each kind of distribution in the framework, depending on the level of sophistication they require. A few key questions to ask include:

**Will I get everything Hadoop has to offer?** All distributions offer Hadoop's core capabilities.

**How do I ensure that my general administrators can work with Hadoop effectively?** For this, you will want a distribution that has a management layer. Additional tools may be needed if integration with other tools is an objective.

**How will Hadoop fit into my environment?** An organization that has an investment in SQL analytics tools needs Open Database Connectivity (ODBC) interfaces. If you want to easily and transparently read and write data stored in a Hadoop cluster from enterprise applications that are not Hadoop-enabled, you need NFS support. To ensure integration with your environment, it's important to consider an Enterprise Reliability and Integration-level vendor.



**How do I ensure that developers and business analysts can access data in Hadoop?** For both of these groups, NFS and ODBC connectivity are important, so arrows point to an Enterprise Reliability and Integration-level vendor.

**How can Hadoop support my data-protection policies?** Hadoop has replication capabilities in the event of node failure. However, if errors are introduced, those errors replicate across the cluster. Enterprise Reliability and Integration level vendors have added point-in-time recovery to Hadoop to support recovery point objectives. Similarly, policies for acquisition of new enterprise systems typically mandate the ability to operate in a disaster recovery ready configuration, which requires mirroring, another feature supported only by Enterprise Reliability and Integration-level vendors.



## Conclusion

Open source software provides building blocks for applications that can serve a variety of purposes the creators may not have even imagined. CITO Research believes that if Hadoop is to become a useful enterprise-grade tool for extracting value and insights from big data, it cannot be more unwieldy than its relational database forebears.

No matter how innovative, very few systems flourish into practical use if they present functional obstacles to adoption. The choice of distribution depends largely on what represents an obstacle to a given organization's plans for Hadoop. For the organization that is interested in participating in the continued development of Hadoop, choosing a distribution that makes limited or no alterations to the trunk is a logical vote. For the organization that needs to connect Hadoop to multiple data analysis platforms, has a requirement for reliability, visibility, and control for mission-critical applications, a comprehensive solution that is available from a vendor in the here and now might make more sense.