

# Cloudera refreshes Hadoop lineup with version 4

**Analyst:** Matt Aslett

5 Jun, 2012

Apache Hadoop software and support specialist Cloudera has launched version 4 of its Cloudera Distribution including Apache Hadoop (CDH), the first from the company to be based on the Apache Hadoop 2.0 code, as well as v4 of its own Cloudera Enterprise support subscription and management software. Among the key capabilities in CDH4 are high availability, security and extensibility, while Cloudera Enterprise v4 adds the ability to manage multiple clusters and distribution versions.

## The 451 Take

We noted at the end of last year that 2012 was likely to be the year in which we would discover whether Cloudera had made enough of its first-mover advantage to thrive in an increasingly competitive market. As we head toward the halfway point, the signs are looking good. The lack of insight into customer traction means we cannot unfortunately fully gauge the company's growth. That aside, Cloudera continues to impress us with its improving products and expanding employee count and partner and customer bases, despite increased competition. The arrival of CDH4 addresses a number of common criticisms of Apache Hadoop and is likely to be an important factor in driving increased adoption of Hadoop among more conservative enterprises that haven't yet moved from testing to production.

## Context

As we noted in January, while the Apache Hadoop project hit a key milestone at the start of the year with the release of version 1.0 of the open source data-processing framework, there was much

more to come this year. That included the Apache Hadoop 0.23 code branch (which it seems is now officially known as Apache Hadoop 2.0), as well as the launch of supported distributions based on that code. It is fitting that the first of the commercial Hadoop distributors, Cloudera, is the first to bring a distribution based on Apache Hadoop 0.23/2.0 to market, with the release of Cloudera Distribution including Apache Hadoop (CDH) v4.

The core advances in CDH4 can be grouped into high availability, security and extensibility. The first has long been a black mark against Hadoop, but has been addressed with the ability to use a secondary NameNode as a hot standby for failover. Security is also an often-cited weakness of Hadoop but has been improved with the addition of table and column permissions for the Apache HBase database, as well as the use of a single Kerberos authentication scheme for all components in the Hadoop distribution. Extensibility is enabled by the delivery of the much-anticipated MapReduce 2.0 (also known as YARN, as well as NextGen MapReduce). Whatever you choose to call it, it is a new architecture that splits the JobTracker into its two major functions: resource management and application lifecycle management. The result is that multiple versions of MapReduce can run in the same cluster, and that MapReduce becomes one of several frameworks that can run on the Hadoop Distributed File System alongside the likes of Apache HAMA – the bulk synchronous parallel computing framework for scientific computations.

Also new in CDH4 are support for HBase coprocessors, which enables the creation of custom real-time applications; REST over HTTP access to HDFS; and a number of claimed performance improvements such as 100% improvement in file system I/O performance, 100% faster HBase random reads and 200% faster data ingestion using Flume. As well as the core CDH distribution, Cloudera has also updated its Cloudera Enterprise offering to v4. Cloudera Enterprise combines subscription-based support with the Cloudera Manager administration software. New in v4 is support for CD4, as well as mixed clusters of CDH3 and 4 (enabling phased migration); the ability to manage multiple clusters from a single Cloudera Enterprise instance; heat maps to provide a visual perspective on cluster health; and federated NameNode management. Also new is a Cloudera Manager API to enable integration with enterprise systems management tools/frameworks, as well as the ability to authenticate administrators against Active Directory via LDAP.

Cloudera is predicting rapid adoption of CDH4 and that its entire customer base is likely to move over to the new version within six to nine months. Unfortunately, the company continues to be shy about disclosing customer numbers, which has been the case since it passed 100 customers in April 2011. It's clear from other metrics that Cloudera is growing well, however. It now has 250 employees, up from about 160 in December 2011. Of those 250, 80 are involved in engineering and 25 in support. The company also now has more than 250 partners, with 100 or so of those being

systems integrators, 70 ISVs and the rest hardware, networking, OS and management players. The Cloudera University training and certification program is helping to seed the market with would-be users. The company has now trained more than 12,000 students, compared with 7,000 in December, and has licensed its training content to over 25 organizations worldwide.

## **Competition**

Cloudera may have been the first Hadoop distributor to market, but it certainly wasn't the last and the company faces a growing challenge from the likes of IBM, EMC, MapR and Hortonworks. We have previously noted that the company's expertise and early-mover advantage have enabled it to position itself front and center of the Hadoop ecosystem. IBM's recent decision to support CDH as part of its big data platform is testament to that, although we expect Big Blue to also support other distributions based on customer demand. With that in mind, we would see Hortonworks as the closest rival to Cloudera based on its expertise and strategy. Or at least it will be when its Hortonworks Data Platform becomes generally available.

Cloudera reports seeing EMC more than IBM in competitive situations, which wouldn't surprise us given that EMC seems the more aggressive of the two regarding Hadoop. Even then, we would expect EMC to compete more directly with Cloudera systems partners such as Oracle, NetApp, Dell and SGI. Cloudera is set to benefit as these and other partners bring Hadoop-based products and services to market, although there is also the potential for Cloudera's role to become marginalized as Hadoop becomes more mainstream.

MapR is also a core rival given the performance advantages claimed for its file system, high availability and workflow functionality, although Cloudera maintains that the Apache Hadoop project has now caught up based on improvements delivered in Apache Hadoop 0.23/2.0. Other potential competitors include DataStax with its combination of Hadoop and Apache Cassandra, LexisNexis' HPC Systems, Amazon's Elastic MapReduce and Mortar Data.

## **SWOT Analysis**

### **Strengths**

Cloudera has undoubted expertise regarding Apache Hadoop, and the recent partnership with IBM is validation of the position it has gained from its first-mover advantage.

### **Opportunities**

We have seen massive interest in Apache Hadoop, which is taking time to translate from development and testing into production. The core capabilities of Apache Hadoop 0.23/2.0 and CDH4 are likely to encourage greater adoption.

### **Weaknesses**

While we are sure that the company is growing well, the lack of insight into customer traction leaves a question mark about its growth.

### **Threats**

As Hadoop becomes increasingly mainstream and the core data management vendors add it to their portfolios, the profile of distributors such as Cloudera could diminish.