the 451 group

451 MARKET INSIGHT SERVICE | 29 June 2010 – Market Development

Cloudera delivers Enterprise extensions to Hadoop distribution

Analyst: Matt Aslett

Having positioned itself to become to the Apache Hadoop data-processing framework what **Red Hat** is to Linux, **Cloudera** has been busy working with early adopters to identify opportunities for delivering additional value beyond its core Cloudera Distribution for Hadoop (CDH). The launch of Cloudera Enterprise sees the company delivering additional value-added capabilities for provisioning and managing Hadoop clusters.

The 451 take

The addition of complementary projects to the Hadoop Core into CDH speaks volumes about Cloudera's desire to position CDH as the de facto enterprise Hadoop distribution. The company continues to take a leading role in driving and supporting the adoption of Hadoop, and with Cloudera Enterprise has also delivered the value-added extensions that should ensure that it can convert that adoption into revenue at a faster and more profitable rate than via support services and training alone.

We reported in early May that Cloudera had been busy strengthening its team and generating new customers for CDH, its distribution of the Apache Hadoop open source distributed data-processing framework. The vendor recently lost founder Christophe Bisciglia, who has decided that the time is right for a new challenge, but has continued to expand at a rapid rate and now has 43 employees, up from 37 at the start of May. It also now has 45 customers, up from 30, and is seeing strong demand for Hadoop and MapReduce as a complement to its existing data-warehousing/analysis techniques, especially among Web application and services providers, as well as financial services firms.

The company has also been hard at work developing CDH version 3 and Cloudera Enterprise, both of which are now available. As expected, CDH3 includes improvements regarding the Pig data analysis tool and Hive data-warehouse platform, which are both subprojects of Apache Hadoop. More than that, though, with CDH3 the company has pulled together a number of the projects associated with the Hadoop Core into CDH3. Besides Hive and Pig, they include the Oozie workflow engine, the ZooKeeper coordination service, the Avro data serialization system and improved support for the HBase distributed database.

Also new in CDH3 is Cloudera Desktop, the desktop UI for file and job browsing, cluster monitoring and job designing. Cloudera Desktop was previously available separately, but is being rolled into CDH and released as open source software as part of the vendor's increased contributions to the Apache project. Cloudera notes that 50% of its development

effort is dedicated to open source projects, and also cites the Flume systems for collecting streaming event data. The 50% of its development effort not focused on open source is concentrated on Cloudera Enterprise, which brings together CDH and the already available production support with new proprietary tools. In the first instance, those include authorization management/user provisioning, resource management, data integration and configuration management.

The Sqoop data import tool also continues to be part of CDH, and Cloudera recently announced a partnership with **Quest Software** to develop a Sqoop-based connector to provide bidirectional data transfer between CDH and **Oracle** databases. Known as Ora-Oop, the connector will support Oracle data types and table variants and will be freely available from both Quest and Cloudera, while Quest will also make CDH available via its website. Ora-Oop is expected to be the first of many targeted data transfer connectors.

Competition

Cloudera may have been the first player set up to build a business around Hadoop, but it certainly wasn't the last. The company maintains that the likes of **Datameer**, **Karmasphere** and **IBM** are potential partners rather than direct competitors and that demand for Hadoop-based deployment expertise and analytics will provide opportunities for everyone. We would agree on the whole, but the lines are beginning to blur: IBM has its own distribution, while Karmasphere has released a Client product that we would see as comparable to Cloudera Desktop (which may partly explain why Cloudera has decided to open source the latter). Datameer offers Hadoop support as well as spreadsheet-based analysis tools.

The true sign that Cloudera has achieved the dominance that its first-mover advantage made potentially possible will be if it can persuade these potential partners to support CDH, rather than the Apache distribution, for example. The expansion of CDH to include Hive, Pig, HBase, Oozie, ZooKeeper and Avro support may be critical to that differentiation. While we would see Hadoop as largely complementary to traditional data-warehousing techniques, there is some overlap with the approaches from **Greenplum** and **Teradata**, for example, to storing and analyzing a company's raw data, rather than simply the data kept in the enterprise data warehouse.