

REPORT REPRINT

IBM drives deployment choices with latest Hadoop distribution release

JAMES CURTIS

07 OCT 2016

Big Blue recently released BigInsights 4.2, showing its ongoing commitment to Hadoop. With three other commercial Hadoop vendors in the market, what is IBM doing to maintain its differentiation?

THIS REPORT, LICENSED EXCLUSIVELY TO IBM, DEVELOPED AND AS PROVIDED BY 451 RESEARCH, LLC, SHALL BE OWNED IN ITS ENTIRETY BY 451 RESEARCH, LLC. THIS REPORT IS SOLELY INTENDED FOR USE BY THE RECIPIENT AND MAY NOT BE REPRODUCED OR REPOSTED, IN WHOLE OR IN PART, BY THE RECIPIENT, WITHOUT EXPRESS PERMISSION FROM 451 RESEARCH.



©2016 451 Research, LLC | WWW.451RESEARCH.COM

IBM is one of four vendors that continue to offer a commercial distribution of Apache Hadoop. In July, the company released its latest version of BigInsights 4.2 with a handful of updated projects, along with the addition of Apache Ranger, Apache Phoenix and Titan. More so, however, IBM has been actively driving a variety of Hadoop consumption options that include cloud and hybrid, as well as driving a broader collection of IBM products and services that can integrate with Hadoop.

THE 451 TAKE

BigInsights was introduced in 2010, clear evidence that IBM has been committed to the Hadoop distribution game for some time. Our belief is that BigInsights lags a bit in terms of market share compared with the other Hadoop distributors. However, that could be changing. The rebranding of BigInsights in 2015 to better delineate the product's proprietary pieces versus the new fully open source distribution is helping IBM earn the respect of the Hadoop community. IBM's greatest strength lies in its broad portfolio of products, which the company can then offer as integration points with Hadoop. While the Hadoop managed service market is competitive, we believe IBM has an opportunity to find success in the Hadoop managed service space, given that the company controls its Hadoop distribution and owns its cloud platform, enabling IBM to greatly optimize the cloud service.

CONTEXT

IBM first offered its Hadoop distribution back in 2010, then referred to as InfoSphere BigInsights. While Hadoop just celebrated its tenth anniversary, the market has seen a number of Hadoop distributions come and go over the past ten years. We recently reported that the market has pared down to just four vendors that offer a commercial Hadoop distribution: Cloudera, Hortonworks, MapR and IBM. Pivotal was the most recent vendor to exit the distribution market in April and now rebrands and resells Hortonworks' distribution.

With IBM being in the Hadoop distribution game for the better part of six years, IBM management reports that there was some confusion around the BigInsights distribution and specifically what the company provided as open versus proprietary components. As such, IBM addressed this point by rebranding its BigInsights distribution in 2015, which we detailed in a previous report. The result was a new 100% open source distribution offering that the company calls the IBM Open Platform with Apache Hadoop (IOP). At the core of the IOP distribution is the ODPI runtime specification that consists of Hadoop common, HDFS, YARN and MapReduce components. While ODPI is run by the Linux Foundation, it began initially with Hortonworks, IBM and Pivotal to implement basic Hadoop standards. However, not all of the Hadoop distribution vendors are aligned with ODPI, notably Cloudera and MapR.

PRODUCTS

In July, IBM announced a number of enhancements to the latest IOP 4.2 distribution. New components include Apache Ranger (v0.5) that provides a security framework for security-related tasks, Apache Phoenix (v4.6.1) that provides an HBase SQL interface enabling OLTP as well as operational analytical queries, and Titan (v1.0.0) that provides a graph database API. All of these components, IBM maintains, were often requested by its customers. Additionally, IBM has identified other open source projects that customers may also add, that IBM has tested, but aren't officially included or supported in the IOP distribution, for instance, Hue, Storm, Cascading and Tez, to name a few.

What about IBM's proprietary, value-add components? IBM offers these components as IBM BigInsights for Hadoop, which is not so much a separate distribution, but a collection of value-add components that are layered on top of the IOP distribution. These value-add pieces – Text Analytics, Big SQL, and BigSheets – essentially extend Hadoop (and the IOP distribution) for enterprise capability as well as enable certain data-driven analytic use cases.

For instance, Spark, which is part of IOP, can be used in conjunction with BigInsights Text Analytics to enable text mining for competitor analysis, insurance claim assessments and sentiment analysis. Big SQL can be used for relational database offload and consolidation, which can then later enable users to query Hive and HBase for further analytics. Still other IBM products, such as IBM SPSS, IBM Streams, IBM BigIntegrate and BigInsights BigReplicate,

can likewise be integrated with BigInsights. Essentially, IBM's strategy is to make Hadoop the core and then build value-add components on top of the IOP distribution as well as drive integration with other IBM products.

In addition to an updated Hadoop distribution, IBM is driving a number of Hadoop deployment choices as well. The on-premises option continues to be available with Avnet and Cisco partnerships. However, IBM also provides several cloud options, available on SoftLayer and IBM Bluemix.

For customers looking to have some control over their Hadoop cluster, there is the option to deploy on SoftLayer infrastructure in a 'bring your own license' scenario. For customers not wanting to deal with cluster administration, there is a fully managed service option available on IBM Bluemix, referred to as BigInsights on Cloud.

Customers can choose a node configuration (small, medium or large) based a particular analytic use case, given that the value-add components of IBM BigInsights are also part of the package. IBM provisions the servers and takes care of cluster security, support and any management, including patches and upgrades. While the service is offered on Bluemix, the underlying infrastructure is driven by SoftLayer.

As such, IBM claims to provide certain performance and cost benefits for customers, such as splitting compute and storage and running on bare-metal hardware instead of virtualized servers. A beta version for a Basic Plan, leveraging Docker containers, was recently introduced that offers the IOP Hadoop distribution, albeit with restricted data requirements and basic HDFS encryption.

COMPETITION

Besides IBM, there are three other commercial Hadoop distributors that include Cloudera, Hortonworks and MapR. Other than Hortonworks, which implements a purely open source business model, the others – Cloudera and MapR – offer proprietary components that are added to their respective distributions.

Regarding the cloud, all of the other distributors have cloud-based Hadoop offerings. Cloudera leverages its Director tool to deploy and manage Cloudera Enterprise in the cloud, including the ability for multi-cloud deployments for AWS, Microsoft Azure and the Google Cloud Platform. Just recently, Cloudera announced AWS S3 support from Impala. Hortonworks has its Cloudbreak tool for provisioning Hadoop in the cloud to AWS, Microsoft Azure, Google Cloud Platform and OpenStack. Hortonworks also maintains a partnership with Microsoft Azure, which offers a managed Hadoop offering as Azure's HDInsight. MapR offers its distribution on Azure and AWS Marketplaces, and is available as an option for Amazon EMR users. MapR can also be deployed on Google Cloud Platform, OpenStack and CenturyLink.

Other than Hortonworks, which partners with Azure, IBM differentiates itself among the Hadoop distributors in that the company owns its managed Hadoop service based on the company's cloud platforms – SoftLayer and Bluemix. However, there are other managed Hadoop offerings. Altiscale (recently acquired by SAP) offers a fully managed Hadoop and Spark service driven by the company's own datacenters. AWS with its EMR offering is another, as is Google with its Cloud Dataproc offering. We can also identify a handful of smaller companies – sometimes referred to as data integration vendors – as possible competitors to IBM, which include Qubole, Treasure Data, Cazena and Xplenty.

While IBM competes with the other Hadoop distributors, the company continues to compete with its traditional rivals in Oracle, Microsoft and Teradata, all of which have significant investments in Hadoop, although none maintains a Hadoop distribution. Oracle has a Big Data Appliance but the company also promotes its Big Data Cloud Service that includes a collection of integrated tools. Teradata likewise maintains a Hadoop appliance and has been active with integrating other tools, such as Aster on Hadoop. Microsoft maintains its cloud-based HDInsight offering on Azure.

SWOT ANALYSIS

STRENGTHS

IBM has the benefit of being a full ecosystem vendor in that not only does the company manage its own Hadoop distribution but it also owns the cloud infrastructure for which it runs its Hadoop deployments. A history of acquisitions and product innovations likewise give IBM a broad portfolio in which to leverage Hadoop.

WEAKNESSES

While BigInsights has been around for six years, IBM rebranded BigInsights in 2015 to better delineate the product's open source and proprietary components, helping to broaden the product's appeal beyond the company's install base.

OPPORTUNITIES

The managed Hadoop service space has a collection of vendors but is not overly competitive. With IBM's resources and the fact that it owns its own cloud platforms, there could be some opportunities for IBM to make a move.

THREATS

IBM not only competes with the three other Hadoop vendors but the company also competes with its larger rivals such as Oracle, Microsoft and Teradata. The challenge for IBM is in staying competitive at all levels to all of its competitors that, to some extent, have significant investments in Hadoop and the analytics space.