

# Evaluating Hadoop in the Data Center

What will make Hadoop an enterprise data center-grade analytics platform?

By John Webster

October, 2012



**Evaluator Group**

*Enabling you to make the best technology decisions*



## Enterprise IT Encounters Hadoop

The demand for Big Data analytics processes by enterprise executives and business group leaders is growing at a rapid pace. However, the challenge to enterprise IT from the Big Data analytics perspective lies in capturing data from multiple new data sources (web, social media, mobile devices, etc.) and performing analytical processes against this data to unlock informational value. Traditional data warehousing technology was not designed to process large volumes of unstructured and structured data in relatively short periods of time. New approaches are required. Enter Apache Hadoop.

Of the numerous ways that have emerged in the last few years to do Big Data Analytics, none has gotten more media attention than Hadoop which began conceptually with a paper that emanated from Google in 2004 that described a process for parallelizing the processing of web-based data—MapReduce—and a supporting file system. Shortly thereafter, Apache Hadoop was born as an open source implementation of the MapReduce framework with its supporting file system—the Hadoop Distributed File System (HDFS). The Hadoop community is now growing dramatically and is producing enhancements and additions that expand its usability within enterprise datacenters.

As Hadoop enters the enterprise data center, IT administrators should review and analyze its architecture from the dual standpoints of: 1.) the infrastructure required to instantiate the platform and 2.) the processes required to make it conform to established data center management practices. While Hadoop is distinctly different from traditional data warehousing platforms, IT administrators are likely to evaluate Hadoop on the basis of established management requirements and suitability to enterprise production data centers—as they would with the infrastructure components that support other data center-resident applications.

## Evaluating Hadoop in the Production Data Center

We believe that it is safe for IT administrators to assume that Hadoop, once established as a viable and potentially valuable business analytics platform for user groups, two things will happen:

1. The number of user groups wanting Hadoop-based applications will grow within the enterprise from one or two to many. For example, the marketing department sees value in analyzing customer data to improve customer targeting. The product group could then see translating the same value they are deriving from Hadoop analytics and using it to improve product quality.
2. As business groups build applications on Hadoop, they will over time, come to increasingly depend on them—a process that will push these applications into the mission-critical category. Email was once a novelty. It is now mission critical and is managed at multiple levels within the enterprise including Auditing, Security and Legal.

It is for these reasons that we believe that enterprise IT begin to evaluate Hadoop and its supporting infrastructure using the same criteria that would apply to other production data center-resident applications that they are responsible for. These criteria include:

### Hadoop Cluster Availability

The Name Node in version 1.0 of Apache Hadoop, which among other things, stores metadata for the Hadoop cluster, has become a well-known single point of failure. If the Name Node fails, data could be lost. Similarly if the JobTracker fails, the job has to be reloaded and rerun from the beginning. Currently, support for a standby Name Node for Apache Hadoop is to be delivered in version 2.0. However, this version is still in test and will protect against a single failure. Alternatives currently exist that eliminate altogether the Name Node and its vulnerability with automated failover and ability to protect against multiple and simultaneous failures and should be considered.

Planned upgrades to Apache Hadoop also currently require outages. During initial phases of Hadoop deployments, this may not be an issue. But if, as mentioned, users come to depend on their analytics applications, outages of any kind will become sources of contention with IT. Again, we note that alternatives are currently available that support rolling upgrades with no downtime.

### Data Protection and Integrity

By default, Apache Hadoop generates and maintains three copies of data: one primary and two clone copies. These copies are placed on different server nodes and processed during the MapReduce process to address potential performance or failures. RAID is typically not used for Hadoop's disk storage that is embedded within Data Nodes. Shared storage (NAS/SAN) is regarded as adding another network between the data and compute layers, and is expensive relative to JBOD (just a bunch of disks) which is typically used. These copies allow the cluster to function without the online presence of a Data Node with the failed disk.

However, in spite of the fact that Apache Hadoop maintains three copies of data available to the cluster at all times, exposure to data loss in Apache Hadoop is still a major concern. For example, if the primary copy of data becomes corrupted, the corruption will be propagated to the other two copies. Therefore, trying to recover the cluster back to a known good state before the corruption occurred won't work using one of Hadoop's copies as would be the case if one or a number of disks within the cluster was responsible for "silent" data corruption.

The ability to use point-in-time snapshots to restart a process from a point before the corruption occurred has long been available to enterprise storage administrators. While this has been planned for some later release of Apache Hadoop for the last four years, it is still not available. To get it now or in the near future, potential enterprise Hadoop users will have to seek an alternative that does support data snapshots. And, aside from the exposure to data loss, enterprise IT administrators could regard the

requirement to maintain three full copies of data available at all times as an inefficient use of cluster resources

Finally, it should be noted that none of Apache Hadoop's copies are mirrored to a physically remote location. Therefore, Hadoop cannot be included under the enterprises' disaster recovery/business continuance plan without having support for remote mirroring. There is currently only one Hadoop distribution that provides snapshot and mirroring support.

## Manageability

Enterprise IT staff encountering Hadoop for the first time will likely encounter a learning curve. Just how steep that learning curve is will depend on a number of factors including:

- Whether or not additional expertise will be needed to integrate Hadoop's many software components (Pig, Hive, Sqoop, etc.)
- Whether or not additional expertise and custom connectors rather than standard interfaces will be needed to integrate data sources with Hadoop to import/export data sets to/from existing systems
- Hadoop's management interface and the relative ease with which IT staff can manipulate familiar data structures such as volumes, problem recovery modes, and data protection processes

Also worth considering is whether or not additional staff with Hadoop expertise need to be hired or outsourced.

## Conclusion

We believe that the newly emerging data analytics processes driven by Hadoop will push well beyond the web-based social media environments where Hadoop has become a standard platform and into the well-established, production data centers of the Fortune 1000 that in some cases will be "lights-out" environments. As this happens, the proponents of Hadoop will encounter a different mindset. Here, IT administrators will bring to the evaluation process a greater appreciation for efficiency, manageability, and quality of service as experienced application users. We note that while the contributors to Apache Hadoop are at least aware of these requirements, progress toward addressing them has moved at a snail's pace. This leaves the door open for alternative implementations of Hadoop to arise and thrive.

## About Evaluator Group

*Evaluator Group Inc. is dedicated to helping **IT professionals** and vendors create and implement strategies that make the most of the value of their storage and digital information. Evaluator Group services deliver **in-depth, unbiased analysis** on storage architectures, infrastructures and management for IT professionals. Since 1997 Evaluator Group has provided services for thousands of end users and vendor professionals through product and market evaluations, competitive analysis and **education**. [www.evaluatorgroup.com](http://www.evaluatorgroup.com) Follow us on Twitter @evaluator\_group*

## Copyright 2012 Evaluator Group, Inc. All rights reserved.

*No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or stored in a database or retrieval system for any purpose without the express written consent of Evaluator Group Inc. The information contained in this document is subject to change without notice. Evaluator Group assumes no responsibility for errors or omissions. Evaluator Group makes no expressed or implied warranties in this document relating to the use or operation of the products described herein. In no event shall Evaluator Group be liable for any indirect, special, inconsequential or incidental damages arising out of or associated with any aspect of this publication, even if advised of the possibility of such damages. The Evaluator Series is a trademark of Evaluator Group, Inc. All other trademarks are the property of their respective companies.*