GIGAOM RESEARCH

How Hadoop passes an IT audit

John Webster

January 16, 2014

This report is underwritten by WANdisco.

TABLE OF CONTENTS

Executive summary

Not originally created for the enterprise environment, Hadoop was built for internet data center environments like Google, Yahoo, Facebook, and Twitter. These are different IT environments, structured, supported, and managed in different ways from enterprise IT. As a result, Hadoop currently lacks many of the functions and internal processes that enterprise IT needs in terms of security, availability, data integrity, and data governance.

There is no question that there are enterprise-level industry segments where Hadoop has taken hold and is flourishing, such as financial services, health care, pharmaceuticals, and energy. Most deployments are in departments with centralized IT becoming involved from the standpoint of providing and integrating infrastructure (servers with embedded storage, networking gear, etc.). In addition, these grassroots Hadoop projects are still mainly at a secondary level and are not yet considered to be critical, production-level IT.

Hadoop must mature further in order to be regarded as a viable enterprise platform capable of supporting critical business functions running real-time applications. As Hadoop matures so will its criticality within the organizations that are now learning its ins and outs. Enterprise IT will become more directly involved with managing and supporting Hadoop -- a process that is by no means a given. In essence, Hadoop has to follow the rules of centralized IT, and therefore the platform will be subject to production data center security levels, management processes, data protection and data integrity guarantees, data governance policies, and, above all, service level agreements (SLAs).

This report will:

- Place Hadoop in the context of enterprise IT and help those managing Hadoop platforms make it responsive to the enterprise's data governance policies and processes
- Outline these policies using the industry segments and data sources mentioned above
- Describe ways in which Hadoop can be made responsive to the enterprise's IT infrastructure, security, auditing, and compliance stakeholders
- Make the point that by addressing these concerns, Hadoop can advance to full production status, including support for real-time applications

IT and the enterprise risk environment

Enterprise IT is usually engaged in trying to reconcile two forces that appear to be in conflict with one another: mitigating risk versus creating new business opportunities. The former looks backward while the latter looks forward. By far, the force that drives IT most frequently is maintaining existing application availability as an element of managing risk. Loss of a critical application for even a short duration can result in lost revenue, lost productivity, user group dissatisfaction, and worse for high-profile companies -- general public awareness of an outage.

The other less powerful force is the one driving IT to create new business applications. This is unfortunate because revenue growth is limited without an ability to capitalize on emerging business opportunities. Without new applications, business can stagnate. Yet IT typically starves this essential function. The phenomenon can be seen in IT budgets where an 80/20 rule applies: 80 percent of the budget goes to maintaining existing infrastructure and applications. The remaining 20 percent goes to new projects that are further down on the list of priorities. The same can be said of IT management staff time, most of which follows the same 80/20 rule.

In a world where enterprise IT budgets are generally flat or only increase incrementally from year to year, tipping the balance toward new revenue generation can have an obvious payback. However, this rebalancing act is done with the understanding that an awareness of risk comes first. The only way to steer more budget dollars toward the opportunity side of the balance sheet is to reduce the overall cost of managing risk and maintaining IT's status quo.

Here, we look specifically at the regulatory and legal environment that envelopes enterprise IT as an element of risk. Next, we apply an understanding of this environment to Hadoop as a platform that is now being used to generate new business opportunity. As a result of this discussion, we hope to make it clear that in order for Hadoop to progress from pilot project status to the production application environment, it cannot appear to add to the enterprise's risk profile. Like other IT production applications and platforms, Hadoop must be responsive to the requirements of the enterprise's data governance policies and procedures that are designed and enforced to mitigate risk. Yes, this whole discussion sounds boring until one understands that failure to comply can and has resulted in fines and settlements exceeding a billion dollars and has landed some executives in jail. The potential risks for enterprise Hadoop users are significant.

Regulations and more regulations for IT

The challenge for IT administrators and IT auditors in particular when addressing regulatory compliance is that there is typically not just one regulation that applies to a given IT-related issue, be it security, records retention, or business continuance. Depending on the industry segment, a number of regulations -- sometimes conflicting -- can apply, resulting in situations where the same data can be subject not only to multiple regulations, but multiple regulations from multiple sources. This is particularly true for public companies in certain industries -- they are subject to both industry-specific requirements as well as SEC regulations that apply to all publically traded companies. And, with IT becoming a critical supporting function across industry segments, issues relating to IT security and continuity as well as data governance and retention are increasingly addressed.

One of the best ways to assess the magnitude of regulatory compliance in enterprise IT is to look at some of the better-known regulations and regulatory bodies. A sampling of agencies, industry organizations, and legislative actions that specifically address IT-related issues include:

Regulatory Agencies and Industry Organizations	Legislative Actions
Securities and Exchange Commission (SEC) -- administers SEC-established regulations	Dodd-Frank (IT security and disaster recovery for the financial services industry)
Health and Human Services (HHS) -- administers HIPAA	Sarbanes-Oxley (financial reporting and records retention for publically held companies)
Federal Energy Regulatory Commission (FERC)	Health Insurance Portability and Accountability Act (HIPAA) for the health care industry
Food and Drug Administration (FDA) -- administers 21 CFR Part 11	California SB 1386 and similar regulations established by most other states that target data breaches (security of personal information held by any organization)
Internal Revenue Service (IRS)	21 CFR Part 11 (health care and life sciences electronic records and signatures)

National Institute of Standards and Technology (NIST) Office of the Comptroller of the Currency (OCC)	
Basel Committee on Capital Accords	
International Organization for Standardization (ISO)	
Uniform Commercial Code (UCC)	

Regulations must be addressed as a function of business risk management that requires additional IT administrative effort, which can often draw budgetary and personnel resources away from other IT projects and initiatives. However, the penalties for noncompliance can be devastating to both companies and individuals, so the investment is necessary.

What Hadoop needs to be responsive

Hadoop must mature in order to be regarded as a platform capable of supporting critical business applications within the enterprise. Maturation in the context of this discussion means an ability to respond to the general requirements of IT administrators as well as audit, security, and records management officers in particular. Specifically, Hadoop must at least begin to address the basic enterprise-level security, disaster recovery and business continuity, and records management requirements that generally apply to real-time business applications.

Security

Data security in a number of different forms is perhaps the most common requirement included in a wide range of legislative and industry-specific efforts to regulate the collection, processing, and storage of certain types of data. Most if not all of these regulations have requirements around information security. Control of who has access to the information is one of the basic requirements. General requirements commonly include:

- **Limit system access to authorized individuals**. Authentication for authorized access to information must be implemented.
- **Audit trail of access**. An audit-trail log for reporting on access to information must be maintained.

Language similar to the above is included in HIPAA, SEC 17a-4, Sarbanes-Oxley, and 21 CFR Part 11. The majority of states in the U.S. have also adopted legislation that requires companies to disclose situations when security has been breached and personal data has been exposed to theft or unauthorized use. And HIPAA is particularly stringent with regard to patient records data.

The lack of native "wire-level" security in Apache Hadoop is acknowledged by its community of developers. They point out that Hadoop users have the ability to use the Kerberos network protocol, which is designed to provide authentication by using secret-key cryptography and the assignment of keys. However, the use of Kerberos in the context of Hadoop can be an issue for large enterprise and public sector IT administrators for at least two reasons. First, login authentication is controlled by a centralized key distribution center (KDC). An attacker could potentially "hack" the KDC and impersonate any authorized user. Second, because of the way Kerberos is architected, a different set of host keys will be required for every node in the Hadoop cluster, adding an additional layer of administrative complexity.

What Hadoop needs is the implementation of native user authentication based on a mechanism that initiates and maintains a secure connection. Server-to-server communication, including intercluster nodes communication and remote procedure calls (RPCs), also needs to be secured. From an IT operational standpoint, native security should be applicable to:

- User operations such as file reads and writes, database manipulations, and MapReduce job submissions
- Intracluster node-node interactions, including RPCs
- Intercluster operations, such as mirroring

Hadoop security should therefore not depend on the implementation by the user of a KDC or any other third-party mechanism. If native authentication keys are used, the Hadoop administrator should have the ability to implement a system whereby the same keys can be used both within and across Hadoop clusters. Access control could also be addressed by implementing full POSIX control on files and directories under Hadoop. Access control lists (ACLs) could be applied to:

- Tables, column families, and columns
- Clusters and volumes
- MapReduce jobs and queues

Thus far, the discussion has focused on user authentication and access control. However, the security of data "at rest" is also an issue for a growing number of enterprises. For them, disk manufacturers now offer disk-level data encryption. This is effective when controlling data exposure after a disk has been removed from a server, for example. Because large Hadoop clusters require frequent disk replacements, we anticipate a growing requirement for the use of disk-level encryption with Hadoop.

Disaster recovery and business continuity

The banking and financial services industry is regulated by a long list of federal, state, and industry-specific agencies. As a result of events like 9/11 and Hurricane Katrina, the agencies that regulate the banking and financial services industry now either require or strongly recommend the implementation of disaster recovery and business continuance capabilities for IT systems.

Health care organizations also see IT disaster recovery and business continuance capabilities written into regulations. HIPAA requires an applications and data criticality analysis, a data backup plan, a disaster recovery plan, an emergency mode operation plan, and a testing and revision procedure. The FDA's *Guidance on Computerized Systems in Clinical Trials* requires "written procedures that describe contingency plans for continuing the study by alternate means in the event of failure of the computerized system." Finally, the Federal Energy Regulatory Commission (FERC) is currently developing an Office of Energy Infrastructure Security that will address a number of threats to electricity, natural gas, and oil delivery systems, including the continued availability and recovery from loss of supporting IT systems.

In enterprise IT it is common to see risk mitigation and compliance processes relating to the management of enterprise data rooted in the storage environment. It is here that some essential functions can be applied directly to data. These essential functions include:

- **Data protection**. Backup copies are created and maintained on either the primary or, more often, secondary storage devices and are used to recover from any event that could result in data loss or corruption.
- **Local data copy**. Clones and snapshots are used to recover from adverse events and to propagate data to other applications and test environments.
- **Remote data copy**. Data is replicated over MAN and WAN distances directly (i.e., without server-to-server transfers of data) to other storage devices. It is used for the same purposes as local data copy and to establish a disaster recovery site under the enterprise's overall disaster recovery plan.
- **Archiving**. An immutable data copy is retained within a storage device (a storage system, for example) and used to satisfy regulatory compliance mandates as noted above and comply with rules covering the management of evidence stored in electronic form as mandated by the Federal Rules of Civil Procedure (FRCP).

For Hadoop clusters, the place where these functions would normally be implemented is in Hadoop Distributed File System (HDFS). However, on its own and without assistance from other entities such as shared storage systems that already have this functionality, HDFS can only address a subset of these functions as follows:

- **Clone copy and snapshot copy.** Hadoop makes local copies of data (three by default), meaning that for every file ingested, additional full copies are created and stored within the cluster. These are essentially clones of the entire file system that Hadoop administrators use to reduce cluster-processing latency and to recover from various types of failures within the cluster.

  However, the practice of maintaining three copies of the entire file system does not offer complete data protection. One example is if the original copy is corrupted as the result of uncorrectable read errors (UREs). The disk detects many, if not most, read errors on a read. These cannot be recreated or propagated. While UREs are therefore statistically infrequent, they are of concern in the context of Hadoop for two reasons. First, large Hadoop clusters can use thousands of disks, increasing the possibility of an undetected occurrence within a given time period. Second, a RAID controller could be used to detect these errors when they go undetected by the disk, but since the disk is typically implemented as JBOD (just a bunch of disks), a RAID controller won't catch them. The error could propagate across to the other copies, which would render them useless as a means to recover from this type of event.

  The ability to create logical (snapshot) copies of data rather full physical copies is supported in version 2 of Hadoop. Using snapshots offers an alternative recovery mechanism to the use of full data copies generated by HDFS to backup data within the cluster, recover from user errors, and in limited disaster recovery scenarios. However, because of the metadata copy mechanism used, HDFS snapshots cannot be used to recover from the event described above.

- **Replication using Hadoop DistCp.** A native Hadoop function called DistCp (distributed copy) can be used to replicate data from one Hadoop cluster to another, either locally or over MAN/WAN distances. It uses MapReduce processes to implement a read-only mirrored copy from source to target. By default, the DistCp process skips over ones that already exist at the target and ones that are being written to when the DistCp job is running. Only a count of skipped files is reported to the administrator at the end of each DistCp job, and even this minimal level of reporting could be inaccurate if DistCp failed for some subset of its files but succeeded on a later attempt. Therefore, administrators must manually run successive DistCp jobs to capture and replicate file updates as well as cross-check a listing of the source and target files to verify that the copy was successful. Even so, because open files are not copied, inconsistencies between the source and target file set will be created unless the cluster is in a state where no writes are taking place when the DistCp job is running.

> In addition, other issues between the source and target clusters could adversely and silently affect the copy, creating inconsistencies between the source and target clusters. Other undetected inconsistencies between source and target files can and do occur because DistCp is not aware of the content of files. It makes the decision to copy or not to copy based solely on file name and size. If the file name and size match, it does not matter if file content is different (e.g., it was updated subsequent to a previous DistCp job). DistCp will not copy it over to the remote read-only mirror.

It is possible to use shared storage systems in conjunction with Hadoop clusters that offer, as part of their of data services, a set of local mirroring, snapshot, and remote replication functions that have been used successfully for decades to support enterprise production-level data protection and disaster recovery processes. These can be used in place of or in addition to the data copy and replication functions available with Apache HDFS that enterprise IT administrators find inadequate.

However, the use of shared storage systems with Hadoop is rare and controversial. Perhaps a more acceptable route to take, at least in the short term, would be to make HDFS more robust with regard to local and remote data copy functions. Doing so could include the implementation of active-active data replication that could be used over LAN, MAN, and WAN distances in such a way that consistency across source and target copies is preserved without manual intervention from the Hadoop administrator. This would also allow a single Hadoop cluster to be "stretched out" over MAN and WAN distances. The solution would also have to avoid the use of a centralized transaction coordinator that could be both a single point of failure and a performance bottleneck.

Records management: the regulatory and legal requirements

Many regulations written for the financial services, health care, and pharmaceutical industries contain language that addresses the storage and retention of electronic records. Requirements, retention periods, and definitions of what constitutes an electronic record vary, but SEC 17a-4 is an often-cited example. In this legislation, records are defined as all documents relating to business activities. Today, the definition would include such items as email and text messages, reports, and transaction logs. Electronic media used to store records must preserve them in a nonrewritable, nonerasable format, such as write once read many (WORM) technology. Records must also be easily searchable and retrievable. Many of these records must be maintained for not less than three years and must be easily accessible to the SEC for the first two years. Some records have longer required retention periods.

Additional requirements

While there are numerous regulations -- some of which have been noted above -- that require the long-term retention and retrieval of records when requested by an administrative authority (like the SEC), there is another category of records retention and retrieval requirements often referred to as e-discovery. In the U.S., the term implies that IT has a mechanism to find and retrieve records that are sensitive from a legal perspective.

In fact, the majority of e-discovery requirements that impact enterprise IT do not emanate from a regulation at all. Rather, they come from legal directives regarding the discovery of evidence that are outlined in the FRCP. In anticipation of litigation, all enterprises are subject to FRCP rules that require the disclosure of evidence when stored in electronic form. This applies to all companies whether public or private and regardless of size.

One of these requirements is referred to as "legal hold," meaning that the records identified as the result of an e-discovery process must now be preserved as evidence. A legal hold requires an organization to gather and preserve data from the entire set of information resources, which includes archives, databases, email, and other information repositories. Furthermore, this data may not be deleted or altered, and any retention expiration date applied to this data must be held in abeyance until the legal hold is removed. Legal hold is an implicit requirement to preserve evidence, including that which exists in electronically stored form. The penalties for noncompliance can be severe. Penalties of millions of dollars can result from the lack of timeliness of the discovery process (FRCP guidelines give 48 hours to produce a list of the information available) and/or deleting or failing to produce information on demand.

Unfortunately, it would appear that the need for data governance and electronic records management functions capable of supporting enterprise-level regulatory compliance and e-discovery requirements were not anticipated by the developers of Apache Hadoop. These features do not currently exist in version 1 or 2. Their absence has become a major stumbling block for the progression of Hadoop into enterprise production IT environments. For example, Hadoop users typically want to use customer data for analytics, but its use often falls under the scrutiny of corporate auditors, security officers, and attorneys whose job it is to ensure that the users of this data comply with the corporation's consumer-data governance policies.

One effective way to satisfy these requirements would be to implement a two-tiered storage mechanism in HDFS that would support a high-performance primary storage tier that is backed up by a high-capacity secondary storage tier. Doing so would allow users needing governance controls to implement them at the

level of the secondary storage tier while leaving the primary storage tier free to support MapReduce processes. Two ways to do this would be:

- **Create the primary and secondary storage tiers within the Hadoop cluster that is managed by HDFS.** One can envision the use of solid-state disk (SSD) distributed across nodes as the primary storage tier and high-capacity disk also distributed across nodes as the secondary tier. Data is either copied or migrated from the primary to the secondary tier, where data is essentially archived and where sensitive data can be secured, locked down, or handled in whatever way necessary to satisfy governance policy.

- **Create a secondary storage tier external to Hadoop that is not managed by HDFS**. While not in common practice, Hadoop users in some instances have used shared storage systems to create the storage tiers. This allows the data services resident in the array (file indexing, WORM, etc.) to be employed without having to wait for the Apache community to develop them or write a custom implementation of HDFS.

Key takeaways

- It is clear that Hadoop is progressing from the internet or webscale data center to the enterprise data center, where its developers and supports will encounter a new set of rules and a different management regime.

- It is also clear that Hadoop is maturing away from a 100 percent batch-processing platform to one that also supports OLTP. However, in order for Hadoop to progress from pilot project status to the production application environment, it cannot add to the enterprise's risk profile.

- As these progressions unfold, Hadoop will need to demonstrate an ability to respond to the needs of not only IT administrators who will be operationally responsible for production-level and possibly mission critical implementations of Hadoop but also the corporate-level audit, security, and legal executives that are stakeholders in this process as well.

- While we have noted that some security, data protection, and compliance measures either already exist or can be added to Hadoop, native implementations that address the requirements will be preferable to the enterprise user. The result of doing so, we believe, will be twofold.

    - First, because the implementation is built into Hadoop, it will be that much easier for IT administrators to deploy and support. They will not be asked to do the complex work involved with integrating Kerberos, for example, with multinode Hadoop clusters to affect security. Therefore, there is a better chance that they will actually use the built-in functions.

    - Second, the inability to easily secure Hadoop environments, to include it under a disaster recovery plan, and to make Hadoop responsive to data governance policies have often been cited as issues that block the progress of Hadoop in the enterprise, from proof-of-concept projects to an application platform IT administrators can support as they would other production applications. Implementing solutions to these issues removes the barriers.

About John Webster

John Webster is a senior partner at Evaluator Group and contributes to the firm's ongoing research into data storage technologies including hardware, software, and services management. His specialties include big data analytics, storage and data management, virtualization, cloud computing, and analysis of storage infrastructure acquisition alternatives.

Mr. Webster has over 30 years experience in IT and is the author of numerous articles and white papers. He is also the co-author of a book titled *Inescapable Data -- Harnessing the Power of Convergence*, published in April 2005. He contributes regularly to Forbes.com and is often quoted in business and trade publications. Mr. Webster has also been a featured speaker at Storage Networking World and Storage Decisions events.

About Gigaom Research

Gigaom Research gives you insider access to expert industry insights on emerging markets. Focused on delivering highly relevant and timely research to the people who need it most, our analysis, reports, and original research come from the most respected voices in the industry. Whether you're beginning to learn about a new market or are an industry insider, Gigaom Research addresses the need for relevant, illuminating insights into the industry's most dynamic markets.

Visit us at: research.gigaom.com.

About WANdisco

WANdisco is a provider of enterprise-ready, nonstop software solutions that enable globally distributed organizations to meet today's data challenges of secure storage, scalability, and availability. WANdisco's products are differentiated by the company's patented nonstop data-replication technology, serving crucial high availability (HA) requirements for Hadoop Big Data and Application Lifecycle Management (ALM), including Apache Subversion and Git. Fortune Global 1000 companies rely on WANdisco for performance, reliability, security, and availability. For additional information, please visit wandisco.com.