FEATURE

Auditing Big Data in the Enterprise

Over the past several years, big data has emerged as one of the top strategic technology priorities for organizations. The concept of leveraging large data sets and analytics to drive business value is not new. For example, data warehousing has been around for several decades, along with associated technologies and techniques such as relational database management systems (RDBMS), business intelligence (BI) reporting systems, online analytical processing (OLAP) and data mining. Experienced information systems (IS) auditors are likely familiar with these types of systems and audit techniques associated with them. So, what makes big data unique, and what are some of the key risk factors that IS auditors should consider?

Big data is commonly described using the three Vs model. Many IS audit and risk professionals are already familiar with this model, which represents the concepts of volume, velocity and variety. The following describes each of the three Vs in more depth:

- **Volume**—Refers to the "bigness" of big data. Typically, big data refers to data sets that are vastly larger than traditional data repositories, perhaps beyond the capacity of traditional database software to store and process. However, "big data" is a subjective term; there is no defined threshold (e.g., how many terabytes) to qualify data as "big." At the high end of the spectrum, Walmart is building a data analytics hub, called the "Data Café," which will contain 40 petabytes of transactional data.[1]
- **Velocity**—Refers to the speed at which data are generated and/or changed. Big data is continuously generated and changed, often in real time, as opposed to a traditional approach to loading data via extract, transform and load (ETL) batch jobs. For example, according to *Forbes* magazine, Facebook's 1.2 billion users update their statuses an average of 293,000 times per minute.[2]
- **Variety**—Refers to the multiple sources and types of data that may be employed in a big data solution. Data may come from internal systems, customers/consumers and/or third parties. Additionally, data may be a mix of structured and unstructured data. Structured data usually conform to a defined data model (e.g., columns and rows), whereas unstructured data might be in the form of raw text, images, or perhaps even audio or video files. For instance, Black Knight Inc. has a big data solution for the mortgage industry called the LoanSphere Data Hub, which combines data from multiple sources, including internal mainframe and client-server-based transaction processing systems and industry data that come from flat files and public records. These data are often in the form of scanned document images.[3]

Joshua McDermott, CISA, CEH, CISSP, PMP
Is the director of IT audit for Black Knight Inc., a premier software provider to the financial industry that maintains the industry-leading US property database, covering 99.9 percent of all US property records. McDermott has been an IT professional for 20 years, including 10 years in IT audit, risk and information security positions. McDermott is also a US reserve Air Force cyberdefense operations officer.

It is important to note that some researchers have also suggested additional Vs to describe big data, such as veracity, value, variability and visualization. While those additional concepts are useful, the three Vs model is more commonly used and is sufficient for a basic understanding of big data.

Key Business Risk

As organizations increase adoption of big data solutions to drive business value and maintain a competitive advantage, it is important for the IS auditor to understand the associated risk and consider approaches to providing assurance that risk is being adequately managed. The following are descriptions of relevant risk:

> "ORGANIZATIONS MAY NOT BE EQUIPPED, PARTICULARLY WITH REGARD TO TALENT, TO CAPITALIZE ON THE BUSINESS OPPORTUNITIES ASSOCIATED WITH BIG DATA."

- **IT strategic alignment and resources**—It is important to understand the organization's overall strategy and how big data might be employed to support that strategy. Organizations may not be equipped, particularly with regard to talent, to capitalize on the business opportunities associated with big data. Indeed, there is an expected shortage of talent in data analytics, as IBM forecasts that by 2020, there will be 2.7 million new data and analytic job openings every year.[4] Such a talent shortage may impair an organization's ability to leverage big data to achieve its business strategy. Additionally, organizations may undertake big data initiatives viewing them primarily as technology projects without sufficient consideration of the business objectives and desired outcomes. The COBIT® 5 framework, particularly the Align, Plan and Organize (APO) processes, underscores the importance of effective human resource management and alignment of IT initiatives with business objectives prior to building, acquiring and implementing technology solutions.[5] Therefore, IS auditors should assess technology strategy and resource management processes to ensure that the organization's big data technology initiatives are aligned with its business strategy, and ensure that sufficient and qualified resources (e.g., qualified IT and development staff and data analysts) are available.
- **Development and implementation**—Big data solutions are no different from traditional information systems in terms of implementation and project management risk associated with developing and implementing complex technology solutions. Big data technology projects may experience challenges with scope, quality, cost and time to market. IS auditors should determine whether big data solutions are acquired and developed in a controlled manner using appropriate project management and system development processes. Significant big data initiatives may warrant formal project planning and periodic oversight by the organization's project management office. In addition, big data initiatives often use iterative Agile development methodologies such as Scrum. IS auditors are likely familiar with traditional waterfall system development processes that require formal documentation of requirements and specifications that are typically defined in great detail. However, Agile system development methodologies are iterative in nature and emphasize working software over comprehensive documentation.[6] IS auditors may be challenged to gain assurance that adequate testing was performed against defined acceptance criteria to determine that the big data solution is functioning as intended. Special emphasis should be given to assessing and providing assurance over the quality of the data. This is often referred to as the fourth V, the veracity of the data. To achieve this objective, IS auditors should assess the organization's big data quality assurance strategy or even determine whether an effective data governance program has been implemented.

- **Open source and cloud technologies**—Organizations may choose to implement big data solutions using open-source technology platforms, e.g., Apache Hadoop, or within third-party cloud computing environments, e.g., Amazon Web Services (AWS). These technologies present unique risk considerations that must be considered by IS auditors. For example, open-source software is highly configurable and may be more susceptible to security vulnerabilities. It was noted in February 2017 that there were more than 5,000 Hadoop clusters with weak security settings exposed to the Internet.[7] Hadoop is highly scalable software designed to run on inexpensive commodity server hardware. As the number of nodes increases, so does the risk of weak security settings. Another risk consideration when implementing technology solutions based on open-source software is the type of license associated with specific open-source technologies that are utilized in an organization's big data analytic solution. There are many different types of open-source licenses that range from permissive to highly restrictive.[8] Depending on the type of license being utilized in a given big data solution, there is a risk of intellectual property infringement or exposure of proprietary code for the organization.[9] IS auditors should assess controls to manage and mitigate these vulnerabilities and monitor compliance with open-source software licenses.

  Additionally, many vendors are offering big data solutions in the cloud. Organizations that choose to implement big data in a cloud environment should be aware of the associated risk. This includes third-party vendor performance, solvency, contractual compliance and security risk. IS auditors should confirm that big data cloud technology providers have adequate security controls and that management provides sufficient oversight of the third-party vendor relationship.

- **Data privacy and security**—A significant concern associated with big data is ensuring that adequate safeguards are in place to protect the data and adhere to privacy requirements, particularly for consumer information. Data can be compromised or stolen due to a number of factors, including inadequate security controls, malicious insiders, external threat actors and weak system security configurations. In June 2017, a data analytics firm that provides consulting services to political campaigns accidentally exposed 1.1 terabytes of sensitive consumer information, including 200 million US voters' names, addresses, dates of birth and voter registration information.[10]

> "IS AUDITORS SHOULD REVIEW BIG DATA INFRASTRUCTURE TO DETERMINE THAT IT IS CONFIGURED ACCORDING TO INDUSTRY OR VENDOR SECURITY CONFIGURATION GUIDELINES."

If sensitive data are being collected and stored in the big data solution, there may be regulatory or industry-specific requirements over how those data are protected, shared, retained and purged. Such regulations include the Gramm-Leach-Bliley Act (GLBA) and Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. Therefore, special consideration should be given for assessing safeguards for information such as consumer financial or health data and nonpublic information/personally identifiable information (NPI/PII).

Big data platforms should be configured in a secure manner, and logical access to the data should be restricted. Particularly sensitive data should be encrypted at rest and in transit. IS auditors should review big data infrastructure to determine that it is configured according to industry or vendor security configuration guidelines, such as the Hadoop security guides provided by BMC Software[11] and security research firm Securosis.[12] System audit logs should be enabled and periodically reviewed or monitored. Operating systems and software platforms that support big data analytic solutions should also be regularly scanned for vulnerabilities

and patched. Additional environmental security controls, such as firewalls, intrusion detection systems (IDS) and data loss prevention (DLP) systems, should be considered. Penetration tests should be conducted for Internet-enabled big data solutions, particularly those that contain sensitive information. IS auditors should review relevant regulations for data protection and privacy based on the organization's industry and assess security controls in detail for the big data platform and associated infrastructure and applications.

Conclusion

Big data analytic solutions are already widely deployed in many industries and will continue to experience tremendous growth in the near future. IS auditors will need to update their skills and knowledge to adapt to the paradigm shift from traditional data warehousing, which uses highly structured data on mature database management systems, to data lakes of vast amounts of unstructured data stored on commodity computing hardware using emerging open-source software. While there is tremendous business value potential with big data, there is also considerable risk that must be properly managed. IS auditors who are big-data-aware will be essential in providing assurance over this emerging technology.

Endnotes

1 Khanduja, J.; "Walmart Creates Largest Private Cloud, Data Café, And Analytics Hub," *TechTarget*, 30 January 2017, *http://itknowledgeexchange.techtarget.com/quality-assurance/walmart/*

2 Marr, B.; "4 Mind-Blowing Ways Facebook Uses Artificial Intelligence," *Forbes*, 29 December 2016, *https://www.forbes.com/sites/bernardmarr/2016/12/29/4-amazing-ways-facebook-uses-deep-learning-to-learn-everything-about-you/#4f6d5e1dccbf*

3 Black Knight, "LoanSphere Data Hub," *www.bkfs.com/Products/Mortgage/LoanSphere/Pages/Data-Hub.aspx*

4 IBM, *The Quant Crunch*, 2017, *https://www.ibm.com/analytics/us/en/technology/data-science/quant-crunch.html*

5 De Haes, S.; R. Debreceny; W. V. Grembergen; "Understanding the Core Concepts in COBIT 5," *ISACA® Journal*, vol. 5, 2013, *https://www.isaca.org/journal*

6 Agile, "Manifesto for Agile Software Development," 2001, *http://agilemanifesto.org/*

7 Millman, R.; "Thousands of Hadoop Clusters Still not Being Secured Against Attacks," *SC Media*, 10 February 2017, *https://www.scmagazineuk.com/thousands-of-hadoop-clusters-still-not-being-secured-against-attacks/article/637389/*

8 Malhotra, B.; "Classification of Open Source Licenses: A Developer's Perspective," Black Duck Software Blog, 30 December 2016, *http://blog.blackducksoftware.com/classification-open-source-licenses-developers-perspective*

9 Pittenger, M.; "Cloudera IPO: An Argument Against Open Source Business?" *Computer Business Review*, 11 April 2017, *www.cbronline.com/news/big-data/analytics/cloudera-ipo-argument-open-source-business/*

10 Bertrand, N.; "GOP Data Firm That Exposed Millions Of Americans' Personal Information Is Facing its First Class-Action Lawsuit," *Business Insider*, 22 June 2017, *www.businessinsider.com/deep-root-analytics-sued-after-data-breach-2017-6*

11 BMC Software, *Introduction to Hadoop Security*, 2017, *www.bmc.com/guides/hadoop-security.html*

12 Securosis, *Securing Hadoop: Security Recommendations for Hadoop Environments*, 21 March 2016, *https://securosis.com/assets/library/reports/Securing_Hadoop_Final_V2.pdf*