IPPF

International Professional Practices Framework

Supplemental Guidance

GTAG®

Global Technology Audit Guide

Understanding and Auditing Big Data

The Institute of Internal Auditors | Global

Table of Contents

Executive Summary

Big data is a popular term used to describe the exponential growth and availability of data created by people, applications, and smart machines. The term is also used to describe large, complex data sets that are beyond the capabilities of traditional data processing applications. The proliferation of structured and unstructured data, combined with technical advances in storage, processing power, and analytic tools, has enabled big data to become a competitive advantage for leading organizations that use it to gain insights into business opportunities and drive business strategies. However, the challenges and risks associated with big data must also be considered.

Increased demand, immature frameworks, and emerging risks and opportunities that are not widely understood or systematically managed by organizations have created a need for more guidance in this area. Internal auditors, in particular, must develop new skill sets and obtain knowledge of big data principles to effectively provide assurance that risks are addressed and benefits are realized.

Risks associated with big data include poor data quality, inadequate technology, insufficient security, and immature data governance practices. Internal auditors working with big data should engage with the organization's chief information officer (CIO) and other key leaders to better understand the risks in terms of data collection, storage, analysis, security, and privacy.

This guidance provides an overview of big data: its value, components, strategies, implementation considerations, data governance, consumption, and reporting, as well as some of the risks and challenges these may present. This guide also explains internal auditors' roles and responsibilities when performing assurance or advisory procedures related to big data efforts.

Introduction

The purpose of this guidance is to assist internal auditors in attaining the requisite knowledge in support of their advisory and assurance services related to big data, in accordance with Standard 1210 – Proficiency and Standard 2201 – Planning Considerations. This document provides an overview of big data to help the reader understand big data concepts and how to align internal audit activities in support of the organization's big data initiatives. In addition, this guidance includes a framework of key risks, challenges, and examples of controls that should be considered when planning an audit of big data (see Standard 2100 – Nature of Work).

This guidance does not address the internal audit activity's role in consuming big data or performing its own analysis to support audit and advisory activities (see "GTAG: Data Analysis Technologies" for additional information in this area). This guidance also does not include specific technical controls and work programs to provide audit coverage of big data technology, as these are dependent on the specific organization and big data systems in use.

Internal auditors should supplement this GTAG with other GTAGs and technical work programs to arrive at the most effective big data coverage model for the organization.

Business Significance

Analyzing data to create business value is not a new concept; however, it is becoming increasingly important to interpret data more quickly and base business decisions on the resulting insights. Organizations that effectively acquire and leverage big data are able to capitalize more quickly on emerging business trends, shifts in customer demands, and operational efficiency opportunities. This ultimately enhances the opportunity to improve customer satisfaction and maximize the organization's success.

According to ISACA (Big Data: Impacts and Benefits White Paper, March 2013[1]), "Enterprises that master the emerging discipline of big data management can reap significant rewards and differentiate themselves from their competitors."

Examples of business benefits from big data include:

- Competitive advantage.
- Increased revenue.
- Innovation and faster product development.
- Market demands predictions.
- Well-informed business decisions.
- Operational efficiency.

Indeed, significant benefits can be realized from big data programs if they are properly executed and well controlled. These programs aid in the consolidation and consumption of large volumes of structured and unstructured data, providing the opportunity for unique analytics and timely insights. (Previously, this may not have been possible or may have taken days or weeks to perform.) By using insights from big data, the organization can make better decisions, target new customers in creative and differentiating ways, service existing customers with a targeted and improved delivery model unique to the individual, and offer new services and capabilities that truly distinguish the company from its competitors. Organizations that capitalize on big data opportunities can develop a lasting competitive advantage.

Additionally, big data efforts can enhance an organization's transparency, improve management analysis and reporting, and drive down costs in support of continuous improvement programs. As data is centralized and consolidated for strategic big data efforts, the cost of performing incremental analytics using this data is greatly reduced, such that the entire organization can benefit from these initiatives.

[1] http://www.isaca.org/Knowledge-Center/Research/ResearchDeliverables/Pages/Big-Data-Impacts-and-Benefits.aspx

However, to fully exploit big data business benefits, organizations need to invest in creating the appropriate environment, hiring and retaining skilled people, defining and implementing repeatable processes, and deploying suitable technologies.

Practical Applications

Every day, organizations are introducing creative and innovative ways to use big data. Practical applications of big data can be found in retail, public, private and nonprofit organizations, for example.

Retailers – Retailers are building sophisticated customer profiles based on purchase history, online browsing, product feedback, geographic location, and demographics to create a data-driven shopping experience. Sales associates can access information about a customer's past purchases to offer personalized service and make recommendations driven by the customer's profile and social media postings, combined with the latest data about fashion trends. If the customer likes a garment that is not available in the store in the customer's preferred size or color, the sales associate can use a smart device to instantly determine whether the product is available in another store, or arrange for it to be shipped to the customer's home.

Another example of retailers using big data to improve customer service is the analysis of existing and potential customers' home addresses and demographic data to identify the ideal location for a new store.

Governments – City governments can collect traffic congestion data using multiple sources (e.g., sensors around the city, buses, and taxicabs) and analyze the data to identify patterns. Traffic analysis can be used to determine public transportation or road expansion needs, or to create websites that offer real-time traffic information to subscribers, which can help them plot the shortest travel routes, avoid traffic jams, and estimate arrival times.

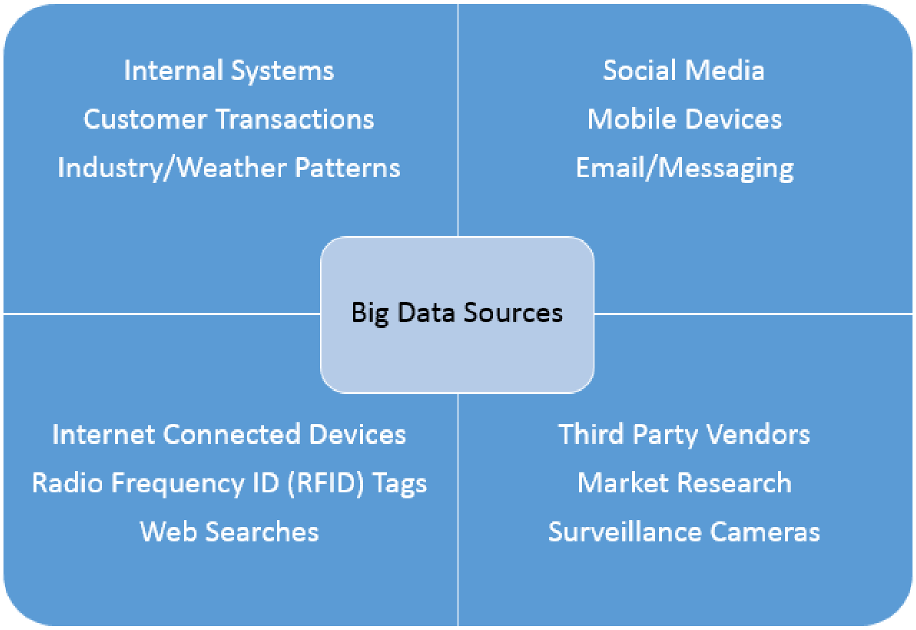
Financial Institutions – Financial institutions can forecast trends, model options, and predict outcomes to increase their customer base and improve customer loyalty. For example, financial institutions can provide more personalized services and tailored products using transactional history and demographic data to determine the right pricing and financial strategies for an individual.

Big data allows organizations to collect data from external sources. For example, a bank can collect data from social media to identify customers who may be dissatisfied with another financial institution and who may be seeking a new service provider. This information can be used to create targeted marketing messages to attract those unhappy customers by promising personalized products and services.

Big Data Basics

Big data is generated within organizations (e.g., transaction data, customer complaint data), industries (e.g., customer adoption rates per product type), societies (e.g., traffic cameras, economic data), nature (e.g., size, location, and frequency of earthquakes), and numerous other sources. In some instances, organizations must purchase data from external sources; in other cases, data sets are available for free use. Some examples of big data sources are outlined in **Figure 1.**

Figure 1—Examples of Big Data Sources

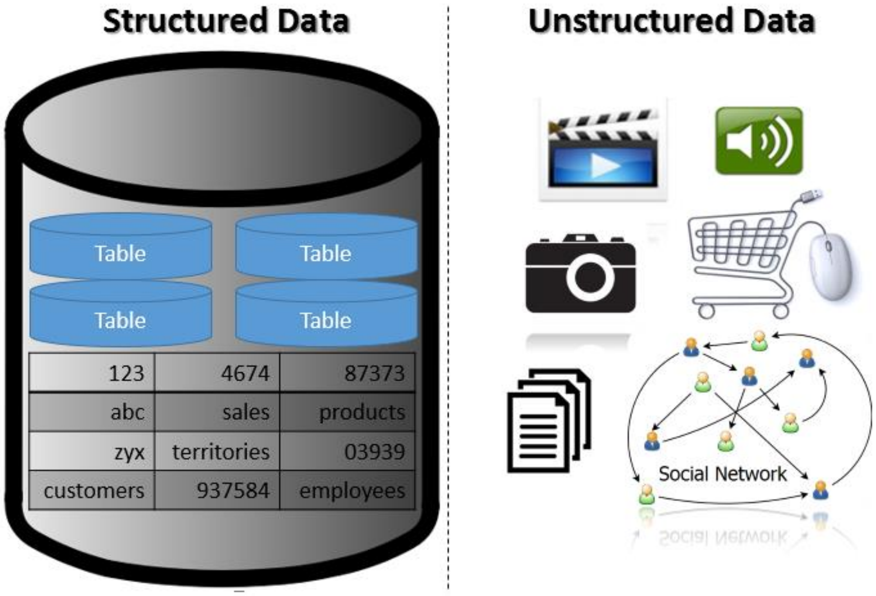


Structured and Unstructured Data

Historically, the majority of data stored within organizations has been structured and maintained within relational — or even legacy hierarchical or flat-file — databases. Structured data is organized and allows for repeatable queries, as much of the data is maintained in relational tables. It is often easier to control than unstructured data, due to defined ownership and vendor-supported database solutions.

However, the use of unstructured data is growing and becoming more common within organizations. This type of data is not confined to traditional data structures or constraints. It is typically more difficult to manage, due to its evolving and unpredictable nature, and it is usually sourced from large, disparate, and often external data sources. Consequently, new solutions have been developed to manage and analyze this type of data. See **Figure 2** for a diagram that shows the difference between structured and unstructured data.

Figure 2—Examples of Structured and Unstructured Data



123	4674	87373
abc	sales	products
zyx	territories	03939
customers	937584	employees

Data Storage

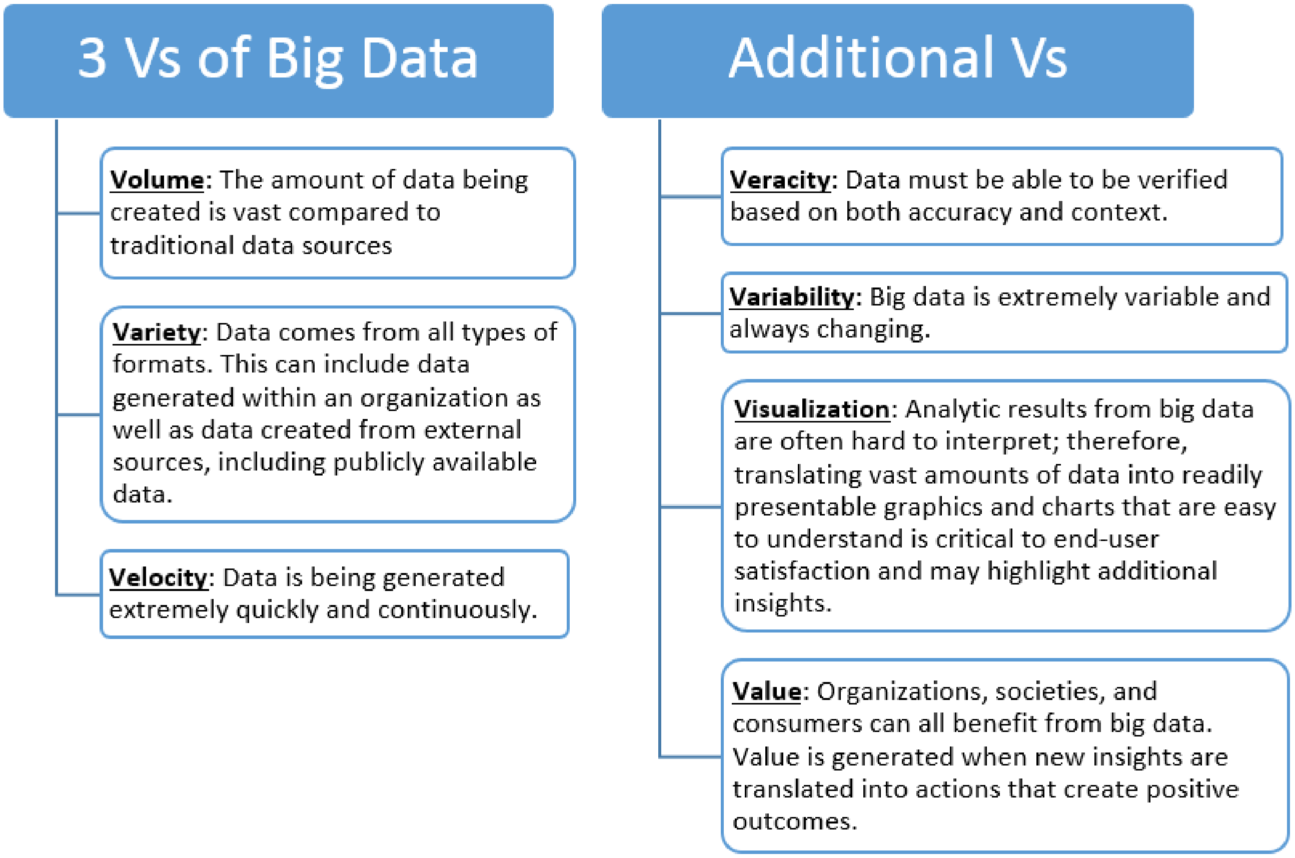
A large repository of enterprisewide data specifically designed for analytics and reporting is known as a *data warehouse*. Data warehouses are typically relational databases that store information from multiple sources. Data is uploaded from operational systems that contain transactional data to data warehouses, which store complete information about one or more subjects. ETL (extract, transform, and load) or ELT (extract, load, and transform) tools are configured to move data from the operational system to the data warehouse. The data is loaded in the format and structure of the data warehouse, which is often aggregated.

Data lakes are becoming an increasingly popular solution to support big data storage and data discovery. Data lakes are similar to data warehouses in that they store large amounts of data from various sources, but they also store additional data attributes from source systems at a level of granularity that would ordinarily be lost in data aggregation for data warehouses. This provides big data solutions with all available data elements at a sufficient level of granularity to perform a complete analysis. Yet, it offers organizations the flexibility to solve unanticipated problems, because it maintains all data in a readily available format.

Three Vs of Big Data

The most common dimensions or characteristics of big data management are *volume*, *velocity*, and *variety* (the 3Vs), but as systems become more efficient and the need to process data faster continues to increase, the original data management dimensions have expanded to include other characteristics unique to big data. In fact, Mark van Rijmenam proposed four additional dimensions in his August, 2013 blog post titled "Why The 3Vs Are Not Sufficient To Describe Big Data,"[2] the additional dimensions are *veracity*, *variability*, *visualization*, and *value.* **Figure 3** illustrates the "Expanded set of Vs of Big Data."

Figure 3—Expanded Vs of Big Data



3 Vs of Big Data

Volume: The amount of data being created is vast compared to traditional data sources

Variety: Data comes from all types of formats. This can include data generated within an organization as well as data created from external sources, including publicly available data.

Velocity: Data is being generated extremely quickly and continuously.

Additional Vs

Veracity: Data must be able to be verified based on both accuracy and context.

Variability: Big data is extremely variable and always changing.

Visualization: Analytic results from big data are often hard to interpret; therefore, translating vast amounts of data into readily presentable graphics and charts that are easy to understand is critical to end-user satisfaction and may highlight additional insights.

Value: Organizations, societies, and consumers can all benefit from big data. Value is generated when new insights are translated into actions that create positive outcomes.

[2] https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Big Data Program Elements

Internal audit is often requested by the board and/or members of senior management to provide insights and perspectives on big data programs as they are being implemented (see Standard 2010.C1). To effectively engage with management and assess big data programs, internal auditors should understand the various components that comprise a big data program, as well as related roles and responsibilities (see Standard 2100 – Nature of Work, Standard 2200 – Engagement Planning, and Standard 2201 – Planning Considerations, as well as their respective Implementation Guides).

As big data programs are implemented and modified over time, internal auditors should remain engaged and monitor these efforts, in accordance with Standard 2010 – Planning and Standard 2030 – Resource Management. In doing so, internal auditors must maintain requisite knowledge and skills, and dynamically alter their risk assessment and audit coverage model to account for any changes (see Standard 1210 – Proficiency). Big data is evolving rapidly and will continue to present risks and opportunities for organizations and internal auditors for the foreseeable future.

Articulated Business Case

For big data programs to be successful, a clear business case must be articulated in alignment with the organization's strategy. The big data program should have defined objectives, success criteria, and executive-level business sponsorship. The business case should also include a cost-benefit analysis of deploying such a significant program versus leveraging existing tools and technologies within the enterprise.

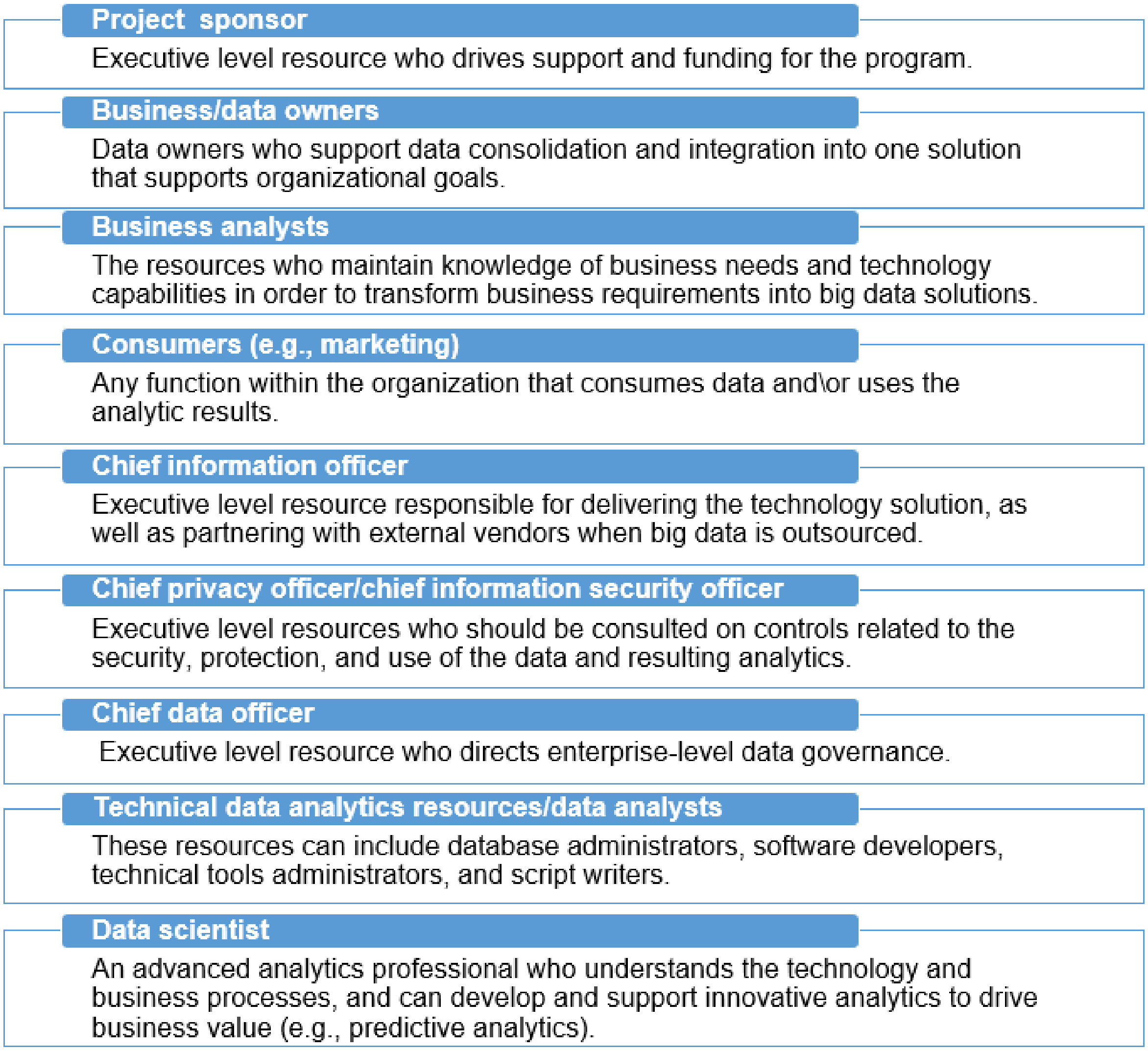
Strong organizational sponsorship is crucial; without this support, the sustained investment of necessary resources and adequate prioritization may not occur. The business case for a big data program should also include technology sponsorship, with appropriate vetting of the options and costs presented, as well as clear ownership for the sustainability of the program post implementation. Multiple organizations now have chief data officers (CDOs) who are senior managers focused on ensuring big data programs have the necessary support.

Defined Roles and Responsibilities

Clearly defining roles and responsibilities across key resources and functions can accelerate and simplify deployment and support for big data programs. For example, if an organization plans to conduct analytics on employee activities and behavior (e.g., employee fraud investigations), human resources and legal may need to be consulted. Marketing personnel may want to drive, or be heavily involved in, pilot analytics of customer engagement to gauge customer response. IT may be unable to support planned efforts due to competing priorities. Risk, security, and privacy functions may want to review and assess the controls in the big data environment.

Without buy-in from key business stakeholders and consumers on their responsibilities in the overall effort, data insights may go unused for long periods of time following the deployment of the big data program. Therefore, prior to making a significant investment in big data efforts, organizations should engage all relevant stakeholders to ensure support, determine value, inquire about requirements, address preferences, gauge bandwidth to advance plans in spite of other priorities, and drive action during and post implementation. A formal stakeholder analysis may be necessary to identify all relevant stakeholders before implementing the program. **Figure 4** shows examples of key stakeholders who may help drive and support big data efforts.

Figure 4— Examples of Big Data Key Stakeholders



Project sponsor
Executive level resource who drives support and funding for the program.

Business/data owners
Data owners who support data consolidation and integration into one solution that supports organizational goals.

Business analysts
The resources who maintain knowledge of business needs and technology capabilities in order to transform business requirements into big data solutions.

Consumers (e.g., marketing)
Any function within the organization that consumes data and\or uses the analytic results.

Chief information officer
Executive level resource responsible for delivering the technology solution, as well as partnering with external vendors when big data is outsourced.

Chief privacy officer/chief information security officer
Executive level resources who should be consulted on controls related to the security, protection, and use of the data and resulting analytics.

Chief data officer
Executive level resource who directs enterprise-level data governance.

Technical data analytics resources/data analysts
These resources can include database administrators, software developers, technical tools administrators, and script writers.

Data scientist
An advanced analytics professional who understands the technology and business processes, and can develop and support innovative analytics to drive business value (e.g., predictive analytics).

Adequate Resources

Organizations must have sufficient resources to support a big data implementation, which includes more than just the funds necessary to purchase equipment to store data and process large, complex, analytical jobs. The use of third-party resources and/or solutions must be considered when building the overall program approach, because depending on the size and planned scale of the big data program, outsourcing may introduce more timely, scalable, and cost-effective options.

The human resources department within organizations often assists with the recruitment and selection of appropriate personnel to fill the big data roles, define compensation and benefits for big data personnel, and assist big data training and development programs. Organizations face many challenges recruiting big data personnel, due to the technical complexity of the positions and the scarcity of talent in the job market.

Performance Metrics

The success criteria established during program design should be tracked through agreed-upon performance metrics. These metrics should present a balance of operational and organizational performance. They should also provide management with insight into the cost, level of adoption, availability, and usage of the big data solution across the enterprise.

Alignment of big data initiatives or pilot projects with meaningful business metrics (e.g., reduced customer acquisition and retention costs, growth in market share, and increased click-through rates) provide tangible demonstration of return on investment. However, qualitative benefits should also be tracked, as they may help demonstrate the program’s value.

Risks often arise when metrics are not clearly and completely designed in alignment with the business case. Individuals may overemphasize or overfund one aspect of the big data program at the expense of others due to inadequately defined performance measures or measures that may serve as an incentive for compensation. Therefore, management must be extremely diligent when establishing success criteria and defining supporting metrics of any key program, including big data initiatives, as metrics can drive appropriate or inappropriate behavior.

Tools and Technologies

It is critical for organizations to identify the most appropriate tools and technologies to fit their current and future needs. These tools should enable the organization to acquire, process, analyze, and use data from sources that produce increasing amounts of structured and unstructured data. For example, reporting, visual analytics, monitoring, and job automation tools may improve the user experience and reduce the time spent on maintaining the environments. Adequate storage is also a must, as the volume of data can grow quickly and significantly as the program expands. Technology solutions must be in alignment with the defined business requirements, which depend on several variables, including the organization’s planned and future uses, the size of the organization, and many other factors.

Storage Solutions

Historically, the most significant challenge with the ever-growing amount of data has been the increasing need for additional storage and sufficient backup capabilities. Due to the additional sources of data and increasing appetite for data, the need for new and more powerful tools and technologies has become vital. Data from new sources and data generated from systems rather than human beings, have challenged traditional technologies and tools by demanding new capabilities and solutions for storing data and making it readily available.

Onsite Vs. Cloud Environments

The platforms and tools available to organizations will change over time; however, key technology considerations often remain consistent across solutions. For example, organizations must choose between onsite and cloud-based big data environments and consider how they will staff analytic development. Onsite solutions require a facility capable of hosting a large number of servers and an IT team to support the infrastructure. The facility should be large enough to support scalability as big data usage increases. Cloud-based big data implementation serves as another option that may be more cost effective and accelerate time to delivery. However, cloud solutions also expose the organization to additional risks, which must be fully assessed prior to deciding whether to use the cloud. Cloud-based “big data as a service” (BDaaS) solutions provide total scalability. BDaaS solutions can include:

- IaaS — Infrastructure as a service (e.g., hardware, storage devices, and network components for big data).
- PaaS — Platform as a service (e.g., operating systems, development platforms, and middleware).
- SaaS[3] — Software as a service (e.g., applications to process big data or to conduct analysis and reporting).

Organizations can chose to use one cloud service (IaaS, PaaS, or SaaS) to supplement in-house systems, or integrate all three cloud services to develop the entire big data solution.

Data Discovery Tools

The vast variety and volume of data available today are the result of exponentially cheaper storage and ubiquitous data collection that have come with the rapid growth of social media platforms, sensors, and multimedia. Unfortunately, traditional data warehouses and business intelligence solutions were not built to meet big data requirements and have been overrun by the wave of data storage and processing requirements. Consequently, organizations that are still using data warehouses may be operating and making decisions based on incomplete data.

[3] For more information about cloud computing, see the National Institute of Standards and Technology (NIST) special publication SP800-145 “The NIST Definition of Cloud Computing”, http://dx.doi.org/10.6028/NIST.SP.800-145

There are three primary elements to big data discovery: (1) understanding what data is available; (2) acquiring it; and (3) learning from it to develop meaningful insights that lead to actionable items. Organizations are at varying levels of maturity in terms of their ability to manage and understand internal structured data. Many organizations struggle significantly with unstructured data or data outside of the organization. Third-party and unstructured data is where big data technology and organizations with effective big data programs thrive. Identifying and acquiring this data often requires creative thinking, development or configuration of application programming interfaces (APIs), and potential fees for subscription to data providers. Acquiring all available data is one approach, but for organizations with limited resources, it may be best to start with a specific-use pilot and grow the program incrementally.

Distributed data processing[4] and enhanced machine learning[5] increase the value of big data. These computing advances can help organizations identify patterns unrecognizable to humans and lower capacity applications. In addition, new data visualization tools are being included as part of big data solutions to provide flexibility, interaction, and ad-hoc analysis capabilities.

Monitoring Tools

It is important to define key performance indicators (KPIs) for big data systems and analytics during implementation to enable ongoing production monitoring. Monitoring tools should be used to report on the health and operational status of the big data environment and provide the information necessary to proactively identify and mitigate the operational risks associated with big data. The monitoring tools should be able to report on anomalies across various aspects of the big data platform, as well as job processing. As stated earlier, KPIs should be created to report on the effectiveness and performance of big data systems.

Software Acquisition

Software development or purchase-and-customization activities for big data are very different from traditional systems. Relevant open-source technology can be downloaded free of charge from many places. Additional product distributions are also available free of charge or for purchase from value-added vendors. Although they may be appealing, free downloadable distributions from value-added vendors come with no product or technical support.

There are differences in the features and functionality of various product offerings and numerous vendor customizations of different platforms, which makes it difficult to understand and differentiate various offerings. Structured query software components, for example, are not a part of all distributions, and some vendors have better security features than others. Understanding these differences and aligning them to a big data program’s requirements is

[4] Distributed data processing refers to multiple computers in the same or different locations sharing processing capacity to speed computer processing.

[5] Machine learning refers to computer programs capable of learning algorithms without the need of human interaction for programming.

imperative for selecting the appropriate software distribution. Whether an onsite or cloud-based solution is implemented, IT departments should carefully evaluate big data requirements and avoid purchasing unnecessary software, processing power, and storage.

Although big data hardware is commoditized for distributed processing, the underlying software complexity increases the importance of the solution design and development phase. Big data platforms almost always have additional software modules installed alongside them. These additional software modules provide extended features on how to manage, interact with, and analyze data, as well as how to present the results. Increasingly, big data programs have specialized data visualization software to present results in dashboards.

Big data vendors have been helping organizations navigate the technical environment, extent of customization, abundance of software tools, numerous data interfaces, and modeling complex data. Even so, organizations are challenged to identify internal resources (e.g., big data program managers) with sufficient knowledge to work with and manage big data vendors through the development lifecycle. Often, data scientists are hired to help develop the analytical models.

Ongoing Program Support

Big data solutions are not meant to be built and remain static, nor are they meant to have a significant production overhead. Still, as with many open-source transformational technologies, the rapid pace of change in the big data landscape creates challenges that often outpace big data architects' ability to keep up with dozens of new tools, plug-ins, and rapid product releases.

As a result, ongoing support from internal resources or vendors is necessary to ensure continued success of the program. This ongoing support includes traditional IT operations, such as capacity planning (i.e., scaling flexibility), production monitoring, and disaster recovery planning. Further, internal and external data sources are consistently being added, removed, or changed. Supporting data storage infrastructures and related data integrations need to be assessed and aligned with these activities. Standard application change and patch management practices also apply (see "GTAG: Change and Patch Management Controls: Critical for Organizational Success, 2nd Edition"). Finally, the analytic models themselves must be monitored and maintained.

Data Governance

The adoption of big data in an organization requires strengthened data governance to ensure that information remains accurate, consistent, and accessible. There are several key areas where data governance for big data is critical; these include metadata (i.e., data about data) management, security, privacy, data integration, data quality, and master data management. Key traditional data governance activities to address these areas include identifying data owners, consumers, critical data elements (CDEs), special handling requirements, lineage, master data, and authoritative data sources. Organizations must implement appropriate controls to ensure that all necessary data quality dimensions (e.g., completeness, validity, uniqueness) are properly maintained and that such controls protect CDEs in the same manner. Some examples of control processes related to data quality and CDE protection include data defect identification and data loss/leakage prevention (DLP).

The key difference with data governance in a big data context, compared to traditional data governance programs, relates to the agility that organizations must have throughout the data lifecycle to meet analysis demands. The risks associated with organizations' need for agility are compounded by characteristics or business requirements for each data set, including those for privacy and security, and the unique characteristics that may be required for particular business operations.

Data owners should take responsibility for the quality and security of their data, with a heightened focus on the riskiest data elements. The riskiest elements should be determined based on the results of a risk assessment process, which may be led by the data owners or other functions within the organization (e.g., information security). Data owners must ensure systems-of-record are appropriately defined and processes to update CDEs are clear. This should include the identification of authoritative data sources (i.e., those that define which system's data takes priority when data elements vary or conflict among two or more systems). Some organizations have assigned "data stewards" to assist in this and other data governance efforts.

Ultimately, the objective is for organizations to be able to move information quickly, while maintaining high quality and security. This requires agile data governance, which ensures appropriate controls are in place to support the sustainability and value proposition of these programs.

Consumer Adoption

Big data analytics can be implemented for an organization's internal use (e.g., to drive business decisions related to operations, marketing, human resources, or IT) or to meet customer's needs (e.g., analyzing customers' past buying behavior can enable organizations to recommend new products or services when customers visit the organization's website). Regardless, the goal of any big data analytic solution should be to provide meaningful information for internal data consumers (i.e., employees) and external data consumers (i.e., customers or suppliers), and to improve decision-making processes. Big data solutions that create value will drive sustained adoption within the organization.

As with other IT projects, this process must begin with gathering comprehensive requirements to understand the questions the consumer is trying to answer or issues the consumer is trying to predict. Involving internal consumers in the design and testing process helps them develop a sense of ownership in the solution. Any post-implementation feedback should be addressed promptly to sustain and increase adoption. Organizations should also plan marketing and training campaigns to share success stories and educate internal consumers on the potential of big data analytics. Stakeholder surveys are an effective tool for obtaining feedback and lessons learned to improve development processes for subsequent implementations.

Analytics and Reporting

Reports should be designed with the appropriate flexibility of input parameters (e.g., start and end dates, customer segments, and products) to allow consumers to narrow or broaden the focus of their analysis. This flexibility enables consumers to ask questions that might not have been anticipated during the initial development phase and supports adoption by empowering consumers with self-service capabilities, which helps minimize the traditionally slow and costly report development lifecycle. The available granularity of report data to support consumers' standard or drill-down reporting should be balanced with consumer requirements, processing capabilities, and data privacy concerns.

Self-service tools are important for activities that involve customers, vendors, or employees who need to make fast decisions. For example, a customer service representative can use big data and a self-service reporting application to view a customer's product and service history across multiple organizational lines on one screen. This would reduce the number of phone calls the customer would need to make to answer product inquiries. Privacy and security concerns can be addressed by restricting access to sensitive data fields to only those consumers who have a valid business need to see those data fields.

Many people are familiar with the concept of predictive analytics, which attempts to explain what will occur next based on historical data. For example, hospitals utilize predictive analytics to determine which patients may be readmitted for additional treatment. Data scientists can apply survivor analysis algorithms to help human resources departments predict employee dissatisfaction, and that information can be used to support workforce management and planning activities.

Analytic reports may also be alert-based, to help consumers identify which actions are needed to address a particular situation. For example, sentiment analysis techniques can be applied to determine a customer's satisfaction with a product or service, based on information shared by the customers via social media. High satisfaction levels can drive new distribution strategies, while poor satisfaction levels may require immediate remediation actions to protect customer loyalty and the organization's reputation.

Other advanced analytical techniques, such as link analysis, can be used by internal auditors and fraud investigators to better understand relationships among employees, vendors, and customers, to uncover money laundering or other fraud schemes. The analytic scenarios are unlimited, providing vast opportunities to add value to the end consumer through various reporting channels and options.

Consistent Processes

Defined and well-controlled processes are necessary for the continued success of a big data program. Without defined and consistent processes, environments can quickly become unstable, and the confidence in the underlying data can be lost. When analytic reports designed to provide strategic insights use inaccurate or incomplete data, their value is immediately undermined. If risks such as these are not addressed, big data programs can quickly lose funding and be eliminated. Therefore, as people and technology are being deployed, careful attention should also be paid to the underlying processes and related controls that support the big data program.

Key Risks

Risks related to big data can arise from many factors, both internal and external to the organization. The following categories represent the primary risk areas:

- Program governance.
- Technology availability and performance.
- Security and privacy.
- Data quality, management, and reporting.

Area	Key Risk	Control Activities
Program governance	Lack of appropriate management support, funding, and/or governance over the big data program can expose the organization to undue risk or failure to meet strategic goals.	• Funding should be adequate to support business needs. • Program objectives should support enterprisewide strategy initiatives. • Management should receive metrics that demonstrate achievement of goals. • The organization should establish a governing entity to manage the big data strategy. • There should be agreed-upon SLAs between the business and IT to describe and measure performance expectations. • Business and technical requirements should be documented, analyzed, and approved. • Executive management should develop a big data strategy that provides solutions across the organization. • Prior to approving the business case, management should conduct a proof of concept to validate that the systems designs align with strategic goals. • Roles and responsibilities should be clear and well defined. • The organization should provide the necessary resources to deploy and maintain the big data strategy. • Third-party vendor management best practices should be used to manage big data suppliers.

Area	Key Risk	Control Activities
		• Data governance should be part of the overall enterprise governance to ensure that big data objectives align with the organization's strategic goals (see Standard 2110 – Governance).
Technology availability and performance	Ineffective technology solutions and/or configurations may result in a negative customer experience, reduced system availability, and/or degraded performance.	• IT operations should be structured in a manner that supports big data service level expectations. • Data lifecycle policies and procedures should be documented and followed. • Big data systems should be part of the maintenance strategy. • Big data systems should be part of the change management strategy. • Big data systems should be included in the patch management strategy. • Big data systems should be procured, built, and/or configured in alignment with the complexity and demands documented in the business case. • Systems and support tools should be configured to provide automatic notifications to support personnel. • Reporting tools should be configured to be flexible, intuitive, and easy to use; and training aids should be provided. • Big data systems should be configured to allow flexibility and scalability without sacrificing performance. • Periodic performance testing should be conducted and weaknesses should be remediated. • The big data systems lifecycle should be managed properly. • IT general controls should be assessed periodically.
Security and privacy	Ineffective information security standards and configurations	• Information security management should be part of the big data strategy. • Data security management should be part of the big data strategy.

Area	Key Risk	Control Activities
	may result in unauthorized access to — and theft of — data, inappropriate modifications of data, and regulatory compliance violations.	• Third-party access should be managed properly. • Data privacy should be part of the big data strategy.
Data quality, management, and reporting	Data quality issues and/or inaccurate reporting may lead to inaccurate management reporting and flawed decision making.	• Policies and procedures should be established to ensure data quality. • Policies and procedures should be established to ensure that data obtained from third parties complies with data quality standards. • Policies and procedures should be established to ensure reporting accuracy. • Access to reports should be granted based on business needs. • Reporting tools and procedures should allow for flexibility and ad-hoc reporting. • Users should be trained periodically to maximize report utility. • The selection of vendors who provide reporting products and services should align with business needs.

Internal Audit's Role in Big Data

Internal audit should consider the role of big data within organizations as part of risk assessment and audit planning (see Standard 2010 – Planning and 2010.A1). If the risks are significant, internal audit can determine an appropriate plan to provide coverage of big data risks and controls. In doing so, internal audit has the opportunity to educate the board on the organization's big data initiatives, the resulting risks and challenges, and the significant opportunities and benefits. Typically, internal audit provides coverage of big data through multiple audits versus a single, stand-alone big data audit.

As big data programs are implemented, similar to other large-scale programs, internal audit should consider involvement through formal and/or informal assessments. These may include advisory projects, pre- or post-implementation reviews, and adequate participation in governance and steering committees. As noted in Standard 2130 – Control, "The internal audit activity must assist the organization in maintaining effective controls by evaluating their effectiveness and efficiency and by promoting continuous improvement." As such, internal audit should assess process and technology controls. Internal audit should also focus significantly on how the data is being consumed and the actions the organization is taking based on results obtained from big data analysis. Internal auditors should play a critical role in an organization's big data initiatives, and this role can adjust over time as solutions are implemented, mature, and evolve (see Standard 2201 – Planning Considerations).

Internal auditors may also leverage big data solutions in support of their data analytic efforts for audit projects. Because the organization has already acquired, consolidated, and integrated the data, internal audit may gain significant efficiencies by consuming data from a data warehouse or data lake, rather than targeting many source systems.

Big data audit programs will vary by organization and usage. Program governance is a key component of big data audit programs. Internal auditors must verify that the objectives of a big data program align with the enterprisewide business strategy. Additionally, internal auditors should perform tests to ensure the big data program provides value and is fully supported by appropriate leadership in the organization. While the specific technology and level of vendor sourcing for big data solutions will vary by organization, internal auditors should ensure the confidentiality, integrity, availability, and performance of big data systems aligns with management's business requirements and needs.

As big data systems require vast amounts of data for analysis, audit programs should include test steps to ensure the quality, security, and privacy of the data used for analysis, as well as analytic outputs. Because big data consumes data from many disparate sources to provide a more complete view of a subject, audit programs should provide reasonable assurance that the data is safe from unauthorized modification and can only be viewed by authorized individuals. Audit programs should also test the controls over the quality of data input into the system as well as the quality of output and reporting from the system; these efforts may include providing substantive test coverage over the quality of key system data and reporting.

Appendix D provides a sample work program to help internal auditors jumpstart a review of big data in their organization. Note that the list of review activities provided in the sample work program is not comprehensive, as work programs must be customized to meet the specific needs of each organization.

Big data systems and reference materials are changing frequently. Therefore, internal audit should remain focused on using the most appropriate guidance when developing and executing audit plans and coverage models of big data within their organizations.

Appendix A. Key IIA Standards

The following selections from The IIA's *International Standards for the Professional Practice of Internal Auditing* (*Standards*) are relevant to big data. It is important to note that information in the *Standards* should be complimented with each standard's respective Implementation Guide.

Standard 1210 – Proficiency

Internal auditors must possess the knowledge, skills, and other competencies needed to perform their individual responsibilities. The internal audit activity collectively must possess or obtain the knowledge, skills, and other competencies needed to perform its responsibilities.

Interpretation:

Proficiency is a collective term that refers to the knowledge, skills, and other competencies required of internal auditors to effectively carry out their professional responsibilities. It encompasses consideration of current activities, trends, and emerging issues, to enable relevant advice and recommendations. Internal auditors are encouraged to demonstrate their proficiency by obtaining appropriate professional certifications and qualifications, such as the Certified Internal Auditor designation and other designations offered by The Institute of Internal Auditors and other appropriate professional organizations.

> **1210.A3 –** Internal auditors must have sufficient knowledge of key information technology risks and controls and available technology-based audit techniques to perform their assigned work. However, not all internal auditors are expected to have the expertise of an internal auditor whose primary responsibility is information technology auditing.

Standard 2010 – Planning

The chief audit executive must establish a risk-based plan to determine the priorities of the internal audit activity, consistent with the organization's goals.

Interpretation:

To develop the risk-based plan, the chief audit executive consults with senior management and the board and obtains an understanding of the organization's strategies, key business objectives, associated risks, and risk management processes. The chief audit executive must review and adjust the plan, as necessary, in response to changes in the organization's business, risks, operations, programs, systems, and controls.

> **2010.A1 –** The internal audit activity's plan of engagements must be based on a documented risk assessment, undertaken at least annually. The input of senior management and the board must be considered in this process.

2010.A2 – The chief audit executive must identify and consider the expectations of senior management, the board, and other stakeholders for internal audit opinions and other conclusions.

2010.C1 – The chief audit executive should consider accepting proposed consulting engagements based on the engagement's potential to improve management of risks, add value, and improve the organization's operations. Accepted engagements must be included in the plan.

Standard 2030 – Resource Management

The chief audit executive must ensure that internal audit resources are appropriate, sufficient, and effectively deployed to achieve the approved plan.

Interpretation:

Appropriate refers to the mix of knowledge, skills, and other competencies needed to perform the plan. Sufficient refers to the quantity of resources needed to accomplish the plan. Resources are effectively deployed when they are used in a way that optimizes the achievement of the approved plan.

Standard 2100 – Nature of Work

The internal audit activity must evaluate and contribute to the improvement of the organization's governance, risk management, and control processes using a systematic, disciplined, and risk-based approach. Internal audit credibility and value are enhanced when auditors are proactive and their evaluations offer new insights and consider future impact.

Standard 2110 – Governance

The internal audit activity must assess and make appropriate recommendations to improve the organization's governance processes for:

- Making strategic and operational decisions.
- Overseeing risk management and control.
- Promoting appropriate ethics and values within the organization.
- Ensuring effective organizational performance management and accountability.
- Communicating risk and control information to appropriate areas of the organization.
- Coordinating the activities of, and communicating information among, the board, external and internal auditors, other assurance providers, and management.

2110.A2 – The internal audit activity must assess whether the information technology governance of the organization supports the organization's strategies and objectives.

Standard 2130 – Control

The internal audit activity must assist the organization in maintaining effective controls by evaluating their effectiveness and efficiency and by promoting continuous improvement.

Standard 2200 – Engagement Planning

Internal auditors must develop and document a plan for each engagement, including the engagement's objectives, scope, timing, and resource allocations. The plan must consider the organization's strategies, objectives, and risks relevant to the engagement.

Standard 2201 – Planning Considerations

In planning the engagement, internal auditors must consider:

- The strategies and objectives of the activity being reviewed and the means by which the activity controls its performance.
- The significant risks to the activity's objectives, resources, and operations and the means by which the potential impact of risk is kept to an acceptable level.
- The adequacy and effectiveness of the activity's governance, risk management, and control processes compared to a relevant framework or model.
- The opportunities for making significant improvements to the activity's governance, risk management, and control processes.

2201.C1 – Internal auditors must establish an understanding with consulting engagement clients about objectives, scope, respective responsibilities, and other client expectations. For significant engagements, this understanding must be documented.

Standard 2210 – Engagement Objectives

Objectives must be established for each engagement.

2210.A1 – Internal auditors must conduct a preliminary assessment of the risks relevant to the activity under review. Engagement objectives must reflect the results of this assessment.

2210.A2 – Internal auditors must consider the probability of significant errors, fraud, noncompliance, and other exposures when developing the engagement objectives.

2210.C1 – Consulting engagement objectives must address governance, risk management, and control processes to the extent agreed upon with the client.

> **2210.C2 –** Consulting engagement objectives must be consistent with the organization's values, strategies, and objectives.

Standard 2220 – Engagement Scope

The established scope must be sufficient to achieve the objectives of the engagement.

> **2220.A1 –** The scope of the engagement must include consideration of relevant systems, records, personnel, and physical properties, including those under the control of third parties.

Standard 2230 – Engagement Resource Allocation

Internal auditors must determine appropriate and sufficient resources to achieve engagement objectives based on an evaluation of the nature and complexity of each engagement, time constraints, and available resources.

Interpretation:

Appropriate refers to the mix of knowledge, skills, and other competencies needed to perform the engagement. Sufficient refers to the quantity of resources needed to accomplish the engagement with due professional care.

Standard 2240 – Engagement Work Program

Internal auditors must develop and document work programs that achieve the engagement objectives.

Standard 2310 – Identifying Information

Internal auditors must identify sufficient, reliable, relevant, and useful information to achieve the engagement objectives.

Interpretation:

Sufficient information is factual, adequate, and convincing so that a prudent, informed person would reach the same conclusions as the auditor. Reliable information is the best attainable information through the use of appropriate engagement techniques. Relevant information supports engagement observations and recommendations and is consistent with the objectives for the engagement. Useful information helps the organization meet its goals.

Appendix B. Related Guidance

GTAG: Continuous Auditing: Coordinating Continuous Auditing and Monitoring to Provide Continuous Assurance, 2nd Edition

GTAG: Assessing Cybersecurity Risk: Roles of the Three Lines of Defense

GTAG: Auditing Application Controls

GTAG: Change and Patch Management Controls: Critical for Organizational Success, 2nd Edition

GTAG: Data Analysis Technologies

GTAG: Developing the IT Audit Plan

GTAG: Identity and Access Management

GTAG: Information Technology Outsourcing, 2nd Edition

GTAG: Information Technology Risk and Controls, 2nd Edition

Practice Guide: Auditing Privacy Risks, 2nd Edition

Appendix C. Glossary

Terms identified with an asterisk (*) are taken from the "Glossary" of The IIA's *International Professional Practices Framework (IPPF),* 2017 edition.

Add value* – The internal audit activity adds value to the organization (and its stakeholders) when it provides objective and relevant assurance, and contributes to the effectiveness and efficiency of governance, risk management, and control processes.

Application programming interfaces (APIs) – A set of routines, protocols, and tools used in software development.

Assurance services* – An objective examination of evidence for the purpose of providing an independent assessment on governance, risk management, and control processes for the organization. Examples may include financial, performance, compliance, system security, and due diligence engagements.

Authoritative data sources (ADS) – Sources that provide "official" data to other systems.

Board* – The highest level governing body (e.g., a board of directors, a supervisory board, or a board of governors or trustees) charged with the responsibility to direct and/or oversee the organization's activities and hold senior management accountable. Although governance arrangements vary among jurisdictions and sectors, typically the board includes members who are not part of management. If a board does not exist, the word "board" in the *Standards* refers to a group or person charged with governance of the organization. Furthermore, "board" in the *Standards* may refer to a committee or another body to which the governing body has delegated certain functions (e.g., an audit committee).

Chief audit executive* – Chief audit executive describes the role of a person in a senior position responsible for effectively managing the internal audit activity in accordance with the internal audit charter and the mandatory elements of the International Professional Practices Framework. The chief audit executive or others reporting to the chief audit executive will have appropriate professional certifications and qualifications. The specific job title and/or responsibilities of the chief audit executive may vary across organizations.

Chief data officer (CDO) – Executive level position responsible for governing and managing data across the organization.

Consulting services* – Advisory and related client service activities, the nature and scope of which are agreed with the client, are intended to add value and improve an organization's governance, risk management, and control processes without the internal auditor assuming management responsibility. Examples include counsel, advice, facilitation, and training.

Control* – Any action taken by management, the board, and other parties to manage risk and increase the likelihood that established objectives and goals will be achieved. Management plans, organizes, and directs the performance of sufficient actions to provide reasonable assurance that objectives and goals will be achieved.

Control environment* – The attitude and actions of the board and management regarding the importance of control within the organization. The control environment provides the discipline and structure for the achievement of the primary objectives of the system of internal control. The control environment includes the following elements:

- Integrity and ethical values.
- Management's philosophy and operating style.
- Organizational structure.
- Assignment of authority and responsibility.
- Human resource policies and practices.
- Competence of personnel.

Critical data elements (CDEs) – Data elements that are critical for users or systems to perform calculations or conduct business.

Data elements – The smallest unit of data that conveys meaningful information (e.g., name, account number).

Data governance – The exercise of authority, control, and shared decision making (planning, monitoring, and enforcement) over the management of data assets. Data governance is high-level planning and control over data management.[6] Data governance is necessary to best manage data while maintaining quality and accuracy.

Data lineage – The visualization of end-to-end data processing. Data lineage documents information about data origin, manipulation, calculation, transformation, and final destination.

Data loss prevention (DLP) – The set of controls in place to ensure that users do not send critical information outside the organization.

Data owner – The person or entity responsible for safeguarding data, data classification, and granting or denying access.

Data quality dimensions – The Enterprise Data Management council has defined seven key dimensions associated with data quality: completeness, coverage, conformity, consistency, accuracy, duplication, timeliness.[7]

[6] The DAMA [Data Management Association International] Guide to the Data Management Body of Knowledge (DAMA-DMBOK), 1st Edition 2009, p.4

[7] http://www.edmcouncil.org/dataquality

Engagement objectives* – Broad statements developed by internal auditors that define intended engagement accomplishments.

Engagement opinion* – The rating, conclusion, and/or other description of results of an individual internal audit engagement, relating to those aspects within the objectives and scope of the engagement.

Engagement work program* – A document that lists the procedures to be followed during an engagement, designed to achieve the engagement plan.

External service provider* – A person or firm outside of the organization that has special knowledge, skill, and experience in a particular discipline.

Extract, load, and transform (ELT) – The process of moving data from the source system into a data warehouse. ELT refers to the order in which the staging steps occur: extract, load and transform. ELT uses the target system for the transformation.

Extract, transform, and load (ETL) – The process of moving data from the source system into a data warehouse. ETL refers to the order in which the staging steps occur: extract, transform, and load.

Fraud* – Any illegal act characterized by deceit, concealment, or violation of trust. These acts are not dependent upon the threat of violence or physical force. Frauds are perpetrated by parties and organizations to obtain money, property, or services; to avoid payment or loss of services; or to secure personal or business advantage.

Governance* – The combination of processes and structures implemented by the board to inform, direct, manage, and monitor the activities of the organization toward the achievement of its objectives.

Key performance indicators (KPIs) – A measure that determines how well a process is performing in enabling a particular goal to be reached.

Master data – The set of attributes used to identify and describe data.

Predictive analytics – The process of attempting to predict the future based on the analysis of data about past events. The process can use multiple techniques to create the predictions (e.g., data mining, statistics, modeling, and machine learning).

Radio frequency ID (RFID) – Small devices that contain an antenna and use electromagnetic fields to tag and track items and transmit information to devices capable of reading their signals.

Risk* – The possibility of an event occurring that will have an impact on the achievement of objectives. Risk is measured in terms of impact and likelihood.

Risk management* – A process to identify, assess, manage, and control potential events or situations to provide reasonable assurance regarding the achievement of the organization's objectives.

Scalability – The ability of systems to change size to accommodate changes in processing demand.

Service level agreement (SLA) – A documented agreement between a service or products provider and the organization that describes the minimum performance objectives, the mechanism to measure performance, and the consequences for missing performance goals.

Streaming data – Data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously.[8]

Structured data – Historically, the majority of data stored within organizations was structured and maintained within relational — or even legacy hierarchical or flat-file — databases. Structured data is organized and allows for repeatable queries, as much of this data is maintained in relational tables. It is often easier to control than unstructured data, due to defined ownership and vendor-supported database solutions.

System of record (SOR) – The authoritative source for a particular data element.

Unstructured data – This data, which is growing and becoming more common within organizations, is not confined to traditional data structures or constraints. It is typically more difficult to manage, due to its evolving and unpredictable nature, and it is usually sourced from large, disparate, and often external data sources. Consequently, new solutions have been developed to manage and analyze this data.

[8] https://aws.amazon.com/streaming-data/

Appendix D. Planning a Big Data Audit Engagement

In order to successfully audit or provide advisory services on big data programs, internal auditors must have an understanding of the related risks and challenges (see Standard 2200 – Engagement Planning and 2201 – Planning Considerations). The severity of risk will vary by organization and will depend on the strategic intent and operational deployment of these initiatives (see Standard 2210 – Engagement Objectives and Standard 2220 – Engagement Scope). The following sections provide additional detail on key risk areas, as well as example risk and control considerations for use in building an audit work program for big data (see Standard 2240 – Engagement Work Program).

Big Data Risks and Challenges

The potential benefits of implementing of a big data program come with significant risks and challenges. Internal audit must help ensure the organization's risks are identified, understood, and appropriately addressed. By managing big data risks to acceptable levels, management increases the likelihood of achieving planned business objectives and realizing the potential benefits of the big data program.

As described above, the primary risk areas impacting big data are:

- Program governance.
- Technology availability and performance risks.
- Security and privacy.
- Data quality, management, and reporting.

Big data programs and environments should also be subject to IT general controls. (See "GTAG: Information Technology Risk and Controls 2nd Edition" for additional information regarding risks and challenges related to IT general controls.)

Engagement Planning

Standard 2200 – Engagement Planning states that for each engagement, internal auditors must develop and document a plan, which must include the engagement's objectives, scope, timing, and resource allocations. One of the most important things internal audit needs to determine when planning the engagement is whether the organization has a unified and cohesive governance structure in place, including policies and procedures, from which clear and consistent guidance can be distributed across the organization. A strong governance model will provide the necessary policies, processes, and tools to consistently manage the environment and control the risks related to big data, which is essential for adequate protection of the organization's information. While multiple organizational functions may own part of the big data strategy, the key to a successful big data audit is to identify a single group of key stakeholders who can provide the necessary information to minimize business disruption and optimize business and audit resources.

Engagement Objective

In accordance with Standard 2210 – Engagement Objectives, internal auditors must establish engagement objectives to address the risks associated with the activity under review. A risk assessment should be performed to assist in defining initial objectives and to identify other significant areas of concern.

The audit objective for a big data audit can be defined in different ways. For example, the objective can be defined as part of the annual audit plan, or as a result of enterprise risk management efforts, past audit findings, regulatory requirements, or specific assurance needs from the board or audit committee.

Engagement Scope and Resource Allocation

Procedures to be performed and the scope (nature, timing, and extent) of the engagement should be determined after the risks have been identified. According to Standard 2220.A1, “The scope of the engagement must include consideration of relevant systems, records, personnel, and physical properties, including those under the control of third parties.”

The audit engagement should encompass strategy and governance (including policies, standards, and procedures), employee awareness, and training. Internal audit must determine the skills necessary to complete the audit engagement and the total number of resources required. The internal audit staff must have the appropriate level of expertise, knowledge, and skills to successfully perform the audit engagement, or external resources with the requisite competencies should be utilized.

It may be difficult to audit the entire big data program. Instead, the scope of the audit engagement can be defined by business unit, location, strategic objective, or any other criteria that are meaningful to the organization.

Engagement Work Program

In accordance with Standard 2240.A1, “Work programs must include the procedures for identifying, analyzing, evaluating, and documenting information during the engagement.”

The following is a sample work program for big data. Internal auditors can use this sample as the baseline to create a specific audit program that meets their organization’s needs.

Objective 1: Plan and Scope the Audit

Review Activities		Comments
1.1 Define the engagement objectives. (Standard 2210) The audit/assurance objectives are high level and describe the overall audit goals.		
1.2 Identify and assess risks. (Standard 2210.A1) The risk assessment is necessary to evaluate where the internal auditors should focus.		
1.2.1	Identify the business risks associated with big data that are of concern to business owners and key stakeholders.	
1.2.2	Verify that the business risks are aligned with the IT risks under consideration.	
1.2.3	Evaluate the overall risk factor for performing the review.	
1.2.4	Based on the risk assessment, identify potential changes to the scope.	
1.2.5	Discuss the risks with management and adjust the risk assessment.	
1.2.6	Based on the risk assessment, revise the scope.	
1.3 Define the engagement scope. (Standard 2220) The review must have a defined scope. The reviewer should understand the big data infrastructure, processes, and applications, and the relative risk to the organization.		
1.3.1	Obtain a list of documents related to big data that can help define the scope. For example: • List of locations using big data. • List of users. • List of reports generated using big data. • List of business processes that depend on big data for strategic decisions.	
1.3.2	Determine the scope of the review.	
1.4 Define assignment success. Success factors need to be identified and agreed upon.		
1.4.1	Identify the drivers for a successful audit.	
1.4.2	Communicate success attributes to the process owner or stakeholder and obtain agreement.	
1.5 Define resources required to perform the audit engagement. (Standard 2230) In most organizations, audit resources are not available for all processes.		
1.5.1	Determine the audit/assurance skills necessary for the review.	

Objective 1: Plan and Scope the Audit		
Review Activities		**Comments**
1.5.2	Estimate the total audit/assurance resources (hours) and time frame (start and end dates) required for the review.	
1.6 Define deliverables. (Standard 2410) The deliverable is not limited to the final report. Communication between the audit/assurance teams and the process owner is essential to assignment success.		
1.6.1	Determine the interim deliverables, including initial findings, status reports, draft reports, due dates for responses or meetings, and the final report.	
1.7 Communicate the process. (Standard 2201) The audit/assurance process must be clearly communicated to the customer/client.		
1.7.1	Conduct an opening conference to: • Discuss the scope and objectives with the stakeholders. • Obtain documents and information security resources required to perform the review effectively. • Communicate timelines and deliverables.	

Objective 2: Identify and Obtain Supporting Documents (Standard 2310)		
Review Activities		**Comments**
2.1	Review policies and standards governing big data.	
2.2	Review the IT infrastructure documentation and identify systems that support big data.	
2.3	Review system design documents.	
2.4	Review the interfaces diagram and identify systems that share data with the big data systems.	
2.5	Review the list of internal and/or external users.	
2.6	Review contracts with service providers.	
2.7	Review SLAs.	
2.8	Review performance metrics and remediation plans.	
2.9	Review the disaster recovery plan and test results.	
2.10	Review the business continuity plan and test results.	

Program Governance Risks

To successfully deploy a big data program, organizations must deploy and appropriately govern the necessary people, processes, and technology. Without adequate program governance, a big data implementation may expose the organization to undue risk, ranging from failed implementation and limited adoption to security and privacy issues. Organizations also face difficulties in designing metrics to measure the cost and value of big data programs. Executive leadership may choose to discontinue funding a big data program if the program value cannot be adequately demonstrated and communicated.

Finding skilled and competent individuals to sponsor, manage, and implement a big data program in a highly evolving technology landscape is a challenge all organizations face. The McKinsey Global Institute predicts that there will soon be a large shortage of data analysts as well as managers and analysts with the ability to harness big data to make effective decisions.[9] Colleges and universities are also having difficulty keeping curriculums aligned with rapidly changing business needs in order to create a pipeline of resources with relevant skillsets.

Technological evolution puts a greater emphasis on organizations to make the right decision on whether to build or buy big data solutions and services. Organizations that outsource some or all of their big data services face additional third-party vendor management, cloud security, and privacy risks.

Even with skilled people and technology in place, companies must have sufficient data governance and management processes to ensure various data quality dimensions are adequate to support organizational decision making. Enterprise data often exists in silos, which increases the complexity of identifying and inventorying critical databases, data elements, and data lineage.

Objective 3: Understand Big Data Program Governance	
Control Objective	**Description**
3.1 Funding should be adequate to support business needs.	Funding model(s) are chosen to support the initial design and implementation, ongoing activities (e.g., sustainable production support resources and technology maintenance through the full lifecycle), and recommended projects that result from the implementation of a big data program.
3.2 Program objectives should support enterprisewide strategy initiatives.	Program objectives and the business case are aligned with the enterprisewide strategy and initiatives to ensure the cost-benefit analysis supports the need to establish a big data program.

[9] http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Objective 3: Understand Big Data Program Governance

Control Objective	Description
3.3 Management should receive metrics that demonstrate goal achievement.	Metrics — both quantitative and qualitative — are designed, implemented, and monitored to demonstrate the value of the program.
3.4 The organization should establish a governing entity to manage the big data strategy.	A governing, cross-organizational structure exists to prioritize big data activities (e.g., order of source system integrations, selection of analytics, report development) to address concerns arising from competing priorities.
3.5 There should be agreed-upon SLAs between the business and IT to describe and measure performance expectations.	SLAs are designed and implemented to ensure consumer expectations are proactively managed (e.g., timing of report availability, frequency of data refresh, downtime windows).
3.6 Business and technical requirements should be documented, analyzed, and approved.	Business and technical requirements are gathered and analyzed to support the decision to build or buy (e.g., interval vs. cloud based) a big data environment and support the ultimate selection of a solution/technology service provider.
3.7 Executive management should develop a big data strategy that provides solutions across the organization.	The big data program is inclusive of all relevant key organizational areas to limit duplication of effort and redundant technology environments in the company.
3.8 Prior to approving the business case, management should conduct a proof of concept to validate that the systems designs align with strategic goals.	A pilot project (with known opportunities and without significant complexity) has been selected, and priority areas are identified for "wins" to further support the build-out of the program.
3.9 Roles and responsibilities should be clearly defined.	An executive sponsor, who provides strong executive-level support and sponsorship for the big data program, has been identified.
	Roles and responsibilities of data owners, stewards, and subject matter experts (SMEs) have been established and defined.
	A responsibility assignment (RACI) matrix has been documented and maintained for enterprisewide data governance roles and responsibilities.
	Roles and responsibilities for business partners that rely on big data solutions and reporting have been established and defined, including the necessary controls that these resources must implement to successfully consume data from these environments.

Objective 3: Understand Big Data Program Governance

Control Objective	Description
3.10 The organization should provide the necessary resources to deploy and maintain the big data strategy.	The organization has identified and funded critical positions to support the big data program, and has introduced appropriate talent into the organization with the requisite skills and experience to make the program successful. The skills assessment is periodically reperformed in alignment with the changing needs of the organization and the technology in use.
3.11 Third-party vendor management practices should be used to manage big data suppliers.	Contracts include adequate provisions on security, availability, support models, pricing, etc. Appropriate legal and control partner feedback is incorporated into the agreement prior to execution.
	Contractual agreements are monitored for third-party vendors who host and/or access big data environments. These contracts appropriately account for dynamically scaling the environment to support increased or decreased demand.
	Vendor roles and responsibilities are documented and approved in a master services agreement.
	SLAs are documented, approved, and monitored to ensure third parties meet minimum performance levels. Adequate penalties are defined and enforced when these SLAs are not met. Management has vendor governance routines in place to formally assess and take action on SLA results.
	Transition and termination considerations were factored into the agreement and overall solution assessment (e.g., What happens to the data when the contract ends? How long will it take to transition critical analytics for the organization?).
	Please refer to “GTAG: Information Technology Outsourcing, 2nd Edition” for additional guidance on third-party vendor risk management considerations.
3.12 Data governance should be part of enterprise governance.	Data management policies and related procedures (e.g., ownership, sharing, privacy, and retention requirements, etc.) are defined, documented, and shared.
	The organization has identified a CDO or equivalent role with responsibility for enterprise data governance.
	An inventory of critical databases, tables, data elements and lineage, and other metadata is created and maintained.
	Enterprise data standards are defined and communicated to all relevant employees as part of annual training requirements.

Objective 3: Understand Big Data Program Governance	
Control Objective	**Description**
	Data quality controls are implemented for critical data elements, with adequate governance in place to ensure remediation of identified data issues.
	Systems of record (SOR) and authoritative data sources (ADS) have been identified, and data is appropriately reconciled between SOR and the data warehouse(s).
	The enterprise data management team tracks and governs business compliance with data governance and management policies.

Technology Availability and Performance Risks

When big data platforms are not scalable for the rapidly increasing amounts of structured and unstructured data necessary for analytics, there may be degraded performance and inaccurate or untimely analytic outputs. As more data is extracted and loaded into big data systems, analytic model assumptions may need to be reviewed to determine whether the model and the model's underlying assumptions are still accurate.

The performance and availability of big data systems becomes increasingly important as the organization's reliance on these systems increases for key executive decision making and revenue generation processes. Big data systems cannot provide analytic outputs when the systems or related data feeds are unavailable. This can adversely affect the timeliness of management decisions, create a negative customer experience, and/or lead to lost revenue. High availability and/or disaster recovery solutions can mitigate system downtime risk, but come at an increased cost.

Big data programs can have significantly varying data storage and retention requirements. Certain streaming data, for example, may never need to be saved or backed up to an organization's systems, but might not be recoverable if the data is not processed and analyzed immediately. Other big data applications, however, may require significant historical data to understand and discover patterns or behaviors over extended periods of time.

Data lakes, whether internal to the organization or in the cloud, also pose significant risks. Data lakes represent a consolidation of detailed data from many different organizational systems in a centralized location, potentially serving as a single point of data exposure. Due to the fact that data lakes accept any and all data in its native format, well-defined metadata may not exist. These scenarios can lead to confusion about the true system of record for authoritative data for use in analytics.

Backup, storage, and retention policies and procedures should be commensurate with the analysis requirements and value of the data to avoid data loss. Programs such as backup jobs and production analytic jobs should be configured to send automatic timely notifications to production support personnel regarding job success as well as batch and real-time job failures. Not resolving job failures quickly can result in lost data and/or inaccurate analytic results, which is especially critical for streaming data.

Ineffective technology configurations may result in a negative customer experience, reduced system availability, and degraded performance. ETL/ELT tools that are not configured appropriately can result in big data platforms missing significant volumes of data. Network bandwidth may inhibit the movement of vast amounts of data when configurations are not optimized. Analyzing the upstream and downstream impacts of platform upgrades and maintenance can be challenging due to the rapid evolution of big data hardware and software technology.

Objective 4: Understand Technology Availability and Performance Risks

Control Objective		Description
4.1	IT operations should be structured in a manner that supports big data production up-time expectations.	Production support models are defined and agreed upon with the appropriate business partners to ensure adequate support of the organization (e.g., whether 24x7 support is needed; time to respond to end user problems or questions).
		High availability and/or disaster recovery solutions are implemented for big data systems to support system availability needs and minimize downtime in the event of an outage.
		Procedures are in place for the execution, monitoring, and recovery of data backups for big data systems.
		A documented production support process is in place to monitor real-time and batch jobs, alert appropriate personnel, and track incidents and problems from notification to resolution.
4.2	Data lifecycle policies and procedures should be documented and followed.	Data storage and retention requirements and procedures are documented and in place to ensure the appropriate storage, retention, and destruction of data.
4.3	Big data systems should be included in the patch management strategy.	IT monitors and patches big data applications and databases to ensure they are kept current and supported by vendors.
4.4	Big data systems should be part of the maintenance strategy.	A documented upgrade and maintenance process is in place to ensure big data environments are upgraded and maintained in accordance with company expectations and vendor support agreements.

Objective 4: Understand Technology Availability and Performance Risks

Control Objective	Description
4.5 Big data systems should be procured/built, and configured in alignment with the complexity and demand documented in the business case.	In accordance with a documented assessment, the organization deploys an environment or solution with the necessary hardware and performance to deliver against the organization’s needs, based on the size and complexity of anticipated analytics.
4.6 Systems and support tools should be configured to provide automatic notifications to support personnel.	ETL/ELT tools are configured to provide automated notifications of job completions and failures. IT examines record counts and control totals of completed jobs for reasonableness. IT resolves job failures with reasonable promptness and documents the reason for, and resolution of, each failure.
4.7 Big data systems should be part of the change management strategy.	As upstream and downstream systems are modified, changes to the impacted data feeds are appropriately incorporated into the project to limit data breaks and/or issues in the big data environment(s).
	Prior to upgrading/changing big data technologies, IT analyzes the impact on upstream and downstream interfaced systems to ensure the continued operation of big data systems.
	Network layer requirements for big data transmissions are documented and analyzed to ensure analytics can be transmitted successfully among various devices.
4.8 Reporting tools should be configured to be flexible, intuitive, and easy to use. Training aids should be provided.	Reporting tools and technologies are adequately configured and integrated with the big data environment to support the needs of end users. Sufficient training is provided on these end user tools.
4.9 Big data systems should be configured to allow flexibility and scalability without sacrificing performance.	Big data systems are designed and implemented to allow for scalability to ensure the necessary levels of transmission, storage, availability, and performance are maintained as system usage increases.
4.10 Periodic performance testing should be conducted and weaknesses remediated.	Performance testing is conducted on new and modified analytics to ensure results are produced within reasonable and expected timeframes.
4.11 The big data systems’ lifecycle should be properly managed.	Management continues to reassess the solution due to changes in the industry, technology, and external landscape, to maintain competitive and high-performing services (e.g., a newly recommended architectural change may drastically improve performance based on new knowledge in the industry).
4.12 The big data analytics model should be part of the maintenance strategy.	Analytics model maintenance is performed on a regular basis to ensure continued accuracy and reliability of big data analytics.

Objective 4: Understand Technology Availability and Performance Risks	
Control Objective	**Description**
4.13 IT general controls must be assessed periodically.	Please refer to "GTAG: Information Technology Risk and Controls 2nd Edition" for information regarding access controls, IT operations, system development life cycle (SDLC), and change management considerations. These controls are critical to the success, sustainability, and security of big data programs.

Security and Privacy Risks

Security and privacy are the risks most people can easily identify and understand. A person or organization can gain incredible insight into a person's life by combining personal information with social media commentary and "likes," publicly posted product and service reviews, and internet browsing history captured through web cookies. Organizations face difficulties ensuring all data collected is used for legitimate purposes and the organization complies with laws and regulations. Customers may feel uneasy about customized marketing campaigns driven by the analysis of personal data, even when the data is used for valid business reasons.

Additionally, news stories about public- and private-sector data breaches resulting in stolen personal data have become all too common, and costs associated with an organization's failure to protect the personal information of its employees, customers, and vendors are constantly increasing. Regulatory compliance sanctions and fines, which vary by location and jurisdiction, can result in significant legal and financial liability for the organization. Further, organizations that experience a data breach may suffer significant brand and reputation damage, leading to declining revenues and increased costs.

The threats and vulnerabilities associated with inappropriate insider access (e.g., employees, consultants, and big data vendors) are often as significant as those associated external threats, given the inherent knowledge and privileges possessed by these groups. Such insider actions may include stealing sensitive and confidential data, obtaining trade secrets, or taking inappropriate actions based on insider knowledge. Knowledge and insights gained from big data systems stolen for personal gain often go undetected because companies focus cybersecurity efforts on external threats and may have inadequate controls to prevent and detect insider activity. Account privileges should be strictly limited to the access needed to perform the individual's job responsibilities, and additional controls should be implemented to monitor and detect suspicious activity.

Ensuring all systems are appropriately and consistently secured becomes more challenging as big data systems become more complex and powerful and house larger volumes of disparate data. Inadequate patching or security configurations may open vulnerabilities that can be exploited to view or modify sensitive data. System disruptions may also occur, resulting in unavailable services and lost productivity.

Please refer to the IIA Practice Guide “Auditing Privacy Risks, 2nd Edition” for additional information regarding privacy risks and challenges, as many of these are quite relevant to big data programs and environments. Additionally, please refer to the “GTAG: Assessing Cybersecurity Risk: Roles of the Three Lines of Defense” for additional security-related risks and considerations.

Objective 5: Understand Big Data Security and Privacy	
Control Objective	**Description**
5.1 Information security management should be part of the big data strategy.	A cybersecurity program exists within the organization to combat internal and external threats.
	A hardened baseline security configuration is established to ensure a consistent and secure operating environment for big data systems and their infrastructure.
	System utilities capable of circumventing operating system, network, and application controls are prohibited or appropriately controlled.
	Access to, and use of, audit tools is segmented and restricted to prevent compromise, misuse, and/or destruction of log data. Log data is reviewed to identify suspicious activity.
	All cloud-based services utilized by the organization are approved for the use and storage of the organization’s data.
	IT evaluates the security of relevant service providers to address concerns regarding shared infrastructure, externally hosted systems, and vendor access to data prior to implementing a cloud-based or other third-party computing solution.
	Patch management processes are documented and implemented to ensure systems are patched with the latest approved patches in a timely manner (see “GTAG: Change and Patch Management Controls: Critical for Organizational Success, 2nd Edition”).
	Please refer to “GTAG: Assessing Cybersecurity Risk: Roles of the Three Lines of Defense” for additional information regarding cybersecurity risks and related controls.
5.2 Data security management should be part of the big data strategy.	Only authorized business users have access to data and reports from big data systems. Access is aligned to job responsibilities and based on the concept of least privilege.
	Only a small group of authorized technical users have privileged access to big data systems, including operating systems, databases, and applications.

Objective 5: Understand Big Data Security and Privacy	
Control Objective	**Description**
	End user reporting tools are appropriately configured to ensure only authorized personnel can view sensitive data.
	Access rights to big data systems are reviewed periodically to ensure their appropriateness.
5.3 Third-party access should be properly managed.	Security, contractual, and regulatory vendor requirements are addressed prior to granting access to data and information systems. Management assesses compliance with these provisions as part of vendor governance routines.
	Please refer to "GTAG: Identity and Access Management" for additional information regarding access controls, provisioning, security administration, and enforcement.
5.4 Data privacy should be part of the big data strategy.	Data is inventoried and classified to ensure the organization's critical data, including personal information requiring protection, is appropriately safeguarded.
	Personally identifiable information and other sensitive data is sanitized or scrambled prior to replication from production to development or test environments.
	An incident response process has been documented, approved, and implemented to ensure data breaches are handled appropriately.
	Please refer to the IIA Practice Guide "Auditing Privacy Risks, 2nd Edition" for information regarding privacy frameworks and principles regarding auditing privacy.

Data Quality, Management, and Reporting Risks

"Garbage in, garbage out" is a common phrase in the IT field. It means putting bad data into a system will result in bad data output from the system. Poor data or data quality issues may lead to inaccurate management reporting and flawed decision making. Databases that are not designed to ensure the integrity of the data can result in incomplete or invalid data. Analytics that rely on invalid data will most likely yield flawed results. Therefore, big data analytics must account for these data quality risks.

Additionally, data that is not obtained and analyzed in a timely manner may result in incorrect analytic outputs, flawed management decisions, or loss of revenue. Data sourced from third parties should be timely, accurate, complete, and from a reputable source. Third-party data that is in an inappropriate format may not be suitable for analysis and may delay management decision making.

After data has been received and analyzed, it may be challenging to ensure that end users manage and protect the data. A lack of end user computing controls may lead to inaccurate reports and data leaks. End user production reports, ad hoc reports, and predictive analytic outputs must all be reviewed and approved to limit flawed management decisions. Big data reports should also adhere to an organization's data classification policies to ensure only appropriate data is shared, both internally and externally. Report options and distribution channels may be appropriate only for data of specific sizes and formats. Organizations may face obstacles when determining the appropriate report options and channels for each analytic result.

Objective 6: Understand Big Data Quality, Management, and Reporting

Control Objective	Description
6.1 Policies and procedures should be established to ensure data quality.	Procedures are established and documented to ensure the quality of enterprisewide data is consistent with organizational data standards.
	Processes exist to formally agree on systems of record for critical data elements to support consistent resolution of data differences between systems.
	Big data databases are designed and implemented to ensure data integrity, including appropriate reconciliation and monitoring of key data elements to source systems or systems of record.
	Processes are in place to ensure the timely receipt of data into big data systems and the timely delivery of analytics to big data consumers.
	For critical end-user-computing (EUC) applications, management implements measures to preserve and protect the integrity of data and calculations that used the data. Examples include access controls, formula identification, check totals, etc.
	Requirements for preparing the data inputs for big data applications and analytics are established and documented.
	Data is available in a timely manner, enabling prompt decision making in line with business requirements and needs.
6.2 Policies and procedures should be established to ensure that data obtained from third parties complies with data quality standards.	Procedures are established and documented to ensure data purchased from third parties is accurate, complete, and from a reputable source.
6.3 Policies and procedures should be established to ensure report accuracy.	Procedures for reviewing and approving the output from analytic reports are established and documented. Predictive analytics that drive action receive heightened focus and attention.

Objective 6: Understand Big Data Quality, Management, and Reporting		
Control Objective		**Description**
6.4	Access to reports should be granted based on business needs.	Reporting elements that are appropriate for internal versus external consumption have been defined and documented. Data owners have approved consumption of these reports and data by authorized individuals.
6.5	Reporting tools and procedures should allow for flexibility and ad hoc reporting.	Procedures exist for when ad hoc reporting is allowed versus when it is permissible to rely on approved production reports, which cannot be altered without working through change management and system development processes.
6.6	Users should be trained periodically to maximize report utility.	As underlying data and reporting methods are modified, end users are trained on the resulting changes and downstream impacts to their reports, end user systems, etc.
6.7	Report vendor selection should be in alignment with business needs.	A process is in place for selecting appropriate and preferred reporting options/channels (e.g., Tableau®, Splunk®, Oracle® Business Intelligence Enterprise [OBIEE]) for analytic outputs.

Contributors

Brian Allen, CISA, CISSP

Stephen Coates, CIA, CISA, CGAP

Brian Karp, CIA, CISA, CRISC

Hans-Peter Lerchner, CIA, CISA

Jacques Lourens, CIA, CISA, CGEIT,CRISC

Tim Penrose, CIA, CISA, CIPP

Sajay Rai, CISM, CISSP

The IIA thanks the Information Technology Guidance Committee (ITGC) for its support in developing this guidance.

About The IIA

The Institute of Internal Auditors (IIA) is the internal audit profession's most widely recognized advocate, educator, and provider of standards, guidance, and certifications. Established in 1941, The IIA today serves more than 190,000 members from more than 170 countries and territories. The association's global headquarters are in Lake Mary, Fla., USA. For more information, visit www.globaliia.org.

About Supplemental Guidance

Supplemental Guidance is part of The IIA's International Professional Practices Framework (IPPF) and provides additional recommended (nonmandatory) guidance for conducting internal audit activities. While supporting the *International Standards for the Professional Practice of Internal Auditing*, Supplemental Guidance is not intended to directly link to achievement of conformance with the *Standards*. It is intended instead to address topical areas, as well as sector-specific issues, and it includes detailed processes and procedures. This guidance is endorsed by The IIA through formal review and approval processes.

For other authoritative guidance materials provided by The IIA, please visit our website at www.globaliia.org/standards-guidance or www.theiia.org/guidance.

Disclaimer

The IIA publishes this document for informational and educational purposes. This guidance material is not intended to provide definitive answers to specific individual circumstances and, as such, is only intended to be used as a guide. The IIA recommends that you always seek independent expert advice relating directly to any specific situation. The IIA accepts no responsibility for anyone placing sole reliance on this guidance.

Copyright

April 17