Introduction

The H-ISAC Big Data Controls Working Group (BDCWG) provides a venue for sharing information regarding the challenges and opportunities associated with big data systems. The working group is focused on identifying both information control best practices for big data systems and data analytics best practices for security and business applications. Additional information regarding the working group can be found in Appendix A.

This report is intended to provide a set of recommended information control best practices for big data environments and also raise awareness regarding some of the challenges associated with protecting information in these environments. The intended audience includes organizations that have implemented or are considering implementing a big data environment. Additionally, the information presented in this report serves to facilitate the alignment of various stakeholders in the big data space, including security professionals, big data engineers, data scientists, and consumers. The consumer groups associated with these systems includes business analysts, security analysts, data scientists, business operators, and decision makers. Unlike traditional database environments, modern big data environments generally support the storage and processing of a large variety of different data feeds and simultaneously serve as an analytic platform for generating new information from data stored within the environment. Appropriate controls must be established to protect information in these systems while still enabling analytic processing and access by a broad set of consumers.

The report begins by providing a brief narrative of big data and a description of the types of systems that are the focus of this report. Next the topic of data analytics and security controls is addressed followed by a discussion of specific risk areas associated with big data environments. An overview of data encryption, including some emerging and novel techniques is provided, followed by a set of general recommendations and a list of suggested policies for big data systems.

Big Data

The term *big data* is used to describe large volumes of structured and unstructured data that exceed the storage and compute capacity of traditional database environments. There are a variety of definitions that exist but most often big data is described in terms of extremely large *volume*, *variety*, *velocity*, or *veracity*. It should be noted that many data sets that we consider *big* by today's standards will not be viewed as such in the future, as storage and computational infrastructures continue to increase and improve. There are many challenges inherent with big data, including data storage, query, transfer, and visualization, and numerous new capabilities have emerged in recent years to address these challenges. Most products in the big data space are concerned with providing enhanced capabilities to increase productivity, provide new efficiencies, and enable new information discovery. Security controls, however, is often one of the last topics discussed by engineers developing these new breed of data management systems. Moreover, because big data technologies are still new and rapidly evolving, most security practitioners lack a sufficient understanding of these environments and are not able to provide sufficient information control recommendations.

Big data is having a significant impact across multiple industries, including healthcare. Within the healthcare space big data applications are being used to create efficiencies – thereby improving profits –

and also improve disease detection, prevent deaths, and improve overall quality of life. [1] Various companies have developed numerous new sophisticated analytic techniques that leverage big data and machine learning to predict diseases and augment the work done by medical professionals (e.g. IBM Watson Health [2]).

There are several different flavors of big data systems available in the market today and many vendors adopt the buzzword *big data* to help market their particular products. However, capabilities that leverage the open source Apache Hadoop ecosystem dominate today's big data space. Hortonworks and Cloudera are the primary distributers of these products; they offer professional services to facilitate the installation and maintenance of these systems for organizations around the world. Given that these software tools are open source with a large developer community the landscape of new capabilities is evolving very rapidly. This creates a significant challenge for security professionals who must help ensure that protected data is properly secured in these environments and appropriate controls are in place to mitigate risk.

While big data environments are becoming increasingly popular, especially for large organizations, it might not be the best fit for everyone. Organizations should consider what constraints exist in their environments that might necessitate the migration from a traditional database environment to a big data system. Usually this is driven by the need to store and analyze more data than what is possible with their existing environments. Cost is also another key factor, as organizations weigh the long term cost of maintaining and expanding their existing infrastructure to support increased storage and computational capacity.

Data Analytics

Equally as ambiguous and overused as *big data* is the term *data analytics*. In the context of big data systems data analytics provide the cognitive functions to help make sense of data that is ingested into these environments and produce actionable output from that data. This is critical given the large volumes of data that can't be analyzed manually. For our purposes we will subdivide the data analytics space into two subcomponents: *visual models* that present data in a form that can be interpreted by an individual and *computational models* that process data and produce some mathematical output (e.g. a risk score).

There are numerous items that need to be considered when implementing and managing data analytics in a big data environment. One challenge is to balance the speed of innovation with appropriate regulatory guardrails to ensure compliance with applicable rules and regulations while safeguarding protected information. Below we outline a few areas that must be considered regarding analytics that operate in big data environments.

Model Governance

Big data systems support data storage and analytic processing to produce new data. In order to understand risk and apply appropriate controls it is necessary to have a governance structure around the creation and deployment of new analytic models. This isn't a popular topic among data scientists, as governance models can stifle innovation and slow down development.

Education is a key component, as data scientists working on behalf of some organization might not be aware of specific rules concerning data handling for different data sets. Creating a simple catalog of analytic models – which includes the data source(s), analytic description, and model output – is an easy way to provide the transparency necessary for managerial as well as security and regulatory oversight.

Derived Data

Through the application of computational models new data sets can be produced in big data environments that are derived from other data sets. One can imagine different scenarios where data stored in different tables in a big data environment (e.g. HR records and employee name linked by employee ID) could be combined to produce a more sensitive data set than the original sources. Moreover, data that is encrypted as part of existing control standards might be decrypted to support some analytic process and the results might not be re-encrypted. Software developers and data scientists must consider data lineage associated with computational models and ensure that derived data is properly labeled and protected.

Automated Decisions

When possible, organizations attempt to automate different tasks to free up resources that can be applied to other projects. Data analytic models often provide multiple efficiencies including automated decision-making. In sophisticated environments analytic models can automatically adjust controls or notify another person or system if an anomaly is detected. Aside from the obvious task of fully testing analytics involved in automated decision-making organizations must also consider potential regulatory violations that may arise by such automated processes (e.g. European Union privacy regulations).

Security Data Analytics

Many organizations use big data systems to support security operations, including incident investigations, fraud detection, and improved identity and access management. While the large scale storage and compute environments in use today were not initially designed for the purpose of improving security, the architectures are very well suited for that purpose and are often used to either augment or replace Security Information and Event Management (SIEM) systems. More specifically, analytics for security that leverage big data environments are becoming increasingly popular and there are a plethora of new vendors in this space that provide security analytic capabilities built on big data environments.

Risks

As new technologies and business practices are introduced into organizations they often come with new risks that must be controlled and mitigated; big data environments are no different. Cybercriminals will attempt to exploit any potential vulnerability for political or financial gain if the payoff is worth their time and effort. Big data environments represent a ripe target for some cybercriminals – data sets from multiple sources are aggregated in one place. For Healthcare organizations this can mean access to large volumes of member and employee Protected Health Information (PHI) and Personally Identifiable Information (PII). Proponents for big data systems argue that while aggregating sensitive data sets from multiple sources in a multi-tenant environment does present some additional risks, it’s much easier to maintain controls and monitor access for one system instead of managing protected data that is stored in multiple disparate repositories across an organization with varying levels of controls that aren’t documented or well understood. Below we outline a few characteristics of big data systems that must be considered when assessing risk and implementing appropriate control standards.

Multi-Tenancy

One key characteristic of modern big data environments is the multi-tenant nature of these systems. In addition to system administrators and privileged users that might have access to the underlying infrastructure, data scientists represent a new type of user that has access to sensitive information but might not understand how their work might negatively impact confidentiality, integrity, and availability. Typically data in big data environments can be protected from unauthorized use through the use of network groups (e.g. Active Directory). Additional controls, including data encryption, adds supplementary layers of protection and further segments data access to smaller groups of individuals. Certain data sets can also be filtered or masked for different user groups using access controls; cell-level data encryption is one technique that is commonly used.

Data Variety

In addition to the large volumes of data from structured and unstructured sources that are typically stored in big data systems careful attention must be paid to the types and variety of data as well, including retention policies and inherent access controls that must be applied. While many traditional infrastructures might segregate customer data, employee data, and security data, it is common to physically collocate and logically separate these different data sources in big data environments. Big data systems can contain potentially harmful data, including privacy regulated data (e.g. PII, PHI) as well as security and infrastructure data that could expose critical information about an organization’s environment and controls. In addition, as noted in the previous section, analytic processes can produce new data sets that are byproducts of multiple other data sources with potentially conflicting retention and access control policies.

Know Your Data and Your Environment

Before assessing risk and developing specific controls organizations must first understand what information they are trying to protect. Below we outline several items to consider related to understanding the data and associated systems in an organization’s environment.

What Is It?

Most organizations manage some type of protected data that needs to be safeguarded against cyber-attack. Organizations in the healthcare sector typically maintain protected data for both their employees (e.g. PII) and also their customers (e.g. PII and PHI). Health-related data sets are very sought-after commodities by cybercriminals due to their value on the black market for resale and for use in other malicious campaigns (e.g. phishing). Infrastructure logs and security event data are also valuable data sets for cybercriminals, as they can provide a mapping of the organization’s network infrastructure and expose potential control gaps that could be exploited. Before attempting to put in place blanket controls a critical first step is to assess one’s environment and identify the critical assets that might be targeted by cyber-attack.

Where Is It?

Whether all protected information resides in one location or it is spread out across multiple repositories it’s critical to identify where protected data is stored and what controls are in place. Adding sophisticated controls to protect data in a big data environment is of little use if that same data is also stored in other environments without sufficient security controls in place. One should also consider whether data is permitted to be stored in a cloud-based environment such as Amazon Web Services (AWS). Many big data vendors offer cloud-based solutions that are easily and quickly deployed. Regardless of the environment or the location, data governance is a key aspect of any security environment, including big data systems, and access to protected data should be limited only to individuals who require it to perform their job duties.

Classification

If data sets have rules that govern the handling and sharing of data, those same access controls must be implemented in the big data environment. Big data systems offer a wide variety of configurations for these types of use cases, including restricting access at the cell (an element within a specific record) level. Automated rule testing can be implemented to help support proper classification and dissemination of protected information.

Data Encryption

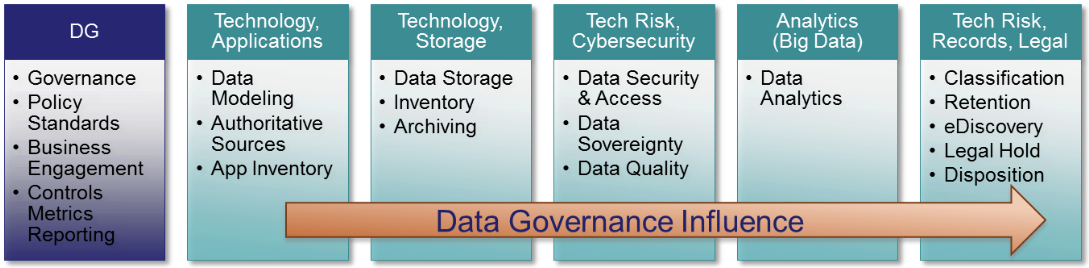
A key component concerning the protection of data in big data environments is data encryption. There exist multiple encryption and key management solutions for encrypting data in big data environments and the best solution will depend on the organization’s specific requirements. Some solutions offer format-preserving solutions that integrate into big data environments and there exist different capabilities within the Hadoop ecosystem to support data encryption for bulk files at rest. However most agree that new solutions need to be developed that support the encryption of protected data while not hindering the development and application of computational models.

As mentioned previously, big data systems support data storage, data query, and the execution of analytic models that process data and produce new information. Unfortunately, typical analytic models that are deployed to run in big data environments cannot make sense of encrypted data without first decrypting it. There are various ways organizations deal with this issue, including only encrypting specific fields and limiting the analytic process to only consider only the unencrypted fields. A technique known as *homomorphic encryption* is an encryption scheme that enables operations to be performed on encrypted data such that the operations are consistent with what would be performed on unencrypted data. There are multiple partially homomorphic systems offered by different vendors that can perform certain specific operations on encrypted text. A *fully homomorphic encryption* (FHE) system supports any arbitrary computation (e.g. search, addition, multiplication) on encrypted text. While FHE systems are not yet practical due primarily to computational limits, many different derivative solutions already exist today. These encryption systems will have a significant impact on data protection in big data systems as they become more computationally feasible.

One risk that often goes overlooked is the monitoring of data decryption and potential exfiltration in a big data environment. If an individual or process is authorized to decrypt data for the purpose of analysis or analytic modeling, controls should be put in place to monitor that activity and ensure that protected data that was once encrypted is not decrypted and exfiltrated from the big data environment.

Data Governance

Data governance plays a critical role in big data systems and serves to drive accountability and transparency necessary for successful operations in these complex environments. An effective data governance strategy will serve to promote effective management of information that optimizes business benefit while maintaining acceptable levels of risk. The figure below illustrates the role of data governance across multiple organizations and business areas.



DG	Technology, Applications	Technology, Storage	Tech Risk, Cybersecurity	Analytics (Big Data)	Tech Risk, Records, Legal
• Governance • Policy Standards • Business Engagement • Controls Metrics Reporting	• Data Modeling • Authoritative Sources • App Inventory	• Data Storage • Inventory • Archiving	• Data Security & Access • Data Sovereignty • Data Quality	• Data Analytics	• Classification • Retention • eDiscovery • Legal Hold • Disposition

An effective data governance strategy will leverage partnerships across multiple teams, including Legal, Compliance, Privacy, Information Technology, and Security. The benefits include increased operational efficiencies, improved regulatory & audit support, and reduced cost of operations. In big data environments, where multiple disparate data sources are aggregated and new information is produced from analytic models, data governance becomes even more critical and must be an integral part of an organization's big data strategy.

Talent Management

There are significant implications to talent management for the cyber professional going forward. As additional controls and enabling capabilities from vendors include new analytic techniques such as machine learning, the cyber security professional needs to understand the role of models and basic data science principles to design and deploy security controls going forward. Conventional cyber security curriculums (which are much improved today from previous years) need to be upgraded with data science principles and techniques to be relevant to meet enterprise requirements as they evolve.

Historically, data science and cyber security curriculums and disciplines have been exclusively independent, and for those enterprises employing both disciplines they have remained independent. Conventional wisdom and common practice is to separate cyber security professionals from data science professionals based on previous attempts to converge them that failed. Data scientists gravitate towards analyzing data results; security professionals gravitate towards actionable intelligence and a focus on enhancing controls. There are many cases where separation between analysts and data science engineers has been found to yield better results. However, as computational capacity increases new machine learning models are being developed that form a growing base of security capabilities that support real-time front line security controls.

Talent management techniques and practices need to be modified to expose security engineers and analysts to data science and the techniques that are most commonly used to analyze data. Many of the enterprise controls deployed in the future will require an understanding of these techniques. Security analysts today are likely to be regularly exposed to security intelligence associated with data breaches involving millions of records exposed publicly or on the Dark Web. Analyzing hashes of credential data to determine enterprise impact has become routine and data analysis techniques are essential to this practice. The concentration of security controls driven by models will continue increasing pressure on security professionals to learn practical data science techniques. Helping them learn and exposing them to the fundamentals of data science is the responsibility of the CISO today and going forward.

General Recommendations

In this section we outline some general recommendations for big data environments.

- Governance: Establish governance and oversight teams with appropriate security and privacy subject matter experts to oversee the following activities:
    - Access to the big data environment and to protected information stored in that environment
    - Analytic model management, to include the oversight of all analytic models used for production
    - Oversight and review of user and process activity in the big data environment, to include monitoring anomalous user or process activity that might pose a security risk
- Activity Monitoring: User activity should be logged and monitored. Emphasis should be placed on high-risk applications and users.
- Vulnerability Scanning: Routine vulnerability scanning should be implemented and assessment by an external vulnerability assessment team should be conducted on a periodic basis.
- System Documentation: Document and publish system architecture designs to include information pertaining to which nodes in the big data environment are accessible to which external applications.
- Data Provenance: Establish procedures for maintaining data provenance for the big data environment to include a catalog of all incoming and outgoing data feeds.
- Model Validation: Develop a structure for analytic model governance to include validation of analytic results, compliance oversight, and performance testing.
- Data Integrity: Establish procedures for maintaining data integrity within the big data environment.

Recommended Policies and Procedures

In order to ensure that big data environments adhere to the recommended guidelines it’s helpful to fully document all policies and procedures associated with various types of activity. Policies represent a set of guidelines that serve to document acceptable practices and ensure transparency of operations across all stakeholders. The associated procedures correspond to specific implementations of the different policies, and can vary depending on the organization. Below we outline some recommended policies for big data environments.

- Access Controls: Establish the roles, responsibilities, and requirements for controlling access to data.
- Data Quality: Establish the roles, responsibilities, and requirements for ensuring data quality and integrity.
- Data Ingestion: Establish processes for bringing data into the environment for storage and processing, to include retention policies and classification.
- Data Encryption: Determine what data requires encryption and the process for encryption.
- Data Decryption: Establish processes for data decryption and governance of decryption activities.
- Environment Monitoring: Establish tools and processes for monitoring the overall system usage and health.
- Protected Data Monitoring: Monitor user activity and metrics associated with protected data.
- Tenancy Guidelines: Establish processes for individuals and tools to gain access to the system.
- Data Cataloging: Document all data stored in the environment.
- User Activity Logging: Log all user activity and monitor activity.

Appendix A

The Big Data Controls Working Group concluded two years with this collaborated white paper on Big Data Security Controls. This group has since decided to focus on analytics and merged with the Cybersecurity Analytics Working Group.The H-ISAC Big Data Controls Working Group (BDCWG) provided a venue for H-ISAC members to share information regarding the challenges and opportunities associated with big data systems, with the objective of identifying security control best practices in this emerging technology space.

Bibliography

[1] "Big Data Security and Priivacy Handbook: 100 Best Practices in Big Data Security and Privacy," Cloud Security Alliance, 2016. [Online]. Available: https://downloads.cloudsecurityalliance.org/assets/research/big-data/BigData_Security_and_Privacy_Handbook.pdf. [Accessed 23 December 2016].

[2] "Big Data," IBM, [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[3] B. Marr, "How Big Data is Changing Healthcare," Forbes, 21 April 2015. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#567bd9622873. [Accessed 4 May 2017].