Developing Security Intelligence in Big Data

Hardik A. Gohel and Himanshu Upadhyay

Abstract In today's world, as the volume of digitized data grows exponentially, the need and the ability to store and computationally analyze large datasets are growing along with it. The term "big data" refers to very large or complex datasets, such that classical data processing software applications are insufficient to manage. A great example of a company that symbolizes the modern mass data-driven world is Google. It is possibly the most successful IT company in the world as well as the largest data processing company of modern times. In April 2004, Larry Page and Sergey Brin wrote their first and now famous "Founders Letter" to their employees which stated "Google is not a conventional company. We do not intend to become one." Twelve years down the line, with a change in leadership, incoming CEO Sundar Pichai wrote a letter to employees in 2016 and concluded it with "Google is an information company. It was when it was founded, and it is today. And it's what people do with that information that amazes and inspires me every day." There are many challenges in the analysis of large volumes of data, including data capture and storage, data analysis, curation, searching, sharing and transfer-ring, data visualization, data inquiry and updating, among others. However, the biggest challenge is information security and privacy of big data [29]. A lack of securi-ty around big data can lead to great financial losses and damage to the reputation for the company. Security threats and attacks are becoming more active in violating cyber rules and regulations. These attacks also affect big data and the information contained in it. Attackers target personal and financial data, or a company's confidential intellectual property information, which greatly affects their competitiveness. The biggest threat is when attackers target personal or consumer financial information stored in big data. Although there are rules and regulations in place to protect data, there are still vulnerabilities in big data that are serious enough to warrant substantial concern. In a recent and highly publicized incident, WikiLeaks released a huge trove of alleged internal documents from the US Central Intelligence Agency (CIA). It is by far the

H. A. Gohel (✉) · H. Upadhyay
Florida International University, Miami, FL, USA
e-mail: hgohel@fiu.edu

H. Upadhyay
e-mail: upadhyay@fiu.edu

S. Margret Anouncia and U. K. Wiil (eds.), *Knowledge Computing and Its Applications*, https://doi.org/10.1007/978-981-10-6680-1_2

largest leak of CIA documents in history. There are thousands of pages describing sophisticated software tools and techniques used by the agency to break into smartphones, computers, and even Internet-connected televisions. Both government and corporate leaks have been made possible due to the ease of downloading, storing, and transferring millions of documents in a very short time. With this state of affairs in mind, there needs to be a comprehensive examination of these threats and attacks on big data, and a study of novel approaches to defend it. This chapter presents an in-depth look into the threats and attacks on big data and inspects the methods of defense and protection. We discuss the vulnerabilities of modern big data systems, and the characteristic methods of intrusion, and unauthorized seizure of data. We present a few case studies of big data weaknesses and their exploitation by attackers. The information offered here is very useful in building proper defenses against potential malicious incidents. We also discuss the specific security demands of big data environments in government and medical sectors.

Keywords Big data · Security threats · Cryptography · Digital footprints
Machine learning

1 Introduction of Big Data Security

Big data is a term for datasets that are so huge or complex that usual data processing software is insufficient to manage them. The possible challenges of big data are to capture, store, analyze, exchange, curate, share, search, visualize, update, and query data. The expression "big data" normally refers to the utilization of predictive analysis, client behavior analysis, or other data analytics strategies that concentrate on extraction of value from information, and occasionally to a specific size of the dataset [2, 28]. There is little uncertainty that the amount of information now accessible is expansive, yet that is not the most relevant feature for this new information ecosystem [3]. Analysis of datasets can discover new connections to figure out business patterns, avert diseases, fight crime, and so on [27]. Business executives, scientists, medical practitioners, the media, and government agencies routinely face challenges with vast datasets in realms such as Internet searches as well as urban and business informatics. Researchers experience several restrictions in e-Science work, including genomics [4], meteorology, complex simulations, connectomics, and biological and ecological research [21].

Datasets can develop quickly since they can be progressively assembled by numerous cheap data-detecting Internet of things (IoT) gadgets, for example, cell phones, aerial(remote sensing), cameras, software logs, amplifiers, radio frequency identification(RFID) devices, and wireless sensor networks (WSNs), [24]. The world's technical per capita ability to store data has almost doubled every 40 months since 1980 [9] with almost 2.5 exabytes (2.5×10^{18} bytes) each day. One question for major organizations is to determine who ought to possess the big data activities that influence the whole organization [20].

Relational database management systems (RDBMS) and desktop statistics and visualization application packages experience issues with managing big data. This task may need massive parallel programs executing on a great multitude (almost one thousand) of servers (Villanova University). What can be considered as big data differs, relying on the abilities of the users as well as their tools, and extending capacities make defining big data a moving target. For a few organizations, managing gigabytes of information may in fact trigger the need to rethink the options for data administration. For others, however, several terabytes may be gathered before the size of data would need to be considered [8].

Big data is thus a term that depicts the huge volume of data—structured, as well as unstructured—that the business experiences every day. However, it is not the specific measure of this data which is imperative; it is, in fact, what the firms do with it that matters. Big data could be broken down for any intelligence that could prompt better choices and key business decisions.

While the expression "big data" is comparatively new, the collaboration of socially available data and its further analysis has long been practiced. However, the extent and scope were limited. This idea became important in the mid-2000s when industrial analysts contemplated the now-standard meaning of big data through the following three Vs:

- Volume—Organizations gather information from an assortment of sources, including business exchanges, online networking, and the data from sensors or even machine-to-machine information. Storage would have been an early issue although new advancements (for example, Hadoop) have reduced this weight.
- Velocity—Data rushes in at enormous speeds and needs to be managed instantaneously. RFID labels, sensors, and smart meters drive the need to manage torrential data in real time.
- Variety—Data comes in a wide range of structures and formats, including an unstructured document containing text or a set of organized numeric information stored in conventional databases. Unstructured data could even extend to forms like e-mail, audio, video, stock ticker information, and monetary transactions.

Another V, Veracity, has been added by a few firms to depict it (Villanova University), with the frequency of revisions being tested by industrial authorities [8]. Further, the 3Vs have been extended to other corresponding qualities of big data.

Digital footprints: Big data usually is a free-of-cost result arising due to the digital interactions (Source: www.bigdataparis.com).

Machine learning: Big data with respect to machine learning doesn't requires reason to identify patterns. In other words, we don't need to worry about the reason to discover patterns and correlations in the data which offers novel and great insights [15].

Big data is gathered from several sources. Sensors used to assemble atmospheric data, social media posts, digital images and video, transaction records (usually purchases), as well as wireless GPS signals, are some examples. On account of cloud computing and Internet socialization, several *peta*bytes of unformatted information are generated on the Web every day and most of this data would have an inherent business value on the off chance that it could be recorded and analyzed.

Mobile communication firms, for instance, gather information from cellular towers; gas companies gather information from seismic investigations and refinery sensors; electricity utilities collect data from not just power plants, but also circulation grids. Organizations gather a huge amount of user-generated data from clients and prospects including debit/credit card numbers, social security numbers, information on purchasing propensities, and utilization patterns of customers. This ingestion of big data and the requirement to circulate it all through the firm has actually provided a potential focus area for cyber-criminals and hackers. This information, which was earlier unusable by the firm, is now extremely important and, being liable to security laws and compliance conditions, needs to be secured.

All in all, as we can see, big data is a very well-known feature; so, what are we truly examining? Here, we look into the security issues, which include two distinct focal points: securing the firm and its clients' data in a big data setting and utilizing big data methods to examine, and even anticipate, security lapses. Be that as it may, security and protection issues are amplified by the velocity, volume, and veracity of big data, for example, massive cloud reserves, diversity of information sources and data formats, streams of data being acquired, and high-volume cloud-based migration. Conventional legacy systems for security, which are customized to smaller static data, are thus deficient.

2 Big Data Architecture Vulnerabilities (www.cisoplatform.com)

2.1 Existing Big Data Architecture

Big data is basically quite different from conventional relational databases in aspects of requisites and architecture. As we have seen, big data is normally described by 3Vs (volume, velocity, and variety). Some of the basic differences of the big data architecture include:

- **Distributed architecture**: The architecture for big data is largely distributed, scaling up to almost 1000 s of storage and data processing nodes. The data is partitioned horizontally, duplicated, and then shared across the available data nodes. Consequently, the big data architecture is quite resilient as well as fault tolerant.
- **Real-time computations**: The processing of such data needs to be continuous in nature, supporting real-time computations, which is expected to succeed the current batch processing supported by Hadoop.
- **Ad hoc queries**: Big data permits data analysts to extract optimal results by executing appropriate queries to analyze the big data sources.

- **Powerful, parallel programming language**: Extremely complex and largely parallel calculations, which are more computation intensive, need to be performed in big data rather than the usual PL/SQL and SQL queries. Hadoop, for example, uses the MapReduce framework, usually written in Java, in order to perform calculations on data processing nodes.
- **Easier code relocation**: With big data, it is easier to move and relocate the code rather than the data.
- **Non-relational data**: Big data is largely non-relational, as compared to the usual databases that follow the traditional relational approach. The primary benefit of data being stored in a non-relational form is in its ability to accept and hold huge volumes of data that exhibit considerable variety.
- **Automatic tiers**: The most frequently accessed (hottest) big datasets are tiered into high performing media, whereas the coldest ones are accordingly sent to cheaper, high-volume disks. Consequently, it is very difficult to precisely know where the data would be placed among the possible data nodes.
- **Data input from various sources**: Data collected from a variety of sources, for example system logs, social media, end-to-point devices, comprises big data.

Keeping these features of big data in mind, we could outline the following vulnerabilities that could render the existing big data architecture inadequate.

I. Insecure computation

An insecure program could pose as a major security challenge for any big data solution. Particularly, these would include any insecure program that could:

- Access sensitive or confidential data including personal particulars, age information, and credit card details.
- Corrupt the data, causing results to be incorrect.
- Present denial-of-service to the solutions proposed to the big data, in turn inflicting financial loss.

II. End-point validation of input/filtering

Big data, as specified earlier, gathers data from a huge variety of sources. As a result, two major challenges arise during the process of data collection:

- **Data filtering**: This approach aims to categorically filter the data which is suspected to be malicious or rogue data.
- **Input validation**: Another mechanism includes a clear understanding of what data can be trusted and what cannot. Thus, there would be efforts made to identify whether the data has been received from valid sources or not.

The massive amount of data being collected in big data renders it impossible to filter or validate the input data. An additional feature of the data (i.e., its behavior) creates another challenge for input data filtering and validation. Usual signature-based filtering of data may not provide a complete solution to the data filtering and input validation problem. A malicious or rogue data source could, for instance, insert large volumes of legitimate although incorrect figures into the system to alter the expected results.

III. Granular access control

Current big data solutions have been designed by keeping the performance and scalability aspects in mind, without focusing much on security. This is quite in contrast to existing relational database management systems that have quite considerable security features such as access control at various levels, such as user, table, row, and even up to cell levels. However, several constraints limit the provision of comprehensive-level access controls for a big data store, such as:

- Big data security is still an ongoing research concern.
- The non-relational character of data does not conform to the traditional realms of access control, viz., tables, rows, and cell levels. Present databases such as NoSQL depend on third-party packages or middleware applications to provide suitable access control solutions.
- Unplanned queries create added challenges regarding access control. For instance, the end user may have submitted SQL queries that would be valid for relational databases.
- Further, access control by default is disabled. This gives a suitable explanation for the practical problems encountered when trying to provide access controls at a global level to the big data.

IV. Insecure storage and communication of big data

The storage and communication of big data pose several additional security challenges:

- **Distributed data nodes**: Big data, in order to optimize its data storage, utilizes a large collection of data nodes that may even be distributed. This introduces a great challenge for the authorization, authentication, and encryption of the data at every node.
- **Auto-tiered data**: The tiering of data, as we saw earlier, is a method to optimize the process of data storage and retrieval. Such a procedure, being performed automatically, could incorrectly save very sensitive data on low-cost, less-sensitive media.
- **Real-time analysis and continuous calculations**: The effectiveness of analytical processing would require low latency to be maintained for the execution of queries. Therefore, the steps of encryption, and further decryption, could inflict overheads to system performance.
- **Transactional logs**: The transactions performed on the big data would create huge volumes of transaction logs, which could be considered as another big data that should be protected in the same way as the data.

V. Invasion of privacy by data mining and analytics

Big data is being visualized as a ready source of data for many kinds of studies, surveys, etc. Such monetization of these big data sources would in turn involve data mining as well as analytics on a large scale. However, this raises several concerns regarding the security of this data, which could in fact be drilled into and misused, thus creating a possible invasion of privacy into the data belonging to the user without their knowledge and consent. Similarly, the related issues of invasive marketing and disclosing of sensitive details need to be addressed. For instance, when AOL released unspecified search logs to be used for academic purposes, the users could be easily identified by the people performing searches. Netflix faced similar issues when the users of their unspecified datasets could be identified by their IMDB scores being correlated with respective Netflix movie scores.

3 Big Data Security Techniques

Numerous organizations now utilize big data for promotion and research, yet might not have the basics right—especially from a security point of view. Big data ruptures can be huge as well, with the potential for causing significantly more reputational harm and legitimate repercussions than at present. An increasing number of organizations are utilizing the innovation to store and dissect petabytes of information including Web logs, click stream information, and online networking substance to increase their knowledge about their clients and their businesses. In this section, we look at the security techniques proposed by federal agencies (www.splunk.com) as well as several corporate firms (www.utdallas.edu) for securing big data.

a. How government IT agencies can counter security threats by analyzing big data (www.splunk.com)

Corporations and government agencies are under almost constant attacks as cyber-gangs, as well as nation states, regularly troll to obtain valuable information. Moreover, the inherent complexity of enterprise IT infrastructure, as well as cyber-threat techniques, makes the detection of such attacks extremely overwhelming. It is of course harder for government agencies, where expertise in information security is usually thin while budgets are tight; for instance, the figures obtained from the White House depict a total growth of just about 1.5% per year in federal IT expenditures since 2009. Thus, a survey of around 300 cyber-security professionals, spread across federal as well as state/local agencies, discovered that it usually takes 16 days on average to identify the threats after intrusions into their systems and networks. Although the majority of the threats are caught quite quickly, the more sophisticated attacks could need several weeks to uncover the malicious plans. Moreover, the most fatal cyber-attacks are those undetected threats that might never be known.

Security research has provided warnings over the years regarding the growing complexity and fatality of advanced cyber-threats. The latest generation of malware contains tricky and highly evasive methods devised to exploit the uncertainties in Internet's core technology, defects in network software stacks, and constraints of security devices. Recent infringements have rested any doubts whether these attacks could be sustained for a substantial amount of time, while collecting and distributing sensitive data that often gets used to commit blackmail, if not being resold, as is usual in most cases. Attackers searching internal networks and systems usually function beneath the purview of security devices and usual security event monitoring systems (SIEMs). For instance, the attack which siphoned almost 80 million social security numbers and the associated health records off the Anthem system took around nine months to be discovered. Similarly, an attack made on a key retailer was not noticed for many days, while it sucked up customers' credit card details during the chain's busiest shopping season, mocking the almost $1.5 million spent by the firm on installing malware detection systems. Similarly, the IRS would need several months, as well as many millions of dollars, to try to recover the damages inflicted when taxpayer data were stolen to file fake returns (Fig. 1).

- Survey Results Show Government IT is Not Ready for Security Challenges

MeriTalk [16] recently conducted a survey of government security professionals, according to which more than 75% consider that their data security team is rather reactive, largely due to the sheer volume of security-relevant data that is coming into the systems. Prevention is always better than cure, thus without significant change in their strategies and processes, these public sector agencies will never be well prepared.

Organizations with traditional security systems usually learn the hard way, since the evidence of any threat, as well as the incriminating data, is not directly obtainable from the security devices. The survey found that the government IT agencies are brimming with data. Vulnerability scans from the intrusion detection systems (IDS), logs from different servers like e-mail servers, VPN, DHCP, proxy servers, show that a huge proportion of such big data often goes unexamined.

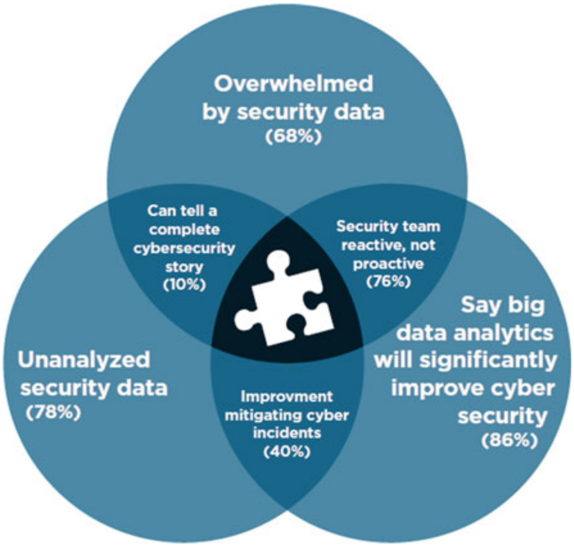
Overwhelmed by security data (68%)
Can tell a complete cybersecurity story (10%)
Security team reactive, not proactive (76%)
Say big data analytics will significantly improve cyber security (86%)
Unanalyzed security data (78%)
Improvment mitigating cyber incidents (40%)

Fig. 1 Big data analytics—The missing piece to the cyber-security puzzle. *Source* www.splunk.com

Barely 10% of the respondents claim to be getting a total security profile from the data analysis, while 78% mention that some security data always remains unchecked either owing to limited time or the lack of capable security analysts.

Government security professionals, however, see a possibility to fix these security issues. The survey results suggest that applying big data analytics to cyber-security issues would generate a big effect (as in Fig. 2). More than 60% say that it would detect breaches in real time, while almost half believe that big data analytics would also allow them to determine the probable causes of a breach. Yet, the survey suggests a cognitive discord (Fig. 3). Although most respondents considered the strategy of analyzing the security data comprehensively, only 28% of

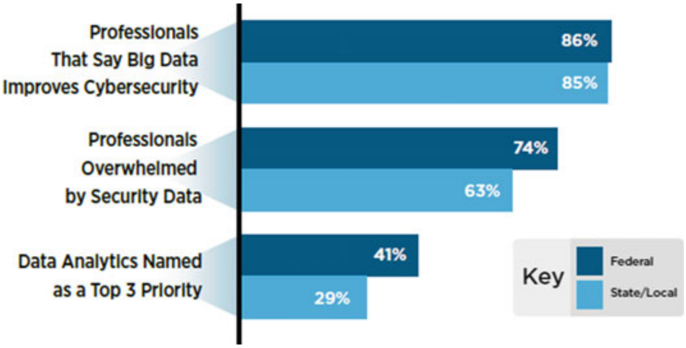

Fig. 2 Big data analytics is not seen as a priority despite the opportunity it provides to improve security. How Government IT Can Counter Security Threats by Analyzing Big Data MeriTalk Survey Analysis

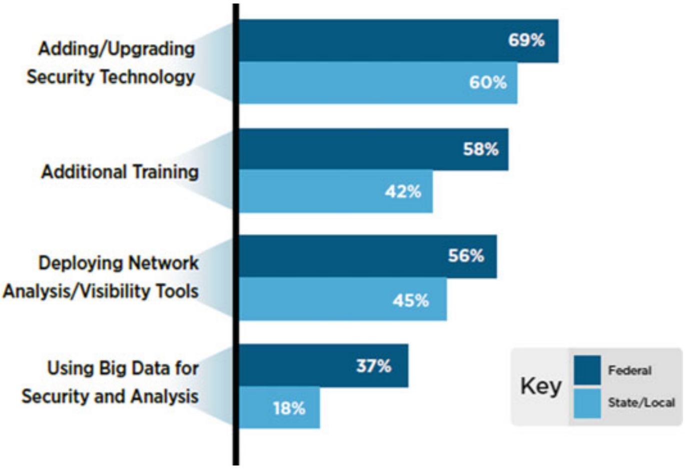

Fig. 3 Breakdown of public sector currently improving security. How Government IT Can Counter Security Threats by Analyzing Big Data MeriTalk Survey Analysis

the organizations want to actually use it to try and connect, as well as correlate, the information. This is evidence of the constricted government IT budgets, making it hard to defend new projects, though it is a myopic view.

The MeriTalk survey discovers that most of the government organizations spend a large portion of their security budget on basics, including upgrading existing systems and hardware deployment for enhanced network visibility, as well as on training the security personnel on the latest threats and associated technologies, as shown in Fig. 3. We see that less than one-third of the respondents hire new security employees. Thus, this lack of proper security measures actually reinforces the need for big data security analysis, as it is a well-proven means to achieve greater efficiency, also allowing the experts, however limited, to be much more effective.

All these issues raise doubts for the confidence people have on public sector data security. In this context, just 40% of government IT agencies can confidently say that they could stop 90% of possible cyber-incidents before the damage occurs. On the inclusion of big data analytics, results of the survey show an increase in effectiveness to almost 60%. The extent of such an improvement in cyber-detection, as well as prevention, would, however, depend on the size of the incident, as well as the information targeted, though recent research approximates the average breach to be costing almost $4 M. Thus, the government agencies should choose if they wish to resolve just one such incident or implement suitable platforms, like Splunk, to provide comprehensive operational intelligence to the organizational big data.

These platforms derive meaning from the machine data that is created by the systems, apps, and security software and applications. Such a comprehensive approach makes this platform well suited to solve problems of cyber-security data analysis for preventive, as well as forensic, uses. Such software tools would ingest, analyze, and review system data, holding terabytes of new data each day, and supporting petabytes of historical searches, like a typical network of cyber-security applications and servers would produce. The key to identifying and revealing cyber-threats is to correlate the events arising from multiple sources. Present attacks are multipronged, as well as distributed, so as to elude traditional security mechanisms and secretly mix with usual network traffic and system activity. They can only be detected after joining snippets of unusual behavior over a period of time across several systems. Typical attacks launch a start through malware which invades an unsuspecting user's PC through Trojan e-mail attachments or through a compromised Web link. After entering the network, this precursor for the multiphase attack begins to download more complex programs which can examine the internal network, launch attacks on systems displaying vulnerabilities, and return the targeted data to their selected destination. Individual events would probably be detected by the existing mechanisms, though they would not be able to trigger an alert.

- Government Action Plan

Government security professionals are now digging into the large stores of their under-utilized data and still continue to be reactive in their approach, rather than proactively trying to resolve problems before they mature into a crisis situation. The need, however, is to aggregate all existing data sources into big data, at a single place, and to further analyze and correlate the data. They could, in fact, filter and examine the data, so that it could search for ad hoc security analysis of all incidents. The MeriTalk survey displays that at a rate of almost 90%, a big data analysis system such as Splunk would significantly enhance cyber-security, thus making it easy for government IT organizations to improve their security position.

b. Strategies by corporate firms to counter threats for big data security (www.utdallas.edu), (in.pcmag.com)

Each business organization is making attempts to gather large amounts of business intelligence (BI), as much as their executives, marketing professionals, and other department personnel can collect. But, once accumulated, the striking problem is not just how to analyze and derive valuable information from such massive data, but also how to secure this big data. So, in addition to executing analytics and data visualization programs on this trove, the firm also needs to ensure that there are no spillages or leaking spots in the data store. As a solution, the Big Data Working Group of the Cloud Security Alliance (CSA) has released a handbook to avoid issues in privacy as well as security. It lists the following 10 best practices in order to secure big data that utilize a cache of storage, data encryption, monitoring, governance, and, of course, security techniques.

1. Safeguarding Distributed Programming Frameworks

The philosophy of having distributed big data stores as well as programming frameworks like Hadoop brings along a strong possibility of data being leaked. There is a probable issue termed untrusted mappers as the data being obtained from various sources might generate erroneous aggregated results.

The CSA suggests that firms should first create trust by employing methods like Kerberos authentication and also ensure conformity to existing security regulations. Then, we must un-normalize the data, by disassociating it from all the possible personal identification information (PII), thus ensuring that the privacy of the persons is not compromised. Further, we implement a predefined security policy in order to authorize access to the files, and in turn ensure that there is no leakage of the information through any of the system resources. This is done by the use of mandatory access control (MAC), for example, the sentry tool for Apache HBase. Then, once the tough part is over, what is left is to just prevent any data leaks through timely maintenance. IT department officials must keep checking the worker nodes as well as mappers into the cloud-based virtual environment, and be vigilant for fake nodes as well as modified duplicate data.

2. Securing Big Non-Relational Data

Non-relational databases, for instance NoSQL, are quite vulnerable to common attacks like NoSQL injection. The CSA provides a range of preventive countermeasures. We could begin with encryption or hashed passwords, as well as ensuring end-to-end encryption using algorithms like advanced encryption standard (AES), Rivest, Shamir, Adleman (RSA), and secure hash algorithm 2 (SHA-256). Other mechanisms, like transport layer security (TLS) and secure sockets layer (SSL), are useful encryption methods as well. Beyond those usual measures, additional layers including data tagging and object level security could be used to secure the non-relational big data, by a scheme called pluggable authentication modules (PAM). This provides a flexible means to authenticate users while ensuring the safety of log transactions by using tools like NIST log. Lastly, there are also *fuzzing methods* that uncover vulnerabilities injected, through cross-site scripting, between NoSQL and HTTP protocols when using automated means for data input at the data node, protocol, and application level of this distribution.

3. Secure Big Data Store and Transaction Logs

Storage management is a vital portion of the big data security concern. The CSA suggests the use of signed message digests (MDs) to assign digital identifiers to every digital document or file, as well as to use a method named secure untrusted data repository (SUNDR) to identify unauthorized modifications to files through malicious server agents. There are also several other techniques, such as key rotation, lazy revocation, broadcast and policy-based data encryption, as well as digital rights management (DRM). Yet, it is of course best to try and securely build the personal cloud storage atop the existing infrastructure.

4. End-Point Filters and Validation

End-point security is of utmost importance to the business. The organization could begin with utilizing trusted certificates, performing resource testing, as well as connecting only trusted devices to the network by using a mobile device management (MDM) solution (in addition to the antivirus and other malware protection software). Further, the firm could use similarity, as well as outlier detection techniques, to filter the inputs into proper and malicious, thus safeguarding the system against ID-spoofing, or Sybil attacks, wherein one entity masquerades as having multiple identities.

5. Real-Time Compliance and Security Monitoring

Compliance is always an issue for firms, and more so when actually dealing with a continuous overflow of data. This can be best tackled with real-time analytics as well as security at each level of the storage stack. The CSA suggests that firms apply big data analytics through the use of tools like Kerberos, secure shell (SSH), and Internet protocol security (IPS) to be able to control big data in real time. Once that is done, further steps include mining event logs, implementing prime security systems for the routers as well as application-level firewalls, and also starting to implement comprehensive security control, at not just the cloud, but even at cluster and application levels. The CSA also warns enterprises to be aware of evasion attacks that try to bypass the big data security infrastructure, commonly known as "data poisoning" attacks.

6. Preserving Data Privacy

Sustaining data privacy in ever-increasing datasets is truly hard. According to the CSA, the key is to exhibit *scalability* and *composability*, by the implementation of methods like ***differential privacy***—to maximize the accuracy of query results with minimal identification of records—and ***homomorphic encryption*** for storing and processing the encrypted information into the cloud. Further, the CSA recommends the incorporation of employee awareness programs that focus on imminent privacy regulations, and to maintain the software infrastructure by employing proper authorization methods. Finally, best practices encourage the implementation of an approach called "privacy-preserving data composition" that controls leakage of data from multiple databases by monitoring and reviewing that infrastructure which links the databases.

7. Cryptography for Big Data

Cryptography has become more advanced. Using a simple mechanism like the searchable symmetric encryption (SSE) protocol, corporate firms can in fact execute Boolean queries on encrypted big data, by just creating a system to be able to search and then filter the encrypted data. The CSA further recommends several cryptographic techniques. For example, relational encryption allows encrypted data to be compared without sharing encryption keys. This is done by matching the identifiers and their attribute values. Another mechanism called identity-based encryption (IBE) makes it easier for key management in public key systems. This is done by letting plaintext be encrypted for any given identity values. Attribute-based encryption (ABE) could be used to integrate all access controls into a comprehensive encryption scheme. Moreover, there is the converged encryption method where encryption keys could assist cloud providers in identifying duplicate data.

8. Granular Access Control

Access control comprises two core steps: restricting and granting user access. A policy to choose the right level of control in any particular situation needs to be built and implemented. The setup of such granular access controls requires the following:

- Normalize elements that are mutable, and de-normalize others.
- Administer secrecy needs and guarantee proper execution.
- Monitor access labels.
- Observe admin data.
- Use single sign-on (SSO), as well as
- Use a proper labeling scheme to uphold appropriate data federation.

9. Auditing

Auditing, especially at a granular level, is a must for big data security, especially after the system has been attacked. It is recommended that corporate firms create a consistent audit analysis following any attack, as well as to be sure to provide a complete audit trail. However, this should not hamper performance, as it is necessary to guarantee quick access to data to reduce response time. The confidentiality and integrity of audit information are also equally essential. Information related to the audit shall be separately stored, with granular protection in the form of user access control as well as regular monitoring. It is also possible to use open source audit layering or even a query coordinator like ElasticSearch that could make the entire process simpler.

10. Data Provenance

Contextual to the requirements, data provenance could have several meanings. Here provenance refers to the origin of the metadata being generated by usual big data applications. Such data truly needs significant protection. This would need to first develop an infrastructure authentication protocol, which would control the access, and also set up regular status updates as well as data integrity verification by using methods like ***checksums***. Additionally, to secure the big data at the source itself, we need to implement scalable, dynamic, granular access control while implementing encryption methods. Just one approach may not be suitable to provide security to the big data across all levels of the organization and so we need to secure every level of the big data infrastructure as well as the application stack.

4 Threats and Attacks Against Big Data

Considering the security and privacy aspects, we need to visualize big data security from several viewpoints. We need to think not just of how to protect the data primarily, but also the processing of big data, and of course the output of this big data processing. Kim et al. [13] provide three major concerns where security is needed in big data: access control, data security, and information security.

Xu et al. [30] have presented a big data security model that considers the user role of security in each of these phases of the big data process. Actually, to secure the big data environment, it is vital to identify the possible threats, and attacks, that big data could experience during its lifecycle. Such an identification of the threats, as well as attacks, would help the security groups to strengthen the defense against such threats. Alshboul et al. [1] present a big data lifecycle model consisting four phases: data collection, data storage, data processing, and knowledge creation. Further, they integrated their model with possible threats and attacks to provide a security threat model, which could further be used to secure the big data infrastructure. They have identified four types of roles the users portray in the big data environment, including data provider, data collector, data miner, and decision maker [30].

Alternatively, various threats as well as attacks threaten the big data technology. Dev et al. [6] have explained a data mining-based threat that utilizes data mining methods to extract important information as well as sensitive data. Big data is also threatened by the aspect of data privacy. This is because the declarations of such sensitive data may tarnish the reputation of individuals or organizations. These include threats of re-identification and wrong results [10]. Wu and Guo [29] have also maintained that information assurance and privacy are major concerns for big data environments where the extraction of personal or sensitive data could harm individuals as well as corporate organizations that could lead to numerous business problems.

Big data technology, as we saw earlier, faces several security threats and attacks that usually are drawn from the usual features of big data technology, in turn relying on data analytics methods and data mining algorithms. Moreover, attackers could also make use of data mining procedures to locate sensitive data and further release it to the public, causing a data breach. Threats as well as attacks to big data can thus be classified in terms of the four phases of the big data lifecycle, which have been depicted in Table 1.

Each of these phases has peculiar characteristics and is allocated different tasks, making each of the phases susceptible to particular threats and attacks. For example, the data collection phase experiences attacks such as phishing and spoofing, which target people engaged in the process of data gathering and distribution. Better awareness and proper compliance to security procedures and policies are one of the ways to enhance security for this phase. Data mining-based attacks, commonly performed by hackers, can be characterized by the use of these extraction techniques to retrieve sensitive data, which is then used illicitly. Such attacks could be restricted by the fragmentation of the datasets either horizontally or vertically, as well as by adopting means for non-centralized storage of data. The risk of physical threats like theft or unauthorized access, however, needs to be dealt with in the usual ways.

The data analytics phase poses a risk to attacks that may result in either the release of sensitive data or in harming the data process. Data mining-based attacks may occur to discover and provide vital information, or associative techniques

Table 1 Threats model (Alshboul, Wang, Nepali)

Phases	Threats and attacks	Description	Suggested defense
Data collection	Phishing	These attacks are hacking data provider and collector to get an access to the data in the collection phase	Security awareness program
	Spamming		
	Spoofing		
Data storage	Data mining-based attacks	Targeted datasets to extract knowledge [6]	Divide datasets (vertically and horizontally) and noncentral data storage framework
	Attacks on data storage devices	Stealing hard disks or make images of them	Physical security measures noncentral data storage framework
	Unauthorized data access	People access data illegally	Access control
Data analytics	Data mining-based attacks	Using data mining methods to extract sensitive knowledge	Divide datasets (vertically and horizontally) and use access control
	Re-identification threat	Identification threats of personal information [10]	Core attribute encryption
	Wrong result threat	Using incorrect analysis process, which leads to incorrect results [10]	Follow correct analysis procedures and document, audit, and review the process
Knowledge creation	Privacy threats	Releasing the resulted knowledge (e.g., rival competitors)	Adopt encrypt the resulted knowledge and adopting access control strategy
	Phishing and spoofing	Decision makers are targeted	Security awareness programs

could help to locate personal attributes which would actually have an impact on data privacy. To protect the big data structure from such attacks, defensive methods such as fragmenting the dataset, either horizontally or even vertically to encrypt the high-weighted core attributes, may be implemented. Another threat possible in this phase is that of obtaining inappropriate results from the data analysis process [13]. Thus, it is important to follow the proper analytics process as well as to document it.

Lastly, in the knowledge creation phase, the knowledge created from the big data process usually comprises of sensitive information that should not be distributed publicly, especially to business rivals. Some threats to privacy, as well as security attacks, may intend to affect those who are related to the final result of this big data process. As a result, effective security policies and their proper implementation are required. Once everyone realizes the extent of damage possible due to security lapses in big data, they will hopefully begin to adhere to established standard procedures for access control. The development of efficient security awareness programs that could not just mitigate, but even prevent the occurrence of any threat, is needed.

5 Case Study of CIA—USA Documents Disclosed on WikiLeaks

Take for example, the WikiLeaks case [25]. WikiLeaks, in what is considered to be one of the greatest leaks, on March 7, 2017, released thousands of pages describing complex software tools as well as techniques used by the CIA to break into computers, smartphones, and even Internet-connected televisions. The documents provide a detailed catalog of highly specialized tools. They also include the instructions for hacking a wide range of common computer-oriented tools used for spying such as Skype, Wi-Fi networks, PDF documents, and even commonly used antivirus programs.

The Wrecking Crew program explains the process to crash a specific computer, while another program is used to steal passwords, by using the auto-complete function on Internet Explorer. Other similar programs called AngerQuake, CrunchyLimeSkies, McNugget, and ElderPiggy have been used. This document dump is the latest attack on the antisecrecy organization and of course a great blow to the CIA that uses its hacking capabilities to conduct espionage against foreign targets.

WikiLeaks remarked that this was just the first release of a larger set of confidential CIA material and includes more than 7800 Web pages with about 940 attachments, most of them partly redacted by WikiLeaks editors to avoid revealing the actual code for cyber-weapons. The entire archive of CIA material, however, consists of several hundred million lines of computer code, the group claimed. In one disclosure that may specifically trouble the tech world if confirmed, WikiLeaks said that the CIA and its allied intelligence services have managed to compromise both Apple and Android smartphones, allowing their officers to bypass the encryption on popular services such as Signal, Telegram, and even WhatsApp. WikiLeaks also confirmed that government agents can also collect audio as well as message traffic even before encryption is done.

The National Security Agency (NSA) documents given by Edward J. Snowden in 2013 to journalists do not include examples of how those tools had actually been used against foreign targets. Even though the liability for such a leak toward compromising national security was limited, the breach was highly embarrassing, specifically for an agency that depends on secrecy. Robert M. Chesney, a specialist in national security law at the University of Texas at Austin, likened the CIA episode, to a group calling themselves as the Shadow Brokers disclosing the set of hacking tools being used by the National Security Agency, last year. There was no public confirmation of the authenticity of the documents, which were produced by the CIA's Center for Cyber Intelligence and are mostly dated from 2013 to 2016. The authenticity of the documents has been confirmed by one government official, while another former intelligence officer said that few of the code names for the CIA programs, one of the organization charts as well as the description of a CIA hacking base, appeared to be genuine.

The CIA was surprised by this kind of a document dump. In some regard, the CIA documents confirmed and provided details on a few capabilities that have been suspected in technical circles for a long time. It is being contemplated by people who know a lot about security and hacking that the CIA was investing in enhancing these capabilities as, otherwise, other parties like China, Iran, Russia, and several private agents would actively examine and exercise the possibilities.

This episode would surely raise concerns in the USA and other nations regarding the trustworthiness of technology where cyber-security can impact human life and public safety. There is no evidence to suggest whether the CIA hacking tools have been in fact used against Americans themselves, although documents suggest the government has knowingly allowed such vulnerabilities in phones and other electronic devices to make spying easier, though these could be effectively used by hackers too. Therefore, patching these security holes immediately would be the best way to make everyone's digital life safer.

In the business world, the so-called Panama Papers and several other large-volume leaks have laid bare the details of secret offshore companies used by wealthy and corrupt people to hide their assets. Both government and corporate leaks have been made possible by the ease of downloading, storing, and transferring millions of documents in seconds or minutes, a sea change from the use of slow photocopying for some earlier leaks, including the Pentagon Papers in 1971.

6 Big Data Security Analytics Tools

Big data security analytics subsequently will be as common as malware detection and vulnerability scans. This is owing to the fact that these big data platforms permit enterprises to collect data from several, diverse data sources, collaborate that data in almost real time, explore patterns, and identify malicious activities, and also observe, report, and perform forensic investigations. The tools used for analyzing big data security are in turn capable of scrutinizing and processing huge volumes of disparate data types. Although most organizations may not require all of the features of big data security analysis tools, these are extremely efficient in providing a comprehensive solution for big data security. The leading tools providing security analytics for big data include solutions from IBM, RSA, Splunk, LogRhythm, Fortscale, Hexis, and Cybereason. Each of these solutions can be evaluated against five essential parameters that would determine the extent of its utility:

i. Unified data management

Any product for big data security analytics is based on the principle of unified data management. This is because the data management platform typically stores and also queries the data all through the enterprise. Moreover, it must also try and adjust the data management features with its cost and scalability features. The most commonly used analytics platforms usually prefer Hadoop, as it is a widely used platform for managing big data and its related ecosystems. For instance, Fortscale, uses a Cloudera distribution of Hadoop, which permits Fortscale to linearly scale up, as there are new nodes being added to the cluster.

A distributed data management system is used by IBM's QRadar. This enables the data storage to be scaled horizontally. Although it is rare, since most applications would need access to only local data, there are possibilities that the data store may be distributed, for instance, in case of forensic analysis. As a result, their corresponding security information management systems, or SIEM, would also need to be distributed. Such a big data SIEM makes use of data nodes to scale up to petabytes of data, rather than using storage area networks (SANs), resulting in substantial economies in costs as well as complexity. Similarly, RSA Security Analytics too uses an architecture that is able to scale linearly, as it is both federated as well as distributed. Events need to be prioritized so as to enhance the efficiency of the tool. Hawkeye Analytics Platform (Hawkeye AP) is built on a data warehouse platform for security event data. It is usually characterized with providing low-level and scalable data management (i.e., it is able store large amounts of data in multiple files across servers). Further, it is also crucial as it comprises of tools to query the data in a structured way. The Hawkeye AP, presented by Hexis Cyber Solutions, interestingly stores data in columnar form, as compared to traditional row-based methods. This, in turn, ensures not just optimized performance, but also tamperproof storage of big data.

ii. Support for multiple data types, including log, vulnerability, and flow

Variety, as we have seen, is one of the basic attributes of big data. This in turn poses considerable challenge to the integration of data, and that too securely. The security analytics tool by RSA employs a modular architecture to collaborate various data types, also allowing data from other sources to be added subsequently. Multiple kinds of data often dictate the need for various security tools. For instance, IBM's QRadar has a component called vulnerability manager that is able to combine data from several vulnerability scanners and enhance the data with situation-relevant details [11]. Incident Forensics is a special module to analyze security-based incidents through flow data and packet captures across the network. This detection tool comprises of a search engine which can effectively process even terabytes of network data. Another popular example of a platform with wide support for almost all types of data (machine data, security activities, system and audit logs, flow data, and application logs) is the Security Intelligence Platform by LogRhythm. This platform generates data regarding file veracity, process validity, network communication efficiency as well as user activity, by analyzing the big data store. The Enterprise Security package by Splunk lets analysts not just search for data but also depicts visual correlations among them to discover and collect data about malicious events.

iii. Scalable data ingestion

Big data security analytics can analyze large volumes of data comprising of a broad range of data types. For this, security products for big data analytics must consume data from servers, terminals, nodes, networks, and any other infrastructural components that vibrantly change states. The major risk of such data ingestion is, however, that the analytics processing cannot match pace with the rate of incoming data. Splunk is quite well recognized for its wide capabilities for data ingestion. This platform also allows for custom connections, in addition to the usual ones. Data here is maintained in a schema-less mode and is indexed at the time of ingestion itself to allow multiple data types and also responding rapidly to queries. IBM QRadar scales up to geographically distributed systems from simpler, single-appliance deployments. It is well designed to fulfill the demands of big enterprises. It can process millions of events in a second for real-world applications. Augmentation is another important point to be considered here. It is the process of supplementing the recorded event with additional information, specifying the context in which it has been collected. RSA security analytics, for instance, qualifies the network data, as it is being analyzed, by adding details of these network sessions, or maybe threat indicators, or any other details, which could help analysts to comprehend the wider picture enclosing the low-level security data. Moreover, the data collection is also a key concern. The time taken by the system to collect the data effectively sets a lower bound on the rate of detecting any security events. The positioning of data collection locations also determines the extent and the type of data being collected. Another platform, named Cybereason, uses sensors that execute in the user space of terminal operating systems, thereby allowing data to be collected without causing any adverse effect on either user experience or the kernel-level functions. These sensors can keep collecting data even if the devices are not linked with the enterprise network.

iv. Information security-specific analytic tools

Big data security analytic tools are required to scale up to be able to meet the amount of data produced by the enterprise. Likewise, the analysts must be able to execute queries on the event data at appropriate levels of abstraction keeping the view of an information security position. Fortscale uses machine learning as well as statistical analysis, commonly known as data science methods, to be able to adapt to the changes occurring in the security environment. As a result, the analysis can be based on the real data than just on predefined rules. Machine learning algorithms can identify changes in the network behavior, doing away with the need of any human intervention to modify the predefined set of rules.

Security analytics largely depends on the intelligence regarding malicious activities. The RSA live service dictates data processing, as well as correlation rules to the deployments of RSA security analytics. New rules could thus be used to analyze the new data that is arriving in real time, as well as the historical data that is stored on the RSA security analytics system. Data science methods are often used to improve the quality of analysis. Analytics workflow by LogRhythm includes the processing, machine analysis, and forensic analytics stages. Processing transforms the data in many ways to raise the chance for detection of useful patterns from raw data. The processing comprises of data classification, time normalization, risk contextualization, and metadata tagging.

v. Compliance reporting, alerting and monitoring

Compliance reporting is an essential component for almost all enterprises today. The reporting rules that are to be included along with big data security platforms would thus need to satisfy the specific compliance requirements of the organization.

The risk manager is an add-on with IBM Security QRadar and provides tools to manage configurations of the network devices for the purpose of risk management and compliance. It has the following features including automated monitoring, compliance policies assessment, multi-vendor product audits, and threat modeling.

Fortscale, as mentioned earlier, utilizes machine learning to regularly assess changes to the baseline activities in order to detect anomalous events. The system can raise alerts and also provide information regarding the context of these events. RSA security analytics provides almost 90 templates to fulfill reporting needs of regulations including SOX, HIPAA, and PCI DSS with minimum efforts needed from the end users. The Cybereason platform has been specifically designed to identify any malicious activity. The platform comprises of an investigation console that coordinates the information and also visualizes the attack timelines and the users and devices affected. Continuous monitoring via dashboards is the strategy followed by Splunk Enterprise Security. The metrics include key security, as well as, performance indicators, along with trend pointers. Prioritized workflows are another strong point for this platform, which also supports the tracking of highly privileged users while reporting on any unauthorized attempts to access any critical applications. The Hawkeye AP package is stocked with 400 reports that can be customized to particular requirements. It provides the option to create custom reports, as it uses the relational technology and also supports the ANSI standard for SQL, and also the JDBC and ODBC drivers. Lastly, the LogRhythm platform consists of prioritized alarms, standardized reports, and a real-time reporting dashboard. It also comprises additional forensic analysis tools such as case management tools, evidence lockers, and incident tracking metrics.

7 Open Source Security Tools for Big Data

As the world is progressing toward IoT, the need for Web as well as infrastructure security has become a prime concern. We need to make our devices, systems, and the networks secure. Industry innovators such as Google, Facebook, and Netflix have looked into this concern and engage in the development of security tools with the open source fraternity. The rapid transition in the network infrastructure, from being closed to now scaling entire enterprises, would in turn increase the possibility of threats or attacks by virus, rootkits, malware, spyware, adware, and so on. Such security threats could in fact result in numerous disruptions, from denial-of-service (DoS) attacks to DNS poisoning, identity theft, etc., across the Internet. The frequency of security breaches occurring on the Web prompts organizations, and even professionals, to take appropriate precaution against such attacks. Several open source tools [5] are available in order to counter such threats and to protect our big data stores from risk. Following are ten open source tools widely used as big data security solutions in the industry.

1. OSQuery

OSQuery has been developed by Facebook. It is a simple tool for Linux and Mac OS X infrastructure. The important features of this tool include hardware changes, monitoring files and network traffic and process creation. This tool allows for easy access to data and also logs system information according to the queries posed. It further allows users to code automation scripts, to apply executive big data security intelligence, as well as to discover novel ways for the enterprise to upgrade servers.

2. Security Onion

Security Onion is a Linux-based solution for intrusion detection systems (IDS), log management, and network security monitoring (NSM). It is basically an intrusion detection system and is extremely simple to set up for an organization. Security Onion comprises of three major functions: full packet capture, network-based IDS (NIDS), and host-based IDS (HIDS) for detecting intrusion and several powerful tools to provide security and analytics for big data [23].

3. Skyline

Skyline provides features to uncover anomalies in the big data infrastructure. Operating in real time, it is built to facilitate offline monitoring of several thousands of metrics. It is generally used to monitor systems where huge quantities of data arise from high-resolution time series. Once any anomalous metric is detected, it floats up the entire time series to the Web app, so that the anomaly can be noticed and rectified.

4. Google Rapid Response

Google Rapid Response (GRR) is an approach where Google aims to examine incident responses instantaneously, even though from a remote site [7]. GRR typically comprises of an agent, or client, to be deployed to the target system, as well as a server infrastructure, which could communicate with and also manage the agent. It is available for most popular OS including Linux, Windows, and Mac OS X. Its major features include live, remote, memory analysis by the use of open source memory packages for Linux, Windows, and Mac OS X, and also a memory analysis framework, called Rekall.

5. OSSEC

OSSEC is an open source host-based IDS (HIDS) with excellent features such as file integrity checking, log analysis, policy monitoring, real-time alerting, rootkit detection, and dynamic responses. It executes on most popular operating systems, such as Linux, Windows, Mac OS, AIX, HP-UX, and Solaris [17].

6. Scumblr and Sketchy

Scumblr is a Web-based application, which allows its users to conduct regular searches and take appropriate actions based on the results generated. It also performs searches by using plug-ins, or APIs, called search providers. Every search provider recognizes the method to perform searches through a particular site or API, such as Google, Twitter, Bing. Moreover, searches could also be pre-configured within Scumblr itself, on the basis of the options offered by the search provider.

7. RAPPOR

RAPPOR, developed by Google, is another interesting privacy mechanism that allows the analysis of big data, such as wide demographic statistics, to make inferences about their populations, while preserving privacy of the individual users.

8. OpenVAS

OpenVAS is a powerful and comprehensive vulnerability management solution. It provides a robust framework, comprising several tools and services to incorporate powerful vulnerability scanning, and in turn, vulnerability management [19].

9. OpenSSH

OpenSSH, as the name suggests, is a free version of SSH (secure shell) association tools, which the technical users of the Internet can rely on. The users of telnet, ftp, or rlogin may not be aware that their passwords are transmitted in an unencrypted form, directly, across the Internet. OpenSSH therefore encrypts all the traffic (even passwords) so as to effectively eliminate attacks such as connection hijacking, eavesdropping. Moreover, OpenSSH also provides secure tunneling capabilities and offers various authentication methods, in addition to supporting all versions of the SSH protocol [18].

10. MIDAS

MIDAS is a framework to develop a Mac intrusion detection analysis system, which is based on the collaborative work and discussions by the Facebook and Etsy security teams. The repository offers a modular framework as well as a number of helper utilities, apart from an example module that can be used to detect alterations to usual OS X persistence mechanisms.

8 Summary

This chapter discussed the impact of big data and the need for securing it. We began the analysis by outlining the vulnerabilities faced by the existing big data architectures and discussed several competent security techniques for big data in the context of government and corporate sectors. We discussed in detail the different factors that make the current big data architecture vulnerable. Further, we discussed the possible techniques by which the government agencies, as well as corporate, could counter this issue. We presented results of a survey by MeriTalk that concludes that big data analytics is not a priority even though it would considerably improve data security. We also talked about the Anthem system that was attacked to steal vital information from medical records. Additionally, we described several best practices for big data security, which suggest the need for granular access control, as well as advanced encryption of big data. Compliance and security monitoring need to be real time, especially at end points. It is suggested that regular auditing is the best precautionary method though. Keeping the security and privacy aspects in view, the threats and attacks against big data have been discussed at length. We have analyzed the big data lifecycle threat model to elaborate on the types risks each phase is susceptible to. We presented a case study that showcased the possibility and impact of illegal access to big data stores of the CIA. The WikiLeaks case clearly demonstrated the vulnerability of big data, implying the need for big data security. We then described several big data analytics tools and explained in detail their desired characteristics. We also analyzed several open source security tools for big data that would provide an affordable and efficient solution to secure the big data stores.

References

1. Alshboul, Y., Wang, Y., & Nepali, R. K. (2015). Big data lifecycle: Threats and security model. In 2015, Twenty-first Americas Conference on Information Systems, Puerto Rico 2015.
2. Bisk. (2017). "What is big data?" Business intelligence by Villanova University. Retrieved on May 22, 2017.
3. Boyd, D., & Crawford, K. (2011). Six provocations for big data. In *Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. https://doi.org/10.2139/ssrn.1926431.2011.
4. Community cleverness required. (2008). *Nature, 455*(7209), 1. https://doi.org/10.1038/455001a. PMID 18769385.2008.
5. Dan, S. (2016) Comparing the top big data security analytics tools. At http://searchsecurity.techtarget.com/feature/Comparing-the-top-big-data-security-analytics-tools. Accessed on May 16, 2017.
6. Dev, H., Sen, T., Basak, M., & Ali, M. E. (2012). An approach to protect the privacy of cloud data from data mining based attacks. In *Proceeding of High Performance Computing, Networking Storage and Analysis, IEEE*, November, (pp. 1106–1115). https://doi.org/10.1109/SC.Companion.2012.133.
7. GRR Rapid Response at https://github.com/google/grr. Accessed on May 22, 2017.
8. Grimes, S. (2017). Big data: Avoid 'Wanna V' Confusion. InformationWeek. Retrieved May 25, 2017.
9. Hilbert, M., López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information". *Science, 332*(6025), 60–65. https://doi.org/10.1126/science.1200970. PMID 21310967.
10. Jensen, M. (2013). Challenges of privacy protection in big data analytics. In *Proceeding of the International Congress on Big Data, IEEE*, June, (pp. 235–238). https://doi.org/10.1109/BigData.Congress.2013.39.
11. Jitendra, C. (2014). Top 5 big data vulnerability classes. At http://www.cisoplatform.com/profiles/blogs/top-5-big-data-vulnerability-classes-1. Accessed on 12, 2017.
12. Kantarcioglu, M. (2017). Securing 'big' data. At http://www.utdallas.edu/~muratk/research-summary.pdf. Accessed on May 18, 2017.
13. Kim, S.-H., Eom, J.-H., & Chung, T.-M. (2013). Big data security hardening methodology using attributes relationship. In 2013 International Conference on Information Science and Applications (ICISA), IEEE, June, (pp. 1–2). https://doi.org/10.1109/ICISA.2013.6579427.
14. Mac Intrusion Detection Analysis System (MIDAS). Available at https://github.com/etsy/MIDAS. Accessed on May 26.
15. Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London: John Murray.
16. MeriTalk. (2015). Survey analysis, How Government IT Can Counter Security Threats By Analyzing Big Data. At https://www.splunk.com/content/dam/splunk2/pdfs/white-papers/how-government-it-can-counter-security-threats-by-analyzing-big-data.pdf.
17. OSSEC—Open Source HIDS Security at http://www.ossec.net/ Accessed on May 22, 2017.
18. OpenSSH. (2017). At http://www.openssh.com/. Accessed on May 24, 2017.
19. OpenVAS. (2017). At http://openvas.org/. Accessed on May 25, 2017.
20. Oracle and FSN. (2017). Mastering big data: CFO strategies to transform insight into opportunity. December 2012.
21. Reichman, O. J., Jones, M. B., Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*. *331*(6018), 703–705. https://doi.org/10.1126/science.1197962. PMID 21311007.2011.
22. Rob, M. (2017). 10 best practices for securing big data. At http://in.pcmag.com/feature/107583/10-best-practices-for-securing-big-data. Accessed on May 12, 2017.
23. Security at http://blog.securityonion.net/p/securityonion.html. Accessed on May 27, 2017.

24. Segaran, T., Hammerbacher, J. (2009). *Beautiful data: The stories behind elegant data solutions* (p. 257). O'Reilly Media. ISBN 978-0-596-15711-1.2009.
25. Shane, S., Rosenberg, M., & Lehren, A. W. (2017). WikiLeaks releases trove of alleged C.I.A. hacking documents. *New York Times* March 7, 2017.
26. Skyline anomaly detection system at https://github.com/etsy/skyline. Accessed on May 28, 2017.
27. The Economist Newspaper. (2010, February 25). *Data, data everywhere*. Accessed on May 28, 2017.
28. What is big data?—Bringing big data to the enterprise. www.ibm.com. Retrieved May 20, 2017.
29. Wu, C., & Guo, Y. (2013). Enhanced user data privacy with pay-by-data model. In *Proceeding of the International Conference of Big Data*, IEEE, Ieee, October, pp. 53–57. https://doi.org/10.1109/BigData.2013.6691688.
30. Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *The Journal for Rapid Open Access Publishing, 2*, 1149–1176. https://doi.org/10.1109/ACCESS.2014.2362522.