UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

June 20, 2017

Data privacy and security in Business Intelligence and Analytics

Victoria Beleuta
Universitat Politècnica de Catalunya
Calle Jordi Girona 1-3
08034 Barcelona

Supervisor: Prof. Jaime M. Delgado Merce

Contents

Abstract

Widespread web application adoption created a large, complex and varied amount of data known as Big Data. These data sets have a great value for many economic and scientific sectors, however they come with additional difficulties when it comes to storing and analyzing them. Big Data Analytics is the term that describes the process of researching this massive amount of information in order to find hidden patterns and correlations. Business Intelligence departments can now support decision-making processes based on this broad range of data points collected throughout the lifetime of an application and the designated user's interaction with it. However, the abundance and extensive use of Big Data comes with a number of security and privacy risks that must be addressed. This work identifies and analyzes these concerns as well as their requirements. Focusing on user privacy, some of the major issues include: over collection of data in mobile applications, misuse of data, and multi-source data analysis. These issues can not always be solved using existing privacy preserving methods. The variety and velocity of Big Data makes it difficult to distinguish between sensitive and non-sensitive information, so traditional anonymization techniques can not always be used. Furthermore, analyzing multi-source datasets can lead to risks of user re-identification. In this paper we investigate proposed solutions for securing Big Data as well as ways to maintain data privacy. We look into two major use cases: healthcare and web analytics, where Big Data is becoming more and more important. We sum up with a comparison of the requirements and solutions used to preserve user data privacy for the statistical and clinical data collected in today's applications.

1 Introduction

The amount of data on the Internet has exploded in recent years. Social network sites, companies that capture trillions of bytes of information about customers and operations, as well as the millions of networked sensors that are being embedded in physical world devices, such as mobile phones and automobiles, will continue to fuel this exponential data growth. Large pools of data that can be captured, communicated, aggregated, stored, and analyzed are now part of every sector and function of the global economy. User data is collected with the help of cookies, events, screen capturing techniques and direct user provision and is stored within the on-premise and cloud-based data warehouses for further processing.

With the rapid increase in the amount of user data that is gathered and exchanged electronically, as well as the ambiguously defined scope for its use, the problem of information privacy is rapidly gaining attention. Dozens of public and private organizations hold parts of our personal information which are subject to a variety of more or less explicit privacy agreements and government laws. The importance and impact of data-related problems that need to be solved in contemporary organizations has lead to the emergence of Business Intelligence and Analytics (BI&A) as an important area of study for both practitioners and researchers. This interest is further fueled from the realization that added value for website visitors is not gained merely through larger quantities of data on a site, but through easier access to the required information at the right time and in the most suitable form. This issue is becoming increasingly important on the Internet, as non-expert users are overwhelmed by the quantity of information available online, and commercial websites strive to add value to their services in order to create loyal relationships with their visitors/customers. Moreover, the emergence of e-services, such as e-commerce, e-learning and e-banking, has changed dramatically the manner and scope in which the Internet is being used, turning websites into businesses and increasing the competition between them.

Business intelligence applications are gaining popularity, due to the desire of officials of public and private companies to monitor, analyze, understand, and eventually improve business processes. BI applications typically extract data from multiple data sources, sanitise them to ensure data quality and consistency, trans-

form them, and then generate various kinds of reports used by managers and officials to analyze the performed processes. For academics, big data and BI is shifting the IT research environment to become more integrated with other fields. Researchers can now triangulate data using an integrated set of market observations, empirical data, and focused survey data with the ability to cross validate and challenge the researcher to develop new methods and techniques to effectively leverage big data.

BI environments typically contain complete and accurate information on all the details of a given organization, and these assets represent a significant security risk, for both the organization as well as their customers. They are also very dynamic, with a broad and frequently changing audience both internal and external to the organization. This makes the task of controlling user access to a subset of data extremely challenging, particularly in higher-order BI applications such as corporate performance management where the true value of BI is derived from a broad data view. In addition, the physical infrastructure is often constructed out of many different types of tools and data is constantly in motion.

In each sector that uses big data, there are different types of information being collected and it is used for various purposes. Because of this we evaluate, in later sections, the privacy and security risks in each category separately as they will differ vastly. Taking into account both the existing body of published research papers, originating from various conferences, as well as the white-papers and case-studies, we intend to classify some of the widespread security and privacy concerns, existent or proposed solutions, as well as areas where future research efforts are required.

In the following section (2) we introduce the field of Business Intelligence and Analytics. In Section 3 we identify, classify and analyse possible security and privacy risks, followed by possible solutions and good practices in Section 4. We explore in more detail, two of the more pressing use cases: healthcare and web analytics. A summary of all the findings as well as some of the proposed solutions is presented in the last section (5).

2 Business Intelligence and Analytics

The term ”Big Data” is used to characterize data sets that are large, varied and rapidly-changing. It requires database management systems with capabilities beyond those seen in standard SQL-based systems. Formally, it is defined by the 4Vs: volume, velocity, variety, and veracity. Volume refers to the large amount of data that is being generated everyday, velocity is the rate of growth and how fast the data is gathered for analysis, and variety provides information about the types of data such as structured, unstructured and semi-structured. The fourth ”V” refers to veracity, which includes availability and accountability. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and distributed techniques. In today's Internet-oriented economies, big data arises from five major sources: large-scale enterprise systems, online social graphs, mobile devices, Internet of Things (IoT) and open or public data[12].

Business analytics refers to the skills, technologies, applications and practices used to explore and investigate past business performance in order to provide actionable insights. The opportunities associated with data and analysis in different organizations have helped generate significant interest in BI&A. In addition to the underlying data processing and analytical technologies, BI&A includes business-centric practices and methodologies that can be applied to various high-impact applications such as e-commerce, market intelligence, healthcare, security, etc.

2.1 Science and Technology

Many areas of science and technology are benefiting from high-throughput sensors and instruments, from astrophysics and oceanography, to genomics and environmental research. The National Science Foundation (NSF) mandated that every project is required to provide a data management plan, to facilitate information sharing and data analytics. Cyber-infrastructure, in particular, has become critical for supporting such data-sharing initiatives. Data used in such applications is fine-grained, high-throughput and instrument collected and has a huge scientific impact.

2.2 E-government and Politics

As government and political processes become more transparent, participatory, online, and multimedia-rich, there is a great opportunity for adopting BI&A research in e-government and politics applications. Selected opinion mining, social network analysis, and social media analytics techniques can be used to support online political participation, e-democracy, political blogs and forums analysis, and process transparency and accountability[11]. Semantic information directory and ontological development can also be developed to better serve their target citizens. Data comes from fragmented information sources and legacy systems, rich textual content and unstructured informal citizen conversations.

2.3 Urban Planning

Mobile devices are always turned on, location-aware, and have multiple sensors including cameras, microphones, movement sensors, GPS, and Wi-Fi capabilities. This has revolutionized the collection of data in the public space and its use. In [31], scientists are working on a collaborative project to analyze mobile phone communications to better understand the needs of the one billion people who live in settlements or slums in developing countries. They explore food shortage prediction models using variables such as market prices, drought, migrations, previous regional production, and seasonal variations.

Big data use within the "smart grid" context also illustrates the benefits of sophisticated data analysis. The smart grid allows electricity service providers, users, and other third parties to monitor and control electricity use. Utilities view the smart grid as a way to precisely locate power outages or other problems, including cyber-attacks or natural disasters, so that technicians can be dispatched to mitigate problems. Consumers benefit from more choices on means, timing, and quantity of electricity they use. Pro-environment policymakers view the smart grid as key to providing better power quality and more efficient delivery of electricity to facilitate the move towards renewable energy[22].

An additional area for data-driven environmental innovation is traffic management and control. Urban planners benefit from the analysis of personal location

data for decisions involving road and mass transit construction, mitigation of traffic congestion, and planning for high-density development[25]. Such decisions can not only cut congestion but also control the emission of pollutants. At the same time, individual drivers benefit from smart routing based on real-time traffic information, including accident reports and information about scheduled roadwork and congested areas.

2.4 Security and Public Safety

Researchers in computational science, information systems, social sciences, engineering, medicine, and many other fields have been called upon to help fight violence, terrorism, cyber crimes, and other cyber security concerns. Facing the critical mission of international security, the need to develop the science of "security informatics" was recognized, with its main objective being the development of advanced information technologies, systems, algorithms, and databases for security related applications, through an integrated technological, organizational, and policy-based approach. BI&A has much to contribute to the emerging field of security informatics.

Big data is very valuable for fraud detection in the payment card industry. With electronic commerce capturing an increasingly large portion of the retail market, the merchants that bear ultimate responsibility for fraudulent card payments must implement robust mechanisms to identify suspect transactions often performed by first-time customers. To this end, some companies have developed solutions to provide merchants with predictive fraud scores for "Card-Not-Present transactions" in order to measure in real time the likelihood that a transaction is fraudulent. To do that, the services analyze buyer histories and provide evaluations, much like a summarized list of references but in the form of a single score. As fraudsters become more sophisticated in their approach, online merchants must remain ever more vigilant in their efforts to protect the integrity of the online shopping experience[30].

Security issues are a major concern for most organizations. Companies of different sizes are facing the daunting task of defending against cybersecurity threats and protecting their intellectual assets and infrastructure. Processing and analyzing security-related data, however, is increasingly difficult. A significant challenge

in security IT research is the information stovepipe and overload resulting from diverse data sources, various data formats, and large data volumes. Current research on technologies for cybersecurity, counter-terrorism, and crimefighting applications lacks a consistent framework for addressing these data challenges. Selected BI&A technologies such as criminal association rule mining and clustering, criminal network analysis, spatial-temporal analysis and visualization, multilingual text analytics, and cyber attacks analysis and attribution should be considered for security informatics research[16].

2.5 E-commerce and Market Intelligence

Significant market transformation has been accomplished by leading e-commerce vendors through their innovative and highly scalable e-commerce platforms and product systems. Tech giants such as Google, Amazon, and Facebook continue to lead the development of web analytics, cloud computing, and social media platforms. The data that e-commerce systems collect from the web are less structured and often contain rich customer opinion and behavioral information. By analyzing it, businesses can obtain targeted and personalized recommendations as well as increase sale and customer satisfaction. Many shoppers use Amazon's "Customers Who Bought This Also Bought" feature, prompting users to consider buying additional items selected by a collaborative filtering tool. The most prevalent business model for the Internet is based on financing products and services with targeted ads whose value correlates directly with the amount of information collected from users. Businesses care not so much about the identity of each individual user but rather the attributes of user's profile, which determines the nature of ads that are shown.

Analytics can also be used in the offline environment to study customers' in-store behavior to improve store layout, product mix, and shelf positioning. Companies are increasingly trying to link online activity to offline behavior, both in order to assess the effectiveness of online ad campaigns, as judged by conversion to in-store purchases, and to re-target in-store customers with ads when they go online[30]. As e-commerce becomes even more widespread, some new research areas are emerging such as: privacy-preserving data mining, network mining and parallel DBMS, etc.

2.6 Web Analytics

Web analytics integrate application usage data, app-centric analytics software and heuristics into the development process. Today, websites commonly use third party web analytics services to obtain aggregate information about users that visit their sites. This information includes demographics and visits to other sites as well as user behavior within their own sites. Web analytics solutions share common: objectives (monetization), requirements (analysis of visitors and conversions), and restrictions (meeting privacy and performance stakes). However, to obtain this aggregate information, individual user browsing behaviors are tracked across the web. These, potentially privacy-violating practices have been strongly criticised, resulting in tools that block such tracking as well as anti-tracking legislation and standards such as Do-Not-Track (DNT). These efforts, while improving user privacy, degrade the quality of web analytics.

Website publishers use web analytics information to analyze their traffic and optimize their site's content accordingly. Publishers can obtain analytics data by running their own web analytics software programs, that provide statistics about users on their site, such as pageviews, clickstreams, browsers, operating systems, plugins as well as frequency of returning visitors. However, they do not provide other potentially useful information, such as user demographics. For this reason, publishers often outsource the collection of web analytics to a third party data aggregator, such as comScore, Google, Quantcast or StatCounter. A data aggregator collects data from users visiting a publisher's website and presents this data in aggregate form to the publisher. This outsourcing is convenient for publishers, because they only have to install a small piece of code (i.e., a JavaScript code snippet) provided by the data aggregator. More importantly, this technique allows publishers to learn statistical information they could not otherwise learn from their own web server logs, such as the demographic profile of their user base and the other websites their users visit. A data aggregator can infer extended web analytics information because it collects user data across many publisher websites. Compiling extended web analytics via these collected data also benefits the data aggregator because it can sell this information to advertisers and publishers alike.

Although this method is beneficial for the publishers and the data aggregators, it raises concerns about users being tracked while browsing the web. Data aggregators are given a lot of information about users' actions on the web and have

to be trusted that they will not abuse it. These criticisms have led to industry self-regulation to provide opt-out mechanisms, the DNT initiative by the W3C, and many client-side tools, either to implement DNT, or to prevent tracking outright. To the extent that these efforts take hold, the ability for data aggregators to provide extended analytics to publishers will be degraded. In addition, even with tracking, inferring accurate user demographics is a difficult task that may produce inconsistent results [9].

Companies using analytics aim to satisfy the customer through early and continuous delivery of valuable software. Web analytics provides empirical evidence of application usage and end-user behavior that, when properly integrated into a development process, provides insight into user requirements, validation of development priorities and measure the accuracy and completeness of a test plan.

Web analytics provide:

- Application adoption and usage metrics within a specific operations framework.
- Production incident alerts from application exceptions.
- Organizational adoption and productivity analysis connecting application investment to enterprise ROI.

The value of web analytics seems obvious, but the details can make it difficult. Collecting, analyzing and acting on application runtime data poses unique challenges both in terms of the kinds of data that need to be gathered and the metrics that measure success. In order to be effective, web analytics implementations must take into account the diversity of today's applications. The emergence of cloud, mobile and distributed computing platforms has also influenced the way analytics should be integrated and analyzed.

An application feature can span one or more methods, incorporate multiple components, run across runtime surfaces and even be implemented multiple times in different languages. Measuring usage and performance is a requirement for monitoring the application and new features across devices and platforms. Many of today's applications are data-driven where the actual behavior itself is encoded in the data. Knowing what templates, workflows and other content is being processed can be more valuable than knowing which workflow or rendering engine processed that data. Session information can be defined differently within an app

server, a mobile session, within a browser or distributed by a cloud-based service. Unhandled exceptions, caught and thrown exceptions, unexpected performance or suspicious user behavior can all constitute a "production event." Applications are often comprised of multiple components, some on-premises and some service-based and are versioned at various times. Another important requirement is calculating the workflow across distributed parts of applications and then aggregating this information over time and across versions.

Consumer, business-to-business, and line-of-business applications each come with their own security and privacy obligations. These obligations are further fragmented by industry and by jurisdiction. Web analytics instrumentation and content management must be extensible and able to enforce these requirements for any individual application. Integration into existing platforms, processes and methodologies is a requisite to effective web analytics implementations.

A users' identity can be defined and tracked by gathering identifying information such as IP addresses, the operating system, browser and other user-related parameters. Web analytics solutions must have the capacity to enforce privacy and security policies at both client and aggregate levels. Ensuring effective data governance is a precursor to effectively analyzing the resulting runtime data.

In Universal Analytics, there are safeguards like IP masking and the Analytics browser opt-out add-on. The information stored in the local first-party cookie is reduced for `analytics.js`. To anonymize user data, only an identifier made up of two randomly generated 32-bit numbers is used. Clearing or deleting cookies from a browser does not necessarily ensure that future visits to a website will be considered as a new sessions in analytics.

2.7 Smart Health

The health community is facing a flood of health and healthcare related content generated from numerous patient care points of contact, sophisticated medical instruments, and web-based health communities. Two main sources of health big data are genomics-driven (genotyping, gene expression, sequencing data) and payer-provider (electronic health records, insurance records, pharmacy prescription, patient feedback and responses). Extracting knowledge from health big data poses significant research and practical challenges. This is in a big part due to the

HIPAA (Health Insurance Portability and Accountability Act) and IRB (Institutional Review Board) requirements for building a privacy-preserving and trustworthy health infrastructure and conducting ethical health-related research[20]. Health big data analytics, in general, lags behind e-commerce applications because it has rarely taken advantage of scalable analytical methods or computational platforms, due to the sensible nature of the information collected.

The potential advantages of big data analytics within the medical field have resulted in public policy initiatives to mine and leverage such data. David Cameron, former Prime Minister of the United Kingdom, recently announced that every National Health Service (NHS) patient would henceforth be a "research patient" whose medical record would be "opened up" for research by private healthcare firms. The Prime Minister emphasized that privacy-conscious patients would be given opt out rights. He added that "this does not threaten privacy, it doesn't mean anyone can look at your health records, but it does mean using anonymous data to make new medical breakthroughs"[5].

Over the past decade, Electronic Health Records (EHR) have been widely adopted in hospitals and clinics worldwide. Significant clinical knowledge and a deeper understanding of patient disease patterns can be gained from such collections. A good example is Google Flu Trends, which predicts and locates outbreaks of the flu making use of this type of information and aggregate search queries. Early detection of disease activity, when followed by rapid response, can reduce the impact of the disease. Yet another example is the National Retail Data Monitor (NRDM), which keeps tabs on sales of over-the-counter healthcare items from 21,000 outlets across the United States. By analyzing the remedies people purchase, health officials can predict short-term trends in illness transmission, with an often significant lead-time[30].

The clear limitations of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule's scope have been exposed through the development of broad new range of sources for healthcare data that are outside the scope of the HIPAA rules. These sources are often referred to as "wearables" and include healthcare mobile applications, healthcare websites and others. There is enormous concern about how this "non HIPAA" healthcare data is being addressed, and how the privacy interests of individuals are being protected.

Indeed, this weak spot for online privacy issues can be seen with the Department of Health and Human Services (HHS) itself, which has had to quickly patch privacy

and security issues on sensitive apps like its own "HIV Services" app on iOS and Android. The risk with health and wellness apps is the huge amount of clinical data collected. Companies may then not realize that they are sharing private data with third parties, such as advertising entities, analytics companies, social networks and hosted solutions[27].

The app transmitted data can make it easy for third parties to deduct if the end-user was pregnant, taking chemotherapy medication, being treated for AIDS, recovering from alcoholism or the like. Health-related apps that aren't regulated by the FDA and aren't covered by HIPAA, which means the vast majority of such apps can present the perfect storm of privacy exposure apps that collect highly sensitive data, which use third parties that are not trained to appropriately handle such data.

Examples of third-party services that routinely get data from health-related apps, but which may lack appropriate privacy or security safeguards, would be push-notification services, analytics packages, social network-sharing SDKs and third-party email services. A significant number of the entities associated with these third-party services have their own back-end proprietary databases of information that can be used to reidentify end-users who may believe themselves to be anonymous in the context of the app. An "anonymous" device ID that is transmitted to a party whose back-end database contains that same ID, plus the end-user's name, address, email or phone number, is not anonymous at all.

In other instances, third-party services may be configured to collect hashes of data such as phone number or email address that can easily be revealed through look-up tables. Still other services especially those of the hosted solution variety may have access to highly detailed, personally identifiable information about end-users. This data is often collected and stored in an unencrypted format and without any meaningful overlay of legal restrictions as to its disclosure or use. Not only can such data potentially be shared for profit, but also represents a significant exposure in the event of a data breach[27].

Finally, the mode of data collection by the app developer/publisher can also raise privacy issues. An app that has no third-party network traffic associated with it, but that nevertheless transmits information in plaintext over HTTP, makes it possible for any person within range of the end-user's Wi-Fi signal to sniff the data. Or, when the app user is accessing any employer or school connection, those organizations have access to that data as well.

Any compliance program involving health and wellness apps must address data collection at a technical level. Association rule mining and clustering, health social media monitoring and analysis, health text analytics, health ontologies, patient network analysis, and adverse drug side-effect analysis are all promising areas of research in health-related BI&A. In addition, due to the importance of HIPAA regulations, privacy-preserving health data mining is also gaining attention.

3 Big Data Security and Privacy Issues

Big data brings with it new privacy and security concerns. Although our data capacity is growing exponentially, we have imperfect solutions for many security issues that affect even local, self-contained data. While data warehousing handles the integration of the big amount of data across multiple business systems, Business Intelligence is concerned with how to use that integrated data to make strategic business decisions. However, whenever private data is accessed, security becomes a concern. Data breaches, location-based services, cloud computing and regulatory changes force organizations to review and revise their current privacy policies.

Data breaches rank high on the priority list and big data breaches will have an even bigger impact. The more information there is, the more likely it is that it includes personal or sensitive information. In addition, sources of information vary greatly, allowing multiple opportunities for infiltration. Furthermore, distributed computing, which is the only way to process the massive quantity of big data, opens up additional opportunities for data breaches. Political and structural issues may have a negative impact on the effectiveness of security policies, and organizations are open to malicious damage from disgruntled or careless employees. Organizations should compartmentalize personal information, restrict access, encrypt data when transmitting it across public networks, encrypt data on portable devices, and in storage to protect it from employees who have been given too much privilege, from rogue administrators and from hackers. Data loss prevention tools, tokenization, data masking and privacy management tools should also be considered.

Location information can be GPS information, the nearest cell tower, informa-

tion about wireless access points, indoor positioning information, speed, altitude, smart meter identifiers and IP addresses. Not every organization processes geolocation data, but the area is evolving rapidly, and a specific way of processing may suddenly surface as a privacy scandal (e.g. smartphones storing more location information than expected). Many providers are still in the "collect" stage rather than the "use" stage. They compile vast amounts of information, often without a clear plan of what to do with it. This is in contradiction with a fundamental privacy principle: collect information only for the purpose for which it is needed[14].

Building big data infrastructure in-house is a major investment of time and money for research, hardware, software, and countless other details, so most organizations will not install their own big data infrastructure. Thus, big data in the cloud security and privacy issues should be examined.

Cloud computing and privacy are innately at odds. Privacy laws apply to one specific country, however, the public cloud, in its ideal form, is not related to any country. Most privacy laws have some flexibility and, in many cases, there are legally acceptable solutions. Organizations should focus on the location of the legal entity of the provider, not on the physical locations of its operation centers. There are cases when sensitive company information should not leave the country but in most cases, and usually under conditions, in-country storage is not mandatory for privacy compliance. In some cases, it will be sufficient to ensure that personal data will not be stored in a specific country that is known for its privacy violations[14].

Table 1: Security and Privacy Issues

1	Infrastructure Security	• Secure computations in distributed programming frameworks • Security best practices for non-relational data stores
2	Data Privacy	• Privacy-preserving data mining and analytics • Cryptographically enforced data centric security • Granular access control
3	Data Management	• Secure data storage and transactions logs • Granular audits • Data provenance and verification
4	Integrity/Reactive Security	• End-point input validation and filtering • Real-time security monitoring

According to the Cloud Security Alliance's report [8], big data security and privacy challenges can be divided into four groups: Infrastructure Security, Data Privacy, Data Management and Integrity/Reactive Security. In Table 1, we enumerate the subpoints for each of these categories, that we will look closer at in this section.

3.1 Infrastructure Security

In this subsection we will look at the infrastructure vulnerabilities that are encountered in big data environments. Table 2 summarizes on one side the challenges for each problem enumerated bellow and on the other side shows the solutions that should be considered, which are further explained in Section 4.

3.1.1 Secure computations in distributed programming frameworks

The distributed computing framework, utilizing parallel computation across multiple workers, creates opportunities for breaches of security. Identifying a malicious or unreliable worker's computer and protecting the data from these unreliable processors is a key step in protecting a big data environment. Compromised worker nodes may tap the communication among other Workers and the Master with the objective of replay, Man-In-the-Middle, and DoS attacks. Rogue data nodes can be added to a cluster, and subsequently receive replicated data or deliver altered code. The ability to create snapshots of legitimate nodes and re-introduce altered copies is a straightforward attack in cloud and virtual environments and is difficult to detect.

3.1.2 Security best practices for non-relational data stores

In finding solutions to big data management, many organizations migrate from a traditional relational database to a NoSQL (Not Only Structured Query Language) database to deal with the unstructured data. However, NoSQL solutions were originally built as solution-specific tools to operate within a larger framework, leaving security to the parent system. Furthermore, the architectural flexibility that made NoSQL a good solution for multi-sourced data leaves it also vulnerable to attack. Across the board, NoSQL uses weak authentication techniques and weak password storage mechanisms. This exposes NoSQL to replay attacks and password brute force attacks, resulting in information leakage. NoSQL uses HTTP Basic- or Digest-based authentication, which are prone to replay or man-in-the-middle attack. REST, which is another preferred communication protocol, is also based on HTTP and is prone to cross-site scripting, cross-site request forgery, injection

attacks, etc. By manipulating the RESTful connection definition, it is possible to get access to the handles and configuration parameters of the underlying database, thereby gaining access to the file system. Although some of the existing NoSQL databases offer authentication at the local node level, they fail to enforce authentication across all the cluster nodes. Easy to employ injection techniques allow backdoor access to the file system for malicious activities. Since NoSQL architecture employs lightweight protocols and mechanisms that are loosely coupled, it is susceptible to various injection attacks like JSON injection, array injection, view injection, REST injection, GQL injection, schema injection, etc. Lenient security mechanisms can be leveraged to achieve insider attacks. These attacks could remain unnoticed because of poor logging and log analysis methods, along with other rudimentary security mechanisms. As critical data is stowed away under a thin security layer, it is difficult to ensure that the data owners maintain control.

Table 2: Infrastructure Security

	Challenges	**Solutions**
Secure computations in distributed programming frameworks	• Malfunctioning data nodes • Infrastructure attacks (Man-In-the-Middle, DoS) • Rogue data nodes	• Trust Establishment • MAC
Security best practices for non-relational data stores	• Transactional integrity • Lax authentication mechanisms • Inefficient authorization mechanisms • Susceptibility to injection attacks • Lack of consistency • Insider attacks	• Hardware appliance-based encryption/decryption and bulk file-based encryption • Using SSL/TLS to establish connections • Intelligent hashing algorithms • Data tagging techniques

3.2 Data Privacy

In this subsection we will look at data privacy concerns when handling big data. Table 3 enumerates challenges faced when protecting identifying user data, as well as presently used solutions. In Section 4 we explain the solutions further on, as well as how they can help avoid other challenges.

3.2.1 Privacy-preserving data mining and analytics

Big data can enable invasions of privacy, invasive and unrequested marketing and threaten civil liberties The amount of information collected on each individual can be processed to provide a surprisingly complete picture of one's life, as well as some future plans. As a result, any organization that owns data are legally responsible for securing it and managing the usage policies. Attempts to anonymize specific data are not successful in protecting privacy because there is so much data available that some of it can be used as a correlation for identification purposes. Users' data is also constantly in transit, being accessed by internal users and outside contractors, government agencies, and business partners sharing data for research[24]. Privacy, for legal reasons, must be preserved even in spite of a higher monetary or system performance cost.

In some cases, an insider in the company hosting the Big Data can abuse his/her level of access and violate privacy policies. If the party owning the data outsources data analytics, an untrusted partner might be able to abuse their access to the data to infer private information from users. This case can apply to the usage of Big Data in the cloud, as the cloud infrastructure (where data is stored and processed) is not usually controlled by the owners of the data. Sharing data for research is another important use. However, ensuring that the data released is fully anonymous is challenging because of re-identification. Re-identification is the process by which anonymized personal data is matched with its true owner.

3.2.2 Cryptographically enforced data centric security

There are two distinct approaches to applying security controls on data visibility: first, managing access to the system, and secondly, applying encryption to the data itself. The first method, which is easier and less costly to implement, is also less effective, as it provides what it calls a larger "attack surface". If the system is breached, then the attacker has access to all the data. Deploying encryption on all data on a granular basis helps ensure that even if there is a system breach, the data itself remains protected.

3.2.3 Granular access control

The biggest challenge for implementing big data security is respecting privacy concerns while still allowing for data usage and analytics to obtain the correct results. This is one of big data's greatest security challenges, as the collection of data is useless without being able to use it; but a data privacy breach has legal and ethical implications, as well as damaging the marketplace. Granular access control acts on each piece of data, thus ensuring a high level of both security and usability[24].

There are three problems with effective implementation of granular access control:

1. Keeping track of the security/secrecy requirements and policies in a cluster-computing environment
2. Keeping track of user access throughout all components of the ecosystem
3. Proper implementation of security/secrecy requirements with mandatory access control

In a shared environment with many different applications, the applications that contribute data need to communicate those requirements to the queriers. This coordination requirement complicates application development, and is often distributed among multiple development teams. An additional challenge in tracking these requirements is to maintain access labels across analytical transformations. An element created from two or more other elements may be restricted at the least upper bound of the restrictions of the other elements, according to some lattice. However, some data access requirements do not follow a lattice, such as aggregated medical records that could be broadly releasable, while the elements that contribute to those aggregates are highly restricted. Once a user is properly authenticated, it is still necessary to pull security-related attributes for that user from one or more trusted sources. LDAP, Active Directory, OAuth, OpenID, and many other systems have started to mature in this space. One of the continual challenges is to properly define authorizations, so a single analytical system can respect roles and authorities that are defined across a broad ecosystem.

Choosing an appropriate level of granularity is on its own a challenge, and requires knowledge of the data storing infrastructure and analytics systems. A row typically represents a single record, and row-level access is often used for data

derived from multiple sources. A column represents a specific field for all records, and a column-level access is often used for sensitive elements, because the identification columns are not necessarily available to users. Cell-level access means a label is applied to every grain of information, and can support a wide range of usages and analytics. However, in order to be effective, such technique must be rigorously applied, and in data sets this big, the overhead could be prohibitive.

Table 3: Data Privacy

	Challenges	**Solutions**
Privacy-preserving data mining and analytics	• Insider worker's violation of power and access rights • Untrusted partner's access and ability to infer private information from users • Preserving anonimity of released data in spite of re-identification	• Differential Privacy • Universal Homomorphic Encryption
Cryptographically enforced data centric security	• Cryptographically-enforced access control method using encryption • Protocol for searching and filtering encrypted data • Protocol for computation on encrypted data • Protocol ensuring the integrity of data coming from an identified source	• Identity and attribute based encryption • Boneh and Waters • Fully homomorphic encryption • Group signatures
Granular access control	• Keeping track of the security/secrecy requirements and policies in a cluster-computing environment • Keeping track of user access throughout all components of the ecosystem • Proper implementation of security/secrecy requirements with mandatory access control	• Appropriate level of granularity required • Model the expectation of change over time

3.3 Data Management

This subsection describes data management privacy and security challenges in big data environments. In Table 4 we can see the risks next to the appropriate solutions, discussed in Section 4.

3.3.1 Secure data storage and transactions logs

In order to deal with petabytes of data, a form of storage called auto-tiering has become a necessity. In auto-tiering, items are assigned a level of storage automatically, based on policies established by the organization. Auto-tiering opens a number of vulnerabilities because of unverified storage services or if some security policies are mismatched. Moreover, because data is moved automatically, rarely accessed, critical information could end up on a lower tier, which typically has less security attached to it. Finally, auto-tiering maintains transaction logs of its activities, which now also have to be protected in order to protect the data of its logs information.

There are specific vulnerabilities associated with big data storage: confidentiality and integrity, data provenance, and consistency. The need for availability presents risks as well: stored data needs to be available on demand, which requires the system to have hooks to the data that can be exploited.

Auto-tiered environments are susceptible to two types of attacks:

- Collusion attacks - service providers exchange keys and access codes and thus gain access to more than the subset of data assigned to them
- Rollback attacks - an outdated dataset is uploaded to replace the latest version

In addition to those attempting to steal sensitive information or damage user data, storage service providers are also assumed to be untrustworthy third parties. Data transmission among tiers in a storage system provides clues that enable the

service provider to correlate user activities and data set. Without being able to break the cipher, certain properties can be revealed. Due to the extremely large size, it is infeasible to download the entire data set to verify its availability and integrity. Lightweight schemes are desired to provide verification that is probabilistically accurate and implies low computing and communication overhead. Auto-tiering also places challenges on the service providers to guarantee constant availability. Not only does weaker security at lower tiers risk Denial of Service (DoS) attacks, the performance gap between lower tiers and higher tiers also extends the backup windows during periods of fast restore and disaster recovery.

It is now typical that data flows among tiers and is shared by multiple users. To maintain consistency among multiple duplicates stored at difference locations is non-trivial. Two issues that need to be addressed carefully are write-serializability and Multi-Writer Multi-Reader (MWMR) problems. While a data owner stores the cipher text in an auto-tier storage system and distributes the key and permission access to the users, each user is authorized to have access to a certain portion of the data set. Also, the service provider cannot interpret the data without the cipher key materials. However, if the service provider colludes with users by exchanging the key and data, they will obtain a data set that they are not entitled to.

In a multi-user environment, the service provider can launch roll-back attacks on users. When an updated version of a data set has been uploaded into storage, the service provider can fool the user by delivering the outdated version. Certain evidence is required to help users ensure that data is up-to-date, and the user should have the capability of detecting the inconsistency. A lack of record keeping will lead to disputes between users and storage service provider, or among users. When data loss or tampering occurs, transmission logs/records are critical to determining responsibility. For example, a malicious user outsources data to a storage system. Later, the user reports data loss and asks for compensation for his claimed loss. In this case, a well-maintained log can effectively prevent fraud[24].

3.3.2 Granular audits

The goal of real-time security monitoring is to raise the alert at the first sign of trouble. However, since that doesn't always happen due to the challenges of identifying real threats among a huge number of false positives, it's important to have frequent, granular audits to identify breaches after the fact. Audit information

also helps identify exactly what happened and why it was not detected by the monitoring system.

An effective audit depends on four factors:

1. Completeness of the information that is required for the audit
2. Timely access to the information
3. Integrity of the information
4. Controlled access to the information, to prevent tampering, thus compromising the integrity

A successful audit will also require information about the proper techniques and technologies in your big data infrastructure, such as application logging and Security Information and Event Management (SIEM). If properly and timely conducted, a granular audit will help avoid a similar attack or problem in the future.

3.3.3 Data provenance and verification

Big data is collected from a wide variety of sources, and in enterprise settings that can mean millions of end-user machines. In this environment, the question of how trustworthy the data might be is of paramount importance. As the volume grows, so does the complexity of the provenance. Provenance information is contained in the metadata attached to each data object and provides information about the object's creation. The provenance metadata includes the origin for the big data infrastructure itself, which is in fact a way of having meta-metadata. As development in this area progresses, provenance metadata will become more complex due to large provenance graphs generated from provenance-enabled big data applications. Furthermore, it is worth noting that analytics for graphs of this size and complexity are very resource-intensive in terms of computational overhead. Malfunctioning infrastructure, and attacks on infrastructure from inside or outside the organization are the biggest threat and the provenance metadata itself must be protected as well in order to allow audits and other detection mechanisms to be effective in verifying data sources.

In applications, when large numbers of components collaboratively generate such large provenance graphs, it is inevitable that some infrastructure components could sporadically malfunction. Once the malfunction occurs, provenance records may be incorrect because they cannot be timely generated. As a result, the malfunctioning infrastructure components will reduce the provenance availability and reliability, which makes it an appealing attack target, because of its pivotal role in the usability of Big Data applications, it naturally becomes a target in Big Data applications. An outside attacker can forge, modify, replay, or unduly delay the provenance records during its transmission to destroy the usability of the provenance, or violate privacy by eavesdropping and analyzing the records.

Compared to the outside attacks, the infrastructure inside attacks are more harmful. An inside attacker could modify and delete the stored provenance records and audit logs to destroy the provenance system in Big Data applications.

Table 4: Data Management

	Challenges	**Solutions**
Secure data storage and transactions logs	• Confidentiality and integrity • Provenance • Availability • Consistency • Collusion attacks • Roll-back attacks • Disputes	• Dynamic Data Operations • Privacy Preservation • Secure Manipulations on Encrypted Data
Granular audits	• Timely access to the information • Controlled access to the information, to prevent tampering, thus compromising the integrity • Completeness of the information that is required for the audit • Integrity of the information	• Forensics/SIEM tool • Audit Layer/Orchestrator
Data provenance and verification	• Malfunctioning infrastructure components • Infrastructure outside attacks • Infrastructure inside attacks	• Granular Access Control • Independent Persistence

3.4 Integrity and Reactive Security

Finally, we explore and present input validation and filtering as well as system monitoring and the concerns that come with these requirements in Table 5. As in the subsections above, the solutions will be further elaborated on in Section 4.

3.4.1 End-point input validation and filtering

Because of the immense pool of data sources, including end-point collection devices, a major challenge facing big data schemes is whether the data is valid from the point of input. Given the size of the data pool, both data collection devices and programs are susceptible to attack and should not always be trusted.

An adversary may interfere with a device which gathers data, or may tamper with the data collection application running on that device to send a malicious input to a central data collection system. An adversary may perform ID cloning attacks on a data collection system by creating multiple fake identities and by then providing malicious input from the faked identities. The challenges of these types of attacks become more acute in a bring-your-own-device (BYOD) scenario. Since an organizations's users and employees are allowed to bring their own devices and use them inside the enterprise network, an adversary may use his/her device to fake the identity of a trusted device and then provide a wrong input to the central data collection system. A more complicated scenario involves an adversary that can manipulate the input sources of sensed data. For example, instead of compromising a temperature sensor, an adversary may be able to artificially change the temperature in a sensed location and introduce malicious input to the temperature collection process.

3.4.2 Real-time security monitoring

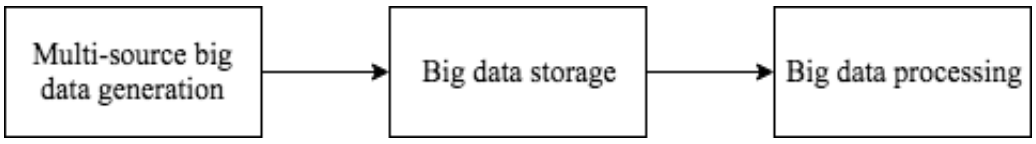
Real-time security monitoring is intended to alert the organization at the very first sign of an attack. However, there is an enormous amount of feedback from Security Information and Event Management (SIEM) systems, whose aim is to provide big-picture feedback of an organization's data security in real time. Few organizations have the resources to monitor this feedback with the kind of oversight and analysis necessary to identify real attacks from false alarms. Privacy considerations drive the need for high security, but make detection delicate, as privacy laws need to be navigated along with the analytics that will identify attacks. Big data analytics itself can be used to identify threats, including differentiating real threats from false positives. Logs can be mined for anomalous connections to the cluster. Such analysis and monitoring tools should be placed within the framework or in a front-end system whose job is primary to provide the analytics necessary to assess the SIEM feedback to identify threats[24].

Security monitoring requires that the infrastructure/platform is intrinsically secure. Threats to a Big Data infrastructure include rogue admin access to applications or nodes, application threats, and eavesdropping on the line. Such an infrastructure is mostly an ecosystem of different components, where (a) the security of each component and (b) the security integration of these components must be considered. The other major threat model revolves around adversaries that will try to evade the big data analytics tools used to identify them. Attackers can create evasion attacks in an effort to prevent being detected, and also launch data poisoning attacks to reduce the trustworthiness of the data sets used to train big data analytics algorithms. Apart from these security threats, other barriers become important, such as legal regulations. Depending on where the monitored data exists, certain privacy laws may apply. This may create hurdles because some data might not be available for security monitoring or may only be available in a certain format (e.g., anonymized).

Table 5: Integrity/Reactive Security

	Challenges	**Solutions**
End-point input validation and filtering	• Device tampering • ID cloning attacks • Malicious or compromised data	• Prevention of manipulation • Detection and filtering of compromised data • Devices with authentication capabilities
Real-time security monitoring	• Security of the public cloud • Security of the Hadoop cluster, the security of the nodes, the interconnection of the nodes, and the security of the data stored on a node • Security of the monitoring application itself • Security of the input sources	• Monitoring and analysis tools

Figure 1: Big Data Life Cycle



4 Solutions and good measures

Business Intelligence aims to use a company's data to derive meaningful insights that can help a business flourish. However, when an organization grants anyone access to their data, it is essential that proper security measures are implemented. Privacy, data security and cybersecurity are core components of the effectiveness and success of any company.

In order to ensure big data privacy, various mechanisms have been developed in recent years. These mechanisms can be grouped based on the stages of the big data life cycle shown in Figure 1. In data generation phase, we look at what we collect, so access restriction and/or de-identifying data techniques are used. The approaches to privacy protection in data storage phase are chiefly based on encryption procedures, which we show in Table 7. Additionally, to protect sensitive information, hybrid clouds are utilized where sensitive data is stored in private cloud. For the data processing phase, we incorporate Privacy Preserving Data Publishing (PPDP) and knowledge extraction from the data. In PPDP, anonymization techniques such as generalization and suppression are utilized to protect the privacy of data. These mechanisms can be further divided into clustering, classification and association rule mining based techniques. While clustering and classification split the input data into various groups, association rule mining based techniques find the useful relationships and trends in the input data. To handle diverse measurements of big data in terms of volume, velocity, and variety, and arriving at very high speed from various sources, there is a need to design efficient and effective frameworks [18].

Data aimed at being used in Business Intelligence tools is normally presented to employees via the corporate intranet or over e-mail. This way, IT workers to use existing security infrastructures to help them to secure their Business Intelligence data. These existing security frameworks, when combined with corporate policies against the distribution of sensitive company information are, in most cases, enough to ensure data privacy. Table 6 shows five critical areas necessary for a comprehensive security policy for a BI environment.

Table 6: Areas of security policy

Data Classification	which BI data is deemed sensitive and what measures are required to protect it; there may be multiple levels of sensitivity and associated protection measures;
User/Role Classification	what BI data users should be able to access based on their role or function in the organization;
Entitlement Standards	how BI applications are allowed to access data and perform specific functions;
Data Transmission	where encryption is required and what levels are needed for user access, file transfer, etc;
Data Storage	where data is allowed to be stored, how data is backed-up, what data retention polices apply;

With Big Data comes the opportunity for many new advances, including additional avenues for research. At the same time, current rules (including but not limited to those in the healthcare industry) are viewed as creating barriers to some kinds of effective research, and it is clear at a minimum that the rules and operating procedures are complicated and confusing enough that certain research efforts are being impeded. In addition, there is a current rulemaking proposal to revise the "Common Rule", the primary set of research regulations related to federal funded "human subjects" research, with a goal of streamlining research requirements where there are no meaningful privacy or other risks to research subjects. Research is not just for universities, and companies in a wide variety of industries need to understand these research rules (including when their own internal data analytics "cross the line" into the regulated research area), and an even larger group of companies have data that can (and should) be made available for beneficial research projects [21].

In the rest of this section we will look closer into some key points that help implement a security policy such as: implementing best-practice IT standards on computer machines, de-identify customer data to ensure data privacy before it enters a BI environment, ensuring that company data and applications do not leave company devices and approved development environments and others.

4.1 IT Standards

All developers and consultants who create applications for an organization should do so using the organization's own computers. This allows the IT department to ensure that a certain minimum standards of security are met. These standards can include password complexity requirements, automatic locking of idle machines, etc. Compliance with ISO/IEC 27001:2013 [4] standards is a good goal for an IT department to strive for. Furthermore, organizations should ensure that the obtained certification relates to the use of Big Data and use components in the Big Data system that follow same security standards to maintain the desired level of security.

When considering the NoSQL databases security concerns, which are shown in Table 2, it is good practice to use SSL/TLS to establish connections between the client and the server and also for communication across participating cluster nodes. Adopting such mechanisms to exchange mutually verifiable keys and establish trust would ensure data confidentiality while data is in transit. The NoSQL architecture should support pluggable authentication modules with the capacity to enforce security at all levels as the situation demands.

Communication across clusters should also be better controlled so that each node can validate the trust level of other participating nodes before establishing a trusted communication channel. The utilization of intelligent hashing algorithms can ensure that the data is replicated consistently across nodes, even during node failures. All NoSQL products usually recommend to run them on a trusted environment, which ensures that only trusted machines can access the database ports. Appropriate data tagging techniques, with time stamps enforced through intelligent algorithms while piping data from its source, will defend against unauthorized modification of data and preserve the authenticity of the data stored.

4.2 Source Filtering

Organizations should also pay close attention to data sources to make sure the data collected is not malicious, and we have identified the validation and filtering risks in Table 5. Solutions must implement prevention of tampering, as well as detect and filter compromised data. However, it is almost impossible to build a complex and extensive system that is completely resistant to tampering. Therefore, there

is no secure way to ascertain the integrity of your data at the input. The big data collection system design must take into account this inherent unreliability and the inevitability of relying on untrusted devices and try to develop the most secure data collection platforms and applications possible. The system should be able to identify likely ID cloning attacks and be prepared with cost-effective ways to mitigate the attacks. In addition it should understand that a determined adversary can infiltrate any existing system with false data, so designers must develop detection and filtering algorithms to find and eliminate malicious input. Only devices with authentication capabilities should be used to ensure that validation of endpoint sources is possible. Assigning confidence levels on the endpoint sources and re-evaluating them regularly, especially after patches or changes in firmware is necessary. If confidence in endpoint source is low, one should use it in combination with other higher confidence endpoint sources for taking actions.

4.3 De-identifying Data

To minimize the impact of data leaks and serious threats to data privacy, companies should consider de-identifying the data before it enters the BI system wherever possible. De-identification is an irreversible process that strips data of elements that represent Personally Identifiable Information (PII). PII is any information that can be used alone or together with other sources to uniquely identify, contact or locate an individual. It consists of a broad range of information, including dates of birth, addresses, driver's license numbers, credit card numbers, bank account numbers, health and insurance records, and much more. According to the U.S. General Accounting Office, 87% of the U.S. population can be uniquely identified using only gender, date of birth and ZIP code. So it's not just the most obvious types of PII, like credit card numbers, that require protection[6]. This is a very common process employed in health information systems where the privacy of Protected Health Information (PHI) is mandated by HIPAA policy.

One technique for de-identifying data is tokenization. This process, popular in the credit card industry to achieve compliance with PCI standards, involves substituting sensitive data with tokens that reference the data in some other external database. This means that only the tokens are exposed to BI users, and additional privileges can be required to access the underlying data.

Another recent technique is differential privacy, which is a framework for formalizing privacy in statistical databases. This method tries to maximize the value of query results by minimizing the risks of revealing individual information, by adding mathematical noise to a small sample of the individual's usage pattern.

De-identification has been a source of enormous and confusing debate. Governments discuss de-identification at the threshold of the privacy debate, in connection with how personal data is defined. Technologists and data scientists debate the potential for re-identification of data subjects in various scenarios, and the media reports cases of re-identification as a serious threat to our privacy. In fact, de-identification presents a potential win-win scenario, where privacy risks are reduced in dramatic ways, while the value of data largely remains. It is also clear that there is an immediate trade off of sorts between privacy and data value, with reduced privacy concern often meaning less valuable, useful or reliable data. So, as the volume of data grows, and the potential public and private benefits of this data grow even more, it will be critical for there to be a thoughtful debate about de-identification, so that the value of data can be preserved and improved at the same time that privacy can be protected.

This also needs to encompass the debate (largely for public entities) about transparency, where true public disclosure of data creates the most substantial privacy concerns. In other environments, where transparency is a secondary value, these risks often can be controlled through legal, contractual and security protections, to preserve data benefits without meaningful privacy risk. Managing an effective de-identification process whether in individual industry segments like healthcare or on a broader basis where the rules are more ambiguous presents both a challenge and opportunity for many companies seeking to capitalize on the value of their data.

With so much data and powerful analytics, it may be impossible to completely ensure that the ability to identify an individual was completey removed, especially if there are no rules established for the use of anonymized data files. For example, if one anonymized data set was combined with another completely separate data base, without first determining if, to protect anonymity, any other data items should be removed prior to combining, it is possible individuals could be re-identified. The important and necessary key that is usually missing is establishing the rules and policies for how anonymized data files can be combined and used together.

If data masking is not used appropriately, by analyzing big data one could easily reveal the actual individuals even by looking at data that has been masked. Organizations must establish effective policies, procedures and processes for using data masking to ensure privacy is preserved. Since big data analytics is so new, many small-, medium- and even some large-scale companies do not realize the associated risks, so they use data masking in ways that could breach privacy. Many resources are available, such as those from IBM, to provide guidance in data masking for big data analytics [17].

4.4 Encryption

Encryption of data in transit and at rest is a must when working with big data to ensure data confidentiality and integrity. This means also looking into proper encryption key management solutions, considering the vast amount of devices that need to be covered. When encrypting, one should consider the timeframe for which the data should be kept, since data protection regulations might require that some data be disposed of, after a certain period of time [23]. When designing databases, one must also consider if some confidential data could be contained in separate fields, so that they can be easily filtered out and/or encrypted.

From previous section, we know that encryption is a necessary step for securing NoSQL databases, as seen in Table 2. Data integrity needs to be enforced through an application or middleware layer. Passwords should never be left in the clear while at rest and in transit, but instead should be encrypted or hashed using secure hashing algorithms. Similarly, data stored in the database should never be left without protection. Considering the already weak authentication and authorization techniques employed in NoSQL databases, it is vital to keep the data encrypted while at rest despite the associated performance impacts. Hardware appliance-based encryption/decryption and bulk file-based encryption are faster and would alleviate some concerns about the performance impacts of encryption. However, hardware-based encryption comes with certain trade offs, as it often leads to a vendor lock-in, low-strength key used in encryption/decryption that can be exploited by attackers. As a result, malicious users who gain access to the file system could directly extract sensitive data from the file system[8].

Encryption is also the most widely used measure for ensuring data privacy. In

Table 3 we indicated which encryption techniques are used for solving the already mentioned challenges and in Table 7 we list the encryption methods and their definition.

Table 7: Types of encryption

Type	**Description**
Attribute based encryption	Access control is based on the identity of a user complete access over all resources.
Homomorphic encryption	Can be deployed in IBE or ABE scheme settings updating cipher text receiver.
Storage path encryption	It secures storage of big data on clouds.
Usage of Hybrid clouds	Hybrid cloud is a cloud computing environment which utilizes a blend of on-premises, private cloud and third-party, public cloud services with organization between the two platforms.

Identity and attribute based encryption methods enforce access control using cryptography. In identity-based systems, plaintext can be encrypted for a given identity and the expectation is that only an entity with that identity can decrypt the ciphertext. Any other entity will be unable to decipher the plaintext, even with collusion. Attribute-based encryption extends this concept to attribute-based access control. A complete homomorphic encryption scheme would keep the data encrypted even while it's being worked on[24].

Boneh and Waters [10] construct a public key system that supports comparison queries, subset queries and arbitrary conjunction of such queries. In a breakthrough result in 2009[15], Gentry constructed the first fully homomorphic encryption scheme. Such a scheme allows one to compute the encryption of arbitrary functions of the underlying plaintext. Earlier results constructed only partially homomorphic encryption schemes. Group signatures enable individual entities to sign their data but remain identifiable only in a group to the public. Only a trusted third party can pinpoint the identity of the individual.

Confidentiality and integrity can be achieved with robust encryption techniques and message-digests. The exchange of signed message-digests can be used to address potential disputes. User freshness and write-serializability can be solved by periodic audit and chain hash or persistent authenticated dictionary (PAD). Secure untrusted data repository (SUNDR) can be used to detect fork consistency attack

and write serializability. Broadcast encryption and key rotation can be used to improve scalability. Researchers have proposed technologies to handle the provenance issues. Data availability can be improved through proof of retrievability (POR) or provable data possession (PDP) methods with high probability.

Regarding collusion attacks, as long as the users do not exchange their private keys, a policy-based encryption system (PBES) can successfully guarantee a collusion-free environment. If the users are willing to exchange their private keys without exchanging the decrypted content, a mediated decryption system can avoid collusion attacks. If the users are willing to exchange the decrypted contents, digital rights management can prevent collusion attacks. Two non-repudiation protocols have been proposed recently to address disputed issues. Digital signatures using asymmetric encryption, regular audits, and hash chaining can help secure the data. Persistent Authenticated Dictionaries (PADs), which allow queries against older versions of the structure, can assist in identifying Rollback attacks.

The problem is not an absence of technologies, but the absence of an all-inclusive systemic approach or framework. In large scale auto-tier storage systems, there is no systematic approach to integrate them into a seamless, holistic solution. The result is a patchwork of security policies, rather than a unified structure in which all the parts work together to provide a sufficient level of security. The non-uniform security policies among different tiers pose an additional challenge to securing inter-tier data transmission. More considerations are necessary to balance tradeoffs among security, usability, complexity, and cost.

4.5 Access Control

The most rudimentary security technique to secure BI is to apply access controls to the data. It is important that only company-authorized developers have access to development environments. In addition, users should only be granted access to data on an "as-needed" basis. Having access to the wrong data means potential security vulnerabilities and can result in erroneous analysis results. Some best practices to achieve this goal include:

- Authorized users must have their own login credentials to the development environment itself.

- Each authorized user should be provided with data warehouse credentials that are different than the user's development environment credentials and also different from other users' data warehouse credentials. However, it is generally more difficult to maintain controls in the data warehouse over the long haul, granting or denying access to users for specific tables, columns, and even rows of data requires a lot of DBA time. It's much easier to manage the controls using presentation/reporting tools. The only drawback to this strategy is if people use different presentation and reporting tools to access the same data, one would have to manage security across different tools and the chance of making a mistake quickly grows.

- Training users on the importance of keeping login credentials confidential (and instituting stiff penalties for violation of these standards, as well as pointing out that any inappropriate action performed with a person's username implies that the activity was performed by that person himself).

- Allowing access to development environment from company personal computers only.

- Requiring a secure VPN connection when users use their company PCs to access development environments remotely.

As we see in Table 2, which shows us the Infrastructure Security problems and solutions, one must ensure the trustworthiness of the workers' computers with the help of trust establishment and Mandatory Access Control (MAC). In trust establishment, employees are stringently authenticated, and given access properties only by masters, who are explicitly authorized to do so. Employee properties must be checked periodically to ensure they continue to conform to predefined standards. In the MAC system, the access of each employee is constrained to a very limited set of tasks, and the ability of a user to control the objects it creates is highly restricted. MAC adds labels to all file system objects defining the appropriate access for each object, and all users' appropriately defined access. However it cannot guarantee privacy for computations based on output keys produced by untrusted workers. To prevent information leakage through the outputs, a recently developed de-identification framework of differential privacy based on function sensitivity can be used. However, when implementing the solutions outlined above, one must consider performance penalties caused by MAC and limitations of differential privacy

in providing guarantees.

To cope with the complexity and extensive scale of tracking and implementation in big data environments, it is recommended to reduce the complexity of granular access controls on the application level. Instead, use the infrastructure to implement as much of the access control as possible, and adopt standards and practices that simplify whatever access controls still exist in the application level. This also helps build a framework to support NoSQL security, and ensure that the access scheme assigns an appropriate level of granularity, balancing the size of the data store with the need for security[24].

Very granular access control is an important starting point for securing provenance and verification metadata. This means that even if a data object is removed, it might be an ancestor of other data; therefore, its provenance should be retained.

Access control should also be dynamic and scalable, as well as use, lightweight, fast authentication mechanisms for reducing overhead. Furthermore, secure channels between infrastructure components and responsive, flexible revocation mechanisms should be included in the architecture.

4.6 Monitoring

It is impossible to completely prevent data from being stolen, but it is crucial that a breach can at least be detected. There are several good approaches for doing so, but here are some security measures recommended that will, together, make a breach detectable in most cases:

- block all outgoing internet access on development environments, which will prevent users from directly removing files from the environment
- control the way in which files can be transferred to/from the development environment by: (a) limit the development environment's network access to only incoming Remote Desktop connections and mapped network drives, and (b) disable all file transfers over RDP
- install and configure an audit tool to monitor your network mapped drives and keep a record of the files that each user transfers to/from the development environment

- in the event of a suspected breach (and whenever an employee who had access to company data leaves the organization), an administrator should review the logs stored by the audit tool for suspicious transfers from the development environment to the user's own company PC, and if any breach is suspected, the administrator should immediately notify the appropriate department within the organization
- enable logging on nodes participating in the Big Data computation, on databases(relational or not), as well as Big Data applications
- detect and prevent modification of logs
- regularly test the restoration of Big Data backups considering the vast amount of data being used in the system

To perform a check, logging options need to be enabled for all components to ensure completeness of information. This would include applications at all layers, including operating systems. And a forensic or Security Information and Event Management (SIEM) tool will collect, analyze, and process the log information, and this will perform granular audits as indicated in Table 4. This should be done outside of the infrastructure used for the data, to ensure that it is not threathened by the same type of attacks.

Monitoring and analysis tools are being developed and announced by different Hadoop providers and vendors. An alternative solution is the implementation of front-end systems to monitor Hadoop requests (e.g., Database Activity Monitoring proxy of firewall). The application security depends on the application itself and whether security controls have been built-in. New solutions and frameworks are slowly entering the Big Data arena to help real-time monitoring. One of the criticisms of these tools in Hadoop is that they are only batch-oriented, which is useful for historical or trend analysis, but not for real-time monitoring. Examples of attempts to overcome this hurdle include Apache Storm[3] and Apache Kafka[2]. Other real-time streaming applications, which are built upon Hadoop, are entering the market, such as Apache Flink[1].

4.7 Data Staying on Company Devices

There is a fine balance between creating a secure development environment and blocking it to the point that a BI tool user is not able to properly extract information. There are many legitimate reasons why a worker would need access to resources outside the company's network and preventing him from getting access to those will, at a minimum, lead to decreased work satisfaction and productivity loss. In addition, if an employee indeed intends to remove data/applications from company computers, there is nothing that can realistically be done to stop him without locking down the development machines. It is, therefore, crucial to not only have well established company policies, but also make sure that the employees can be trusted.

It is worth mentioning, that the explosive growth in mobile devices has lead to the adoption of Bring You Own Device (BYOD) in corporate IT, which can have a major impact on BI security. Users are starting to expect mobile access to everything they have in their office, and companies are finding it more efficient for employees to provide their own devices, including laptops, mobile phones and tablets. For a growing segment of users, mobile will be the exclusive way they consume BI. This means, however, that sensitive BI data will leave the safe confines of the corporate network, and personal and business data will end up intermixed on the same device. Another major concern for mobile access is lost and stolen devices. In addition, if a mobile device has offline capabilities (data is cached locally) the risk of data theft is very high. Policies need to be created and enforced for authentication of mobile devices, and BI applications should be architected to avoid retention of local data copies. Data encryption should also be high on the priority list whenever mobile devices are involved.

4.8 Data Staying on Development Machines

Most enterprise organizations will want to ensure that they have a robust centralized development environment in place. For example, an organization could direct developers to use a specialized development server, which they can access remotely. Advantages to having a centralized development environment include automated backups, insurance against a stolen devices, and ease of collaboration among developers. This helps with big data sets as well, which which may be

too resource-demanding to be handled by personal computers, due to resource constraints.

In addition, it is essential that both raw and transformed data reside within a properly-architected data warehouse as much as possible. Having a properly architected data warehouse comes with a myriad of advantages including: a central source, so that various departments and business units can produce reports using identical data; added layers of security, allowing for granular control of company data by database, table, column, and/or row; audit capabilities to allow administrators to see exactly which user and machine requested certain data in the event of a security breach. Connections to company data warehouses should only be permitted from approved development environments, by installing necessary drivers and performing necessary configuration on those machines only.

4.9 Plan for Security Breaches

It is important that organizations create a plan of action to deal with data being stolen, which may include: (i) government requirements on reporting certain sensitive stolen data (HIPAA-protected data, for instance), (ii) whether there is any risk that the stolen data could wind up in the hands of competitors, and (iii) implications to a company's image by announcing that a data breach has occurred.

5 Ensuring Data Privacy in Different Use Cases

In this section, we will look at two use cases of big data, Healthcare and Web Analytics, and how the BI applications or tools ensure data privacy. These two areas deal with different types of data. In clinical data, one keeps the relation between the patient's personally identifiable information and their medical history. Whereas in statistical data, usually gathered from web analytics concerning user behaviour, the application developers and/or publishers are not interested in personally identifiable information, and they only identify users with the help of IDs. However, with such type of data, a user can still be identified with the help of correlation with other databases or geo-location data. We want to look at how different privacy ensuring approaches are applied for these types of data to ensure user information is protected during data processing. Table 8 gives a comparison of these two use cases. We look at how some of the solutions mentioned in Section 4 to protect user privacy are implemented for each type of data: clinical and statistical.

Table 8: Techniques for ensuring data privacy in clinical and statistical data

Solution	**Clinical Data**	**Statistical Data**
De-identification	Safe Harbor Method, Expert Determination Method	Not necessary
Anonymization	Masking, Randomisation, Generalisation	Anonymization of IP Addresses
Access Control	Private/Public Cloud	User Permissions
Encryption	ABE	HTTP over SSL (HTTPS)

Research challenges arise from the complications of applying existing methods of privacy preservation to big data. First, the multi-source of the data, increases the risk of re-identification. Secondly, anonymization techniques like, generalization, suppression, anatomization, etc. can not be directly applied to big data because it is difficult to distinguish between sensitive and non-sensitive attributes in unstructured data. And lastly, it is hard to maintain trade off between privacy and data utility to help advance research.

5.1 Clinical Data

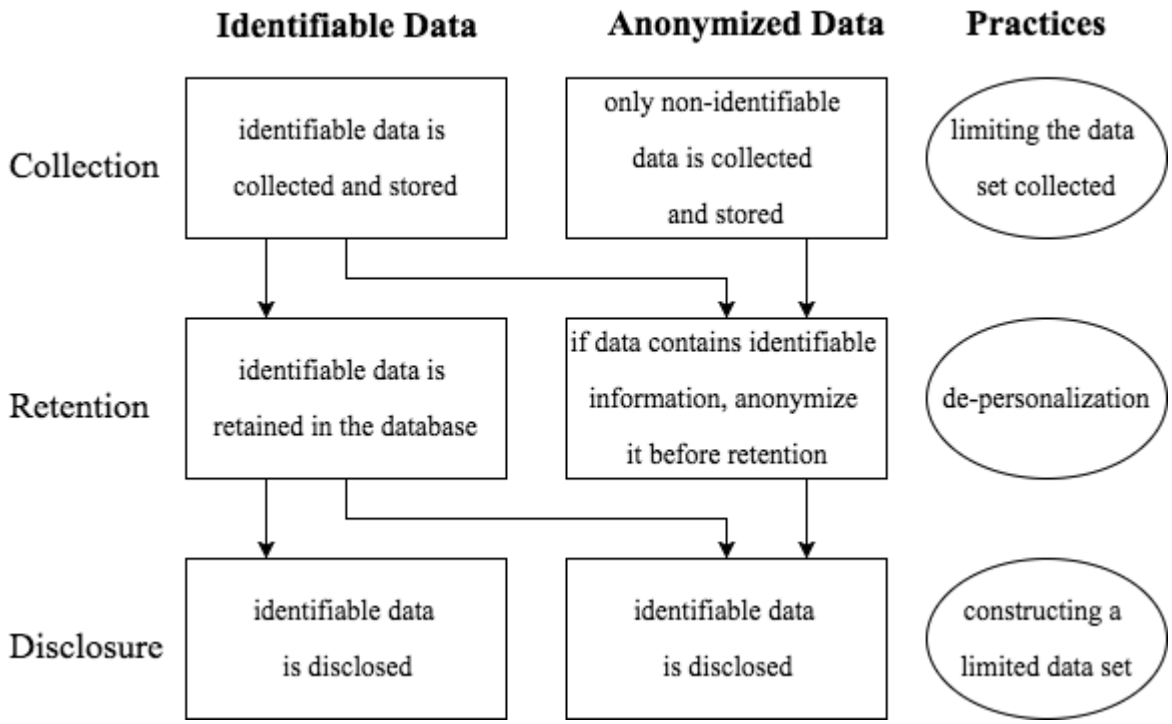
There is increasing adoption of IT in health research and practice. The adoption of Electronic Data Collection (EDC) and Electronic Medical Records (EMRs) is growing. Researchers are increasingly turning to EMRs as a source of clinically relevant patient data. Therefore, there is a trend towards collecting, storing, and exchanging health information electronically. The ease of storage and exchange of large volumes of health data electronically has raised privacy concerns, since existing medical training does not provide yet adequate coverage of practices for protecting EMR privacy. Basic practices for protecting patient privacy are not followed. Serious vulnerabilities in Internet-based health data collection systems have been reported.

One of the mechanisms to safeguard PHI is to anonymize it. This means remove or obfuscate any identifying information about the individual patients in the data set, hence making the re-identification of those individuals very difficult. Data anonymization can be applied during any of three different stages of a clinical research study: collection, retention, and disclosure. These three activities are sequential and the diagram in Figure 2 shows the workflow between these stages[13]. If data is collected anonymously, then by definition it is anonymized during retention and disclosure. If the data is anonymized during retention then that data will be anonymized during disclosure. Similarly, if data is identifiable when it is collected then it will be identifiable until something specific is performed to anonymize it.

In healthcare, de-identification and anonymization techniques can be applied to individual patient data (IPD) in order to fulfill transparency, disclosure and research requests while safeguarding the privacy of individuals and conforming to existing directives and regulatory guidance. Data privacy laws consider that if personal data is removed or de-identified and subject code identifiers cannot be linked back to specific individuals, then it is no longer considered personal data.

There are a number of techniques that can be used by data providers to adequately de-identify datasets prior to sharing. De-identified PHI is defined in the HIPAA Privacy Rule as 'Health information that does not identify an individual

Figure 2: Data anonymization practices in different stages of processing flow



and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information. De-identifying the data means removing or recoding identifiers, removing free text verbatim terms, and explicit references to dates. Participants' identification code numbers are de-identified by replacing the original code number with a new random code number. Anonymization is a step subsequent to de-identification that involves destroying all links between the de-identified datasets and the original datasets. The key code that was used to generate the new identification code number from the original is irreversibly destroyed (i.e., destroying the link between the two code numbers) [7].

The standard in (HIPAA) CFR - Title 45: Public Welfare, Subtitle A 164.514 provides 2 alternative approaches to de-identification, which we describe in the following subsections.

5.1.1 Harbor Method

This method describes 18 types of identifiers that must be removed (eg, by deleting, recoding or redacting) in order for the resulting datasets can be considered de-identified. The identifiers most commonly collected in clinical studies are: names (eg, investigators and vendors), contact numbers and addresses (eg, investigator and vendor telephone and fax numbers, and postal and email addresses), dates, device identifiers and serial numbers, photographic images, characteristics (eg, verbatim text including reported adverse events, medical history, concomitant medications and other comments), and any other unique identifying number (eg, treatment kit numbers), or code, except a random identifier code[7]. The 'Safe Harbor' method is not focused on clinical trial data and therefore data providers using a de-identification method based on this approach also remove and review other personal information that may be present in the dataset.

5.1.2 Expert Determination Method

The Expert Determination method is a risk management exercise that incorporates both direct and quasi-identifiers. It satisfies both the need to protect the

identity of individuals, and allows organizations to perform deep analysis on data used for secondary purposes.

Both the Safe Harbor and Expert Determination approaches start with a shared principle of identifying direct (e.g. ID number) and quasi (e.g. date of birth and date of death) identifiers, and applying de-identification techniques. Direct identifiers need to be completely anonymised, while quasi identifiers can be offset or aggregated. In the context of providing data in a secure controlled access model where data requests are reviewed and subject to data sharing agreements, data providers may decide that a statistical assessment of the risk of re-identification (a key part of the Expert Determination approach) is not necessary in most cases.

To check if a dataset is anonymised, one must demonstrate that after anonymisation it is no longer possible to isolate some records of an individual in the dataset; there is no ability to link, at least, two records concerning the same data subject or a group of data subjects (in the same database or in two different databases); and there is no possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes. Examples of anonymisation techniques that could be applicable to clinical reports are: Masking, Randomisation (noise addition and permutation), Generalisation (aggregation and k-anonymity).

When going through an anonymisation process, one should go through the following steps:

1. Determination of direct identifiers and quasi-identifiers
2. Identification of possible adversaries and plausible attacks on the data
3. Data utility considerations
4. Determining the risk of re-identification threshold and evaluation of the actual risk of re-identification
5. Anonymisation methodology
6. Documenting the anonymisation methodology and process

Any reports made from clinical data should include information about the anonymisation process, including the techniques used and the reasons for using

them, as well as the outcome of the analysis of the risk of re-identification or alternatively confirmation that the three criteria for anonymisation have been fulfilled. It is also recommended to encrypt health data both in transition and at rest, since companies in the health space have become popular targets for cyber criminals due to the vast amount of valuable data being gathered daily.

For delivering electronic health services with ubiquitous network access, scalability and cost savings, cloud-based platforms are the most desirable. However, transfering electronic health records (EHRs) to the cloud poses major threats to privacy, data integrity, and confidentiality. To manage some of these concers, there are several regulations and standards proposed, which provide guidelines and frameworks for sharing and exchanging health information via digital representation of clinical data. These are the HITECH Act, HIPAA, HL7 CDA, CEN 13606, ISO 22600 EHRcom, IHE XDS, and openEHR. Even so, there are many non-standardized communication architectures, hence, healthcare providers are the ones responsible for protecting their data to ensure that proper access controls are in place.

With the growing popularity of cloud computing, EHR data can be stored in the cloud and shared among authorized parties. Shared data may include patient-sensitive and personal information, such as chronic diseases, mental health issues, psychiatric care, sexual behavior, fertility issues, abortion status, and HIV status, which require implementations that guarantee data privacy. Encryption is used as a common practice, to provide a simple form of access control that protects the system against unauthorized access.

Some studies such as [19], propose solutions for privacy-preserving data sharing with attributed-based encryption (ABE) in the cloud to encrypt data and to provide the hierarchical access structure for fine-grained data sharing. With ABE, fine-grained access control in the EHR can be easily managed, and this has been seen as a promising approach for cloud-based EHR systems. One of the challenges of data sharing though is key management. Yu et al. [29] pointed out data security and access control issues in the EHR sharing within the public domain because of heavy computation overhead in key distribution and data management, which occurs in applying fine-grained access control. They used Key-Policy ABE (KP-ABE), Proxy Re-Encryption (PRE), and lazy re-encryption in order to define and enforce access-control policies, but secure and dynamic access rights are

demanding. Similarly, in [28], a fine-grained access control and searchable public-key encryption technique were applied in an EHR system. A hierarchical access structure was demonstrated to ensure common trust for information sharing.

Another aspect that should be taken into consideration is the fact that EHR data could be stored in multiple clouds due to the need for scalability and privacy. To better manage the system, researchers in [26] propose the use of different types of clouds, i.e. a public cloud and a private cloud. EHRs stored in the private cloud can only be accessed by the authorized medical professionals, whereas those in the public cloud can be used by medical researchers, pharmaceutical companies, insurance companies, public health agencies, commercial or government agencies, etc. Each cloud requires a RBAC policy that is based on a special type of ABPRE, namely Attribute-Based Proxy Re-encryption. Electronic health record system data sharing is based on the technology of threshold encryption.

5.2 Statistical Data

In this work, we refer to statistical data as the data gathered from applications that show how a user interacts with a given site or application. The most widely used way to gather such data is with the help of Google Analytics, which can be easily embedded into any given webpage. Usually, the collected user data is identified with a clickId and/or IP address, not with any information that can be considered PII. However, recently, there have been more and more concerns related to the huge amount of data gathered about users' online browsing patterns, which lead to new approaches of processing this data.

Anonymizing Google Analytics implies changing the IP address of visitors of any given website to such that it can no longer be directly assigned any other tracking information collected through Google Analytics. Google Analytics has triggered serious concerns with German data protection advocates in the past. The assumption was that too much data was being collected that could be assigned to each user. Despite the fact that oftentimes IP addresses reside behind various, potentially multi-layered, NAT setups, they may just as often be regarded as personal data due to certain amount of user-inference that they contain. In certain cases

users may also obtain a globally unique IP address. As a result, the anonymity of the IP addresses of users has become the basic requirement for legally compliant use of the tracking service. The IP anonymization implies seting the last octet of IPv4 user IP addresses and the last 80 bits of IPv6 addresses to zeros in memory. The Google Analytics API is extended for this purpose with the `anonymizeIP` function. To anonymize user data, only an identifier made up of two randomly generated 32-bit numbers is used. In addition, each user has the right to disagree with the use of their personal data. To this end, each site operator should provide a link that allows users to opt-out of data acquisition by Google Analytics. Existing old data, collected without anonymization, should, in principle, be deleted.

To maintain access control, one can define user profiles with the help of Google Analytics. This ensures that data is only accessed by employees with the correct user permissions to collect, store, modify or download data. You can assign user permissions at the account, property, and view levels. Edit permission is required at the account level to create filters and edit permission is required at the view level to apply filters.

To implement secure transmission, Google Analytics uses HTTP Strict Transport Security (HSTS), which instructs browsers to use that encryption protocol for all traffic between end users, websites, and Google servers, if it is supported. To ensure security for clients, there is no method by which one can opt out of HTTPS encryption of Google Analytics traffic.

From a privacy perspective, this scenario poses interesting and very concrete research challenges. Data sources used by BI application often reside in different systems, different departments, even in different companies. This implies that data in the sources of the BI applications is subject to different constraints. It was collected under different privacy agreements with the citizens in mind, and in addition, the different institutions may further regulate the use of the information they obtained.

The biggest issue is to define the privacy requirements the BI application must obey when processing the data provided by the source. Privacy laws and agreements are typically defined at a very high level and with a certain degree of ”fuzziness”. However, BI developers need to know which data can be extracted from the source databases, whether this data can be used to clean/refine data from other providers (e.g., entity resolution), which report users can view the data with, whether data can be shown in aggregate form, at which level of aggregation, and

so on. This degree of precision is needed to know how to develop and test the BI application and also how to audit and to resolve possible disputes. Privacy policy languages and purpose-based access control languages are of general applicability and can be used in different contexts where data is released to third parties. However, their generality makes it hard to express privacy requirements that are testable during the BI data lifecycle.

Three areas of risk where PII can be sent through Google Analytics are:

- On-site form submissions
- Site search
- Inbound traffic

One of the most common cases of PII in Google Analytics is the result of form submissions that pass information via the URL which is automatically recorded by GA when the page loads. This happens often in sign-up and registration forms. It could be as seemingly harmless as an email or password capture from a newsletter or login portal. Forthermore, if for example an ecommerce site passes credit card numbers or full billing information via the URL, the risks are more severe.

Another case is the search field. Often people will mistake the site search for a login box and type in their username or password. Web applications should be designed by keeping in mind the search usability to make sure it isn't misinterpreted, for example by ensuring there is a placeholder such as: "search this site" rather than something more ambiguous like a blank space. On-site search tool should also filter out names, email addresses, usernames, or passwords in your site search reports.

Bloggers may use trackbacks/linkbacks/pingbacks to communicate between an organization website and theirs. In some cases, PII may be hidden in the URL or Full Referrer string, such as email addresses.

In order to ensure that PII does not leak into Google Analytics, organizations should be productive in monitoring for PII leakage and adjust your testing to detect leakages before launch. With analytics becoming such a fundamental part of business decisions, organizations can not afford to risk being forced to delete your data. PII data integrity is a business requirement during the design and upgrades of website and/or applications. In addition, ongoing PII audits should be regularly scheduled in the areas mentioned above to identify leakages and fix minor

problems before they become a much bigger issue.

By using these two different use cases, we have shown that the data privacy problem is different from one application to another. The privacy aspects of Clinical and Statistical data are distinct in the type of data that each domain collects, stores and processes. There is a difference in how critical the potential consequences of a data exposure may end up being. There is also a difference in how relevant the produced data remains with time. In some scenarios, the historical dataset of statistics collection domain may not reflect the most recent behavioural and preferential traits of the very same user. On the other hand, some of the historical clinical records of the patient, when exposed, tampered with or altogether destroyed may pose just as severe, if not even more, consequences for the patients. As patients accumulate their health record history, having later in life evaluations made on the basis of false records may have hazardous consequences. At the same time, both the Clinical and Statistical data field strive for prevention of private data exposure. Both domains may often have real or near-real time processing requirements and stringent processes in place. Any research and successful solutions that are undertaken in either of the fields could prove to be beneficial for the other one. The benefits are not scoped exclusively to these two fields, and instead may as well extend to the other domains, like urban planning, e-government and public-safety. It is only through gathering and processing user-related data for the common benefit that we can expect to live in smart cities, and be surrounded by ubiquitous IoT items at our homes, work and public spaces.

6 Conclusion

Along with the benefits of using big data in a great variety of applications including healthcare, e-commerce and infrastructure, comes a lot of concern for how this vast collection of data is kept secure and how our privacy as users is ensured. This work provided an analysis of the big data world within the research environment as well as different domains of our global, interconnected, economy. We started our research by reading recently published papers and reports that provided a general overview of the current Big Data and Business Intelligence status. We followed up by looking into the references of the most relevant papers to explore further work. Next, we identified the motivation for this work and created an outline for the research. The first step was finding the privacy and security concerns in modern organizations and categorizing them, as seen in Section 3. Next, knowing what problems need to be addressed, we looked at the methods that can be implemented to help tackle these concerns, as presented in Section 4. We established that the risks to user data and privacy are strongly correlated with the type of the application and the collection environment of each data set. In Section 5, we explored the difference in the amount of personally identifiable information across different fields, by looking closer at the data collected, stored and processed in healthcare and web analytics.

Business Intelligence and Analytics tries to make sure that a balance is reached between drawing beneficial information out of big data while minimizing the privacy and security risks within an organization. One of the hardest challenges that still needs to be overcome is the multi-source nature of data. There is a huge amount of applications that collect their own sets of data. Each has a different way of ensuring privacy for their customers, as well as different data stores and security policies. And while each may satisfy all necessary requirements, most applications nowadays share their data with third party organizations. As a result, the risk of re-identification is always present, since this data comes from different sources and it may provide a much more complete user picture.

We discussed some of the unresolved or partially-resolved concerns. One of such partially-resolved concerns is the lack of standards and intra- and inter-governmental regulations and well-established common practices. This is mainly

due to the fact that big data has seen an exponential increase in usage in recent years while the necessary guidelines did not have enough time to catch up. In the case of healthcare, there are more standards in place, such as the Health Insurance Portability and Accountability Act (HIPPA), and more are being developed and adjusted as the requirements change. However, due to the rise in IoT devices, the global amount of big data available will only increase, which will make it easier for even non-clinical data to be just as easy to correlate and use to identify individuals. This poses new privacy challenges and there is ample space for further collaborative, across the fields, research efforts in both academia and industry. It is through raising the general public awareness and academic research that we can learn to understand the potential bottlenecks of the existing privacy domain corner-cases. And, in the process, we may obtain the skills that will help us avoid making duplicate mistakes in the new, yet non-existent, technology use cases and emerging domains.

References

[1] Apache Flink. https://flink.apache.org/.

[2] Apache Kafka. https://kafka.apache.org/.

[3] Apache Storm. https://storm.apache.org/.

[4] ISO/IEC 27001:2013 information technology – security techniques – information security management systems – requirements. https://www.iso.org/standard/54534.html.

[5] Everyone 'to be research patient', says David Cameron. http://www.bbc.co.uk/news/uk-16026827, 2011.

[6] Protecting personally identifiable information: What data is at risk and what you can do about it. *A Sophos White Paper*, 2011.

[7] Data de-identification and anonymization of individual patient data in clinical studies - a model approach. *TransCelerate BioPharma Inc.*, 2013.

[8] Expanded top ten big data security and privacy challenges. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf, 2013.

[9] Istemi Ekin Akkus, Ruichuan Chen, Michaela Hardt, Paul Francis, and Johannes Gehrke. Non-tracking Web Analytics. 2012.

[10] D. Boneh, C. Gentry, and B. Waters. Collusion resistant broadcast encryption with short ciphertexts and private keys. 2005.

[11] H. Chen. AI, E-Government, and Politics 2.0. *IEEE Intelligent Systems*, 2009.

[12] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Q.*, 36(4), 2012.

[13] Khaled El Emam. Data anonymization practices in clinical research. 2006.

[14] Gartner. Top five privacy issues organizations must tackle. http://www.itbusinessedge.com/slideshows/show.aspx?c=91946.

[15] Craig Gentry. A fully homomorphic encryption scheme. 2009.

[16] Chen H. *Intelligence and Security Informatics for International Security: Information Sharing and Data Mining.* 2006.

[17] Rebecca Herold. 10 big data analytics privacy problems. https://www.secureworldexpo.com/industry-news/10-big-data-analytics-privacy-problems, 2014.

[18] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. *Journal of Big Data*, 2016.

[19] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou. Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. *IEEE Transactions on Parallel and Distributed Systems*, 2013.

[20] K. Miller. Big data analytics in biomedical research. *Biomedical Computation Review*, 2012.

[21] Kirk J. Nahra. The top ten privacy and security issues companies need to watch in 2016. https://www.bna.com/top-ten-privacy-n57982066294/, 2016.

[22] Dr. S.L. Nalbalwar, Jayesh D. Ruikar, and Shailesh R. Sakpal. Smart Grid: A modernization of existing power grid. *International Journal of Advanced Engineering Research and Studies*, 2012.

[23] European Union Agency For Network and Information Security. Big data security: Good practices and recommendations on the security of big data systems. 2015.

[24] Jason Parms. More info, more problems: Privacy and security issues in the age of big data. https://www.business.com/articles/privacy-and-security-issues-in-the-age-of-big-data/.

[25] Carlo Ratti. Mobile landscapes: Using location data from cell-phones for urban analysis. 2006.

[26] Fatemeh Rezaeibagha and Yi Mu. Distributed clinical data sharing via dynamic access-control policy transformation. *International Journal of Medical Informatics*, 2015.

[27] Steven Roosa. A deep dive into the privacy and security risks for health, wellness and medical apps, 2015.

[28] J. Sun and Y. Fang. Cross-domain data sharing in distributed electronic health record systems. *IEEE Trans. Parallel Distrib. Syst.*, 2010.

[29] S.Yu, C.Wang, K.Ren, and W.Lou. Achieving secure, scalable, and fine-grained data access control in cloud computing. *INFOCOM*, 2010.

[30] Omer Tene and Jules Polonetsky. Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, 2013.

[31] Amy P. Wesolowski and Nathan Eagle. Parameterizing the dynamics of slums. *AAAI Spring Symposium*, 2010.

Additional Reading

[32] Ahmed Abbasi, Conan Albrecht, Anthony Vance, and James Hansen. MetaFraud: A meta-learning framework for detecting financial fraud. *MIS Q.*, 36(4):1293–1327, 2012.

[33] D. P. Acharjya and Ahmed P Kauser. A survey on big data analytics: Challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications*, 2016.

[34] Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, and Frederic Stahl. A survey of data mining techniques for social network analysis. *Journal of Data Mining and Digital Humanities*, 2014.

[35] Elmustafa Sayed Ali Ahmed and Rashid Saeed. A survey of big data cloud computing security. *International Journal of Computer Science and Software Engineering (IJCSSE)*, 2014.

[36] Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, et al. The stratosphere platform for big data analytics. *The VLDB Journal*, 23(6), 2014.

[37] Bart Baesens, Ravi Bapna, James R. Marsden, Jan Vanthienen, and J. Leon Zhao. Transformational issues of Big Data and analytics in networked business. *MIS Q.*, 38, 2014.

[38] Maged N. Kamel Boulos, Antonio P. Sanfilippo, Courtney D. Corley, and Steve Wheeler. Social web mining and exploitation for serious applications: Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. *Computer Methods and Programs in Biomedicine*, 2010.

[39] Alex G. Büchner and Maurice D. Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Rec.*, 27(4), 1998.

[40] Michael Chau and Jennifer Xu. Business intelligence in blogs: Understanding consumer interactions and communities. *MIS Q.*, 36(4):1189–1216, 2012.

[41] Annamaria Chiasera, Fabio Casati, Florian Daniel, and Yannis Velegrakis. Engineering privacy requirements in business intelligence applications. *Proceedings of the 5th VLDB Workshop on Secure Data Management*, 2008.

[42] Brian Demilia, Michael Peded, Kenneth Jorgensen, and Ramesh Subramanian. The ethics of bi with private and public entities. *Communications of the IIMA*, 2012.

[43] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM Trans. Internet Technol.*, 2003.

[44] Adel S. Elmaghraby and Michael M. Losavio. Cyber security challenges in smart cities: Safety, security and privacy. *Journal of Advanced Research*, 2014.

[45] Federico Michele Facca and Pier Luca Lanzi. Mining interesting knowledge from weblogs: a survey. *Data and Knowledge Engineering*, 2004.

[46] Xin James He. Business Intelligence and Big Data Analytics: An overview. *Communications of the IIMA*, 2014.

[47] Daning Hu, J. Leon Zhao, Zhimin Hua, and Michael C. S. Wong. Network-based modeling and analysis of systemic risk in banking systems. *MIS Q.*, 36(4):1269–1291, 2012.

[48] A. Juan-Verdejo, B. Surajbali, H. Baars, and H. G. Kemper. Moving business intelligence to cloud environments. 2014.

[49] Raymond Y. K. Lau, Stephen S. Y. Liao, K. F. Wong, and Dickson K. W. Chiu. Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. *MIS Q.*, 36(4):1239–1268, 2012.

[50] Jay Lee, Behrad Bagheri, and Hung-An Kao. Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics. *International Conference on Industrial Informatics (INDIN)*, 2014.

[51] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, 2011.

[52] Kirsten E. Martin. Ethical issues in the big data industry. *MIS Q.*, 2015.

[53] Sung-Hyuk Park, Soon-Young Huh, Wonseok Oh, and Sang Pil Han. A social network-based inference model for validating customer profile data. *MIS Q.*, 36(4):1217–1237, 2012.

[54] DIMITRIOS Pierrakos, GEORGIOS Paliouras, CHRISTOS Papatheodorou, and CONSTANTINE D. Spyeopoulos. Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction*, 2003.

[55] Nachiketa Sahoo, Param Vir Singh, and Tridas Mukhopadhyay. A hidden markov model for collaborative filtering. *MIS Q.*, 36(4):1329–1356, 2012.

[56] Andreas Seufert and Josef Schiefer. Enhanced business intelligence - supporting business processes with real-time business analytics. *Conference Paper 16th International Workshop on Database and Expert Systems Applications*, 2005.

[57] Yunchuan Sun, Junsheng Zhang, Yongping Xiong, and Guangyu Zhu. Data security and privacy in cloud computing. *International Journal of Distributed Sensor Networks*, 2014.

[58] Kim Verkooil and Marco Spruit. Mobile business intelligence: Key considerations for implementations projects. *Journal of Computer Information Systems*, 2013.

[59] Mary J. Wills. Decisions through data: Analytics in healthcare. *Journal of Healthcare Management*, 2014.

[60] S. Yakoubov, V. Gadepally, N. Schear, E. Shen, and A. Yerukhimovich. A survey of cryptographic approaches to securing big-data analytics in the cloud. pages 1–6, 2014.

[61] Kudakwashe Zvarevashe, Mainford Mutandavari, and Trust Gotora. A survey of the security use cases in big data. *International Journal of Innovative Research in Computer and Communication Engineering*, 2014.