Studies in Big Data 11

Hrushikesha Mohanty
Prachet Bhuyan
Deepak Chenthati *Editors*

Big Data

A Primer

Springer

Studies in Big Data

Volume 11

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

Security and Privacy of Big Data

Sithu D. Sudarsan, Raoul P. Jetley and Srini Ramaswamy

Abstract Big data, with its diversity (voice, video, structured and unstructured), has brought in unique challenges to security and privacy due to its shear scale. They are expected to be distributed, cloud-based and hosted by service providers. Security challenges in terms of cryptography, log/event analysis, intrusion detection/prevention, and access control have taken a new dimension. Privacy of online and cloud data is being addressed by governments, researchers, policy makers, as well as professional/standards bodies. The book chapter would cover challenges, possible technologies, initiatives by stakeholders and emerging trends with respect to Security and Privacy.

Keywords Security · Privacy · Big data · Cloud · Internet of things (IOT) · Anonymity · Critical information infrastructure (CII) · Confidentiality · Integrity · Availablity · Online privacy · Offline privacy

1 Introduction

Data security and privacy issues are no strangers to us. What is it that makes "big data" security and privacy different? And, why do we need to have a dedicated chapter on this? Just as a kingdom and an empire have different characteristics and need to be governed with different strategies, even though it is all about administering the land with its resources and inhabitants, big data differs in its characteristics from traditional data which warrants a detailed discussion.

S.D. Sudarsan (✉) · R.P. Jetley
ABB Corporate Research, Bangalore, India
e-mail: sudarsan.sd@in.abb.com

R.P. Jetley
e-mail: Raoul.Jetley@in.abb.com

S. Ramaswamy
US ABB, Cleveland, USA
e-mail: srini@ieee.org

H. Mohanty et al. (eds.), *Big Data*, Studies in Big Data 11,
DOI 10.1007/978-81-322-2494-5_5

Security mechanisms to safeguard data of a given type, say voice or email, have reached a level of maturity, and one can talk about accepted common practice or best practices. Here, in big data, we are dealing with possibly a diverse set of data ranging from voice, video, images, and structured as well as unstructured text. Even in the absence of scale, accepted security mechanisms for such diverse data do not exist today. Similar to the kingdoms in an empire, where each kingdom will have its own law, sometimes in conflict with another kingdom within the same empire, security, and privacy requirements of specific data within a big data need not be same but could be even at loggerheads.

Data storage mechanisms, till recently, were based on normalization, codification, extraction, and so on. The underlying criteria were based on high cost of storage and transport which were true to the technology of twentieth century. With time, practices based on these criteria have become the norm. In today's world, the cost of storage is close to zero and availability of connectivity is a foregone conclusion. We are now looking at data being stored as they are created, with sufficient redundancy to ensure availability in a distributed way. Obviously, the security mechanisms that were relevant to normalized single point storage do not suffice for distributed and redundant storage. A summary of key considerations of the past as against the emerging trends is provided in Table 1.

Cloud provides resources on demand and at least, in theory, expected to be fully elastic. Elasticity implies elimination of resource constraints in terms of platform, software, and infrastructure. In the absence of resource constraints, availability is improved through redundancy. While redundant resources increase the availability, they throw additional challenges to confidentiality and integrity. Even if confidentiality is managed, handling integrity with multiple copies is a challenge as all the copies need to be in sync. Cloud is also expected to involve third-party providers, perhaps

Table 1 Design considerations of past and future

Criterion	Waning (past)	Emerging (futuristic)
Storage cost	High. Minimize data by extraction, encoding, and normalizing	Close to zero. Retain as much source data as possible
Storage security	Uncommon. Mostly physical access control. Once accessed, entire file system is accessible. Granular secure storage only on need basis	Secure and safe storage is expected by default. Strong logical access control mechanisms
Communication cost	High	Connectivity is a given and cost is close to zero
Communication security	Optional. Security using crypto algorithms only when needed	Default, e.g., IPv6
Computation	Limitations due to single core and serial algorithms	Multi-core is default and advantages of parallel algorithms
Accessibility	From a single access point with optional backup access/redundancy	From multiple access points and as much redundancy as possible
Data synchronization/ integrity	As provided by the storage provider. Explicit mechanisms needed only if backup/redundancy storage exists	Challenge to be addressed due to redundant multiple copies

except for few fully owned private clouds. Security while dealing with third-party at a global level is another challenge. To achieve scale and redundancy, use of cloud as well as service providers have become imperative. Security and privacy issues applicable to cloud are thus very relevant and part and parcel of handling big data.

Internet of things (IoT) enables any device to be able to connect any other device using the internet. To highlight any device aspect, "internet of everything" is also used. At the same time, to specifically support industrial devices, "industrial IoT" is used. Irrespective of the universalization or restriction, IoT is a key ingredient of the upcoming industrial revolution "industry 4.0." This revolution promises seamless communication across any and every connected device. Each physical device has corresponding cyber device ensuring cyber-physical system in the true sense. IoT is a delight to research and business community at large with the need for new protocols, security solutions, applications, products, and systems. Many of the IoT proponents expect every device to be uniquely identifiable, which is fast tracking IPv6 adoption. The current transition is reflected in the increasing support for dual IP stack, which supports both IPv4 and IPv6. The presence of dual stack is a security challenge starting from the limitations of traditional firewalls to intrusion detection systems. If new communication protocols are added to the mix, security challenge compounds multifold.

Cloud and IoT along with other factors are catapulting the amount of data, and we have "big data" to deal with. Early database designers had to deal with the high cost of storage and communication. They went to the extent of not just storing only basic data but also encoded and normalized them to have the least amount of storage. This also ensured that only necessary data were transmitted and information was derived from this data as computation was considered relatively inexpensive. This also augured well for security practitioners since confidentiality was achieved typically with access control and cryptography, while integrity was achieved with error checking mechanisms. The key challenge was to ensure the availability as data were stored only at one location in a secure way. Later, backup and standby systems came up as the cost of storage and communication reduced over time. With ability to provide elastic storage with as much redundancy as needed, cloud has changed the very assumptions of security practitioners. IoT-enabled devices would generate and transmit so much data that security issues as well as managing the life cycle of those data are other dimensions that need to be addressed.

As devices become part of the cyber world with cloud and IoT, the criticality of cyber security takes another dimension. During cyber-attack, cyber-physical systems by their very nature create damages both in cyber as well as physical world. This is very important to understand, since in typical IT systems, cyber-attacks seldom cause physical damage. One can compare the effects of cyber-attack on a laptop to that of a surgical robot or an automobile resulting in their malfunction.

Yet another aspect of big data is about the nature of data itself. Often one hears about the quality of data, and efforts are made to get clean data. In fact, if the data are not clean or as per the quality requirements, they tend to be rejected. Redundant data are ignored. Any copy of the data is only for the purpose of backup and disaster recovery. However, with big data, one expects any data and every data to

be stored. Redundancy is expected to improve the availability. One may appreciate the challenge of keeping the integrity of data and synchronizing multiple copies.

Security has traditionally focused on confidentiality, integrity, and availability (CIA) in descending order of importance. Confidentiality has been handled using access control mechanisms and cryptography. Integrity is achieved using signatures, digests, and error detecting/correcting codes. Availability is managed with redundant communication links and standby storage systems. With big data and cyber-physical systems, the current practice of decreasing order of importance of CIA is unlikely to hold water.

To summarize, the emergence of cloud and IoT and the resultant big data brings in new security dimensions and paradigms. Challenges from cloud architecture to third-party reliance to making every device uniquely identifiable in the cyber-physical domain to big data security and analytics are all there. Exciting times ahead!

2 Security Versus Privacy

Security and privacy seem to be at loggerheads at all times and is a highly debated subject for quite sometime among researchers and lawmakers. One of the important reasons for this inconclusive debate is the subjective nature of the definitions attributed to these terms. Security aims to reduce risk. To reduce risk, specific information about a resource is often called for. As soon as the resource is related to entities such as individuals and corporates, privacy concerns creep in. One way this issue is typically addressed is by anonymizing data.

Yet with big data if at all anonymity can be provided has become a question mark. We do have standing examples of using anonymized data and yet identify

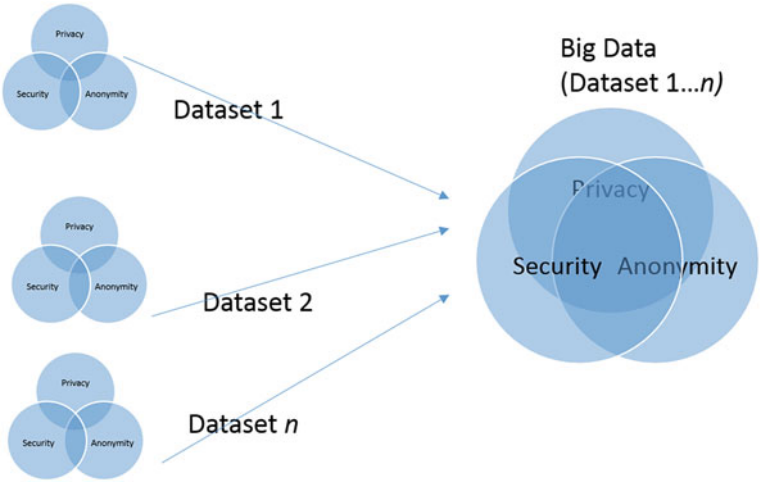

Fig. 1 Effect of big data on privacy and security

individuals with certainty. As early as in 2002, Latanya Sweeney was able to correlate data from the Group Insurance Commission (GIC) and voter registration list for the city of Cambridge to identify Massachusetts Governor Weld unambiguously [1]. In this case, the GIC data had personally identifiable information (PII) removed and yet it was possible to recover by correlating two different data sets made available for good reasons. With big data, we are going to have several such data sets that could be correlated and hence ensuring privacy, while enabling security is going to be one of the grand challenges to mankind. As we access more and more data sets, then privacy keeps losing its place, while anonymity also becomes questionable with security becoming the lone criteria as depicted in Fig. 1. As it stands today, unless new ways of preserving PII are found, we are staring at a world where there is no such thing as privacy.

3 Security

Security incidents happen when risk materializes. Risk is the chance of occurrence of hazard. Corollary is that only when all potential hazards and their chances of occurrence are known will it be possible to provide mitigation to avoid/handle the risk resulting in appropriate security. However, it is often not possible to identify all possible hazards. Secondly, even if certain hazards are known, mitigation may not be possible or feasible. Further, hazards themselves are not constant and set in stone. They keep changing with time and situation. This cannot be more true than in the connected internet world.

A system is considered secure when it satisfies the requirements of CIA. This is not, of course, the only measure. Another popular measure is based on the privacy, authentication, integrity, and non-repudiation also known as PAIN.

Availability and the resultant accessibility are to be taken seriously. Open Security Foundation[1] reports that while 58 % of data loss events involving PII is by outsiders, the rest 42 % is by insiders including 19 % accidentally. As much as 31 % of data loss is attributed to hackers over the last decade; however, in 2013, this went up to 47 %. While these data are based on reported incidents in USA, the rest of the world cannot turn a blind eye.

3.1 Big Data Security in IT

Information system security is commonly referred as IT security the scope of which includes several sectors such as business, finance, and banking. Confidentiality of data is of prime importance here. In case of a pharmaceutical industry, formula of a

[1] http://www.datalossdb.org.

specific drug could be a trade secret. In case of banking sector, customer data is to be protected and several laws enforce such a requirement. Espionage is one of the key threats to security, and elaborate studies have been made on this subject. System-level security policies specifically address this issue. With the advent of big data, as we increase availability, the number of ways in which espionage attempts could be made increases by orders of magnitude. Integrity of information is another aspect to be addressed. Several security mechanisms are in place including hash and signature. Yet another important characteristic in case of transactions is non-repudiation. Non-repudiation is a property that ensures that the owner or sender of the information cannot dispute it later. Even defense information systems have similar requirements in a stricter way. In this case, secrecy is supreme which keeps the adversary guessing and when needed provides the "surprise" element, so often decisive.

So much water has flown in case of IT security that not only we have the best practices, but also we have several standards and certifications such as Federal Information Processing Standards (popularly known as FIPS), ISO/IEC 27001 information security management system, ISO/IEC 15408 (popularly known as "Common Criteria"), Control Objectives for Information and related Technology (COBIT), and Information Technology Information Library (ITIL). However, as big data comes in, these standards need to be upgraded to handle the scale and complexity.

3.2 Big Data Security in Critical Infrastructure

Importance of critical infrastructure has gained traction over the years, and several countries have taken explicit steps to protect them. For our discussion, critical information infrastructure (CII) part of the critical infrastructure is particularly interesting, and big data has important ramifications as one looks to protect CII. Identification and definition of "critical infrastructure" has varied from country to country. Certain sectors such as banking and finance, energy/electricity, transportation, and health services are common across several countries. A handbook by ETH covering CII and its protection was published almost a decade ago [2]. Defense has always been a critical sector, and while some countries protect it discreetly, other countries make it part of critical infrastructure. When it comes to infrastructure, the biggest challenge is not confidentiality but availability. In fact, confidentiality may not be an issue at all. As a case in point, if we take a passenger reservation system, the most important requirement would be availability. Same thing could be said about the availability of a grid supplying electricity.

For long CII was protected by not connecting it to public internet. Where such a connection became essential, air gap and other mechanisms were put in place to address security requirements. Those days are now receding past and more and more CII is accessible from public internet. This has added to the data volume. Yet another important factor is the digitization of data collection. For example, manual metering of electricity usage is giving way to automatic/smart meters. The metering

Table 2 Comparison of IT and CII systems

Criterion	IT system	CII
Priority of security requirement	Confidentiality, integrity, and availability (CIA)	Availability, integrity, and confidentiality (AIC)
Application type	General purpose; can also be used for special purpose	Special purpose
Attack surface	Reasonably well understood; public	Not very well understood; not public
Security awareness	Openly discussed	Opaque

information, instead of being maintained and processed locally, is becoming part of a large data repository and processed in some form of centralized or clustered manner. This enables consolidation, trend analysis, and various types of report generation. At the same time, availability of such information also enables vested interests to play with, for example, pricing based on demand and create artificial shortage or surplus. Hence, securing the big data is a key aspect as we become more "smart" with smart home, smart plant, smart appliance, etc.

In contrast to IT system standards that focused on security, CII standards focused on safety and availability. Commonly referred as industrial control protocols, e.g., IEC 61850[2] and Modbus/TCP,[3] a closer study clearly indicates that safety and timing issues are addressed and traditional security is outside the scope of these protocols. Several industrial control protocols are currently under revision which will address some of the security issues. However, existence of legacy systems and backward compatibility requirements are the challenges to overcome.

Another way of looking at CII is that they are in one way or another cyber-physical systems (CPS). This distinction and understanding is critical from security point of view as a compromise and security incident of an IT system results in information disclosure or financial loss, whereas in case of CII the result could be physical as well as financial. For example, security issues related to implantable medical devices are outlined in [3].

A quick comparison of IT systems and CII systems is depicted in Table 2. From a security point of view, the order of CIA requirements is in the opposite order. As an example, confidentiality is perhaps the most important characteristic in a business tender, while availability takes the top priority in case of a defibrillator during a medical emergency. Integrity remains at the center. Figure 2 provides a pictorial view of confidentiality versus availability across different types of systems. As far as hardware and application types go, IT systems typically prefer general purpose, commercial off the shelf solutions to keep the cost low and avoid vendor lock-in. However, for CII systems, special purpose hardware and industrial-type systems are used that are tailor made for specific applications. For example, the mission computer in a space application and a single board computer in a submarine will have different

[2] http://www.iec.ch/smartgrid/standards/.

[3] http://www.modbus.org/.

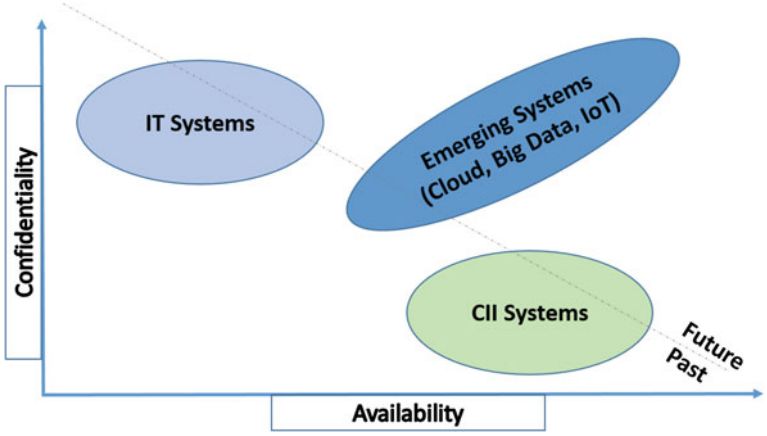

Fig. 2 Generalized view of confidentiality versus availability needs

requirements and are designed specifically based on the use case. General purpose solutions do not fit the bill here. Attack surface and attacker characteristics are reasonably well understood and documented for IT systems. Also, being non-mission critical, public declaration of vulnerability and patching them within a reasonable time frame is an accepted norm. However, for CII systems, due to the strategic nature, attacks are not disclosed. When they do become public, the system will become non-operational till the same gets fixed. For example, if a vulnerability resulting in potential brake failure in an automobile is discovered, then no vehicle owner will drive that type of vehicle till it is fixed to the satisfaction of the user. This also forces the vendor to work behind the doors and come up with a fix to deliver, possibly without even disclosing the real reason for delivering the fix. This restricts pooling in of resources. However, of late just like CERT[4] advisory for general purpose IT systems, ICS-CERT[5] for industrial control systems has become active. Hopefully, the benefit of community effort and pooling together of resources will happen.

3.3 Confidentiality

Confidentiality is the most discussed topic when it comes to data security. It implies placing a set of rules or restrictions with the aim of limiting access to the data in question. Confidentiality is hence achieved through access control mechanisms. While confidentiality relates to any type of data, if the same is associated with PII, then this may be referred as privacy.

[4] http://www.cert.org/.

[5] https://ics-cert.us-cert.gov/.

Traditional security mechanisms have relied based on restricting data access in one or more of the following ways as required:

- Store data in plain text, but use cryptographic techniques during transmission for security. Here, the assumption is that the data in transit is vulnerable.
- Store data in plain text, but some authentication is needed to access. Password protected files fall under this category.
- Store data in an encrypted way and decrypt whenever access is required, often used in portable devices.

The mechanism used to implement confidentiality services is usually referred as authentication, authorization, and access control (AAA). Authentication refers to the establishment of user identity. Authorization refers to determining the resources that the authenticated user can or cannot have access to. Access control is the process of enforcing access permissions to authorized resources while restricting access to other resources. Such an access control could be with additional constraints such as time of the day or IP address from which connection is made.

Use of passwords for authentication has proven to be futile in most cases.[6] Multifactor authentication is recommended but with caveats.[7] For example, compromising point of sale units is highlighted in [4].

With the quantity of data going northwards and redundancy becoming the norm rather than exception, access control is a daunting task. Parts of data are likely to be provided with anonymity, security, privacy as well as public in any repository. The amount of anonymization or sanitation that is needed mandates development of automated and intelligent tools. To address access control during transit, IPv6 has some kind of encryption as part of the protocol as against IPv4 which does not care about it. Of course, IPv4 was designed with the idea enable sharing and IPv6 aims at secure sharing.

Crypto algorithms have become synonymous with confidentiality. The strength of crypto algorithms has relied on key strength and difficulties in brute forcing to break them. However, analysts agree that it is possible to decrypt any encrypted message given sufficient samples. Big data just does that—provide sufficient, perhaps more than sufficient samples.

3.4 Integrity

Integrity implies trust. In terms of security, it means that the data did not get modified and is the same as the time at which it was created/edited by authorized entities. This could be source data or a result of certain calculation or editing in an authorized manner. Integrity also implies consistency of data across multiple

[6] http://www.hongkiat.com/blog/keeping-online-data-safe/.

[7] http://www.asd.gov.au/publications/csocprotect/multi_factor_authentication.htm.

copies. While in paper world, original or authenticated copy is used to confirm integrity of any new copy, in cyber world finding the origin of data itself is a challenge.

Data integrity issues could occur because of one or more of the following:

- Hardware errors,
- Software errors,
- Intrusions, and
- User errors

Preserving data integrity is an issue with storage mechanisms themselves and several tools such as *fsck* utility in Unix operating system. At a file system level, transactional file system could help [5]. They help identify data corruption. If we consider a distributed file system such as the Google File System [6], then it employs a checksum-based technique called chunkserver to detect data corruption. Mirroring, RAID, and checksum are commonly employed techniques to handle data integrity. Several host intrusion detection systems perform integrity verification[8,9] to detect intrusions.

Data loss or theft or breach is another issue that affects integrity. Once copied, it could be tampered and released to create confusion and integrity issues. Data loss prevention (DLP) is a common technique used to avoid data loss [7]. However, it is designed for traditional systems where specific computers implement DLP. As soon as data is accessed from a non-DLP system, the purpose is defeated. In addition to security, data loss often compromises privacy as well.

3.5 Availability

Availability implies the ability to access authorized data as and when required. Availability of critical data is achieved mostly by designing "high availability" (HA) systems. HA systems are designed with backup servers and alternate communication links. Disaster recovery and standby systems address HA needs. With big data and cloud, availability in itself is unlikely to be the issue due to the redundancy of data and multiple access routes. The challenge would be ensuring access control to authorized users and entities. Ensuring controlled access in the world of IoT does look intimidating.

With the emergence of IoT, entities such as sensors and appliances also need access to data. The resource constrained entities such as sensors and point-of-sale terminals being part of the mix, breaking into the network, and achieving privilege escalation to overcome access controls is a distinct possibility. In a way, this complexity is creating challenges for security teams to detect security incidents.

[8] http://www.tripwire.com/.

[9] http://www.la-samhna.de/samhain/.

This can be seen from the fact that in 2012, about 37 % organizations were able to detect intrusions by themselves, while in 2013, it came down to about 33 % [4].

4 Privacy

Privacy is essential in many ways. Personal elements in life such as intimacy, friendship, role play, and creative experimentation need to remain private. However, IT-enabled and IT-networked world of today is posing challenges in maintaining privacy. In data world, privacy typically refers to PII. In several countries, privacy is protected legally, e.g., Health Insurance Portability and Accountability Act (HIPAA 1996 US) and Sarbanes–Oxley Act (SOX 2002 US). Privacy aims to protect information that is considered personal and should not be shared without informed consent. Even when shared, use of PII is often restricted to specific purposes. However, to protect citizens, countries have enacted laws that define constituents of PII. Whenever there is a need to share information which also contains PII, then the PII components are anonymized or sanitized before sharing. For example, it is a very common practice to understand drug effectiveness on patients by studying medical records. While such a study is essential to understand and develop more effective treatment, PII part of the medical record needs to be removed by anonymization to protect individual patients.

Accessing anonymized data is becoming easier by the day as connectivity gets better. Even if accessible, limitations in terms of compute capability and knowledge to process data were challenges, not so long ago. However, it is a thing of past. With free online tools and powerful personal computers, processing capability has reached one and all.

What can big data do to privacy? In another world, not long ago, a person can go for shopping and come back while pretty much remaining anonymous. Today, with online monitoring, one can be traced from the time one leaves his/her home, the route taken and items bought in a store. The security/surveillance system at home records the car leaving the garage. Traffic monitoring cameras record the route. The store card and/or the credit card used to buy the items ensures recording of shopping info. And this is for real! Today we capture so much data that were never captured and stored in the past, which is what big data is doing. What was anonymous is suddenly traceable to individuals and privacy issues crop up.

Big data has impacted privacy so much that the USA is contemplating changes to the "Consumer Privacy Bill of Rights" of 2012 [8] and initiated big data and privacy review in January 2014 [9].

4.1 Online Privacy

Availability of data online takes away the ability of individuals or organizations to decide by themselves sharing of information about them. Personalizing, it could be

re-phrased as my losing control over what, when, how, and how much I would like to share information about me. Any data with PII online is at once a security and privacy issue. While traffic monitoring is a security issue, tracing the route taken by an individual is a privacy issue. Tracing the route and owner of the vehicle may be acceptable from security point of view for crime investigation, the same will be violation of privacy as often is the case, e.g., with paparazzi. As shown in Fig. 1, one can argue that there is no such thing as privacy, particularly once data is online.

What constitutes PII is generally as defined by the law(s), which typically is a list of information types, e.g., first/middle/last name, driving license number, date of birth, and credit/debit card number. The very definition of what constitutes PII permits use of other information which sufficiently enables identifying specific target audience. Value of PII has been understood by businesses, and hence, today data with PII has gained so much traction that a plethora of branded credit/debit cards, loyalty cards, frequent flier programs, and the like have emerged. They do provide some benefit to the customer in terms of discounts, and receiving targeted information. Yet it should be mentioned and noted that several online agencies and service providers sell their products and services by highlighting their ability to target identified users. Anyone browsing internet today can see advertisements popping up that are very specific which is possible only if some uniquely identifying data is available.

Very often, customers are provided with notice as well as terms of use and similar options while installing or using services. The information shared based on the agreement or consent given is for that instance and isolated. However, the moment multiple independently consented data are correlated and analyzed over time, what gets revealed is not what was consented for! Because big data reveals patterns and information that could not be found otherwise. A more interested reader can refer to [10].

4.2 Offline Privacy

Is there an issue with offline data privacy? After all, the data is not online! Let us consider a paper document containing PII. As such, it is perhaps safe to assume that without physical access to the document PII remains protected. Governments own a large number of data containing PII. Laws like the Freedom of Information Act enacted in USA or the Right To Information Act enacted in India enable access to public records. It is therefore reasonable to assume that at least data lying with government agencies offer only limited protection of PII. At the same time, as countries enact one or another form of paper reduction acts as well as physical storage constraints along with environmental issues, paper documents are continuously being replaced by electronic versions and hence making them online. As we move forward, with smart home, smart appliances, surveillance, and monitoring in public and work places, there will be nothing offline. In a nut shell, offline privacy is evaporating fast and sooner or later offline privacy will become a thing of past.

4.3 Privacy Post Archival

Archiving data for historic preservation to legal mandate is not uncommon. The moment archived data is accessed and become "actionable," the question of privacy crops up. Predicting future behavior based on past behavior is a trivial task. Such predictions need observation or access to the past behavior. By analyzing archived data, we are just enabling predicting future behavior, and in a sense identifying "stereotypes." Till few years back, most email providers offered a limited mail storage option to the users. This forced the user to delete emails that are no longer relevant or possible to manage without. However, of late, unlimited storage option is provided by email providers and data will continue to be archived. The traditional life cycle management of data where there is a definite phase to destroy is becoming less relevant. With social networks having become so popular, a childhood prank may make someone a suspect or unsuitable for certain positions at a much later stage in life, due to potential "similar" future behavior due to the past behavior!

The challenges of archival and disposal has been discussed by the National Electronic Commerce Coordinating Council in their report [11].

5 Emerging Trends

5.1 The Story So Far

Emergence of cloud has taken virtualization and elasticity to new heights. Service providers are already offering software as a service (SaaS) to provide on-demand software, platform as a service (PaaS) to provide platforms on demand to build/run applications, and infrastructure as a service (IaaS) to provide on-demand provisioning of resources (e.g., virtual machines). On demand and elasticity are there to use.

The addition of "smart" to everything enabling IoT has catapulted the number of entities that could communicate in the cyberspace while erasing boundary lines across several segments of industry. Smart devices and broadband access have resulted in the amount of data generated to unthinkable limits, and no amount of prediction seems to be close to reality resulting in "big data." Such a rapid change in the infrastructure, communication, and data generation has left system designers grappling for an understanding of potential security and privacy threats. Increase in availability has resulted in potential access to a large number of entities hitherto not part of the design considerations.

5.2 On the Horizon

As we look forward, it is clear that several things are unraveling. Cloud is moving from a monolithic option to offer variety in terms of private, public, community,

and hybrid clouds. Business verticals are expected to use community clouds to leverage on consortia approach. Private cloud would be the choice for strategic areas such as defense. The current offerings will mostly fall into public cloud and will become more of a commodity offering. Hybrid clouds are likely to be adopted by large organizations to keep critical operations on their private cloud and keep the majority on public or community cloud to optimize on the cost.

IoT itself will get better defined with new protocols defined. Variants of IoT such as industrial IoT would emerge. The current internet protocols may get replaced by new ways of networking. Techniques to uniquely identify devices connected to IoT will emerge. The question of whether the streams of data generated by smart devices and sensors need to be stored, and if so, how and where to do are all the questions that will be debated for possible solutions. Replacing compromised and/or defunct sensors will be major challenge that will be addressed by researchers and industry.

Big data requires big time crunching. We see plans announced by the National Security Agency (NSA) that their Utah Data Center has plans to have exaflop (1018 bytes), zettaflop (1021 bytes), and yottaflop (1024 bytes) by 2018, 2021 and 2014, respectively [12]. The compute capabilities will not only enable analytics, but also challenge strength of crypto algorithms. We can expect novel crypto or even new technologies from unchartered territories may emerge.

5.3 Research Challenges

Characteristics of big data and the possibility of global access to such data have not been understood sufficiently. Ability to mine patterns from autonomous sources has resulted in privacy issues. Anonymization of individual data sets is proving to be insufficient. Key research challenges include:

- Understanding characteristics of large volumes of data.
- Understanding characteristics of large and diverse data sets.
- Identifying duplicate data sets.
- Security requirements of big data.
- Access control mechanisms when availability is difficult to control.
- Keeping integrity of multiple copies of data.
- Challenges in maintaining anonymity of data.
- Anonymizing data containing PII.
- Designing of crypto algorithms whose strength does not depend on sample size.
- Identifying best practices for security and safety, including new type of data sources and sets.
- Real-time monitoring with encrypted streaming data.
- Achieving reliable computations in distributed multi-core environments.
- Reducing the information overload for security and information managers.

- Privacy preserving data analytics.
- Audit and compliance verification of information systems.

The list of challenges above is not exhaustive but indicative. We can look forward to researchers trying to achieve technical solutions to these issues. Government and standards bodies will be fully occupied with defining policies, drafting laws and standards, identifying best practices, and promoting their adoption. Academics will define new curricula. Businesses will need to re-orient and redefine strategies as they get access to more data, while their own data gets shared faster and wider.

5.4 Summary

In this chapter, we covered several security-related topics in a concise manner. We started with a discussion on cloud, IoT, and big data as well their effect on security. This was followed by a short discussion on security and privacy as well as the effectiveness or otherwise of anonymization with the emergence of big data. Security was then addressed in some detail. In particular, the contrasting requirements on confidentiality and availability with the convergence of IT and CII was highlighted. We briefly covered CIA aspects as well. Privacy, being an issue that affects everyone, was discussed. Going further emerging trends including a partial list of challenges were highlighted.

Even though we do hear pessimistic view like there is no such thing as privacy in an online world, it may be worth remembering and recalling that mankind has seen much tougher challenges and it has risen to the occasion each time and has come up with a solution to each challenge. This time also we can trust our scientists and engineers to come up with workable solutions to the issues at hand.

Questions

1. List the reasons stating why big data security concern is different than that of conventional data.
2. Explain the association between cloud technology and big data and state the security challenge the association brings in.
3. Discuss on trade-off between security and anonymity of big data.
4. Draw a comparison between IT and CII systems on security issue.
5. List the mechanisms used to assure data confidentiality and illustrate on each.
6. Why big data could be vulnerable for confidentiality?
7. Present a scenario how big data can invade one's privacy.

References

1. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002)
2. Dunn, M., Wigert, I.: Critical Information Infrastructure Protection. https://www.emsec.rub.de/media/crypto/attachments/files/2011/03/ciip_handbook_2004_ethz.pdf (2004)
3. Gollakota, S., Hassanieh, H., Ransford, B., Katabi, D., Fu, K.: They can hear your heartbeats: non-invasive security for implanted medical devices. In: Proceedings of ACM SIGCOMM, Aug 2011
4. Threat Report: M-Trends beyond the breach. https://www.mandiant.com/resources/mandiant-reports/ (2014)
5. Gal, E., Toledo, S.: A transactional flash file system for microcontrollers. In: Usenix '05: Proceedings of the Usenix Annual Technical Conference, pp 89–104. Usenix, Anaheim, CA, USA (2005)
6. Ghemawat, S., Gobioff, H., Leung, S.T.: The Google file system. In: Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP '03), pp 29–43. Bolton Landing, NY, Oct 2003
7. Kanagasingham, P.: Data Loss Prevention. SANS Institute. Aug 2008
8. The White House Washington: Consumer data privacy in a networked world: a framework for protecting privacy and promoting innovation in the global economy. http://www.whitehouse.gov/sites/default/files/privacy-final.pdf (2012)
9. Podesta, J., Pritzker, P., Moniz, E.J., Holdren, J., Zients, J.: Big data: seizing opportunities, preserving values. http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf (2014)
10. Sloan, R.H., Warner, R.: Unauthorized Access: The Crisis in Online Privacy and Security. CRC Press, Boca Raton (2013). ISBN:978-1-4398-3013-0
11. NECCC: Challenges in managing records in the 21st century. National Electronic Commerce Coordinating Council 2004. https://library.osu.edu/assets/Uploads/RecordsManagement/Challenges-in-21st-e-recs-neccc.pdf (2004)
12. Bamford, J.: The NSA is building the country's biggest spy center (watch what you say). WIRED. http://www.wired.com/2012/03/ff_nsadatacenter (2012)