Big Data Analytics and Security: A Big Choice and Challenge for the Generation

Gebremichael Girmay and D. Lalitha Bhaskari

Abstract The concept of big data comes into concern when challenges have been identified in digital capacity, velocity, and type of the data gathered. The domain of data source is diverse: social networking, mobile phones, sensors, satellites, and different types of organizations. The data is collected/generated at high rate, and its type is complex (structured, unstructured, and semi-structured). Big data analytics, which is the process to reveal hidden patterns and secret correlations that is to predict and capture insights from this overwhelming data, forced standard technologies to be upgraded or replaced by big data technology, and has a multitude benefits and applications. In this paper, the benefits and particularly the challenges of privacy and security in big data are dealt and solutions are suggested. Parallel to capturing insights from such massive and mixed data which also is sourced from huge varieties of IoT devices, the principles of big data pose advanced security solutions.

1 Introduction

As our world is now in a new age of information existence, huge aspects or matters necessary to our life are becoming dependable on automated computing technologies. Particularly private and government organizations, institutions, and enterprises of all categories are highly characterized by whether they are equipped and integrated with information computing technology. Almost for most of these organizations, the pending issue within the current and coming decade(s) is that of storing huge size (or binary voluminous) of data to be collected, the mix type of

G. Girmay (✉) · D.L. Bhaskari
Department of Computer Science & Systems Engineering, AUCE(A),
Andhra University, Visakhapatnam, India
e-mail: micgirmay@gmail.com

D.L. Bhaskari
e-mail: lalithabhaskari@yahoo.co.in

S.C. Satapathy et al. (eds.), *Smart Computing and Informatics*, Smart Innovation,
Systems and Technologies 78, https://doi.org/10.1007/978-981-10-5547-8_22

data format sourced from countless IT devices, locations and vendors, the throughput or rate at which these data are piped, analyzed and gain valuable insights-all together these and other concerns are coined by the engulfing term *Big Data*.

Big data can lend new benefits but also raises issues with respect to management, technological setup, and analytical process, privacy, and security threats. In other words, new brands of algorithms and programs have to be designed and developed to handle big data, as well a firm policy and plan in compliance with this new era is required to protect and safeguard all sort of data processing.

A number of global and local enterprises have already envisioned and implemented big data setup. Apache Hadoop, MapReduce, and others are the prominent platforms available for performing big data analytics that apply parallelism and distributed algorithm on clusters or nodes.

The essential works done and being tackled (by companies, researchers, IT stakeholders, throughout the world) under the big data era are to mention some, big data analytics, big data security analytics, big data privacy, big data forensics.

2 Big Data and Big Data Analytics

Big Data—refers to the voluminous data sets that are too complex to manage and process using the existing or regular database management tools.

As articulated by industry analyst Doug Laney, big data spans three dimensions: Volume, Velocity, and Variety.

Volume—The size of data is excitingly very large and will continue to increase enormously, measured as terabytes, petabytes, exabytes, zettabytes, and so on. Some of the issues emerging include how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

Variety—Data collected is of different types of formats, i.e., structured (such as numeric data in traditional databases), unstructured (e.g., Videos), and semi-structured. The domain of data source is diverse: Internet, social media, business organizations, sensors and mobile phones, space, health centers, defence, weather forecasting, etc. Managing, merging, and governing different varieties of data are something many organizations still tied with.

Velocity—Data is streaming in at unprecedented speed and must be dealt with in a timely manner. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

Figure 1 generalizes the meaning of the three Vs. However, the 3Vs is replaced to more number of Vs, in some research papers and articles, example 4Vs, 5Vs, even up to 11Vs. Some of these terms are veracity, value, variability, verification, validity, volatility, and visibility. The orders of priority for these terms also vary among the articulators who want to escalate the Vs. What we can generalize from this is that big data is really a manifold issue and that needs a more automated and customized solution.

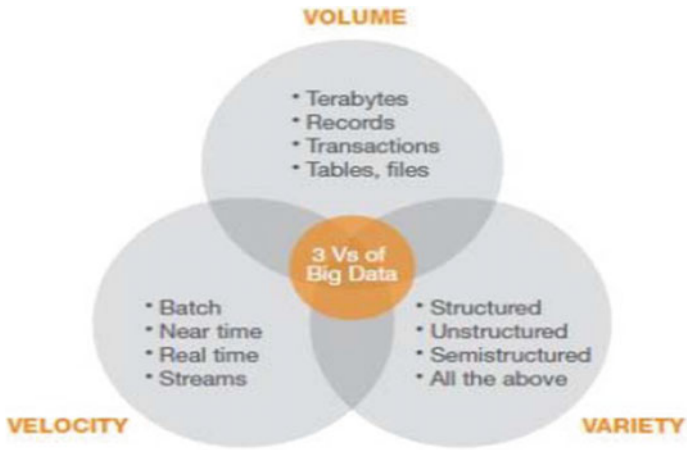

Fig. 1 The three Vs of Big Data [1]

Big data analytics—is the process of collecting, arranging, and analyzing huge data sets to reveal hidden patterns, correlations and other helpful information and deep insights that is important to take enhanced decisions. As the concept of big data has been around for years; many organizations now understand that if they capture all the data that streams into their businesses, they can apply analytics and get significant value from it [2]. Using big data algorithms which lend high-performance data mining, predictive analytics, and optimization enables to continuously derive innovation and make the best possible decisions.

That is, implementing big data analytics companies can make more informed business decisions that cannot be achieved solely by conventional business intelligence (BI) processing architectures and tools. With big data analytics, data scientists and others can analyze not only huge volumes but also complicated set of data that conventional analytics solutions cannot touch. That is the data size is now in exabytes or zettabytes (1ZB = 10^{21} bytes = 10^{3} exabytes = 10^{6} petabytes = 10^{9} terabytes = 10^{12}gigabytes), which is the mix of structured and unstructured data type, and as mentioned in the previous section, the domain of data source is diverse and includes transaction data, web server logs and Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things.

The processing pipeline can be organized into the following phases [3]:

- Data Acquisition and Recording
- Information Extraction and Cleaning
- Data Integration, Aggregation, and Representation
- Query Processing, Data Modeling, and Analysis
- Interpretation.

3 The Potentials and Challenges of Big Data

Big data has a number of potential benefits. In the first place, with the right setup of analytics platforms, an enterprise can collect a large enough variety of data at expected slot of time. The big data analytics platform itself endows scalability, fault-tolerant, high speed, better efficiency, reduced latency, by merely using moderately computing devices in the mappers and reducers nodes, for example.

Then comes the predictive power owed by big data analytics, example weather forecasting, healthcare, retail, smarts cities, future business decisions to earn high profits by examining their customer dreams, and more advancements. In general, the analysis of big data is often valuable to companies and to consumers, as it can guide the development of new products and services, predict the preferences of individuals, help tailor services and opportunities, and guide individualized marketing. Due to these and other benefits, the world is really forced to conceive and implement big data analytics.

The benefits of big data themselves are either challenges or result in challenges. Broadly, the challenges in big data environment include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy and security, and the architectures, methods and platforms that support for these scenarios.

For most organizations, big data analysis is a challenge. Consider a large company which is well established with the pre-big data technologies and platforms such as the conventional Business Intelligence (BI). Here, on behalf of the company, one can pose several challenges with regard to the transition to big data environment. Can this company resolve this challenge by reshuffling of the resources (including professional workers), or should the data warehouse be replaced with big data analytics tools and established from scratch? For sure additional investment is required, keeping in mind necessity of huge storage, the complexity of data types (structured and unstructured), applications and operations used to find patterns and deep insights through the analytics; that is big data requires high-performance analytics and thus the benefits be gained.

As outlined in [4], three alternatives of implementation are possible to migrate from the current conventional platform to big data platform. These alternatives are: *Revolutionary*, *Evolutionary*, and *Hybrid*.

The challenges regarding big data privacy and security are the main topic of this paper and are discussed in subsequent sections.

4 Big Data and the IoT

Big Data and IoT are two crucially influenced digital techno-analytical eras. The Internet of Things (IoT) is defined in [5] as "IoT is a term used to describe the ability of devices to communicate with each other using embedded sensors that are linked through wired and wireless networks". Therefore, within IoTs there are

thousands of types of sensors that act as a source of data by sensing and colleting from their intended function and transmit through the Internet networking system. These devices include everything from computers, smartphones, tablets, headphones, wearable devices, car, digital cameras, and almost anything else we can think of that can embed sensor(s). These devices are to be connected and use the Internet to transmit, compile, and analyze data. The IoT applications are tremendous, for example in transportation, retail, agriculture, weathercasts, security, and so forth.

Data generated from the IoT will grow exponentially as the number of connected nodes increases. Gartner predicts that the number of "connected things" will reach 25 billion by 2020 [6].

So, what is the relation between big data and IoT? Yes as discussed in previous sections, big data has the capability of storing the data at rest and the data at motion which are sourced from IoT as well. At the same time it is the task of big data analytics to dig out the hidden insights from the real values, and use it for better business, commercial or political decisions. Therefore, IoT and big data are two critically related projects, as if there are two sides of the same coin.

5 Privacy and Security Challenges with Big Data

Privacy is considered a purely legal issue [7, 8]. Security is the process of actions to make practical the privacy laws and legislations planed to protect user's data and information.

When we focus on privacy and security of data or big data in general, there are many issues to be raised. For example

- how an organization should secure the big data infrastructure (hardware and software),
- how the data is collected, accessed and used,
- which of the data is sensitive,
- what is the users/customers perception on the data or information they exchange,
- who is the responsible to handle these and other issues, etc.?

Security and privacy of potentially sensitive data, information, and IT infrastructure as a whole is always a challenging concern in all level and type of IT and related enterprise systems. The issues are worsen specially as enterprises move into the big data environment; this is because the architecture and platform (e.g. open source framework with distributed data processing across clustered computer environment as in HDFS) used, the type of data (structured, unstructured and hybrid), the millions or billions of data source (devices, vendors, customers), and the rate at which these data are flooding in and out, can result in several forms of vulnerability, one of which is the security and privacy vulnerability.

It is difficult enough to judge whether the data or information shared among millions of IT users is originated and uploaded for free or whether it is abused for monetization or moral satisfaction or dissatisfaction. To more elevate the issues regarding privacy, data from overwhelming sources are fastly circulating our universe, that is if a sensitive personal data is seized or breached in one location, there is a great chance of dissemination of that item and appearing in almost all locations of the world as instantly in real time or near-real time, and hence it is essential to locate and identify the real location of privacy breach to regulate the multi-source and multi-branch fountain of illegal acting, hence guarding the big data environment.

The big question here for researchers, legislatures, regulators is how to provide optimistic guarantee, for single or group of users, for their data privacy, of being secluded from the presence or view of others. As an example of privacy challenges, as referenced in [9], "the dilemma facing regulators is how they can regulate the collection, storage and trading of personal data on the internet, when all of these activities, and the corporations themselves, operate across multiple continents and jurisdictions". This and other phenomenon can force us to abstract that privacy and security breach in this new era of data management is not a simple issue, rather it potentially affects a much larger number of people in the globe which extends to economical, political and cultural intricateness as well.

One thing, organizations and other stakeholders of this issue, have to accept by default is that security and privacy threats are advancing in parallel to the current advancement in big data technological and operational management. The world of IT is now in a period of advanced persistent threats (APTs), or intelligence threat and thus organizations must take advantage of new technologies to protect the whole big data platform.

5.1 General Approach for Big Data Privacy and Security Solutions

As part towards the solution, the research designs, legislations, and regulations regarding privacy in big data environment have to be outlined, structured and implemented at several levels(for their inevitable reasons and backgrounds, in varity of orders), at

- user/individual level,
- institutional/organization level, and
- global level.

The vendors of data are now a day not only the licensed organizations, in other words individual users are not only consumers but also producers of huge amount of variety of data formats and transmitters of data at high speed automated IOTs. Some of the privacy issues to put as a burden on individuals may be summarized as:

awareness—about multitude types and ways of data stealers/disturbers; **trust**—to use or not to use network systems/apps; and **responsibility**—no to blame others on careless handling personal data. These and the other privacy techniques used at personal and organization level helps to avoid the collection of *personally identifiable information*.

Organizations share huge amount of data, information, technology and so forth, locally/globally, for the sake of enhanced business operations, for example. Hence the existence of one organization is critically dependent on other ones when we consider IT security in general, because if there is a vulnerable infrastructure in either of these parts then the security breach will be reflected in one or the other way in some or all the cooperating systems. Therefore it is a mandatory to have common/global IT security resolutions, regulations, legislations, and frameworks of transparency and guidelines.

Since big data is a relatively new concept; it is hard to say there is a list of best practices that are widely recognized by the security community. Of course, the principles of big data are the key to advanced security intelligence [10]. Anyway, several concerned intellectuals and practitioners have listed out a number of general security considerations and recommendations that can be applied to big data.

Methodology (**approaches**)—As a best overall practice, it will be helpful to sketch and clarify the roadmap of tackling big data threats. First, the *Big Data asset taxonomy* and the *Big Data treats taxonomy* have to be identified, then, followed by mapping the threats to big data asset taxonomy. Big data assets taxonomy can be categorized as data, infrastructure, analytics, security and privacy techniques, and roles. Similarly, big data threats taxonomy can be classified as organizational, legal, eavesdropping, nefarious, and unintentional damage, etc.

The security and privacy techniques, i.e., security-related assets have to be outlined, example as in [11, 12], which are categorized into four aspects of big data ecosystem, as *infrastructure security*, *data privacy*, *data management*, and *integrity and reactive security*. These can be further break down into several levels of aspects. This, of course, summarizes the challenges pointed out in preceding paragraphs. As threat asset taxonomy is vast one can select a specific weak point (example, intrusion detection) and provide intended solution with the help of taxonomy roadmap, algorithms, technologies and applications, that apply to big data platform.

Some of the basic big data security consideration includes

- Data anonymous,
- Data encryption,
- Cryptographically enforced access control and monitoring,
- Policy and compliance, and
- Governance frameworks.

Of course not all data is same and not all coming in data is to be stored. Sensitive data/information has to be anonymized or encrypted, compressed and stored in a

way of disparate. Authentications and authorizations are to be more efficient that may include voice, image or some kind of patterns taking into account scalability, time (low and high latency), according to the big data ecosystem.

As a matter of consequence, knowledge of machine learning, AI, advanced cryptography, etc., is highly required to support effective analytical models that identify potentially fraudulent transactions, identity theft, and malwares in general. This will help to look for solutions to improve or change the security posture from a reactive to a predictive model.

5.2 *Big Data Security Technologies and Applications in Use*

Of course dozens of security packages or tools have been in test and introduced as practical application this day, though they are not full-fledged security solutions towards the new era of security intelligence of data acquisition and analysis. Some of the leading big data security analytics tool vendors include Cybereason, IBM, RSA and Splunk. For example, Splunk/ELK, HP ArcSight feature (called Correlation Optimized Retention and Retrieval (CORR) Engine) serve as a foundation for threat detection, security analysis, and log data management as pointed out in [13].

Some of the existing or conventional applications for handling security include security incident event management (SIEM), and intrusion detection system (IDS). These may not compete for all type of data analysis capability (batch, near-time, and real time) within large business organization. They basically target on real threats and vulnerabilities in small and medium organizations.

For instance, the target of traditional SIEM is mainly for batch systems. SIEM tools basically accomplish their intended security function based on centralized collected security log data from different type of security controls, OS and other software used by an enterprise. That is their basic function is to collect, analyze, and report security breach issues based on logs and events, though some of the SIEM products may have the task of stopping attacks on the way. These SIEM are blamed for important big data issues such as scalability, incident response, though some recent SIEM tools can comply for that.

Big data security analytics (BDSA) is a specialized application of the more general concept of big data. Some of the key features that distinguish big data security analytics from other information security domains include scalability, reporting and visualization, persistent big data storage, information context, breadth of functions [14]. Real time is the newest business of concern. That is big data security analytics tools are designed to collect, integrate, and analyze large volumes of data based on context and correlation fashion, in near-real time, which requires several additional advancements.

6 Conclusion

In this paper, the definition of big data has been explained based on the 3Vs dimension followed by the benefits acquired from big data analytics. Accordingly the world is migrating towards big data environment due to the all round advancement in IoT, data size, data format, data type, latency, and the taxonomy. Though big data environment is a must to adapt, several challenges are facing researchers, data scientists, and organizations. Among these big data challenges are the big data privacy and security issues. Some tools, methods, and procedures are being in use already, though yet to reach the optimistic satisfaction. In general, tackling privacy and security issues in big data is a mandate of data generators, reservoirs, disseminators, and users as a whole. This issue may not as simple as in the traditional data and information management. This is because the multi-dimensional taxonomy of big data environment necessitates implementation of this concern at several levels of the big data ecosystem and using advanced and appropriate securing algorithms.

References

1. Big Data Analytics, https://www.researchgate.net/profile/Sachchidanand_Singh3/publication/261456942_Big_Data_analytics.
2. Big Data Analytics: What it is and why it matters, http://www.sas.com/en_us/insights/analytics/big-data-analytics.html.
3. Big Data Whitepaper: Challenges and Opportunities with Big Data, www.purdue.edu/…/assets/pdfs/BigDataWhitePaper<….
4. Dr. Arvind Sathi: Big Data Analytics: Disruptive Technologies for Changing the Game. First Edition, MC Press Online, LLC (2012).
5. Big Data: Seizing Opportunities, Preserving Values.
6. Driving Real-Time Insight: The Convergence of Big Data and the Internet of Things. Oracle White Paper (July 2016).
7. Rebecca Herold: What Is The Difference between Security and Privacy? (2002).
8. Wikipedia: Information Privacy, https://en.wikipedia.org/wiki/Information_privacy.
9. The Guardian: Little Privacy in the Age of Big Data, https://www.theguardian.com/technology/2014/jun/20/little-privacy-in-the-age-of-big-data.
10. A Randy Franklin Smith Whitepaper: Top 5 Truths about Big Data Hype and Security Intelligence.
11. José Moura, Carlos Serrão: Security and Privacy Issues of Big Data.
12. Cloud Security Allniance: Top Ten Big Data Security and privacy Challenges. (2012).
13. Slashdotmedia: 10 Ways to Build a Better Big Data Security Strategy. IT Manager's Journal (January 2014).
14. Dan Sullivan: Introduction to Big Data Security Analytics in the Enterprise.