A Review on Big data Lifecycle and Security Challenges

Parminder Kaur
Research Scholar, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana
Email: parminder.kaur.pam@gmail.com

Abstract
Big data is term referring to the process of managing huge amount of data from different sources, such as, DBMS, log files, postings of social media. The big data passes through different developmental stages order to become useful knowledge –life cycle of big data. The big data technology also raises many security concerns and challenges. In this paper, the focus is on big data lifecycle and the various possible security threats and attacks.

Keywords: Big data, data acquisition, data storage, knowledge generation, lifecycle, phishing, spamming, spoofing.

1. Introduction

Before big data, databases were used as a processing platform because the data generated was small and data was conventional. But when database is faced non-relational or large-scale, complex data which is primarily heterogeneous, it gave rise to difficulty dealing with them. Big data not only augment the related computing services technologies but at the same time change the traditional mode of many industries.

Big data is the latest word in the IT industry, it is followed by data warehouse, data analysis and data mining.

Big data helps in acquiring knowledge from the massive, complex data. IBM, Amazon, Microsoft and other companies are continuously committed to develop and efficiently utilize the big data, triggering the development boom of big data.

Every second, Big Data gets bigger. Alone social media generates endless streams of data, flowing in from Facebook, Twitter, and other social sites like never before.

Facebook is social networking site, having 1.59 billions active user monthly, 1.038 billions active user daily, 300 petabytes of user data is stored by it. Facebook has a treasure reservoir for us to survey the human society, social events, disasters, public affairs, and so on, only if we have the power to control the huge amount of data. This is perfect example for Big Data processing, as the information comes from multiple, heterogeneous sources.

2. Characteristics of big data

The various characteristics of big data can be represented as:

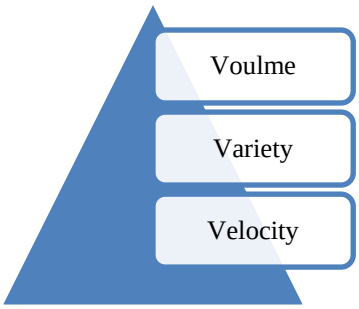

Figure 1 Characteristics of Big data

2.1 Volume

The volume of the data determines the value and prospective uses of the data under consideration, and whether it can actually be considered big data or not. Social media plays a important role: Twitter generates more than 7 terabytes of data every day.

2.2 Variety

Big data is both structured and unstructured so it is not always easy to put them in relational databases. Data analyst must be aware about the type of content. Most of the s information we have today is unstructured and it is usually too big to manage well.

2.3 Velocity

Velocity refers to the speed at which data is generated and it moves around. For example the social media posts get viral within seconds.

3. Big data life cycle

Before concluding or getting knowledge from the data generated from million and billion sources today, data needs to processed and cleaned so that only useful information comes out. This task is achieved by dividing the tasks into various phases of the big data life cycle.

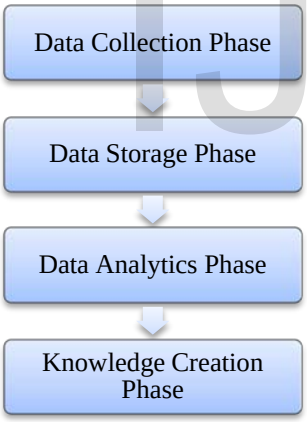

Figure 2 Phases of Big Data Lifecycle

3.1 Data Collection Phase

In data collection phase, data from innumerable sources and in different formats: structured, semi-structured, and unstructured, is captured. Earlier, data was acquired mainly using standalone applications built around the database, or was protected in spreadsheet documents. SQL made it feasible to abstract away the direct connection (somewhat). If data is to be moved from one system to another, a dedicated ETL tool (Extract, Transform and Load) was used. These tools alter the contents of one database to a script, converting it to a different table format if required, then making some basic decisions about data.

3.2 Data Storage Phase

In data storage phase, the gathered data is stored and prepared for being used in the next phase. Cloud environments have also provided alternative environments to SQL environments, with Microsoft Azure's DocumentDB and Google Cloud Store all providing scalable alternatives to relational databases, typically running in parallel with SQL-oriented data stores. Hadoop's map reduce algorithm made the management of data easier task.

3.3 Data Analytics Phase

After gathering data and storing it in secured storage solutions, data processing analysis is performed to generate useful knowledge. In this phase, data mining methods-clustering, classification, and association rule mining are used over the data.

3.4 Knowledge Creation Phase

At last, the analytics phase comes up with new information and prized knowledge to be used by decision makers. Knowledge created can be used by the business leaders and managers to establish a baseline of "business rules". In other words good governance can be established by making an efficient use of the knowledge generated.

4. Security Challenges and Possible attacks

There could be various threats to the security of the data during different phases. Most common threats and attacks are:

4.1 Spoofing

Spoofing is a false or malicious practice in which communication is sent from an unknown source masked as a source known to the receiver.

4.2 Phishing

Phishing is the efforts to acquire insightful information such as usernames, passwords, and

credit card or debit card details (and sometimes, indirectly, money), often for malicious reasons, by imitating as a trustworthy entity in an electronic communication.

4.3 Spamming

Spamming is sending unsolicited e-mail, often of a commercial nature, sent indiscriminately to multiple mailing lists, individuals, or newsgroups; junk e-mail.

4.4 Physical Protection of data storage location

The actual location where the data is stored is also vulnerable to attacks and intrusion by unauthorized people.

4.5 Breach of Privacy

The privacy of the people should not be breached and all the cyber ethics and laws should be followed. No such mining technique should be used which violates the legal system.

5. Conclusion

There is a flood of data on the internet, data is generated each and every second, the data generated is not completely useless or useful, but by using efficient tools and analysis models the “Big data” can be filtered and only useful data can be used by the various fraternities - banking, education, government to enhance their respective jobs and also satisfy the needs and expectations of the people.

References

[1] R. K. N. W. Yazan Alshboul, "Big Data LifeCycle: Threats and Security Model," in *Twenty-first Americas Conference on Information Systems*, Puerto Rico, 2015.

[2] B. Purcell, "The emergence of “big data” technology and analytics," *Journal of Technology Research.*

[3] K. P. Prashant Kumar, "Big Data and Distributed Data Mining: An Example of Future Networks," *International Journal of Advance Research and Innovation*, pp. 36-39, 2013.

[4] M. K. Shilpa, "Big Data Visualization tool with Advancement of Challenges," *International Journal of Advanced Research in Computer Science and Software Engineering*, pp. 665-668, 2014.