MANTECH Publications

Recent Trends in Computer Science and Software Technology
Volume 8, Issue 3, September-December 2023

Navigating the Data Deluge: Unraveling the Architecture, Applications, and Security Landscape of Big Data

Arayan Gupta[1], Puneet Gola[2]
Student[1], Professor[2]
Department of CSE
Aligarh Polytechnic Institute (API)
Corresponding Author's Email id: *arayangupta453@gmail.com[1]*

Abstract

In the era of information explosion, Big Data has emerged as a pivotal paradigm revolutionizing the way data is collected, processed, and utilized. This paper explores the multifaceted dimensions of Big Data, focusing on its architecture, diverse applications, and the paramount concern of security. The architecture section delves into the intricate frameworks and technologies that enable the storage, processing, and analysis of vast datasets. Highlighting the transformative power of Big Data, the applications section showcases its versatile role across various domains, including healthcare, finance, e-commerce, and more. However, as the prominence of Big Data grows, so does the need for robust security measures. The security aspect scrutinizes the challenges and strategies associated with safeguarding sensitive information within the Big Data ecosystem. This paper aims to provide a comprehensive overview of Big Data, shedding light on its architecture, applications, and the imperative need for security measures in the ever-evolving digital landscape.

Keywords: *Big Data, Architecture, Applications, Security, Data Analytics, Information Technology, Data Storage, Cybersecurity, Data Privacy, Machine Learning.*

INTRODUCTION

Big Data are set of large data. Big data contains high volume of small data sets. Big data are high velocity and high variety information assets. New forms of processing are required for big data. Using this database system we can analyze complicated data very easily[2]. Complicated analysis of any data may make a small data to appear as big data. Any data that exceeds current capabilities of data processing can be considered as big data. Velocity, volume and variety characteristics can be viewed as follow[3][6].

Traditional data processing applications are inadequate to deal with big data. Traditional applications need to deal with some problems or challenges like data capturing, data analysis, search, sharing data, storage, querying and privacy protection. Big data can be simply used to refer to advanced data analytics methods to extract value from data. Indeed large size of data is not most relevant characteristics of this new data system. Data sets are rapidly growing in the field of internet and informatics. Many authorities are finding difficulties with large data-sets in areas like internet and urban informatics. Now a day's data are gathered by cheap and mobile information sensing devices like wireless sensor networks, RFID, microphones and such number of devices[4].

Big data provides system with data-driven decision-making facility for databases. With the creation of value from Big data, various problems are dealt and are more progressive with each pipeline step. Such problems may include Scalability, Complexity, timeliness, Heterogeneity and privacy for data. Here are some reasons for what we need Big Data[5].

1. 80% data in world is unstructured
2. Rate of growth for unstructured data is 15 times the structured data
3. Raw Computational power is growing at enormous rate that today's computers are showing the processing power that a supercomputers showed half a decade ago.
4. Democratization to information (accessible to all)
5. The volume of data being stored today is increasing exponentially. Eg: Twitter generates more than 7TB data every day, Facebook 10TB and some enterprises generate some terabytes of data every hour.

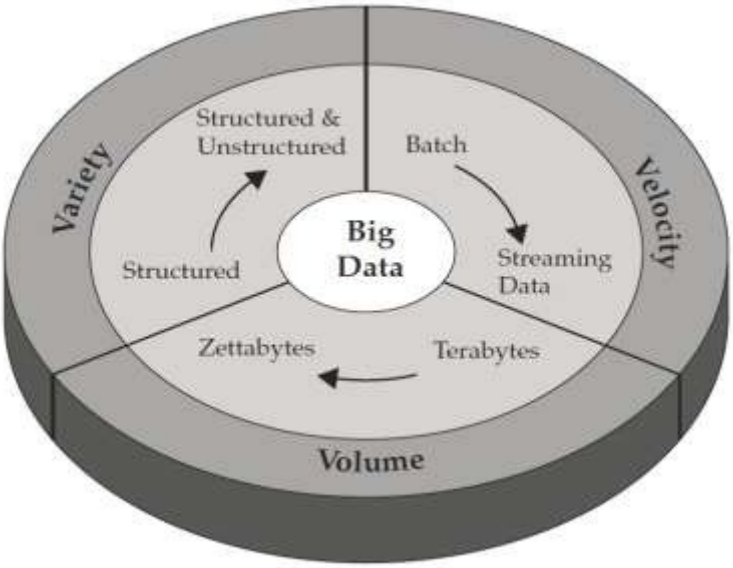

Fig 1 Characteristics of Big Data

ABOUT BIG DATA

When we think about it we find that, we can track and record something. Apart from this we can talk about analysis of stored data. This analysis is going to become theme of Big Data. We store every data like environment data, financial data, medical data, sensor data, surveillance data, survey data, social data, and the list goes on and on. For example – opening commuter train's doors, check in for a plane, purchasing a song on iTunes, TV channel package selection, selecting or changing a TV channel, accessing stored data, accessing the software or tools online; everything of this generates data[6].

As implied by term "Big-data", many organizations are facing massive volume of data. To find useful information to organization from massive amount of data, we need to analyze complete data. Mastery is required to analyze and to find useful information from unstructured data on web. This data is in the form of text, images, videos, social media posts and files. To analyze this unstructured, semi-structured or structured data, one need advance tools and systems. Conventional system is not sufficient to perform data analysis in Big-Data. Big data comes with a lot of opportunities to deal with many areas like health, education, and business, but as conventional methods or system are not sufficient to deal with. Here we can discuss some Big

data challenges and can overview some computational models for efficient analysis of data[7][8][9].

A. Big Challenges [10]

i. Scale: As the name suggests, Big Data is a set of large volume of data sets. To handle such a huge amount of data is very tough task. Earlier it was managed by designing faster processors. But now days we are in the era of cloud technology development for sharing large amount of data and applications. Development of processors is done with the speed as that of data is produced. Hard disks are used to store this high volume of data, but hard disk are slow I/O devices. To replace hard disk we can use solid state drives and these devices can perform to some extend with the speed of analysis of data.

ii. Heterogeneity and Incompleteness: To analyze any data we need data in single format and that too in structured format. As we deal with Big Data, we find data in structured and unstructured format as well. Heterogeneity is Big challenge for analyst. For example, maintaining information about Employees and customers. Customers might have purchased different products of company. Each customer information contents will be different as may be the case. In case of employees, data values will be different. This storage design is not properly structured. So analysts need to manage both heterogeneity and incompleteness of data.

iii. Privacy: Privacy is another Big challenge in Big Data. In some countries there are strict laws for user data privacy, where as in some countries it is not. For example, in USA there are strict laws that prevent open access to health related information, but in case of others it is less. To perform sentimental analysis, we cannot get private posts of users from social media.

iv. Human Collaborations: With advances in technologies, still it is not possible to detect some patterns with the help of computer. Crowd-sourcing is new method of harnessing human-ingenuity. Ex-Wikipedia, blogs, etc. We need to depend on information given by strangers, which is most times correct. But some people intentionally provide false information. We need a method or system to deal with this information. We need systems that should be intelligent to take decision.

v. Timeliness: Another big challenge is speed of analysis. As the amount of data is huge, time required to analyze is very high. System that deals with huge size should be able to deal effectively in terms of speed. Eg: System to detect fraud transaction. Before completion of transaction, any fraud transaction should be detected and dealt with.

B. Data Mining Technique

Big data is useful for many organizations, researchers and Government. Getting important information from big data is Data Mining. There are many data mining techniques that can be used with Big data, some of them are [11][12]:

1. Classification Analysis
2. Cluster Analysis
3. Evolution Analysis
4. Outlier Analysis

HADOOP is open source software used for Big Data processing. HADOOP is used by various organizations and researchers. Distributed Computing Environment is basis for large data processing in HADOOP. Data processing is the result of various techniques like Google File system, MapReduce, Hadoop Kernel, HDFS, Apache Hive, Base and other components.

APPLICATIONS

As like new technologies, big data brought cost reduction, substantial improvements in processing, or new products or offerings for customers, sharing features, and privacy. Some of the applications of Big Data are discussed here:

A. Privacy Preserving for Smart Home Data:

Main objective of privacy preserving is to ensure that private data is protected always. Data protection is provided during processing or data release time. A framework is required for security and privacy maintenance for smart homes. The framework should not compromise data utility for security and privacy preserving. Numbers of sensors are embedded in smart homes to identification of events. Each event generates different type of data and that data is then used for processing. Each Data need to be encrypted by some method before sending to user or to central

system outside the home. This encryption is used for privacy preserving. Certain results can be made identifiable for limited users who directly take care of home.

Number of researchers had worked on this phenomenon and had achieved some results. Various concerns about privacy of smart home data have been raised in various literatures [13][14][15]. Many technical solutions have been designed by number of researchers for preserving privacy throughout data lifecycle. S. Moncrieff et. al. [16] proposes a solution, in which dynamic privacy level alteration method is used. Based on environment context an intrusive nature of technology is decreased without affecting original functionality. S. Meyer, et. al. [17] suggests a interactive system which disclose selected information. This system works on context-aware module.

B. Privacy preserving data mining (PPDM)[18] :

Growing development and popularity of data mining has brought some threats to user's sensitive data. Recent developments in new technique, called as Privacy Preserving Data Mining, have been studied and are still studied by researchers for more developments in data security throughout data mining task. The basic idea behind PPDM is to perform data mining task without compromising security of information from any data. Current work in PPDM is carried out to reduce the privacy risk brought by data mining operations. Privacy may be bleached in processes like data collections, data publications, data transformation and so on.

Basically there are four types of users namely data provider or supplier, data collector, data miner, and decision maker. Data mining discover different patterns from huge amount of data. Data mining find its applications to number of areas / domains like business intelligence, Internet search, Library automation in digital format, scientific applications, social media, and the list goes on and on.

C. Smart Grid [19]:

Smart grid refers to next generation power grid. Next generation power grids will combine conventional power grid, communication networks, and pervasive computing capability. It will improve efficiency, reliability, safety and control of grid. By the use of smart grid, it is possible to setup bidirectional flow infrastructure. Flow will carry electricity and information. It will help

to satisfy stakeholder demands, coordinate functionality in power generation, control power consumption and control power market.

Smart grid use real time data processing to provide various features as stated in above paragraph. Smart grid system integrates Automation system, Telecommunication sector, Electric network control and many other sub sectors. To take future decisions and to support requirements, huge amount of data is needed to be fetched and stored in minimum time span. Diverse measurements acquired by Intelligent Electronic Devices (IEDs) in the smart grid generates huge amount of data from various sources. Some these sources are:

a) Data about Power utilization habits of users
b) Phasor measurement Units Data
c) Energy Consumption data
d) Energy market pricing and bidding data generated by ARM (Automated Revenue Metering)
e) Data about management, control and maintenance of equipments used in generation, transmission and distribution of electricity
f) Financial data and large data sets obtained indirectly from other sources

All above resources are generating huge amount of data that leads to Big Data. According to [20], by 2009the amount data generated by electric utilities is in terabytes (TB). As years are passing away, more number of TB data is generated and which is stored at data centers of company. This fact leads to need to apply big Data technology to smart grid system.

SECURITY

With the tremendous demand of analysis of large data set and increase in application use, Big data has found various opportunities in industries and organizations. Until introduction of web-based applications, all the applications and data were internal. With web-based systems, it became possible to access data anywhere by using internet. The risk was very low with internal data utilization as no external threat is possible with internal utilization. With the adoption of Web-based application, cloud-computing, mobile computing, social media, etc. data became more sensitive and unsecure. As now a day's huge amount of data is collected and processed by various organizations, it is vital to ensure security of data. With the help of Big data i.e. large

amount of data and variety of data, researchers students, organizations, scientist and many others can start taking decisions, plan new processing, can discover new opportunities for analysis & optimization, and can deliver very large amount of new innovations.

To design Big Data security, system designer and administrator has to consider following things:

Big Data Framework: Within Big Data itself, we can manage massive amount of data at any point with any type of structure. Big Data can be powered by Hadoop, NoSQL, MongoDB, TeraData or any other system. Sensitive data can be reside on Big data nodes, and in the form of log files, configuration files and error logs as well.

Analytics: The ultimate outcome of Big data utilization is optimization and new innovations by organizations. Data is accessible wherever and whenever is demanded with the help of queries. Data can be presented in the form of reports and in dashboards. Most sensitive asset can be represented by Big Data, in the case data falls in wrong place or wrong hands.

Data Sources: To explore most advantage of big data, organizations can leverage data in various forms. Data exists in structured, unstructured form. We can identify various resources of data to leverage data in structured or unstructured format. These resource include enterprise resource planning systems, video files, social media posts, spreadsheets, text files, customer relationship management information, surveys, and so on. In coming days new resources will arise and will have to compete with more alternatives. In coming time there will be more diversity to accommodate. In order to address security policies and compliance all above mentioned resources need to be compiled.

The security measures provided by Antivirus and Firewall do not providing required security to data and information over web-system, cloud computing and mobile computing. Big data are not limited to hard disk and isolated systems. Big data goes beyond these constraints and expands across number of remote machines.

Big data face some major security challenges as listed below:

1. Single level of protection implemented in computation by distributed systems'
2. NoSQL is used effectively and on large platform, which make security measures difficult to keep up with demands.
3. Some organizations cannot or do not implement different levels of confidentiality within organization.
4. Data trustworthy and accuracy maintenance for newly arrived large amount of data
5. Security measures for automated transactions
6. Routine audits on Big data are not performed due to volume of data
7. Origins of data are unmonitored and not tracked
8. Inaccessibility by IT people to control encryption and connections security
9. Data gathering by Unethical IT practicing people without users permission

CONCLUSION

This paper has presented Big data concept, its basic terminologies and some security concerns. Traditional databases support storage of small number of data sets with number of security measures. But when the term analysis comes in consideration, traditional databases are not sufficient. Traditional database works on structured data or homogenous data, where as current data sets are heterogeneous and exists in various forms. Traditional security measures are not sufficient and are not sufficient for security in Big data. This paper has summarized some security concerns, Big challenges and some applications of Big data. As a future development, researchers can work on analysis methods, optimization methods, security over connections and faster processing.

REFERENCES

1) ChengXiang Zhai, "Basic Concepts in Big Data", Department of Computer Engineering,, University of Illinois at Urbana-Champaign.

2) Boyd, dana; Crawford, Kate, "Six Provocations for Big Data", Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.,2008.

3) Reichman, O.J., Jones, M.B., Schildhauer,"Challenges and Opportunities of Open Data in Ecology", Science. 331 (6018): 703–5. doi:10.1126/science.1197962. PMID 21311007.

4) Hellerstein, Joe,"Parallel Programming in the Age of Big Data", Gigaom Blog,9 November 2008.

5) Paul C. Zikopoulos ,Chris Eaton, Dirk deRoos, Thomas Deutsch, And George Lapis, "Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data ", McGraw-Hill book.

6) Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review", martinhilbert.net. Retrieved 2015-10-07.

7) Rahul Beakta, "Big Data and Hadoop : A Review Paper", RIEECE-2015,ResearchGate, e-ISSN:1694-2329, vol.2, Spl.Issue 2(2015), pp.13-15.

8) Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

9) Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions in International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

10) Divyakant Agrawal, Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the United States.

11) Smitha.T, Dr.V.Sundaram, "Classification Rules by Decision Tree for disease prediction" International journal for computer Application, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975-8887; pp- 35-37.

12) Mucherino A. Petraq papajorgji P.M.Paradalos 1998. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560.

13) M. Chan, E. Esteve, et. al., “A review of smart homes—Present state and future challenges,” Computer Methods and Prigramns in Biomedicine, vol.91. no.1, pp.55-81, Jul. 2008

14) K. Courtney, G. Demiris, et. al., “Needing smart home technologies: the perspectives of older adults in continuing care retirement communities,” Informatics in Primary Care, vol.16, pp.195-201, 2008

15) G. Demiris, B. Hensel BK, et. al, “Senior residents’ perceived need of and preferences for smart home sensor technologies,” Int J Technol Assess Health Care, vol.24. no.1, pp.120-1024, 2008

16) S. Moncrieff; S. Venkatesh, et. al., "Dynamic Privacy in a Smart House Environment," IEEE International Conference on Multimedia and Expo, pp.2034-2037, Jul. 2007

17) S. Meyer, A. Rakotonirainy, “A survey of research on context-aware homes,” Australian Computer Society, vol.21, pp.159-168, 2003.

18) Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, And Yong Ren, “Information Security in Big Data: Privacy and Data Mining”, IEEE Transaction, Digital Object Identifier 10.1109/ACCESS.2014.2362522, pp. 1149-1176.

19) Hui Jiang1, Kun Wang1, Yihui Wang1, Min Gao2, And Yan Zhang3, “Energy Big Data: A Survey”, IEEE SPECIAL SECTION ON THEORETICAL FOUNDATIONS FOR BIG DATA APPLICATIONS: CHALLENGES AND OPPORTUNITIES, 15 August 2016.

20) K.Wang, Y. Shao, L. Shu, Y. Zhang, and C. Zhu, ``Mobile big data fault tolerant processing for eHealth networks," IEEE Netw., vol. 30, no. 1,pp. 1_7, Jan. 2016.