International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(4), pp.043-048
DOI: http://dx.doi.org/10.21172/1.104.09
e-ISSN:2278-621X

BLAZING ISSUES IN BIG DATA

Manisha Valera[1], Yash Patel[2]

Abstract- An enormous depository of terabytes of data is generated each day from contemporary information systems and digital technologies. Analysis of these massive data requires a lot of efforts at multiple levels to extort facts for decision making. Big data analysis is a current area of research and development. The basic intention of this paper is to explore the potential impact of big data challenges and open research issues.As an outcome, this paper provides a platform to discover big data at numerous stages. in addition, it opens a new scope for researchers to develop the solution, based on the challenges and open research issues.

Keywords – Big Data, MapReduce Framework, Security, Privacy

1. INTRODUCTION

The amount of data in world is rising because of use of internet now a days. Big data is a collection of data sets which is very huge in size. Traditional database systems are not able to capture, amass and analyze this huge amount of data. As the internet is growing, amount of big data continue to rise. Now days, big data is one of the burning topics in IT industry. It will play essential role in future. Big data changes the way that data is managed and used.. The present paper highlights important concepts of Big Data. here we discuss various aspects of big data. We define Big Data and discuss the parameters along which Big Data is defined. This includes the five V's of big data and reviews the security aspects of Big Data.

2. FIVE V'S OF BIG DATA

There are many properties associated with big data. The prominent aspects are Volume, Variety, Velocity, Variability and Value.[2]

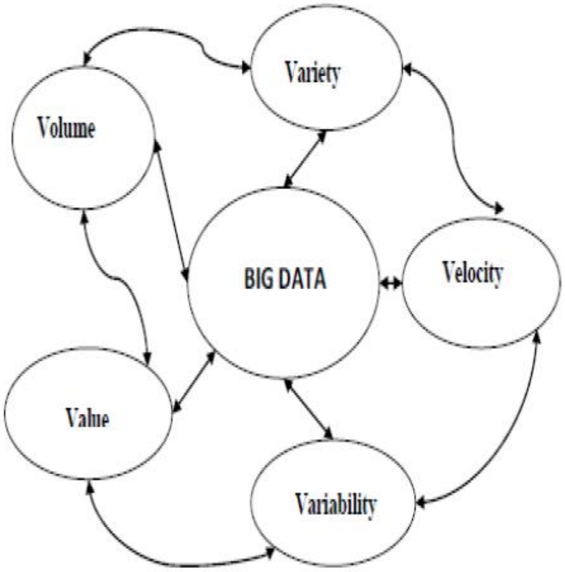

Figure1: Five V's of Big Data

2.1 Volume:

The volume of big data is exploding exponentially day to day. The data gathered through social websites and sensor networks going to cross from petabytes to Zeta bytes. Many factors contribute to the increase in data volume. Transaction-based data stored through the years. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

[1] Assistant Professor, Department of Computer Engineering, Indus University, Gujarat, India
[2] Student, Department of Computer Engineering, Indus University, Gujarat, India

2.2 Variety:

Data produced are from different categories, consists of unstructured, standard, semi structured and raw data which are very difficult to be handled by traditional systems. Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, and financial transactions.

2.3 Velocity:

This concept indicates the speed at which the data generated and become historical. Big data is capable enough to handle the incoming and outgoing data rapidly. Data is streaming in at unprecedented speed and must be apportioned with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

2.4 Variability:

It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

2.5 Value:

All enterprises and e-commerce systems are acute in improving the customer relationship by providing value added services. For that, study on customer attitudes and trends in the market are to be analysed. Moreover, users can also query the data store to find business trends and accordingly they can change their strategies. By making big data open to all, it creates transparency on functional analysis. Supporting real time decisions and experimental analysis in different locations datasets can do wonderful things for enterprises.

[3] Presents A Comprehensive Survey On Security Issues Of rising Technology Which Is Related To Big Data. The Analysis Of Big Data Involves Multiple divergent Phases Which consist of Data Acquisition And Recording, Information Extraction And Cleaning, Data Integration, Aggregation And Representation, Query Processing, Data Modelling And Analysis And Interpretation. Each Of These Phases Introduces Challenges. Heterogeneity, Scale, Timeliness, Complexity And Privacy Are Certain Challenges Of Big Data.

There Are Four Different Aspects Of Big Data Security:

Infrastructure Security

□ □Security For Hadoop

□ □Availability

□ □Architecture Security

□ □Group Communication

□ □Communication Security

□ □Authentication

Data Privacy

□ □Cryptography

□ □Access Control

□ □Confidentiality

□ □Privacy-Preserving Queries

□ □Privacy In Social Networks

□ □Anonymization

□ □Differential Privacy

Data Management

□ □Security At Collection Or Storage

□ □Policies, Laws, Or Government

□ □Sharing Algorithms

Integrity And Reactive Security

□ □Integrity

□ □Attack Detection

□ □Recovery

3. CHALLENGES IN BIG DATA SECURITY AND PRIVACY ISSUES

Daily Big Data faces high level of Problems while dealing with the privacy and security of enormous and diverse data. Data are shared on a large scale by different people such as researchers, scientists, doctors, business officials, government agencies etc. Although the tools and technologies that have been developed till date to handle these huge bulk of data are not able

enough to provide enough security and privacy to data. Recently, CSA (Cloud Security Alliance) released the top ten big data security &privacy challenges [4]. The top researchers from CSA's Big Data working group compiled such relevant challenges in the perspective of big data security and privacy which can be categorized into four aspects and further these categories are subdivided into 10 discrete security challenges as follows [5]:
Infrastructure Security
Data Privacy
Data Management
Integrity and Reactive Security

4. ISSUES AND CHALLENGES IN BIG DATA

Big Data Issues and Challenges Related to Characteristics of Big Data

• Data volume: When data volume is thought of the very first issue that occurs is storage. As data volume increases so the amount of space required to store data proficiently also increases. Not only that the enormous volumes of data needs to be retrieved at a fast speed to extract results from them. Networking, bandwidth, cost of storing like in-house versus cloud storing are other areas to be looked after [6]. With the increase in volume of data the value of data records tend to reduce in proportion to age, type, richness and quality [7].

• Data velocity: Computer systems are creating more and more data, both operational and analytical at growing speeds and the number of consumers of that data are growing. People want all of the data and they want it as soon as possible leading to what is trending as high-velocity data. High velocity data can mean millions of rows of data per second. Traditional database systems are not able enough of performing analytics on such volumes of data and that is constantly in motion. Data generated by both devices and actions of human beings like log files, website click stream data like in E-commerce, twitter feeds can't be collected because the state of the art technology can't handle that data [7].

• Data variety: Big data comes in many a form like messages, updates and images in social media sites, GPS signals from sensors and cell phones and a whole lot more. Many of these sources of big data are virtually new or rather as old as the networking sites themselves, like the information from social networks, Facebook, launched in 2004 and Twitter in 2006. Smart phones and other mobiles devices can be bracketed in the same category. As these devices are omnipresent the traditional databases that store most corporate information until recently are found to be ill suited to these data. Much of these data are unstructured and unmanageable and noisy which requires scrupulous technique for decision making based on the data. Better algorithms to analyze them are an issue too [10].

• Data value: Data are stored by different organizations to gain insights from them and use them for analytics for business intelligence. This storing produces a gap between the business leaders and the IT professionals. The business leaders are concerned with adding value to their business and obtaining profits from it. More the data more are the insights. This however doesn't go fine with the IT professionals as they have to deal with the mechanics related to storing and processing the enormous amounts of data [7].

Big Data Management, Human Resource and Man Power Issues and Challenges

Big data management deals with organization, administration and control of large volumes of structured and unstructured data. It aims to ensure a high level of data quality and ease of access for business intelligence and big data analytics applications.

Efficient data management helps companies, agencies and organizations in locating important information from large sets of the order of terabytes and petabytes of amorphous or semi structured data. Sources may range from social media sites, system logs, call details and messages. There are however some challenges with big data and its management:

• Being new to big data and its management is the major challenge users of big data face. As organizations are new to big data it typically has insufficient data analysts and IT professionals having the skills to help understand digital marketing data [11].

• The sources of big data are assorted with respect to size, format and method of collection. Digital data comes from many medium as at ease to humans, like documents, drawings, pictures, sounds, video recordings, models and user interface designs, with or without metadata describing what the data is and its origin and how it was collected. Immaturity with these new data types and sources and inadequate data management infrastructure are a big problem.

• The expertise of a data analyst must not be limited to the technical field. It should be expanded to research, analytical, interpretive and creative skills [7].

• IT investments are also deficient like purchasing modern analytical tools to manage bigger data and analyze with better efficiency more complex data [11].

• Due to lack of governance or stewardship, business sponsors and a compelling business case it is difficult for new projects to begin [12].

Big Data Technical Issues and Challenges

• Fault Tolerance: With the beginning of technologies like cloud computing the aim must remain such that whenever breakdown occurs the damage done must occur within acceptable threshold rather than the entire work requiring to be redone. Fault-tolerant computing is tedious and requires extremely complex algorithms.

To reduce the probability of failure to an acceptable level we can do:

• Divide the entire computation to be done into tasks and assign these tasks to different nodes for computation.
• Keep a node as a supervising node and look over all the other assigned nodes as to whether they are working properly or not. If a problem occurs the particular task is restarted. There are however certain scenario where the entire computation can't be divided into separate tasks as a task can be recursive in nature and requires the output of the previous computation to find the present result. These tasks can't be restated in case of an error. Here checkpoints are applied to keep the state of the system at certain intervals of time so that computation can restart from the last checkpoint so recorded [7].
• Data Heterogeneity: 80% of data in today's world are amorphous data. Working with unstructured data is inconvenient and expensive too. Converting these to structured data is impractical as well [7].
• Data Quality: storage of big data is very expensive and there is always a disagreement between business leaders and IT professionals regarding the amount of data the company or the organization is storing. The quality of data is an vital factor to be looked into here. Ensuring whether the amount of data is sufficient for a particular conclusion to be drawn or whether the data is relevant at all are further queries [7].
• Scalability: The challenge in scalability of big data has led to cloud computing. It is able of aggregating multiple different workloads with different performance goals into very large clusters. This needs high level of sharing of resources that is quite costly and brings along with it various challenges like executing various jobs so that the goal of every workload is met successfully. It also has to deal with system failures in an efficient manner as it is quite common when working with large clusters. Hard disk drives being replaced by solid state drives and phase change technology do not have the same performance between sequential and random data transfer. The kind of storage device to be used is thus a large question threatening around big data storage issue [7].

5. BIG DATA STORAGE AND TRANSPORT ISSUES AND CHALLENGES

Big data processing issue has been well explained by the author of [9] by a very good example. Each time a new storage medium is invented the quantity of data becomes more and more. The capability of current disks are about 4 terabytes per disk so 1 exabyte requires 25000 disks. Even if a single computer system is capable enough of processing 1 exabyte, to directly work with that many number of disks is well beyond its capacity. Accessing this surge of data overwhelms current communication networks. If 1 gigabyte per second network has an effective sustainable transfer rate of 80% its sustainable bandwidth is about 100 megabytes. This boils down to transferring 1 exabyte for 2800 hours, provided the sustainable transfer rate is maintained. This is actually transferring from the storage point to the processing point for a longer duration than actually processing it [9].

6. BIG DATA PROCESSING ISSUES AND CHALLENGES

Effectual processing of big data requires vast parallel processing and new analytics algorithms so as to provide rapid information. Often it may be unknown how to deal with a very large and varied volume of data and whether all of it needs to be analyzed. Challenges also include finding out data points that are really of importance and how to use the data to mine maximum advantage from it [7].

7. REASONS FOR SECURITY AND PRIVACY ISSUES AND CHALLENGES IN BIG DATA

Security and privacy are big concerns as far as big data are concerned and as big data grows by volume every day, every minute, every second so are these concerns on the rise [8].

A principal reason for security and privacy concern in big data is because big data is now widely easily reached. However the tools and technologies that have been developed till date to handle these massive volumes of data are not proficient enough to provide enough security and privacy to data [8].
• The technologies lack adequate security and privacy maintenance features and the reason for this is because there is a lack of basic understanding about how to provide protection to these huge volumes of data and sufficient training is not provided regarding how to provide security and privacy to these large scale data [8].
• The data security and privacy maintenance regarding big data lacks adequate policies that ensure agreement with current approaches to security and privacy [8].
• The present technologies have fragile security and privacy maintenance capability so they are continuously being breached both accidentally and intentionally. Thus reassessing and updating current approaches to prevent data leakage has to be done on a continuous basis [8].
• There is lack of spending on IT security to protect big data by the companies. About 10% of a company's IT budget should be spent on security but below 9% is spent on an average thus making it tougher for themselves to protect their data.

Privacy and Security Issues and Challenges with Big Data
• Secure Computations in Distributed Programming Frameworks Distributed programming frameworks use parallel computing and data storage for huge amounts of data. An example of this is MapReduce framework, which divides an input file into many chunks and then a mapper for each chunk reads the data, does computations and provides outputs in the form of key/value pairs. A reducer then combines the values belonging to each unique key and outputs the results. The main concerns here are: securing the mappers and securing the data from a malicious mapper. Mappers returning incorrect results are difficult to detect and it eventually results in incorrect aggregate outputs. With very large data sets malicious mappers are

too hard to be detected as well and they eventually harm essential data. Mappers leaking, intentionally or unintentionally, private records are also an issue of concern. MapReduce computations are often subjected to replay attack, man-in-the-middle attack and denial-of-service attack Rogue data nodes can be added to a cluster, and in turn receive replicated data or deliver altered MapReduce code. Creating snapshots of legitimate nodes and reintroducing altered copies is an easy attack in cloud and virtual environments and is difficult to detect [13].

• Security Best Practices for Non-Relational Data Stores Non-relational databases used to store big data, mainly NoSQL databases, handle many challenges of big data analytics without concerning much over security issues. NoSQL databases consist of security embedded in the middleware and no explicit security enforcement is provided. Transactional integrity maintenance is very lenient in NoSQL databases. Complex integrity constrains can't be inculcated in NoSQL databases as it hampers with its functioning of providing better performance and scalability. NoSQL databases have weak authentication techniques and weak password storage mechanisms. They use HTTP Basic- or Digest- based authentication and are subjected to man-in-the-middle attack. REST (Representational State Transfer) based on HTTP is prone to cross-site scripting, cross-site request forgery and injection attacks like: JSON injection, array injection, view injection, REST injection, GQL (Generalized Query Language) injection, schema injection and others. NoSQL is unsupportive of blocking with the help of third party as well. Authorization techniques in NoSQL provide authorization at higher layers only. It provides authorization on a per database level rather than at the level where the data are collected. NoSQL databases are subjected to inside attacks as well due to lenient security mechanisms. They may go unnoticed due to poor logging and log analysis methods along with other fundamental security mechanisms [13]

• Secure Data Storage and Transaction Logs Data and transactions logs used to be kept in multi-tiered storage media. It doesn't keep track of where the data are stored unlike in previous multi-tiered storage media where IT managers knew which data resided where and when. This gave rise to many new challenges for data security storage. Unreliable storage service providers often search for clues that help them correlate user activities and data sets and get to know certain properties, which can well prove essential to them. They however are not able to break into the data overcoming the encipherment. As the data owner stores the cipher text in an auto-storage system and distributes the private key to each user, he gives the right to access data of certain portions to certain users, he being unauthorized to access the data. However he may conspire with users by exchanging the key and data hence he can obtain data to which he is not allowed to. The service provider can instigate roll back attack on users in case of a multi-user environment. He may serve archaic versions of data while the updated ones are already uploaded in the database. Data tampering and data loss resulted by malevolent users often results in disputes between the data storage provider or amongst users [13].

•End Point Input Validation/ Filtering Organizations collect data from a variety of sources including hardware devices, software applications and endpoint devices. As and when collecting these data, validation of the data as well as the source is a challenge. Often mischievous users tamper with the device from where the data are collected or tamper with the data collecting application installed in the device so that malicious data gets input into the central data collecting system. Fake IDs may be created by malicious users and provide malicious data as input into the central data collecting system. ID cloning attacks like Sybil attacks are predominant in a Bring Your Own Device (BYOD) scenario where a malicious user brings his own device, faked as a trusted device and provides malicious input from there into the central data collecting system. Input sources of sensory data can be manipulated as well like artificially changing the temperature from a temperature sensor and inputting malicious input into the temperature collection process. GPS signals can be manipulated much the same way. The malicious user may change data while it is in transmission from a generous source to the central data collection system. It's a man-in-the middle attack in a sense [13].

• Real-Time Security Monitoring Real-time security monitoring has been an ongoing challenge in the big data analysis scenario mainly due to the number of alerts generated by security devices. These alerts, may be co-related may be not, lead to many false positives and due to human being's incapability to successfully deal with such an huge amount of them at such a speed, results in them being clicked away or ignored [14]. Security monitoring requires that the Big Data infrastructure or platform be inherently secure. Threats to a Big Data infrastructure include rogue admin access to applications or nodes, (web) application threats, and eavesdropping on the line. Infrastructure which is mostly an ecosystem of different components, the security of each component and the security integration of the components must be considered. In case of a Hadoop cluster run in a public cloud the security of the public cloud, itself being an ecosystem of components consisting of computing, storage and network components, needs to be considered. The security of the Hadoop cluster, the security of the nodes, the interconnection among the nodes and the security of the data stored in a node needs to be considered. The security of the monitoring application including applicable correlation rules that should follow secure coding principles, must be considered as well. The security of the input source from where the data comes from too must be taken into account [13].

• Scalable and Composable Privacy-Preserving Data Mining and Analytics Big data are subjected to appropriation of privacy, invasive marketing, reduction of civil liberty and increase in state and corporate control. An employee of a company in charge of the big data store can misuse his power and violate privacy policies. For example: He can stalk people by monitoring through chats, if the company is a social networking one that facilitates chatting. An untrustworthy business partner can infiltrate into private information and take it up into the cloud as cloud infrastructure is handled by the owners of data [13].

• Cryptographically Enforced Data-Centric Security There exist two fundamental approaches of controlling visibility of data to individuals, organizations and systems. The first one being restricting access to underlying systems like operating systems or hypervisor. The second is encapsulating the data itself in a protective shell by virtue of cryptography. The first approach or the system-based approach provides a larger attacking surface. There are many attacks like buffer overflow and privilege escalation attack that bypass access control implementations and access the data. Protecting data end-to-end by encryption provides a much smaller well-defined attacking surface. It is vulnerable to covert side-channel attack and can extract secret keys, though it is an insurmountable task. Various threats associated with cryptographically enforced access control method using encryption are: It should not be identifiable by the adversary, the corresponding plaintext data looking at the cipher text even if he has to choose between a correct and an incorrect plain text. For a cryptographic protocol for facilitating searching and filtering encrypted data the adversary should not be able to learn anything about the encrypted data beyond the corresponding predicate, whether satisfied or not. The cryptographic protocol must also ensure that adversary must not be able to forge data that came from the claimed source for this may well be false hence affecting integrity of data [13].

• Granular audits Real-time security monitoring notification at the very moment an attack takes place is a real challenge. There may often be new attacks or missed true positives. In order to discover a missed attack audit information is required. Audit information from any device must be complete or rather it must give us details about what exactly happened and what went wrong. It must give timely access, so that it serves the purpose of compliance, regulation and forensic investigation. It must not be tampered and must be accessible only in authorized areas [1].

8. CONCLUSIONS

It is observed from the study that data will keep on rising as the year run by so it is very imperative to make sufficient arrangement on how to protect such essential information. Furthermore for the future trend of the ever growing data which is expected to be doubling-up on a yearly basis, research should continue in these areas to see how these concepts can be improved and how the issues and challenges can be reduced to the barest minimum. Encouraging progresses have been made in the area of big data, but much work still needs to be done. Therefore it is imperative to constantly improve the security strategies for securing the Big data.

9. REFERENCES

[1] Getaneh Berie Tarekegn, Yirga Yayeh Munaye," BIG DATA: SECURITY ISSUES, CHALLENGES AND FUTURE SCOPE" , International Journal of Computer Engineering & Technology (IJCET) Volume 7, Issue 4, July–Aug 2016, pp. 12–24, Article ID: IJCET_07_04_002

[2] Manisha Valera, Ankit Virparia, Om Mehta, "AN EXHAUSTIVE STUDY: BIG DATA", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 6, June 2016

[3] Trupti V. Pathrabe, "Survey On Security Issues Of Growing Technology: Big Data", IJIRST, National Conference On Latest Trends In Networking And Cyber Security, March 2017.

[4] https://cloudsecurityalliance.org/media/news/csa-big-data-releases-top-10-security-privacy challenges/.

[5] A Cloud Security Alliance Collaborative research, Expanded Top Ten Big Data Security and Privacy Challenges. 2013.

[6] www.coursera.org, Introduction to Big Data, University of California, San Diego. https://www.coursera.org/learn/big-data-introduction

[7] http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report. Published: 4th January 2014 in Technology, Education Harsh Kishore Mishra. Center for Computer Science and Technology. School of Engineering and Technology, Central University of Punjab, Bhatinda

[8] Schmitt, C., Shoffner, M., Owen P., Wang, X., Lamm, B., Mostafa, J., Barker, M., Krishnamurthy, A., Wilhelmsen, K., Ahalt, S., & Fecho, K. (2013)," Security and Privacy in the Era of Big Data: The SMW, a Technological Solution to the Challenge of Data Leakage", RENCI, University of North Carolina at Chapel Hill. Text: http://dx.doi.org/10.7921/G0WD3XHT Vol. 1, No. 2 in the RENCI White Paper Series, November 2013.Created in collaboration with the National Consortium for Data Science. (www.data2discovery.org)

[9] Stephen Kaisler, i_SW Corporation. Frank Armour, American University. J. Alberto Espinosa, American University. William Money, George Washington University, "Big Data: Issues and Challenges Moving Forward", 2013 46th Hawaii International Conference on System Sciences

[10] Andrew McAfee and Erik Brynjolfsson, "Big Data: The Management Revolution", October 2012. Harvard Business Review

[11] http://www.dataversity.net/common-big-data-management-issues-solutions/ The Most Common Big Data Management Issues (And Their Solutions). By: A.R. Guess. July 15 2014.

[12] TDWI Research. TDWI Best Practices Report. Managing Big Data. Fourth Quarter 2013. By Philip Russom

[13] Expanded Top Ten Big Data Security and Privacy Challenges Big Data Working Group. April 2013. © 2013 Cloud Security Alliance – All Rights Reserved

[14] Reena Singh. Kunver Arif Ali ",Challenges and Security Issues in Big Data Analysis", IJIRSET. Volume: 5. Issue: 1. January 2016.