Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.142 – 147

Big Data Security: A Review of Big Data, Security Issues and Solutions

Pooja Bisht[1], Kulvinder Singh[2]

Computer Science, Uttarakhand Technical University, India

[1] poojabisht001@gmail.com; [2] kulvinder.taak@gmail.com

Abstract— Big Data is an important issues in the recent years, enables computing resources to be provided as Information Technology services with high efficiency and effectiveness. The amount of data in world is growing day by day. Data is growing because of use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. Generally size of the data is Petabyte and Exabyte. Traditional database systems are not able to capture, store and analyze this large amount of data. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the limits. However, the fast growth rate of such large data generates numerous challenges, such as the rapid growth of data, transfer speed, diverse data, and security. The paper shows the fundamental concepts of Big Data. These concepts include the overview of big data, its characteristics, big data technologies, and security issues with big data with discussing new approaches for Security used in Big Data.

Keywords— Big Data, Hadoop, MapReduce, HDFS, NoSQL, IoT, Security

I. INTRODUCTION

Big Data is the word used to describe massive amounts of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. Due to the advent of new technologies, devices, and communication like social networking sites, the amount of data produced by mankind is growing rapidly every year. Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not just a data; rather it has become a complete subject, which involves various tools, techniques and frameworks.

The origin of the term 'Big Data' is used due to the fact that we are creating a huge amount of data every day. The growth rate of data is expected to double every two years, from 2500 Exabytes in 2012 to 40,000 Exabytes in 2020. At the KDD BigMine 12 Workshop Usama Fayyad in his invited talk presented following data numbers about internet usage, that is each day Google has more than 1 billion queries, Twitter has more than 250 million tweets per day, Per day Face book has more than 800 million updates, and YouTube has more than 4 billion views per day. Big Data is a heterogeneous mix of both structured data (traditional datasets –in rows and columns like DBMS tables, CSV's and XLS's) and unstructured data like PDF documents, e-mail attachments, images, manuals , medical records such as x-rays, ECG and MRI images, rich media like graphics, audios and videos, contacts, forms and documents. Businesses are primarily concerned with managing unstructured data, because about 80 percent of enterprise data is unstructured.

II. CHARACTERISTICS OF BIG DATA

Big data has certain characteristics with which we can separate it from the normal data. There are six key characteristics that define big data. These characteristics are also known as Six V's of big data.

Volume: It refers to the vast amount of data generated every second. Many factors contribute towards increasing volume such as storing transaction data, live streaming data and data collected from sensors, human interaction on systems like social media etc. New big data tools use distributed system so that we can store and analyse data across databases that are dotted around anywhere in the world. The quantity of data generated is not in terabytes but in petabytes or zettabytes.

Variety: Variety Refers to the type of data that is being stored. Today data comes in different types of formats from different or many sources. With the explosion of sensors, smart devices and social media technologies, data is being generated in countless forms, including text, web data, tweets, sensor data, audio, video, click streams, log files and more. There are three categories on the variety of data, structured data (relational databases), semi-structured data (xml data) and unstructured data (text and multimedia contents). In fact, 80% of the world's data is unstructured (text, images, video, voice, etc).

Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand. It refers to the speed at which new data is generated and the speed at which data moves around. Like social media messages going viral in seconds. New Technology allows us now to analyze the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Veracity: It refers to the level of reliability associated with certain types of data. Veracity means data uncertainty. The quest for high data quality is an important Big Data requirement and challenge, but even the best data cleansing methods cannot remove the inherent unpredictability of some data, like the weather, the economy, or a customer's buying decisions.

Value: Every data is important and carries Value. Good information may be hidden in unstructured non-traditional data. The challenge is identifying what is valuable and then transforming and extracting that data for analysis.

Variablity: It refers to the messiness, inconsistency or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable but technology now allows us to work with this type of data. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent.

III. BIG DATA TECHNOLOGY

Here is an overview of important technologies to know about for context around big data infrastructure.

A. Hadoop

Hadoop is an open source; Java-based programming framework supports the processing of large sets across clusters of computers in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Using Hadoop, big amount of data sets can be processed over cluster of servers and applications may be run on system with thousands of nodes involving terabytes of information. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This lowers the risk of system failure even when a large number of nodes fail. It enables a scalable, cost effective, flexible, fault tolerant computing solution. A distributed file system spanning all nodes in a Hadoop cluster for data storage links the file systems on local nodes to make it onto a very large file system thus improving the reliability.

The Hadoop framework is used by popular companies including Google, Yahoo Amazon and IBM, largely for applications involving search engines and advertising. The preferred operating systems are Windows and Linux but Hadoop can also work with BSD and OS X.

Hadoop framework includes following four modules:

- Hadoop Common: It contains a set of components Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- Hadoop YARN: This is a framework for resource management and schedules the jobs across the cluster.
- Hadoop Distributed File System (HDFS): A distributed file system that is used to store data across cluster of commodity machines while providing high availability and fault tolerance.
- Hadoop MapReduce: This is YARN-based system for parallel processing of huge data sets.

Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow (workflow) and configuration administration. HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on many input and output data nodes to make them into one big file system. Doug Cutting, Hadoop's creator, who was working at Yahoo at the time named the framework after his child's stuffed toy elephant. The present Hadoop ecosystem consists of the Hadoop kernel, MapReduce, the Hadoop distributed file system (HDFS) and a number of related components such as Apache Hive, HBase, Oozie, Pig and Zookeeper.

B. MapReduce

MapReduce is a core component of the Apache Hadoop framework which was developed by Google in response to the problem of creating web search indexes. It is a software framework used to write applications for large-scale data processing in parallel on clusters of commodity computer resources in a reliable and in fault-tolerant manner. A Map Reduce job initially divides the data into individual chunks which are processed by individual Map jobs in parallel. Then the outputs of the maps sorted by a framework are input to the reduce tasks. Generally the input and the output of the job are both stored in a file-system. Monitoring, Scheduling, re-executing failed tasks are taken care by the framework.

The MapReduce algorithm contains three important tasks, namely Map, Sort and Reduce.

- The Map task takes a set of data and converts it into key-value pairs, where individual elements are broken down into tuples.
- The individual key-value pairs are sorted by key into a larger data list. The data list groups the equivalent keys together so that their values can be iterated easily in the Reducer task.
- The Reduce task takes the grouped key-value paired data as input and runs a Reducer function on each one of them. Here, the data can be aggregated, filtered, and combined (key-value pairs) into a smaller set of tuples.

The output of the reduce task is typically written to the FileSystem.

C. Hadoop Distributed File System (HDFS)

Hadoop distributed file system is design to span all the nodes in a Hadoop cluster for data storage. It provides reliable and scalable data storage and is a Java based file system. It links together file systems on local nodes to make it into a large file system. It improves reliability by replicating data across multiple sources to overcome node failures. There are no restrictions on the data that HDFS stores the data can be unstructured and schema-less while relational databases require that data be structured and schemas be defined before storing the data. HDFS also makes applications available to parallel processing. An important feature of HDFS is that it provides file permissions and authentication.

D. NoSQL

NoSQL (usually referred to as "Not Only SQL") designs to handle extremely large data sets and allows high-performance, agile processing of information at massive scale. It is a column oriented database infrastructure that has been flexibly adapted to the full demands of big data. The efficiency of NoSQL can be achieved because it is associated with unstructured data unlike relational databases that are highly structured. NoSQL databases are unstructured in nature, trading off stringent consistency requirements for speed and agility. NoSQL centralizes on the concept of distributed databases, where unstructured data may be stored across multiple processing nodes, and multiple servers. This distributed architecture allows NoSQL databases to be horizontally scalable; as data continues to grow, just add more hardware to keep up, without slowing down in performance. The NoSQL distributed database framework has been the solution to handling some of the biggest data warehouses i.e. the likes of Google, Amazon, and the CIA.

E. Internet of things (IoT)

The Internet of Things is the network of everyday objects, computing devices, mechanical and digital machines, buildings, vehicles, objects, animals or people embedded with electronics, software, sensors, and network connectivity that enables these objects to collect and exchange and transfer data over a network without requiring human-to-human or human-to-computer interaction. These objects have built-in sensors that can be assigned an IP address and provided with the ability to transfer data over a network. Internet of Things allows objects to be sensed and controlled remotely across existing network. The output of these sensors is machine-generated data which is well structured from simple sensor records to complex computer logs. As sensors proliferate and amounts of data grow, it is becoming an increasingly important component of the information stored and processed by many businesses. A thing, in the Internet of Things, can be a person with a heart monitor implant, a farm animal with a biochip transponder, an automobile that has built-in sensors to alert the driver when tire pressure is low or any other natural or man-made object. Therefore the Internet of Things is the future of technology that can make our lives more efficient data.

IV. SECURITY ISSUES WITH BIG DATA

Big data is used by many enterprises, organizations for marketing and research but they may not have fundamental assets particularly for security perspective. If a security threat occurs to big data, it would become even more serious issues. Nowadays, many enterprises are using this to store and analyze petabytes of data about the company, business and their customers. As a result classification of information becomes more critical. Big data deals with storing, processing and retrieval of data. Many technologies are used for these purposes just like memory management, transaction management, virtualization and networking. Hence security issues of

these technologies are also applicable for big data. The four important security issues of big data are authentication level, data level, network level and generic issues.

Issues on Authentication level: The issues that can be categorized under user authentication level deals with encryption/decryption techniques, authentication methods such as administrative rights for nodes, authentication of applications and nodes, and logging. There are many clusters and nodes present and every node have a different priorities or rights. Nodes with administrative rights can access any data. But sometimes if any malicious node got administrative priority then it will steal or manipulate the critical user data. For faster execution with parallel processing, many nodes join clusters. In case of no authentication any malicious node can disturb the cluster.

Issues on Network level: The issues that can be categorized under a network level deal with network protocols and network security, such as distributed nodes, distributed data, Internodes communication.

There are many nodes present in clusters and processing of data is done in these nodes. This processing of data can be done anywhere among the nodes in cluster. So it is difficult to find on which node data is processing. Because of this difficulty on which node security should be provided is going to be complicated. Two or more nodes can be communicate with each other or share their data/resources through network. Many times RPC (Remote Procedure Call) is used for communicating via network. But RPC is not securing until and unless it is encrypted.

Issue on Data level: The issues that can be categorized under data level deals with data integrity and availability such as data protection and distributed data. Data is very important part and also plays essential role in big data. Data is important and personal information about us by the government or social networking sites. To improve efficiency, big data environments like Hadoop store the data as it is without encryption. If hacker access the machines, then there is impossible to stop him. In distributed data store, information is stored in many nodes with replicas for quick access. But if any replica or information from other node is deleted or manipulated by hacker then it will be difficult to recover that data.

Issues on Generic level: The issues that can be categorized under general level are traditional security tools, and use of different technologies. In big data environment many technologies are used for processing the data also some traditional security tools for security purposes. Traditional tools are developed over years ago. So these tools may not be performed well with new distributed form of big data. As big data uses many technologies for data storing, data processing and data retrieval, there may be some complexities occur because of these various technologies data.

V. APROACHES FOR SECURITY IN BIG DATA

Big data technologies were not initially designed with security in Mind. Data security means security of data from unauthorized (intentional, unintentional or malicious) alteration, destruction, or exposure. It can also define as protecting a database from critical forces and the unwanted actions from users who are not unauthorized. Security should be the first priority when it comes to putting up the big data in organizations. Understanding the data risks, recognize the common attacks and retain security to protect the data. The data that organizations hold is essential for their success. Tragically, an organization data may also hold significance to other organizations or individual person. We need to make sure that it is secure from unauthorized access. But there will be more refined efforts to steal data and destroy reputations of organizations. These problems are constantly evolving and we can't ignore these problems.

A. A new emphasis on encryption

Hackers can easily hack the data which is unencrypted but even when encryption is used, there's a question mark over the effective use of encryption keys, with many businesses failing to manage them effectively. The role of encryption can be integral in big data, as long as it is performed using suitable encryption algorithms and key sizes, and the encryption keys are adequately secured. Still, big data analytics need to allow for search and other computations over the stored data. There are various new encryption techniques used today such as Homomorphic encryption, Attribute-Based-Encryption, Functional encryption etc.

B. Security as a service

Enterprises need to protect their properties, but they also need to be profitable to stay in business market. Protecting information properties has become a priority for enterprises that need to meet obedience necessities or need to protect sensitive data. The cloud environment provides a solution called security as a service also known as SecaaS offers a way for organizations to access security services that are robust, scalable and cost effective. With reward comes risk, and enterprises should consider advantages and risk when assessing SecaaS goods and their suppliers. Especially, enterprises need to know that they can outsource responsibility but they cannot outsource accountability; so organizations should implement an assertion plan that includes assessing the services obtained from SecaaS providers. When an audit is not possible, enterprises must still gain proof that

controls used to protect organization information properties are working efficiently. It will provide protected email, improved identity management and new emphasis on security from small and medium enterprises.

C. Real-time gathering of data

Today's, the era of Big Data, and everybody is getting in on the act. From marketing enterprises to content providers, Big Data is driving a more personalized digital involvement.

As we acquire more and more wearable, and connect objects in our atmosphere via the Internet of Things, there will be more data being streamed back to these businesses, in real time. There are still problem with the use of this data; many users don't yet understand how it is stored or used. But the devices we purchase are efficiently data gathering devices, and we all need to be more aware of the significances.

D. Privacy by Design

The concept of privacy and data protection by design is essential, as a mechanism to report the privacy risks from the very beginning of the processing and apply the necessary privacy preserving solutions in the different stages of the big data value chain. Increasingly, Businesses will need to be able to demonstrate that they have actively considered privacy and adequately addressed any associated information security risks and that this is built into the DNA of their organization.

E. Data Protection

Data Protection will become mandatory for organizations whose primary purpose involves processing sensitive personal data or who monitor data subjects regularly on a large scale.

F. Log management

It is a process of collection of logs from any device, aggregation of logs into a single searchable format, analysis of logs through data enrichment, and long-term retention of log data. To detect attacks, diagnose failures, or investigate unusual behavior, we need a record of activity. Unlike less scalable data management platforms, big data is a natural fit for collecting and managing event data. Many web companies start with big data particularly to manage log files. It gives us a place to look when something fails, or if someone thinks you might have been hacked. So to meet the security requirements, we need to audit the entire system on a periodic basis.

G. Authentication

Traditional strong authentication methods built on top of passwords do not address the liability and risk of the insecure password layer, and the shared secret architecture of tokens and one-time passwords (OTP) used on top of passwords is cryptographically inferior when compared to their asymmetric counterparts used within modern public-key crypto systems. Such outdated methods are vulnerable to many attack vectors and create a cumbersome experience that users dislike and often avoid. In fact, overly simple authentication is a significant source of consumer mistrust of brands. Most importantly, none of these methods are compatible with many of the devices and "things" that will require user authentication in the (near) future, but lack the requisite input mechanisms. The ubiquity of smart phones and connectivity combined with emerging biometrics technology provides opportunities to reinvent authentication, bringing control and convenience to our fingertips.

H. Data anonymization

Anonymization is the process of modifying personal data in such a way that individuals cannot be re-identified and no information about them can be learned. Perfect anonymization is difficult in practice without compromising the utility of the data set. In big data this problem increases due to the amount and variety of data. Appropriate anonymisation techniques can still play a role in ensuring the safe use or sharing of data within an organization, among different organizations or when data is made publicly available, such as in case of 'open data' projects. Anonymize data fields such that sensitive information cannot be pinpointed to an individual record. For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconduct by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals.

I. Use secure communication

Secure communications are required to protect data-in-transit. There are multiple threat scenarios that in turn mandate the necessity for https and prevent information disclosure or elevation of privilege threat categories. Using the TLS protocol (which is now available in all Hadoop distributions) to authenticate and ensure privacy of communications between nodes, name servers, and applications.

VI. CONCLUSIONS

Nowadays many companies and organizations like banking, educational system, government institute, insurance companies etc. are using big data for analysis purpose. Huge amount of digital information collected from these sources. Data deluge is going to keep on increasing throughout the next years, and each data scientist will have to handle much more quantity of data every year. This data is becoming more diverse, larger, and faster. When deal with big data, security is one of the challenges that arise when systems try to handle the concept of big data. More researches required to overcome the security of big data instead of current security algorithms and methods. The Paper has elaborated a brief overview on the big data its characteristics, big data technologies, and security issues with big data with new approaches for Security used in Big Data. The purpose of this paper is to study and investigate the principle of big data and security issues concern with big data. The future of big data is unlimited and the evolution is unimaginable in E-commerce and services.

REFERENCES

[1] Ahmed and Saeed, "A Survey of Big Data Cloud Computing Security", *International Journal of Computer Science and Software Engineering (IJCSSE),* Volume 3, Issue 1, December 2014.
[2] ENISA, "Privacy by design in big data", Final 1.0, Public, December 2015.
[3] European data protection supervisor, "Meeting the challenges of big data", Opinion 7/2015.
[4] Carlo Vaccari, "Big Data in Official Statistics", PhD Thesis in Computer Science, University of Camerino, JULY 2014.
[5] Cooper, Mell, "Tackling Big Data", NIST Information Technology Laboratory Computer Security Division.
[6] Gaddam, "Securing your Big Data Environment", Black Hat USA 2015
[7] Geethakumari, Srivatsava, "Big Data Analysis for Implementation of Enterprise Data Security", *International Journal of Computer Science and Information Technology & Security (IJCSITS),* ISSN: 2249-9555 Vol. 2, No.4, August 2012.
[8] Inukollu, Arsi and Ravuri, "Security Issues Associated With Big Data in Cloud Computing", *International Journal of Network Security & Its Applications (IJNSA),* Vol.6, No.3, May 2014.
[9] Khan, Yaqoob, Hashem, Inayat, Ali, Alam, Shiraz, Gani, "Big Data: Survey, Technologies, Opportunities, and Challenges", Volume 2014, Article ID 712826.
[10] K.U. and David, Issues, Challenges, and Solutions: "Big Data Mining", M.E.S College, Marampally, Aluva, Cochin, India.
[11] https://datajobs.com/what-is-hadoop-and-nosql
[12] https://en.wikipedia.org/wiki/Internet_of_Things
[13] http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT
[14] http://www8.hp.com/h20195/V2/GetPDF.aspx/4AA5-0863ENW.pdf
[15] https://www.ciphercloud.com/blog/cloud-encryption-trends-predictions-for-2016-taking-a-proactive-approach-to-data-protection/
[16] http://www.fruitionit.co.uk/2016/02/our-5-key-predictions-for-information-security-in- 2016/
[17] [http://www.isaca.org/Knowledge-Center/Research/ResearchDeliverables/Pages/Security-As-A-Service.aspx
[18] http://www.sas.com/en_us/insights/big-data/internet-of-things.html
[19] [http://techspective.net/2016/02/04/stronger-security-requires-advanced-authentication/
[20] http://www.tutorialspoint.com/hadoop/index.htm
[21] Venkatesh H, Perur, Jalihal, "A Study on Use of Big Data in Cloud Computing Environment", *International Journal of Computer Science and Information Technologies,* Vol. 6 (3), 2015, 2076-2078.
[22] Vinit Gopal Savant, "Approaches to Solve Big Data Security Issues and Comparative Study of Cryptographic Algorithms for Data Encryption*", International Journal of Engineering Research and General Science* Vol 3, Issue 3, May-June 2015.

