A Review on Security Challenges of Big Data

[1]A V S M Adiseshu, [2]S Lavanya Reddy, [3]P Hasitha Reddy, [4]Roshini K

[1]Assistant Professor, [2]Associate Professor, [3]Assistant Professor, [4]Assistant Professor
[1,2,3,4] Department of Computer Science & Engineering, Sree Dattha Institute of Engineering & Science

Abstract: Innovations in technology and greater affordability of digital devices have presided over today's Age of Big Data, an umbrella term for the explosion in the quantity and diversity of high frequency digital data. These data hold the potential as yet largely untapped to allow decision makers to track development progress, improve social protection, and understand where existing policies and programmes require adjustment. Big data implies performing computation and database operations for massive amounts of data, remotely from the data owner's enterprise. Since a key value proposition of big data is access to data from multiple and diverse domains, security and privacy will play a very important role in big data research and technology. Hence Big Data faces security issues and challenges for controlled and assured sharing for efficient direct access. The objective of this paper is to make effective use of big data requires access from any domain to data in that domain, or any other domain it is authorized to access with security is a challenge.

Key Words: Big Data, Analytics, security, hadoop, mapreduce.

1. INTRODUCTION

The term *Big Data* refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety). Human beings now create 2.5 quintillion bytes of data per day. The rate of data creation has increased so much that 90% of the data in the world today has been created in the last two years alone. This acceleration in the production of information has created a need for new technologies to analyze massive data sets. The urgency for collaborative research on Big Data topics is underscored by the U.S. federal government's recent $200 million funding initiative to support Big Data research.

Managing big data and navigating today's threat environment is challenging. The rapid consumerization of IT has escalated these challenges. The average end user accesses myriad websites and employs a growing number of operating systems and apps daily utilizing a variety of mobile and desktop devices. This translates to an overwhelming and ever-increasing volume, velocity, and variety of data generated, shared, and propagated. The threat landscape has evolved simultaneously, with the number of threats increasing by orders of magnitude in short periods. This evolving threat landscape, the number of sophisticated tools and computing power that cybercriminals now have at their disposal, and the proliferation of big data mean software security companies are wrestling with challenges on an unprecedented scale. Protecting computer users from the onslaught of cyber threats is no easy task. If threat detection methodologies are weak, the result is inadequate. Successful protection relies on the right combination of methodologies, human insight, an expert understanding of the threat landscape, and the efficient processing of big data to create actionable intelligence. Understanding how data is organized, analyzing complex relationships, using specialized search algorithms, and employing custom models are critical components.

2. THE EVOLUTION OF DATA MANAGEMENT

It would be nice to think that each new innovation in data management is a fresh start and disconnected from the past. However, whether revolutionary or incremental, most new stages or waves of data management build on
their predecessors. Although data management is typically viewed through a software lens, it actually has to be viewed from a holistic perspective. Data management has to include technology advances in hardware, storage, networking, and computing models such as virtualization and cloud computing. The convergence of emerging technologies and reduction in costs for everything from storage to compute cycles have transformed the data landscape

and made new opportunities possible. As all these technology factors converge, it is transforming the way we manage and leverage data. Big data is the latest trend to emerge because of these factors. So, what is big data and why is it so important? Later in the book, we provide a more comprehensive definition. To get you started, big data is defined as any kind of data source that has at least three shared characteristics:

✓Extremely large *Volumes* of data
✓Extremely high *Velocity* of data
✓Extremely wide *Variety* of data

Big data is important because it enables organizations to gather, store, manage, and manipulate vast amounts data at the right speed, at the right time, to gain the right insights. But before we delve into the details of big data, it is important to look at the evolution of data management and how it has led to big data. Big data is not a stand-alone technology; rather, it is a combination of the last 50 years of technology evolution.

DEFINING BIG DATA

Big data is not a single technology but a combination of old and new technologies that helps companies gain actionable insight. Therefore, big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction. As we note earlier in this chapter, big data is typically broken down by three characteristics:

✓**Volume:** How much data
✓**Velocity:** How fast that data is processed
✓**Variety:** The various types of data

Although it's convenient to simplify big data into the three *V*s, it can be misleading and overly simplistic. For example, you may be managing a relatively small amount of very disparate, complex data or you may be processing a huge volume of very simple data. It is critical that you don't underestimate the task at hand. Data must be able to be verified based on both accuracy and context. An innovative business may want to be able to analyze massive amounts of data in real time to quickly assess the value of that customer and the potential to provide additional offers to that customer. It is necessary to identify the right amount and types of data that can be analyzed to impact business outcomes. Big data incorporates all data, including structured data and unstructured data from e-mail, social media, text streams, and more. This kind of data management requires that companies leverage both their structured and unstructured data.

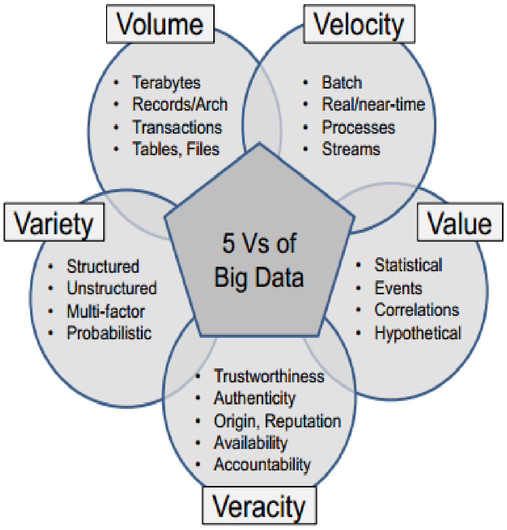

Fig.1 Big Data Architecture

3. THE BIG DATA ARCHITECTURE

To understand big data, it helps to lay out the components of the architecture. A big data management architecture must include a variety of services that enable companies to make use of myriad data sources in a fast and effective manner. In 2000, Seisint Inc. develops C++ based distributed file sharing framework for data storage and querying. Structured, semi-structured and/or unstructured data is stored and distributed across multiple servers. Querying of data is done by modified C++ called ECL which uses apply scheme on read method to create structure of stored data during time of query. In 2004 LexisNexis acquired Seisint Inc. and 2008 acquired ChoicePoint, Inc. and their high speed parallel processing platform. The two platforms were merged into HPCC Systems and in 2011 was open sourced under Apache v2.0 License. Currently HPCC and Quantcast File System[50] are the only publicly available platforms capable of analyzing multiple exabytes of data. In 2004, Google published a paper on a process called MapReduce that used such an architecture. The MapReduce framework provides a parallel processing model and associated implementation to process huge amount of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step).

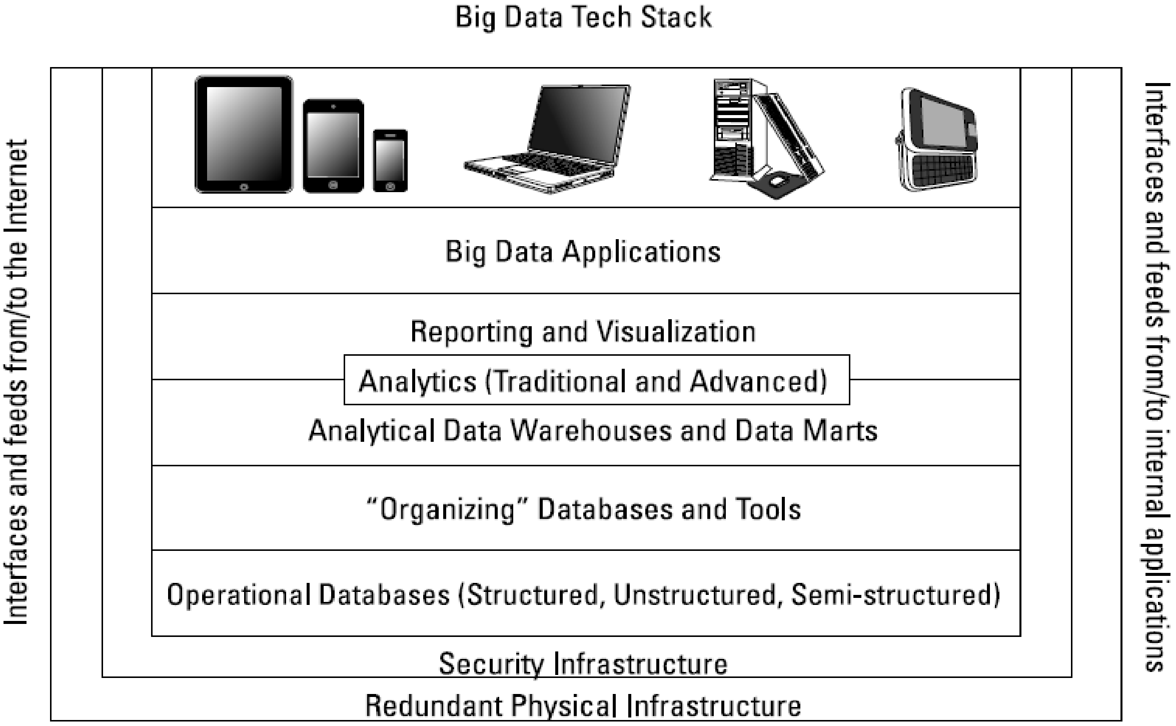
Big Data Tech Stack
Big Data Applications
Reporting and Visualization
Analytics (Traditional and Advanced)
Analytical Data Warehouses and Data Marts
"Organizing" Databases and Tools
Operational Databases (Structured, Unstructured, Semi-structured)
Security Infrastructure
Redundant Physical Infrastructure
Interfaces and feeds from/to the Internet
Interfaces and feeds from/to internal applications

Fig.2 Big Data Architecture

The framework was very successful, so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open source project named Hadoop. MIKE2.0 is an open approach to information management that acknowledges the need for revisions due to big data implications in an article titled "Big Data Solution Offering".[53] The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.

4. BIG DATA SECURITY

With the advent of Big Data comes the risk of greater security breaches as data volumes increase. Many companies are still trying to evaluate the potential of Big Data, let alone investigate the risks associated with Hadoop and the Cloud. In the quest for new ways to house and exploit increasing amounts of unstructured data, companies need to ensure they have mechanisms in place which allow them to meet government compliancy regulations for data protection. Concerns about the security of stored data represent a significant barrier to the widespread adoption of Big Data, and in response, a number of companies are emerging with new products that secure data in ways which are practically transparent to the user. One fundamental method is software and hardware encryption technology that operates on selected data on the fly or across an entire disk. However, software-based encryption adds significant extra load on a database server's CPU. This increases costs and overall complexity, particularly when the solution is required to scale. Before addressing the security issues related to Big Data, it may be interesting to note that you can actually use Big Data Analytics to secure it: "Big Data Security for Dummies" by Solera Networks goes into the details of how to harness the power of Big Data to detect advanced threats and targeted attacks by collecting digital evidence in order to streamline incidence response and help companies integrate Big Data into existing security fabrics. According to Solera Networks, security-conscious organizations are turning to Big Data Security as the newest weapon in their cybercrime arsenals. By collecting all available digital evidence, including raw packets, flow data and files, organizations can uncover advanced targeted attacks traditional security defenses sometimes miss. Organizations are learning to use internal data sources they never knew existed and to extend the value of known data sources by integrating their Big Data Security solutions into their existing security fabric.

ENHANCING BIG DATA SECURITY

In the quest for new ways to store and exploit Big Data, companies need to ensure they have mechanisms in place which allow them to meet government compliancy regulations for data protection, especially for data at rest. Implementations must essentially involve two things. First, secure encryption technology must be used to protect confidential data, in particular Personally Identifiable Information (PII) and Protected Health Information (PHI), but also a company's own Intellectual Property (IP). Second,, careful management of access to the cryptography keys which unlock the encrypted data must be put in place. Growing concerns about the security of stored data are creating new opportunities for IT vendors, and a number of companies are emerging with new products that secure data in ways which are practically transparent to the user. One method being applied is software and hardware encryption technology operating on selected data on the fly or across an entire disk of data at rest. However software-based encryption adds significant extra load on a database server's CPU and costs notably increase, along with complexity, when a solution is required to scale.

DATA PROTECTION METHODS FOR APACHE HADOOP

To address this, the Intel Distribution for Apache Hadoop software includes built-in support for enterprise-class access controls by providing a flexible and efficient framework for managing and controlling user access to data and services by means of existing Kerberos authentication solutions. Administrators can use Intel® Manager for Apache Hadoop software to create and manage access control lists (ACLs) and to authorize individual users for specific data tables and services. A variety of integrated features, such as wizard-based setup and encrypted key exchange, simplify the otherwise complex task of establishing strong, cluster-wide security safeguards.

The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential. The following are only some of the questions that need to be addressed:

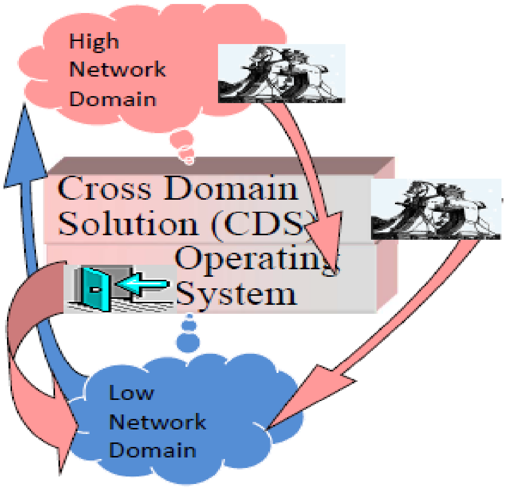

Fig.3.Big Data Security Network Reality

- Data provenance: authenticity and integrity of data used for analytics. As Big Data expands the sources of data it can use, the trustworthiness of each data source needs to be verified and the inclusion of ideas such as adversarial machine learning must be explored in order to identify maliciously inserted data.

- Privacy: we need regulatory incentives and technical mechanisms to minimize the amount of inferences that Big Data users can make. CSA has a group dedicated to privacy in Big Data and has liaisons with NIST's Big Data working group on security and privacy. We plan to produce new guidelines and white papers exploring the technical means and the best principles for minimizing privacy invasions arising from Big Data analytics.

- Securing Big Data stores: this document focused on using Big Data for security, but the other side of the coin is the security of Big Data. CSA has produced documents on security in Cloud Computing and also has working groups focusing on identifying the best practices for securing Big Data.

- Human-computer interaction: Big Data might facilitate the analysis of diverse sources of data, but a human analyst still has to interpret any result. Compared to the technical mechanisms developed for efficient computation and storage, the human-computer interaction with Big Data has received less attention and this is an area that needs to grow. A good first step in this direction is the use of visualization tools to help analysts understand the data of their systems. We hope that this initial report on Big Data security analytics outlines some of the fundamental differences from traditional analytics and highlights possible research directions in Big Data security.

5. CONCLUSION

Concerns about the security of Big Data have slowed the uptake of analytics projects across many businesses. However suppliers are now emerging with new products which secure data in ways which are almost transparent to the user. One of the methods being applied is software and hardware encryption technology operating on selected data on the fly or across entire disks of data at rest. However software-based encryption adds significant extra load on a database server's CPU and notably increases costs, along with complexity, when the solution is required to scale. Intel Distribution for Apache Hadoop software now offers a solution, providing an enterprise-ready software platform for Big Data analytics that is highly optimized for performance, stability, manageability, and security when run on systems powered by the Intel Xeon processor E5 family. By taking advantage of Intel AES-NI technology, the Intel Distribution for Apache Hadoop software accelerates data encryption by up to 5.3 x and data decryption by up to 19.8x, so IT organizations no longer have to choose between performance and security. Advantech platforms based on Intel Architecture for Communications Infrastructure provide a broader product offering with greater configurability, scalability and performance for hardware acceleration across multiple form factors. This enables business to achieve the competitive advantages of Big Data analytics with less risk and with the confidence that their most sensitive data is protected.

6. REFERENCES

[1] www.cloudsecurityalliance.org/research/big-data, Cloud Security Alliance Big Data Analytics for Security Intelligence

.[2] Grasping the Fundamentals of Big Data

[3] Addressing Big Data Security Challenges: The Right Tools for Smart Protection, Trend Micro, Incorporated.

[4] Security – A Big Question for Big Data, *Prof Roger R. Schell, University of Southern California,* Keynote LectureIEEE BIgData 2013, Santa Clara, CA, October 9, 2013.

[5] Enhancing Big Data Security, Networks & Communications Group, www.advantech.com.

[6]Big Data Working Group, Big Data Analytics for Security Intelligence, September 2013.

BIBLIOGRAPHY

Mr. A V S M Adiseshu received his M.Tech (CSE) from Acharya Nagarjuna University, Andhra Pradesh**.** His interested area is Big Data and Data Mining. Presently he working as an Assistant Professor in CSE Department at Sree Dattha Institute of Engineering & Science, Hyderabad, Telangana.

Ms. S. Lavanya Reddy received her M.Tech (CSE) from Jawaharlal Nehru Technological University (JNTU) Hyderabad, Telangana**.** Her interested area is Big Data and Data Mining. & Data Warehousing .Presently she working as an Associate Professor in CSE Department at **Sree** Dattha Institute of Engineering & Science, Hyderabad, Telangana.

Ms. P Hasitha Reddy received her M.Tech (CSE) from Jawaharlal Nehru Technological University (JNTU) Hyderabad, Telangana**.** Her interested area is Big Data, Mobile Computing and Human Computer Interaction. Presently she working as an Assistant Professor in CSE Department at Sree Dattha Institute of Engineering & Science, Hyderabad, Telangana.

Ms. Roshini K, received her M.Tech (CSE) from Jawaharlal Nehru Technological University (JNTU) Hyderabad, Telangana**.** Her interested area is Data Mining and Computer Networks. Presently she working as an Assistant Professor in CSE Department at Sree Dattha Institute of Engineering & Science, Hyderabad, Telangana.