International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7406-7411
© Research India Publications. http://www.ripublication.com

A Detailed Study of Security and Privacy Concerns in Big Data

Palak Mittal
Department of Computer Science and Engineering, MRIU, Faridabad

Mansi Sharma
Department of Computer Science and Engineering, MRIU, Faridabad

Dr. Prateek Jain
Department of Computer Science Accendere KMS Pvt. Ltd.

Abstract

In today's world we are generating tons of data every single day, that much data will just go waste if it's not analyzed and evaluated efficiently. Big data comes up with a number of advantages when it comes to predictive analysis, customer satisfaction, product improvement and many more but also brings challenges like security, complex data and language problems and data security and privacy is one of them. Hence it becomes necessary to identify such problems and look up for solutions so as to achieve the benefits that big data offers. This paper focuses on data security and privacy challenges in big data and provides some solution to improve data's security.

Keywords: Big Data, Security

INTRODUCTION

With the rapid increase of internet, we have been moving from using traditional data like documents and texts to the complex form of data consisting a larger number of videos, photos, complex maps, location data, high quality audio and many more. Data is getting bigger and bigger every single second. Big data is worthless if it is not analyzed from a decision-making perspective area like business, science, engineering, defense, education, healthcare and society at large.

1,820 TB of data creation, 11 million instant messages, 698,445 Google searches, 168 million+ emails and many more forms big data. It would not be a good act of wasting all this raw data and getting nothing out of it. Several organizations, small and big, uses this huge amount of data to know their customer needs, to know what profits their company more, increasing requirements and many more Today all organizations whether huge or little utilize different techniques like hereditary calculations, neural systems, and opinion mining to think about various sorts of information that may help in item and process disclosure, profitability, and strategy making.[2] The data is collected from various sources like social networking sites. Advances in information stockpiling and mining advances make it conceivable to protect expanding measures of information produced straightforwardly or in a roundabout way by clients and investigate it to get significant new bits of knowledge. Let us understand the importance of big data examples, these you must have noticed a new change in Gmail. When the user is willing to reply an email, he is provided with some options which in a way matches the context and topic of mail like Thank you, reply me etc. Another example could be social networking sites, we are uploading millions of posts, photos, videos, putting so much about us, where do you think all this data go, its analyzed. What about all the online surveys we answer on YouTube, Google Rewards, Websites, even this information is collected and computationally analyzed. Big data isn't only used for finding trends and patters but also it helps any organization to understand its customer's need in a better way. Today many software companies provide their customers platform where people can discuss issues as well as benefits of a certain software or any recently update or change released by company. This all data generated by active users acts as a feedback and provides company with larger number of options. As discussed above the increasing amount of big data collection is the result of increasing connectivity among people through internet. Today companies are getting real time information from users with the blessing of advances in internet and connectivity from all kinds of electronic devices like mobile phones, laptops, personal PCs, and many more. Advances in technology and sciences have given us tools that help us with storing all this raw data and convert it into something fruitful. So now we know why is big data considered to be of such great importance and hence big data analyzing becomes important too.

It becomes equally important to introduce this term "Data Mining" when we are discussing big data and getting efficient information from it. Data Mining also called "Knowledge Discovery" helps us recognizing different patterns, trends and relations among different chunks of data. Just like some people mine to get precious stones, rocks, minerals and ores, data mining provides us with the precious and amazingly helpful information out of raw data. It improves understanding and boosts decision making of organizations. The more one analyzes the raw data the better results are obtained. "Data Warehouse" is like a repository where data can be stored from different data sources. Then another term comes "OLAP", OLAP or Online Analytical processing helps users to select data from databases and analyze it. Suppose a sports shoe manufacturing company's employee wants to view the analyzed data of the number of shoes sold and their model number, and calculate the revenue, loss and profits, OLAP simples the task of employee as the data in OLAP is stored in multidimensional database. But with this huge amount of data collection comes big challenges too. Big data is considered complex and is also one of its characteristics which makes it little challenging to work with big data.

THE V'S OF BIG DATA

These V's are considered to be important when it comes to characteristics of big data. All these V's tells us the different aspect and feature

1. Volume: The term "Volume" here is concerned with the amount or quantity of data. This huge amount of data becomes more of a challenge and less beneficial it comes to analyzing and accessing this all data, this means we need more tools and resources to deal with this huge data.

2. Velocity: Velocity is associated with the speed with which the data transfers from one point (server or any user interface) to another point, analyzed and collected. The speed with which data these days is being analyzed gives us real time information. [8]

3. Variety: Big data is collection of a huge and various type and this is what "Variety" here means. Today we don't have only limited type of data like documents or just audio. Data today is unstructured.

4. Value: Value is also a characteristic of big data. The huge amount of data available is useless unless it gives useful information. Value means the worth or importance of the data or information being extracted.

5. Veracity: Now, if some data is lost in big data then it is not considered to be an issue as there is a possibility of that that data loss can be covered by some other data and this characteristic is termed as Veracity.

6. Validity: Veracity will only be efficient if the data has validity characteristic. Here Validity means the correctness and accuracy of the data. Incorrect data can sometimes lead to the incorrect decision making.

7. Visualization: big data is only useful when one can extract valuable results from it. Visualization means representing data using graphs, charts, and plots which is much easier to understand and derive conclusions. [6]

8. Volatility: If data is no longer capable or efficient for giving analytical results then that data should no longer be used. One needs to know at what point the data available is of no use. Using the same old data for a long period of time will give ineffective and insufficient results**.** This information is basically data which can be structured, semi-structured or unstructured.

Figure 2: Structured and Unstructured Data [5]

Structured data means the data contained in a database. Semi-structured data is metadata. Unstructured data means audio files, video files, power point presentations, collaboration software, and instant messaging. Many data mining techniques are closely related to some of the machine learning techniques.

Big data is associated with large volume, different types of data structured or maybe unstructured with known sources or sometimes unknown (security issue) and having decentralized control and can give us all sort of information for present technology and can even lead us to further more innovations on technology all we have to do is explore all those possibilities that big data holds but to explore all the data is not a handy task hence now we will discuss briefly about systems that let us explore those possibilities.

TOOLS FOR BIG DATA

There are several tools available that helps in the managing of big data. Due to the volume and variety of big data it becomes quite difficult to manage data and to operate on it. Cassandra, Plotly, Apache Hadoop, Apache Spark, Wolfram Alpha, Open Refine, Neo4j, Rapidminer are such tools which helps us make efficient use of big data.

Hadoop

Apache Hadoop is a java programming based open source framework used for processing big data sets. Hadoop 0.1.0 was first released in April 2006. MapReduce, Hadoop Common, Hadoop Distributed File System (HDFS), Yarn, Hive are some of the Hadoop components.

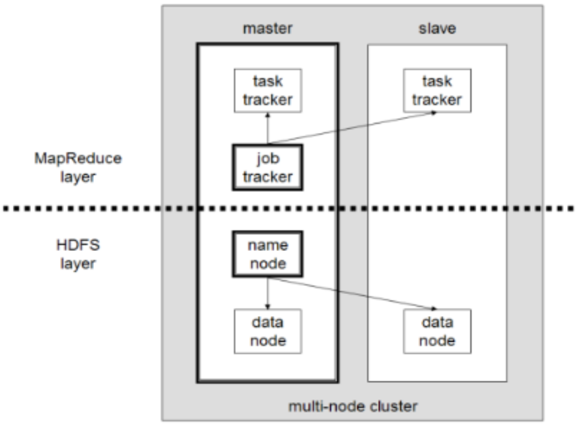


Figure 3: Hadoop Architecture [8]

1. HDFS is java based the primary storage system of Hadoop ecosystem. It provides fault tolerance, scalable and cost-efficient data storage.
2. Yarn provides scalability, compatibility and cluster utilization. It provides dynamic allocation of cluster allocation.
3. Hive provides data summarization, queries and analysis. It is a SQL like interface through which you can query data from database.

Benefits of Hadoop

Hadoop has proved to be extremely beneficial in terms of data sets processing at big level. Hadoop's flexibility is one of the many reasons that make it a good choice. One does not have to create structured schemas to store data. Hadoop being open source solves the money issue. Hadoop is a way of storing enormous data sets across distributed clusters of servers. It is robust i.e. it will continue to work even when your individual cluster or server fails as it will send the work to other nodes. Diagram below shows the components of Hadoop ecosystem.

Apache Spark

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It depends on Hadoop MapReduce and it stretches out the MapReduce model to effectively utilize it for more sorts of calculations, which incorporates interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application. Spark is intended to cover an extensive variety of workloads, for example, batch applications, iterative algorithms, interactive queries and streaming.

Spark Core, Spark SQL, Spark Streaming, Machine Learning Library (MLlib), GraphX are components of Apache Spark.

1. **Spark Core**: Spark core has responsibility for monitoring and scheduling the tasks on spark cluster. It also handles basic input-output functionalities.
2. **Spark SQL**: Spark SQL introduces SchemaRDD, a new data abstraction which supports semi-structured and structured data processing. It supports full compatibility with Hive data.
3. **Spark Streaming**: It allows fault-tolerant and scalable stream processing for live streams of data.
4. **MLlib**: MLlib is machine learning library and consists of high-speed and high-quality algorithms. It is nine times as fast as the Hadoop disk based version of Apache Mahout before Mahout gained a Spark interface. [9]
5. **GraphX**: It is a graph processing framework. And it has application program interface (API) for graph computation.

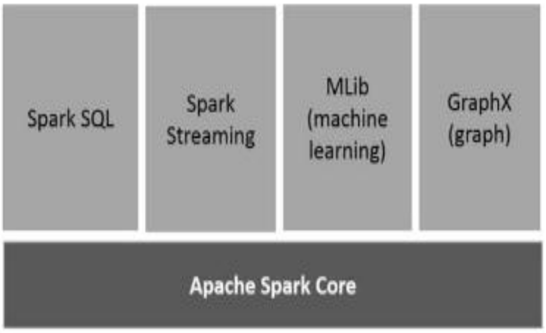


Figure 4: Apache Spark Components. [9]

Apache Cassandra

It is a NoSQL database. It is scalable, and has high-performance distributed database to handle large amounts of data. We can store and retrieve data other than tabular relations with the help of a NoSQL database. The qualities of this database are that it is schema free, has a simple API, is consistent, supports easy replication, and can handle large amounts of data.

Node, data center, cluster, commit log, Mem-table, SSTable, and Bloom Filter are the components of Apache Cassandra.

1. **Node**: A node is an essential part of Cassandra Architecture, this is where information is stored and queried upon.

2. **Data Center**: Data center can be of two types- physical or virtual data center. It is a collection of related nodes.

3. **Cluster**: It contains one or many data centers. It can span physical locations.

4. **Commit Log**: Ever node in the cluster has a commit log. It is a kind of crash recovery system. Data is written in these logs for durability, it can be archived, deleted, or recycled.

5. **Mem-table**: The data is written in mem-tables after commit logs. There will be multiple mem-tables for single- column families.

6. **SSTable**: When the contents reach a threshold value data from the mem- table is sent to a disk file which is SSTable.

7. **Bloom Filter**: These are accessed after every query. Bloom Filter is an algorithm to test if an element is a member of the set.

CHALLENGES WITH BIG DATA

Big data offer organization with enormous however terabytes or petabytes of information streaming each day to an association have uncovered that present foundations and designs are not adequate to address the difficulty. Big Data is very useful for organizations but there are also many challenges of big data. The enormous volume of data from different sources makes it vulnerable to thefts, or malfunctions.All information stored in a company's database may not be of use. It's a challenging thing to sort the useful data from the useless data. This makes the analysis work very time consuming and incurs significant costs. Data privacy and security and another one of the concerns in big data. Details of employee of any organization can be compromised if security of that data is compromised. In 2014, when a group of hackers got into the Sony data they had access to a large number of employee information and personal details. This incident shows us the importance of securing data.

As briefly mentioned above while dealing with big data one can face some challenges. It becomes necessary to understand these difficulties and look for solution. The figure above shows some existing challenges.

The characteristics of big data namely Volume, Velocity, Veracity and Variety also pose a challenge.

1. **Language**: It would be easier to handle and analyze big data if we have an easy language just like languages present in other fields.
2. **Reliable Data**: As by definition big data is a collection of huge data chunks, sometimes sources of the data can be unknown and not reliable. Hence it is not advisable to completely rely on such data.
3. **Complex Data**: As mentioned above, structured and unstructured data, the combination of both makes data quite complex and hence sometimes while analyzing data one can face troubles.
4. **Technologies**: Once realized the amount of information travelling across the internet, specialists started to question how to handle this amount of data. Hence there is need of better technologies.
5. **IT Infrastructure**: To store data efficiently and safely there is need of a proper and a robust architecture.
6. **Data Privacy**: Data is important to all whether a company or any individual. Hence securing the data becomes important. No organization wants to lose data that it collected over years.

DATA SECURITY AND PRIVACY

Today, almost all our electronic gadgets (be it mobile, laptops or tablets) are always associated with web and consequently sending and sharing data has a major part thus has the security and protection. With increment in gathering of information the subject of security likewise rises. Do we have enough and productive advancements to handle with the security and protection issues. The following are specified couple of focuses which clarifies why huge information has security and protection issues.

1. **Access Controls**: It is fundamentally vital to give a framework in which encoded verification/approval confirms that clients are who they say they are, and figure out who can perceive what.
2. **Non-relational data stores**: Think No SQL databases, which without anyone else's input normally need security.
3. **Data provenance**: fundamentally concerns metadata (data about data), which can be greatly useful in figuring out where data originated from, who got to it, or what was done with it. More often than not, this sort of data ought to be examined with excellent speed to limit the time in which a breach is active.

Presently utilized security arrangements, for example, firewalls and DMZs may not be able to provide big data with the required security strength.

Cloud Secure Alliance (CSA), a non-profit organization has categorized the security and privacy problems into four categories which cover the entire range of the big data lifecycle:

1. **Infrastructure security**
   - Secure distributed processing of data
   - Security best actions for Non-Relational Data-Bases
2. **Data privacy**
   - Data analysis through data mining preserving data privacy
   - Cryptographic solutions for data security
   - Granular access control
3. **Data management and Integrity**
   - Secure data storage and transaction logs
   - Granular audits
   - Data provenance
4. **Reactive security.**
   - End-to-End filtering and validation
   - Supervising the security level in real time

Sources of data production (devices), the data itself, data processing, data storage, data transport and data utilization on different devices. [4] A part of Big Data security and protection is identified with Internet of Things. Internet of Things is a proposed advancement of the Internet in which regular objects have network connectivity, enabling them to send and get data.

The gigantic increase in the number of connected devices like cars, lighting systems, refrigerators, telephones, glasses, traffic control systems, health monitoring devices, SCADA systems, TVs, home security systems, home automation systems and so forth has prompted makers to push to the market, in a brief timeframe, an extensive set of devices, cloud frameworks and mobile applications to exploit this opportunity.

HP led an investigation on market available IoT solutions and concluded that 70% of them contain security issues to be specific privacy issues, insufficient authorization, absence of transport encryption, unreliable web interface and inadequate software protection.

HP began a venture titled "OWASP Internet of Things Top Ten" which intends to assist IoT providers with identifying the best ten security IoT device issues and how to keep away from them.

The ten security issues which were recognized are:

1. **Insecure web interface**: which can enable an aggressor to exploit an organization web interface and acquire unapproved access to control the IoT device.
2. **Insufficient Authentication/Authorization:** can enable an aggressor to misuse a bad password policy, and break weak passwords.
3. **Insecure network services:** This can lead to an attacker exploiting unnecessary or weak services running on the device.
4. **Lack of Transport Encryption**: enabling an attacker to spy information in travel between IoT devices and support systems.

5. **Privacy Concerns**: most IoT devices and support systems gather personal information from clients and fail to protect that data.
6. **Insecure cloud interface:** without appropriate security controls an aggressor can utilize different attack vectors to get to information or controls by means of the cloud website.
7. **Insecure mobile interface**: without appropriate security controls an attacker can utilize various attack vectors to get to information or controls by means of the mobile interface.
8. **Insufficient security configurability:** due to lack or poor configuration mechanisms an attacker can access data or controls on the device.
9. **Insecure software/firmware**: attackers can take advantage of unencrypted and unauthenticated connection to hijack IoT devices updates, and perform pernicious update.

10. **Poor physical security**: in the event that the IoT device is physically accessible then an attacker can utilize USB ports, SD cards or other storage means to get to the device OS and any data put away on the device.

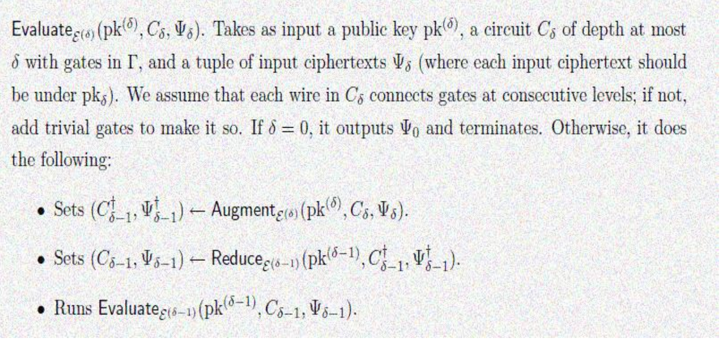
SOLUTION

It can be a little hard to find magical solutions that will make all the problems go in one go. Traditional encryption methods are satisfactory to protect static data, however are not sufficient when data analyzing is involved.

Homomorphic Encryption

Homomorphic encryption invented by Craig Gentry is a type of encryption which enables calculations to be completed on cipher text and produce outcome which, after decoding matches the aftereffect of operations performed any without encrypted data.

Let's understand this with an example. Suppose, Anita wants to multiply two numbers 6 and 7 but does not know the multiplication process. She asks John to help who knows multiplication but also, she doesn't trust John with her data. So, she encrypts her data 6 and 7 to 89 and 50 respectively. Now John gets data 89 and 50, multiplies them and give that (encrypted) result to Anita. Now Anita can decrypt the result of John and finds the answer 42. This way she got the answer and also didn't lose any data. This is what Homomorphic encryption does to your data. This method basically encrypts plaintext and it supports arbitrary computation on ciphertext.



$\mathsf{Evaluate}_{\mathcal{E}^{(\delta)}}(\mathrm{pk}^{(\delta)}, C_\delta, \Psi_\delta)$. Takes as input a public key $\mathrm{pk}^{(\delta)}$, a circuit C_δ of depth at most δ with gates in Γ, and a tuple of input ciphertexts Ψ_δ (where each input ciphertext should be under pk_δ). We assume that each wire in C_δ connects gates at consecutive levels; if not, add trivial gates to make it so. If $\delta = 0$, it outputs Ψ_0 and terminates. Otherwise, it does the following:

- Sets $(C^\dagger_{\delta-1}, \Psi^\dagger_{\delta-1}) \leftarrow \mathsf{Augment}_{\mathcal{E}^{(\delta)}}(\mathrm{pk}^{(\delta)}, C_\delta, \Psi_\delta)$.
- Sets $(C_{\delta-1}, \Psi_{\delta-1}) \leftarrow \mathsf{Reduce}_{\mathcal{E}^{(\delta-1)}}(\mathrm{pk}^{(\delta-1)}, C^\dagger_{\delta-1}, \Psi^\dagger_{\delta-1})$.
- Runs $\mathsf{Evaluate}_{\mathcal{E}^{(\delta-1)}}(\mathrm{pk}^{(\delta-1)}, C_{\delta-1}, \Psi_{\delta-1})$.

Figure 3: Homomorphic Encryption [7]

Benefits

1. Cloud Security: cloud providers no more have to decrypt data to run queries on customer's demand.
2. Individual's data is more secure.
3. Because of Homomorphic encryptions a user can store its data on any server (untrusted or trusted). It also allows users to search the data i.e. retrieving the encrypted data. IBM received patent and created an open source Homomorphic encryption method called "HELib" which stands for Homomorphic encryption library.

CryptDb

CryptDb is an intermediate (proxy) between the client and database. The user send ordinary (plaintext) query to the database and cryptDb sends a new query (performs encryption) to the database. CryptDb receives the encrypted result back from the database and then it decrypts it and sends it the user who can now easily understand the result. CryptDb assembles several Homomorphic encryption techniques so that user can perform SQL (database queries) operations with his data. The query executing system for encrypted data is almost similar to that of plain or simple query although queries like selection, aggregates, orders and many such are performed on ciphertext using modified operators. CryptDB's proxy stores a secret master key MK, the database schema, and the current encryption layers of all columns. The DBMS server sees an anonymized schema (in which table and column names are replaced by opaque identifiers), encrypted user data, and some auxiliary tables used by CryptDb. [7]

Benefits:

1. In case the user feels that the encryption (onion layer) type is not sufficient according to the sensitivity of the data then the data will be processed in some different fashion for example by using SQLite.
2. The developer can specify the minimum level of encryption types according the sensitivity of data and hence can ensure that the data will not be exposed.
3. It does not expose the additional data ensuring the security of the database.

ISSUES WITH HOMOMORPHIC ENCRYPTION AND CRYPTDB

Homomorphic Encryption	**CryptDb**
1. Immense computational requirements: this method is not practical, sometimes be not efficient for the technology that we use these days. 2. Implementations: there are some other encryption schemes provided by some companies which are believed to be un-hackable and more secure. 3. Customer never knows: Sometimes when a customer demands some queries to be run on data, he has to believe the results of cloud provider as he cannot see the plaintext himself.	1. Space requirement can be an issue as every encryption type for all columns requires space also each onion layer being larger than the previous one. 2. It also does not provide the logged-in information that who and how many are logging in the database and are accessing it. 3. Also since the decryption of data depends on the key if the key is lost the data accessing power is lost.

CONCLUSION

Big data is vast collection of data transferring across over world every single day. This information is analyzed by organizations to discover different patterns and trends which helps them to improve their services. Many tools like Apache Hadoop, storm, Neo4j helps organizations helps in data management at vast level. With big data comes big challenges, Language, lack of IT architecture, data complexity is one of them. This paper focuses on data privacy and security issue. No organization wants to lose the important data collected over many years and hence taking care of data becomes important. Concept of homomorphic encryption gives us a way of completely encrypting our data into ciphertext and the operating with the same encrypted data which also has raised some implementation issues that whether this method id that convenient as it is said to be. We discussed the benefits as well as issues. Then there came CryptDb which promises to retrieve the encrypted data and decrypt it for user without showing the entire database.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. Prateek Jain, Accendere Knowledge Management Services Pvt. Ltd., for his valuable comments that led to substantial improvements on an earlier version of this manuscript

REFERENCES

[1] Katina Michael and Keith W. Miller, ”Big Data: New Opportunities and New Challenges”, Computer (Volume: 46, Issue: 6, June 2013)

[2] Xindong Wu and Gong-Qing Wu, “Data Mining With Big Data”, IEEE Transactions on Knowledge and Data Engineering (Volume: 26, Issue: 1, Jan. 2014)

[3] José Moura[1,2] and Carlos Serrão[1], “Security and Privacy Issues of Big Data”, 1 ISCTE-IUL, Instituto Universitário de Lisboa, Portugal 2 IT, Instituto de Telecomunicações, Lisboa, Portugal.

[4] Raluca Ada Popa, Catherine M. S. Redfield, Nickolai Zeldovich, and Hari Balakrishnan, “CryptDB: Protecting Confidentiality with encrypted Query Processing”, MIT CSAIL page 4(section 3), October 2011

[5] Chanpreet Kaur, “Data and Data Warehouse: Scope, Challenges, Future”, February 18, 2015.

[6] Suhail Sami Owais and Nada Sael Hussein, “Extract Five Categories CPIVW from 9V’s Characteristics of the big data”, IJACSA, Vol. 7, No.3, 2016.

[7] Craig Gentry, ”A Fully Homomorphic Encryption Scheme”, September 2009.

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

[9] V Srinivas Jonnalagadda, P Srikanth, Krishnamachari Thumati, and Sri Hari Nallamala, ”A Review Study of Apache Spark in Big Data Processing”, International Journal of Computer Science Trends and Technology (IJCST)–Volume 4 Issue 3, May - Jun 2016.