Study on Big Data Security

Sayali Suresh Sherkar

Student

Department of Master of Computer Application

ASM IMCOST, Thane, Mumbai, India

Abstract:

This research paper talks about the study done on security of data in big data. As the name suggests Big Data means data in the large volume. Which is impossible to store in our traditional databases. The data can be in any form like audio, images, text, documents, and multimedia files etc. Most of the data is generated by social media like Facebook, Instagram. There are other industries who generates data in large amount. Real time data contains redundant data, noise, incorrect, incomplete data. Therefore, data cleaning is required. In this process, duplicate, incorrect & incomplete data is removed. After that further processing is done on data. Security comes next. When it comes to data, security is most important. Despite the increasing abundance of data, businesses face several challenges and opportunities. While better analysis allows businesses to make better decisions, it also has certain drawbacks, such as security concerns. In this paper, we discuss security issues faced in big data.

Big Data

Big Data in layman terms is defined as Large amount of data. It can be in structured, semi structured, and unstructured form. Here structured means data properly in the form of table and rows. Example of semi structured data can be XML file. Unstructured data is data which is stored in its raw format i.e. Its original form. For example, Audio, video, images files.

To increase profit businesses, use big data in to study customer behavior, pattern which can help them provide better services to their customer. Businesses who use it effectively can make faster and more informed business decisions, giving them a possible competitive advantage over those that don't.

Big data, for example, gives important customer insights that businesses can utilize to improve their branding, sales, and campaigns to increase customer engagement and conversion rates. Historical and real-time data can

be evaluated to assess changing consumer or corporate buyer preferences, allowing organizations to become more responsive to client demands and needs.

The three Vs of big data

1. **Volume:** first v refers to the volume of the data. Quantity of data. Which is in terabyte & petabyte.
2. **Variety:** As the name suggests, the data generated is in various form. Traditionally data used to be in table & row format. But with the rise of big data now data is available in audio, video, image meaning in unstructured format.
3. **Velocity**: speed of data generation. Large amount of data usually comes from social media.

Benefits of Big Data –

1. **Cost optimization:** big data helps in reducing the cost of data storing, data processing and data analyzing. It can be helpful in finding cost effective solutions.

2. **Understanding the Market:** Big data can help us understand the current market situation. By studying the large amount data, finding patters in it we can analyze the situation and business can plan their strategy.

3. **Improvement in Efficiency:** Big Data approaches have the potential to significantly improve operational efficiency. Big Data technologies can acquire massive amounts of usable customer data by engaging with customers/clients and soliciting their valuable input. This data can then be analyzed and interpreted to identify relevant trends (client tastes and preferences, pain points, purchase behaviors, and so on), allowing firms to create personalized goods and services.

Issues In Big Data:

Security

As the data of an organization increases, it becomes expensive & hard to store it. As per **Risk-Based Security Mid-Year Data Breach report** , nearly 4.1billion of data was exposed due to breach in data.

Privacy

To make sure data is secured from theft and prevent it from misusing it, we have laws. But organization who you have trusted to keep your information safe can use it for their analyzing or studying purposes which can violate the laws.

Wrong Analysis of Data

Reading the data in wrong way can cause the problem in planning the strategy. A good example can be Google's Flu trends. This project was made to get the correct details of flu outbreaks in various places based on the google searches. It gave great results but later the predications were not accurate, and it could not pick up the 2009 pandemic.

Security In Big Data:

Big Data security can be said as the equipment & precautions taken to guard data & analytics. The main aim security is to defend against all attacks, thefts, and other malicious activities that can damage precious data in 2019. Data security is crucial. In today's world no data is secure. We need to take some extra measure to secure the data. These threats encompass the robbery of records stored on-line, ransomware, or DoS attacks that could crash a server. the problem may be even worse while groups keep facts that is sensitive or private, which includes customer statistics, credit card numbers, or even surely touch details. moreover, attacks on an employer's massive facts storage should motive serious economic repercussions consisting of losses, litigation expenses, and fines or sanctions.

Common Security Issues:

1.False Data Generation: cybercriminals can feed false data into the Data Lake. This way they gain the opportunity to penetrate your system and hack it or it maybe used for wrong purpose.

2.Untrusted Mapper: if untrusted & unauthorized people have access to the organization's code, they can make changes to mapper which is responsible for data processing & storing. They can make mappers to function incorrectly and have the desired results produced.

3.EndPoints – the security logs are drawn from an endpoint. What is important is that we need to verify whether these are authenticate or not.

How can we keep our data secure?

1. **Securing Distributed Programming Framework:**

   The Cloud Security Alliance (CSA) suggests that methods such as Kerberos Authentication should be use while compiling the security policies. Then, we should conceal all the personal data by separating it from the all the data so that privacy won't be invaded. Then we can access Files.

2. **Securing Non-Relational Data:**
   To avoid attack on non-Relational databases, we can make use of encryptions or password hashing, using algorithms to secure data. (AES, RSA etc.)

3. **Data Storage Security and Transaction Logs:**
   Storage management is a vital component of Big Data security. The CSA advises employing signed fully utilized to give a unique identity for each digital file, and using a technique known as secure untrusted data repository (SUNDR) to identify unauthorized file alterations by hostile server agents. There are so many techniques noted we can make use of those.

4. **Endpoint Verification and Sorting:**
   Endpoint security is critical, and your enterprise may begin by utilizing trustworthy certificates, doing resource testing, and linking only trusted endpoints to your network using a smart phone management platform (on top of antivirus and malware protection software). From there, you can leverage statistically similarity detection methods and outlier identification techniques to filter dangerous inputs while guarding against Sybil attacks and ID-spoofing assaults.

5. **Cryptography:**
Cryptography is a technique for safeguarding information and interactions by using codes so that only the people who need the information can comprehend and process it. As a result, Unauthorized access to information is prohibited. The suffix graphy is "writing" in English and the word "crypt" in English is "hidden. Cryptography techniques are derived from mathematical principles and a collection of rule-based calculations known using algorithms to alter messages in ways that make decoding challenging. These algorithms are used to generate cryptographic keys, digitally sign documents, verify data privacy, and safeguard secret transactions such as credit card and debit card transactions.

Security Tools

Key Management Centralized

The process of safeguarding secret key against loss or misuse is known as key management. Centralized key management performs better than distributed or application-specific key management. Centralized management solutions use a single point of access to safeguard keys, audit logs, and policies. A dependable key management system is critical for businesses that handle sensitive information.

Detection and prevention of attacks

Big data's dispersed architecture is advantageous for infiltration attempts. By monitoring network traffic, an Intrusion Prevention System (IPS) allows security teams to defend large data platforms from vulnerability attacks. The intrusion prevention system (IPS) frequently resides just behind the firewall and isolates the infiltration before it does actual damage.

Control of User Access

A basic network security tool is user access control. Inadequate access control mechanisms can be fatal for large data systems. A strong user control policy must be built on automated role-based settings and policies. Policy-driven access control safeguards big data platforms from insider threats by managing complicated user control levels, such as numerous administrator settings, automatically.

Encryption

Big data encryption software must protect data at rest and in transit across massive data volumes. Companies must additionally encrypt both user-generated and machine-generated data. As a result,

encryption technologies must support several big data storage formats, such as NoSQL databases and distributed file systems such as Hadoop.

Conclusions:

In today's world, where billions of data is generated every hour, big data is prominently used to study the customer behavior, to plan next strategy for business, to check the progress of product etc. But we should make sure that the data we are storing is secured enough. The customer trusts organization to keep it safe from external attacks**,** social engineering, data theft.
Following are the points that needs to be keep in mind:

- Big Data is very important to big organizations to study their customers.
- Securing Data is crucial. For that, necessary measures need to be taken.
- Encryption algorithms like 'Quantum Proof Encryption', 'Homomorphic Encryption' which makes it impossible to hack.
- Following CSA guidelines to keep data secure.
- Make sure that IT team keeps an eye for unusual activity in network to avoid breach.

Reference:

1. https://www.epcgroup.net/what-is-big-data-security-challenges-for-organization-data/
2. https://www.dataversity.net/big-data-security-challenges-and-solutions/#
3. https://techvidvan.com/tutorials/big-data-security/
4. https://www.integrate.io/blog/big-data-security-concerns/
5. https://www.sisense.com/glossary/big-data-security