Ramla Suhra, 2024, 12:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science,
Engineering and Technology

An Open Access Journal

Enterprise-Wide Data Security Strategies for Big Data Analytics

Ramla Suhra
Staff Software Engineer, H-E-B Digital, H-E-B, TX, USA

Abstract- With the technological advancement in this era, data is growing at the pace more than ever, so does the value of data in decision making. When data driven decision making has become the strategic priority, organizations want to do collect the data which includes personal information which can help them make better decisions through. Customers on the other end are concerned about the security of their personal information and habits. Moreover, government has started strengthening the regulations on data management. Companies are now forced to be responsible and transparent about their data practices. Hence data privacy has become a crucial part of business operation. The whitepaper reviews the common areas to addressed while dealing with secure data to ensure that the data is protected and compliant. There is also an analysis on the common the risks faced by organizations while storing and processing big data with personal information for further analytics. The paper also explores options to mitigate the risks and suggests opportunities to improve in future. There are also attempts to share some business use cases and examples that can be referred by the readers interested in this topic.

Keywords- Data analytics, Data privacy, Big data, Data governance.

I. INTRODUCTION

Humans and machines generate massive amounts of data daily. Big data refers to extremely large and complex data which cannot be processed by traditional processing tools and cannot be stored and analyzed easily. Big data includes structured data like transaction logs and inventory databases, unstructured data like social media data, IoT data etc. With the advent of artificial intelligence and machine learning there are also semi structured datasets which are used for Model training. With the introduction of storage in cheap distributed storages, the data population and storage are growing at an unpredictable pace. As the data explosion continues, the need for analysis of this data has become more and more crucial for any organization.

Big Data Analytics helps companies with understanding the latest trends and patterns in their business domain and enables them to remain competitive. Unlike reactive approaches towards decision making, big data has paved its way to empowering organizations in futuristic predictions by integrating Artificial intelligence to the big data. For example, learning customer behavior can predict an increase in the sales of the company. Social media data collection helps companies with valuable insights into the customer as well as market trends.

Data collection in enormous volumes to gain value out of it has started raising concerns from the customers as they fear intrusion on their privacy if security of this data is compromised. As if to prove this right, the exploding amount of data has increased privacy breach incidents. There have been recurrent series of data breach incidents in the past couple of years, some examples being Equifax and Yahoo data breaches [2][3]. As a response to the rising concerns and questions, several laws and regulations have been brought into effect by governments, example being GDPR and CCPA regulations. Due to their irresponsible handling of sensitive information, the companies were forced to pay significant settlement fees because of the data breaches they caused. It has become imperative that companies take more strident efforts to roll out privacy-friendly products and thereby reducing the risks associated with sensitive data processing.

This research paper reviews data security risks, their challenges and solutions for stakeholders as well as future improvements that can be adopted by organizations to ensure that they are safe from the potential breaches and their aftermaths.

II. COMMON RISKS AND CONSEQUENCES

Storage and processing of sensitive data can cause significant risks for companies. Primary risks are as listed below:

Privacy Breaches

Unregulated data management can cause sensitive data including personal information exposure to unauthorized parties. The consequences of privacy breaches can be significant for companies, both financially and reputationally.

Cyber Attacks

Unprotected data and systems are targets for cyber-attacks. They are malicious acts with the intent of stealing data and infrastructure and cause damage to the company’s operations. Some examples are ransomware and malware attacks.

Compliance Violations

If data shared or exposed contains sensitive information, sharing them without proper controls could violate data protection regulations.

The consequences of all the above risks are severe for companies. They include the following:

Financial Loss

- Companies may face hefty fines imposed by legal and regulatory departments.
- It will result in revenue loss when customers lose their trust in them.
- Insurance premiums can be increased in the occurrence of an incident.

Damage of Reputation

- Customers may lose trust in the company’s data management practices and their ability to secure their sensitive data.
- There will be negative publicity due to such incidents and it could harm the brand value of the company.

Business Interruption

- In the case of data breach, company’s operations can be interrupted to research and recover from the breach.

Decreased Competitive Advantage

- Customers or vendors will be hesitant to work with them after the incident.
- Competitors can take advantage of this incident to their benefit.

Big Data Security Issues

Modern Big Data has several major security issues in relation to their sensitivity. We must first review them thoroughly to understand how to secure them efficiently while analyzing them.

Protecting Sensitive Data

In addition to the data volume, big data storage and analysis can also help us gain more associated insights. For example, retail companies store massive datasets of customer purchase information. They also collect market trends which are popular in some product categories. These two datasets when combined, they can create targeted marketing campaigns on the new trending product in that category for customer segments who usually purchase the product category frequently. While this might seem to a business problem being solved effectively, such data analysis often generates more information from the original datasets.

User’s Personal Information

Social media sites often know more about their users than any others. Many social media ads predict interest in music or language learning before a person starts searching for courses. If leaked, this information can reveal extremely sensitive data.

Identity Disclosure

Many big data services capture customer’s behaviors. Hiding one’s information becomes extremely difficult over a period due to the trackability created by such data collection.

III. DATA SECURITY CHALLENGES

Data security challenges can be broadly classified into four categories.

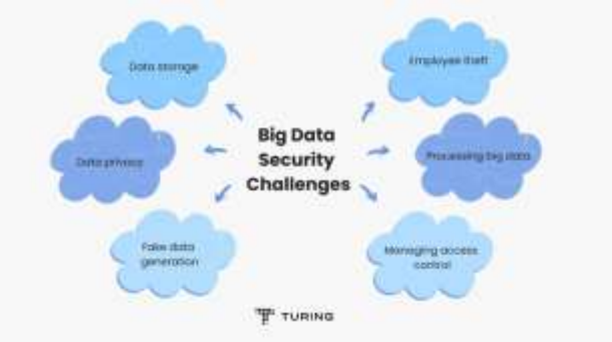

Figure 1: Data Security Challenges [4]

Data Management Challenges

Data usually gets stored in raw format if not labelled with proper sensitivity indicators can be treated as public data which does not require protection from unauthorized access. Performing continuous analytics on datasets can transform them into different formats. After these cycles there is an increased chance of losing sensitivity tracking on the dataset. Due to the nature of big data, it becomes difficult to protect all the data which has evolved over a period.

There are several big data processing tools to help companies process data quickly and efficiently. The tools are provided as Software as a service (SaaS). This can lead to storage of data in the product company’s cloud data storage. Controlling data security issues on infrastructure not owned by the organization is a challenge.

Distributed nature and large volumes of big data make it difficult to implement effective access controls. The sharing of data with third-party applications and services further complicates the issue, increasing the risk of unauthorized access.

Simulated fake data is another challenge in big data security as they can corrupt the data and cause data loss. Employees who have access to the data also in some cases turn into malicious actors by stealing data or causing data corruption which will have severe consequences for the enterprise. Data encryption implementation is often overlooked if they are not reviewed.

Regulatory and Compliance Adherence

A lack of understanding of regulatory and compliance standards can lead to data analytics teams not following these requirements. Data infrastructure teams often provision infrastructure without adhering to established standards.

Technological Complexity

The modern big data technology stack is growing by data making it hard to secure all the components and the data stored within it effectively. The latest evolution of large language model is an example. They follow complex architecture and deal with sensitive large volume data. Since they also rely on third party libraries and frameworks, they are vulnerable to security threats. Moreover, there is a rapid change of technological landscape when it comes to big data analytics, machine learning and artificial intelligence.

Organizational Approach

Lack of awareness of responsibilities related to data security risk prevention and control across various teams in the organization is a major challenge.

Compliance and risk management departments frequently struggle to communicate the relevance of data security to other organizational stakeholders. Allocation of enterprise budget is not well understood and hence there can be scrutiny on various initiatives.

Data Security Stakeholders

We would like to explore the relevant solutions that can help an organization navigate through these challenges and become successful while trying to tighten their data security control and measures. Before diving deeper understanding the stakeholders involved in the data security and compliance landscape is important.

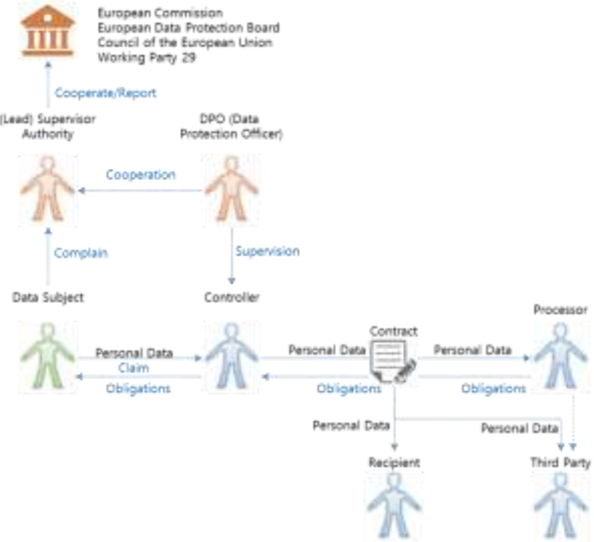

Figure 2: Data Security Stakeholders with Compliance Workflow[5]

Data Subject

A data subject is any person whose personal data is being collected, held, or processed. "'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;"[6]

Data Controller

Data controllers own and operate the data storage and processing systems. Data controllers encompass different teams and they as a unit oversee data retrieval, storage, processing and sharing the data within the organization and as well as with third party entities.

Data Protection Officer

The primary role of the data protection officer is to ensure that their organization stores, processes and shares the personal data of its staff, customers, providers or any other individuals in compliance with the data protection rules.

Lead Supervisor Authority

An independent public authority established by a member state. They are responsible for monitoring the application of data security. Controllers must notify any personal data breach to their national supervisory authority without undue delay.

Data Controller Role and Strategies

Data Organizations form the data controller unit in the data security landscape. They are the legal owners of the data and its security aspects.

Every organization must form different subunits as suggested by the below design and the following section reviews the subunits, their roles and the right strategies which can be followed within each unit to enable the organization with the best possible data security and compliance.

Data Security Officers

Data security officers are one of the key team(s) within an organization which should be formed as early as possible, possibly even before forming the data processing teams.

This team is headed by the company's chief information security officer (CISO) who is the leader and face of data security in an organization. The person in this role is responsible for creating the policies and strategies to secure data from threats and vulnerabilities, as well as devising the response plan if the worst happens.

In some organizations, they also employ a separate role called "data security analyst" to focus on the key responsibilities. This team is responsible for the activities below.

It is essential to educate employees about data security requirements within the organization. They perform security design reviews for new system integrations and publish findings with clear action items. The team publishes all their key findings that can potentially cause issues and tracks the action items to the closure.

They implement intrusion detection systems (IDS), and intrusion prevention systems (IPS) that can monitor the network and detect and block suspicious activities, reducing the risk of data breaches.

They also frequently test the system vulnerability using security exercises like penetration tests which help find and alert vulnerabilities.

They work in collaboration with vendor teams to get external third-party risk assessments (TPRM) done to identify any data security or compliance issues being brought in through the interactions.

Conduct or facilitate Privacy Impact Assessments (PIAs) which can help organizations to determine what privacy risks big data projects might pose. This helps find and put in place the right safety measures and limits.

While they make recommendations for enhancing data systems security, they also engage in reviewing attempted breaches of data security and rectifying security weaknesses. Performing security audits and risk assessments and analyses.

They maintain proactive monitoring of the data and integrations and alert any data security concerns and facilitate regular security audits to make sure that the organization is compliant.

IV. IT DIRECTOR

The primary responsibilities of IT director include enacting the strategies put forth by the CISO, monitoring all activity that occurs within the IT infrastructure, implementing and managing any security technology in use across the organization, maintaining regulatory and compliance posture and collaborating with the CISO on evaluating new security technology and defining components of the company's incident response plan.

Data Governance

Data governance team plays a key role in data safeguarding sensitive data. Their responsibilities include creating data policies, maintaining and enforcing them. They conduct training sessions for employees on data security best practices.

The team has a great understanding of data and the domains which helps them define data roles and their permissions. The team classifies the data based on the data sensitivity and assign ownership to the data sets. They help the organization set up the data access permissions for user roles within an organization. These permissions are regularly reviewed and audited by data governance.

They also have active participation in vendor and third-party management requiring contractual commitments to data security standards. As part of this, they work with the organizations' legal team and make sure that all the legal agreements are compliant.

This team defines data archival and retention policies so that data is retained only for as long as it is required. The data archival policies ensure that the data is archived or purged in a timely manner to minimize the impact of security incidents.

Data Platform Administrators

Data platform responsibilities encompass managing data access requests, ensuring compliance with security protocols, implementing robust security measures like encryption and monitoring, automating security tasks, managing the data lifecycle, implementing strong access controls, identifying and reporting security threats, enabling secure data sharing, protecting data integrity with backup and recovery, monitoring cloud activity for suspicious behavior, and analyzing network traffic for anomalies.

Data Processing/Analytics Teams

Data owners or analytics teams need to ensure data quality and integrity. They implement data validation and standardization techniques to prevent corruption, including data anonymization and encryption of sensitive information. They oversee maintaining data cleanliness, accuracy, consistency, and reliability. They protect sensitive data while preserving its utility for analysis, using techniques like labeling or tagging datasets. Additionally, continuous monitoring and peer review of implementation changes can be used to help identify and mitigate potential data vulnerabilities.

Best Practices

After making sure that the enterprise has set up the above team and defined the roles and responsibilities, teams can also set up some standards and best practices to help them achieve the data security goals seamlessly.

- Implement a Zero Trust Approach.
- Secure unstructured as well as semi structured data.
- Ensure that the data security and compliance features are enabled on the
- data processing layers.
- Secure data across all the different data stores.
- Enable security rules for endpoints.
- Always secure APIs using strong authentication and authorization features.
- Enable real-time compliance monitoring and alerting.
- Enable fine grained access control whenever possible.
- Always store the passwords in secured storages like key vaults
- Always enable security scanning templates in the GIT CICD pipelines.
- Use log analysis and continuous monitoring.

Cloud data storage is becoming increasingly popular, and here are some best practices for ensuring its security [7]

- Choose a reliable and reputable cloud provider.
- Understand the data security responsibilities owned by the enterprise vs the cloud service provider.
- Enable strong authentication and authorization mechanisms like Multi factor Authentication (MFA), OAuth etc. during cloud connectivity and access.
- Use customer managed keys in the cases where the data is highly classified and is required to be secured with the highest encryption.
- Enable encryption on object storage.
- Never open any public IPs for any services.

Future State

Automated scanning and identification of sensitive data using Artificial intelligence (AI) techniques can be a future state implementation that can be integrated. AI-driven adaptive access control systems dynamically adjust user access privileges based on various factors to protect sensitive data and resources from unauthorized access.

Monitoring user and entity behavior within a network to detect deviations from normal patterns using AI may indicate suspicious or malicious activity. AI tools can automate incident response processes by providing real-time alerts, prioritizing security incidents based on severity, and orchestrating response actions.

AI powered vulnerability management tools can automatically scan networks, systems, and applications to find potential security weaknesses and prioritize fixing the most critical issues first. This will thereby help with reducing data security threat incidents. [8]

V. CONCLUSION

In today's rapidly evolving and data-centric world, the importance of data security risk control cannot be overstated. It's challenging to predict future security threats, but businesses should stay informed about industry trends and try to anticipate potential risks. Gearing up towards the organization’s security and governance using all the above stated expertise and vigilance can help the organization go further miles with great success.

REFERENCES

1. PricewaterhouseCoopers, “Seven privacy megatrends: A roadmap to 2030,” PwC. https://www.pwc.com/us/en/services/consultin

g/cybersecurity-risk-regulatory/library/seven-privacy-megatrends.html

2. “Press releases | About Us | Equifax UK.” https://www.equifax.co.uk/about-equifax/press-releases/en_gb/-/blog/equifax-ltd-uk-update-regarding-the-ongoing-investigation-into-us-cyber-security-incident
3. “Largest data breach in history costs Yahoo another $85M | NAFCU,” NAFCU, Sep. 07, 2022. https://www.nafcu.org/newsroom/largest-data-breach-history-costs-yahoo-another-85m
4. “Big data security: advantages, challenges, and best practices.” https://www.turing.com/resources/big-data-security
5. WiKi Security.” https://rura.wikisecurity.net/blog/tips/summary-eu-gdpr-stakeholders
6. “Accelerating cloud shift means narrowing opportunity for providers,” Gartner, Feb. 09, 2022. https://www.gartner.com/en/newsroom/press-releases/2022-02-09-gartner-says-more-than-half-of-enterprise-it-spending
7. P. Bhatnagar and A. Kudrati, “11 best practices for securing data in cloud services,” Microsoft Security Blog, Sep. 12, 2024. https://www.microsoft.com/en-us/security/blog/2023/07/05/11-best-practices-for-securing-data-in-cloud-services/.
8. “The Future of Data Security: Predictions and new trends.” https://www.alumio.com/blog/the-future-of-data-security-predictions-and-new-trends
9. “Data Governance Overview | Institutional Research & Effectiveness | Wright State University.” https://www.wright.edu/financial-operations/institutional-research-and-effectiveness/data-governance/data-governance-overview