DATA GOVERNANCE AND SECURITY IN BIG DATA ENVIRONMENTS

Ravi Shankar Koppula
Satsyil Corp, Herndon, VA, USA
Ravikoppula100@gmail.com

Abstract

This paper explores the multifaceted challenges and emerging strategies in data governance and security within big data environments. With the exponential growth in data volume, velocity, and variety, traditional data governance frameworks are under strain, necessitating innovative approaches to data classification, metadata management, and access control. The paper emphasizes the crucial role of data governance in ensuring data quality, security, and regulatory compliance, particularly under the stringent demands of the General Data Protection Regulation (GDPR). It discusses the use of metadata for enforcing data privacy policies, alongside conventional monitoring and policy application methods. Furthermore, the paper delves into data quality management, highlighting the adverse impacts of poor data quality and presenting data cleansing techniques as a solution. It examines the efficacy of data encryption and dynamic masking in protecting sensitive information, with a focus on the importance of sophisticated key management. Additionally, the paper outlines best practices for data retention, archiving, and destruction to manage costs and comply with evolving regulations effectively. The need for flexible, metadata-driven retention policies and strategies for secure data destruction in distributed storage systems is articulated, underlining the critical need for new tools and methodologies in addressing the unique challenges of big data governance and security. Through a comprehensive analysis, the paper contributes valuable insights into creating adaptable, efficient, and secures data management frameworks in the era of big data.

Keywords–Data governance, Big Data, GDPR, Data privacy, Metadata management, Data quality, Data encryption, Data retention, Archiving, Data destruction.

I. INTRODUCTION

Data governance is essential in a big data environment to ensure data can be easily found, understood, and relied upon. Implementing data governance helps determine data quality, minimize risks, maximize data value, and ensure efficient data processing. It also aids in IT decision making and requires commitment from senior IT management. In a big data environment, traditional governance approaches may be impractical, making urgency and issue-driven stages more common. Data security is crucial in big data, as it deals with large volumes of diverse, rapidly moving, unstructured, uncertain data. Protecting data from harm and ensuring accurate conclusions are vital for maximizing the value of big data.

Overall, aircraft engine fan blades are sophisticated components that require careful design, maintenance, and inspection to ensure the safety and reliability of modern jet engines.

1.1. Definition of Data Governance

Data governance maximizes and ensures high-quality data throughout its life cycle. It provides decision rights to meet organization's needs in data quality, integration, compliance, and risk management. Data governance exercises decision-making and authority for data management, with stewardship being a key determinant of its formality. It encompasses people, processes, and technology for consistent data handling. Big data adds complexity with its volume and variety, requiring documented and authorized decision rights. Collaboration is crucial due to self-service and dispersed nature of big data. Proper governance avoids dehumanized decision-making and identifies and controls undesired consequences. Personal data used for analytics carries higher risk and must be used appropriately.

1.2. Importance of Data Governance in Big Data Environments

There has been an increasing interest in big data as an opportunity and an operational challenge for industry in the past year. The economic value of data is a frequent topic in the business press, but what does the term "big data" actually mean and why are its characteristics so important? This topic will be explored in Data Governance and Security in a Big Data Environment. In order to place the concepts of data governance and big data in context, it is necessary to begin with an understanding of the term "big data". In general, big data can be defined along the following dimensions: volume, velocity, and variety. That is, big data is data that exceeds the processing capacity of conventional database systems. The size of big data is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes. Data is also arriving quickly and unpredictably. This is the velocity dimension and often data is streamed in real or near real time. Finally, big data encompasses all data types; both structured and unstructured. Often, the unstructured data types have attributes that are revealed only when analyzed using statistical methods, machine learning, or graph algorithms. This heuristic rather than deterministic approach to discovering relationships makes data mining with unstructured data an iterative process and further extends the depth of analysis. Big data often spans tens to hundreds of terabytes, which again exceeds the capacities of conventional systems. Finally, the tools and methods used must scale to the increasing size and complexity of data to remain economically viable. A first impression might be that big data sounds a lot like the central issue in database management, and in fact, it shares many of the same problems. However, the necessity to leverage the potential of big data as a strategic asset has created a need for new tools and methods independent of those found in existing data analysis and management technologies. This point will be reinforced throughout the text as it will become evident that the risk of losing the potential value in big data due to non-conformance of legislation, mishandling, poor quality, and mismanagement of data has been a central driver for this work.[1]

1.3. Overview of Data Security in Big Data Environments

A major criticism of traditional data quality, data management, and data governance schemes is that they operate as a separate system or scheme to what they're trying to govern. For example, a data migration project will involve the assignment of additional resources to move data from system A to system B and then even more resources to document the content of the data. This metadata would prove beneficial to a data governance scheme, yet it is probable that this metadata will be lost or input into the wrong system. In modern complex data environments, the resource and the data are often one and the same, so any scheme on the resource must be tightly coupled with the management of the data. Failure to do this has led to governance initiatives appearing as overhead costs with little in the way of direct return on investment.

Big Data is a technology initiative that enables organizations to derive value from data. It has emerged as a major element of the modern IT ecosystem with the increased use of data-driven insights throughout the decision-making process. The term Big Data does not just refer to the volume of data being managed but also to the fact that it is coming from a wider variety of sources, including unstructured data. In the context of Data Governance, unstructured data is an important concept as a significant portion of business-critical information is held in this form. However, unstructured data is arguably the type of data over which organizations have the least visibility and control. The shift towards unstructured data has also led to a blurring of lines between data warehousing, content management, and data management, creating a complex environment in which there are many systems that manage and store data. This has big implications when we consider that the data-driven insight which Big Data delivers is often about the content that is buried in these systems.[2]

II. DATA CLASSIFICATION AND METADATA MANAGEMENT

An effective data classification scheme simplifies understanding of data location, including in big data environments. Data classification provides metadata for visibility into data storage, aiding policy enforcement and automation. This ensures security and supports governance.

Big data relies on automation and classification to process and analyze data. Data classification is crucial in big data environments to ensure efficient data ingestion and secure storage. Hadoop's file system, HDFS, allows for data classification to maintain consistent security and compliance policies. By classifying data upon ingestion, users can assign appropriate access controls and limit data access to authorized applications.

Data classification is the process of organizing data into categories for its most effective and efficient use. Data is classified in order to apply the right level of security to the data throughout its lifecycle. With effective data classification, data is understood, business rules are applied, and data is kept in the appropriate storage location.[16]

2.1. Importance of Data Classification in Big Data Environments

Proper data classification is crucial for effective data security. If data isn't classified correctly, ensuring its security becomes nearly impossible. Incorrectly classified data may be treated as if it doesn't need protection, posing risks of theft, destruction, or tampering. Data protection measures should align with data value and risk level. Tracking data location and movement is essential for accurate classification and has implications for activities like data retention and deletion.

Data classification is crucial in big data environments. It organizes data into categories, enhancing its usability and accessibility. The main aim is to make data easier to locate when needed. It also prioritizes sensitive or important data for analysts' use.

2.2. Strategies for Effective Data Classification

Data identification is the first step in classifying data. It can be done manually or with automated methods, which are more efficient and cost effective. Automated methods can be applied to a wide range of content repositories and can be easily scaled. After data identification, data inventory is necessary to document the data location. Again, automated methods are more efficient and cost effective, especially for large organizations with data stored in multiple locations. Manual methods are time consuming and quickly become outdated.

2.3. Role of Metadata in Data Governance and Security

Metadata encompasses data about the data. It includes information such as who, what, when, where, why, and how of the data. It is crucial for data governance and security. Metadata can be in various forms, such as business terms, policies, data quality, etc. Effective metadata management enhances data governance and reduces costs. In data security, metadata is used for monitoring data access and assessing security policies. It is also critical for data loss prevention systems in identifying sensitive data and monitoring its movement. Accurate metadata is essential for the success of these programs.[3]

III. DATA ACCESS CONTROL AND AUTHORIZATION

With traditional access control, switching between different policies requires system configuration changes. Sharing data securely between systems and controlling access and authorization can pose challenges. For instance, a big data system may need to push data to a third-party analytics system while governing what data is accessed and for how long. Integrity of the accessed data and auditable access activity are vital.

Access control and authorization for big data systems vary and can change over time. Flexible and easily configurable mechanisms are necessary. Sharing data between instances or systems requires fine-grained control.

Traditionally, database access control mechanisms cannot handle big data requirements due to limited expressiveness and scalability. Therefore, addressing access and authorization needs early in the design process is crucial to avoid security concerns.

Access control and authorization mechanisms for big data cover determining allowed activities for legitimate users or systems and granting necessary permissions for authenticated entities to perform operations on requested data.

3.1. Role-Based Access Control (RBAC)

RBAC is effective for access control, but has downsides. Single-role assignment may cause over-authorization. Broad roles may access unnecessary data. Additional roles may limit data access, but managing roles can be complex. Users may perform tasks outside their role and need role changes to access data. "Role explosion" increases administration costs. Despite downsides, RBAC will be used with other models in the future.

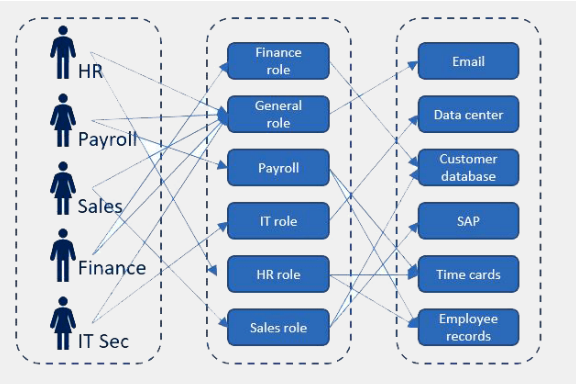

Fig.1 [9]

Role-based access control (RBAC) allows employees to access only the data that is necessary to perform their job duties. Each employee has a role in the organization. A role is defined as a set of responsibilities, and an employee is authorized to perform a certain set of tasks. RBAC restricts access to data by job function, which has many advantages for data security. The set-up is simple, as only the role of the employee and the data that is accessible to that role must be defined. Access to data can be easily controlled for each employee by assigning them to different roles, or changing the access rights of a given role. In comparison to access control lists, RBAC is much easier to administer, as rules can be defined to automate changes in role permissions. RBAC is widely used in the industry today and there are many commercial and open-source solutions available.

3.2. Attribute-Based Access Control (ABAC)

Also called the policy-based access control, an access control system in computing is a mechanism that is used to regulate which users are allowed to access what resources in the system, and what operations they can perform on those resources. Its main function is to authenticate users, restrict the access of certain users to some resources, and keep the access to resources within the usage constraints. Traditional access control systems use an access matrix, which specifies the rights for each user over each resource. However, the attribute and role-based access control systems use a more compact and easily manageable representation of user access rights.

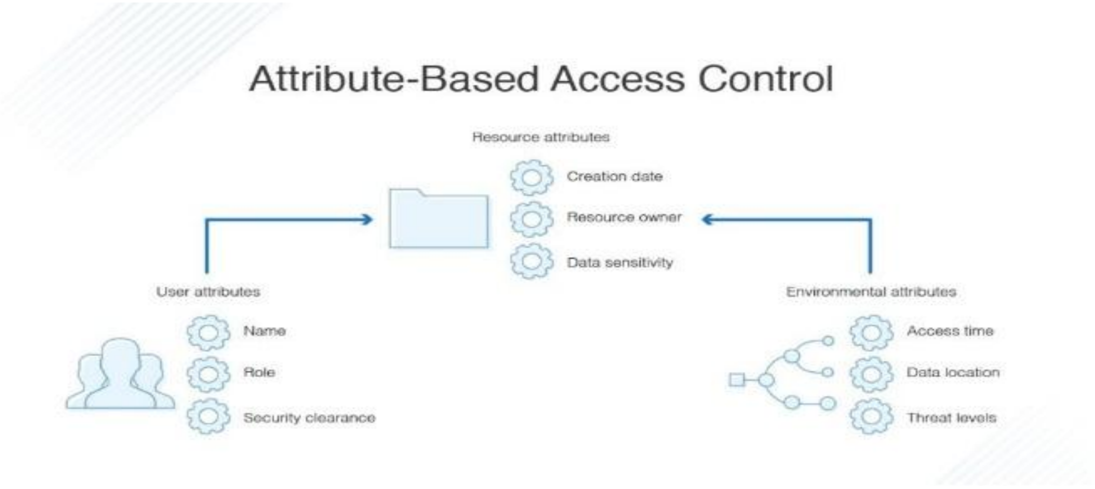

Fig. 2 [10]

Users access control in a system can be done with moral or immoral intentions. Many times, users may access resources out of curiosity or because they may have stumbled upon something that they would not normally access. However, users may also want to access sensitive company information with the intention of selling the information to other companies. ABAC is a heavy-duty access control mechanism that is designed to protect sensitive information resources from users trying to access them with evil intentions.

3.3. Access Control Policies and Enforcement

Users with admin privileges can query and generate reports in various ways. Access control policies can be applied to specific types of queries and data. For instance, only medical professionals should execute decision support queries on medical data. Current database systems cannot track user programs and queries or enforce access control policies automatically. This requires unavailable system features and ongoing research and development.

The DBMS ideally enforces the access control policy, but today's models only allow limited control. SQL commands like Grant and Revoke can enable or disable user access to specific data, but cannot express conditional access. For example, finance department employees should access salary records, but others should not. Current SQL access control cannot enforce this policy.

Enforcement of access control policies is anything but simple. In many cases, it is not clear at all how to enforce a certain access control policy using available database system facilities. Often, the best that can be done is to enforce the access control policy using ad hoc methods, making it the responsibility of the DBA and system managers to monitor system usage and configuration to guard against unauthorized data accesses.

Even though it is possible for an organization to define an access control model for their data and have a strong set of access control rules, it is not sufficient to just assume that the access control rules will be enforced. Hence, it is necessary for the database or data warehouse system to provide flexible means to express the access control policy and to support enforcement of the policy.[4]

IV. DATA PRIVACY AND COMPLIANCE

Complying with data privacy laws in big data is difficult due to the nature of technologies and analysis. Traditional structured data is easier to manage and secure compared to diverse data sources, like social media and IoT devices. The use of Hadoop and clustered environments further complicates understanding data storage, processing, and access. This gap in understanding raises concerns about protecting individuals' privacy, leading some to argue that big data and privacy are incompatible.

Data privacy laws protect personal information and allow individuals to monitor how it is used. These laws are complex and adapt to technological changes. They apply to citizens' data, regardless of where it is processed. In the EU, Directive 95/46/EC has been replaced by the GDPR. The GDPR harmonizes data privacy laws, empowers EU citizens, and changes how organizations handle data. It imposes higher penalties, encouraging companies to prioritize data governance and security.

Fig.3 [11]

4.1. Overview of Data Privacy Regulations

Enforcing data governance policies, addressing legal and regulatory guidelines, ensuring data protection, and building data privacy-preserving systems are the high-level steps to ensure privacy. Regulatory guidelines and countries' legal systems have many considerations for privacy such as sensitive data rules, data locality constraints, cross-border data flow rules, data retention policies, and laws relating to individual data subjects. These constraints are guidelines for data processing, and following them means a company is on the right path to data privacy.

Privacy is crucial in determining how personal information is used by third parties. It can be grouped into four categories: data subject, communication, information, and individual privacy. In today's information age, privacy is essential but threatened by technology. Protecting information is crucial, especially in the big data environment. An entity's success relies on customer trust, which can be eroded by uncertainty about privacy. [14]

4.2. Impact of GDPR on Big Data Environments

The European Union (EU) General Data Protection Regulation (GDPR) will replace the Data Protection Directive 95/46/EC in spring 2018. Organizations which keep data on EU citizens will need to overhaul their operations in order to comply or face hefty fines. GDPR is a regulation that requires businesses to protect the personal data and privacy of EU citizens for transactions that occur within EU member states. Compliance with this regulation has been mentioned as a major concern and secure line of business for all organizations. It will affect not only all organizations within the EU, but also organizations outside of the EU. This is mainly due to the regulation applying to any organization that processes and holds data of EU citizens, irrespective of the company's location. Any business found to be non-compliant will be faced with fines up to 4% of annual global revenue or 20 million Euros, depending on which is greater. This is a considerable increase in fines than the previous directive, which signifies how the EU are taking stronger action in ensuring that organizations are held accountable for neglecting data privacy and protection.

Big data technologies and processing of data has evolved greatly over the years since the previous Data Protection Directive was established. The volume of data held has increased considerably and data can be stored in various locations. Personal data also covers a broader scope of information and defines it as: "Any information related to a natural person or 'Data Subject', which can be used to directly or indirectly identify the person." This can be anything from a name, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer IP address. The changes in the definition of personal data and broad range of data it covers is a clear contrast to previously and has created uncertainty as to whether big data technologies are still in compliance with new regulation.

4.3. Strategies for Ensuring Data Privacy and Compliance

An innovative approach to enforcing data protection policies is to use metadata to embed protection state with the data. As data is accessed and processed, it can be automatically intercepted and have control checks and enforcement procedures applied to ensure that the intended data usage is compliant with the data protection policy. Measures can include blocking data access or usage, alerting an administrator, quarantining data for further assessment, or applying a security or privacy enhancing transformation. This approach to data protection control is proactive and can be highly effective in preventing policy violations.

Traditional data protection strategies have focused on building barriers to keep others out, but in big data environments, it is neither practical nor effective to lock down data. A much more effective strategy is to closely monitor data usage – knowing who is using the data, how it is being used, and to what effect. This can be achieved with policy-based approaches to monitoring data usage. As policies are executed, data about data usage can be collected and used to continually assess and improve the effectiveness of the data protection strategy. Monitoring should be pervasive, and incorporate real-time monitoring of data in motion and near real-time monitoring of important data processing operations. An effective strategy to monitor data usage is to make use of metadata to define where and how sensitive data is used, and to then deploy monitoring operations that are aligned with data processing routines.

Ensuring data privacy and compliance with data privacy regulations is often the most challenging of the data quality dimensions. The reason is that as data circulates and is transformed, it becomes increasingly difficult to control who has access to the data and how it is being used.[5]

V. DATA QUALITY MANAGEMENT

Data quality management is a hot topic because no program can succeed completely when working with bad or low-quality data. Data quality can be defined in numerous ways. Generally, data is considered high quality if it is "fit for its intended uses in operations, decision making, and planning". Primarily, data is of high quality if it is clean, correct, and up to date. In the context of big data, data might be of high veracity and variable in quality. Veracity (in the context of data) refers to the credibility of the data and the truth of its content. Usually, data quality problems are caused by human error in the data entry process and are detected after the processing of that data requested and is too late to fix. This can pose a problem with big data where the errors are extremely magnified.

One of the main reasons big data has issues with data quality is that it can sometimes completely overlook the data operations planning. This is a much-needed step to make sure the data project has a direction and achievable goals. Data operations plans mostly resolve data quality issues. The reason it's often overlooked is that programmers and developers consider it a lower priority task and start data processing too early. Data that has no plan and has yet to stake out a goal often becomes meaningless and is easily corrupt.[15]

5.1. Importance of Data Quality in Big Data Environments

The amendments made to the data could potentially cause errors elsewhere, or replace one piece of bad data with another of equal or greater badness. In the worst-case scenario, the data preparer will find that additional information is unavailable, leaving gaps which will also be filled on an ad-hoc basis. The preparer may even seek similar data from another source, effectively duplicating bad data. The result is wasted time, and an increase in the amount of bad data which will need to be rectified at a later date.

It is often claimed that business users spend 80% of their time preparing data, and only 20% actually performing analysis. While the 80-20 rule may not be accurate in all cases, there is usually a significant grain of truth in this statement. Often the data being prepared is in a poor state to begin with, but the preparer makes the best of what is available, making manual amendments where necessary, or seeking additional information to compensate for data that is missing. In some cases one may not even realize that the data being used is of poor quality. Any amendments that are made are likely to be based on business knowledge, rather than any knowledge of the true state of the data. This can be likened to the scenario in which a doctor prescribes medicine to alleviate the symptoms of an illness, without actually examining the patient in order to diagnose the illness itself. The doctor will appear to be doing something beneficial, but the patient may not actually get any better. Similarly, a business user's actions may appear to be effective, but the data may not be improving. In both cases, more damage may be done.

This is the crux of the matter: data that is of poor quality, or that is not trusted, cannot be used for decision making. However, in traditional data analysis, the actual impact of poor data quality can be hard to quantify, and it is common to find that data quality is only really addressed reactively, on a case by case basis. The argument can often be made that the cost of preventing poor data quality is more than the cost of dealing with the errors as and when they become apparent. But this argument is fundamentally flawed, based on an assumption that data quality can actually be restored. In many cases it cannot, and the cost of poor data quality is much more than many might think.

5.2. Data Cleansing Techniques

A common requirement is to enhance an incomplete set of data such as that obtained from a legacy system. In this case, there is more missing data than error and the aim is to infer the missing data from that which is present. This is quite a difficult operation and there are no guarantees of success. Finally, when the cleansing process is done, it is always good to measure the results of cleansing in order to give confidence to the data users.

This stage sometimes leads to the need to correct the data or even discard it and recapture it. Correcting the errors can be done a number of ways and in general the choice depends on the effort involved and the benefit expected. In any case, it is likely to be a complex operation. In desperate cases, one might need to recapture or repurchase the data.

Normally, the process of the data cleansing took a long time from the initial step of discovering the error until the last step of correcting the data. Duplicate records identification is crucial because to check each record for duplication against all other records could be impractical when there are millions of records. In many cases, the data is transformed as a part of its movement and loading. Unfortunately, the transformation often causes an error (sometimes, a complex one) which is detected later. In such cases, it is convenient to have direct facilities to locate the errors.

5.3. Data Quality Monitoring and Measurement

Monitoring and measuring data quality in big data environments involves tracking and assessing data to determine quality gaps and the impact of poor data quality. This informs data quality improvements and provides visibility on quality levels, trends, and variations. In big data, new solutions are emerging to address unique challenges in monitoring and measuring data quality. Traditional methods use quality dimensions and metrics, but these can be difficult in big data. New approaches include profiling data to automatically assess quality and compare it to benchmarks. Data quality visualization is a powerful tool for understanding the implications of data quality in big data.

In big data environments, data quality problems are compounded due to the high volume, velocity, and variety of information. It is difficult to control data quality at the point of entry and monitor it over time. Big data introduces new forms of data that are less understood and managed in terms of quality. The structure of the data may change in different analytical processes. Quality expectations may not always be clear or sacrificed for speed. Understanding what constitutes "good quality data" for big data is challenging. These factors make it hard to ensure and measure data quality, which can have catastrophic effects for an organization.[6]

VI. DATA ENCRYPTION AND MASKING

Encrypted data has no value for applications of big data. Trend analysis and statistical analysis can be performed on the data without decryption. On-demand decryption can protect stored data that is not analyzed for long periods. Map Reduce programming can facilitate computation on encrypted data. This enables easy development and maintenance of data processing tasks on both encrypted and standard data.

Data encryption has been of interest since its early implementation. Applying encryption to data in a Hadoop cluster poses challenges in at-rest, in-motion, and during processing. Traditional database systems have offered encryption solutions, but Hadoop's complex storage environment requires a different approach. Transparent encryption for Hadoop storage automatically encrypts data upon entry in the file system. Though still in its early stages, transparent encryption appears to be the more practical method, saving administrators time, cost, and headaches.

6.1. Encryption Techniques for Big Data

Encryption can be classified as either symmetric key encryption or asymmetric key encryption. In symmetric key encryption, the same key is used for both encryption and decryption of the data. The sender uses the key to encrypt the data, and the receiver uses the same key to decrypt the data. The key needs to be stored securely between the sender and the receiver. If an attacker gains access to the key, then the attacker can decrypt and make use of the data. In asymmetric key encryption, separate keys are used for encryption and decryption of the data. A public key is used to encrypt the data, and a private key is used to decrypt the data. The public key can be shared with anyone as it can only be used to encrypt the data. The private key is kept hidden and secure by the receiver. Asymmetric key encryption is more secure than symmetric key encryption due to the use of separate keys for encryption and decryption. At the present time, the most common encryption technique for Hadoop data uses symmetric key encryption with a single key.

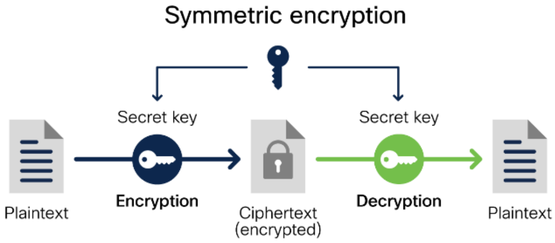

Fig.4 [12]

Data encryption is the process of changing the original data into transformed data (referred to as cipher text) such that only authorized entities can understand the transformed data. The encryption process uses an encryption algorithm and an encryption key. In traditional enterprise systems, encryption techniques were used to protect data both at rest and in motion. Encryption of data at rest involved databases, file systems, and backup tapes. On the other hand, encryption of data in motion involved the usage of SSL and TLS protocols for data transfer. Encryption of data in motion is quite similar to encryption of data in Hadoop.

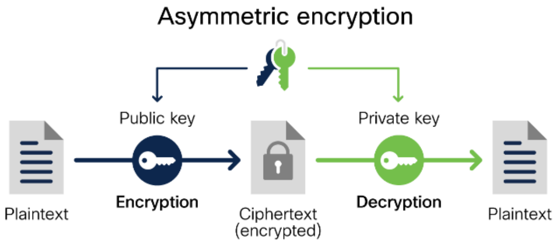

Fig.5 [12]

6.2. Masking Sensitive Data in Big Data Environments

There are various data masking techniques used by organizations. Static data masking involves cleaning and overwriting data with conceal values while maintaining the original characteristics. Dynamic data masking involves continuously securing original data during transit using technology services. Field-level data masking changes visible data at the database field level to unreadable text for unauthorized users. Format-preserving encryption changes values while preserving the data format. The choice of method depends on data type and user access. Reversible and irreversible encryption methods can also be used. Reversible masking assigns a unique value to the original and allows for undoing the process, while irreversible masking is deterministic and masks the same value consistently.

6.3. Key Management for Data Encryption

The most significant impact of encryption is decryption. During decryption, data is most vulnerable. Trust is required in the party performing decryption. Keys must be kept secure to ensure effective encryption. One method is keeping encrypted data and keys on the same system. However, this doesn't fulfill separation of duties requirements. More advanced systems involve a key server or HSM. Users access encrypted data on a database, while keys are accessed separately. This method ensures different access privileges and prevents key exposure. HSMs provide a secure location for key storage and cryptographic operations, eliminating unauthorized decryption opportunities.[7]

VII. DATA RETENTION AND ARCHIVING

In most traditional data management scenarios, when the useful life of data expires or it outlives its application, the data is deleted. However, in the world of big data, this is not always the case. When data is relatively cheap to store and the potential benefit of holding the data at a later stage is high, there is often a desire to keep all data indefinitely despite its perceived current value. A lack of an effective retention policy for stored data, however, can lead to so-called data hoarding and far outweigh the cost benefits of data storage. Data hoarding can lead to privacy concerns, as the stored data may breach privacy laws or it may be held longer than the actual purpose for which it was collected, as well as the increased cost of long-term storage, compliance, and e-discovery on a potentially vast and unmanaged dataset. Therefore, the implementation of a strategy to keep only the data necessary to meet regulations, litigation, and business requirements is crucial for organizations looking to effectively manage their data in a big data environment. This, however, can be problematic due to the continually changing nature of data and analytics, as what is deemed necessary to keep is constantly shifting. Therefore, a flexible, metadata-driven retention policy that includes all data whether it is in production or archived, that can adapt to the continuous changes in the environment, is most suitable for big data.

7.1. Retention Policies for Big Data

Organizations must define retention policies for big data to manage its diverse nature and volume. It's important to store data for a specified time and decide what to do with it afterward. Nonetheless, storing all data indefinitely is costly and risky, while deleting everything after a set number of days may not comply with regulations. The ideal approach is to apply retention policies based on content and context, although this is challenging with structured and unstructured data. New technological approaches can help understand and tag data, making its lifecycle and retention more manageable. Some policies may involve analyzing data and migrating it to another storage medium instead of immediate deletion.

7.2. Archiving Strategies for Long-Term Data Storage

The value of data can change over time, so archived data may need future access. Historical data is often used in big data analytics for predicting the future and analyzing specific subsets from a specific time. Data should not be hindered by slow access from offline storage like tape. Instead, consider moving data to a cheaper storage instance within the same cluster. Implementing a data archiving platform can manage data lifecycle and provide a single view for all data, whether in use or archived.

Key to successful archiving is the ability to maintain and easily access data throughout its entire lifecycle. Immutable data storage is essential to the preservation of data, ensuring that it cannot be altered or deleted for a specific retention period. WORM (write once, read many) storage is a common solution for this, with tape and optical disk being the most cost-effective mediums. However, cloud storage is becoming increasingly popular in the age of big data, and a viable alternative if the service provider can guarantee that WORM capabilities will be in place for the specified time.

Archives store data that is not needed for everyday business activities but may be required in the future. It is important to have an efficient archiving strategy that considers data management, quality, preservation, security, and cost control. While some archived data can be deleted, a significant portion of big data is subject to retention policies and requires long-term storage. The emergence of big data has complicated storage and access processes, making it more efficient to move data from high-cost to low-cost storage through "tiered storage".

7.3. Data Destruction and Disposal

Data destruction methods range from clearing and purging data for reuse to physically destroying the media, making it unrecoverable. The chosen method depends on costs and the value of the data. For instance, leaked data from the US Veterans Affairs Department in 2006 was found to be recoverable using standard techniques. This led to the use of costly cleansing methods like overwriting the data 7 times or physically destroying the drives.

Data that is no longer needed or is of no value should be destroyed. This is, in fact, a requirement of European Union data protection law (Directive 95/46/EC). The destruction of big data is a complex task that requires rigorous methods to ensure that the data is completely disposed of and is not recoverable. This is because a mix of big data sources is stored on different platforms and storage systems with interdependencies that can be hard to identify and trace. Moreover, the distributed nature of big data means that data may be replicated in many places; destroying the metadata of data may make it hard to find all copies.[8]

VIII. CONCLUSION

In conclusion, managing big data presents a complex web of challenges and opportunities, particularly in the realms of data governance, security, privacy, quality, and retention. Below points encapsulate the core conclusions from the research paper, highlighting the multifaceted approach required for effective data governance and security in big data environments.

Importance of Data Governance: Effective data governance is essential to ensure the accuracy, availability, and protection of data in big data environments. It helps organizations maintain data quality, comply with regulations, and build trust among stakeholders.

Security Challenges: Big data environments face unique security challenges due to their scale, variety, and the velocity of data. Ensuring data security requires robust encryption, access controls, and continuous monitoring to protect sensitive information.

Role of Policies and Standards: Establishing clear policies and standards is crucial for managing data governance and security. These policies should cover data classification, access management, and incident response to provide a framework for consistent data handling practices.

Technological Solutions: Leveraging advanced technologies such as AI and machine learning can enhance data governance and security. These technologies can automate data classification, detect anomalies, and respond to security threats in real time.

Collaboration and Training: Successful data governance and security require collaboration across different organizational units and ongoing training for employees. This ensures that everyone understands their roles and responsibilities in protecting data assets.

Regulatory Compliance: Adhering to regulatory requirements is a critical aspect of data governance. Organizations must stay updated with regulations like GDPR, CCPA, and others to avoid penalties and protect user privacy.

Continuous Improvement: Data governance and security are not one-time efforts but require continuous evaluation and improvement. Regular audits, updates to policies, and adoption of new technologies are necessary to address evolving threats and data management needs.

Strategic Importance: Integrating data governance and security into the organization's strategic objectives can provide a competitive advantage. It helps in building a robust data infrastructure that supports business goals and fosters innovation.

REFERENCES

[1] S. Leonelli, "Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science," Issue 1, Jun. 2019, doi: https://doi.org/10.1162/99608f92.17405bb6

[2]Nikolai Janoschek, "BI Survey," BI Survey, 2018. https://bi-survey.com/data-governance

[3]"Data Classification System - Definitions | University of Missouri System," www.umsystem.edu. https://www.umsystem.edu/ums/is/infosec/classification-definitions

[4]P. Colombo and E. Ferrari, "Access control technologies for Big Data management systems: literature review and future trends," Cybersecurity, vol. 2, no. 1, Jan. 2019, doi: https://doi.org/10.1186/s42400-018-0020-9.

[5]M. Mostert, A. L. Bredenoord, B. van der Slootb, and J. J. M. van Delden, "From Privacy to Data Protection in the eu: Implications for Big Data Health Research," European Journal of Health Law, vol. 25, no. 1, pp. 43–55, Dec. 2018, doi: https://doi.org/10.1163/15718093-12460346.

[6] I. Taleb, M. A. Serhani, and R. Dssouli, "Big Data Quality: A Survey," 2018 IEEE International Congress on Big Data (BigData Congress), Jul. 2018, doi: https://doi.org/10.1109/bigdatacongress.2018.00029.

[7]"SACA: A Study of Symmetric and Asymmetric Cryptographic Algorithms | Request PDF," ResearchGate. https://www.researchgate.net/publication/330555888_SACA_A_Study_of_Symmetric_and_Asymmetric_Cryptographic_Algorithms

[8]M. Grecco, "Disk vs Tape vs Cloud: What Archiving Strategy is Right for Your Business?," ProStorage, Feb. 20, 2018. https://getprostorage.com/blog/disk-vs-tape-vs-cloud

[9]M. Bozman, "Role Based Access Control (RBAC) | Explanation & Guide," BetterCloud Monitor, Jan. 08, 2018. https://www.bettercloud.com/monitor/the-fundamentals-of-role-based-access-control/

[10]"RBAC vs. ABAC Access Control: What's the Difference?," DNSstuff, Oct. 31, 2018. https://www.dnsstuff.com/rbac-vs-abac-access-control

[11] "The GDPR and Privacy: What Security Leaders Need to Know | 2018-09-24 | Security Magazine," www.securitymagazine.com. https://www.securitymagazine.com/articles/89443-the-gdpr-and-privacy-what-security-leaders-need-to-know

[12] “How to Enable Secure Authentication in Mobile Apps,” Infopulse, Mar. 12, 2018. https://www.infopulse.com/blog/how-to-enable-secure-authentication-in-mobile-applications

[13] “User Data Privacy,” AMB Law. https://amblaw.com/user-data-privacy/

[14]M. Nadeau, “General Data Protection Regulation (GDPR): What you need to know to stay compliant,” CSO Online, Jun. 12, 2018. https://www.csoonline.com/article/562107/general-data-protection-regulation-gdpr-requirements-deadlines-and-facts.html

[15] U. Nayak and U. H. Rao, The InfoSec Handbook: An Introduction to Information Security. Apress, 2014. Accessed: Jun. 11, 2024. [Online]. Available: https://books.google.com/books?id=Qe9lBAAAQBAJ

[16]“What is Data Classification: Types, Applications, and Best Practices,” levity.ai. https://levity.ai/blog/data-classification-types-applications