Survey on Security and Privacy Issues in Cloud-based Big Data Applications

Vedant Kharche
vedant.kharche@und.edu

Jun Liu
jun.liu@und.edu

School of Electrical Engineering and Computer Science
College of Engineering and Mines
University of North Dakota
Grand Forks, 58202

Abstract

Big data has emerged as a new paradigm for data applications and has attracted interest in many industry fields. The widespread usage of big data relies not only on the promising solutions and mechanisms of data analytics but also on security protection and privacy preservation. Big data are widely adopted in many security-sensitive sectors to support decision-making, which includes finance, marketing, politics, and healthcare systems. The five basic characteristics of Big Data (Volume, Variety, Velocity, Value, and Veracity) have been targeted raised security and privacy issues. It is critical to identify and analyze the various security and privacy issues targeting each of the five basic characteristics. Security protection for big data is even more important as distributed frameworks and cloud-based storage solutions become increasingly incorporated in big data applications. The integration of big data and cloud-based systems with smartphones has resulted in the development of numerous applications that generate real-time and sensitive data, raising concerns about data privacy. New security and privacy concerns arise with the adoption of Hadoop and MapReduce in storing and processing huge volumes of data. It is necessary to identify security and privacy issues associated with the essential components of the infrastructure used for hosting cloud-based Big Data applications, ranging from devices used for supporting data flows to access control policies. This paper aims to provide a survey on security and privacy issues associated with the five basic characteristics of Big Data applications, together with a detailed review of the new security and privacy changes for cloud-based Big Data applications.

1. Introduction

“Big data refers to large and complex datasets that typical software is inadequate for managing [1]. Big Data can be described based on five basic characteristics: Volume, Variety, Velocity, Value, and Veracity, which are commonly called 5V characteristics. 5Vs are typically used to characterize of Big Data as volume, velocity, variety, veracity, and value [2,3]. Volume is the size of data; velocity is the high speed of data; variety indicates heterogeneous data types and sources; veracity describes consistency and trustworthiness of data; and value provides outputs for gains from large data sets.” Big Data can be generally described by 5 characteristics of volume, veracity, velocity, variety, and value. These characteristics significantly impact big data security. Before establishing security solutions, we must understand how these characteristics impact big data security by analyzing the challenges it poses:

1) Volume: The sheer volume of data in big data contexts makes implementing effective security controls difficult. Scalable security procedures that can be deployed consistently across the entire system are required to manage the security of such a massive amount of data.

2) Velocity: Because big data processing is sometimes required in real-time, it can be difficult to impose security measures without compromising performance. To keep up with the increasing data processing speeds, security measures must be quick and efficient.

3) Variety: Big data is derived from a wide range of sources, including organized and unstructured data. Protecting this various data involves different security methods based on the data type, which makes implementing a consistent security approach difficult.

4) Veracity: Big data is frequently plagued by data quality challenges such as mistakes, inconsistencies, and inaccuracies. This can make implementing security measures that rely on correct data, such as access control, difficult.

5) Value: The importance of big data makes it an appealing target for hackers. Protecting important data necessitates strong security procedures that guard against a wide range of security risks, including as cyberattacks, insider threats, and data breaches.

Due to the challenges posed by the characteristics of big data organizations need specific tools and technology such as Hadoop, Spark, NoSQL databases, and data warehouses that can efficiently store and process huge volumes of data to handle big data. These technologies offer distributed processing and storage capabilities that let businesses scale their data processing capacity as necessary. Big data is playing a vital role in business, healthcare, finance, and government, use big data in diverse ways. Big data is also driving innovation in industries, including artificial intelligence, machine learning, and data analytics. This innovation enables businesses to process, analyze, and process huge volumes of data in real time, allowing them to make data-driven decisions and helping businesses to better understand client behavior, run more efficiently, develop fresh goods and services, and make informed decisions. Ensuring appropriate security measures is crucial with the widespread use of big data technologies across various industries. These technologies allow organizations to collect, analyze and store vast amounts of data from sources such as mobile devices, social media apps, and IoT devices, including sensitive information like financial, health, personal identity, and company secrets. Breaches at different data lifecycle stages can lead to fraud, identity theft, and cyberattacks. Big data environments are complex and distributed, making data protection challenging. Enterprises must implement strong security measures, such as access controls, encryption, monitoring, and threat detection, to secure data confidentiality, integrity, and availability.

2. Security Challenges for Big data

Data, devices, networks, and cloud infrastructure all play a critical role in big data management and analysis. Data is at the heart of big data as insights are gained through collecting and analyzing large and complex datasets. This data is gathered in the form of structured or unstructured data from large number of devices such as mobile phones, computers, servers. The data collected from devices is transferred across networks, allowing organizations to share data between their teams. Big data technologies have also integrated cloud infrastructure as it provides scalable and flexible resources for storage, processing, and analysis, allowing organization to handle large volumes of data. Data, devices, networks, and cloud technologies all work in tandem to collect, store, process and analyze volumes of data, ultimately providing insights that can help organizations make informed decisions. It is, therefore, necessary to understand the various security challenges concerning these components.

2.1 Data Related Security Challenges

The primary goal of big data is to extracts insights from massive volumes of data. This leading to big data platforms holding vast volumes of sensitive data, which are appealing targets for hackers. Insiders with privileged access can either purposefully or inadvertently misuse or steal data. Big data systems must adhere to a variety of standards, including those governing data protection, privacy, and security. Many big data systems are cloud-based, which brings new security vulnerabilities such as data breaches, account hijacking, and denial of service assaults. Organizations must establish a complete security strategy that includes robust authentication and access control, encryption, monitoring and auditing, disaster recovery, and business continuity planning to handle these security problems.

2.2 Device Related Security Challenges

Massive volumes of data gathered from many sources, including sensors, IoT devices, and mobile devices, are processed and analyzed in big data settings. Device vulnerabilities, data privacy, authentication, access control, monitoring and auditing, and physical security are all security considerations in big data contexts. Cyberattacks, viruses, and other security dangers can befall devices utilized in big data contexts. It is vital to ensure the privacy and security of the data acquired by these devices to preserve user confidence and comply with data protection requirements. Unsecured devices can be used to gain unwanted access to the system, potentially leading to data breaches and other security concerns. It is critical to monitor and audit devices used in big data settings to detect and prevent security problems. Physical security of equipment used in big data contexts is crucial for preventing data theft, manipulation, or destruction. By addressing device security issues in big data settings, businesses may reduce the risk of security events while also protecting the confidentiality, integrity, and availability of their data.

2.3 Network Related Security Challenges

To move, store, and analyze the enormous volumes of data required for big data activities, a dependable and fast network is crucial. Finding insights and value in the data is impossible without a reliable and scalable network. As a result, network security considerations are important in large data contexts. Some network security considerations in big data contexts include network vulnerabilities, data privacy, authentication, and access control, monitoring and auditing, and cloud security. Because networks can be exposed to assaults, viruses, and other security concerns, network vulnerabilities are a major problem in big data contexts. In large data contexts, data transported over networks can be intercepted, hacked, or modified, making data privacy a key problem. Unsecured networks can be used to gain unwanted access to the system, potentially resulting in data breaches and other security concerns. In large data settings, monitoring and auditing network activity is critical for detecting and preventing security problems.

2.4 Cloud Integration Related Security Challenges

Cloud computing is becoming more popular for storing, processing, and analyzing large amounts of data. Its design, however, has specific security concerns that must be addressed to assure data privacy, confidentiality, and availability. One of the primary security issues in cloud computing is the possibility of data breaches. Because data is hosted on remote servers held by third-party cloud providers, attackers may be able to obtain unauthorized access to sensitive data via weaknesses in the cloud architecture or apps. Cloud companies also have access to consumer data, prompting issues about data ownership and management. Another major security risk is data privacy. Data sent to the cloud can be intercepted, hacked, or modified, possibly revealing critical information. Cloud computing also raises questions regarding access control. It is critical to ensure that access to data and cloud resources is correctly authorized and verified, with appropriate rights provided to individuals based on their roles and responsibilities. Maintaining the security of cloud infrastructure and services is crucial to ensuring data confidentiality, integrity, and availability in big data settings.

2.5 User Related Security Challenges

Big data security presents considerable difficulties for enterprises, especially when it comes to security threats associated with users. Insider attacks, which can happen when employees or contractors with access to sensitive data inadvertently or intentionally compromise security, are one of the most important user-related security challenges. This may occur because of data theft or the unintentional release of private information. Weak passwords pose a serious concern as well because they can be quickly cracked or deduced, providing illegal access to massive data platforms. Another popular strategy used by hackers to access large data systems is phishing assaults, which frequently fool users into disclosing important information or downloading malware. Those that purposefully or inadvertently divulge private information outside of the company may be causing data leakage.

3. Data Related Security Solutions

From a data perspective, securing big data entails safeguarding the data as well as the systems and procedures that handle and process the data. With the increasing amount of data being generated and processed, it is critical to protect sensitive data from unauthorized access or exposure. Organizations need to ensure that data is protected at all stages of the data lifecycle, from storage and transmission to processing and analysis. By using techniques such as data masking, data anonymization, data encryption, and access controls, organizations can reduce the risk of data breaches, protect sensitive information from cyber threats, and comply with regulatory requirements. Therefore, it is important to understand these techniques and how they can be implemented to ensure effective data-related security.

3.1 Access Control

Access control security solutions in big data environments are focused on ensuring that only authorized users have access to sensitive data. There are several mechanisms organizations can employ to ensure security. Big data systems must have reliable access control measures to guarantee their confidentiality, integrity, and availability. To make sure that only authorized individuals may access sensitive data and that unauthorized access attempts are immediately discovered and dealt with, access restrictions should be routinely evaluated and updated.

- **Role-Based Access Control (RBAC)**
  RBAC is a popular access control technique in which individuals are assigned roles based on their job duties and responsibilities. Each role relates to a set of permissions that govern the actions that a user may do within the system. RBAC is a versatile mechanism that may be controlled and scaled as the company expands.

- **Attribute-Based Access Control (ABAC)**
  ABAC is an access control method that leverages characteristics such as job title, department, or location to determine access capabilities. This technique enables businesses to set more detailed access controls based on specific qualities rather than just roles.

- **Discretionary Access Control (DAC)**
  A sort of access control method in which data owners or administrators have the power to give or prohibit access to their data. DAC is often utilized in smaller situations where data owners can successfully regulate access to their data.

- **Mandatory Access Control (MAC)**
  MAC is a type of access control technique in which access to data is regulated by security labels applied to the data and users. MAC is often utilized in government and military situations where tight security requirements are required.

In large data settings, a variety of access control measures should be utilized to guarantee that only authorized individuals have access to sensitive data. The technique utilized will be determined by the organization's security needs, the scale of the environment, and the complexity of the access control regulations.

3.2 Data Masking

Data masking and anonymization are two related approaches used in big data security to safeguard sensitive information while keeping its value for legitimate reasons. Both strategies are intended to safeguard the confidentiality and privacy of sensitive data by blurring or masking the actual data. Data masking is a technique for disguising or hiding sensitive data in a dataset. The purpose of data masking is to restrict unwanted access to sensitive information while retaining the data's overall usefulness. Data masking techniques include:

- **Substitution**
  With this strategy, sensitive data is replaced with non-sensitive data. For instance, substituting a social security number with a number produced at random. When the data is not required for analysis or processing, this strategy is beneficial.

- **Redaction**
  This approach includes limiting sensitive data from a dataset. For instance, eliminating all personally identifying information from an email message. When the data is not required for analysis or processing, this strategy is beneficial.

- **Format-Preserving Encryption**
  This technology encrypts sensitive data while retaining its original format. Encrypting a social security number, for example, while keeping its length and structure. This approach is useful when data must be handled or evaluated while remaining secure.

3.3 Data Anonymization

Anonymization is a process that includes eliminating or changing identifying information from a dataset. The purpose of anonymization is to make it difficult or impossible to identify individual users or sensitive information. Anonymization techniques include:

- **Generalization and Suppression**
  This mechanism reduces the amount of information that is shared or released and thus reduces the risk of breach of privacy. Generalization makes changes in the sensitive data by replacing it with less precise data while still maintaining the original meaning of the dataset. This reduces the information that is shared while still being useful to the application. In case of suppression the data is removed entirely when the application does not require extremely sensitive data for its analysis. Both generalization and suppression are used in various privacy-preserving techniques such as anonymization, differential privacy, and k-anonymity.

- **Perturbation**
  Perturbation refers to adding of noise or random data in a dataset. Perturbation allows the privacy to be preserved by adding noise or distortions while still being useful for data analysis. Some common methods of perturbation include rounding values, adding noise to data or aggregating data to higher level of abstraction.

- **Differential privacy**
  In differential privacy mechanisms Individual privacy is protected by adding random noise to the data in a controlled manner while maintaining the probability distribution of the data. This random noise makes determining the values of individual data items difficult for an attacker while still allowing meaningful analysis to be performed on the data.

3.4 Data Encryption

Data encryption is critical in big data security because it protects sensitive data from illegal access, theft, or change. Given the heightened danger of data breaches and cyber-attacks in today's data-driven society, encryption is critical for protecting sensitive data. Encryption adds an extra layer of protection to help prevent data breaches and protect organizations from reputational and financial harm.

- **Transparent Data Encryption**
  Some relational database management systems (RDBMS) employ transparent data encryption to encrypt data at rest. TDE encrypts data in database data files and safeguards it when the files are backed up. The database engine handles the encryption and decryption processes, so no modifications to the application or user interface are required. This implies that once TDE is configured, the application and user interface do not require any changes.

- **File-Level Encryption**
  Encrypting individual files or directories on a file system is referred to as file-level encryption. This can be accomplished using encryption software or tools, such as PGP (Pretty Good Privacy), which can encrypt and decode data using a public-private key pair. File-level encryption is an excellent choice for large data situations where data is stored in files, such as log files or CSV files. It allows you

to encrypt data at rest and secure specific files or directories with various encryption keys.

- **Application-level Encryption**
  Encrypting data within an application is what application-level encryption entails. This can be accomplished using encryption libraries or APIs (Application Programming Interfaces) like OpenSSL or Bouncy Castle, which give developers with encryption and decryption methods that can be implemented into their programs. Application-level encryption is a viable choice for large data settings that use bespoke programs to handle data. It enables developers to include encryption and decryption functions directly into their apps, adding another degree of protection.

- **Database-level Encryption**
  Encrypting certain columns or tables within a database is known as database-level encryption. This can be accomplished using database-specific encryption capabilities or third-party encryption solutions that connect with the database. Database-level encryption is an excellent choice for large data settings that contain sensitive data in databases. It enables the encryption of columns or tables, preserving important data even if the database is compromised.

4. Network Security Solutions

Network security solutions are critical in protecting big data systems. These solutions are intended to safeguard the network infrastructure, which consists of the hardware, software, and protocols that allow data to be delivered and received across the network. With the increasing reliance on computer networks and the internet for daily operations, the risk of cyber-attacks and data breaches has increased significantly. Organizations need to protect their networks from external and internal threats to ensure the confidentiality, integrity, and availability of their data. By implementing network security solutions such as firewall, network segmentation, IDS/IPS, and network access control, organizations can reduce the risk of cyber threats, prevent unauthorized access to their networks. To ensure successful network security solutions, it is important to understand these strategies and how to use them.

4.1 Firewall

A firewall is a network security device that is used to block unwanted network access. It serves as a firewall between the trusted internal network and untrustworthy external networks such as the Internet. Firewalls can be hardware- or software-based, and they operate by evaluating incoming and outgoing network traffic in accordance with pre-defined security standards. If the traffic does not comply with the security policy, it is denied or stopped. Several levels of security can be provided by firewalls. Simple firewalls merely allow or deny traffic based on IP addresses, but more sophisticated firewalls may scan packet content and do deep packet inspection. In addition to intrusion detection and prevention, content filtering, and antivirus protection, firewalls can provide other security functions.

4.2 Intrusion Detection and Prevention Systems (IDS/IPS)

IDS/IPS systems are intended to detect and block illegal network access. IDS systems monitor and analyze network traffic for signals of malicious activity, such as network scans, denial-of-service attacks, or efforts to exploit vulnerabilities. When an intrusion detection system detects an attack, it sends an alert to the network administrator. IPS systems go a step further by automatically identifying and blocking harmful traffic. IPS systems are more proactive than IDS systems because they may act automatically to prevent an attack from succeeding.

4.3 Network Segmentation

In big data network security, network segmentation is a crucial technique that involves dividing a network into smaller sub-networks or segments, each with its own set of security measures. By doing so, organizations can limit the attack surface and the impact of a security breach. One of the security technologies that can be used for network segmentation is a Virtual Private Network (VPN), which establishes a secure and encrypted tunnel between two endpoints, allowing two networks to communicate securely across the internet. VPNs can also be used to connect distant users to a business network or to link two geographically isolated networks. VPNs provide security by encrypting all network traffic, making it impossible for attackers to intercept and read the data. Other technologies such as VLANs, firewalls, and routers can also be used for network segmentation, ensuring that each network segment is separated from the others, and access between segments is governed by strict security standards.

4.4 Network Access Control

Network-related access control is an important aspect of big data security solutions, as it allows organizations to control access to their network and prevent unauthorized access, modification, or theft of data. Network access control (NAC) is a security tool which restricts access to a network depending on the user's identification, the device they are using, and whether they are following security guidelines. By limiting access to just conforming users and devices, NAC systems can be used to prevent illegal access to a network. NAC operates by defining rules that control network access. In accordance with these regulations, network permissions, device security settings, and user authentication requirements may all be necessary. The NAC solution will check a user's identification and device security status when they try to connect to the network to make sure they comply with the set policies. NAC solution can be implemented in different ways.

- **Endpoint NAC**
  Endpoint NAC solutions check the security status of devices attempting to connect to the network. These solutions can ensure that devices meet specific security requirements, such as having updated antivirus software, firewalls, or the latest security patches.

- **Network NAC**
  Network NAC solutions check the identity and security status of users attempting to connect to the network. These solutions can ensure that users have appropriate permissions to access specific network resources, and that they meet specific security requirements.

- **Hybrid NAC**
  Hybrid NAC solutions combine endpoint and network NAC capabilities to provide a comprehensive security solution. These solutions can ensure that both users and devices meet specific security requirements before being allowed to access the network.

5. Device Security Solutions

Solutions for device security are crucial for safeguarding the availability, confidentiality, and integrity of data held on devices. Implementing efficient device security solutions is essential given the rise in the number of devices used to store and process sensitive data. Device security can be ensured via a variety of methods, such as the usage of trust certificates, device management, and corruption-free software. Only trusted software is loaded on devices thanks to the usage of trust certificates, which are used to confirm the reliability and authenticity of software. Device management entails keeping track of devices to make sure they are current, correctly set up, and in line with security regulations. In order to guarantee that the software operating on devices is devoid of bugs, malware, or other vulnerabilities that could be exploited by attackers, it's crucial to use corruption-free software solutions. To effectively defend against cyber-attacks, firms must understand these tactics and how to put them into practice.

5.1 Trust Certificates

In a big data system consisting of many connected devices, it's crucial to make sure that each one is permitted and authenticated to access the data on the network. The identification of every device on the network can be confirmed using trust certificates. By authenticating and confirming the identification of devices on a network, these certificates guard against data breaches and unwanted access. When a device attempts to access data on the network, the device's trust certificate, which is specific to that device and provides information about the device's identification, is verified. Trust certificates can also be used to monitor network activities, regulate access to data on the network, and encrypt data sent between devices. When it comes to protecting devices in large data systems, trust certificates are a potent option. Trust certificates can aid in preventing unauthorized access, safeguarding against data breaches, and ensuring the integrity of the data on the network by offering authentication, encryption, access control, and monitoring capabilities.

5.2 Utilization of Corruption-Free Software

Maintaining security in a big data system depends on the devices' software not being corrupted. System outages, security breaches, and other security risks can be caused by corrupted software. Use software solutions that are created to be secure and free from corruption to prevent these issues. There are numerous ways to accomplish this. First off, selecting reputable software vendors can assist guarantee that the offered software is dependable and safe. Second, keeping software up to date can shield it against flaws and guarantee that the most recent security fixes are used. Thirdly, routine software testing can spot potential flaws or places where corruption might happen. Fourthly, putting security measures in place helps stop illegal access and safeguard sensitive data. These protocols include access control, encryption, and multi-factor authentication. By taking these measures, organizations can ensure the integrity of their big data system, prevent security breaches, and protect sensitive data.

5.3 Device Management Policies

In big data systems, device management policies ensure devices are managed, updated, and secured properly through centralized management, access control, monitoring, patch management, and encryption policies. These policies prevent security breaches, protect sensitive data, and ensure system integrity by controlling devices, tracking device behavior, and constantly updating security patches and software. Organizations can enforce these policies to stop unauthorized users from accessing critical data and prevent potential security holes and data loss. By doing so, they can guarantee uniform application of security requirements and comply with legal requirements.

6. Distributed Integration Oriented Security Solutions

In order to guarantee the security and protection of data that is shared across numerous systems and networks, distributed integration-oriented security solutions are crucial. The demand for efficient security solutions has grown more critical than ever with the rise of distributed systems and cloud computing. Data governance regulations, monitoring, and logging are a few of the strategies that can be used to guarantee the security of distributed systems. Organizations may manage their data efficiently and identify and address possible security issues in real-time by putting these approaches into practice. By implementing these techniques, organizations can establish clear policies for managing data, monitor and analyze data activity to identify potential security risks, and respond promptly to any security incidents.

6.1 Data Governance Policies

Effective management of data assets in a distributed environment requires a strong data governance policy. This policy should cover the management of data assets, including policies and processes, to ensure responsible data maintenance and access. To maintain data accuracy, completeness, and consistency, compliance rules must be created and adhered to in accordance with relevant laws, regulations, and industry standards. To achieve this, collaboration across multiple teams and departments and the use of technology tools to monitor and enforce policies is essential. A comprehensive data governance strategy can help ensure the successful use of data assets in a distributed environment and mitigate the risks of data breaches.

6.2 Monitoring and Logging

Monitoring and auditing systems are crucial for maintaining the security of a distributed environment. These systems help to detect and respond to security breaches, track network activity, and ensure compliance with regulatory requirements. There are several components to a robust monitoring and auditing system in a distributed environment:

- **Security Information and Event Management (SIEM) Systems**
  SIEM systems gather, analyze, and report on security data from diverse sources in real time. These sources might include network devices, servers, apps, and security devices. SIEM systems correlate data from these sources to identify security risks and provide alerts to warn security professionals of any questionable behavior. SIEM systems analyze data using a variety of methodologies such as signature-based detection, behavioral analysis, and machine learning. Signature-based detection includes comparing data to established patterns of malicious activity, whereas behavioral analysis searches for abnormalities in user behavior. Machine learning algorithms may be used to detect patterns of behavior that signal a security danger, even if the threat has not previously been discovered.

- **Data Loss Prevention (DLP) Systems**
  DLP systems are used to monitor data while it is accessed and to prevent unwanted access or data exfiltration. DLP systems may monitor data while it is at rest, in use, or in transited systems monitor data using a variety of methodologies, including content-based inspection, context-based inspection, and behavior-based inspection. The process of evaluating the content of data to identify sensitive information, such as credit card numbers or social security numbers, is known as content-based inspection. Context-based inspection is examining the environment in which data is utilized, such as the location or device used to access the data. The process of studying user behavior to identify security concerns is known as behavior-based inspection.

7. User Related Security Solutions

User-related security solutions are critical to mitigating user-related security risks in big data. User Behavior Analytics (UBA), authentication, employee training, and audit trails can be implemented. These solutions are designed to ensure the privacy and security of users' personal information, as well as to prevent unauthorized access to their accounts. By utilizing UBA, organizations can monitor user behavior and detect potential threats before they cause damage. Strong authentication mechanisms are essential for preventing unauthorized access to user accounts. Providing regular employee training can help ensure that employees understand how to identify and prevent security threats. Audit trails are important for tracking user activity and identifying potential security vulnerabilities. By implementing above solutions organizations can protect user data and prevent security breaches.

7.1 Authentication

Authentication is a critical aspect of security in big data environments. Authentication procedures in big data security relate to the processes, protocols, and technologies used to validate the identity of persons or systems accessing the data. These procedures are critical for maintaining data confidentiality, integrity, and availability in large data situations. Some of the authentication mechanisms are described as follows.

- **Multifactor authentication (MFA)**
  MFA is an authentication technique that requires users to give multiple authentication factors to access the system. These elements can include a password, a security token, or biometric data such as a fingerprint or face recognition. This approach provides an extra layer of protection to the authentication process, making it more difficult for unauthorized users to obtain access to the system.

- **Single Sign On (SSO)**
  SSO is an authentication technique that allows users to utilize a single set of credentials to access different systems. By minimizing the number of passwords that users must remember, this approach streamlines the authentication process and enhances security. It also minimizes the chance of password reuse and password-related security problems.

- **Public key infrastructure (PKI)**
  PKI is a system that employs public key cryptography to deliver authentication and encryption services. It employs digital certificates to authenticate users' identities and protect communications between them. PKI can be used in a big data environment to authenticate users accessing Hadoop clusters or to secure communication between nodes in a Hadoop cluster.

7.2 User Behavior Analytics (UBA)

UBA technologies monitor user activity to discover abnormalities and possible security issues, such as illegal access attempts or odd data access patterns. UBA tools may be used to monitor many sorts of data, such as network traffic, system logs, and application logs. UBA tools evaluate user behavior using a variety of methodologies such as machine learning, statistical analysis, and rule-based analysis. Machine learning algorithms may be used to find patterns of activity that suggest a security problem, whilst statistical analysis can be used to identify abnormalities in user behavior. Rule-based analysis entails defining rules or thresholds that signal a security issue. UBA is a valuable tool for organizations looking to improve their big data security. By analyzing user behavior in real-time, UBA can provide insights that help organizations detect and respond to potential threats before they become major security incidents.

7.3 Employee Training

A key security measure for large data protection is employee training. Data breaches and cyberattacks are on the rise in today's increasingly digital world, and enterprises must take preventative measures to safeguard their data. Organizations may lower the risk of security incidents and safeguard sensitive data from unauthorized access by training employees on security best practices and how to handle data securely. A wide range of subjects can be covered in employee training, including as password policies, phishing awareness, data handling protocols, compliance rules, and incident response. There are numerous ways to give this training, including in-person classes, online learning modules, and role-playing security situations. Maintaining employee knowledge of the most recent security dangers and best practices requires ongoing training. By investing in employee training, organizations can improve their overall security posture and create a culture of security awareness and accountability among employees.

7.4 Audit Trails

Audit trails are crucial in big data for monitoring user activity and detecting potential security incidents. They record system events and activities, including data access and user behavior, allowing organizations to quickly identify anomalies and investigate potential security problems. Audit trails also help organizations comply with legal requirements and prove they are taking the necessary precautions to secure sensitive data. To develop an effective audit trail, organizations should ensure proper logging methods and retention periods, establish policies and procedures for reviewing audit logs, and secure audit trails against illegal access or manipulation.

8. Conclusion

The emergence of big data has brought about unprecedented benefits to organizations, but it has also given rise to new security challenges. This paper has explored various security challenges related to data, devices, networks, cloud integration, and user-related issues. Moreover, we have reviewed several solutions for these challenges that are aimed at protecting the confidentiality, integrity, and availability of big data. It is important for organizations to take a holistic approach to address big data security challenges by implementing appropriate technical and organizational solutions, and to ensure that security measures are regularly reviewed and updated to mitigate emerging threats. Ultimately, the successful implementation of security solutions will enable organizations to leverage the full potential of big data while maintaining the confidentiality, integrity, and availability of sensitive information.

References

[1] D. S. Terzi, R. Terzi, et al., *A Survey on Security and Privacy Issues in Big Data*[C], in Proc. 2015 IEEE International Conference on Internet

[2] ISO/IEC JTC 1, *Big Data* [R], Preliminary Report, 2014.

[3] Technology and Secured Transactions(ICITST' 2015), 2015.

[4] Kaustav Ghosh, and Asoke Nath. "Big Data: Security Issues, Challenges and Future Scope." International Journal of Research Studies in Computer Science and Engineering 3.3 (2016): 1-9.

[5] A.A. Hussein (2020) Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). Journal of Information Security, 11, 304-328.

[6] Dongpo Zhang. "Big data security and privacy protection." 8th international conference on management and computer science (ICMCS 2018). Atlantis Press, 2018.

[7] Sitalakshmi Venkatraman, and Ramanathan Venkatraman. "Big data security challenges and strategies." AIMS Math 4.3 (2019): 860-879.

[8] Taran Singh Bharati. "Challenges, Issues, Security And Privacy Of Big Data." *International Journal of Scientific & Technology Research* 9 (2020): 1482-1486.

[9] M. Binjubeir, A. A. Ahmed, M. A. B. Ismail, A. S. Sadiq and M. Khurram Khan, "Comprehensive Survey on Big Data Privacy Protection," in *IEEE Access*, vol. 8, pp. 20067-20079, 2020, doi: 10.1109/ACCESS.2019.2962368.

[10] M Kantarcioglu and E Ferrari (2019) Research Challenges at the Intersection of Big Data, Security and Privacy. Front. Big Data 2: 1.doi: 10.3389/fdata.2019.00001

[11] S. Subbalakshmi, and K. Madhavi. "Security challenges of Big Data storage in Cloud environment: A Survey." *International Journal of Applied Engineering Research* 13.17 (2018): 13237-13244.

[12] J. Koo, G. Kang, Y.-G. Kim, Security and Privacy in Big Data Life Cycle: A SurveyandOpenChallenges. *Sustainability* 2020, *12*,10571.https://doi.org/10.3390 /su122410571

[13] A Comprehensive Survey on Security and Privacy for Electronic Health Data" (2021) by Se-Ra Oh, et. al.

[14] Big Data Security Issues with Challenges and Solutions (2023) by Santanu Koley

[15] A Survey of Security and Privacy in Big Data (2016) by Haina Ye, et. Al

[16] Danda B. Rawat, Ronald Doku, and Moses Garuba. "Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security." IEEE transactions on services computing 14.6 (2021): 2055–2072. Web.

[17] Bardi Matturdi et al. "Big Data Security and Privacy： A Review." China communications 11.2 (2014): 135–145.

[18] KamtaNath Mishra et al. "Cloud and Big Data Security System's Review Principles: A Decisive Investigation." Wireless personal communications 126.2 (2022): 1013–1050.

[19] Adel Al-Zahrani, and Mohammed Al-Hebbi. "Big Data Major Security Issues: Challenges and Defense Strategies." Tehnički Glasnik 16.2 (2022): 197–204.

[20] Shahab KAREEM et al. "Big Data Security Issues and Challenges." Journal of applied computer science & mathematics 15.2 (2021): 28–36. Web.

[21] Julio Moreno, Manuel Serrano, and Eduardo Fernández-Medina. "Main Issues in Big Data Security." Future internet 8.3 (2016): 44–. Web.