Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2019, 6(12):297-306**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Analysis on Big Data to Data Drive Security

Kartheek Pamarthi

Kartheek.pamarthi@gmail.com

Abstract: The users of a big data platform deposit enormous volumes of sensitive data on the platform. Businesses can enhance the quality of the services they offer and reduce the expenses of providing users with tailored services through the sharing of sensitive data. However, there is a problem with safe data exchange. This research aims to provide a structure for safe data sharing on a big data platform. Data protection throughout transmission, storage, use, and deletion on an untrusted big data sharing platform is an integral aspect of this architecture. In our work, we provide a virtual machine monitor-based user process protection approach and a proxy re-encryption algorithm that use heterogeneous ciphertext transformation. Both of these approaches can be useful while setting up the system's features. The framework not only protects users' sensitive data, but it also makes sure that data shared by users is secure. On the other hand, data owners are able to maintain full control over their own data inside a secure environment that prioritises the protection of current Internet transactions.

Keywords: secure sharing; sensitive data; big data; proxy re-encryption

Introduction

Information systems in a Big Data setting often function from outside sourced, diverse, and unstructured data and engage in sophisticated information exchanges. Therefore, data quality that flows across information systems can rapidly decline over time without adequate control over input and process quality [1]. There are several elements that contribute to security risks. These include processing data "on the move," the variety of data sources, the constant acquisition of huge volumes of data, and so on [2].

As a result, we conclude that data security and quality face numerous obstacles due to the Big Data context's restrictions. Studies addressing data security and data quality have covered most of these bases in their respective literature reviews. We were unable to find any research that thoroughly discussed this topic, and even fewer that acknowledged the potential impact of data security on data quality [3].

The complexity increases when we think about potential conflicts between the two systems; as a result, we need to come up with new solutions. The goal of this article is to discuss potential problems with implementing a "Data Quality Management System" and guaranteeing "Data Security" at the same time. Research on the interplay of the two systems is necessary. We were unable to find any research that thoroughly discussed this topic, and even fewer that acknowledged the potential impact of data security on data quality [3].

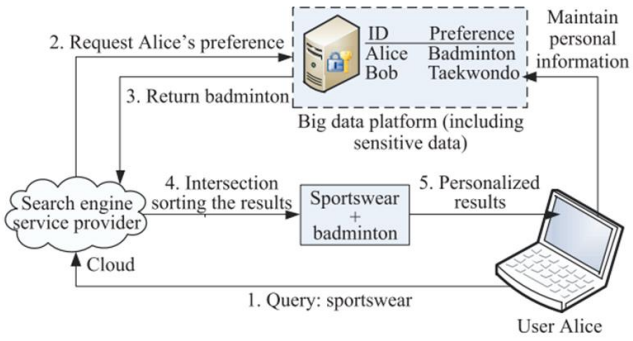

Fig. 1: Utilisation of sensitive information (the preferences of the user).

User preferences are depicted in Fig. 1 as sensitive data. First thing the Search Engine Service Provider (SESP) does when Alice enters a query (sportswear) is checks the big data platform for Alice's preference. When Alice does a search using the keyword "badminton"—which the big data platform has recorded and shared—the engine provides results that are tailored to her specific interests (sportswear + badminton). Alice has a good shopping experience whenever she finds her preferred badminton apparel. As a result, everyone comes out ahead [5]. Data sharing does boost company assets, however there are security concerns with sharing sensitive information due to Internet vulnerability and the possibility of data leaking. There are four main security considerations for the safe transfer of sensitive information. Before anything else, moving private information from a data owner's own server to a remote big data platform raises security concerns. Secondly, the big data platform may have security issues with the computation and storage of sensitive data. The third concern is the security of sensitive data stored in the cloud. Concerns about safe data deletion come up in the fourth place [6]. A number of domestic and international research organisations and academics have made significant strides in the investigation and study of potential solutions to these security issues.

Literature Review

The criteria for judging a property are laid forth by a quality metric. It may be based on objective metrics that account for the wants, needs, and experiences of stakeholders, or it can be based on subjective evaluations that consider the opinions, needs, and experiences of stakeholders, such as a feedback questionnaire or user surveys.
References [7, 8]. Quality evaluation metrics are classified into three groups by Bizer [9] based on the information type employed as a quality indicator. One kind of metric is content-based, which uses the information itself as an indicator of quality; another is context-based, which uses metadata to indicate quality regarding the circumstances of the information's creation or use; and a third type is rating-based, which relies on explicit ratings for the information, its sources, or its providers.
Assessing the direct and indirect expenses of poor data quality and developing plans and methods to meet quality goals are the primary concerns of those working to improve data quality. Two primary approaches to enhancing product quality are outlined in [10].
Altering the values of the data is the first kind of tactic. Examples include fixing inaccurate values, removing duplication, and updating obsolete data values. Data cleansing, record linking, data integration, standardising values, acquiring fresh data, and standardising schemas are all examples of ways that fall under this category of strategy. Data creation and modification process re-design is the second kind of strategy. A data format check-up before storage, a validation phase to ensure data source trustworthiness, etc., are all examples of such improvements. Traditional data quality evaluation approaches do not work in Big Data settings, despite the fact that there are many such models. All data quality management systems face these four main obstacles:
High Volume: Due to the massive amount of data, it is challenging to assess and enhance data quality in a fair amount of time [11]. Also, a scalable solution has to be put in place because the amount of data is always going up. Data quality strategies' scalability is a measure of how well they handle bigger and more complicated datasets over time [12].
• Heterogeneity: Most data generated nowadays is not structured and is either semi-structured or unstructured, which makes processing it more difficult than structured data. It is difficult to analyse the connections and semantics in unstructured data [13]. It is also not uncommon for semi-structured data to be impracticable or extremely complex to transfer into structured formats [14].
The data we have today changes at a dizzying rate and can become irrelevant at any moment. Making ill-informed decisions due to a lack of up-to-date information is possible if data collection is not done quickly and accurately [15]. Protecting sensitive information: To assess data quality and make required modifications, flexible and easy access to all data is required. On the other hand, data security measures could make data quality management more difficult and time-consuming. Data access can be hindered, for example, by privacy and confidentiality safeguards [16].
The three pillars of traditional security—availability, integrity, and confidentiality—remain unchanged. There are three tenets of data security: availability, integrity (data cannot be changed without authorization), confidentiality (data cannot be accessed by unauthorised parties), and integrity (data cannot be altered without authorization). When it comes to data security, ISO/IEC 27001 [17] covers attributes including dependability,

authenticity, accountability, and non-repudiation. Many studies in the field of Big Data have highlighted the importance of securing sensitive and personal information as the main goal of security. Managing user consent for personal data and complying with legal and regulatory requirements are aspects of confidentiality that pertain to privacy. In this part, we will go over the main obstacles encountered by security systems and the dangers that could compromise data at any point in the Big Data lifecycle.

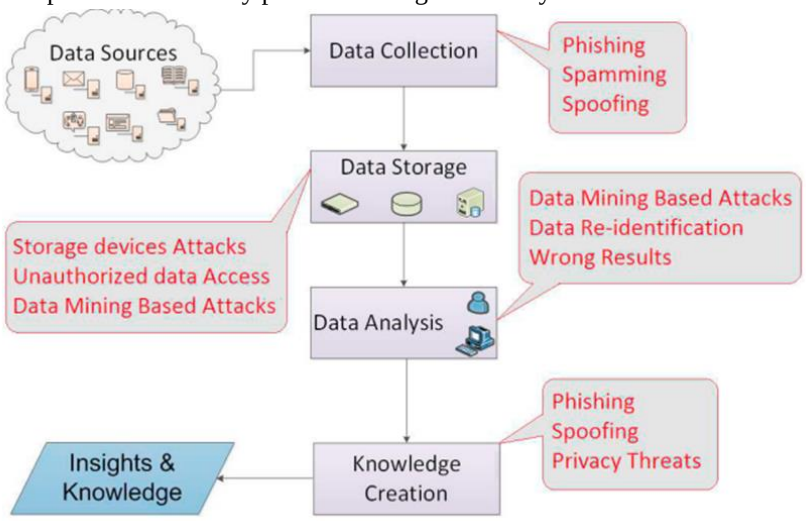

Fig. 2: Life Cycle Model for Big Data Security.

Collecting information and data from various sources is the first step. Data storage security is the next step after data collection. In the last stage, we apply data mining tools to uncover significant insights. Every one of these steps could be dangerous from a security perspective. The many security threats linked to these processes are illustrated in Figure 2.

Big Data Security: Advantages, Challenges, And Best Practices

In today's interdependent digital world, large data poses risks owing to security flaws but also brings opportunities for innovation. Information about consumer habits, company efficiency, and industry tendencies can be found in the mountains of data created by companies and people. With this new knowledge, companies may improve their decision-making, operations, and product development processes. However, organisations still face a significant difficulty in guaranteeing the security of large data, since breaches can result in significant losses for both individuals and corporations.

What is big data security?

Big data security is an umbrella term for a set of procedures and policies put in place to safeguard massive amounts of data, or "big data," from dangers like viruses and illegal access. All of this work is geared at making ensuring that data is secure, accessible, and private. Implementing strong access controls, authentication, and authorization procedures; continuously monitoring; detecting and responding to threats; providing thorough employee training; and more are all essential parts of large data security management. It is critical for businesses to secure large data in order to protect intellectual property, financial information, and personal consumer data. This aids in making well-informed decisions, builds confidence with customers, and helps organisations comply with data protection rules.

Benefits of big data security

Organisations can fully utilise big data with the help of big data security, which helps to reduce risks, build trust, and drive innovation and growth. The main advantages of large data security will be examined here and are shown in figure 3.

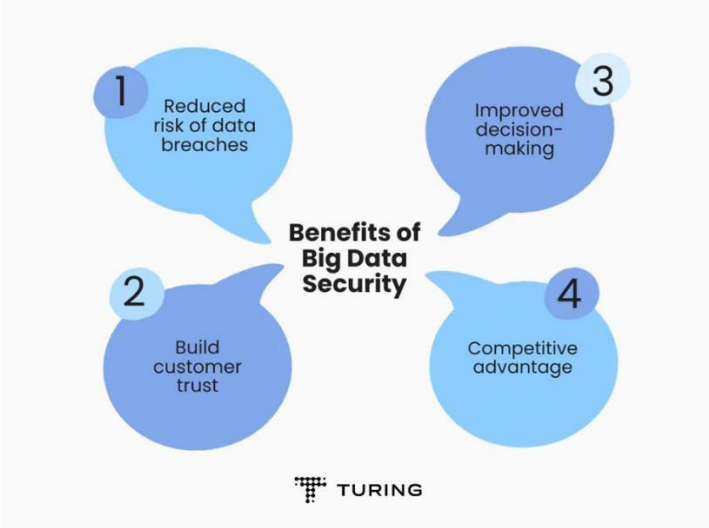

Fig 3: Benefits of Big data Security

a. Reduced risk of data breaches

The installation of various protective measures aimed at preserving data confidentiality, integrity, and availability is crucial for big data security in limiting the danger of data breaches. Important strategies for reducing the possibility of data breaches include strong data encryption, role-based access control, proactive threat detection, and continuous real-time monitoring. To further improve data security and decrease the likelihood of breaches, big data security solutions use technologies such as firewalls, intrusion detection systems (IDS), and intrusion prevention systems (IPS) to keep an eye on network activity, spot possible dangers, and quickly prohibit suspicious activities.

b. Increased customer trust

Ensuring strong data security is essential for building trust with customers in the modern digital age. Customers are becoming increasingly wary of how companies handle their private data in light of the prevalence of data breaches. Statista reports that just 46% of American customers have faith in banking institutions to keep their personal information secure. This number highlights the fact that consumers and companies do not trust one another when it comes to matters of data security and privacy. In order to recover and keep customers' trust in the responsible management of their data, organisations must establish and adhere to robust data security procedures. When it comes to protecting client information from prying eyes, big data security is important. Customers have more faith in a business and are more likely to remain loyal if they see that the firm values the security of their personal information. To validate their dedication to data security and reassure their clients that their data is safe, many organisations hire trustworthy third parties to do security audits. This preventative measure does double duty: it helps the organisation retain customers' trust while simultaneously enhancing its image as a trustworthy steward of confidential data.

c. Improved decision-making

By preventing unauthorised parties from gaining access to sensitive information, big data security is vital in maintaining the reliability and correctness of data. Only authorised persons can access critical information thanks to measures like encryption, robust authentication processes, and strict access controls. Organisations may empower stakeholders to make data-driven decisions by ensuring a safe data environment allows for confidently deriving correct insights and patterns.

To better serve clients with excellent credit histories, financial institutions use big data to improve their risk management and fraud detection systems. The safety and correctness of the underlying data, however, are crucial to the success of these applications. Safeguarding data helps keep private information safe and maximises the use of big data analytics to boost company results and customer satisfaction.

d. Competitive advantage

Businesses get an advantage with big data security because it protects important assets and makes data-driven decisions easier. Companies can increase customer retention rates by fostering trust and loyalty among their

client base through the prioritisation of data protection and privacy assurances. Strategic partners who help a company grow and expand are more likely to work with a company that has strong security measures in place. All of these benefits put the business ahead of rivals who haven't put money into big data security analytics quite yet, and they're helping the company develop. Businesses can improve operational efficiency, expand product offerings, and maintain a competitive advantage in the marketplace by utilising safe data procedures. This also helps limit risks connected with data breaches.

Common big data security challenges

Figure 4 shows big data security challenges. In today's digital environment, attackers use sophisticated tools and creative strategies, making big data security a serious concern. In order to adequately safeguard their data, companies must understand the major obstacles in big data security. In order to properly safeguard their data, firms must handle the following key challenges:

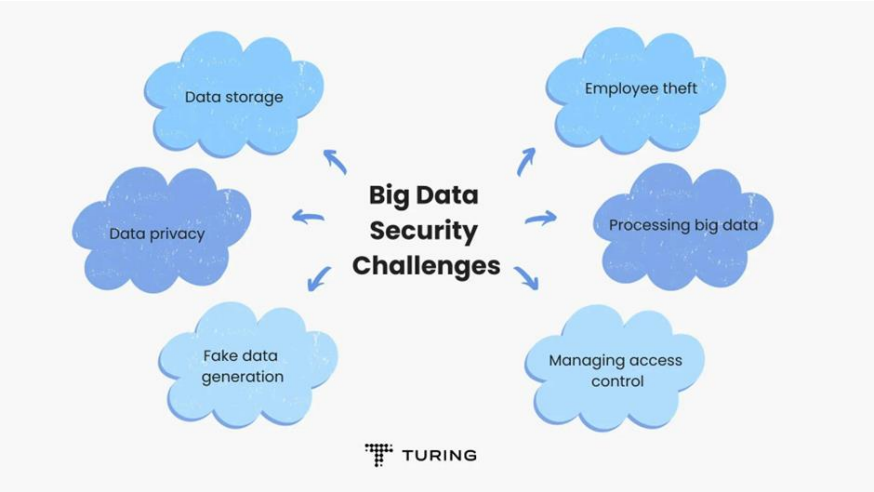

Fig 4: Big data security challenges

a. Data storage

Securing massive amounts of data is a challenge when dealing with big data. Implementing appropriate security measures for all data kinds becomes challenging when dealing with big data systems, which store a variety of data types, including structured, semi-structured, and unstructured data. Furthermore, sensitive information may exist in more than one place due to data redundancy and replication, which is typical in big data architecture and raises the possibility of illegal access.

b. Data privacy

Since big data systems frequently gather and retain vast quantities of personal data, data privacy poses a substantial threat to big data security. Businesses face challenges in securing and maintaining data privacy due to the fact that it gathers data from various sources, including online and offline actions. Data breaches and illegal access are further concerns with big data platforms because they include exchanging data with third-party apps and services.

c. Fake data generation

Another issue with big data security is the fabrication of fake data, which can trick and manipulate big data systems. Businesses may end up making poor choices due to erroneous data and insights caused by this difficulty. In order to influence the buying decisions of potential clients, criminals may, for instance, create false product reviews. In addition, attackers can more easily steal sensitive information if they utilise phoney data to hide real data.

d. Managing access control

The data stored by big data systems spans numerous servers and storage facilities, making them extremely complicated and dispersed. The difficulty in implementing and managing access restrictions that are compatible with all data types is a direct result of this. Additionally, big data platforms may store and share massive amounts of data with other apps and services. Unauthorised access is always a concern when dealing with data that is both large and diverse, making access management a formidable obstacle.

e. Processing big data

Due to the data's exposure to numerous third-party programmes and servers, processing big data—defined as complicated and scattered data across numerous systems—involves substantial risk. In a big data system, data is created and processed quickly, frequently in real time. Because of this rapid pace, security concerns are hard to keep an eye on and react to quickly. Processing the data while maintaining security measures needs meticulous planning and the implementation of strong security standards, especially as the amount increases.

f. Employee theft

Even more so for individuals engaged in big data analysis, all employees have access to the data to a certain degree. Even more concerning is the fact that some workers have first-hand knowledge of the company's security procedures, passwords, and access controls for its data systems. The ability to obtain sensitive information can be used by an employee who has access to a big data system. They can also damage the company's finances and image by manipulating data.

Top ten big data security best practices

Protecting sensitive information and ensuring data integrity requires strong big data security procedures to be implemented. For the purpose of keeping large data safe, here are ten guidelines that companies should follow.

1. Encryption

Encryption is essential for the security of huge data because it changes the data's format into ciphertext, which is unintelligible to anyone without the right key. Data held on computers, servers, and inside a network can be protected by encrypting it so that it stays confidential both while in transit and at rest. Even if an intruder were to obtain the data, they would be unable to decipher it. Encryption also makes it harder to alter data by keeping it intact.

2. Effective user access control

Businesses must prioritise the protection of big data since it contains important and sensitive information. To lessen the likelihood of data breaches, theft, or unauthorised access, effective user access management makes sure that only authorised people may see, edit, or remove this data. User access control for large data can be implemented in various ways. One popular method is role-based access control, which enables administrators to establish roles and then grant access based on those responsibilities.

3. Monitoring cloud security

The on-demand scalability offered by cloud platforms allows organisations to expand their big data infrastructure in response to increases in data volume, which is a major benefit for enterprises utilising big data analytics. When dealing with the unpredictable workloads that come with big data analytics, this adaptability is crucial. It is critical to detect risks and protect big data assets since cloud infrastructure is susceptible to cyber assaults owing to exposed API keys, tokens, and misconfigurations. To guarantee cloud security, organisations can use monitoring technologies that can identify unauthorised attempts to access data or exfiltration.

4. Network traffic analysis

Unusual data transfers or sudden spikes in traffic are examples of abnormalities in network behaviour that can be detected using network traffic monitoring. These events could indicate possible security issues, such as data breaches or insider assaults. Improving the capacity to detect and mitigate hazards prior to their infliction of substantial harm, network traffic analysis can also reveal patterns associated with particular forms of attack, such as malware, phishing, distributed denial of service, or mitM assaults. On top of that, it's useful for keeping tabs on how well companies are following security guidelines and industry rules in real time.

5. Vulnerability management

The intricate and sensitive data stored in big data platforms makes them easy prey for cybercriminals. To steal information, halt operations, or even drain funds, criminals might leverage big data platforms. Because it helps discover and patch vulnerabilities proactively, vulnerability management is vital for big data security. This reduces the chance of data breaches, leaks, and unauthorised access to critical information.

6. Employee training and awareness

To ensure the safety of big data, it is crucial to educate employees on the risks they face and how to mitigate them. A study conducted by cybersecurity firm Tessian and Stanford University professor Jeff Hancock found that 88% of data breach instances are caused by human error, according to a report by Security Today. Big data security best practices, such as strong password creation, phishing email identification, and suspicious activity

reporting, can be taught to employees through proper training. Data protection standards are complex, but with proper training, your staff will be able to follow them.

7. Insider threat detection

Identifying insider threats early on helps avert any large security incidents that may occur later on, which is crucial for big data security. Employees, independent contractors, and anybody else with access to a company's systems and data might pose a hazard. Financial gain, vengeance, or some other malevolent purpose might be driving them. The individuals committing insider threats frequently have legal access to the information and systems they aim to compromise, making detection a challenging task. On the other hand, insider risks can be significantly reduced with the help of tools like behaviour analytics, anomaly detection, user profiling, and data access monitoring.

8. Prompt incident response plan

Big data security relies on having a plan in place to respond quickly to incidents. In the event of a cyberattack, it is important for organisations to have a plan in place that outlines how to respond swiftly and efficiently in order to limit damage and recover data as soon as possible. Data validity and restoration following an incident is facilitated by an expedited reaction plan, further assuring the data's dependability and accuracy. In order to formulate a strategy for swiftly responding to incidents, organisations should think about the following:

• Identify the types of incidents that can occur
• Develop a specific response plan for each type of incident
• Assign roles and responsibilities
• Test the incident response plan regularly

9. Real-time compliance and security monitoring

Big data security in the modern digital age requires constant monitoring for compliance and security as organisations work with massive amounts of data. In order to prevent harm, organisations should constantly check safety and compliance issues. This way, they can catch any questionable actions in the early stages. Compliance rules and regulations like GDPR, HIPAA, and PCI-DSS apply to big data because of the sensitive information it routinely manages about companies and customers. To keep big data processing activities in line with rules and prevent penalties and reputational harm, organisations can use real-time monitoring to generate notifications in the event of a violation.

10. Regular data backup

A lack of data backup protection leaves you vulnerable to security incidents like malware assaults and data breaches, which can result in irretrievable data loss. Always be ready for the worst, even if you take every precaution to avoid cyberattacks. Ensuring data security is of the utmost importance. By regularly backing up data, organisations can recover corrupted or lost data, minimising business disruptions and potential financial losses. Furthermore, this gives customers faith that you can restore important data in the event of a data security breach, protecting your credibility and reputation.

Privacy Preserving Apriori Algorithm in Mapreduce Framework

Hiding a needle in a haystack

Current association rule algorithms that aim to protect user privacy add noise to the actual transaction data. Since the objective is to avoid a decline in data utility and a breach of privacy, this effort preserved the original transaction within the noised transaction. As a result, the approach still has the chance that an unreliable cloud service provider deduces the actual collection of frequently used items. The "hiding a needle in a haystack" concept is the basis of this privacy-preserving technique, therefore it is important to provide sufficient privacy protection despite the possibility of association rule leakage. The underlying principle of this approach is that, as illustrated in Figure 5, it is challenging to spot a needle in a haystack of data, such as a huge dataset. The necessity to address the privacy-data utility trade-off means that existing approaches cannot inject noise randomly. On the contrary, this method requires more computational resources to generate the "haystack" that will conceal the "needle." Thus, it's important to weigh the pros and downsides of using the Hadoop architecture in the cloud to address various issues. The original association rule is shown by the dark diamond dots in Figure 5, whereas the noised association rule is represented by the empty circles. Having so many noised association rules makes it difficult to uncover original rules.

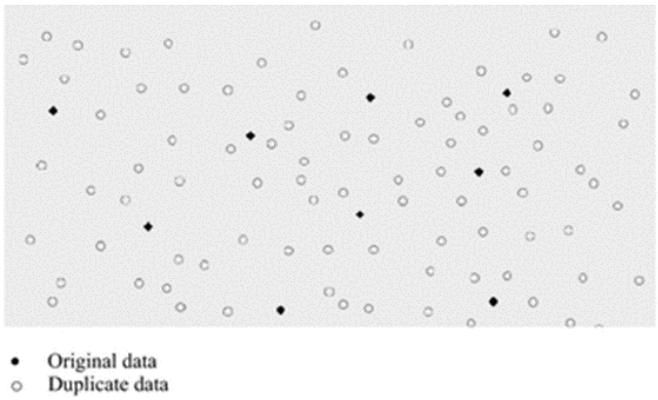

Fig. 5: Hiding a needle in a haystack Mechanism of hiding a needle in a haystack is shown

Figure 6 shows how the service provider introduces a fake item into the original data set of transactions in order to make it more noisy. After then, both the real and fake goods are given a distinct code. After an external cloud platform extracts frequently used items, the service provider keeps the coding information to filter out the dummy item. The data given by the service provider is used by the external cloud platform to execute the Apriori algorithm. To the service provider, the external cloud platform returns the value of the frequently used item set and support. The service provider uses a code to filter out the often used items that are impacted by the dummy item, and then uses the frequently used items without the dummy item to extract the right association rule. Since there isn't a lot of computation involved in extracting the association rule, the operation isn't a strain on the service provider.

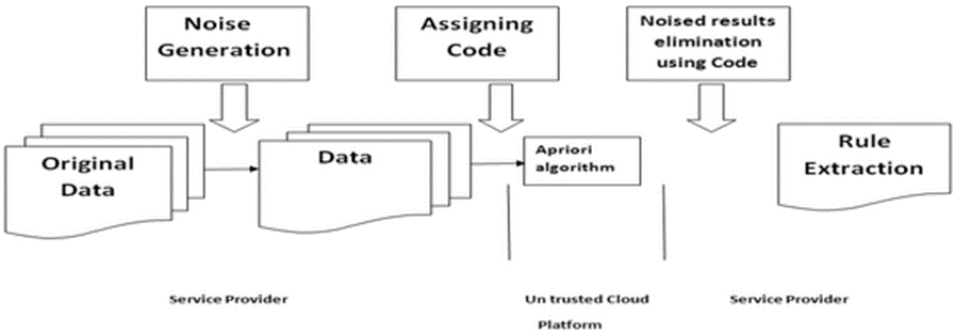

Fig. 6: Overview of the process of association rule mining the service provider adds a dummy item as noise to the original transaction data collected by the data provider

Privacy-preserving big data publishing

Commercial, academic, and medical applications rely heavily on the production and sharing of raw data. As the number of open platforms, such social networks and mobile devices, from which data can be collected grows, so does the volume of this data. Input privacy and output privacy are the two main categories into which privacy-preserving models often fall. Publishing anonymised data with models like k-anonymity and l-diversity is the main concern in input privacy. Problems like association rule concealing and query auditing, which involve perturbing or auditing the output of various data mining algorithms to maintain privacy, are typically of interest in output privacy. There has been a lot of research on privacy that has concentrated on how useful the public data is and how well privacy is preserved (vulnerability quantification). Separating the data into smaller pieces (fragments) and then anonymizing each one separately is the best solution. Due to the absence of variation in the sensitive property inside the equivalence class, k-anonymity is unable to protect against attribute disclosure attacks, even though it can prevent identity attacks. According to the l-diversity paradigm, there should be a minimum of l sensitive values present in every equivalence class. In order to achieve significant performance improvements, it is common practice to process huge data sets using distributed platforms like the MapReduce framework. This allows for the distribution of a costly process among numerous nodes. Consequently, privacy model enhancements are implemented to address the inefficiencies.

An integral part of trust management is trust evaluation. A trust value is a continuous or discrete number that is obtained by evaluating the elements influencing trust using evidence data. This is a technological way to describing trust for digital processing. It suggests two methods for protecting personal information during trust assessments. To accomplish the goal of privacy preservation and evaluation result sharing among different requestors while reducing computation and communication costs, it is proposed to implement two servers. Here we have two separate service providers who, for reasons of their own company, do not conspire with one another. To protect the confidentiality of the companies under review, one option is to use an authorised proxy (AP), which is in charge of managing collected evidence and controlling access to it. The alternative is an evaluation party (EP) that processes the data obtained from several trust evidence suppliers; this might be provided by a cloud service provider, for example. An encrypted trust pre-evaluation result is generated by the EP after processing the encrypted data. First, the EP verifies the user's access eligibility with AP when they seek the pre-evaluation result.

Conclusion

In conclusion, we presented a model for the systematic sharing of sensitive data on big data platforms that uses the heterogeneous proxy re-encryption algorithm to secure the submission and storage of sensitive data and the VMM to secure the use of clear text in the cloud by the private space of user processes. Protecting users' sensitive data is a top priority for the proposed framework. As a workable approach to equitably distributing advantages under semi-trusted circumstances, data owners retain full ownership of their data. To further enhance encryption efficiency, we intend to optimise the heterogeneous proxy re-encryption technique in the near future. Furthermore, a significant area for future research is the optimisation of interface overhead. We cover the privacy issues that arise at each stage of the big data life cycle and go over the pros and cons of current privacy-preserving solutions as they pertain to big data applications. Both older and more modern methods of protecting personal information in large data are covered in this paper.

References

[1]. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys 2009; 41
[2]. Crosby PB. Quality is free. New York:McGraw-Hill 1979
[3]. ISO, ISO/IEC 25012:2008—Software engineering. Software product quality requirements and evaluation (SQuaRE). Data quality model, Report, International Organization for Standarization 2009
[4]. Laranjeiro N, Soydemir SN, Bernardino J. A Survey on Data Quality: Classifying Poor Data. IEEE 21st Pacific Rim International Symposium on Dependable Computing 2015; 179-188
[5]. Bertot JC, Choi H. Big Data and e-Government: Issues, Policies, and Recommendations. The Proceedings of the 14th Annual International Conference on Digital Government Research 2014
[6]. Müller H, Freytag JC. Problems, Methods, and Challenges in Comprehensive Data Cleansing. Technical Report HUB-IB-164, Humboldt University, Berlin 2003
[7]. Redman TC. Data's Credibility Problem. Harvard Business Review 2013
[8]. Cappiello C, Francalanci C, Pernici B. Data quality assessment from the user's perspective. The 2004 international workshop on Information quality in information systems 2004; 68-73
[9]. Bizer C. Quality-driven information filtering in the context of web-based information systems. Ph.D. Thesis. Freie Universit, Berlin 2007 [10] Merino J, Caballero I, Rivas B, Serrano M, Piattini M. A Data Quality in Use model for Big Data. Future Generation Computer Systems 2015
[10]. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal 2015; 1-10
[11]. Chen M, Mao S, Liu Y. Big Data: A Survey. Springer Science+Business Media 2014
[12]. Barna S, Divesh S. Data Quality: The other Face of Big Data. IEEE, the 30th International Conference on Data Engineering 2014
[13]. Katal A, Wazid M, Goudar RH. Big Data: Issues, Challenges, Tools and Good. IEEE, the 6th International Conference on Contemporary Computing (IC3) 2013

[14]. Strong DM, Yang WL, Wang RY. Data uamity in context. Communications of the ACM 1997; 41:103-110

[15]. ISO, ISO/IEC 27001:2013-Information technology -- Security techniques -- Information security management systems -- Requirements, International Organization for Standarization 2013

[16]. Hakuta K, Sato H. Cryptographic Technology for Benefiting from Big Data. Springer, The Impact of Applications on Mathematics 2014;85- 95