International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Data Security and Privacy in Data Engineering

Sultan Yerbulatov

Lead Data & Analytics Engineer, Chevron Eurasia Business Unit LLP Tengizchevroil
Atyrau, Republic of Kazakhstan
Email: *sultan.yerbulatov[at]gmail.com*

Abstract: ***Due to the increased amount of data, it has become a key resource for strategic decision - making, data security and privacy issues are becoming more relevant than ever. In the context of Data Engineering, where vast amounts of information are collected, processed and stored, ensuring reliable data protection and maintaining their confidentiality becomes a top priority. Effective data management involves not only their technical processing, but also a guarantee that the entire information lifecycle is accompanied by appropriate security measures. The purpose of the work is to consider such an important aspect when working with data as data security and confidentiality in Data Engineering. The methodological foundations are: scientific works of domestic and foreign authors, articles, expert opinions.***

Keywords: Data Engineering, data, information, modern technologies, IT, security, confidentiality

1. Introduction

The rapid development of Internet technologies has united people all over the world. Nowadays, gathering people virtually is more feasible than physically. This virtual connection between people led to the emergence of the cyber community. The daily interactions of people in the cyber community generate a huge amount of data, known as “Big Data”.

Big Data is information of large volume, high transmission speed and great variety, which requires new forms of processing to improve decision making, information retrieval and process optimization. A data set can be called "Big Data" if it is collected, analyzed, stored, filtered and visualization of data is difficult with traditional or modern technologies. In general, big data technology involves the process of collecting, storing and extracting valuable information from data, as shown in Fig.1.

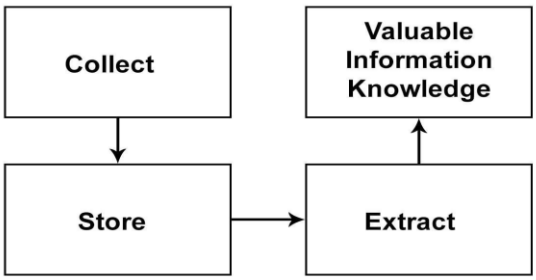

Figure 1: Technological process of big data processing

The global big data analytics market is projected to be worth $103 billion in 2023, highlighting the growing importance and integration of big data into business operations. The volume of data worldwide is expected to continue to grow exponentially, with projections suggesting that the world will generate a staggering 181 zettabytes of data by 2025. This growth is fueled by the ubiquity of digital technologies and the Internet of Things (IoT), with the number of IoT users projected to grow to 41.6 billion by 2025.

The digital universe currently spans 64 zettabytes of data, highlighting the enormous amount of digital information available. This data is predominantly generated by users and accounts for 70% of the world's data, including social media posts, emails, customer reviews and more. Interestingly, by 2024, the ratio of original to duplicate data is projected to be 1: 10, indicating significant redundancy in the digital information industry [7, 8].

However, despite the significant benefits, the field of “big data” also has problematic aspects in the field of data security and confidentiality. Key issues in this area include aspects such as confidentiality, integrity, availability, monitoring and auditing, as well as key management and data confidentiality. Achieving confidentiality requires the use of AAA principles - authentication, authorization and access control. Data integrity can be ensured through data source verification, information validation, data loss prevention, and deduplication.

Given the huge volume of data generated by various sources, cloud technologies are widely used as external storage. Ensuring the protection of data transmitted by third party providers requires the implementation of an access control system that ensures compliance with privacy and security principles. Using attribute - based encryption is a common method of protecting transmitted data. This may include encryption based on key policy attributes or encryption based on ciphertext policy attributes [1].

1) General characteristics of data development

Data engineering is the active processing, cleaning, and readiness of data for analytics, data science, and artificial intelligence implementation. Mainly, this includes creating an ETL data infrastructure and an ELT pipeline for machine learning, as well as implementing data quality checks and a data pipeline. The job of a data engineer is to assist data scientists and analysts in creating and implementing data - centric platforms.

The need for data engineering is important to businesses for several reasons:

- Technology stack.
- A team providing support for data science projects.
- Configuration that makes it easier to make data - driven business decisions.

- Applying data models to create predictive and prescriptive analytics for better results.

Organizations implementing analytics and data science projects prefer to deploy qualified data engineers. Following the guidelines of data architects, data engineers use a variety of tools and technologies to perform various tasks, including:

- Setting up connections to data sources.
- Creation of data warehouses for intermediate storage of information.
- Extract data from various sources.
- Managing large amounts of data.
- Ensuring data quality and resolving disputes.
- Process data to create standardized data.
- Setting up and maintaining data pipelines.
- Batch processing of streaming and real - time data.

Data engineering relies on a variety of big data processing technologies, including Hadoop cluster, Apache Spark, Splunk, Apache Flink, Azure HDInsight; NoSQL data stores such as Apache Cassandra database, MongoDB; databases with in - memory cache, such as Redis, SAP HANA; data processing tools such as Apache Kafka, Apache NiFi, Informatica Cloud services; cloud tools such as ASES data pipelines, Google Big Query and Azure Data Factory; standard DBMS and file systems; various scripts for specific operating systems, such as Linux Shell scripting, Windows batch scripting and Power shell scripting; cloud storage such as S3; API - based tools such as AWS API gateway; time series data warehouses; IoT - specific tools such as Node - Red.

Apart from data, data engineers become proficient in BI tools such as Tableau, MS Power BI, which makes it easier to provide data in a suitable format and structure.

Data science experts are also familiar with cloud and DevOps tools such as Jenkins and Docker to create effective implementations.

The field of data engineering mainly involves data pre - processing, which reduces the overhead of data scientists and analysts during the data preparation stages. To better understand this, below is a high - level overview of the data engineering setup framework for data scientists [2].

- It collects raw data from various sources such as applications, file systems, IoT sensors, and other file stores using ETL (extract, transform, load) or ELT (extract, load, transform) pipelines.
- While ETL is primarily aimed at creating data warehouses, ELT is designed for big data processing frameworks.
- Data quality processes and transformation techniques are an integral part of data development.
- The pre - processed data is stored in a data warehouse or data warehouses for later use.
- The setup provides input to the Data Science Framework.
- Data analysts and data scientists perform the primary analysis to develop objects.
- The data is used to create reports, business intelligence charts, and machine learning applications.
- Feature development is an iterative process to optimize data processing using machine learning.
- Data scientists iteratively apply different machine learning models to create the best model for a particular use case.
- Input data plays a key role in training and testing models during the development process.

If we talk about the advantages, then they include:

- Pre - process data from various sources and formats into a standard format and structure.
- Pipeline automation to obtain additional data or latest data using batch automation and scheduling tools.
- Support for real - time analytics using advanced technologies such as Apache Kafka, Spark and data - bricks.
- Enforce governance policies and ensure compliance with data security requirements using masking and encryption of sensitive information.
- Generate ready - to - use data to quickly complete analytics projects.
- Setting up the data structure in accordance with the machine learning algorithm, based on the recommendations of data processing specialists [3].

2) Differential privacy

Differential privacy is a mathematical definition of the concept of “having privacy. ” It is not a specific process, but rather a property that a process can have. For example, one can calculate (prove) that a given particular process satisfies the principles of differential privacy. Simply put, for every person whose data is included in the analysis set, differential privacy ensures that the result of the differential privacy analysis will be virtually indistinguishable regardless of whether your data is in the set or not.

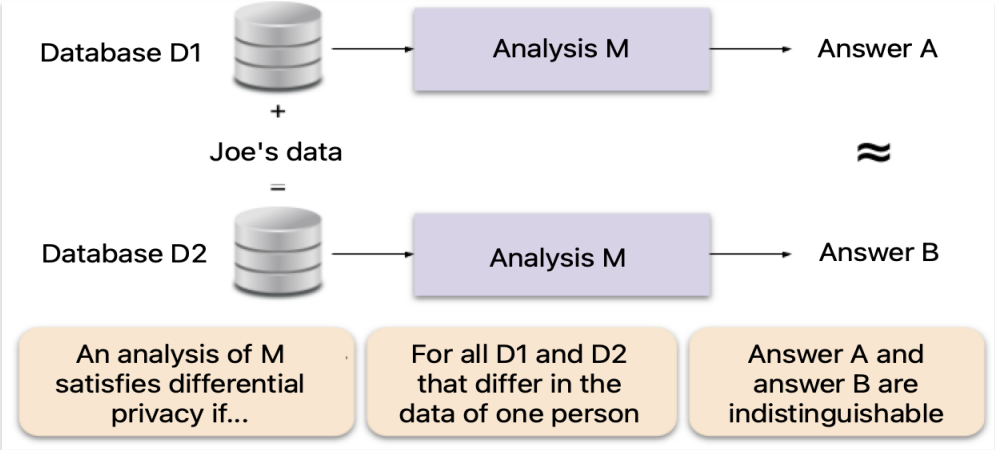

Figure 2: Schematic representation of differential privacy

The principle of differential privacy is shown in Figure 2. Answer A is calculated without Joe's data, and answer B is calculated with his data. And it is argued that both answers will be indistinguishable. That is, whoever looks at the results will not be able to tell in which case Joe's data was used and in which it was not used.

You can control the required level of privacy by changing the privacy parameter ε, which is also called privacy loss or privacy budget. The smaller the ε value, the less distinguishable the results and the more protected the data of individuals.

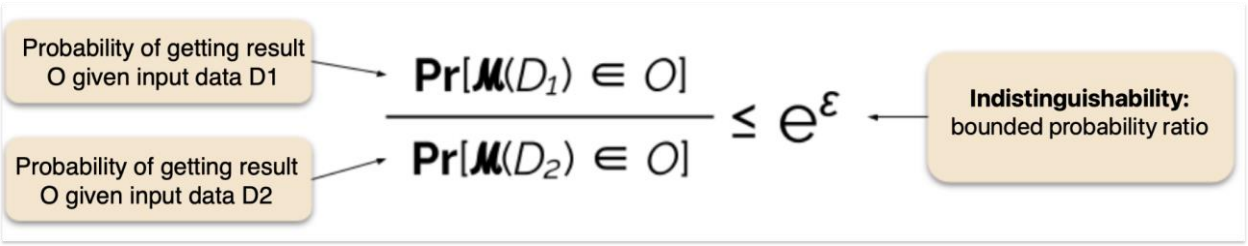

Figure 3: Formal definition of differential privacy

Differential privacy stands out among other techniques by providing a number of important advantages.

1) Extensive identification: This approach assumes that any information is considered identifying, freeing us from the difficult (and sometimes impossible) task of identifying all identifying attributes in a data set.
2) Ancillary information resistance: Differential privacy is resistant to ancillary information attacks and prevents connections to anonymized data.
3) Compositionality: This approach is compositional, which allows us to determine the overall loss of privacy when performing two differentially private analysis operations on the same data. This characteristic provides reasonable privacy guarantees even when analyzing the same data multiple times, in contrast to techniques such as anonymization, which lack compositionality and can lead to catastrophic privacy losses.

Taking advantage of these advantages, using differential privacy in practice is preferable to some other techniques. However, it should be noted that this technique, despite its effectiveness, is still relatively new. Outside the academic community, it is difficult to find proven tools, standards and proven approaches. But we are confident that with growing interest in reliable and simple solutions for ensuring data privacy, the situation will soon change [4].

3) Privacy, security and social implications of data mining

Data mining is a unique technique for finding valuable information in vast data to solve real - world problems. This is not simply a combination of the words “data” and “analysis”, but rather a masterful fusion where data representing a set of events is passed through a mining filter to highlight essential information. Also known as knowledge discovery in databases (KDD), this type of analysis involves a multifaceted approach to data processing and classification. There is a wide range of techniques used in data mining that transform raw data into usable information. This approach has applications in areas such as anomaly detection, anti - fraud, counter - terrorism, and assisting in crime investigations through identification of lies. Data mining makes a significant contribution to improving customer service, customer satisfaction and ultimately the overall look of our daily lives, whether we realize it or not.

Importance of Data Mining:

- It allows you to conduct research in large databases, highlighting only reliable information with improved segmentation.
- An effective and cost - effective solution to detect risk and fraud, helping to drive profitability.
- Occasionally assists clients facing purchasing difficulties in making decisions and increasing sales.
- Data mining techniques help organizations plan in real time and save time.
- A means to save money through fraud detection.

Application areas of data mining:
- Healthcare of the future.
- Market basket analysis.
- Manufacturing Engineering.
- Fraud detection.
- Intrusion detection.
- Customer segmentation.
- Financial banking.

Data Mining Architecture: The architecture of this approach reveals the process of data extraction. It includes data sources, analysis engine, data warehouse server, template evaluation, user interface and knowledge base. All of these steps work in concert to ensure effective data analysis.

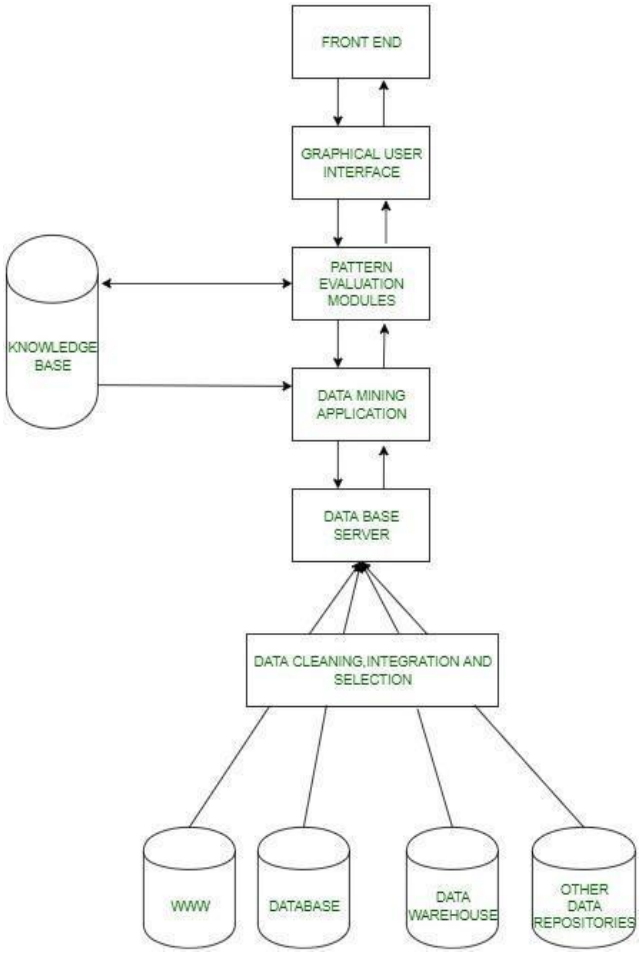

Figure 4: Data mining architecture

Privacy, Security and Social Impact of Data Mining:
Maintaining security and confidentiality has always been a priority. Previously, this was done through forecasting future events based on the analysis of previous data. Let's say we purchase a product and previous purchases are used to predict what becomes a threat to our personal information. The ongoing development of data mining poses significant risks to data security and privacy that require robust protection. The real danger is the possibility of unauthorized access to information, making it difficult to prevent misuse. Thus, a system is needed to ensure that data and its resources are protected in terms of authenticity and integrity.

Ways to ensure data security:
- Installing a multi - level security system to minimize data loss by reducing security settings.
- Providing access only to those who have permission to access the data.
- Verify user identity for data.

However, some privacy techniques can keep data private and extract useful information from a dataset.

Privacy Preserving Data Mining (PPDM): The main goal of PPDM is to ensure data privacy and extract only the necessary information. PPDM includes various methods divided into categories:
- Data Hiding Method: Transforming data so that sensitive information is not visible to others. Techniques include cryptography, data corruption, and anonymization.
- Knowledge hiding technique: Extracting sensitive content from data using data mining algorithms.
- Hybrid method: Combining two methods with limitations to improve efficiency.

Social Impact of Data Mining: Data mining is innovatively impacting our daily lives, influencing work processes, shopping, information searching and saving us time. The presence of data mining is noticeable in the fields of healthcare, finance, marketing and social media. Thus, data mining plays a key role in ensuring data privacy and security, while contributing to the improvement of life and processes in various fields [5, 6].

2. Conclusion

Clearly, privacy technology is essential to modern data use. As the volume of data increases and its critical role in business processes, it becomes clear that insufficient protection can lead to serious consequences, such as data leaks, violations of legislation and loss of trust from users.

Data Engineering professionals must continually improve their security practices and strategies in response to ever - changing threats and standards. This includes not only technical measures such as encryption and access monitoring, but also ensuring legal compliance and creating a security culture within the organization.

References

[1] Big Data Security and Privacy Issues – A survey. [Electronic resource] Access mode: https: //ukdiss. com/examples/big - data - security - and - privacy - issues. php. – (accessed 25.01.2024).
[2] What is data mining? [Electronic resource] Access mode: https: //www.educba. com/what - is - data - engineering /. – (accessed 25.01.2024).
[3] What is date engineering and its role. [Electronic resource] Access mode: https: //sky. pro/media/chto - takoe - data - inzhiniring - i - ego - rol /. – (accessed 25.01.2024).
[4] Differential privacy — data analysis with confidentiality (introduction to the series). [Electronic resource] Access mode: https: //habr.

com/ru/companies/domclick/articles/526724 /. – (accessed 01/25/2024).
[5] Privacy, security and social implications of data mining. [Electronic resource] Access mode: https: //www.geeksforgeeks. org/конфиденциальность, security and social implications of data mining/. – (accessed 25.01.2024).
[6] The growing role of privacy technology in the future of data security. [Electronic resource] Access mode: https: //www.immuta. com/blog/privacy - engineerings - emerging - role /. – (accessed 25.01.2024).
[7] Beckmann J.30 Impressive Big Data Statistics for 2023 // techreport.2023. [Electronic resource] Access mode: https: //techreport. com/statistics/big - data - statistics/
[8] Shemale R. Big Data Statistics For 2024 (Growth, Market Size & More) // demandsage.2024. [Electronic resource] Access mode: https: //www.demandsage. com/big - data - statistics/