Effective Anonymization of Sensitive Data in the Large-Scale Systems Using Privacy Enhancing Technology

MSc Research Project
Cloud Computing

Tarini Beeruka
Student ID: 21117314

School of Computing
National College of Ireland

Supervisor: Dr. Punit Gupta

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Tarini Beeruka
Student ID:	21117314
Programme:	Cloud Computing
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Punit Gupta
Submission Due Date:	01/02/2023
Project Title:	Effective Anonymization of Sensitive Data in the Large-Scale Systems Using Privacy Enhancing Technology
Word Count:	5327
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Effective Anonymization of Sensitive Data in the Large-Scale Systems Using Privacy Enhancing Technology

Tarini Beeruka
21117314

Abstract

Data Anonymization is the process of securing confidential or sensitive information by hiding or encrypting identifiers that connect a specific person to stored data. Various privacy-preserving strategies, methodologies, frameworks, and prototypes have been proposed or developed for disclosing data while preserving user privacy. Data owners, such as hospitals, financial institutions, and social networking sites, analyze data and conduct business operations on the dataset after applying anonymization techniques to protect users' privacy. In the analysis of the data, not all columns are utilized. When analyzing data, it is discovered that sensitive columns are also included, which is sometimes unnecessary because it not only consumes more computational resources but also exposes existing personal information in the dataset to risk. Data Anonymization techniques can overcome this issue, and further encryption can be applied. The author developed an optimal anonymization tool that fetches the dataset, splitting the data based on sensitive and personal attributes, only sending essential columns for computation while adding encryption for the files for added security. For greater performance and security, the Amazon S3 bucket is used for storing and retrieving data and CryptPandas has been used for the encryption and decryption of pandas data frames. The author compares the datasets after using a data anonymization tool and the raw dataset. Execution time and memory consumption are the parameters considered in this study, the well-being of patient is shown as output after performing computation of datset. According to the results, the dataset showed a 45.5 % decrease in execution time and a 33.3% decrease in memory consumption after utilizing the data anonymization tool.

Keywords— Amazon S3, CryptPandas, Data analysis, Data anonymization, Sensitive data

Table of Content

1 Introduction

Due to the increase in information digitization, very vast amounts of structured, semi-structured, and unstructured data are being produced quickly. By acquiring, organizing, sorting, analyzing, and mining these data, an organization can gain access to a large amount of sensitive user data. When the data are stored, they serve the purposes of the company as well as provide services to other businesses. In conventional cloud storage, plain text or encrypted data are continuously kept. These data can be considered ”dead” because they are not used during computation. The term ”personal data” only refers to pertinent details or required to locate a real person's data file. When aggregated, various pieces of information might be able to identify a person. Also included is private data about a particular person. The mere fact that data was anonymized through the use of privacy-enhancing methods does not, however, indicate that it is no longer connected to a specific individual and does not constitute personal information. Data that has been anonymized might still be linked to a specific person and be regarded as personal information.

Furthermore, any personal data that has been rendered anonymous in a manner that makes it difficult or impossible to pinpoint the person. Data that falls under a particular category or is sensitive must be handled accordingly. Personal information that falls under a special category requires a higher level of security since it is delicate. Between sensitive and non-sensitive personal data, GDPR clearly distinguishes. Special groups defined by Article 9 of the GDPR demand further care. Any data that reveals a subject's information is considered sensitive data, or special category data, in accordance with GDPR. Examples of sensitive data include racial or ethnic origin, political and religious convictions, genetic information, mental and sexual health, and sexual orientation(Gal and Aviv; 2020)

Data anonymization is one strategy companies can use to abide by strict data privacy rules that demand the protection of personally identifiable information, such as patient records, phone numbers, and financial details. Figure 1 describes the flow of performing Data anonymization. Even after the identifier's data has been removed, attackers can use de-anonymization techniques to trace the anonymization process. As larger datasets are sent for computation, processing times and memory requirements increase. However, the research study demonstrates that not all columns are equally helpful when carrying out specific computations. Because they include sensitive and personal information, adding unnecessary columns like name, religion, and email address slows down the procedure and raises the risk. To solve this problem, the author developed a tool for data anonymization that divides the data into two groups: datasets with sensitive data and computational data. By doing this, there are fewer rows to evaluate when the dataset is transmitted for processing, and by encrypting it, sensitive datasets are shown to be secure.

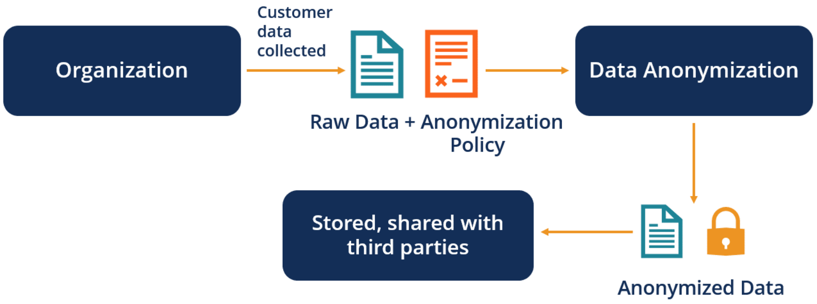


Figure 1: Flow of Performing Data Anonymization (Corporate Finance Institute; 2022)

1.1 Research Question

As the existing anonymization techniques are linear, how can we determine the highest level of privacy achievable by maximizing permitted data loss when dealing with large-scale data?

For the thesis to achieve its objectives, it is necessary to answer the above research question. Aiming to preserve privacy while obtaining precise results, the above-mentioned question provides useful columns it prioritizes while dealing with large volumes of data.

In the author's opinion, the Data Anonymization tool can be practical when analyzing data using cloud services in the following ways:

- By encrypting and delivering only the necessary columns for computation, the developers can prioritize user privacy while minimizing overall load.
- The Data Anonymization Tool allows researchers and developers to evaluate memory consumption and execution times while conducting analysis.

1.2 Ethics Consideration

Table 1 provides information regarding the ethics declaration. The dataset used for the research uses Medical Cost Personal Datasets, from the website Kaggle. The information about the copyright issues, Terms of use, and Privacy policy of the company have been mentioned following:

Medical Cost Personal Datasets

Copyright Dispute Policy

Kaggle Terms of Use

Kaggle Privacy Policy

Table 1: Table of Ethics Consideration Declaration

This project involves human participants	Yes / ***No***
The project makes use of secondary dataset(s) created by the researcher	Yes / ***No***
The project makes use of public secondary dataset(s)	***Yes*** / No
The project makes use of non-public secondary dataset(s)	Yes / ***No***
Approval letter from non-public secondary dataset(s) owner received	Yes / ***No***

1.3 Paper Structure

Following is an overview of the paper's content.

- ***Section 2*** covers GDPR, legalization and identification, along with details on how sensitive data can be eliminated during analysis.
- ***Section 3*** follows with the process flow of the data anonymization tool.
- ***Section 4*** deals with design aspects of the application through data flow diagram with performance indicators.

- ***Section 5*** depicts implementation of tool, dataset used and calculations performed along with output results.
- ***Section 6*** has all the bar graphs and output displays along with discussion.
- ***Section 7*** states the conclusion and future possible work in the field of data anonymization through cloud services.

2 Literature Survey

An understanding of Data Anonymization and the importance of eliminating sensitive data is essential for researching and evaluating performance. The following section overviews a brief description of existing research projects. A significant part of this section explains what is needed to conduct this project and helpful information from trustworthy sources to support the author's research.

2.1 Legalization and GDPR

A GDPR regulation from the EU specifies how companies and other organizations must use personal data while preserving its integrity. Personal data includes any information that could be used to identify a living person directly or indirectly. Personal information frequently includes names, phone numbers, and addresses. Interests, past purchases, health information, and online conduct are all examples of things that are recognized as personal information since they can be used to identify an individual. Instead of introducing security and privacy as a future addition, businesses should protect the information of their clients who are fundamentally connected to the service they offer. GDPR Ideas and Regulations: When a person can be identified directly or indirectly, such as by name or identifying number, that individual is said to be identifiable. Electronically generated identifiers, geographical information, and a thorough account of a person's anatomical, biological, psychological, socioeconomic, religious, or social identity. Suppliers or customers, business emails or cell numbers, visit cards, images of staff in internal physical places, and IP connected with each collaborator are a few examples of internal data that can be mentioned (Voigt and Von dem Bussche; 2017).

2.1.1 Personal and Sensitive Data

Any piece of information that can be used to identify a person is considered personal data. Name, phone number, residence, age, email address, etc. can be included. Even information that categorizes the presence of someone might be considered personal data. Differentiating between sensitive personal data and personal data in some circumstances may be difficult. Names and initials in relation to locations and dates of birth are examples of personal data but instead of sensitive personal data. But since surnames are frequently connected to a particular nationality or culture, or even to both, more personal details like religious beliefs and ethnicity may be inferred from these traits. This does not imply that in order to preserve their identities on client databases, you would have to follow the guidelines for handling sensitive personal data. But since particular surnames are frequently connected to a specific religion or ethnicity, or even to both, more private information like ethnicity and religion may be inferred from these traits. This does not necessarily mean that to preserve their names on client databases, you would have to follow the guidelines for handling sensitive personal data. (Voigt and Von dem Bussche; 2017). However, this would be regarded as the calculation of sensitive personal data if the data processor is utilizing these names for a specific purpose, such as to send marketing or promotional items for goods or services that are only meant for individuals of a specific religion or race.

2.1.2 Identifying and Eliminating Personal and Sensitive Data

An organization should delete sensitive data from its database for a number of reasons. Protecting sensitive information is important not only because it is required by law in the sector, but also out of consideration for the customers and other private individuals. The top four reasons why a company should delete sensitive information are as follows: Ensure compliance, reduce security risks, fulfill legal duties, and obtain insurance. Data breaches pose serious hazards that shouldn't be taken lightly. To make sure to have strong protection in place to avert significant harm for the firm and everyone involved. Sensitive information must be taken out of documents in order to accomplish this. Additionally, the suppression of information can guarantee that you abide by the GDPR regulations of the nation or business. If the business requires 100% accuracy, using human-assisted automation might be a good idea. The database does not automatically save data; instead, it first goes through a human review process. This method can reduce errors caused by subpar image or documentation quality while increasing accuracy. This solution boosts your company's productivity and efficiency by combining the best features of human and artificial intelligence (Tirza; 2022).

2.1.3 Preserving Sensitive Data on Cloud

It is becoming more and more imperative to use the cloud for processing data on cloud premises due to the growing amount of sensitive and personal data that data controllers are collecting. Sensitive data should not, however, be outsourced to public clouds without protection due to security issues with regular data breaches and recently updated legislative data protection laws such as the General Data Privacy Regulation of the European Union. Technologies that enable the methodical outsourcing of sensitive data processing and storage to the cloud. Due to the confluence of two events, this topic is particularly researched right now. First, the sheer amount of sensitive or otherwise personal information being gathered makes it increasingly important to not just store but also process it on the cloud. Before sending user data to the cloud, local proxies may employ a variety of security measures(Domingo-Ferrer et al.; 2019). Numerous masking approaches have been offered by the research community because of the wide range of data protection settings. Although it may seem confusing, this diversity offers the benefit of making it feasible to accommodate a variety of privacy and functionality requirements.

2.2 Importance of Data Anonymization in Medical Field

An important step in guaranteeing patient privacy and safeguarding sensitive information is the anonymization of medical data. Personal identifiers from medical records, such as names, addresses, and social security numbers, are removed or obscured. This preserves patient privacy while enabling the exchange and analysis of medical data for research and other uses.

De-identification, pseudonymization, and aggregation are just a few of the methods utilized to anonymize medical data. De-identification is the process of stripping the data of any identifying information, rendering it difficult to connect the data to a specific person. The data may become less helpful for research if this is the most secure way of anonymization. Pseudonymization is the process of swapping out personally identifying information for a pseudonym. If necessary, this can be used to trace the data back to a specific person, but it is more challenging to do so. Although less secure than de-identification, this method provides more relevant data for scientific inquiry. Aggregation is the practice of combining data in ways that make it difficult to identify specific patients, such as by geography or age. Although less secure than de-identification or pseudonymization, this approach has its uses, such as when researching a particular demographic. Samarati and Sweeney (1998)

Anonymization of medical data is important for several reasons. It allows for the sharing and analysis of medical data for research and other purposes while maintaining patient privacy. It also enables data to be used for secondary research, improving the quality of healthcare while minimizing the risk of data breaches. Additionally, it allows for the development of new treatments and therapies, as well as the identification of healthcare trends and patterns.

Overall, anonymization of medical data is a crucial step in ensuring patient privacy and protecting sensitive information while still allowing for the sharing and analysis of data for research and other purposes. The choice of the method of anonymization will depend on the purpose of the research, the data and the level of risk of identification of individuals. Vovk et al. (2021)

2.3 Data Anonymization

Data anonymization is the practice of protecting sensitive information by erasing or encrypting identifiers that link specific people to the data that is being stored. Policies for data anonymization make sure that a business is aware of and upholds its responsibility to secure sensitive, private, and confidential data. The capacity to derive confidential details from the findings would be constrained by the collection of anonymous data and the removal of individuals from the database. The General Data Privacy Regulation does not apply to fully ”anonymized” data because it does not comply with the requirements for private data and therefore is not liable to the same processing limitations (GDPR). When people cannot be identified in data, it is said to be ”anonymized.” It is crucial to remember that someone can be recognized without being identified. An individual may still ”be identified” if there is additional information that makes it possible for them to connect to data concerning them that cannot be about another member of the group. In this situation, it’s crucial to think about what ”identifiers” the information possessed has. (Mahesh and Meyyappan; 2013)

2.3.1 Concepts of Data Anonymizations

The following are anonymization methods are Randomization, Using pseudonyms and Re-identification Risk.

Randomization: First, notice is added. It involves a few minor dates and number corrections. The next action is to shuffle. It entails randomly permuting the values of an attribute within a single database. K-anonymity: It requires altering the scale of orders of magnitude. No key individuals, also known as tuples, are ever represented by the exact same ranges of quasi-identifiers thanks to k-anonymity. By grouping crucial information into categories, a range of values, or comparable groups Each equivalence class in the L-diversity methodology must have an adequate number of diverse traits. The K-anonymity approach has evolved into L-diversity and T-closeness.

Pseudonymization: It is typically used to replace or conceal personal data and refers to code. The following information aids in. It is regularly used to replace or conceal personal data and refers to replacements or coding. Figure 2 gives a brief description on the difference between anonymization and pseudonymization. An illustration would be the person’s name for a predetermined text for an item randomly selected from such a list of values. Contents is frequently encoded using a predefined key that now the user sets up when utilizing a symmetric cryptography method or either key for encryption; nevertheless, if the user uses a code and an intruder discovers the code, the dataset’s content will be exposed. It entails directional encoding of data with the ideal use of a salt entropy key. Character masking, which consists of substituting text characters with asterisks so that readers can distinguish them from the specified characters, might be the easiest option

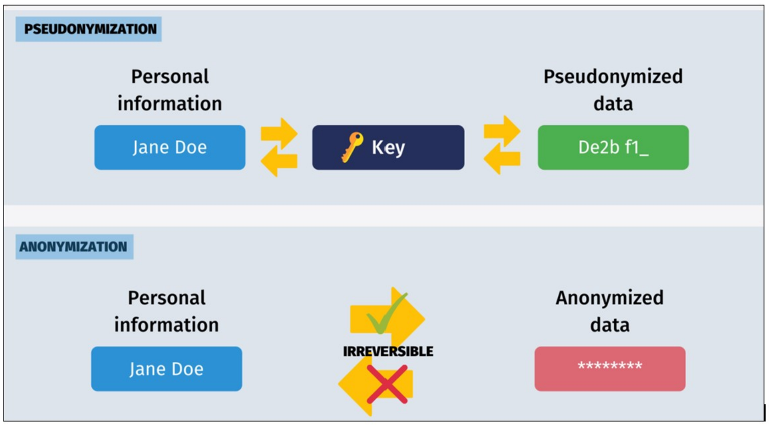


Figure 2: Anonymization vs. Pseudonymization (Admin; 2021)

Risk of Reidentification: The possibility of re-identification as a violation of privacy grows each time an assault targets a certain target. The setting within which the content is disclosed and the procedures employed for anonymization have a significant impact on this risk. It is theoretically possible to re-identify a person no matter how stringent the de-identification procedure is, even if doing so requires having access to a variety of outside information by merging it with anonymized data. It is sufficient to apply endless resources to uncover associations with data or individuals who have previously been recognized from other databases, hence no masked dataset can be totally protected unless all of the data is meaningless.

2.4 Data Anonymization Tools

Organizations working with personal information must prioritize data anonymization. Utilizing data anonymization has the goal of enhancing the integrity of data sharing. Organizations should consider the choice stated below if they want to add this level of integrity and security to their data.

2.4.1 ARX

configuration viewpoint, investigation point of view, utility assessment perception, and risk analysis perspective—are used to categorize the functionality of the ARX data anonymization tool. Data from the configuration perspective is imported, and some data transformation rules are made. It is also decided here which confidentiality and reliability models for material will be used. Data may be absorbed from a wide range of sources, and you can even make notes about the taxonomy, data types, and structure of the data. An optimal solution among the potential alterations for the supplied data is established here from the perspective of exploration. For a specific model, risk thresholds and quality management need to be examined.Utility analysis' primary objective is to determine whether a particular transformation is appropriate for a particular situation. Several mathematical techniques are used in the setting of risk analysis to generate different metrics related to privacy concerns. Models for categorization can be developed using data flow and descriptive statistics. This point of view also makes it possible to pinpoint qualities that need to be improved in order to boost safety.

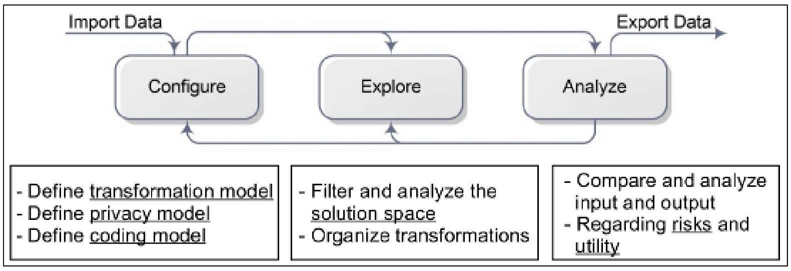

Import Data
Export Data
Configure
Explore
Analyze
- Define transformation model
- Define privacy model
- Define coding model
- Filter and analyze the solution space
- Organize transformations
- Compare and analyze input and output
- Regarding risks and utility

Figure 3: WorkFlow of Data Anonymization through ARX Tool(*Anonymization tool*; 2022)

2.4.2 Clover DX's Data Anonymization Tool

The vital production-level data is transformed into an anonymised data set by Clover DX's data anonymization tool. The quantity of production data is removed of its sensitive components but still retains the necessary information thanks to its anonymization operation. The tool's comprehensive approach to data anonymization results in a set of data that is extremely trustworthy, among other major advantages. Strong privacy measures include introducing random data and allowing users to select how anonymous the output should be compared to the source data. Here, it takes the least amount of time possible to implement an anonymization method. Once set up, the Autonomous Privacy-preserving Engine can produce anonymized data as needed (Sánchez et al.; 2020).

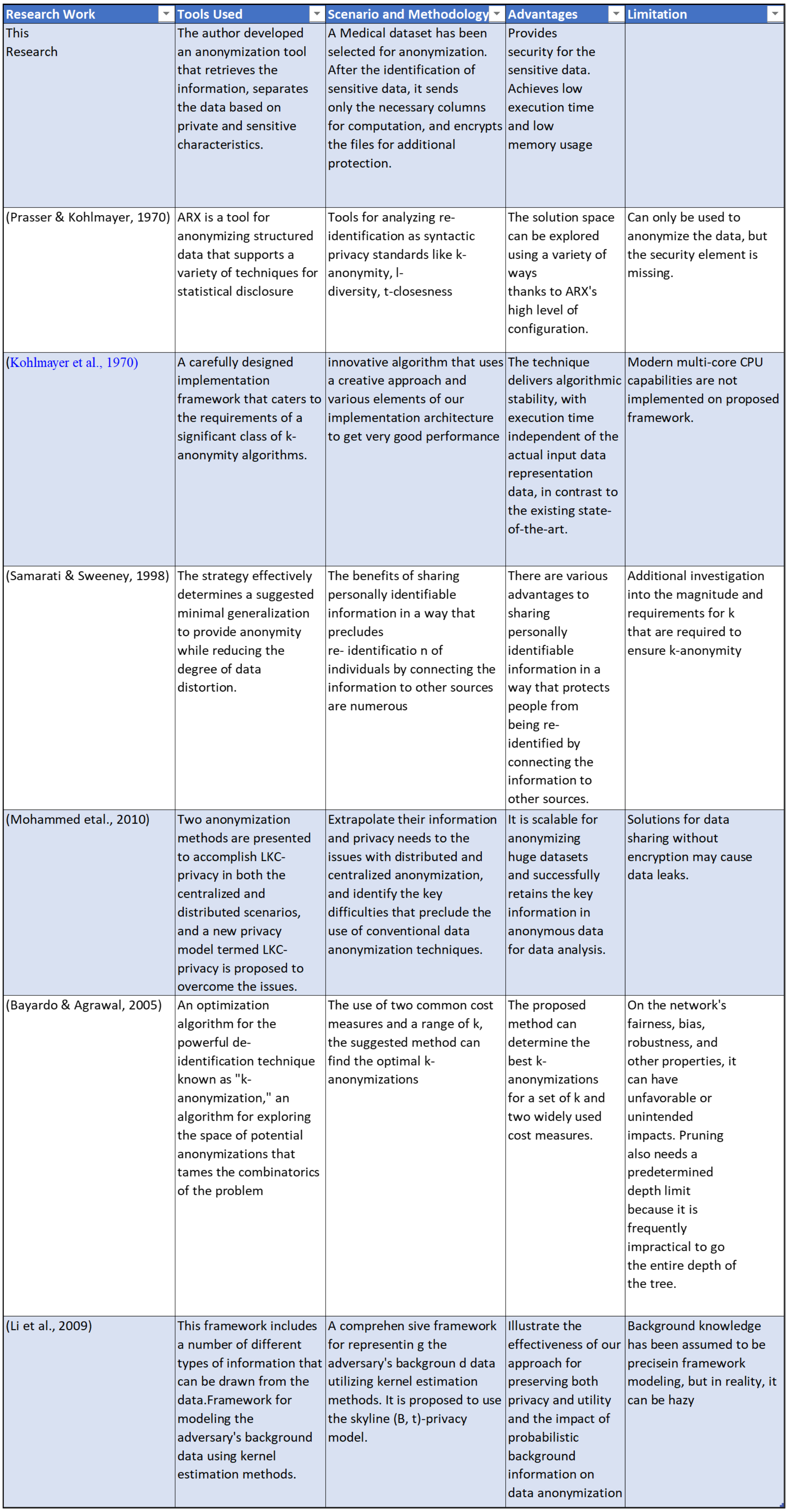
2.4.3 BizDataX

The anonymization tool from BizDataX is skilled at processing a variety of data formats. However, if a user thinks of a particularly specific circumstance, the user can use a straightforward extension where you can add your own logic. The three fundamental stages are: removing data from the memory of the BizDataX application; processing the data to make it anonymous; and publishing the database with the data. They typically write roughly 1 billion records every hour of data. The integrity of the main database frequently suffers when data masking activities are taking place. BizDataX keeps an eye out for any potential problems(Prasser et al.; 2020).

2.5 Comparative Study of Research Study

For the purpose of conducting research and assessing performance, it is crucial to comprehend Data Anonymization and the significance of removing sensitive data. The section-2 in the report provides a quick overview of ongoing research initiatives on Data Anonymization, Sensitive Data, tools and methodologies to perform anonymization of data. This section devotes a sizable portion to outlining the requirements for carrying out the project and providing valuable data from reliable sources to back up the author's study.

The following figure 4 gives the brief summary while performing a comparative study of various research studies on the bases of tools used, scenario, advantages, and limitations.



Research Work	Tools Used	Scenario and Methodology	Advantages	Limitation
This Research	The author developed an anonymization tool that retrieves the information, separates the data based on private and sensitive characteristics.	A Medical dataset has been selected for anonymization. After the identification of sensitive data, it sends only the necessary columns for computation, and encrypts the files for additional protection.	Provides security for the sensitive data. Achieves low execution time and low memory usage	
(Prasser & Kohlmayer, 1970)	ARX is a tool for anonymizing structured data that supports a variety of techniques for statistical disclosure	Tools for analyzing re-identification as syntactic privacy standards like k-anonymity, l-diversity, t-closesness	The solution space can be explored using a variety of ways thanks to ARX's high level of configuration.	Can only be used to anonymize the data, but the security element is missing.
(Kohlmayer et al., 1970)	A carefully designed implementation framework that caters to the requirements of a significant class of k-anonymity algorithms.	innovative algorithm that uses a creative approach and various elements of our implementation architecture to get very good performance	The technique delivers algorithmic stability, with execution time independent of the actual input data representation data, in contrast to the existing state-of-the-art.	Modern multi-core CPU capabilities are not implemented on proposed framework.
(Samarati & Sweeney, 1998)	The strategy effectively determines a suggested minimal generalization to provide anonymity while reducing the degree of data distortion.	The benefits of sharing personally identifiable information in a way that precludes re- identificatio n of individuals by connecting the information to other sources are numerous	There are various advantages to sharing personally identifiable information in a way that protects people from being re-identified by connecting the information to other sources.	Additional investigation into the magnitude and requirements for k that are required to ensure k-anonymity
(Mohammed etal., 2010)	Two anonymization methods are presented to accomplish LKC-privacy in both the centralized and distributed scenarios, and a new privacy model termed LKC-privacy is proposed to overcome the issues.	Extrapolate their information and privacy needs to the issues with distributed and centralized anonymization, and identify the key difficulties that preclude the use of conventional data anonymization techniques.	It is scalable for anonymizing huge datasets and successfully retains the key information in anonymous data for data analysis.	Solutions for data sharing without encryption may cause data leaks.
(Bayardo & Agrawal, 2005)	An optimization algorithm for the powerful de-identification technique known as "k-anonymization," an algorithm for exploring the space of potential anonymizations that tames the combinatorics of the problem	The use of two common cost measures and a range of k, the suggested method can find the optimal k-anonymizations	The proposed method can determine the best k-anonymizations for a set of k and two widely used cost measures.	On the network's fairness, bias, robustness, and other properties, it can have unfavorable or unintended impacts. Pruning also needs a predetermined depth limit because it is frequently impractical to go the entire depth of the tree.
(Li et al., 2009)	This framework includes a number of different types of information that can be drawn from the data.Framework for modeling the adversary's background data using kernel estimation methods.	A comprehen sive framework for representin g the adversary's backgroun d data utilizing kernel estimation methods. It is proposed to use the skyline (B, t)-privacy model.	Illustrate the effectiveness of our approach for preserving both privacy and utility and the impact of probabilistic background information on data anonymization	Background knowledge has been assumed to be precisein framework modeling, but in reality, it can be hazy

Figure 4: comparative study of various research studiesBayardo and Agrawal (2005) Prasser and Kohlmayer (2015) Mohammed et al. (2010) Li et al. (2009) Kohlmayer et al. (2012)

3 Research Methodology

Correct and logical methodologies are needed in developing an application. With a structured plan, applications can be run easily, including configuration and customization. This section explains how to keep private data from being accessed by just anyone, as well as improve the performance of the application itself.

3.1 Process Overflow

Figure 4 shows the flow of the Data Anonymization Tool. It is necessary to follow the right procedures, by which, if the proper method is adopted, the results will be more accurate and the customizing process will be easier to handle. This section provides a general overview of the data anonymization Tool used to retrieve the data, encrypt the files, and show the appropriate outcome.

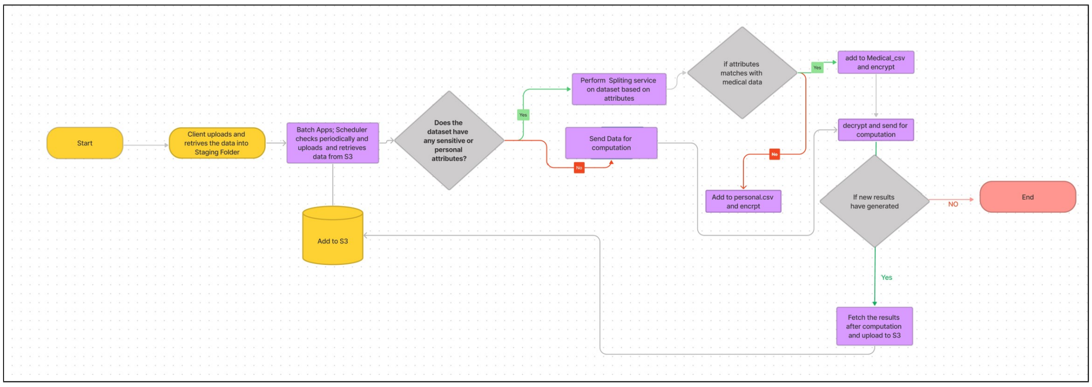

Start
Client uploads and retrives the data into Staging Folder
Batch Apps; Scheduler checks periodically and uploads and retrieves data from S3
Add to S3
Does the dataset have any sensitive or personal attributes?
Yes
No
Perform Spliting service on dataset based on attributes
Send Data for computation
if attributes matches with medical data
Yes
No
add to Medical_csv and encrypt
Add to personal.csv and encrpt
decrypt and send for computation
If new results have generated
NO
End
Yes
Fetch the results after computation and upload to S3

Figure 5: Process Flowchart for Performing Data Anonymization Tool

The Staging folder and scheduler are used to send the dataset from the client to the Amazon S3 bucket. The service next determines whether any of the submitted dataset's properties qualify as sensitive or personal data. The author has taken into account a medical dataset in this study. If the dataset contains sensitive information, it will be sent to an additional service that separates the data into medical and personal datasets and encrypts them. Only the medical data is transmitted for computation when the data has been decrypted. To increase the security of the files, two separate encryption techniques have been utilized to encrypt them. If the uploaded data does not contain any characteristics deemed to be sensitive or personal data, it will be transmitted straight to the computation. Through crypt Pandas, the files are encrypted. The tool is quite helpful when there are many columns in the dataset that won't be used for computation since it allows them to be segmented and only the necessary attributes supplied for analysis while also securely encrypting personal information files (Sedayao and Enterprise Architect; 2012).

Based on the parameters "BMI" and "Smoker," the study appraised the patient's health. Execution time and memory utilization of the raw dataset and the dataset after using the Data Anonymization tool are compared extensively. The user can retrieve the evaluated results from S3 Bucket after the results have been obtained.When the dataset is assessed and the results are uploaded to S3, the Tool process is complete.

4 Design Specification

Designing an anonymization tool requires a thorough understanding of sensitive attribute identification and best encryption approaches. The high-level method for evaluating a medical dataset utilizing a Data Anonymization Tool is succinctly described by the author. After the results are obtained, the performance indicators are explained in this section.

4.1 Diagram

An overview of the application process is shown in the section below. Usually, diagrams like this are for general use, so everyone can understand them. These diagrams are also utilised in certain businesses to ease communication between technical and commercial divisions.

4.1.1 Flowchart Diagram

Figure 7 illustrates the workflow of evaluation of medical dataset through Data Anonymization Tool from fetching the data form user to uploading the results back to S3.

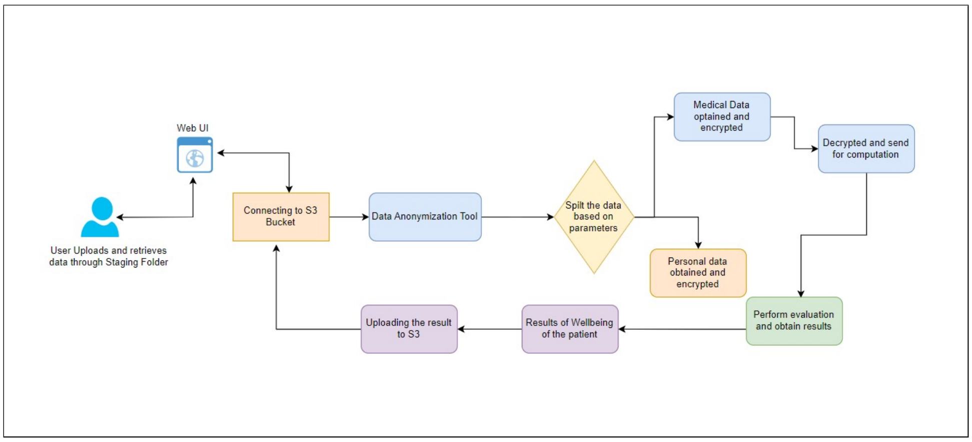

Web UI
User Uploads and retrieves data through Staging Folder
Connecting to S3 Bucket
Data Anonymization Tool
Spilt the data based on parameters
Medical Data optained and encrypted
Decrypted and send for computation
Personal data obtained and encrypted
Perform evaluation and obtain results
Results of Wellbeing of the patient
Uploading the result to S3

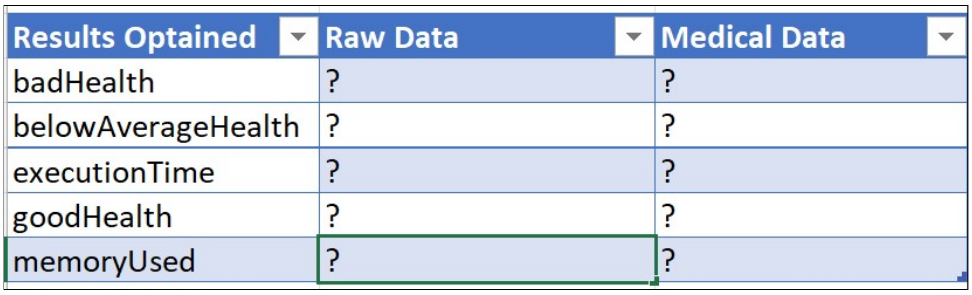
Figure 6: Evaluation of Medical Dataset through Data Anonymization Tool Flow

A ”staging area in the IT industry is a location where the assemble all of the tools and supplies so they are prepared to be used on the job. A staging folder is a folder where it is put together all the items and folders required for a particular project so it’s available to compete, email, or burn to CD in the presenting industry. The folder containing the PowerPoint file is treated as a staging folder when you run FixLinks. All linked files are duplicated and placed in the same location as the PPT file. A dataset in CSV format is uploaded for this study.The term ”REST” refers to a form of software architecture that establishes a standard for client-server communication over a network (Van Rossum and Drake Jr; 2002) REST offers a set of restrictions for software architecture that support the system’s performance, scalability, durability, and simplicity. A cloud - based storage resource in Amazon Web Services, Simple Storage Service, an object storage service, is called an Amazon S3 bucket. The objects that are stored in Amazon S3 , which resemble file folders, are made up of data and the metadata that describes it. Anonymization technology was developed to deal with the growing volume of sensitive data that organizations use and store. Modern anonymization techniques, a branch of Natural Language Processing, use dictionaries and techniques to precisely identify every term that could be interpreted as personal data. Anonymization results in non-identifiable datasets that are no longer considered to be personal info and can therefore be used and shared without additional legal authorization (Chan et al.; 2019).

Data splitting is a protective strategy based on the fragmentation of sensitive data and storage of the bits in clear form in several places. The arrangement of the fragments should be such that no one fragment permits re-identification of the person to whom it corresponds or divulges sensitive information that can be connected to a specific subject. A list of diagnostics, for instance, is actually useless to an invader if a fragment only contains the values of the property ”Diagnosis” since he cannot connect the diagnoses to the persons to which they apply. In the case of the cloud, data can be outsourced via a local proxy conducting data splitting to either numerous cloud identities within the same CSP or to other clouds, each of which is administered by a different CSP and offers the same kind of cloud services. Regardless of their content, CryptPandas enables you to quickly encrypt and decrypt pandas dataframe. The results are ultimately uploaded to S3 following the computation-based evaluation of the dataset.

4.2 Performance Indicators

Two significant results are found during evaluation for the performance measures. The patient’s health comes first, followed by memory requirements and the time needed to execute the findings. The comparison is based on the assessment of raw data while accounting for the patient’s well-being, obtaining its memory consumption and execution time, and using medical data that has already been anonymized while also obtaining the same comparison-based metrics.



Results Optained	Raw Data	Medical Data
badHealth	?	?
belowAverageHealth	?	?
executionTime	?	?
goodHealth	?	?
memoryUsed	?	?

Figure 7: Evaluation of Health with Memory Usage and Execution Time

Based on the variables BMI and Smoker, the health of the patient is determined. the memory utilization is estimated, although Flask will correctly update the content-length if response.set data is used. Please take note of the encoding and the use of byte strings as well. And the memory profiler library is used to calculate the memory utilization. Python typically takes a different route when handling requests. Asynchronous and synchronous request processing are the two main models. However, when it comes to measuring memory utilization per request, they both face the same challenge. One Python worker processes a lot of requests (concurrently or sequentially) during the course of his lifetime, which is why. So it’s challenging to determine a request’s precise memory use. The author calculates the increase in percentage between two sets of raw information and the anonymized data using the original formula. The results are then examined based on each parameter to comprehend how the tool should be used. Further more, the parameters are individually compared for greater comprehension.

5 Implementation

This section explains how the patient’s health is calculated and how the proposed design will be implemented. In addition to explaining, this part offers a description of the application’s high-level architecture and data flow. The application’s source code is available here.

5.1 Dataset

To conduct the evaluation, the author used a public dataset from the Kaggle website. Figure 8 tells us about the field names of the different datasets after performing splitting, The collection includes medical field IDs made up of both personal and medical information.

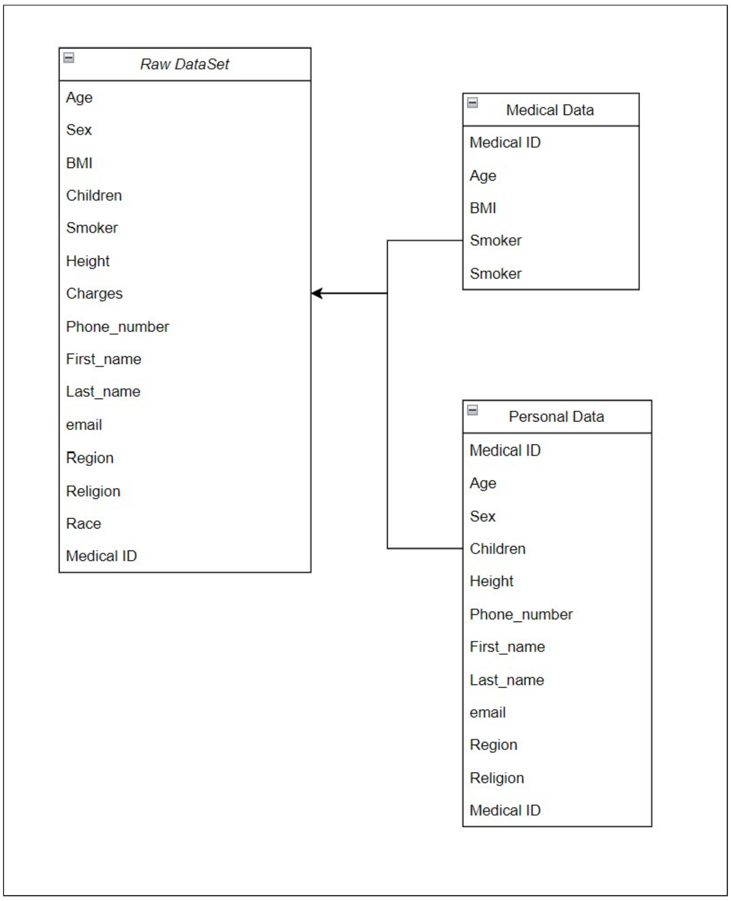


Figure 8: Class Diagram of Raw, Medical and Personal Dataset

Even after the identifier's data has been deleted, attackers can use de-anonymization techniques to follow the anonymization process. As more datasets are provided for calculation, processing times and memory needs increase. The results of the study, however, demonstrate that not all columns are equally helpful when doing specific computations. Because they include sensitive information, adding new columns like name, religion, and email address slows down the procedure and raises the risk. To solve this problem, the author developed a data anonymization tool and implemented encryption on split data.

To perform the encryption and decryption of the files Crypt Pandas has been used. Regardless of their content, CryptPandas enables to quickly encrypt and decrypt pandas dataframe. The following code explains the execution of encryption and decryption.

5.2 Architecture Overview

Figure 9 displays the high-level of evaluation of data using Data Anonymisation Tool. Companies can adhere to stringent data privacy regulations that mandate the protection of personally identifiable information, such as patient records, phone numbers, and financial details, by using data anonymization as one tactic. The process of performing data anonymization is shown in

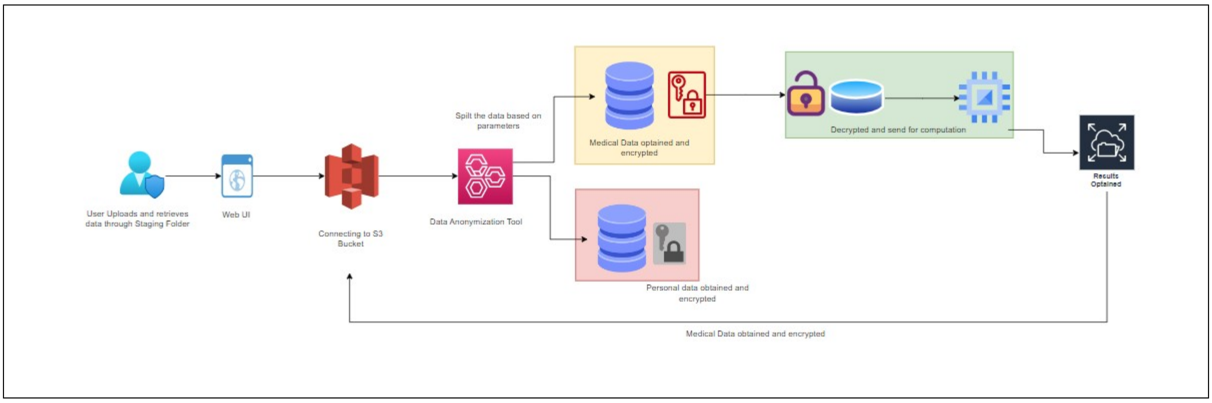

Spilt the data based on parameters
User Uploads and retrieves data through Staging Folder
Web UI
Connecting to S3 Bucket
Data Anonymization Tool
Medical Data optained and encrypted
Personal data obtained and encrypted
Decrypted and send for computation
Results Optained
Medical Data obtained and encrypted

Figure 9: High-Level Architecture of Evaluation of Data Anonymization Process

Figure 8. Attackers can employ de-anonymization methods to track the anonymization procedure even after the identifier's data has been erased. Processing times and memory requirements grow as more datasets are sent for calculation. However, the study's findings show that not all columns are equally useful when doing particular computations. Including extra columns like name, religion, and email address slows down the process and increases risk because they contain sensitive information. The author created a tool for data anonymization to address this issue. The result is displayed as JSON format on Postman (Zhang and Zhang; 2017)

The following code helps in understanding the encryption and decryption process:

```
#Personal data encryption using CryptPandas
if output_key_personal_encrypt:

crp.to_encrypted(df_personal_record, password=local_key_personal_encrypt_pwd,
    ↪ path=local_key_personal_encrypt)

s3_resource.Bucket(bucket_name).upload_file(local_key_personal_encrypt,
    ↪ output_key_personal_encrypt)

#Performing decryption using CryptPandas @app.route('/data/calculate', methods
    ↪ =['POST']) def calculateData():
bucket_name = request.json['bucket_name'] bucket_key = request.json['
    ↪ bucket_key'] is_encrypt = request.json['is_encrypt']

local_data_decrypt_crypt = 'decrypted_input/data.crypt' good_health_count = 0
below_average_health_count = 0

bad_health_count = 0 conn = client('s3')
s3_resource = boto3.resource('s3')
```

5.3 Performing Calculation

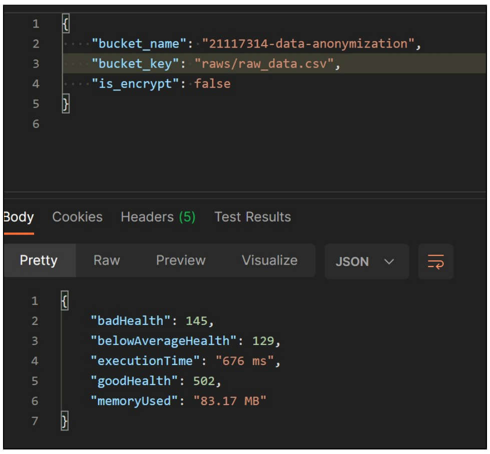
The following are the metrics, the author has considered to perform evaluation of health of patient: Good health, Bad health, and below average health. It is calculated based on two parameters, BMI and smoking.

```
#Calculation formula
```

```
good_health_count = df.query("bmi < 30 & smoker == 'no'").shape[0]
below_average_health_count = df.query("bmi < 30 & smoker == 'yes'").shape[0]
bad_health_count = df.query("bmi >= 30 & smoker == 'yes'").shape[0]
```

5.4 Results Obtaining through Postman

Figure 10 gives the output of the raw data, which has performed evaluation without data anonymization tool and the following results are obtained through Postman.



```
{
    "bucket_name": "21117314-data-anonymization",
    "bucket_key": "raws/raw_data.csv",
    "is_encrypt": false
}
```

```
{
    "badHealth": 145,
    "belowAverageHealth": 129,
    "executionTime": "676 ms",
    "goodHealth": 502,
    "memoryUsed": "83.17 MB"
}
```

Figure 10: Results Displaying through Postman

6 Evaluation

This section is filled with three summary bar charts for each category with health of the patient, Execution and Memory usage in Performance Indicators. Subsequently, the author discusses the finding of the results of the tool usage.

6.1 Comparison of Execution Time

Figure 11, When the raw data is directly executed the result has been 1055ms and the anaonymized data show 355ms. It clearly says, as larger data set is being processed the exeution time is more. There fore it is suggestable to use data anonymization tool.

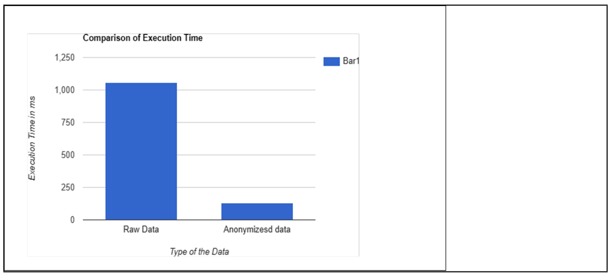


Figure 11: Comparison of the Execution Time in ms

6.2 Comparison of Memory Usage

The comparison of memory usage of the raw and anonymized data bar chart illustrated in Figure 12 The clearly shows that memory usage of raw data was more than anonymized data.

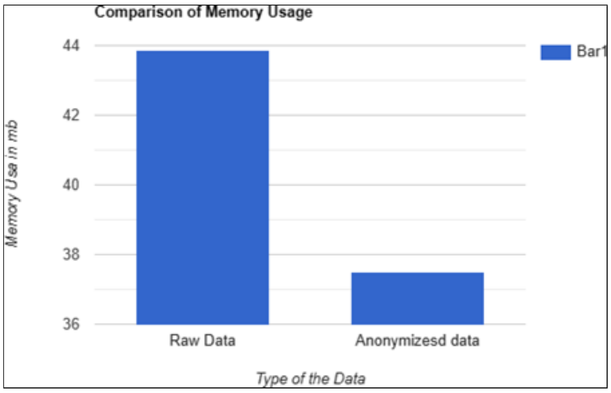


Figure 12: Comparison of Memory Usage Table in MB

6.3 Comparison of Patient Health with Different Cases of the Raw and Anonymized Data

Comparison of memory usage of the raw and anonymized data are shown in the bar chart in Figure 13. This is the prime example that, results have been accurate which tells us that it is no need to have personal data items in the dataset, and it is better to eliminate if the evaluation does include personal data.

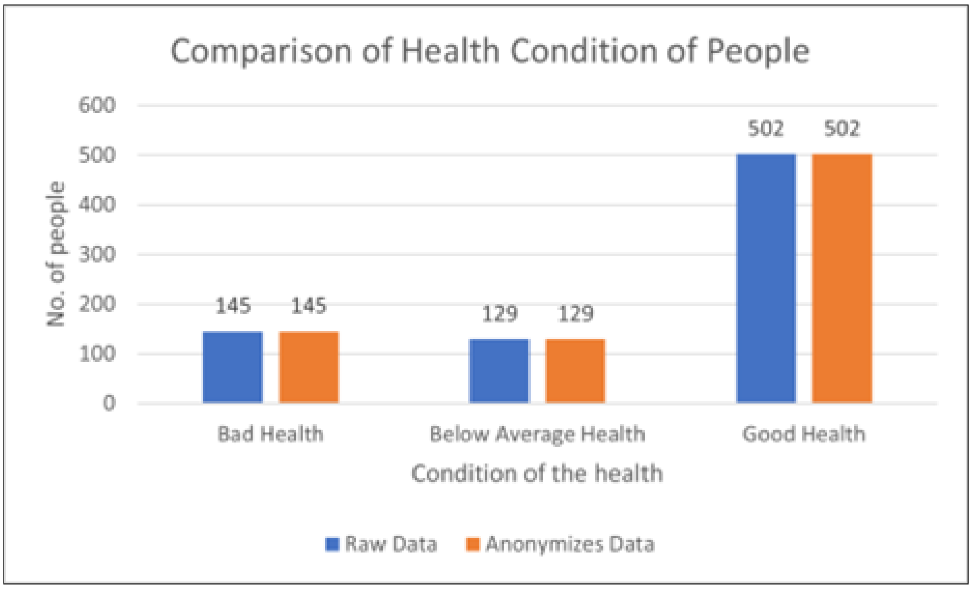
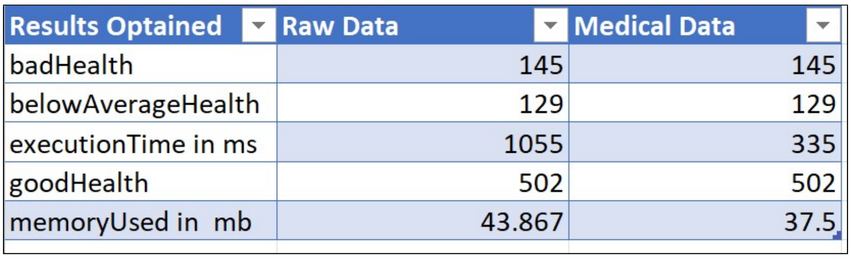


Figure 13: Result Summary on comparing the results obtained by Raw and Anonymized Data

6.4 Discussion

Each time a line is executed, the Python interpreter uses memory. Increment: The memory usage difference between the current line and the previous line. It basically indicates how much memory a specific line of Python program uses. Occurrences: the amount of times a specific line of code is run. The time module can be used to determine how long it took to run a programme. Using time(), save the time at the start of the code. At the conclusion of the code, save the timestamp. The time for execution is determined by the distance between the beginning and the end. Here while performing the health of the patient, author has taken cases; good health, bad heath and below average health. Figure 13 gives us the reult of both data as well as the execution and memory usage.



Results Optained	Raw Data	Medical Data
badHealth	145	145
belowAverageHealth	129	129
executionTime in ms	1055	335
goodHealth	502	502
memoryUsed in mb	43.867	37.5

Figure 14: Values of Performance of Raw and Anonymized Data

Memory Usage We all are aware that large dateset take longer time to execute. But here we have a chance to eliminate the data items which are not use to lesser the memory usage and get faster results. As Figure 14 gives us the result that raw data makes use of more data than the anonymised data.

Execution Time There is much large difference between execution time when compared between raw and anonymized data.To is again suggestable to split and perform the analysis there by always providing the encryption to the files.

Health of the patients There has been same results while evaluating from Figure 14. So there has not been any compramise with fairness and accuracy of the result. As we know that perfomace also plays an importanct role in utilization of the tool. In over all, it has shown that usage of tool shows greater result in this case. As to answer the research question, highest level of privacy can be achievable without compramising the accuracy of the result by provididng high security and better computational logic.

7 Conclusion

The task of hidden relevance is data anonymization. It requires not only enough work to be successfully implemented, but also in-depth research and comprehension of the material and the numerous impersonation approaches.The danger of reidentification is considered to be low. There is always a potential that the assailant knows more about the problem than we do, despite our best efforts to make it anonymous. Additionally, there might be a viable option that would allow for the maintenance of the data set's existing structure. if it is not necessary to maintain the accuracy of that data. Additionally, the outcomes of the process won't always be entirely satisfactory, thus GDP-permitted processes, including security controls, should be put in place to maintain confidentiality in the handling of personal data. And the same is valid for sets of data that we stated a desire to anonymize in our business.

Consumer privacy is at danger when such information is divulged by the utilities with outside stakeholders. In this essay, we contrast a number of privacy-preserving anonymization methods that businesses might use to mask their data. Examples based on the data anonymization methods were provided by the author.For large businesses with numerous data silos, protecting data privacy is a challenge that is getting more and harder. New data rules from the EU, the US, and China show that this problem is just just getting started. This pattern emphasizes the value of anonymization, one of the most crucial techniques in the ”privacy arsenal” of a data scientist. Data anonymization is a method that can be used to safeguard sensitive information contained in the information while maintaining, to variable degrees, its usefulness. However, as we'll see, this tool works best when used in conjunction with other methods rather than on its own as a means of data protection.Since this dataset contains both relational and transactional data, a case study including patient data is also included. Most anonymization techniques now in use focus particularly on relational or transactional data. For datasets used in contemporary applications, this does not, however, provide a solution. The author may have access to a great deal of sensitive information because they work with medical data. Therefore, it is always best practice to protect the privacy of the individual and by introducing new technologies that can be helpful in the field of cloud computing that performs data analysis. In the future, the patient data set can also be made anonymous utilizing techniques that are currently available to guarantee great security. conducting the investigation in which the computational logic makes use of sensitive data.conducting the investigation in which the computational logic makes use of sensitive data. The fairness and accuracy of the ethics of before and after anonymization when using cloud services can serve as the foundation for a future investigation.

References

Admin (2021). Pseudonymization according to the GDPR [definitions and examples].
URL: *https://dataprivacymanager.net/pseudonymization-according-to-the-gdpr/*

Anonymization tool (2022).

Bayardo, R. and Agrawal, R. (2005). Data privacy through optimal k-anonymization, *21st International Conference on Data Engineering (ICDE'05)*, pp. 217–228.

Chan, J., Chung, R. and Huang, J. (2019). *Python API Development Fundamentals: Develop a full-stack web application with Python and Flask*, Packt Publishing Ltd.

Corporate Finance Institute (2022). Data Anonymization.
URL: *https://corporatefinanceinstitute.com/resources/business-intelligence/data-anonymization/*

Domingo-Ferrer, J., Farras, O., Ribes-González, J. and Sánchez, D. (2019). Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges, *Computer Communications* **140**: 38–60.

Gal, M. S. and Aviv, O. (2020). The competitive effects of the gdpr, *Journal of Competition Law & Economics* **16**(3): 349–391.

Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A. and Kuhn, K. A. (2012). Flash: efficient, stable and optimal k-anonymity, *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, IEEE, pp. 708–717.

Li, T., Li, N. and Zhang, J. (2009). Modeling and integrating background knowledge in data anonymization, *2009 IEEE 25th International Conference on Data Engineering*, IEEE, pp. 6–17.

Mahesh, R. and Meyyappan, T. (2013). Anonymization technique through record elimination to preserve privacy of published data, *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pp. 328–332.

Mohammed, N., Fung, B. C., Hung, P. C. and Lee, C.-K. (2010). Centralized and distributed anonymization for high-dimensional healthcare data, *ACM Transactions on Knowledge Discovery from Data (TKDD)* **4**(4): 1–33.

Prasser, F., Eicher, J., Spengler, H., Bild, R. and Kuhn, K. A. (2020). Flexible data anonymization using arx—current status and challenges ahead, *Software: Practice and Experience* **50**(7): 1277–1304.

Prasser, F. and Kohlmayer, F. (2015). Putting statistical disclosure control into practice: The arx data anonymization tool, *Medical Data Privacy Handbook*.

Samarati, P. and Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information, *PODS*, Vol. 98, pp. 10–1145.

Sánchez, D., Martínez, S., Domingo-Ferrer, J., Soria-Comas, J. and Batet, M. (2020). μ-ant: semantic microaggregation-based anonymization tool, *Bioinformatics* **36**(5): 1652–1653.

Sedayao, J. and Enterprise Architect, I. I. (2012). Enhancing cloud security using data anonymization, *White Paper, Intel Coporation* p. 17.

Tirza (2022). Automatically remove sensitive information.
URL: *https://www.klippa.com/en/blog/information/automatically-remove-sensitive-information/*

Van Rossum, G. and Drake Jr, F. L. (2002). Python/c api reference manual, *Python Software Foundation* .

Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr), *A Practical Guide, 1st Ed., Cham: Springer International Publishing* **10**(3152676): 10–5555.

Vovk, O., Piho, G. and Ross, P. (2021). Anonymization methods of structured health care data: a literature review, *Model and Data Engineering: 10th International Conference, MEDI 2021, Tallinn, Estonia, June 21–23, 2021, Proceedings*, Springer, pp. 175–189.

Zhang, X. and Zhang, D. (2017). Research on encryption algorithm based on python, *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC)*, IEEE, pp. 586–588.