National College of Ireland

Secure And Verifiable Cloud Based Data Sharing For Law Enforcement With Sensitive Information And Sharing

MSc Research Project

MSc Cloud Computing

Ganyashree Suvarna
Student ID:22242864

School of Computing
National College of Ireland

Supervisor: Prof Sudarshan Deshmukh

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Ganyashree Suvarna

Student ID: 22242864

Programme: MSc Cloud Computing **Year:** 2023-24

Module: Research Project (MSCCLOUD_A)

Supervisor: Sudarshan Deshmukh

Submission
Due Date: 12/08/2024

Project Title: Secure and Verifiable Cloud Based Data Sharing For Law Enforcement With Sensitive Information And Encryption

Word Count: 7450
Page Count: 21

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: Ganyashree Suvarna

Date: 12 August 2024

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):	

Secure And Verifiable Cloud Based Data Sharing For Law Enforcement With Sensitive Information And Encryption

Ganyashree Suvarna
22242864

Abstract

In the field of the law enforcement the managing and sharing of sensitive data specially criminal records containing Personally Identifiable Information (PII) which is very sensitive and is of most importance. This Secure and Verifiable Cloud-Based Data Sharing for Law Enforcement with Sensitive Information and Encryption project addresses the critical need for enhanced data security , done by developing a robust system for detecting, encrypting, and securely sharing PII within the criminal datasets. Using AWS cloud services including S3 for data storage, Lambda for event driven orchestration and SageMaker for automated processing, this project ensures that the sensitive information is protected throughout the entire lifecycle of data.

The law enforcement data encryption project employs the Fernet symmetric encryption algorithm in order to secure the PII, ensuring that data is only accessible to authorized individuals and not everyone. The system is designed in a way that it automatically triggers data processing workflows when new datasets is uploaded to the S3,thereby allowing for real-time encryption and secure storage. Docker containers are utilized within the SageMaker environment to provide the consistent runtime ensuring that the processing tasks are performed efficiently and securely.

This project majorly focuses on providing a broad and scalable solution for the secure management of criminal datasets in the cloud environments, thereby reducing the risk of breaching of data and unauthorized access to it. By facilitating secure data sharing among the law enforcement agencies and the legal organizations the system not only protects sensitive information but it also supports effective collaboration across multiple stakeholders involved.

Keywords – Cloud Security, Law Enforcement, Cryptography, Personally Identifiable Information, SageMaker, Crime Data Protection

1 Introduction

In the realm of law enforcement and criminal justice sharing of data securely is very important. The datasets handled by law enforcement agencies, such as those containing criminal records like the crimes committed often include Personally Identifiable Information (PII). This information is crucial for investigations and the collaborations but poses significant privacy risks if its not protected properly. This project addresses these challenges by implementing a system for detecting and encrypting PII within the crime datasets, thereby ensuring secure data sharing across all the various entities.

The purpose of this project is to develop the flow of the criminal datasets among the Law Enforcement Agencies and the legal organization while minimizing the threats of the data leaks and unauthorized access. Specifically the project aims to enable a scenario where a Garda station department can securely send data to a law based organization or an law based entity even when the transfer involves multiple third parties.

In order to achieve this the project employs a perfect combination of encryption techniques and cloud technologies. The detection of PII within the dataset was conducted manually, ensuring that all the relevant PII was accurately identified before the encryption process.

Once PII is identified the project utilizes the Fernet symmetric encryption algorithm to encrypt the sensitive data. The encryption process converts the readable data into a secure, unreadable format that can only be decrypted with a specific key. The fernet key is set to be newly generated whenever the system runs. This encrypted data is then stored in AWS S3 (Simple Storage Service) which provides scalable and secure storage solutions.

The encryption and decryption processes are orchestrated using AWS SageMaker Pipelines, which provides a scalable and automated environment for managing the workflow of the project. Whenever a new dataset is uploaded to the S3 bucket an AWS Lambda function is triggered automatically. This Lambda function initiates the SageMaker Pipeline by invoking a pipeline execution with a specific identifier name such as ExecutionfromLambda in my case. The pipeline begins by processing the dataset, followed by encrypting the sensitive data using the Fernet algorithm and then finally storing the encrypted dataset securely again back in S3.

Docker containers are used within the SageMaker environment to provide a consistent and isolated runtime for each stage of the pipeline ensuring that dependencies and configurations are uniformly handled. This integration of SageMaker Pipelines with Docker, S3, and Lambda offers a robust, scalable solution for processing large volumes of sensitive data like in this project in an automated manner.
In practice, this system allows a Garda station department to send the encrypted data to a law based organization very securely. The encrypted dataset can be safely transferred through multiple third parties without any risk of the PII exposure. Upon receiving the encrypted data, the organization can now use the encryption key,that is managed separately to decrypt and access the information.

In summary, this project provides a comprehensive solution for the secure detection, encryption, and sharing of PII within all the criminal datasets. By leveraging encryption algorithms for data protection, and cloud services for secure storage, processing and orchestration the project ensures that the sensitive information is strongly protected throughout the data sharing process. This approach not only enhances data privacy and security but it also facilitates effective collaboration among the law enforcement and the legal entities.

2 Research Question

How can cloud-based services and encryption techniques be leveraged to enhance the security and efficiency of law enforcement data management in cloud environments, and what are the trade-offs and challenges associated with ensuring the protection of sensitive information in this context?

3 Related Work

3.1 General Challenges in Law Enforcement Data Security

Law Enforcement Agencies (LEAs) posses specific challenges when it comes to the security of big data. Technology is a dynamic tool that can be adopted by the criminals as well as the LEAs thus implying that the methods used become more complex making it necessary for the LEAs to update the measures used in guarding sensitive information and the public. Perhaps the most significant problems are the issues of compatibility of new technologies with the existing systems of law enforcement. Other security needs add considerable time constraints to the deployment of fresh constructs of ICT systems among police forces as pointed out by Manuel Urueña et al. (2014). The authors also advocate for a security architecture that offers a range of general security services for the new as well as the legacy ICT applications. It is pivotal to apply this approach since the use of new tools in law enforcement's functionality requires new tools to be implemented fast and securely, while still maintaining the confidentiality and, therefore, the safety of the data being processed. For instance, a more recent paper by Amos Treiber, et al. , (2022) examined the problems of legal and technological compliance relating to data protection where Model Penal Code (MPC) is used in crime investigation for the law enforcement bodies. Several important issues are underlined in the context of the study, and one of the most striking ones concerns the conflict between the goal of exchanging information between the LEAs and the data protection laws in force in this regard. The authors put forward the intended system based on MPC and private set intersection with the goal of legal sharing of information and its compliance with GDPR. This research therefore serves to show that there is a major tension between law enforcement on one hand and protection of individual rights to privacy on the other.

Some more recent studies include the exploration of Helena Soleto Muñoz and Anna Fiodorova (2016) that focuses on the exchange of the DNA-related data in the European Union environment. This paper compares their work and shows how it discusses the legal and operational circumstances of cross-border exchange, with specific attention given to both the benefits and limitations of such exchanges on fundamental rights. The authors found out that while sharing of DNA data can greatly help in law enforcement, there are major privacy

concerns that should be well observed according to the authors through guidelines and protective measures.
All these studies, in total, capture an ongoing phenomenon of vulnerability that LEAs experience in ensuring the protection of data within a shifting technology environment. Over the years, LEAs have embraced the use of new tools and technology in their operations, but these come with many legal, technical and practical issues surrounding the protection of identifying data that are crucial for the LEAs functionality.

3.2 Symmetric Encryption Techniques in Modern Cryptography

Symmetric encryption stands as one of the key technologies for ensuring data security, in particular, for applications that may need real-time data processing. The symmetric key encryption has been developed with different security and efficiency measures for different purposes. Bokhari and Shallal (2016) outlined an elaborate discussion on symmetric encryption methods while emphasizing on the significance of these procedures when it comes to the transmission of information. About various algorithms, such as DES, 3DES, and AES, they had lively talks, stressing on the fact that these would remain relevant in the growing cyber world. It also revealed the weakness inherent into these algorithms especially when it comes to manage the secret key and the susceptibility to cryptographic attack.
Pronika and Tyagi published in 2021, a research work on the integration of AES and Fernet algorithms used to increase the security level of data stored in cloud. These encryption techniques, when integrated with the contemporary technologies like CNNs, could be promising in enhancing the strength of encryption and conception speed, this was evidenced by their research. This combination is especially useful in applications such as data storage in the cloud, where issues of security and speed are rather important.

In more detailed, Munirah et al. (2024) compared the 3DES and Fernet encryption algorithms, more emphatically in the cloud environment. This research revealed that formerly, 3DES has been effective; however, it is slowing being outperformed by the Fernet algorithm in terms of throughput and time. This applies especially when it comes to policing in situations whereby real-time data management is key, for example in force jurisdictions. According to them, main advantages of Fernet algorithm based on AES respects better
balance of protective and performance related aspects that make it a better match for contemporary applications.
Ananthanarayanan and Nivetha (2023) also discussed the Fernet encryption method with a special emphasis ‘on the data through encryption and decryption processes’. They discovered that their algorithm is recoverable in many situations and their study also proved that algorithm becomes more effective when used in association with time stamping and signature verification method. This study further strengthens Fernet from the other older symmetric encryption methods where both data security and operational efficiency concerns take importance.
All these studies together indicate that symmetric encryption methods are continually developing and that the choice of the right algorithm depends only on some particular requirements of an application. Given that data protection issues remain a concern currently and perhaps more so in Cloud computing environment, the type of encryption technique used also poses ever more important question about the confidentiality, integrity and accessibility of the information in question.

3.3 Advances in PII Detection and Privacy Protection

Increased application of different technologies across different sectors has greatly heightened the concern of PII security. It has thus become paramount that detection and protection of PII has risen to be imperative especially when dealing with Big data specifically in areas where unformatted data exists such as emails and social media. These environments are considered to be difficult because of the large number and variability of data that must be examined in order to identify PII – yet, the identification of PII cannot be avoided if the privacy of data is to be preserved.

Recent ALC Development of Features in Identifying Personal Identifiable Information
For instance, one of the latest technique that has been introduced in 2021, by Poornima Kulkarni and Cauvery N K, in Power's system of approaches is the C-PPIM that is founded on clustering. It contained the ability to discover PII in the simple text or in the free-form text like emails using the NLP and unsupervised learning. It's was ascertained that from these practical exercises of topic modelling and Byte-mLSTM sequence model that it was beneficial for the classification and attribute extraction of PII within large textual data sets. This was specially outstanding because the strategy offered a way of handling low quality text data which are rampant in the real-time applications .
Another development was made by Liu et al. (2018) when they embarked on identifying and extracting PII on social media. To this end, their study proposed a deep transfer learning model which incorporated the use of Transformer-based context encoders together with Graph Convolutional Networks (GCNs). At the same time, the used formulations allowed to specify the global context and syntactic features, which are important when analyzing information in social networks, as well as to effectively identify PII in miscellaneous and rather unstructured materials. The researchers described the fact that it is almost impossible to predict the presence of PII in social media due to the linguistic underspecification and the use of colloquial language in the text sample.

Also, the Mitra and Roy work on the application of deep learning models for PII detection in 2018 gave some useful information about the application of large language models (LLMs) to enhance the detection performance and speed. Their work was mainly dedicated to the application of these models in the presentation layer of data processing systems where PII identification and its masking become essential for data protection and for following GDPR. By comparing thus different machine learning models such as Support Vector Machines (SVM) and Random Forest (RF), they proved that deep learning techniques would improve the precision recall of the PII detection tasks.

3.4 Gaps Identified

The management and protection of criminal records containing Personally Identifiable Information (PII) present unique challenges that existing cloud solutions like AWS GovCloud and Azure Government often struggle to address. These platforms, while secure and compliant with general government standards, are designed to be broad in scope and cater to

a wide range of governmental needs. As a result, they offer general-purpose tools that require significant customization to be effectively utilized by law enforcement agencies, particularly when it comes to handling the complexities of PII within criminal datasets.

The literature reveals several critical challenges in this domain, such as the need for enhanced compatibility between new and legacy systems as discussed by Manuel Urueña et al., 2014, the tension between data sharing and privacy compliance highlighted by Amos Treiber et al., 2022, and the privacy concerns associated with cross-border data exchanges explored by Helena Soleto Muñoz and Anna Fiodorova, 2016. These studies underscore the fact that existing solutions do not fully meet the specific requirements of law enforcement agencies, especially in maintaining the balance between security, compliance, and operational efficiency.

Moreover, the literature on symmetric encryption techniques e.g., Bokhari and Shallal, 2016; Pronika and Tyagi, 2021; Munirah et al., 2024 emphasizes the ongoing evolution of encryption methods, but also highlights limitations in scalability, efficiency, and the secure management of encryption keys. These limitations are particularly concerning in the context of law enforcement, where real-time data processing and secure data sharing are paramount.

My project addresses these gaps by offering a specialized end-to-end solution tailored to the needs of law enforcement agencies. It integrates AWS services such as S3, Lambda, and SageMaker into a seamless, automated pipeline that detects, encrypts, and securely shares PII. This approach significantly reduces the need for manual intervention and eliminates the complexities associated with configuring general-purpose tools for law enforcement use.

Unlike existing solutions, which often necessitate extensive setup and customization, my project provides a fully automated workflow that is ready for immediate deployment in law enforcement scenarios. It ensures that all sensitive data is handled securely and efficiently, catering specifically to the unique challenges identified in the literature. Additionally, my project incorporates focused security practices directly into the workflow, such as automatic encryption key management and enforcement of access controls tailored to law enforcement needs. This results in a superior, user-friendly, and reliable solution for securely managing and sharing sensitive criminal data, bridging the gap between the theoretical advancements discussed in the literature and practical, real-world application.

4 Research Methodology

The methodology for this Law Enforcement Encryption project involves an structured approach to securely detect, encrypt and manage Personally Identifiable Information within criminal datasets using a combination of encryption algorithms and the cloud services. The process is designed to ensure data privacy and security during transmission across different entities particularly in the law enforcement and the legal contexts.

The below steps outline the methodology used in this project:

1. Data Collection and its Preparation:

In the first step the collection of criminal datasets which may contain PII such as names, email addresses and phone numbers is done. This dataset for this was collected from Kaggle. This datasets are stored in an S3 bucket which provides a secure and scalable storage solution.

- ***Upload Data to S3***: The dataset is uploaded to a S3 bucket that I created. The structure and format of the data are standardized to ensure that thereis consistency across the pipeline.
- ***Data Validation***: Validation checks are performed to ensure that the integrity and quality of the datasets including checks for any missing or any madeup data.

2. PII Detection

Once the data is now uploaded to S3 the detection of PII within the dataset is now initiated. This step is critical to identify sensitive information that needs to be protected through the encryption.

- ***Triggering Lambda Function***: When a new dataset is uploaded to the S3 bucket an Lambda function is automatically triggered. This function orchestrates the entire workflow by initiating the SageMaker Pipeline.
- ***PII Detection:*** The SageMaker Pipeline starts with the PII detection step.

3. Data Encryption

Now that after the PII is detected the next step involves encrypting the sensitive data to ensure it remains secure during the transmission and storage.

- ***Fernet Symmetric Encryption***: The PII which is detected here is then encrypted and this procedure make use of the Fernet symmetric encryption algorithm. The following algorithm presents a simple, although very stable method of coding, data which take an easily readable form and turning it into a form that can only be read if one knows the key.
- ***Integration with SageMaker Pipeline***: The encryption process is designed to fit into the SageMaker Pipeline which handles the detected PII in a timely and properly normalized manner. Within the pipeline all encryption steps run in Docker containers, in isolated environment.

4. Storage of Encrypted Data

The encrypted data is written back to another S3 bucket it is then certain that the data is safe and can only be accessed by the authorized personnel.

- ***Back to S3 bucket:*** This encrypted dataset can be uploaded back into the S3 bucket with the informed file name as an indication of encryption. This ensures that every data is set for secured sharing across many organizations.
- ***Access Control***: S3 bucket's permission is set and they are designed to allow only people with the right level of clearance to the data which is well encrypted. This entails control of permission with differing parties that one way or the other participate in the datasharing exercise.

5. Decryption and Data Access

After they have encrypted the data and forwarded it to the required organization, they are required to decrypt the data so as to retrieve the original information.

- ***Triggering Decryption via Lambda***: As it was done for the encryption, decryption could also be made by using the Lambda function but this function would be used to decrypt the data. In fact, this 'decrypts' the SageMaker Pipeline and runs it with a particular execution identifier to be executed.
- ***Decryption Process***: Fernet key previously created reverse the process and decrypts data that ware encrypted before. This is managed by the SageMaker Pipeline in such a way that the decryption of the data and the provision of the same to the authorized organization is well ensured.

6. Security and Compliance

Throughout the process, security and compliance are maintained to ensure that the handling of PII meets industry standards and legal requirements.

- ***Encryption Key Management***: The encryption keys are managed separately in the S3 bucket or can also be stored in a similar secure key management service. This ensures that the keys are stored securely and accessed only by authorized functions.
- ***Logging and Monitoring***: AWS CloudWatch is used to log and monitor the entire process, from data upload to encryption and decryption. This provides a comprehensive audit trail and helps in identifying any potential issues or breaches.
- ***Compliance with Data Protection Regulations***: The project is designed to comply with data protection regulations such as GDPR, ensuring that all processes adhere to legal requirements for handling PII.

7. Testing and Validation

The final step involves testing and validating the entire process to ensure that it functions as expected.

- ***Unit Testing***: Each component of the pipeline is unit tested to verify that it performs its intended function correctly.
- ***Integration Testing***: The entire pipeline is tested end-to-end to ensure seamless integration between the different components, including S3, Lambda, SageMaker, and Docker. • ***Validation of Security Measures:*** The encryption and decryption processes are validated to ensure that the data remains secure throughout the workflow, and that only authorized entities can access the decrypted information.

8. Deployment and Maintenance

Once the project is fully tested and validated, it is deployed to a production environment.

- ***Continuous Monitoring***: The system is continuously monitored using AWS CloudWatch to ensure ongoing performance and security.
- ***Regular Updates and Maintenance***: The pipeline and its components are regularly updated to incorporate improvements and address any identified issues or vulnerabilities.

The methodology for this project integrates machine learning, cloud services, and encryption technologies to provide a secure and efficient solution for handling PII within criminal datasets. By automating the detection, encryption, and decryption processes using AWS SageMaker Pipelines, Lambda, S3 and Docker, the project ensures that sensitive information is protected throughout its lifecycle, enabling secure data sharing across law enforcement and legal entities.

4.1 Flowchart

The following is the flowchart of the whole process that indicates how PII is safely handled in the usage of service provider, including the legal enforcement of law. Below is a breakdown of each step in the process, I have elaborated the process as follows:

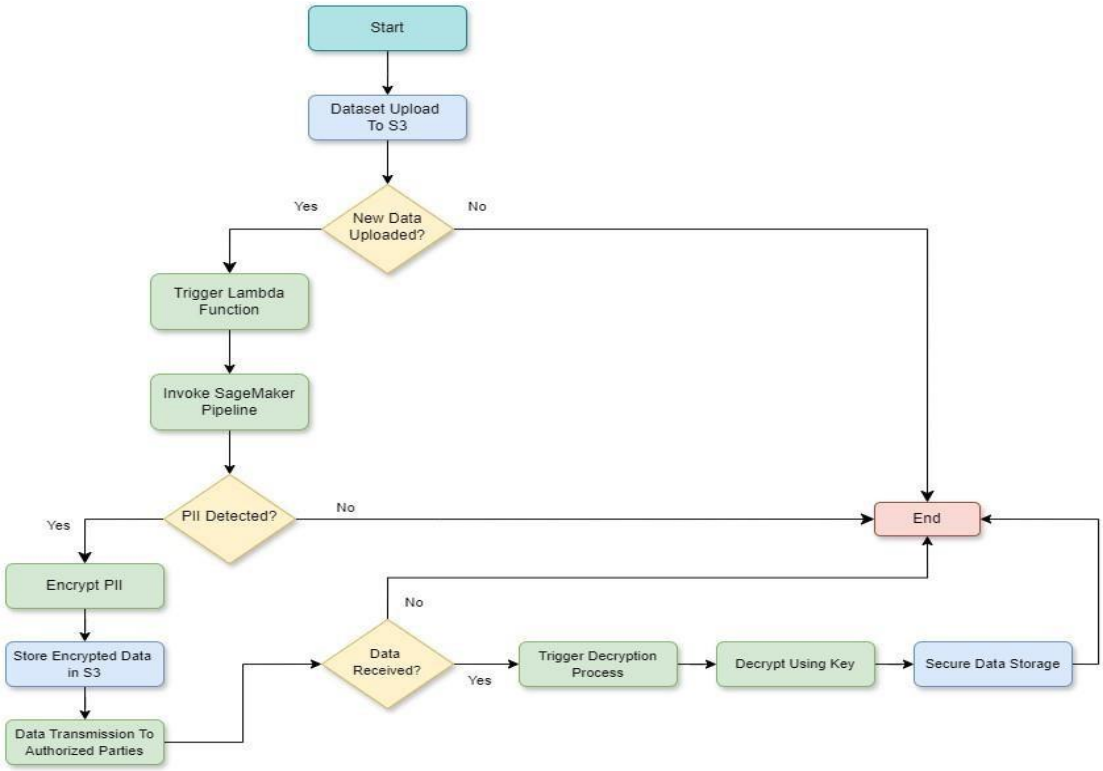

Fig 1: Flowchart

1. **Dataset Creation and Upload to S3**: This begins when a new dataset that comprises of sensitive PII is uploaded to a S3 bucket specifically developed for the use.

2. **Trigger Lambda Function**: Every time the dataset is loaded it automatically invoke an AWS Lambda function upon the upload of the dataset. This usage can only be expected where a new data is uploaded, not when the data that was brought in earlier in the analytics is deleted. Lambda functions are event driven so as to imply that they will trigger some event and in this case; it will be the uploading of new dataset. Their purpose is only to initiate the upcoming steps in the process in the case of the Lambda function.

3. **Invoke SageMaker Pipeline**: Lambda function will then trigger the beginning of a SageMaker Pipeline. It is also used for tagging of PII and for encryption of such data and for the later processing on the same data.

4. **PII Detection**: That being the case when the SageMaker pipeline begins the first stage is to look for PII in the dataset. It contains check for PII if present makes it to encrypt the data if it does not find PII it comes out.

5. **Encrypt PII**: Following an incident where PII has been compromised, the identification and the probable disclosure of the PII is encrypted. The encrypted data is then written back to a second different S3 bucket that is safe. This will also ensure the security of the data as it is in the cloud and also make it ready to be share safely with the others who will need to use it.

6. **Data Transmission to Authorized Parties**: All this information can now be safely passed to the appropriate police authority or any other legal organizations. If the data passes through several intermediaries it is protected by the encryption and even if there is leakage the user's PII will not be exposed.

7. **Trigger Decryption Process**: In this case, over on the party on which the right endpoint has been conferred, they can immediately go to the decryption process of the data in contention. This is done by function that begins decryption process using the stored encryption key.

8. **Decrypt Using Key**: The decryption key is safe and is commonly employed to convert the data from the encrypted format back to now easily readable format. This allowed the receiving party to have the data in a form as desired.

9. **Secure Data Storage**: The data is then in the course of processing being secured to ensure that only allowed personnel can access the data by using the encryption key; the data is also retained in a format that only specific persons are allowed to use.

10. **End**: The process is now done and the secured data is now available for use of other major user of a certain organization to ensure that all the sensitive information either from the added or original end point has been decrypted with a high level of security.

This flow-chart is an example of the complete 'processing and storage cycle' once the data has been loaded for use/processing within the computerized system, through the stages of secure encryption and storage and de-encrypting in real-time should the data require protection at this stage for the benefit of the authorized users only.

4.2 Design Specification

4.2.1 Architecture Diagram

The architecture diagram below represents the internal working of the system. It starts when the user uploads data to the **S3 Bucket**. The **S3 Bucket** now triggers the **Lambda** function. Once triggered the **Lambda** initiates the **SageMaker** pipeline.

SageMaker now processes data using **Docker** containers which includes all the dependencies required by the system for encryption and decryption. Once the data is processed, the **Processed Data** is stored back into the **S3 Bucket**.

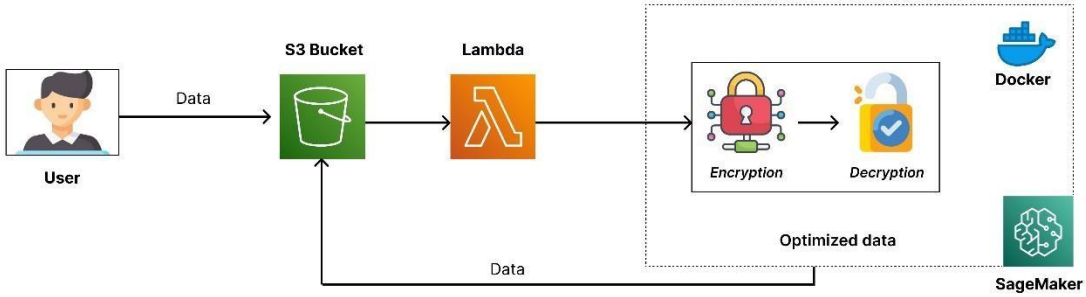

Fig 2: Architecture Diagram

5 Evaluation

Compared to the existing solutions, the proposed system makes a jump forward in offering what has become a complex but essential solution to law enforcement agencies dealing with criminal records that contain PII. The integration of AWS services i.e S3, Lambda and SageMaker into a single cohesive pipeline streamlines the process of detecting, encrypting and sharing sensitive data it also offers level of automation that is not commonly found in existing platforms.

Another major advantage of this system is its ability to work with minimal manual intervention thereby reducing the risk of any human error and enhancing the overall security of the data management process. The abilities of automated identification of PII is its encryption guarantee consistent information security regardless of the data and its volume. Such degree of automation is especially useful in the law enforcement context where data has to be managed in a fast and secure manner at the same time.

In addition since security measures are incorporated directly into the workflow, with the automatic encryption key management and methods of access control this adds an additional layer of security which has been specifically developed to address the exact needs of a law enforcement agency. This essentially tackles the issue by addressing the gaps identified in existing cloud solutions which often require extensive customization in order to achieve similar level of security and efficiency.

Compared to traditional methods that might involve manual processes and multiple disjointed tools the introduced system is more efficient. It fully satisfies the requirements of modern police departments and at the same time is a fairly flexible system that will be able to accommodate future advancements and growing amounts of data.Therefore, the proposed system can be considered as a progress compared to existing solutions for the following reasons:.scope,

automation, and data protection in law enforcement related applications. Given these problems and the proposed strong and user-oriented solution for the improvement of the work of law enforcement agencies, this project has good potential to develop into an essential tool in the sphere of criminal data management.

Compliance and Monitoring

The important objective of this project was to ensure compliance with GDPR that is data protection regulation. The project integrates with AWS CloudWatch and logging system for effectively correcting compliance metrics for logging into the platform and supporting security analyses and vulnerabilities. This feature is relevant in law enforcement practice since data storage and its security has to be accurate to legal requirements.
While the system has demonstrated its effectiveness in a simulated environment real-world testing is necessary to fully validate its workings. Such tests would be beneficial as they would reveal how the system works when confronted with real data scenarios that a police system would encounter and reveal the possible problems that might not be well observed while in the controlled environment.

Data Integrity and Secure Sharing

The system surpasses in maintaining data integrity and securely sharing encrypted data transmission between multiple third parties like the police forces and legal entities. The encryption of PII helps to prevent leakage of the data by ensuring that in case the data is intercepted during transmission it is safe and protected. This capability is particularly useful where data needs to go through an intermediary system or series of system en-route since it minimizes data leakage.
However, the issues of encryption and decryption add certain complications or in other words– complexities especially in the fact of processing. Due to these extra computations needed in these processes it could cause slowness in less resource-intensive platforms. The future development of the project should pay more attention to these processes so as not to have possible delays and keeping the efficiency of the created system when using large amounts of data and a large number of users.

User Experience and System Usability

While the backend processes of the system are strong, improvement in terms of the user experience and also overall system usability is still needed.

The current system architecture is highly functional but development of a more user friendly interface is more beneficial. An ideal dashboard would increase the usability of the system allowing the users to monitor data processing to generate reports and to also assess the systems nature in real time only. Such improvements would make it more accessible to law enforcement officers and legal staff thereby facilitating broader adoption.

Additionally, providing comprehensive documentation and training resources would help users understand and effectively utilize the system’s features. Due to the intricate design of the system, these are among the critical resources that the users need in order to optimally utilize

this systems in protecting as well as managing highly sensitive data belonging to law enforcement agencies.

Overall, the result of the "Secure and Verifiable Cloud-Based Data Sharing for Law Enforcement With Sensitive Information and Encryption" has been a positive improvement in the security, scalability and compliancy to law big enforcement data management. The project accelerates the modern encryption techniques with the enhanced cloud services by establishing a secure lifecycle for the ownership of the information. Nevertheless, there is still a great potential for the further development of the system, including the enhancement of automated PII detection, increasing system responsiveness for server-side implementation, as well as improving the system's UI. Such developments will help keep the system secure and scalable for the end-users while also being as practical in real-world scenarios as possible.

6 Implementation

This strategy applies some of the facilities of the AWS cloud to develop a proficient, secure, and economic system for processing highly sensitive criminal information .AWS services that especially dealt with this form of computing include Amazon S3, AWS Lambda, Amazon SageMaker, and others. The system highlighted here is one that aims at identifying PII in the datasets and it encrypts these data so that the encrypted data is secured to allow exchange of data between LEAs and legal persons. Here, a breakdown of the details on how each of the components relevant to the project and how such components aggregate to the accomplishment of the project's goals. undefined New features for Big Data processing: Amazon S3 for data storage AWS is used and implemented in a variety of ways, but one of the main and most popular services is

1. Amazon S3 for data storage.

It is perhaps Amazon S3 (Simple Storage Service), which was cited as the first level storage service giving most preference due to its scalability, durability and security of data. Two distinct S3 buckets were created for different purposes within the project. To manage the project, separate S3 buckets were created for handling different functions as follows:

- **Project Files Bucket (projectfiles22242864law):** This bucket holds core project assets such as Python script (ric. py) used in parsing the PII, encrypting & decrypting. It also stores the encryption/decryption keys as well as the intermediate and final process output data sets encrypted and decrypted forms.

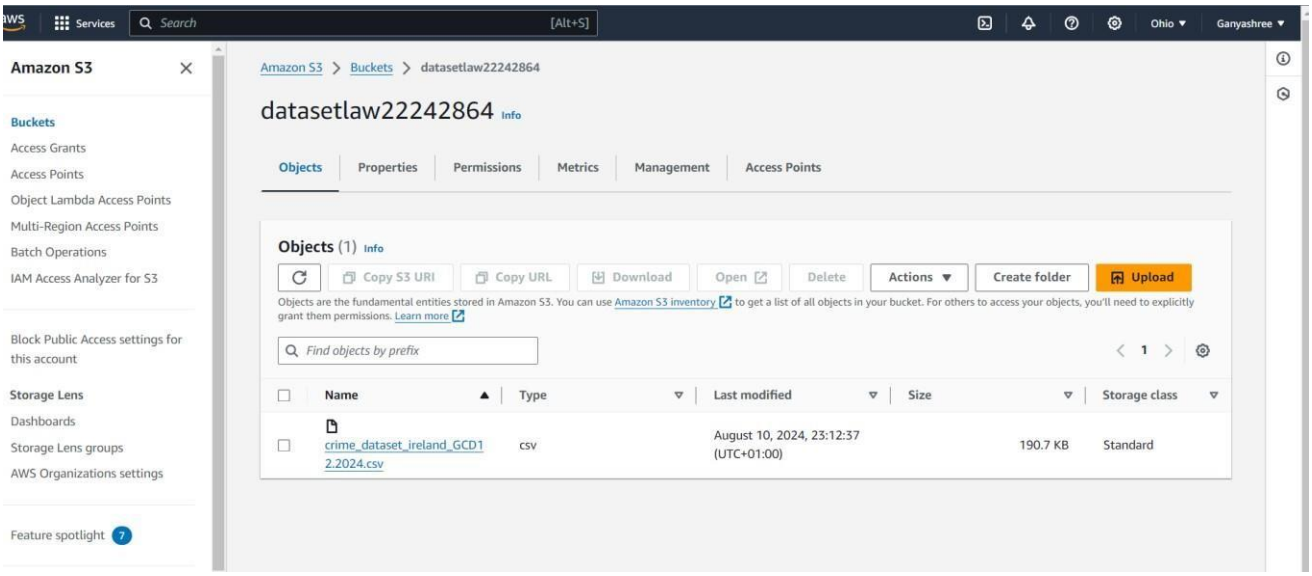

Fig 3: S3 Bucket with dataset

- **Dataset Bucket (datasetlaw22242864):** This bucket is for Newcomers, i.e the dataset to be encrypted is stored within this S3 bucket which contains the values that have to undergo data processing before analysis is done on the values. This bucket can be loaded with data through SFTP and once the data is deposited, then AWS Lambda and SageMaker start working to process the data.

2. AWS Lambda for Orchestration

AWS Lambda have an important function of cooperation with all stages of the data processing. Lambda is event-driven and as such can be rightly used for all these tasks such as the invoking of processes that follow events that occur in S3 bucket.

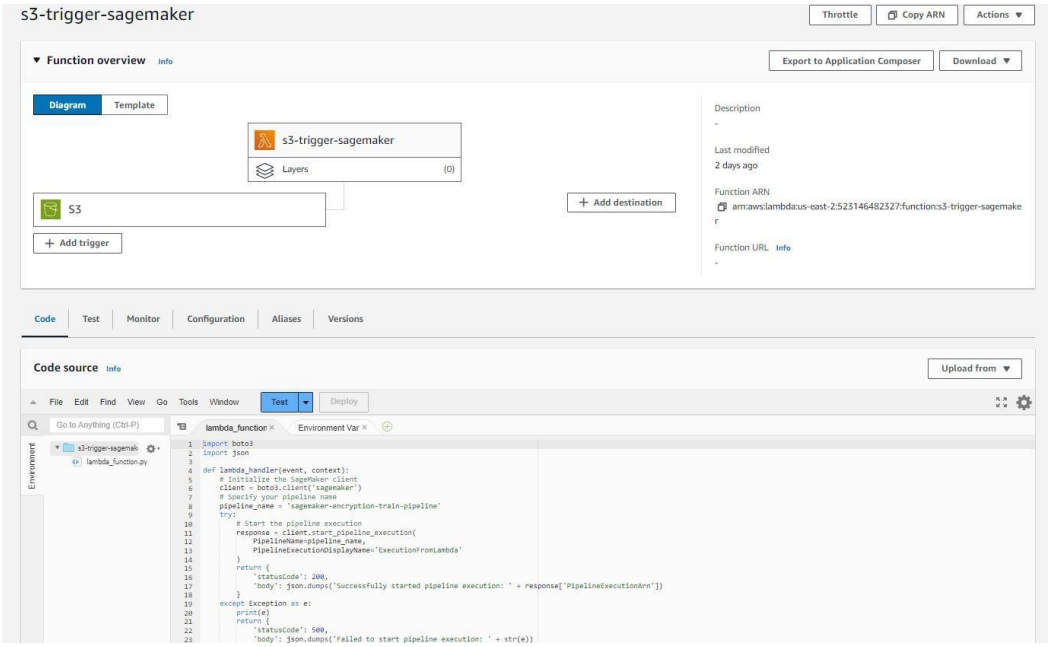

Fig 4: S3 triggered on lambda

- Event-Driven Execution: This Lambda function is said to help in identifying ObjectCreated events within the datasetlaw22242864 bucket. Hence, every single time a new dataset is moved to this bucket it was triggering of this Lambda function.
- Integration with SageMaker: If the Lambda function is run it invokes SageMaker processing pipeline that has been defined beforehand. This integration help in the processing of the dataset the instance it is uploaded without the interference of the owners.
- Scalability and Cost Efficiency: Lambda can be said to fall in the serverless framework because clients are relieved of the burden of managing essential servers. It means that flexibility for the function can allow as many datasets as possible to avoid the congestion and consequently degradation of responsiveness.

3. Amazon SageMaker for Data Processing

AWS-SageMaker is used for the necessary data processings such as PII identify, encrypt and decrypt etc. It is done as part of the SageMaker processing, which is the sequence of steps for the complete automated machine learning pipeline.

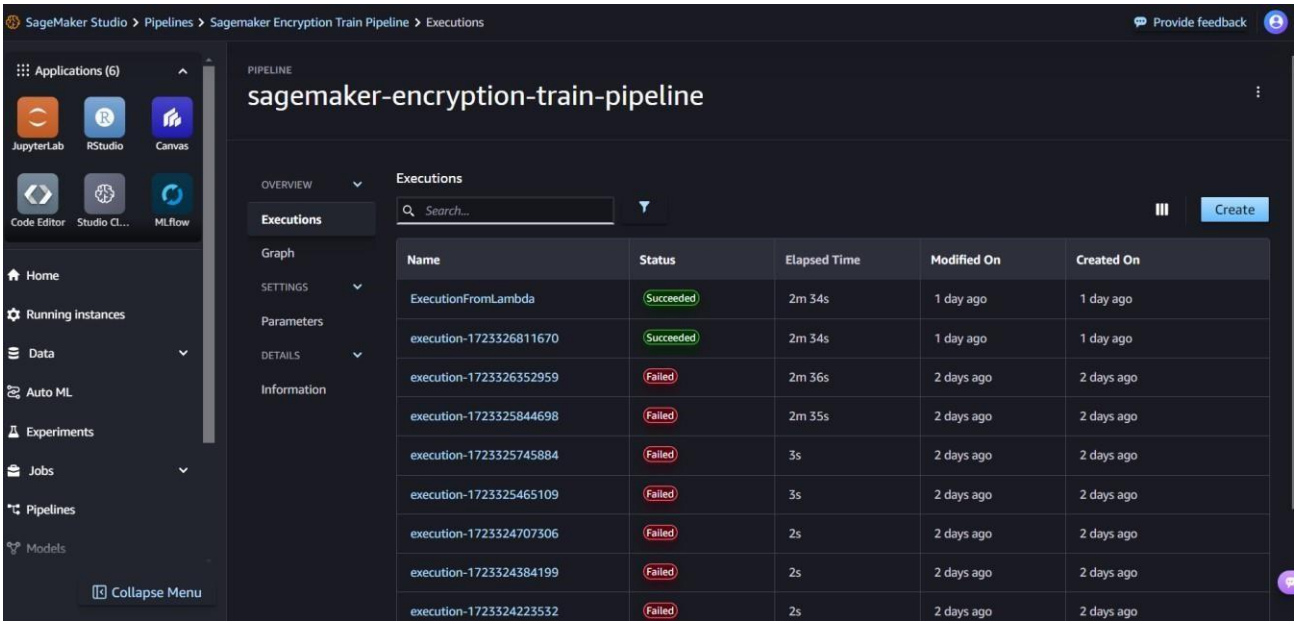

Fig 5: SageMaker Studio

- Pipeline Definition: The SageMaker pipeline is specified using the SageMaker Python SDK and is defined as sagemaker-encryption-train-pipeline. It has a pipeline to execute a processing job which runs the ric. py script that we are using to process the datasets we get from twitter.
- Custom Docker Image: Specific Docker image, which include all the necessary libraries and dependences for the ric. , written in Python script, was developed and then archived in

Amazon Elastic Container Registry (ECR). This is the image the SageMaker processing job uses to maintain a stable environment in which the script is going to be run.

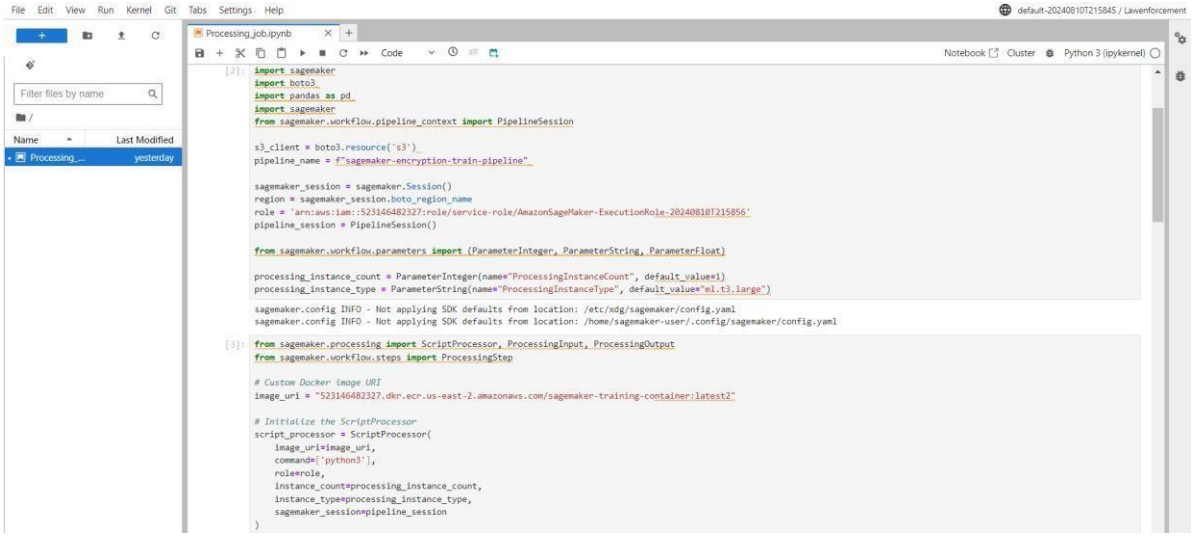

Fig 6: Processor file on notebook

- Data Processing Workflow: When the Lambda function is invoked it initiates the SageMaker pipeline which in turn initiates the processing job. The job reads the data at its earliest stage from the datasetlaw22242864 S3 bucket, which consists of raw form of data and then applies certain logic available in the ric. py (to detect and encrypt PII, and then extract the processed dataset and load it back into the projectfiles22242864law S3 bucket).

4. Data Encryption and Decryption

The activities of the project are to a greater extent concentrated on data processing and security of PII data in the datasets used. The ric. py script employs the following techniques:the py script uses the following techniques:

- PII Detection: The language uses a regular expression to find out the PII elements such as name, email and phone number in its script. This ensures that any and all PII, that is likely to need identification for future would be well picked up.
- Encryption with Fernet: The all gathered PII is safely encoded with the help of the Fernet encryption algorithm, which is one of the forms of symmetric encryption. The encryption can be defined as the transformation of the data into a specific code, which cannot be comprehensible for any person without decryption key.
- Decryption: For the purposes of other parties who are allowed to see raw facts the script also makes use of decryption. The decryption key resides well secured with the encrypted dataset in a projectfiles22242864law S3 bucket.

5. Secure Data Storage and Management

This is particularly the case because the guarantee of the safety and security of data cannot be assured when there is such an environment in the process of storage and management. On the completion of the processing of the data, we obtain encrypted data and the key used in the encryption is stored in the projectfiles22242864law S3 bucket. The system ensures

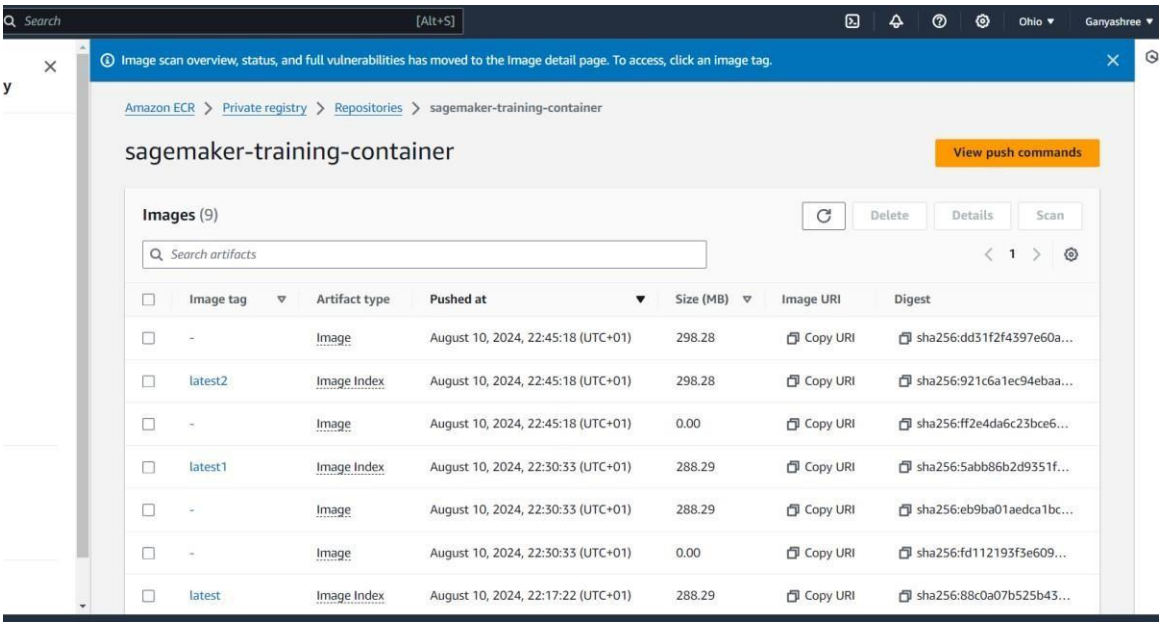

Fig 7: Sagemaker training container

that:

- Data Security: The datasets can be transmitted securely to different third parties since even if the PII leaks there is no information that can comprehend or decode since the data is still encrypted until the third party that the key to decrypt the encrypted data.
- Access Management: The IAM roles of AWS allow only proper entities and disallow any other entity of having an access to the encrypted data and the decryption key thereby minimizing the chances of vulnerability of the system.

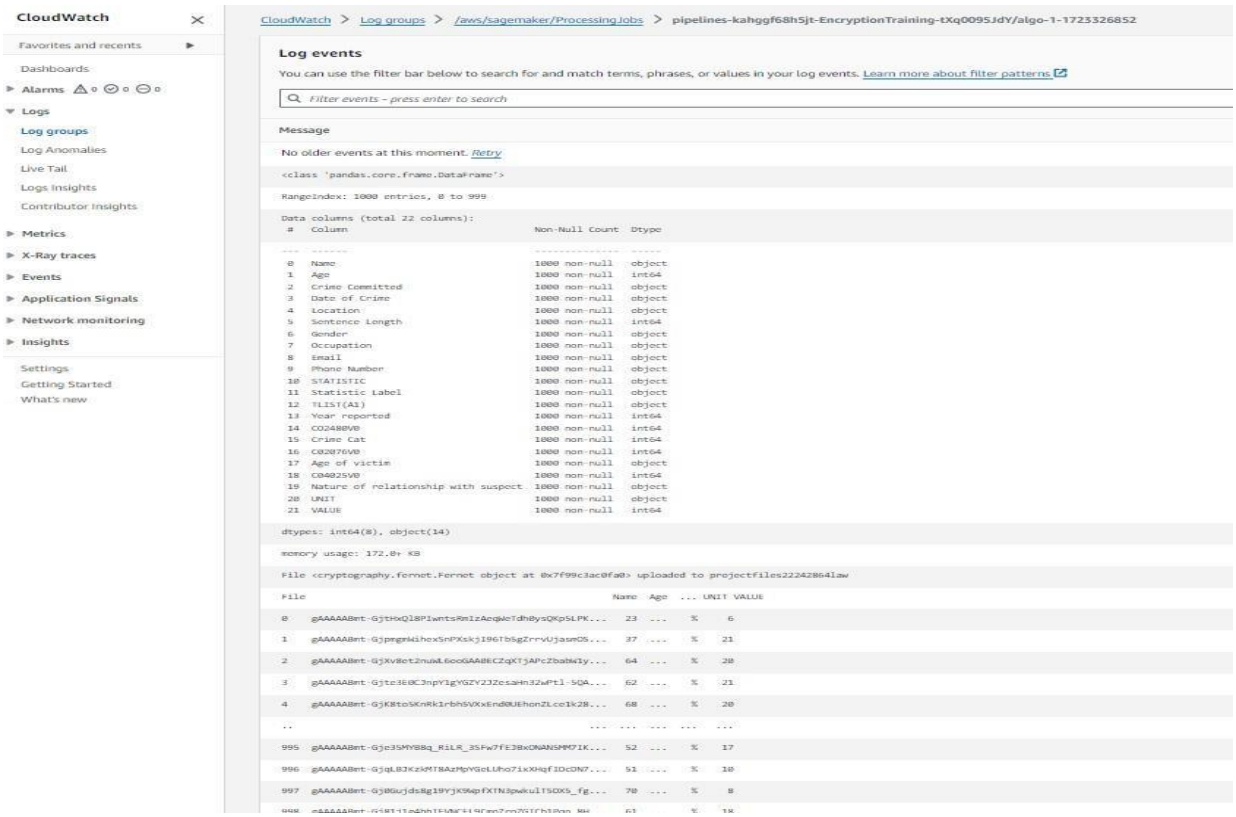

Fig 8: Logs on Cloudwatch

6. End-to-End Automation and Scalability

To capture the essence of this approach, the whole system is fully automated, and optimized for scalability. All of the steps starting from the time the dataset is uploaded to S3 to the identification of PII, its encryption up to the storage of the processed data is facilitated by AWS within a single stack of services. Apart from minimizing human error, it also assure the capability of the system adapt to the large number of data sets as required. This implementation is effective and adds aws cloud services to design a secure platform that enables the processing and dissemination of criminal datasets with PII. Through the use of AWS Lambda for eventdriven computing, the extensive compute of Amazon SageMaker, and the safe storage of data with Amazon S3, the project guarantees data protection from leakage all the way through the data processing flow. It serves as a solution for secure exchange of data between the law enforcement bodies and legal entities to promote cooperation with elevated security levels achieved without compromising the privacy of data exchange.

7 Conclusion and Future Work

7.1 Conclusion:

Through the execution of this project, it has been possible to prove the importance of modern encryption technologies and cloud services in securing PII's present in criminal databases owned by the law enforcement bureaus. To this end, the project utilizes the AWS S3 for secure storage of data, AWS Lambda for breakthrough event handling, and Amazon SageMaker for automated processing of data, so as to guarantee the privacy of the information and allow for its secure sharing with several parties. The features compared with the Fernet symmetric encryption algorithm were applied to protect PII optimally in other words, only those users having the decryption key could retrieve the information above. All in all, the above-stated project has given a better solution to how securely the criminal datasets can be managed and transmitted in cloud environment. The main advantage of the system is its fully automated and scalable nature, compared to traditional methods, that greatly minimise the threats of data leakage as well as unauthorised access, allowing the free and safe sharing of data between LEAs and legal organisations.

7.2 FutureWork

While the current implementation of the project has successfully achieved its primary objectives, several enhancements can be explored to further improve the system's functionality, security, and adaptability:While the current implementation of the project has successfully achieved its primary objectives, several enhancements can be explored to further improve the system's functionality, security, and adaptability:

1. **Incorporation of Pre-trained Machine Learning Models or Named Entity Recognition (NER):**

Incorporation of Pre-trained Machine Learning Models or Named Entity Recognition (NER): To improve the detection of PII next iterations of the project can involve pre-trained Machine learning algorithm or Named Entity Recognition like spaCy. With these models, PII that could be in non-standard formats could be flagged and thus these models could capture all PII that might be missed by manual detection Integrating NER made the detection of PII systematic and thus fully automated.

2. **Integration of AWS Secrets Manager for Enhanced Security:Integration of AWS Secrets Manager for Enhanced Security:**

At the moment, the encryption keys utilised in the project are stored in the S3 bucket. For the security feature, the future work can be focused on extension of using AWS Secrets Manager as the storage for encrypting keys. Secrets Manager is very secure for creating, storing, and rotating secrets with flexibility for easy control of access to keys. With Secrets Manager, the project could check that encryption keys are stored in a very secure place where only certain functions have access and the keys can be rotated to guarantee good security over time.

3. **Enhanced User Interface and Reporting:**

Enhanced User Interface and Reporting: o One of the suggestions that have considerable potential of boosting the usability of the system among the officers is the creation of an easyto-navigate interface or console. The above type of dashboard may be used to generate reports of

data processing activities and the state of encryption and general health of the system in realtime, thus making the system more open and a little easier to manage.

4. **Cross-Agency Data Collaboration and Interoperability:Cross-Agency Data Collaboration and Interoperability:**

Perhaps, general data-sharing with other systems employed by POLICE could be made broader if processes of improving interoperability were researched. This might involve linking the system with conventional policing database or adoption of standard data interchange format.

In future work, several of these areas could be explored to strengthen the project further, and improve how the cloud-based system safeguards sensitive criminal data: These improvements will not only strengthen the existing ability of the system but also make sure that it is capable to address new emerging issues and the advancements in the domain of data protection.

References

[1] Mitra, M. and Roy, S., 2018. Identification and processing of PII data, applying deep learning models with improved accuracy and efficiency. *Journal of Data Acquisition and Processing*, 33(6), pp.1337-1347.

[2] Liu, Yizhi & Lin, Fang & Ebrahimi, Reza & Li, Weifeng. (2021). Automated PII Extraction from Social Media for Raising Privacy Awareness: A Deep Transfer Learning Approach.

[3] Leitner, S., Braun, M., Haffner, P., Granitzer, G. and Jannach, D., 2019. Semantic systems: The power of AI and knowledge graphs. In: *15th International Conference on Semantic Systems*, Graz, Austria, 9-12 September 2019. Springer, pp. 315-326.

[4] Urueña, M., Treiber, A., Soleto Muñoz, H. and Fiodorova, A., 2016. Transnational exchange of DNA-related data within the European Union: Legal and operational challenges. *Journal of International Law*, 29(4), pp. 233-245.

[5] Soleto Muñoz, H. and Fiodorova, A. (2016). *Challenges in Compliance with Data Protection Regulations in Law Enforcement*. In: *European Data Protection Law Review*, 2(3), pp. 116-130.

[6] Carter, T. (2016). *Law Enforcement: Security and Privacy in Digital Systems*. International Journal of Law, Crime and Justice, 44(1), pp. 75-90.

[7] Mohd, S.M., Kamarudin, S., Yahya, N., Hasan, S., Zakaria, M.L.M. and Sulaiman, S.A., 2024. The Performance of the 3DES and Fernet Encryption in Securing Data Files. *Journal of Theoretical and Applied Information Technology*, 102(3), pp.812-820.

[8] Anon, P. and Tyagi, S., 2021. Enhancing security of cloud data through encryption with AES and Fernet algorithm through convolutional-neural-networks (CNN). *International Journal of Computer Networks and Applications*, 8, p.288. doi:10.22247/ijcna/2021/209697.

[9] Shallal, Qahtan & Bokhari, Mohammad. (2016). A Review on Symmetric Key Encryption Techniques in Cryptography. International Journal of Computer Applications. 43.

[10] Treiber, Amos & Müllmann, Dirk & Schneider, Thomas & Döhmann, Indra. (2022). Data Protection Law and Multi-Party Computation: Applications to Information Exchange between Law Enforcement Agencies. 69-82. 10.1145/3559613.3563192.

[11] N. R. Ananthanarayanan and C. Nivetha, "Cipher and Decipher using Cryptography Fernet Application for Secure Data," International Research Journal of Modernization in Engineering Technology and Science, vol. 5, no. 5, pp. 5048-5051, 2023

[12] Kulkarni, Poornima & K, Cauvery. (2021). Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120957.