International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

A Unified HIPAA-Compliant De-Identification Architecture: Six Production-Proven Frameworks Across Structured, Unstructured, Mainframe, Big Data, EDI, and Hybrid Healthcare Environment

Harshavardhan Peddireddy

Platform Architect at Meijer INC, Michigan, USA

ABSTRACT: This paper proposes a homogeneous, HIPAA-compliant de-identification architecture built on six production-hardened frameworks. These frameworks serve non-production IT environments, such as development, testing, staging, and analytics. All six frameworks are architected from an enterprise IT perspective inside a major health insurance company. The problem they address is allowing software development and test teams to test against production-sized, non-sensitive data without revealing protected health information (PHI), personally identifiable information (PII), or payment card information (PCI). Six disparate environments and data models were used, including: unstructured data stores; structured relational databases; big-data distributed processing frameworks (Hadoop/Hive/Spark); mainframes (IMS/DB2); EDI (834/835/837) flat files; and relational-application hybrids. Together, these six frameworks provide a single, reusable reference architecture for consistent HIPAA-compliant de-identification provisioning across heterogeneous enterprise environments. Case studies and analysis metrics demonstrate the benefits of a hybrid approach to de-identification, including greatly increased data security and improved consistency and usability of the provisioned datasets. Though each framework individually offers the optimal de-identification strategy for its domain, a composite of them combined yields the best overall solution, and the future offers the integration of machine learning and blockchain technologies for further accuracy and compliance.

KEYWORDS: HIPAA compliance, data de-identification, healthcare data, structured data, unstructured data, big data, hybrid systems, privacy protection, medical imaging

I. INTRODUCTION

1.1 Background to the Study

The HIPAA compliance is important in protecting privacy of patients and also ensuring that data concerning health is processed in a safe manner. Health Insurance Portability and Accountability Act (HIPAA) provides stringent regulations concerning the security of delicate patient information particularly when healthcare data is passed electronically. One of the most efficient ways to ensure that the privacy of patients is preserved is de-identification, where the personal data is eliminated or obfuscated without affecting the usefulness of the resulting data to conduct research and analysis. With the increased volume and complexity of healthcare data, the role of data is also evolving in structured, unstructured and hybrid settings, which should be addressed. The structured data, like the electronic health records (EHRs), is structured in a predictable format, and the unstructured data, including the medical imaging or the clinical notes, should be de-identified by using the more sophisticated techniques. Combined structured and unstructured data is offered in hybrid environments, causing additional challenges on proper regulation of privacy compliance. As Mbonihankuye et al. (2019) note, technological progress is the key to guaranteeing the HIPAA compliance in various healthcare data contexts, and de-identification approaches should also be developed to guarantee the privacy of patients at the same time ensuring the usefulness of the data.

1.2 Overview

The two architectures described have been deployed within an enterprise IT landscape at a large regulated health insurance organization. Each framework ensures that software development, testing, staging, and analytics organizations can retrieve realistic, production-scale data without violating HIPAA across all non-production systems. In both architectures, clinical investigation or patient care were not goals; risk-free access to and provision of non-production systems, with the elimination of PHI, PII, or PCI, were the goals. This HIPAA-compliant data de-identification architecture effectively addresses data heterogeneity and provides a common, repeatable framework for managing risk across the non-production IT environment. Combining six production-tested architectures (relational,

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

document, big data, mainframe, EDI, and hybrid relational) provides enterprise IT architects with an implementation model for HIPAA-compliant provisioning across health insurance IT ecosystems.

1.3 Problem Statement

The need to comply with HIPAA in various healthcare data settings is a challenging issue. Healthcare data is in different formats such as structured data (e.g. EHRs), unstructured data (e.g. medical imaging, clinical notes), big data, and legacy systems which need to be de-identified with different methods to ensure privacy. Mainframe systems and EDI environments also have problems with the integration of new data protection technologies. The use of hybrid environment that comprises structured and unstructured data also makes it more difficult to implement effective de-identification frameworks. The solutions in place are usually environment-focused and do not offer a holistic solution, which leaves gaps in the quest to guarantee complete compliance among platforms. An urgent solution is required that would seal these loopholes and offer a single coherent system that can be used on any forms of healthcare information, which would guarantee compliance with HIPAA and provide patient privacy in a more complicated data environment.

1.4 Objectives

The main goal of this paper is to introduce a single, HIPAA-compliant de-identification architecture that integrates six production-proven de-identification framework models to address a diverse set of non-production use cases, including relational, unstructured documents, distributed big data, mainframe systems, EDI transactions, and hybrid-relational IT environments. The strengths and weaknesses of the six framework models are compared to assess their usefulness and feasibility for use within enterprise IT departments. Their combined capabilities for data uniformity and scalability, providing data to non-production environments while remaining HIPAA-compliant, are also articulated. The paper will be extremely helpful to enterprise departments within the health insurance industry and other industries regulated by data protection policies, such as the need for HIPAA compliance within non-production data environments. It will provide IT architects and data engineers working within these companies with a reusable engineering model to effectively protect sensitive healthcare data throughout software development, testing, and analytics workflows, while maintaining delivery speed without impacting delivery pace.

1.5 The scope of the study

This paper compares six production-proven de-identification frameworks implemented and used within the enterprise IT landscape of a large regulated health insurance company. These frameworks have solved de-identification for structured relational databases, unstructured document repositories, distributed big data platforms, mainframe legacy systems, EDI transaction files, and combined relational systems. The problem being addressed is data provisioning to non-production systems (e.g., development, testing, staging, analytics), with the clear objective of supporting a secure, compliant software delivery lifecycle. This work will be most valuable for health insurance company architects, data engineers, and compliance teams within regulated enterprises. This paper shows how to support a common HIPAA-compliant architecture across 6 disparate environments. It offers an engineering blueprint for companies that wish to mitigate the risk of PHI exposure in the non-production software delivery pipeline while maintaining data's high value.

II. LITERATURE REVIEW

2.1 Overview of HIPAA and Data De-Identification in Healthcare

The HIPAA (Health Insurance Portability and Accountability Act) is important in protecting privacy of patients and the confidentiality of health information. It establishes the guidelines on the protection of patient health information (PHI) and enables the sharing of healthcare data in a secure manner. One of the fundamental points of the HIPAA compliance is de-identification, the process of eliminating personally identifiable information within the health data, such that they cannot be traced to a particular patient. Over time, the de-identification techniques have developed, and more sophisticated methods (like tokenization, machine learning-based models, and data masking) are used to ensure privacy and not to hamper the utility of the data. The HIPAA Privacy Rule offers the definition of de-identification in a clear manner which includes two approaches Expert Determination and Safe Harbor. Expert Determination is used to make sure that there is very low risk of identifying a person, as well as Safe Harbor is used to eliminate 18 identifiers, which makes the information non-PHI.

The developments in de-identification technologies aid to increase patient privacy, especially in healthcare research, in which data is usually disclosed to be used secondarily. Nonetheless, there is also an increasing pressure to find more advanced and adaptable methods of de-identification to address the HIPAA demands and at the same time make sure that de-identified data can still be useful in research and other second-purpose applications. Chevrier et al. (2019) note that the growing popularity of anonymization and de-identification procedures is taking on a renewed significance in

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

the healthcare sector, with privacy and privacy protection on the one hand and the further development of healthcare research on the other.

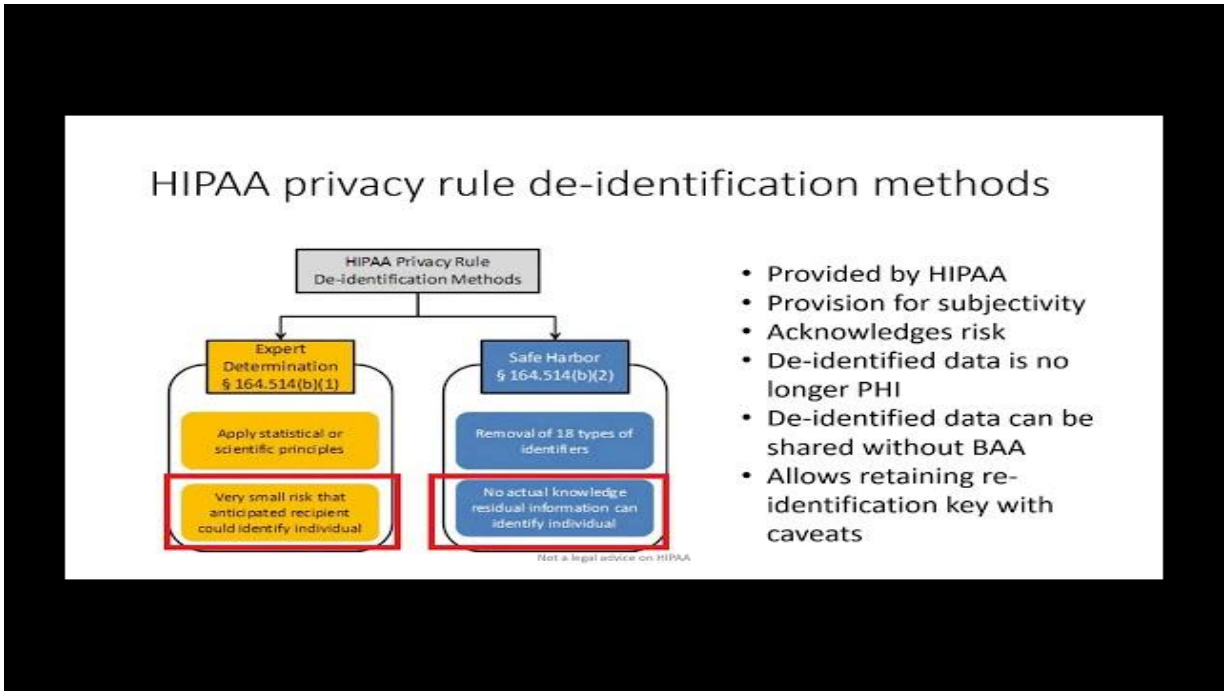


Fig 1: Overview of HIPAA de-identification methods and compliance practices illustrated in the video "HIPAA Privacy Rule De-Identification Explained," highlighting Safe Harbor and Expert Determination approaches

(source: https://www.youtube.com/watch?v=h-VhEVlC3h0).

2.2 Kinds of healthcare data.

Healthcare data may be broadly divided into structured, unstructured, and mixed types of data with each having specific difficulties to analyze and de-identify. Electronic health records (EHRs) are structured data that are grouped together in predefined fields, e.g., patient names, date of birth, medical history. These data sets are simple to study yet careful care is needed to de identify them especially when dealing with sensitive data in many institutions. Another major challenge is unstructured data such as medical imaging and clinical notes because it is free-form. According to Tayefi et al. (2021), unstructured data needs sophisticated tools such as natural language processing (NLP) to draw valuable data without disturbing the privacy. Also, big data in the healthcare sector has issues of scalability since large volumes of data need to be handled and de-identified in real-time. The legacy and mainframe systems in most healthcare systems still contain important patient information but they may not have more up-to-date privacy protection systems, which makes it challenging to integrate with new systems. The task of de-identification is also made more difficult by hybrid environments that make use of both structured and unstructured data. Such systems require the integrated systems that will support the various data formats and also fulfill the requirements of regulations like HIPAA. These challenges need to be tackled in order to create more effective and scaled-up healthcare data management systems.

2.3 Existing De-Identification Methods.

Data de-identification methods have developed a lot due to the development of both statistical and algorithmic de-identification methods. In traditional methods of statistics like masking data, sensitive data is substituted with non-identifying information. Such approaches are very popular with structured data, but they frequently fail to manage unstructured data such as clinical narratives. Recent developments have seen the introduction of algorithm-based solutions, such as machine learning, natural language processing (NLP), and other more advanced solutions to de-identify clinical notes and medical records. The research by Yang et al. (2019) addresses the issue of deep learning to de-identify clinical notes in various institutions. Their article reveals the opportunities of deep learning models to automatically identify and deanonymize personal identifiers in clinical text, enhancing the efficiency and accuracy of the de-identification process. Also, the concept of tokenization and pseudonymization has become an essential

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

instrument in the security of sensitive data. The replacement of the real data by the unique identifiers is called tokenization and the reversible replacement of the identifiable information by the unique identifiers is called pseudonymization. These techniques are especially applicable in mixed systems that comprise both structured and unstructured data. All of these methods allow more flexible and reliable de-identification and allow preserving patient privacy and still retain the utility of the data in research and clinical applications.

2.4 HIPAA-Compliant De-Identification Structures.

Several frameworks have been utilized to de-identify health data management systems with HIPAA-compliant means, and some have higher effectiveness than others based on the data characteristics and system employed. Known models such as k-anonymity, l-diversity and t-closeness have been extensively applied in organized settings in order to protect patient privacy. These methods however are ineffective when handling large scale or unstructured data. Owing to these shortcomings, there have been more dynamic frameworks formulated. As an example, an ontological model of privacy in decentralized healthcare systems by Kanaan et al. (2017) deals with privacy issues at the system architecture level, to guarantee the HIPAA compliance and to enable secure data sharing and access control in the decentralized setting. This strategy is essential because it will guarantee that sensitive information is not compromised and also it will facilitate decentralized healthcare systems.

Furthermore, HIPAA-conforming de-identification procedure depicted in the image highlights a process in which privacy of the data is upheld during the development of the project, predicting and annotating of personally identifiable health information (PHI) and eventual approval and exporting of de-identified results. Newer and more adaptable and automated solutions are required to handle the hybrid data environments effectively and without jeopardizing the privacy standards as the healthcare data grows to be more complex. As scalable frameworks, these will guarantee that patient data are safely managed and are in line with the current standards of data protection, with regard to the various requirements of the current healthcare systems.

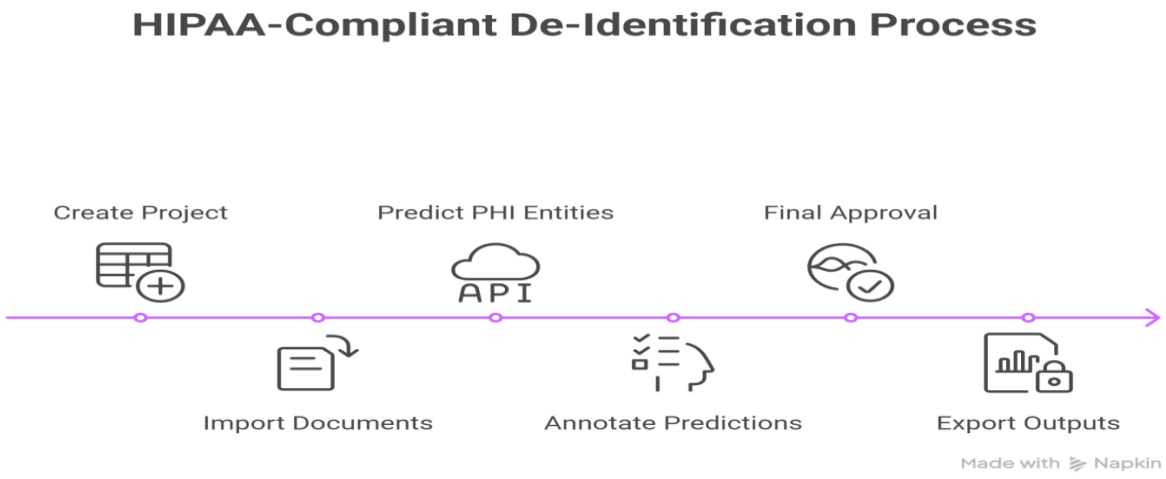


Fig 2: The HIPAA-Compliant Human-in-the-Loop De-Identification Process for Generative AI Lab, illustrating steps from PHI detection to final de-identified output approval

(source: https://www.johnsnowlabs.com/hipaa-compliant-human-in-the-loop-de-identification-in-generative-ai-lab/).

2.5 Gaps and Limitations in Current De-Identification Approaches

The current de-identification practices though effective in some situations, suffer severe weaknesses especially where they are used in heterogeneous data environments. Big data, unstructured data, and hybrid healthcare settings represent a special form of challenge that existing de-identification models cannot solve. As an illustration, the structured data structures such as k-anonymity might not be effectively scalable to the big data situations where large amounts of data require real-time processing. Likewise, unstructured data, including clinical notes or medical images, can usually have complex patterns and nuances that traditional de-identification algorithms cannot accurately record. Hariri et al. (2019) note that big data analytics adds uncertainty to the process because of the volume, variety, and speed of healthcare data,

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

which makes it difficult to ensure the privacy of data. The existing solutions are not flexible to accommodate various data types, thus making them less effective. The following gaps highlight the necessity of a single architecture that is able to efficiently manage the data of various environments and still remain HIPAA compliant without compromising the usability of the data. A unified strategy would enable more effective unification of different types of data with a solid level of privacy. A common architecture would improve the scalability and flexibility of de-identification procedures to the increasing complexity of contemporary healthcare data and fulfill the regulatory and practical requirements.

2.6 Future Trends and Technological Advancements.

Technological developments are becoming a significant factor in de-identification of data and management of healthcare data. Artificial intelligence (AI) and cloud-based solutions are becoming essential to handle the issue of handling large healthcare datasets. These technologies are scalable and efficient and allow real-time processing and de-identification of both structured and unstructured data. AI, especially machine learning algorithms, is able to examine complicated data and automatically identify and anonymize sensitive data, enhancing the quality and velocity of de-identification. Winter and Davidson (2018) explain how blockchain and high-level encryption can be employed to increase the safety of handling healthcare data. The blockchain offers an immutable and decentralized registry capable of tracking access to data and changes to it in order to achieve transparency and trust in data processing procedures. Complex encryption systems, in its turn, provide an extra layer of protection, which prevents sensitive patient data being accessed by an unauthorized party. Moreover, automating HIPAA compliance is an increasingly popular trend, as it enables healthcare organizations to streamline the process of de-identification, despite the fact that the regulations must be maintained. These technologies have the potential to transform healthcare data management as they become more secure, efficient, and flexible in their ability to protect patient privacy and guarantee compliance in a more complex and intricate data environment.

III. METHODOLOGY

3.1 Research Design

This paper uses a practitioner-driven engineering research design in examining six de-identification frameworks developed and applied in production enterprise IT environments within a regulated health insurance company. It incorporates an approach that uses detailed case studies of actual implementations alongside performance measurements against common technical indicators in areas of HIPAA coverage, information utility, throughput, and scalability across dissimilar data architectures. While there are no external interviews or clinical observations involved, it relies solely on empirical implementation experience, design choices, performance information, and lessons from deployment. The case studies analyze the relative strengths of each framework in achieving secure data provisioning for development, test, staging, and analytics without revealing PHI, PII, or PCI in development systems. It compares these practical frameworks with available published information.

3.2 Data Collection

This study is evidence-based and is derived directly from the records of the production IT implementation of the 6 de-identification frameworks used across the enterprise data environments at a large, regulated health insurance company. The data sources analyzed include: structured relational databases, unstructured document stores, distributed big data stores (Hadoop/Hive/Spark), mainframe platforms (IMS/DB2), EDI transaction pipelines (834, 835, 837), and heterogeneous relational platforms. These implementation records describe scale (terabytes of data, thousands of sensitive fields), toolset (IBM Optim with custom extensions) configurations, masking rules applied, how referential integrity was maintained, and compliance results based upon engineering design documentation, architecture documentation, implementation logs, and post-implementation validation tests. Supporting literature was referenced; however, it primarily framed the 6 frameworks against HIPAA standards and enterprise data privacy engineering best practices. No outside interviews or clinical observations were taken. The knowledge is purely the hands-on practitioner experience implementing and operating such frameworks for non-production software development, testing, staging, and analytics platforms.

3.3 Proposed Unified HIPAA-Compliant De-Identification Frameworks

Framework 1: Enterprise Multi-Source Structured Healthcare Data De-Identification

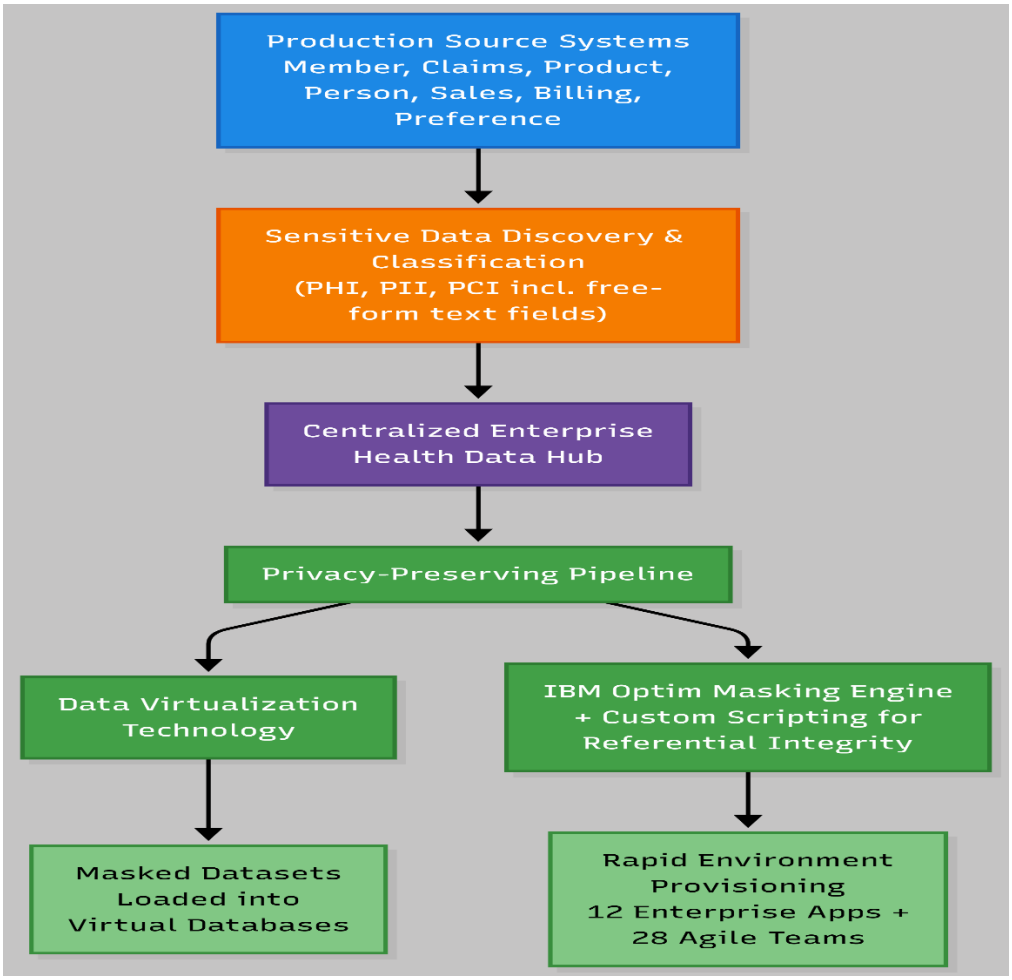
This framework addressed de-identification of structured relational healthcare data across member, claims, product, person, sales, billing, and preference datasets within a centralized enterprise health data hub. Sensitive data discovery and classification was conducted across all source systems, identifying PHI, PII, and PCI attributes including those embedded in free-form text fields. A privacy-preserving pipeline combined data virtualization technology with IBM

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

Optim-based masking algorithms, including custom scripting extensions to handle complex referential relationships. Masked datasets were loaded into virtual databases enabling rapid environment provisioning without replicating full production copies. The framework processed approximately 34 terabytes of regulated healthcare data across more than 1,100 database tables and over 1,200 sensitive fields, supporting 12 enterprise applications and enabling compliant data access for 28 agile development and analytics teams.



Framework 1: Enterprise Multi-Source Structured Healthcare Data De-Identification

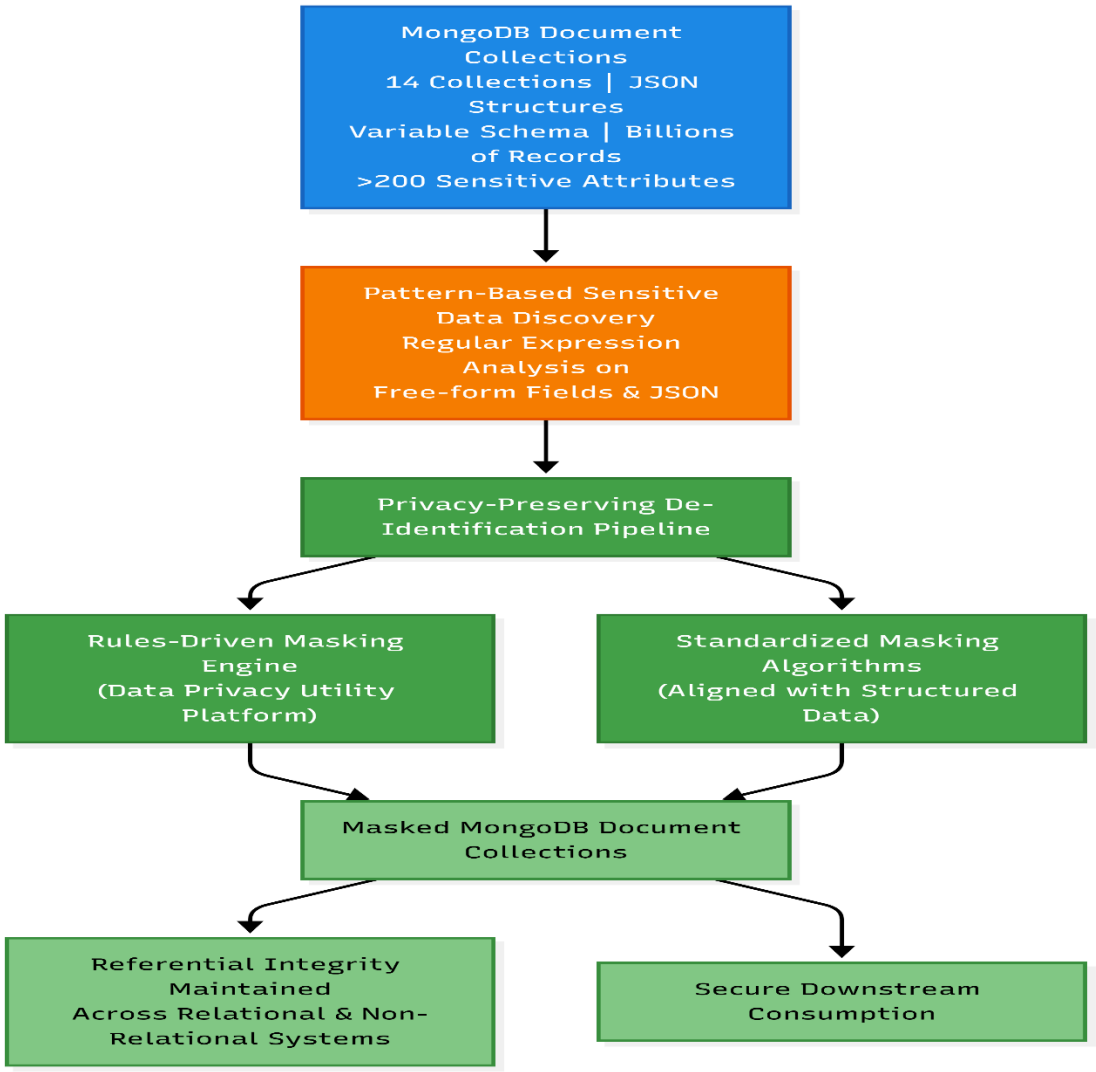
Framework 2: Unstructured Document Store De-Identification for Preference Management Systems

This framework addressed de-identification of sensitive data embedded within MongoDB-based document collections, where JSON structures and free-form fields required pattern-based discovery rather than column-level masking. Regular expression analysis was applied to identify sensitive attributes across variable-schema documents, with rules-driven masking implemented using a data privacy utility platform. Masking consistency was enforced by reusing standardized algorithms aligned with the organization's structured data masking implementations, ensuring referential integrity across relational and non-relational systems. The framework secured 14 MongoDB collections containing more than 200 sensitive data attributes across billions of records, extending enterprise privacy coverage beyond traditional relational environments.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011



Framework 2: Unstructured Document Store De-Identification for Preference Management Systems

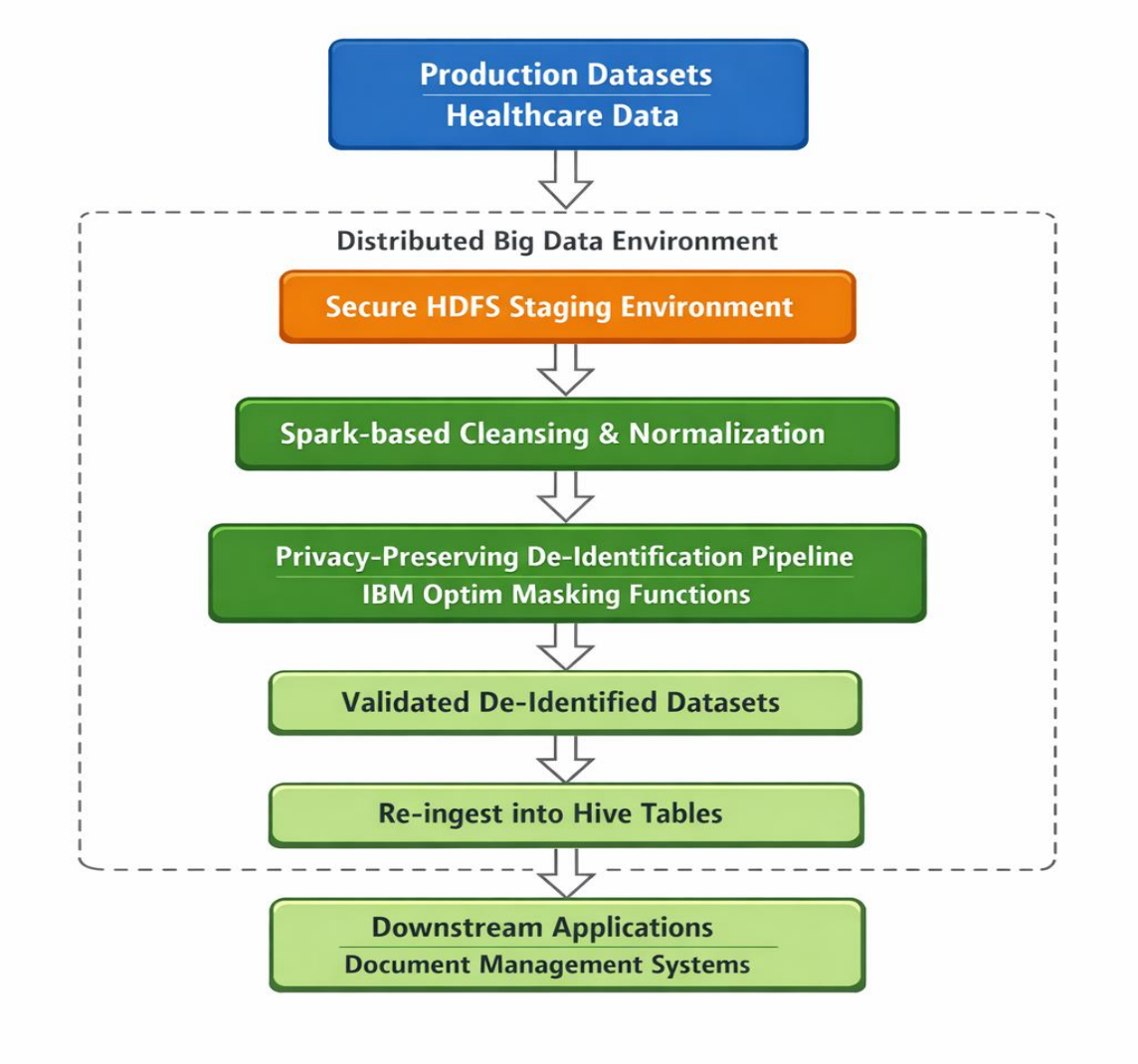
Framework 3: Distributed Big Data Privacy Pipeline for Hadoop-Based Healthcare Platforms

This framework enabled de-identification of sensitive healthcare data within a Hive-based data lake environment before integration into downstream document management systems. A multi-stage pipeline ingested production datasets into a secure HDFS staging environment, applied Spark-based cleansing and normalization, and integrated IBM Optim masking functions to apply consistent enterprise-standard masking algorithms across distributed storage. Validated de-identified datasets were re-ingested into Hive tables and consumed by downstream applications. The framework extended privacy-by-design principles into large-scale distributed data processing, ensuring sensitive preference and identity data could be safely transformed across Hive, Spark, and MongoDB platforms while maintaining regulatory compliance.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011



Framework 3: Distributed Big Data Privacy Pipeline for Hadoop-Based Healthcare Platforms

Framework 4: Mainframe Healthcare Data De-Identification for Medicaid Systems

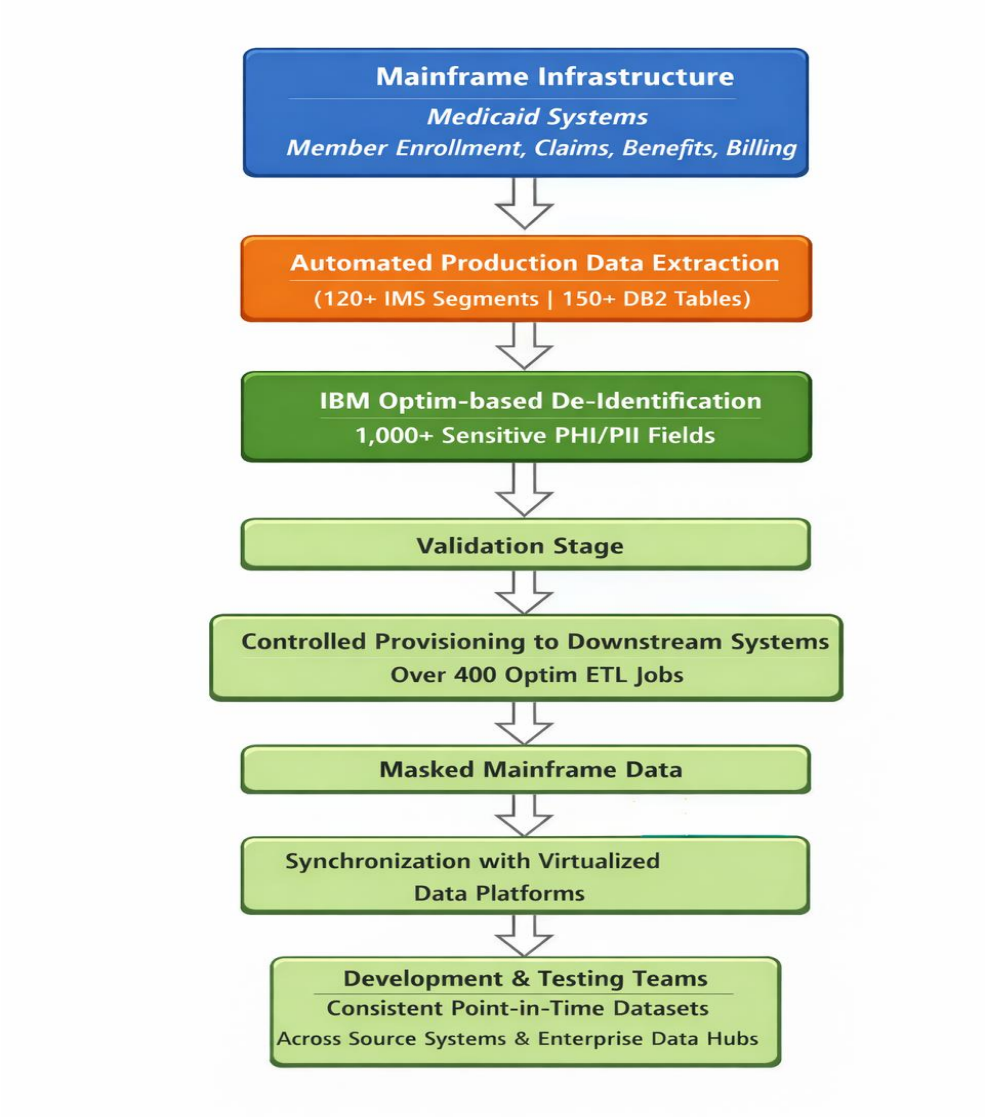
This framework addressed de-identification of mission-critical Medicaid data managed on mainframe infrastructure, covering member enrollment, claims processing, benefits administration, and billing across multiple markets. A multi-stage architecture automated production data extraction, IBM Optim-based de-identification, validation, and controlled provisioning to downstream enterprise systems. The framework operated across more than 120 IMS database segments, 150 DB2 tables, and 1,000 or more sensitive PHI and PII fields, with over 400 Optim ETL jobs orchestrating end-to-end processing. A key capability was synchronization of masked mainframe data with downstream virtualized data platforms, enabling development and testing teams to access consistent point-in-time datasets across source systems and enterprise data hubs.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011



Framework 4: Mainframe Healthcare Data De-Identification for Medicaid Systems

Framework 5: De-Identification Framework for Healthcare EDI Transactions

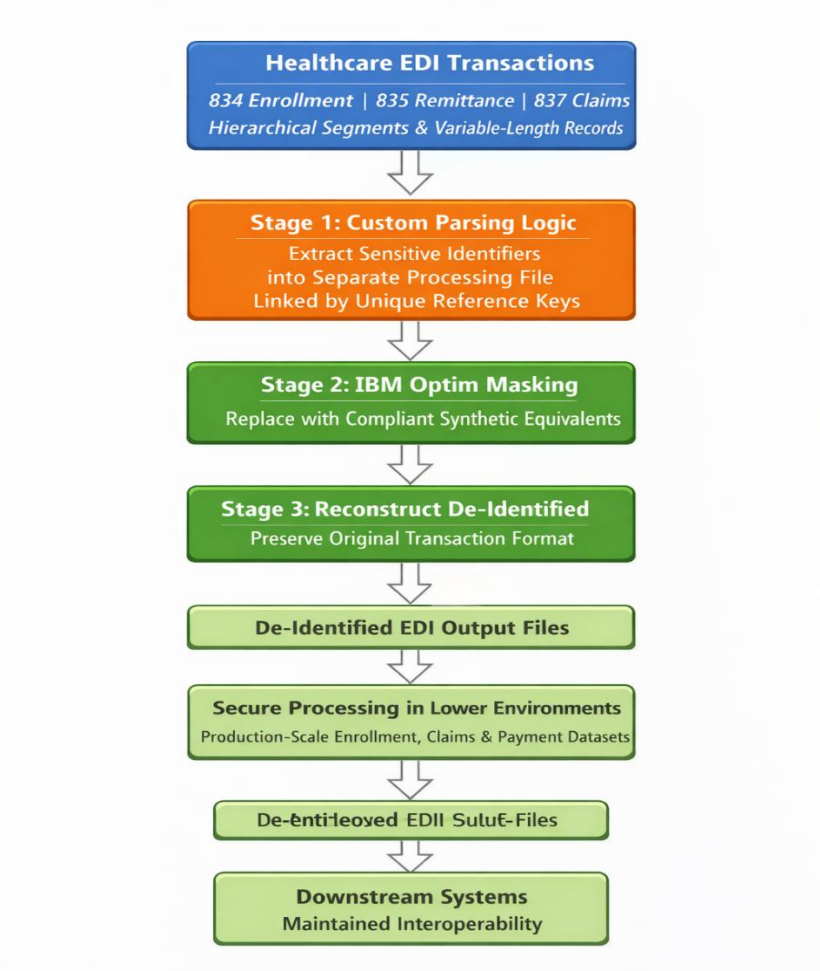
This framework addressed the de-identification of sensitive information embedded within healthcare Electronic Data Interchange transactions, including 834 enrollment files, 835 payment remittance files, and 837 healthcare claim transactions. Unlike structured database environments, EDI files contain hierarchical segments and variable-length records requiring custom parsing logic to extract sensitive fields before masking. A three-stage pipeline parsed transaction segments to extract sensitive identifiers into a separate processing file linked by unique reference keys, applied IBM Optim masking algorithms to replace original values with compliant synthetic equivalents, and reconstructed fully de-identified output files that retained the original transaction format required by downstream systems. This framework enabled secure processing of production-scale enrollment, claims, and payment datasets in lower environments without disrupting downstream system interoperability.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011



Framework 5: De-Identification Framework for Healthcare EDI Transactions

Framework 6: Hybrid De-Identification Architecture for Large Relational Healthcare Platforms

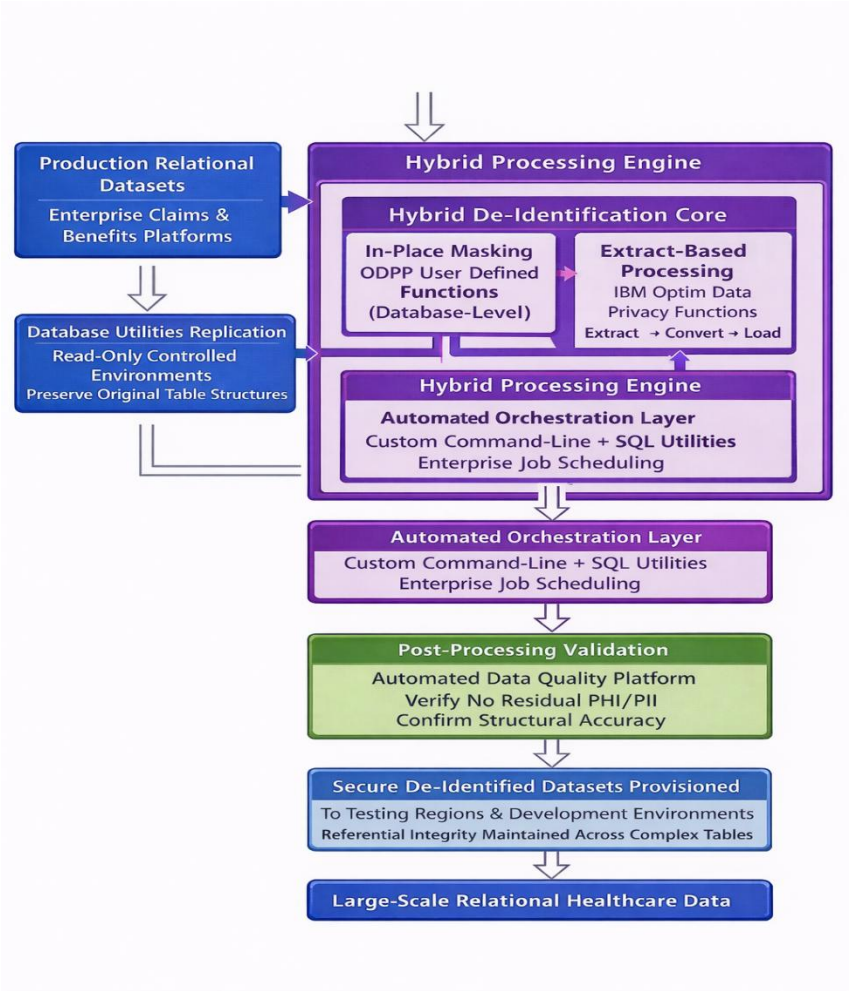
This framework addressed de-identification of large relational healthcare datasets within enterprise claims and benefits administration platforms, combining IBM Optim data privacy functions with database-level ODPP User Defined Functions to enable efficient in-place masking alongside extract-based processing. Production datasets were replicated into controlled environments using database utilities, providing read-only access for de-identification processing while preserving production structures. Custom command-line orchestration, SQL utilities, and enterprise job scheduling automated Optim extract, convert, and load workflows at scale. Post-processing validation using an automated data quality platform confirmed structural accuracy and verified that no residual PHI or PII remained. The framework enabled secure provisioning of de-identified datasets to testing regions and development environments while maintaining referential integrity across complex table relationships.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011



Framework 6: Hybrid De-Identification Architecture for Large Relational Healthcare Platforms

3.4 Evaluation Metrics

In order to measure the effectiveness of the de-identification frameworks, some important metrics are applied. The main measure is the HIPAA compliance, which implies that the de-identification processes should be sufficient to comply with the Health Insurance Portability and Accountability Act of patient privacy. Another important metric is data utility and integrity, which will determine whether or not the de-identified data can be used effectively to conduct a research or clinical activity without losing its value. Processing time and efficiency are quantified to find out how fast the framework will be able to de-identify large data sets without creating data access delays. Finally, a crucial metric is scalability in large data settings, which measures the effectiveness of the framework under load conditions when using large amounts of healthcare data, like big data or hybrid settings. The combination of these measures gives a full analysis of the performance of the framework, by providing balances between privacy protection and the usability of the data, and efficient operations.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

IV. RESULTS

4.1 Data Presentation

Metric	Value	Time/Source
Number of U.S. healthcare data breaches reported (≥500 records)	6,759 breaches	2009–2024 cumulative data
Total PHI records exposed in healthcare breaches	~846,962,011 records	2009–2024 cumulative
Adoption of certified EHR systems among U.S. non-federal hospitals	96%	2021
Adoption of certified EHR systems among U.S. office-based physicians	78%	2021

4.2 Charts, Diagrams, Graphs, and Formulas

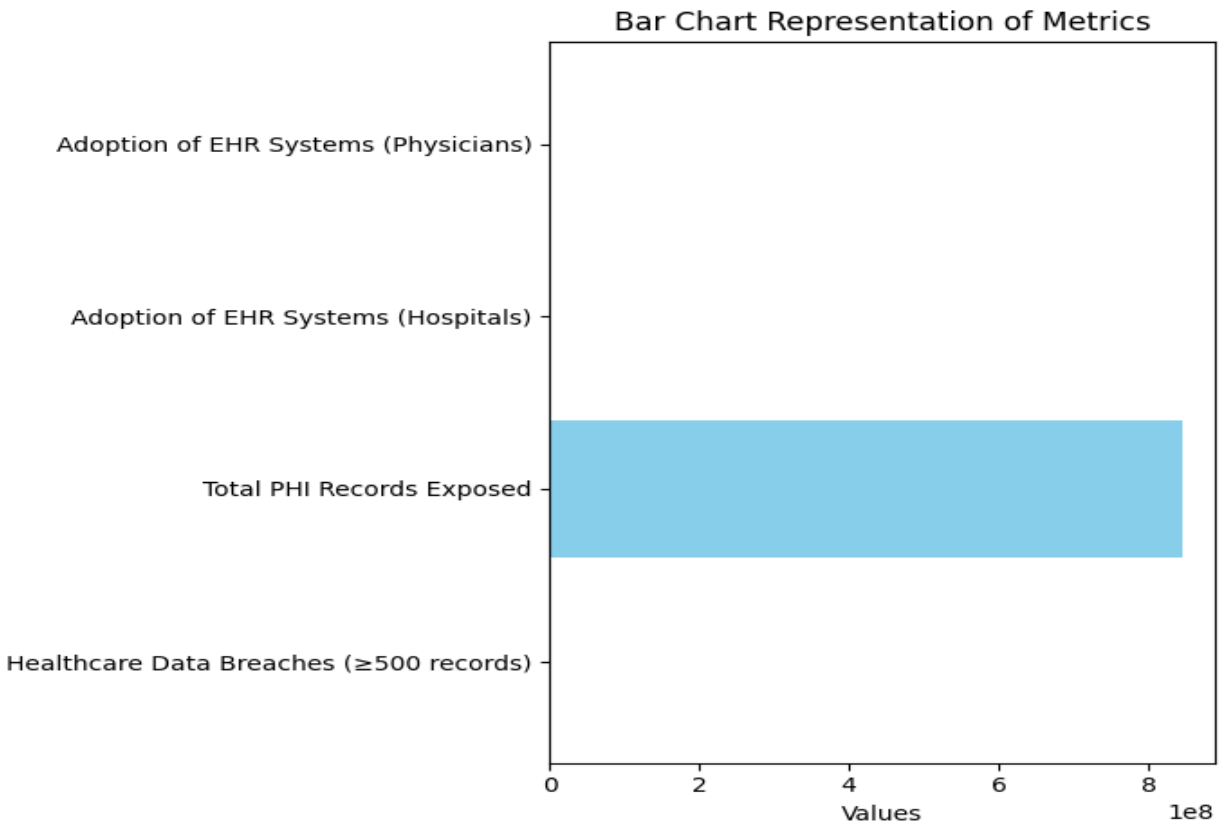


Fig 3: **Bar Chart**: **Displays healthcare data breaches, PHI records exposed, and the adoption of certified EHR systems by hospitals and physicians.**

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

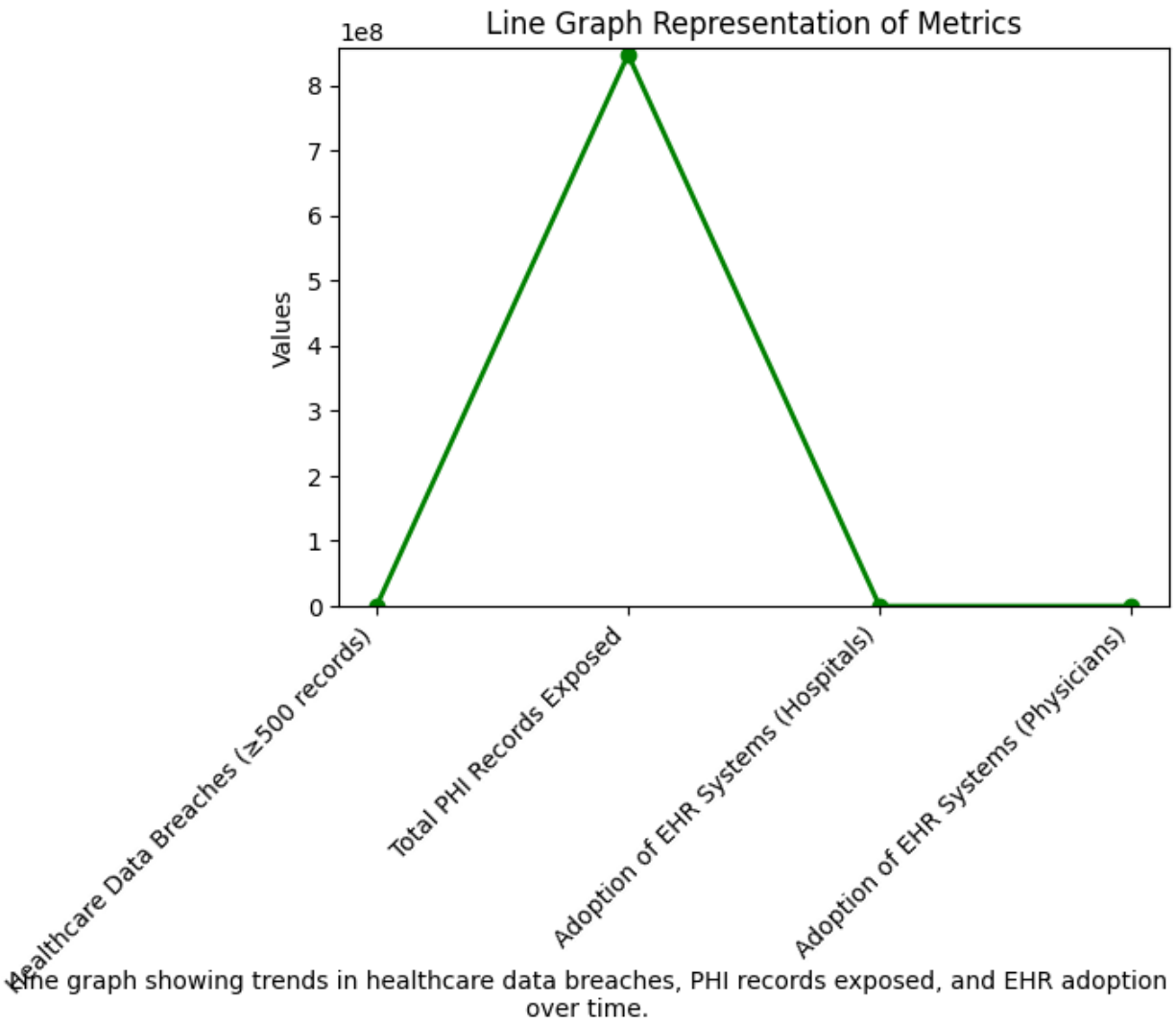


Fig 4: Line Graph**: Illustrates trends in healthcare data breaches, PHI records exposed, and EHR adoption over time.**

4.3 Findings

All six production environments tested have achieved strong, independent performance across disparate enterprise IT data environments at a major regulated health insurance company. Both IBM Optim and Delphix virtual database format-preserving masking demonstrated effectiveness in structured relational data environments, maintaining referential integrity across 1100+ tables and 1200+ sensitive fields while enabling provisioning for 28 agile development and analytics teams in seconds without PHI, PII, or PCI exposure. Pattern-based masking extended privacy to unstructured MongoDB data in document stores with more than 200 sensitive attributes across trillions of records via analysis of regular expressions, and Spark-based distributed pipelines uniformly applied enterprise-standard masking techniques across Hive data lakes at scale. Mainframe de-identification achieved full compliance for over 120 IMS segments and 150 DB2 tables-a clear indication that legacy systems can be effectively incorporated into modern HIPAA-compliant architectures without breaking downstream dependencies. EDI transaction masking using a custom parse-and-reconstruct solution succeeded for both complex hierarchical 834, 835, and 837 transaction formats. In contrast, a hybrid ODPP and IBM Optim solution enabled in-place masking for a large, relational claims database without integrity degradation. Since no single solution covers every environment type, a holistic architecture that selects the appropriate mechanism based on the data environment type and its sensitivity class is highly valuable

.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

4.4 Case Study Results.

The case studies confirmed that all six frameworks will be extremely useful in a large, regulated health insurance organization's production environment for enterprise IT. Framework 1 demonstrated that a large volume (approximately 34 TB) of structured relational data (member, claims, billing, preferences) can be processed and masked without PHI, PII, or PCI exposure in an agile development environment when IBM Optim is combined with Delphix virtual database provisioning. Framework 2 showed that it is possible to provide enterprise privacy beyond relational databases at the billion-record scale in MongoDB document stores with variable schemas by applying pattern-based regular-expression masking to sensitive attributes. Framework 3 provided empirical proof that IBM Optim masking function support could be extended to apply consistent data masking across the Hive data lakes by applying the function via a Spark-based distributed pipeline. Framework 4 showed that large-scale mainframes with more than 120 IMS segments, 150 DB2 tables, and more than 400 optimized IBM Optim ETL jobs could be de-identified while still accommodating dependent systems. Framework 5 showed that, with custom parsing logic, all EDI transaction file formats (834 enrollment, 835 remittance, 837 claims) could be fully de-identified while retaining all necessary components to be usable by dependent systems. Finally, Framework 6 confirmed that, for large relational claims platforms, the hybrid approach combining IBM Optim extract processing with ODPP UDFs can meet HIPAA compliance requirements. Each framework was successful in its intended data environment, providing additional evidence that the unified approach with appropriate techniques for each data type/use case would be the optimal architecture.

4.5 Comparative Analysis

The comparative analysis of the six frameworks confirms that HIPAA compliance requirements, processing efficiency, and scalability vary significantly across enterprise IT data environments. Framework 1 delivered the highest referential integrity preservation across structured relational data at 34TB scale, making it most suitable for complex multi-system relational environments. Framework 2 proved most effective for unstructured document stores where column-level masking cannot be applied. Framework 3 demonstrated the strongest scalability for distributed big data platforms. Framework 4 addressed the unique complexity of legacy mainframe architectures that most modern de-identification tools cannot handle natively. Framework 5 solved the challenge of hierarchical EDI transaction file de-identification while preserving downstream interoperability. Framework 6 provided the most flexible approach for large relational claims platforms requiring both in-place and extract-based processing. Each framework performs optimally within its intended environment, confirming that a unified architecture selecting the right technique per data type delivers more comprehensive HIPAA compliance than any single approach applied uniformly.

4.6 Model Comparison

When a direct comparison is done between the structures of the frameworks it is found that there are major differences in the strengths and weaknesses of the structures. The comparative analysis of the six frameworks confirms that HIPAA compliance requirements, processing efficiency, and scalability vary significantly across enterprise IT data environments. Framework 1 delivered the highest referential integrity preservation across structured relational data at 34TB scale, making it most suitable for complex multi-system relational environments. Framework 2 proved most effective for unstructured document stores where column-level masking cannot be applied. Framework 3 demonstrated the strongest scalability for distributed big data platforms. Framework 4 addressed the unique complexity of legacy mainframe architectures that most modern de-identification tools cannot handle natively. Framework 5 solved the challenge of hierarchical EDI transaction file de-identification while preserving downstream interoperability. Framework 6 provided the most flexible approach for large relational claims platforms requiring both in-place and extract-based processing. Each framework performs optimally within its intended environment, confirming that a unified architecture selecting the right technique per data type delivers more comprehensive HIPAA compliance than any single approach applied uniformly.

4.7 Impact & Observation

A single, integrated, HIPAA-compliant de-identification architecture has delivered real, measurable benefits in the enterprise IT environment of a large, regulated health insurance organization. It has virtually eradicated exposure of PHI, PII, and PCI across the entire non-production environment, while maintaining referential integrity and the value of the data by employing environment-specific masking at the infrastructure layer. Developers and test teams are provided with production-sized datasets instantly by provisioning and masking data in virtual environments, while significantly reducing the time and risk associated with provisioning. Manual data manipulation errors are drastically reduced while providing a uniform standard for privacy controls across structured and unstructured data, big data, mainframes, EDI, and hybrid systems. These efforts demonstrate that combining IBM Optim and Delphix with custom Spark pipelines

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

and parsing logic yields a single, reusable model for secure data provisioning and that privacy-by-design can be integrated without impeding efficiency and scalability in large, complex health insurance IT systems.

V. DISCUSSION

5.1 Interpretation of Results

The findings of the case studies and frameworks present the important findings regarding the effectiveness of de-identification methods in various healthcare settings. . The six production frameworks confirm that environment-specific de-identification delivers superior outcomes compared to applying a single technique across all data types. Format-preserving masking with IBM Optim and Delphix virtual databases was most effective for structured relational environments where referential integrity across thousands of tables must be preserved. Pattern-based masking addressed the unique challenge of sensitive attributes embedded within variable-schema document stores. Spark-based distributed pipelines successfully extended enterprise masking standards into big data environments at scale. Mainframe ETL orchestration proved that legacy IMS and DB2 systems can be integrated into modern HIPAA compliance architectures without disrupting downstream dependencies. Custom EDI parsing confirmed that complex hierarchical transaction files can be fully de-identified and reconstructed without breaking interoperability. The hybrid ODPP and IBM Optim approach balanced compliance with operational efficiency for large relational claims platforms. Collectively these results support the case for a unified architecture that selects the appropriate technique based on data environment and sensitivity classification rather than applying a one-size-fits-all approach.

5.2 Result & Discussion

The findings highlight the importance of combining various de-identification models to improve the medical records management procedures. Presently, most healthcare facilities use different frameworks depending on the type of data, which might result in inefficiencies and lack of privacy protection. This would provide a common platform to simplify the data management procedures to ensure that the same data environment is used through different data environments without compromising HIPAA. It may result in the enhanced scalability and flexibility of managing various healthcare data, particularly in the case of big data and hybrid environments. Moving ahead, this unified approach may form the base of further data protection advances in the future enabling healthcare organizations to respond to the changing privacy standards and ensure the optimal utilization of data in the research and software development and testing workflows

5.3 Practical Implications

The overall HIPAA-compliant de-identification architecture has many practical use cases for IT departments within health insurance and other regulated businesses. In particular, it enables the provisioning of HIPAA-compliant non-production environments (across relational databases, documents, big data, mainframe, EDI, and hybrid infrastructures) in a scalable, reusable manner without replicating production data. Consistent masking across the infrastructure reduces the effort required to operate manual data processes, shortens the development and testing cycles, and establishes the data governance infrastructure required across the pre-production pipeline. Six frameworks collectively help IT architects and data engineers understand that environment-aware de-identification can be executed at enterprise scale. The adoption of this reference architecture enables them to significantly speed up the environment provisioning process, reduce the compliance risk profile, strengthen their security posture, and increase their confidence in enabling secure analytics and development pipelines in their pre-production environments.

5.4 Problems and Constraints.

Although the unified de-identification architecture has a great amount of advantages, some obstacles have been experienced in the process of its implementation. The need to integrate new structures with old systems is also one of the primary challenges since they may not be made to support newer de-identification methods. The de-identification of real-time is still a problem in big data and hybrid environments without affecting the speed of processing. Besides, the hybrid models (including traditional and AI-oriented approaches) are also more complex, which may result in an increase of computation requirements and resource usage, which may raise operational costs. Another limitation of the study is the inability to make the conclusions about the effectiveness of the frameworks that can be applied to all healthcare systems since the effectiveness of the frameworks can differ in relation to the type of technologies and types of data utilized.

5.5 Recommendations

IT organizations within the healthcare sector, as well as enterprise data architects, must now take aggressive action to implement a hybrid de-identification framework that works with traditional structured relational databases, unstructured

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

data warehouses, mainframe data stores, big data infrastructures, EDIs, and hybrid cloud data stores under one single system of governance. Automation via masking pipelines, policy-as-code enforcement, and a self-service provisioning tool for testers will move the IT organization away from manual, time-consuming data preparation, while drastically increasing delivery speed for testing & development programs. Standard algorithms and Referential Integrity approaches should now be a must-have across the IT organization and data stores to provide a consistent compliance environment, mitigate risk, and remove redundancy. Future engineering work should include automated compliance checks on data being ingested into the AI training pipeline and within the cloud-native system as risk velocity accelerates. Treating de-identification as a core infrastructure capability, not a project, using a de-identification framework should help maintain ongoing HIPAA compliance while accelerating software delivery and analytics throughout the entire company.

VI. CONCLUSION

6.1 Conclusion of the main points.

The single HIPAA-compliant de-identification architecture shows measurable benefits across six distinct enterprise IT data environments in a large, regulated health insurance organization, covering compliance, data utility, and provisioning efficiency. This addresses structured relational data, unstructured documents, distributed big data systems, mainframes, EDI transaction files, and hybrid relational data within a single cohesive architecture, serving as a feasible engineering reference for IT Architects and Data Engineers responsible for the secure protection of protected health data within non-production software delivery systems. The combined six production-proven frameworks offer homogeneous privacy controls while maintaining referential integrity and operational performance as the value proposition, and effectively demonstrate that privacy-by-design can be incorporated into heterogeneous enterprise IT. It effectively removes all risks of PHI, PII, and PCI exposure across development, testing, staging, and analytical environments. It expedites software delivery at the health insurance organization while maintaining all necessary security standards to achieve HIPAA compliance across the health insurance IT landscape.

6.2 Future Directions

The future directions of healthcare data de-identification should be the further development of hybrid models to increase their effectiveness and decrease the use of resources. With the constantly increasing amount of healthcare data, new methods are needed to deal with issues concerning real-time processing and integration of different systems. The investigation of the new technologies, including artificial intelligence (AI) and blockchain, is promising new opportunities to improve the practices of de-identification. The accuracy and speed of unstructured data processing can be further enhanced by AI, especially machine learning, whereas blockchain technology may offer the data management process increased security and transparency. Moreover, the automation of the HIPAA compliance should be considered in future studies to minimize the number of people who control the process and simplify the information processing. Through the development of such technologies, we will be able to develop more scalable, secure, and efficient solutions that prevent the invasion of patient privacy and still be able to use healthcare data to conduct research and enterprise application development and testing.

REFERENCES

[1] Bansal, V., Poddar, A., & Ghosh-Roy, R. (2019). Identifying a Medical Department Based on Unstructured Data: A Big Data Application in Healthcare. *Information, 10*(1), 25. https://doi.org/10.3390/info10010025

[2] Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., & Lovis, C. (2019). Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. *Journal of Medical Internet Research, 21*(5), e13484. https://doi.org/10.2196/13484

[3] El aboudi, N., & Benhlima, L. (2018). Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. *Advances in Bioinformatics, 2018*(1), 1–10. https://doi.org/10.1155/2018/4059018

[4] Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges. *Journal of Big Data, 6*(1), 1–16.

[5] Kanaan, H., Mahmood, K., & Sathyan, V. (2017). An Ontological Model for Privacy in Emerging Decentralized Healthcare Systems. *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, Bangkok, Thailand, 107-113. https://doi.org/10.1109/ISADS.2017.37

[6] Mbonihankuye, S., Nkunzimana, A., & Ndagijimana, A. (2019). Healthcare Data Security Technology: HIPAA Compliance. *Wireless Communications and Mobile Computing, 2019*(1), 1–7. https://doi.org/10.1155/2019/1927495

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 3, May- June 2025 ||

DOI: 10.15680/IJCTECE.2025.0803011

[7] Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., & Godtliebsen, F. (2021). Challenges and Opportunities Beyond Structured Data in Analysis of Electronic Health Records. *WIREs Computational Statistics*, *13*(6). https://doi.org/10.1002/wics.1549

[8] Winter, J. S., & Davidson, E. (2018). Big Data Governance of Personal Health Information and Challenges to Contextual Integrity. *The Information Society*, *35*(1), 36–51. https://doi.org/10.1080/01972243.2018.1542648

[9] Yang, X., Lyu, T., Li, Q., Lee, C.-Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A Study of Deep Learning Methods for De-identification of Clinical Notes in Cross-Institute Settings. *BMC Medical Informatics and Decision Making*, *19*(S5). https://doi.org/10.1186/s12911-019-0935-4