Towards a Privacy, Secured and Distributed Clinical Data Warehouse Architecture

Ranul Deelaka Thantilage[1,2](✉), Nhien-An Le-Khac[1](✉), and M-Tahar Kechadi[1,2]

[1] School of Computer Science, University College Dublin, Dublin, Ireland
ranul.thantilage@ucdconnect.ie, {an.lekhac,tahar.kechadi}@ucd.ie
[2] Insight Centre for Data Analytics, Dublin, Ireland
https://www.insight-centre.org/, https://www.ucd.ie/

Abstract. Reputed organisations are always prompting Data Warehouses (DWs), which are essential for storing and mining their historical datasets. When it comes to the healthcare industry, DWs are becoming ever so imperative, as efficient storage for medical data is vital for one's health while mining it and seeking new insights. While clinical datasets are very complex, their timely integration and analysis are crucial to providing excellent care for patients. This research aims to provide an efficient data warehousing solution with multiple privacy and security measures integrated by design. Securing the data at all stages: during data input, exploration, pre-processing, selection, analysis, and presentation, is very challenging. This research explores data security from a holistic perspective and possible distributed analysis mechanisms while streamlining data sharing between healthcare centres to increase efficiency and better patient treatments. This study also considers security and privacy issues at all stages of the data warehousing process (data lifecycle) to ensure its correct handling and use. We focus on distributed clinical data warehouse architectures. We also describe the main requirements of a clinical data warehouse for the whole data lifecycle, Data Capture, Acquisition Management, Archiving, Sharing, Reporting, Analysis, and Privacy and Security. The proposed architecture is evaluated considering existing state-of-the-art concerning data analysis and sharing capabilities while ensuring data security and privacy.

Keywords: Data warehouse · Clinical data · Cardiology · Architectures · Data security · Data privacy · Clinical data warehouse requirements

This project is part of the Eastern Corridor Medical Engineering centre (ECME). It is supported by the EU INTERREG VA Programme, managed by the Special EU Programmes Body (SEUPB).

T. K. Dang et al. (Eds.): FDSE 2022, CCIS 1688, pp. 73–87, 2022.
https://doi.org/10.1007/978-981-19-8069-5_5

1 Introduction

It is imperative to perform healthcare analytics on data collected from multiple sources. Data analytics will help identify different trends and patterns for improvised healthcare behaviour. In addition, it will help forecast future drug usage patterns, measure healthcare centre efficiency, and analyse decisions and results of doctors and clinical staff. Moreover, it will aid in better healthcare treatment for patients while increasing the accuracy of treatments and minimizing costs where possible. Healthcare data is complex and collected from various heterogeneous sources. As per the current healthcare standards [x], [20], a clinical data warehouse would need to store a wide variety of data. For example, image files of digital tests and blood reports will be just additions on top of the general patient information and medical history. Especially in the healthcare industry, the technology updates as soon as new research brings up varying digital tests to make treatment much more efficient. These new additions would mean new data types for the DW. Therefore, a key concern would be that the DW architecture should be future-proof to support different data types from varying sources, which are constantly evolving.

Flexibility, scalability, data integration, and software system compatibility are all addressed by current state-of-the-art healthcare data warehousing solutions. Recent healthcare research projects have focused on building GDPR [22] compliant data gathering techniques in the form of a consent management system. However, these systems lack the integration of privacy and security by design for clinical data warehouses. Even though some frameworks have suggested interesting techniques for data security and privacy, they are not always efficient. They need high-end computing power, which is not always available during some critical times of care.

The main research objective is to integrate privacy and security in clinical DW while ensuring data quality and value are not lost during the data transition to secure forms. To summarise, healthcare big data warehouse architecture should integrate security and privacy by design. Data safety and privacy, as well as the capacity to share, analyse and extract meaningful insights to enhance healthcare research and practices, are the main goals of building a healthy clinical DW. In this paper, we propose a distributed clinical DW architecture to tackle the aforementioned security and privacy concerns while ensuring data sharing and availability at critical healthcare timelines. Identifying stakeholders and their requirements is crucial to help develop privacy-by-design and security-by-design clinical DW architecture. It will support efficient data sharing and analysis of healthcare data.

The paper is organised as follows. We discuss related work on DW security and privacy in Sect. 2. In Sect. 3, we showcase the system requirements of a clinical DW and the data flow of the proposed DW architecture. Section 4 proposes privacy and security preserving clinical DW architecture. Section 5 discusses the data processing, privacy, and security components of the system. In Sect. 6, we evaluate the proposed architecture. We include and give some future directions in Sect. 7.

2 Related Work

DWs are a desirable target since they hold the healthcare institute's most valuable assets. The healthcare industry has been substantially affected by the digital revolution. Many aspects of our life are made easier by big data, however, electronic health records (EHRs) include some of the most important knowledge and sensitive patient information. In 2018, around 500 data breaches resulted in the exposure of over 15 million patient healthcare records. Halfway through the next year, the figure had risen to 25 million records according to Davis of Health IT Security [5]. There are several ways to ensure data security. Similar to other computing security principles DW security also follows the CIA principle of Confidentiality, Integrity, and Availability.

Several solutions have been proposed and implemented to prevent such breaches and attacks to happen in distributed big data warehouses. Sebaa et al. [17] have proposed a Hadoop-based architecture and conceptual data model for a medical data warehouse but they have not looked at security or privacy solutions. Using the Advanced Encryption Standard (AES) and the One-Time Pad (OTP) encryption technique, a secure data warehouse architecture is proposed by Gupta et al. [7]. The proposed architecture does not look into privacy-preserving or healthcare-specific DWs. Shahid et al. [18] propose a three-layered big data warehouse architecture. Data access control, secure storage, and data anonymisation are developed in the architecture. They do not have a secure emergency authentication mechanism for sharing critical clinical data. Mia et al. [13] propose a privacy-preserving clinical DW architecture. By combining three data sources, they create a prototype with a whole pipeline from data gathering through analytics. They do not look at emergency data authentication or sharing of critical data. The key properties of these existing solutions were taken into account and further enhanced when developing the proposed clinical DW architecture.

Blockchain has become a key and revolutionary concept in data security and hence integrating it into healthcare data warehousing will give many benefits. Secure Views protect data from unauthorized user access. It shields users from potentially seeing information from data records that have been filtered by the view. De-identification, a type of dynamic data masking, refers to severing the connection between the data and the person with whom it was originally linked. Data anonymisation is a technique to protect data privacy that maintains the data but conceals the source, by removing personally identifying information like names, social security numbers, and addresses from data sets.

2.1 Blockchain

According to Crosby et al. [4] a blockchain is a distributed database of records, or public ledger, of all transactions or digital events that have been completed and shared among participants. Each transaction in the public ledger is double-checked by a majority of the system's members. Haleem et al. [8] show the four main taxonomy of blockchain systems. Public blockchains provide a fully decentralized network, while private blockchains are restricted to a single entity. A consortium blockchain is a permissioned network and public only to a specific group, while hybrid blockchains combine the benefits of both private and public blockchains. In this study, the authors discuss blockchain technology and its major benefits in healthcare, as well as fourteen key uses of blockchain in healthcare. It is critical to ensure that a health blockchain is “fit-for-purpose.” This notion serves as the foundation of the study by Mackey et al. [12], which includes perspectives from a diverse set of practitioners at the vanguard of blockchain conception, development, and deployment. If implemented successfully blockchain would be an ideal solution for security issues such as ransomware attacks, and data breaches. Blockchain-integrated healthcare applications are still in their development, and more effort in terms of technical discovery and research is required. It would be inefficient and costly to store huge documents on the blockchain, such as complete electronic medical records or genetic data sets. In addition, querying data within a blockchain is challenging, restricting clinical, statistical, and research applications. Hence further research should be conducted on how to efficiently use blockchain technology within implementations of minimal data processing resources as the base of a new generation of health information exchange.

2.2 Secure Views

Secured views are designed to aid data scientists in inspecting and understanding the dataset from several angles while keeping patient-sensitive data hidden. When views are particularly specified for data privacy limiting access to sensitive data that should not be disclosed to all users, they should be declared as secure views. Views built only for query convenience, such as views developed to ease queries for which users do not need to comprehend the underlying data format, should not be utilized as secure views may take longer to perform than non-secure views. Shahid et al. [18] break down into three types of secured views; statistical view giving measurements for characteristics that are automatically calculated, such as standard deviations, domain ranges, and value statistics, and anonymised view providing a comprehensive view of shared datasets protected by several techniques, and anatomized view providing both broad and detailed views of quasi-identifiers. Murphy [14] looks into an approach where data is frequently exported into a user-friendly data mart. This also restricts the number of patients a client may see (secure views), which is significant from the standpoint of patient privacy. Kalio et al. [9] discuss a framework for data warehouse security and privacy that uses a hybrid method. To allow secure viewing, the system uses three layers of authentication, and the user must be a registered user. The result set of a query constructed on top of one or more tables is referred to as a view. Views do not retain data; nevertheless, view requirements are evaluated during runtime, and the result is displayed to the user. Secure views are more complicated queries that need more processing power. In addition, a single view becoming compromised can lead to a data breach expanding to all users using that particular secure view.

2.3 De-identification

De-identification masks the true identity of data owners, by deleting all fields that might directly identify a person. Individuals are named using a new random identifier when the data is de-identified and shared. In some circumstances, keeping the relationship between the old and new IDs is needed to update de-identified data. The de-identification procedure can be achieved in two ways, according to the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [21]. They are Expert Determination and the Safe Harbor Method. De-identification and anonymisation go hand in hand. It's frequently mistaken with de-identification and used interchangeably. Although both anonymisation and de-identification attempt to safeguard the privacy of the data subject, they are conceptually distinct. Shukla et al. [19] describes a method for de-identifying electronic healthcare data that uses chained hashing to generate short-lived pseudonyms to reduce the impact of inference attacks, as well as a re-identification strategy that emphasizes information self-determination. Kayaalp [10] looks into modes of de-identification in terms of clinical data, and breaks into 08 distinct modes. Repository-wide batch de-identification, on-demand cohort-specific de-identification, on-demand de-identification of query results, de-identification with patient and provider identifiers, scientist-involved de-identification, patient-involved de-identification, physician-involved de-identification, and online de-identification by honest brokers. Rahmani et al. [15] use a novel bio-inspired algorithm based on the natural phenomenon of apoptotic cells in the human body to solve the challenge of concealing sensitive clinical data in big data warehouses. De-identification can be accomplished in a variety of ways. The effectiveness of the de-identification process is dependent not only on the automatic de-identification systems' ability but also on the users of de-identification systems.

2.4 Anonymisation

Anonymisation is the process of deleting or encrypting identifiers that link an individual to stored data to secure private or sensitive information. The individually identifiable message is relayed via a data anonymisation procedure, which maintains the data but hides the source. Data anonymisation techniques include masking, pseudonymization, generalization, swapping, perturbation, and synthetic data. Santos et al. [16] present a transparent data masking method for numerical values in DWs based on the mathematical modulus operator that may be employed without requiring changes to the user application or source code. Existing pseudonymization models rely on external trusted third parties, making de-pseudonymization a multistage process requiring an additional interpersonal connection, which might result in significant delays in patient treatment. Hence, Aamot et al. [1] suggest an improved technique based on an asymmetric encryption scheme that separates the pseudonymization and de-pseudonymization tasks. Kumar et al. [11] studies dynamic data masking and aid to analyse the level of security needed for real-time applications. A realistic categorization module for a built-in data masking architecture is proposed by Ali et al. [2]. The suggested module would identify sensitive data and choose the optimum masking format to increase data privacy and security at rest. By mapping sensitive data characteristics with the appropriate irreversible or reversible masking techniques, this module allows sensitive data attributes to be securely utilized and conforms with privacy regulatory standards inside the healthcare data warehouse. Anonymisation too has its limitations. As shown by Guerra-Balboa et al. [6] although structural concepts can give high-value data in general, they are vulnerable to vulnerabilities such as background knowledge and attribute-linkage attacks. According to Chen et al. [3] standard k-anonymity approaches cannot adequately safeguard a data set with significant sequential correlation. Healthcare data warehouses contain location-based data in some instances, these approaches are challenged with data sets with sparse or short trajectories, as trajectories might have minimal overlap, resulting in inescapable data and value loss. Further constraints mentioned in related work include the distinction between synthetic and anonymised data, as well as the loss of data quality in some cases.

3 System Modelling

3.1 Requirements of Clinical DWs

Clinical DWs have varying requirements based on their stakeholders. The advent of numerous new forms of organizations, such as physician medical groups and medical research institutions, has resulted from the growth of Big Data. There are an increasing number of stakeholders offering diverse medical services, and their resource input and behavioural engagement have varying implications on the services they provide. Stakeholders in the healthcare industry have a significant impact on the industry's trajectory. By studying the literature, and analyzing the requirement of users and stakeholders such as healthcare organizations, patients, doctors, clinical staff, administrators, researchers, and regulatory organizations, this research identified the requirements of a clinical DW architecture which was further summarized in 05 categories as shown below.

01 - Data Capture Requirements. Clinical data is available in varying file formats (image files, video files, document files, etc.), both open source, and proprietary. Therefore it is a must for the DW to support import from *multiple data types.* In healthcare centres, varying digital devices can be found. Radiography, cardiology, ultrasound, etc. A DW warehouse should support the data import/export from/to *multiple devices.* Doctors tend to write notes on observations done on a patient, these *clinical notes* should be imported to the DW. Data capture should be supported from *non-standard* and *legacy devices.*

02 - Acquisition Management Requirements. Incorrect data might pose serious health risks for patients and place a major burden on practitioners, leading to fraud, misbehaviour, poor treatment, and data theft. Hence *maintaining record integrity* is a must. *Metadata tagging* would allow faster query time and make sure the DW is efficient. *Data ownership tagging* and usage rights would ensure proper governance and user authorization during access. *Dynamic data normalizing* will be helpful for records with continuous numerical values. *Cloud access* will aid in maintaining ease of access to the system globally. *Interoperability between clinical data standards (HL7, DICOM, etc.)* will facilitate smooth information transfer between health information systems.

03 - Archive Requirements. To reap the benefits of EHR breakthroughs, a healthcare data transfer plan that encourages a scalable health information system with *system migration capabilities* is essential. A tiered approach to data management with *multi-tier storage* with *standards based archiving* will ensure that all information about a patient is available within conventional clinical processes. In addition, *disaster management and data recovery* plan with backups are essential for healthcare institutes as downtime should be minimized.

04 - Sharing, Reporting and Analysis Requirements. Specific *data views based on job role* would ensure only necessary data is available and in easy to access form that is already preferred by the user. *Side-by-side comparison* of reports for physicians will aid in identifying differences in one or more reports and understanding the disease's progress. *Geo-location-based disease analysis* is important in understanding disease risk factors, incidence, and consequences, according to spatial epidemiology. *Restricting and alert if allergic drug diagnosed* can prevent life-threatening mistakes. *Drug side effect-based and community-based disease analysis* can aid in for better treatments.

05 - Security and Privacy Requirements. *Data anonymisation* safeguards a patient's right to privacy while also helping to prevent data breaches and identity theft. *User access control* is a critical component of data security because it maintains user rights so that legitimate users may only access data in the system that corresponds to their rights. *Encryption in data transport and data storage* is one of the most helpful data protection solutions. Even if attackers obtain access to the data, healthcare providers can make it more difficult for them to read patient information by encrypting data in transit and at rest.

3.2 Data Flow of Proposed Architecture

The DW will be protected with an IDPL, Intrusion Detection, and Prevention Layer. This layer will ensure the data warehouse is not attacked by hackers and other unauthorized users for attacks such as denial of service. At the stage of data input, all users are authenticated. Data input will then be funnelled through the MSIL, Malicious Software Identification Layer. This layer will scan all input files and ensure no malicious software or codes are entered into the data warehouse.

Next, the data will be transported through encrypted channels to the processing area. At this stage data will be cleaned, modelled, meta-data added, and standardized. User IDs and other personal identifiers of the data will pass through a de-identification phase and then all data will be funnelled through the anonymisation layer. Technologies such as blockchain can be utilized in this layer to further secure the data. The anonymised data will then be encrypted and passed to the data marts for storage. Each data mart is linked with the central demographic data mart where all demographic data is stored.

At the level of data analysis or presentation, the data is first passed through a data filtration layer, which ensures to filter and release of data that is only intended for the specific user. All users have to pass through the previously mentioned IDPL and the user authentication layer before gaining access. In the architecture, each user or user group is defined with an anonymity level as shown below.

1. **Anonymity Level 01:** Full Access to own records. This level of anonymity is assigned to patients so that they could have access to any of their records. The data will be decrypted and de-anonymised.
2. **Anonymity Level 02:** Full Access to specific users' data. This level of anonymity is assigned to doctors so that they have access to all necessary data of the patients they treat. The data will be decrypted and de-anonymised.
3. **Anonymity Level 03:** Limited Access to specific users' data. This level of anonymity is assigned to nurses so that they have access to only the necessary data of the patients they are assigned with. The data will be decrypted and de-anonymised.
4. **Anonymity Level 04:** Limited Access to specific anonymised data. This level of anonymity is assigned to clinical researchers and other staff who are authorized to access data that are anonymised and given consent by the data owner. The data will be decrypted but not de-anonymised.

The data warehouse has a specific data access mechanism granted by patient consent to access specific sets of their data for globally recognized emergency treatment centres. This will be similar to a Digital Emergency Medical Card (DEMC). For example, the allergies of a patient, critical conditions, specific surgery details, and other necessities of a patient will be made available here. An emergency can occur in any part of the World, at this stage, it is unsure if the patient will be responsive or unresponsive. Hence a new mechanism of secure three-factor user authorization is proposed to enable access to this data. Emergency Treatment Units (ETUs) globally can pre-register with the platform. The individual can enter their travel plan to the platform which will ensure that the ETUs in the area have access to the DEMC. In addition, a patient-related access control such as; a smart card, bio-metric or similar authentication mechanism can authenticate access on behalf of the individual even at times when he/she is unresponsive. This would ensure the relevant critical medical information is accessible throughout their journeys, at times of emergencies and this mechanism will be a breakthrough feature in the proposed clinical data warehouse framework. A usual data warehouse focuses on secure data storage and efficient analysis, but when it comes to a clinical data warehouse it is essential to aid in efficient treatments and improving healthcare. Stored and secured data, if not possible to be accessed during a patient's last breath, to save his/her (the data owner's) life, might be the most secure storage solution but would be an inefficient technology in an industry such as healthcare. The proposed architecture solves this critical issue in clinical data warehousing.

4 System Architecture

The proposed clinical data warehouse adopts a distributed big data architecture as clinical data is distributed in its nature and contains both structured and unstructured data. Clinical data is stored on various platforms in the healthcare industry. Each of these follows different formats, some even following proprietary formats. Therefore, it is a must for the proposed data warehouse to allow cross-platform standardization. These include interoperability with clinical data standards such as HL7, and DICOM. As most individuals travel the World nowadays, their clinical records should be given access at any given time from any geographic location. Therefore, it is most efficient to use a cloud-based data warehouse ensuring easy access to medical records globally. Rather than a healthcare centre governing the data, each individual should be given ownership to govern their own clinical data. This ensures that the patient itself can choose which data to be shared and when. Anonymous data can be made available for research and analysis purposes and patients can electronically provide consent to anonymise and use their data for research and analysis purposes.

To achieve this the architecture shown in Fig. 1 ensures to enforce data protection at all stages of the data life-cycle; data input, processing, storage, and analysis or presentation. Section 5 explains each component of the DW architecture in terms of data, security, and privacy.

5 Data, Security and Privacy Components

Even though data security and data privacy are connected, they serve different goals and require separate countermeasures. Confidentiality, integrity, availability, and non-repudiation are all goals of data security. In other words, data security ensures that only authorized people have access to data. While the primary goal of data privacy is to prevent shared data from revealing sensitive information about data owners. Data privacy rules govern this obligation. Due to the importance of the quality of shared data, privacy preservation is a difficult issue. The proposed DW architecture ensures data security and data privacy by using the following components.

Central Data Components

- **Data Processing Layer.** This is a key layer of the DW that takes care of functions such as data cleaning, data modelling, adding meta-data, and data standardization. These are key components of any DW.
- **Data Marts.** It is these data marts that collectively form the distributed cloud DW-based encrypted system. Each data mart is linked with the demographic data mart which stores all individual identifier-related data.

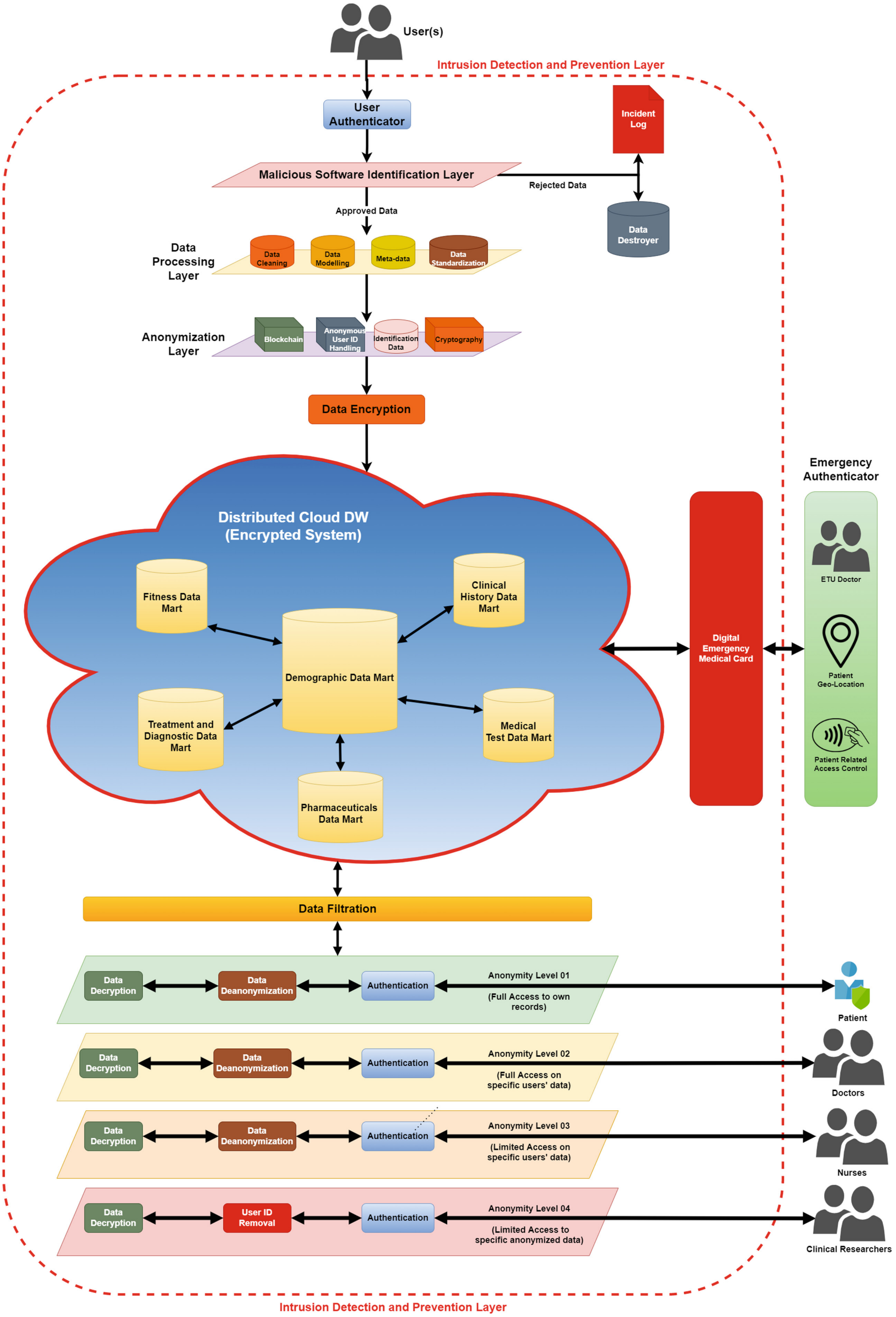

Fig. 1. Proposed Data Warehouse Architecture.

- **Digital Emergency Medical Card.** This is a key component that facilitates the treatment of patients that undergo accidents and other emergencies. This is a collection of all data that should be accessed by an emergency treatment centre.

Data Security Components

- **User Authenticator.** This ensures that only authorized users are given access to the DW.
- **Malicious Software Identification Layer.** This component runs a malware scan on all files that are entered into the DW to ensure there are no malware attacks.
- **Incident Logs.** Any data detected as malware is logged on the incident log with information about the user/IP address which entered the data.
- **Data Destroyer.** Data identified as malware are destroyed/removed using this component
- **Data Encryptor/Decryptor.** The encryptor interface has methods for performing typical encryption, random number generation, and hashing operations on data content. The decryptor restores the original form of the encrypted data.
- **Encrypted Data Transport Channels.** In addition to content encryption, transport encryption ensures all data transport channels are secure from third-party attacks.
- **Emergency Authentication.** This ensures the digital emergency medical card, when accessing at a time patient is non-responsive by only passing three levels of authentication, ETU doctor, patient geo-location, and patient-related access control like a smart card.
- **Intrusion Detection and Prevention Layer.** This is the process of monitoring and eliminating events in the DW regarding malicious software and cyber attacks.

Data Privacy Components

- **Anonymisation Layer.** The anonymisation layer is the key layer concerning data privacy control of the DW. This uses technology such as blockchain, anonymising user IDs, de-identification, and performing other cryptographic operations to ensure data privacy.
- **Anonymity Level based De-anonymisation.** As each user is granted access based on anonymity levels, this component ensures that only the relevant data is de-anonymised for the respective user.
- **Data Filtration.** This layer ensures only relevant data can be accessed as per the authorization given to each user.
- **Anonymity Level based Secure Views.** Each user is given their own secure view based on the anonymity level assigned by using this component.

6 Evaluation

In terms of evaluating the proposed architecture, we will use a case study to showcase how the proposed system would be an effective solution in the healthcare industry.

Person X is a patient with a critical heart condition, that requires special attention during any surgery performed. In addition, X has severe allergic reactions to some drugs. X is from Europe and travels on a tour to a different continent. X meets with an accident, is found unconscious, and gets airlifted to the nearest hospital. The doctors diagnose that his condition is critical and needs to undergo immediate surgery.

If X's medical data were not available to be accessed by the doctors at the emergency treatment centre as most typical clinical data warehouses do not facilitate secure data sharing, but rather are kept within the healthcare facility, the doctors would not be aware of the critical heart condition of X nor his allergies. A higher possibility is doctors start the open surgery immediately but with a lower success rate as the previous critical heart conditions would come as a surprise. Similarly, if an allergy-based drug is diagnosed not knowing it can be life-threatening. Therefore, the chances of survival for X might be reduced. This would be the case with a general clinical data warehouse.

The proposed architecture is equipped with an emergency authenticator component that specializes in rapid multi-factor authentication to access critical data for global emergency treatment centres. Hence the doctors would have access to the necessary data of X which can be carefully reviewed before the surgery. Therefore, the clinical staff and doctors would be aware of the current conditions and allergies and allow them to prepare as necessary for a successful surgery without surprises.

The clinical DW architecture that is being presented would guarantee that the necessary and urgent medical information is available during the patient's travels and in an emergency. This mechanism will be a ground-breaking feature. When it comes to a clinical data warehouse, it is crucial to support effective therapies and enhance healthcare. Unlike a typical data warehouse, which focuses on secure data storage and efficient analysis. The most secure data storage option might not be suitable in a sector like healthcare if the data would not be able to be retrieved during a patient's final breath to save his or her (the data owner's) life. The suggested architecture includes a specialized data access mechanism and emergency authorization system to address this important clinical data warehousing problem.

Table 1 shows the feature-wise comparison of the proposed clinical DW architecture with existing state-of-the-art. The feature list was developed based on the identified key requirement of clinical DWs and their stakeholders.

The Table 1 further proves that the proposed clinical DW architecture addresses all the privacy and security requirements of healthcare data.

In terms of security, the intrusion detection, and prevention layer, user & emergency authenticator, malicious software identification layer, and encryption ensures security-by-design DW architecture. Privacy is assured by a separate

Table 1. Feature comparison with existing DW architectures.

Features	[17]	[7]	[18]	[13]	Ours
Healthcare specific DW architecture	✓	✗	✓	✓	✓
Intrusion detection & prevention	✗	✗	✗	✗	✓
Malicious software identification	✗	✗	✗	✗	✓
De-identification of patient data	✗	✗	✓	✓	✓
Secure data encryption	✗	✓	✓	✓	✓
Data authentication for emergency treatment	✗	✗	✗	✗	✓
Anonymity level based secure views	✗	✗	✓	✗	✓

anonymisation layer that integrates blockchain, anonymous user IDs, and the de-identification of data. In addition, secure anonymity level-based views are added to ensure a privacy-by-design DW architecture.

The DW architecture proposed is of a distributed system and supports the naturally distributed nature of clinical data sets. Data mart distribution is based on the identified categorization of clinical data sets such as demographic data, medical test data, pharmaceutical data, etc. Data sharing and analysis components of the DW are made possible by having specific secure views based on the needs and requirements of the particular researcher. In addition, the emergency authenticator component facilitates a multi-factor geo-location-based authentication to release critical data for emergency treatment centres at times when the client or patient is unresponsive. The deployment of the system can be cloud-based which would add ease in access from healthcare-related institutes globally and aid in facilitating central cloud access for heterogeneous clinical data from multiple sources.

7 Discussion and Conclusion

Clinical data is distributed in nature itself. Therefore, according to identified requirements of clinical data warehousing distributed storage is a must. Distribution of the data can happen from different perspectives, e.g. sensitivity of data, location of a healthcare facility, etc. Clinical data happens to be both in the form of structured and unstructured data of multiple formats. Additionally, support for multiple schemas would be necessary. Therefore, the data warehouse architecture should support high-end privacy of the data, high scalability, high traceability, and data distribution.

Clinical data would need specific security and privacy measures to be taken. It would be an added challenge to efficiently distribute the data as most data is kept solely by different healthcare centres/facilities. A streamlined process would need to be adapted to better facilitate data sharing across facilities for efficient treatments. Therefore, automated data anonymisation is essential while ensuring privacy and cross-compatibility of the data. To recapitulate, healthcare big data warehouse architecture should be built which integrates security and privacy by design. Data safety and privacy, as well as the capacity to analyze and extract meaningful insights to enhance healthcare research and practices, are the major goals of building a healthy healthcare data warehouse.

The architecture developed was formulated focusing on the key components of data security and data privacy. A data component known as the 'Digital Emergency Medical Card' and a privacy and security component known as the 'Emergency Authenticator' was added to the data warehouse architecture. This ensures access to crucial medical data for doctors at emergency treatment centres globally. This access is granted using three levels of authentication, doctor user authentication, patient geo-location, and patient-related access control like a smart card. Security in all stages: data input, processing, transport, storage, and analysis was assessed, in developing a privacy-by-design clinical data warehousing architecture.

A prototype of the proposed architecture will be further developed in future works. It will then be tested based on the key identified features to ensure it meets the outlined requirements of a clinical DW architecture. Furthermore, we will enhance the proposed framework to increase scalability.

References

1. Aamot, H., Kohl, C.D., Richter, D., Knaup-Gregori, P.: Pseudonymization of patient identifiers for translational research. BMC Med. Inform. Dec. Mak. **13**(1), 1 (2013). https://doi.org/10.1186/1472-6947-13-75, BMC Medical Informatics and Decision Making
2. Ali, O., Ouda, A.: A classification module in data masking framework for Business Intelligence platform in healthcare. In: 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE IEMCON 2016, December 2016. https://doi.org/10.1109/IEMCON.2016.7746327
3. Chen, R., Fung, B.C., Mohammed, N., Desai, B.C., Wang, K.: Privacy-preserving trajectory data publishing by local suppression. Inf. Sci. Inform. Comput. Sci. Intell. Syst. Appl. Int. J. **231**, 83–97 (2013). https://doi.org/10.1016/J.INS.2011.07.035
4. Crosby, M., Nachiappan, Pattanayak, P., Verma, S., Kalyanaraman, V.: Blockchain Technology. Technical report (2015). http://www.blockchaintechnologies.com/blockchain-definition
5. David, J.: The 10 Biggest Healthcare Data Breaches of 2019, So Far (2019). https://healthitsecurity.com/news/the-10-biggest-healthcare-data-breaches-of-2019-so-far
6. Guerra-balboa, P., Pascual, M., Parra-arnau, J., Forn, J., Strufe, T.: Anonymizing Trajectory Data : Limitations and Opportunities (2013)
7. Gupta, S., Jain, S., Agarwal, M.: DWSA: A secure data warehouse architecture for encrypting data using AES and OTP encryption technique. Adv. Intell. Syst. Comput. **742**, 505–514 (2019). https://doi.org/10.1007/978-981-13-0589-4_47/COVER, https://link.springer.com/chapter/10.1007/978-981-13-0589-4_47
8. Haleem, A., Javaid, M., Singh, R.P., Suman, R., Rab, S.: Blockchain technology applications in healthcare: an overview. Int. J. Intell. Netw. **2**(May), 130–139 (2021). https://doi.org/10.1016/j.ijin.2021.09.005

9. Kalio, Q.P., Nwiabu, N.D.: A framework for securing data warehouse using hybrid approach. Int. J. Comput. Sci. Math. Theory **5**(1), 44–55 (2019). https://www.iiardpub.org
10. Kayaalp, M.: Modes of De-identification Modes of De-identification, April, 2018
11. Kumar, G.K.R., Rabi, B.J.D., Manjunath, T.N.: A study on dynamic data masking with its trends and implications. Int. J. Comput. Appl. **38**(6), 19–24 (2012). https://doi.org/10.5120/4612-6828
12. Mackey, T.K., Kuo, T.T., Gummadi, B., Clauson, K.A., Church, G., Grishin, D., Obbad, K., Barkovich, R., Palombini, M.: 'Fit-for-purpose?' - challenges and opportunities for applications of blockchain technology in the future of healthcare. BMC Med. **17**(1), 1–17 (2019). https://doi.org/10.1186/s12916-019-1296-7
13. Mia, M.R., Hoque, A.S.M.L., Khan, S.I., Ahamed, S.I.: A privacy-preserving national clinical data warehouse: architecture and analysis. Smart Health **23**(2021), 100238 (2022). https://doi.org/10.1016/j.smhl.2021.100238
14. Murphy, S.: Data Warehousing for Clinical Research, pp. 679–684. Springer, US, Boston, MA (2009). https://doi.org/10.1007/978-0-387-39940-9_120
15. Rahmani, A., Amine, A., Hamou, R.M.: De-identification of health data in big data using a novel bio-inspired apoptosis algorithm. Int. J. Organ. Collective Intell. **5**(3), 1–15 (2015). https://doi.org/10.4018/ijoci.2015070101
16. Santos, R.J., Bernardino, J., Vieira, M.: A data masking technique for data warehouses. In: ACM International Conference Proceeding Series, May 2014, pp. 61–69 (2011). https://doi.org/10.1145/2076623.2076632
17. Sebaa, A., Chikh, F., Nouicer, A., Tari, A.K.: Medical big data warehouse: architecture and system design, a case study: improving healthcare resources distribution. J. Med. Syst. **42**(4), 1–16 (2018). https://doi.org/10.1007/s10916-018-0894-9
18. Shahid, A., Nguyen, T.A.N., Kechadi, M.T.: Big data warehouse for healthcare-sensitive data applications. Sensors **21**(7) (2021). https://doi.org/10.3390/s21072353
19. Shukla, A., Sahni, M.K., Aggarwal, S., Rai, B.K.: Real-time De-identification of Healthcare Data, April 2018 (2018)
20. Sravanthi, K., Reddy, T.S.: Applications of big data in various fields. Int. J. Comput. Sci. Inf. Technol. **06** (2015). https://www.ijcsit.com
21. US Department of Health and Human Services: Health Insurance Portability and Accountability Act of 1996 (HIPAA)—CDC (1996). https://www.cdc.gov/phlp/publications/topic/hipaa.html
22. Wolford, B.: What is GDPR, the EU's new data protection law? - GDPR.eu. https://gdpr.eu/what-is-gdpr/