Scientific Research Publishing

Int. J. Communications, Network and System Sciences, 2022, 15, 111-125
https://www.scirp.org/journal/ijcns
ISSN Online: 1913-3723
ISSN Print: 1913-3715

Addressing the Security Challenges of Big Data Analytics in Healthcare Research

Mohamed Sami Rakha, Lucas Lapczyk, Costa Dafnas, Patrick Martin

Queen's University, Kingston, Canada
Email: martin@cs.queensu.ca

How to cite this paper: Rakha, M.S., Lapczyk, L., Dafnas, C. and Martin, P. (2022) Addressing the Security Challenges of Big Data Analytics in Healthcare Research. *Int. J. Communications, Network and System Sciences*, **15**, 111-125.
https://doi.org/10.4236/ijcns.2022.158009

Received: April 15, 2022
Accepted: July 31, 2022
Published: August 3, 2022

Abstract

Big data and associated analytics have the potential to revolutionize healthcare through the tools and techniques they offer to manage and exploit the large volumes of heterogeneous data being collected in the healthcare domain. The strict security and privacy constraints on this data, however, pose a major obstacle to the successful use of these tools and techniques. The paper first describes the security challenges associated with big data analytics in healthcare research from a unique perspective based on the big data analytics pipeline. The paper then examines the use of data safe havens as an approach to addressing the security challenges and argues for the approach by providing a detailed introduction to the security mechanisms implemented in a novel data safe haven. The CIMVHR Data Safe Haven (CDSH) was developed to support research into the health and well-being of Canadian military, Veterans, and their families. The CDSH is shown to overcome the security challenges presented in the different stages of the big data analytics pipeline.

Keywords

Big Data, Analytics Pipeline, Security, Data Safe Haven, CIMVHR, Health Data, Data Repository, Restricted Data Environment

1. Introduction

Over the last decades, the amount of data that has been transferred and collected by organizations in the healthcare domain has substantially increased [1]. Healthcare data is currently stored across multiple distributed sources, including local hospital and clinic databases as well as central government databases. While this approach provides individual organizations with control of their data in terms of accessibility, security, and privacy regulations, it presents obstacles to researchers seeking to exploit the data by carrying out more comprehensive studies on

data sets formed by linking and combining data from the diverse data sources [2].

One obstacle is the heterogeneity of the data, which forces researchers to deal with data of various types and storage formats. A second obstacle is the volume and variety of the data requiring researchers to access advanced analytic tools and significant computational and storage resources. A third obstacle is the strict security and privacy constraints that must be enforced by the owners of health data [3]. As a result, acquiring access to health data by researchers becomes a problematic or prohibitive process that can entail substantial administrative overhead related to ethics board and possibly government ministry approval [4] [5].

Big data and associated analytics, which offer new tools and methods to address the first two of these obstacles faced by researchers, have been described as some of the most powerful transformative forces affecting healthcare today [6]. They have the potential to revolutionize healthcare in various ways, such as the more routine practice of evidence-based medicine, which leads to better decision-making in patient care; the creation of new, personalized approaches to medicine based on both genomics and vastly improved data availability, and the development of new and innovative strategies for improved chronic illness and disease management. Big data analytics provide researchers with opportunities to solve new or existing problems by discovering new patterns, trends, and relationships. Addressing the security and privacy challenges related to big data analytics is an essential step towards overcoming the third obstacle stated above.

One approach to effectively managing and exploiting big data in the health domain is through the creation of "Data Safe Havens" [1] [7] [8]. A data safe haven is a repository in which valuable but potentially sensitive data is maintained securely under governance and informatics systems suited to the nature of the data and how it is being utilized [1]. Data safe havens provide researchers with hardware and software support for advanced big data analytics as well as a secure environment for the transport, storage, use, and maintenance of the data.

The paper has two main objectives. The first objective is to examine the security challenges related to big data analytics in healthcare research from a new perspective, namely the location of these challenges within the big data analytics pipeline. Based on this perspective, we can gain new insights into how to meet the challenges. The second objective is to discuss how the challenges can be overcome using a data safe haven approach for big data analytics.

The remainder of the paper is structured as follows. Section 2 presents a general model of the big data analytics pipeline. Section 3 introduces the security challenges in big data analytics for healthcare research and locates them relative to each pipeline stage. Section 4 introduces the data safe haven approach to managing health data and then describes the structure and security mechanisms of a specific system that was developed to support research into military and Veteran health in Canada. Section 5 discusses how the data safe haven approach

can overcome the security challenges of big data analytics. Section 6 describes related work, and Section 7 concludes the paper.

2. Big Data Analytics

Big data consists of extensive datasets, primarily in the characteristics of volume, variety, velocity, and/or variability, requiring a scalable architecture for efficient storage, manipulation, and analysis [9]. The ongoing transition to big data in the health domain drives a fundamental change in the type of analytics possible. Big data analytics emphasize computation value across the entire dataset, which gives researchers better chances to determine causation rather than just correlation. Discovering the cause, in turn, aids researchers in changing a trend or outcome.

A model of a big data analytics pipeline is shown in Figure 1. The pipeline flows through several stages, and each stage moves data in and out of the underlying **Data Storage**. The stages in the pipeline model are as follows:

- **Data Collection**: Data is collected for analysis from one or more sources and may be in various formats.
- **Data Linking**: Data from multiple sources may be linked at the record level before analysis to combine all attributes related to a single logical entity.
- **Data Transformation**: The data is prepared for analysis, which can involve a number of tasks, including data cleaning, data conversion, and reformatting, and the derivation of new attributes from those present in the data.
- **Data Modelling**: Data mining and machine learning techniques like clustering, classification, association, and deep learning are applied to the data to create predictive models.

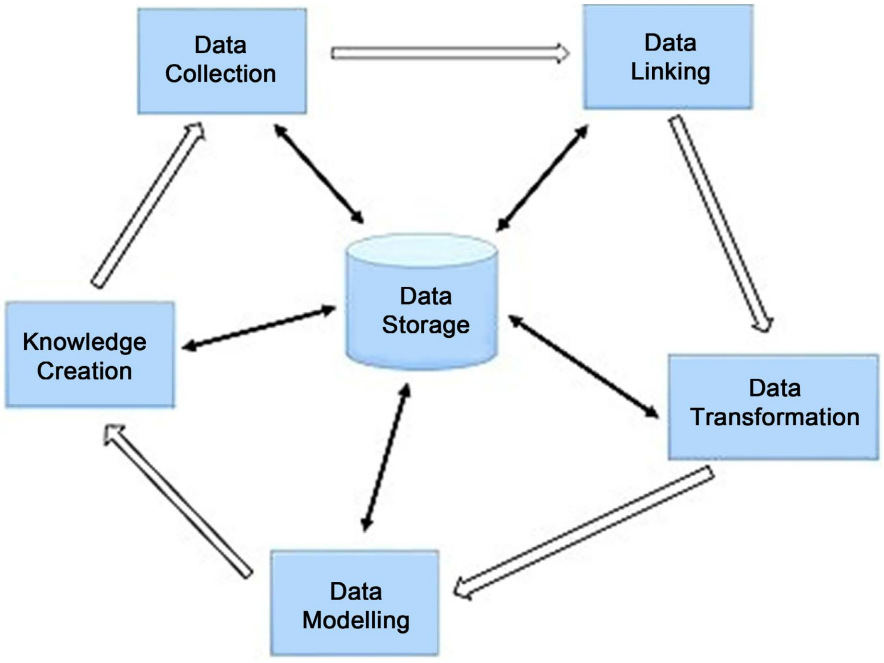

Figure 1. Big data analytics pipeline.

- **Knowledge Creation**: The models are used to generate new knowledge and inform decision-making.

3. Security Challenges

The use of big data in healthcare research presents several security challenges. Previous surveys of the security challenges in big data analytics, such as those by Abouelmehdi *et al.* [3] and de la Torres *et al.* [10] discuss big data analytics at a high level and try to identify a set of security or privacy techniques that can be applicable at that level. Our examination below provides a new perspective by going deeper into the big data analytics pipeline and looking at the challenges present in each stage of the pipleline.

The two main objectives of the security mechanisms employed in an implementation of a big data analytics pipeline in the healthcare setting are the following:

- To ensure that only permitted operations are performed with the data by authorized users.
- To ensure the privacy of patients whose data is being analyzed.

Implementations of the pipeline, which attempt to achieve the objectives through some combination of security mechanisms and security policies, face the following challenges:

- **Data Storage**: The data storage must protect data from unauthorized access and operations.
- **Data Collection**: The data collection stage must protect patient privacy when interacting with data sources and during data transmission.
- **Data Linking**: The data linking stage must provide a secure environment for linking and preserve patient privacy during and after linking.
- **Data Transformation**: The data transformation stage must provide a secure environment to ensure the safety of the data and preserve patient privacy during transformation.
- **Data Modelling**: The data modelling stage must provide a secure environment for modelling operations.
- **Knowledge Creation**: The knowledge creation stage must ensure that insights and results extracted from the modelling protect patient privacy.

4. Data Safe Havens—Addressing the Security Challenges

As mentioned above, data safe havens are a popular approach to big data in healthcare research because they provide researchers with hardware and software support for advanced big data analytics as well as a secure environment for the transport, storage, use, and maintenance of the data [1]. Prior research has proposed several safe haven systems for health data [1] [7] [8]. These systems are mainly customized to specific use cases and data sources and are commonly not generalized to be used or licensed to external organizations. The shared health data usually undergoes anonymization and de-identification processes as an es-

sential step to protect the privacy and security of patients. This allows the data safe haven operators and the original data owners to maintain data privacy requirements.

While the existing data safe havens tend to support similar objectives and functionality, there is not a commonly accepted architecture or structural model for these systems. Therefore, in the remainder of the paper, we examine how our implementation of a data safe haven for military and Veteran health research addresses the security challenges of big data analytics.

CIMVHR Data Safe Haven

The Canadian Institute for Military and Veteran Health Research (CIMVHR) [11] supports research projects that aim to improve the health and well-being of Canadian military members, Veterans, and their families. The CIMVHR Data Safe Haven (CDSH) was developed to support these researchers by providing a secure repository and analytics platform for data acquired and held by individual research projects and their affiliated organizations or institutions. The CDSH is the first data safe haven built for health data of Canadian military, Veterans, and their families. It is housed at the Queen's University Centre for Advanced Computing (CAC) [12], so all data in the safe haven remains within Canada. The CAC operates a high security, high availability data centre specializing in secure advanced computing resources and support for academic and medical clients. A discussion of the rationale and objectives of the CIMVHR data safe haven project is given elsewhere [13].

When a project is approved to use the CDSH, agreements are put in place to transfer the project's de-identified data into the safe haven and to establish the working environment required by the research team members to conduct their analytics work. Each project has its workspace within the CDSH, and data is not shared between project workspaces unless specifically authorized by the project leaders. The CDSH administrator assigns a user login identifier and access privileges to each research team member. The project leader specifies users' access privileges to data within a project workspace. Within each project, one or more users are designated with a *data manager* role and given the authority to approve and perform data import and export from the workspace.

The architecture of the CDSH is shown in **Figure 2**. The CDSH provides researchers, and administrators access using Remote Desktop Protocol (RDP) [14] via a secure virtual private network (VPN) connection [15] to the internal network of the safe haven through a firewall [16]. All the VPN connections are authenticated and authorized by their respective VPN servers (VPN Servers with a key symbol in **Figure 2**). The firewall strictly controls traffic to and from the internal network of the CDSH and ensures only authorized users are allowed in the CDSH. After authentication, users are presented with a remote desktop giving access to the databases, applications, and services for which they are authorized. The data for a project is not accessible to users of the CDSH beyond the

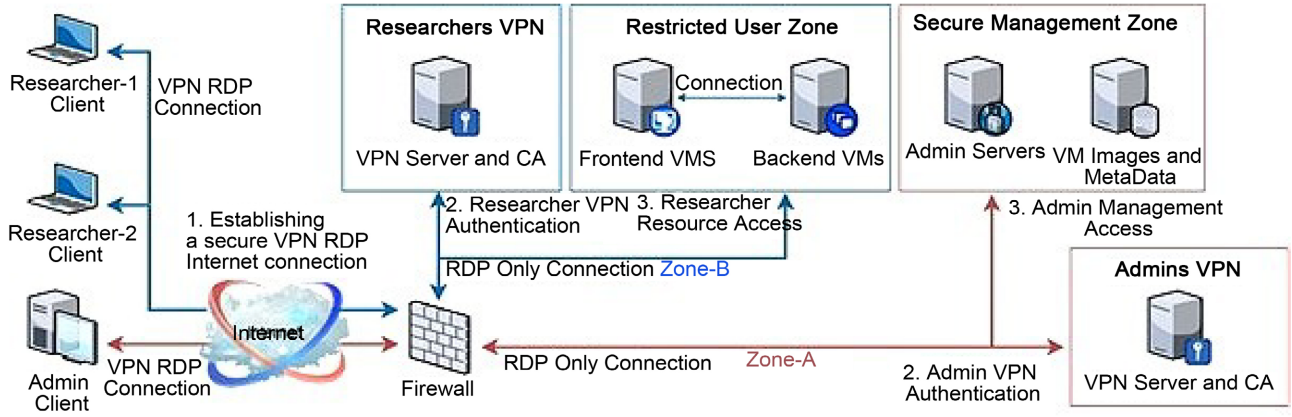

Figure 2. CDSH architecture.

members of that project unless explicit data sharing agreements are established. CDSH administrators are also able to enforce different levels of restrictions on data access within a project.

The CDSH components run on multiple VMs. A backend VM hosts the databases and analytic software (Backend VMs with document symbol in **Figure 2**), and a frontend VM hosts the user interfaces (Frontend VMs with monitor symbol in **Figure 2**). The system is built using both open source and commercial software. The CDSH currently provides users with a secure login to a remote desktop on Windows Server [17] and gives them the choice of using SAS [18], Cognos Analytics [19], Python [20], and R [21] to perform analytics and machine learning experiments. Additional software can be easily included if needed by a research project. A federated database is implemented in the backend VMs using IBM DB2 [22], providing access to various data sources, including databases and structured files.

The internal network of the CDSH is segmented into two zones. Zone-A (red connectors in **Figure 2**) includes the Secure Management Zone and the Admin VPN Servers. Zone-A network contains systems only accessible by the administrators. Such systems have virtualization infrastructure, as well as images of Virtual Machines, that are required for deployment [23]. This zone also includes a set of administration servers and tools for user management, access configurations, and system auditing. The VPN servers in this zone manage the VPN accessibility of the authenticated administrators.

Zone-B (blue connectors in **Figure 2**) includes Restricted User Zone and VPN Access Servers. Zone-B is the network space where systems accessible by the researchers are hosted. The systems include front-end virtual machines, which allow researchers to log in and submit batch jobs, as well as back-end virtual servers, which perform the analytic processing and health data hosting. Zone-B is the only zone accessible to the researchers, and their level of access is further restricted to specific services. Researchers can only access particular IP addresses on preset TCP/IP ports from their VPN account, which enhances security by precluding researchers from connecting to machines or services outside their

interests or authority.

The CDSH implements network segregation between the different zones [24]. This ensures resources that require different levels of access are well-isolated and controlled. Network segregation also helps prevent accidental disclosure and, therefore, potential breaches. In particular, because the networks are segregated, Researchers should never be found inside the Administrator's area, *i.e.*, the Secure Management Zone.

Since the CDSH is intended to accept internet connections, it employs Cybersecurity [25] concepts in the system setup. Generally, cybersecurity aims to safeguard the well-known "CIA Triad", that is, Confidentiality, Integrity, and Availability, of networks, systems, and data [26]. Security incidents occur because of oversights or misconfigurations related to any one of the three properties. The internal network of the CDSH, which is referred to as the *Restricted Data Environment Network* (*RDEN*), is designed with a "defense-in-depth" strategy [27] in mind and deploys security controls at multiple levels. The strategy ensures that, in the case of a failure of the primary controls, other compensating controls are used as preventive, or at least detective, means to hinder potential breaches or attacks.

Each network zone is built on an isolated subnet using separate physical network devices (e.g., switches). The physical isolation of subnets has certain advantages over logical separation, such as Virtual Local Area Networks (VLANs) [28]. The physical devices are only interconnected via the network firewall, allowing firewall-level security to manage the two separate subnets or zones.

To build a secure remote connection to the system, referred to as VPN/RDP connection in **Figure 2**, users connect via encrypted VPN tunnels between their devices and the network firewall. Besides, connections from users' endpoints to their RDEN hosts have a second level of encryption via the Remote Desktop Protocol (RDP) [29] and its built-in TLS encryption [30].

The implemented network firewall is a physical device with a dedicated network interface connected to each of the zones. This avoids the use of VLANs for network segregation and improves the overall security posture. By default, the machines on the RDEN Restricted User Zone (See **Figure 2**) are blocked from Internet access and from accessing any resources unless explicitly allowed. The Secure Management Zone is allowed to access the Internet and hosts in the Restricted User Zone. The contacted hosts are authorized to communicate back, provided that the traffic has been initiated by the Secure Management Zone and are established in the firewall states table. Both of the zones are allowed to contact the firewall for three services: echo request (or ping), DNS (UDP port 53) [31], and NTP (UDP port 123) [32]. Enabling this set of communications allows hosts in the zones to verify basic connectivity (ping the firewall), use a single point for DNS name resolution, and ensure that all the hosts have synchronized time. Outgoing rules from the Restricted User Zone require explicit specification of a destination IP address and port to precisely identify the target endpoint and

service that will be contacted. Administrators must implement any system change within a specific time window, whenever any outgoing rules are necessary. Outgoing rules would involve situations to enable software patching and updating, which require meticulous planning and procedures. For example, maintenance is handled outside of business hours because the users must be logged out, and VPN access is temporarily blocked. This countermeasure prevents a situation where users can be logged in and intentionally or not, establish unwanted outgoing connections.

RDEN is built upon the idea of "least privilege" [27], which means that users only gain the privileges necessary to carry out their work. This idea is further extended to installed systems and networks, as only a minimal subset of services and traffic rules is enabled to limit the attack space and avoid configuration errors. For example, if a researcher needs read-only access to a specific directory, they are only granted read access and only in that particular directory. Giving any unnecessary privileges could increase systems' exposure and vulnerability levels, which could lead to them being misused or abused.

To further enhance the level of security, the firewall is configured to log the traffic whenever any outgoing rule is enabled, whether it is to download data or to install or update software packages, any firewall system-level events, such as hardware change or system error or reconfiguration, and VPN authentication details. The logs are recorded locally as well as transported via SysLog [33] to a central Security Information and Event Management (SIEM) [34] server for analysis and long-term retention. In addition to the SysLog, NetFlow [35] traffic summaries are used to export information on traffic flowing through the firewall. The SIEM system is used to correlate events, such as intrusion alerts, suspicious flows, and failed or successful authentication events. That tool improves detection capabilities within the RDEN network. The logs and NetFlow traffic summaries are retained for one year by default. Having a second copy of logs at SIEM ensures that, in the case of a breach or failure of a researcher's systems, it is possible to see the chain of network communications and events leading up to the event.

The CDSH contains two separate VPN server instances installed on the firewall—one for Researchers and another for Administrators. The two instances are separate to avoid potential errors and enhance the level of security as, by default, Researchers never have the same implicit permissions as Administrators, who require a greater level of access to the network to implement changes, perform updates and resolve emerging issues. Researchers, however, are not provided with any implicit access other than DNS and NTP from the network firewall. The firewall creates a virtual subnet for each connected Researcher that only contains two host IP addresses. The first address is assigned to the firewall, and the second is given to the Researcher. Static IP addressing is implemented to ensure that a Researcher receives a specified IP address, based on which explicit firewall traffic and routing rules are defined.

VPN connections are certificate-based with X509 certificates [36] to identify and authenticate a user. Therefore, to address the need for two separate VPN instances, two separate Certification Authorities (CA) are run, one to supply and verify certificates for each VPN Server instance. Depending on the issued certificate, a user is authenticated by the Administrator or Researcher VPN server. Once a VPN connection is established, a user communicates with Microsoft Windows 10 or Windows Server 2016 VMs over the secure RDP connection.

A RDP connection provides users with various capabilities, including the ability to connect to remote systems and interact with a remote display on local machines, mount local disks to enable file transfers, do printer redirection and copy data between systems by using Clipboard Copy/Paste. Since these capabilities are not acceptable in a safe haven context, VMs deployed in RDEN have them disabled via Group Policy Objects (GPO) [29]. As a result, the risk of researchers copying sensitive contents, intentionally or not, is significantly mitigated. Furthermore, researchers must sign confidentiality agreements since there is still a risk of users taking screenshots of the displayed screen. Still, such a process must be intentionally malicious and requires significantly more time to acquire data.

In addition to all the confidentiality and integrity aspects, secure backups are available to research groups that require them. Secure backups are performed by a client software, configured to backup designated directories over an encrypted network session. This ensures that data is available in case of any disaster affecting the network. The backup systems store sensitive data on encrypted tape media, as it is secure, relatively low cost, and durable. The backup or archive data is encrypted in transit with TLS protocol [30] negotiated between backup client and server.

5. Security Evaluation

The CDSH addresses the security challenges in the different stages of the big data analytics pipeline, which were outlined in Section 3, as follows:

- **Data Storage**: Data is protected through two mechanisms. First, users must be authenticated before logging into the CDSH, and data access is controlled by the privileges assigned to the user by the system administrator. Second, further access control is provided by the backend database management system.
- **Data Collection**: Secure data collection is provided by secure connections like VPN/RDP, which employs multiple levels of encryption.
- **Data Linking**: Data linking can be performed within the secure environment of the CDSH using privacy-preserving approaches or through secure connections to a trusted third party to perform the linking [37].
- **Data Transformation**: The CDSH provides a secure and isolated environment with only administrator-approved tools for data transformation.
- **Data Modelling**: The CDSH provides a secure and isolated environment

with only administrator-approved tools for data modelling.

- **Knowledge Creation**: The CDSH provides mechanisms to control the export of data and results outside of the safe haven environment.

The quality of the security of a system is commonly evalauated based on a qualitative analysis of how the system handles common types of attacks [38] [39]. We therefore can further evaluate the security strength of the CDSH by performing a qualitative analysis of the system's ability to prevent the following common security attacks:

- **VLAN Hopping Attack** [40]: In this attack, hackers can escape their restricted VLAN to other networks. Thus, the attacker can access resources that were not supposed to be reachable.
  - **Prevention**: CDSH uses a separate physical switch for network zones (*i.e.*, Zone-A and Zone-B). The zones separation and physical network segregation prevent the VLAN Hopping attack. Hence, VLAN miss-configurations are not a threat as the VLAN concept is not leveraged for Zone-A or Zone-B.
- **Man-in-the-Middle (MITM) Attack** [41]: This attack occurs when an attacker begins to route legitimate user's traffic via an attacker's machine. The consequences could result in an attacker performing sniffing (i.e., eavesdropping on sensitive data in transit) or session hijacking (i.e., tampering with data in transit).
  - **Prevention**: Besides the physical segregation, MITM attacks are further thwarted by ensuring that CDSH users are assigned VPN addresses inside of/30 subnets. Subnets with/30 prefix have only two bits dedicated to their IP address space, and as a result, only two host IP addresses can exist there [28]—namely the user's and the firewall's IP addresses. This design ensures that a network firewall controls any communications to and from individual users' VPN accounts. Since that effectively results in users being placed in separate broadcast domains (one per each VPN user), it is not possible in this design to perform the "ARP poisoning" [42] necessary to run MITM attacks. To better illustrate this concept, in **Figure 2**, suppose Researcher-1 and Researcher-2 log into the system simultaneously via VPN. If Researcher-1 is assigned an address on subnet 192.168.0.0/30, he will be assigned address 192.168.0.2, and his gateway will be set to 192.168.0.1. Researcher-2, in such a scenario, will be given another/30 range and relevant configuration—meaning that he would be locked inside of 192.168.0.4/30 network with 192.168.0.6 as the IP address and 192.168.0.5 as his gateway. These two researchers cannot directly communicate with each other's machines as they reside on different subnets. Researcher-2 at 192.168.0.6 could only communicate with Researcher-1 at 192.168.0.2 or any other address on the network via the central firewall, as those clients reside on different isolated subnets. Consequently, since all communication has to go via the firewall, the firewall rules control all network traffic that is taking place within the network. Such a setup strengthens the security of the proposed system when compared to alternate designs.

- **Malware Attack** [43]: This attack occurs when malicious code enters the victim system and performs unauthorized harmful activities. Malware attacks often leverage network communication channels so that hackers can control their malicious activities.
  - **Prevention**: The CDSH system leverages an Intrusion Prevention System (IPS) [16] that is enabled on the firewall to block network attacks. The firewall performs IP list-based blocking and Deep Packet Inspection (DPI) on network packets' payload to recognize potentially malicious patterns.

6. Related Work

There are several research efforts to develop safe havens in the health domain. Ainsworth *et al*. [44] propose an electronic space eLab in which researchers can post and share their data, analysis scripts, and final results following the concept of Virtual Research Environments. eLab is mainly designed for the health domain with options to extend to other domains. Later, eLab became part of a safe haven managed by the Farr Institute Health eResearch Centre in North England [45].

Burton *et al*. [1] discuss how the meaning of the term "Data Safe haven" emerged over time. They show how that there is no formal definition of the term Data Safe haven. Instead, to resolve that confusion, they describe 12 criteria that may help in providing clarity to the term's meaning.

Robertson *et al*. [7] describe work underway to provide a clear and logical architecture of a data safe haven. The safe haven aims to maximize the benefits of data within jurisdictions to medical science and health care. They discuss a proposed architecture for two jurisdictions—one in Scotland and another in Italy. Robertson's system architecture follows seven fundamental principles and is divided into two levels. The first level of the safe haven includes data directly moved from the core system after some de-identification and only accessed by approved managers. The second level may consist of a subset of the data available in the first level that may be accessed by a group of approved researchers.

Vaccarino *et al*. [8] propose and implement a system called Brain-CODE as a safe haven for brain disorders data within Ontario, Canada. Brain-CODE provides the researchers with access to aggregated data sources that were isolated and not shared before. Brain-CODE uses linking mechanisms for provincial, national and international databases while ensuring the privacy and security of patients.

In contrast to the safe haven systems proposed by prior research, CDSH provides a novel design from perspectives such as network security, services, and analytical tools.

7. Conclusions

Big data and associated analytics have the potential to revolutionize healthcare research. They provide researchers with opportunities to solve new or existing

problems by discovering new patterns, trends, and relationships across the large and heterogeneous data sets maintained by various health organizations. Significant security and privacy challenges, however, remain to be overcome before this potential can be realized.

Data safe havens, which have been proposed to manage and exploit big data in the health domain effectively, are one approach to overcoming the associated security challenges. They can provide researchers with hardware and software support for advanced big data analytics as well as a secure environment for the transport, storage, use, and maintenance of the data.

The paper first identifies the security challenges related to the different stages of the big data analytics pipeline. The challenges identified derive from the two main objectives of a security mechanism, namely, to ensure that only permitted operations are performed with the data by authorized users and to ensure patients' privacy whose data is being analyzed.

The paper then presents the CIMVHR Data Safe Haven and the security features it provides. It is shown how these features address the security challenges presented by big data analytics in healthcare research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Burton, P.R., Murtagh, M.J., Boyd, A., Williams, J.B., Dove, E.S., Wallace, S.E., Tasse, A.-M., Little, J., Chisholm, R.L., Gaye, A., Hveem, K., Brookes, A.J., Goodwin, P., Fistein, J., Bobrow, M. and Knoppers, B.M. (2015) Data Safe Havens in Health Research and Healthcare. *Bioinformatics*, **31**, 3241-3248. https://doi.org/10.1093/bioinformatics/btv279

[2] Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S. (2019) Big Data in Healthcare: Management, Analysis, and Future Prospects. *Journal of Big Data*, **6**, Article No. 54. https://doi.org/10.1186/s40537-019-0217-0

[3] Abouelmehdi, K., Beni-Hessane, A. and Khaloufi, H. (2018) Big Healthcare Data: Preserving Security and Privacy. *Journal of Big Data*, **5**, Article No. 1. https://doi.org/10.1186/s40537-017-0110-7

[4] Price, W.N. and Cohen, I.G. (2019) Privacy in the Age of Medical Big Data. *Nature Medicine*, **25**, 37-43. https://doi.org/10.1038/s41591-018-0272-7

[5] Krishna, R., Kelleher, K. and Stahlberg, E. (2007) Patient Confidentiality in the Research Use of Clinical Medical Databases. *American Journal of Public Health*, **97**, 654-658. https://doi.org/10.2105/AJPH.2006.090902

[6] Daschle, T.A. (2015) Academic Medicine in a Transformational Time. *Academic Medicine*, **90**, 11-13. https://doi.org/10.1097/ACM.0000000000000577

[7] Robertson, D., Giunchiglia, F., Pavis, S., Turra, E., Bella, G., Elliot, E., Morris, A., Atkinson, M., McAllister, G., Manataki, A., Papapanagiotou, P. and Parsons, M. (2016) Healthcare Data Safe Havens: Towards a Logical Architecture and Experiment Automation. *The Journal of Engineering*, **2016**, 431-440. https://doi.org/10.1049/joe.2016.0170

[8] Vaccarino, A.L., Dharsee, M., Strother, S., Aldridge, D., Arnott, S.R., Behan, B., Dafnas, C., Dong, F., Edgecombe, K., El-Badrawi, R., El-Emam, K., Gee, T., Evans, S.G., Javadi, M., Jeanson, F., Lefaivre, S., Lutz, K., MacPhee, F.C., Mikkelsen, J., Mikkelsen, T., Mirotchnick, N., Schmah, T., Studzinski, C.M., Stuss, D.T., Theriault, E. and Evans, K.R. (2018) Brain-Code: A Secure Neuroinformatics Platform for Management, Federation, Sharing and Analysis of Multi-Dimensional Neuroscience Data. *Frontiers in Neuroinformatics*, **12**, Article No. 28. https://doi.org/10.3389/fninf.2018.00028

[9] Chang, W.L. and Grady, N. (2015) NIST Big Data Interoperability Framework: Volume 1, Big Data Definitions. NBD-PWG NIST Big Data Public Working Group.

[10] de la Torre, I., García-Zapirain, B. and López-Coronado, M. (2017) Analysis of Security in Big Data Related to Healthcare. *Journal of Digital Forensics, Security and Law*, **12**, 39-46. https://doi.org/10.15394/jdfsl.2017.1448

[11] CIMVHR Canadian Institute for Military and Veteran Health Research (2019). https://cimvhr.ca

[12] CAC The Centre for Advanced Computing (2020). http://cac.queensu.ca

[13] Martin, P., Rakha, M. and Whitnall, J. (2021) Data Safe Haven for Military, Veteran, and Family Health Research. *Journal of Military, Veteran and Family Health*, **7**, 102-107. https://doi.org/10.3138/jmvfh-2020-0035

[14] Microsoft (2018) Remote Desktop Protocol. https://docs.microsoft.com/en-us/windows/win32/termserv/remote-desktop-protocol?redirectedfrom=MSDN

[15] Venkateswaran, R. (2001) Virtual Private Networks. *IEEE Potentials*, **20**, 11-15. https://doi.org/10.1109/45.913204

[16] Surana, J., Singh, K., Bairagi, N., Mehto, N. and Jaiswal, N. (2017) Survey on Next Generation Firewall. *International Journal of Engineering Development and Research*, **5**, 984-988.

[17] Microsoft (2022) Microsoft Windows Server Documentation. https://docs.microsoft.com/en-us/windows-server

[18] SAS (2020) The SAS Platform. https://www.sas.com/en_ca/software/platform.html

[19] IBM (2020) Cognos Analytics. https://www.ibm.com/ca-en/products/cognos-analytics

[20] Python (2020) Python Programming Language. https://www.python.org

[21] The R Foundation (2020) The R Project for Statistical Computing. https://www.r-project.org

[22] IBM (2021) DB2 Supported Data Source. https://www.ibm.com/support/pages/data-source-support-matrix-federation-bundled-db2-luw-v115

[23] Morabito, R., Kjallman, J. and Komu, M. (2015) Hypervisors vs. Lightweight Virtualization: A Performance Comparison. 2015 *IEEE International Conference on Cloud Engineering*, Tempe, 9-13 March 2015, 386-393. https://doi.org/10.1109/IC2E.2015.74

[24] Padhy, R.P., Patra, M.R. andSatapathy, S.C. (2011) Cloud Computing: Security Issues and Research Challenges. *International Journal of Computer Science and Information Technology and Security*, **1**, 136-146.

[25] Von Solms, R. and Van Niekerk, J. (2013) From Information Security to Cybersecurity. *Computers and Security*, **38**, 97-102. https://doi.org/10.1016/j.cose.2013.04.004

[26] Samonas, S. and Coss, D. (2014) The CIA Strikes Back: Redefining Confidentiality, Integrity, and Availability in Security. *Journal of Information System Security*, **10**, 21-45.

[27] Conrad, E., Misenar, S. and Feldman, J. (2012) CISSP Study Guide. Newnes.

[28] Lammle, T. (2016) CCNA Routing and Switching Complete Study Guide: Exam 100-105, Exam 200-105, Exam 200-125. John Wiley and Sons, Hoboken.

[29] Dautis, B. (2018) Installing and Configuring Windows 10: 70-698 Exam Guide. PACKT Publishing Limited, Birmingham.

[30] Dierks, T. (2008) The Transport Layer Security (TLS) Protocol Version 1.2. https://tools.ietf.org/html/rfc5246 https://doi.org/10.17487/rfc5246

[31] Mockapetris, P. and Dunlap, K.J. (1988) Development of the Domain Name System. *Symposium Proceedings on Communications Architectures and Protocols*, Stanford, 16-18 August 1988, 123-133. https://doi.org/10.1145/52324.52338

[32] Mills, D.L. (1991) Internet Time Synchronization: The Network Time Protocol. *IEEE Transactions on Communications*, **39**, 1482-1493. https://doi.org/10.1109/26.103043

[33] Gerhards, R. (2009) The Syslog Protocol. https://tools.ietf.org/html/rfc5424 https://doi.org/10.17487/rfc5424

[34] Miller, D., Harris, S., Harper, A., VanDyke, S. and Blask, C. (2011) Security Information and Event Management (SIEM) Implementation. McGraw-Hill, New York.

[35] Claise, B. (2004) Cisco Systems NetFlow Services Export Version 9. https://tools.ietf.org/html/rfc3954 https://doi.org/10.17487/rfc3954

[36] Chokhani, S., Ford, W., Sabett, R., Merrill, C. and Wu, S. (1999) Internet x.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework. https://tools.ietf.org/html/rfc3647 https://doi.org/10.17487/rfc2527

[37] Christen, P. (2012) Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science and Business Media, Berlin.

[38] Elahi, G., Yu, E. and Zannone, N. (2011) Security Risk Management by Qualitative Vulnerability Analysis. *Proceedings of* 2012 *Third International Workshop on Security Measurements and Metrics*, Banff, 21 September 2011, 1-10. https://doi.org/10.1109/Metrisec.2011.12

[39] Clark, K., Tyree, S., Dawkins, J. and Hale, J. (2004) Qualitative and Quantitative Analytical Techniques for Network Security Assessment. *Proceedings of the* 2004 *IEEE Workshop on Information Assurance and Security*, West Point, 10-11 June 2004, 321-328.

[40] Daryabar, F., Dehghantanha, A., Norouzi, F. and Mahmoodi, F. (2011) Analysis of Virtual Honeynet and VLAN-Based Virtual Networks. *Proceedings of* 2011 *International Symposium on Humanities*, *Science and Engineering Research*, Kuala Lumpur, 24-27 June 2012, 73-77. https://doi.org/10.1109/SHUSER.2011.6008503

[41] Callegati, F., Cerroni, W. and Ramilli, M. (2009) Man-in-the-Middle Attack to the HTTPS Protocol. *IEEE Security and Privacy*, **7**, 78-81. https://doi.org/10.1109/MSP.2009.12

[42] Bull, R.L., Matthews, J.N. and Trumbull, K.A. (2016) VLAN Hopping, ARP Poisoning, and Man-in-the-Middle Attacks in Virtualized Environments. *DEF CON*

24, Las Vegas, 4-7 August 2018, 9.

[43] Khouzani, M.H., Sarkar, S. and Altman, E. (2012) Maximum Damage Malware Attack in Mobile Wireless Networks. *IEEE/ACM Transactions on Networking*, **20**, 1347-1360. https://doi.org/10.1109/TNET.2012.2183642

[44] Ainsworth, J., Cunningham, J. and Buchan, I. (2012) Elab: Bringing Together People, Data, and Methods to Enhance Knowledge Discovery in Healthcare Settings. *Studies in Health Technology and Informatics*, **175**, 39-48.

[45] Bechhofer, S., De Roure, D., Gamble, M., Goble, C. and Buchan, I. (2010) Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*. https://doi.org/10.1038/npre.2010.4626.1