data

MDPI

Article

Investigating the Adoption of Big Data Management in Healthcare in Jordan

Hani Bani-Salameh [1,*,†], Mona Al-Qawaqneh [1,†] and Salah Taamneh [2,†]

[1] Department of Software Engineering, The Hashemite University, Zarqa 13133, Jordan; bme_monaqawaqneh@hotmail.com
[2] Department of Computer Science and Applications, The Hashemite University, Zarqa 13133, Jordan; taamneh@hu.edu.jo
* Correspondence: hani@hu.edu.jo
† These authors contributed equally to this work.

Abstract: Software developers and data scientists use and deal with big data to easily discover useful knowledge and find better solutions to improve healthcare services and patient safety. Big data analytics (BDA) is getting attention due to its role in decision-making across the healthcare field. Therefore, this article examines the adoption mechanism of big data analytics and management in healthcare organizations in Jordan. Additionally, it discusses health big data's characteristics and the challenges, and limitations for health big data analytics and management in Jordan. This article proposes a conceptual framework that allows utilizing health big data. The proposed conceptual framework suggests a way to merge the existing health information system with the National Health Information Exchange (HIE), which might play a role in extracting insights from our massive datasets, increases the data availability and reduces waste in resources. When applying the framework, the collected data are processed to develop knowledge and support decision-making, which helps improve the health care quality for both the community and individuals by improving diagnosis, treatment, and other services.

Keywords: big data; big data analytics (BDA); healthcare; FODA; Hadoop; VistA; framework; CHISTAR; MapReduce

check for updates

Citation: Bani-Salameh, H.; Al-Qawaqneh, M.; Taamneh, S. Investigating the Adoption of Big Data Management in Healthcare in Jordan. *Data* **2021**, *6*, 16. https://doi.org/10.3390/data6020016

Received: 14 December 2020
Accepted: 31 January 2021
Published: 6 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Day by day, healthcare data are expanding. Data exponentially grow in number using health information systems and dealing with patients' records. Huge amounts of health big data are created every day. This opens a debate about how to draw meaning from this exponentially growing amount of data. The analysis of such data is important for extracting information, gaining knowledge, and discovering hidden patterns [1–3]. Additionally, analysis of health big data improves the quality of services and reduces costs [4].

1.1. Big Data Characteristics

Big data was defined by the Big Data Commission at the TechAmerica Foundation as "large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information" (TechAmerica Foundation, 2012). The characteristics of big data are defined by the V family. The three most-known Vs are volume, variety, and velocity [5,6]. Other Vs were recently added, veracity [7,8], visualization [9], variability [10], and value [11].

- Volume: the large volume of data that is collected/generated.
- Variety: the multiple types of data extracted from many sources. Big data in health care can be classified into three main types: structured, semi-structured, and unstructured [4].

- Velocity: the speed at which the big data are being collected, generated, and processed.
- Veracity: whether the data are trusted or not (the quality or reliability of data).
- Variability: refers to types of data whose meaning and dimensions are constantly changing.
- Visualization: a way to simplify the numerous and complex data by representing them in graphs and charts.
- Value: refers to the value of the generated/collected data.

1.2. Health Big Data

Big data are generated in different sources and sectors; healthcare is one of those domains. Health big data are generated from millions of patients' medical records which are collected and stored in an electronic format [12]. Health big data vary in sources and types. Data sources can be either human beings (doctors, nurses, or recorders) or sensors. Types of data can be classified as the following:

1. Structured (such as Logical Observation Identifiers Names and Codes (LOINC), and laboratory tests);
2. Semi-Structured (such as data generated by the conversion of paper records to electronic records and the readings of the instrument);
3. Unstructured (such as nurses' handwritten notes, data from SNSs, and medical images).

Health big data contain personal medical records, personal genetics, population data, patient's signals (ECG, EEG, EMG, etc.), test results, and medical images [10]. The best non-invasive method to evaluate a patient and determine whether a medical condition exists is medical imaging [13,14].

1.3. Big Data Analytics

Big data analytics means taking the big data collected in experiments and converting it to valuable data [15]. Without analysis, big data is not sufficient [16]. Thus, to get the benefits, we have to use both of them together, "big data and big data analytics." Big data analytics plays a role in extracting insights from massive data sets [17]. It increases the availability of data and analytic capabilities and reduces waste of resources; thus, it can improve health care cost-efficiency, quality, and outcomes [14]. Applying big data analytics lets value arise to develop actionable information and supports decision-making [6,18–22].

For almost a decade, the Hakeem healthcare program (https://ehs.com.jo/hakeem-program) has been adopted in the process of applying the electronic society in Jordan, and since then, data have been collected and stored in an electronic format from more than five million patients. In addition to that, private sectors started to convert their systems to electronic systems using different programs and applications.

Therefore, in Jordan, the healthcare system is growing fast and makes use of huge, varying, and sensitive data. The collected data carry all big data's characteristics, but without utilizing it or even naming it "big data." To get the benefits and insights from these data, it must be stored, managed, and analyzed efficiently.

This article proposes a framework that aims to make use of the generated medical big data in Jordan and benefits from analyzing it. It helps by making data more valuable and subject to study and analysis; that can be done by designing a future prediction and policy decision model that utilizes medical big data that comes from various sources. Therefore, it can help the community and support individuals with better healthcare services [23]. This article aims to give insights into the big data research state-of-art in Jordan. Thereafter, we study and analyze any previous attempts to realize big data in Jordan by answering two research questions:

RQ1: What is the role of big data in our health system? What are the challenges?
RQ2: How can one design an efficient, applicable solution for our health system using big data analytics?

The rest of this article is organized as follows. Section 2 provides an overview of the related studies. Additionally, it reviews the health big data frameworks, their analytics, and their challenges. Section 3 performs domain engineering to better understand the context of the framework. Section 4 represents the proposed conceptual framework that utilizes the collected healthcare data. The proposed framework's challenges are addressed in Section 5. The framework's design and integration processes are presented in Section 6. Finally, Section 7 presents the conclusions drawn from the study and proposes some recommendations for further research and study.

2. Background Research

There is a considerable amount of literature on health big data. This section provides a brief overview of the existing related work and studies. Previous work and studies are divided into two areas: first, work related to health big data that shows the advantages and benefits of using big data in healthcare systems; second, work related to health big data analytics and frameworks. Additionally, this section introduces a list of health big data challenges as mentioned by others.

2.1. Research Related to Health Big Data

Health big data provide multiple benefits. Kim et al. [12] mentioned that health big data can serve government citizens and address major national challenges involving healthcare. Feldman et al. [8] showed that health big data and analytics serve many goals, such as reducing costs, allowing better and effective care for patients, providing real-time data access and a transparent tool for the providers, improving tool quality for researchers, assisting medical companies in integrating data collected from hospitals and home devices for safety monitoring and better services, and reducing waste.

Many researchers believe that big data can improve both public and personalized healthcare [7,8,10,17,24]. It can help individuals to improve predictions, diagnosis, and treatment, and provide them with information regarding their health, which in turn will empower them, and help physicians make clinical decisions [24]. In terms of public health [8,10], the integration of medical records with health big data analytics tools can predict the occurrence of highly prevalent or deadly diseases. Additionally, it helps in the monitoring and tracking of such disease spreading. Wong et al. [14] proposed that the reported number of infected cases can be improved once adopting big data in public health surveillance.

Lee and Yoon [16] believe that the ability (1) to get information from the electronic health record in real-time, and (2) to meet the physicians' and researchers' needs at the point of care, are the major reasons for using big data in healthcare. ur Rehman et al. [9] mentioned other benefits of health big data, such as security in saving and sharing the data and improving the trust between patient and physician.

Hermon and Williams [11] followed Groves et al. [25] in dividing health big data usages into four broad categories; they also added a fifth category. The categories are administration and delivery, clinical decision support, clinical information, behavior/consumer, and support information.

2.2. Big Data Analytics and Frameworks

To address the benefits and challenges of health big data, it is important to develop new techniques and technologies which help overcome the challenges [26]. Traditional tools and techniques cannot handle big data issues [7,8,12,24]. This section presents studies and results that make use of big data analytics to help communities and researchers make better decisions.

Bahga and Madisetti [27] proposed a framework as a solution for healthcare data problems. CHISTAR (Cloud Health Information Systems Technology Architecture) is a way to reach semantic interoperability in the e-health world. It makes the health data

exchange easier. CHISTAR provides us with an integration engine that solves the data integration problem.

Demchenko et al. [28] focused on modern technologies' challenges. They defined the components of big data ecosystems (BDEs) and the challenges related to big data. Additionally, they proposed a big data architecture framework (BDAF) that would support the extended big data in all sectors, especially medical and healthcare big data. The proposed framework supports the main components and processes of BDEs.

Crawford and Schultz [29] proposed a framework focusing on big data reduction at the customer end. This framework aims to: (a) trust enhancement between enterprises and customers, (b) enable secure data sharing, (c) conserve the privacy of customers, (d) lower the service utilization cost, and (e) empower data sharing control to customers.

Madhavi and Ramana [30] used Hadoop and MapReduce to provide a solution that de-identifies personal health information. An advanced mechanism of using Map Reduce, UDF (User Data File), and Pig Queries was used in their application to be executed on the healthcare dataset and to protect the healthcare dataset. Additionally, they considered that Hadoop was suitable for de-identifying demographic and social data.

Sheriff et al. [31] proposed a framework that can be used as a reference for implementing a complete healthcare analytics ecosystem. They considered that big data, Complex Event Processing (CEP), and the Internet of Things (IoT)) are the bases of any healthcare analytics framework.

Moreover, Song and Ryu [32] provided a framework for healthcare and social sectors in Korea. The proposed framework was presented by applying it to the avoidance and detection of suicide in Korea. The authors mentioned several issues that affect the utilization of big data in healthcare and social factors.

Furthermore, Cunha et al. [33] proposed a generic functional architecture with Mahout and Hadoop framework for storing and analyzing health big data. They applied the architecture on Twitter health data.

Kavitha and Kannudurai [34] in their work focused on improving the quality of health big data analytics with Hadoop by improving MapReduce performance quality. They used the fuzzy c-means clustering algorithm, which allows one to avoid duplications and repetitions entering data in the health system, maintains the low cost of all kinds of health data, and provides the right care to patients efficiently. They believe that big data helps in giving great results, and in maintaining all types of health care analysis.

Moreover, ur Rehman et al. [9] focused on the idea of data reduction. They proposed a novel concept of big data reduction at the customer end and a framework that aims to reach that concept. The goals of this work are: lowering the costs, enhancing security and privacy, and enhancing the trust between customers and enterprises. Additionally, they mapped their big data reduction framework with various application areas including healthcare.

Roden et al. [35] conducted four studies to investigate the effect of using big data on the operation models of multi sectors. Then, they developed a tentative framework for big data changes on the operational models. Their framework shows different results according to the nature of big data.

Wang and Hajli [36] developed a model that links big data analytics capabilities with business value. Their model is based on the resource-based theory (RBT) and capability building view to explain how big data analytics capabilities can be developed and what potential benefits can be obtained using these capabilities in the health care industries.

Additionally, Wang et al. [37] developed a model that links explanatory variables to practices, then the benefits, and finally the business value. They call their model the big data analytics enabled transformation model (BDET). The model is based on practice-based-value (PBV) that links IT with healthcare to help practitioners and researchers in understanding how the interaction between the critical elements of practice and IT tools. The authors introduced the idea of big data impacting transformational practices in health care industries.

Many challenges are related to health big data and health big data analytics [38–41]. Challenges are technologies, collecting, transforming, and sharing, storing, analyzing, visualizing health data, security, and automation.

2.3. Challenges of Health Big Data

Based on the existing studies and frameworks, this section presents a set of challenges that are related to health big data and health big data analytics. Challenges belong to technologies, collection, storing/storage, aggregation, analysis, sharing, and visualization of health data [5,17,42–44]. None of the previous works merged their existing systems with the big data analytics system or the National Health Information Exchange. Section 4 presents the proposed conceptual framework and its challenges. The following are brief descriptions of the challenges as addressed in the literature and previous studies.

- Technologies: The challenges include hardware tools (e.g., sensors and computers as data collection and storage devices), software tools (e.g., data collection and processing software), and the infrastructure (e.g., internet access (communications) and power supply) [43,45].
- Collection: The data collection challenges involve the way of collecting and transferring vast amounts and different types of data to the data repository. Collection methods include both devices (sensors and computers) and human resources. Data transferring to the repository is related to network challenges [28,46].
- Transforming and Sharing: Data transformation includes different stages, levels, and data structure forms. At the lower level (first stage), the raw data need to be transformed from the collection devices to the data repository, where one of the biggest challenges is due to integration limitations and lack of system interference capabilities, which may require special data adaptation tools. In later stages, the valuable data will be transformed out of the repository based on certain queries [28,46]. To share such data, a high network bandwidth capacity is needed which is considered a challenge in this phase.
- Storing: Once the data are collected, another challenge arises—the storage capacity. The huge amount of data needs large space and high-performance I/O devices, which is costly.
- Analysis: The analysis is required to convert the raw data into valuable data. In big data analysis, data must be scalable, well-structured, secure, and consistent. The challenges are both data pre-processing and the lack of experts in the analysis tools.
- Visualization: Health big data faces another challenge which is how to visualize such huge, varied, and different structures of valuable data [9,47].
- Security: Security is one of the most important challenges. Therefore, cybersecurity should be applied to protect systems, networks, and all data from digital attacks and to ensure the secure operation of the complex research, production infrastructures for creating trusted secure environments, and cooperating groups of researchers and technology specialists [9].

3. Domain Engineering

This work performs domain engineering (also known as product line engineering) to better understand the domain and build the model. According to Aharoni and Berger, domain can be defined as a set of applications that use common concepts to describe requirements, problems, and capacities [8]. It is a set of applications or systems from the same domain which have recurrent features that can be analyzed and identified to promote the goal which is software reuse [19]. The domain engineering ensures that the available variability allows producing the application [25].

The domain engineering process consists of three phases, steps, or domains [25]. Domain analysis is the first phase and is the main focus of this work. The outputs (analyzed model) of this phase are the common features and relevant information of the domain collected by studying the underlying theory, emerging technologies, existing systems,

and data captured from experts [28]. The second phase is the domain design. The output of the domain analysis becomes the input for this phase to make the specifications (design model). The last phase is domain implementation, which involves the activities of development of reusable components [19].

This work applies FODA (feature-oriented domain analysis) [48–50] since it best fits the purposes and goals of the study. FODA identifies and analyzes the systems based on their supported features/tools. Additionally, FODA supports software reuse at the functional and architectural levels [28]. This method performs three major activities:

- Context analysis: This phase focuses on defining the domain of the product (Challenges of Health Big Data in Section 3.1).
- Domain model development (Section 4).
- Domain design and implementation (Section 6).

3.1. Context (Domain) Analysis

The development (building from scratch) of a large and complex software system requires an understanding of the system features and the capabilities of the software required to implement those features. Domain analysis explores software systems to define and exploit commonality. It defines the features and capabilities of a class of related software systems. Additionally, it provides a common understanding of the domain [42].

The first step in FODA is defining the context. In this article, the context is defined as "health big data challenges." Health big data face a lot of challenges that limit the ability to apply them in many countries (Section 5).

4. Domain Model Development—Proposed Framework

"In practice, we are now data rich, but information poor" [51]. As discussed earlier, in Jordan there are different health systems for the private and public sectors. Public sectors adopted the Hakeem health program (electronic health solution) in the process of applying the electronic society, and since then data from over five million patients have been collected and stored in an electronic format. The private sectors have been computerized and are using multi-systems. In Figure 1, X represents local repositories for each army hospital. The arrows represent the raw data generated from the medical centers being transferred either to their local repositories or to Hakeem's repository. A, B, and C represent private medical centers. A′, B′, and C′ represent their local repositories.

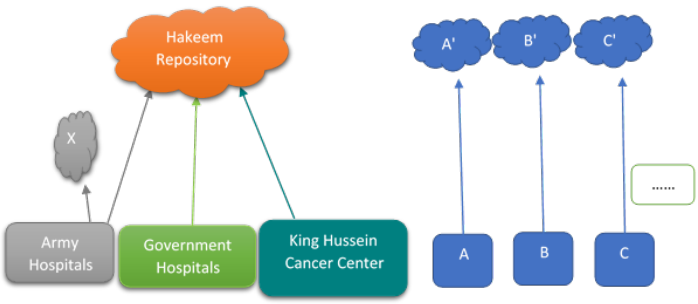

Figure 1. Current health data systems in Jordan.

The collected data vary in sources and types. Sources can be either: individuals (e.g., doctors, nurses, or recorders) or sensors. Types are [33]:

1. Structured (such as laboratory tests and Logical Observation Identifiers Names and Codes (LOINC)).
2. Semi-structured (such as data generated by converting the paper records to electronic health and medical records and instrument readings).
3. Unstructured (such as handwritten nurse notes, medical images, videos, electrocardiogram (ECG), and data from SNSs).

Our health system grows fast, and it involves huge and varying sensitive data (big data). Data needs to be stored, managed, and analyzed efficiently.

4.1. Discussion

In general, a BD framework consists of a set of special technologies, tools, and programs (see Figure 2).

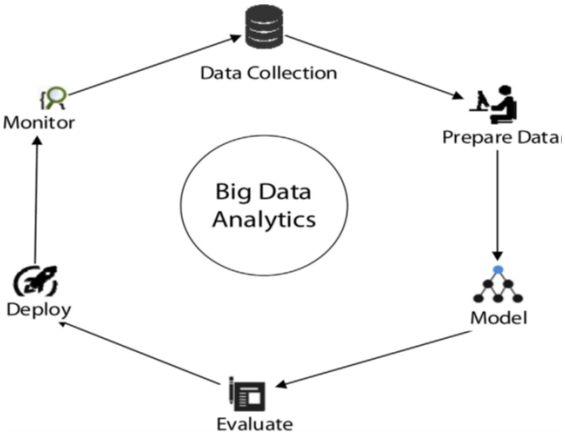

Figure 2. BD analytics framework [9].

It is designed to collect valuable data and get insights from the raw data. It focuses on how to analyze the data that could not be handled using the common analysis techniques, due to their volume, velocity, and the other big data characteristics (V family mentioned in Section 1.1) [52]. It is difficult to collect, store, and process the BD to reach the beneficial target of data utilization.

The conceptual framework includes Hadoop/MapReduce and other related tools. It focuses on how to analyze the raw data collected from our medical centers to get insights and use it to provide better healthcare. The proposed framework is inspired by other related works. ur Rehman et al. [9] proposed a BD analytics framework that includes collecting, preparing, modeling, evaluating, deploying, and monitoring data (Figure 2). Song and Ryu [32] suggested a conceptual framework and applied it to suicide medical data (Figure 3). It explains the process of extracting and analyzing social big data. First, it collects big data (unstructured data). Second, data are analyzed (buzz analysis, keyword analysis, opinion analysis, and account analysis). Then, data are classified and converted to structured data. After that, the structured data are connected to the government and public sectors. Finally, data are analyzed using structural equation modeling, the multi-level model, and data mining analysis [32]. Moreover, Wang and Hajli [36] presented a BD architecture that consists of three components: (1) data aggregation, (2) data processing, and (3) data visualization.

Figure 4 shows the proposed conceptual framework. The left boxes (level 1) represent the medical centers and the medical data sources. BD analytics (level 2) represents the whole difference between the current situation and the proposed one. The BD repository (level 3) is the only medical repository (private cloud) that allows the sharing of medical data between all kinds of medical centers. The dashed arrows represent the raw data that comes from the medical centers. The solid arrows represent the valuable data generated from BD analytics. The National Health Information Exchange helps in improving the quality, safety, efficiency, and exchanging patients' data. The High Health Council and the World Health Organization get their valuable information from the National Health Information Exchange, so they can issue medical reports and put together strategies to enhance the health services. Applying this framework ensures that Jordan will become a pioneer in the health care sector, regionally and globally.

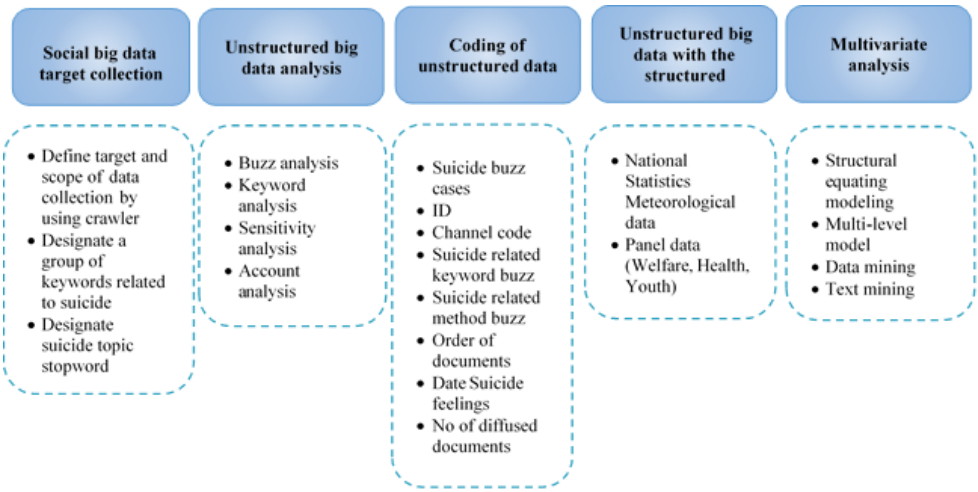

Figure 3. Social big data analysis process and method: application to suicide [30].

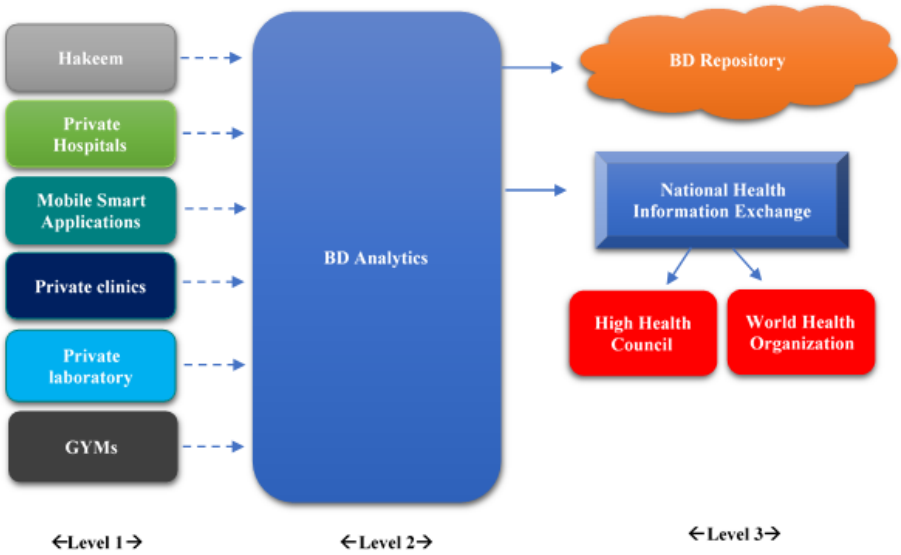

Figure 4. The proposed framework.

4.2. Big Data Analytics

Big data analytics examines the raw data to help provide better health services for society and individuals, and to better use healthcare real-time data [36].

Due to its volume, variety, and other characteristics, it is difficult to analyze and process big health data with the traditional techniques. Thus, appropriate tools and techniques are needed to help handle such data. Hadoop is designed to handle and process such numerous and varied data based on the MapReduce concept, with which it splits the input data sets into multiple chunks, and then processes them in parallel among various nodes [33]. Figure 5 shows the major phases of the proposed framework as it has been merged with our existing health system. Data collection has been covered by the existing system. Data preparation, data analysis, and visualization are phases that will be integrated into the system. Storing, the last phase, is where the valuable data is stored in a shared repository.

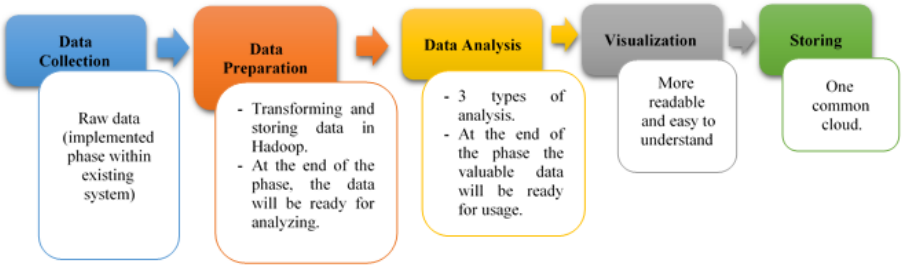

Figure 5. Analytics phase.

- Data Collection: Raw and heterogeneous data are collected from varied sources (individuals or sources) and types (structured, semi-structured, or unstructured) as mentioned earlier. At the end of this phase, data are converted into a structured format.
- Data Preparation: Preparation is a major phase where data are prepared and analyzed using Hadoop/MapReduce [37]. The data passes through two components: transforming and storing. The transforming component is where the structured data are transformed into certain data formats [36]. The healthcare data (e.g., patients' records) must be extracted from EHR and then converted into a specific standard data format (Hadoop Distributed File system), and finally sorted according to a specific criterion (such as patient's gender, name, address, or medical history). The storing component is where the data are cleaned. Data are loaded into the hive or in the Hadoop cloud for more analysis and processing. The implementation and completion of the data storage methods can be either in real-time or in batch processes [53].
- Data Analysis: The generated output can be described as follows:
  - Meaningful Information about the health of the local's community.
  - Prevention and prediction information that is related to diseases.
  - Meaningful information that can be used in diagnosing individuals' health status.

  To reach the expected results, an appropriate analysis was performed which depends on two properties (data type and goal of the analysis). Existing studies and related works [12,36,37,53] adopted three major types of data analysis (see Table 1). This work follows the other existing studies and related works by adopting the same used types of data analysis.

  The descriptive analysis helps to understand the behavioral history of the patient and find the effects of that behavior on the outcomes from their EHR database.

  Hadoop/MapReduce is the most widespread programming model in predictive analysis, which integrates analytical approaches, such as text mining, natural networks, and natural language processing (NLP) in a massively parallel processing (MPP) environment. In predictive analytics, huge amounts of data can be processed in batches, which allows the analysis of both structured data and unstructured data, and supports data processing in near real-time. The data are tracked in motion (in real-time), which allows quick responses to any unexpected events. Moreover, users can use predictive analysis in developing predictive models that can predict patterns and hidden relationships between the target variables.

  Based on the predictive model, the prescriptive analysis provides the optimal solution of actions to be done and offers reports based on the static prediction.
- Data Visualization: This means interpreting the data as graphs, shapes, tables, diagrams, or charts [43]. It supports physicians, nurses, analysts, and health care managers by making data more readable [37,40]. Visualization is done on the results gained during the analysis phase (visualizing only relevant data) [11,54]. Traditional visualization tools have failed with big data due to its volume [54]. Many big data visualization tools run on the Hadoop platform, such as Hadoop YARN, Hadoop Distributed File System (HDFS), Hadoop Common, and Hadoop/MapReduce [33].
- Data Storing/Storage: What makes this phase different is the way data are stored. The storage process is in one shared cloud that is used to store data that come from

both public and private sectors. This makes it easier to transfer patients among medical centers.

Table 1. Types of analysis supported by the proposed framework.

Analysis Types	Questions Need Answers	Comments
Descriptive	- What has happened in the past?	High-speed parallel process.
Predictive	- Why did it happen? - What will happen?	Hadoop/MapReduce is the most used programming model of this type.
Prescriptive	- What should be done?	Based on the predictive model.

5. Proposed Framework Challenges

The proposed framework may face or add challenges to the current healthcare system. This section discusses the challenges.

- Transforming and Sharing: Data transforming involves different stages, levels, and different data structure forms. The raw data is transformed from the collection devices into the data repository. Then, the valuable data will transform out of the repository based on some queries [28,46]. Date sharing requires a high network bandwidth capacity which can be considered a challenge in this phase.
- Storing/Storage: The proposed framework suggests storing the data in a private cloud, which adds a challenge along with the government policies.
- Analysis: Analysis challenges could be solved by using the appropriate tools. The only challenge is to learn how to deal with these tools.
- Visualization: Many visualization tools could be used with big data. The only challenge is to learn how to deal with these tools.

The proposed framework focuses on adding tools and techniques to overcome such challenges. Table 2 introduces a comparison of the proposed framework with the existing frameworks based on the challenges faced and shows that all challenges are addressed in the framework. The proposed framework addresses the same challenges addressed by Demchenko et al. [28].

Table 2. The proposed framework compared to the other frameworks based on challenges faced.

	Challenges						
Frameworks	**Technologies**	**Collection**	**Transforming and Sharing**	**Storing/Storage**	**Analysis**	**Visualization**	**Security**
(Demchenko et al., 2013)	√	√	√	√	√	√	√
(Crawford, and Schultz, 2014)					√		√
(Madhavi and Ramana, 2015)		√	√	√	√		
(Cunha et al., 2015)		√	√	√	√		
(Sheriff et al., 2015, January)		√	√	√	√	√	
(Song and Ryu, 2015)	√	√	√	√	√		√
(ur Rehman et al., 2016)	√	√	√	√	√		√
(Kavitha and Kannudurai, 2016)		√	√	√	√		
(Roden et al., 2017)			√	√			
(Wang and Hajli, 2017)			√		√	√	
Wang et al., (2018)			√		√	√	
Proposed framework	√	√	√	√	√	√	√

6. The Framework's Integration Design

This section presents the framework's integration process with the current medical systems (see Figure 6). Left boxes represent the data sources and their databases that are connected as examples. The dashed lines represent indirect communications, and the solid lines represent direct access communications. The BD analytics phase is represented

by Hadoop/MapReduce. The right side represents the BD repository, the national data exchange, the High Health Council, and the World Health Organization. The left side arrows represent raw data that come out of the medical sources. Right-side arrows represent valuable data that are generated in the BD analytics phase.

The proposed framework interfaces with an intelligent project to computerized healthcare, which is used by Hakeem. This project is called VistA (Veterans Health Information Systems and Technology Architecture) [55].

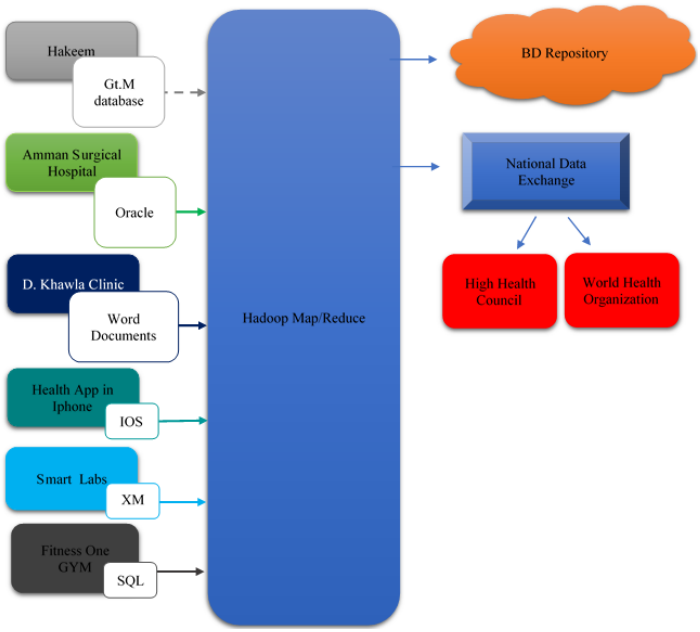

Figure 6. Proposed framework as integrated with the current health system.

6.1. Hakeem and our Proposed Framework

As mentioned earlier, the proposed framework involves five phases. This section introduces the first two phases (data collection and data preparation) and explains how they work. The remaining phases are common between all clinical centers and have been discussed (see Figure 5).

6.1.1. Data Collection

An existing step in Hakeem's health system. It gathers different types of raw and unstructured data from different sources. Data is converted into a structured format based on the data types. For example, medical images are converted into structured data by Vista imaging system. The system uses DICOM (Digital Imaging and Communications in Medicine) [56] standard which allows transferring images from/into PACS (Picture Archiving and Communication System), and any other DICOM compliant devices [57].

6.1.2. Data Preparation

Data is prepared using Hadoop/MapReduce [37]. Data passes through two components: transforming and storing/storage. In the transforming component, the structured data is transformed into certain data formats [36]. The healthcare data such as patient records are extracted from EHR in the Hakeem system, and then converted into a specific standard data format (Hadoop Distributed File system) sorted based on a specified criterion (e.g., patient gender, patient name, address, or medical history). Then, the prepared data is stored. Data is loaded into HDFS or Hadoop cloud for more analysis and processing. The implementation of the data storage methods can be either in batch processes or in real-time [14].

As shown in Figure 6, Hakeem's database (Gt.M) cannot communicate directly with the Hadoop framework; therefore the solution to such limitation is to use another technol-

ogy called CHISTAR [27]. CHISTAR makes the health data exchange easier and provides an integration engine that solves the data integration problem from several data sources into the cloud.

Semantic Interoperability

To be able to interpret and share the data, and to make the information exchanged effectively; CHISTAR uses a two-level modeling approach [27] that separates clinical knowledge from data. This allows CHISTAR to get the benefit of making the system more robust as the software needs not to be changed whenever there is a change in the clinical knowledge.

Data Integration

The data integration process in CHISTAR is converting the EHR data, which comes from different resources and with different formats, to flat files stored in HDFS. Therefore, the integration engine is based on Hadoop/MapReduce.

Data integration engine receives different data formats (HL7 messages, HL7 CDA documents, text files, images files, video files, etc.) from different sources, and performs different processes/techniques (meta-data lookup, semantic matching, input splitting, and data aggregation) on the data formats.

Figure 7 illustrates the architecture of CHISTAR's data integration engine, the meta-data lookup, and semantic matching steps. The engine allows the interoperability with the health information systems regardless of their different sources and formats. *Data integration* consists of six steps as follows:

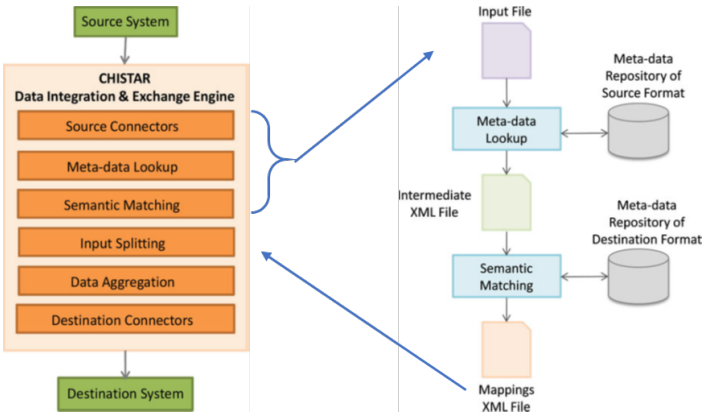

Figure 7. Illustration of CHISTAR's data integration engine architecture.

1. The first step (source connectors) is a connection between the source connectors and the external systems, and the source connectors configuration against the data sources.
2. The second step (metadata lookup) helps in discovering the third step (semantic matching). In metadata lockup, the input file (source file) is analyzed and the metadata repository is looked up to retrieve the semantics of all the data elements in the input file and produce annotations. The output of this step is an intermediate XML file that contains all the data elements in the input file along with the annotations.
3. Step three (semantic matching): in this step, the search process is done relying on annotations for source data elements in the intermediate XML file. The search process is done on the metadata repository of the destination format, to produce a list of candidate mappings for each data element in the intermediate file. The output of this step is a mapping XML file that is used in guiding the data importing process that means the same mappings can be used to import several source files of the same format.
4. In the fourth step (input splitting), when entering more than one data file, the input is split to parallelize the data importing process.

5. The fifth step (data aggregation) means transforming the data from input splits to the destination format, and this done by MapReduce jobs.
6. The sixth step (destination connectors) is storing the data files on HDFS storage via the destination connectors.

To integrate EHR data from Hakeem or VA Vista, CHISTAR builds upon the File Manager (FM) Projection toolkit and the Mirth Connect. FM projection projects and structures the FM data; therefore, they can be viewed through standard database query and reporting tools.

Limitations

This study has a certain number of limitations.

1. Hakeem's internal policies require official approaches and their top management involvement to facilitate the integration of the framework with their system, which made it difficult to integrate and validate the framework.
2. Dealing with sensitive data, which requires a huge effort to convince the stakeholders of the capability of our proposal and the high level of security that ensures and maintains data privacy.
3. Gt.M database is not widely used. This is a consequence of the direct execution of the source code, which is needed to save memory and run fast. Additionally, the coding style lags best practices for readability [58].
4. FM projection (related to CHISTAR) suffers from low performance. FM projection projects and structures the data; therefore, they can be viewed through standard database query and reporting tools, and examines VistA FM data and structure by using SQL like representations. The limitation is due to its nature as a storage plugin to MySQL, which allows FileMan files to appear as SQL tables. Additionally, and as a code base, it is rather old and does not compile on the modern MySQL forks such as MariaDB (https://mariadb.org/), which introduces a performance challenge.

7. Conclusions and Recommendations

Big data analytics has emerged as an evolving field in the health domain. The importance of health big data analytics and management is intended to improve individual and community health care. It improves health quality and reduces the overall cost.

This article proposes the design of a healthcare big data framework that can merge with the existing health system in Jordan. Also, it suggests a connection with the National Data Exchange to improve the quality of society's health services. The article performed a domain engineering process to understand the context of the model and the posed challenges. Then, a conceptual design for the framework is presented. Applying the proposed framework helps the community and supports individuals with better healthcare services.

Future work might involve building an application to demonstrate the effectiveness of the framework. Also, applying this framework with the existing health system, and conducting case studies to examine how the integration will improve healthcare and improves services.

Author Contributions: Methodology, H.B.-S. and M.A.-Q.; writing—original draft, H.B.-S., M.A.-Q. and S.T.; writing—review and editing, H.B.-S., M.A.-Q. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ricciardi, C.; Amboni, M.; De Santis, C.; Improta, G.; Volpe, G.; Iuppariello, L.; Ricciardelli, G.; D'Addio, G.; Vitale, C.; Barone, P.; et al. Using gait analysis' parameters to classify Parkinsonism: A data mining approach. *Comput. Methods Programs Biomed.* **2019**, *180*. [CrossRef]
2. Ricciardi, C.; Cantoni, V.; Green, R.; Improta, G.; Cesarelli, M. Is It Possible to Predict Cardiac Death? In Proceedings of the XV Mediterranean Conference on Medical and Biological Engineering and Computing—MEDICON 2019, MEDICON 2019, IFMBE Proceedings, Coimbra, Portugal, 26–28 September 2019; Henriques J., Neves N., de Carvalho, P., Eds.; Springer: Cham, Switzerland, 2019; Volume 76. [CrossRef]
3. Improta, G.; Ricciardi, C.; Cesarelli, G.; D'Addio, G.; Bifulco, P.; Cesarelli, M. Machine learning models for the prediction of acuity and variability of eye-positioning using features extracted from oculography. *Health Technol.* **2020**, *10*, 961–968. [CrossRef]
4. Reddy, S.R.I.P. Deepak and Udaya Kumar Ramanadham. Big Data Analytics for Healthcare Organization , BDA Process , Benefits and Challenges of BDA: A Review. *Adv. Sci. Technol. Eng. Syst. J.* **2017**, *2*, 189–196. [CrossRef]
5. Bizer, C.; Boncz, P.; Brodie, M.L.; Erling, O. The meaningful use of big data: Four perspectives–four challenges. *ACM Sigmod Rec.* **2012**, *40*, 56–60. [CrossRef]
6. Chen, D.Q.; Preston, D.S.; Swink, M. How the use of big data analytics affects value creation in supply chain management. *J. Manag. Inf. Syst.* **2015**, *32*, 4–39. [CrossRef]
7. Bollier, D.; Firestone, C.M. *The Promise Additionally, Peril of Big Data*; Aspen Institute, Communications and Society Program: Washington, DC, USA, 2010; pp. 1–66.
8. Feldman, B.; Martin, E.M.; Skotnes, T. Big Data in Healthcare Hype and Hope. *Dr. Bonnie* **2012**, *360*. Available online: https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf (accessed on 5 February 2021).
9. Ur Rehman, M.H.; Chang, V.; Batool, A.; Wah, T.Y. Big data reduction framework for value creation in sustainable enterprises. *Int. J. Inf. Manag.* **2016**, *36*, 917–928. [CrossRef]
10. Zhang, X.; Pèrez-Stable, E.J.; Bourne, P.E.; Peprah, E.; Duru, O.K.; Breen, N.; Denny, J. Big data science: Opportunities and challenges to address minority health and health disparities in the 21st Century. *Ethn. Dis.* **2017**, *27*, 95. [CrossRef] [PubMed]
11. Hermon, R.; Williams, P.A. Big data in healthcare: What is it used for? In Proceedings of the 3rd Australian eHealth Informatics and Security Conference, Edith Cowan University, Joondalup Campus, Perth, WA, Australia, 1–3 December 2014; pp. 40–49.
12. Kim, G.H.; Trimi, S.; Chung, J.H. Big-data applications in the government sector. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
13. Luo, J.; Wu, M.; Gopukumar, D.; Zhao, Y. Big data application in biomedical research and health care: A literature review. *Biomed. Inform. Insights* **2016**, *8*, BII-S31559. [CrossRef] [PubMed]
14. Wong, H.T.; Chiang, V.C.L.; Choi, K.S.; Loke, A.Y. The need for a definition of Big Data for nursing science: A case study of disaster preparedness. *Int. J. Environ. Res. Public Health* **2016**, *13*, 1015. [CrossRef] [PubMed]
15. Bram, J.T.; Warwick-Clark, B.; Obeysekare, E.; Mehta, K. Utilization and monetization of healthcare data in developing countries. *Big Data* **2015**, *3*, 59–66. [CrossRef]
16. Lee, C.H.; Yoon, H.J. Medical big data: Promise and challenges. *Kidney Res. Clin. Pract.* **2017**, *36*, 3. [CrossRef]
17. Iwashyna, T.J.; Liu, V. What's so different about big data? A primer for clinicians trained to think epidemiologically. *Ann. Am. Thorac. Soc.* **2014**, *11*, 1130–1135. [CrossRef] [PubMed]
18. Bansal, A.; Deshpande, A.; Ghare, P.; Dhikale, S.; Bodkhe, B. Healthcare data analysis using dynamic slot allocation in Hadoop. *Int. J. Recent Technol. Eng.* **2014**, *3*, 15–18.
19. Duggal, R.; Khatri, S.K.; Shukla, B. Improving patient matching: Single patient view for Clinical Decision Support using Big Data analytics. In Proceedings of the 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2–4 September 2015; pp. 1–6.
20. Shahbaz, M.; Gao, C.; Zhai, L.; Shahzad, F.; Hu, Y. Investigating the adoption of big data analytics in healthcare: The moderating role of resistance to change. *J. Big Data* **2019**, *6*. [CrossRef]
21. Wang, Y. Leveraging Big Data Analytics to Improve Quality of Care in Healthcare Organizations: A Configurational Perspective. *Spec. Issue Big Data Firm Perform.* **2018**, *30*, 362–388. [CrossRef]
22. Kumar, Y.; Sood, K.; Kaul, S.; Vasuja, R. Big Data Analytics and Its Benefits in Healthcare. In *Big Data Analytics in Healthcare*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 3–21.
23. Chawla, N.V.; Davis, D.A. Bringing big data to personalized healthcare: A patient-centered framework. *J. Gen. Intern. Med.* **2013**, *28*, 660–665. [CrossRef]
24. Auffray, C.; Balling, R.; Barroso, I.; Bencze, L.; Benson, M.; Bergeron, J.; Del Signore, S. Making sense of big data in health research: Towards an EU action plan. *Genome Med.* **2016**, *8*, 71. [CrossRef]
25. Groves, P.; Kayyali, B.; Knott, D.; Van Kuiken, S. The 'Big Data' Revolution in Healthcare: Accelerating Value and Innovation. McKinsey & Company. 2013. Available online: https://www.ghdonline.org/uploads/Big_Data_Revolution_in_health_care_2013_McKinsey_Report.pdf (accessed on 5 February 2021) .
26. Jee, K.; Kim, G.H. Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system. *Healthc. Inform. Res.* **2013**, *19*, 79–85. [CrossRef]
27. Bahga, A.; Madisetti, V.K. A cloud-based approach for interoperable electronic health records (EHRs). *IEEE J. Biomed. Health Inform.* **2013**, *17*, 894–906. [CrossRef] [PubMed]

28. Demchenko, Y.; Ngo, C.; Membrey, P. Architecture framework and components for the big data ecosystem. *J. Syst. Netw. Eng.* **2013**, 1–31. Available online: https://docplayer.net/1631822-Architecture-framework-and-components-for-the-big-data-ecosystem.html (accessed on 5 February 2021)
29. Crawford, K.; Schultz, J. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *BCL Rev.* **2014**, *55*, 93.
30. Madhavi, D.; Ramana, B.V. De-Identified Personal health care system using Hadoop. *Int. J. Electr. Comput. Eng.* **2015**, *5*, 1492–1499. [CrossRef]
31. Sheriff, C.I.; Naqishbandi, T.; Geetha, A. Healthcare informatics and analytics framework. In Proceedings of the Computer Communication and Informatics (ICCCI), 2015 International Conference on IEEE, Coimbatore, India, 8–10 January 2015; pp. 1–6.
32. Song, T.M.; Ryu, S. Big data analysis framework for healthcare and social sectors in Korea. *Healthc. Inform. Res.* **2015**, *21*, 3–9. [CrossRef]
33. Cunha, J.; Silva, C.; Antunes, M. Health twitter big bata management with hadoop framework. *Procedia Comput. Sci.* **2015**, *64*, 425–431. [CrossRef]
34. Kavitha, V.; Kannudurai, S. Health Care Analytics with Hadoop Big Data Processing. *Int. J. Adv. Res. Biol. Eng. Sci. Technol.* **2016**, *2*, 264–266.
35. Roden, S.; Nucciarelli, A.; Li, F.; Graham, G. Big data and the transformation of operations models: A framework and a new research agenda. *Prod. Plan. Control* **2017**, *28*, 929–944. [CrossRef]
36. Wang, Y.; Hajli, N. Exploring the path to big data analytics success in healthcare. *J. Bus. Res.* **2017**, *70*, 287–299. [CrossRef]
37. Wang, Y.; Kung, L.; Wang, W.Y.C.; Cegielski, C.G. An integrated big data analytics-enabled transformation model: Application to health care. *Inf. Manag.* **2018**, *55*, 64–79. [CrossRef]
38. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 54. [CrossRef]
39. Bora, D.J. Chapter 3—Big Data Analytics in Healthcare: A Critical Analysis, In *Advances in Ubiquitous Sensing Applications for Healthcare, Big Data Analytics for Intelligent Healthcare Management*; Academic Press: Cambridge, MA, USA, 2019; pp. 43–57, ISBN 9780128181461. [CrossRef]
40. Kaur, J.; Mann, K.S. AI based HealthCare Platform for Real Time, Predictive and Prescriptive Analytics using Reactive Programming. *J. Physics Conf. Ser.* **2018**, *933*, 012010. [CrossRef]
41. Khanra, S.; Dhir, A.; Islam, A.N.; Mäntymäki, M. Big data analytics in healthcare: A systematic literature review. *Enterp. Inf. Syst.* **2020**, *14*, 878–912. [CrossRef]
42. Bellazzi, R. Big data and biomedical informatics: A challenging opportunity. *Yearb. Med. Inform.* **2014**, *9*, 8. [CrossRef] [PubMed]
43. Chen, C.P.; Zhang, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347. [CrossRef]
44. Kaisler, S.; Armour, F.; Espinosa, J.A.; Money, W. Big data: Issues and challenges moving forward. In Proceedings of the System Sciences (HICSS), 2013 46th Hawaii International Conference on IEEE, Wailea, HI, USA, 7–10 January 2013; pp. 995–1004.
45. Jagadish, H.V.; Gehrke, J.; Labrinidis, A.; Papakonstantinou, Y.; Patel, J.M.; Ramakrishnan, R.; Shahabi, C. Big data and its technical challenges. *Commun. ACM* **2014**, *57*, 86–94. [CrossRef]
46. Kuo, M.H.; Sahama, T.; Kushniruk, A.W.; Borycki, E.M.; Grunwell, D.K. Health big data analytics: Current perspectives, challenges and potential solutions. *Int. J. Big Data Intell.* **2014**, *1*, 114–126. [CrossRef]
47. Olshannikova, E.; Ometov, A.; Koucheryavy, Y.; Olsson, T. Visualizing Big Data with augmented and virtual reality: Challenges and research agenda. *J. Big Data* **2015**, *2*, 22. [CrossRef]
48. Kang, K.C.; Cohen, S.G.; Hess, J.A.; Novak, W.E.; Peterson, A.S. *Feature-Oriented Domain Analysis (Foda) Feasibility Study*; Tech. Rep. CMU/SEI-90-TR-21; Software Engineering Institute, Carnegie Mellon University: Pittsburgh, PA, USA, 1990.
49. Oliveira, L.S.; Gerosa, M. Collaborative features in content sharing Web 2.0 social networks: A domain engineering based on the 3C collaboration model. In *Proceedings of the Collaboration Researchers' International Workshop on Groupware (CRIWG'11)*; Springer: Berlin/ Heidelberg, Germany, 2011; Volume 6969, pp. 142–157.
50. Bani-Salameh, H.; Jeffery, C.; Hammad, M. Developers' social networks-tools analysis based on the 3Cs model. *Int. J. Netw. Virtual Organ.* **2013**, *13*, 159–175. [CrossRef]
51. Sakr, S.; Elgammal, A. Towards a comprehensive data analytics framework for smart healthcare services. *Big Data Res.* **2016**, *4*, 44–58. [CrossRef]
52. Luna, D.R.; Mayan, J.C.; Garcìa, M.J.; Almerares, A.A.; Househ, M. Challenges and potential solutions for big data implementations in developing countries. *Yearb. Med. Inform.* **2014**, *23*, 36–41.
53. Wang, Y.; Kung, L.; Byrd, T.A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Chang.* **2018**, *126*, 3–13. [CrossRef]
54. Wang, L.; Wang, G.; Alexander, C.A. Big data and visualization: Methods, challenges and technology progress. *Digit. Technol.* **2015**, *1*, 33–38.
55. VistA. Available online: https://worldvista.org/AboutVistA (accessed on 5 February 2021).
56. Bidgood, W.D., Jr.; Horii, S.C.; Prior, F.W.; Van Syckle, D.E. Understanding and using DICOM, the data interchange standard for biomedical imaging. *J. Am. Med. Inform. Assoc. JAMIA* **1997**, *4*, 199–212. [CrossRef] [PubMed]

57. Strickl, N.H. PACS (picture archiving and communication systems): Filmless radiology. *Arch. Dis. Child.* **2000**, *83*, 82–86. [CrossRef] [PubMed]
58. YottaDB: What Was Old Is New Again. 2018. Available online: https://docs.yottadb.com/Presentations/WhatWasOldIsNewAgain.pdf (accessed on 5 February 2021).