Hindawi
Journal of Healthcare Engineering
Volume 2020, Article ID 8865808, 12 pages
https://doi.org/10.1155/2020/8865808

Review Article

A Semantic-Based Approach for Managing Healthcare Big Data: A Survey

Rafat Hammad, Malek Barhoush, and Bilal H. Abed-alguni

Yarmouk University, Irbid, Jordan

Correspondence should be addressed to Rafat Hammad; rafat.hammad@yu.edu.jo

Received 29 September 2020; Revised 2 November 2020; Accepted 9 November 2020; Published 23 November 2020

Academic Editor: Saverio Maietta

Healthcare information systems can reduce the expenses of treatment, foresee episodes of pestilences, help stay away from preventable illnesses, and improve personal life satisfaction. As of late, considerable volumes of heterogeneous and differing medicinal services data are being produced from different sources covering clinic records of patients, lab results, and wearable devices, making it hard for conventional data processing to handle and manage this amount of data. Confronted with the difficulties and challenges facing the process of managing healthcare big data such as volume, velocity, and variety, healthcare information systems need to use new methods and techniques for managing and processing such data to extract useful information and knowledge. In the recent few years, a large number of organizations and companies have shown enthusiasm for using semantic web technologies with healthcare big data to convert data into knowledge and intelligence. In this paper, we review the state of the art on the semantic web for the healthcare industry. Based on our literature review, we will discuss how different techniques, standards, and points of view created by the semantic web community can participate in addressing the challenges related to healthcare big data.

1. Introduction

Big healthcare data refers to the process of collecting, integrating, managing, processing, and analyzing different kinds of medical data, which are excessively complex and inefficient to be processed and managed using existing database management systems and tools [1–3]. Big data outperforms conventional systems in the amount of data and processing capacity. There are many definitions available for big data; the most recognized is the definition that was given by Douglas Laney [4], who presented three features of big data: volume, velocity, and variety (known as the 3 Vs). Many researchers have introduced other Vs to this definition [5], but the 3 V definition model stays the most widely accepted definition.

Volume refers to the amount of data generated by different information systems. Velocity refers to the speed at which data are being produced, processed, stored, and analyzed [6]. Variety refers to the types of data being captured and processed, which can have different degrees of organization (e.g., structured, unstructured, and semi-structured) and different formats (e.g., plain text, video, audio, and images).

1.1. Motivation for This Survey. The data generated from biomedical research and the process of digitization the healthcare sector have already generated and will continue to produce large amounts of data. This data is generated from different resources such as clinical records, hospitals, patient monitoring devices, medical images, and lab results. With the present continuously improving innovation of technologies, it has become simpler to gather, manage, and analyze these different types of medical data to infer meaningful insights. [7, 8]. For example, Santis et al. [9] used reverse engineering approaches with laser scans and surface texturing to evaluate shift and reduction of the tibiofemoral contact area after meniscectomy. Medical image management remains an energizing field of exploration and applications for healthcare and biomedical research. It includes

all the techniques that provide the efficient storage, transmission, and retrieving of image data [10].

One of the significant difficulties in utilizing large data in healthcare is interoperability. Clinical information is spread across numerous sources administered by various districts, emergency clinics, and managerial divisions. The reconciliation of these information sources would require building up a new framework where all information suppliers integrate with each other. Data coming from different resources will have many challenges because of the irregularity in naming, structure, organization, and format. A significant prerequisite is to catch applicable information and make it broadly accessible and available in a clean and consistent configuration for easy integration with other information systems [7].

In the last few years, the semantic web was introduced by the World Wide Web Consortium (W3C) to enable a simpler method for searching, reusing, integrating, and sharing information [11]. Semantic techniques have demonstrated to be the most pertinent for comprehending and solving a lot of difficulties and challenges which face the healthcare big data community. This explains why many researchers and scientists are centered on finding semantic-based solutions for big healthcare data [12]. Semantic techniques have profited healthcare communities by improving their effectiveness and efficiency. The following are a few applications which list some of the motivations for using semantic techniques to allocate big data in the healthcare domain [13]:

(i) Allowing medical doctors to understand as plenty as they can about an affected person and as early in their existence as possible, to pick out up caution signs and symptoms of significant contamination as they rise.

(ii) Allowing patients to wear medical smart devices to keep them away from hospitals. These devices produce considerable amounts of data, and analysts can use different big data tools to handle this large volume of data in real-time mode.

(iii) Improving care personalization by analyzing all available healthcare data. This allows delivering the right treatment to the right patient at the right time.

(iv) Allowing hospitals to enhance their security by monitoring any unusual changes in the network traffic to stop any cyber attacks.

(v) Monitoring patients with partial disabilities to analyze their activities and know their requirements for coexistence with the community.

1.2. Organization of the Paper. This paper is an endeavor to concentrate on the impact of consolidating healthcare big data with the semantic web to make it more intelligent. The remainder of the paper is organized as follows: Section 2 discusses the concepts of semantic modeling and ontology. The contribution of semantics to healthcare big data acquisition is discussed in Section 3. Section 4 discusses the role of semantic techniques in integrating healthcare data. Semantic healthcare repositories are discussed in Section 5. Finally, our findings and concluding remarks are summarized in Section 6 and Section 7.

2. Background and Related Work

In this section, we study various ideas identified in managing healthcare big data and give broad insights concerning these ideas to assist readers with understanding the basic concepts introduced in this paper.

2.1. XML, RDF, SPARQL, and OWL Languages. Semantic modeling uses the following languages to represent and model the contained data: Extensible Markup Language (XML), Resource Description Framework (RDF), SPARQL, and Web Ontology Language (OWL) [14].

XML is a simple text-based markup language that defines a set of rules for representing and describing documents in a format that is both human-readable and machine-readable. It is one of the most widely used formats for sharing structured information between computers and between people [15]. Figure 1 shows a sample of XML document:

RDF is a language for representing information about resources on the Web. It was designed to provide a common way to describe information so it can be exchanged between different types of computers regardless of their operating systems or programming languages. RDF documents are written in XML and the language used by RDF is called RDF/XML. RDF identifies things using Uniform Resource Identifier (URI) and describes resources with properties and property values [16]. Figure 2 shows a sample of RDF which describes the resource "http://www.example.com/rdfsource.html."

SPARQL is a semantic query language for retrieving and manipulating data stored in RDF format [17]. Figure 3 shows SPARQL query which returns the names and emails of every patient in the RDF dataset.

OWL is an ontology language designed to represent knowledge about things and the relations between these things. It includes a set of operators for forming concept descriptions that can be used to share knowledge between applications [18].

2.2. Big Data Platforms and Tools. To scale and accommodate a large amount of healthcare data, healthcare management systems must use distributed computing platforms. Many of the available distributed platforms are built on top of Apache Hadoop, which is considered the base. Hadoop is an open-source framework that was designed to support distributed storage and processing using simple programming models. It consists of two main components: HDFS and MapReduce. HDFS is a distributed file system that stores data on a cluster of machines. MapReduce is a computational model that spreads data and calculations over any number of servers in the cluster.

On the other hand, NoSQL, which simply means "not only SQL, was created especially as a distributed database

```
<patient>
    <name>
        <firstName> Mike </firstName>
        <lastName> Dale </lastName>
    </name>
    <birthDate> 3/2/2005 </birthDate>
    <gender> Male </gender>
</patient>
```

Figure 1: A sample of XML document.

```
Property      Value      Resource
<RDF>
  <Description about = "http://www.example.com/rdfsource.html">
    <author> Mike Dale </author>
    <title> RDF Source Example </title>
  </Description>
</RDF>
```

Figure 2: A sample of RDF document.

```
PREFIX ds : < http://informatics.com/rdf >
SELECT  ?name
        ?email
WHERE
   {
     ?patient    a ds : Patient .
     ?patient    ds : name   ?name .
     ?patient    ds : email  ?email .
   }
```

Figure 3: A sample of SPARQL query.

framework where data can be stored in multiple processing nodes.

There are a large number of available NoSQL databases these days. Based on the used data model, these databases are usually divided into four categories and shown in Table 1.

The NoSQL database must have a set of features and characteristics to perform their work in a highly efficient way, especially when dealing with big data applications, and these characteristics are summarized as follows [28]:

(i) Horizontal scalable: NoSQL data need to spread their data among multiple servers; this data needs to be treated efficiently

(ii) Data replication: NoSQL data need to be replicated to improve performance

(iii) Suitable consistency model: NoSQL replicated data need to be consistent with suitable concurrency

(iv) Simple graphical user interface: to attract NoSQL users, the GUI must be simple and easy to use, and all operations performed in SQL databases can be done in NoSQL databases easily and conveniently

(v) Powerful data store: the data stored must be close to the user, which increases performance

(vi) Flexible attribute resizing: different records within the NoSQL database may have different attributes, and this feature helps to reduce the tables in a database to the minimum

The semantic model is a method for arranging and organizing data so that it can be interpreted by computers without human mediation. It is a conceptual model that incorporates semantics and relations to data [29, 30]. Ontology is used as a common representation of knowledge, and this gives the flexibility to share and reuse knowledge between distributed and heterogeneous systems and databases [31]. It is considered as one of the main building blocks of semantic modeling and it consists of the following components [32]:

(i) Classes represent sorts of things in the world. For example, the human "Leg" represents a class.

(ii) Instances of classes are individuals fulfilling the classes' intension. For example, the sentence "My Leg" represents an instance.

(iii) Relations between instances emerge from the interactions of individuals. For example, "My Leg Is Part of Me" is considered a relation.

(iv) Axioms specify our knowledge of the domain. For example, we can conclude the following sentence: "Every Instance of Foot Is a Part of an Instance of Leg."

2.3. Related Work. The area of managing healthcare big data has recently grabbed a lot of attention and has been taken into consideration. Table 2 sums up the key related research papers including our survey. Among these related surveys, none of these papers address the managing of healthcare big data from semantic respective.

2.4. Ontologies for Healthcare Data. Many ontologies have been created in the context of the healthcare domain. The greater part of these ontologies has been made to a particular area in healthcare such as drug development, human disease, rehabilitation, and human hereditary [7]. The list of ontologies is consistently developing and increasing and many of them are available at BioPortal [39]. Table 3 includes some examples of available healthcare ontologies with their features.

3. Semantics for Healthcare Data Acquisition

Data acquisition is the method of gathering, extracting, and transforming data before storing it in the repository, which will be used later for analysis. The acquisition of big data is regularly controlled by the three Vs which include volume, velocity, and variety [12, 46].

Powerful analysis is based on storing the right information. The semantic technologies can be used in data acquisition to extract related and important data. This allows the process of discovering and excluding unnecessary information that contains errors or irregularities before storing it in its final repository [12, 47].

Table 1: Categories of NoSQL data models.

Data model	Description	Platform/tools
Key-value	It represents and stores data as a collection of key-value pairs.	Aerospike [19] Redis [20] RocksDB [21]
Document-oriented	It represents and stores data as documents, such as XML and JSON.	MongoDB [22] CouchDB [22] PostgreSQL [23]
Graph	It represents and stores data as graph structures with nodes, edges, and properties.	Neo4j [24] AllegroGraph [24] ArangoDB [24]
Column	It is a type of the key-value data model, but it uses the notions of rows and columns. It is different from relational databases because the names and format of the columns in the same table can fluctuate from row to row.	HBase [25] Cassandra [26] Accumulo [27]

Table 2: Related research papers.

Paper	Year	Description
[12]	2018	This study presents an overview that compares different approaches for analyzing big data using semantic techniques. The comparison includes the description of every methodology in addition to its downsides. This paper did not cover the issues of capturing and managing data, it focuses only on the analytical part of big data.
[33]	2012	This study presented and compared different RDF storage solutions using a predefined set of characteristics. This study did not cover RDF storage from distributed respective which is needed by big data management.
[11]	2018	This study covered several approaches for integrating big data coming from different sources using semantic techniques. This study did not cover the issues capturing and storing the integrated data in its final target storage repository.
[34]	2018	This study is a review of the exploration subjects in the field of the semantic web over its first 20 years of presence. The study includes recognizing the primary current research trends, challenges, and future research directions in the area of semantic web and linked data.
[35]	2014	This study presented an overview of different frameworks and methodologies which are used to analyze big data in the healthcare domain. This paper did not cover the management part of healthcare big data.
[36]	2015	This study presented the benefits, opportunities, and challenges of using big data in the healthcare field. It also covers a set of applications and methodologies which are used by healthcare and the medical community.
[37]	2010	This research paper presented a framework that uses cloud computing to connect and share information across hospitals. The proposed solution used distributed resources (hardware and software) to process large amounts of medical images.
[38]	2016	This study presented an overview that summarized the challenges which face the big healthcare data in terms of volume, velocity, variety, and veracity. It proposed a systematic data-management pipeline approach for extracting, storing, analyzing healthcare data. This survey did not cover the data management from semantic respective.
Current study	2020	This study reviews the state of the art of semantic web technologies in healthcare management systems. It reviews the role of semantic technologies in extracting, integrating, and storing healthcare big data in distributed environments.

Table 3: List of some ontologies in healthcare domain.

Reference	Ontology	Description	Format	Classes	Properties
[40]	MedDRA	Used for data entry, information retrieval, analysis, and visualization. It covers drug development, health consequences, and malfunction of gadgets.	UMLS	73,429	18
[41]	DOID	Used to represent human illness.	OBO	12,694	15
[42]	PMR	Used for decision support in rehabilitation.	OWL	1,597	52
[43]	HP	Used to give an organized vocabulary for the phenotypic highlights experienced in human genetic and focus on monogenic diseases.	OBO	18,407	0
[44]	ATC	Used to classify drug's ingredients according to the organ on which they act and their chemical characteristics.	UMLS	6,358	3
[45]	ICF	Used to classify health domains: body, individual, and cultural points of view.	OWL	1,596	67

Most of the semantic techniques which are used in healthcare data acquisition are based on using ontologies to represent the healthcare data by following these three steps: (1) converting source data sets to RDF, (2) using ontologies to apply conversion business rules to the RDF data, and (3) loading the processed data into their final repositories.

Ding et al. [48] developed a semantic web portal that enables patients to access all their medical information such as prescription, lab results, doctors, and diseases. This portal allows users to effectively disclose, search, discover, and visualize semantic data easily and smartly. It also enables the process of transforming data from one format, such as relational databases, into an RDF format [7].

Data provenance is utilized for depicting data evolution, which records the entire data process, including all changes and transformations which have been applied to change data from one state to another [49]. Zhao et al. [50] introduced different methods for modeling, capturing, and querying provenance data by identifying mapping joins between data generated from different sources related to genomics. The authors show the utilization of named RDF graphs with various degrees of granularity to make provenance declarations about connected data [51].

The Ambient Assisted Living (AAL) system attempts to make better living conditions for older and disabled people. Forkan et al. [52] proposed a cloud-based solution called CoCaMAAL to handle the process of data gathering and processing in AAL systems. They used ontologies to enhance AAL services by providing a single virtual community that includes patients, devices, and computational servers. The proposed model implements a service-oriented architecture (SOA) for unified context generation. This is achieved by processing and integrating the collected sensor data by choosing the ideal fitting services using a context management system (CMS).

Jiang et al. [53] proposed a context-awareness wearable sensor system. The proposed solution can handle large amounts of data generated from the continuous monitoring of devices wearable by the elderly, sending alerts to the right people when necessary, and sending valuable information for analysis using big data solutions.

Tilahun et al. [54] presented a set of Web semantic, called Linked Open Data (LOD), to publish and connect public heterogeneous health data. All healthcare data were stored in RDF graphs where the triples are connected using the Silk, an open-source framework for integrating different data sources [7].

Michel Dumontier and Villanueva-Rosales [55] introduced their knowledge base structured using semantic technologies to capture pharmacogenomics and related information such as genes, medications, and therapeutic. They represented their semantic information using XML markup languages. Utilizing semantic techniques to capture and model neuroradiological knowledge in a head injury situation and using an ontology to retrieve clinical neurological from different data sets were studied by Garcia et al. [56].

Ullah et al. [57] introduced a Semantic Interoperability Model for Big-Information in IoT (SIMB-IoT) to convey semantic interoperability between data generated from different healthcare information systems. They used annotations for big data, stored the data in an RDF format, and used SPARQL to query the data from the RDF graphs. Yoon et al. [58] proposed a web-based automated extraction system, called DiTex to extract disease-related topics using natural language processing and semantic similarity ranking algorithms. Pacaci et al. [59] developed a semantic transformation approach to extract data from electronic health record systems by converting source datasets to RDF and then loading the processed data into their final repository.

4. Semantics for Healthcare Data Integration

Data can exist over numerous datasets, which requires data to be consolidated and merged using some common fields [12]. Data integration can be defined as the process of integrating data from various sources in a standard format, storing data in proper repositories, and providing a unified view of the data to be used for retrieval and analysis [60].

Using semantic techniques for data integration has become progressively pivotal and has gained a lot of consideration in both database and Web communities. The utilization of ontology as a data broker can make data integration easier in different ways, such as providing global concepts to represent data, automating the process of data integration, and providing the ability to query data semantically [61].

Ethier et al. [62] built up a core ontology to deal with semantic interoperability across heterogeneous biomedical datasets within TRANSFoRm project. This project was designed as an infrastructure for a learning healthcare system in European Primary Care. Keller et al. [63] proposed a semantic framework for consolidating heterogeneous air traffic data using a shared ontology, which is used to transform the original source data into a unified RDF representation. The integrated RDF store can then be queried using SPARQL to retrieve information semantically.

HBase is a NoSQL column-oriented distributed database that can store considerable amounts of data from terabytes to petabytes of data [64]. Kang et al. [65] used HBase repository, ontologies, and data mapping to develop a semantic big data model that integrates heterogeneous data from different resources. Yu et al. [66] presented a framework to integrate data that are continuously generated from different healthcare providers. They used Kafka, a real-time streaming framework, to collect the stream data and store it in NoSQL database. They developed a semantic lifting engine to generate the RDF triples and store it in the Virtuoso RDF repository. They used Apache Jena RDF semantic reasoning framework to analyze the data to find any health risks.

Using ontologies to automatically generate SQL statements during the data extraction process was studied by Mate et al. [67]. They introduced an ontology-based approach to represent concepts of medical data. Schoppenhauer et al. [68] developed a model called Ontology-Based Data Access (OBDA), which maps biomedicine classes and relationships to database entries by generating equivalent SQL statements to retrieve data from heterogeneous relational databases. The data can be retrieved using SPARQ queries, and the result data can then be available as materialized or virtualized RDF triples.

Livingston et al. [69] created a common ontology-based semantic model to integrate and query heterogeneous

biomedical data sources. They devolved a knowledge base system called KaBOB (the Knowledge Base of Biomedicine), which integrates data from 18 different biomedical datasets that produce millions of RDF triples.

Manning et al. [70] developed a metadata ontology to extract knowledge from biological data using semantic techniques. The RDF triples were generated by mapping metadata from different datasets into this ontology. A user-friendly interface was implemented to provide the capability to answer complex queries and retrieve data from both RDF and RDBMS sources and then display the results.

Issa et al. [71] used the Apache spark framework with semantic modeling to integrate, store, process, and analyze sensor big data. An ontology is generated first based on Semantic Sensor Network (SSN) to model the use of the sensor data, and then large amounts of raw sensor data are transformed to semantic data using the resulting SSN-based ontology.

Liang et al. [72] developed an ontology that is used to integrate data from genes, symptoms, diseases, and phenotype. They extracted data from various sources into the ontology and performed semantic reasoning to help select the right medications that will be effective for some diseases such as bipolar and epilepsy.

Mapping very large and diverse datasets (stored in XML, KML, JSON, structured plain text files, or relational databases) into a common shared domain ontology was researched by Knoblock and Szekely [73]. Their methodology depends on extracting, modeling, and storing data in a system called Karma, which will perform their data integration, visualization, and analysis in a distributed environment over the whole dataset. Ruttenberg et al. [74] introduced Neurocommons prototype knowledge base for integrating and querying biomedical knowledge from numerous sources and disciplines. The prototype allows users to evaluate, showing the practicality, and scalability of the current semantic tools.

Chisham et al. [75] developed an RDF-based store framework, called CDAO, to store phylogenetic data. The developed framework provides a web service to allow programmers to access data in the store. The framework also contains a friendly user interface and visualization capabilities, which allows users to execute different kinds of domain-specific queries and view the results. CDAO also has the capability to import different formats such as PHYLIP, MEGA, NeXML, and NEXUS and store them in the store.

HL7 (Health Level 7) is a set of standards that were developed to enable the exchange of data between different healthcare information systems. On the other side, IEEE 1451 is a set of standards that are used in the context of sensor data to enable communication between different transducers. HL7 and IEEE 1451 have a different format, which makes the process of integration difficult between them. Kim et al. [76] implemented a simple software interface engine that can send and receive messages in IEEE 1451 and HL7 formats and provide interoperability between the two formats.

In the context of biological data science, computational analyses require integrating numerous datasets coming from different sources to enable data analysis and knowledge discovery. Garcia et al. [56] built a semantic web-based system called LinkHub, which extracts the graph of relationships between biological entities that are stored in different data sources including both RDF and relational datasets. The system provides various interfaces to interact with and query this graph data.

Integrating continuously changing biosciences data sources (because their schema changes over time) was studied by Marenco et al. [77] who developed a framework called Query Integrator System (QIS). The framework is based on an ontology server that maps data elements to concepts in an ontology. It includes different tools and utilities such as a graphical interface to design a distributed query.

Cheung et al. [78] implemented a web-based prototype that allows interoperability between different types of yeast genome data that have different formats. The prototype uses a native RDF database to store the integrated data generated from various heterogeneous data sources. It supports the mapping and conversion process of relational databases to RDF format. It also supports the retrieval of data from the RDF database stored using RDF-based queries.

5. Semantics for Healthcare Big Data Storage

The number of healthcare RDF data collections surpasses billions of triples and keeps continuously increasing beyond the performance capacity of traditional RDF management systems running on a single machine. In the last few years, big data techniques started to inspire researchers to develop new distributed methods to handle this large amount of emerging data. Big data management techniques can assist healthcare organizations to build knowledge-based systems to extract and infer meaningful insights from different data sets [79].

Most of the semantic techniques which are used in healthcare data storage are based on using the MapReduce paradigm, which was widely used to store and query healthcare data stored in Hadoop Distributed File System (HDFS). MapReduce is a distributed framework that is intended to scale up from one single machine to thousands of machines and to handle considerable amounts of data using simple programming models. Data is stored as RDF triple graph and SPARQL query language is used to retrieve and query the RDF store.

Rohloff and Schantz [80] used a MapReduce framework to implement a triple-store, called SHARD, which was built on top of Hadoop and deployed in Amazon EC2 cloud. SHARD was designed to be distributed and scalable and has the ability to store and query datasets with billions of triples. It stores the data as RDF triple graph and query data using SPARQL query language. Husain et al. [81] presented a framework which is able to handle a large amounts of RDF data stored in Hadoop Distributed File System (HDFS), a highly fault-tolerant system. A greedy approach algorithm was introduced to generate a query plan for answering SPARQL queries using Hadoop's MapReduce framework.

HBase is a column-oriented nonrelational distributed database, which is used to store RDF triples. Jena is a Java

framework which was used to provide a programmatic interface for RDF, SPARQL, and ontologies. Khadilkar et al. [82] presented Jena-HBase which uses HBase and Jena framework to manage RDF datasets. The implemented framework supports end-users with APIs to store, query, and reason over large amounts of RDF data in a cloud-based environment.

Zeng et al. [83] introduced a distributed in-memory RDF management system, called Trinity.RDF, that physically models and stores RDF data in its native graph form. The proposed system is scalable and has the ability to handle large amounts of RDF data that could reach trillions of triples. Because the data is stored as a graph, the system supports large scope of complex graph analytics on RDF data and processes SPARQL queries efficiently.

Galarraga et al. [84] implemented Partout, a distributed framework to manage a large volume of RDF triples in a cluster of machines. The framework depends on dividing RDF triples into fragments, and based on the query log, assigning the fragments that are used with each other to the same node in the cluster. There is a central coordinator node that is accountable for distributing the RDF triples among the hosts and running SPARQL queries with the best execution plan.

Hose and Schenkel [85] presented a distributed SPARQL engine, called WARP, for a large scale of RDF datasets. The proposed engine is based on partitioning and replication of RDF triples across cluster nodes, enabling efficient query execution. Huang et al. [86] presented a scalable management system for RDF data which is based on the Hadoop MapReduce framework. The proposed architecture is based on using a graph partitioning algorithm to store triples that are close to each other on the same machine in the cluster. Partitioning RDF graphs in this way will reduce the amount of network communication greatly during query execution.

Lee and Liu [87] introduced a semantic hash partitioning framework, called SHAPE, which is based on applying a bassline hash partition on the RDF graph, and then replicating the vital triples only to improve the utilization of data access locality. The proposed method reduced the query execution time significantly by minimizing the internode communication during the query distributed processing.

Quilitz and Leser [88] implemented an engine, called DARQ, for querying data from multiple distributed RDF repositories using a simple interface, transparent to end-users. The implementation is based on using the Service Description Language (SDL), which is a set of standards provided by the semantic community. SDL enables the query engine to decompose SPARQL queries into subqueries, which will then be sent to the individual data sources and the results are integrated back to end-users.

Saleem et al. [89] implemented a comprehensive and open repository for storing biomedical information that will be used to categorize genetic mutations responsible for cancer. Their approach is based on transforming their data (more than 20 billion triples) into an RDF format, which will be distributed across multiple SPARQL endpoints. The data can be queried using a federated SPARQL query processing engine.

Harth et al. [90] presented an approach for executing queries over linked data, published as RDF triples. Their framework is based on building an index structure, called QTree, to summarize the content of RDF graph-structured source data. The generated index structure was used to answer conjunctive queries over linked data in an efficient way.

Vocabulary of Interlinked Datasets (VoID) is a RDF schema for representing metadata about RDF datasets. It is designed by the semantic web community to simplify the process of publishing, discovering, and querying on a graph of interlinked datasets [91]. Görlitz and Staab [92] proposed a framework, called SPLENDID, for querying distributed RDF data using statistical information obtained from the VoID. SPLENDID uses the generated statistics to build a good query execution plan to improve the SPARQL query execution performance.

Hammad and Banikhalaf [93] introduced a parallelization framework for storing and querying XML-based healthcare medical records in a distributed XML repository. They used MapReduce to execute certain types of queries called containment queries. Schwarte et al. [94] presented FedX, a framework that incorporates parallelization techniques to enable efficient querying of multiple distributed heterogeneous RDF datasets. During query execution, subqueries are generated and the execution plan is evaluated at the relevant node machines, reducing the remote intermediate requests, and, consequently, improving the query performance. The retrieved partial results are aggregated locally and returned to end-users.

Min et al. [95] used edge computing to propose a reinforcement learning offloading scheme which enables healthcare IoT devices to improve their computation performance while preserving user privacy. Their schema used a Dyna architecture to accelerate the learning speed process of the healthcare IoT devices.

6. Challenges and Future Research Directions

Research in the healthcare domain has grown fundamentally; many challenging problems remain to be solved regarding dealing with healthcare big data semantically. In this section, we feature some research challenges and opportunities to help different scientists and researchers know the issues that need to be solved and investigated in the healthcare domain. The challenges and opportunities are featured as follows:

(i) Data quality considerations: when it comes to big data, it is not just about volume; the successful semantic management systems must maintain the following five characteristics during the data integration step: accuracy, completeness, reliability, relevance, and timeliness. Unfortunately, most of the existing semantic healthcare systems were developed without maintaining all these quality characteristics during the data integration step. Moreover, these solutions were created and developed targeting an answer for a specific domain of

Table 4: References are organized by data acquisition, data integration, data storage, and semantic techniques.

Reference	Data acquisition	Data integration	Data storage	Semantic techniques	Solution
Ding et al. [48]	X	X	—	X	A semantic web portal that allows patients to access all their medical information
Zhao et al. [50]	X	X	X	X	An infrastructure prototype for modeling, capturing, and querying RDF provenance genomics data.
Forkan et al. [52]	X	—	X	X	A cloud-based solution to gather and process data in AAL systems.
Jiang et al. [53]	X	X	—	—	A context-awareness system to handle a large amount of generated data from wearable healthcare devices.
Tilahun et al. [54]	—	X	X	X	Linked open data (LOD) to publish and connect public heterogeneous health data.
Dumontier et al. [55]	X	—	X	X	Knowledge base structured to capture pharmacogenomics information and store them in XML format.
Garcia et al. [56]	X	X	—	X	Used ontologies to capture and integrate neuroradiological data.
Ullah et al. [57]	—	X	X	X	Integrated different healthcare data and used SPARQL to query the data from the RDF graphs.
Yoon et al. [37	X	—	—	X	Proposed DiTex to extract disease-related topics using natural language processing and semantic similarity ranking algorithms.
Pacaci et al. [59]	X	—	—	—	Used semantic techniques to extract data from electronic health record systems and store the data in RDF
Ethier et al. [62]	—	X	—	X	Developed TRANSFoRm project which uses ontologies to integrate biomedical datasets
Yu et al. [66]	X	—	X	X	Used kafka to collect healthcare stream data and store it in the NoSQL database.
Mate et al. [67]	X	—	—	X	Used ontologies to automatically generate SQL statements during the data extraction process for medical data.
Schoppenhauer et al. [68]	X	X	X	X	Developed a model that maps biomedicine classes and relationships to database entries. This model was used to extract data and store it as RDF triples.
Livingston et al. [69]	—	X	X	X	Devolved a knowledge base system called KaBOB to integrate data from many different biomedical datasets and store the results in RDF triples.
Liang et al. [72]	—	X	—	X	Developed an ontology that integrates data from genes, symptoms, diseases, and phenotype.
Chisham et al. [75]	—	—	X	X	Developed an RDF-based store framework, called CDAO, to store phylogenetic data.
Rohloff et al. [80]	—	—	X	X	Used a MapReduce framework to implement a distributed scalable RDF triple-store, called SHARD.
Galarraga et al. [84]	—	—	X	X	Implemented partout, a distributed framework to manage a large volume of RDF triples in a cluster of machines.

medicinal services with a shortage of providing a complete and comprehensive semantic healthcare solution. Further research is required regarding assessing the quality of healthcare system. To build up credibility, healthcare data are progressively supposed to show that it has the required quality by automating the process of data quality assessment.

(ii) Semantic metadata management: metadata management includes setting up strategies and cycles that guarantee data can be incorporated, referenced, shared, connected, investigated, and kept up to meet the interests of the corporation's stakeholders. Metadata is produced every time new data is generated, modified, or deleted. Metadata management for healthcare systems faces many challenges and issues that need to be resolved. Healthcare big data processing engine should have the capability to automate the process of discovering and selecting the most appropriate data sources based on predefined business rules. Different datasets may store the same data in multiple locations even if they have the same semantic meaning. The semantic processing engine should have the ability to select the right datasets based on different criteria and dimensions, such as query performance and data quality: accuracy, completeness, currency, and consistency.

(iii) Managing uncertainty in healthcare big data: the massive amount of data collected from different healthcare sources such as wearable devices and online social media naturally contains a certain amount of uncertainty because of noise, irregularity,

and missing information. On the other hand, the process of analyzing such collected uncertain data requires progressed explanatory methods for effectively anticipating future blueprints with high exactness and progressed dynamic procedures. Leaving the data with uncertainty will cause stakeholders to untrust any future results based on this data. Unfortunately, little work has been done in the area of uncertainty with semantic healthcare big data. New methods and procedures based on machine learning (ML) and natural language processing (NLP) must be introduced to design uncertainty data models. Integrating ML and NLP with semantic healthcare big data should have the ability to give more exact quicker, and adaptable outcomes for analyzers and decision-makers.

(iv) Securing healthcare data using blockchain: precise and complete healthcare data are one important resource for patients. Any changes in data by cyber-attackers can cause dangerous health problems, such as giving incorrect medicines or therapies to patients. Therefore, the protection of privacy and securing medical data records have consistently been a worry for everybody during healthcare data administration. The rise of blockchain innovation carries novel ideas to secure healthcare data which are stored on the host servers. Further research is required since blockchain requires a lot of computational resources to create blocks, which is inapplicable for medical sensor devices.

7. Summary and Conclusion

The healthcare community is constantly generating massive, high-speed, heterogeneous, and disparate data that includes structured, semistructured, and unstructured. Many challenges are facing the process of managing healthcare data such as volume, velocity, and variety. Unfortunately, traditional information systems are not capable of exploiting such data powerfully and efficiently.

In the previous sections, we have reviewed existing solutions found in the literature related to semantic healthcare big data management from different perspectives including data acquisition, data integration, and data storage. We summarize our findings in Table 4.

Semantic web technologies have the opportunity to transform the way healthcare providers utilize technology to gain insights and knowledge from their data and make decisions. Both big data and semantic web technologies can complement each other to address the previous challenges and add intelligence to healthcare management systems. This paper reviews some of those challenges and discusses the role of semantic Web technologies in healthcare management systems. The review was conducted from four viewpoints. First, we reviewed some of the semantic ontologies which have been created and used by the healthcare community. Second, we reviewed the role of semantic technologies in extracting and transforming healthcare data before storing it in repositories. Third, we conducted a review of the different approaches for integrating heterogeneous healthcare data. Finally, we reviewed the different semantic methods and approaches for storing healthcare data in distributed environments.

In our survey, we found some issues and challenges that need to be answered regarding dealing with healthcare big data semantically. We reported these limitations and challenges, and we discussed some of the potential research directions. We intend to lead an exploratory study with some of the potential solutions for managing big healthcare data semantically [96].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] What is Healthcare Big Data?, 2006, https://www.evariant.com/faq/what-is-healthcare-big-data.

[2] V. Jirkovský and M. Obitko, "Semantic heterogeneity reduction for big data in industrial automation," *ITAT*, vol. 1214, 2014.

[3] P. Russom, "Big data analytics," *TDWI best practices report*, vol. 19, no. 4, pp. 1–34, 2011.

[4] D. Laney, "3D data management: controlling data volume, velocity and variety," *META Group Research Note*, vol. 6, no. 70, p. 1, 2001.

[5] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, vol. 65, 2016.

[6] A. Gandomi and M. Haider, "Beyond the hype: big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.

[7] X. Zenuni, B. Raufi, F. Ismaili, and J. Ajdari, "State of the art of semantic web for healthcare," *Procedia-Social and Behavioral Sciences*, vol. 195, pp. 1990–1998, 2015.

[8] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *Journal of Big Data*, vol. 6, no. 1, p. 54, 2019.

[9] R. De Santis, A. Gloria, S. Viglione et al., "3D laser scanning in conjunction with surface texturing to evaluate shift and reduction of the tibiofemoral contact area after meniscectomy," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 88, pp. 41–47, 2018.

[10] T. Deserno, *Medical Image Processing:Optipedia*, SPIE Press, Bellingham, WA, 2009.

[11] J. Ahmed and M. Ahmed, "Semantic web approach of integrating big data- A review," *International Journal of Computer Sciences and Engineering*, vol. 6, pp. 529–532, 2018.

[12] A. Taouli, a. b. Djamel, N. Keskes, K. Bencherif, and B. Hassan, "Semantic for big data analysis: a survey," in *Proceedings of the INTIS2018: BigData & Internet of things IoT*, Marrakech, Maroc, December 2018.

[13] The datapine Blog, [online]. Available: . Available: https://www.datapine.com/blog/big-data-examples-in-healthcare/. Visited 26 April 2020.

[14] I. Merelli, H. Pérez-Sánchez, S. Gesing, and D. D'Agostino, "Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives," *BioMed Research International*, vol. 2014, Article ID 134023, 13 pages, 2014.

[15] XML (extensible markup language), [online]. Available: https://whatis.techtarget.com/definition/XML-Extensible-Markup-Language (Visited 19 August 2020).
[16] XML RDF, [online]. Available: https://www.w3schools.com/xml/xml_rdf.asp (Visited 19 August 2020).
[17] SPARQL, [online]. Available: https://en.wikipedia.org/wiki/SPARQL (Visited 19 August 2020).
[18] Web ontology language (OWL), [online]. Available: https://www.w3.org/OWL/(Visited 19 August 2020).
[19] V. Srinivasan, B. Bulkowski, W.-L. Chu et al., "Aerospike," *Proceedings of the VLDB Endowment*, vol. 9, no. 13, pp. 1389–1400, 2016.
[20] C. A. Baron, "NoSQL key-value DBs riak and redis," *Database Systems Journal*, vol. 6, no. 4, pp. 3–10, 2016.
[21] Z. Cao, S. Dong, S. Vemuri, and D. H. Du, "Characterizing, modeling, and benchmarking RocksDB key-value workloads at Facebook," in *Proceedings of the 18th {USENIX} Conference on File and Storage Technologies ({FAST} 20)*, pp. 209–223, Santa Clara, CA, USA, February 2020.
[22] N. D. Bhardwaj, "Comparative study of couchdb and mongodb–nosql document oriented databases," *International Journal of Computer Applications*, vol. 136, no. 3, pp. 24–26, 2016.
[23] R. M. Lerner, "At the forge: PostgreSQL, the NoSQL database," *Linux Journal*, vol. 247, 2014.
[24] D. Fernandes and J. Bernardino, "Graph databases comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB," in *Proceedings of the 7th International Conference on Data Science, Technology and Applications*, pp. 373–380, DATA, Porto, Portugal, January 2018.
[25] M. N. Vora, "Hadoop-HBase for large-scale data," in *Proceedings of the 2011 International Conferenceon Computer Science and Network Technology*, vol. 1, pp. 601-605, IEEE, Bangalore, India, December 2011.
[26] E. Dede, B. Sendir, P. Kuzlu, J. Hartog, and M. Govindaraju, "An evaluation of cassandra for hadoop," in *Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing*, pp. 494–501, IEEE, Santa Clara, CA, USA, June 2013.
[27] J. Kepner, W. Arcand, B. Bill et al., "Lustre, hadoop, accumulo," in *Proceedings of the 2015 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–5, IEEE, Waltham, USA, July 2015.
[28] R. Cattell, "Scalable SQL and NoSQL data stores," *Acm Sigmod Record*, vol. 39, no. 4, pp. 12–27, 2011.
[29] What is semantic data?, [online]. Available: . Available: http://www.semagix.com/what-is-semantic-data.htm. Visited 26 April 2020.
[30] Semantic data model, [online]. Available: . Available: https://www.techopedia.com/definition/30489/semantic-data-model. Visited 26 April 2020.
[31] What are Ontologies?, [online]. Available: https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/(Visited 26 April 2020).
[32] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, "The anatomy of phenotype ontologies: principles, properties and applications," *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 1008–1021, 2018.
[33] D. C. Faye, O. Curé, and G. Blin, "A survey of RDF storage approaches," *ARIMA Journal*, vol. 15, pp. 11–35, 2012.
[34] F. Gandon, "A survey of the first 20 years of research on semantic Web and linked data," *Ingénierie Systèmes Information*, vol. 23, 2018.
[35] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, 2014.
[36] L. Wang and C. A. Alexander, "Big data in medical applications and health care," *American Medical Journal*, vol. 6, no. 1, p. 1, 2015.
[37] C. He, X. Jin, Z. Zhao, and T. Xiang, "A cloud computing solution for hospital information system," in *Proceedings of the 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 2, pp. 517-520, IEEE, Xiamen, China, December 2010.
[38] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. S. Iyengar, "Computational health informatics in the big data age," *ACM Computing Surveys*, vol. 49, no. 1, pp. 1–36, 2016.
[39] BioPortal, [online]. Available: https://bioportal.bioontology.org/(Visited 26 April 2020).
[40] MedDRAMSSO, [online]. Available: https://www.meddra.org/about-meddra/organisation/msso (Visited 26 April 2020).
[41] Human disease ontology, [online]. Available: https://bioportal.bioontology.org/ontologies/DOID (Visited 26 April 2020).
[42] Physical Medicine and Rehabilitation, [online]. Available: https://bioportal.bioontology.org/ontologies/PMR (Visited 26 April 2020).
[43] The human phenotype ontology, [online]. Available: https://www.human-phenotype-ontology.org (Visited 26 April 2020).
[44] Anatomical therapeutic chemical classification, [online]. Available: http://www.human-phenotype-ontology.org (Visited 26 April 2020).
[45] International Classification of Functioning, "Disability and health," 2020, http://www.who.int/classifications/icf/en/.
[46] K. Lyko, M. Nitzschke, and A.-C. N. Ngomo, "Big data acquisition," in *New Horizons for a Data-Driven Economy*, pp. 39–61, Springer, Berlin, Germany, 2016.
[47] X. Zhang, Y. Hu, K. Xie, W. Zhang, L. Su, and M. Liu, "An evolutionary trend reversion model for stock trading rule discovery," *Knowledge-Based Systems*, vol. 79, pp. 27–35, 2015.
[48] Y. Ding, B. Chen, Y. Ding et al., "Semantic web portal: a platform for better browsing and visualizing semantic data," in *Proceedings of the International Conference on Active Media Technology*, pp. 448–460, Springer, Berlin, Heidelberg, New York, August 2010.
[49] R. Hammad and C.-S. Wu, "Provenance as a service: a data-centric approach for real-time monitoring," in *Proceedings of the 2014 IEEE International Congress on Big Data*, pp. 258–265, IEEE, Washington DC, USA, June 2014.
[50] J. Zhao, S. S. Sahoo, P. Missier, A. Sheth, and C. Goble, "Extending semantic provenance into the web of data," *IEEE Internet Computing*, vol. 15, no. 1, pp. 40–48, 2010.
[51] K.-H. Cheung, E. Prud'hommeaux, Y. Wang, and S. Stephens, *Semantic Web for Health Care and Life Sciences: A Review of the State of the Art*, Oxford University Press, Oxford, UK, 2009.
[52] A. Forkan, I. Khalil, and Z. Tari, "CoCaMAAL: a cloud-oriented context-aware middleware in ambient assisted living," *Future Generation Computer Systems*, vol. 35, pp. 114–127, 2014.
[53] P. Jiang, J. Winkley, C. Zhao, R. Munnoch, G. Min, and L. T. Yang, "An intelligent information forwarder for healthcare big data systems with distributed wearable sensors," *IEEE Systems Journal*, vol. 10, no. 3, pp. 1147–1159, 2014.

[54] B. Tilahun, T. Kauppinen, C. Keßler, and F. Fritz, "Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation," *JMIR Medical Informatics*, vol. 2, no. 2, p. e31, 2014.

[55] M. Dumontier and N. Villanueva-Rosales, "Towards pharmacogenomics knowledge discovery with the semantic web," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 153–163, 2009.

[56] A. Garcia, Z. Zhang, M. Rajapakse, C. Baker, and S. Tang, "Capturing and modeling neuro-radiological knowledge on a community basis: The head injury scenario," in *Proceedings of the Health and Life Sciences workshop at the WWW2008*, Beijing, China, April 2008.

[57] F. Ullah, M. A. Habib, M. Farhan, S. Khalid, M. Y. Durrani, and S. Jabbar, "Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare," *Sustainable Cities and Society*, vol. 34, pp. 90–96, 2017.

[58] J. Yoon, J. W. Kim, and B. Jang, "DiTeX: disease-related topic extraction system through internet-based sources," *PloS One*, vol. 13, no. 8, Article ID e0201933, 2018.

[59] A. Pacaci, S. Gonul, A. A. Sinaci, M. Yuksel, and G. B. Laleci Erturkmen, "A semantic transformation methodology for the secondary use of observational healthcare data in post-marketing safety studies," *Frontiers in Pharmacology*, vol. 9, p. 435, 2018.

[60] B. Elsharkawy, H. Ahmed, and R. Salem, "Semantic-based approach for solving the heterogeneity of clinical data," *IJCI. International Journal of Computers and Information*, vol. 5, no. 1, pp. 35–45, 2016.

[61] H. Zhang, Q. Li, G. Yi et al., "An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival," *BMC Medical Informatics and Decision Making*, vol. 18, no. 2, p. 41, 2018.

[62] J. F. Ethier, M. McGilchrist, A. Barton et al., "The TRANSFoRm project: Experience and lessons learned regarding functional and interoperability requirements to support primary care," *Learning Health Systems*, vol. 2, no. 2, p. e10037, 2018.

[63] R. M. Keller, S. Ranjan, M. Y. Wei, and M. M. Eshow, "Semantic representation and scale-up of integrated air traffic management data," in *Proceedings of the International Workshop on Semantic Big Data*, pp. 1–6, California, USA, June 2016.

[64] HBase Tutorial for Beginners, [online]. Available: https://www.guru99.com/hbase-tutorials.html (Visited 26 April 2020).

[65] L. Kang, L. Yi, and L. Dong, "Research on construction methods of big data semantic model," in *Proceedings of the World Congress on Engineering*, vol. 1, pp. 2–4, London, U.K, July 2014.

[66] H. Q. Yu and F. Dong, "Semantic lifting and reasoning on the personalised activity big data repository for healthcare research," *International Journal of Web Engineering and Technology*, vol. 14, no. 2, pp. 103–121, 2019.

[67] S. Mate, K. Felix, T. Dennis et al., "Ontology-based data integration between clinical and research systems," *PloS One*, vol. 10, no. 1, Article ID e0116656, 2015.

[68] A.-K. Kock-Schoppenhauer, C. Kamann, H. Ulrich, P. Duhm-Harbeck, and J. Ingenerf, "Linked data applications through ontology based data access in clinical research," *Studies in Health Technology and Informatics*, vol. 235, pp. 131–135, 2017.

[69] K. M. Livingston, M. Bada, W. A. Baumgartner, and L. E. Hunter, "KaBOB: ontology-based semantic integration of biomedical databases," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–21, 2015.

[70] M. Manning, A. Aggarwal, K. Gao, and G. Tucker-Kellogg, "Scaling the walls of discovery: using semantic metadata for integrative problem solving," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 164–176, 2009.

[71] H. Issa, L. van Elst, and A. Dengel, "Using smartphones for prototyping semantic sensor analysis systems," in *Proceedings of the International Workshop on Semantic Big Data*, pp. 1–6, San Francisco, California, June 2016.

[72] C. Liang, J. Sun, and C. Tao, "Semantic web ontology and data integration: a case study in aiding psychiatric drug repurposing," *Studies in health technology and informatics*, vol. 216, p. 1051, 2015.

[73] C. A. Knoblock and P. Szekely, "Semantics for big data integration and analysis," in *Proceedings of the 2013 AAAI Fall Symposium Series*, Arlington, Virginia, USA, November 2013.

[74] A. Ruttenberg, J. A. Rees, M. Samwald, and M. S. Marshall, "Life sciences on the semantic web: the Neurocommons and beyond," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 193–204, 2009.

[75] B. Chisham, B. Wright, T. Le, T. Son, and E. Pontelli, "CDAO-store: ontology-driven data integration for phylogenetic analysis," *BMC Bioinformatics*, vol. 12, no. 1, p. 98, 2011.

[76] W. Kim, S. Lim, J. Ahn, J. Nah, and N. Kim, "Integration of IEEE 1451 and HL7 exchanging information for patients' sensor data," *Journal of Medical Systems*, vol. 34, no. 6, pp. 1033–1041, 2010.

[77] L. Marenco, T.-Y. Wang, G. Shepherd, P. L. Miller, and P. Nadkarni, "QIS: a framework for biomedical database federation," *Journal of the American Medical Informatics Association*, vol. 11, no. 6, pp. 523–534, 2004.

[78] K.-H. Cheung, K. Y. Yip, A. Smith, R. Deknikker, A. Masiar, and M. Gerstein, "YeastHub: a semantic web use case for integrating data in the life sciences domain," *Bioinformatics*, vol. 21, no. 1, pp. i85–96, 2005.

[79] N. M. Elzein, M. A. Majid, I. A. T. Hashem, I. Yaqoob, F. A. Alaba, and M. Imran, "Managing big RDF data in clouds: challenges, opportunities, and solutions," *Sustainable Cities and Society*, vol. 39, pp. 375–386, 2018.

[80] K. Rohloff and R. E. Schantz, "High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store," in *Proceedings of the Programming Support Innovations for Emerging Distributed Applications*, pp. 1–5, Reno/Tahoe, Nevada, USA, October 2010.

[81] M. Husain, J. McGlothlin, M. M. Masud, L. Khan, and B. M. Thuraisingham, "Heuristics-based query processing for large RDF graphs using cloud computing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 9, pp. 1312–1327, 2011.

[82] V. Khadilkar, M. Kantarcioglu, B. Thuraisingham, and P. Castagna, "Jena-HBase: a distributed, scalable and efficient RDF triple store," in *Proceedings of the 11th International Semantic Web Conference Posters & Demonstrations Track, ISWC-PD*, vol. 12, pp. 85–88, Bethlehem, USA, January 2012.

[83] K. Zeng, J. Yang, H. Wang, B. Shao, and Z. Wang, "A distributed graph engine for web scale RDF data," in *Proceedings of the VLDB Endowment*, vol. 6, no. 4, pp. 265–276, Riva del Garda, Trento, Italy, February 2013.

[84] L. Galárraga, K. Hose, and R. Schenkel, "Partout: a distributed engine for efficient RDF processing," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 267-268, Seoul, Republic of Korea, April 2014.

[85] K. Hose and R. Schenkel, "WARP: workload-aware replication and partitioning for RDF," in *Proceedings of the 2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pp. 1–6, IEEE, Brisbane, Australia, April 2013.

[86] J. Huang, D. J. Abadi, and K. Ren, "Scalable SPARQL querying of large RDF graphs," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 1123–1134, 2011.

[87] K. Lee and L. Liu, "Scaling queries over big RDF graphs with semantic hash partitioning," *Proceedings of the VLDB Endowment*, vol. 6, no. 14, pp. 1894–1905, 2013.

[88] B. Quilitz and U. Leser, "Querying distributed RDF data sources with SPARQL," in *Proceedings of the European Semantic Web Conference*, pp. 524–538, Springer, Tenerife, Canary Islands, Spain, January 2008.

[89] M. Saleem, S. S. Padmanabhuni, A.-C. Ngonga Ngomo et al., "TopFed: TCGA tailored federated query processing and linking to LOD," *Journal of Biomedical Semantics*, vol. 5, no. 1, p. 47, 2014.

[90] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich, "Data summaries for on-demand queries over linked data," in *Proceedings of the 19th International Conference on World Wide Web*, pp. 411–420, Raleigh, North Carolina, USA, January 2010.

[91] Describing linked datasets with the VoID vocabulary, [online]. Available: https://www.w3.org/TR/void/(Visited 26 April 2020).

[92] O. Görlitz and S. Staab, "Splendid: sparql endpoint federation exploiting void descriptions," in *Proceedings of the Second International Conference on Consuming Linked Data*-, vol. 782, pp. 13–24, CEUR-S. org, Bonn, Germany, July 2011.

[93] R. Hammad and M. Banikhalaf, "A parallel approach for managing XML-based electronic medical records," in *Proceedings of the 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–5, IEEE, Aqaba, Jordan, October 2018.

[94] A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt, "Fedx: optimization techniques for federated query processing on linked data," in *Proceedings of the International Semantic Web Conference*, pp. 601–616, Springer, Bonn, Germany, October 2011.

[95] M. Min, X. Wan, L. Xiao et al., "Learning-based privacy-aware offloading for healthcare IoT with energy harvesting," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4307–4316, 2018.

[96] NoSQL Database: About Quality Attributes. Understanding first before choosing, [online]. Available: http://www.tisa-software.com/news/blog/219-nosql-database-about-quality-attributes-understanding-first-before-choosing (Visited 26 April 2020).