Received June 29, 2019, accepted July 4, 2019, date of publication July 9, 2019, date of current version July 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927491

In Search of Big Medical Data Integration Solutions - A Comprehensive Survey

HOUSSEIN DHAYNE[1], (Student Member, IEEE), RAFIQUL HAQUE[2], RIMA KILANY[1], AND YEHIA TAHER[3]

[1]Faculty of Engineering, ESIB, Saint Joseph University, Beirut 1104, Lebanon
[2]Cognitus, 75008 Paris, France
[3]David Laboratory, 78035 Versailles, France

Corresponding author: Rafiqul Haque (rafiqul.haque@cognitus.fr)

ABSTRACT In recent years, the radical advancement of technologies has given rise to an abundance of software applications, social media, and smart devices such as smartphone, sensors, and so on. More extensive use of these applications and tools in various industrial domains has led to data deluge, which has fostered enormous challenges and opportunities. However, it is not only the volume of the data but also the speed, variety, and uncertainty, which are promoting a massive challenge for traditional technologies such as data warehouse. These diverse and unprecedented characteristics have engendered the notion of ''Big Data.'' The data-intensive industries have been experiencing a wide variety of challenges in terms of processing, managing, and analysis of data. For instance, the healthcare sector is confronting difficulties in respect of integration or fusion of diverse medical data stemming from multiple heterogeneous sources. Data integration is critically important within the healthcare sector because it enriches data, enhances its value, and more importantly paves a solid foundation for highly efficient and effective healthcare analytics such as predicting diseases or an outbreak. Several data integration technologies and tools have been developed over the last two decades. This paper aims at studying data integration technologies, tools, and applications within the healthcare domain. Furthermore, this paper discusses future research directions in the integration of Big healthcare data.

INDEX TERMS Big data, data integration, healthcare data.

I. INTRODUCTION

Healthcare is a highly data-intensive industry [1]. The ever-increasing trend of healthcare data has already led to a massive growth of the volume. It was predicted that the data of the U.S.A healthcare sector alone would soon reach the Zettabyte scale and, not long after, the Yottabyte [2]. The increased usage of the term *Big Data* in healthcare literature is also an indicator of the emerging importance of large-scale data sets in healthcare and biomedicine [3], and there is also an increasing awareness of the role that Big data can play in scientific and clinical research [4].

The exponential growth in healthcare data has been forecasted to continue expanding in various forms, such as electronic health records (EHR), patient-reported outcomes, biometric data, medical imaging, biomarker data, wearable devices, and genomic information. These data are primarily stemming from multiple heterogeneous sources. Figure 1 shows some of the primary sources of healthcare data, which include medical service providers, pharmaceutical industries, public healthcare organizations, researchers, and medical insurance *etc*. The integration of such vast, real world, clinical data sets from Electronic Medical Record (EMR) with omics data, as well as targeted biochemical and hormonal analyzes, makes it possible to discover new diagnostic and therapeutic tools [5] as well as capture the full complexity of diseases [6]. In one elegant example [7], the integration of continuous sensing of blood-glucose along with the evaluation of the gut microbiome, anthropometrics, drugs, dietary habits, and a variety of lab tests on 800 individuals, as illustrated in Figure 2, was used to predict postprandial glycemic index, which has provided accurate information on dietary regimens to improve metabolic homeostasis.

Feldman *et al.* [8] presents multiple unintegrated medical data pools controlled by six stakeholders: providers, payers, researchers, developers, consumers and marketers,

The associate editor coordinating the review of this manuscript and approving it for publication was Zehong Cao.

FIGURE 1. Sources of big data in healthcare.

and government. These sources have been producing healthcare data for many years; consequently, the storage of healthcare data has become an ever-increasing container. Furthermore, lately, the social media mainly, the Twitter[1] has been explored as a data source that contains many different types of data of value to healthcare research on many various diseases [9]. These data are insightful, and therefore, Twitter is becoming a vital source of healthcare data. However, it is worth noting that Twitter data flow with high velocity (speed).[2]

Data stemming from multiple heterogeneous sources increase data *variety* and *uncertainty*. Uncertainty is concerned with the quality of data. The quality of healthcare data is of critical importance to perform effective data analysis to extract meaningful intelligence that helps in decision making. The assessment of the quality of evidence to be derived is crucial; it will depend on the data sources to be integrated such as social network [11] or public repositories [12], together with the standard quality indicators as selection bias, sample size, and measurement noise [13]. Variety is a well-known issue in Big Data. A wide range of unstructured data is available in the healthcare sector, including MRI image, surgical video, text, recorded conversation with patients. Also, there are structured data, such as EHR data. This essentially means that the technologies used in healthcare domain must be able to process data with diverse types.

In the light of the above discussion, we outline that healthcare data has four properties volume, huge velocity, significant variety, and substantial uncertainty (veracity) which constitute the notion of *Big Healthcare Data*. These characteristics foster a wide variety of challenges or barriers for the users of healthcare data and also for the technologies that are used in the healthcare domain. The major challenges involved in Big Healthcare Data include: *data integration*, *data processing* and *anlaysis*. In fact, for healthcare data, integration is a huge obstacle, mainly due to the variety and velocity of data. According to Martin *et al.* integrating unstructured data is a huge challenge for the Big Data analyst [4]. Even with large scale structured EHR data, there are still many integration issues [14].

Over the last decade, an exhaustive number of Big Data tools and approaches have been proposed. Many solutions are available for dealing with significant data issues and challenges. These solutions have been reviewed in a large body of literature. For example, Merelli *et al.* [15] focus on technological aspects related to Big Data analysis in biomedical informatics including architectural approaches for big data, solutions for data annotation, data access measures, and security for health data. Priyanka and Kulennavar [16] discuss the definition of big data and characteristics of big data analytics in healthcare, and describe various sources and data types. Luo *et al.* [17] review the recent progress and breakthroughs of big data applications in four healthcare domains: public health informatics, clinical informatics, bioinformatics, and imaging informatics. Jee and Kim [18] explain how reforming the healthcare system based on big data analytics, could effectively reduce health concerns such as the selection of appropriate treatment paths, improvement of health systems, etc. However, big data analytics in healthcare require data integration to be successful. Unfortunately, the issues of healthcare data integration and utility have been largely overlooked in the current literature [19].

Lenz *et al.* [20] review and discuss integration technologies in healthcare systems, where they identify technological integration (based on technological infrastructure) and semantic integration (based on the meaning of the data). Authors propose a document-based approach to support integration in healthcare networks. Zhang *et al.* [21] discusses several approaches proposed for data integration in bioinformatics, which is classified into five groups: data warehousing, federated databasing, service-oriented integration, semantic integration, and wiki-based integration. As Big data is often heterogeneous, noisy, unreliable, and dynamic; in this context, these traditional approaches do not apply to the non-relational, schema-less dataset.

What we instead aim to provide in this paper is a comprehensive survey of Big Data integration from different standpoints. We will study the developments in advanced solutions for Big healthcare data integration. We will present the design and development of varying integration strategies that are commonly adopted by the research communities. We will discuss current challenges concerning Big Healthcare data integration. Furthermore, we will discuss some possible future research directions in this area.

This remainder of the paper is organized as follows. In section II, we provide a background of concepts, tools, and technologies related to Big Healthcare Data integration.

[1]Twitter: https://twitter.com/?lang=en

[2]Every second, on average, around 6,000 tweets are tweeted on Twitter [10]

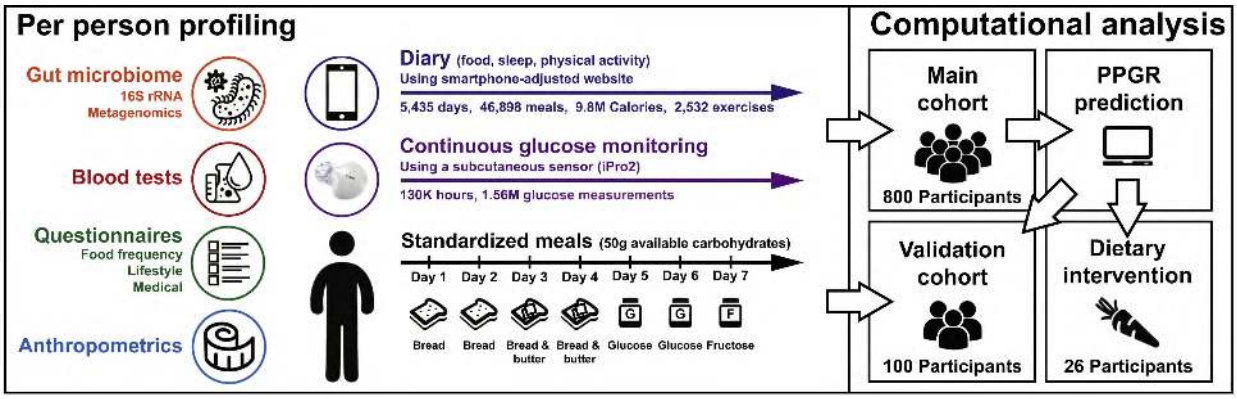

FIGURE 2. Illustration of various data sources to predict personalized nutrition [7].

In Section III, we put forward our comprehensive survey. In section IV we discuss our findings followed by research directions presented in Section V. We conclude in Section VI.

II. BACKGROUND

In this section, we study different concepts related to Big Healthcare Data integration and provide extensive details about these concepts. This will help the readers to understand the underlying challenges and techniques of data integration.

A. BIG DATA

Over the past two decades, a deluge of data has been brought with the technological advancements from several fields (e.g., medical data and scientific sensors, user-generated data, financial data, and other) [22]. Big Data has been defined in a large volume of literature. It has been defined as datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within an acceptable time [23]. However, the concept of Big data has been presented through the 3V model, which refers to high-volume, high velocity, and high-variety information assets [24]. Lately, this notion has been extended to a 5V model by including two new "Vs.": Value and Veracity which are incorporated into the Big data definition [25].

Within the Healthcare sector, various definitions of Big data are found [26]; this term was introduced in [27] as "Big data in healthcare encompasses high volume, high diversity in biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points." Several solutions have been proposed to tackle these challenges. These solutions rely on the most recent and advanced technologies: Apache Hadoop framework [28], NoSQL [29] and Cloud computing [30].

The advent of Big Data has given rise to a new concept called data lake. A data lake is a repository to store a vast amount of raw data in its native format. The term data lake is often associated with Apache Hadoop-oriented object storage. Hadoop provides these techniques through Apache Hadoop YARN and HDFS [31]. YARN presents the next generation of Hadoop compute platform and offers a pluggable architecture and resource management for data processing engines to interact with data in HDFS [32]. Data lake can be a powerful approach to resolve the problem of accessibility and integration of Big data [33].

The data lake is different from the traditional Data Warehouse from various aspects. Even though relational data warehouses have led the complex integration and analytics, their slow-changing data models and rigid field-to-field mappings are too brittle to support Big data volume and variety. By contrast, data lake approach circumvents these problems, because data lake does not enforce a rigid metadata schema as do relational data warehouses [34]. Instead, data lake support a flexible "schema-on-read" access to all enterprise data, through multi-use and multi-workload data processing on the same sets of data, from batch to real-time [35]. The data flow in the data lake aims at decoupling the metadata from the raw data and storing them separately. In this way, end-users have the potential to query data from multiple perspectives (see Figure 3).

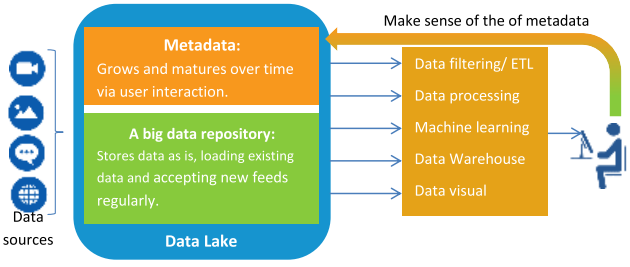

FIGURE 3. Data flow in the data lake.

However, despite the robustness and availability of Big data tools, building and deploying Big data solutions is difficult. Therefore, domain-specific solutions are still needed. These solutions often depend on different data dimensions as well as the type of data and the target to be studied.

B. DATA INTEGRATION

The goal of data integration is the provisioning of unified access to data that requires information from multiple sources

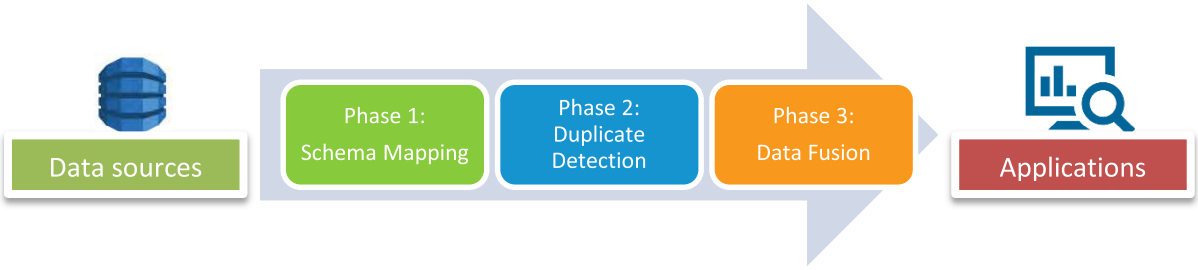

FIGURE 4. Traditional data integration process.

and providing users with a unified view of the data [36]. Ziegler and Dittrich [37] explain the reasons behind integrating multiple data sources, which are twofold:

- Facilitate information access by providing an integrated view to a set of existing information systems.
- Gain a more comprehensive basis by combining data from different complementing information systems.

The traditional data integration process is assumed to be a three-step process (figure 4) where the last step is referred to, as data fusion. In this step, the duplicate representations of data are combined and fused into a single image while inconsistencies are resolved. The two other steps are schema mapping and duplicate detection [38].

There is a wide variety of challenges regarding data integration. One of the key challenges is dealing with the problems emerging from the heterogeneity of data sources [39]. Dong and Naumann [40] introduce several challenges of data integration:

- *heterogeneity* at the schema level, where different data sources often describe the same domain using different schemas, as well as heterogeneity at the instance level, where different sources can represent the same real-world entity in different ways. Several solutions proposed to address the heterogeneity challenge, at the schema level, as well as schema mapping and matching.
- *Data conflicts* that can arise because of incomplete data, incorrect data, and out-of-date data. Data fusion addresses this second challenge by fusing records on the same real-world entity into a single record and resolving possible conflicts from different data sources.

To sum up, data integration is the process of linking and connecting systems and giving users the illusion of interacting with one single information system. This process often encompasses a fusion step. However, the integration methods of such a traditional integration process, which focus on structured data sources, need to be significantly expanded to integrate a variety of data sources, both structured and unstructured. In particular, pairwise matching of schemas and entities is not scalable enough [41].

C. HEALTHCARE STANDARDS FOR DATA INTEGRATION

Interoperability is the ability of two or more components, applications, or systems to exchange and use information. In health care, interoperability is the ability of the technologies to facilitate the integration of patient data from different systems. To achieve interoperability, healthcare organizations (management staff, vendors, service provider entities, etc.) have created Healthcare standards. Below is a list of key standards organizations relevant to the health sector [42]:

- *OpenEHR*: The objective of the OpenEHR Foundation is to leverage ICTs, in particular, life-long interoperable EHRs to improve the quality of healthcare and research. The main work of the openEHR Foundation is performed by four 'programs' which respectively focus on specifications, clinical modeling, software, and education.
- *Health Level Seven (HL7)*: It provides a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information.
- *International Health Terminology Standards Development Organisation (IHTSDO)*: determines global standards for health terms, an essential part of improving the health of humankind. It is committed to maintain and grow its leadership as the global experts in healthcare terminology, ensuring SNOMED CT, its world-leading product, is accepted as the universal common language for health terms.
- *World Health Organisation (WHO):* The WHO is the directing and coordinating authority for health within the United Nations system. It is responsible for providing leadership on global health matters, shaping the health research agenda, setting norms and standards, articulating evidence-based policy options, providing technical support to countries, and monitoring and assessing health trends.

However, data is one of the most critical aspects of the healthcare system. Therefore, interoperability of healthcare systems requires the integration of standard Data Models, Terminologies, and Messaging standards. [43].

1) HEALTHCARE DATA MODELS

The healthcare data models define the structure of the information to be stored in Electronic Health Records (EHRs). The most popular and recognized clinical data models are:

- *openEHR clinical model* [44]: openEHR clinical models are composed of archetypes and templates. An archetype is a computable specification of the data points and groups of a specific clinical topic, such as Fetal heart rate, ECG result, or diagnosis. Whereas templates are composed of elements of one or more archetypes, such as templates created specifically for the Diabetic review' and Antenatal visit.
- *HL7 Clinical Document Architecture (CDA)* [45]: It is an interchange standard for any document classified as a clinical document such as discharge summaries and evaluation or operative notes. It defines a clinical document as having the following six characteristics: Persistence, Stewardship, Potential for authentication, Context, Wholeness, and Human readability. This data model is governed by formalized reference schema HL7 Reference Information Model (RIM).
- *Fast Healthcare Interoperability Resources(FHIR)* [46]: FHIR defines a set of ''Resources'' representing granular clinical concepts. Resources can be managed in isolation, or aggregated into complex documents. Technically, FHIR is designed for the web. The resources are based on simple XML or JSON structures, with an HTTP-based RESTful protocol, in which each resource has a predictable URL.

2) HEALTHCARE TERMINOLOGY

Healthcare terminology or coding systems are structured list of terms, which provide specific codes for clinical concepts such as diseases, operations, allergies, drugs, and diagnoses. These vocabularies can be used to support the recording and reporting of patient care at different levels of detail. Examples of terminology standards are :

- *Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)* [47]: SNOMED CT is a medical terminology that includes terms of all medical domains and provides the general core terminology for the EHR. The concepts in SNOMED CT are divided into hierarchies as diverse as body structure, clinical findings, geographic location, and pharmaceutical/biological product.
- *The International Classification of Diseases (ICD)* [48]: The standard diagnostic tool for epidemiology, health management, and clinical purposes. It is used to monitor the incidence and prevalence of diseases and other health problems. There are two major revisions of ICD in use; ICD-10 and ICD-9, which are represented as entirely separate code systems.
- *Logical Observation Identifiers Names and Codes (LOINC)* [49]: LOINC is a database and universal standard for identifying medical laboratory observations. Since its inception, the database has expanded to include not just medical and laboratory code names, but also: nursing diagnosis, nursing interventions, outcomes classification, and patient care data set.

3) HEALTHCARE MESSAGES

To provide integrated patient care, the different clinical systems of a hospital need to communicate with each other. Message standards provide a consistent data flow among systems and organizations, specifying the format, data elements, and structure.

- *HL7 messaging standard* [50]: It is arguably the most widely implemented standard for interoperability in healthcare across the world and allows for the exchange of clinical data between disparate systems. HL7 is used for transmitting data related to patient charts, files, and other associated documents and audio recordings.
- *Digital Imaging and Communications in Medicine (DICOM)* [51]: DICOM is the de facto standard for exchanging medical images. It defines the formats for medical images that can be exchanged with the data and quality necessary for clinical use.

Unfortunately, these standards did not solve the problem of data integration. Brooks [52] pointed out some issues that organizations confront due to inconsistencies with data standards. For example, issues related to data exchange arise because the standards and terminologies are not designed to serve multiple purposes. Moreover, there are overlaps among standards; many standards have been named by one or more authoritative body, for example, HL7 and Accredited Standards Committee X12 (ASC X12) have some duplication in standards used for reporting of clinical data associated with the claims process.

III. INTEGRATION OF BIG HEALTHCARE DATA - THE SURVEY

With current trends in technology, Big data integration is turning into a complex process. Subsequently, there is no single methodology to suit all these data formats and requirements each provider brings. Effective data integration from heterogeneous massive amounts of data often requires a complex framework involving various methodologies. Putting together such a framework would be complicated. Therefore, we introduce a data integration framework (Fig. 5) a combination of Approaches, Technologies, Workflow, Tools, and Privacy, which ultimately helps to combine data from disparate sources into meaningful and valuable information.

An exhaustive number of methodologies have been proposed in large bodies of literature. Using the proposed framework as a blueprint, we studied these methodologies and reported in the following sections.

A. INTEGRATION METHODOLOGIES FOR BIG HEALTHCARE DATA

There are two main approaches for data integration: eager and lazy [53].

- *Eager*: In the eager or movement approach, the data from each source that may be of interest is extracted in advance, translated and iterated as appropriate, fused

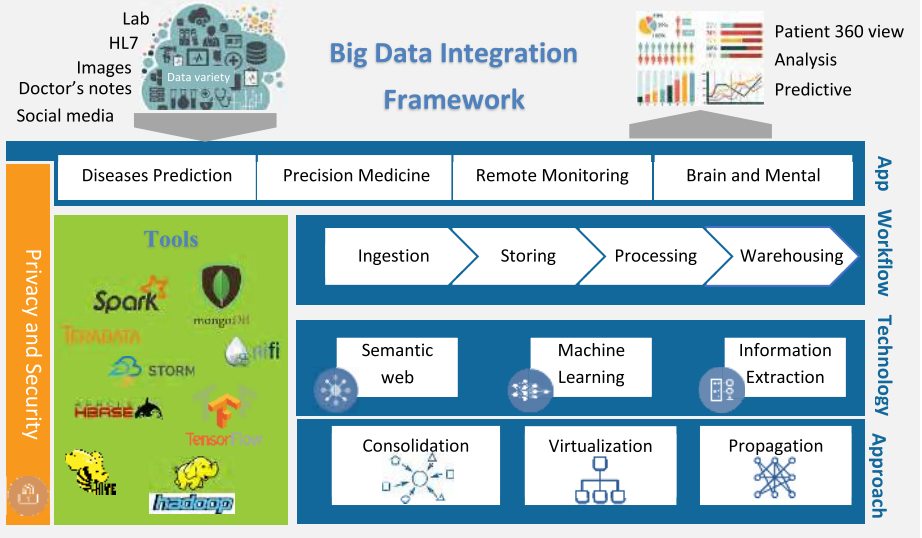

FIGURE 5. The reference framework consists of a set of components. The bottom component is the Approach component that specifies the integration Approach for combining data from different sources. The technology component processes data to discover and build links between datasets. An important component is the workflow component that describes the different stages of data integration. There are also two cross-cutting components, Big Data Tools and Security, which provide common functions that span other components in the framework. Application component uses all Framework components to produce a packaged data integration solution.

with relevant data from other sources, and stored in a centralized warehouse. When a query is posed, the query is evaluated directly at the repository, without accessing the original information sources.

- *Lazy*: In the lazy or mediated approach, the data is extracted from the sources only when queries are posed. First, the query determines the appropriate set of information sources to answer the query and generates the appropriate sub-queries or commands for each information source. Second, the query obtains results from the information sources, performs appropriate translation, iterating, and merging of the information, and returns the answer to the user or application.

There are effectively three Approaches of Big Data Integration. Table 1 introduces these Approaches as follows:

1) DATA CONSOLIDATION

Data consolidation refers to the collection and merging of data from multiple sources systems into one integrated place. Data consolidation phases include: (1) analysis of data models and datasets of the source and target environments; (2) transformation of the source data set; and, (3) merging of the data sets [54]. The data warehouse is such an example. Bill Inmon described a data warehouse as being a subject-oriented, integrated, time-variant, and nonvolatile collection of data [55] that is considered a core component of business intelligence. Data warehouse was built specifically for relational databases. However, the complex characteristics of data, including Volume, Variety, Velocity, Veracity, promote the need for new technology to handle new demands. Today, multiple data warehousing techniques rely on Hadoop

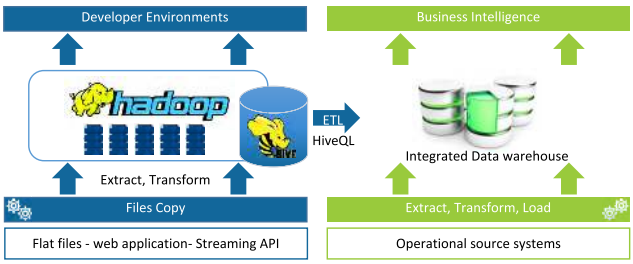

FIGURE 6. Modern data warehouse architecture.

platform to meet the new requirements of Big Data [56], [57]. Figure 6 shows a modern architecture of Big data warehouse where the raw data stream is stored in the HDFS system and then loaded into the data warehouse to complement the information that is already gathered.

Two different approaches could be found in the Big data consolidation architecture [58], which combine BDW(Big data warehouse) and EDW(Enterprise data warehouse) to implement data integration solutions (see Figure 7). In application architecture approach A, data ingestion mechanism is performed by the Hadoop platform, then specialized data integration tools are used to move data into RDBMS. Therefore, this approach is flexible in ingesting any data and also to address scale issues. In architecture B, the data that appear with Big data characteristics are stored and processed in the Hadoop platform, whereas RDBMS is used to store and process small and structured data. In the final stage, the information is available from both data stores.

Several Big data warehouse based solutions have been proposed in the literature, such as specialized disease

TABLE 1. Various approaches used in integrating big medical data.

Approaches	Pros	Cons
Data consolidation	• The imported data may be filtered and cleaned. • The retrieved data are converted and transformed into a more precise structure.	• The data must be refreshed frequently to ensure users have access to up-to-date content. • structure may not accommodate questions that arise at any given time.
Data virtualization	• Data remains stored in the component data source, instead of copying a huge amount of data into a single data store. • The user can see online information all the time.	• Any changes in the sources schema require updates to the federated schema. • Data cleansing is difficult and must be done on-the-fly. • Performance can be an issue because it is dependent on the query load capacities of the other data sources of the federation.
Data propagation	• Near-real-time updating of data changes throughout the data sources. • ETL process could be used with data propagation for the real-time data warehouse. • Data sources are integrated with transparency of location, source, and data structures.	• To attain high performance and to handle frequent synchronization, specialized tools and technologies are required.

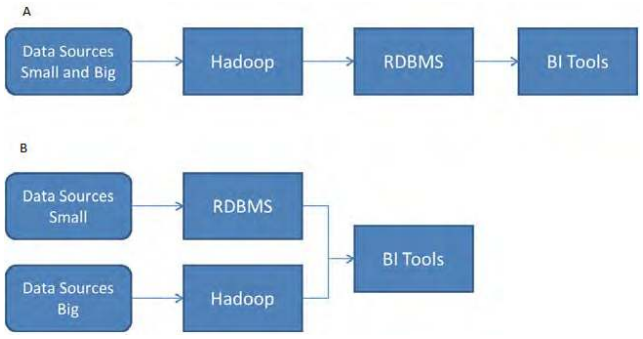

FIGURE 7. Architecture patterns involving Hadoop and RDBMS [58].

clinics [59], combining clinical and genomics queries [60], and a semantic warehouse to support the digital cancer patient [61] (see figure 8).

Apache Hive [63] often regarded as a distributed data warehouse infrastructure that enables easy data ETL from HDFS or other data storage like HBase. It provides HiveQL as a high-level query tool for accessing data. Cloud-based approach for interoperable EHRs [64], biomedical data integration [65] and medical Big data processing system [66] are examples of Hive Data warehouse.

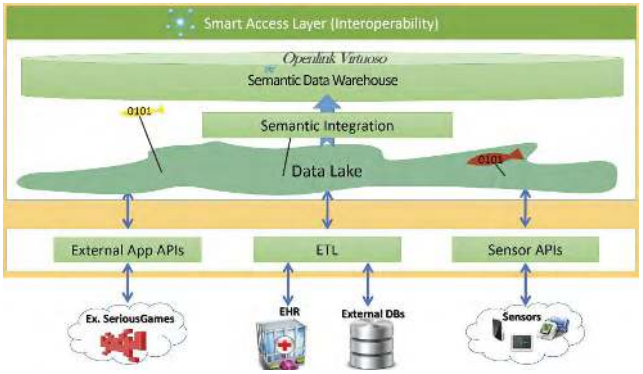

FIGURE 8. The data access layer of the iManageCancer platform [62].

Data variety is common in healthcare data. Therefore, the data lake is deemed as a potential solution for integrating diverse medical data. For instance, the classical data warehouse was often used to analyze the cost of care. The advanced cost analysis requires to integrate EHRs data with claims data. In the classical data warehouse, the process began with unloading the data in the warehouse then execute a new extract, transform, and load process. However, a data lake makes it more simple to enter a new data source or add advanced operation (queries, algorithms, etc.) [67]. Krause [68] summarized data lake as a matured Big data consolidation solution.

To sum up, conventional technology *data warehouse* can consolidate data with straightforward characteristics such as being small and structured. However, for Big Healthcare datasets specifically, data that comes with diversity pose a significant challenge on these technologies, and hence advanced technologies such as data lake for storing data in a scalable ecosystem and Apache Hive for warehousing data can be more efficient. However, Hive is a batch style warehousing solution; similarly, Apache Hadoop that is used in building scalable warehouse relies on batch style operational mode.

2) DATA VIRTUALIZATION

In the last decade, data sources have increased beyond the traditional structured world of databases and data warehouses to the stage of Big data where semi-structured and unstructured data are more common as well as some of it is stored outside of the local system. Moreover, the data warehouse is designed to host structured and mainly internal data from operational and transactional systems. Besides, building an enterprise data warehouse is a costly initiative that takes long to implement, and it is not always practical to move massive data from one source to another. Therefore, the data virtualization technique could enhance data integration and help to adopt new Big data source and modern formats [69].

Data virtualization is a process that federates different data sources, such as website data sources, relational databases, file repositories, document files, and data service providers into a single data access layer. This new abstraction layer makes all integrated data sources seem as one extensive

database is being accessed. However, to merge these different data sources, data virtualization uses several abstractions, transformation techniques, and data access mechanisms [70]. SAP HANA [71] is a Smart Data Access data-virtualization approach that enables unified access to heterogeneous data sources with a massive volume of data in real-time using in-memory processing. It provides Big data integration, allowing connection to data stored in Hadoop and NoSQL systems as SQL tables. Figure 9 illustrates the key components of SAP HANA architecture.

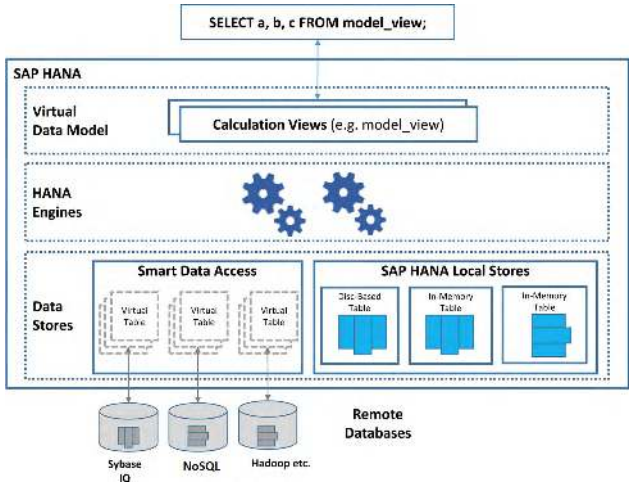

FIGURE 9. SAP HANA data virtualization architecture [71].

Data federation is a type of data virtualization where on-demand integration is used for combining data residing within different data stores [72]. For instance, Hadoop offers an HDFS Federation [73] technique to federate different instances at the HDFS level through a separation of namespace and storage, enabling a generic block storage layer. In the healthcare domain, data federation technique was used in many research projects such as e-Health Service [74], BioFed [75], and Genomic Computing [76].

3) DATA PROPAGATION

Traditional batch-oriented ETL processing cannot satisfy the real-time requirements when data is integrated continuously and concurrently. Hence, the approach called data propagation was proposed. Data propagation is usually referred to as active data integration [77], where a copy of the data from a data source or multiple data sources to discrete locations is done, often to make data more accessible to users. This process usually operates online and with event-driven architecture(push mode). In general, the process of change propagation could have multiple stages such as transformations and filtering of the changed data that is delivered to the dependent systems [78]. Constantinescu *et al.* [79] propose SparkMed as a data integration framework for mobile healthcare. An automated process (daemon) of the framework could be attached to applications, collects the multimedia data (such as the hospital information system, picture archiving, and reporting systems) and prepares them to be propagated in a convenient, reliable manner. SparkMed can integrate a wide range of different medical software and database systems into a cloud-like peer-to-peer multimedia data store.

B. INTEGRATION TECHNOLOGIES FOR BIG HEALTHCARE DATA

The complexity of the variety in Big data is well known, including complex heterogeneous data types (structured data, unstructured data, and semi-structured data, etc.), complex intrinsic semantic associations in data (clinical, genomic, etc.), and complex relationship networks among data (relationships between entities) [80]. For instance, integrating diverse data types such as clinical data, gene expression, DNA methylation, miRNA expression, and copy number alterations (CNA) has improved the prognostic prediction of glioblastoma multiforme [81]. The heterogeneity of different types of data mentioned above creates a challenge for the data integration process. The integration process requires a specific mechanism for aggregating data that develops specifically to deal with the nature of healthcare data to overcome these complexities. Data integration technologies such as semantic web, machine learning, information extraction, and linked data can enhance understanding the context of information. In this section, we discuss the existing technologies for data integration.

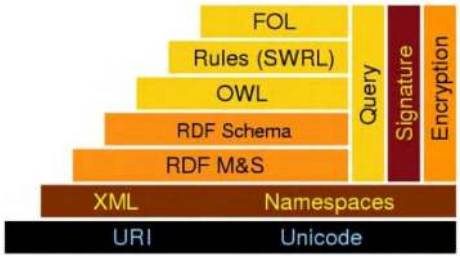

FIGURE 10. Semantic web layer cake.

1) SEMANTIC WEB

Semantic web standards are a perfect fit for data integration capabilities [82], with multiple aspects to integrate data from globally separate, distributed, and heterogeneous data sources. Semantic web technology is a robust and extensible data model used for global naming, with the ability to reason based on Description Logic [83]. The semantic web comprises publishing information in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL) and SPARQL (a query language for semantic web data sources), Figure 10 shows the different layers of the semantic web framework. According to this framework, information is represented in statements, called RDF triples. The three parts of each triple are called subject, predicate, and object. Besides that, OWL and Linked Open Data, which use RDF as the data model, have gained popularity in data integration.

- *Web Ontology Language (OWL)*: is a W3C standard [84], that offers great machine interpretability for the web content. Moreover, OWL is used for the harmonization of integrated data from diverse sources. Several ontologies have already been developed in the disciplines of healthcare. CNTRO [85] A Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives has been created to annotate and query the temporal information of clinical data semantically and using inference to expose new temporal features and relations. Open Biomedical Ontologies (OBO) project [86] aims at creating controlled vocabularies for shared use across different biological and medical domains. PCPO [87] integrates multiple drug resources and maps two well-known drug class resources, Anatomical Therapeutic Chemical classification system (ATC) and National Drug File Reference Terminology (NDF-RT). SNOMED CT [88], the comprehensive, multilingual clinical healthcare terminology from the International Health Terminology Standards Development Organization (IHTSDO), has its OWL ontology. Another favorable development is the ICD-11 standard [89] that was created using OWL and RDF. In fact, OWL would be a potential solution for Big data integration [90]–[92], since it has the ability to merge multiple data sources and publish them onto the cloud [93].
- *Linked Open Data (LOD):* describes a method of publishing and linking structured data coming from different data sources that can be interlinked and published on the web. An increasing number of data providers have adopted Linked Data principles as a data structure. Linked data is resulting in the emergence of global data space on the web containing billions of RDF triples [94]. The linked data paradigm can integrate Big data by the mean of annotating unstructured data with open linked data from the cloud, which leads to linking those heterogeneous datasets to each other. Linked data is one of the widely used technologies for data integration in the healthcare sector. A variety of genomic and drug-related datasets as Linked data were published by members of the W3C Healthcare and Life Sciences Interest Group (HCLS IG) [95]. LOD datasets have been crawled by the Semantic Web Search Engine (SWSE) and can be accessed via a faceted browsing interface. Certain datasets are interconnected through semantic vocabularies such as ''sameAs,'' ''seeAlso'' (see figure 11), where others remain challenging to define a methodology for linking them with other Linked Data sources [96].

Described above techniques are capable of addressing numerous problems of data integration in the healthcare context. Many projects already used OWL and LOD to enhance data integration such as integrating heterogeneous wearable data in healthcare [97], support of digital cancer patient [61], obesity surveillance [98] and in the interoperability of electronic health records [64].

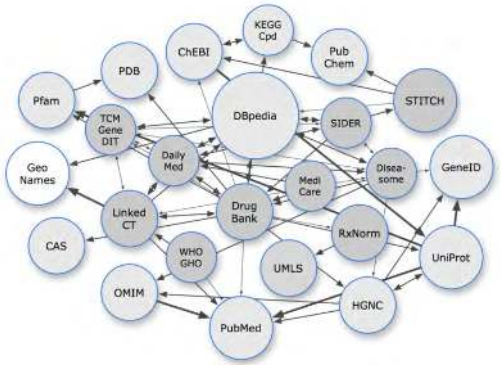

FIGURE 11. A graph of some of the LODD datasets (dark grey), related biomedical datasets (light grey), related general-purpose datasets (white) and their interconnections [95].

Big linked cancer data [99] presents a scalable Linked Data-driven solution for the continuous integration of biomedical data sources. The integration process relies on three types of datasets that are loaded into various SPARQL endpoints: Linked data version of TCGA and PubMed metadata in RDF and a set of mappings between these datasets. The latter aims at establishing a bridge between the structured data contained in TCGA and the constant flow of RDF data generated by analyzing PubMed. The bridging process is performed by matching the synonyms for every disease and gene found in Linked TCGA with PubMed article's abstract. The scalability of the framework is ensured through the novel TopFed federated query engine.

Searching and exploring data about medicinal products and drugs from different data sources are essential requirements for physicians, to cover those information requirements, Kozak *et al.* [100] identify drug data sources such as MeSH, ATC, NDF-RT, NCI DrugBank, CZ-Drugs, and FDA, to integrate them. The data sources have a different format with structured and unstructured data. In [96], the authors provide a solution which enables the analysis of existing Linked Data representations of each considered data source then it allows the creation and publishing of new Linked data for those with no Linked Data representation. Authors recommend linking the dataset to an accepted reference terminology, allowing anyone else to connect their datasets to this reference terminology and thus enabling integration with other datasets.

Accordingly, technical challenges for broader adoption of the Semantic Web standards for Big Data include large-scale data reasoning and performance optimization of semantic-based systems. Another limitation of the Semantic Web technologies is that they are purely for graph data representation. Therefore, if the data is unstructured, it would not be enough to use these techniques alone for building a data integration solution.

2) MACHINE LEARNING

Automatically learning mappings between two datasets through machine learning, can remove much of the

development effort involved in data integration. Machine learning has been proving itself in many diverse domains such as image processing, social network mining, finance, and computer vision [101]. Moreover, using machine learning to integrate various datasets, such as Medical Imaging, Electrophysiological monitoring, Clinical Information, has improved the accuracy of diagnosis and prediction of disease outcomes [102]–[104]. Generally, Machine learning methods can be divided into two main classes: supervised learning (predictive) and unsupervised learning (descriptive):

- Supervised learning takes samples of training data with known labels as input to learn a general prediction rule mapping new data samples to existing labels. For example, the input data may include affected and unaffected diabetic patients. Thereby, a model is learned to help and accelerate the diagnosis of Diabetes disease [105]. Classification and regression are the main methods of supervised learning: in regression, the output variable is numeric (or continuous), while the one used for classification is categorical (or discrete). Some of the widely used classification techniques include Support Vector Machines (SVMs) that are widely applied to large medical datasets [106]. Its main drawback lies in colossal time and memory complexities, which depends on the training set size cardinality and heterogeneity [106]. Random Forest is another useful and easy to understand classifier. Due to its space limitation and overfitting problem, Random Forest may not be applicable for huge datasets [107]. Besides, regression algorithms are also widely used in health applications to model relationships between objects and targets [108] [109].
- Unsupervised learning includes clustering which takes as input an unlabeled dataset. A model is learned by finding the structure of unknown input data and looking for the similarity between entities. Many algorithms can be used for clustering purpose, including partitioning, hierarchical, and density-based [110]. Only the partitioning algorithm is capable of handling large datasets, while hierarchical and density-based are slow for large datasets. These methods are often used in pattern discovery in gene expression data and molecular subtyping of cancer patients.

Artificial Neural Networks ANN is a family of machine learning approaches whose models are hierarchical representation of supervised and unsupervised learning models. In a neural network, associations between result and input variables are described using multiple combinations of hidden layers of pre-specified functions. The purpose is to estimate weights using input data and results to minimize the average error between the results and their predictions. However, the disadvantages of the neural network when dealing with the Big data are the requirement for constant memory consumption and the computational time [111]. ANN has been used in various health care applications, such as diagnose cancer [112], diagnose Parkinson's disease [113], etc. A modified version of ANN called Deep Learning, which builds neural networks with a large number of hidden layers. Deep learning architecture (i.e., deep neural networks DNN, convolutional neural networks CNN, recurrent neural networks RNN) provides better capabilities for dealing with Big Data issues, such as volume and velocity [114]. Deep-learning systems can accept multiple data types as input, an aspect of particular relevance for heterogeneous healthcare data (figure 12).

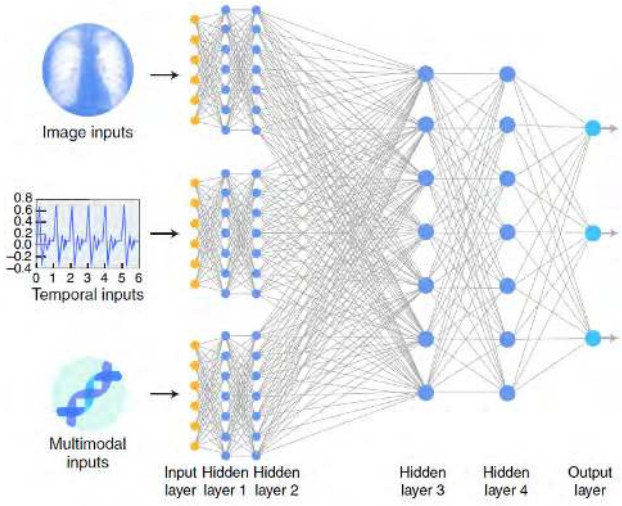

FIGURE 12. Deep learning can be trained on a variety of data types (images, time-series, etc.) [115].

From the beginning, supervised, unsupervised, and hybrid machine learning approaches have been applied to the data integration field to support the integration process [116]. Examples include entity resolution that uses Decision trees, Logistic regression, and SVM [117], as well as applying Deep learning model (e.g., Word2Vec, Par2Vec) to compare a long biomedical text [118]. At the same time, schema alignment adopted machine learning algorithms such as Naive Bayes and stacking to match types and attributes [119].

Since there is no single traditional machine learning technique which can perform well for healthcare data integration, one application often employs a combination of several methods.

In [120], a variety of data integration techniques from a machine learning view have been reviewed. It has concluded the following:

- Feature concatenation: With the modern high dimensionality of data and rich structural information, feature concatenation is often impracticable.
- Bayesian models: In general, these models can use prior information and model measurements with various distributions.
- Tree-based methods: These models can be applied in two strategies, 1) build a tree with all features, 2) collectively make a final decision based on trees learned from each view.
- Kernel methods: In a first step, Metric learning aims at fusing the similarity matrices learned from personal

views together, then a final Kernel learning model combines similarity between results.
- Network-based fusion methods: they can infer direct and indirect connections in a heterogeneous network.
- Multi-view matrix factorization models: the model begins by extracting new features from each data view first and then incorporate these new features together. Finally, a classification or clustering algorithm can be applied to the combined features. These models have the potential to learn interactions among features from different views.
- Deep learning: Different deep learning models can be applied to individual data views, and then the result is integrated with multi-modal learning for capturing the complex mechanism of systems.

Moreover, this study has emphasized the importance of methods such as multi-view matrix factorizations and multi-modal deep learning for data integration in the Bioinformatics domain.

Zolfaghar *et al.* [121] used machine learning techniques to study the 30-day risk of readmission for congestive heart failure patients. In this study, the income factor represents the primary predictor variable for risk of readmission (RoR). The first step is to map income value which is available in National Inpatient Sample (NIS) dataset (with 8 million records and more than 100 features) to the MultiCare Health Systems (MHS) data. However, due to privacy restrictions, it is not possible to link patients in NIS to patients in MHS data. To achieve this, K-means is used to cluster the NIS dataset by relying on three variables; age, gender, and elective hospitalization. Then the average income is calculated for each cluster, and the computed value is used to map each record of data in MHS to the closest cluster based on the Euclidian distance function.

A hybrid machine learning method was applied to classify schizophrenia and control individuals by integrating fMRI and single nucleotide polymorphism (SNP) data. Two SVMs were used, one on fMRI data and one on SNP data, and then the results were combined into a single module using majority voting [102].

Napolitano *et al.* [122] have combined drug and protein structures, disease states, and drug toxicity using a kernel-based (KB) method. Each data is represented by a kernel matrix in a drug-centered feature space. After combining these kernel matrices into a single kernel matrix, the authors applied SVM for classification. The result was used for repurposing and sensitivity prediction.

In [123], three independent deep neural networks (DNN) were trained in clinical data, gene expression, and copy number for predicting the prognosis of breast cancer. The final multi-modal deep network was obtained by joining the predictive scores of each independent model.

However, the characteristics of Big data create scalability challenges for traditional techniques [124]. For example, estimating model parameters through iterative procedures used by various machine learning methods, including ANN, SVM, and DT, could not be easily scaled across large data sets. Therefore, non-iterative training algorithms are increasingly used by Big data applications [125]. As well, the volume challenge can be reduced using dimensionality reduction. Methods such as principal component analysis (PCA), Multimodal deep learning, Isomap have proven their effectiveness in dimensionality reduction [126].

3) INFORMATION EXTRACTION

In Big Data, Variety problem refers to different types of data collected via different data sources, such as text, videos, images, data logs, audio, and so on. Therefore, this data will not be in a format ready for integration. Thus, the need for a process that gets information from the underlying data sources and explores the relations between entities. The term 'Information Extraction' refers to the process by which structured, useful information such as entities, relationships between entities, and attributes describing entities are automatically derived from unstructured or semi-structured raw data. Moreover, information extraction prepares and facilitates different types of sources to be integrated and queried [127]. Subsequently, information extraction must be incorporated into the data integration workflow to make use of the extracted knowledge [128].

Information extraction is interdisciplinary, involving data mining, statistics, computing linguistics, and machine learning. Several methods have been introduced for the extraction of information. Sarawagi *et al.* [127] categorized extraction methods along two dimensions:
- Hand-coded or Learning-based: A hand-coded system needs experts to determine the appropriate rules or regular expressions or program snippets to perform the extraction. Conversely, learning-based extraction relies on unstructured data labeled to train machine learning models.
- Rule-based or Statistical: Rule-based methods are driven by hard predicates (rules), while statistical learning methods make decisions based on a weighted sum of rule firings.

The EHRs contain text dictations from several physicians, structured data from sensors and measurements, and unstructured data from video and image (MRI, X-RAY). These resources require the development of strategies to transform unstructured data in a structured form suitable for integration. Several methods of information extraction have been applied in textual parts of EHRs. Information extraction was used to detect drug safety signals by transforming clinical notes into a feature matrix encoded using medical terminologies [129]. Furthermore, i/n conjunction with the warning of increased cardiovascular mortality resulting from Cilostazol medication, Leeper *et al.* [130] employed a novel text-analytics pipeline to quantify the adverse events associated with Cilostazol use in a clinical setting.

Information extraction approaches for image data are different from textual data. For instance, a method was designed and implemented in [131] to manipulate 240GB of brain

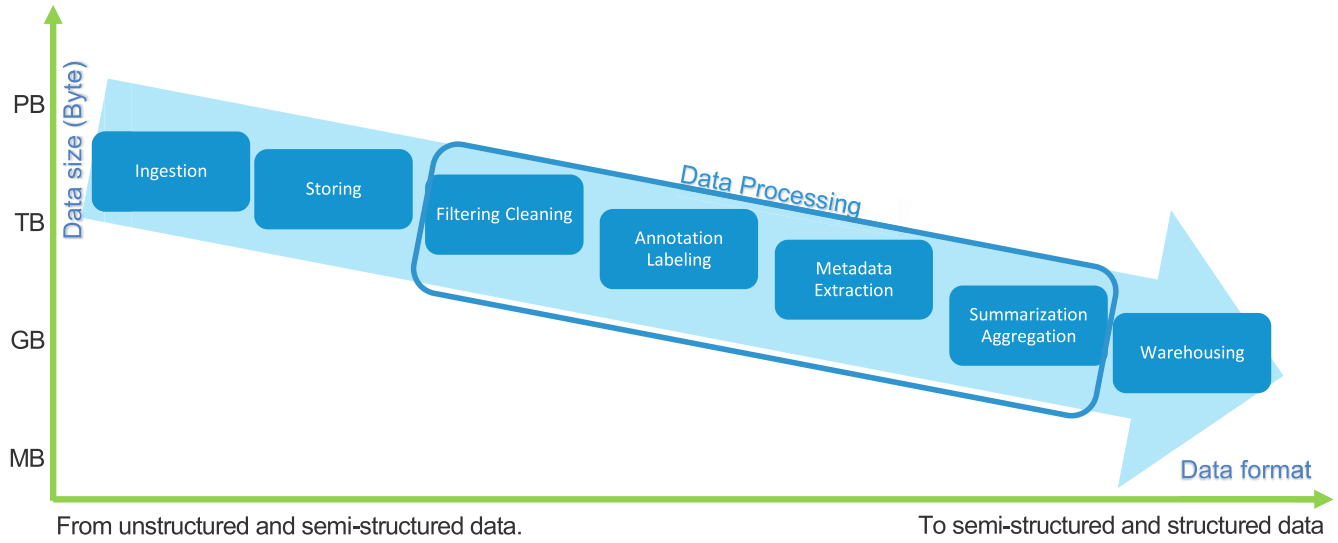

FIGURE 13. Generic workflow for Big data integration: Huge data from heterogeneous sources are received as input and valuable and manageable data are output. During the workflow, various processing stages are applied to reduce the size of data on the one hand, on the other hand, extract structured data from semi-structured and unstructured data. Each stage could benefit from techniques such as entity extraction and enrichment with semantic web technologies, and dimensionality reduction and classification with machine learning technologies. Under certain circumstances, some Big Data integration problems could be readily solved in such way.

image data for 1200 patients stored by the Alzheimer's Disease Neuroimaging Initiative (ADNI), to predict to what degree a patient has Alzheimer's. Different steps are executed, which include spatial normalization, extraction of features, feature selection, and patient classification. The proposed filter features selection method is based on mutual information as relevance measure and redundancy between the features through minimal-redundancy-maximal-relevance criterion (mRMR).

Accuracy is the biggest challenge in healthcare information extraction. Precisely, we mean that when the information is extracted, it should be extracted correctly, with acceptable accuracy for clinical staff that should be approximately 95%. This adds a performance overhead over current treatment capabilities.

C. INTEGRATION WORKFLOW FOR BIG HEALTHCARE DATA

Organizations need data in a consumable format to assist analysts in the decision-making process. The main challenges are the following:

- how to convert the raw data from multiple sources affected by heterogeneity format into clear and coherent information?
- which workflow is enough to take structured and unstructured ones and convert them into insight?
- how the data will be stored in the cluster?
- what tools to use to process and integrate the data?
- how to provide access to the end consumer?

To an extent, the technologies used for Big medical data integration are similar to that of traditional Big data integration projects. The main difference lies in how integration is performed [22]. Several frameworks are being developed for medical Big data integration and fusion, as mentioned in subsection III-A. However, the majority of existing frameworks cannot cope with those challenges, which require multiple processing steps during integration. Therefore, we provide a generic workflow, depicted in figure 13, that summarizes the stages of Big data integration process.

At each step, there is work to be done [132], and there are Big data tools and technologies to be used [133]. The workflow process begins with data ingestion from data-sources, followed by data storing, data filtering and cleaning, data annotation and labeling, metadata extraction, information summarization, and aggregation, and ends in the Data Warehousing step where the information is ready to analyze or visualize.

1) DATA INGESTION

The first step involves loading data from different sources in various formats into a single or clustered store. Data ingestion is modeled as a pipeline consisting of several incremental steps with clearly defined interactions. Services for data ingestion must be compatible with standards such as HL7, CDA, DICOM, and IHE XDS. Through this pipeline, all patient identifying data may be anonymized to ensure data privacy. Moreover, much of this data is of no interest, and so, can be filtered and compressed by orders of magnitude. For instance, a Big Data platform [134] was implemented for the analysis of medical data in the Mayo Clinic. The platform can ingest and store 62 ± 4 million HL7 messages per day. The data ingested by the platform can be of any medical data type, be it structured, semi-structured, or unstructured data.

2) DATA STORE

After the ingestion phase, the raw data is pooled into a centralized and scalable storage location(the data lake), which is an

intermediate storage area and working environment for data that typically represents source data in its original format. Due to the different data formats, Data Lake will consist of various systems, such as relational database for storing and managing structured data(demographic information, vital signs, etc.), NoSQL for semi-structured data(medical device reports) and file system for unstructured data(clinical narratives, notes, letters, reports, images and omics data). For instance, CancerLinQ [135] (Cancer Learning Intelligence Network for Quality) is designed as an oncology rapid-learning health care system. CancerLinQ accepts all data in any format that a practice chooses to send, then stores and maintains the original data in a data lake.

3) DATA FILTERING AND CLEANING

EMR illustrates well the need for data cleaning as it may provide noisy data containing incomplete information [136]. Processing raw data without preparation routines may require additional computing resources that are not affordable in the context of Big data. Data filtering is achieved by removing unnecessary information for health care monitoring based on a defined criterion, while data cleaning is accomplished using several components such as noise reduction, missing data management, and normalization. In [137], k Nearest Neighbour and K-means are used to remove the noise from diabetes dataset and thereby improving the quality of data.

4) DATA ANNOTATION AND LABELING

At this stage, the workflow has to explore the different data formats like clinical notes, images, and scientific publications and must be able to discover, extract and annotate them with actual labels such as the name of the entity, relations between them, etc. Ontologies are applied to clarify the meaning of the concepts by using standardized terms across various data sources. A typical step in medical data processing that aggregates unstructured clinical notes is the identification of these medical concepts from UMLS using MetaMap [138]. UMLS uses as well, the notion of a Concept Unique Identifier (CUI) to map terms with similar meaning in different terminologies [139]. The difficulty with current labeling techniques is that they do not understand model relationships between classes. Ayala *et al.* [140] proposed a two-phase machine learning approach that computes novel features that take into account the relationships.

5) METADATA EXTRACTION

At the heart of Big data integration process is Metadata. The lack of well-defined schemas characterizes big data. Besides, data integration in such an environment is subject to frequently changing requirements. Therefore, this step should extract structural information describing the schema of the data and clarifying the semantics of metadata. The result of this stage will help the end user to query over structured and semi-structured data, and to discover associated schema between various datasets [141].

6) INFORMATION SUMMARIZATION AND AGGREGATION

Frequently, health data have volumes of hundreds of millions or billions of records per day and are not in size ready for analysis. Thus, it is essential for business intelligence to summarize and aggregate the required information from the heterogeneous sources and express them in a structured form appropriate for analysis. Summarized and aggregated data and their associated metadata are then used to create abstractions or pattern representations. For instance, summarization of an extensive medical record, allows a set of events within a facet to be recursively aggregated and replaced with summary events, such as a series of atenolol and propanolol prescriptions can be aggregated into the beta-blockers category [142].

7) DATA WAREHOUSING

Data lake tend to be complicated to navigate for users unused to working with unprocessed data. Furthermore, researchers and clinicians tend to favor data warehouses. Conversely, data scientists could apply dimensionality modeling from the data lake to prepare the datasets and then feed them back into a traditional data warehouse for decision analysis. This final step facilitates the integration of different data sources and reduces data movement and latency. For instance, supporting critical precision medicine use cases, an automated workflow has been implemented for incorporating sequencing results from both structured and unstructured sources into a research-centric clinical data warehouse [143].

The successful completion of the workflow enables the users to use the data for analysis.

D. INTEGRATION TOOLS FOR BIG HEALTHCARE DATA

Scalability, Reliability, and Maintainability are a vital consideration when it comes to Big data integration tools. While the available tools are mostly open source and wrapped around Hadoop and related platforms, there are many trade-offs that developers and users of Big data analytics in healthcare must consider. While the development costs may be lower since these tools are open source and free of charge, the downsides are the lack of technical support and minimal security [2].

Various frameworks and tools have been implemented to meet the management of the ever-growing size of complex heterogeneous data, from data ingestion to data visualization. So far, most Big data tools do not provide a complete process for data integration, but they can be a part of an integration architecture to store and process data.

1) INGESTION TOOLS

Since data are collected from a variety of sources and formats, ingestion tools need to take into account the volume and velocity of structured and unstructured data. Flume[3] is a reliable and distributed service for efficiently collecting large amounts of log data. Sqoop[4] is a tool that imports structured data from traditional RDBMS database and provides

[3]https://flume.apache.org/
[4]https://sqoop.apache.org/

methods for transferring data to HDFS or Hive. Apache NiFi[5] is a reliable and scalable tool to load and collect data from different sources then dump it into other sources. NiFi is highly configurable and includes an easy to use user interface.

For instance, Tilve *et al.* [144] proposed a tool to integrate information from research processes from different fields. Notably, the information generated in the areas of proteomics, genomics, cell cultures, and histomorphology. The data is collected by acquiring a wide range of gross information from different local databases using the Sqoop tool and stored inside the data storage system HDFS.

2) STORAGE TOOLS

The complicated nature of health data means that today's healthcare sector cannot rely solely on traditional data storage methods. Data storage should continue to be innovated to accommodate data nature and growth. Methods must be scalable while maintaining high performance in data access.

- *File System:* HDFS is a highly fault-tolerant distributed file system that stores data on the clusters, it can handle large amounts of data, regardless of format [145]. Amazon S3[6] (Simple Storage Service) is an online service that allows storing large amounts of data. S3 is free to join and is a pay-as-you-go service.
- *NoSQL:* The traditional relational database has faced many challenges to store and process Big data effectively. Therefore, to solve these challenges, a variety of "NoSQL" databases appeared with many aspects such as reading and writing data quickly, supporting mass storage, ease of expansion, and low cost. Furthermore, NoSQL Databases are classified into three basic categories: Key-value (HBase,[7] Redis[8]), Column-oriented (Cassandra,[9] Hypertable[10]) and Document database (MongoDB,[11] CouchDB[12]) [146]. As one of NoSQL data stores, Graph databases (Neo4J,[13] AllegroGraph,[14] Openlink Virtuoso [15]) provide an enterprise-grade RDF triple store [147].
- *Data Warehouse:* Apache Hive[16] is a data warehousing infrastructure that provides a SQL-like interface: Hive QL [148]. It enables easy data ETL from HDFS or other data storage like HBase. Teradata[17] is one of the well-known RDMS, best suited for database warehousing application dealing with a considerable amount of data. Teradata has a patented PDE (Parallel database extension) software that enables parallel processing and allows faster processing with a large margin over traditional databases.

For example, HMBDPS [66] is a distributed Hadoop-based Medical Big Data Processing System which aims at integrating and processing Big medical data(structured, semi-structured and unstructured data that are produced by HISs) to study some features of user behaviors and provide personalized recommendations based on public behavior for each user. Figure 14 shows the data access layer of HMBDPS, where the physical unit of storage is built based on the Hadoop cluster and the logical unit of storage is implemented using Hive. This logical unit could be accessed by Hive QL (HQL) and user-defined functions (UDF) to store and manage data efficiently.

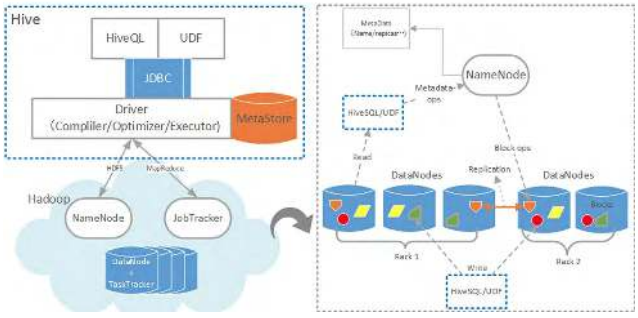

FIGURE 14. Architecture of the Big Data Warehouse in HMBDPS [66].

3) PROCESSING TOOLS

Scalable data processing is essential for Big data integration that involves a complex sequence of processing steps with different medical data formats. Various types of processing tools are employed in Big Data integration:

- *Batch Processing:* To process Big Data, MapReduce is a well-accepted method to perform parallel computing and distributed storage [145]. MapReduce is a programming model and an associated implementation for processing and generating large datasets, and is suitable for semi-structured or unstructured data. Around MapReduce, many projects, such as Pig,[18] use a high-level language that spends less time writing mapper and reducer programs. Spark[19] project introduced a cluster computing engine for Big data applications that offers scalability, flexibility, and speed to deal with Big Data challenges. Spark has the power to process and hold data in memory across the cluster. Spark SQL[20] is a module for structured and semi-structured data processing, it is used to query data, both inside a Spark program and from an external repository [149].
- *Stream Processing:* Real-time or near-real-time data processing requires a different processing paradigm than

[5]https://nifi.apache.org/
[6]https://aws.amazon.com/
[7]https://hbase.apache.org/
[8]Redis: https://redis.io/
[9]http://cassandra.apache.org/
[10]http://www.hypertable.org/
[11]https://www.mongodb.com/
[12]http://couchdb.apache.org/
[13]https://neo4j.com/
[14]https://allegrograph.com/
[15]https://virtuoso.openlinksw.com/
[16]https://hive.apache.org/
[17]https://www.teradata.com/
[18]https://pig.apache.org/
[19]https://spark.apache.org/
[20]https://spark.apache.org/sql/

the batch mode. Stream processing operates each entity data item as soon as it enters the system. In this respect, several distributed computing systems can manage and process Big Data in near real-time. Storm[21] is a low latency distributed stream processing framework that could handle very high stream data rates and deliver results with less latency than other solutions [150]. Spark Streaming[22] is a distributed batch processing framework (over a sliding window) with stream processing capabilities that speeds up batch processing workloads by offering full in-memory computation and processing optimization [151]. Kafka[23] is a real-time message publish-subscribe system that combines the benefits of traditional log aggregators and messaging systems. It is designed as a kernel for data stream architecture. A Kafka Streams library that provides a stream processing capability has been added to the Kafka client library. Kafka is built to be high-throughput, horizontally scalable, fault-tolerant, and allows geographic distribution of data streams and processing [152]. Flink[24] is a fully-fledged and efficient batch processor that lies on top of a streaming runtime. Flink follows a paradigm that encompasses data flow processing as a unifying model for real-time analysis, continuous streams, and batch processing, both in the programming model and in the execution engine [153].

- *Machine Learning:* Scalability, speed, coverage, usability and extensibility are the main factors to evaluate when choosing machine learning tools, with the note that the prioritization of these factors largely depends on the applications they are being used for [124]. For example, Mahout[25] enables the distributed implementation of machine learning algorithms for Big data, providing scalable feature selection, data sampling, and classification. MLlib[26] from Spark provides a scalable and distributed implementation of popular machine learning methods such as k-means clustering, regression models, SVM, Naïve Bayes. Google's TensorFlow[27] is another tool successfully used for deepening approaches, including long short-term memory (LSTM) algorithms, convolutional neural networks (CNN), etc. TensorFlow allows distributed implementation of Deep learning model on many CPUs or GPUs for large scale analysis.

As a case in point, Panahiazar *et al.* [154] discussed how to store multiple datasets from different resources including EHRs, Medical, and Genomics Images into the Hortonwork repository [155] and then used Pig to clean and prepare data. The authors performed a simple operation like AVERAGE to compare the performance of Pig with other tools like SQL. SQL took 18 minutes to run, but Pig ran in less than two minutes on two nodes. In this respect, Ding *et al.* [156] proposed a Shared Nearest-Neighbor Quantum Game-based Attribute Reduction (SNNQGAR) algorithm for performing consistent segmentations of cerebral cortical surfaces of the complex neonatal brain regions. SNNQGAR is parallelized using a new hierarchical coevolutionary Spark model combined with an improved MapReduce. This architecture provides improved attribute reduction solutions for big data processing.

[21] https://storm.apache.org/
[22] https://spark.apache.org/streaming/
[23] https://kafka.apache.org/
[24] https://flink.apache.org/
[25] https://mahout.apache.org/
[26] https://spark.apache.org/mllib/
[27] https://www.tensorflow.org/

E. PRIVACY AND SECURITY FOR BIG HEALTHCARE DATA

The integration of Big data raises many privacy concerns, particularly in the health care sector, due to the promulgation of the Health Insurance Portability and Accountability Act (HIPAA). Because the data integration process aggregates data into a centralized repository, it is extremely vulnerable to attack. Therefore, security and privacy policies should be considered as part of the design of the health data integration platform. Besides, legislation and regulation should often be regarded as re-evaluate emerging technologies and capabilities [157].

Traditional security and privacy mechanisms are insufficient to protect Big data. Nevertheless, new technologies also host unknown back doors. Therefore, integrity, confidentiality, and availability of data must be carefully considered.

1) SECURITY

Security is defined as protection against unauthorized access. Therefore, for providing secure access to clinical data, healthcare information systems must provide the following policies: authentication, access control, confidentiality, integrity, attribution/non-repudiation [158]. However, the diversity of data sources, data formats, streaming, and infrastructure can lead to unique security vulnerabilities. In this regard, Alshboul *et al.* [159] proposed a Big data security lifecycle model, designed to take into account the phases of the Big data lifecycle and correlate threats and attacks that face Big data environment within four phases:

- Data collection phase: It is essential to collect data from reliable sources and to use specific security measures, such as the encryption of individual data fields (patient identifier).
- Data storage phase: the collected data may contain sensitive information. Thus, some security measures can be used, such as the data anonymization approach, the permutation, and partitioning of data to ensure the safety of the collected data.
- Data analytics phase: In this phase, machine learning methods such as clustering, classification, and association rule are used for link extraction and feature selection, which can extract sensitive data. Therefore, this phase needs to be protected while making sure only authorized staff can be engaged in this phase.
- Knowledge creation phase: The created knowledge is treated as sensitive information, especially in the

health sector. Therefore, organizations must ensure that this information (e.g., patient information) is not to be publicly released.

2) PRIVACY

Health care data sharing allows early detection of epidemics, but without evident privacy protection, it is difficult to extend these surveillance measures nationally or internationally [160]. Privacy is often defined as the ability to protect sensitive information about personally identifiable healthcare information. Various traditional methods guarantee some degree of privacy in Big Data, but their disadvantages have led to the emergence of newer methods [161].

- *De-identification* is a traditional method for privacy-preserving, in which, data must first be sanitized with generalization and suppression before the publishing for data processing to protect patient privacy. K-anonymity, L-diversity, and T-closeness are three traditional methods of De-identification. Many scalable anonymization solutions within the MapReduce framework have been proposed to improve these traditional techniques of protecting Big data privacy.
- *Hybrid execution model* is used for guaranteeing privacy in cloud computing. It utilizes public clouds only for an organization's non-sensitive data, whereas for an organization's sensitive, private data and computation, the model executes within their private cloud.
- *Privacy-preserving aggregation* is built on homomorphic encryption as a widespread data collecting technique for event statistics. These encrypted texts can be aggregated, and the aggregated result can be retrieved with the corresponding private key. Thus, privacy-preserving aggregation can protect the privacy of the patient during the Big data collection and storage phases.

To address the scalability problem of big data privacy, Gheid and Challal [162] presented a general architecture of Big data analytics for multi-source Big data. The architecture introduced an efficient and privacy-preserving cosine similarity computing protocol in response to the efficiency and privacy requirements of data mining in the Big data era. Moreover, Zhang *et al.* [163] proposed a scalable two-phase top-down specialization approach for the anonymization of large data sets using the Map Reduce framework in the cloud.

F. INTEGRATION APPLICATIONS AND PLATFORMS FOR BIG HEALTHCARE DATA

The integration of Big data presents new opportunities to create novel applications in the field of healthcare, which provide many benefits to clinicians who can seamlessly search across healthcare systems to get the complete picture of a patient. In this section, we discuss these applications.

1) DIABETICS DATA INTEGRATION

MOSAIC system [164] has been designed to be potentially used in any context dealing with Type 2 Diabetes Mellitus (T2DM) patients. In MOSAIC system, the i2b2 [164] Data Warehouse (DW) allows integrating clinical information coming from hospital EHRs, administrative data from the local health care agencies, and environmental data collected from satellites. A common data model was defined and implemented using the i2b2 technology to query and integrate these heterogeneous huge data, [165]:

- A query engine, implemented as back-end service, relies on a MongoDB and provides a logical layer between the user and the data.
- A Temporal Abstraction module takes raw quantitative data stored in the i2b2 DW, analyzes the evolution of diabetes by taking advantage of different datasets, and then stores the integrated result in the DW.

2) DATA INTEGRATION IN DISEASES PREDICTION SOLUTION

IManageCancer project [61] aims to provide a cancer-specific self-management platform, with particular emphasis on avoidance, early detection, and management of adverse events of cancer. The high-level architecture of iManageCancer (shown in figure 8) is based on the Data Lake concept to store and manage heterogeneous, structured, and unstructured data sources. This Data Lake includes various databases such as PostgreSQL for storing patient information, Cassandra DBs for staging Big data available (e.g., activity monitoring data, sensor data, etc.). The data is queried, transformed into triples, and loaded into a semantic data warehouse where it is available for further analysis. This architecture can specify which of the available data have to be semantically linked and integrated by selecting the suitable mappings to a modular ontology. Furthermore, The Semantic Warehouse provides a 'semantically enriched' and 'search-optimized' index to fill the limited flexibility of query mechanisms for unstructured content of data lake. Based on this approach, the data warehouse has the flexibility to be created from scratch at any time.

Fang *et al.* [59] demonstrated how medical Big data integration and analysis could be used to construct early prediction and intervention models. A data center based on the Hadoop platform has been built to integrate healthcare data acquired from existing patient-related information systems, such as HIS, LIS, PACS, EMR, ECG, into a data warehouse. Data standardization and consistency are achieved using extract-transform-load (ETL) technology to integrate the vast amount of unstructured data. The main difference is that methods, such as MapReduce [166], can be applied in each processing link to carry out parallel processing in those Big data.

3) DATA INTEGRATION IN PUBLIC HEALTHCARE

Obesity is a public health problem that has raised concern worldwide, and this problem requires the systematic collection, analysis, and interpretation of all factors affecting weight gain to drive health policy and promote a lifestyle, environmental and socioeconomic changes. Figure 15 shows a 'semantic ETL' service proposed in [98] which connects multiple information retrieved from different data sources:

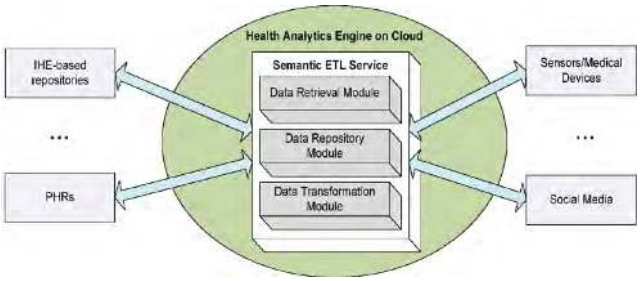

FIGURE 15. The main modules in the semantic ETL service [98].

a) IHE-based documents with patient information from the IHE repository, b) CCD-based documents with patient information from patient PHR, c) stream data from sensing devices and d) messages from various web sources. The retrieved data is stored to NoSQL databases (a combination of MongoDB and HBase) in a schema-less format to provide flexibility. Finally, the data transformation module transforms data into RDF documents, and through ontology reasoning, high-level context data is derived and transformed into documents compliant with the integration schema. RDF documents are then exported from the ontologies and stored in the data repository module.

4) DATA INTEGRATION IN WEARABLE HEALTHCARE

Mezghani *et al.* [97] proposed a collaborative semantic web platform that copes with heterogeneous Big data analysis which comes from different wearable devices, based on the Knowledge as a Service (KaaS) approach. This architecture extended NIST Big data model with a Semantic Knowledge Layer that offers a common understanding of data. The platform produces more accurate and valuable information by fusing Big heterogeneous medical data. The proposed Wearable Healthcare Ontology (WH_Ontology) is designed to deal with the heterogeneity of wearable data to ensure semantic interoperability and to allow creating more accurate knowledge about the patient, such as detecting and predicting anomalies.

5) DATA INTEGRATION IN INTEROPERABLE EHR SYSTEMS

Most medical information systems store clinical information about patients in proprietary formats. Therefore, Bahga and Madisetti [64] presents a Cloud-based approach to integrating electronic health record systems named Cloud Health Information System Technology Architecture (CHISTAR). CHISTAR reference model extends and adapts the OpenEHR and HL7 v3.0 data types. Data integration is performed to throw a data integration engine and achieved in two steps. In the first step, a source connector connects to an external system where a meta-data lookup is performed to discover the semantics of the data elements in the source file. In the next step, semantic matching is done with the meta-data repository of the destination to find and retrieve a list of candidate mappings in an intermediate file. The data loaded by the integration engine is stored as a flat file in HDFS distributed storage; therefore, a MapReduce based bulk loader loads the data from flat files into HBase. This work has been extended by a cloud-based information integration and informatics (III) framework to facilitate the collection and analysis of heterogeneous and distributed healthcare systems within a scalable cloud infrastructure [167].

6) DATA INTEGRATION IN ADVANCED PRECISION MEDICINE

Many questions for clinical research (such as how cancer arises, how much complex diseases are dependent on personal genomic traits or environmental factors) could be answered by modern genomics. Moreover, a vast, intricate and incompatible raw data (Encyclopedia of DNA elements (ENCODE), Cancer Genome Atlas (TCGA), 1000 Genomes Project, etc.) have been produced by computational efforts in primary and secondary genomic data management coupled with the progress of RNA sequencing technology. Ceri *et al.* [76] proposed a data model that ensures the interoperability between different produced formats and allows merging datasets with different schemas. They also defined a new federated query language GMQL that has the ability of computing distance-related queries along the genome, seen as a sequence of positions. The defined query is executed based on simple interaction protocol, such as 1)requesting information about remote datasets, 2)transmitting a query in high-level format and obtain data about its compilation, 3) launching query execution and then controlling the transmission of results, so as to be in control of staging resources and communication load.

A processing method is used in [65] for complex biological data processing operations to understand gene-disease associations. Both structured and unstructured data are loaded onto HDFS, and they are queried and integrated using both a commodity hardware-software cluster and a commercial Big Data System to find the records that match the given query. The data that were acquired consisted of 20 million literature abstracts obtained from PubMed in XML format, mRNA expression data and miRNA expression data from a single Glioblastoma patient downloaded from TCGA, along with a gene and disease lexicon, using EntrezGene and NCI Thesaurus, respectively. The result suggests that available technologies within the Big Data domain can reduce the time and effort needed to utilize and apply distributed queries over large datasets in practical clinical applications in the life sciences domain.

G-DOC Plus [168] is used to integrate multiple datasets such as health data selected from private and public resources, Cancer Genome Atlas (TCGA) and recently added datasets from REpository for Molecular BRAin Neoplasia DaTa (REMBRANDT), caArray studies of lung and colon cancer, ImmPort and the 1000 genomes data sets. G-DOC Plus aims to support physician, scientists and researchers to understand the mechanisms of cancer and non-cancer diseases to drive new hypothesis for precision medicine. G-DOC Plus uses MongoDB to store variant data from sequencing studies. Images are stored in a DICOM Clinical Data Manager "Dcm4chee" system, and metadata are stored in a MySQL

database. All engines are hosted on an EC2 server instance on the Amazon cloud.

7) DATA INTEGRATION IN MENTAL HEALTH CARE

Brain and mental health research from Big brain data have grown as an emerging area for both data analyst and neuroscience community. Electroencephalography (EEG) is by far the most commonly used technique to study brain function. It has proven its usefulness with advanced sensing technology and signal processing algorithms to support people with healthcare needs, such as identifying ketamine responses in treatment-resistant depression using a wearable forehead EEG [169], exploring resting-state EEG complexity before migraine attacks [170], indexing brain cortical dynamics and detecting driving fatigue and drowsiness [171]. Alternatively, functional magnetic resonance imaging (fMRI) is used extensively to identify regions linked to critical functions such as speaking, moving, sensing, or planning. Clinicians also use fMRI to further understand neurobehavioral disorders, such as Alzheimer's disease, epilepsy, brain tumors, stroke, traumatic brain injury, and multiple sclerosis [172].

Since EEG and fMRI are the two most commonly used noninvasive functional neuroimaging techniques, and because they exhibit highly complementary characteristics, their multimodal integration has been actively sought [173]. Hosseini *et al.* [174] introduce a new method for epileptogenic network definition and prediction of the seizure (ictal) onset, by integrating multimodal fMRI and EEG Big data. A deep-learning approach was developed to extract high order features for identification of interictal epileptic discharge (IED) and nonIED time intervals in electrographic data, leveraging the emerging mobile-edge computing platform. Figure 16 illustrates the pipeline of this integration.

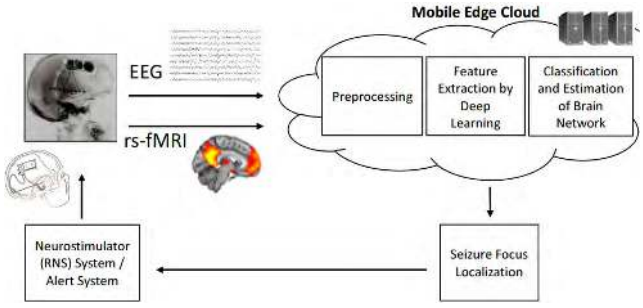

FIGURE 16. Integration workflow of the approach for the analysis of independently acquired EEG and RS-fMRI data to locate the epileptogenic site.

IV. DISCUSSION

In the previous section, We have extensively studied existing solutions found in the literature related to Big Healthcare Data integration. We investigated the ins and outs of these solutions from different perspectives including the mode of operations (real-time or batch style or both), what specific problems it solves, which of the Big Data characteristics it deals with (volume or velocity), does it tackle the quality problems, etc. We summarize our findings in Table 2. In what follows, we will assess the weaknesses and challenges of Big Healthcare Data Integration solutions:

A. WEAKNESSES

We found some powerful solutions which can integrate data (section III-F). However, we found a few critical problems which are unsolved in existing technologies:

- *Lack of powerful solution*: There is no ready to use solution (similar to commercial off the shelf) for medical data integration. The existing solutions were developed, aiming at building a solution for a specific area of the healthcare domain. For instance, the integration of cancer data. Such solutions can integrate some particular types of data such as image and text but cannot integrate surgical videos. Also, some solutions integrate a kind of data that relies on a specific data model such as RDF.
- *Lack of solution for Integrating Medical Data Streams*: In our study, we have not found an efficient solution which can be used to integrate medical data streaming from sources such as social media and sensors. Such a solution is enormously useful for performing a critical analysis on the fly.
- *Lack of solution for integration with data quality consideration*: Medical data integration faces a quality challenge. Integration of heterogeneous data may produce messy data or data that are not meaningful. This is why integration solutions must solve quality issues. Unfortunately, we did not find any solution which can address both issues effectively.
- *Lack of solution for integration with real-time responses*: Many of the existing solutions are built on the Hadoop-MapReduce framework, which mostly solved the data volume challenge. However, due to the extravagant sorting algorithm that Hadoop relies heavily on for performing reduce function, the performance can be a bottleneck.

B. CHALLENGES

Our survey found several challenges regarding processing Big Data [132], [175], and specifically for healthcare Big data [1], [13], [176]. Integrating Big healthcare data has its challenges. Recent studies show that Big healthcare data could not be efficiently integrated using traditional techniques, technologies, and tools. Therefore, many issues have not been addressed and need to be answered. The most important challenges are briefly presented below:

- *Heterogeneous data:* Data integration and fusion of the noisy, heterogeneous and longitudinal data generated by different technologies such as medical imaging, physiological signal, genomic data [177], and social media [178] constitute the Big challenge for healthcare informatics. These data may be diverse in terms of Data types, file formats, data encoding, the data model (syntactic heterogeneity) as well as they may have differences in meanings and interpretations (semantic heterogeneity).

TABLE 2. Summary of data integration applications in Healthcare context.

Name of the solution	Type of data integrated by the solution	The Big Data characteristics it deals with	Type of operations it performs	Main strength	Major Weakness
Big Data Technologies: New Opportunities for Diabetes Management [165]	Billing healthcare information, EMR of the FSM diabetes unit, Environmental data from satellite Landsat.	volume, variety, velocity	Gathering and storing data in multidimensional data star schema	Integration of large volume with different types of data: environmental, and medical	Use data warehouse technique, that could have a limitation with Big data
iManageCancer: Developing a Platform for Empowering Patients and Strengthening Self-Management in Cancer Diseases [61]	EHR, Serious games, Sensors.	volume, variety, velocity, veracity	The data is stored in Data Lake and loaded into a semantic warehouse as a batch process	Semantic data warehouse has the flexibility to be created from scratch at any time	Using semantic layer approach with a static build phase is too slow to keep up activity monitoring and sensor data
A study on specialist or special disease clinics based on Big data [59]	HIS, LIS, PACS, EMR, ECG, ultrasound.	volume, variety	ETL batch processing	MapReduce is used in the ETL process to deal with the volume of data	Multidimensional cube used to store data has a limitation concerning scalability
A health analytics semantic ETL service for obesity surveillance [98]	IHE-based documents, CCD-based documents, sensing devices, text from various web sources.	volume, variety, velocity, veracity	Semantic ETL service on the cloud	Use the semantic web to handle the data quality	The proposed semantic ETL services do not handle the volume issue
A cloud-based approach for interoperable electronic health records, (EHRs) [64]	EHR from heterogeneous, and distributed healthcare systems	volume, variety	A cloud-based distributed batch processing	supports advanced security features	Semantic interoperability could have many limitations with huge data.
Data Management for Next Generation Genomic Computing [76]	Encyclopedia of DNA, TCGA, 1000 Genomes Project	volume, variety	Use federate technique to integrate data	Define a new federated query language GMQL	Merging the transmission of results from multiple data sources could be an issue
Knowledge and Theme Discovery across Very Large Biological Data Sets,Using Distributed Queries [65]	PubMed, mRNA and miRNA from Glioblastoma	volume, variety, veracity	Hybrid solution with distributed queries and batch processing	Map-reduce based software	The proposed extraction method could have an issue with huge data
A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare [97]	Wearable data from multiple sources, with multiple formats	volume, variety, velocity	Adopt the NIST Big Data reference architecture and store wearable data in distributed clusters deployed in the cloud	Wearable Healthcare Ontology facilitating the aggregation of distributed heterogeneous data	The proposed SPARQL endpoint could become a bottleneck
G-DOC Plus: an integrative bioinformatics platform for precision medicine [168]	Patients data, Gene Expression Omnibus, Cancer Genome Atlas, Brain Neoplasia Data, 1000 genomes data sets	volume, variety	Researchers explore, search and link data as a cohort of samples and save results in G-DOC Plus for further downstream analysis.	Offers advanced computational tools to integrate a variety of biomedical Big data	The integration of patients data in the systems requires transformation, not only of the format but of the model as well.
Deep Learning with Edge Computing for Localization of Epileptogenicity using Multimodal rs-fMRI and EEG Big Data [174]	Clinical dataset of fMRI and EEG	volume, variety, velocity	Deep learning structures use a hierarchical multilevel learning approach to extract meaningful features.	Autonomic edge-computing platform that supports deep learning	The integration method does not take advantage of the patient's medical history, which could affect the evaluation of epilepsy

- *Unstructured data:* Clinical context produces unstructured data (or at least in a semi-structured form) such as handwritten doctor notes, images, audio, and video streams. These unstructured resources contain a richness of information relevant to understanding human health [8].
- *Problems with data standards:* Medical data usually lack consistent data standards and are often fragmented or generated in legacy IT systems with an incompatible structure, which exacerbates the inconsistency of medical terminology. Integration problems in healthcare include gaps in data standards, overlapping standards, and multiple data standards development organizations [52].
- *Health-monitoring data:* Real-time integration of health-monitoring data such as vital signs monitoring devices, environmental exposure, and pharmacological profiles, into the existing medical data poses several technical challenges [179].
- *Patient privacy and data security:* There is a growing interest in the security of electronic medical information that is distributed in confidential silos owned by a multitude of stakeholders [8]. The patient has the right to determine when, how, and to what extent his health

information is shared with others. That is, during data integration processes, essential factors related to patient privacy and consent and other legal issues related to these data need to be considered [180]. These statutory and regulatory aspects could create a potential barrier to the proper implementation of the data integration process.

- *Non-expert users:* Data integration should be an automatic process. Due to the scale and heterogeneity of medical data, automatic integration is not completely accurate. End-users of integrated data are generally physicians, nurses, and health professionals with limited informatics training [181]. It would, therefore, be difficult to help non-expert users access heterogeneous data sources via data integration systems.

V. RESEARCH DIRECTIONS

We discovered some promising research directions related to the integration of Big Healthcare Data. We briefly explain these directions in the following:

A. BIG HEALTHCARE DATA FUSION

The advent of Big Data has given rise to a new notion called *Data Fusion*, which is an extended concept of data integration. In data integration, data are gleaned from multiple heterogeneous sources, whereas fusion consists of data integration followed by data reduction or replacement operations. Fusion adds different levels of uncertainty to support a more narrow set of application workloads, which is critical specifically to perform analysis efficiently. Therefore, Big data fusion has become an important issue to healthcare industry practitioners and researchers and is gaining popularity as a concept for building efficient solutions. According to our study, Big Healthcare data fusion is an open and critical issue which can be dealt with by producing novel fusion techniques.

B. INFORMATION EXTRACTION

Data annotation technique aids data integration process to understand and fuse medical data such as images (MRI, radiology image, CT scan, molecular image), text (medical report, academic article), EHRs. These different types of data formats, make the annotation process of medical data very difficult compared to other domains. Moreover, the evolution of medical data has created a large heterogeneity of data sources and increased complexity to extract the needed information [182]. Therefore, hot opportunities have been arising in developing new information extraction and labeling methods.

C. DEVELOPING EFFICIENT INTEGRATION TECHNIQUES

Machine learning is a possibly feasible way to improve traditional data reduction techniques to process or even preprocess Big data. These emerging techniques may help to understand the trends of data, classify Big data, and detect similarities [175]. Recent developments in deep learning and artificial neural networks emerged as the preferred machine learning approach in machine perception such as computer vision, speech recognition, and natural language processing. Therefore, combining these models will open new chances to address many of the challenges of integrating structured and unstructured data, making better leverage of the information-rich yet unstructured data in EHRs [183].

D. REAL-TIME INTEGRATION OF DATA STREAMS

The increasing trend of using smart devices in the healthcare sector for carrying out several tasks such as change detection in real-time monitoring of EEG signals [184] and using smart equipment in surgical procedures [185], requires integration of data streams on the fly. The wearable device technology is becoming popular and is gaining importance for long-term health monitoring [186]. The integration of patient-generated fitness data with Big medical data such as EMR can be compelling and robust. This integration can help clinicians in making more informed decisions about patient health [187]. Therefore, real-time data capturing from wearable medical devices regardless of the data format and the integration of all these new formats with medical data is a newly emerging research domain.

E. ADVANCED INTEGRATION TECHNOLOGIES

In recent years, in addition to acquiring healthcare data, it became possible to obtain data from various data sources (social networks, monitoring, IoT devices, etc.). This perspective raises new questions about the quality of the data. Therefore, the ADR-PRISM project [188] identified 21 criteria for evaluating social media to select the most informative data elements to support medical domain research.

Thereby, the need for new integration technologies to benefit from the methodology of Big data integration to obtain a fully integrated picture of disease causality and help in effective early detection [189]. The following are some of the research works that can be done.

- Precision Medicine: Combining genetic data with diseases, therapies, and outcomes can help to improve the selection of the best treatment. Also, integrating historical patient data about lifestyle and environmental exposure has the potential to determine the causes triggering the onset of a disease state.
- Infectious diseases early detection: Combining data from web-based searches, social information, travel, trade, climate changes, etc., with syndromic surveillance and diagnostic data including the next generation sequencing, can improve the detection of early signs of disease outbreaks (e.g. influenza, bacterial-caused food poisoning) and coordinate quarantine and vaccination responses.
- Chronic diseases detection: Combining data from social and physical behaviors, nutrition, genetic factors, environmental factors and the development of mental/physical diseases, can help to better understand the triggers of chronic diseases for effective early detection.

TABLE 3. Definitions of all acronyms used in the paper.

ADNI	Alzheimer's Disease Neuroimaging Initiative
ANN	Artificial Neural Networks
ASC X12	Accredited Standards Committee X12
ATC	Therapeutic Chemical classification system
BDW	Big Data Warehouse
CCD	Continuity of Care Document
CDA	Clinical Document Architecture
CNA	Copy Number Alterations
CNN	Convolutional Neural Networks
CUI	Concept Unique Identifier
CZ-DRUGS	Medicinal products registered in the Czech Republic
DICOM	Digital Imaging and Communications in Medicine
DNA	Deoxyribonucleic Acid
DT	Decision Tree
DW	Data Warehouse
ECG	electrocardiogram
EDW	Enterprise Data Warehouse
EMR	Electronic Medical Record
ETL	Extract, Transform, Load
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
fMRI	Functional Magnetic Resonance Imaging
HCLS	Semantic Web Health Care and Life Sciences
HDFS	Hadoop Distributed File System
HER	Electronic Health Record
HIS	Hospital Information System
HiveQL	Hive Query Language
HL7	Health Level Seven
HWO	World Health Organisation
i2b2	Informatics for Integrating Biology and the Bedside
ICD	The International Classification of Diseases
ICT	Information and communications technology
IHE	Integrating the Healthcare Enterprise
IHTSDO	International Health Terminology Standards Development Organization
KaaS	Knowledge as a Service
LIS	laboratory information system
LOD	Linked Open Data
LOINC	Logical Observation Identifiers Names and Codes
LSTM	Long Short-term Memory
MeSH	Medical Subject Headings
MHS	Multicare Health Systems
miRNA	microRNA
MRI	Magnetic resonance imaging
mRMR	minimal-redundancy-maximalrelevance criterion
mRNA	Messenger Ribonucleic Acid
NCI	National Cancer Institute
NDF-RT	National Drug File - Reference Terminology
NIS	National Inpatient Sample
OBO	Open Biomedical Ontologies
OWL	Web Ontology Language
PACS	Picture Archive and Communication System
PDE	Parallel database extension
PHR	personal health records
RDF	Resource Description Framework
RIM	Reference Information Model
RNN	Recurrent Neural Networks
RoR	Risk of Readmission
S3	Simple Storage Service
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SNP	Single Nucleotide Polymorphism
SPARQL	Simple Protocol and Rdf Query Language
SVM	Support Vector Machines
SWSE	Semantic Web Ssearch Engine
T2DM	Type 2 Diabetes Mellitus
TCGA	The Cancer Genome Atlas
UDF	User-defined Functions
UMLS	Unified Medical Language System
XDS	Cross Enterprise Document Sharing
Yarn	Yet Another Resource Negotiator

F. SCALABLE DATA INTEGRATION PLATFORM

Since medical data hold the characteristics of Big data [176], there is a need for a scalable solution that is based on the Hadoop framework. The platform should have the ability to store a huge amount of heterogeneous data and deliver a wide range of automated data integration processors like data filtering, cleaning, summarizing and link discovery, as well as it should provide a fusion solution that supports healthcare standards such as HL7, OpenEHR and others. It must also be able to merge data from internal and external sources. The main idea behind this platform is to summarize and integrate data graphically and interactively, enabling domain experts to find key information for supporting decision making interactively.

VI. CONCLUSION

The rapid growth of healthcare data has given rise to the notion of Big Healthcare Data. A wide variety of data are available in the healthcare sector, including text (e.g., prescriptions), images (e.g., MRI scanning reports), video (recorded operations in a surgical room), etc. Furthermore, in recent years, sensors and social media data are heavily used within the healthcare sector for various purposes. These data flow with high-speed. Uncertainty is a common issue of data when it flows from outside of the well-known and reliable internal repositories. The complex nature of data, which include volume, variety, speed, and uncertainty raised a massive challenge for traditional technologies, e.g., spreadsheet-based tool, relational database, and single machine programming. The advent of Big Data promoted the growth of Big Data technologies for various operations, including collection, integration, processing, analysis, and visualization, over the years. In this paper, we studied existing technologies extensively. Our study focused on finding the strength and weaknesses of these technologies.

In our survey, We found limitations in existing model technologies that are based on the Hadoop platform, NoSQL, parallel programming. We reported these limitations in this paper. We analyzed our findings concerning different parameters such as the ability of current integration technologies to deal with variety, speed, uncertainty. Furthermore, We discussed a few promising research directions. We planned to conduct an experimental study with some of the potential solutions for integrating Big Healthcare Data.

APPENDIX

All the acronyms used in the paper are enlisted in Table 3.

REFERENCES

[1] M. H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, "Health big data analytics: Current perspectives, challenges and potential solutions," *Int. J. Big Data Intell.*, vol. 1, nos. 1–2, pp. 114–126, Jan. 2014. doi: 10.1504/ijbdi.2014.063835.

[2] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Inf.Sci. Syst.*, vol. 2, no. 1, p. 3, Dec. 2014. doi: 10.1186/2047-2501-2-3.

[3] I. de la Torre Díez, H. M. Cosgaya, B. Garcia-Zapirain, and M. López-Coronado, "Big data in health: A literature review from the year 2005," *J. Med. Syst.*, vol. 40, no. 9, p. 209, Sep. 2016. doi: 10.1007/s10916-016-0565-7.

[4] K. Verspoor and F. Martin-Sanchez, "Big data in medicine is driving big changes," *Yearbook Med. Informat.*, vol. 23, no. 1, pp. 14–20, Aug. 2014. doi: 10.15265/iy-2014-0020.

[5] N. Perakakis, A. Yazdani, G. E. Karniadakis, and C. Mantzoros, "Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics," *Metabolism*, vol. 87, pp. A1–A9, Oct. 2018. doi: 10.1016/j.metabol.2018.08.002.

[6] K. J. Karczewski and M. P. Snyder, "Integrative Omics for health and disease," *Nature Rev. Genet.*, vol. 19, no. 5, pp. 299–310, Feb. 2018. doi: 10.1038/nrg.2018.4.

[7] D. Zeevi, *et al.*, "Personalized nutrition by prediction of glycemic responses," *Cell*, vol. 163, no. 5, pp. 1079–1094, Nov. 2015. doi: 10.1016/j.cell.2015.11.001.

[8] B. Feldman, E. M. Martin, and T. Skotnes. (Oct. 2015). *Big Data in Healthcare Hype and Hope*. [Online]. Available: https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf

[9] M. J Paul and M. Dredze, "You are what you tweet: Analyzing Twitter for public health," in *Proc. ICWSM*, vol. 20, 2011, pp. 265–272.

[10] *Twitter Usage Statistics–Internet Live Stats*. Accessed: Apr. 17, 2019. [Online]. Available: http://www.internetlivestats.com/twitter-statistics/

[11] F. Gesualdo, G. Stilo, E. Agricola, M. V. Gonfiantini, E. Pandolfi, P. Velardi, and A. E. Tozzi, "Influenza-like illness surveillance on twitter through automated learning of naïve language," *PLoS One*, vol. 8, no. 12, Dec. 2013, Art. no. e82489. doi: 10.1371/journal.pone.0082489.

[12] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D109–D114, Nov. 2011. doi: 10.1093/nar/gkr988.

[13] R. Bellazzi, "Big data and biomedical informatics: A challenging opportunity," *Yearbook Med. Informat.*, vol. 23, no. 1, pp. 08–13, Aug. 2014. doi: 10.15265/iy-2014-0024.

[14] R. Chisholm, J. Denny, and D. Fridsma, "Opportunities and Challenges related to the use of Electronic Health Records data for research," Nat. Inst. Health, Bethesda, MD, USA, Feb. 2015. Accessed: Jun. 6, 2019. [Online]. Available: https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/opportunities-challenges-electronic-health-records.pdf

[15] 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 Res. Int.*, vol. 2014, pp. 1–13, Sep. 2014. doi: 10.1155/2014/134023.

[16] K. Priyanka and N. Kulennavar, "A survey on big data analytics in health care," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5865–5868, 2014.

[17] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: A literature review," *Biomed. Informat. Insights*, vol. 8, Jan. 2016, Art. no. BII.S31559. doi: 10.4137/bii.s31559.

[18] K. Jee and G.-H. Kim, "Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system," *Healthcare Informat. Res.*, vol. 19, no. 2, p. 79, 2013. doi: 10.4258/hir.2013.19.2.79.

[19] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: Promise and challenges," *Nature Rev. Cardiol.*, vol. 13, no. 6, pp. 350–359, Mar. 2016. doi: 10.1038/nrcardio.2016.42.

[20] R. Lenz, M. Beyer, and K. A. Kuhn, "Semantic integration in healthcare networks," *Int. J. Med. Inform.*, vol. 76, nos. 2–3, pp. 201–207, Feb. 2007. doi: 10.1016/j.ijmedinf.2006.05.008.

[21] Z. Zhang, B. V. Bajic, J. Yu, K. H. Cheung, and J. P. Townsend, "Data integration in bioinformatics: Current efforts and challenges," in *Bioinformatics-Trends and Methodologies*. Rijeka, Croatia: InTech, Nov. 2011. doi: 10.5772/21654.

[22] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, Jul. 2014.

[23] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[24] D. Laney, "3d data management: Controlling data vol. velocity, and variety," META, San Francisco, CA, USA, Tech. Rep. 670, 2001.

[25] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016. doi: 10.1016/j.inffus.2015.08.005.

[26] E. Baro, S. Degoul, R. Beuscart, and E. Chazard, "Toward a literature-driven definition of big data in healthcare," *BioMed Res. Int.*, vol. 2015, pp. 1–9, Jul. 2015. doi: 10.1155/2015/639021.

[27] C. Auffray, *et al.*, "Making sense of big data in health research: Towards an EU action plan," *Genome Med.*, vol. 8, no. 1, p. 71, Jun. 2016. doi: 10.1186/s13073-016-0323-y.

[28] T. White, *Hadoop: The Definitive Guide*, 4th ed. Beijing, China: O'Reilly, 2015. [Online]. Available: https://www.safaribooksonline.com/library/view/hadoop-the-definitive/9781491901687/

[29] R. Cattell, "Scalable SQL and NoSQL data stores," *ACM SIGMOD Rec.*, vol. 39, no. 4, pp. 12–27, Dec. 2017. doi: 10.1145/1978915.1978919.

[30] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, Jan. 2015. doi: 10.1016/j.is.2014.07.006.

[31] A. Luckow, K. Kennedy, F. Manhardt, E. Djerekarov, B. Vorster, and A. Apon, "Automotive big data: Applications, workloads and infrastructures," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 1201–1210. doi: 10.1109/bigdata.2015.7363874.

[32] V. K. Vavilapalli, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, E. Baldeschwieler, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, and H. Shah, "Apache Hadoop YARN: Yet another resource negotiator," in *Proc. 4th Annu. Symp. Cloud Comput.*, Oct. 2013, p. 5. doi: 10.1145/2523616.2523633.

[33] H. Fang, "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem," in *Proc. IEEE Int. Conf. Cyber Technol. Automat., Control, Intell. Syst. (CYBER)*, Jun. 2015, pp. 820–824. doi: 10.1109/cyber.2015.7288049.

[34] B. Stein and A. Morrison, "The enterprise data lake: Better integration and deeper analytics," *PwC Technol. Forecast Rethinking Integr.*, vol. 1, nos. 1–9, p. 18, 2014.

[35] *Microsoft Word—Modern Data Architecture Hadoop*. Accessed: Jun. 1, 2019. [Online]. Available: http://www.stratebi.es/todobi/Sep16/WP_EN_BD_Talend_ModernDataArchitecture_Hadoop.pdf

[36] L. Haas, "Beauty and the beast: The theory and Practice of information integration," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2006, pp. 28–43. doi: 10.1007/11965893_3.

[37] P. Ziegler and K. R. Dittrich, "Data integration—Problems, approaches, and perspectives," in *Conceptual Modelling in Information Systems Engineering*. Berlin, Germany: Springer, 2007, pp. 39–58. doi: 10.1007/978-3-540-72677-7_3.

[38] J. Bleiholder and F. Naumann, "Data fusion," *ACM Comput. Surv.*, vol. 41, no. 1, pp. 1–41, Dec. 2008. doi: 10.1145/1456650.1456651.

[39] S. Bergamaschi, D. Beneventano, F. Guerra, and M. Orsini, *Data Integration*, D. W. Embley and B. Thalheim, Eds. Berlin, Germany: Springer, 2011. doi: 10.1007/978-3-642-15865-0.

[40] X. L. Dong and F. Naumann, "Data fusion," *VLDB Endowment*, vol. 2, no. 2, pp. 1654–1655, Aug. 2009. doi: 10.14778/1687553.1687620.

[41] E. Rahm, "The case for holistic data integration," in *Advances in Databases and Information Systems*. New York, NY, USA: Springer, 2016, pp. 11–27. doi: 10.1007/978-3-319-44039-2_2.

[42] *Delivering Ehealth Ireland*. Accessed: Jun. 1, 2019. [Online]. Available: http://www.ehealthireland.ie/Library/Document-Library/Standards-Catalogue-v1_0.pdf

[43] R. Qamar and A. Rector, "Semantic mapping of clinical model data to biomedical terminologies to facilitate data interoperability," in *Proc. HealthCare Computing Conf.*. San Francisco, CA, USA: Citeseer, 2007, pp. 1–9.

[44] D. Kalra, T. Beale, and S. Heard, "The openEHR foundation," in *Studies in Health Technology and Informatics*, vol. 115. Amsterdam, The Netherlands: IOS Press, 2005, pp. 153–173.

[45] R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J. E. Mattison, "The HL7 clinical document architecture," *J. Amer. Med. Inform. Assoc.*, vol. 8, no. 6, pp. 552–569, Nov. 2001. doi: 10.1136/jamia.2001.0080552.

[46] D. Bender and K. Sartipi, "HL7 FHIR: An agile and RESTful approach to healthcare information exchange," in *Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2013, pp. 326–331. doi: 10.1109/cbms.2013.6627810.

[47] P. L. Elkin, S. H. Brown, C. S. Husser, B. A. Bauer, D. Wahner-Roedler, S. T. Rosenbloom, and T. Speroff, "Evaluation of the content coverage of SNOMED CT: Ability of SNOMED clinical terms to represent clinical problem lists," *Mayo Clinic Proc.*, vol. 81, no. 6, pp. 741–748, Jun. 2006. doi: 10.4065/81.6.741.

[48] H. Quan, B. Li, L. D. Saunders, G. A. Parsons, C. I. Nilsson, A. Alibhai, and W. A. Ghali, "Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually Coded database," *Health Services Res.*, vol. 43, no. 4, pp. 1424–1441, Jan. 2008. doi: 10.1111/j.1475-6773.2007.00822.x.

[49] S. M. Huff, R. A. Rocha, C. J. McDonald, G. J. E. D. Moor, T. Fiers, W. D. Bidgood, A. W. Forrey, W. G. Francis, W. R. Tracy, D. Leavelle, F. Stalling, B. Griffin, P. Maloney, D. Leland, L. Charles, K. Hutchins, and J. Baenziger, "Development of the logical observation identifier names and codes (LOINC) vocabulary," *J. Amer. Med. Inform. Assoc.*, vol. 5, no. 3, pp. 276–292, May 1998. doi: 10.1136/jamia.1998.0050276.

[50] *Hl7 Standards Product Brief—Hl7 Version 2 Product Suite | Hl7 International*. Accessed: Nov. 4, 2019. [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185

[51] O. S. Pianykh, *Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide*. New York, NY, USA: Springer, 2009.

[52] P. Brooks, "Standards and interoperability in healthcare information systems: Current status, problems, and research issues," in *Proc. 5th MWAIS Conf.*, 2010, pp. 1–9. [Online]. Available: https://aisel.aisnet.org/mwais2010/18

[53] J. Widom, "Integrating heterogeneous databases: Lazy or eager?" *ACM Comput. Surv.*, vol. 28, no. 4es, p. 91, Dec. 1996. doi: 10.1145/242224.242344.

[54] D. Austin, K. E. Noble, R. J. Lotero, and S. Chadalavada, "Method and mechanism for data consolidation," U.S. Patent 6 615 220 B1, Sep. 7, 2009.

[55] W. H. Inmon, "What is a data warehouse?" *Prism Tech Topic*, vol. 1, no. 1, pp. 1–5, 1995. [Online]. Available: http://repository.binus.ac.id/2009-2/content/M0584/M058459913.pdf

[56] T. K. Das and A. Mohapatro, "A study on big data integration with data warehouse," *Int. J. Comput. Trends Technol.*, vol. 9, no. 4, pp. 188–192, Mar. 2014. doi: 10.14445/22312803/ijctt-v9p137.

[57] S. O. Salinas and A. C. Lemus, "Data warehouse and big data integration," *Int. J. Compt. Sci. Inf. Tech.*, vol. 9, no. 2, pp. 1–17, 2017.

[58] S. Mohanty, M. Jagadeesh, and H. Srivatsa, *Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics*. New York, NY, USA: Apress, 2013.

[59] Z. Fang, X. Fan, and G. Chen, "A study on specialist or special disease clinics based on big data," *Frontiers Med.*, vol. 8, no. 3, pp. 376–381, Sep. 2014. doi: 10.1007/s11684-014-0356-9.

[60] S. N. Murphy, P. Avillach, R. Bellazzi, L. Phillips, M. Gabetta, A. Eran, M. T. McDuffie, and I. S. Kohane, "Combining clinical and genomics queries using i2b2—Three methods," *Plos One*, vol. 12, no. 4, Apr. 2017, Art. no. e0172187. doi: 10.1371/journal.pone.0172187.

[61] H. Kondylakis, A. Bucur, F. Dong, C. Renzi, A. Manfrinati, N. Graf, S. Hoffman, L. Koumakis, G. Pravettoni, K. Marias, M. Tsiknakis, and S. Kiefer, "iManageCancer: Developing a platform for Empowering patients and strengthening self-management in cancer diseases," in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 755–760. doi: 10.1109/cbms.2017.62.

[62] H. Kondylakis, L. Koumakis, M. Tsiknakis, K. Marias, and S. Kiefer, "Big data in support of the digital cancer patient," *Tackling Big Data Life Sci.*, vol. 1, p. 104, Jan. 2016.

[63] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy, "Hive - a petabyte scale data warehouse using hadoop," in *Proc. IEEE 26th Int. Conf. Data Eng. (ICDE)*, Mar. 2010, pp. 996–1005. doi: 10.1109/icde.2010.5447738.

[64] A. Bahga and V. K. Madisetti, "A cloud-based approach for interoperable electronic health records (EHRs)," *IEEE J. Biomed. Health Inform.*, vol. 17, no. 5, pp. 894–906, Sep. 2013. doi: 10.1109/jbhi.2013.2257818.

[65] U. S. Mudunuri, M. Khouja, S. Repetski, G. Venkataraman, A. Che, B. T. Luke, F. P. Girard, and R. M. Stephens, "Knowledge and theme discovery across very large biological data sets using distributed queries: A prototype combining unstructured and structured data," *PLoS One*, vol. 8, no. 12, Dec. 2013, Art. no. e80503. doi: 10.1371/journal.pone.0080503.

[66] Q. Yao, Y. Tian, P.-F. Li, L.-L. Tian, Y.-M. Qian, and J.-S. Li, "Design and development of a medical big data processing system based on Hadoop," *J. Med. Syst.*, vol. 39, no. 3, p. 23, Feb. 2015. doi: 10.1007/s10916-015-0220-8.

[67] J. Roski, G. W. Bo-Linn, and T. A. Andrews, "Creating value in health care through big data: Opportunities and policy implications," *Health Affairs*, vol. 33, no. 7, pp. 1115–1122, Jul. 2014. doi: 10.1377/hlthaff.2014.0147.

[68] D. D. Krause, "Data lakes and data visualization: An innovative approach to addressing the many challenges of health workforce planning," *Online J. Public Health Informat.*, vol. 7, no. 3, p. 7, Dec. 2015. doi: 10.5210/ojphi.v7i3.6047.

[69] M. Karpathiotakis, I. Alagiannis, T. Heinis, M. Branco, and A. Ailamaki, "Just-in-time data virtualization: Lightweight data management with ViDa," in *Proc. 7th Biennial Conf. Innov. Data Syst. Res.(CIDR)*, 2015, p. 25.

[70] A. Chandramouly, B. D. D. Owner, I. N. Patil, C. Engineer, I. R. Ramamurthy, I. S. R. Krishnan, B. S. Lead, and I. J. Story, "Integrating data warehouses with data virtualization for BI agility," in *Proc. Intel IT–Big Data Businedd Intell.*, 2013, pp. 1–8.

[71] A. Pattanayak, "Data virtualization with SAP HANA smart data access," *J. Comput. Commun.*, vol. 5, no. 8, pp. 62–68, Jun. 2017. doi: 10.4236/jcc.2017.58005.

[72] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Comput. Surv.*, vol. 22, no. 3, pp. 183–236, Sep. 1990. doi: 10.1145/96602.96604.

[73] S. Srinivas, *An Introduction to HDFs Federation*. Santa Clara CA, USA: Hortonworks, 2011.

[74] W. Liu and E. Park, "Big data as an e-health service," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2014, pp. 982–988. doi: 10.1109/iccnc.2014.6785471.

[75] A. Hasnain, Q. Mehmood, S. S. E. Zainab, M. Saleem, C. Warren, D. Zehra, S. Decker, and D. Rebholz-Schuhmann, "Biofed: Federated query processing over life sciences linked open data," *J. Biomed. Semantics*, vol. 8, no. 1, p. 13, Mar. 2017. doi: 10.1186/s13326-017-0118-0.

[76] S. Ceri, A. Kaitoua, M. Masseroli, P. Pinoli, and F. Venco, "Data management for next generation genomic computing," in *Proc. 19th Int. Conf. Extending Database*, Jul. 2016, pp. 485–490. doi: 10.5441/002/edbt.2016.46.

[77] J. Schiefer, J.-J. Jeng, and R. M. Bruckner, "Managing continuous data integration flows," in *Proc. CAiSE Workshops*, 2003, pp. 1–9.

[78] C. Constantinescu, U. Heinkel, and H. Meinecke, "A data change propagation system for enterprise application integration," *Inf. Syst.*, vol. 1, p. 9, Aug. 2002.

[79] L. Constantinescu, J. Kim, and D. Feng, "SparkMed: A framework for dynamic integration of multimedia medical data into distributed m-health systems," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 1, pp. 40–52, Jan. 2012. doi: 10.1109/titb.2011.2174064.

[80] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014. doi: 10.1109/tkde.2013.109.

[81] Y. Zhang, A. Li, C. Peng, and M. Wang, "Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning," *IEEE Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 825–835, Sep. 2016. doi: 10.1109/tcbb.2016.2551745.

[82] N. Shadbolt, T. Berners-Lee, and W. Hall, "The semantic Web revisited," *IEEE Intell. Syst.*, vol. 21, no. 3, pp. 96–101, May 2006. doi: 10.1109/mis.2006.62.

[83] A. Langegger and W. Wöß, and M. Blöchl, "A semantic Web middleware for Virtual data integration on the Web," in *The Semantic Web: Research and Applications*. Berlin, Germany: Springer, 2008, pp. 493–507. doi: 10.1007/978-3-540-68234-9_37.

[84] *Owl Web Ontology Language Overview*. Accessed: Apr. 21, 2019. [Online]. Available: https://www.w3.org/TR/owl-features/

[85] C. Tao, W.-Q. Wei, H. R. Solbrig, G. Savova, and C. G. Chute, "CNTRO: A semantic Web ontology for temporal relation inferencing in clinical narratives," in *Proc. AMIA Annu. Symp.*, 2010, p. 787.

[86] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca–Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnol.*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007. doi: 10.1038/nbt1346.

[87] Q. Zhu and C. Tao, "Pharmacological class data representation in the Web ontology language (OWL)," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2014, pp. 77–84. doi: 10.1109/bigdata.2014.7004397.

[88] *Snomed International*. Accessed: Nov. 5, 2019. [Online]. Available: https://www.snomed.org/snomed-ct/
[89] *Who | ICD-11 Revision*. Accessed: Jan. 6, 2019. [Online]. Available: http://www.who.int/classifications/icd/revision/en/
[90] C. A. Knoblock and P. Szekely, "Semantics for big data integration and analysis," in *Proc. AAAI Fall Symp. Series*, 2013, p. 12.
[91] D. Ostrowski, N. Rychtyckyj, P. MacNeille, and M. Kim, "Integration of big data using semantic Web technologies," in *Proc. IEEE 10th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2016, pp. 382–385. doi: 10.1109/icsc.2016.101.
[92] H. Dhayne, R. K. Chamoun, and M. Sokhn, "Survey: When semantics meet crowdsourcing to enhance big data variety," in *Proc. IEEE Middle East North Afr. Commun. Conf. (MENACOMM)*, Apr. 2018, pp. 1–6. doi: 10.1109/menacomm.2018.8371011.
[93] S. K. Bansal, "Towards a semantic extract-transform-load (ETL) framework for big data integration," in *Proc. IEEE Int. Congr. Big Data*, Jun. 2014, pp. 522–529. doi: 10.1109/bigdata.congress.2014.82.
[94] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, Jul. 2009. doi: 10.4018/jswis.2009081901.
[95] *Hclsig/lodd/data - W3c Wiki*. Accessed: Nov. 5, 2019. [Online]. Available: https://www.w3.org/wiki/HCLSIG/LODD/Data
[96] M. S. Marshall, R. Boyce, H. F. Deus, J. Zhao, E. L. Willighagen, M. Samwald, E. Pichler, J. Hajagos, and E. Prud'hommeaux, and S. Stephens, "Emerging practices for mapping and linking life sciences data using RDF—A case series," *J. Web Semantics*, vol. 14, pp. 2–13, Jul. 2012. doi: 10.1016/j.Websem.2012.02.003.
[97] E. Mezghani, E. Exposito, K. Drira, M. D. Silveira, and C. Pruski, "A semantic big data platform for integrating heterogeneous wearable data in healthcare," *J. Med. Syst.*, vol. 39, no. 12, p. 185, Oct. 2015. doi: 10.1007/s10916-015-0344-x.
[98] M. Poulymenopoulou, D. Papakonstantinou, and F. Malamateniou, "A health analytics semantic ETL service for obesity surveillance," *Studies Health Technol. Inform.*, vol. 210, pp. 840–844, May 2015. doi: 10.3233/978-1-61499-512-8-840.
[99] M. Saleem, M. R. Kamdar, A. Iqbal, S. Sampath, H. F. Deus, and A.-C. N. Ngomo, "Big linked cancer data: Integrating linked TCGA and PubMed," *J. Web Semantics*, vols. 27–28, pp. 34–41, Aug. 2014. doi: 10.1016/j.Websem.2014.07.004.
[100] J. Kozák, M. Nečaský, J. Dědek, J. Klímek, and J. Pokorný, "Linked open data for healthcare professionals," in *Proc. Int. Conf. Inf. Integr. Web-based Appl. Services*, 2013, p. 400. doi: 10.1145/2539150.2539195.
[101] L. Deng and D. Yu, "Deep learning methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, 2014.
[102] H. Yang, J. Liu, J. Sui, G. Pearlson, and V. D. Calhoun, "A hybrid machine learning method for fusing fMRI and genetic data: Combining both improves classification of schizophrenia," *Frontiers Hum. Neurosci.*, vol. 4, p. 192, Oct. 2010. doi: 10.3389/fnhum.2010.00192.
[103] W. Ding, C.-T. Lin, M. Prasad, Z. Cao, and J. Wang, "A layered-coevolution-based attribute-boosted reduction using adaptive quantum-behavior PSO and its consistent segmentation for neonates brain tissue," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1177–1191, Jun. 2018. doi: 10.1109/tfuzz.2017.2717381.
[104] W. Ding, C.-T. Lin, and Z. Cao, "Deep Neuro-cognitive co-evolution for fuzzy attribute reduction by quantum leaping PSO with nearest-neighbor memeplexes," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2744–2757, Jul. 2019. doi: 10.1109/tcyb.2018.2834390.
[105] K. Polat, S. Güne , and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 482–487, Jan. 2008. doi: 10.1016/j.eswa.2006.09.012.
[106] S. Oh, M. S. Lee, and B.-T. Zhang, "Ensemble learning with active example selection for imbalanced biomedical data classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 2, pp. 316–325, Mar. 2011. doi: 10.1109/tcbb.2010.96.
[107] E. A. Manilich, Z. M. Özsoyoglu, V. Trubachev, and T. Radivoyevitch, "Classification of large microarray datasets using fast random forest construction," *J. Bioinf. Comput. Biol.*, vol. 9, no. 2, pp. 251–267, Apr. 2011. doi: 10.1142/s021972001100546x.
[108] M. G. Jaffe, G. A. Lee, J. D. Young, S. Sidney, and A. S. Go, "Improved blood pressure control associated with a large-scale hypertension program," *JAMA*, vol. 310, no. 7, p. 699, Aug. 2013. doi: 10.1001/jama.2013.108769.
[109] E. N. de Vries, H. A. Prins, R. M. Crolla, A. J. den Outer, G. van Andel, S. H. van Helden, W. S. Schlack, M. A. van Putten, D. J. Gouma, M. G. Dijkgraaf, S. M. Smorenburg, and M. A. Boermeester, "Effect of a comprehensive surgical safety system on patient outcomes," *New England J. Med.*, vol. 363, no. 20, pp. 1928–1937, Nov. 2010. doi: 10.1056/nejmsa0911535.
[110] S. K. Popat and M. Emmanuel, "Review and comparative study of clustering techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 805–812, 2014.
[111] B. Chandra and R. K. Sharma, "Fast learning for big data applications using parameterized multilayer perceptron," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2014, pp. 17–22. doi: 10.1109/bigdata.2014.7004351.
[112] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Med.*, vol. 7, no. 6, pp. 673–679, Jun. 2001. doi: 10.1038/89044.
[113] T. J. Hirschauer, H. Adeli, and J. A. Buford, "Computer-aided diagnosis of parkinson's disease using enhanced probabilistic neural network," *J. Med. Syst.*, vol. 39, no. 11, p. 179, Sep. 2015. doi: 10.1007/s10916-015-0353-9.
[114] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014. doi: 10.1109/access.2014.2325029.
[115] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019. doi: 10.1038/s41591-018-0316-z.
[116] X. L. Dong and T. Rekatsinas, "Data integration and machine learning," in *Proc. Int. Conf. Manage. Data*, Jun. 2018, pp. 1645–1650. doi: 10.1145/3183713.3197387.
[117] H. Köpcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," *VLDB Endowment*, vol. 3, nos. 1–2, pp. 484–493, Sep. 2010. doi: 10.14778/1920841.1920904.
[118] G. Soğancoğlu, H. Öztürk, and A. Özgür, "BIOSSES: A semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, Jul. 2017. doi: 10.1093/bioinformatics/btx238.
[119] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin, "Relation extraction with matrix factorization and universal schemas," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, Jun. 2013, pp. 74–84.
[120] Y. Li, F.-X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings Bioinf.*, vol. 19, no. 2, pp. 325–340, Dec. 2016. doi: 10.1093/bib/bbw113.
[121] K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S.-C. Chin, and B. Muckian, "Big data solutions for predicting risk-of-readmission for congestive heart failure patients," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 64–71. doi: 10.1109/bigdata.2013.6691760.
[122] F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. Dämato, and D. Greco, "Drug repositioning: A machine-learning approach through data integration," *J. Cheminformat.*, vol. 5, no. 1, p. 30, Dec. 2013. doi: 10.1186/1758-2946-5-30.
[123] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 841–850, Jun. 2018. doi: 10.1109/tcbb.2018.2806438.
[124] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017. doi: 10.1109/access.2017.2696365.
[125] X. Wang and W. Cao, "Non-iterative approaches in training feed-forward neural networks and their applications," *Soft Comput.*, vol. 22, no. 11, pp. 3473–3476, Jun. 2018. doi: 10.1007/s00500-018-3203-0.
[126] N. Sokolovska, K. Clément, and J.-D. Zucker, "Deep kernel dimensionality reduction for scalable data integration," *Int. J. Approx. Reasoning*, vol. 74, pp. 121–132, Jul. 2016. doi: 10.1016/j.ijar.2016.03.008.
[127] S. Sarawagi, "Information extraction," *Found. Trends Databases*, vol. 1, no. 3, pp. 261–377, 2008.
[128] M. B. Habib and M. van Keulen, "Information extraction, data integration, and uncertain data management: The state of the art," CTIT, New York, NY, USA, Tech. Rep. TR-CTIT-11-06, 2011.

[129] P. LePendu, S. V. Iyer, A. Bauer-Mehren, R. Harpaz, J. M. Mortensen, T. Podchiyska, T. A. Ferris, and N. H. Shah, "Pharmacovigilance using clinical notes," *Clin. Pharmacol Therapeutics*, vol. 93, no. 6, pp. 547–555, Mar. 2013. doi: 10.1038/clpt.2013.47.

[130] N. J. Leeper, A. Bauer-Mehren, S. V. Iyer, P. LePendu, C. Olson, and N. H. Shah, "Practice-based evidence: Profiling the safety of cilostazol by text-mining of clinical notes," *PLoS One*, vol. 8, no. 5, May 2013, Art. no. e63499. doi: 10.1371/journal.pone.0063499.

[131] F. Estella, B. L. Delgado-Marquez, P. Rojas, O. Valenzuela, B. S. Roman, and I. Rojas, "Advanced system for automously classify brain MRI in neurodegenerative disease," in *Proc. Int. Conf. Multimedia Comput. Syst.*, May 2012, pp. 250–255. doi: 10.1109/icmcs.2012.6320281.

[132] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, Aug. 2012. doi: 10.14778/2367502.2367572.

[133] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: A survey," *Comput. Sci. Rev.*, vol. 17, pp. 70–81, Aug. 2015. doi: 10.1016/j.cosrev.2015.05.002.

[134] D. Chen, Y. Chen, B. N. Brownlow, P. P. Kanjamala, C. A. G. Arredondo, B. L. Radspinner, and M. A. Raveling, "Real-time or near real-time persisting daily healthcare data into HDFS and ElasticSearch index inside a big data platform," *IEEE Trans. Ind. Inform.*, vol. 13, no. 2, pp. 595–606, Apr. 2017. doi: 10.1109/tii.2016.2645606.

[135] A. M. Carbonell, J. A. Warren, A. S. Prabhu, C. D. Ballecer, R. J. Janczyk, J. Herrera, L.-C. Huang, S. Phillips, M. J. Rosen, and B. K. Poulose, "Reducing length of stay using a robotic-assisted approach for retromuscular ventral hernia repair," *Ann. Surgery*, vol. 267, no. 2, pp. 210–217, Feb. 2018. doi: 10.1097/sla.0000000000002244.

[136] X. Zhu and X. Wu, "Class noise vs. Attribute noise: A quantitative study," *Artif. Intell. Rev.*, vol. 22, no. 3, pp. 177–210, Nov. 2004. doi: 10.1007/s10462-004-0751-8.

[137] P. C. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 1, pp. 8–13, Jan. 2015.

[138] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program," in *Proc. AMIA Symp.*, Jul. 2001, p. 17.

[139] H. Dhayne, R. Kilany, R. Haque, and Y. Taher, "SeDIE: A semantic-driven engine for integration of healthcare data," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 617–622. doi: 10.1109/bibm.2018.8621243.

[140] D. Ayala, I. Hernàndez, D. Ruiz, and M. Toro, "TAPON: A two-phase machine learning approach for semantic labelling," *Knowl.-Based Syst.*, vol. 163, pp. 931–943, Jan. 2019. doi: 10.1016/j.knosys.2018.10.017.

[141] G. Vemuganti, "Metadata management in big data," in *Big Data: Countering Tomorrow's Challenges*. 2013.

[142] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman, "LifeLines: Using visualization to enhance navigation and analysis of patient records," in *The Craft of Information Visualization*. Amsterdam, The Netherlands: Elsevier, 2003, pp. 308–312. doi: 10.1016/b978-155860915-0/50038-x.

[143] D. J. Foran, W. Chen, H. Chu, E. Sadimin, D. Loh, G. Riedlinger, L. A. Goodell, S. Ganesan, K. Hirshfield, L. Rodriguez, and R. S. DiPaola, "Roadmap to a comprehensive clinical data warehouse for precision medicine applications in oncology," *Cancer Informat.*, vol. 16, Jan. 2017, Art. no. 117693511769434. doi: 10.1177/1176935117694349.

[144] C. M. A. Tilve, A. P. Ayora, C. R. Ruiz, D. G. Llamas, L. G. Carrajo, F. J. G. Blanco, and G. G. Vázquez, "Integrating medical and research information: A big data approach," in *Studies in Health Technology and Informatics*, vol. 210. Amsterdam, The Netherlands: IOS Press, 2015, pp. 707–711. doi: 10.3233/978-1-61499-512-8-707.

[145] R. C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," *BMC Bioinf.*, vol. 11, no. S12, p. S1, Dec. 2010. doi: 10.1186/1471-2105-11-s12-s1.

[146] J. Han, E. Haihong, G. Le, and J. Du, "Survey on NoSQL database," in *Proc. 6th Int. Conf. Pervas. Comput. Appl.*, Oct. 2011, pp. 363–366. doi: 10.1109/icpca.2011.6106531.

[147] G. E. Modoni, M. Sacco, and W. Terkaj, "A survey of RDF store solutions," in *Proc. Int. Conf. Eng., Technol. Innov. (ICE)*, Jun. 2014, pp. 1–7. doi: 10.1109/ice.2014.6871541.

[148] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy, "Hive—A petabyte scale data warehouse using hadoop," in *Proc. IEEE 26th Int. Conf. Eng.*, Mar. 2010, pp. 996–1005. doi: 10.1109/icde.2010.5447738.

[149] M. Armbrust, A. Ghodsi, M. Zaharia, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, and M. J. Franklin, "Spark SQL," in *Proc. 2015 ACM SIGMOD Int. Conf. Manage. Data*, Jul. 2015, p. 25. [Online]. Available: https://doi.org/10.1145/2723372.2742797

[150] N. Marz, "Storm: Distributed and fault-tolerant realtime computation," in *Proc. Strata Conf. Making Data Work*, 2013, p. 28.

[151] A. Spark, "Spark streaming programming guide," *Acessado Em*, vol. 4, no. 2, p. 2017, Apr. 2014.

[152] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A distributed messaging system for log processing," in *Proc. NetDB*, Jun. 2011, pp. 1–7.

[153] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache flink: Stream and batch processing in a single engine," in *Proc. Bull. IEEE Comput. Soc. Tech. Committee Data Eng.*, May 2015, p. 36.

[154] M. Panahiazar, V. Taslimitehrani, A. Jadhav, and J. Pathak, "Empowering personalized medicine with big data and semantic Web technology: Promises, challenges, and use cases," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2014, pp. 790–795. doi: 10.1109/bigdata.2014.7004307.

[155] T. Singh and V. S. Darshan, "A modern data architecture with apache hadoop," in *Proc. Int. Conf. Green Comput. Internet Things (ICGCIoT)*, Oct. 2015, pp. 574–579. doi: 10.1109/icgciot.2015.7380530.

[156] W. Ding, C.-T. Lin, and Z. Cao, "Shared nearest-neighbor quantum game-based attribute reduction with hierarchical coevolutionary spark and its application in consistent segmentation of neonatal cerebral cortical surfaces," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2013–2027, Jul. 2019. doi: 10.1109/tnnls.2018.2872974.

[157] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: Preserving security and privacy," *J. Big Data*, vol. 5, no. 1, p. 17, Jan. 2018. doi: 10.1186/s40537-017-0110-7.

[158] D. R. Masys and D. B. Baker, "Patient-centered access to secure systems online (PCASSO): A secure approach to clinical data access via the world wide Web," in *Proc. AMIA Annu. Fall Symp.*, 1997, p. 340.

[159] Y. Alshboul, R. Nepali, and Y. Wang, "Big Data LifeCycle: Threats and security model," in *Proc. 21st Amer. Conf. Inf. Syst.*, 2015.

[160] C. Clifton, M. KantarcioÇ lu, A. Doan, G. Schadow, J. Vaidya, A. Elmagarmid, and D. Suciu, "Privacy-preserving data integration and sharing," in *Proc. 9th ACM SIGMOD Workshop Res. Issues Data Mining Knowl. Discovery*, 2004, pp. 19–26. doi: 10.1145/1008694.1008698.

[161] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: A technological perspective and review," *J. Big Data*, vol. 3, no. 1, p. 25, Nov. 2016. doi: 10.1186/s40537-016-0059-y.

[162] Z. Gheid and Y. Challal, "An efficient and privacy-preserving similarity evaluation for big data analytics," in *Proc. IEEE/ACM 8th Int. Conf. Utility Cloud Comput. (UCC)*, Dec. 2015, pp. 281–289.

[163] X. Zhang, L. T. Yang, C. Liu, and J. Chen, "A scalable two-phase top-down specialization approach for data anonymization using MapReduce on cloud," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 363–373, Feb. 2014. doi: 10.1109/tpds.2013.48.

[164] *i2b2: Informatics for Integrating Biology & the Bedside*. Accessed: Jun. 1, 2019. [Online]. Available: https://www.i2b2.org/

[165] R. Bellazzi, A. Dagliati, L. Sacchi, and D. Segagni, "Big data technologies," *J. Diabetes Sci. Technol.*, vol. 9, no. 5, pp. 1119–1125, Apr. 2015. doi: 10.1177/1932296815583505.

[166] J. Dean and S. Ghemawat, "MapReduce," *Commun. ACM*, vol. 51, no. 1, p. 107, Jan. 2008. doi: 10.1145/1327452.1327492.

[167] A. Bahga and V. K. Madisetti, "Healthcare data integration and informatics in the cloud," *Computer*, vol. 48, no. 2, pp. 50–57, Feb. 2015. doi: 10.1109/mc.2015.46.

[168] K. Bhuvaneshwar, A. Belouali, V. Singh, R. M. Johnson, L. Song, A. Alaoui, M. A. Harris, R. Clarke, L. M. Weiner, Y. Gusev, and S. Madhavan, "G-DOC Plus–an integrative bioinformatics platform for precision medicine," *BMC Bioinf.*, vol. 17, no. 1, p. 193, Apr. 2016. doi: 10.1186/s12859-016-1010-0.

[169] Z. Cao, C.-T. Lin, W. Ding, M.-H. Chen, C.-T. Li, and T.-P. Su, "Identifying ketamine responses in treatment-resistant depression using a wearable forehead EEG," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1668–1679, Jun. 2018. doi: 10.1109/tbme.2018.2877651.

[170] Z. Cao, K.-L. Lai, C.-T. Lin, C.-H. Chuang, C.-C. Chou, and S.-J. Wang, "Exploring resting-state EEG complexity before migraine attacks," *Cephalalgia*, vol. 38, no. 7, pp. 1296–1306, Sep. 2017. doi: 10.1177/0333102417733953.

[171] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, "Multi-channel EEG recordings during a sustained-attention driving task," *Sci. Data*, vol. 6, no. 1, Apr. 2019, Art. no. 201546. doi: 10.1038/s41597-019-0027-4.
[172] V. Subbaraju, S. Sundaram, and S. Narasimhan, "Identification of lateralized compensatory neural activities within the social brain due to autism spectrum disorder in adolescent males," *Eur.J. Neurosci.*, vol. 47, no. 6, pp. 631–642, Aug. 2017. doi: 10.1111/ejn.13634.
[173] R. Abreu, A. Leal, and P. Figueiredo, "EEG-informed fMRI: A review of data analysis methods," *Frontiers Hum. Neurosci.*, vol. 12, p. 28, Feb. 2018. doi: 10.3389/fnhum.2018.00029.
[174] M.-P. Hosseini, T. X. Tran, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh, "Deep learning with edge computing for localization of epileptogenicity using multimodal rs-fMRI and EEG big data," in *Proc. IEEE Int. Conf. Autonomic Comput. (ICAC)*, Jul. 2017, pp. 83–92. doi: 10.1109/icac.2017.41.
[175] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, "Big data challenge: A data management perspective," *Frontiers Comput. Sci.*, vol. 7, no. 2, pp. 157–164, Apr. 2013. doi: 10.1007/s11704-013-3903-7.
[176] C. H. Lee and H.-J. Yoon, "Medical big data: Promise and challenges," *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, Mar. 2017. doi: 10.23876/j.krcp.2017.36.1.3.
[177] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *BioMed Res. Int.*, vol. 2015, pp. 1–16, 2015. doi: 10.1155/2015/370194.
[178] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control," *PLoS Comput. Biol.*, vol. 7, no. 10, Oct. 2011, Art. no. e1002199. doi: 10.1371/journal.pcbi.1002199.
[179] K. Shameer, M. A. Badgeley, R. Miotto, B. S. Glicksberg, J. W. Morgan, and J. T. Dudley, "Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams," *Briefings Bioinf.*, vol. 18, no. 1, pp. 105–124, Feb. 2016. doi: 10.1093/bib/bbv118.
[180] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, p. 1351, Apr. 2013. doi: 10.1001/jama.2013.393.
[181] Y. L. Lin, A.-M. Guerguerian, J. Tomasi, P. Laussen, and P. Trbovich, "Usability of data integration and visualization software for multidisciplinary pediatric intensive care: A human factors approach to assessing technology," *BMC Med. Inform. Decis. Making*, vol. 17, no. 1, p. 122, Aug. 2017. doi: 10.1186/s12911-017-0520-7.
[182] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, May 2012. doi: 10.1038/nrg3208.
[183] A. Rajkomar, E. Oren, and J. Dean, "Scalable and accurate deep learning with electronic health records," *Digit. Med.*, vol. 1, no. 1, May 2018, Art. no. 18. doi: 10.1038/s41746-018-0029-1.
[184] Z. Cao, C.-T. Lin, K.-L. Lai, L.-W. Ko, J.-T. King, K.-K. Liao, J.-L. Fuh, and S.-J. Wang, "Extraction of SSVEPs-based inherent fuzzy entropy using a wearable headband EEG in migraine patients," *IEEE Trans. Fuzzy Syst.*, to be published. doi: 10.1109/tfuzz.2019.2905823.
[185] L. Yu, Y. Lu, and X. Zhu, "Smart hospital based on Internet of Things," *J. Netw.*, vol. 7, no. 10, p. 1654, Oct. 2012. doi: 10.4304/jnw.7.10.1654-1661.
[186] M. Haghi, K. Thurow, and R. Stoll, "Wearable devices in medical Internet of Things: Scientific research and commercially available devices," *Healthcare Inform. Res.*, vol. 23, no. 1, p. 4, 2017. doi: 10.4258/hir.2017.23.1.4.
[187] I. K. Al-Azwani and H. A. Aziz, "Integration of wearable technologies into Patient's electronic medical records," *Qual. Primary Care*, pp. 151–155, Aug. 2016. [Online]. Available: http://hdl.handle.net/10576/4713
[188] C. Bousquet, B. Dahamna, S. Guillemin-Lanne, S. J. Darmoni, C. Faviez, C. Huot, S. Katsahian, V. Leroux, S. Pereira, C. Richard, and S. Schück, J. Souvignet, A. L.-L. Loüt, and N. Texier, "The adverse drug reactions from patient reports in social media project: Five major challenges to overcome to operationalize analysis and efficiently support pharmacovigilance process," *JMIR Res. Protocols*, vol. 6, no. 9, p. e179, Sep. 2017. doi: 10.2196/resprot.6463.
[189] *Big Data Technologies in Healthcare: New Whitepaper Released by BDVA | BDVA*. Accessed: Jun. 1, 2019.[Online]. Available: http://bdva.eu/?q=node/629

HOUSSEIN DHAYNE received the bachelor's degree in fundamental and applied mathematics from Lebanese University, Lebanon, in 2003, and the M.Sc. degree in cooperation in information processing sciences from Lebanese University, Lebanon, and Paul Sabatier University, France, in 2005. He is currently pursuing the Ph.D. degree in computer engineering with Saint Joseph University, Lebanon. His research interests include big data integration, healthcare data, and the semantic web.

RAFIQUL HAQUE received the Ph.D. degree in computer science and information system from the University of Limerick, Ireland. He is the Lead Scientist and Chief Technology Officer of Cognitus—a research, innovation, and development startup in Paris. He is a Guest Lecturer with Telecom Sud Paris and an Invited Professor with Lebanese University. He has worked in various areas of computer science for almost 10 years in different academia and research centers in different countries, including Germany, The Netherlands, Ireland, and France. He has published several articles in conferences and journals sponsored by the IEEE, ACM, and Springer. His research interests include data analytics, distributed database, scalable and distributed computing, service-oriented computing (SOC), and cloud computing.

RIMA KILANY received the Ph.D. degree in computer and communication from the Ecole Nationale Supérieure de Télécommunications, Paris, France. She is currently an Associate Professor with the Ecole Supérieure des Ingénieurs de Beyrouth, Saint-Joseph University, Lebanon. Her current research interests include enterprise application integration, data integration, semantic web technologies, cloud, and digital business transformation.

YEHIA TAHER received the Ph.D. degree in computer science from the Université Claude Bernard Lyon 1. He is currently an Associate Professor of computer science with the Université de Versailles Saint-Quentin-en-Yvelines, France. His research interests include service-oriented computing, complex event processing, cloud computing, and business process management. In recent years, his research is largely dedicated to real-time big data analytics. He worked in different research institutes within Europe. He works in collaboration with academia and industries. He collaborates with academia in France and other European countries, including The Netherlands and Ireland. In addition, he has established collaboration with universities in middle-east. He has published around 50 conference papers and journals in top tier conferences and journals.