IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.8, Issue.8, August 2020 E-ISSN: 2347-2693

Data Integration Techniques For Healthcare – A Comprehensive Survey

R. Thirumahal[1*], G. Sudha Sadasivam[2]

[1,2]Dept. of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamilnadu, India

[*]Corresponding Author trk1193@gmail.com, Tel.: +91-98209-38174

DOI: https://doi.org/10.26438/ijcse/v8i8.2329 | Available online at: www.ijcseonline.org

Received: 03/Aug/2020, Accepted: 21/Aug/2020, Published: 31/Aug/2020

***Abstract*—** Data is the most valuable asset. As a strategy, integration is the first step towards transforming data into meaningful and valuable information. Data integration provides the ability to manipulate data transparently across multiple data sources. Healthcare sector in particular has been hindered by the diversity of the biomedical data. A framework to unify the sources of such diverse data can facilitate diagnosis and plan for treatment. According to Experian, 66% of companies lack a centralised approach to data resulting in data silos. The data integration market is expected to grow annually at the rate of 12.5% since 2018. This paper discusses the need for data integration, the challenges in implementing a data integration framework, various approaches for data integration, their strength and weakness. The research directions which act as additional add-on or improvements to the existing system have been discussed**.**

***Keywords*—** Data integration, Big Data, Data Integration Methods

I. INTRODUCTION

Data deluge in the current era has resulted in huge volume of readable, downloadable and usable data. Internet acts as a huge repository drowning the users with the required information. Growth of the Internet and its mass accessibility have been a huge advantage in collection and dispersion of information. Since most of this information is free, it is also being used in many researches. Data is the most valuable asset, utilizing it correctly can allow one to make intelligent business decisions, drive growth and improve profitability. Most of the organizations and institutions find the process of unifying the data that they hold in different data silos complex and decide to look past any analytic activity on that data. Difficulty arises when the user decides to operate on multiple data sources rather than a single source. If mutually exclusive information is available in these data sources then they can be combined and worked on efficiently. But if they contain common information, then problems arise during the process of information retrieval. In order to perform effective fusion of the heterogeneous data sources, it becomes mandatory to choose a language that is common to all the data sources. There is no universal language available to retrieve the information. This necessitates design a common query language while designing a data integration system. The efficiency of the query language and the efficiency of the processing components determine the overall efficiency of the system.

Having a unified framework to access data from different sources is said to reduce manual fetch and organization time by hours. As a strategy, integration is the first step towards transforming data into meaningful and valuable information. The data from the different sources need to be integrated so that a unified access to all the collective data is available to the end-applications. This paper describes the challenges faced in data integration and in a broad sense describes the methods that are used to integrate data. Benefits of data integration include data integrity and data quality, availability of the data, streamline business operations, increase in productivity, and improvement in decision making.

The rest of the paper is organized as follows. Following the introduction in section I, the data integration challenges are presented in section II. The data integration methods are discussed in section III. Section IV presents the comprehensive survey followed by research directions in section V. The conclusion is presented in section VI.

II. RELATED WORK

In Public Health ontology design, patient’s health and medical status are in the form of electronic health record is modelled using UML class diagram [1]. Next the public health ontology design was built in RDFS/OWL and linked into related ontology wherever relevant. Then the defined properties are enhanced with OWL constructs to specify facets such as cardinality, relational characteristics such as symmetry, transitivity, inverse properties and functional values and the instances were populated in the ontology knowledge base. Finally, the contents of the ontology are analyzed through a set of metrics based on the OntoQA measures [2]. The schema metrics that address the design of the ontology schema and instance metrics that address the way instances are organized within the ontology are verified. The schema is analyzed using metrics such as relationship diversity, schema depth and attribute richness. The instances are evaluated using measures such as class specific metrics and relationship specific metrics. Java based scripts are used to collect the required statistics from the knowledge base.

The ontology was rich in detail and its knowledge base have high cohesion internally as its entities are strongly related. The semantic representation enables an interoperable, structured patient health record which is machine processable, easily accessible across the Internet. This, in turn, will allow queries that yield richer and more relevant search results. But, due to the limited connectivity to external ontologies the knowledge base is fairly isolated. A universal patient identifier is crucial and it must be ensured that all health care providers use the same identifier while recording health records. Only then data integration of all health information for a patient can be guaranteed.

A descriptive study on open source electronic medical record integration model for clinical data exchange between health care facilities was carried out [3]. Respondents were doctors, nurses, pharmacists, laboratory staffs, and person in charge of hospital information system as informant for content analysis. A web-based service portal was created to implement clinical data integration that can be accessed by clinician registered within the Ministry of Health. The patient’s clinical history was stored in the hospital database with a unique Open Identity Medical Record (OpenIDRM) code. This OpenIDRM was stored on the Health Service Server to integrate it. OpenIDRM contains all the record number of patient’s. One patient may have several different medical record numbers in several hospitals. In conclusion, clinician can access the patient’s clinical history by opening a web portal system through a unique OpenIDRM code. The major advantages of the system are, all the different data sources are integrated into a single megastore of data and execution of queries on the megastore is faster compared to a data federation model. The major issue faced by the system is the latency taken to design the model and integrate all the sources. It is difficult to create a database model that can accommodate all kinds of data.

An ontology-oriented architecture for Telemedicine Systems was created in which the core ontology was defined as a knowledge base [4]. Ontology mapping of the different sources were done in two phases. First one was data collection phase and the second one was running phase. In data collection phase the data is collected from various sources using, specialized domain ontologies, rules extraction, natural language processing techniques for unstructured information, data mining techniques for structured data sources. All the collected data were integrated using ontology mapping. Rules output from the data collection phase are specified using the concepts of the core ontology. In running phase, the telemedicine system accepts the information collected from the sensors and the integrated data mining rules. It sends an alarm to the respective medical team if the system detects any abnormal measurement. The physician analyses the alarm based on the integrated Web rules. Finally, either the alarm is rejected because it is an exception, or the alarm is accepted. The telemedicine system permits communication between different data sources each with its own ontology, personalized treatment of each patient and the improvement of traditional AI systems. The system can be enhanced using social networks and NLP techniques.

First the local medical ontology was created by extracting the abstract relationship from heterogeneous data sources. Similarity Detection Algorithm based on Medical Ontology (SDAMO) was applied on the local medical ontology in order to form the global medical ontology which carry out the ontology fusion by calculating the similarity between concepts, and guide ETL process [5]. By introducing the medical similarity algorithm into the hybrid medical ontology, not only the multi-semantic problem was solved, but also the heterogeneity of mass medical data was eliminated. The data integration scheme with SDAMO algorithm improves the efficiency of the medical heterogeneous data integration. This scheme helped the medical workers and patients, and promotes the development of the medical informatization. The medical domain ontology UMLS and integrated Chinese language system was not enough for ontology construction process. There is a need to construct learning tools for medical ontology and training model using ontology data. Constructed ontology can be mapped automatically by machine learning.

Indian healthcare does not currently use technology extensively while implementing new approaches may be most significant in the rural areas of India where healthcare infrastructure and resources are not as readily available [6]. The proposed model can be applied in the India healthcare system to support Emergency response services, Healthcare monitoring, and Understanding healthcare needs. Some of the barriers and challenges to healthcare delivery in rural areas can be alleviated by mHealth and big-data analytics. mHealth can motivate health initiatives and policy makers. In conjunction with a model for mHealth, leveraged outcomes such as personalised prescriptions, access to trained medical professionals and enabled hospitals can hold all that mHealth promises for every global citizen. A mHealth and big-data integration model was proposed that utilises generated data from mHealth devices for big-data analysis that could result in providing insights into the India population health status [7]. The major advantages of the system include the following. Biomedical, behavioural and lifestyle data from individuals may enable customised and improved healthcare services to be delivered. The analysis of data from mHealth devices can reveal new knowledge to effectively and efficiently support national healthcare demands.

Comparison of Data integration techniques in Healthcare Domain

Table 1 provides a comparative survey on the strengths and weaknesses of Healthcare Data Integration solutions.

Table 1 : Summary of data integration applications in Healthcare context

Title	**Method**	**Strength**	**Weakness**
A semantic big data platform for integrating heterogeneous wearable data in healthcare.	Ontology creation using SPARQL endpoint for wearable data which is stored in distributed clusters.	This ontology provides the aggregation of distributed heterogeneous data.	SPARQL endpoint could become a bottleneck.
Knowledge and theme discovery across very large biological data sets using distributed queries [8]	Hybrid solutions with distributed queries and batch processing	Investigates how various data structures and data models are best mapped to the proper computational framework.	Extraction method may have an issue with large amount of data
Data management for next generation genomic computing [9]	Data federation method is used for integration	Geno Metric Query Language (GMQL) - a new-generation query language inspired by relational algebra and extended with orthogonal, domain-specific abstractions for genomics.	Combining the results from multiple data sources may be an issue
A health analytics semantic ETL service for obesity surveillance.	Semantic ETL service aims at semantically integrating big data for use by analytic mechanisms	Assist in identifying patterns and contributing factors for this social phenomenon and, hence, drive health policy changes.	Difficult to handle the volume issue by semantic ETL service
Big Data Technologies [10]	Star schema to store the multidimentional data.	Integration of huge heterogeneous data such as environmental and medical	Data warehouse method has some limitation with bigdata
Using ontologies to improve semantic interoperability in health data [11]	Extended Ontology Toolkit for Chronic Disease Management	Ontologies have proved to be more dynamic than other methods.	Building large ontologies can be time consuming and can require considerable amount of input from domain experts.
Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources [12]	Hybrid machine learning approach to construct side effect classifiers	Different types of drug side effects are considered to achieve better predictive performance.	Need to examine the model interpretability.
A Large-Scale Clinical Validation of an Integrated Monitoring System in the Emergency Department. [13]	Integrated, automated system that interfaces to a peer-to-peer network of medical devices.	Improve patient outcomes in the busy environment of a major emergency department and other high-dependence areas of patient care.	Requires some sort of supervised training in order to perform useful analyses.
Prostate Cancer Information System (PCIS) [14]	Data Federation supported by an ontology	Well-structured entity –to-entity mapping with clearly defined relationships	The time of query execution is directly depended on the individual speeds of the different data sources
The ACGT Master Ontology and its applications – Towards an ontology-driven cancer research and management system [15]	Ontology-based data integration	A stable conceptual interface to the database systems because the ontology provides a rich and predefined vocabulary	Domain experts need to construct, merge, and maintain the domain ontologies. Each dataset needs to be registered to the ontology

III. DATA INTEGRATION CHALLENGES

The task of integrating heterogeneous data sources faces a number of challenges as discussed in this section.

II.I Heterogeneity in Data Sources

One of the key challenges is dealing with the problems emerging from the heterogeneity of data sources. Dong and Naumann introduce several challenges of data integration [16]. Data sources can be heterogeneous in syntax, schema, or semantics, thus making data interoperation a difficult task. Syntactic heterogeneity is caused by the use of different models or languages. Schematic heterogeneity results from structural differences. Semantic heterogeneity is caused by different meanings or interpretations of data in various contexts. It is also difficult to model storage and organization of data by data source and sink. Logically mapping the heterogeneous data between the source and target systems needs a carefully prepared ontology that defines even the most granular of entities and relationships so that data is not misinterpreted while mapping. Before the data in the source is moved to the target, it needs to be transformed to fit into data organization structures of the target system. This transformation will be time intensive and there is no guarantee that all the source data will fit target schema.

II.II Heterogeneity in Data Sources

Business analysis, domain expertise, and technical knowledge about source systems are required while mapping data from one format to another. While differences in naming conventions and data formats is a good first step, this task also requires understanding the relationships of one data set to another. The business rules embedded in the source system that produce the data must be considered when applying transformation logic to create the integrated data set.

II.III Integrity of the Data

Data Integration is an essential factor to be considered while integrating different data sources. Data Integration syncs huge quantities of heterogeneous data generated from different sources. Integrity of data has to be preserved when it is converted from legacy format to another format. Also, the requirement in the number of tables on the source and target sides can be mismatched. The structure of columns and data types might differ as well. Data Integrity is compromised when the data is altered in transit between the source and target. Transformation and translation of data needs to be relevant to the business context and should add value to the existing source data. Any discrepancies while accommodating such data integration provisions will lead to errors in the integration system. The tool or framework for integration needs to have facilities to correct these errors and get the system back online.

II.IV Performance

Performance is a major concern for any data integration system. Richness of the data and the time taken to process are to be optimally balanced to design a good data integration system. The processing time and the response time of the source and target system are to be minimal. Database structure, variety and quantity of the data also affect the performance. A poorly performing integration system can suffer from issues like slow data processing and latency in synchronising source and target in real time.

II.V Data Security

Data security is one of the top priorities in a data integration solution. Organizations want to ensure that all data are stored securely and confidentially. Organizations face severe consequences when data integration is performed in an insecure environment. Organizations can suffer from one or many of these consequences such as Loss of Revenue, Data Breach, Data Leakage, Loss of Trade Licenses of Organizations, Government Penalties and Lawsuits, Loss of Organizational Reputation.

To overcome these challenges, a suitable approach to data integration is essential. Section III discusses these approaches.

IV. DATA INTEGRATION METHODS

A data warehouse is a centralised collection of electronically stored data. Data are extracted, transformed, and loaded (ETL) into a central repository from the different sources. The data warehouse also has provisions to retrieve and analyze data, and to manage it. While having all the data under a single roof sounds optimal for query processing, the cost of building and updating the warehouse with data from the different sources is not a great trade-off. This section details on various techniques suitable for data integration

III.I Data Federation

A federated database system is a single conceptual view of the integrated database. It integrates multiple autonomous database systems using a meta-database management system. The data sources are called ‘‘federated” because the data is not copied into a central repository, rather, the federation server maintains indices to the data of interest in the source systems. Geographically decentralized data sources are connected via a computer network. Performance bottleneck of this approach includes network performance, schema design, and the availability of the source database systems. In the healthcare domain, data federation technique was used in many research projects such as e-Health Service [17], BioFed [18], and Genomic Computing [19].

III.II Ontology Based Integration

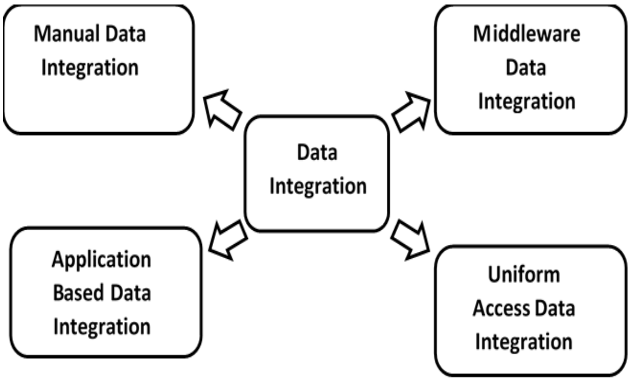
Manual Data Integration
Middleware Data Integration
Data Integration
Application Based Data Integration
Uniform Access Data Integration

Figure. 1. Data Integration Techniques

Figure 1 shows the data integration techniques which can be categorized as manual integration, middleware integration, application-based integration and uniform access integration. In manual data integration all phases of integration are done under the supervision of data manager without automation whereas in middleware data integration the software connects applications and transfers data between them and databases. In application-based integration software applications are used to locate, retrieve, clean, and integrate data from disparate sources that are compatible with one another. Uniform access integration technique retrieves and uniformly displays data by allowing the data to stay in its original location. This requires less storage and easier data access and simplified view of data. Recent research in data integration focuses on the semantic integration problem such as semantic conflicts among the heterogeneous data sources. A common strategy addresses the semantic conflicts is through the use of an ontology with explicitly defined schema terms. This approach is called ontology-based data integration.

Ontology is a data model that represents a set of concepts within a domain and the relationships among these concepts. In short, ontology is a specification of a conceptualization [20]. A conceptualization is an abstract, simplified view of the world. The specification is a declarative representation of the conceptualization in a concrete form. It appropriately encodes the concepts and relationships in a computer understandable language by interpreting the knowledge.W3C recommends the Web Ontology Language (OWL) to represent the web ontologies.

Ontology-based data integration efficiently uses the ontology to combine the data from the multiple heterogeneous sources. A semantic layer is provided on the top of the underlying data. The main goal of the ontology is to provide a set of techniques to solve the semantic heterogeneity problems. Ontology-based data integration has unique advantages. The ontology provides a rich and predefined vocabulary. It has a stable conceptual interface to the database systems. It requires a mediator system to represent all objects in the domain of the interest. Queries are then fired against the mediator, which in turn analyze the query and split it in to smaller queries. The queries are then directed to the source systems. This model of data integration is classified as a Local-as-View as displayed in Figure 2, to denote that queries on the local databases are reformulated in terms of the global mediation [21]. Local-as-View model scales better and is easier to maintain. It provides formal definitions of the terms used in the data sources, and renders the implicit meaning of the relationships among the different terminologies of the data sources explicitly. For example, one can determine whether two classes of the data items from two different database systems are equivalent, or whether one is a subset of another by using ontology. Ontology also allows the users to query different database systems as one by tying them together at asemantic level. Many projects already used ontology to enhance data integration such as integrating heterogeneous wearable data in healthcare [22], support of digital cancer patient [23], obesity surveillance [24] and in the interoperability of electronic health records [25].

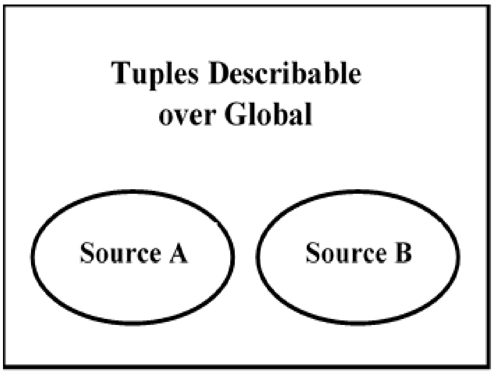

Figure. 2. Local AS View

III.III Peer-to-peer Data Management System(PDMS)

PDMS implements a decentralized view of data integration. In this method there is no centralized mediator to collaborate the data sources. Each data source can play different role at different time. A data source can play a role as mediator, a data server and a client for other data source. So, each participant to the system is a peer. Data from the different peers are related using suitable mapping. The PDMS architecture shown in Figure 3 does not follow the global schema pattern, where each participating data source heeds to follow a set of terms which is defined in advance. New data sources can join or leave just by including or removing the mapping between them. A query to a PDMS is posed using the peer schema of one of the peers. If a query is asked to a particular peer, it is reformulated using the peer mappings into a set of queries that may refer to other peer relations. Here, the information available in the entire P2P system is used to answer the query.

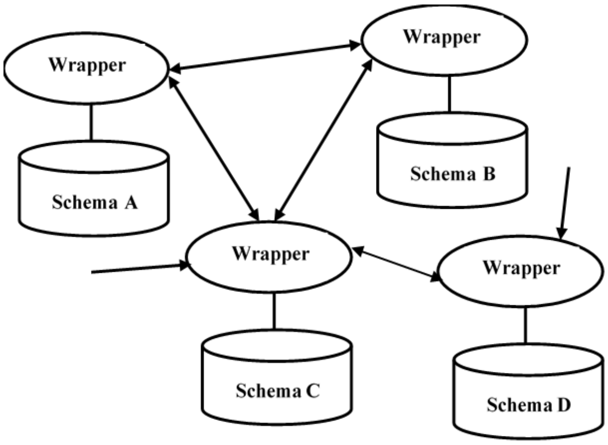

Figure 3. Peer-to-peer Data Management System

Healthcare data is produced from a variety of resources and this data is huge and the growth of the data is exponential. This data needs to be shared among different health care system. With the methods available such as data warehousing, data federation and peer to peer data management it is difficult to maintain and reuse semantically complex data extraction and transformation. In data warehousing, the data collected from various sources are transformed into specific format and then it is loaded into the data warehouse. This procedure consumes more time. The federation server maintains indices to the data sources which are connected via a computer network in data federation. PDMS does not follow the global schema pattern. So, ontology-based integration is most suitable for healthcare domain because here the global schema is created from the available data sources. User can query the global schema which will retrieve the data from the respective data sources. Ontologies can be used to organise and describe the medical concept of both the source and target system. Different health information system can share these domain specific ontologies.

V. RESEARCH DIRECTIONS

Based on the study further research can be done in the area of data integration. Some are listed below:

1. Since SQL based syntax is followed by most of the databases, the wrapper used to create the mediated schema can be integrated to get the data from different sources. If the data sources consist of dynamic data then wrapper can be generated dynamically according to the data source being used.

2. Query optimization will improve the efficiency and gives fast results. So, the given query is split into multiple subqueries according to the required data sources. Then the queries can be optimized in the native optimizer and the results are sent back to the mediator schema that will combine the results of multiple data sources and present to users. While querying, duplicate data can be eliminated and semantically correlated data can be analysed and stored appropriately.

3. If the conflicts are dealt in the beginning itself then more accurate results can be provided to the users. Performance of the system can be improved because it does not involve the user's query time. Conflicts identification is one area with a large scope for research [26][27]. Identifying conflicting attributes when a query is presented is the basis for working with multiple databases [28]. Though this process seems simple, it has a semantic dimension to it, which increases its complexity manifold. In order to identify such attributes, machine learning methods can be used [29][30]. To identify the most important data from the conflicting attributes, input from the user can be considered using the analytic hierarchy process.

4. To address the challenges on mHealth and big-data analytics technologies, a framework can be created for integrating the mHealth big-data and applying the results in India's healthcare. concentrate on the healthcare delivery problems faced by rural and low-income communities in India to illustrate more general aspects and identify key issues. India healthcare system and challenges faced particular issues, including inadequate healthcare resources, insufficient funding, poor healthcare infrastructure and rural–urban disparity. Data integration frame work with NoSQL technology could resolve integration, transformation, inconsistencies, noise challenges in big data.

VI. CONCLUSION

It is evident that data integration as a field of study is evolving with time. A wide variety of data are available in the healthcare sector, including text, images, video, sensors and social media data are heavily used within the healthcare sector for various purposes. This paper presents a detailed study of the data integration system by describing the challenges and integration methods. It also discusses the strength and weakness of different methods. The research directions which act as additional add-on or improvements to the existing system have been discussed.

REFERENCES

[1] Rohini R. Rao, Krishnamoorthi Makkithaya and Neha Gupta, " Ontology based semantic representation for Public Health data integration", International Conference on Contemporary Computing and Informatics (IC3I), **4799-6629**, IEEE **2014**.

[2] Samir Tartir, I. Budak Arpinar and Amit P. Sheth , "Ontology evaluation and validation", Theory and Applications of Ontology: Computer Applications, pp. **115-130**, **2010**.

[3] ArifKurniadi and RetnoAstuti, "Patient Clinical Data Integration In Integrated Electronic Medical Record System for Health Care Facilities in Indonesia," Jurnal Kesehatan Masyarakat, Vol. **13-2,** No. 11/**2017**, pp. **239-246**.

[4] Jesus Peral, Antonio Ferrandez, David Gil, Rafael Munoz-Terol, and Higinio Mora, "An Ontology-Oriented Architecture for dealing with Heterogeneous Data applied to Telemedicine Systems", IEEE Access PP(99):**1-1**, **July 2018**.

[5] Shuxia Ren, Xu Lu and Teng Wang, "Application of Ontology in Medical Heterogeneous Data Integration", 3rd International Conference on Big Data Analysis, 978-1-5386-**4794**, IEEE, **2018**

[6] IBM Corporation. The India cure: Remedying the challenges of the healthcare landscape. Somers, NY: IBM Institute for Business Value, **2017**

[7] SamanehMadanian, Dave T Parry, David Airehrour and Marianne Cherrington, "mHealth and big-data integration: promises for healthcare system in India", BMJ Health Care Inform: 10.1136/bmjhci-2019-100071, september **2019**.

[8] U. S. Mudunuri et al., "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. **2**, **2013**.

[9] S. Ceri, A. Kaitoua, M. Masseroli, P. Pinoli, and F. Venco, ``Data management for next generation genomic computing," In the Proceedings of the 19th International Conference on Extending Database, pp. **485-490**,**2016**.

[10] R. Bellazzi, A. Dagliati, L. Sacchi, and D. Segagni, ``Big data technologies," Journal of Diabetes Science and Technology, Vol. 9, No. 5, pp. **1119_1125**, **2015**.

[11] Harshana Liyanage, Paul Krause and Simon de Lusignan, "Using ontologies to improve semantic interoperability in health data," Journal of Innovation in Health Informatics, Vol. **22**, No. **2**, **2015**.

[12] Wei-Po Lee, Jhih-Yuan Huang, Hsuan-Hao Chang,King-Teh Lee, And Chao-Ti Lai, "Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources," IEEE Access. 2169-3536, Vol. **5**, **2017**.

[13] David A et al., "A Large-Scale Clinical Validation of an Integrated Monitoring System in the Emergency Department," IEEE Journal of Biomedical and Health Informatics, Vol. **17**, No. **4**, **2013**.

[14] Hua Min, Frank J. Manion, Elizabeth Goralczyk, Yu-Ning Wong, Eric Ross and Robert Beck J, "Integration of Prostate Cancer Clinical Data using an Ontology," Journal of Biomedical Informatics, Vol. 42, No. 06, pp.1035-1045, 2009.

[15] Mathias Brochhausen et al., "The ACGT Master Ontology and its applications – Towards an ontology-driven cancer research and management system," Journal of Biomedical Informatics, Vol. 44, No. 10, pp.**8**–**25**, **2010**.

[16] X. L. Dong and F. Naumann, ``Data fusion," VLDB Endowment, Vol. 2, No. 2, pp. **1654-1655**, **2009**.

[17] W. Liu and E. Park, ``Big data as an e-health service,", International Conference on Computing, Networking and Communications (ICNC), pp. **982-988**, **2014**.

[18] A. Hasnain et al., "Biofed: Federated query processing over life sciences linked open data," Journal 0f Biomedical Semantics, Vol. **8**, No. 1, p. **13**, **2017**.

[19] S. Ceri, A. Kaitoua, M. Masseroli, P. Pinoli, and F. Venco, ``Data management for next generation genomic computing," In the Proceedings of 19th International Conference on Extending Database, pp. **485-490**, **2016**.

[20] Buccella A, Cechich A and Brisaboa NR, "An ontology approach to data integration." Journal of Computer Science and Technology, Vol. **3**, No. **2**, pp. **62**–**68**, **2003**.

[21] Levy AY, Rajaraman A and Ordille JJ, "Querying heterogeneous information sources using source descriptions," In the Proceedings of the twenty-second international conference on very large data bases (VLDB'96). Mumbai, India; **1996**

[22] E. Mezghani, E. Exposito, K. Drira, M. D. Silveira, and C. Pruski, "A semantic big data platform for integrating heterogeneous wearable data in healthcare," Journal of Medical Systems, Vol. **39**, No. **12**, p. **185**, **2015**.

[23] H. Kondylakis et al., "iManageCancer: Developing a platform for Empowering patients and strengthening self-management in cancer diseases," In the Proceedings of the IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), pp. **755-760**, **2017**.

[24] M. Poulymenopoulou, D. Papakonstantinou, and F. Malamateniou, "A health analytics semantic ETL service for obesity surveillance," Studies Health Technology and Informatics, Vol. 210, pp. **840-844**, **2015**.

[25] A. Bahga and V. K. Madisetti, "A cloud-based approach for interoperable electronic health records (EHRs)," IEEE Journal of Biomedical and Health Informatics, Vol. 17, No. 5, pp. **894-906**, **2013**.

[26] Carol I and Kumar SBR. "Conflict resolution and duplicate elimination in heterogeneous datasets using unified data retrieval techniques," Indian Journal of Science and Technology, Vol. **8**, No. 22, pp.**1-6 2015**.

[27] Khazalah F, Malik Z and Rezgui A, "Automated conflict resolution in collaborative data sharing systems using community feedbacks," Information Sciences, Vol. **298**, pp. **407-424**, **2015**.

[28] Weiguo F, Lu H, Madnick SE and Cheung D, "Discovering and reconciling value conflicts for numerical data integration," Information Systems, Vol. **26**, No. 8, pp. **635-656**, **2001**.

[29] Ramkumar T, Hariharan S and Selvamuthukumaran S, "A survey on mining multiple data sources," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. **3**, No. **1**, pp. **1-11**, **2013**.

[30] Ramkumar T, Srinivasan R and Hariharan S, "Synthesizing global association rules from different data sources based on desired interestingness metrics," The International Journal of Information Technology and Decision Making, Vol. **13**, No. **3**, pp. **473-495**, **2014**.

AUTHORS PROFILE

R. Thirumahal is currently an Assistant Professor (Selection Grade) in the Department of Computer Science and Engineering, PSG College of Technology. She has a total of 22 years of teaching experience. Initially, she worked at TULEC, a division of Tata Infotech Ltd, Mumbai for 3 years and then at Thadomal Shahani Engineering College, Mumbai for 16 years. She has also published one book and thirteen papers in international journals. Thirumahal has completed her BE degree in Computer Science and Engineering from Bharathidasan University, Tamil Nadu, and her ME degree in Computer Engineering from Mumbai University. Ms Thirumahal is currently pursuing PhD in Information and Communication Technology under Anna University, Chennai. Her areas of expertise include data integration, big data analytics, and machine learning.

G. Sudha Sadasivam is Professor and Head of the Department of Computer Science and Engineering at PSG College of Technology, Tamil Nadu. She has 22 years of teaching experience. She has published 5 books and 20 papers in indexed journals. She has coordinated two AICTE-RPS projects and a UGC-sponsored project in the areas of distributed computing. Dr Sadasivam is the coordinator of PSG-Yahoo Research on Grid and Cloud Computing, Nokia Research on Personalisation, Xurmo Research on Social Networking, and Cloudera project on VM Migration in Federated Cloud Environment. Her team has set up PSG-Yahoo/Nokia lab on big data analytics at PSG College of Technology. Her areas of interest include distributed computing and big data analytics.