International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 5, October 2020, pp. 4853~4862
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i5.pp4853-4862

Framework for efficient transformation for complex medical data for improving analytical capability

Chandru A. S.[1], Seetharam K.[2]
[1]Department of Computer Science and Engineering, Visvesvaraya Technological University (VTU), India
[2]Department of Computer Science and Engineering, Jnanavikas Institute of Technology, India

Article Info

Article history:

Received Aug 14, 2019
Revised Mar 3, 2020
Accepted Mar 18, 2020

Keywords:

Data transformation
Healthcare analytics
Medical
Structurization
Text mining

ABSTRACT

The adoption of various technological advancement has been already adopted in the area of healthcare sector. This adoption facilitates involuntary generation of medical data that can be autonomously programmed to be forwarded to a destined hub in the form of cloud storage units. However, owing to such technologies there is massive formation of complex medical data that significantly acts as an overhead towards performing analytical operation as well as unwanted storage utilization. Therefore, the proposed system implements a novel transformation technique that is capable of using a template based stucture over cloud for generating structured data from highly unstructured data in a non-conventional manner. The contribution of the propsoed methodology is that it offers faster processing and storage optimization. The study outcome also proves this fact to show propsoed scheme excels better in performance in contrast to existing data transformation scheme.

Corresponding Author:

Chandru A. S.,
Recearch Scholar, Department Computer Science and Engineering,
Visvesvaraya Technological University (VTU),
Belagavi, India.
Email: chandru.as.research@gmail.com

1. INTRODUCTION

With the increasing adoption of the mobile network and ubiquitous computing, there are also tremendous increase in application as well as services that facilitate the user with extensive data availability [1, 2]. Cloud computing makes this process easier by their different types of standard services [3]. The beneficial factor of this technologies is that there is a greater deal of synchronicity among different information capturing devices, protocols used for processing the information, and network devices to route the data to defined servers [4, 5]. However, there are also extremely challenging situation for handling such massive forms of evolving data. The proposed study considers a case study of healthcare sector which autonomously gather data from different forms of analog, digital, and hybrid devices and fetches this data to the hub where it is further stored in a physical unit or forwarded to data center. A typical smart healthcare unit can be considered to handle multiple forms of data using different types of sensing devices where diversified information about the patient health is collected [6].

However, the challenges in this smart technological advancement are i) each sensor captures a continuous data and the size of the data is ever increasing which cannot be stored in physical storage over the premises of healthcare units, ii) the forms and formats of the data are highly diversified and there could be more inclusion of the such data from the same healthcare units leading to formation of complex medical data [7], iii) these data when thought of aggregated in one place is quite difficult to be stored as they form a condition of highly unstructured data (in such form, conventional knowledge extraction method cannot be applied), iv) existing approaches has larger degree of usage of distributed software framework like Hadoop

and MapReduce; however, they are very much baseline standard and there is always a need to amend the complex architecture of it when the application demands changes. Besides there are many reported loopholes of such framework which cannot be recommended to be used for a critical application e.g. healthcare which involves sensitive clinical data to be used in futuristic analysis [8].

Existing mining approaches are applied on complex data but with highly defined and controlled environment and their cost effectiveness in the field of medical science is not yet been proven [9]. It is essential to ensure that a complex form of medical data be identified of its unique data structure so that there is a good possibility to perform structurization of the complex medical data. Once the complex medical data is formulated in efficient data structure than it opens up many ways to perform analysis over the structured data. However, it is not easy way as such data are always in the form of stream and there are various associated problems [10]. Another challenging scenario of complex medical data will be to perform storage optimization too over cloud. Not all the data should be stored over a cloud which also exponentially increases in size. Hence, a better storage optimization can be done if only the mined data are stored in the distributed cloud with a superior form of indexing mechanism. Therefore, the proposed system presents a solution where a unique transformation system is implemented as a solution toward curving unstructured data to the highly structured data without using any of the system that has been implemented till date or using system which has higher resource dependencies. The focus is towards incorporating cost effectiveness towards normalizing the complex medical data.

The paper discusses this solution in terms of methodology, algorithm, and benchmarked outcomes. The organization of this paper is as follows: Section 1 discusses about the existing literatures where different techniques are discussed for detection schemes used in power transmission lines followed by discussion of research problems and proposed. Section 2 discusses about algorithm implementation followed by discussion of result analysis in Section 3. Finally, the conclusive remarks are provided in Section 4.

The Background. This section discusses about the existing data transformation and processing schemes carried out towards healthcare sector. Study considering the EHR-based data was carried out by Muslim et al. [11] focusing towards reducing the discretion in the existing standards of medical data. Emphasis towards adoption of ETL (Extract/Transform/Load) towards processing the heterogeneous data is also discussed by Diouf et al. [12] where the authors claimed that existing transformation scheme still encounters bigger challenges in from of present state of complex data. Wang et al. [13] have developed a data sharing scheme using conventional ETL process over varied dataset to show that their work is capable of initiating better transformation process. Different work has been carried out towards making the healthcare system smart for precise diagnosing [14-17]. The work of Magarino et al. [18] has presented a prototype-based model for analyzing the action of sleep state. Existing system has also being carried out towards leveraging the diagnosis system using machine learning approach as well as focusing on data privacy as seen in the work of Zhong et al. [19] and Chen et al. [20].

Incorporation of the smart intellectual system using big data approach was carried out by Zhang et al. [21] for assisting effective diagnosis. Study towards EHR data was also carried out by Wu et al. [22] as well as Viceconti et al. [23] where big data approach has been used for monitoring the health-factor towards leading to an effective diagnosis system. Study towards existing approaches of big data analytics towards improving the healthcare system is carried out by Shafquat et al. [24] towards evolving out with cost effective schemes of existing times. A layer-based approach was carried out by Chen et al. [25] where a specific case study of diabetes has been considered with an intention for evolving up with highly customized treatment. The work of Garattini et al. [26] has presented discussion of significator attribute of using big data approach for managing diagnostic analysis over lethal disease. Utilization of Hbase for medical database management is carried out by Chrimes and Zamani [27] where big data analytics has been formulated towards contribution on varied clinical services. Nearly, the similar types of the approach were also discussed by the Tawalbeh et al. [28].

The work of Sarkar et al. [29] has presented a discussion of a novel security model towards medical dataset. The work of Zhang et al. [30] have discussed about the significance of machine learning towards smart clinical services. The work of Srinivasan and Arunasalam [31] has discussed about a unique form of data analytical system towards emphasizing over the offering securing big data in medical data. A closer look into the existing system towards transformation and analytical based approach considering medical data shows that it has adoption of machine learning or big data approach where the idea is to find an ellite solution as a part of the data transformation system. The approaches are claimed to offer better outcome towards addressing the problems stated in their respective literature. They existing works are associated with both advantageous facts as well as limiting factors too.

The work of Bhalla and Bagga [32] introduces a model known as opinion mining for the text categorization. This model has been developed by using RB-bayes model. The RB-Bayes compatation is having accuracy 83.34. The work of Aich et al. [33] introduces a model called Neural-network to categorize

the web-based text. This model is useful for resolving the real-time problem for different types of text mining approaches. The next section outlines the problem that are extracted after reviewing and is considered as the essential reseach problems to be solvedand is addressed in proposed study.

The problems that are yet unsolved in leveraing the analytical processing of medical data are:

a. Existing studies are specific to disease-based approach which narrows down the scope of the existing approaches towards generalized diagnosis over medical data.
b. There is no wide range of consideration to the fact that EHR data are now generated in the form of data stream where it is quite challenging to apply online analytics on the top of it.
c. The granuliaty of the data and its internal structure is something which is significantly missing from the existing system that causes degradation in the accuracy in the analytical process.
d. Apart from frequently used ETL scheme, there is no much novelty towards leveraging data transformation scheme towards facilitating better knowledge extraction process.

Based on the above unsolved problem from existing system, the framing of the problem statement can be as follows "*Developing a computationally cost effective data transformations scheme for facilitating processing the complex medical data for better analytical operation.*"

The Proposed Solution. The prime aim of the proposed system is to evolve up an effective data transformation scheme that is capable of rendering the suitability of the complex medical data in order to optimize the storage as well as leverage the analytical processing of the medical data. The proposed system adopts an analytical research methodology in order to achieve its aim. The pictorial representation of the methodology adopted in proposed solution in shown in Figure 1.

The proposed system takes the input of the unstructured medical data which leads to generation of suitably structured data in a non-conventional manner in order to maintain cost effectiveness of medical data analysis. The proposed system considers Electronic Health Record (EHR) data which consists of both clinical as well as non-clinical information about the patient. Assuming that a healthcare system hosts its storage services over cloud as well as assuming that healthcare system adopts various advanced technologies to captured autonomous essential EHR information of the patient; it can be considered that all these EHR data are forwarded to the datacenter in the form of data stream which is further subjected to storage followed by analysis. However, different from any existing system, the proposed system considers that raw medical data should not be directly stored in cloud storage unit and instead a processing is applied on the top of it to make the data more suitable for storage and analysis. In this regards, the proposed system constructs an analytical environment that aggregates all the unstructured data in a very unique form which is further subjected to next part of processing.

The proposed sstem extracts data chunks (or individual patient data as a cell and index its values and all the core fields (also known as headers) followed by a unique indexing operation. This indexing operation assists in connecting all the respective values of each core fields in a temporary buffer over cloud while all the static header information are stored permanently in the cloud storage units. The benficial factor of this process is to resist the iterative identification of core fields and storing the same which considerably saves the processing effort and time. The proposed system also uses a virtual template based approach in order to facilitate the process of indexing and data structurization using a matrix-based operation.

The proposed system uses a syntactical parsing where semantic-based approach is applied with higher degree of customization over the values of the core fields. The unstructured data is thereby transformed to semi-structured and finally to structured medical data where finally semantics are applied to extract the correct inference of the medical data in the form of knowledge. The next section discusses about the implementation of proposed system with respect to algorithm and its respective execution.

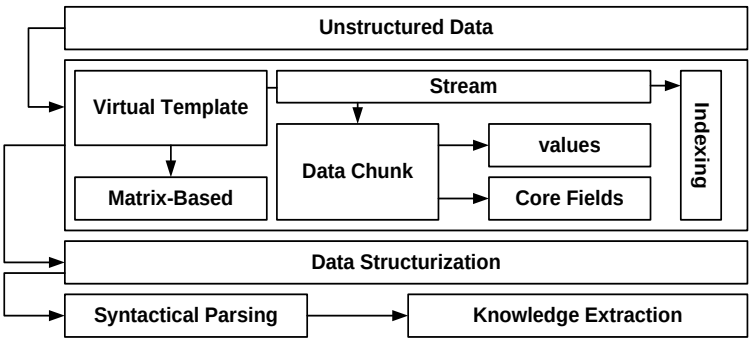


Figure 1. Adopted methodology for proposed transformation and analytical system

2. SYSTEM IMPLEMENTATION

The proposed system introduces a unique form of smart EHR system that is capable of storing the mined information witin the cloud distributed storage system inspited of storing the raw medical data in it. Hence the target is to obtain higher degree of storage optimization and assisting a smooth analytical operation for the EHR data. This section discusses about the assumption and dependencies associated with the design construction followed by implementation strategy and execution flow.

2.1. Assumption and dependencies

The *primary assumption* of the proposed system is that there are various numbers of heathcare units which has an autonomous mechanism to upload the EHR data which are free from any errors or artifacts. It will mean that there are no problems within the EHR data from the local source of origination. The *secondary assumption* of the proposed system is that all these healthcare chains are connected to cloud-based a service which extracts the diversified EHR data from multiple sources and resposit it in distributed order in the cloud clusters. The *tertiary assumption* of the proposed system is that there are diverse forms of EHR data being collected and when the problems aggregated or artifacts are introduced over a cloud resposition process. Therefore, the goal of the proposed system will be to address this problem by introducing a technique that can rectify this data integration problem in cloud in order to reduce the adverse effect of unstructured data. The *core dependencies* of the proposed system is all the EHR data is in the format of text and could be resposited in any format; however, for easiness in computing, the proposed system considers the text is in plain text format.

2.2. Implementation strategy

It should be noted that proposed system is a computational model that targets to optimize the distributed storage system for healthcare unit in order to facilitate smart analytical operations. The study of the proposed system demands a stream of incoming data from the source of the EHR data; however, such mechanism is quite a complex Therefore, the proposed system constructs an analytical data aggregation model as the primary implementation strategy so that near-time traffic flow of textual EHR data could be mapped in proposed system. After the process of data aggregationis analytically design, the next process will be to perform transformation operation from unstructured data to semi-structured data so that analytical operation can be actually carried out on the top of it. The next part of the study implements a context-based mechanism which extracts the mined information on the basis of the correlated data obtained from semantics of the text extracted. For this purpose, the proposed system performs initial classification of the text data into two types viz. i) static data and ii) dynamic user data. For faster processing, the proposed system extracts the static data and reposit over the cloud storage units while for mainintaining non-redundant data over the distributed cloud storage system, the proposed system extracts the dynamic data, index it and then reposit over the temporary buffer. The proposed algorithm is then applied to this temporary buffer in order to facilitate data transformation operation followed by extraction of mining approach. In the entire process, the proposed system offers significant amount of cost effective analytical operation over the medical data. The execution process is as shown in next sub-section.

2.3. Execution flow

The first step of the execution flow is to process the input stream of text data. However, it is priorly discussed that extraction of streams of text data is challenging and needs an access to the buffer of cloud storage points which maintains an adaptive queue system. However, the proposed system attempts to realize this problem by constructing its own buffer of stream data d_s over a sampled period of time t_s, see Figure 2.

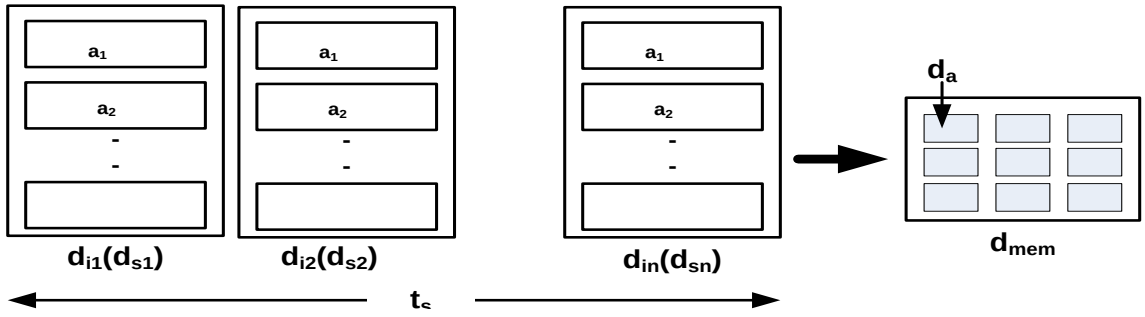


Figure 2. Structure of incoming data

Therefore, the first step of the proposed execution is to construct a synthetic unstructured data over the proposed buffer system. The algorithm takes the input of data stream ds over a definite sampled period of time ts which after processing yields tem_matrix (temporary matris of unstructured data).

```
Algorithm for Synthetic Unstructured Data Construction
Input: ds (data stream)
Output: tem_matrix (temporary matris of unstructured data)
Start
1. For i=1:ts
2.      di→extract(ds)
3.      da→agg(di)
4.      dmem→reposit(da)
5.      For j=1:a
6.            extract dstat→(h, po)
7.            reposit dstat→cloud_template
8.            extract ddyn
9.            reposit ddyn→tem_matrix
10.   End
11. End
End
```

The discussions of steps of the algorithm are as follows: The algorithm considers that there is an individual data d_i for each respective stream of data d_s (Line-1). The system than aggregates all the sampled incoming stream d_s and store it in an aggregated matrix d_a (Line-2). All these data are the cumulatively aggregated followed by repositing da in order to form a sampled buffer d_{mem} (Line-4). Therefore, the buffer dmem can be considered to posses all the aggregated streams of sampled data d_a, which can be now subjected to further processing. The proposed system considers that each individual data di consists of individual patient EHR records *a*, where the variable *a* will number of unique EHR record of individual patient. There are good possibilities that number of *a* in one individual data d_1 could differ in another individual d_2 as well as they could also be same, it all depends upon the density of the text data in the traffic.

Therefore, considering all the values of *a* (Line-5), the algorithm extracts static data d_{stat} with respect to the headers *head* and *pointers* p_o (Line-6). Basically, headers are the prime attribute of the field that represents the complete column of text data with respect to its type while pointer connects all the header with its respective value d_{dyn}. It will mean that proposed system make use of a template based data reposition technique where dstat represents static data over the template while d_{dyn} represents patient EHR data fed by the personel of the healthcare sector. It interprets that d_{dyn} is a direct value to represent individual headers (h_1, h_2, …) where pointers segregates header with individual value of EHR (Line-7 and Line-8). Finally, all the dstat data are resposited over cloud-template storage while user EHR data i.e. d_{dyn} is stored in a temporary matrix *tem_matrix* (Line-9) which has all the unstructured data. The constribution of this algorithm is that it offers a sandboxing mechanism to construct a memory system where the unstructured data can be stored in order to facilitate upcoming processing of data. Figure 3 pictorially represents the processing of the data stream and followed by repositing it.

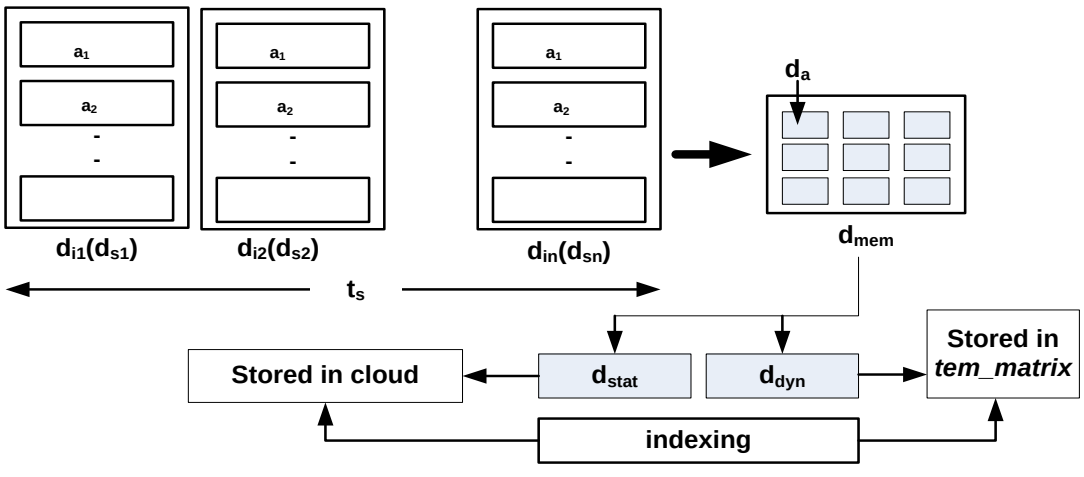


Figure 3. Mechanism of repositing data stream

The next part of the implementation is associated with the extracting knowledge from the data stored in temporary matrix. The algorithm takes the input of *h* (headers) and d_{dyn} (dynamic data) which after processing yields and outcome of d_{sem_struc} (semi-structured data) and $term_{know}$(mined data).

Algorithm for Mined Data Extraction using Semantic
Input: *h* (headers), d_{dyn} (dynamic data)
Output: d_{sem_struc} (semi-structured data), $term_{know}$(mined data)
1. **For** h=1:max
2. t1→add < and > tags to *h* and d_{dyn}
3. reposit in d_{sem_struc}
4. [term loc]→semantic(t1)
5. extract $term_{know}$→term
6. **End**

The discussions of steps of the algorithm are as follows: In order to perfoming mining operation, it is necessary for the algorithm for ensure proper confirmation of the user EHR records. As there are also possibilities of redundant data (which may affect the future part of analytical operation), there is a need to perform a proper indexing operation. For this purpose, the algorithm first accesses the temporary matrix which has all the user-fed EHR records of the data; however, there is no static data in the temporary matrix. Therefore, the algorithm will need to check the indexes of the each headers present in the variable *a* and construct and index. The construction of the index is carried out on the basis of the term used in the individual headers with respect to its individual location.

The proposed system calculates the number of entries for each individual sampled data streams in order to initially confirm the ownership of respect ddyn for the respective individual headers *h*. For an example, consider that there are three headers, in such case, $h_{1(loc1)}$→d_{dyn1}, $h_{2(loc2)}$→d_{dyn2}, and $h_{3(loc3)}$→d_{dyn3} with respect to three different locations loc_1, loc_2, and loc_3 respectively. This index of the respective headers and locations are stored in cloud unit and is respectively indexed with the respective user fed EHR data. The algorithm initially considers all the maximum *max* headers present in data streams, followed by adding start tag < and end tag /> (Line-2). This operation is carried out for header *h* and user fed EHR data d_{dyn}, which results in semi-structure data stored in matrix d_{sem_struc} from unstructured data (Line-3). It should be noted that this operation is carried out wthout using any existing tools of any distributed framework and yet they are quite faster and efficient in their usage.

The next part of the proposed system is about using the semantic operation in order to obtain the corrected meaning and context of the medical EHR data (Line-4). This operation allows the proposed system to construct customized number of semantics on the basis of the problems in diagnosis and hence this operation with minor finetuning in the semantics can result in highly improve scope of automated diagnosis and knowledge extraction process. The proposed system can constructs semantics on the basis of the name of the prominent disease or the highly significant review of the disease made by certain physician. The constructed semantics are applied over the terms t_1 from Line-2 which connects both semi-structured headers and user fed value of EHR record of individual patient with respect to its terms and location (Line-4). Finally, the exracted *term* is considered to be final outcome of the proposed algorithm in the form of knowledge. The Figure 4 introduces the complete process of proposed system.

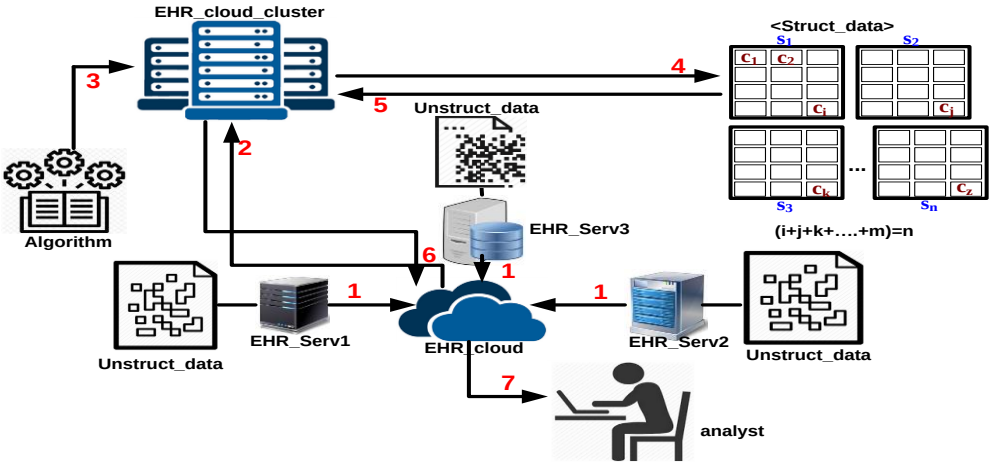


Figure 4. Complete process of proposed system

3. RESULT ANALYSIS

In order to carry out analysis of the proposed system, the challenging step is the primary process of taking a precise input of EHR data. This is because a complex form of EHR data is required for this purpose where the data is larger in dimension as well as there are various forms of heterogeneuity in the data. Therefore, the proposed system reviewed some of the publically available dataset [34-37] in order to visualize the patterns of possible medical data in the big data. Hence, the first process is to ensure that proposed system also consider EHR data which is similar to the pattern of big data. The study considered the EHR data from the existing standard dataset of [36, 37] where each files are of size of megabytes with different accessible formats that are widely supported on any machine. However, such available dataset doesn't have consistencies and hence, there is a need to construct an artificial EHR data and hence, the proposed system constructs a synthetic EHR data which consists of multiple headers with all headers with different values in order to assess the effectiveness of mining approach. The proposed dataset consists of discrete information about the patients as well as it also has various clinical inferences from the physical of the respective patient. The purpose of the analysis will be to obtain analytical information about the disease criticality of the respective patient.

3.1. Assessment environment

The scripting of the proposed system has been carried out in MATLAB considering normal windows environment with core-i3 machine. Adoption of MATLAB offers various benefits of carrying out transformation operation using the matrix-based mechanism in order to assess the proposed system. The complete input data is splitted into smaller versions to generate individual data d_i where the size of one individual data may differ from each other. The idea is also to assess the overall data processing time to see if different file size has any effect over the throughput of analytical operation. The analysis of the proposed study is carried out considering time for data structurization and memory saturation state.

3.2. Results obtained

The assessment of proposed system is carried out for overviewing the individual outcome as well as comparative outcome. The computation of the time for data structurization is carried out by obtaining the difference between the times for previous record with the current data structurization record. The analysis is carried out over 10 experimental trials, where the allocation of input datastream is increased randomly in order to map with the near-real world data streaming process over the network.

The outcome in Figure 5 shows that data structurization time increase with increase in experimental trials, which is very much natural in order. It is because every experimentals trials have varying allocation of size of data streams with 10-20% definitive increment and rest 80-90% allocation of data is carried out in random fashion. An interesting observation to see in this part of analysis are two folds viz. i) the proposed system offers a gradual increase in time, which is quite deterministic in order and shows that it offers a stabilized data processing capability, and ii) irrespective of fluctuating size of data streams, the proposed system offers better transformation performance where abnormal traffic increase doesn't have a significant effect towards the performing of performing analytical operation.

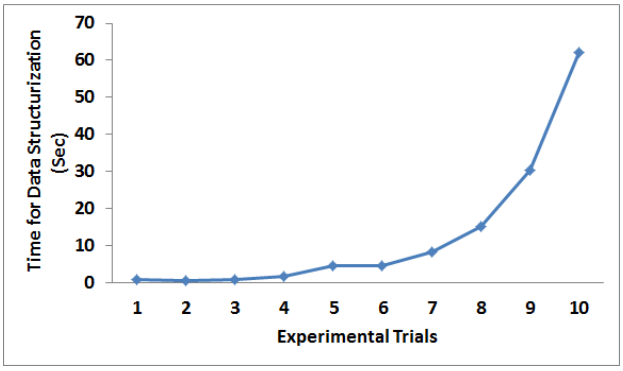


Figure 5. Outcome of data structurization time

The next part of the analysis is check for the time consumption for the core analytical process as well as data structurization process. Bascically, an effective data structurization always lead to higher accuracy of the mined data; however, higher value of the knowledge extraction time can be expected. It is

because after the data structurization is carried out than the system will be always required to cross check with the indexes from the cloud resources in order to confirm the position and term obtained from the dynamic data. This could possibly increase the knowledge extraction time to some extent. Figure 6 highlights the same fact where it is shown that knowledge extraction time is slightly more than the time for daa structurization.

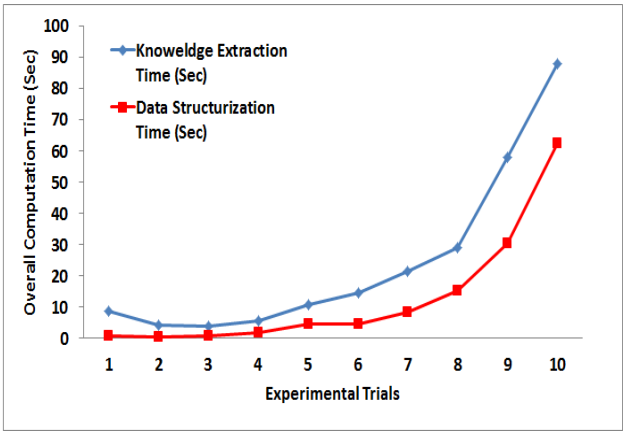
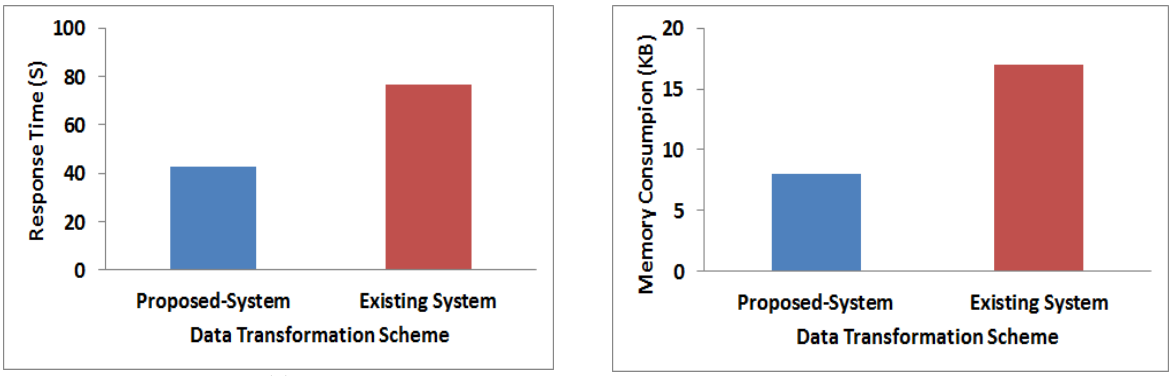


Figure 6. Analysis of overall computational time

Apart from this, the proposed system has been compared with the existing ETL scheme (Extract/Transform/Load) which is the frequently used data transformation scheme in existing system (e.g. [11-13]). The comparative analysis is carried out with respect to response time and memory consumption as shown in Figure 7.



(a) (b)

Figure 7. Comparative analysis, (a) Analysis of response time, (b) Analysis of memory consumption

A closer look into the outcome of the comparative analysis in Figure 7 shows that proposed system offers significantly better performance as compared to the existing system in terms of response time as shown in Figure 7(a) and memory consumption time as shown in Figure 7(b). The prime reason behind this is ETL scheme is shrouded with a problem of parsing with the source data especially of the data is massive and heterogeneous in itself. However, proposed system offers a very lightweight scheme without any dependencies of third party applications or plugins as well as uses a simplified semantics for extracting the knowledge. The outcome of memory consumption also shows that propsed system occupies a very low amount of memory. The core reason behind this is that proposed system uses a sandboxing mechanism where a temporary buffer is constructed which can significantly save the consumption of the memory. Hence, there is no record of the underlying process which makes the analytical operation faster as well as memory efficient. Therefore, the proposed sysem offers a cost effective data transformation process for EHR data.

4. CONCLUSION

The paper presents a unique approach of transformation where the complex unstructured stream of medical date is converted to highly structured data in different mechanism that is free from any dependencies. The contribution of the proposed system are as follow: i) the proposed system reduced the storage complexity as well as increases faster processing time, ii) the proposed system introduces a novel mechanism to identify static and dynamic data where the static data is stored in cloud and dynamic data is stored in temporary buffer using template based approach, iii) the proposed system doesn't store any raw data over cloud but stores only the mined data which reduces higher degree of complexity and increases enrichness of data.

REFERENCES

[1] L. Pallavi, *et al*., "ERMO² algorithm: an energy efficient mobility management in mobile cloud computing system for 5G heterogeneous networks," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, pp. 1957-1967, 2019.

[2] C. Dobre and F. Xhafa, "Pervasive Computing: Next Generation Platforms for Intelligent Data Collection," *Academic Press*, 2016.

[3] T. Francis and M. Madhiajagan, "A Comparison of Cloud Execution Mechanisms: Fog, Edge and Clone Cloud Computing," *Proceeding of the Electrical Engineering Computer Science and Informatics*, vol. 4, pp. 446-450, 2017.

[4] F. M. Groom and S. S. Jones, "Enterprise Cloud Computing for Non-Engineers," *CRC Press*, 2018.

[5] K. L. Jackson and S. Goessling, "Architecting Cloud Computing Solutions: Build cloud strategies that align technology and economics while effectively managing risk," *Packt Publishing Ltd*, 2018.

[6] A. Rghioui and A. Oumnad, "Challenges and opportunities of internet of things in healthcare," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 2753-2761, 2018.

[7] B. Ristevski and M. Chen, "Big Data Analytics in Medicine and Healthcare," *Journal of Integrative Bioinformatics*, vol. 15, no. 3, 2018.

[8] K. Abouelmehdi, *et al*., "Big healthcare data: preserving security and privacy," *Journal of Big Data*, vol. 5, no. 1, 2018.

[9] Z. A. Mdaghri, *et al*., "Study and analysis of data mining for healthcare," *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, Tangier, pp. 77-82, 2016.

[10] Z. Karakaya, "Software engineering issues in big data application development," *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, pp. 851-855, 2017.

[11] A. Muslim, *et al*., "Web services of transformation data based on OpenEHR into Health Level Seven (HL7) standards," *2017 Second International Conference on Informatics and Computing (ICIC)*, Jayapura, pp. 1-4, 2017.

[12] P. S. Diouf, *et al*., "Variety of data in the ETL processes in the cloud: State of the art," *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, pp. 1-5, 2018.

[13] Q. Wang, *et al*., "Research of ETL on university data exchange platform," *2011 IEEE International Symposium on IT in Medicine and Education*, Cuangzhou, pp. 285-288, 2011.

[14] Y. Wang, *et al*., "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technological Forecasting and Social Change*, vol. 126, pp. 3-13, 2018.

[15] J. A. Perez, *et al*., "Big Data for Health," in *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1193-1208, Jul. 2015.

[16] S. T. Prasad, *et al*., "Diabetic data analysis in big data with predictive method," *2017 International Conference on Algorithms, Methodology, Models & Applications in Emerging Technologies (ICAMMAET)*, Chennai, pp. 1-4, 2017.

[17] N. Das, *et al*., "Big Data Analytics for Medical Applications," *International Journal of Modern Education and Computer Science*, vol. 10, no. 2, pp. 35, 2018.

[18] I. G. Magariño, *et al*., "Agent-Based Simulation of Smart Beds With Internet-of-Things for Exploring Big Data Analytics," in *IEEE Access*, vol. 6, pp. 366-379, 2018.

[19] H. Zhong and J. Xiao, "Enhancing health risk prediction with deep learning on big data and revised fusion node paradigm," *Scientific Programming*, 2017.

[20] M. Chen, *et al*., "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in *IEEE Access*, vol. 5, pp. 8869-8879, 2017.

[21] Y. Zhang, *et al*., "Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data," in *IEEE Systems Journal*, vol. 11, no. 1, pp. 88-95, Mar. 2017.

[22] P. Wu, *et al*., "Omic and Electronic Health Record Big Data Analytics for Precision Medicine," in *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263-273, Feb. 2017.

[23] M. Viceconti, *et al*., "Big Data, Big Knowledge: Big Data for Personalized Healthcare," in *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1209-1215, Jul. 2015.

[24] S. Shafqat, *et al*., "Big data analytics enhanced healthcare systems: a review," *The Journal of Supercomputing*, pp. 1-46, 2018.

[25] M. Chen, *et al*., "5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds," in *IEEE Communications Magazine*, vol. 56, no. 4, pp. 16-23, Apr. 2018.

[26] C. Garattini, *et al*., "Big Data Analytics, Infectious Diseases and Associated Ethical Impacts." *Philosophy & Technology*, vol. 32, pp. 69-85, 2017.

[27] D. Chrimes and H. Zamani, "Using Distributed Data over HBase in Big Data Analytics Platform for Clinical Services," *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1-16, 2017.

[28] L. A. Tawalbeh, *et al*., "Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications," in *IEEE Access*, vol. 4, pp. 6171-6180, 2016.
[29] B. K. Sarkar, "Big data for secure healthcare system: a conceptual design," *Complex & Intelligent Systems*, vol. 3, no. 2, pp. 133-151, 2017.
[30] Y. Zhang, *et al*., "IEEE Access Special Section Editorial: Big Data Analytics for Smart and Connected Health," in *IEEE Access*, vol. 4, pp. 9906-9909, 2016.
[31] U. Srinivasan and B. Arunasalam, "Leveraging Big Data Analytics to Reduce Healthcare Costs," in *IT Professional*, vol. 15, no. 6, pp. 21-28, 2013.
[32] R. Bhalla and A. Bagga,"Opinion mining framework using proposed rb-bayes model for text classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 477-485, 2019.
[33] S. Aich, *et al*., "Convolutional neural network-based model for web-based text classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 5185-5191, 2019.
[34] "Census Data," United State Census, [Online]. Available: http://www.census.gov/data.html. [Retrieved 11-04-2020].
[35] "Physionet," [Online]. Available: http://www.physionet.org/physiobank/database/. [Retrieved on 11-04-2020]
[36] "Registry of Open Data on AWS," [Online]. Available: https://registry.opendata.aws/, [Retrieved on 11-04-2020].
[37] "NHS Continuing Healthcare Activity," data.gov.uk. [Online]. Available: https://data.gov.uk/dataset/f259a7b7-6c97-4632-93d0-38b93afb324f/nhs-continuing-healthcare-activity, Retrieved on 11-04-2020

BIOGRAPHIES OF AUTHORS

Chandru A S, Assistant Professor, Department of Information Science & Engineering, NIE Institute of Technology, Mysuru, India. He is pursuing his PhD from VTU, Belagavi, India. His area of interest is Big data analytics. He has 8 years of work experience.

Seetharam Keshavarao, Professor, Department of Computer Science & Engineering, Jnanavikas Institute of Technology, Bidadi, Bengaluru, India. He has completed his PhD from IIT Bombay. His area of interest was artificial intelligence. He has around 34 years of work experience.