A SURVEY ON RESOURCE OPTIMIZATION AND PREDICTION TECHNIQUES WITH BIG HEALTHCARE DATA

[1]S. Arun Kumar,[2]M. Venkatesulu

ABSTRACT—*Big data (BD) analytics is utilized to gathers and analyzes large volume of data to find useful information. BD is utilized to mine valuable information for predictive analytics. Healthcare data classification classifies the patient details into many columns based on the user requirement. Resource optimizations are collection of processes to match with the accessible resources to attain the goal. Prediction examines the present and historical events to forecast about the future events. In recent times, many research works are carried out for improving the performance of classification and prediction process with minimal resource utilization. However, the prediction time and resource utilization remained challenging issue in healthcare applications. The key objective of the paper is to present comparative literature survey of existing resource optimization and prediction techniques. The contribution of the survey is two-fold. In first one, a brief description of storage techniques, resource optimization and prediction techniques like NoSQL database (DB), data management methodology (DMM), two-stage scheduling policy, data parallelism (DP), external scheduler (ES), internal scheduler (IS), Fuzzy Linguistic Summarization (LS) approach and thus highlights the significant features, merits, and shortcomings. In second one, one of the existing schemes was selected to test whether it is adequate for real systems. Consequently, our theoretical analysis is compared with experimental results to look forward the results and to afford valuable insight to future researchers.*

Keywords*-- Big data analytics, healthcare data classification, Data traffic, resource utilization, Prediction*

I. INTRODUCTION

BD includes huge development in healthcare applications. Precise analysis of medical data is performed using BD development in biomedical and healthcare communities which leads to early prediction of disease, patient care and community services. Healthcare BD are portrayed as electronic health DS which are huge and difficult to manage with conventional software and/hardware. BD is applied for real-time disease tracking, forecasting disease outbreaks and developing personalized health care. BD contributes to an evidence-based medicine, device/remote monitoring and patient profile analytics.

[1] *Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Sikkim Manipal University, Majitar, Sikkim, arunsphd@gmail.com*

[2] *Department of Information Technology, Kalasalingam Academy of Research and Education Krishnan koil, India, m.venkatesulu@klu.ac.in*

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

A. Contribution

Contribution of paper is twofold. First, we present a theoretical analysis of the existing storage techniques, data access methods, resource optimization techniques and predictive analytics. By considering this examination, the inefficiencies of existing schemes were described while developing our analysis by experiencing issues. Finally with results of our analysis, we selected to test one of the existing schemes to predict whether it is adequate for usage in real systems. Consequently, our theoretical analysis is compared with experimental results. We look forward that the results of this work provide valuable insight to new technique designers and will create additional researches in the area.

B. Organization of the Survey

This paper is ordered as follows: Section II discusses review on different resource optimization and prediction techniques with big healthcare data, Section III portrays the study and analysis of the existing resource optimization and prediction techniques, Section IV explains the possiblecomparison between them. In Section V, the discussion and limitations of the existing techniques are studied and Section VI concludes the paper. The key area of research is to improve the performance of resource optimization and prediction techniques

.

II. LITERATURE REVIEW

To store the health data, Relational DB is utilized which does not handle huge and different nature. For storage of healthcare data, a designed model utilizes NoSQL databases in [1]. For accessing to distribution properties, the presented model was executed in Cloud environment. However, space complexity (SC) remained unaddressed during healthcare data storage. To parallelism in bioinformatics workflows, a DMM was presented in [2] through reducing the data-interdependent file transfers. By combining with new two-stage scheduling approach (TSA), it performs load estimation and balancing heterogeneous distributed computational resources (HDCS). However, the load balancing efficiency was not improved using DMM.

In [3], a data analytics and visualization framework was presented for health-shocks prediction with large-scale health informatics dataset (DS). The designed framework with cloud computing (CC) services based on Amazon web services (AWS). But, prediction accuracy was not enhanced using data analytics and visualization framework. A new data- driven, hierarchical and interactive phenotyping tool called PHENOTREE in [4] allowed the physicians and medical researchers to contribute in phenotyping process of large-scale EHR cohorts. PHENOTREE used sparse principal component analysis (SPCA) to recognize the key clinical characteristics that classify the population. But, the classification accuracy was not enhanced using PHENOTREE.

A BD analytics-enabled business value model was designed in [5] where resource-based theory (RBT) and capability building view elucidate how BD analytics abilitiesare improved. Though resource utilization was optimized, the load balancing was not carried out in efficient manner. In [6], - omic and EHR data characteristics are connected with data analytics with data pre-processing, mining and modeling. BD analytics assign precision medicine with disease biomarkers detection from multi-omic data and integrate -omic information into EHR. But, the scheduling of data was not performed with EHR data characteristics. A new methodology was designed in [7] for automatic classification of intracardiac electrograms (EGMs) with lesser signal preprocessing. For lessening computational complexity, a Compression-based Similarity Measure (CSM) was designed. Though the classification was carried out, the prediction time was not reduced.

To detect the generation of testable new hypothesis, a standardized method was presented in [8] using pre-existing amounting of biological information. But, the classification accuracy was not improved using standardized methods. The large quantity of chemogenomics data presented to design model depending on BD in [9]. High quality DS preparation is an essential step to gather chemogenomics DS. But, the storing and accessing of BD were not carried out in efficient manner. To enhance scaling and processing complex and heterogeneous data, a novel innovative technology was presented in [10]. Stakeholder investments in data acquisition identify enormous potential of BD to gather conventional information benefits and build knowledge assets. Though the classification was carried out, resource utilization was not minimized.

In [11], a novel frontier in technological developments and it applications of cardiovascular medicine was determined. Issue of BD science on cardiovascular proteomics study and translation to medicine were explicated. With the structured and unstructured data, a convolutional neural network (CNN)- based multimodal disease risk prediction algorithm was designed in [12]. Though the classification technique was employed, the prediction of disease was not carried out in efficient way.

III. EXISTING RESOURCE OPTIMIZATION AND PREDICTION TECHNIQUES WITH BIG HEALTHCARE

BD is high-volume which needs cost-efficient, novel forms of information processing for improved imminent and decision making. Major issue with BD is continuously increasing requires of computational resources and storage services. It outcome in enhancement of large-scale high performance computing (HPC) models namely cluster, grid and CC. CC is elucidated as HPC environment comprises virtual machines (VMs) to scale resources consistent with computational need. Most of the recent research works aimed to increase the performance prediction time and resource optimization in big healthcare data for reducing the risk level of patients. In order to increase the performance of prediction and resource optimization, their existing techniques are extensively studied.

A. Cloud Healthcare Systems

In [13], public health management systems were introduced using My Personal Health Record eXpress (MphRx). The personal data of patients are stored and preserved through online access by using the designed systems. Users can retrieve the data using every device at anytime and anywhere. The designed system takes care of themselves through systems consistent with health status. Cloud services are category of Software as a Service (SaaS) platform. But, the data access took large amount of time for accessing the data from the health management system. The ensemble classifiers are employed for classify and store the data which reduces the time for accessing. These classification techniques help in easy data access with minimal time consumption.

Based on O'Driscoll, CC and Hadoop was presented in [14] for processing and analyzing large genomic DSs, using the proper technologies. GENOME data are significant in healthcare branch that include more value. Through the petabyte scale, the designed model in [14] utilizes thedistributed and parallelized infrastructure for

DS. Apache HBase approach and Map-Reduce programming were designed to store and process the clinical data. In [15], the designed system was integrated with web-based layer, to parallelize computation process. In above mentioned techniques, the Map-Reduce functions were used to reduce the SC. However, the SC was higher. Hashing techniques are utilized to lessen the SC. Hashing technique stores the data with help of hash values.

Through observing open and visual environments, the gap between potential and actual data usage was evaluated in [16]. By combining MIMIC DB in RapidMiner environment, a framework was introduced for efficient exploitation of healthcare data. For data analysis, a Hadoop and analytic algorithms was presented. For Decision Support Service, a structural design of healthcare SaaS Platform is designed in [17]. In Cloud service model, Microsoft's Azure was introduced. To analysis the biomedical data, a cloud-based system was presented in [18]. To select the optimal predictive algorithms, the designed system combined the metalearning framework and open source BD technologies for analysis. A communicational framework was designed in [19] that connect key segments of health world.

Table 1: Could health systems studies

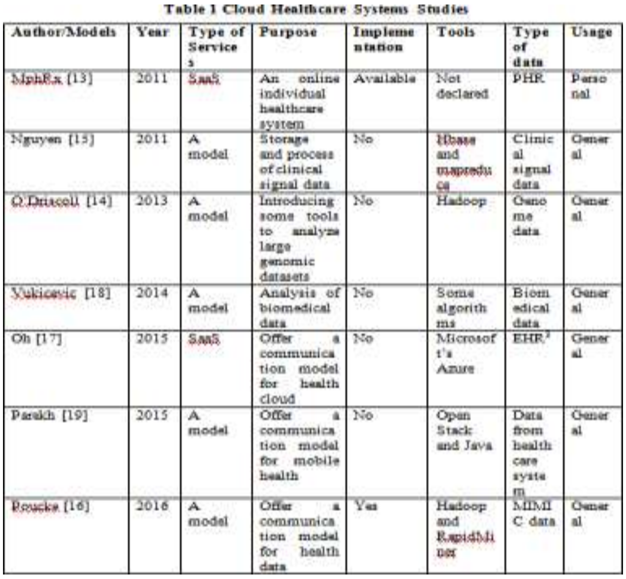


Table 1 Cloud Healthcare Systems Studies

Author/Models	Year	Type of Services	Purpose	Implementation	Tools	Type of data	Usage
MphRx [13]	2011	SaaS	An online individual healthcare system	Available	Not declared	PHR	Personal
Nguyen [15]	2011	A model	Storage and process of clinical signal data	No	Hbase and mapreduce	Clinical signal data	General
O.Driscoll [14]	2013	A model	Introducing some tools to analyze large genomic datasets	No	Hadoop	Genome data	General
Vukicevic [18]	2014	A model	Analysis of biomedical data	No	Some algorithms	Biomedical data	General
Oh [17]	2015	SaaS	Offer a communication model for health cloud	No	Microsoft's Azure	EHR^{2}	General
Parekh [19]	2015	A model	Offer a communication model for mobile health	No	Open Stack and Java	Data from health care system	General
Roucke [16]	2016	A model	Offer a communication model for health data	Yes	Hadoop and RapidMiner	MIMIC data	General

The data mining algorithm mine the data entered in presented framework. The comparison of the Cloud Healthcare Systems Studies is studied in table 1.From the table, few restrictions are present for BD. Healthcare data comprises large volume of data from several sources. It is essential to select model with distribution capability. Existing approaches are limited to SaaS and not utilized for backend users. Healthcare data models include many data as additional aspect. Therefore, a comprehensive model needed to be designed to address the issues of healthcare's data.

B. NoSQL DBs

For storing the healthcare data, NoSQL DB is utilized. It is vulnerable to Cloud environment and exploits Cloud capabilities. NoSQL DB utilization is used in many sectors due to their ability to manage application needs. In CC, for storing huge amount of redundant data, it is a preferred selection. NoSQL DBs presents a novel storage design with greater scalability, availability and fast retrieval needs controlling unstructured and partially structured data. NoSQL DBs are open source and is low cost per terabyte than traditional DBs. For web based data, DBs is appropriate one.

Applying data model, it includes four types for DB, namely

i. Key-value DB,
ii. Columnar DB,
iii. Document-Based DB and
iv. Graph-Based DB

1) *Key-value based Databases*

Key-value DBs includes minimal complex structure. Using pre-defined key, Data's are retrieved and stored with independent value. Key/value pair is a unique value in a set employed for accessing the data.

2) *Columnar–based Databases*

Column is significant in Columnar DBs and includes related data grouped closely. Data is accumulated in column- family basis which illustrated in configuration or startup time. It stores every data types efficiently.

3) *Document–based Databases*

Document–based DBs organizes large and complex documents. It controls types of documents to include no. of fields in any length. Document is attained as whole object and it failed to split name/value pairs. For indexing of documents on primary identifier and properties, it is exploited.

4) *Graph-based Databases*

Graph-based DB exploits graph structures with nodes, edges and properties to store data.

5) *NoSQL DBs characteristics*

NoSQL DB comprises diverse data models. Users choose one of them using their application necessitates. Each type of data are detected namely medical findings, artifacts x- ray images and electroencephalography wave recordings. Huge quantity of data is document-based and unstructured. The major necessity is frequent modification of definitions. Continuous reads and writes employ the Column-based DB. Through the particular pattern, the data are recovered and few columns are engaged in query. The key-value models are used for access the enhanced performance writing. The designed model functions with flat data model and query is executed in defined keys.

By Document-based DB, the larger range of access patterns and data types is operated. Several reads and writes are managed by DB. It handles difficult structure and number of columns without any necessity to build scheme again. It supports complex queries in many fields. To identify the relationships between objects, Graph-based models are used. NoSQL DBs are prepared for huge amount of data in distributed way. Data were distributed in several machines in various geographical regions. In Cloud environment, NoSQL DBs are applicatory which intensifies their extensibility.

With assumption of query time, Document-based DBfunction improved than SQL Server in 'Write' operation for DB sizes. SQL server executes 'Read' operation for severalqueries and DB sizes enhanced than Document-based DB due to its indexing property. The shared features are totaled to Document based DB, then the speed of data recovery process is enhanced. However, Sharding affects writing speed and lessens the 'write' operations than prior state. 'Write' speed is greater than SQL Server. In NoSQL DBs, the data access is a

difficult task. To increase analysis of NoSQL DBs, classifier with hashing techniques helps to access the data from the large DB with minimal time consumption.

Table 2: health care data and nosql dbs characteristics comparision

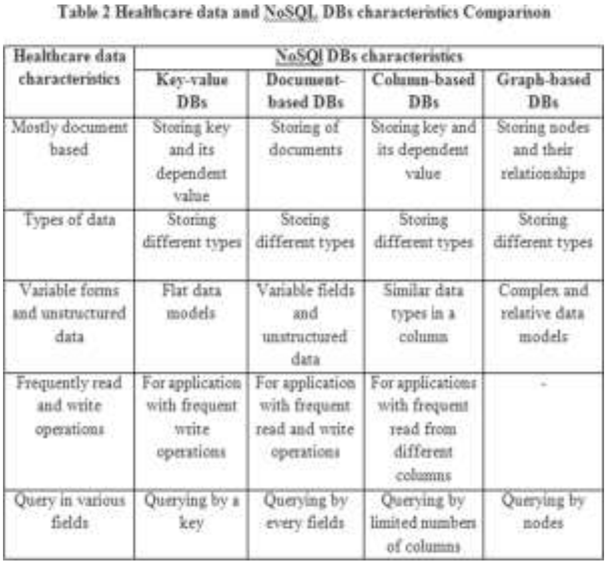


Table 2 Healthcare data and NoSQL DBs characteristics Comparison

Healthcare data characteristics	NoSQl DBs characteristics			
	Key-value DBs	Document-based DBs	Column-based DBs	Graph-based DBs
Mostly document based	Storing key and its dependent value	Storing of documents	Storing key and its dependent value	Storing nodes and their relationships
Types of data	Storing different types	Storing different types	Storing different types	Storing different types
Variable forms and unstructured data	Flat data models	Variable fields and unstructured data	Similar data types in a column	Complex and relative data models
Frequently read and write operations	For application with frequent write operations	For application with frequent read and write operations	For applications with frequent read from different columns	-
Query in various fields	Querying by a key	Querying by every fields	Querying by limited numbers of columns	Querying by nodes

C. Data aware optimization in hybrid clouds

While reducing the data-interdependent file transfers, DMM is designed to attain parallelism in bioinformatics workflows. Novel TSA with designed methodology is exploited for carry out the load estimation and balancing within HDCS. Through minimizing the file transfer among sites the designed methodology increases the time and cost efficiency. For mapping their execution into heterogeneous distributed resources, DMM arranges the workflow into pipelines with lesser data interdependencies with scheduling policy. But, the workflow was not reduced using DMM. The data parallesim can be used for improving the workflow execution performances.

D. Data Parallelism (DP) Approach

DP is carried out in bioinformatics workflows that accelerate the workflow execution. It contains input fragmentation into chunks where it processed separately. DPapproach is presented for bioinformatics workflows namely sequence alignment and mapping of short reads to attain high degree of parallelism in multiprocessor architectures and distributed computing environments. Data consistency requires the result of separately processed chunks. In distributed computing environment where data is position on various sites, the designed approach solves the data interdependency issues where data necessity transferred from multiple sites is recombined.

A sensible approach addresses the data interdependencies for reducing or removing the unnecessary file transfers whose output are recombined on same site. It is responsible for processing the recombined output. It is finished on same site and operates on output of previous in same site. It is apparent where the recursive process considered the anticipated data dependencies of analysis. Without data interdependencies, the segments of original workflow are split into workflow ensembles. Future tasks operating on similar data are grouped back-to-back that form pipeline. Data input space includes instances. Instance is single data entry, where the data exist individually.

Input data are fragmented into the chunks for maintain data dependencies of several subsequent analysis. Pipeline is created through combining future tasks operation on same data with back-to-back end. Data input space comprises no. of instance. Instance is single data entry where data exist individually. Instances are systematized in organization units (OU) where group of instances discover data dependencies of one or more tasks. OU is group of Insts which address data dependencies with number of successive process for formation of OU pipeline. But, the workflow execution consumes large amount of resources. The scheduling techniques can be used to reduce the resource utilization during file transfer.

E. Two-stage scheduling approach

The designed process includes Insts identification in input data and groups them into OUs consistent with workflow data interdependencies. An identifier comprises OU it belongs to and presented for every Inst. Identifier is connected in respective data and protected indefinitely. Initial integrity of input data is guaranteed to protect workflow execution. Moreover, it is exploited for recombination process and ensures the accessibility of information in forthcoming stages for analysis. A new 2-stage scheduling approach connects an ES at stage 1 mapping OU pipelines into sites internal to every site scheduler at stage 2 to attain data and task parallelism (TP) when managing OU pipeline.

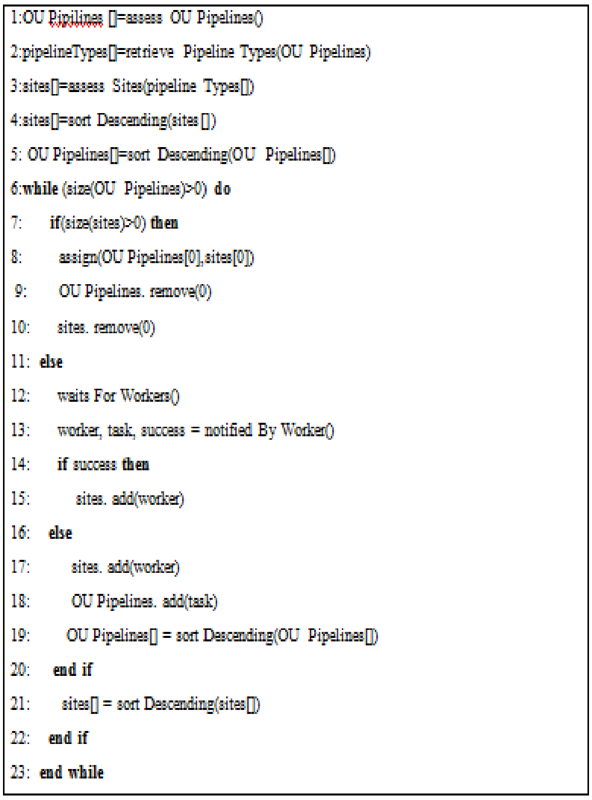
1) External scheduler

Through the ES, load balancing of OU pipelines are performed. The first step is implemented, while OU pipelines and computational sites are different and estimation concerning OU pipeline loads and processing power of sites. Second step includes the exploitation of aforementioned estimations via scheduling algorithm grouped with allocating OU pipelines to computational resources. The algorithmic process of ES is explained in Algorithm 1.

Load of OU pipelines and evaluation type are executed by ES. It decides the capabilities of accessible sites in processing the pipelines through retrieving the targeted benchmarks or functioning new on fly. The OU pipelines are submitted to the sites consistent with the fastest processor largest task (FPLT) algorithm and job failures are managed by resubmission.

2) Internal scheduler

IS are local for every site. During OU pipeline process, it is responsible to achieve data and TP. TP involves independent tasks execution directly in parallel while DP requires detection of tasks whose input are fragmented in chunks and processed. For automatic identification, Tasks aremarked as appropriate one for fragmentation in workflow description stage or conserve list of tasks. The algorithm of IS operation is described in algorithm 2 and 3.



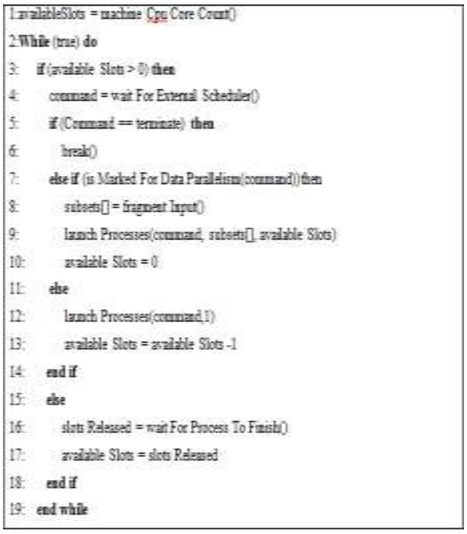
```
OU Pipilines []=assess OU Pipelines()
pipelineTypes[]=retrieve Pipeline Types(OU Pipelines)
sites[]=assess Sites(pipeline Types[])
sites[]=sort Descending(sites [])
OU Pipelines[]=sort Descending(OU Pipelines[])
while (size(OU Pipelines)>0) do
    if(size(sites)>0) then
        assign(OU Pipelines[0],sites[0])
        OU Pipelines. remove(0)
        sites. remove(0)
    else
        waits For Workers()
        worker, task, success = notified By Worker()
        if success then
            sites. add(worker)
        else
            sites. add(worker)
            OU Pipelines. add(task)
            OU Pipelines[] = sort Descending(OU Pipelines[])
        end if
        sites[] = sort Descending(sites[])
    end if
end while
```

Algorithm 1 External scheduler Algorithm

The no. of CPUs on computational site is discovered by IS and sets no. of simultaneous processing slots. Master sends commands to IS and allocates threads to carry out in parallel. Where it gathers task where DP is feasible, it divides the input data into individual chunks or subsets. Furthermore, it launches threads to process in parallel. Through number of fragments, a decision was made and includes tradeoff among process initialization overhead and load balancing among threads. If all data present in similar site, then it unable to allocate data processing load in earlier stage between threats. Moreover, the several threats access the data in anytime without additional cost.

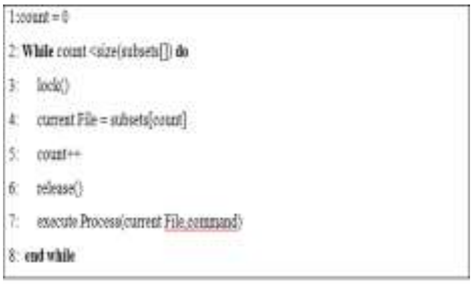
3) FPLT Algorithm

To detect the time problems, FPLT is utilized while every task is accessible from start with no adding computational burden. If, the computational power of processors is unequal then the FPLT algorithm is more complex. The Processor assigns a task to exceed its ability and it causes delay in makespan workflow. This process is carried out when a few processors are smaller than average participating in workflow.



```
availableSlots = machine Cpu Core Count()
While (true) do
  if (available Slots > 0) then
    command = wait For External Scheduler()
    if (Command == terminate) then
      break()
    else if (is Marked For Data Parallelism(command))then
      subsets[] = fragment Input()
      launch Processes(command, subsets[], available Slots)
      available Slots = 0
    else
      launch Processes(command,1)
      available Slots = available Slots -1
    end if
  else
    slots Released = wait For Process To Finish()
    available Slots = slots Released
  end if
end while
```

Algorithm 2 Internal scheduler Algorithm



```
count = 0
While count <size(subsets[]) do
  lock()
  current File = subsets[count]
  count++
  release()
  execute Process(current File,command)
end while
```

Algorithm 3 Launch Data Parallel Process

4) Versatile framework

In [2], a versatile framework was designed to enhance parallel implementation of data-intensive bioinformatics workflows. It attains surpassing time and cost efficiency through lessens the file transfers among sites. It accomplishes through combination of DMM that systematizes workflow into pipelines with minimal data interdependencies with scheduling policy for mapping execution into collection of heterogeneous distributed resources with hybrid cloud. Though the cost efficiency and surpassing time performance was improved, the resource optimization was not carried out in effective way. For optimizing the resource, scheduling algorithms can be used to balance the load during the file transfer in cloud.

5) Cloud enabled data analytics and visualization framework

For health-shocks prediction, the framework is designed with large-scale health informatics DS. Depending on AWS integrated with geographical information systems (GIS), the framework offers the CC services for BD capture, storage, index and visualization of data. A predictive model is designed for health-shocks and gathered data from 1000 households in rural and distantly accessible regions. Via fuzzy rule summarization method, the gathered data produce predictive model of health-shock. The designed method provides stakeholders with interpretable linguistic rules to elucidate causal factors concerning health-shocks. For categorizing the health-shock results, the interpret-ability and generated data models accuracy are utilized.

Through the fuzzy LS approach, the pre-processed data was utilized to generate fuzzy rule based classification model for health-shocks prediction. The designed technique generated an interpretable rule based model to visualize and forecast level of health-shocks experienced by individuals. Extracted fuzzy rules utilize quality measures which found strength of each rule in ability to model data. For ranking and interpretation, it offers stakeholders with quality of rules. Based on interpretability of rules, the generated fuzzy model is computed in their ability factors affecting diverse levels of health-shocks to predict health-shocks levels from unlabeled data. The generated rules offer sensible and meaningful profiles with factors equivalent to several health-shocks. Depending on k-fold cross-validation of data samples, the prediction accuracies of fuzzy model are designed. LS approach attained better prediction accuracies with larger data samples.Through CC, Large-scale health data analytics failed to assist healthcare professionals to generate and conduct surveys with lesser human and financial resources. It identifies socio-economic, environmental and cultural norms to directly or indirectly cause health-shocks. It observes and detects the healthcare system and occurrence of health-shocks in rural and tribal areas of Pakistan.

To manage real world information indistinctness, Fuzzy Logic Systems (FLSs) offers transparent and flexible model through linguistic quantifier’s utilization like Poor or High. It denotes methodology to calculate with words, then linguistic quantifiers explicated by fuzzy sets with human interpretable If–Then rules. Fuzzy rules suggest clear LS patterns linking independent input variables and dependent target output decision. Also, mined fuzzy classification rules comprises quality measures associated with every rule to calculate strength of patterns in data and top rules are ranked with exact output conditions. Figure 1 describes the Fuzzy LS approach 1.

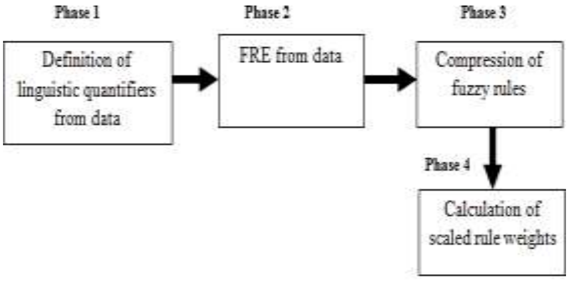

Phase 1
Definition of linguistic quantifiers from data
Phase 2
FRE from data
Phase 3
Compression of fuzzy rules
Phase 4
Calculation of scaled rule weights

Figure 1 Fuzzy Linguistic Summarization approach

Figure 1: fuzzy linguistic summarization approach

a) Definition of linguistic quantifiers from data

The input/output data in phase 1 has four independent variables and one dependent variable representing strictness of health-shocks which are mapped to predefined linguistic quantifiers. Derived variables were acquired from pre- processed data gathered in user study. Doubts concerning linguistic quantification over data values in numerical and continuous valued data attributes advised the requisite to exploit fuzzy sets. Using the fuzzy membership function(MF), the generalization of crisp set which assigns to estimate membership of an element. Preprocessed data is partitioned into group of MFs which measure values of data attributes into linguistic labels and partitions data space into fuzzy regions.

b) Fuzzy rule extraction (FRE) from data

In phase 2, FRE was designed. For extracting fuzzy rules from sampled data, FRE is exploited which is termed as single-pass method. Data is joined to fuzzy sets for antecedents and consequents of rules in phase 1. With duplicate and contradictory rules, If–Then profile rule for every data instances was generated via this process.

c) Compression of fuzzy rules

In phase 3, the data instance based profile rules are compressed to evaluate data instances into distinctive end rules. The process comprises modified computation of two rule quality and acquires scaled weight of unique summarization rule. Quality measures are depends on generality and reliability. For assessing rule generality, the fuzzy rule support is employed and rule reliability is based on confidence.

IV. COMPARISON OF RESOURCE OPTIMIZATION AND PREDICTION TECHNIQUES WITH BIGHEALTHCARE DATA & SUGGESTIONS

In order to compare theresource optimization and prediction techniques, no. of patient data is taken to perform the experiment. Various parameters are used for improving the resource optimization and prediction techniques with big healthcare data.

A. Model Space Complexity

SC is measured as the amount of memory space consumed for storing and accessing patient data. It is calculated in megabytes (MB). SC is formulated as,

$$SC = nunberofpatientdata * nenoryspaceconsunedofonepatientdata$$

Minimal SC, the method is said to be more efficient.

B. Model Resource Utilization Rate

Resource utilization rate (RUR) is measured as amount of resources consumed for load balancing the patient data. It is evaluated in percentage (%).RUR is formulated as,

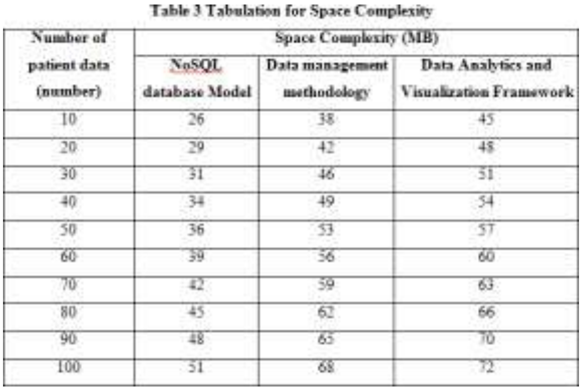


Table 3 Tabulation for Space Complexity

Number of patient data (number)	Space Complexity (MB)		
	NoSQL database Model	Data management methodology	Data Analytics and Visualization Framework
10	26	38	45
20	29	42	48
30	31	46	51
40	34	49	54
50	36	53	57
60	39	56	60
70	42	59	63
80	45	62	66
90	48	65	70
100	51	68	72

RUR

= NuNber of patient data correctly

scheduled with available re

Total NUNBEr of

patiend data

Higher RUR, the method is said to be more efficient.

The SC with number of patient data ranges from 10 to 100 is portrayed in Table 3. SC is compared with existing NoSQL DB Model, DMM and Data Analytics and Visualization Framework. From table 3, it is evident that SC using NoSQL DB Model is lesser as compared to DMM and Data Analytics and Visualization Framework. The graphical representation of SC is shown in figure 2.

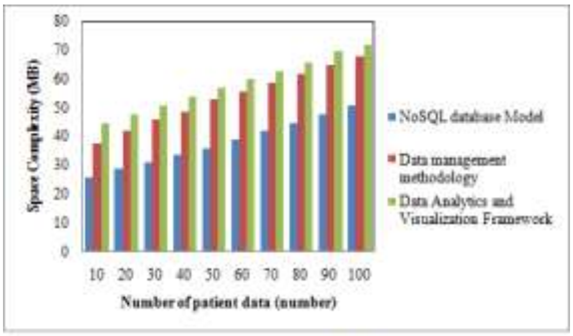


Figure 2 Measure of Space Complexity

Figure 2: measure of space complexity

From figure 2, SC based on the different number of patient data is described. From the figure 2, NoSQL DB Model consumes lesser SC than DMM and Data Analytics and Visualization Framework. Research in DMMutilizes 30% lesser memory space than NoSQL DB Model and has 36% lesser memory space than Data Analytics and Visualization Framework.RUR with number of patient data is portrayed in table

4. RUR is compared with existing NoSQL DB Model, DMM and Data Analytics and Visualization Framework. From table 4, it is evident that RUR using DMM is higher as compared to NoSQL DB Model and Data Analytics and Visualization Framework. The graphical illustration of RUR is shown in figure 3.

Figure 3 describes the RUR with different number of patient data. From the figure 3, DMMhas higher RUR than NoSQL DB Model and Data Analytics and Visualization Framework. Research in DMMhas 6% higher RUR than NoSQL DB Model and has 19% higher RUR than Data Analytics and Visualization Framework.

Table4:tabulationforreasourceutilizationrate

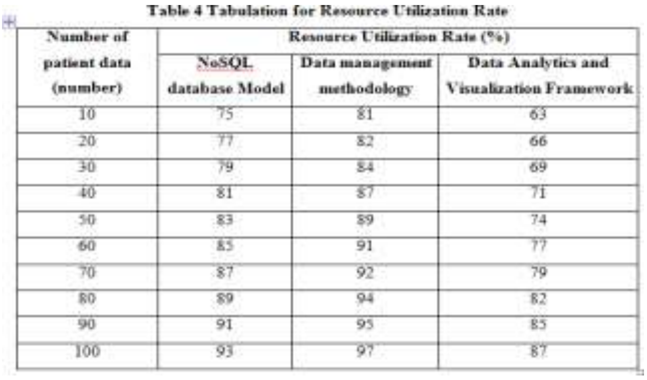


Table 4 Tabulation for Resource Utilization Rate

Number of patient data (number)	Resource Utilization Rate (%)		
	NoSQL database Model	Data management methodology	Data Analytics and Visualization Framework
10	75	81	63
20	77	82	66
30	79	84	69
40	81	87	71
50	83	89	74
60	85	91	77
70	87	92	79
80	89	94	82
90	91	95	85
100	93	97	87

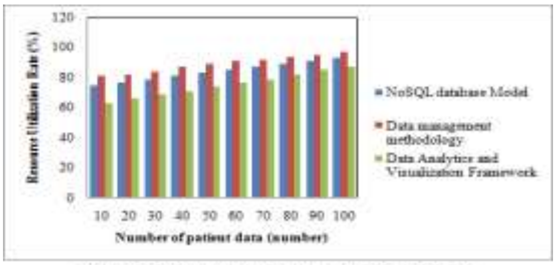


Figure 3 Measure of Resource Utilization Rate

Figure 3: measure of reasource utilization rate

Table 5: tabulation for predition time

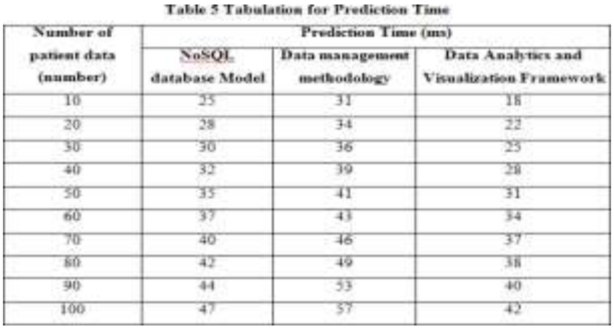


Table 5 Tabulation for Prediction Time

Number of patient data (number)	Prediction Time (ms)		
	NoSQL database Model	Data management methodology	Data Analytics and Visualization Framework
10	25	31	18
20	28	34	22
30	30	36	25
40	32	39	28
50	35	41	31
60	37	43	34
70	40	46	37
80	42	49	38
90	44	53	40
100	47	57	42

C. Model Prediction Time

Prediction time (PT) is measured as the amount of time consumed to predicting disease from patient data. It is evaluated in milliseconds (ms). The prediction time is mathematically formulated as,

Prediction TINe
= Ending tiNe
− Starting tiNE for predicting patient disease

Lower PT, the method is said to be more efficient.

The PT with number of patient data is portrayed in table 5. PT is compared with existing NoSQL DB Model, DMM and Data Analytics and Visualization Framework. The result provided in table 5 confirms that, the PTusing Data Analytics and Visualization Framework is lower as compared to NoSQL DB Model and DMM. The graphical representation of prediction timeis shown in figure 4.

From figure 4, prediction time based on the different number of patient data is described. From the figure 4, Data Analytics and Visualization Framework consume lesser prediction time than NoSQL DB Model and DMM. Research in DMMconsumes 13% lesser prediction time than NoSQL DB Model and consumes 27% lesser prediction time than Data Analytics and Visualization Framework.

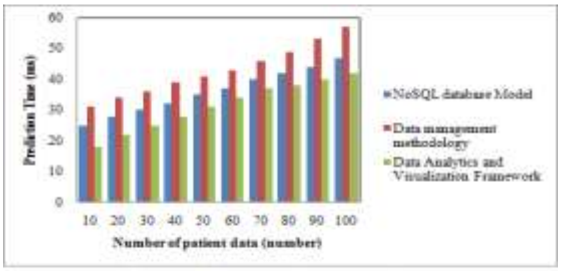


Figure 4 Measure of Prediction Time

Figure 4: measure of prediction time

V. DISCUSSION ON LIMITATION OF RESOURCE OPTIMIZATION AND PREDICTION TECHNIQUES WITH BIG HEALTHCARE DATA

The healthcare data storage exploits the model derived from NoSQL DB. But Sharding affects writing speed in DB and lessened the 'write' operations than prior state. 'Write' speed is better than SQL Server. In addition, SC remained unaddressed during healthcare data storage. DMM is presented for parallelism in bioinformatics workflows when reduce the data-interdependent file transfer. The designed methodology validates scalability and speed-up the process. However, it does not distribute data processing loads among threads in earlier stage. The Data gets accessed by several threads at any time without extra cost. The load balancing efficiency was not improved using DMM. FPLT Algorithm causes delay during the workflow optimization. The Data analytics and visualization framework was established with the huge-scale health informatics DS for health-shocks prediction. In rural and distantly accessible regions of Pakistan, a predictive model is employed for health-shocks and collected unique data. But, prediction accuracy was not improved using data analytics and visualization framework.

A. Future Direction

The future direction of resource optimization and prediction techniques for predicting the disease with minimum resource can be carried out with help of machine learning techniques and ensemble classification techniques.

VI. CONCLUSION

A comparison of different existing resource optimization and prediction techniques with big healthcare data is studied. From the study, it is observed that the existing techniques uses large amount of resources and enhanced prediction time of disease from patient data. Survival review shows that existing CNN-based multimodal disease risk prediction algorithm failed to predict disease in effective manner. In addition, high quality DS preparation consumed large amount of memory space for storing and accessing the information. The wide range of experiments on existing methods computes the performance of the many resource optimization and prediction techniques with its limitations. This survey is interesting direction for future researchers to attain practical solutions in real applications. Finally, from the review, the research work can be done via using machine learning and classification techniques to enhance disease prediction accuracy and minimizing the resource utilization.

REFERENCES

1. ZohrehGoli-Malekabady, Mohammad KazemAkbari, MortezaSargozaei-Javan, “An Effective Model for Store and Retrieve Big Health Data in Cloud Computing”, Computer Methods and Programs in Biomedicine, Elsevier, Volume 132,August 2016, pp 75-82
2. Athanassios M. Kintsakis, Fotis E. Psomopoulos and Pericles A. Mitkas, “Data aware optimization of bioinformatics workflows in hybrid clouds”, Journal of Big Data, Volume 3, Issue 20, 2016, pp 1-26.
3. Shahid Mahmud, RahatIqbal and Faiyaz Doctor, “Cloud enabled data analytics and visualization framework for health-shocks prediction”, Future Generation Computer Systems, Elsevier, Volume 65, December 2016, pp 169-181.
4. Inci M. Baytas, Kaixiang Lin, Fei Wang, Anil K. Jain, Life Fellow, and Jiayu Zhou “PHENOTREE: Interactive Visual Analytics for Hierarchical Phenotyping from Large-Scale Electronic Health Records”, IEEE Transactions on Multimedia, Volume 18, Issue 11, November 2016, pp 2257 – 2270.
5. Yichuan Wang and Nick Hajli, “Exploring the path to big data analytics success in healthcare”, Journal of Business Research, Elsevier, Volume 70, January 2017, pp 287-299.
6. Po-Yen Wu, Chih-Wen Cheng, Chanchala D. Kaddi, JananiVenugopalan, Ryan Hoffman, and May D. Wang, “-Omic and Electronic Health Records Big Data Analytics for Precision Medicine”, IEEE Transactions on Biomedical Engineering, Volume 64, Issue 2, February 2017, pp 263 – 273.
7. J.M. Lillo-Castellano, I. Mora-Jimenez, R. Santiago-Mozos, F. Chavarria-Asso, A. Cano-Gonzalez, A. Garcia-Alberola, and J.L. Rojo-Alvarez, “Symmetrical Compression Distance for Arrhythmia Discrimination in Cloud-Based Big-Data Services”, IEEE Journal of Biomedical and Health Informatics, Volume 19, Issue 4, July 2015, pp 1253 – 1263.
8. Andreas Schmidt, IgnasiForne and Axel Imhof, “Bioinformatic analysis of proteomics data”, BMC System Biology, Springer, Volume 8, Issue 2, March 2014, pp 1-7.
9. Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jorg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, Thomas J. Ashby and Hongming Chen, “ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics”, Journal of Cheminformatics, Springer, Volume 9, Issue 17, 2017, pp 1-9.
10. Ivo D. Dinov, “Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data”, GigaScience, Springer, Volume 5, Issue 12, 2016, pp 1-15.
11. Maggie P. Y. Lam, Edward Lau, Dominic C. M. Ng, Ding Wang and Peipei Ping, “Cardiovascular proteomics in the era of big data: experimental and computational advances”, Clinical Proteomics, Springer, Volume 13, Issue 23, 2016, pp1-14.
12. Min Chen, YixueHao, Kai Hwang, Lu Wang and Lin Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, IEEE Access, Volume 5, April 2017, pp 8869 – 8879.
13. *Mphrx*: http://mphrx.com/.

14. AislingO'Driscolla, J.D., Roy D. Sleator, "Big data Hadoop and cloud computing in genomics", Journal of Biomedical Informatics, Elsevier, Volume 49, Issue 5, 2013, pp 774–781.

15. Andrew V. Nguyen, Rob Wynden and Yao Sun, "HBase, MapReduce, and Integrated Data Visualization for clinical signaldata", Computational Physiology, Spring Symposium, 2011,pp 40-44.

16. Sven Van Poucke, Zhongheng Zhang, Martin Schmitz, Milan Vukicevic, Margot Vander Laenen, Leo Anthony Celi, Cathy De Deyne, "Scalable Predictive Analysis in Critically Ill Patients using a Visual Open Data Analysis Platform", PloS ONE, Volume 11, Issue 1, 2016, pp 1-21.

17. Sungyoung Oh, Jieun Cha, MyungkyuJi, Hyekyung Kang, Seok Kim, EunyoungHeo, Jong Soo Han, Hyunggoo Kang, HoseokChae, Hee Hwang, and SooyoungYoo, "Architecture design of healthcare software-as-a-service platform for cloud based clinical decision support service", Healthcare informatics research, Volume 21, Issue 2, 2015, pp 102-110.

18. Milan Vukicevic, SandroRadovanovic, Milos Milovanovic, and MiroslavMinovic, "Cloud Based Metalearning System for Predictive Modeling of Biomedical Data", The Scientific World Journal, Hindawi Publishing Corporation, Volume 2014, 2014, pp 1-10.

19. Parekh, M. and B. Saleena, "Designing a Cloud Based Framework for HealthCare System and Applying Clustering Techniques for Region Wise Diagnosis", Procedia Computer Science, Elsevier, Volume 50, 2015, pp 537-542.