# A MapReduce Based Distributed Framework for Similarity Search in Healthcare Big Data Environment

Hiren K.D. Sarma[1], Yogesh K. Dwivedi[2(✉)], Nripendra P. Rana[2], and Emma L. Slade[2]

[1] Department of Information Technology, Sikkim Manipal Institute of Technology, Rangpo Sikkim, India
hirenkdsarma@gmail.com
[2] School of Management, Swansea University, Swansea, SA2 8PP, UK
{y.k.dwivedi,n.p.rana}@swansea.ac.uk, emmalslade@hotmail.co.uk

**Abstract.** Similarity search in the big data environment is a challenging task. Patient Similarity search (PaSi) is an important issue in healthcare network and data. The results of PaSi search may be highly useful for drawing different conclusions and decisions to improve healthcare systems. Such findings can also be useful for choosing the treatment paths for new patients. In this paper, we propose a MapReduce based framework as a solution to the PaSi problem in the context of a healthcare network imagined to be implemented considering the healthcare centers of India. It is assumed that such a healthcare network will be implemented in future over the Government of India cloud known as GI cloud or 'MeghRaj'. The paper also discusses the associated implementation challenges of the proposed framework and the query handling approach for the proposed framework to solve the PaSi problem is stated. Finally, the paper outlines the future scope of the work.

**Keywords:** Big data · MapReduce · Similarity search · Patient similarity (PaSi) · Cloud · Framework

## 1 Introduction

In today's world, the volume of digital data generated by different information and communication technology (ICT) related applications, and other applications in the domain of meteorology, scientific instruments, healthcare or medical networks, etc., is enormous [1]. Such huge volume of data, which is generated largely over networks, is not practical to handle through classical database management approaches [3]. For these systems, capturing, storing, processing and retrieval of appropriate data in a timely manner are some extremely important issues. Centralized solutions to these problems are not suitable and distributed solutions have their own problems [2]. Some problems of distributed processing include network bottlenecks, requirements of global information locally, extra communication overheads, etc. Such applications have given birth to the concept of big data. Big data has attracted the attention of researchers and data scientists in recent times. Novel solution approaches are required to handle big data related issues [1].


M. Janssen et al. (Eds.): I3E 2015, LNCS 9373, pp. 173–182, 2015.
DOI: 10.1007/978-3-319-25013-7_14

Healthcare network is one example, which generates a huge volume of data every day. One of the processing issues connected to healthcare data is to find out Patient Similarity (PaSi). It is defined as the rate of similarity between two or more patients in terms of their symptoms, treatment procedures, personal information, etc. [1]. A typical PaSi solution will find out those patients who have the greatest amount of information in common. Then the treatment paths followed for such patients can be adapted for new patients. These data related to patients are stored in different databases of patient information systems maintained across healthcare networks. There could be several issues in solving PaSi. The data format used for different patients could be different. This is due to the lack of predefined record structure applicable to all patients. The volume of such data to be processed will be colossal. Moreover, some data related to patients can be uncertain and the data will be generated at a very high rate across the healthcare network. As a result, we need big data solutions to address the PaSi problem of healthcare networks. Hence, there is a need to think of some distributed and scalable solution approaches in order to address this problem. MapReduce is a tool that can be used to develop distributed and scalable solutions against big data problems [2, 3]. MapReduce has also been used to solve some healthcare problems [1].

'MeghRaj' is a cloud computing environment developed by the Government of India [4]. There is scope to implement a healthcare network connecting different health centers or hospitals spread across the country over this cloud. If such a system is implemented, it is going to generate big data. Thus, we need different big data solutions to handle different issues related to data processing and storage of these data.

In this paper, we consider a healthcare network that can be implemented over the 'MeghRaj' cloud and address the issue of finding PaSi. We propose a framework, which is based on MapReduce, to address the PaSi problem. We assume that the patient information will be stored in an unstructured manner. Even in the same machine or data source, two different patients' information can be differently structured. This framework is a proposal and its performance evaluation through simulation is undertaken by us.

The rest of the paper is organized as follows. Section 2 presents the background on big data followed by Sect. 3 in which related works are mentioned and the problem undertaken here is stated formally. In Sect. 4 the proposed framework is discussed and Sect. 5 describes the implementation challenges present in the proposed framework. Finally, Sect. 6 concludes the paper with an outline of the future scope of this work.

## 2 Background

MapReduce is a programming model and an associated implementation for processing and generating large datasets [3]. It is possible to handle big data through MapReduce, programmers find the system easy to use, and this parallel data processing tool has been made popular by Google. It is a scalable and fault-tolerant data processing tool that makes it possible to process a massive volume of data in parallel with the association of many low end computing systems [2]. Users specify the computation task at hand in terms of a map and a reduce function, and then the underlying runtime system processes the given task by distributing the computation tasks across large scale clusters of

computation nodes. This tool can handle machine (i.e. computation node) failures and can also make efficient use of network and disks by appropriate scheduling mechanisms. A decomposable algorithm, partitionable data, and sufficient small data partitions are required for effective use of MapReduce [6]. There are some enhancements to MapReduce. For example, in the work [5], classic MapReduce was optimized to decrease the data transformation load. A shared area for information was considered in this approach. Such an approach is suitable for solving problems like k-nn and top k queries. In [8], a method was developed to handle workloads in hierarchical MapReduce architecture. Haloop proposed in [7] is another type of MapReduce structure suitable for handling iterative problems. iMapreduce proposed in [10] also supports iterative processes. The work presented in [12] is aimed at reducing the amount of data transferred in the MapReduce network. Here, MPI (Message Passing Interface) was used for message passing in a MapReduce structure. The work presented in [11], replaces Hadoop File System (HDFS) with a concurrency optimized data storage layer. This layer is based on the BlobSeer data management service. It is essential to estimate the input/output (I/O) behavior of MapReduce applications and the work presented in [9] is a model that can be used to estimate I/O behavior of MapReduce applications.

In this section, we also establish the relationship of healthcare data with big data. If we look at the networked environment considering the hospitals across a country like India, then we visualize that patients' data will be generated at an exponential rate. These data will have different formats and standards. In healthcare networks, various data related to patients' health, diseases and recovery processes could be made available. Such data will be of great help for the treatment of other patients. Of course, for this to happen, there is a need of processing these data from different perspectives.

Big data is characterized by four 'Vs' namely volume, variety, velocity, and veracity [1]. Data generated through healthcare networks exhibit all the above four characteristics. In such systems huge volume of data is generated in various formats with a high velocity. Moreover, for many patients we get uncertain data in the data generated by healthcare networks and this fact leads to veracity of healthcare data. Thus all four Vs of big data are present in healthcare data. Therefore, big data solutions are required to solve different data processing problems of healthcare data.

Looking at the high volume of data in healthcare networks, big data solutions are necessary for data analysis [1]. According to [13], processing costs can be reduced by using big data analytics in healthcare. In the work presented in [14], problem like selection of appropriate treatment paths is addressed and solution for improvement of healthcare systems has been proposed. A scalable knowledge discovery platform for healthcare big data is proposed in [15].

## 3 Related Work and Problem Statement

Finding Patient Similarity (PaSi) is the major task considered in this paper. We consider a very specific healthcare network system, which is yet to be built but must be a reality in the near future. The system under consideration is the healthcare network of India. Different healthcare units i.e., hospitals spread across the country will be connected in a hierarchical

manner. One has to find out PaSi of two or more patients in terms of their symptoms, treatments, personal information, etc. The objective in PaSi is to identify those patients who have the greatest amount of information in common. Using the result of PaSi, new patients can be treated by following treatment processes adapted for those previous patients.

PaSi solutions can be found out by either of the two approaches as mentioned in [1]. First is the use of machine learning and data mining algorithms and the second being information retrieval by simple search or by entity-relationship graphs. A brief survey related to these two techniques can be found in [1].

In [1], a MapReduce based method for finding PaSi solution is proposed. The method is scalable and distributed - named as ScaDiPasi - takes small execution time and is implementable over big data related to healthcare networks. The experimental results reported in the paper show that the ScaDiPasi would be able to produce PaSi solutions over big data of healthcare networks.

### 3.1 Cloud of Government of India

With an aim of exploiting the benefits of cloud computing, the Government of India has initiated a very ambitious GI cloud project. This cloud has been named as 'MeghRaj' [4]. As mentioned in [4], the objectives of GI cloud are: optimum utilization of infrastructure; speeding up the development and deployment of e-governance applications; easy replication of successful applications across different states of the country to avoid duplication of effort and cost in development of similar applications; and, making the certified applications following common standards available in one place.

The GI cloud consists of multiple national and state clouds. A detailed discussion on the architecture of GI cloud and projects to be implemented under GI cloud can be found in [4].

This GI cloud infrastructure can definitely be used for implementing a healthcare network across the country. Such a healthcare network will generate healthcare big data. Any stakeholder of the healthcare network can share the benefits of processing such big data with different objectives within less time. Therefore, the advantages and benefits of such a system can be significant.

### 3.2 Problem Statement

Finding Patient Similarity (PaSi) is an important problem. As already discussed, there are different approaches for finding PaSi solutions. Interestingly, there exists no single solution applicable to all kinds of problems. Although there are several PaSi solutions already proposed for different situations, we believe, those will not be directly applicable to the big data to be generated by the cloud based healthcare network system of India that is being considered in the current research. Therefore, we need a novel solution to solve the PaSi problem of the healthcare network over GI cloud. Hence, the problem statement of this paper is:

> To design a big data based framework for addressing the Patient Similarity (PaSi) problem considering the future healthcare network that can be built over the GI cloud.

## 4 Proposed Framework

In this section, we provide a framework for similarity search related to the healthcare big data environment. The proposed framework is based on MapReduce [3]. We consider the cloud environment 'MeghRaj' [4]. It is assumed that a healthcare network considering all health centers of India spread across the country will be implemented over 'MeghRaj'. The structure of the healthcare network will be hierarchical as shown in Fig. 1. Different layers in this hierarchy are the hospitals at different levels such as (i) at panchayat level, (ii) at block level, (iii) at sub-division level, (iv) at district level, (v) at state level, (vi) at region level, and (vii) at national level. Although the type of hospitals can be categorized as elementary health center, public health center, medical college and hospital, private hospital, etc., at this stage we do not discriminate the hospitals based on type. We assume that all hospitals are similar with respect to generation of patients' data. We mainly need to work on the patient data irrespective of the facilities and infrastructure of the hospitals, which are generating such data.

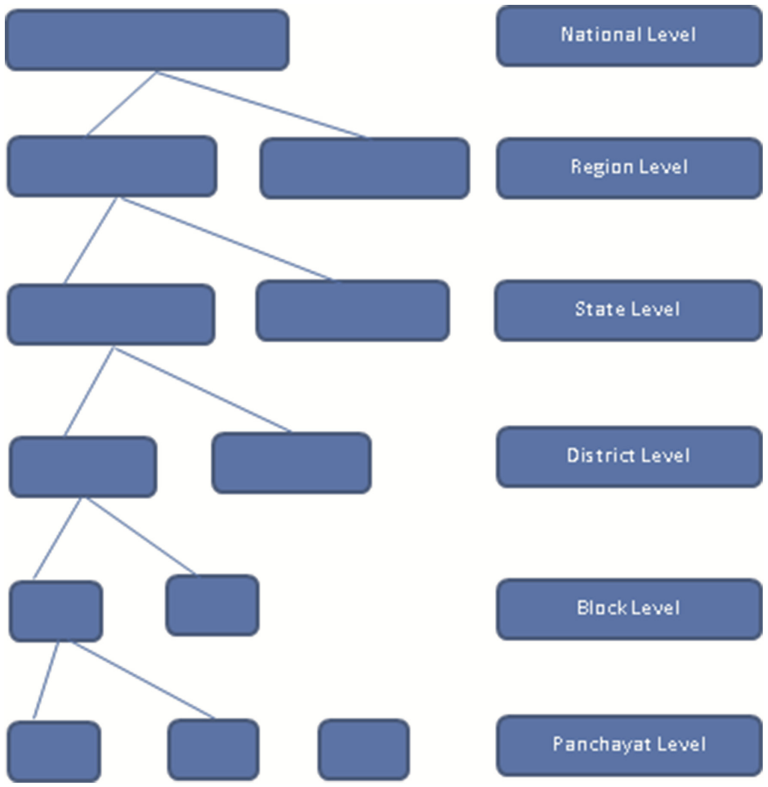


**Fig. 1.** Hierarchical organization of health centers

The data generated by such a healthcare network system will be stored in the cloud 'MeghRaj'. It is assumed that the healthcare network will be implemented over the Internet. As an end result, a legitimate user of the healthcare network should be able to throw a query related to similarity search problem to the healthcare network and should receive the result back from the network as quickly as possible. This scheme is shown in Fig. 2.

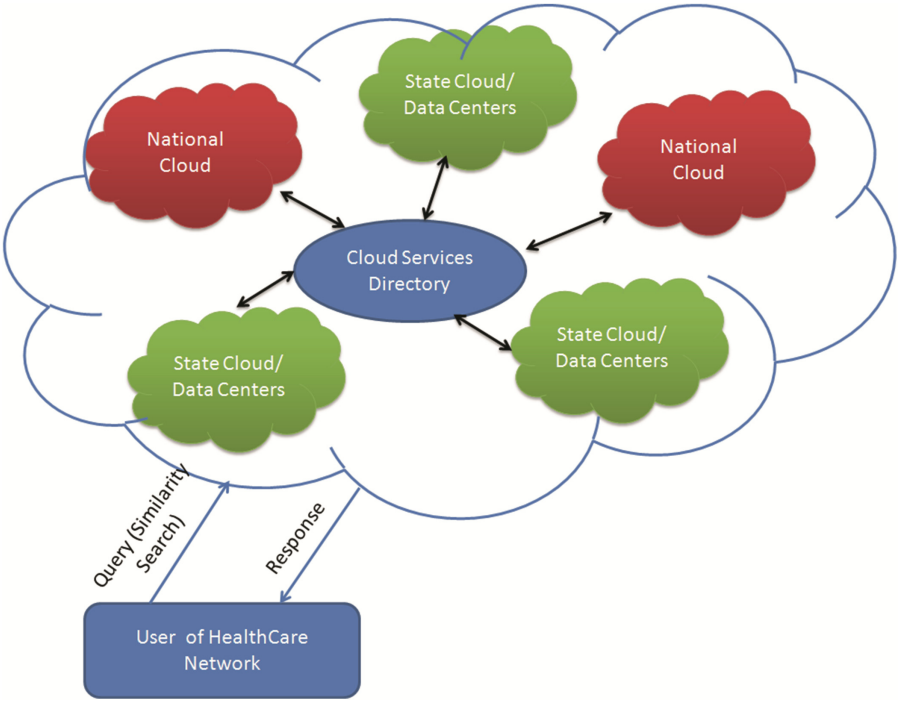


**Fig. 2.** Query sending to healthcare network to be implemented over MeghRaj cloud

### 4.1 How MapReduce Will Work

In order to optimize the solution process of similarity search problem, we propose a load-balancing module, which keeps track of distributed load among the computing nodes of the network. Moreover, this module tries to balance the computing load among the nodes.

As we assume that the data sources can be of heterogeneous nature, we propose to have a data format middleware, which brings different data formats to a homogenous common format before any kind of processing task takes place.

Bottlenecks due to maximum message exchanges in distributed processing are an issue, which is also unavoidable. We propose a bottleneck assessment and control module that takes care of the possibilities of bottleneck occurrence. In the presence of

a bottleneck in certain nodes, this module will divert the necessary workload to some other nodes so that the bottleneck can be controlled temporarily.

Data aggregation module will aggregate different data against different queries into some aggregated state. In an aggregated state the volume of data will be reduced significantly and intermediate code to represent data will be generated. This aggregated data in encoded form with reduced volume will be moving across the network reducing amount of data traffic in the network.

All these four modules mentioned above i.e., load balancing module, data format middleware, bottleneck assessment and control module, and data aggregation module, will be working outside the MapReduce environment. The output of these modules will be integrated with MapReduce for optimal performance of the entire system. This framework is depicted in Fig. 3.

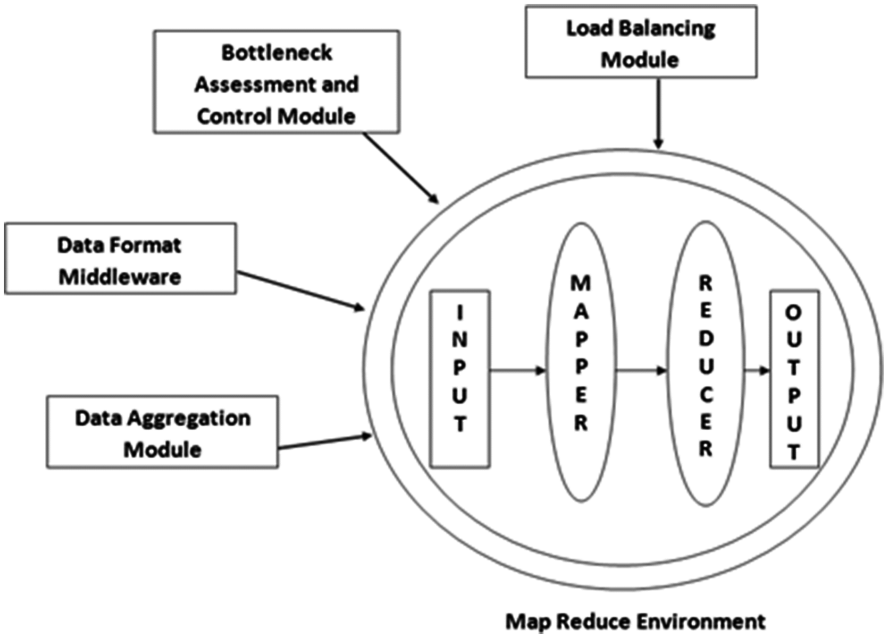


**Fig. 3.** Proposed MapReduce based framework for similarity search in healthcare big data

## 5 Implementation Challenges

The major challenges in total implementation of the proposed framework are as follows:

**Challenge Set 1.** Appropriate technique for load estimation and appropriate algorithm for load balancing are to be designed and associated theoretical complexity analysis is to be carried out. Proper task scheduling algorithm for load balancing is to be designed and analyzed.

**Challenge Set 2.** Data format middleware is to be designed, which will be highly specific to the structures of different databases present in the healthcare network.
**Challenge Set 3.** Proper algorithm for bottleneck assessment is to be designed. Moreover, bottleneck has to be controlled and this may lead to migration of processes or computation tasks from one node to another lightly loaded node. Thus there is a necessity to design proper process migration algorithm.
**Challenge Set 4.** Data aggregation algorithm considering the patient databases is to be designed and analyzed for its performance. Proper encoding mechanism has to be designed for it.

### 5.1 Solution for Similarity Search

The Patient Similarity (PaSi) search is the similarity search problem considered here. A user throws a query, and then this query is translated into an appropriate uniform format through an intermediate query building process. It works on the MapReduce framework. We propose to have three phases of query processing to solve the PaSi problem. In each phase, Mapper and Reducer functions are to be implemented along with a Ranker function. The Ranker function at each phase evaluates the similarity level of the output of each phase with the input query. This is proposed to be a dynamic decision regarding forwarding of the output of the previous phase to the next phase for further processing in search of more similar results. A threshold level of similarity can be set as per the wish of the user with respect to his/her query. The second phase will continue to be executed until the similarity equates or exceeds the threshold level set by the user. This evaluation is carried out by the Ranker function to be implemented in each phase and a decision is made dynamically regarding the forwarding of the processing task to the next phase. This scheme is shown diagrammatically in Fig. 4.

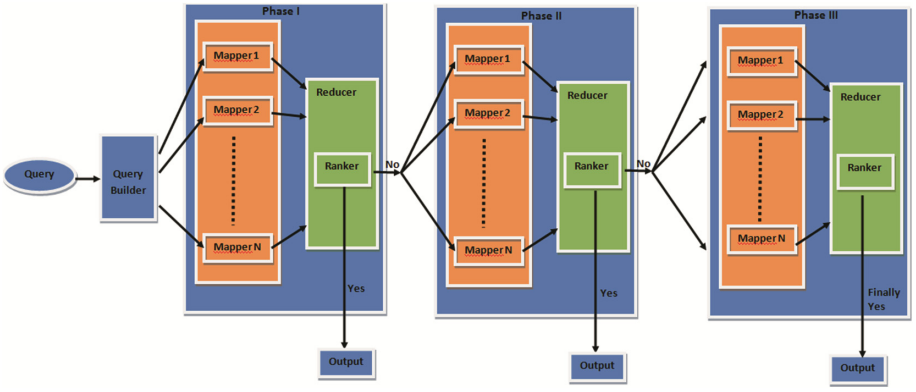


**Fig. 4.** Proposed query processing model based on MapReduce

## 6 Conclusion

In this work we address the problem of similarity search in a healthcare network using 'big data'. We focus on the Patient Similarity search popularly known as PaSi, and propose a framework for addressing the PaSi problem in healthcare data. Implementation issues are discussed thoroughly and a MapReduce based model of query handling is also proposed. The proposed framework and the query handling model are designed considering the Government of India cloud also known as 'MeghRaj'. It is assumed that a healthcare network will be implemented considering all the hospitals spread across India and will be deployed over 'MeghRaj' cloud. As for future scope of this work, it is noteworthy that various algorithms can be designed to address the implementation challenges outlined in Sect. 5. Moreover, the query handling model can be implemented over MapReduce framework considering some suitable patients database and various performance parameters, like execution time and accuracy, can be measured and analyzed.

## References

1. Barkhordari, M., Niamanesh, M.: ScaDiPaSi: an effective scalable and distributable MapReduce-based method to find patient similarity on huge healthcare networks. Big Data Res. **2**(1), 19–27 (2015)
2. Lee, K.H., Lee, Y.J., Choi, H., Chung, Y.D., Moon, B.: Parallel data processing with MapReduce: a survey. AcM sIGMoD Rec. **40**(4), 11–20 (2012)
3. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
4. GI Cloud Initiative (2015). http://deity.gov.in/content/gi-cloud-initiative-meghraj
5. Ding, L., Xin, J., Wang, G., Huang, S.: ComMapReduce: an improvement of MapReduce with lightweight communication mechanisms. In: Lee, S.-G., Peng, Z., Zhou, X., Moon, Y.-S., Unland, R., Yoo, J. (eds.) DASFAA 2012, Part II. LNCS, vol. 7239, pp. 150–168. Springer, Heidelberg (2012)
6. Highland, F., Stephenson, J.: Fitting the problem to the paradigm: algorithm characteristics required for effective use of MapReduce. Procedia Comput. Sci. **12**, 212–217 (2012)
7. Bu, Y., Howe, B., Balazinska, M., Ernst, M.D.: HaLoop: efficient iterative data processing on large clusters. Proc. VLDB Endowment **3**(1–2), 285–296 (2010)
8. Martha, V.S., Zhao, W., Xu, X.: h-MapReduce: a framework for workload balancing in MapReduce. In: 27th International Conference on IEEE Advanced Information Networking and Applications (AINA), pp. 637–644 (2013)
9. Groot, S.: Modeling I/O interference in data intensive Map-Reduce applications. In: 12th International Symposium on IEEE/IPSJ Applications and the Internet (SAINT), pp. 206–209 (2012)
10. Zhang, Y., Gao, Q., Gao, L., Wang, C.: Imapreduce: a distributed computing framework for iterative computation. J. Grid Comput. **10**(1), 47–68 (2012)
11. Nicolae, B., Moise, D., Antoniu, G., Bougé, L., Dorier, M.: BlobSeer: bringing high throughput under heavy concurrency to hadoop Map-Reduce applications. In: 2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS), pp. 1–11 (2010)
12. Mohamed, H., Marchand-Maillet, S.: MRO-MPI: MapReduce overlapping using MPI and an optimized data exchange policy. Parallel Comput. **39**(12), 851–866 (2013)

13. Srinivasan, U., Arunasalam, B.: Leveraging big data analytics to reduce healthcare costs. IT Prof. **15**(6), 21–28 (2013)
14. Jee, K., Kim, G.H.: Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc. Inf. Res. **19**(2), 79–85 (2013)
15. Metaxas, O., Dimitropoulos, H., Ioannidis, Y.: AITION: a scalable KDD platform for Big Data Healthcare. In: 2014 International Conference on IEEE-EMBS Biomedical and Health Informatics (BHI), pp. 601–604 (2014)