ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Performance Evaluation Of Map Reduce Work Loads On Medical Dataset.

R Tamilarasi*, M Indu Maheswari.

Department of Information Technology, Sathyabama University, India.

ABSTRACT

On large datasets the parallel communications will be supported by the Map Reduce, which is a very popular and simple programming model. For easy interface, Map tasks number of applications from several areas like machine learning, graph processing and mining which are suitable. The breaking of input into minor sub-problems and producing solution for every smaller sub-problem is completed by the simple interface Map tasks. The results of map tasks are processed by the reduce tasks and these are combined into a solution for a given exact problem. Every reduce task should regain and should process the outcome that are formed by all the maps. In order to calculate the average performance number of applications in various situations, one of the method called MVA (Mean Value Analysis) is been applied. The reduce tasks related to the similar Map Reduce job and synchronization with map have precedence constraints workloads and hence the MVA cannot be applied directly to this. The exact alternative solutions in which queuing network model is used to calculate the transition rates among states and which together exploit Markov Chains, to characterize the feasible states of the system are also obtainable. As the state space develops exponentially with the numerous tasks, the approaches will not be balanced properly. The analysis of disease can be compared in future by the means of the admin which will manage every report of the patient and drug validation which is provided to the patient by it, since the patient will not have the knowledge about the prescription like whether the medicine which is prescribed is perfect or not. The users will give their diagnostic reports and symptoms to the system. Hence diagnosis will be done with two steps. They are a) patient login b) patient signup. They deal with the process of working of the intra job works in pipeline parallelism which analyze the delays of synchronization and mean value for the HADOOP work by mean value analysis algorithm. Naïve Bayesian classifier which is the machine learning algorithm is proceeded further to overcome the existing approach for the purpose of dataset of lung cancer and predicts analysis of drugs used, feedback of various patients and the resultant output.

Keywords: Admin tasks, MVA mean value analysis and machine learning algorithm.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

The powers of the huge clusters of computers are tied together by a very popular way which is emerged by the Map Reduce. By map reduce the programmers will assume in a data-centric fashion and they focus on permitting the information of distributed execution, fault tolerance, network communication that are to be handled by framework of Map reduce and applying transformations to sets of the data report. The Map Reduce is applied on the broad domain of pipelining problems. For applications like steam processing [1] and event monitoring the map reduce is been used. A run-time system which is associated, and a programming representation for processing the big scale data are the relatively young frameworks of map reduce. HADOOP is the very familiar open-source implementation of a Map Reduce framework which follows the design that is laid out in the original document. The combinations of features that are related to increasing popularity of the Hadoop are data-local scheduling, handling of straggler tasks, customizable and modular architecture [2], fault tolerance and capability to work in a varied environment. Though the performance is noticed to be suboptimal inside the database context, it is familiar for its regained fault tolerance and elastic scalability. An open source implementation of MapReduce that is HADOOP is slower than two state-of-the-art parallel database systems that perform a several range of analytical. The improved performance can be achieved by the Map Reduce by the means of allocating several number of compute nodes obtained from the cloud and this helps to fasten up the computation. Hence in pay-as-you-go environment the 'renting of more nodes' approach is not a cost elective one. There is a kind of interest in checking whether both efficiency and the elastic scalability can be provided by the Map Reduce and also elastically scalable system of data processing that are economical are desired by the users. By the means of block-level scheduling the Map Reduce can achieve the elastic scalability. The datasets of input are splitted in to data blocks of even sized by the runtime system routinely. The data blocks are scheduled to the obtainable compute nodes dynamically for processing. In real systems [3] the Map reduce is confirmed to be extremely scalable. The simple computations that are to be performed are expressed by design of a new abstraction as a response for this complexity. But the disorganized details of data distribution, fault-tolerance, load balancing in a library and parallelization are hidden by it. The several functional languages and the primitives in the Lisp are reduced and the abstraction is motivated by the Map. We realize that the majority of our computations are involved by applying a map operation to every logical proof. The huge computations can be parallelized easily and the main mechanisms for fault tolerance [4] can be re-executed by the usage of the functional representation with reduce operations and user spiced map. The main aim of this review is to offer a timely remark on the status of the related work of present research which is aimed at enhancing and improving the Map Reduce frame work and status of the Map Reduce studies. An overview is given about the main approaches and they are classified according to their strategies.

RELATED WORK

In order to accomplish the several previous challenging tasks the suitable work loads are used by the cluster operators of the Map Reduce and further two latest capabilities are demonstrate here. The computational dimension or workload growth is anticipated by operators. Instead of having the uniform configurations that are optimized for a normal case which does not exist, the operators can select extremely specific configurations optimized for various kind of jobs inside a workload. The impact of consolidating various workloads into the same cluster can also be expected by the operator scan. By means of the vocabulary of workload description it is been introduced that operators can quantify the superposition of several workloads across many workload characteristics **[**6**]** systematically. The runtime scheduling scheme is adopted by the Map Reduce. In order to process one at a time the scheduler assigns data blocks to the available nodes. Hence the runtime cost is introduced by the scheduling strategy and slows down the Mara-duce job execution whereas on the other side, the parallel database system is bent from a compiling-time scheduling plan. The query optimizer produces a spread query plan for every available node whenever a query is submitted. At the time of executing the query of each node will know its logic of processing as per the distributed query plan. After the production of the query plan that is distributed, the scheduling cost is not introduced. The authors noticed that run-time scheduling strategy of Map Reducer is extremely costly than the compiling-time scheduling of DBMS. The runtime scheduling strategy enables the Map Reduce to excess of elastic scalability, specifically the capability of adjusting resources dynamically during the job execution [7]. The overall efficiency and the performance of every application which are running on the top of the system is understood by the system of Map reduce. In order to evaluate the performance of the system the present users of Map reduce system should run the benchmarks in the system. Unless the system is built, the latest hypothetical system cannot be evaluated. The most achievable system configuration before committed to an

ISSN: 0975-8585

optimal solution becomes more difficult since the scale of the system keeps on becoming larger. In various cases, the incapability to evaluate the hypothetical system will prevent the design innovation in frameworks and systems [8].

OVER VIEW OF MAP REDUCE WORK LOADS

MVA are not supposed to be applied directly to workloads as they have priority constraint, such as reduce tasks related to the similar Map Reduce job and synchronization with map. The exact alternative solutions, which together exploit Markov Chains, are to calculate the transition rates among states and to signify the possible system state, and also to represent the models of queuing network. As the state space develops exponentially with the numerous tasks, the approaches will not be balanced properly. The analysis of disease can be compared in future by the means of the admin which will manage every report of the patient and drug provide to the patient by it since the patient will not have the knowledge about the prescription like whether the medicine which is prescribed is perfect or not. The below figure explains that the symptoms obtained from the patient is used to extract the disease which is related to the symptoms and also offer a particular dataset called as lung dataset. The lung disease symptoms and drugs details are stored in the health care management administrator.

OVERALL ARCHITECTURE

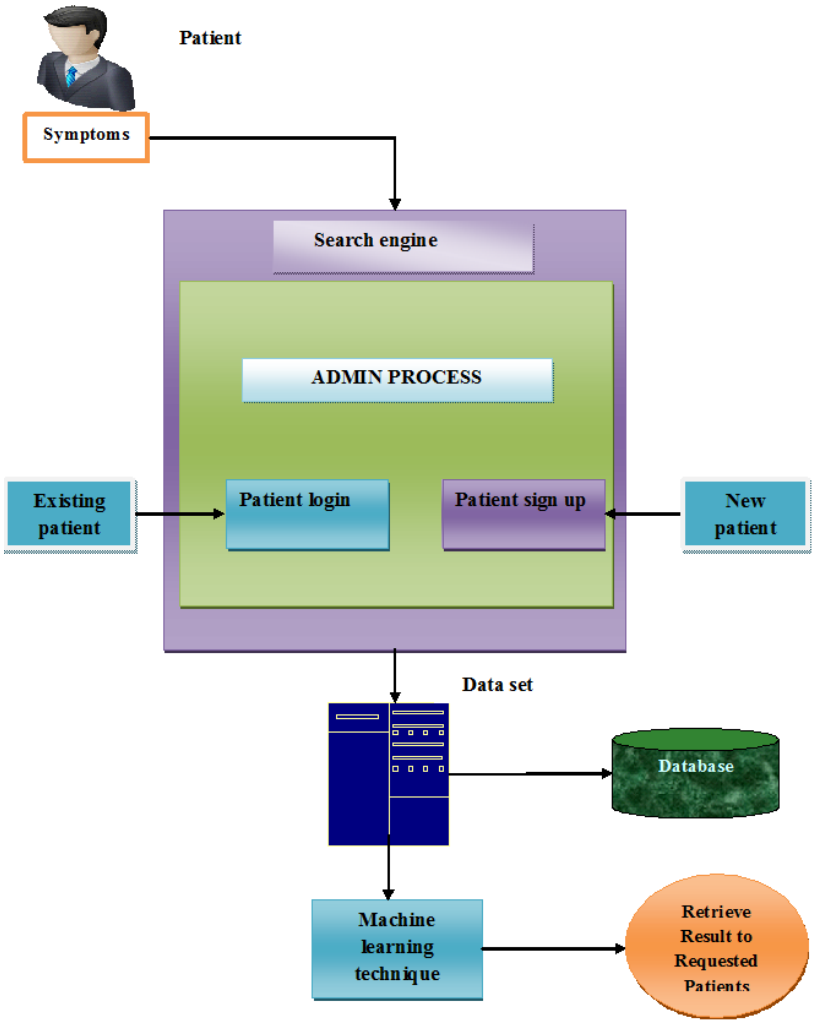

Figure 1 Over All Working Process Of Map Reduce Work Loads Via Medical Applications.

The dataset contain patient details and lung disease related details and drugs details. The patients login healthcare administrator to access. Using machine learning algorithm a patient can easily search the disease symptoms and get accurate and more suitable drug for that disease.

WORKING METHODOLOGIES

Admin process

Admin plays a major role. Once the admin activity is completed; the details of the patient will be accessed so that patient can know about his/her disease. All the reports of the patient will be managed by the admin for comparing the disease analysis in future and to offer the drug for the patient.

Lung dataset

The lung dataset contain all information about lungs. The lung disease symptoms like which type of lung disease and drugs detail are stored in the lung dataset. Based on given patient symptoms compare with dataset and recognize then suggest exact matching drugs to the patient.

Machine learning technique

The 'machine learning algorithm' predicts the Formulating Questions, analysis and gathers the Patient Health Condition and the output result. Machine learning technique of naïve Bayesian classifier compares the patient formulating questions with dataset and produce matching result.

NAIVE BAYESIAN CLASSIFIER ALGORITHM

Naive Bayesian classifier is used for prediction or classification. And it is generality, robust and simple this procedure organizes for different application like classification of land soils in agricultural, detection of material damage, machine learning applications and classification of web application. Naïve Bayesian classifies the dataset and minimizes the computational cost.

Input:

Training Data Set D with their associated class labels

Output:

Classification of Groups.

1. Training Set D, Initialize X with one component.
2. If R (B_i Y)> R(B_j/Y) for all k≤ j≤ m ; j ≠i
Maximize R(B_i/Y)
3. Compute
R (B_i/Y) =$\frac{P\left(\frac{Y}{Bi}\right)R(Bi)}{R(Y)}$
4. R(Y/B_i)R(B_i) need be maximized.
5. R(Y/B)= $\prod_{z=1}^{n} R(Y_k/B_i)$=R($Y_1$/$B_i$) X R($Y_2$/$B_i$)X……R(Yn/$B_i$) value of attribute A_z , for Dataset Y
6. If (A_z = categorical) then R(Y /B_i) / R(B_i)
Else R(Y/B $_i$) = g(Y_z, μB_i, ρB_i)
7. To predicate the class label Y, R(Y/B_i)R(B_i)
R(Y/B_i) R (B_i) > R(Y/Bj) R(Bj) for all k≤ l ≤ m ; j≠ i
8. Output the classifier

ISSN: 0975-8585

RESULT AND DISCUSSION

A series of experiments on the below particular dataset are conducted for evaluating the proposed approach performance. The proposed methods are evaluated and implemented in the following configuration in these experiments.

S.NO	Requirements
1.	Operating System: windows 7/8
2.	Processor: Intel core i5
3.	RAM: 2-4 GB
4.	Hard Disk Drive: 500 GB
5.	JAVA-JDK.1.7.0_17, Eclipse, Apache Tomcat-7.0.47
6.	MySQL, Apache Hadoop-2.3.0

Table 1 Performance Requirements

EXPERIMENTAL RESULTS

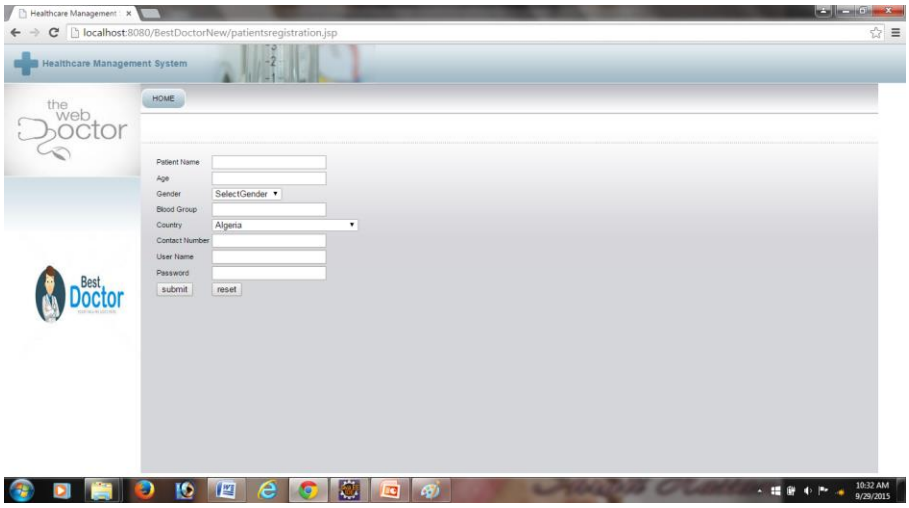

Figure 2 Patient Sign up and Login Page

In Patient Sign up and Login Page, If the patient is a new user they have to signup by giving their details and if the patient is an existing user then he/she will login with the username and password to access further pages.

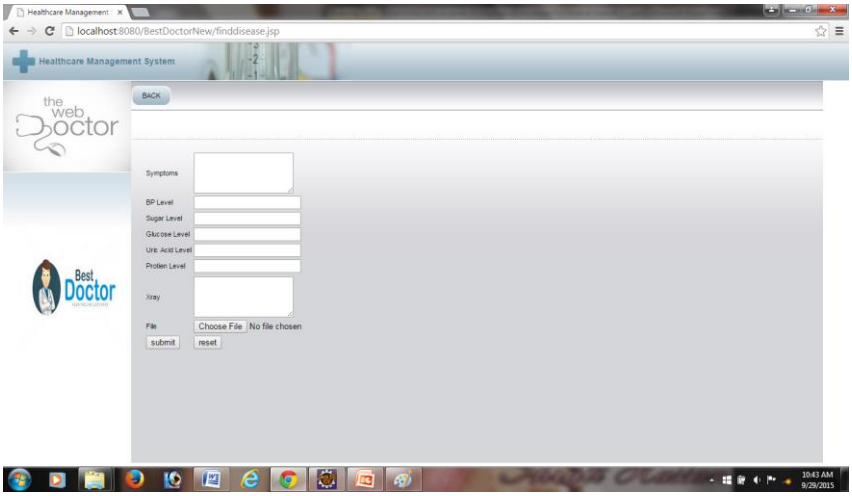

Figure 3 Update patient symptoms and find disease

ISSN: 0975-8585

Figure 3 represent the page where the patients need to upload the symptoms like BP Level, Sugar Level, Glucose level, Uric acid level, Protein level, X-rays etc. to find the disease.

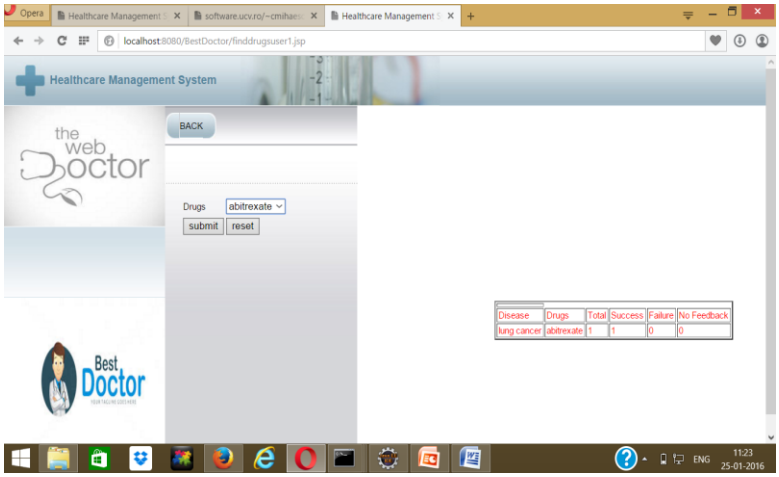

Figure 4 Suggest drugs to patient

Once the symptoms are updated, next the disease for the particular symptoms will be displayed and drugs for the particular disease will be suggested from the dataset.

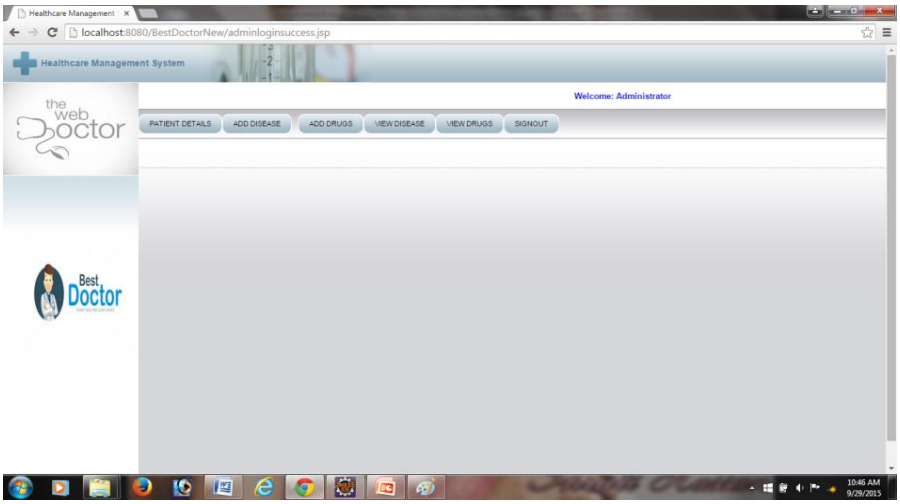

Figure 5 Administrator pages

In administrator page, the administrator will login and can views the patient details; disease details and drug details. Incase the symptoms are not available in the dataset; the administrator will add disease for the patient symptoms and drugs for the disease.

PERFORMANCE EVALUATION

Figure 4 presents the performance evaluation of existing technique and proposed technique. The reliability, efficiency and security are compared in this graph. When compare to Mean Value Analysis the Machine Learning Algorithm is more reliable, high efficiency and security.

ISSN: 0975-8585

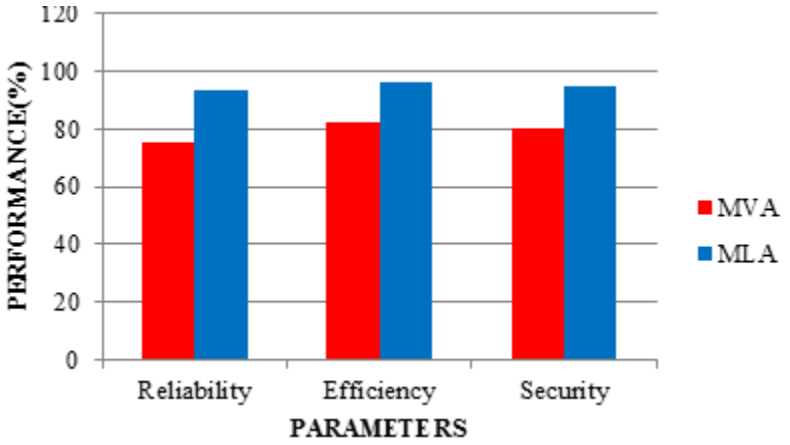

Figure 4 Performance evaluation of existing and proposed technique

QUERY RETRIVAL TIME

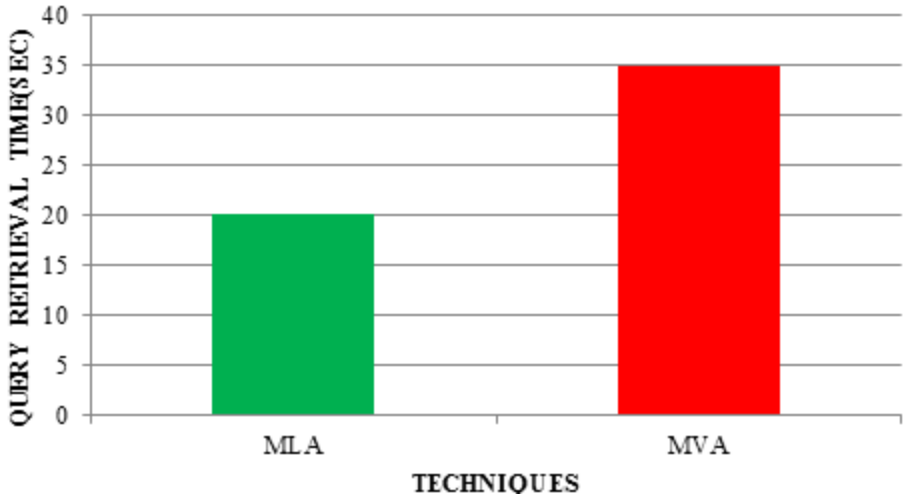

Figure 5 Query Retrieval Time

Figure 5 presents the query retrieval time of existing technique and proposed technique. When the patient enters a query in to search engine MVA takes more time to retrieve the drugs for the query given when compared the MLA tales less time to retrieve the drugs.

CONCLUSION

This paper proposes machine learning technique to provide immediate answer for a patient according to the patient question. The patient searches the lung related questions through search engine and the search engine provides the required result to the patient from the dataset where the expert have provided the solutions for the health seeker questions and suggested some drugs for the disease in the trained dataset. The expert post entire lung related problems and drugs for that particular disease. In future enhancement we try to develop by uploading lung image with query and upload normal lung image to the dataset. The image which is uploaded by patient is compared with normal image in the dataset and gives exact output.

REFERENCES

[1] Tyson Condie, Neil Conway, Peter Alvaro, Joseph M.Hellerstein, "MapReduce Online", Russell SearsYahoo! Research.

[2] Herodotos, "Hadoop Performance Models Technical Report", CS-2011-05Computer Science Department, Duke University.

[3] Jeffrey Dean and Sanjay, "MapReduce: Simplified Data Processing on Large Clusters"

[4] Kyong-HaLeeYoon-JoonLee, “Parallel Data Processing with Map Reduce: A Survey”
[5] Yanpei Chen, Archana Ganapathi_, Rean Griffith, Randy Katz, “The Case for Evaluating Map Reduce Performance Using Workload Suites”.
[6] Dawei Jiang Beng Chin Ooi Lei Shi, “The Performance of Map Reduce: An Indepth Study”.
[7] Guanying Wang, “Evaluating MapReduce System Performance: A Simulation Approach”, Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree