Enhancing Big Data Analysis by Using Map-reduce Technique

Alaa Hussein Al-Hamami*, Ali Adel Flayyih
Faculty of Computer Science and Informatics, Amman Arab University, Amman, Jordan

Article Info

Article history:

Received Nov 29, 2017
Revised Jan 30, 2018
Accepted Dec 13, 2018

Keywords:

Big data
EEg data
Hadoop
Cloud-wave
NoSQL

ABSTRACT

Database is defined as a set of data that is organized and distributed in a manner that permits the user to access the data being stored in an easy and more convenient manner. However, in the era of big-data the traditional methods of data analytics may not be able to manage and process the large amount of data. In order to develop an efficient way of handling big-data, this work enhances the use of Map-Reduce technique to handle big-data distributed on the cloud. This approach was evaluated using Hadoop server and applied on Electroencephalogram (EEG) Big-data as a case study. The proposed approach showed clear enhancement on managing and processing the EEG Big-data with average of 50% reduction on response time. The obtained results provide EEG researchers and specialist with an easy and fast method of handling the EEG big data.

Corresponding Author:

Ala’a Al-Hamami,
Faculty of Computer Science and Informatics,
Amman Arab University,
Jordan Street–Mubis-P.O Box. 2234-Amman 11953, Jordan.
Email: a.alhamami@psut.edu.jo

1. INTRODUCTION

Using Hadoop in Cloud-computing as an environment for this kind of applications is, so efficient for-at least- four reasons. These reasons are the following: (a) the highly fault tolerance it has, (b) the automated data distributed it performs with balancing of the computation load across different nodes it performs, (c) parallel computation property it has and (d) as close as possible the computation location from data position property it has that reflects in network overhead of transferring [1].

Vast developing in technologies (1-huge data collection, 2-powerful multiprocessor computing (dual core, quad core) 3-data mining algorithm) will support Data Mining in business application community [2]. Big data is use to illustrate massive datasets consisting of 4-V definitions: Volume, Velocity, Variety and Value (such as electronic medical records, biomedical image & signal and biometrics data) [3].

Electroencephalogram (EEG) data is a kind of biomedical signal data sets and clinical Big-Data. EEG is a test that is use to evaluate and record the electrical activity of the brain. EEG is widely used in the diagnosis and analysis of critical diseases. Electrophysiological data is another domain, where Big Data implemented and contains approximately 100 multi-channel signals. With records obtained from each patient generating at 5 to 10 Gigabytes (GB) data and by utilizing standalone tools such as Markonis [4] were found to be ineffective to meet the growing demand of data and needs to update multi center collaborative studies with real time and interactive access [5]. In this paper, Hadoop engine will be used to conduct the Map-Reduce processes, to process EEG Big-data, where the Map coverts the data to list with indexing and thereby, makes the comparison operations among values much easier than it used to be before. In the Reduce step, the programmer or the developer can choose the data that they are interested in, so that this will minimize the amount of data that we have and focus on the data that is of their main interest.

2. RESEARCH METHOD

In order to enhance the efficiency of analyzing EEG Big-Data for more understanding and ease of studying patient cases, EEG Big-Data needs to implement along with Hadoop by using Map-Reduce technique. The use of Map-Reduce and Hadoop on distributed systems, in the Cloud Computing environment can contribute to the significant advance in clinical Big-Data processing and utilization. In addition, this will offer new opportunities in the emerging era of Big-Data analysis and enhance the outcome of clinical EEG Big-Data analytical tools.

To use the proposed method (EMRT), we follow the following steps:

a. Store the datasets (EEG-Text-Files) in identified folder (input folder) that will be the root for programmers.
b. The input folder is located in a network and called H. Work that it considers as an environment path for Hadoop folder if it downloads under test environment, which contains the required data for programmer. It is possible to change the path through required configuration.
c. The next important step is to create Database from folder, which it has too, many field that separated by columns and every line represents a record. To run Map-Reduce, which must to have same structure and data type that every Database (SQL server, Image, Text file, Oracle etc) must have specific Map-Reduce.
d. Then converting these records to list, the main feature in the list, it has an index that enable programmer to choose interests columns of folder that will organize the data easily. In this paper the first value patient's number ID, Age, Time turnoff, signal analysis.

The Map has constant steps that aim to convert the text to a list. Then, by using the Hadoop commands, the different functions will be the reference for these operations is the Map. The functions of the Hadoop automatically create the Map, then the procedures entered that is wanted, and thereby obtaining the data required. Then, these data are export to the output folder. The output folder need to create as well as the input folder before the Java classes run which encapsulated into Jar file.

In this paper, the four columns of EEG-data are taken (ID, Time turnoff, Age, Signal analysis) the first column will be considered as the key and the rest columns are considered as the values. Thus, the comparison will be done on the values which be interested so that obtain the key. The key starts from zero in the index until it gets to the last value in the record. Through Java commands Read of Line (ROL) and End Of File (EOF) the problem of number of lines has been solved by ROL to make looping on list and another command for moving from one file to another [6]. The Hadoop automatically creates a folder that has a value in it but the programmer needs to provide the name of the input and output folder to the Hadoop. In addition, when exporting the Java file, the name of the input and the output folder will have needed for providing. Thus, the user sends the folder to the Hadoop, where the programmer sends an empty folder and specifies its place (could be on another server). After conducting the Map and Reduce processes, the values will have sent automatically to the output folder [7]. The programmer needs to specify the path with the ID address and other information. Thus, the function of the Map is to convert the data to list and the function of the Reduce is optional according the programmer’s mains interest.

The programmer reads the text file line by line and takes the column in it by choosing the data type, where the Hadoop is flexible in conducting these operations. The string differs from the text in that the text is larger than the string. The Map takes the text that originated from pictures and converts it to a list and the list takes the text as objects, where Map can store any type of data in it. The load collects data that turns the light off. It could give extra information on which once turned the light first and which one turned the light last, where the sensor contributes to the function of turning off the light. The main class, which is the public class, the static class that is branched, Mapper class and Reduces class from Java.

3. MAPPER FUNCTION

The EEG files will split by Job-Tracker in the number of blocks and each block will have processed by one Task-Tracker. Java has a feature read of line to select the interested column and in the research paper has chosen the ID of patient, Age, time turnoff and signal analysis. Consequently, the four columns of selecting will be transformed in the indexing list from another side the size of data is compressed that will decrease the response time and efficient retrieval will get.

4. IMPLEMENTAION AND EXPERIMENT EVALUATION

The first step of building a Map is to determine the type of the file in the form of a text file that can read and by moving the cursor to the end of line. Then, the programmer needs to determine the values that needed by taking them from the columns, where the columns need to be constant to build Map on it. After moving to the Java step, the comparison of time done in the form of a text, where text converted to time automatically. The text will have converted to time and the time will have converted to list, the Reduce deals with two texts and take back two other texts. In addition to the Java, files (import files); it made a configuration on the imported files inside the Hadoop. The final step is to generate the output file, which has all the values that have specified by the user. As shown from Table 1, the EMRT using Map-Reduce technique on Hadoop and distributing the big data on cloud.

Table 1. Comparison Table

Approach	Response time (s)	Accuracy
Schultz, 2013	0.71	80.6%
Mohammed et al., 2014	0.99	70.1%
Wang et al., 2014	1.01	93.0%
Markonis et al., 2015	0.69	68.7%
Proposed	0.59	96.5%

After gaining the required EEG text files, HDFS split these files across multiple of computer (nodes) in Cloud-Computing. Then Hadoop runs Map-Reduce model for processing EEG-Big- Data in real time and return back the result to user. The features of Hadoop reliable, fault tolerance by cloning the block data on three other nodes, scalable, parallel computing and high throughput access for files that make Hadoop so efficient for managing large Data sets.

At the last step, the EEG-Data that contains the filtration values will have sent to output folder that resides into HDFS and then will retrieved by client then will be loaded in the client's device. The final form of EEG list has ready for taking a decision by medical specialists depending on the result that have been analyzed in typical response time, accuracy analysis, and without redundancy as in Table 2.

Table 2. Experiment Result

Patient ID	Proposed Approach Response Time (s)	Hit	Miss	Old Data Structure Response Time (s)	Hit	Miss
1	*0.38*	√		2.4	√	
3	*0.24*	√		2.33	√	
5	*0.35*	√		2.35	√	
5	*1.22*	√		3.19	√	
7	*0.36*	√		2.36		√
7	*0.41*	√		2.44	√	
11	*1.14*	√		2.62	√	
12	*0.5*	√		2.65		√
12	*1.03*	√		2.68	√	
17	*0.37*	√		2.72	√	
19	*0.57*	√		2.75		√
19	*1.5*	√		2.78	√	
23	*0.32*		√	2.81	√	
23	*0.51*	√		2.85	√	
25	*0.32*	√		2.88	√	
26	*0.2*	√		2.91	√	
30	*0.2*	√		2.94	√	
32	*0.48*	√		2.97	√	
42	*0.22*	√		3.01		√
42	*1.3*	√		3.04	√	
43	*0.47*	√		3.07	√	
43	*1.3*	√		3.1	√	
45	*0.26*	√		3.14	√	
45	*0.34*	√		3.17		√
46	*0.3*	√		3.2	√	
46	*0.38*	√		3.23	√	
61	*0.2*	√		3.27	√	
61	*1*	√		3.3	√	
76	*1.29*	√		3.33	√	
Average	*0.59*			*2.87*		
Hit Ratio		*96.55%*			*86.20%*	

5. CONCLUSION

The previous results show clear enhancements on the response time and accuracy of the retrieved data, with average overall enhancement of 50%, which reflect on the performance of big-data management and process.

Table 2 shows a comparison of the results between previous used techniques and EMRT, by applying the same used dataset. EMRT using Map-Reduce technique on Hadoop and distributing the big data on cloud. It showed clear enhanced performance over previous related works, which will definitely reflect on the efforts and output of the clinical researchers and experts. Moreover, this enhanced results, will make it easy for the global societies to adopt the (IoT) concept.

a. The use of cloud computing is useful and effective regarding the cost and efforts.
b. Map-Reduce technique is more efficient in big data management and processing than traditional data structure techniques.
c. Clinical data, especially EEG data should be distributed on the cloud, for more reliability and ease of use.
d. The EEG data management should use Map-Reduce on Hadoop in order to make it easy and efficient for researchers and experts to retrieve their reed information, and do their studies.

REFERENCES

[1] Schultz, T. (2013). Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle. *Bulletin of the American Society for Information Science and Technology*, 39(5), PP: 34-40.

[2] Al-Hamami A H, Mohammad A Al-Hamami, and Soukaena H Hashem, *"A Proposed Technique for Medical Diagnosis Using Data Mining"*, International Conference on intelligent computing and Information systems, CICIS, March 19-22, 2009, Cairo, Egypt. http://icicis.edu.eg

[3] Bigdely-Shamlo, N., Makeig, S., & Robbins, K. A. (2016). Preparing laboratory and real-world EEG data for large-scale analysis: A containerized approach. Frontiers in neuroinformatics, 10.

[4] Markonis, D., Schaer, R., Eggel, I., Müller, H., &Depeursinge, A. (2015). Using MapReduce for large-scale medical image analysis. arXiv preprint arXiv:1510.06937.

[5] Wang, W., & Krishnan, E. (2014). Big data and clinicians: a review on the state of the science. JMIR medical informatics, 2(1).

[6] Mohammed, E. A., Far, B. H., &Naugler, C. (2014). Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData mining, 7(1), 1.

[7] Vadivel, M., and Raghunath, V. (2014). Enhancing Map-Reduce Framework for Bigdata with Hierarchical Clustering: Internation

BIOGRAPHIES OF AUTHORS

Ala'a Al-Hamami received his BS in Physics from University of Baghdad, Iraq in 1970 and an MS in Computer Science from University of Loughborough Technology, England in 1979. In 1983, he received his Ph. D from the University of East AngliaGeorge Mason, England. He is a Professor of Computer Science at Princess Sumaya University, Amman, Jordan. Prof. Al-hamami is interested in Computer Security, Computer Networks, and Internet of Things.

Mr. Ali has a Master degree in Computer Science from Amman Arab University; GPA= 3.81/4.0. The Title of his Master thesis is “Electroencephalography (EEG) Data Analysis by using Map-Reduce Technique”. His work experience was in the following:

- Middle East Bank, Internship assignment for 1 month, July, 2009
- The Smooth Well Company, Manager Assistant, one full year, 2010