

Runtime Optimization of a New Anomaly Detection Method for Smart Metering Data Using Hadoop Map-Reduce

Farid Fathnia, Mohammad Reza Barazesh and Mohammad Hossein Javidi Dasht Bayaz

Department of Electrical Engineering

Ferdowsi University

Mashhad, Iran

Farid.fathnia@mail.um.ac.ir

Abstract— In this paper, we will try to speed up the identification of abnormalities and disturbances that are generated in the data sent from smart meters to the control center by cyber attackers. Two important points in this discussion are the high precision of the proposed method and the speed of its identification. High accuracy is a separate topic and has been documented by same authors in another article. The speed of the method should also be considered in order to prevent possible losses from the onset of a cyber-attack during the operation of the power grid. Therefore, by involving a new computing environment with the term "Cloud Computing", we try to speed up the detection. How to do this process and how to simulate the hardware and software of this topic will come through this article.

Keywords— *Anomaly Detection; Electricity Market; Smart Meter; Cloud Computing; Hadoop; Map-reduce*

I. INTRODUCTION

The power industry is integrating electricity transmission and information systems with telecommunication networks to form a bi-directional electricity and information infrastructure, which is a precondition for network intelligence. In this network, power control systems will have a close relationship with data transfer systems. Also, due to the use of new telecommunication systems and technologies, the structure of the smart grid will be highly efficient and optimal in terms of cost and management. However, aggregating these two structures, while having many benefits, has disadvantages in terms of system security and protection.

Smart network communication system has a layered structure and it has the task of collecting information and controlling power. A telecommunications system in an intelligent network is the result of the integration of several control centers, each of these control centers must be responsible for monitoring several power generation and substation centers. A control center should also be responsible for managing operational data and controlling electricity market orders.

Several issues, such as checking and recording customers' energy consumption, failure reporting, meeting customer

requirements, and so on, all relate to telecommunication fields for information exchange, thus, a special effort has to be made in implementing security on telecommunications and remote systems. In other words, the reliability of smart networks depends on the reliability of the telecommunication and control systems involved in the intelligent system. Due to the expansion of smart grids, telecommunication networks are upgrading day by day, and at the same time, they require better control and higher reliability. Smart grid requires multiple connections between components to meet the expected goals, and with these multiple connections, advanced cybercriminals need to deal with security vulnerabilities and threats.

In this paper, our focus will be on smart meters. Smart meters that transmit power consumption data to power distributions companies should be subject to strict security measures, because the change in these data by security cyber-attackers has resulted in the issuance of false bills, inaccurate statistics, false predictions, and inaccurate decisions. It is priced, and the result is nothing but a deterioration in electricity consumers. An example of this disadvantage with the relevance of demand dispatch is set out in [1]. Therefore, solutions should be developed to detect the turbulence and impact of the attackers with high accuracy and speed.

With regard to communications system intelligence, the electric network that has minimal communications with the Internet is subject to many risks. These include security attacks by enemy groups and hackers to interrupt the production, transmission and distribution of electricity, or manipulate and corrupt data sent to the smart grid. Different layers of cyber security should be designed in such a way as to minimize the threat posed by attacks. All connections to an Internet network need to be very secure. Intruder detection is required not only on the network connection to the Internet, but also within the network, and especially in the wireless data transmission environment, so that the system can detect intrusion. At the same time, in the event of an incident, the system should provide the appropriate response as soon as possible and with minimal delay. For this cloud computing is predicted to play a leading role in the design of future smart grids.

Cloud computing is an emerging technology that has come up with a number of relevant features, ease-of-use and demand-based bundles to access a large amount of computing resources, with minimal management issues for providers of such services. Using cloud-based infrastructure, subscribers can access their applications at any moment, anywhere, and through the Internet.

With the emergence of the Big Data concept, the calculations involved in energy consumption in the future smart grid will be a very big challenge for the centralized structure. In order to repair itself and respond quickly to disturbances during the operation of the power grid, control centers should receive and analyze information very quickly. Also, many offline and online software runs simultaneously on the servers, which is due to the heavy burden on the computing resources of the control centers. So the concept of cloud computing can be helpful in solving these problems, especially for detecting very high rate abnormalities.

From the cloud-based applications in the smart grid, we can point out virtual storage of energy and information storage devices [2]. In this case, intelligent network components interact with the cloud instead of interacting directly with each other, and important energy management decisions are made in this direction [3]. In [4], it is shown that the information sent from the smart meters to the control center during the peak time is more than the other times, so the dynamic bandwidth allocation mechanism, which is a sub-branch of cloud computing technology, can create more bandwidth over time, in order to all tasks are done in parallel and without problems. In [5], the use of cloud computing services, as the basis of communication and management mechanisms, has been outlined for providing robust and efficient intelligent network monitoring and monitoring systems. In [6], the parallel processing framework (including cloud computing services) is provided on a local server for a peak shaving issue to prevent the voltage from being exceeded from its permitted range under different load conditions. The cost optimization method with the help of cloud computing is presented in [7], in order to economize information management in the study agenda of researchers. Infrastructure as a cloud computing service has been used in [8] to have instantaneous and online communication without the least delay when reviewing and analyzing bulk data sent by smart meters.

Due to the growing number of security threats, various authors have provided cloud computing security technologies [9-10]. For example, in [9], the protection and privacy analysis of information in the smart grid is performed in a variety of clouds, especially private cloud. In this paper, the speed and efficiency of the techniques used to deal with malicious software is shown. In [10], cloud computing has been used to implement it in state estimation. In this paper, with the help of public cryptography, security vulnerabilities are detected. The state estimate in this situation uses the highest bandwidth in the cloud to provide a secure and secure structure for the smart grid.

In [11], innovation is in the implementation of the two-stage algorithm for detecting malformations of PMU data through programming in the form of map-reduce. In this

study, a multi-core system was used to implement the process on the data of 4500 PMUs (equivalent to 18,000,000 measurements).

Our approach in this paper is in line with article 11. This means that in the first step, a method is proposed that can give us the right precision in detecting cyber-attacks. Secondly, in order to achieve the goal of optimizing information management by expanding the concept of large data with a minimum delay and performing step 1 as quickly as possible during the operation of the smart grid, the emerging capabilities of cloud computing technology, which is map-reduce programming on the platform Hadoop is used. The presentation of the method for detecting malformations with high precision in article [12] is presented by the authors of this paper, and so in this article we will simply optimize the implementation time of the program, which can be a great innovation in the field of intelligent network and the link between electrical and computer science. Implementing the Hadoop Platform for map-reduce programming is explained later.

The remaining of this paper is as follows. In Section II, a brief overview of the basic concepts in the method presented in Article [12] will be made. In section III, the Hadoop architecture and the map-reduce programming method are expressed. The simulation results are presented in Section IV, and finally, in Section V, we will summarize all of the content.

II. ANOMALY DETECTION METHOD

In [12], we tried to address the problem of detecting anomalies in the advanced intelligent network infrastructure by improving the OPTICS algorithm, which is a dense-axis approach to data mining studies. Because maintaining the security of data in the smart grid and protecting the privacy of customers is one of the most important issues in the operation of an intelligent network. It is worth noting the advantages of using the OPTICS algorithm, is to analyze the data in the multidimensional space. Initially, the diagnosis of abnormal data was done by the LOF index, due to the close relationship between the variables and the concepts of the two approaches mentioned. Then, with regard to the fact that all points are not clearly distinguishable from conventional and non-anomalous points, which also makes detection of anomalies more difficult, it was attempted to use the statistical concept of the coefficient of variation. With this concept and its used feature which is proved in the text of the paper, the relationship between the parameters of the OPTICS method and the LOF index was obtained and through it, we began to identify abnormalities. In order to test, real data from the smart meters in London was used, and we showed that the accuracy of the proposed method is high and it has just a slight error in various scenarios [12].

In Figure 1, the flowchart of the algorithm presented in [12] is shown. See the article for more details. For the sake of completeness, the important parameters shown in Figure 1 and indicators for assessing the accuracy of the proposed method are explained briefly below.

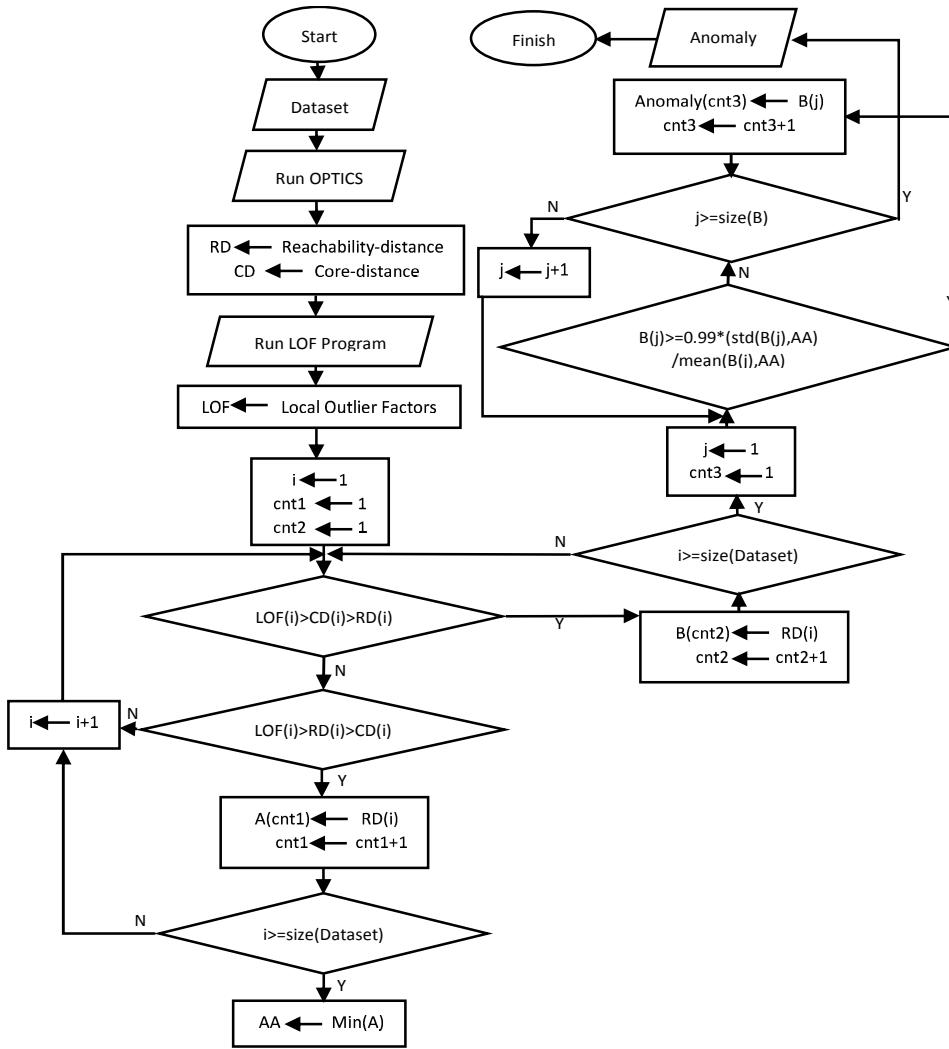


FIGURE 1 FLOWCHART OF ANOMALY DETECTION METHOD [12]

A. Core Distance (CD)

The core-distance of an object p is the smallest ϵ value that makes $\{p\}$ a core object. If p is not a core object, the core-distance of p is undefined.

B. Reachability Distance (RD)

The reachability-distance of an object q with respect to another object p is the greater value of the core-distance of p and the Euclidean distance between p and q . If p is not a core object, the reachability-distance between p and q is undefined.

C. Local Outlier Factor (LOF)

LOF is the average of the ratio of the local reachability density of p and those of p 's M_{inpts} -nearest neighbors. It is easy to understand that the lower p 's local reachability density is, and the higher the local reachability density of p 's M_{inpts} -nearest neighbors are, the higher LOF (p) is.

D. Evaluation Indicators

- 1) TPR or Sensitivity is the percentage of anomaly instances correctly detected.
- 2) FPR is the percentage of normal instances incorrectly classified as anomaly.
- 3) "Precision" is the percentage of correctly detected anomaly instances over all the detected anomaly instances.
- 4) "Accuracy" is the percentage of all normal and anomaly instances that are correctly classified.
- 5) The "F-measure" is the equally-weighted (harmonic) mean of precision and sensitivity.
- 6) "Specificity" is the value of $1 - \text{FPR}$.

These measures determine how the proposed method performs in identifying anomaly instances. In Table 1, various scenarios for assessing the performance of the proposed method are shown [13].

III. BIG DATA CHALLENGE

Parallel and distributed processing platforms have always been one of the most important platforms used in the current technology era. The reason for this is the ever-increasing amount of data. On the other hand, the lack of specific structure in the data has made processing and exploration of them difficult and complex. But with the emergence of parallel and distributed processing platforms such as Hadoop, as well as new programming methods such as map-reduce, the complexity of this area is minimized, since an important part of the tasks of parallel processing and the transfer of them related to processing machines is within the scope of these platforms and they have done well.

In the smart grid, as mentioned earlier, the amount of generated information is added day by day, and the adoption of quick systems for management decision becomes more evident. In this paper, as a goal, the investigation and inventory of intelligent smart meter data is in order to identify cybercrime abnormalities, it is very necessary to provide a system that can, at the same time with its high accuracy, make the controller more aware of events faster. In addition, because we are faced with the problem of operation, this approach requires a minimum of delay in work and eliminates cybercrime effects in the least possible time, and a solution to deal with them.

Below is a brief description of the Hadoop and map-reduce programming.

A. Hadoop

In the current era of software, due to the increasing amount of information, the need for advanced systems for immediate access to information and their accurate recovery, is one of the most important issues. In this regard, many software platforms have been created to address the needs of users and experts in the field. Meanwhile, the Hadoop software framework is one of the best frameworks produced by professionals, which generates many of the leading issues, including mass storage of information at a low pace, high-speed data recovery, intelligent categorization Information and most importantly accurate analysis of information has been solved [14]. But in the meantime, the Hadoop itself has been made up of very complex and progressive components. Among the important parts of the Hadoop software, it is possible to mention the YARN, the map-reduce technology, the job tracker, the task tracker, and the Hadoop distributed file system [15,16].

It's better to know that Hadoop is not a database. Hadoop is also not a software program. Hadoop is a framework or suite of software and libraries that provides the processing mechanism for a huge amount of distributed data. In fact, the Hadoop can be likened to the operating system, designed to handle and manage a large amount of data on different machines.

One of the important parts of the Hadoop is its file system, known as the Hadoop Distributed File System (HDFS). The HDFS features great attributes including design for very large files, the ability to retrieve the second version of the, the ability to work on regular hardware without the need for

specific hardware, and the ability to distribute a large number of computers and display them as a single system. HDFS splits and stores files as 64MB or 128MB blocks. The structure of HDFS is such that there is a name node and a series of data nodes. The name node's task is to store the file metadata information, and the data node holds the file block information [16].

B. Map-Reduce Programming

Map-reduce is a simple programming model that is used to solve large-scale computational and distributed problems. This concept was presented by Google in 2004. Map-reduce is a software framework that provides a safe and scalable platform for the development of distributed applications [15]. In this method, there are two main steps called mapping and reduction.

1. Mapping step: The main node takes the input and divides it into smaller issues. Then they distribute them between nodes that are tasked with doing things. This node may also repeat the same thing, in which case we have a multi-level structure. Finally, these subtasks are processed and the original response is sent to the original node.
2. Reduction step: The main node, which receives responses and results, combines them to provide output. In the meantime, actions such as sifting, summarizing or converting may be done. Go to the results.

These two main actions are applied on a ordered pair (key, value). The mapping function takes one from data and converts it to a list of ordered pairs:

$$Map(k_1, v_1) \rightarrow List(k_2, v_2) \quad (1)$$

Then, the map-reduce collects all couples with the same key from all lists and integrates them together. So, for each key generated, a group is created. Now the reduction function applies to each group:

$$Reduce(k_2, List(v_2)) \rightarrow List(v_3) \quad (2)$$

The map-reduce approach allows us to overcome the traditional approach issues by moving the processing unit towards the data. So, as shown in Figure 2, the data is distributed among several nodes, and each node processes a portion of its data.

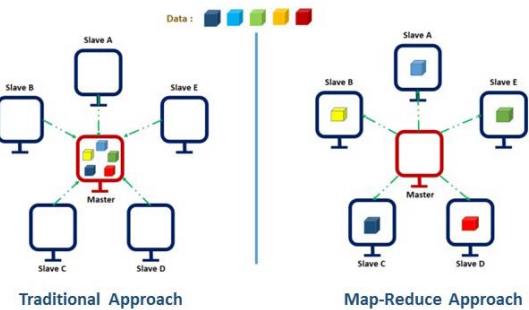


FIGURE 2 COMPARISON OF DATA TRANSFER METHOD

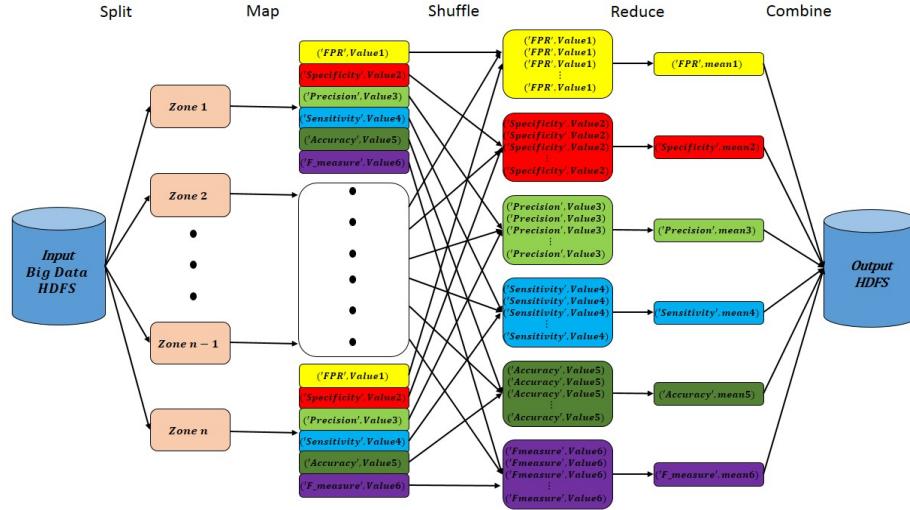


FIGURE 3 PROPOSED ALGORITHM BASED ON MAP-REDUCE PROGRAMMING

IV. SYSTEM MODELING AND SIMULATION RESULTS

The data used in this paper includes customers' energy consumption and associated load factor. It is assumed anomalies are modeled by security invaders with normal distribution functions. More details of Required information can be obtained from the references [12] and [10].

A. System Modeling

Our main goal in this paper is to increase the speed of the process of identifying abnormalities in smart meter data and reduce the overall run time of the algorithm. The main parameter for increasing the speed of the algorithm is reducing the I/O operation and the optimal use of the system resources. In most of the methods and algorithms discussed in similar studies, input documents are analyzed individually and the rest of the work is done. In the main idea of the proposed method, the algorithm presented in [12] on the Hadoop platform is implemented using the map-reduce method so that multiple documents can be examined simultaneously and in parallel. Input data splitting, mapping and reduction are the main sections of the algorithm shown in Figure 3.

1) Input Data Splitting Step: It is assumed that all consumers covered by a control center are expanded in several regions or districts. For example, consumers in each region can be identified on the basis of a 20 KV feeders, or that the regionalization criterion can be based on the Demand and non-Demand subscribers. Each area is named with a special code. This category is included in the data used.

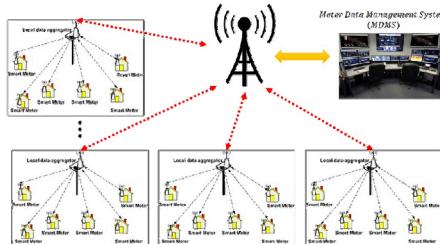


FIGURE 4 CONSUMERS CLASSIFICATION

Map-reduce programs support incoming data in a variety of formats. In our proposed algorithm, since input data is a text document, $n \times m$ matrices are made for each region in which n refers to the number of consumers of each region and m to the number of consumers' data properties and are constant numbers for all of them in all Areas. In fact, these matrices are separated from the first input matrix (which belongs to the entire consumers), and consequently their composition forms the initial input matrix. It should be noted that the data set will be stored initially for use as an input on the HDFS.

2) Mapping Step: The significance of the mapping step is because it performs the main part of the implementation of the proposed method. Each separate matrix of the splitting step is introduced into the proposed algorithm for detecting abnormalities and ordering pairs (key, value) based on system evaluation indicators. As shown in Figure 3, six evaluation indicators are introduced as mapping step keys and their values will be inserted in front of them. Considering the correct use of resources and reducing input/output operations are important issues in improving the implementation time of the algorithm.

3) Reduction Step: The main goal of the reduction step is the aggregation of values that have the same key. These keys are the same evaluation indicators derived from the input documents as obtained at the mapping step. In this case, aggregation is meant to the mean of all values with the same key. After this step, the final file generated, which is the evaluation indicators of the proposed algorithm and will be stored on the HDFS.

B. Implementation Setting

Because the rand function has been used to test the proposed method, we have used the Monte Carlo algorithm to obtain more precise and comprehensive results. To apply this mechanism to the map-reduce method in the Hadoop platform, we also divide each Monte Carlo repeat for parallel processing

into the splitting step to double the speed of conducting investigations and so we'll see some kind of nesting map-reduce program. Focusing on Figure 5 makes this clearer.

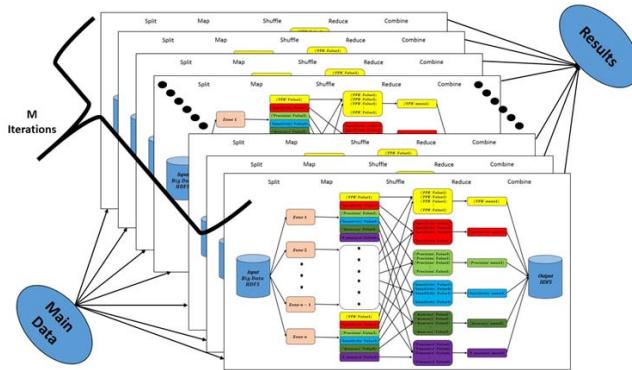


FIGURE 5 PROPOSED ALGORITHM WITH MONTE CARLO SIMULATION

First of all, to use these features, the algorithm written in MATLAB language was rewritten again with the Java language. To implement and run the program on the Hadoop platform, basic settings are required, as shown in Tables 1 through 2. All evaluations have been carried out on a clustered assembly that connects through the network. This network has 2 nodes; one is considered to be the master and another as a slave. All tested machines are virtual and their connection through the virtual network is at 1 GB/s speed. In order to investigate more, we will discuss Hadoop platform tests in a different scenario. In fact, in addition to the first scenario, which is multi-node mode and HDFS is used as the data management bank, in other case, it only uses virtual disk space memory and it's kind of a single-node issue and will be considered locally and HDFS There will not have a role.

TABLE 1 CHARACTERISTICS OF USED SYSTEMS

Type	Master	Slave	PC
CPU	4Cores/ 2.5 GHz	4Cores/ 2.5 GHz	4Cores/ 2.5 GHz
RAM	2 GB	4 GB	8 GB
DISK	20 GB	20 GB	1000 GB
OS	Ubuntu 14.04	Ubuntu 14.04	Windows 10
Hadoop	Version 2.6	Version 2.6	-
Count	1	1	1

TABLE 2 SYSTEM SETTING

Property	Value
HDFS Block Size	256 MB
Dfs..replication	2
Mapreduce.job.maps	14
Mapreduce.job.reduces	7
Map.reduce.reduce.shuffle.parallelcopies	50
Mapreduce.map.memory.mb	1700

C. Simulation Results

TABLE 3 CONSTANT PARAMETERS FOR SIMULATION

No. of Data	3000	10000	20000	29000
Minpts	500	500	500	500
No. of Attacks	10	100	2000	3000
No. of Consumers in each region	1000	1000	1000	1000

1) Scenario 1: 1 iteration

TABLE 4 SPEED COMPARISON IN SECONDS FOR SCENARIO 1

Information		MATLAB	Hadoop Multi Node	Hadoop Single Node
Number of Data	3000	73.66	23	4.44
	10000	2739.28	16.26	8.541
	20000	24249.96	21.88	13.448
	29000	50400.23	30.75	16.634

2) Scenario 2: 100 Iterations (Monte Carlo Algorithm)

TABLE 5 SPEED COMPARISON IN SECONDS FOR SCENARIO 2

Information		MATLAB	Hadoop Multi Node	Hadoop Single Node
Number of Data	3000	241.67	202.71	146.99
	10000	5057.26	1154.4	442.849
	20000	64080.65	2124.45	865.118
	29000	>86400	2867.08	1294.641

D. Results Analysis

According to the assumptions in Table 3, the results of comparing the speed of the proposed algorithm in the MATLAB environment and the Hadoop platform in both single-machine (without HDFS) and two-machine modes (with HDFS) are presented in Tables 4 and 5. As can be seen, the implementation of the proposed algorithm on the Hadoop platform differs significantly from its implementation in the normal way using the software MATLAB. For example, if the number of data is 29,000 and the number of Monte Carlo iterations is 100, the speed required by MATLAB software to perform computations is more than a day, but using Hadoop platform, this time can be reduced to about 21 minutes. So it emphasizes the need to use such platforms in similar situations. In the small number of data, as seen for 3000 data, there is no significant difference between the results achieved by MATLAB software and Hadoop platform, because the purpose and the basis of parallel processing platforms are to deal with massive and big data, and if the volume of data is not high there will be no particular difference, and these platforms will not be able to show their performance efficiently. Another important point in Tables 4 and 5 is that the results of the mode without HDFS are better than this database is involved in the calculation. This is due to the fact that in these tests, for the implementation of the anomaly

detection algorithm, the virtual machines containing the Linux operating system are installed on the Windows operating system with the specifications in Table 1. The HDFS data management system should use its own available memory that is related to the machine, and prepare its own processes, such as formatting the specified space, blocking, and so on to begin the process of anomaly detection. However, if the local environment uses a virtual machine and the time-consuming processes that the HDFS needs does not work, it's clear that the algorithm's execution time will be faster. In fact, the software constraints that advance HDFS do not allow the mechanism to show its optimal performance in a low-volume data set compared to other methods. The larger the data size (on a gigabyte and terabyte scale), the HDFS will gradually show off on its powerful, and will definitely record its best performance.

Figure 6 shows the elapsed times for each platform in four different states in terms of the number of data. As you can see, because there is a very significant difference between the, the bar diagrams for Hadoop are not clear and are not as apparent as the MATLAB software output.

Fig. 7 shows the elapsed time to run the anomaly detection algorithm, assuming 20,000 data in multi node and Local modes in the map-reduce programming. This test has been recorded for a number of consumers in each region. From this figure we notice that the approach should be directed towards more multi-regional indicators of the total number of consumers for increasing the speed of the algorithm.

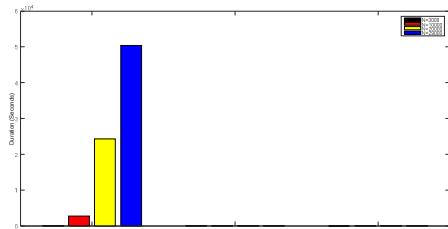


FIGURE 6 SPEED COMPARISON OF DIFFERENT PLATFORMS

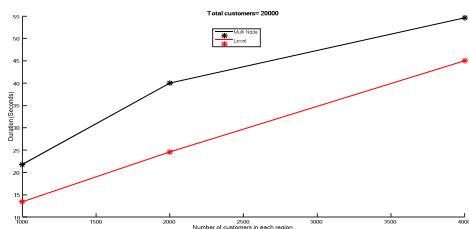


FIGURE 7 SPEED COMPARISON BASED ON REGION CONSUMERS NUMBER

V. CONCLUSION

In this paper, the focus was on optimizing the anomaly detection algorithm from the run-time point of view, and attempted to access this important task through the implementation of modern computer science. By implementing the emerging map-reduce program in the Hadoop platform, parallel processing was proposed to solve the anomaly detection problem. Initially, the area under the control of each control center is divided into multiple regions, and based on this, the number of parallel processes is

determined. Another parallelism that formed was based on the repetitions of the Monte Carlo algorithm, which in fact resulted in nested map-reduce program. Results in this section show the beneficial effect of implementing this method on the speed of the algorithm. Under the same conditions and with a specific system, the implementation of the algorithm in the context of the MATLAB and Hadoop applications under Multi Node and Single Node modes was compared.

REFERENCES

- [1] F. Fathnia, F. Daburi Farimani, F. Fathnia, and M. H. Javidi Dash Bayaz, "The Effect of Cyber Attacks on the Demand Dispatch Application and Identify Them by OPTICS," 4TH International Conference on Knowledge-Based Engineering and Innovation (KBEI), December 2017.
- [2] L. Ji, W. Lifang, and Y. Li, "Cloud Service-Based Intelligent Power Monitoring and Early-Warning System," in Proc. IEEE Conference of innovative Smart Grid Technology, pp. 1-4, 2012.
- [3] S. Bera, S. Misra, and J. P. C. Rodrigues, "Cloud Computing Applications for Smart Grid: A Survey," IEEE Transactions on Parallel and Distributed Systems, vol. 26, no. 5, May 2015.
- [4] S. Misra, S. Das, M. Khatua, and S. M. Obaidat, "QoS-Guaranteed Bandwidth Shifting and Redistribution in Mobile Cloud Environment," IEEE Transactions on Cloud Computing, 2013.
- [5] W. Tushar, W. Saad, H. Poor, and D. Smith, "Economics of Electric Vehicle Charging: A Game Theoretic Approach," IEEE Transactions on Smart Grid, vol. 3, no. 4, pp. 1767-1778, December 2012.
- [6] Y. Xu, Z. Y. Dong, F. Luo, R. Zhang, and K. P. Wong, "Parallel-Differential Evolution Approach for Optimal Event-Driven Load Shedding Against Voltage Collaps in Power Systems," IET Generation, Transmission and Distribution, vol. 8, no. 4, pp. 651-660, April 2014.
- [7] S. Rusitschka, K. Eger, and C. Gerdes, "Smart Meter Data Cloud: A Model for Utilizing Cloud Computing in the Smart Grid Domain," in Proc. IEEE International Conference on Smart Grid Communication, pp. 483-488, 2010.
- [8] H. Goudarzi, S. Hatami, and M. Pedram, "Demand-Side Load Scheduling Incentivized by Dynamic Energy Prices," in Proc. IEEE International Conference on Smart Grid Communication, Dec. 2012.
- [9] Y. Yang, L. Wu, and W. Hu, "Security Architecture and Key Technologies for Power Cloud Computing," in Proc. IEEE International Conference on Transportation Mechanics and Electronics, 2011.
- [10] S. Subashini and V. Kavitha, "A Survey on Security Issues in Service Delivery Models of Cloud Computing," Journal of Networking and Computing Applications, vol. 34, no. 1, pp. 1-11, 2011.
- [11] S. J. Matthews and A. Leger, "Leveraging MapReduce and Synchrophasor for Real-Time Anomaly Detection in the Smart Grid," IEEE Transactions on Emerging Topics in Computing, April 2017.
- [12] F. Fathnia and M. H. Javidi Dash Bayaz, "Anomaly Detection in Smart Grid With Help of an Improved OPTICS Using Coefficient of Variation," 26th Iranian Conference on Electrical Engineering (ICEE), 2018.
- [13] S. R. Gaddam, V. V. Phoha, and K. S. Balagani, "K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, March 2007.
- [14] I. Hashem, I. Yaqoob, and S. Mokhtar, "The Rise of Big Data on Cloud Computing: Review and Open Research Issues," Information Systems, vol. 47, pp. 98-115, 2015.
- [15] Available :<http://hadoop.apache.org/docs/r2.6.0/hadoopmapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>
- [16] Available: <http://hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoophdfs/HdfsDesign.html>
- [17] Available: <https://data.london.gov.uk/dataset/>