2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems

Statistical Analysis for Detection of Sensitive Data using Hadoop Clusters

Binod Kumar Adhikari
College of Computer Science and Technology
Jilin University
Changchun 130012, China
binodkumaradhikari14@mails.jlu.edu.cn

Wanli Zuo
College of Computer Science and Technology
Jilin University
Changchun 130012, China
wanli@jlu.edu.cn

Ramesh Maharjan
Amrit Campus
Tribhuvan University
Kathmandu, Nepal
ramesh.anahcolus@gmail.com

Xuming Han
School of Computer Science and Engineering
Changchun University of Technology
Changchun 130012, China
hanxvming@163.com

Prakash Bahadur Amatya
Patan Multiple Campus
Tribhuvan University
Kathmandu, Nepal
prak_amatya@yahoo.com

Wajid Ali
College of Computer Science and Technology
Jilin University
Changchun 130012, China
aliwajid2117@mails.jlu.edu.cn

***Abstract* — The omnipresence of internet technology and the advent of smart devices accumulate varieties of voluminous, viscous real-time data in a network from varieties of sources and also facilitates a way for criminal, intruders to attack the network which persuades information theft, financial loss, cyber-attack, and cyberwar. It is the major challenge for researchers to determine sensitive data from large scale real-time data so that the right action can be taken at the right time. Therefore, it is important to purpose a framework to handle massive data and to detect sensitive data from big data. In this regard, the collected data from the network are stored in cloud drives, other storage devices and then transferred to Hadoop distributed file system and processed with MapReduce processing using distributed computing concepts, Machine learning algorithms, and advanced statistical methods. Statistical analytical tools like Descriptive Analysis, Regression, ANOVA, Parametric Levene's test, Pearson Correlation, and Kruskal-Wallis Test present that recent big data analytical tools are effective than the traditional method for retrieval of sensitive data from Big Data.**

Keywords—Big Data, Distributed Computing, Correlation, Regression, Hadoop Distributed File System.

I. INTRODUCTION

The era of big data propounded many opportunities and challenges for the international community related with security that remain unnoted in sectors, such as civil society, banking, education, telecommunication, manufacturing, transportation, particularly, the area of concern are data processing and analysis, network architecture, modeling, criminal activities, and information. It is critical to immediately execute the collected data to detect sensitive data from the huge mass of data, however, the existing traditional statistical methods are incapable of handling such a huge amount of data. Hence, it has become indispensable to beat the bottlenecks of the existing traditional system by applying a hybrid of the machine learning algorithm, advanced statistical methods along with big data technologies, to efficiently execute real-time big data to detect sensitive data.

Big data is shaped by the generation, storage, accumulation of the huge amount of structured, unstructured data with the rapid growth of internet techniques and computer science technology and it creates new age [1]. The concept of big data has been implemented in extensive areas of healthcare, business, management, science, engineering, tourism, etc. along with the conceptual and technological innovations [2]. Big data has been categorized into four elementary elements, which are volume (size of the data), variety (different types of data), velocity (data collection speed of data), and veracity (uncertainty of data) and value (importance of data to various academic and industrial fields) [3]. Manogaran et al. [4] introduce some additional characteristics such as variability (fluctuation of data), viscosity (transmission of data between source and destination), validity (correctness of data) and visualization (a graphical representation of data to identify the most relevant data to the users). In spite of additional features of big data, Kitchin [5] has defined the 5V model as the fundamental feature of the Big Data. Presently, the research on big data has been done in a full-fledged way in different areas of applications using big data analytical tools and machine learning techniques.

Terrorism is a complex and evolving phenomenon. There is an increase in terrorist incidents in the world from the last few decades. Terrorism group has threatened the securities and stabilities of many countries [6]. Contrary to this, many effective efforts have been made to collect, store and analyze terrorist data around the world using the latest technologies. Sun et al. [7] proposed document selection strategies based on information extraction patterns and selects one document at a time so that terrorism-related maximum information can be obtained. Kengpol and Neungrit [8] designed and developed a practical decision support methodology for terrorism insurgency situation with the help of prediction modeling and risk assessment analysis. Nie and Sun [9] carried out systematic research to counter terrorism by using quantitative analysis method and demonstrated the effect of big data from data collection, processing, data mining and monitoring. Plaksiy et al. [10] proposed a financial investigation analysis and processing technique that solves criminal cases related to the money laundering and counter financing terrorism by the use of Big Data application.

Big data analytics uses advanced mathematical and statistical modeling techniques to recognize useful trends or characteristics in the data that support business decision making based on the relationships, anomalies, and patterns [11]. The accepted methods for analyzing big data include

978-1-7281-2058-4/19/$31.00 ©2019 IEEE
DOI 10.1109/HPCC/SmartCity/DSS.2019.00330

splitting of massive data into manageable chunks and compressing data to reduce its storage cost to analyze data by using statistical methods to provide business insights on time [12]. Generally, business managers face big challenges to determine relevant tools and techniques to develop a model that suits well to support decision making [13]. Statistical Modelling supplements data mining to provide validity, credibility, and reliability for decision making [14]. Big data analytics produces visual models which are easier to represent and communicate the information by using labeled chart or diagram to describe the results [15].

Big data is creating new challenges as well as opportunities for a data analyst to handle the massive amount of structured, semi-structured and unstructured data sets by the use of Apache Hadoop [16], which is written in Java Programming Language. Hadoop uses MapReduce architecture [17] for processing and Hadoop Distributed File System [18] for storage of data on a cluster which is made up of few computers to several commodities machines. Verma et al. [19] demonstrated the working mechanism of Hadoop MapReduce and Spark architecture with supportive operations. Spark performs its job on Resilient Distributed Datasets (RDD) and directed acyclic graph (DAG) execution engine. Yang et al. [20] investigated the potential of MapReduce to perform statistical calculations and uses the calculated results to visualize data in an efficient and scalable way. The purpose of this article is to answer the following questions:

- Are big data methods are suitable for the retrieval of sensitive data compared with the traditional existing system?
- Are machine learning algorithms and statistical methods working with Hadoop clusters in a distributed computing environment?
- Is statistical analysis able to analyze the differences between traditional methods for information retrieval with Hadoop clusters?

The rest of the paper is organized as follows: section II defines the problems of analysis related to the big data. In Section III, statistical methods implemented with big data are discussed. Then, there is the conceptual framework for the retrieval of sensitive data from big data which are outlined in section IV. For the experimental setup, section V describes server configuration, client configuration, cluster size and also discusses statistical tools used for the analysis of the result. Section VI presents the experimental result obtained after the processing of data. In Section VII, there is a detailed description of statistical analysis implemented with different statistical methods. The conclusion of the article and future work for the retrieval of sensitive data from big data are summarized in section VIII.

II. PROBLEMS OF ANALYSIS OF BIG DATA

Variety of the huge amount of data is collected in the modern age of technology from social media, climate, health status, financial series, retail contract notes, surveillance, smartphones, networking, cloud computing, etc. These data are in different forms and have exponential growth. To analyze such a massive amount of data by the tradition statistical method have decisive importance.

Big data have exceptional features that traditional data doesn't have. The data present in the big data are unstructured, semi-structured and structured. These data are collected from various sources so its nature and rate of the collection are also high. It creates problems such as computational complexity, instability of algorithms, a collection of noise, wrong correlations. It arises heterogeneous problems and generates significant errors in experimental and statistical analysis [21]. At the same time, it is very difficult to apply traditional statistical methods. Sometimes, it explores problems to rectify medium sized collected data to big data and it creates a hurdle to apply promising statistical methods for lower size data during the analysis of high dimensional data. So, it is tough to effectively process voluminous and varieties of data by the use of traditional statistical methods.

Some of the magnificent problems generated during statistical analysis of big data are:

- An unstructured large volume of data
- Privacy issues and data protection
- Deficiency of time
- Statistician does not have operational extraction practices of information from big data.

Therefore, it does not feel a rapport to expect the productivity from traditional statistical methods for the solution of the problems created by big data. New computational methods should be created by the use of distributed computing, clustering, and machine learning algorithms to solve the problems of big data.

III. STATISTICAL METHODS FOR BIG DATA

With the detonation of "Big Data", statistical learning and statistical computing have grown into a very hot topic in numerous scientific as well as finance, marketing, and another business area. Computational statistics is the alliance between computer science and statistics. Statistical aggregation breaks the entire data set into smaller chunks, compresses each chunk into low dimensional statistical summary and then again combines summary statistics to get the desired result based on the entire data.

The scientific computations are extensively used with big data by the iterative algorithms in the existing society. The models for iterative algorithms are Markov chain Monte Carlo (MCMC) algorithms [22] and Expectation-maximization (EM) algorithm [23] that require a number of iterations and a complete record of the dataset for each iteration. The EM algorithm has been implemented with scientific computation in the presence of missing data for parameter estimation. While Monte Carlo averages are calculated from complete datasets to approximate the quantities by using general principal for big data analysis.

Bennett et al. [24] derived a number of effective formulas for a single pass and provided incremental updates for arbitrary ordered statistical moments and co-moments. They have developed open source parallel statistics framework that uses principal component analysis to compute descriptive, correlative and multi-correlative statistics. The statistical analysis solves the problems of parallel scalability at the large scale data sets. Monotonic regression (MR) is a powerful tool for evaluating the functions that are monotonic to input variables. The MR problems have been implemented in operational research, statistics, signal processing, biology, and other areas.

Unfortunately, these algorithms only solve the problems related to a small number of observations so they cannot provide effective results for medium or large scale data within a given time frame. Burdakov et al. [25] offered a fast and effective algorithm, called Generalize PAV algorithm for monotonic regression (GPAV). The GPAV solves the problems related to the large-scale multivariate monotonic regression and is based on segmentation of large scale problems to small scale.

IV. Conceptual Framework

The exponential growth of innumerable emerging technologies, such as mobile cloud, 5G communication media, connected devices, multimedia, and virtual reality, autonomous automobiles, smart home appliances, and social media contribute to the accumulation massive amount of real-time data in a network. A report of Ali et al. [26] predicted that the Internet world might create a huge community of 50.1 billion connected devices. In this huge amount of data, there may be the chance of sensitive data created by the criminals. This expected growth of data and the determination of those sensitive data from big data arises a challenging task for big data analytics.

The biggest challenge for organizations like the government sector, banks, health cares, research center, and civil society to monitor criminal activities and control future unexpected events. These organizations spend a lot of money to protect and secure data from hackers and infrastructures using various tools and techniques. However, traditional methods for retrieval of sensitive data are inefficient to detect, control and monitor the existing system. As the rate of growth of data speeds up, the new tools and techniques should be implemented by big data analytical methods. It monitors critical network deeds in real time and produces an alarm if sensitive data found in the network.

Fig. 1 provides a conceptual architecture for the determination of sensitive data from big data which is based on the bottom-up approach. Varieties of voluminous data are collected from smart devices which are communicated through network technologies. These data are stored in the cloud and other storage devices through network and storage infrastructure. The huge amount of collected data from smart devices are processed with big data analytical tools like Hadoop, Spark, and Apache Storm. In Hadoop, data are first stored in the Hadoop distributed file system and then processed by MapReduce architecture.

As per requirements for the processing of huge amount of data, a number of nodes in a cluster may vary. There may be one node cluster or may be more that is based on the size of data, and requirements of the system, and real-time processing of data. If a number of nodes are more, the time taken for processing of data will be less. With the use of machine learning techniques like Term Frequency and Inverse Document Frequency (TF-IDF), Neural Network (NN), Support Vector Machine (SVM), K- means clustering and advanced statistical methods like Markov chain Monte Carlo (MCMC) algorithms, Expectation-maximization, and Generalize PAV algorithm for monotonic regression (GPAV) sensitive data are detected from big data.

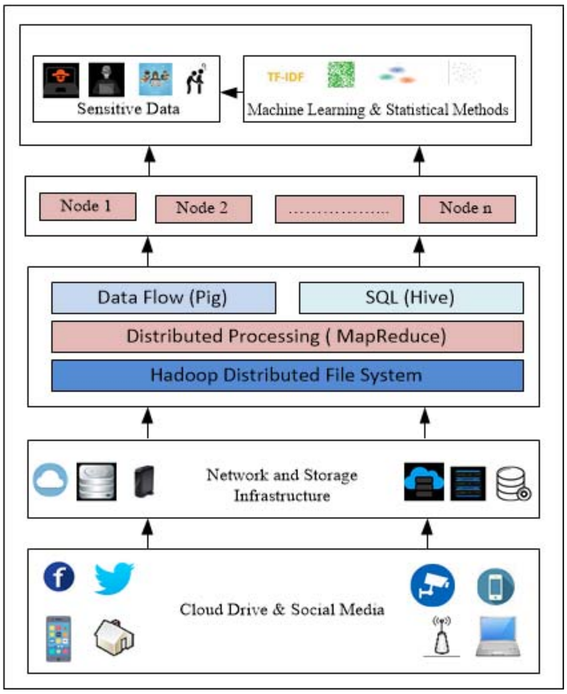
Sensitive Data
TF-IDF
Machine Learning & Statistical Methods
Node 1
Node 2
....
Node n
Data Flow (Pig)
SQL (Hive)
Distributed Processing (MapReduce)
Hadoop Distributed File System
Network and Storage Infrastructure
Cloud Drive & Social Media

Fig. 1. Conceptual framework for retrieval of sensitive data

V. Experimental Setup

The cluster is setup using seven commodity computers in which one computer is set up as a server and remaining six computers are configured as a client. Server computer has configuration of Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz 3.60 GHz 64-bit as CPU with 16 GB RAM and 1 TB storage, and Ubuntu 16.04 as operating system while client computer has configuration of Intel i5 quad core 64-bit with 8 GB RAM and 500 GB Storage, and Ubuntu 16.04 as operating system. Hadoop 2.7.3 with YARN is configured on all seven computers with Java OpenJDK version of "1.8.0 121". The experiments were performed with 1 node, 3 node, 5 node, and 7 node clusters size. SPSS 25 was used as a tool for the statistical analysis of the result.

VI. Experiments and Results

The theme of the experiment is to provide the statistical analysis of the performance of Hadoop cluster by increasing the input data size and number of nodes. The big data analytical methods along with Hadoop clusters are implemented for the calculation of a number of repetition of words present in the document.

The experiment is initiated with 1 node, 3 nodes, 5 nodes, and 7 nodes Hadoop cluster by varying input sizes i.e. 1 GB, 3 GB, 5 GB, and lastly 7 GB. First of all, the experiment had successfully completed for a single node and then the size of the clusters was increased and the time taken for calculation were recorded in Seconds and the result was summarized in table 1 and plotted in graphical form in fig. 2.

Table1: Summary of Experiment

	Input data size				
Node	**1GB**	**2GB**	**3GB**	**4GB**	**5GB**
1	55s	83s	113s	154s	190s
3	43s	69s	97s	131s	147s
5	37s	48s	69s	88s	101s
7	36s	45s	55s	65s	73s

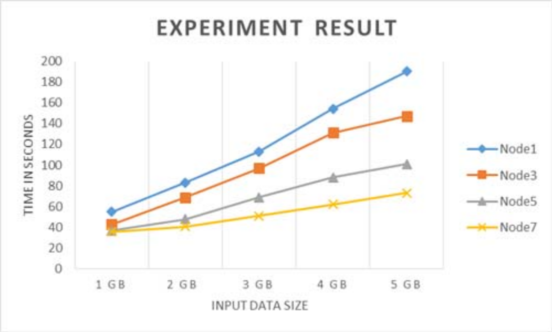

Fig. 2. Time taken for retrieval of information

Looking at the above graph we can say that as the number of nodes increases in the Hadoop cluster, the time taken for retrieval of information decreases. The line generated in the graph is approximately straight lines. It shows that the slope of the line is near to one. From the figure, we can say that the straight lines generated by the use of 7 nodes clusters are less steep than 5 nodes cluster and in the same way line generated by 3 node clusters is less steep than 5 nodes. The figure also shows that the line generated by a 1 node cluster is comparatively steeper than all other straight lines. Time taken for processing of 1 GB data by 7 nodes clusters has taken less than all another cluster. As the data size increases, there are more differences in time for processing by different cluster size.

VII. Statistical Analysis

Statistical analysis characterizes the nature of the data to be inspected and interpreted, establishes the relation of the data among the population, conceive a model to summarize the perception to correlate the data the underlying population, substantiate the validity of the pattern and handle predictive analytics to run master plan that will boost to guide future actions. The goal of the statistical analysis is to pinpoint the trends among the data. Some of the statistical methods implemented are discussed below:

A. Descriptive Analysis

Table II provides the details about input data size and its execution time is taken for a single node cluster.

Table II: Execution time is taken for 1 Node Cluster	
x	y
1000	55
2000	83
3000	113
4000	154
5000	190

Where x is for input size (in MB), y is for the time taken in Seconds. Their means values are $\bar{x} = 3000$ and $\bar{y} = 119$ Based on the formula $\bar{x} = \frac{\sum x}{n}$ and $\bar{y} = \frac{\sum y}{n}$, where n represents a number of points used for calculation. The standard deviations calculated for one node cluster are $\sigma_x = 1581.14$ and $\sigma_y = 54.07$ by the use of formula

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}, \; \sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}.$$

The correlation coefficient for a single node is 0.997178335, which is almost equal to 1. This shows that the line is nearly a straight line. The correlation coefficient is calculated by using formula.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

In the same way, the slope of the straight line (0.0341) and y-intercept (16.7) are obtained for a single node cluster.

The same process is carried out for the rest of the clusters i.e. 3 nodes, 5 nodes, and 7 nodes. The standard deviation (SD), correlation coefficient, slope, and y-intercept were calculated and the summarized results are presented in the given table III.

Table III. Summary of SD, Correlation, Slope, and y-intercept

Node	SD	Correlation	slope	Intercept
1	54.0694	0.997178	0.0341	16.7
3	42.88123	0.995558	0.027	16.4
5	26.68895	0.995286	0.0168	18.2
7	14.87279	0.999322	0.0094	26.6

The value obtained for the standard deviation shows that there is too much time gap between 1 node cluster and 7 node clusters. 1 node has taken more time comparative to 7 node clusters. The correlation coefficient shows that as the number of nodes increases, the time is taken for execution also increases. Thus from the above table, we can conclude that the slope of 7 nodes < slope of 5 nodes < slope of 3 nodes < slope of 1 node.

B. Regression

Table IV provides regression statistics.

Table IV: Regression Statistics	
Multiple R	0.774353699
R Square	0.59962365
Adjusted R Square	0.55252055
Standard Error	154.6978725
Observations	20

The required regression equation of y (time took) on x_1(number of the node) and x_2 (volume of data) is found to be $y = 294.21 - 69.64x_1 + 0.055x_2$ with R^2 =0.59962 and p value =0.00031 and standard error of 15.469. Hence it can be concluded that the regression is significant. Table V gives coefficient, standard error and p-value for interpretation.

Table V : Coefficient, Standard Error, P-value

	Coefficients	*SE*	*t Stat*	*P-value*
Intercept	294.210	102.030	2.884	0.010
Number of Node	-69.640	15.470	-4.502	0.000
Size of Data in MB	0.056	0.024	2.279	0.036

Normal probability plot for the time taken for execution is demonstrated in Fig. 3. The chart indicates that the data are approximately normally distributed because most of the points fall along a curve line.

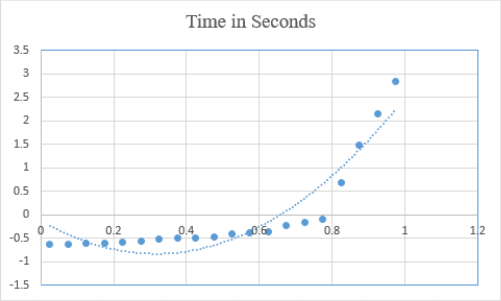

Fig. 3: Normal probability plot of execution time

It can be interpreted that the time taken decreases by 69.64 units per unit increase in a node with standard error =15.4697 and t=-4.50 and p-value 0.000315 and C.I. is 37.001 to 102.27. It is seen that the time is taken and a number of nodes have a significant relation.

It can be interpreted that the data volume increases by 0.055 units per unit increase in data volume in GB with standard error=0.0244 and t=2.2792 and p-value 0.03584and C.I. is 0.004144 to 0.107355. It is seen that the time is taken and data volume in GB have significant relation.

C. ANOVA

1) Based on the number of nodes

ANOVA result is presented on table VI based on a number of nodes.

TABLE VI: ANOVA

	df	Mean Square	F	Sig.
Between Groups	3	239033.93	12.79	.000
Within Groups	16	18689.25		
Total	19			

Since the F-value is 12.79 and its associated significance (.000) is less than 0.05, we accept the null hypothesis and say that variances are not equal for all the three groups. This implies that changes in the number of nodes in a cluster affect the retrieval of sensitive data from big data.

2) Based on the size of data

Table VII gives ANOVA result based on the size of data.

TABLE VII: ANOVA

	df	Mean Square	F	Sig.
Between Groups	4	31140.700	.524	.720
Within Groups	15	59437.800		
Total	19			

Since the F-value is 0.529 and its associated significance (.720) is more than 0.05, we reject the null hypothesis and say that variances are equal for all the three groups.

D. Parametric Levene's test

1) Based on the number of nodes

Table VIII provides Levene Homogeneity test for variances based on a number of nodes.

TABLE VIII: TEST OF HOMOGENEITY OF VARIANCES

Time in second			
Levene Statistic	df1	df2	Sig.
8.927	3	16	.001

The null hypothesis for the parametric Levene's test shows that there is an equality of variance. If the p-value is greater than 0.05, we keep the null hypothesis and assume equality of variance. In this example, the p-value is 0.01, which is below 0.05, so we reject the null hypothesis. The groups have not equal variance.

2) Based on the size of data

Based on the size of data, Levene Homogeneity test for variances is summarized in Table IX.

TABLE IX: TEST OF HOMOGENEITY OF VARIANCES

Time in seconds			
Levene Statistic	df1	df2	Sig.
2.136	4	15	.127

In this example, the p-value is 0.127, thus above 0.05, so we keep the null hypothesis. The groups have equal variance.

E. Correlations

1) Pearson Correlation

Pearson Correlation test is summarized in Table X.

TABLE X: PEARSON CORRELATIONS

		Number of Nodes	Time in Second	Size of Data in MB
Number of Nodes	Pearson Correlation	1	-.691**	.000
	Sig. (2-tailed)		.001	1.000
	N	20	20	20
Time in Second	Pearson Correlation	-.691**	1	.350
	Sig. (2-tailed)	.001		.131
	N	20	20	20
Size of Data in MB	Pearson Correlation	.000	.350	1
	Sig. (2-tailed)	1.000	.131	
	N	20	20	20

In the Pearson correlation, there is a linear positive correlation between the size of data and time taken for execution. The correlation coefficient is 0.350 and its significance value is 0.131. So, it is not statistically significant (p>0.05). It pinpoints that as the data size increases, the execution time for retrieval of sensitive data also increases. There is a negative correlation between a number of nodes and time taken for execution because its correlation coefficient is -0.691. It is statistically significant because its p-value (0.01) which is less than 0.05. It proves that as the number of nodes increases in a cluster, sensitive data are retrieved with high accuracy in minimal execution time.

F. Kruskal – Wallis Test

1) Based on the number of nodes

Based on a number of nodes, table XI outlines the test statistics of the Kruskal – Wallis Test.

TABLE XI: TEST STATISTICS

	Time in Second
Kruskal-Wallis H	12.360
df	3
Asymp. Sig.	.006

Since Kruskal-Wallis value is 12.360 and its associated significant value is 0.006 which is less than 0.05, there is a significant difference in time for execution based on the number of nodes.

2) Based on the size of data

Table XII summarizes the result of the Kruskal-Wallis test.

TABLE XII: TEST STATISTICS

Kruskal-Wallis H	5.786
df	4
Asymp. Sig.	.216

Since Kruskal – Wallis value is 5.786 and its associated significant value is 0.216 which is more than 0.05, there is no significant difference in time for execution based on the size of data.

Therefore, mathematically we can conclude that as the number of nodes increases, performance, such as speed and efficiency, for retrieval of sensitive information also increases in the Hadoop cluster.

VIII. Conclusion and Future Work

In this article, we have explored the possibility of sensitive data detection using big data analytical tools, machine learning algorithms and advanced statistical methods by the use of distributed computing concepts. We have inspected the recent work completed with big data along with statistical methods, machine learning algorithms in different fields. We have developed a framework, for detection of sensitive data from varieties of voluminous data, which fulfills the shortcomings and challenges of the existing methodologies. Moreover, we have implemented big data analytical tools to retrieve sensitive information from a large scale of data by using distributed computing and Hadoop clusters. Then, we have statistically analyzed the result using ANOVA, Parametric Levene's test, Pearson Correlation, and Kruskal-Wallis Test and proved that big data analytical methods are more applicable to retrieve sensitive information than traditional statistical methods.

At this point, we are highly interested to investigate other important researches related to the same directions in the future. First, we are planning to generate more frameworks that provide guidelines and new concept to the researcher in the related field. Second, we are planning to implement and execute machine learning algorithms along with big data analytical tools to retrieve sensitive data accurately. Third, we are planning to determine sensitive data from big data in a real-time environment by comparing phrasal words against massive data sets by using the latest technologies.

Acknowledgment

Project supported by the Nature Science Foundation of Jilin Province (No. 20180101330JC), National Nature Science Foundation of China (No. 61472049), the Outstanding Young Talent Project of Jilin Province (No. 2017052005954), the Key Scientific and Technology Projects of Jilin Province. (No. 20130206051GX).

References

[1] K. Kambatla, G. Kollias, V. Kumar, A. J. J. o. P. Grama, and D. Computing, "Trends in big data analytics," vol. 74, no. 7, pp. 2561-2573, 2014.

[2] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. J. I. s. Khan, "The rise of "big data" on cloud computing: Review and open research issues," vol. 47, pp. 98-115, 2015.

[3] A. B. Ayed, M. B. Halima, and A. M. Alimi, "Big data analytics for logistics and transportation." pp. 311-316.

[4] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K. M. Abbas, and R. Sundarsekar, "Big data knowledge system in healthcare," *Internet of things and big data technologies for next-generation healthcare*, pp. 133-157: Springer, 2017.

[5] R. Kitchin, "Big data—Hype or revolution," *The SAGE handbook of social media research methods*, pp. 27-39, 2017.

[6] I. Toure, and A. Gangopadhyay, "Real-time big data analytics for predicting terrorist incidents." pp. 1-6.

[7] Z. Sun, E.-P. Lim, K. Chang, T.-K. Ong, and R. K. Gunaratna, "Event-driven document selection for terrorism information extraction." pp. 37-48.

[8] A. Kengpol, P. J. C. Neungrit, and I. Engineering, "A decision support methodology with risk assessment on prediction of terrorism insurgency distribution range radius and elapsing time: An empirical case study in Thailand," vol. 75, pp. 55-67, 2014.

[9] S. Nie, and D. Sun, "Research on counter-terrorism based on big data." pp. 1-5.

[10] K. Plaksiy, A. Nikiforov, and N. Miloslavskaya, "Applying Big Data Technologies to Detect Cases of Money Laundering and Counter Financing of Terrorism." pp. 70-77.

[11] Z. Sun, K. D. Strang, and J. Yearwood, "Analytics service oriented architecture for enterprise information systems." pp. 508-516.

[12] N. Kandalkar, A. J. I. j. o. e. t. Wadhe, and technology, "Extracting large data using big data mining," vol. 9, pp. 576-582, 2014.

[13] R. J. Kauffman, J. Srivastava, J. J. E. C. R. Vayghan, and Applications, "Business and data analytics: New innovations for the management of e-commerce," vol. 11, no. 2, pp. 85-88, 2012.

[14] K. D. Strang, "Selecting research techniques for a method and strategy," *The Palgrave Handbook of Research Design in Business and Management*, pp. 63-79: Springer, 2015.

[15] N. R. Vajjhala, K. D. Strang, and Z. Sun, "Statistical modeling and visualizing open big data using a terrorism case study." pp. 489-496.

[16] C. Lam, *Hadoop in action*: Manning Publications Co., 2010.

[17] M. Laclavík, M. Šeleng, and L. Hluchý, "Towards large scale semantic annotation built on MapReduce architecture." pp. 331-338.

[18] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system." pp. 1-10.

[19] A. Verma, A. H. Mansuri, and N. Jain, "Big data management processing with Hadoop MapReduce and spark technology: A comparison." pp. 1-4.

[20] X. Yang, S. Liu, K. Feng, S. Zhou, and X.-H. Sun, "Visualization and adaptive subsetting of earth science data in HDFS: A novel data analysis strategy with Hadoop and spark." pp. 89-96.

[21] J. M. Jordan, and D. K. J. 中. Lin, "Statistics for big data: Are statisticians ready for big data?," vol. 52, no. 1, pp. 133-149, 2014.

[22] M.-H. Chen, Q.-M. Shao, and J. G. Ibrahim, *Monte Carlo methods in Bayesian computation*: Springer Science & Business Media, 2012.

[23] A. P. Dempster, N. M. Laird, and D. B. J. J. o. t. R. S. S. S. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," vol. 39, no. 1, pp. 1-22, 1977.

[24] J. Bennett, R. Grout, P. Pébay, D. Roe, and D. Thompson, "Numerically stable, single-pass, parallel statistics algorithms." pp. 1-8.

[25] O. Sysoev, O. Burdakov, A. J. C. S. Grimvall, and D. Analysis, "A segmentation-based algorithm for large-scale partially ordered monotonic regression," vol. 55, no. 8, pp. 2463-2476, 2011.

[26] A. Ali, W. Hamouda, and M. J. a. p. a. Uysal, "Next generation M2M cellular networks: challenges and practical considerations," 2015.