ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

Analysing Errors in a Conducive Approach on Sensor Data

Priyasree.K[1], Prof.Jayashubhaj.J[2], Damini.V.S[3], Prof.Rabindranath[4]

PG Student, Department of CSE, AMC Engineering College, Bengaluru, India

Associate Professor, Department of CSE, AMC Engineering College, Bengaluru, India

PG Student, Department of CSE, AMC Engineering College, Bengaluru, India

Associate Professor, Department of CSE, AMC Engineering College, Bengaluru, India

ABSTRACT: Big sensor data is pervasive in both industry and scientific applications where the data is created with high volume and speed, it is hard to process utilizing available database administration tools or conventional data handling applications. Hadoop provides a promising platform to support the addressing of this challenge as it provides a flexible stack of massive computing, storage, and software services in a scalable manner at low cost. A few systems have been created as of late to process sensor data on cloud, for example, sensor-cloud. However, traditional techniques do not provide productive support on speed detection and locating of errors in big sensor data sets.For speed data error detection in big sensor data sets, in this paper, we develop a novel data error detection approach which exploits the full computation potential of Hadoop platform. Firstly, Differentiation and definition of Different sensor error types are done. Based on that classification, the network feature of a clustered Wireless sensor network is introduced and analysed to support speed error detection and location. Explicitly, in the proposed approach,a large portion of identification operations can be directed in restricted temporal or spatial data blocks rather than an entire big data set. Thus, the detection and location process can be dramatically accelerated. Additionally, the detection and location tasks can be distributed to Hadoop platform to completely exploit the computation power and massive storage.Through the analysis on our Hadoop platform, it is exhibited that our proposed methodology can significantly recede the time for error detection and location in big data sets generated by large scale sensor network systems with acceptable error detecting accuracy.

KEYWORDS: Big data, Hadoop, data abnormality, error detection, time efficiency, sensor networks, complex network systems

I.INTRODUCTION

RECENTLY, we enter a new era of data explosion which brings about new challenges for big data processing. In general, big data [1], [2] is a collection of data sets, All in all, enormous information, is a gathering of information sets so huge and complex that it gets to be hard to handle with available database administration frameworks or conventional information preparing applications. It represents the progress of the human cognitive processes, usually includes data sets with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time [1], [2], [12], [13], [14], [15], [16], [17],. BigData has typical characteristics of five 'V's, volume, variety, velocity, veracity and value. Big data sets come from many areas, including meteorology, connectomics [1], [2]. According to literature [1], [2], since 1980s, generated data doubles its size in every 40 months all over the world. Hadoop, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. One of important source for scientific big data is the data sets collected by wireless sensor networks (WSN). Wireless sensor networks have capability of essentially improving individuals' capacity to screen and communicate with their physical surroundings. For a WSN application to deduce an appropriate result, it is vital that the information got is clean, accurate, and loss- less. However, effective detection and cleaning of sensor big data errors is a challenging issue demanding innovative solutions. Therefore, the question of how to find data errors in complex network systems for improving and debugging the network has attracted the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

interests of researchers. Since these techniques were not designed and developed to deal with big data on hadoop, they were unable to cope with current dramatic increase of data size. WSN big data error detection commonly requires powerful real-time processing and storing of the massive sensor data as well as analysis in the context of using inherently complex error models to identify and locate events of abnormalities. In this paper, we aim to develop a novel error detection approach by exploiting the massive storage, scalability and computation power of Hadoop to detect errors in big data sets from sensor networks. However, fast detection of data errors in big data with Hadoop remains challenging.

II.RELATED WORK AND PROBLEM ANALYSIS

To address various challenges of big data, research works can be found intensively from the database view. However, the problem can be also discussed from the perspective of parallel systems and cloud .

- **Big Data Processing on Hadoop**

With the quick advancement of data innovation, we enter another period of information. Hence, the technique to process big data has become a fundamental and critical challenge for modern society. Hadoop can be regarded as an ingenious combination of a series of developed or developing ideas and technologies, establishing a pay-as-you-go business model by offering IT services using economies of scale [5], [6], [7], [8], [9], [10], [11]. Kienzler et al. [8] designed a "stream-as-you-go" approach to access and process on incremental data for data-intensive cloud applications via stream-based data management architecture. The extension of the traditional Hadoop framework [11] to develop a novel framework named Incoop by incorporating several techniques like task partition and memorization-aware schedule. Olston et al. [9] present a continuous workflow system called Nova on top of Pig/Hadoop through state full incremental data processing. MapReducehas been widely revised from a batch processing framework into a more incremental one to analyze huge-volume of incremental data on hadoop. It is a framework for processing parallelizable problems across big data sets using a large number of computers (nodes), collectively referred to as a cluster in which all computers (nodes) are on the same local network and use similar hardware; or a grid in which the nodes are shared across geographically and administratively distributed systems. It can sort a petabyte of data in only a few hours. The parallelism also provides some possibility of recovering from partial failure of servers or storage during the operation.

- **Analyzing Errors in Sensor Networks and Complex Networks**

As an imperative logical enormous information source, scientific sensor systems and wireless sensor network applications produce a variety of enormous data sets in real time through various monitored activities in different application domains, such as healthcare, military, environment, and manufacturing. With the sensational increment of huge information produced from complex system frameworks, to find and locate the errors in big data sets turns to be difficult with normal computing and network systems. Wang et al. [22] provide a classification for errors on social networks based on error scenarios analysis. Hence, the error models and types presented in [22] can be extended forthe errors in complex network systems. Xiong proposed an approach which can be used to detect the text data errors in data sets of social network. Mukhopadhyay proposed a model based error correction method for WSN. It is conducted over intelligent sensor network itself. This technique is based on the correction with data trend prediction. The primary goal of this location error analysis is to demonstrate the practical use of the location errors for optimal resource consumption. It can be concluded that current data error detection techniques for complex network systems focus on in-network detecting with intelligent nodes or offline analysis at the root. They ignore the scalability, massive resource and powerful computation capability provided by Hadoop. The proposed approach in this paper aims to address this issue by utilizing the inherent features of Map-Reduce to realize fast error detection on Hadoop.

III.CONDUCIVE APPROACH ERROR DETECTION ON BIG SENSOR DATA ON HADOOP

According to the above analysis, it is clear that complex network systems have a similar clustered network topology. During the filtering of big data sets, whenever an abnormal data is encountered, the detection algorithm needs to finish two tasks. They are depicted as two functions here. "$f_d(n/e, t)$" is a decision making function which determines whether the detected abnormal data is a true error. In other words, "$f_d(n/e, t)$" has two outputs, "false negative" for detecting a true error and "false positive" for selecting a non-error data. "$f_i(n/e, t)$" is a function for tracking and

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

returning the original error source. With the results from the above two functions, the error detection process can be successfully finalized.

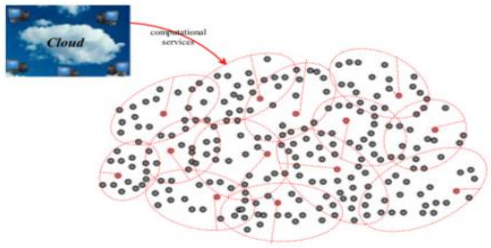

Fig. 3. Cluster based error detection strategy on cloud.

As shown in Fig. 3, there is a complex network and Hadoop platform for running error detecting algorithms. Without any consideration of network features and data characteristics, the error detection algorithm needs to filter the whole big data set from the network. Only a few nodes in the whole network have large sets of links to other nodes. So, based on these nodes, the whole networks can be partitioned into a group of clusters (red circles). If there is certain abnormal data occurs for a certain node k, the high opportunity is that most of the related data for $f_d(n/e, t)$ and $f_i(n/e, t)$ will be located in the clusters where the node k locates. As a result, $f_d(n/e, t)$ and $f_i(n/e, t)$ only need to navigate the related clusters for error detection result. This is because of the fact that except for a few central nodes, most of nodes only have limited links within themselves in their clusters. Hence, the proposed clustering can significantly reduce the time cost error locating and final decision making by avoiding whole network data processing. In addition, with this detection technique, HDFS resources only need be distributed according to each partitioned cluster in a complex network.

IV.ALGORITHMS

To deploy the proposed error detection model and identifying the location of the error, the algorithm can be divided into two parts, detection and location. In this section, we will introduce the big data error detection/location algorithm, and its combination strategy with Hadoop.

- **Error Detection**

We propose a two-phase approach to conduct the computation required in the whole process of error detection and localization. At the phase of error detection, there are three inputs for the error detection algorithm. The first is the graph of network. The second is the total collected data set D and the third is the defined error patterns p. The output of the error detection algorithm is the error set D'.

- **Error Localization**

After the error pattern matching and error detection, it is important to locate the position and source of the detected error in the original WSN graph G(V, E). The input of the Algorithm 2 is the original graph of a scale-free network G (V, E), and an error data D from Algorithm 1. The output of the algorithm 2 is G'(V', E') which is the subset of the G to indicate the error location and source.

V.EXPERIMENTS

To verify the time efficiency and the effectiveness of our approach for detecting errors in big data with hadoop, experiments are conducted on Hadoop [12], [13], [14], [15], [16]. There are three purposes for this experiment.1)Demonstrate that the significant time-saving is achieved in terms of detecting errors from complex network bigdata sets.2)Demonstrate the effectiveness of our proposed error detection approach in terms of different error types. 3) Demonstrate that the false positive ratio of our proposed error detection algorithm is limited within a small value.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

- **Experiment Environment and Process**

The Hadoop is set up as shown in Appendix C.1, available in the online supplemental material. The total testing data set size is around 2,000,000 KB, including temperature, sound, light and vibration. The data sampling from each real world node is heterogeneous. Before the experiment, we conduct the normalization forthe testing data set. In Fig. 1, the testing results show the time performance of our proposed scale-free error detection algorithm on U-Cloud after 740 seconds. Specifically, 10 different error rates are imposed into the experimental data set and tested independently. The testing error rate changes from 1 to 10 percentin 10 repetitive experiments. After about 100 seconds, the proposed algorithm can detect more than 60 percent errors whatever the testing error rate is within the domain between 1 and 10 percent . During the time duration between 0 and 100 second, all error detection rates increase dramatically with a steep trend. After the time point of 300 second, the error detection rates increase slowly with a flat trend. At the time of 740 second, the proposed error detection algorithm on cloud can find and locate more than 95 percent imposed errors from the testing data sets. When testing error rate is 1 percent , the best performance gains are achieved, as about 99.5 percent total errors detection. With the increase of the testing error rate, the error detection rate decreases.

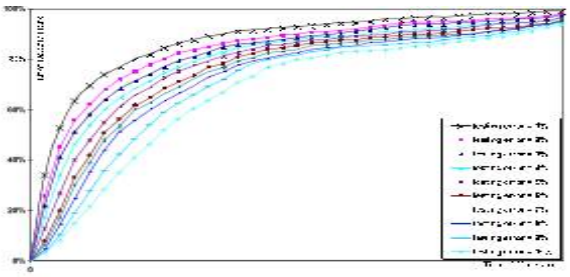

Fig. 1. Time cost for detecting errors from the testing data set

- **Experiment Results**

In order to test the false positive ratio of our error detection approach and time cost for error findings, we impose five types of data errors into the normalized testing datasets with a uniform random distribution. These five types of data errors are generated equally. Hence, the percentage of each type of errors is 20 percent from the total imposed errors for testing.As shown in Fig. 2,when the testing data error rate changes from 1 to 10 percent, at any time slot, our proposed scale-free error detection algorithm achieves significant error detection performance gains compared to non scale-free error detection algorithms. Our proposed scale-free detection on cloud can fast detect most of error data (more than 80 percent)after 740 seconds time duration. However, the non scalefreeerror detection algorithm can only achieve as much as 44 percent error detection rate as the best case. So, it can be concluded from the experiment results in Fig. 2 that the scale-free detection algorithm on cloud for big data can significantly outperform non scale-free error detection algorithms in terms of error finding time cost.The first imposed error type is the flat line error. The second imposed error type is out of bound error. The third imposed error type is the spike error. The forth imposed error type is the data lost error. Finally, the aggregate & fusion error type is imposed By imposing the above listed five types of data error types, the experiment is designed to measure the error selection efficiency and accuracy during the on-cloud processing of data set.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

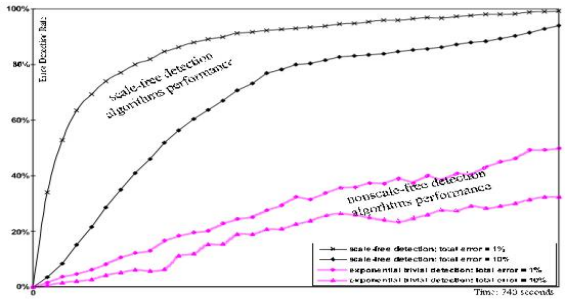

Fig. 2. Comparison of two error detection strategies

VI.CONCLUSIONS AND FUTURE WORK

In order to analyze errors in big data sets from sensor network systems, a novel approach is developed with Hadoop. Firstly error classification for big data sets is presented. Secondly, the correlation between sensor network systems and the scale-free complex networks are introduced. According to each error type and the features from scale-free networks, we have proposed a Conducive strategy for detecting and locating errors in big data sets on Hadoop. With the experiment results from our Hadoop environment, it is demonstrated that 1) the proposed scale-free error detecting approach can noticeably decrease the time for fast error detection in numeric big data sets, and 2) the proposed approach achieves similar error selection ratio to non-scale-free error detection approaches. In accordance with error detection for big data sets from sensor network systems on Hadoop, the issues such as error correction, big data cleaning and recovery will be further explored.

REFERENCES

[1] S. Tsuchiya, Y. Sakamoto, Y. Tsuchimoto, and V. Lee, "Big Data Processing in Cloud Environments," FUJITSU Science and Technology J., vol. 48, no. 2, pp. 159-168, 2012
[2] "Big Data: Science in the Petabyte Era: Community Cleverness Required," Nature, vol. 455, no. 7209, p. 1, 2008.
[3]M. Armbrust, A. Fox, R. Griffith, A.D.Joseph,R. Katz, A. Konwin- ski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," Comm. the ACM, vol. 53, no. 4, pp. 50-58, 2010.
[4] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and Emerging it Platforms: Vision, Hype, and Reality for Delivering Computing As the 5th Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599616,2009.
[5] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?" IEEE Trans. Paralleland Distributed Systems, vol. 23,, no. 2, pp. 296-303, Feb.2012.
[6] S. Sakr, A. Liu, D. Batista, and M. Alomari, "A Survey of Large Scale Data Management Approaches in Cloud Environments," IEEE Comm. Surveys & Tutorials, vol. 13, no. 3, pp. 311336, Third Quarter 2011.
[7] B. Li, E. Mazur, Y. Diao, A. McGregor, and P. Shenoy, "A Platform for Scalable One-Pass Analytics Using MapReduce," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD'11), pp. 985-996,2011.
[8] R. Kienzler, R. Bruggmann, A. Ranganathan, and N. Tatbul, "Stream AsYou Go: The Case for Incremental Data Access and Processing in the Cloud," Proc. IEEE ICDE Int'l Workshop Data ManagementintheCloud(DMC'12), 2012. [9] C. Olston, G. Chiou, L. Chitnis, F. Liu, Y. Han, M. Larsson, A.Neumann, V.B.N. Rao, V. Sankarasubramanian, S. Seth, C. Tian, T. ZiCornell, and X. Wang, "Nova: Continuous Pig/Hadoop Workflows," Proc. the ACM SIGMOD Int'l Conf. Management of Data (SIGMOD'11), pp. 1081-1090,2011.
[10] K.H. Lee, Y.J. Lee, H. Choi, Y.D. Chung, and B. Moon, "Parallel Data Processing with MapReduce: A Survey," ACM SIGMOD Record, vol. 40, no. 4, pp. 11-20, 2012.
[11]"Hadoop,"http://hadoop.apache.org, accessed on March 01,2013.
[12] X. Zhang, C. Liu, S. Nepal, and J. Chen, "An Efficient Quasi-Iden-tifierIndex Based Approach for Privacy Preservation over Incre- mental Data Sets on Cloud," J. Computer and System Sciences,vol. 79, pp. 542-555,2013.
[13] X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-effec- tive Privacy Preserving of Intermediate Datasets in Cloud," IEEE Trans. Parallel and Distributed Systems, vol. 24, no. 6, pp. 1192-1202, June 2013
[14] X. Zhang, T. Yang, C. Liu, and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using Systems, in MapReduce on Cloud," IEEE Trans. Parallel and Distributed,vol.25,no.2,pp.363-373, Feb. 2014.
[15] C. Liu, J. Chen, T. Yang, X. Zhang, C. Yang, R. Ranjan, and K.Kotagiri, "Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates," IEEE Trans. Parallel and Distributed Systems, vol. 25, no. 9, pp. 2234–2244, Sept. 2014.
[16] W. Dou, X. Zhang, J. Liu, and J. Chen, "HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications," IEEE Trans. Parallel and Distributed Systems, 2013.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

[17] C. Yang, X. Zhang, C. Zhong, C. Liu, J. Pei, K. Kotagiri, andJ. Chen, "A spatiotemporal compression based approach forEfficient big data processing on cloud," J. Computer and System Sciences, vol. 80, no. 8, pp. 1563–1583, 2014.
[18] J. Conhen, "Graph Twiddling in a MapReduce World," IEEE Computing in Science & Eng., vol. 11,, no. 4, pp. 29-41, 2009.
[19] K. Shim, "MapReduce Algorithms for Big Data Analysis," Proc. VLDB Endowment, vol. 5, no. 12, pp. 2016-2017, 2012.
[20] "Big Data BeyondMapReduce: Google's Big Data Papers," http://architects.dzone.com/articles/big-data-beyond-mapreduce, accessed Mar. 2013.
[21] R. Albert, H. Jeong, and A. L. Barabasi, "Error and Attack Tolerance of ComplexNetworks," Nature, vol. 406, pp. 378-382, July 2000.
[22] D.J. Wang, X. Shi, D.A. Mcfarland, and J. Leskovec,"Measurement Error in Network Data: A Re-Classification," Social Networks, vol. 34, no. 4, pp. 396-409, Oct. 2012.
[23] D. Xiong, M. Zhang, and H. Li, "Error Detection for Statistical Machine Translation Using Linguistic Features," Proc. 48th Ann. Meeting of the Association for Computational Linguistics (ACL'10),
pp. 604-611, 2010.
[24] S. Mukhopadhyay, D. Panigrahi, and S. Dey, "Model Based Error Correction for Wireless Sensor Networks," IEEE Trans. Mobile Computing, vol. 8, no. 4, pp. 528-543, Sept. 2008.
[25] S. Slijepcevic, S. Megerian, and M.Potkonjak, "Charaterization of Lacation Error in Wireless Sensor Networks: Analysis and Application," Proc. the Second Int'l Conf. Information Processing in Sensor Networks (IPSN '03), pp. 593-608, 2003.
[26] K. Ni, N. Ramanathan, M.N.H. Chehade, L. Balzano, S. Nair, S. Zahedi, G. Pottie, M. Hansen, M. Srivastava, and E. Kohler,"Sensor Network Data Fault Types," ACM Trans. Sensor Networks,vol. 5, no. 3, article 25, May 2009.
[27] A. Alamri, W.S. Ansari, M.M. Hassan, M.S. Hossain, A. Alelaiwi, and M.A. Hossain, "A Survey on Sensor-Cloud: Architecture, Applications, and Approaches," Int'l J. Distributed Sensor Networks,
vol. 2013, pp. 1-18, 2013.