ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

TIME EFFICIENT APPROACH FOR DETECTING ERRORS IN BIG SENSOR DATA ON CLOUD

[1]Kiruthikadevi K, [2]Sabeena Parvin M, [3]Shanmugapriya R
[1]Assistant Professor, Department of Computer Science and Engineering,
Nandha College of Technology
[2,3]UG Scholar, Department of Computer Science and Engineering,
Nandha College of Technology
[1]krithime@gmail.com, [2]sabi694@gmail.com, [3]rrpriyacse@gmail.com

ABSTRACT

Big sensor data is established in both industry and scientific research applications where the data is generated with high volume and velocity it is difficult to process using database management tools or traditional data processing applications. Cloud computing provides a promising platform to support the addressing of this challenge as it provides a flexible load of massive computing, storage, and software services in a scalable manner at low cost. Some techniques have been developed in latest years for processing sensor data on cloud, such as sensor-cloud.

However, these techniques do not offer efficient support on fast detection and locating of errors in big sensor data sets. For fast data error detection in big sensor data sets, in this project, develop a novel data error detection approach which exploits the full computation possible of cloud platform and the network feature of WSN. Firstly, a set of sensor data error types are classified and defined. Based on that ordering, the network feature of a clustered WSN is introduced and analyzed to support fast error detection and location. Especially, in our proposed approach, the error detection is based on the scale-free network topology and most of detection processes can be conducted in limited temporal or spatial data blocks instead of a entire big data set. Hence the detection and location process can be dramatically accelerated.

The detection and location tasks can be scattered to cloud platform to fully exploit the computation power and massive storage. Through the research on our cloud computing platform of U-Cloud, it is demonstrated that our proposed approach can significantly decrease the time for error detection and location in big data sets generated by large scale sensor network systems with acceptable error detecting accuracy.

By doing so, it transforms the problem from analyzing real-valued sample points to binary codes, which releases the door for coding theory to be incorporated into the study of anonymous sensor networks. In addition, to moderate the limitations of previous schemes, the project proposes a zone-based node error detection scheme in big sensor networks. The main idea of the proposed scheme is to use sequential hypothesis testing to detect suspect regions in which error detection nodes are likely placed. A fast and effective error replica node detection scheme is proposed using the Successive Probability Ratio Test. The application is designed by using Microsoft Visual C# .NET 2005 as front end and MS SQL SERVER 2000 as back end.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

INTRODUCTION

Recently, we enter a new age of data explosion which brings about new challenges for big data processing. In general, big data is a group of data sets so large and complex that it becomes difficult to process with onhanddatabase management systems or traditional data handling applications. It represents the progress of the human cognitive processes, usually includes data sets with sizes beyond the skill of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time.

Big data has typical characteristics of five 'V's, volume, variety, velocity, validity and value. Big data sets come from countless areas, including meteorology, connectomics, complex physics simulations, genomics, biological study, gene analysis and environmental research According to literature, since 1980s, generated data pair its size in completely 40 months all over the world. In the year of 2012, there were 2.5 quintillion (2.5 _ 1018) bytes of data being generated every day. Hence, how to process big data has become a fundamental and critical challenge for modern society. Cloud computing provides a gifted platform
for big data processing with powerful computation skill, storage, scalability, resource reuse and low cost, and has attracted significant attention in alignment with big data.

One of important source for technical big data is the data sets collected by wireless sensor networks (WSN). Wireless sensor networks have latent of expressively enhancing people's ability to monitor and interact with their physical atmosphere. Big data set from sensors is often subject to bribery and losses due to wireless medium of statement and presence of hardware inaccuracies in the nodes. For a WSN application to realize an appropriate result, it is needed that the data received is clean, accurate, and lossless. However, effective finding and cleaning of sensor big data errors is a puzzling issue demanding innovative solutions.

EXISTING SYSTEM

The existing system introduces the concept of "interval in distinguish ability" and illustrates how the problem of statistical source privacy can be mapped to the problem of interval in distinguishability. It proposes a quantitative measure to calculate statistical source secrecy in sensor networks.

By introducing real and fake interval concept, the messages are differentiated by the proper observer. The unauthorized observer fails in distinguishing the messages and failed to find the node location.

DRAWBACKS OF EXISTING SYSTEM

- Replica attack made by closest nodes cannot be identified.
- Extra messages need to be communicated between sender and observer nodes.
- Nodes awakening time is more.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

PROPOSED SYSTEM

The proposed system work is motivated from modifying the limitations of previous schemes. In particular, the new system proposes a reputation-based trust management scheme that is designed to facilitate fast detection and revocation of bargained nodes. The key idea of our scheme is to detect untrustworthy zones and perform software attestation against nodes in these zones to notice and revoke the ones that are compromised.

Specifically, it first divides the network into a set of zones, create trust levels for each zone, and detect untrustworthy zones by using the Sequential Probability Ratio Test (SPRT).

Once a zone is determined tobe unreliable, the base station or the network operator performs software verification against all nodes in the untrustworthy zone, detects compromised nodes with subverted software modules, and physically withdraws them.

In addition, a novel mobile replica detection scheme is proposed based on the Sequential Probability Ratio Test (SPRT). The new system uses the statistic that an uncompromised mobile node should never move at speeds in excess of the system-arranged maximum speed. As a result, a benign mobile sensor node's measured speed will nearly always belessthan the system-organized maximum speed as long as it employs a speed measurement system with a low error rate.

On the other hand, model nodes are in two or more places at the same time. This makes it appear as if the fake node is moving much faster than any of the benign nodes, and thus the replica nodes' measured speeds will often be over the system-molded maximum speed.

ADVANTAGES OF THE PROPOSED SYSTEM

- The main benefit of this zone-based detection approach lies in achieving fast node compromise detection and revocation while saving the large amount of time and effort that would be incurred from using periodic software attestation.
- By detecting an entire zone at once, the system can identify the approximate source of bad behavior and react quickly, rather than waiting for a specific node to be identified.
- When multiple nodes are compromised in one zone, they can all be detected and revoked at one time.
- The proposed system validates the effectiveness, efficiency, and robustness of the scheme through analysis and simulation experiments.
- The new system finds that the main attack against the SPRT-based scheme is when replica nodes fail to provide signed location and time information for speed measurement.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

- To overcome this attack, the new system employs a quarantine defense technique to block the noncompliant nodes.

- It provides analyses of the number of speed measurements needed to make replica detection decisions, which shows is quite low, and the amount of overhead incurred by running the protocol.

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 cloud.

1. Error Detection

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

2. Error Localization

After the error design matching and error detection, it is important to locate the position and source of the discovered fault 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 fault 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.

Experiment Environment and Process

The U-Cloud system is set up as shown in Appendix C.1, available in the online supplemental material. Four types of data values collected by a real WSN (scale-free complex network system) are used as the testing data set. The total testing data set size is around 2,000,000 KB, including temperature, sound, light and vibration. Even only considering one node, four types of testing data are gathered with different frequency. In other words, the data sampling from each real world node is heterogeneous. Before the experiment, we conduct the normalization for the testing data set.

CONCLUSIONS AND FUTURE WORK

In order to detect errors in big data sets from sensor network systems, a novel approach is developed with cloud computing. 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 time-efficient strategy for detecting and locating errors in big data sets on cloud. With the experiment results from our cloud computing environment U-Cloud, it is demonstrated that 1) the proposed scale-free error detecting approach can significantly reduce 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 future, in accordance with error detection for big data sets from sensor network systems on cloud, the issues such as error correction, big data cleaning and recovery will be further explored.

REFERENCES

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. 311-336, Third Quarter 2011.

B. Li, E. Mazur, Y. Diao, A. McGregor, and P. Shenoy, "A Platform for Scalable One-Pass Analytics Using Map Reduce," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD'11), pp. 985-996, 2011.

R. Kienzler, R. Bruggmann, A. Ranganathan, and N. Tatbul, "Stream As You Go: The Case for Incremental Data Access and Processing in the Cloud," Proc. IEEE ICDE Int'l Workshop Data Management in the Cloud (DMC'12), 2012.

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.

K.H. Lee, Y.J. Lee, H. Choi, Y.D. Chung, and B. Moon, "Parallel Data Processing with Map Reduce: A Survey," ACM SIGMOD Record, vol. 40, no. 4, pp. 11-20, 2012.

X. Zhang, C. Liu, S. Nepal, and J. Chen, "An Efficient Quasi-Identifier Index Based Approach for Privacy Preservation over Incremental Data Sets on Cloud," J. Computer and System Sciences, vol. 79, pp. 542-555, 2013.

X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-effective Privacy Preserving of Intermediate Datasets in Cloud," IEEE
Trans. Parallel and Distributed Systems, vol. 24, no. 6, pp. 1192-1202,June 2013.

X. Zhang, T. Yang, C. Liu, and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using Systems, in Map Reduce on Cloud," IEEE Trans. Parallel and Distributed, vol. 25, no. 2, pp. 363-373, Feb. 2014.

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.

W. Dou, X. Zhang, J. Liu, and J. Chen, "Hire Some-II: Towards Privacy-
Aware Cross-Cloud Service Composition for Big Data Applications," IEEE Trans. Parallel and Distributed Systems, 2013.
C. Yang, X. Zhang, C. Zhong, C. Liu, J. Pei, K. Kotagiri, and J. Chen, "A spatiotemporal compression based approach for efficient big data processing on cloud," J. Computer and System
Sciences, vol. 80, no. 8, pp. 1563–1583, 2014.

J. Conhen, "Graph Twiddling in a Map Reduce World," IEEE Computing
in Science & Eng., vol. 11,, no. 4, pp. 29-41, 2009.

K. Shim, "Map Reduce Algorithms for Big Data Analysis," Proc. VLDB Endowment, vol. 5, no. 12, pp. 2016-2017, 2012.

R. Albert, H. Jeong, and A. L. Barabasi, "Error and Attack Tolerance of Complex Networks," Nature, vol. 406, pp. 378-382, July 2000.

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.

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.

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.

S. Slijepcevic, S. Megerian, and M. Potkonjak, "Characterization 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.