A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud

Damini V S
Computer Science & Engineering
AMC Engineering College Bangalore,
India

Mr. Rabindranath S.
Computer Science & Engineering
AMC Engineering College Bangalore,
India

Abstract—**Big sensor data is prevalent in both industry and scientific research applications where the data is generated with high volume and velocity it is difficult to process using on-hand 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 stack of massive computing, storage, and software services in a scalable manner at low cost. Some techniques have been developed in recent years for processing sensor data on cloud, such as sensor-cloud. However, these techniques do not provide 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 paper, we develop a novel data error detection approach which exploits the full computation potential of cloud platform and the network feature of WSN. Firstly, a set of sensor data error types are classified and defined. Based on that classification, the network feature of a clustered WSN is introduced and analyzed to support fast error detection and location. Specifically, in our proposed approach, the error detection is based on the scale-free network topology and most of detection operations can be conducted in limited temporal or spatial data blocks instead of a whole big data set. Hence the detection and location process can be dramatically accelerated. Furthermore, the detection and location tasks can be distributed to cloud platform to fully exploit the computation power and massive storage. Through the experiment on our cloud computing platform of U-Cloud, it is demonstrated that our proposed approach can significantly reduce 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, cloud computing, 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 so large and complex that it becomes difficult to process with on hand database management systems or traditional data processing 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], [4], [5], [3]. Big data has typical characteristics of five 'V's, volume, variety, velocity, veracity and value. Big data sets come from many areas, including meteorology, connectomes, complex physics simulations, genomics, biological study, gene analysis and environmental research [1], [2]. According to literature [1], [2], since 1980s, generated data doubles its size in every 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 promising platform for big data processing with powerful computation capability, storage, scalability, resource reuse and low cost, and has attracted significant attention in alignment with big data. One of important source for scientific big data is the data sets collected by wireless sensor networks (WSN). Wireless sensor networks have potential of significantly enhancing people's ability to monitor and interact with their physical environment. Big data set from sensors is often subject to corruption and losses due to wireless medium of communication and presence of hardware inaccuracies in the nodes. For a WSN application to deduce an appropriate result, it is necessary that the data received is clean, accurate, and lossless. However, effective detection and cleaning of sensor big data errors is a challenging issue demanding innovative solutions. WSN with cloud can be categorized as a kind of complex network systems [2]. In these complex network systems [1], [2], [3], [4], such as WSN and social network, data abnormality and error become an annoying issue for the real network applications [5], [6], [7]. Therefore, the question of how to find data errors in complex network systems for improving and debugging the network has attracted the interests of researchers. Some work [8], [3] has been done for big data analysis and error detection in complex networks including intelligence sensors networks. There are also some works related to complex network systems data error detection and debugging with online data processing techniques [7], [8]. Since these techniques were not designed and developed to deal with big data on cloud, they were unable to cope with current dramatic increase of data size. For example, when big data sets are encountered, previous offline methods for error detection and debugging on a single computer may take a long time and lose real time feedback. Because those offline methods are normally based on learning or mining, they often introduce high time cost during the process of data set training and pattern matching. 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 cloud to detect errors in big data sets from sensor networks. Some work has been done about processing sensor data on cloud [8], [3]. However, fast detection of data errors in big data with cloud remains challenging. Especially, how to use the computation power of cloud to quickly find and locate errors of nodes in WSN needs to be explored. Cloud computing, a disruptive trend at present, poses a significant impact on current IT industry and research communities. Cloud computing infrastructure is becoming popular because it provides an open, flexible, scalable and reconfigurable platform.

II. RELATED WORK AND PROBLEM ANALYSIS

With the fast development of modern information technology, we enter a new era of data. Hence, the technique to process big data has become a fundamental and critical challenge for modern society. Cloud computing 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], [1], [1]. Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing provides an ideal platform for big data storage, dissemination and interpreting with its massive computation power [3], [4]. In many today’s real world applications, such as social networks, complex network monitoring, the scientific analysis of protein interactions and wireless sensor networks self monitoring, it is unavoidable to encounter the problem of dealing with big data and big data streams on cloud. At present, some work has been done for processing big data with cloud. Amazon EC2 infrastructure as a service is a typical cloud based distributed system for big data processing. Amazon S3 supports distributed storage. Map Reduce [7], [1], [8], [1], [2] is adopted as a programming model for big data processing over cloud computing.

Recently, wireless sensor network systems have been used in different areas, such as environment monitoring, military, disaster warning and scientific data collection. In order to process the remote sensor data collected by WSN, sensor-cloud platform [8], [8], [9] has been developed including its definition, architecture, and applications. Due to the features of high variety, volume, and velocity, big data is difficult to process using on- hand database management tools or traditional sensor-cloud platform. Big data sets can come from complex net- work systems, such as social network and large scale sensor networks. In addition, under the theme of complex network systems, it may be difficult to develop time-efficient detecting or trouble-shooting methods for errors in big data sets, hence to debug the complex network systems in real time [1], [2], [3], [2]. Sensor-Cloud [7] is a unique sensor data storage, visualization and remote management platform that leverages powerful cloud computing technologies to provide excellent data scalability, fast visualization, and user programmable analysis. Initially, sensor-cloud was designed to support long-term deployments of Micro- Strain wireless sensors. But nowadays, sensor-cloud has been developed to support any web-connected third party device, sensor, or sensor network through a simple Open Data API.

III. ERROR TYPES IN WSN BIG DATA SETS

Many systems in nature can be described as large networks (nodes or vertices connected by links or edges): Friendship networks, Social networks, computer networks, Internet, metabolic networks, power grids, scientific citations, neural networks and large scale sensor networks. Network analysis has been troubled by the issue of measurement of error for a long time [1], [2], [3]. Before deploying an error detection approach on cloud, the error models for big data sets from wireless sensor network systems perspective should be presented first.

Under the theme of the big data sets from real world complex networks, there are mainly two types of data generated and exchanged within networks. (1) The numeric data sampled and exchanged between network nodes such as sensor network sampled data sets. (2) The text files and data logs generated by nodes such as social network data sets. In this paper, our research will focus on the error detection for numeric big data sets from complex networks. In the previous work [2], the errors of complex networks can be classified as six main types for both numeric and text data as Appendix A.1, which can be found on the Computer Society Digital Library.

IV. TIME-EFFICIENT ERROR DETECTION FOR BIG SENSOR DATA ON CLOUD

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. “fd ðn=e; tÞ” is a decision making function which determines whether the detected abnormal data is a true error. In other words, fd ðn=e; tÞ has two outputs, “false negative” for detecting a true error and “false positive” for selecting a non-error data. “fl ðn=e; tÞ” is a function for tracking and returning the original error source. Without any consideration of network features and data characteristics, the error detection algorithm needs to filter the whole big data set from the network. Whenever, an abnormality defined in Section 3 is encountered, the algorithm will call fd ðn=e; tÞ and fl ðn=e; tÞ to traverse the whole network big data set for the final decision making and error source location. However, based on the analysis of scale-free network systems, it has been proved that scale-free networks have a clustering and hierarchical topology. 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 fd ðn=e; tÞ and fl ðn=e; tÞ will be located in the clusters where the node k locates. As a result, fd ðn=e; tÞ and fl ðn=e; tÞ 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 addition, with this detection technique, cloud resources only need be distributed according to each partitioned cluster in a scale-free complex network.

V EXPERIMENTS

To verify the time efficiency and the effectiveness of our approach for detecting errors in big data with cloud, experiments are conducted on U-Could (cloud computing environment at the University of Technology Sydney) [2], [3], [4], [5], [6], [18]. There are three purposes for this experiment. 1) Demonstrate that the significant time-saving is achieved in terms of detecting errors from complex network big data 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.

A 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.

B. Experiment Result

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 following the definition in Section 3 into the normalized testing data sets 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. 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.

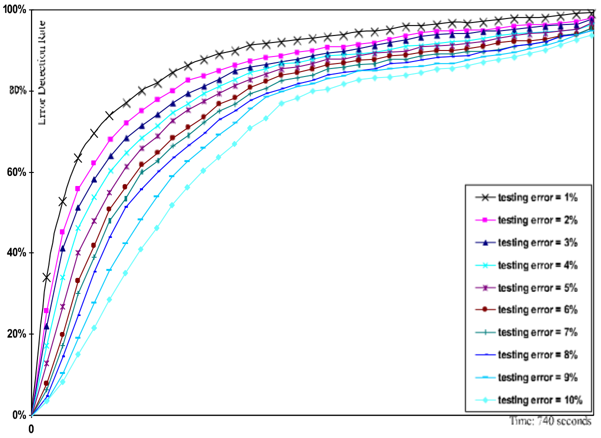

Fig.1: Time cost for detecting errors from the testing data set

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 percent in 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.

VI. 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

[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. 599-616, 2009.

[5] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?" IEEE Trans. Parallel and Distributed Systems, vol. 23,, no. 2, pp. 296-303, Feb.

[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. 311-336, Third Quarter 2011.

[7] 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.

[8] 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

[9] Workflows," Proc. the ACM SIGMOD Int'l Conf. Management of Data (SIGMOD'11), pp. 1081-1090, 2011.