International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Special Issue 3, October 2020 ||

An Efficient Detection and Recovery of Big Sensing Data— A Review

Namya Musthafa[1], Shahida K K[1], Bhavya K Bharathan [2]

Masters in Computer Science & Engineering, RCET, Akkikavu, Kerala, India[1]

Assistant Professor, RCET, Akkikavu, Kerala, India[2]

ABSTRACT*:* Big sensing knowledge is often encountered by **varied sensing systems. Sampling and transferring errors are commonly** encountered during each stage of sensing processing. How to recover from these errors with accuracy and efficiency is **quite challenging because of high sensing data** volume and an unrepeatable wireless communication environment. While big sensing data is processed in the platform provided by Cloud, however scalable and accurate error recovery solutions are still needed. In this paper, we perform a detailed review of the Detection and Recovery of erroneous data present in the sensor **data** from WSN. **This** paper **mainly** focuses on two types of WSNs, such as social network sensors and atmospheric sensors. A well defined comparative analysis was conducted to draw a conclusion on how to improve an error recovery system for sensor data from WSN.

KEYWORDS: WSN; cloud; sensor data; clustering; social network sensor; atmospheric sensor; key exchange management (key words)

I. INTRODUCTION

Recently we come across the term Big data frequently. It is not a complex rocket science but simply is a collection of data set with enormous size. Due to its massive size and extreme complexity, it has become impossible to process with existing popular database management tools or traditional data processing applications. Big data consists of all type of data regardless of whether it is structured, unstructured or semi- structured data. Big data is very much popular due to its impressive characteristics that are termed as V's of Big data. Collecting, managing, and processing big data within a tolerable elapsed time become a challenging research problem for our modern society. Data are collected from a lots means, One of the most important sources for big data comes from sensing systems which are widely deployed almost everywhere in our real world. It is well known that sensing data can have many errors and redundancies which need to be detected and fixed. The quantity of sensing data will grow day by day. In other words, more errors and confusions will be introduced. So, to discover how to manage it, what to keep and how to mine it for useful information is becoming a crucial task. In order to overcome these snags, it is necessary to recover those data errors with the support of Cloud to make big sensing data sets clean.

II. SENSOR DATA AND TYPE OF ERRORS

Sensor data is the output of sensor devices that detects and responds to some type of input from the physical environment. The output may be used to accord information or input to another system or to escort a process. Sensors are often wont to detect almost any physical element. Here are a few examples of sensors: An accelerometer detects changes in gravitational acceleration in a device it's installed in, a photo sensor detects the presence of visible light, infrared transmission (IR) and/or ultraviolet (UV) energy, Lidar, a laser-based method of detection, range finding and mapping, a charge-coupled device (CCD) stores and displays the data for an image, and so on.

One of the important sources of sensor data is Wireless sensor networks, which combine specialized transducers with a communications infrastructure for monitoring and recording conditions at diverse locations. Temperature, humidity, pressure, wind direction and speed, illumination intensity, vibration intensity, sound intensity, power-line voltage, chemical concentrations, pollutant levels, vital body functions, etc. . . are some of commonly used parameters. In this paper, we will mainly focus on WSN.

We all are fascinated by the world of social media. Even we use to access the public as well as private wifi router to not spoil our fun in the social media. Surprisingly these are sensor data, which comes under WSNs. Do you know where these data are sent to? What all the challenges they face? Now we shall dig more into its details. In

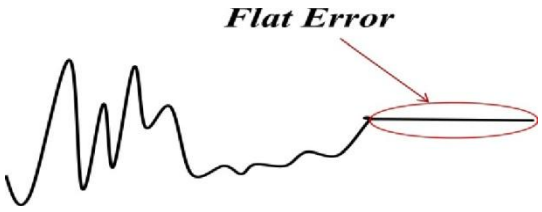
| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Special Issue 3, October 2020 ||

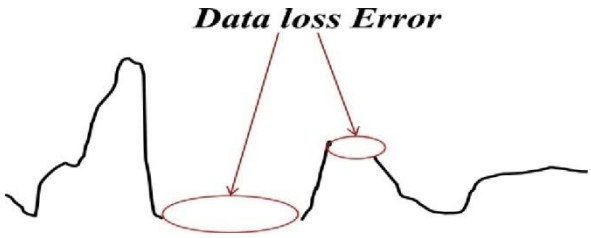
one of the studies on Social networks and sensor networks [11] we can observe how are sensors networks and social networks are correlated to each other. Sensors are becoming more massively common in mobile devices in recent years. Both the Apple iPhone, and therefore the Nokia N95 now contain GPS and accelerometer sensors. Coupled with Bluetooth and wifi communication stacks, the mobile has now become a sensor gateway for the individual. A wide range of Bluetooth sensors, like heart monitors, and environmental monitors can now be related to these mobile phones enabling a replacement paradigm, the private sensor network, in which the individual becomes the sensor hub.

These sensor data are been sent to clouds using pipelines. Some of these pipelines may be faulty and can lead to the possibility of sending erroneous data to respective clouds. Sometimes source of errors may be sensors itself or due to some external intrusion or some other reasons. The errors we encounter may vary in different aspect. In this paper we are mainly focusing on 4 distinct types of errors. Such as spike error, flat line error, data loss error and data bound error. If the data set consists of any error, they can be differentiated according to its recovery method adopted. While defining each error briefly we can guide the idea about the errors we are dealing with. They are:

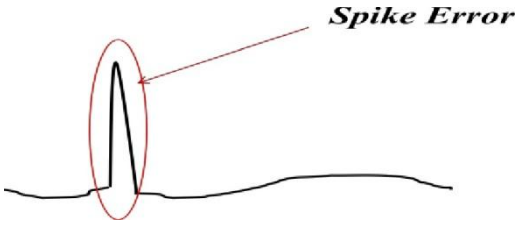
• Flat error: When the data received from the sensors remain unchanged for a while due some interior or exterior issues, they termed as flat error. These errors are difficult to detect as they have a value within reasonable range.



• Data loss error: These types of errors are relatively easier to identify as their value will be Null or zero. The difference of any two consecutively adjacent values in increasing and decreasing order must be equal as well as within bound. Otherwise there is a data loss error.



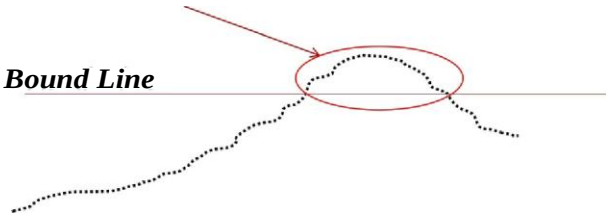
• Spike error: Similarly the spike error gives an erroneous value with sudden jump to relatively higher values. For example, for a temperature sensor current temperature value is marked as 45 degree Celsius in time slot ti and the temperature value marked at ti^1 must not exceed a difference of 5, that is 35 or 56 are recorded. If it does then there is a spike error in the signal produced by sensors.



• Bound error: This type of error is somewhat similar to spike error but here the difference is the value recorded will be beyond the bound or the range..

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Special Issue 3, October 2020 ||



III. DIFFERENT METHODS ADOPTED FOR CLEANING THE DATASET

As we know that the process of cleaning dataset can be processed in 2 main steps. First, Detection phase, in this we will determine whether any detection is present or not. Following the detection we need to localize the errors so that we can perform next phase easily. Second, Recovery phase, as the name indicates here w will recover the errors or faults which exist in the dataset with help of localized dataset. The number of phases in the task of clearing the errors in the dataset may be different in each and every method we adopt. Therefore, we can take a quick tour through some of the methods that can be useful for the above mentioned task.

First of all we can consider methods to deal with detection phase. Error detection using neural network is one among them. Neural network is used to train the data which is received from the sensors and then it detects incorrect information [1], [9]. Here the collected data is automatically stored in the cloud as encrypted data.

Next one is a common method using Hadoop platform [3], [4], [5], [7]. Here 3 inputs are provided for detection. 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 outcome of the error detection algorithm is an error set D'. As a conclusion of a previous study [6], it is stated that the combination of Map reduce algorithm and Key Exchange algorithm can enhance the security and accuracy of detection of errors in Big data.

While heading to another method we realize that here Error detection is conducted in limited time period on big data set. So that Clustering is done over the big data for error detection and localization [8]. It does not consider the whole data set instead it uses clustering. Therefore, we can achieve a faster error detection and location process comparatively. To viably deploy proposed algorithm on cloud, it is important to segment the data set before feeding to the algorithm on cloud. There are two points should be mentioned when carrying out partitioning. Firstly, the partition process could not bring new data errors into a data set or change and influence the original errors in a data set. Secondly, due to the scale-free network systems being a special topology, the partition has to form the data clusters according to the real world situation of scale free network or Cluster-head based WSN.

Till now we have gone through some of the detection methods. Now let's head towards methods for recovery. After error detection, it is crucial to pinpoint the position and source of the detected error in the original WSN graph G(V, E). The original graph of a scale-free network G(V, E), and an error data D are the respective inputs. Upon completing the process we attain an output, G"(V", E") which is the subset of the G to reveal the error location and source. The BCH decoding is complicated because it has to locate and correct the errors. Suppose we have a received codeword r(x)=r0^rlx+r2x2+....+rn- lxn-1, then r(x)m(x)+e(x), where, v(x) is correct codeword and e(x) is the error. First, we must compute a syndrome vector s- (s1,s2,. ..,s2t), which can be achieved by calculating r.HT, where, H is parity-check matrix and can be defined, hence here, o is belong to the GF field and can be pointed out in the GF table. The location numbers for the errors will be achieved by finding roots of o(x).

While considering another method we understood that 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 [3], [4], [5], [7].

Recovery can be performed by replacing the erroneous data by a new value. This new value can be predicted by different method. Prediction of time series by Euclidean distance based approximation [10], CNN, ARIMA, RNN, [12] etc...

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Special Issue 3, October 2020 ||

IV. ANALYSIS REVIEW ON RECENT METHODOLOGIES

To conduct a review analysis and comparative study on "Error detection and Recovery on Big data" we have considered several recent studies on the topic.

T. A. Sharanya, et. al. in Optimized Error Detection in Cloud User for Networking **Services[1].**

In this paper, a novel data error detection approach is developed which exploits the full computation potential of cloud platform and the network feature of WSN. In the proposed approach, the error detection is based on the scale-free topology and most of detection operations are often conducted in limited temporal or spatial data blocks instead of entire big data set. Hence the detection and location process can be fastened. Moreover, the detection and location tasks can be distributed among clouds units to completely exploiting the Computation power and massive storage. The experiment was conducted on user define cloud platform of U-Cloud, it is stated that this proposed approach can significantly reduce the time for error detection and site in big data sets produced by large scale sensor network systems with understandable error detecting accuracy.

They specifically aim to develop a completely unique error detection approach by exploiting the massive storage, scalability and computation power of cloud to detect errors in big data sets from sensor networks.

Fast detection of knowledge 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 must be explored.

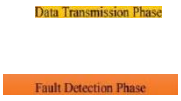
As an important scientific big data source, scientific sensor systems and wireless sensor network applications produce a spread of huge data sets in real time through various monitored activities in different applications. Data errors are unavoidable in real life scenarios. Due to rapid increase of such data it became a crucial challenge to seek out and locate the errors in big data sets with normal computing and network systems. The classification for errors is based on social network's error scenarios analysis. Robustness of four node-level network measures is considered, that is, clustering coefficient, network constraint, and centrality. It performs as a relevant system for error finding and detecting techniques for social networks. The error models and kinds presented in are often extended for the errors in complex network systems. Xiong proposed an approach which may be wont to detect the text data errors in data sets of social network. A model based error correction method for WSN. It is conducted over intelligent sensor network itself. The work is in- network fast error detection by intelligent sensors, its processing capability and time performance are extremely limited when encountering big data sets. Similar work can also be conducted with the consideration of knowledge awareness and low cost according to the description and identify their location, then analyze them. The primary goal of this location error analysis is to demonstrate the sensible use of the location errors for optimal resource consumption.

D. P. Mishra, **et. at. in Fault Tolerant Clustering** Mechanism for **Wireless** Sensor **Network [2].**

Wireless sensor networks consist of a large number of tiny sensor nodes deployed in harsh environment for unattended operation to sense and forward some data to base station through either single-hop or multi-hop transmission since self- organized capabilities are the major peculiarities of sensor nodes. Fault is an unintended defect that ultimately channelizes to the explanation for a mistake. Error is a sign of false (incorrect) state of the system. Imperfection quality of the system state caused by error, ultimately leads to the failure. A failure is the condition where the system becomes ineffective to perform the intended regulated functionalities, due to error. First step to create a WSN fault tolerant system will closely relate various faults; inspect the variability and nature of faults. WSN faults are categorized into three major categories and that they are Sensor reading faults, Software faults and Hardware faults.

Fault prevention is an act of pre-assessment and finding of an abnormal fault causing activities that usually takes place in WSN applications. Since WSN experiences perpetual changes, stringent fault prevention enrolment might not ensure 100% prevention of fault invasion. A primary fault diagnosis system is usually needed to detect and isolate the generated faults. Fault recovery phase is that the primary in-charge to evacuate the consequences of faults through all the phases and that would be achieved with help of appropriate redundancy techniques. The common redundancies such as information, physical, time and software redundancies are applied at several levels.

There are four phases during this scheme — Advertising, Data Transmission, Fault Detection and Fault Recovery.



| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Special Issue 3, October 2020 ||

First phase i.e. advertising phase, the clusters are prepared and selection of cluster heads (CHs) is completed. After selection, the CHs advertise their selection to all or any neighboring or remaining nodes. All concerned nodes select their nearest CH supported the received signal strength during advertisement. Following to the step, CHs assign a TDMA schedule to their cluster members. The second phase is data transmission phase where all subordinate nodes can start sensing and transmitting data to the cluster-head. After receiving data, the cluster-head accumulate it before sending it to the Base-Station (BS). The third phase is the fault detection phase. In bellicose environments, unexpected failure of CH may dissect the network or degrade application performance. If no response comes from CH to BS or subordinate nodes within a interval, BS marks or put flag for concern CH as a faulty node and forwards this information to the remainder of the network and initiate fault recovery process. At last in the final phase, cluster head immediately starts fault recovery process right after detection. When a faulty CH node is identified, all the cluster members related to it are gradually informed about the CH failure. For the CH recovery operation, the sink node chooses a replacement CH from the cluster members, supported cluster member's sensor nodes residual energy. According to this scheme, simply replace the faulty cluster- head by subsequent highest energy node within the cluster. In proposed mechanism, normal nodes does not require any recovery but they switch them-self to lower computational mode by informing their cell managers, and back up secondary cluster heed is employed which can replace the cluster heed just in case of failure. The proposed algorithm it observed that greedy algorithm expends the maximum energy.

et. al. [3], [4], [5], [7], as a result of the study on detection methodologies we came across two distinct datasets with 4 methods to detect the errors in the respective datasets. While considering atmospheric sensor data, the detection was performed by 2 methods, first, the MapReduce algorithm [4], and second, classification method with respect to the clustering of WSN features [7]. These methods are somewhat similar to previously discussed methods but the proposed algorithms are modified to overcome the drawbacks they possessed. The steps involved in detection are exactly the same in all the methods and it is already discussed in section III. MR algorithm used in
[4] is time-efficient as well as faster. But they didn't propose a method to recover the detected errors. Similarly, classification algorithm [7] is also possessed fast performance but they also lack a recovery method or algorithm in their proposed system. The other two methods are performed on the social network sensor dataset [3], [5]. One of them made use of Hadoop [3] to perform detection and localization phases. Whereas the other introduced the social network partition algorithm in MapReduce [5]. They claim that these methods are faster but not successful in proposing the recovery methods.

when a detailed review on recovery also was conducted [8], [9], [10], and concluded that the method mentioned in [8] is good in detection but not accurate in recovery. But they could minimize the time taken for detection to a pleasing extent. Detection using neural networks was one of this study [9]. Forward Error Correction (FEC) is used in the detection and recovery of data. As a result, BCH encodes k data bits into n code bits by adding n-k parity check bits. The length of the codes is defined as $n = 2m -1$ for any integer $m > 3$. We have t as the bound of the error correction. BCH can only correct any combination of errors fewer than t in the n-bit-codes. But comparatively, they are faster than recovery using KNN. Finally, we are considering a recovery method which is an enhancement of detection on social network sensor data [5]. Here Euclidean distance-based approximation [10] is proposed to calculate a time series prediction. And this predicted value is replaced in the location where the error is detected. They are scale-free, faster but not flexible to the different erroneous environment.

V. COMPARATIVE ANALYSIS AND DISCUSSION

Table 1: Comparative analysis of existing studies

Technique	Comparative aspects	
	Advantages	*Disadvantages*
Detection usin g neural network & encrypted data.	Secure and robust	Key management complexity
Fault tolerant for cluster head	Effective compared to previous algorithms	Data errors cannot be recovered
Detection usin	Scale-free and faster	No recovery phase

g hadoop		
Detection usin g Map-Reduce algorithm	Scale-free and time efficient	No recovery phase
Detection using social network partition algorithm with Map-Reduce	Faster and efficient	Non scale-free
Detection usin g clusters	Minimize time for error detection accuracy	Recovery is not accurate
Detection usin g neural network	Better than detection using KNN	Limited range of error correction<t<2 "-1
Recovery by Euclidean distance based approximation	Scale-free and accurate	Not flexible in nature

[6] Ankur Goyal et. a1. has already presented a thorough review analysis about the different techniques of big data cloud storage and various algorithms of error detection and correction process in big data. In this paper, we have considered some of recent error detection- recovery systems to undergo the analysis.

VI. CONCLUSION AND DISCUSSION

The Big Data have acquired a major attention in current era. Sensors produce a huge amount of data day by day. These are managed and analyzed with the concept of Big Data computation. The major challenge faced by the sensor Big Data is the erroneous data present in the sensor datasets. As a result, with the evidence provided by the papers that mentioned in this paper we enhanced our study through some of recent methods for error detection and recovery. From the above study we can conclude that an ideal error recovery system is a system that is scale- free, accurate, time efficient, minimum execution time, and faster. We can achieve this goal by leverage several prediction algorithms with help of automation algorithm to choose appropriate one for the specific situation.

REFERENCES

[1] T. A. Sharanya, “Optimized Error Detection in Cloud User for Networking Services”, International Journal on Future Revolution in Computer Science & Communication Engineering, | January 2018.

[2] D. P. Mishra & Ramesh Kumar, “Fault Tolerant Clustering Mechanism for Wireless Sensor Netw” , IJEE, June 2017.

[3] Priyasree.K & Prof.Jayashubhaj.I, “Analysing Errors in a Conducive Approach on Sensor Data”, International Journal of Innovative Research in Science, Engineering and Technology , May 2016.

[4] Chinthoju Jyothsna & N ittala Swapna Suhasini, “An Approach for Detecting and Analyzing Errors of Big Sensor Data on Cloud”, International Journal of Advanced techonolgy and Innovative research, August-2016.

[5] Chi Yang & Chang Liu, “A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud”, IEEE Transactions On Parallel And Distributed Systems, February 2015.

[6] Ankur Goyal & Vinita Nareda, “A Review Studies on Burst Error Detection and Correction in Big Data”, International Journal of Science and Research , May 2017.

[7] P. Sreenivasulu, “Data Error Detection And Time Reduction For Big Sensor Data Sets”, International Journal of Future Revolution in Computer Science Engineering and Scientific Technology, July 2016.

[8] Vishakha Dattatraya Bajad, “Effectively Error Detection on Cloud by using Time Efficient Technique”, International Journal of Computer Science Trends and Technology, Oct 2016.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Special Issue 3, October 2020 ||

[9] P.Kavya & M.Ananthi "Efficient Algorithm For Detecting And Correcting Errors In Big Sensor Data", International Journal Of Research In Computer Applications And Robotics, April 2016

[10] Chi Yang, Xianghua Xu, "A Scalable Multi-Data Sources based Recursive Approximation Approach for Fast Error Recovery in Big Sensing Data on Cloud", IEEE Transactions On Knowledge And Data Engineering, 2018.

[11] John G. Breslin & Stefan Decker, "Integrating Social Networks and Sensor Networks", Integrating Social Networks and Sensor Networks, 16 January 2009

[12] Teresa Pamua, "Impact of Data Loss for Prediction of Traffic Flow on an Urban Road Using Neural Networks", IEEE transactions on intelligent transportation systems, 2018.