ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 4, May 2016

Real Time Error Detection Service for Scale Free Network Systems using MapReduce

[1] Dakshata Supadu Patil, [2]Prof. Rahul Gaikwad
[1] [2] GF's Foundation Godavari College of Engineering Jalgaon, Maharashtra, India
[1] patildakshata4@gmail.com, [2]gaikwad005@gmail.com

Abstract: ***A new era of data explosion introduces the new problems for processing big data. With the new emerging technologies and fast development of modern world it becomes essential to keep data secure. A collection of data sets so large and composite that it becomes difficult to process with traditional data processing and management applications is called Big data. Big data represents the progress of the human intellectual capability, method to capture, manage, and process the data within a elapsed time [7]. Big data is identified by the characteristics of variety, volume, velocity, value and veracity. Debates on the usage of cloud still exist even though development of cloud computing is rapid. Major concerns in the adoption of cloud computing are security of data and privacy. Real time error detection service reduce the time for error detection and location in big data sets with error detecting accuracy.***

Keywords*: Big data; development;security; privacy; time.*

I. INTRODUCTION

Big data differs from traditional data in many dimensions:

(i) Quantity of data sources (ii) Heterogeneous nature of data sources (iii) Dynamic nature of data sources that is updating rapidly (iv) Qualities of data sources varies in various aspects. Cloud computing provides a best platform for processing data which is complex. Short term usage and capacity on demand are important properties of cloud which makes it powerful for processing big data.. For processing big data applications, security is important which is provided using cloud.

Wireless sensor networks have seen extremely large advances and use in the past two decades. Starting from mining, weather and battle operations, all of these require various sensor applications. Wireless sensor network can work in remote areas without manual interference in other sources. Only one thing user has to do is to collect the data sent by the sensors and extract information from them. The characteristics of sensor nodes are as follows:

i. Resource Constraint
ii. Unknown topology before deployment
iii. Unattended and unprotected once deployed
iv. Unreliable wireless communication

Above characteristics made WSN easily exposed to attacks. Hence Providing security solutions to WSN is difficult. Wireless sensor networks have ability of significantly augmenting people's ability to interact and monitor with their physical environment. Big data set from sensors is susceptible to corruption and losses due to wireless medium of communication. To conclude a suitable result, it is necessary that the data received is lossless, clean, and precise [9]. However, cleaning and detection of sensor big data errors is a difficult concern which requires innovative solutions. Powerful real-time processing is required for WSN big data processing applications. Our proposed error detection approach is specifically optimized for finding errors in big data sets of sensor networks. The main importance of proposed detection is to achieve noticeable time performance improvement in error detection without compromising accuracy of error detection.

II. RELATED WORK AND PROBLEM ANALYSIS

For Analyzing performances of models from different aspects, related literature for detection of error, big data processing on cloud, for complex network systems will be reviewed and compared.

A. Big data processing

A widespread issue in these works is their scalability to large amounts of data. Algorithms have increased their complexity to overcome more sophisticated methods. This makes scope of algorithms limited to offline detection. The Big

Data requirements are found not only in giant corporations such as Amazon or Google, but in various small industries that require querying, storage and retrieval over very large scale systems. Algorithm for handling big data should be able enough to work for large distributed architectures and now Big Data requirements are important to general public, it is essential that algorithm can be scalable. It is now necessary to view the algorithm in parallel; using concepts such as MapReduce [4] for better improvement.

Cloud computing provides an ideal platform for propagation of big data, storage and interpreting with its massive computation power [3], [4]. it is unavoidable to encounter the problem of dealing with big data in many real world applications. Nowadays various kind of work has been done for processing big data with cloud. A typical cloud based distributed system for big data processing is Amazon EC2 infrastructure as a service. A distributed storage is supported by Amazon S3. MapReduce [8], is adopted as a programming model for big data processing over cloud computing. The issue of processing incremental big data is researched at various points from various perspectives.

B. WSN processing

When data from large sensor networks is need to be collected and monitored remotely sensor-Cloud is useful for many applications. For environmental monitoring, health-care, business transactions, transportation, WSN enables innovative solutions. Wireless sensor network systems have invented various solutions in different fields such as disaster monitoring, disaster warming, environmental reviewing, business development process and data collection. Sensor cloud platform has been developed to process the remote sensor data collected by WSN. Architecture of sensor cloud is useful in various applications mainly when the data is located remotely. Big data is difficult to process using on-hand database management tools because volume of big data is increasing rapidly with variety in data sets. Big data sets can come from complex network systems, such as large scale sensor networks and social network. It may be difficult to develop time efficient detecting methods for errors in big data sets in case of complex network systems. Hence to debug the complex network systems in real time, it is difficult to find tools and methods which are precise in generating results. Main challenging situation in wireless sensor networks is to provide reliable data collection and low cost. Model-based error correction provides reliability against transient errors.

C. Error detection in networks

Data error is unavoidable in many real world complex network systems. The previously done research and classification include 6 types of common errors with missing data or erroneous data. A model based error correction method for Wireless sensor network is proposed by Mukhopadhyay [11]. Intelligent sensor networks are used in this correction method. This technique is based on the correction with data trend prediction. To find the root cause of errors is as important as detecting and correcting error. To diagnose root cause of error, a tool a sensor network troubleshooting, is used. But the things which need to be improved are user interface, scalability and time performance. Complex network topology features will be explored with the computation power of cloud for error detection efficiency, low cost and scalability compared to the previous sensor data error detection and localization approach.

D. Classification of errors

In large networks nodes are connected by links or edges such networks are metabolic networks, friendship networks, Social networks, computer networks, Internet, scientific citations, neural networks. Measurement of error takes long time in large scale sensor networks. From the various research done before error classification is done as: In a network, a time series of a node remains same for unacceptable long time duration and this kind of scenario is known as “flat line faults”. Sometimes impossible data values are generated in big data processing which is known as “out of data bounds faults” scenario. Also in whole process of communication some data values are missing that scenario indicates “data lost fault”. Data cleaning is the solution in data lost faults situation. “Spike faults” indicates a rate of change much greater than expected over a short period of time which may or may not return to normal afterwards. Spike faults becomes difficult to search sometimes in large datasets because of its rate of change.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 4, May 2016

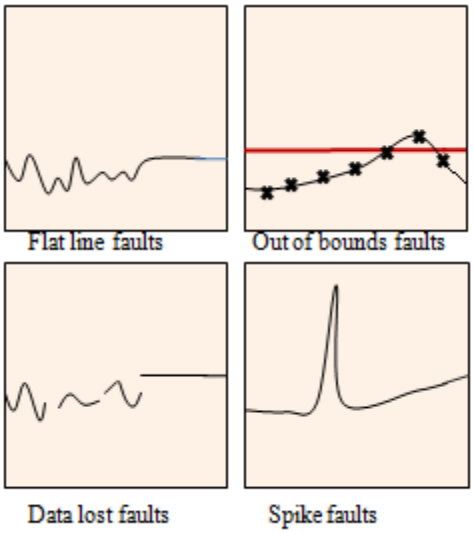

Fig. 1. Error scenarios

III. SYSTEM OVERFLOW

Real time error detection system not only focuses on error detection but it also focus on data recovery and data integrity. System overflow is shown below:

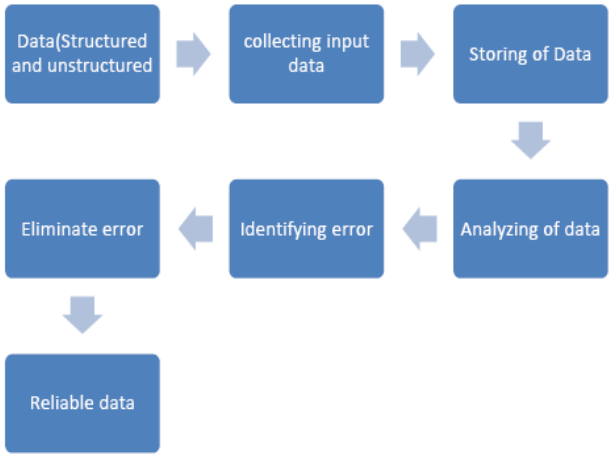

Fig. 2 Overview

MODULES-

In this project we are using 4 modules
1. User module
2. TPA Module
3. Admin module
4. Error definition module

1. User module: User register with his/her complete details. In this user can upload a file and user can process the file like update, delete or append. File details are shown to the respective user after their proper login procedure.
2. Admin module: Admin has responsibility of checking the details of the registered users. Admin also has errors information occurred in file.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 4, May 2016

3.TPA Module: Third Party Auditor(TPA) will verify the files which are processed by users, view all files which are uploaded by users. TPA can also generate token requests.

4. Error definition module: In order to test the false positive ratio of our error detection approach and time cost for error findings, we impose 4 types of data errors.

Flat Line Fault: The "flat line faults" indicates a time series of a node in a network system keeps unchanged for unacceptable long time duration.

Out of data bounds fault: The "out of data bounds faults" indicates impossible data values are observed based on some domain knowledge.

Data lost fault: The "data lost fault" means there are missing data values in a time series during the data generation or communication.

Spike faults: The "spike faults" indicates in a time series data items which are totally out of the prediction and normal changing trend.

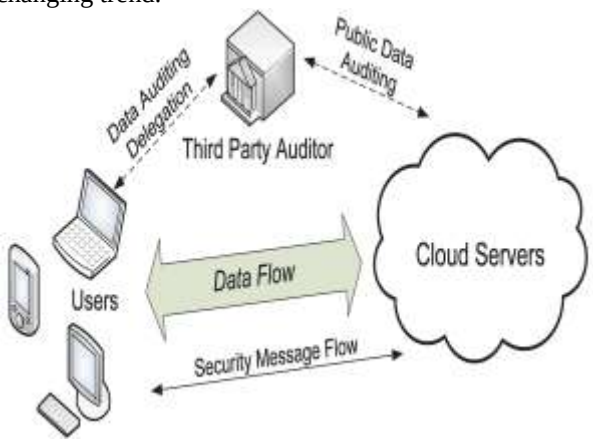

Fig. 3 General architecture

IV. ALGORITHMS

Algorithms are reviewed to deploy error detection model as well as for finding where the error is situated. Detection and location are two sections such as big data error detection/location algorithm, and its combination strategy with cloud.

A. Error Detection

Error detection algorithm has some inputs such as network graph, data sets and patterns of errors. For more secure error detection technique, Key exchange algorithm is required.

1. Key Generation algorithm

For better security purpose, Secret key generation is necessary. Secret key is generated using RSA algorithm which is provided to user at the time of login.

2. Map reduce algorithm

First we go through the mapping phase where we go over input data and create intermediate values such as records from source for data are given as input to the map function<key, value> pairs. The map generates 1 or more values and outputs a key from the given input. When the mapping phase is over Reduce phase is processed. All the intermediate values and output key are taken together into a list and given as an input to the reduce function. Reduce function combines those values into 1 or more final values for same output key. The proposed system which uses this method has to look after about:

a) Initialize a set of workers for map and reduce functions to run successfully.

b) Take input data(documents) and pass them to map function.

c) Streamline values given by map function to the reduce function.

d) Handle errors and preserves reliability of the system.

The map-reduce example is as shown below:

function map(String name, String document):

```
// name: document name
// document: document contents
for each word w in document:
emit (w, 1)
function reduce(String word, Iterator partialCounts):
// word: a word
// partialCounts: a list of aggregated partial counts
sum=0
for each pc in partialCounts:
sum+=ParseInt(pc)
emit (word, sum)
```

B. Error Localization

Finding location of error is important after detection. Error localization algorithm is used to locate position and source of error in original network graph. Error localization algorithms helps in diagnosing the root cause of error.

C. Data Recovery

Data recovery plays role as important as error detection and localization in complex network systems. Third party auditor-Public query services are offered by Trusted Third Party (TTP). TTP is responsible for storing verification parameters. In our system the Trusted Third Party, view the user data blocks and uploaded to the distributed cloud. In distributed cloud environment each cloud has user data blocks. As user has its own block then security is promising there and hence a alert is send to the user if any modification has been made in the file uploaded by the user.

1. Seed Block algorithm-

Data recovery is necessary for getting user's data back in good condition as uploaded by user. Whenever client creates the file in cloud first time, it is stored at the main cloud. When it is stored in main server, the main file of client is being EXORed with the Seed Block of the particular client. And that EXORed file is stored at the remote server in the form of file'. If either unfortunately file in main cloud damaged or file is been mistakenly deleted, then the user will get the original file by EXORing file' with the seed block of the corresponding client to produce the original file and return the resulted file i.e. original file back to the requested client.

The Remote backup services covers the following issues:

- Data Integrity
- Data security
- Data Confidentiality
- Trustworthiness
- Cost efficiency

Seed block algorithm helps in recovering the files in case of the file deletion.

V. EXPERIMENTAL EVALUATION

To test the false positive ratio of our error detection approach and time cost for error findings, we impose five types of data errors that generally occurs in data. Proposed system works for securing data , protecting it against loss and provides data recovery. Real time error detection service system mainly based on error detection service in which after uploading the file, it is divided into 3 blocks- first block, middle block and last block. After diving the file into blocks, it is verified by third party auditor and then accordingly alerts are generated to the corresponding users. Errors are detected and the information of every user is stored at the admin.

As shown in fig. 4, Proposed system reduces the computation overhead compared to existing system. Result analysis shows that Seed Block algorithm also focuses on the security concept for the back-up files stored at remote server, without using any of the existing encryption techniques. The time related issues are solved by Seed Block algorithm by taking minimum time for data recovery.

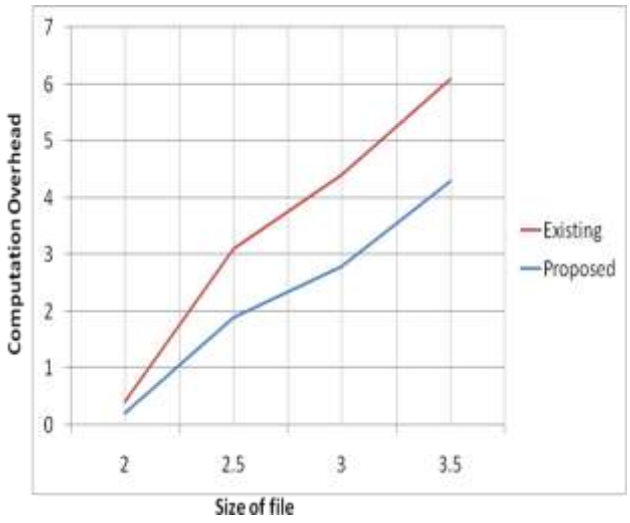

Fig. 4 Comparison graph

TABLE I. COMPARATIVE STUDY

Technique	**Table Column Head**		
	Working principle	***Advantages***	***Disadvantages***
Model based error correction method	Correction with data trend prediction	Fast error detection by intelligent sensors	Processing capability and time performance are extremely limited when encountering big data sets
Decentralized fault diagnosis system	Efficient management of WSN by diagnosing root cause	Requires minimal data collection	Scalability and performance needs to be improved
Sensor network troubleshooting tool	Diagnose root cause of error	Finds interaction bugs	Time performance and User interface needs to be improved
Real time error detection service using MapReduce	Error detection and diagnose location of error	Explores network security features with computation power of cloud	Feature of big data cleaning is to be improved

Based on the above experiment results and performance analysis, it can be stated as proposed real time error detection approach for data processing can increase the error detecting speed without losing error selecting accuracy.

VI. CONCLUSION AND FUTURE WORK

From the above analysis it is concluded that, a novel approach is proposed to preserve integrity of data. The proposed scheme is developed to reduce the computational and storage overhead of the client as well as to minimize the computational overhead of the cloud storage server. In this according to each error type, error detection strategy and error recovery strategy is developed. In accordance with the error detection and error recovery technique, big data cleaning technique will further be explored.

References

R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, ‘‘Cloud Computing and Emerging IT Platforms: Vision, Hype, Reality for Delivering Computing as the 5th Utility,’’ Future Gen. Comput. Syst., vol. 25, no. 6, June 2009.

M.Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.Katz, A.Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, andM. Zaharia, ‘‘A View of Cloud Computing,’’ Commun. ACM, vol. 53, no. 4, pp. 50-58, Apr. 2010.

Q. Wang, C.Wang, K. Ren,W. Lou, and J. Li, ‘‘Enabling Public Auditability and Data Dynamics for Storage Security in Cloud Computing,’’ IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 5, pp. 847-859, May 2011.

C. Wang, Q. Wang, K. Ren, and W. Lou, ‘‘Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing,’’ in Proc. 30st IEEE Conf. on Comput. and Commun. (INFOCOM), 2010, pp. 1-9.

G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, ‘‘Scalable and Efficient Provable Data Possession,’’ in Proc. 4th Int’l Conf. Security and Privacy in Commun. Netw. (SecureComm), 2008, pp. 1-10.

[1] Chi Yang, Chang Liu, Xuyun Zhang, Surya Nepal, and Jinjun Chen, “A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud,” IEEE Trans. Parallel and Distributed Systems, vol. 26, no. 2, February 2015.

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.

K. Shim, “MapReduce Algorithms for Big Data Analysis,” Proc. VLDB Endowment, vol. 5, no. 12, pp. 2016-2017, 2012.

M.H. Lee and Y.H. Choi, “Fault Detection of Wireless SensorNetworks,” Computer Comm., vol. 31, no. 14, pp. 3469-3475, 2008.

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.

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.

C. Liu, J. Chen, T. Yang, X. Zhang, C. Yang, K. Kotagiri, and, R. Ranjan, “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.

A. Sheth, C. Hartung, and Richard Han, “A Decentralized Fault Diagnosis System for Wireless Sensor Netowrks,” Proc. IEEE Second Conf. Mobile Ad-hoc and Sensor Systems (MASS ’05), Nov. 2005.