Locating Data Flow Anomaly With Statistical Distance

Ruoyu Wang

Abstract—[abstract]

I. INTRODUCTION

Big data industry has been blooming within this decade, reshaping the form of life and work across the globe. According to [11], 1.2ZB data had been produced from 2012 to 2014. And the amount doubles every two years. Currently, there are totally 2.7ZB data in the digital universe [1]. Big data analysis has been widely adopted in scientific experiments [12], electric business [2, 19, 3], healthcare [5], governments [10] and many other fields.

However, it will be harder in the future to harness the exploding volumes of data since problems have already appeared in data management and engineering, threatening trustworthiness and reliability of data flows inside working systems. Data error rate in enterprises is approximately 1% to 5%, and for some, even above 30% [14].

Those data anomalies may arise due to both internal and external reasons with respect to a certain system. From one hand, components inside the system may generate problematic source data. For example, in a sensor network, some sensors may generate erroneous data when it experiences power failure or other extreme conditions [13]. Data packages will be lost if sensor nodes fail to connect to network or some sensor hub goes down [6]. Also, human operators act as a heavily vulnerable part to bugs and mistakes. Some even deliberately modify system configurations for malicious compromises [15]. A study [21] found that 65% of organizations state that human errors are the main cause of data problems.

On the other hand, data manipulation [8] from outside hackers composes another potential threat of data quality and reliability. Taken Apache Hadoop as an example. It's security issues has long been discussed within communities and industry [16, 20, 9, 17]. As is shown in Figure 1, the two basic vulnerabilities: *lack of access control* and the *absence of encryption* expose the entire cluster to dangerous threats. Data flows can be intercepted and modified; services can also be altered and blocked [7]. Although there are several frameworks(e.g. Kerberos, Sentry, Knox etc.) and algorithms providing basic protection [24, 18, 22, 23, 4], clever attackers can always bypass the barriers and sneak into the core of data pipelines. Several approaches are developed as sentinels to detect probable infiltrations. However, these approaches are not able to locate corrupted data under carefully planned manipulations. Nor can they figure out the exact reasons and recover the original records. To locate and diagnose anomalies
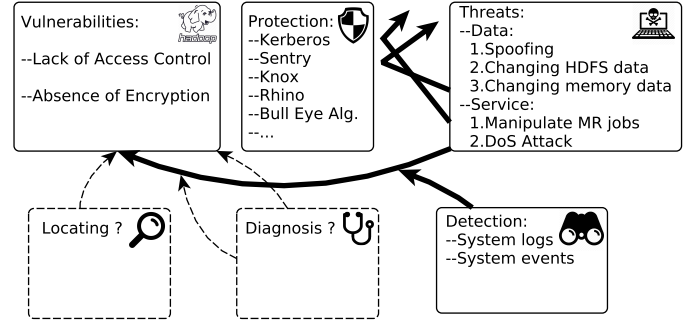

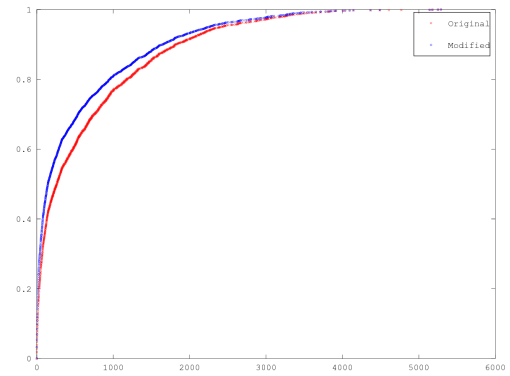Fig. 1. Security Issues of Hadoop Clusters

Fig. 2. Example Cumulative Distribution Function of Original And Modified Data

in data pipelines under carefully planned and disguised data manipulations are still on demand by industry and academia.

According to our observation, typical data manipulations on numerical data will lead to the drift of its distribution. For example, in a sampled data set from Melbourne city pedestrian volume along main streets[1], 20% records were randomly selected and modified, the value set as a tenth of the original. The empirical cumulative distribution function of the two data sets are shown in Fig. 2, where there is clearly a gap between these distributions. These anomalies inside data pipelines will severely affect mining and learning algorithms and further change the final decision given by the entire system.

In order to address issues resulted from data manipulation and pipeline errors, we propose a novel mechanism to detect and locate corrupted data within a data pipeline via statistical distance. As far as we know, this algorithm is the first attempt

[1]https://data.melbourne.vic.gov.au/Transport-Movement/Pedestrian-volume-updated-monthly-/b2ak-trbp

against data anomaly via statistical distance. Evaluations are performed on synthetic and real-world data sets, demonstrating the correctness and effectiveness of the mechanism.

The rest of the paper is organised as follow: Section II introduces recent works on data anomaly detection and pipeline error detection. Some background knowledge is provided in section III. Then in section IV, details of the detection and locating algorithm is proposed. Section V presents evaluation results and further findings of the algorithm. And all contents are concluded in section VI.

II. RELATED WORK

A. Data Anomaly Detection

B. Pipeline Error Detection

C. Provenance

III. PRELIMINARIES

A. Statistical Distance

B. TDA?

IV. ALGORITHM DETAILS

A. Technical Points

[Can be divided and settled inside later two subsections]
1) Which classifier should be chosen?
2) How to determine the classifier threshold?**[fixed value, 3σ]**
3) How to locate the compromised component?
4) How to deal with slightly drifting distribution?
5) .[to be continued . . .]

B. Basic Algorithm

C. Dynamic Algorithm

V. EVALUATION

A. Experiment Environment

B. Methodology

Raise and answer some research questions. Present test background and methods.

C. Experiment on Synthetic Data

Test basic properties of statistical distance and algorithm

D. Experiment on Real World Data

Test performance of algorithm

VI. CONCLUSION

ACKNOWLEDGEMENT

REFERENCES

[1] *Big Data Statistics And Facts for 2017*. 2017. URL: https://www.waterfordtechnologies.com/big-data-interesting-facts/.

[2] Nathan Bronson, Thomas Lento, and Janet L Wiener. "Open data challenges at facebook". In: *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE. 2015, pp. 1516–1519.

[3] Guoqiang Jerry Chen et al. "Realtime data processing at facebook". In: *Proceedings of the 2016 International Conference on Management of Data*. ACM. 2016, pp. 1087–1098.

[4] Jason C Cohen and Subrata Acharya. "Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection". In: *Journal of Information Security and Applications* 19.3 (2014), pp. 224–244.

[5] Peter Groves et al. "The'big data'revolution in healthcare: Accelerating value and innovation". In: (2016).

[6] Herodotos Herodotou et al. "Scalable near real-time failure localization of data center networks". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 1689–1698.

[7] Jingwei Huang, David M Nicol, and Roy H Campbell. "Denial-of-service threat to Hadoop/YARN clusters with multi-tenancy". In: *Big Data (BigData Congress), 2014 IEEE International Congress on*. IEEE. 2014, pp. 48–55.

[8] *Is Data Manipulation the Next Step in Cyber Crime*. URL: https://www.cloudmask.com/blog/is-data-manipulation-the-next-step-in-cybercrime.

[9] Masoumeh Rezaei Jam et al. "A survey on security of Hadoop". In: *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*. IEEE. 2014, pp. 716–721.

[10] Gang-Hoon Kim, Silvana Trimi, and Ji-Hyong Chung. "Big-data applications in the government sector". In: *Communications of the ACM* 57.3 (2014), pp. 78–85.

[11] Emmanuel Letouzé and Johannes Jütting. "Official statistics, big data and human development: towards a new conceptual and operational approach". In: *Data Pop Alliance and PARIS21* (2014).

[12] Frank Austin Nothaft et al. "Rethinking data-intensive science using scalable analytics systems". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM. 2015, pp. 631–646.

[13] Murad A Rassam, Mohd Aizaini Maarof, and Anazida Zainal. "Adaptive and online data anomaly detection for wireless sensor systems". In: *Knowledge-Based Systems* 60 (2014), pp. 44–57.

[14] Barna Saha and Divesh Srivastava. "Data quality: The other face of big data". In: *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*. IEEE. 2014, pp. 1294–1297.

[15] Felix Schuster et al. "VC3: Trustworthy data analytics in the cloud using SGX". In: *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE. 2015, pp. 38–54.

[16] Ather Sharif et al. "Current security threats and prevention measures relating to cloud services, Hadoop concurrent processing, and big data". In: *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE. 2015, pp. 1865–1870.

[17] Priya P Sharma and Chandrakant P Navdeti. "Securing big data hadoop: a review of security issues, threats and solution". In: *Int. J. Comput. Sci. Inf. Technol* 5.2 (2014), pp. 2126–2131.

[18] Biplab Sikdar. "Spatio-Temporal Correlations in Cyber-Physical Systems: A Defense Against Data Availability Attacks". In: *Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security*. ACM. 2017, pp. 103–110.

[19] Roshan Sumbaly, Jay Kreps, and Sam Shah. "The big data ecosystem at linkedin". In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM. 2013, pp. 1125–1134.

[20] Duygu Sinanc Terzi, Ramazan Terzi, and Seref Sagiroglu. "A survey on security and privacy issues in big data". In: *Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference for*. IEEE. 2015, pp. 202–207.

[21] TowerData. *4 Steps to Eliminating Human Error in Big Data*. 2013. URL: http://www.towerdata.com/blog/bid/113787/4-Steps-to-Eliminating-Human-Error-in-Big-Data (visited on 06/30/2017).

[22] Zhang Xu et al. "High fidelity data reduction for big data security dependency analyses". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, pp. 504–516.

[23] Xianqing Yu, Peng Ning, and Mladen A Vouk. "Enhancing security of Hadoop in a public cloud". In: *Information and Communication Systems (ICICS), 2015 6th International Conference on*. IEEE. 2015, pp. 38–43.

[24] Zhiyuan Zheng and AL Reddy. "Towards Improving Data Validity of Cyber-Physical Systems through Path Redundancy". In: *Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security*. ACM. 2017, pp. 91–102.