ATLANTIS PRESS
Advances in Computer Science Research, volume 74
2nd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2017)

Study on the Countermeasure of Security Problems of Big Data Caused by Data Fragmentation

Peng Ni[1, a], Shibo Xu[1, b, *], Hang Ma[2, c]

[1]China waterborne transport research institute, Beijing 100088, China

[2]Peking University, Beijing 100871, China.

[a]winbal@126.com, [b]nipeng@wti.ac.cn, [c]mahang003@126.com

Keywords: Security of big data, Data fragmentation, Security model algorithm

Abstract. The analysis and application of big data has become the frontier science at this era. People make universal attentions on the values that are existing or potential brought by the application of big data, but they make little attentions on the security threats behind the application of big data. At present, some researches have realized the security threats generated by the data leakage, its significance is not less than the application of big data. In this article, we put forward the security problems of data leakage caused by data fragmentation of big data, we also put forward some corresponding solutions and mathematical models, and finally, we conduct theory verification through the actual cases, which can effectively prevent the security problems of data leakage caused by data fragmentation.

1. Introduction

Big data is a sign of information technology in the 21st century, the government regards big data as a new method to stimulate economic growth and enhance the overall strength. But at the same time we should also see that the security problems of big data continuously emerge. Among these security issues, data fragmentation is a new and common form. For example, recently there are some applications that as long as the user inputs email address or telephone number, the application will be able to query the website that user using the email or phone number registered. This is a case of data fragmentation. Data sets of different websites exist in the form of fragmentation and are linked up via a unique identifier (email or phone number), draw a conclusion beyond the original data set and expose the users' private life information.

2. Data Fragmentation

2.1 The meaning of data fragmentation

Data that have different types, different amount and derive from different data sets are distributed on a big data platform in the broad sense in the form of fragmentation. This kind of situation is called data fragmentation. Fragmentation data may come from the same platform, it also may come from different data platforms, and there are often more complex relationships between these fragments which are difficult to integrate. We use data fragmentation network diagram in figure 1 to describe the existence form of data fragmentation.

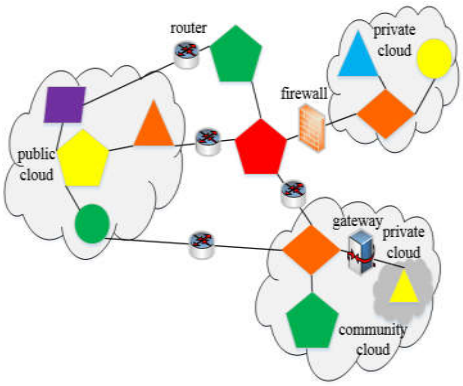

Figure. 1 Network of data fragmentation

In figure 1, we use five graphics with different kinds of shapes and colors to symbolically represent different types, different amount of data, and the data are distributed on the different clouds. There are four kinds of clouds, including private cloud, community cloud, public cloud and hybrid cloud [3]. The safety degree of cloud has negative correlation with its openness and there exists complex link relations between each cloud. In general, between public clouds, between public and community clouds, they can access to each other just needing the router conducting secure network access. And private clouds interacting with other clouds should go through a firewall. Fragmented data not only exists in the cloud, it also exists independently in other data collections, the red and green pentagons node outside the cloud are nodes of this nature.

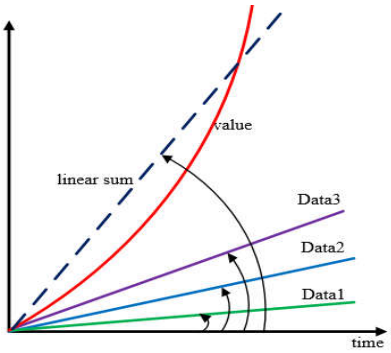

Figure. 2 The relationship of volume and value

Data fragmentation is a process of quantitative change causing the qualitative change, figure 2 describes the phenomenon. In figure 2, Data1, Data2 and Data3 represent different data sets with different sources, different data types and quantity. If the value of a data set is the same as its data leakage amount, increases linearly with the increase of time, so we get the black dotted line in the figure. But in fact, after the three data sets are crossed and validated, the value is not linearly superposition as data volumes, but takes on the form of exponential growth and transcends linear superposition sum quickly. As you see, the value of the data is not in the data itself, but in the secondary utilization after integration and interaction with other data.

2.2 The security problem of data fragmentation

The security threats of data fragmentation are just from its means of data analysis. First of all, it is hard to get useful conclusion from the data of fragmentation, only after cross integration can data have value, so big data analysis and security must be a process of balance (trade-off). Secondly, we do not know what kind of security threat can be produced by data's future secondary utilization when collecting, so we cannot carry out the limitation on laws and restrictions. Thus, the danger of data security problems caused by data fragmentation is even worse.

3. Security scheme model for Data Fragmentation

Big data analysis is an analytical method mining statistical rule and predicting through statistical rule. Therefore, fundamentally speaking, big data analysis does not be needed to find a cause-and-effect relationship and also not be emphasized on the analysis results can accurately restore

subject. Data fragmentation makes data fragments to piece together the almost complete subject, not only deviates from the direction of the large data analysis, but also brings unnecessary security threats. Therefore, with regard to data fragmentation, cleverly hiding key attributes of the subject, keeping general attributes of the subject for statistical prediction, at the same time increasing partly confusion that the error range is allowable, is the main program to solve the threat posed by data fragmentation. To this end, we introduce the Hash algorithm to achieve the key properties of the hidden body. Hash algorithm is a kind of irreversible mapping, arbitrary length digital or character is mapped to a fixed length of the digital space by an operator. In the process of data fragmentation, we use the irreversibility of Hash algorithm to make each piece of the key attributes map, which hides the key attribute information and effectively protects the data security. In addition, we take another confusion measure toward mapping, as shown in type 1.

$$result = hash(key)\%K \qquad \text{(Type 1)}$$

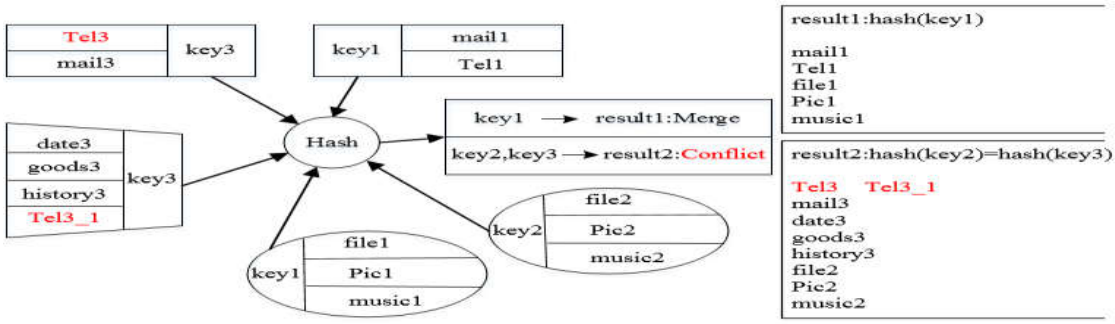

Figure. 3 The application of Hash algorithm in data fragmentation

In the type, key is the key attributes of pieces, after mapped by hash value, calculated through modulus K, and mapped for the final result. In essence, type 1 is a kind of encryption algorithm, it confuses the results caused by the irreversible Hash algorithm and produces more irreversible results in the case of not affect the analysis, thus effectively protects the data security. We describe in figure 3, pieces of data are stored in the NoSQL unstructured database and be queried by keywords.In the formula, K value can be custom, its size determines the proportion of fragmentation merging and shrinking after the mapping. Assuming that the number of original pieces is N, after operation, the debris is K, so the shrink ratio is $K/N \times 100\%$. In the diagram, we get two mapping results, one is result1 generated in key1 mapping, the data of result1 is obtained by two key1 fragments merging; another kind is result2, due to key2 and key3 are mapped to result2, therefore three fragments of key2 and key3 are merged, but the data has produced conflict in the Tel3 and Tel3_1, so whether this map is still reasonable?

The answer is yes. The reason is, first, if from the perspective of statistics, we focus on data information of all pieces, not the precise results of result2. Therefore, Tel3 and Tel3_1 can be analyzed coexistence in the data set. Second, the big data analysis is in great amount, losing a small number of data accuracy, getting the security of the data in return and obtaining a result as predicting credentials is still a very viable option.

4. Empirical case analysis

In this part, we use the classification integration of two data sets to make simulation for data fragmentation. Among that, the first data set is about surrounding temperature, humidity, electromagnetism, vibration, location and other information recorded by vehicular communication system in the process of vehicle running. Inner it, it also can be subdivided into four interrelated date sets, respectively are GPS Info, mainly recording the vehicle location information; Sender Data, mainly recording each sensor record and transmitting data; Sensor Info, mainly recording sensor-created data records information; Sensor Type, mainly recording sensor types, functions and other information. Another data set is about 20 million accommodation records, which includes the resident identity card number, name, phone number and other personal information. Among the five data sets, there is no relation between some data sets, such as 20 million data has no relation with any

other data sets; there exists strong correlation, such as GPS Info and Sender Data, which can be related by Record ID.

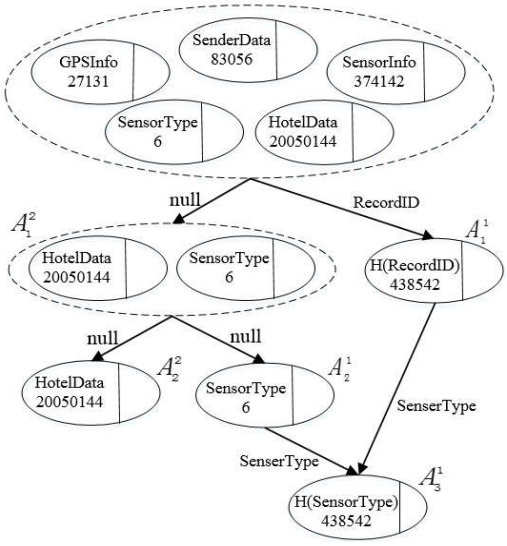

Figure. 4 Case analysis on data fragmentation

Therefore, we write a program to make simulation analysis so as to grow the decision tree (see figure 4). In the first integrated classification, by calculating the sixty-one attributes of five data sets, we come up with the conclusion that when classify and integrate in accordance with the Record ID, the information entropy production is the largest about 0.11 (for Hotel Data quantity is too big), coming up with the child node of first level. In the second classification integration, as Hotel Data has not the same attribute with Sensor Type, we finish it in different types, coming up with three nodes, respectively are: A_1^1, which is integrated from GPS Info, Sender Data and Sensor Info, amounting to 438542 data, whose curtail ratio is 90.55%; A_2^1, which is mapped from the original Sensor Type; A_2^2, which is mapped from the original Hotel Data. As for the gather formed by $\{A_1^1, A_2^1, A_2^2\}$, we make pruning and cumulating its information entropy. We find A_1^1, A_2^1 are integrated and combined by Sensor Type, which can further increase information entropy and finally get two leaf nodes A_2^2, A_3^1. Then the decision tree grows over. The strongest relation among the five data sets immediately reflects on this tree. During this process, all classification integration will base on the method of 3.2, so it effectively protects the data security and privacy.

5. Conclusion

Under the environment of data fragmentation, big data analysis is a process of balance (trade-off). We need to find the correlation among data, also ensure the data security. Experimental results show that by the method of generating the decision tree, it can effectively classify and integrate the data sets according to the strongest relation. At the same time, through the Hash algorithm, it can effectively protect the data security and realize the purpose of finding data relations and protecting data security.

Acknowledgments

This work was partly financially supported by National Science and technology project (2015BAG20B03) fund.

Reference

[1] McAfee A, Brynjolfsson E. Big data: the management revolution [J]. Harvard business review, 2012, 90 (10): 60-68.

[2] Viktor Mayer-Schonberger, Kenneth Cukier, Big Data: A Revolution that Will Transform How We Live. Work and Think, Boston: Houghton Mifflin Harcourt, 2013.

[3] Mell P and Grance T. The NIST Definition of Cloud Computing [R/OL]. http://c src. ni st. gov/group s/SNS/cloud-computing/cloud-def-v 15.doc, 2010-02-11.

[4] DASGUPTA A, KUMAR R, SARLOS T. Fast locality-sensitive hashing [C]// IEEE International Conference on Knowledge Discovery and Data Mining. 2011: 1073—1081.

[5] Kamber M, Winstone L, Gong W.Generalization and decision tree induction: efficient classification in data mining [C] / / Proc. of 1997 Int. Workshop Research Issues on Data Engineering. Birmingham: [s.n.],1997:111 -120.