International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876; Impact Factor: RJIF 5.44
www.newengineeringjournal.in
Volume 3; Issue 1; January 2017; Page No. 96-98

Information security in big data

Syed Ahad, Murtaza Alvi
Lecturer, College of Computer and Information Sciences, King Saud University, Riyadh - Saudi Arabia

Abstract
In today's world, several kinds of data is accumulated in a cloud environment as the cost of devices of information and communication technology are decreasing day by day. There is an urgent need to analyze this massive data so that it can be helpful for the business and society. A new technology needs to be adapted as the quantity of data is so massive which is far more than tens of terabytes or tens of petabytes. Also, these days, social infra-structure services run for 24 hours and 7 days a week. Hence, there is an urgent need to change the configuration of system dynamically. The current paper highlights the information security in big data.

Keywords: big data, information, technology

Introduction

Many laboratories are developing fundamental technologies for processing big data in a cloud environment. In this research, we introduce two basic technologies: Distributed Data Store & Complex Event Processing and Workflow Description for Distributed Data Processing.

The most significant point of cloud computing is that the resources and data are accumulated into data centers on the internet. These days, the cloud services like IaaS, PaaS & SaaS, have been improved in execution as application execution environments are aggregated at several levels for sharing. A new methodology has been introduced to create cloud by aggregating data. So now there is a need to change the role of cloud from application aggregation to data aggregation and utilization. A new technology other than information and communication technology is needed to use this kind of big data which is of more than tens of petabytes.

In this current research work, we will focus on two technologies, that are, distributed data store and complex event processing to process data in a cloud environment. For instance, Global Positioning System (GPS) is mounted on mobile phones with the help of which, the exact location of the user is obtained and the information with exact time is transferred to the cloud center. This big data is stored in cash registers. Then, this stored data is analyzed with the help of the time-series. Hence, the behavior like purchasing behavior of individuals is analyzed from this big data. According to a report, about 7 million pieces per second are accumulated at cloud centers. This big data is not equivalent to that is obtained in reality because of the fact that much of the data is lost while moving to the cloud centers. Many research are going on in order to reduce this data leakage. The most significant point of cloud computing is that the resources and data are accumulated into data centers on the internet. These days, the cloud services like IaaS, PaaS & SaaS, have been improved in execution as application execution environments are aggregated at several levels for sharing. Big data is large amount of data which is obtained as a result of surveys done on very large scales. This big data is not equivalent to that is obtained in reality because of the fact that much of the data is lost while moving to the cloud centers. Many research are going on in order to reduce this data leakage. During of processing of big data on clouds, a large number of read and write requests are generated and it comprises a number of application servers accessing data. So, tens to hundreds of servers are used for data storage so that if a server stops working then other servers can be used efficiently or in other situation, extra servers can be included in the main system.

One representative example of distributed data store is Distributed Key Value Store (KVS). In this technique, a data structure comprises of keys and values is distributively stored in a number of servers and one server is selected to perform read and write operations. It enhances the efficiency of the system as one server takes over the load of all kind of read and write operations and other servers are used to perform other activities. Another feature of this technique is that as data is copied in many servers so if the main server stops working then data can be accessed from other servers.

Definitions Terms

Cloud environment

Cloud environment is the better option to analyze big data as it offers benefits like temporary availability of a large number of computational resources and cost reduction by allowing resources to share data.

Data Provider

The user who owns some data that are desired by the data mining task.

Review of related literature

B. He *et al.* (2011) [1] presented four different architectures which were based on classic multi-tier database application architecture. These four architectures are: Partitioning, Replication, Distributed Control and Caching Architecture.

Ranger *et al.* (2012) [2] observed that different providers have different business models and different kinds of applications are targeted by them. For example, Google, mostly, launches small applications having light work load whereas Azure launches the applications which are efficient for medium to

large services. These days, most of the cloud service providers utilize hybrid architecture. This hybrid architecture has the potential to satisfy the actual service requirements. Kossmann *et al.* (2012) [3] proposed BoW method. In this method, MapReduce is used to cluster very large and multi-dimensional datasets. Dean *et al.* (2013) [4] proposed a method that permits the automatic and dynamic communication between Disk Delay and Network Delay. MapDup Reducer is a MapReduce based system which has the capability to detect near duplicates over massive datasets effectively.Ekanatake *et al.* (2012) [5] implemented the MapReduce framework on a number of processors in a single device. Recently, B. He at al. 2011, develop Mars which is a MapReduce framework and is based on GPS. It enhances the efficiency of the system. Jain *et al.* (2012) [6] proposed a sharing framework which is known as MRShare. MRShare is used to convert a new group that can be executed more effectively by aggregating tasks into groups and evaluating each group as a single query. John *et al.* (2012) [7] proposed a method to reduce the data transfer cost. This method divides a MapReduce task into two sub-tasks: Sampling MapReduce Task and Expected MapReduce Task. In first task, input data is obtained, keys are distributed and a good partition scheme is prepared. In second task, expected MapReduce task is used to perform the partition scheme to group the intermediate keys quickly. R. Vernica *et al.* (2014) [8] proposed a method, Twister, which is an incremented MapReduce runtime which supports Repetitive MapReduce calculations efficiently. It is used to add an extra Combine stage after Reduce stage. Thus, the output of data moves from Combine stage to next iteration's map stage. Robert *et al.* (2013) [9] proposed another method called, HaLooP, which is quite similar to Twister. HaLoop is in fact, a modified version of the MapReduce framework which supports the iterative applications by adding a 'Loop Control'. It permits to save more input and outputs during iterations. There exists a lot of iterations during the processing of data. Nylael *et al.* (2012) [10] proposed a method, Pregel, which is used to implement a programming model. In this model, each node has its own input and transfers only some messages which are needed for the next iteration to other nodes. Nature *et al.* (2014) [11] proposed a 3-stage approach for end-to-end set-similarity joins. They efficiently partition the data across nodes in order to balance the workload and minimize the need for replication. Laney *et al.* (2012) [12] investigated how to perform kNN join using MapReduce. Mappers cluster objects into groups, then Reducers perform the kNN join on each group of objects separately. Kraska *et al.* (2010) [13] proposed a method to reduce shuffling and computational costs, they design an effective mapping mechanism that exploits pruning rules for distance filtering. In addition, two approximate algorithms minimize the number of replicas to reduce the shuffling cost.

Gobioff *et al.* (2012) [14] described that PoS data analysis is also used to analyze the data information. There are many variables like Processing Event Stream (Flow) and Accumulated Numeric Data which are used to get the data like position of a constantly moving person and purchasing behavior of an individual etc. Dean *et al.* (2013) [15] described that for data analysis processing, Distributed Parallel processing with thousands of servers so as to perform statistical analysis of big data in a short time period. Borthakur *et al.* (2014) [16] described that complex event processing (CEP) refers to technology that processes and analyzes in real time complicated and massive event series that are constantly generated in real-world activities and operations. Katz *et al.* (2014) [17] acknowledged the importance of technology for dynamically distributing event processing load in cloud environments to ensure an ability to operate in real time without stopping services. Batista *et al.* (2010) [18] prototyped the basic operation of dynamic load balancing of CEP. To summarize the important points, we used a method in which the current and the extra systems run in parallel. Chen *et al.* (2011) [19] described that to ensure the order of event arrival, the manager that manages the execution environment sends instructions to the individual agents to synchronize the entire processing work at the start and end of configuration changes. This prevents the order of event arrival from being disturbed.

Chang *et al.* (2013) [20] described that critical social systems such as those for detecting signs of disasters or disaster prevention will have to run 24 hours a day, every day. Accordingly, positioned it as technology that differentiates us from our competitors and are conducting research and development.

Procedure

Anonymization technique was used so as to analyze its working in securing the big data. We identify the new challenges in privacy preserving publishing of social network data comparing to the extensively studied relational case, and examine the possible problem formulation in three important dimensions: privacy, background knowledge, and data utility. In privacy preserving data publishing, in order to prevent privacy attacks, data should be anonymized properly before it is released. Anonymization methods should take into account the privacy models of the data and the utility of the data. Generalization and perturbation are the two popular anonymization approaches for relational data. Although privacy preservation in social network data is a relatively new problem, several privacy preserving methods have been developed. Similar to privacy preservation methods in relational data, specific anonymization methods are developed for specific privacy models of social networks and specific utility goals of anonymized data.

Significance of the study

Privacy-preserving data publishing provides methods to hide identity or sensitive attributes of original data owner. Despite the many advances in the study of data anonymization, there remain some research topics awaiting to be explored. Here we highlight two topics that are important for developing a practically effective anonymization method, namely personalized privacy preservation and modeling the background knowledge of adversaries. The objective of data anonymization is to prevent the potential adversary from discovering information about a certain individual (i.e. the target). The adversary can utilize various kinds of knowledge to dig up the target's information from the published data. From previous discussions on social network data publishing and trajectory data publishing we can see that, if the data collector doesn't have a clear understanding of the capability of the adversary, i.e. the knowledge that the adversary can acquire from other resources, the knowledge which can be learned from the published data, and the way through which the knowledge can help to make an inference about target's

information, it is very likely that the anonymized data will be de-anonymized by the adversary.

References

1. B. He. The importance of 'big data': A definition, 2011.
2. C Ranger. Big data: science in the petabyte era, Nature. 2012; 455(7209):1.
3. Kossmann D, Kraska T, Loesing S. An evaluation of alternative architectures for transaction processing in the cloud, in Proceedings of the 2012 international conference on Management of data. ACM. 2012, 579-590.
4. Cordeiro F, Dean J, Ghemawat S, Hsieh W, Wallach D, Burrows M. *et al*. Big table: A distributed structured data storage system, in 7th OSDI. 2013, 305-314.
5. Ekanatake J. The hadoop distributed file system: Architecture and design, Hadoop Project Website. 2012, 11.
6. Jain. A survey of large scale data management approaches in cloud environments, Communications Surveys & Tutorials, IEEE. 2012; 13(3):311-336.
7. John Dean, Ghemawat S. Mapreduce: simplified data processing on large clusters, Communications of the ACM. 2012; 51(1):107-113.
8. Vernica R. Es2: A cloud data storage system for supporting both oltp and olap, in Data Engineering (ICDE), 2014 IEEE 27th International Conference on. IEEE. 2014, 291–302.
9. Robert, Katz R. Chukwa: A system for reliable large-scale log collection, in USENIX Conference on Large Installation System Administration. 2013, 1–15.
10. Nylael T. The Google file system," in ACM SIGOPS Operating Systems Review, ACM. 2012; 37(5):29-43.
11. Big data: science in the petabyte era, Nature. 2014; 455 (7209):1.
12. Douglas, Laney. The importance of 'big data': A definition, 2012.
13. Kossmann D, Kraska T, Loesing S. An evaluation of alternative architectures for transaction processing in the cloud, in Proceedings of the 2010 international conference on Management of data. ACM. 2010, 579–590.
14. Ghemawat S, Gobioff H, Leung S. The google file system," in ACM SIGOPS Operating Systems Review, ACM. 2012; 37(5):29–43.
15. Dean J, Ghemawat S. Mapreduce: simplified data processing on large clusters, Communications of the ACM. 2013; 51(1):107–113.
16. Borthakur D. The hadoop distributed file system: Architecture and design," Hadoop Project Website. 2014, 11.
17. Rabkin, Katz R. Chukwa: A system for reliable large-scale log collection, in USENIX Conference on Large Installation System Administration. 2014, 1–15.
18. Sakr S, Liu A, Batista D, Alomari M. A survey of arge scale data management approaches in cloud environments, Communications Surveys & Tutorials, IEEE. 2010; 13(3):311–336.
19. Cao Y, Chen C, Guo F, Jiang D, Lin Y, Ooi B et al. Es2: A cloud data storage system for supporting both oltp and olap, in Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE. 2011, 291–302.
20. Chang F, Dean J, Ghemawat S, Hsieh W, Wallach D, Burrows M. *et al*. Bigtable: A distributed structured data storage system, in 7th OSDI. 2013, 305–314.