International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue VII, July- 2016

Cluster based Anonymization and Knowledge Discovery on Big Data under Clouds

Dr. P. Sumitra,
Assistant Professor, Department of Computer Science
Ms. R. Krishnaveni,
M.Phil Research Scholar, Department of Computer Science
Vivekanandha College of Arts and Sciences for Women (Autonomous),
Elayampalayam, Tiruchengode, Tamilnadu, India

Abstract

The cloud data centers are installed to share high scalable big data values. Data analytics and mining operations are carried out with the support of big data applications and services. The big data values can be accessed by the public domain members. The big data values are composed with sensitive attributes. Data Anonymization and user identity protection are performed in the big data privacy preservation process. The data Anonymization methods are applied for the data publishing in big data environment. The big data privacy is provided with the K-anonymity, differential privacy and local recoding methods.

The big data services are protected with generalization based anonymity techniques. Proximity relations are estimated for the multiple sensitive attributes. Proximity aware clustering mechanism is adapted as local recoding scheme for the big data privacy. The initial phase of big data Anonymization is carried out with t-ancestors clustering algorithm and proximity-aware agglomerative clustering algorithm. The initial data partition is carried out using the t-Ancestor clustering algorithm with quasi identifier analysis. The local recoding process is executed to anonymized the sensitive attributes. Quasi and sensitive attributes are compared using the Proximity-aware distance measure. The clustering process is carried out with the MapReduce mechanism to improve the scalability.

The privacy preserved big data mining scheme is constructed to perform knowledge discovery on privacy preserved big data values. The system supports big data mining and analytics tasks. The service providers are identified with their performance measures. The pattern mining process is integrated with the MapReduce technique to discover the knowledge from big data values.

Index Terms: *Big Data, Cloud Computing, MapReduce, Data Privacy and Security, Anonymization Process, Clusters and Pattern Discovery*

1. Introduction

A powerful underlying and enabling concept is computing through service-oriented architectures (SOA) – delivery of an integrated and orchestrated suite of functions to an end-user through composition of both loosely and tightly coupled functions, or services – often network based. Related concepts are component-based system engineering, orchestration of different services through workflows and virtualization.

In an SOA environment, end-users request an IT service at the desired functional, quality and capacity level and receive it either at the time requested or at a specified later time. Service discovery, brokering and reliability are important and services are usually designed to interoperate, as are the composites made of these services [7]. It is expected that in the next 10 years, service-based solutions will be a major vehicle for delivery of information and other IT-assisted functions at both individual and organizational levels, e.g., software applications, web-based services, personal and business “desktop” computing, high-performance computing.

The key to a SOA framework that supports workflows is componentization of its services, an ability to support a range of couplings among workflow building blocks, fault-tolerance in its data- and process-aware service-based delivery and an ability to audit processes, data and results, i.e., collect and use provenance information.

Component-based approach is characterized by reusability, substitutability, extensibility and scalability, customizability and composability. There are other characteristics that also are very important [4]. Those include reliability and availability of the components and services, the cost of the services, security, total cost of ownership, economy of scale and so on. Many categories of components are distinguished in the context from differentiated and undifferentiated hardware, to general purpose and specialized software and applications, to real and virtual "images", to environments, to no-root differentiated resources, to workflow-based environments and collections of services and so on.

2. Related Works

Service-Oriented Computing (SOC) enables the composition of services provided with varying Quality of Service (QoS) levels in a loosely coupled way. Selecting a set of services for a optimal composition plan in terms of QoS is crucial when many functionally equivalent services are available [9]. Therefore, service composition is a classic issue in service computing domain. Quality-aware composition of web services has been fully investigated to name a few.

The authors propose a per-service-class optimization as well as a global optimization using integer programming. As opposed to integer programming a genetic algorithm based approach is proposed, where the genome length is determined by the number of abstract services that require a choice to be made. GA based approach focuses on dealing with non-linear constraints. It has the advantage that it is scalable when the number of concrete services per abstract service increases [10]. Considering that more and more functionally equivalent services are available on Internet, Alrifai et al. proposes an interesting mechanism for cutting through the search space of candidate web services, by using skyline queries offline [1]. Skyline queries identify non dominated web services on at least one QoS criteria. A non-dominated web-service means a web-service that has at least one QoS dimension where it is strictly better than any other web service and at least equal on all other QoS dimensions.

Technically, linear programming model is often recruited in service composition evaluation. In practice, various composition styles, e.g., sequential, parallel, alternative and loops can be engaged in a composition plan [11]. In this paper, we focus on investigating the sequential composition model, as other styles can be reduced or transformed into the sequential model by present mature techniques as mentioned.

Generally speaking, service composition is promoted in an open web environment. For a private cloud, the privacy and security are crucial issues in cloud service access [2]. It often leads to an awkward situation that some QoS information may be unavailable in cross-cloud composition evaluation. It is just the reason that although it is assumed that the history records can be obtained through some monitoring mechanism, there is few general QoS dataset widely recruited for testing the performance and accuracy of history record-aware service composition as mentioned.

3. Big Data Processing with MapReduce

Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial, software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data were created; The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Work with big data is necessarily uncommon; most analysis is of "PC size" data, on a desktop PC or notebook that can handle the available data set. Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data [3]. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools and expanding capabilities make Big Data a moving target. Thus, what is considered "big" one year becomes ordinary later. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations. A MapReduce program is composed of a Map() procedure that performs filtering and sorting and a Reduce() method that performs a summary operation [5]. The "MapReduce System" orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system and providing for redundancy and fault tolerance.

The models are inspired by the map and reduce functions commonly used in functional programming, although their purpose in the MapReduce framework is not the same as in their original forms [6]. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved for a variety of applications by optimizing the execution engine once. As such, a single-threaded implementation of MapReduce will usually not be faster than a traditional implementation, any gains are usually only seen with multi-threaded implementations. The use of this model is beneficial only when the optimized distributed shuffle operation and fault tolerance features of the MapReduce framework come into play. Optimizing the communication cost is essential to a good MapReduce algorithm.

MapReduce allows for distributed processing of the map and reduction operations. Provided that each mapping operation is independent of the others, all maps can be performed in parallel – though in practice this is limited by the number of independent data sources and/or the number of CPUs near each source. Similarly, a set of 'reducers' can perform the reduction phase, provided that all outputs of the map operation that share the same key are presented to the same reducer at the same time, or that the reduction function is associative [8]. While this process can often appear inefficient compared to algorithms that are more sequential, MapReduce can be applied to significantly larger datasets than "commodity" servers can handle – a large server farm can use MapReduce to sort a peta byte of data in only a few hours [12]. The parallelism also offers some possibility of recovering from partial failure of servers or storage during the operation: if one mapper or reducer fails, the work can be rescheduled – assuming the input data is still available.

4. Problem Statement

Generalization based data anonymization models are used for privacy preservation in big data services. Proximity privacy model is constructed with semantic proximity of sensitive values and multiple sensitive attributes. Local recoding mechanism is modeled as proximity aware clustering process. The clustering process is divided into two phases. They are t-ancestors clustering algorithm and proximity-aware agglomerative clustering algorithm. T-ancestor clustering algorithm splits an original data set into t partitions with quasi identifier based similar records. Data partitions are locally recoded by the proximity-aware agglomerative clustering algorithm in parallel. Proximity-aware distance measure is applied on quasi-identifier and sensitive attributes. MapReduce is integrated with the clustering process to achieve high scalability with parallel computation. The drawbacks are following from the existing system

- Big data mining operations are not integrated with the system
- Private cloud data sharing tasks are not supported
- Service composition factors are not considered
- Big data analytics are not provided

5. Cluster based Anonymization and Knowledge Discovery on Big Data

The big data sharing is achieved with privacy and security features. Big data mining operations are carried out on privacy preserved data values. MapReduce techniques are adapted for the big data mining operations. The system is divided into four major modules. They are Big Data Services, Clustering Process, Privacy Protection Process and Mining on Big Data. Big data and applications are provided under the big data services. Clustering process is applied to identify the transaction proximity levels. Privacy protection is applied on sensitive data values. Pattern discovery process is carried out under the big data mining process.

Big data provides huge volume of data items. Big data values are provided from the cloud data centers. Big data applications are constructed to process the big data values. Big data are shared with meta data values. Two phase clustering technique is applied in the system. T-ancestors clustering algorithm and proximity-aware agglomerative clustering algorithm are used for the data privacy. Proximity-aware distance measure is used to estimate the relationship levels. MapReduce technique is integrated with the clustering process.

Local recoding method is employed for the privacy preservation process. Local recoding is initiated on the clustered data values. Sensitive and quasi attributes are protected with privacy preservation process. Privacy preservation process is improved with big data analytics method. Pattern discovery process is applied on big data values. Frequent patterns are identified with association rule mining methods. Mining operations are carried out on privacy preserved big data values. Public cloud resources are allocated for the mining process.

The big data values are provided in the data centers under the cloud environment. The sensitive attributes are protected with Anonymization methods. K-anonymity models are used for the data anonymization process. It is not suited for the high scalable big data values. The data Anonymization on big data are carried out with proximity aware agglomerative cluster model. The local recoding method is adapted for the big data privacy preservation process. The data values are clustered in the initial phase. The local recoding operations are carried out on the partitioned data values.

The service provider identification is carried out with the access log information. The service provider performances are measured in the throughput and response time parameters. The big data processing tasks are performed with suitable service provider identification process. The big data mining process is initiated to discover the frequent patterns on the census data values. The Apriori algorithm is employed to discover the frequent patterns. The frequent pattern mining process is carried out on the anonymized data values. The big data mining process protects the sensitive attributes with Anonymization support.

6. Conclusion

Cloud computing environment provides resources for big data applications and services. Two phase clustering approach is adapted for proximity aware privacy preservation on big data values. The system is enhanced to perform mining and analytics on privacy preserved big data values. Service composition methods are adapted to share service components on cross cloud environment. Big data privacy is ensured with proximity aware local recoding mechanism. Data mining and analytics operations are carried out on privacy preserved big data values. Big data services are tuned to support public and private cloud environment. Service composition methods are adapted for service components selection process.

REFERENCES

[1] S. Chaudhuri, "What Next?: A Half-Dozen Data Management Research Goals For Big Data And The Cloud," in Proc. 31st Symp. Principles Database Syst., 2012, pp. 1–4.

[2] L. Wang, J. Zhan, W. Shi and Y. Liang, "In Cloud, Can Scientific Communities Benefit From The Economies Of Scale?" IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 2, pp. 296–303, Feb. 2012.

[3] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data Mining With Big Data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.

[4] X. Zhang, C. Liu, S. Nepal, S. Pandey and J. Chen, "A Privacy Leakage Upper Bound Constraint-Based Approach For Cost-Effective Privacy Preserving Of Intermediate Data Sets In Cloud," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1192–1202, Jun. 2013.

[5] B. C. M. Fung, K. Wang, R. Chen and P. S. Yu, "Privacy-Preserving Data Publishing: A Survey Of Recent Developments," ACM Comput. Survey, vol. 42, no. 4, pp. 1–53, 2010.

[6] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi, S. Khuller and A. Zhu, "Achieving Anonymity Via Clustering," ACM Trans. Algorithms, vol. 6, no. 3, 2010.

[7] X. Zhang, L. T. Yang, C. Liu and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach For Data Anonymization Using Mapreduce On Cloud," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 2, pp. 363–373, Feb. 2014.

[8] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou and J. Chen, "A Hybrid Approach For Scalable Sub-Tree Anonymization Over Big Data Using Mapreduce On Cloud," J. Comput. Syst. Sci., vol. 80, no. 5, pp. 1008–1020, 2014.

[9] Joonsang Baek, Quang Hieu Vu, Joseph K. Liu, Xinyi Huang and Yang Xiang, "A Secure Cloud Computing Based Framework for Big Data Information Management of Smart Grid", IEEE Transactions On Cloud Computing, Vol. 3, No. 2, April/June 2015

[10] Kaitai Liang, Willy Susilo and Joseph K. Liu, "Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage", IEEE Transactions On Information Forensics And Security, Vol. 10, No. 8, August 2015.

[11] S.Hemalatha and S.Alaudeen Basha, "Enabling for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud", International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013

[12] Esma Yildirim, Engin Arslan, Jangyoung Kim and Tevfik Kosar, "Application-Level Optimization of Big Data Transfers Through Pipelining, Parallelism and Concurrency", IEEE Transactions on Cloud Computing, Jan-Mar 2016.