International Journal of scientific research and management (IJSRM)
||Volume||2||Issue||12||Pages||1773-1776||2013||
Website: www.ijsrm.in ISSN (e): 2321-3418

A View on providing anonymicity using top-down specialization on Cloud and Big Data by applying Parallelization

Prateek T R[1], Nandini Prasad K S[2]

[1,2] Department of ISE
Dr. Ambedkar Institute of Technology, India.
nandiniks1@gmail.com

Abstract: *A large number of cloud services require users to share private data for data analysis or mining, bringing privacy concerns. So anonymizing of data sets via generalization is done to satisfy certain privacy requirements. At present, the scale of data in many cloud applications increases tremendously in accordance with the Big Data trend, thereby making it a challenge for commonly used software tools to capture, manage, and process such large-scale data within a tolerable elapsed time. Cloud computing provides massive computation power and storage capacity. Map reduce is a widely adopted parallel data processing framework, to address the scalability problem of the top-down specialization (TDS) approach for large-scale data anonymization. This paper provides an overview of cloud and Big Data. It describes how these concepts are used in providing anonymicity. It also compares the top-down and bottom-up specialization. We close by are sharing our opinions on what some of the important open questions are in this area as well as our thoughts on how the anonymicity algorithms can be improvised so that might best seek out answers.*

Keywords: Cloud, Anonymizing, Big data, Top-down, Scalability.

1. Introduction

Virtually everyone, ranging from big web companies to traditional enterprises to physical science researchers to social scientists, is either already experiencing or anticipating unprecedented growth in the amount of data available in their world, as well as new opportunities and great untapped value that successfully taming the "Big Data" beast will hold. Data are becoming the new raw material of business. Big Data is defined as data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage and extract value and hidden knowledge from it. Lots of data is being collected and warehoused through Web data, e-commerce, purchases at department/grocery stores, Bank/Credit Card transactions, Social network etc.

The progress and innovation is no longer hindered by the ability to collect data. But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion is important to leverage utilization. Fig. 1 indicates how data is getting accumulated in a minute.

Some of characteristics of Big data include: Sparse data and inconsistent data, Data quality (duplicate and erroneous data), Different formats (structured, semi-structure and unstructured), Different contents (e.g., images, videos, and interactive maps) and Trustworthiness (e.g., errors, integrity, reputable source).

MapReduce is a framework or a programming model that allows carrying out tasks in parallel across a large cluster of computers. It mainly consists of two functions namely Map and Reduce.

Cloud Computing is a popular topic for blogging that has been featured in various workshops, conferences, and even magazines. Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The term 'Cloud Computing' refers to applications delivered as services. The datacenter hardware and software is called a *Cloud*. *Public Cloud* is when a Cloud is made available in a pay-as-you-go manner to the general public; the service being sold is *Utility Computing*. *Private Cloud* refers to internal datacenters of a business or other organization that is not made available to the general public. Cloud computing is an evolving paradigm with tremendous momentum.

A large number of cloud services require users to share private data like electronic health records for data analysis or mining, bringing privacy concerns. Anonymizing data sets via generalization to satisfy certain privacy requirements such as k-anonymity is a widely used category of privacy preserving techniques. At present, the scale of data in many cloud applications increases tremendously in accordance with the Big Data trend, thereby making it a challenge for commonly used software tools to capture, manage, and process such large-scale data within a tolerable elapsed time. As a result, it is a challenge for existing anonymization approaches to achieve privacy preservation on privacy-

sensitive large-scale data sets due to their insufficiency of scalability. *Data anonymization* refers to hiding identity and/or sensitive data for owners of data records. Then, the privacy of an individual can be effectively preserved while certain aggregate information is exposed to data users for diverse analysis and mining.

Data anonymility uses TDS approach and uses Mapreduce which solves scalability issue by using job level and task level parallelization. *Job level parallelization* means that multiple MapReduce jobs can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand. *Task level parallelization* refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits. To make full use of the parallel capability of MapReduce on cloud can be used.

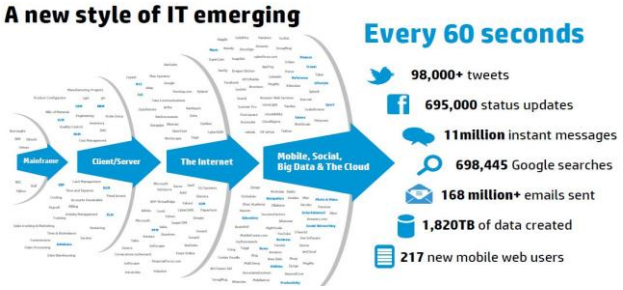

Figure 1: Data accumulation

Cloud computing, a disruptive trend at present, poses a significant impact on current IT industry and research communities .Cloud computing provides massive computation power and storage capacity via utilizing a large number of commodity computers together, enabling users to deploy applications cost-effectively without heavy infrastructure investment. Cloud users can reduce huge upfront investment of IT infrastructure, and concentrate on their own core business. However, numerous potential customers are still hesitant to take advantage of cloud due to privacy and security concerns .Privacy is one of the most concerned issues in cloud computing, and the concern aggravates in the context of cloud computing although some privacy issues are not new.

Personal data like electronic health records and financial transaction records are usually deemed extremely sensitive although these data can offer significant human benefits if they are analyzed and mined by organizations such as disease research centers. For instance, Microsoft Health Vault, an online cloud health service, aggregates data from users and shares the data with research institutes. Data privacy can be divulged with less effort by malicious cloud users or providers because of the failures of some traditional privacy protection measures on cloud .This can bring considerable economic loss or severe social reputation impairment to data owners. Hence, data privacy issues need to be addressed urgently before data sets are analyzed or shared on cloud. Data anonymization has been extensively studied and widely adopted for data privacy preservation in noninteractive data publishing and sharing scenarios.

Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. Developers with innovative ideas for new Internet services no longer require the large capital outlays in hardware to deploy their service or the human expense to operate it. They need not be concerned about over-provisioning for a service whose popularity does not meet their predictions, thus wasting costly resources, or under-provisioning for one that becomes wildly popular, thus missing potential customers and revenue. Moreover, companies with large batch-oriented tasks can get results as quickly as their programs can scale, since using 1000 servers for one hour costs no more than using one server for 1000 hours. This elasticity of resources, without paying a premium for large scale, is unprecedented in the history of IT. Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. From a hardware point of view, three aspects are new in Cloud Computing.

1. The illusion of infinite computing resources available on demand, thereby eliminating the need for Cloud Computing users to plan far ahead for provisioning.
2. The elimination of an up-front commitment by Cloud users, thereby allowing companies to start small and increase hardware resources only when there is an increase in their needs.
3. The ability to pay for use of computing resources on a shortterm basis as needed (e.g., processors by the hour and storage by the day) and release them as needed, thereby rewarding conservation by letting machines and storage go when they are no longer useful.

Armburst et.al [3] has concluded that Cloud Computing will grow, so developers should take it into account. Regardless whether a cloud provider sells services at a low level of abstraction like EC2 or a higher level like AppEngine, Moreover:

1. Applications Software needs to both scale down rapidly as well as scale up, which is a new requirement. Such software also needs a pay-for-use licensing model to match needs of Cloud Computing.
2. Infrastructure Software needs to be aware that it is no longer running on bare metal but on VMs. Moreover, billing needs to be built from the start.

Cloud computing has generated significant interest in both academia and industry, but it's still an evolving paradigm. Essentially, it aims to consolidate the economic utility model with the evolutionary development of many existing approaches and computing technologies, including distributed services, applications, and information infrastructures consisting of pools of computers, networks, and storage resources. Confusion exists in IT communities about how a cloud differs from existing models and how these differences affect its adoption. Some see a cloud as a novel technical revolution, while others consider it a natural evolution of technology, economy, and culture. Nevertheless, cloud computing is an important paradigm, with the potential to significantly reduce costs through optimization and increased operating and economic efficiencies. Furthermore, cloud computing could significantly enhance collaboration, agility, and scale, thus enabling a truly global computing model over the Internet infrastructure. However, without appropriate security and privacy solutions designed for clouds, this potentially revolutionizing computing paradigm could become a huge failure. Several surveys of potential cloud adopters indicate

that security and privacy is the primary concern hindering its adoption.

2. Objectives

Scalability is a problem of existing TDS approaches when handling large-scale data sets on cloud. The centralized TDS approach exploits the data structure to improve the scalability and efficiency by indexing anonymous data records and retaining statistical information in TIPS. The data structure speeds up the specialization process because indexing structure avoids frequently scanning entire data sets and storing statistical results circumvents recomputation overheads.

On the other hand, the amount of metadata retained to maintain the statistical information and linkage information of record partitions is relatively large compared with data sets themselves, thereby consuming considerable memory. Moreover, the overheads incurred by maintaining the linkage structure and updating the statistic information will be huge when date sets become large. Hence, centralized approaches probably suffer from low efficiency and scalability when handling large-scale data sets. There is an assumption that all data processed should fit in memory for the centralized approaches.

Unfortunately, this assumption often fails to hold in most data-intensive cloud applications nowadays. In cloud environments, computation is provisioned in the form of virtual machines (VMs). Usually, cloud compute services offer several flavors of VMs. As a result, the centralized approaches are difficult in handling large scale data sets well on cloud using just one single VM even if the VM has the highest computation and storage capability.

The main objectives of this paper includes: Scalability and Parallel computation.

Scalability: Provides high scalability by using job level and task level parallelization.Job level parallelization means that multiple MapReduce jobs can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand. Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits.

Parallel computation: To make full use of the parallel capability of MapReduce on cloud, specializations required in an anonymization process are split into phases.

3. Related Work

The basic idea behind Cloud computing is that resource providers offer elastic resources to end users. Lei et al. [2] contributions are three-fold: first, an enhanced scientific public cloud model (ESP) that encourages small or medium scale research organizations rent elastic resources from a public cloud provider; second, on a basis of the ESP model, Dawning Cloud is designed which can consolidate heterogeneous scientific workloads on a Cloud site; finally a, innovative emulation methodology .Lei et al. conducted experiments for two typical workloads: high throughput computing (HTC) and many task computing (MTC). Borkar et al. [5] have reviewed the history of systems for managing "Big Data" in the database world and the systems world, as well as examining recent "Big Data" activities and architectures, all from the perspective of three "database guys". Focus here has been on architectural issues, particularly on the components and layers that have been developed in open source and how they are being used to tackle the challenges posed by today's "Big Data" problems. They also presented the approach being taken in ASTERIX project at UC Irvine, and lead to answers to the questions regarding the "right" components and the "right" set of layers for taming the modern "Big Data" beast. Fevre et al . [6] describes that quality is best judged with respect to the workload for which the data will ultimately be used and also provides a suite of anonymization algorithms that incorporate a target class of workloads, consisting of one or more data mining tasks as well as selection predicates. In addition, Fevre et al. considered the problem of scalability and describes two extensions that allow us to scale the anonymization algorithms to datasets much larger than main memory. The first extension is based on ideas from scalable decision trees, and the second is based on sampling.

k-anonymity [7] provides a measure of privacy protection by preventing re-identification of data to fewer than a group of k data items. While algorithms exist for producing k-anonymous data, the model has been that of a single source wanting to publish data. Due to privacy issues, it is common that data from different sites cannot be shared directly. Therefore, Wei et al. presented a two-party framework along with an application that generates k-anonymous data from two vertically partitioned sources without disclosing data from one site to the other. The framework is privacy preserving in the sense that it satisfies the secure definition commonly defined in the literature of Secure Multiparty Computation.

Generalization is done for preserving privacy in publication of sensitive data. The existing methods focus on a universal approach that exerts the same amount of preservation for all persons, without catering for their concrete needs. The consequence is that we may be offering insufficient protection to a subset of people, while applying excessive privacy control to another subset. To overcome this above stated problem Xiao et al. [9] have presented a new generalization framework based on the concept of personalized anonymity. Personalized anonymity technique performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the microdata. Xiao et al. have done mathematical analysis which reveals the circumstances where the previous work fails to protect privacy, and establishes the superiority of the proposed solutions. The theoretical findings are verified with extensive experiments.

MapReduce programming model has simplified the implementation of many data parallel applications. The simplicity of the programming model and the quality of services provided by many implementations of MapReduce attract a lot of enthusiasm among distributed computing communities. From the years of experience in applying MapReduce to various scientific applications they have identified a set of extensions to the programming model and improvements to its architecture that will expand the applicability of MapReduce to more classes of applications. Ekanyake et al. [10] have developed the programming model and the architecture of Twister an enhanced MapReduce runtime that supports iterative MapReduce computations

efficiently. They have compared performance of Twister with other similar runtimes such as Hadoop and DryadLINQ for large scale data parallel applications.

Airavat, a MapReduce-based system which provides strong security and privacy guarantees for distributed computations on sensitive data. Airavat is a novel integration of mandatory access control and differential privacy. Data providers control the security policy for their sensitive data, including a mathematical bound on potential privacy violations. Users without security expertise can perform computations on the data, but Airavat confines these computations, preventing information leakage beyond the data provider's policy. The prototype designed by Roy et al. [11] is efficient, with run times on Amazon's cloud computing infrastructure within 32% of a MapReduce system with no security. Airavat is based on the popular MapReduce framework, thus its interface and programming model are already familiar to developers. Differential privacy is a new methodology for ensuring that the output of aggregate computations does not violate the privacy of individual inputs. It provides a mathematically rigorous basis for declassifying data in a mandatory access control system. Differential privacy mechanisms add some random noise to the output of a computation, usually with only a minor impact on the computation's accuracy. Airavat is the first system that integrated mandatory access control with differential privacy, enabling many privacy-preserving MapReduce computations without the need to audit untrusted code.

4. Conclusion

Cloud computing involves large-scale, distributed computations on data from multiple sources. The promise of cloud computing is based in part on its envisioned ubiquity: Internet users will contribute their individual data and obtain useful services from the cloud. MapReduce programming model has attracted a great deal of enthusiasm because of its simplicity and the improved quality of services that can be provided. Unlike the classical distributed mainly based on the availability of the computation resources, MapReduce takes a more data centered approach supporting the concept of "moving computations to data". This paper provides an overview of related work about how cloud computing and Big data can be used for the process of parallelization.

References

[1] Xuyun Zhang, Laurence T. Yang, Chang Liu, and Jinjun Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud", IEEE Transactions on Parallel and Distributed Systems, Vol. 25, No. 2, pp. 363-373, February 2014.

[2] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 2, pp.296-303, Feb. 2012.

[3] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53, no. 4, pp. 50-58, 2010.

[4] H. Takabi, J.B.D. Joshi, and G. Ahn, "Security and Privacy Challenges in Cloud Computing Environments," IEEE Security and Privacy, vol. 8, no. 6, pp. 24-31, Nov. 2010.

[5] V. Borkar, M.J. Carey, and C. Li, "Inside 'Big Data Management': Ogres, Onions, or Parfaits?," Proc. 15th Int'l Conf. Extending Database Technology (EDBT '12), pp. 3-14, 2012.

[6] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization Techniques for Large-Scale Data Sets," ACMTrans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.

[7] W. Jiang and C. Clifton, "A Secure Distributed Framework for Achieving k-Anonymity," VLDB J., vol. 15, no. 4, pp. 316-333, 2006.

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

[9] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '06), pp. 229-240, 2006.

[10] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox, "Twister: A Runtime for Iterative Mapreduce," Proc. 19th ACM Int'l Symp. High Performance Distributed Computing (HDPC '10), pp. 810-818, 2010.

[11] I. Roy, S.T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and Privacy for Mapreduce," Proc. Seventh USENIX Conf. Networked Systems Design and Implementation (NSDI '10), pp. 297-312, 2010.

Author Profile

Dr. Nandini Prasad K S received B.E degree from PESIT, M.Tech. in Computer Science and Engineering from VTU and Ph.D in Engineering from Kuvempu University. Her area of interest includes: Wireless networks, Big data, Cloud Computing, Automata and Compiler Design. She has over 13 years of teaching and industry experience. She has been awarded with "Bharath Jyothi Award" in 2012. She has presented more than 50 papers at national and international conferences in India and other countries. She is currently working as Associate Professor at Dr. AIT.

Prateek T R received B.E degree from JNNCE, Shimoga and is pursuing M.Tech (CNE) at Dr. AIT. His area of interests include: Cognitive radio, Cloud computing, Big data and Networks.