A Survey on Fast and Secured Retrieval of Data from cloud using metadata as a service

Vibha Nigam
Computer Science Department,
Bhopal MP, India

Navneet Kumar
Computer Science Department,
Bhopal MP, India

Abstract- Fast and secured data retrieval has become pivotal to the large scale distributed storage systems like cloud computing technology. The design of a distributed storage system itself is a challenging task, in particular when scalability, availability, consistency and security are required. This thesis explores three important aspects within this context and examines the role of metadata in large scale data scenario, which improves the performance of file retrieval in fast and secured manner with reduced latency and increased security. In this thesis, a generic approach of using metadata in cloud, named "MaaS – Metadata as a Service", is presented. In this approach various methodologies have been exploited for reducing the latency during data retrieval. Security aspect is also investigated and improved mechanisms have been proposed. The efficacy of the approaches is tested through experimental studies using KDD Cup 2003 dataset. In the experimental results, proposed MaaS has outperformed other methods.

Keywords – Cloud Computing; MaaS; KDD cup 2003; Security; Metadata

I. INTRODUCTION

Cloud computing a buzz word, is defined as "A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet" (Foster et al 2010). In the year 2011, a standard definition for cloud computing is provided by National Institute of Standards and Technology (NIST). According to NIST cloud is an evolving paradigm and is defined as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". This cloud model promotes availability and is composed of three service models, four deployment models and six essential characteristics (Mell & Grance2011).

1.1Cloud Service Model

The basic services offered by the cloud are categorized as

- Software as a Service,
- Platform as a Service
- Infrastructure as a Service.

1.1.1 Software as a service (SaaS)

SaaS features the pre-built software application hosted as service and is a single application delivered to thousands of users from a single vendor. The components of the application are on plug and play fashion where the users can customize their application on their own interest. This service is implemented on multi-tenancy mode where single instance of the application serves multiple end users. Each user works on their own application in an isolated environment, although they are executed from single instance. For cloud users, SaaS requires no upfront investment in servers or software licensing.

1.1.2 Platform as a service (PaaS)

PaaS offers the software platform that can be used to build higher level services i.e. the whole of the development environment is offered as a service. The platform consists of integrated OS, middleware, application software, and is provided to a customer as a service along with the environment. PaaS offerings are targeted at software developers. PaaS supplies all the resources required to build applications and services completely through Internet, without having to download or install any software.

1.1.3. Infrastructure as a service (IaaS)

IaaS delivers basic storage and computing resources as standardized services over the network. Servers, storage, resources and other systems are pooled to make the services available to users. The key characteristics that differentiate standard enterprise computing and cloud computing is that the infrastructure itself is provided as a service and is customizable. Typically, by virtualization, hardware level resources are abstracted, encapsulated and are exposed to end users effectively.

1.2 Types of Cloud

Cloud can be deployed as Public Cloud, Private Cloud, Hybrid Cloud, and Community Cloud.

1.2.1 Public cloud

In public cloud, the cloud services are offered by the third party service providers. The public can access services over internet. This type of cloud is highly scalable, where security is major drawback. The user and service provider works on the basis on the Service Level Agreements.

1.2.2 Private cloud

Private clouds are built exclusively for a single client, mostly an organization, providing the utmost control over data, security and quality of service. The organization owns the infrastructure and has complete control over the data. The private cloud is more secured when compared to the public cloud. Private clouds may be deployed in an enterprise datacenter mostly for institutions and industries. The scalability of cloud depends on the capacity of the data center upon which cloud services are hosted.

1.2.3 Hybrid cloud

Hybrid clouds are those which combine both public and private cloud models. Private clouds are benefited by hybrid clouds in case of high demand. When private cloud lacks in resources, the brokers fetch resources from public cloud and make service available to users in spite of varying workloads. Hybrid clouds work efficiently based on the service level agreements between the public and the private cloud.

1.2.4 Community cloud

Community cloud is a type of cloud used by a group of organizations like sports teams in school organizations. Any time any group of people in the community can communicate and collaborate. The members of the community share access to data and applications of their own community in cloud.

II. CLOUD METADATA ARCHITECTURE

2.1 INTRODUCTION

The continuous increase of computational power in the last decades has brought in a paradigm shift in the computing architecture scenario. The overwhelming production of data necessitated the birth of large scale data processing mechanisms. This, in turn, has forced the cloud scenario to provide for large scale storage and computing infrastructure. This new turn requires cloud computing to address challenges pertaining to scalability, consistency, economical processing of large scale data. Applications storing Exabyte's of distributed data face the problem of efficient access in the absence of appropriate networked environments and this is affecting the Quality of Service (QoS) in the cloud. Deploying such data intensive applications on cloud environment also is a daunting task. Armbrust et al (2010) argued that the growth of cloud computing applications is dependent on two factors: 1.

Availability 2. Data Con denasality. Hence, there is an acute need for an architecture that can address these problems and paves way for using data intensive applications in a cloud environment while still having maintained the QoS. Such architecture should be equipped to address the possible issues, both existing as well as futuristic such as uptime, access time, privacy, security and latency.

2.2 CLOUD METADATA ARCHITECTURE

The proposed cloud metadata architecture comprises of three layers, each layer having its own functionalities. The overall cloud metadata architecture is as shown in the Figure 3.1. The data moves from one defined layer to another and each layer isolates the portion of services. Each layer performs its job. The upper layer is the application layer which acts as an application interface. The middle layer is a vital layer called, 'MaaS'. In this layer, metadata related mechanisms are implemented. The proposed mechanism uses special data structure called cloud bloom filter. This layer plays an important role both for the application layer and the physical storage layer, by mapping the user queries from the cloud portal to the exact physical location, where the data resides. The last layer implements the physical storage data. The architecture represents the standard way of interacting with the top level layers.

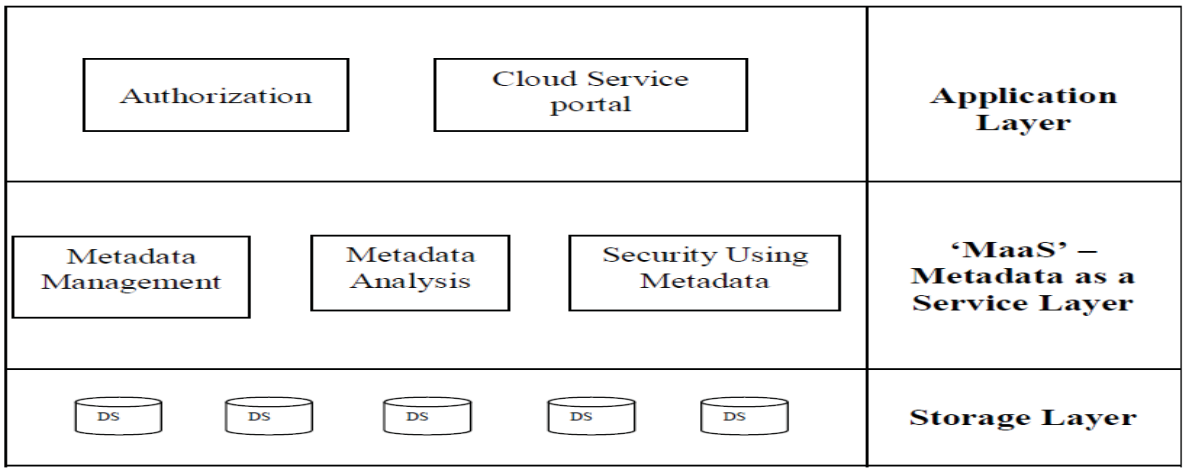

Figure 3.1 Cloud metadata architecture

III. RELATED WORK

In continuation with this attribute based encryption and key distribution mechanism, some kind of verification schemes have also been incorporated in providing high security. In this connection, Zhu et al (2012) has proposed a hashing technique called Co-operative Provable Data Possession (CPDP) for data security by means of verification of data integrity in multi cloud environment.

Wang et al (2011);Sanka t al (2010) have discussed about the key problems of this approach which includes storing data in an encrypted form, establishing access control for the encrypted data, and revoking the access rights from users when they are no longer authorized to access the encrypted data.

Damgård et al (2013) has discussed the notion of key management and distribution techniques and he also suggested that a proper key management and key distribution mechanism strengthens the security of the data stored remotely without TPA.

Due to the results of collision in the CPDP model in his continuation, Ramane & Elangovan (2012) has discussed the use of metadata in data verification scheme. He has discussed that the metadata blocks can be successfully used for verifying the data access i.e. whether the data is accessed by unauthorized users and he concludes that metadata can play a major role in providing security to the data. DePalma et al (2012) has discussed in detail about the self-

protection, which refers to the ability of a system to secure on its own without any third parties indulging into security mechanism.

A survey by researchers (Gantz & Reinsel 2011) estimates that by 2015, there will be 7.9 zeta bytes (7.9 trillion gigabytes) of stored data and, thanks to the increasing use of cloud services, nearly 20% of this data (1.4 zetabytes) will at least go through some cloud service, and around 10% (0.8 zetabytes) will be stored and maintained in the cloud. As more people and organizations now publish their data on cloud, ending relevant information in the digital universe through the use of existing search engines has been an increasingly complex challenge for end users.

IV. EXPERIMENTAL SETUP

The The proposed model is analysed by executing set of experiments. The experiments are carried out in a cloud setup using eucalyptus which is contains cloud controller and walrus as storage controller on a 5 node cluster. Each node has two 3.06 GHz Intel (R) Core TM Processors, i-7 2600, CPU @ 3.40GHZ, 4 GB of memory and 512 GB hard disks, running eucalyptus. KDD Cup 2003 dataset is used for our experiments. In our experiments, KDD Cup 2003 dataset files are uploaded in the cloud server and the performance are investigated. KDD Cup 2003 is an effective benchmark data set which is also a leading data mining and Knowledge Discovery dataset used worldwide for research purpose. KDD cup 2003 is used with the goal of identifying several techniques capable of rapidly building predictive models and scoring new entries on a large database. The KDD Cup 2003 offered the opportunity to work on huge amount of dataset which consists of several thousands of files, size ranging from 25KB to 45KB. The data set has the characteristics of maximum transactions of about 1, 15, 597 with 1057 distinct Item-set.

V. CONCLUSION

This Paper investigates a bloom filter based metadata service and discussed in detail all issues the primary issues leading to the design of MaaS model. Due to the presence of large amount of data, the proposed metadata model has a promising, prospect in this growing and challenging domain. The references made here are by no means exhaustive as many research works in this domain are still under way. This research proposes that the success of the

data retrieval in large scale distributed systems like cloud is dependent on the metadata model, which improves the efficacy of data retrieval by means of reduction in latency, increase in throughput and improved security. The present study has identified “MaaS”, Metadata as a Service for the utilization of large scale distributed systems in cloud, which provides an efficient way of retrieving data in a secured manner from huge amount of data servers dispersed geographically.

A security scheme is proposed which is best suited for cloud storage systems with huge volume of data stored and guarantees data security and user privacy during the data retrieval process. The key generation and issuing protocol is handled at various levels, User level, MDS level and DS level. The model also makes the data owner, feeling confident about the complete security of the data stored, since the encryption and decryption keys cannot be compromised without the involvement of data owner. Our security model allows key based policies to be enforced on the encrypted data stored at cloud servers. Based on the proposed scheme, the thesis presents a secure metadata model architecture that allows the user to store data securely in a cloud scenario. Analysis of the proposed scheme shows that the keys are efficiently distributed and the proposed scheme is analyzed in terms of correctness, security, and efficiency.

VI. FUTURE WORK

Thus, this paper explored the role of metadata which has a large impact on reducing latency and providing security to the data through metadata. There also exists much scope for the future work in the form of combining homomorphic encryption technique with the proposed metadata inspired retrieval model. The process of refining the security scheme is to perform data operations directly on cipher text in the cloud without the need to decrypt the data and hence improves the security. Homomorphic security in cloud storage needs a greater level of investigation, and has great future when combined with metadata.

REFERENCES

[1] Aalst, V & Wil, MP 2013, 'Decomposing petri nets for process mining: A generic approach', Journal on Distributed and Parallel Databases, vol. 31, no. 4, pp. 471-507.

[2] Abad, CL, Luu, H, Roberts, N, Lee, K, Lu, Y & Campbell, R 2012,'Metadata traces and workload models for evaluating Big storage systems.' Proc. of the IEEE/ACM International Conference on Utility and Cloud Computing, pp. 125-132.

[3] Allcock, B, Bester, J, Bresnahan, J, Chervenak, AL, Foster, I, Kesselman, C, Meder, S, Nefedova, V, Quesnel, D & Tuecke, S 2002, 'Data management and transfer in high-performance computational Grid environments', Journal of Parallel Computers, vol. 28, no.5, pp.749-771.

[4] Aloisio, G & Fiore, S 2009, ''Towards exascale distributed data management', International Journal of High Performance Computing Applications, vol. 23, no. 4, pp. 398-400.

[5] Anna, H 2008, 'Similarity measures for text document clustering', Proc. Computer science research student conference, pp. 49-56.

[6] Armbrust, M, Fox, A, Griffith, R, Joseph, AD, Katz, R, Konwinski, A & Lee, G 2010, ' A view of cloud computing', Communications of the ACM, vol.53, no. 4, pp.50-58.

[7] Ayushi, L & Shah, R 2013, 'An enhance approach for cluster analysis for large datasets', International Journal of Engineering Research & Technology (IJERT).

[8] Bloom, BH 1970, 'Space/time trade-offs in hash coding with allowable errors', ACM Communications, vol.13, no.7, pp. 422-426.

[9] Bogdanov, A & Shibutani, K 'Analysis of 3-line generalized Feistelnetworks with double SD-functions', Information Processing Letters, vol. 111, no. 13, pp. 656-660.

[10] Braam, PJ 2003, The Lustre storage architecture, 'Lustre White Paper,' http://www.lustre.org/docs.

[11] Brandt, SA, Xue, L, Miller, EL & Long, DDE 2003, 'Efficient metadata management in large distributed storage systems', Proc. International Conference on Mobile Ad-Hoc and Sensor Systems, pp. 290-298.

[12] Brine, S, Motwani, R & Silverstein, C 1997, 'Beyond market baskets: Generalizing association rules to correlations', ACM SIGMOD, vol. 26, no.2.

[13] Broder, A & Mitzenmacher, M 2004, 'Network applications of bloom filters: A survey', Internet mathematics, vol.1, no.4, pp.485-509.

[14] Burnett, K, Ng, KB & Park, S 1999, 'A comparison of the two traditions of metadata development', Journal of the American Society for Information Science, vol. 50, no.13, pp. 1209-1217.

[15] Cammert, MG, Kramer, J &Seeger, B 2007, 'Dynamic metadata management for scalable stream processing systems', Proc. IEEE International Conference on Data Engineering Workshop, pp.644-653.

[16] Chen, L & Guo, G 2011, 'An efficient remote data possession checking in cloud storage', Proc. International Journal of Digital Content Technology and its Applications, vol.5, pp.43-50.

[17] Damgård, I, Jakobsen, TP, Nielsen, JB & Pagter, JI 2013, 'Secure key management in the cloud', In Cryptography and Coding, Springer Berlin Heidelberg, pp. 270-289.

[18] DePalma, N, Hagimont, D, Boyer, F & Broto, L 2012, 'Self-protection in a clustered distributed system', IEEE Transactions on Parallel and Distributed Systems, vol.23, no. 2, pp.330-336.

[19] Dillinger, PC & Manolios, P 2004, 'Bloom filters in probabilistic verification.' In Formal Methods in Computer-Aided Design, Springer Berlin Heidelberg, pp. 367-381.