Kalpa Publications in Engineering
Volume 3, 2020, Pages 287–292
Proceedings of International Symposium on Applied Science 2019

Design Of Concept And Draft Scenarios Of DLP For DataLake Based On Abyss Distributed Storage

ByungRae Cha[1], Sun Park[1], ByeongChun Shin[2], Sun-Kuk Noh[3], JongWon Kim[1]
[1]School of EECS, GIST GwangJu, Korea
[2]Dept. of Mathematics Chonnam National Univ., angju Metrocity, Korea
[3]SW Convergence Education Institute, ChoSun Univ. GwangJu, Korea
brcha@smartx.kr

Abstract

Data loss prevention (DLP) is a set of tools and processes used to ensure that sensitive data is not lost, misused, or accessed by unauthorized users. The software and tools of DLP monitor and control endpoint activities, filter data streams on corporate networks, and monitor data in the cloud to protect data at rest, in motion, and in use. And in this study, we have designed the concept and draft scenarios of DLP for DataLake based on Abyss distributed storage cluster.

1 Introduction

Recently, the problems caused by the leakage accidents of internal information of companies, including personal information leakage accidents at domestic and external are becoming more serious. As frequent large-scale information leakage, the incidents adversely affect individual and corporate and national competitiveness, the interest and importance of internal information security such as customer information and internal core information are more increasing. Most companies in the past have centered their core capabilities on information security activities to protect internal information assets from outside malicious hackers. And the tendency to overlook the possibility of information leakage by internal employees has been, and rather, the internal employees have been considered to use the network without inconvenience. However, more than 80% of cases of industrial confidentiality leaks were found to be leaked by internal and current employees, and the damage caused by the continuous increase of internal information leaks has become a serious problem. Due to the seriousness of the issue, such as causing or irreparable damage to the external image, more attention is required to the problem.

Data Loss Prevention technology is one of the various ways to prevent information leakage by ensuring that information is stored and distributed more safely by various information protection

T.T. Truong, T.N. Tran, Q.K. Le and T.N. Nguyen (eds.), ISAS 2019 (Kalpa Publications in Engineering, vol. 3), pp. 287–292

solutions. This paper describes the design and brief scenarios of a draft concept of DLP for DataLake framework based on Abyss distributed storage.

2 Related Work

2.1 . DataLake based on Abyss Storage Cluster

DataLake framework of Abyss storage cluster is shown in Fig. 1. Abyss storage clusters are essentially SDS (Software Defined Storage) using open source Ceph.

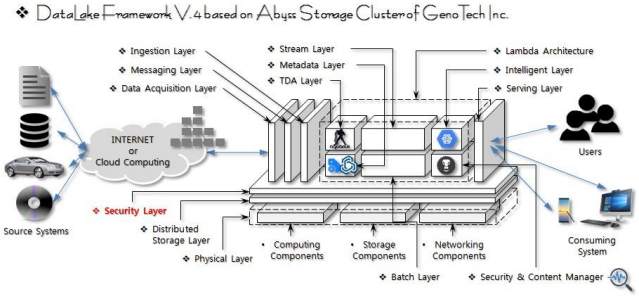

Figure 1: Diagram of DataLake framework based on Abyss storage cluster

Ceph [1, 2] is a free software storage platform designed to present object, block, and file storage from a single distributed computer cluster. Ceph's main goals are to be completely distributed without a single point of failure, scalable to the exa-byte level, and freely-available. The data is replicated, making it fault tolerant. Ceph software runs on commodity hardware. The system is designed to be both selfhealing and self-managing and strives to reduce both administrator and budget overhead.

SDS [3] is a term for computer data storage technologies which separate storage hardware from the software that manages the storage infrastructure. The software enabling a software-defined storage environment provides policy management for feature options such as deduplication, replication, thin provisioning, snapshots and backup. By definition, SDS software is separate from hardware it is managing.

DataLake Framework based on Abyss storage cluster [4, 5] considered the requirements for building an Enterprisewide Data Lake, and is configured as shown in Fig. 1. The physical layer of DataLake is located at the bottom of the DataLake Framework and consists of physical resources such as computing, storage, and networking to support the logical functions of DataLake. And in this study, Abyss storage cluster plays a role to support physical resources.

2.2 Data Loss Prevention

DLP (Data Loss Prevention) [6, 7] is a set of tools and processes used to ensure that sensitive data is not lost, misused, or accessed by unauthorized users. DLP software classifies regulated, confidential and business critical data and identifies violations of policies defined by organizations or within a predefined policy pack, typically driven by regulatory compliance such as HIPAA (Health Insurance Portability and Accountability Act) [8], PCI-DSS (Payment

Card Industry Data Security Standard) [9], or GDPR (General Data Protection Regulation) [10]. Once those violations are identified, DLP enforces remediation with alerts, encryption, and other protective actions to prevent end users from accidentally or maliciously sharing data that could put the organization at risk. DLP software and tools monitor and control endpoint activities, filter data streams on corporate networks, and monitor data in the cloud to protect data at rest, in motion, and in use. DLP

also provides reporting to meet compliance and auditing requirements and identify areas of weakness and anomalies for forensics and incident response.

2.3 C. Shamir's Secret Sharing

Shamir's Secret Sharing [11] is an algorithm in cryptography created by Adi Shamir. It is a form of secret sharing, where a secret is divided into parts, giving each participant its own unique part. To reconstruct the original secret, a minimum number of parts is required. In the threshold scheme this number is less than the total number of parts. Otherwise all participants are needed to reconstruct the original secret.

3 Draft Concept Design Of Dlp For Datalake Based On Abyss Distributed Storage

The proposed DLP method will be logically applied to the security layer of the DataLake framework as shown in Fig. 1 and applied to the media of physically distributed storage. Four scenarios of DLP are briefly described to illustrate the concept and functionality of DLP in security layer of DataLake framework.

3.1 Scenario #1 of DLP

In scenario #1 of the proposed DLP concept, as shown in Fig. 2, the input data is stored in the physical media of distributed nodes through the sequential procedure of Shared Data, Replicate Shards, and Distribute Shards.

3.2 Scenario #2 of DLP

Scenario #2 of the proposed DLP concept applies encryption to the data entered before the procedure of the proposed DLP scenario #1 and sequentially consists of Encrypt Data, Shared Data, Replicate Shards, and Distribute Shards as shown in Fig. 3

3.3 Scenario #3 of DLP

Scenario #3 of the DLP concept applies encryption to the shards entered after shard data in the procedure of the proposed DLP scenario #1 and sequentially consists of Shared Data, Encrypt Shards, Replicate Shards, and Distribute Encrypted Shards as shown in Fig. 4.

3.4 Scenario #4 of DLP

Scenario #4 of DLP in Fig. 5 has almost the same procedures and encryption as in Scenario #3. Compared to scenario #3, the Shamir's secret sharing technique is used for shard data. Instead of spending more computing resources and time than the previous three technologies, security for DLP will be enhanced.

In addition, the proposed DLP scenarios #1 ~ #4 are expected to be extensible and compatible by the Connected Data Architecture (CDA) [12] of the DataLake framework based on Abyss storage cluster.

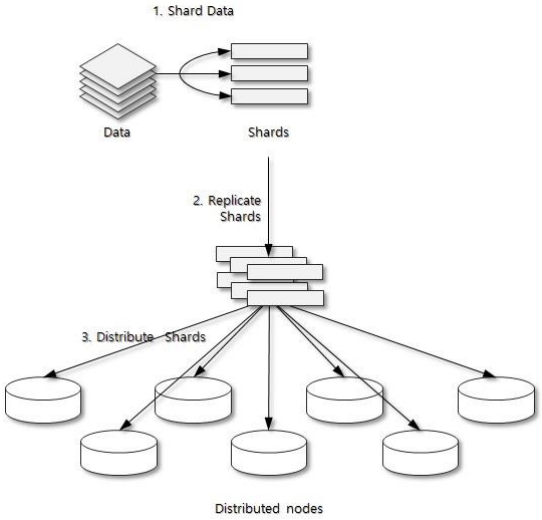

Figure 2: Scenario #1 of draft concept DLP

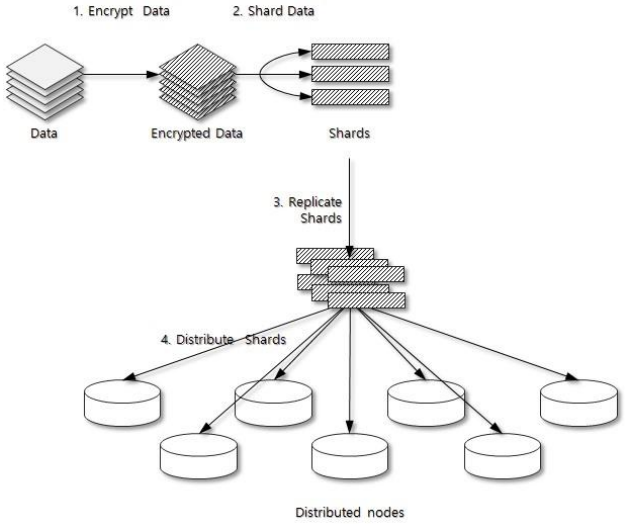

Figure 3: Scenario #2 of draft concept DLP

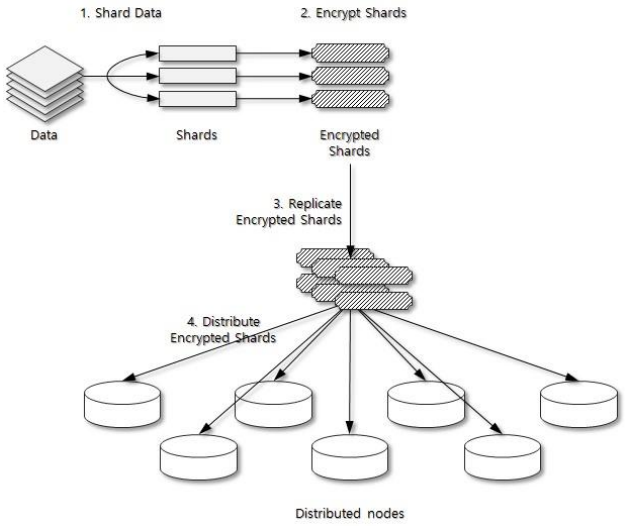

Figure 4: Scenario #3 of draft concept DLP

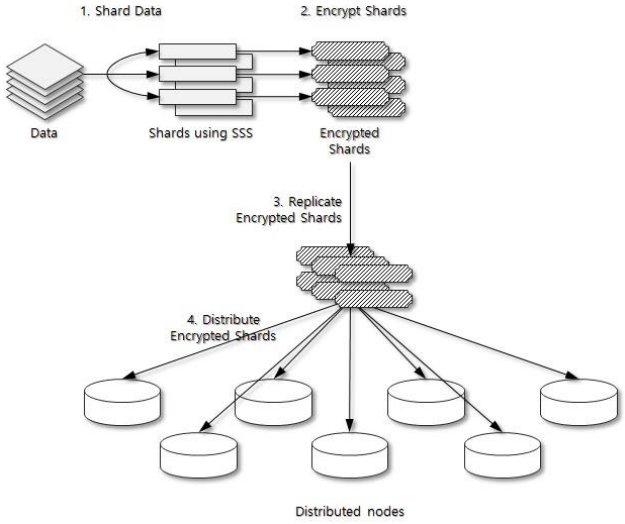

Figure 5: Scenario #4 of draft concept DLP

4 Conclusion

Recently, large-scale information leakage accidents at home and abroad adversely affect not only individuals but also corporate and national competitiveness, and the interest and importance of internal information security such as customer information and internal core information are more increasing.

DLP technology is one of the various ways to prevent the leakage of internal information of an organizations. The DLP concept designs for the DataLake based on Abyss storage cluster and four simple scenarios have been described and proposed.

Acknowledgment

This research was supported by the MIST(MInistry of Science & ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute for Information & communications Technology Promotion) (2017-0-00137).

References

Ceph, https://ceph.io/

Ceph, https://en.wikipedia.org/wiki/Ceph_(software)

Margaret Rouse, "Definition: software-defined storage," SearchSDN. Tech Target. Retrieved November 7, 2013.

ByungRae Cha, Sun Park, JongWon Kim, SungBum Pan, and JuHyun Shin, " International Network Performance and Security Testing Based on Distributed Abyss Storage Cluster and Draft of Data Lake Framework," Hindawi, Security and Communication Networks, Vol. 2018, Article ID 1746809, 2018. doi:10.1155/2018/1746809.

ByungRae Cha, Sun Park, Byeong-Chun Shin and JongWon Kim, "Draft Design of DataLake Framework based on Abyss Storage Cluster," Smart Media Journal, Vol.7, no.1, pp.9-15, March 2018.

DLP (Data Loss Prevention), https://digitalguardian. com/blog/whatdata-loss-prevention-dlp-definition-data-loss-prevention, Jan 2019.

DLP, https://en.wikipedia.org/wiki/Data_loss_prevention_software

HIPAA, https://en.wikipedia.org/wiki/Health_Insurance_Portability _and_Accountability_Act

PCI-DSS, https://en.wikipedia.org/wiki/Payment_Card_Industry _Data_Security_Standard

GDPR, https://en.wikipedia.org/wiki/General_Data_Protection_ Regulation

Shamir's Secret Sharing, https://en.wikipedia.org/wiki/Shamir%27s_ Secret_Sharing

ByungRae Cha, Sun Park, Byeong-Chun Shin, JongWon Kim, "Design and Verification of Connected Data Architecture Concept employing DataLake Framework over Abyss Storage Cluster,"Smart Media Journal, Vol. 7, no.3, pp.57-63, Sept 2018