Volume 6, Issue 7, July 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Study on Big Data Security and Data Storage Infrastructure

Supriya Haribhau Pawar

Department of Computer Engineering, Bharati Vidyapeeth Deemed University, Pune,
Maharashtra, India

Abstract— Big data technology is changing the traditional technology domain and provides the security model as well as security design to address emerging security challenges. This paper intends to provide the security to infrastructure in big data. The paper starts with the definition of big data and discusses why security is used in big data infrastructure, and how the security is improved. In this paper some of the data security and data protection techniques have been proposed for cloud infrastructure and Hadoop infrastructure.

Keywords— Cloud computing, Hadoop, Infrastructure, Data security, firewall security

I. INTRODUCTION

Big data is a word which describes a huge amount of data for both structured and unstructured data. Big data is huge data set with volume, velocity, and variety. Big data improve the operation and make applications to faster. With the increased access to the web-based, mobile and cloud-based application, sensitive data is accessed from different platforms by different users. These platforms are vulnerable to hacking, mainly if they free or low cost. Nowadays, companies are collecting and processing a huge amount of data or information. Data which are stored by the user ensure that this data is secure. The loss in data security leads to company's financial loss and decreases company's reputation. Therefore security is most important in big data. Big data security is improved by providing application security or device security. Challenges in big data security emerged when the system receives a large amount of data from authorized users and data received is accurate.

II. LITERATURE REVIEW

A. Infrastructure

Infrastructure is a framework which supports a superstructure by substructure. Infrastructure combines the hardware, software, network resources and services for performing operations. It provides the solution to its employees, users. Big data provides different types of infrastructures: Cloud infrastructure and Hadoop Infrastructure

III. CLOUD INFRASTRUCTURE

In Cloud infrastructure, software components and hardware components are used such as storage, servers, virtualization and networking. These components are used to support the computing requirements of a cloud computing model. Cloud computing environment include the front end and back end components, cloud infrastructure consists of the back end components. Cloud infrastructure is present in cloud computing modes, i.e., Infrastructure as Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). IaaS is a foundation; PaaS is the middle layer and SaaS is the top layer of the cloud computing stack.

1) IaaS: IaaS provides the cloud computing infrastructures like servers, storage, network and operating systems as on demand service. Rather than purchasing software's, network equipment, data center space services as on demand of users. Characteristics of IaaS: Allows for dynamic scaling, Resources are distributed as service, IaaS has a variable cost and utility model, and it includes multiple users on a single piece of hardware.

Security requirements in IaaS: Cloud protection, communication security, images security.

2) SaaS: SaaS delivered applications to end user over the web. SaaS provides the software licenses to customers and software delivery model on a subscription basis and it's hosted centrally. It is specified to as on demand software . Characteristics of SaaS: SaaS provide the Central location for software delivery, One too many models are used for delivered the software, For doing the communication between different software's Application Programming Interfaces (APIs) is used, Software upgrades and patches are not handling by users.

Security requirements in SaaS: Data protection, Access control, service availability, software security.

3) PaaS: PaaS is the set of Services and tools which are used to make coding and to deploy those applications efficient and quick. Services in PaaS are constantly updated with additional features added and with existing features are upgraded. Characteristics of PaaS: PaaS provide the services to test, host, develop, deploy and maintain applications in the same development environment, PaaS provides web-based user interface creation tools, which is used for creating, modify, test different user interface scenarios, PaaS provides the features to software like load balancing, scalability, and failover, PaaS provides communication tools and project planning for software development team.

Security requirements in PaaS: Application security, access control.

IV. SECURITY IN CLOUD INFRASTRUCTURE

Cloud computing produces services on demand basis. Characteristics of cloud computing are network access, self-service location independent resource pooling, transference of risk. Cloud computing is important for both the academic research word and the industrial world. Cloud computing is important for IT applications. Cloud computing provides different services and resources to different sites and organizations. Cloud computing share the services and resources in distributed environment via network thus it makes security problem.

V. DATA SECURITY

In cloud computing environment data is distributed in different machines and storage devices like mobile devices and PCs, because of this reason data security is a major issue in cloud computing environment. Two important functions in cloud computing environment: data storage and computing of data. Customers of cloud services can get access to their data and finish their computing task just through the Internet connectivity. Clients have no idea where the data are stored and which machine execute the task of computing. In the research of the cloud computing, the some of the data security and data protection techniques have been proposed.

Data confidentiality: Data confidentiality technique is designed for protecting sensitive information from unauthorized users. Outsourced data is stored in a cloud computing environment, from these environment owners directly control data. Only authorized users can access the data services for, e.g., data computing, data sharing, data search. Data encryption method is used for ensuring confidentiality.

Data Access Controllability: In cloud computing environment all access of the data is controlled by the data owner. When the user accesses the data, owner checks authority of that user. Without authorization of the user, the owner cannot get the permission to access the data. Owner provides different access privileges to different users for controlling the access data.

Data Integrity: Data integrity maintains the accuracy and completeness of data. An owner always regards that data which comes from different sources in a cloud can be stored correctly and reliable. This shows that data should not be illegally modified, improperly tempered, maliciously fabricated or deliberately detected. If unwanted operation deletes or corrupt the data, the data owner should be able to detect the loss or corruption data.

A. Firewall Security

To protect the private network from unauthorized access firewall system is used. A Firewall can be implemented in software and hardware form or a combination of both. Firewall work like a filter between computer and internet. Firewall management program can be configured in two ways:

A default-deny policy: This firewall administrator allowed network services, and everything else is denied.

A default-allow policy: This firewall administrator provides the network services which are not allowed, and everything else is accepted.

A default-deny approach to the firewall is more secure than default-allow approach, but the difficulty occurs in configuring and managing the network.

The Cloud-based firewall provides three basic things: availability, scalability, and extensibility. Availability in cloud-based firewall provides high availability (>99.99%) through the infrastructure with the power redundant, network services as well as backup strategies when site failure. High availability is depending on the manufacturer. Scalability in cloud-based firewall delivered services to multiple users. Scalability comes in an enterprise when bandwidth increases. For increasing the bandwidth in accessing data firewalls are designed in cloud environments. Extensibility in cloud-based firewalls, network manager provides the secure communication path. Cloud-based firewalls extend the new functionality in network security.

VI. HADOOP INFRASTRUCTURE

Hadoop is an open source framework. In Hadoop, using simple programming model big data is stored and process from distributed environment. Hadoop framework defines some modules: Hadoop Common-It contains libraries and utilities needed by other Hadoop modules; Hadoop Distributed File System- it stores data and provides stored data across the clusters, Hadoop Map Reduce- an implementation of this model for large-scale data processing.

VII. DATA SECURITY

Data Masking: Data masking is technique for secure sensitive data; it is used for developing data and analytic data. But there are some limitations:

1) Mapping tables: When data transformation is happen in Hadoop masking process generate mapping tables, this process in not expandable and fast. Mapping table generate another space to protect sensitive data and this process is expensive.

2) Joining operation: Data masking eliminate the integrity and generate duplicate data, which does not provide correct result of joining operation.

3) Security: There is no assurance of data masking take correct sensitive data for provide the security.

Because of these limitations, Hadoop environment proposed different security requirements.

Data Encryption: It translates data into secret code. Encryption is most useful technology for data security. For read the encrypted file need secret key or password that enables to decrypt it. Encrypted data is called ciphertext, and unencrypted data is called plain text. There are two main types of encryption called asymmetric encryption and symmetric encryption. In symmetric algorithm use the same key for encryption and decryption. The same key has also called "secret-key". In asymmetric algorithms use the different key for encryption and decryption. Different key has also called "public key".

Data-centric security: Data-centric security provides the sensitive data elements which are replaced with de-identified, usable and in equivalents format. This means only sensitive data is modified. This approach is used with both structured and semi-structured data. This is also called an end to end data protection and provides an enterprise-wide solution for data protection. This protected data is used in applications, data transfers, data storage and analytic engines.

Authorization: It provides access privileges to system and user. Hadoop provides authorization via file permission in HDFS and resources for Map Reduce and gained access control at a service level. Authorization provides security mechanism on client or user over accessing files, system resources, computer program, services and application features. Authorization verifies user identity when they access the data. System administrator gives the permission to all users and system for accessing data. During authorization, system administrator verifies authenticated user's access and grants the permission for accessing data.

B. Firewall security

Firewall is added in Hadoop environment for secure the network traffic. Network traffic happens when data is transfer to one network to another network. Firewall filters the data between computer and internet. This process defines several methods to filter the incoming and outgoing network traffic. The main purpose of a firewall is eliminating the duplicate data which is generating when communication has happened between different networks. There are two types of firewall:

1) Software firewall: This firewall is a network protection firewall for home users. It provides antivirus software for protecting the data. Still providing inbound and outbound security, software protects against worm and Trojan application.

2) Hardware firewall: This firewall construct into network devices such as routers and protect single machine on the network. This firewall uses packet filtering technique to provide the security in data.

VIII. CONCLUSION

Cloud and Hadoop environments are widely used in industry and research aspects; therefore, security is an important aspect for organizations running in a cloud environment. Using proposed approaches, Cloud and Hadoop environment can be secured for business operation and big data applications.

REFERENCES

[1] Raval, K.S., Suryawanshi, R.S., Naveenkumar, J. and Thakore, D.M., 2011. The Anatomy of a Small-Scale Document Search Engine Tool: Incorporating a new Ranking Algorithm. International Journal of Engineering Science and Technology, 1(3), pp.5802-5808.

[2] Archana, R.C., Naveenkumar, J. and Patil, S.H., 2011. Iris Image Pre-Processing And Minutiae Points Extraction. International Journal of Computer Science and Information Security, 9(6), p.171.

[3] Jayakumar, M.N., Zaeimfar, M.F., Joshi, M.M. and Joshi, S.D., 2014. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor, 5(1), pp.46-51.

[4] Naveenkumar, J. and Joshi, S.D., 2015. Evaluation of Active Storage System Realized through MobilityRPC.

[5] Jayakumar, D.T. and Naveenkumar, R., 2012. SDjoshi,“. International Journal of Advanced Research in Computer Science and Software Engineering,” Int. J, 2(9), pp.62-70.

[6] Jayakumar, N., Singh, S., Patil, S.H. and Joshi, S.D., Evaluation Parameters of Infrastructure Resources Required for Integrating Parallel Computing Algorithm and Distributed File System.

[7] Jayakumar, N., Bhardwaj, T., Pant, K., Joshi, S.D. and Patil, S.H., A Holistic Approach for Performance Analysis of Embedded Storage Array.

[8] Naveenkumar, J., Makwana, R., Joshi, S.D. and Thakore, D.M., 2015. OFFLOADING COMPRESSION AND DECOMPRESSION LOGIC CLOSER TO VIDEO FILES USING REMOTE PROCEDURE CALL. Journal Impact Factor, 6(3), pp.37-45.

[9] Naveenkumar, J., Makwana, R., Joshi, S.D. and Thakore, D.M., 2015. Performance Impact Analysis of Application Implemented on Active Storage Framework. International Journal, 5(2).

[10] Salunkhe, R., Kadam, A.D., Jayakumar, N. and Thakore, D., In Search of a Scalable File System State-of-the-art File Systems Review and Map view of new Scalable File system.

[11] Salunkhe, R., Kadam, A.D., Jayakumar, N. and Joshi, S., Luster A Scalable Architecture File System: A Research Implementation on Active Storage Array Framework with Luster file System.

[12] Jayakumar, N., Reducts and Discretization Concepts, tools for Predicting Student’s Performance.

[13] Jayakumar, M.N., Zaeimfar, M.F., Joshi, M.M. and Joshi, S.D., 2014. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor, 5(1), pp.46-51.

[14] Kumar, N., Angral, S. and Sharma, R., 2014. Integrating Intrusion Detection System with Network Monitoring. International Journal of Scientific and Research Publications, 4, pp.1-4.

[15] Namdeo, J. and Jayakumar, N., 2014. Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts. International Journal, 2(2).

[16] Naveenkumar, J., Keyword Extraction through Applying Rules of Association and Threshold Values. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), ISSN, pp.2278-1021.

[17] Kakamanshadi, G., Naveenkumar, J. and Patil, S.H., 2011. A Method to Find Shortest Reliable Path by Hardware Testing and Software Implementation. International Journal of Engineering Science and Technology (IJEST), ISSN, pp.0975-5462.

[18] Naveenkumar, J. and Raval, K.S., Clouds Explained Using Use-Case Scenarios.

[19] Naveenkumar J, S.D.J., 2015. Evaluation of Active Storage System Realized Through Hadoop. International Journal of Computer Science and Mobile Computing, 4(12), pp.67–73.

[20] RishikeshSalunkhe, N.J., 2016. Query Bound Application Offloading: Approach Towards Increase Performance of Big Data Computing. Journal of Emerging Technologies and Innovative Research, 3(6), pp.188–191.

[21] Sagar S lad s d joshi, N.J., 2015. Comparison study on Hadoop’s HDFS with Lustre File System. International Journal of Scientific Engineering and Applied Science, 1(8), pp.491–494.

[22] Salunkhe, R. et al., 2015. In Search of a Scalable File System State-of-the-art File Systems Review and Map view of new Scalable File system. In nternational Conference on electrical, Electronics, and Optimization Techni ques (ICEEOT) - 2016. pp. 1–8.

[23] BVDUCOE, B.B., 2011. Iris Image Pre-Processing and Minutiae Points Extraction. International Journal of Computer Science & Information Security.

[24] P. D. S. D. J. Naveenkumar J, “Evaluation of Active Storage System Realized through MobilityRPC,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 3, no. 11, pp. 11329–11335, 2015

[25] N. Jayakumar, S. Singh, S. H. Patil, and S. D. Joshi, “Evaluation Parameters of Infrastructure Resources Required for Integrating Parallel Computing Algorithm and Distributed File System,” IJSTE, vol. 1, no. 12, pp. 251–254, 2015.

[26] N. Jayakumar, T. Bhardwaj, K. Pant, S. D. Joshi, and S. H. Patil, “A Holistic Approach for Performance Analysis of Embedded Storage Array,” Int. J. Sci. Technol. Eng., vol. 1, no. 12, pp. 247–250, 2015.

[27] J. Naveenkumar, R. Makwana, S. D. Joshi, and D. M. Thakore, “Performance Impact Analysis of Application Implemented on Active Storage Framework,” Int. J., vol. 5, no. 2, 2015.

[28] N. Jayakumar, “Reducts and Discretization Concepts, tools for Predicting Student’s Performance,” Int. J. Eng. Sci. Innov. Technol., vol. 3, no. 2, pp. 7–15, 2014.

[29] J. Namdeo and N. Jayakumar, “Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts,” Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 2, no. 2, 2014.

[30] R. Salunkhe, A. D. Kadam, N. Jayakumar, and S. Joshi, “Luster A Scalable Architecture File System: A Research Implementation on Active Storage Array Framework with Luster file System.,” in ICEEOT, 2015.