RESEARCH ARTICLE OPEN ACCESS

Authentication System for Hadoop in Cloud Environment- A Collaborative Approach

Shashank Raghunath, Mohanish Satam, Kiran Pugaonkar, Nikhil Joshi, Gayatri Naik
Department of Computer Science and Engineering
Yadavrao Tasgaonkar Institute of Engineering and Technology
Chandai, Karjat, Mumbai
Maharashtra - India

ABSTRACT
This article focusses on the necessity of an independent authentication system for a Hadoop framework in a Cloud environment and concludes with a brief description of its implementation. Hadoop does not authenticate the client and therefore, data nodes can be accessed using block location. A classic PHP based authentication mechanism is used to grant access to the user. PHP being the most popular web development script, serves well as an authenticator using hashing techniques like SHA-256 for storing password. Hadoop being deployed in a cloud environment, acts as web service and hence PHP authentication becomes much simpler.
Keywords:- Authentication, Hadoop, Cloud.

I. INTRODUCTION

The Apache Hadoop projects provide a series of tools designed to solve big data problems. The Hadoop cluster implements a parallel computing cluster using inexpensive commodity hardware. The cluster is partitioned across many servers to provide a near linear scalability. The philosophy of the cluster design is to bring the computing to the data. So each data node will hold part of the overall data and be able to process the data that it holds. The overall framework for the processing software is called Map Reduce.

Apache Hadoop can be useful across a range of use cases spanning virtually every vertical industry. It is becoming popular anywhere that you need to store, process, and analyze large volumes of data. Examples include digital marketing automation, fraud detection and prevention, social network and relationship analysis, predictive modeling for new drugs, retail in-store behavior analysis, and mobile device location-based marketing [1].

Authorization, the process of granting access to requested resources, is pointless without suitable authentication. Both the cloud provider and the enterprises must consider the challenges associated with credential management and implement cost effective solution that reduce the risk appropriately. Password authentication is considered as one of the simplest and most convenient authentication mechanisms [2].

II. AUTHENTICATION ARCHITECTURES

The architecture for authentication systems fall under four categories as described below:

A. Single Server model

A single server (Fig. 1) maintains a database of user passwords. Most of the existing authentication systems follow this single-server model. The main drawback of single server is the single point of vulnerability. It leads to offline dictionary attacks against the user password database.

B. Simple Multi-Server model

In the simple multi-server model depicted (Fig.2), the server side comprises of multiple servers to overcome the single point of vulnerability; the servers are equally exposed to users and a user has to communicate in parallel with several or all servers for authentication. The main problem with the simple multi-server model is the demand on communication bandwidth and the need for synchronization at the user side since a user has to engage in simultaneous communications with multiple servers.

C. Gateway Augmented Multi-Server model

In the gateway augmented multi-server model (Fig.3). A gateway is positioned as a relaying point between users and servers and a user only needs to contact the gateway. Apparently, the introduction of the gateway removes the demand of simultaneous communications by a user with multiple servers as in the plain multi-server model. However, the gateway introduces a redundant layer in the architecture, to relay messages between users and servers. Gateways also reduce system reliability.

.

D. Two Server model

The two-server model (Fig. 4) comprises of two servers at the server side, one of which is a public server exposing itself to users and the other is a back-end server staying behind the scene. Users contact only the public server, but the two servers work together to authenticate users.

III. RELATED WORK

Lishan Kang [13] has proposed an Identity-Based Authentication (IBA) scheme over traditional mutual authentication. In cloud storage sharing, mutual authentication between users and between user and Cloud environment is critical in ensuring data security. However, traditional mutual authentication using public-key operation unleashes cloud storage system load, computation and communication overhead and reduces scalability. An IBA scheme has short key size, is identity-based and no interactive. This scheme divides the sharing users between domains. In the domain global master key is shared to exercise mutual authentication. By the analysis of performance, this scheme improves the computational and communicational efficiency over two times. This scheme is enabled by an emerging cryptographic technique from the bilinear pairing and its security can be assured by the Bilinear Diffie-Hellman Problem (BDHP).

In IBA scheme, the master key of some do-main becomes the bottleneck of Cloud Storage System's security. Once the master key of some domain is leaked, the domain's security will be wrecked. In addition, if a user wants to share another user's data, they must be in the same domain Zhidong Shen [14] have proposed Trusted Computing Platform (TCP) to aid the process of authentication in cloud computing. The TCP is based on the Trusted Platform Module (TPM). The TPM is a logic independent hardware. It can resist the attacks from both software and the hardware. The TPM contains a private master key to protect for other information stored in cloud computing system. Because the hardware certificate is stored in TPM it is hard to attack it. So TPM provides the trust root for users. Since the users have full information about their identity, the cloud computing system can use some mechanism to trace the users and get their origin.

In TCP the user's identity is proved by user's personal key and this mechanism is integrated in the hardware, such as the BIOS and TPM. So it is very hard to deceive a user-id. Each site in the cloud computing system will record the visitor's information. By using the TCP mechanism in cloud computing, the trace of participants can be known by the cloud computing trace mechanism. The TCP provides cloud computing a secure base for achieving trusted computing. Integration of hardware modules with cloud computing system is a challenging research issue. Hadoop currently uses Kerberos protocol [3] for its authentication.

When organizations begin to utilize applications in the cloud, authenticating users in a trustworthy and manageable manner becomes an additional challenge. Our proposed work on a dual server authentication protocol utilizes dual servers for authentication to enhance the cloud security. The significance of this protocol is the usage of the fundamental concepts and basic elements of the triangle for authentication.

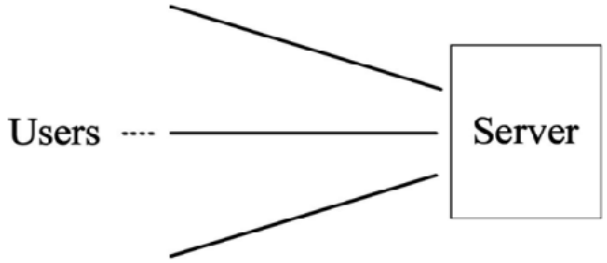


Fig (1). Single Server Model

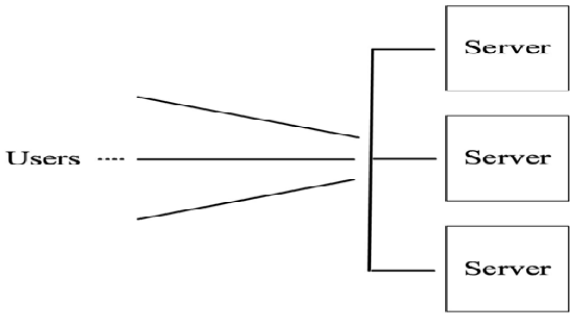
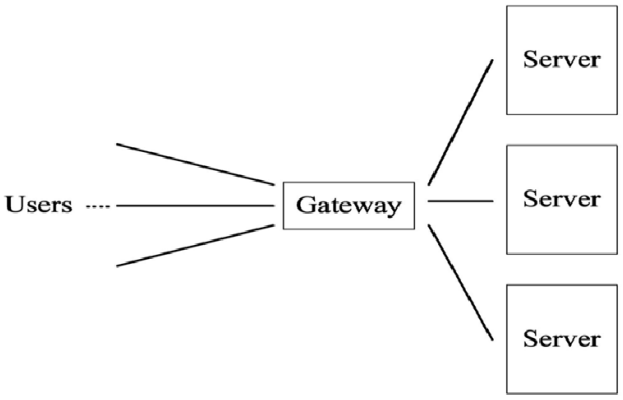
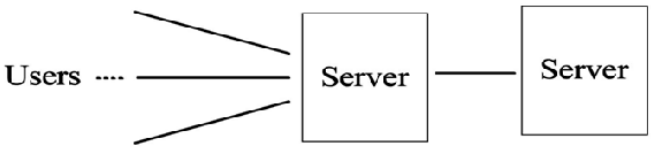


Fig (2). Simple Multi-Server Model



Fig(3). Gateway Augmented Multi Server Model



Fig(4). Two Server Model

IV. PROPOSED ARCHITECTURE

The proposed architecture is an implementation of the two server model. It includes an independent authentication server which authenticates the user into the cloud service which contains Apache Hadoop. The following diagram illustrates the basic architecture:

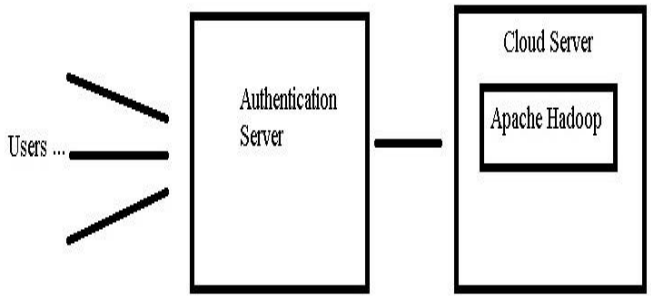


Fig (5). Proposed Architecture

As shown in Fig (5), the architecture is comprised of two servers, an authentication server and a cloud server.

*1) **Authentication Server:*** The authentication server is an Apache-PHP-MySQL server. The users log into the server using their username and password given during registration process.

2) ***Cloud Server:*** Cloud server particularly in this case is one which provides Hadoop as service. Typical choices for cloud server containing Hadoop would be Openstack 6.0 with Sahara project, or Amazon EC2 server.

V. IMPLEMENTATION

For implementation in a local machine with limited resources, WAMP server is used along with Hortonworks Sandbox.

A typical use case scenario is as follows:

1) **Registration:** The user registers into the server by providing details such as username, password, email id and by verifying him/herself by entering the correct CAPTCHA code.

2) **Login:** The login is done by entering the correct username and password.

3) **Hadoop:** After a successful login, the user is provided with an encrypted link that redirects the user into the cloud where Hadoop services are active.

To increase the efficiency of this simple username password authentication mechanism, the following measures have been undertaken:

1) **CAPTCHA:** CAPTCHA stands for Completely Automated Public Turing Test to Tell Computers and Humans Apart. In simple terms it is a mechanism which can differentiate between a human and a machine. This is done by asking simple questions or by showing images that a human can easily identify but a machine cannot. Spams and BOT registration is brought under control by using CAPTCHA.

2) **SHA256 and SALT:** SHA256 is a password hashing function which takes random text as input and generates a 64 bit hexadecimal text. This output is then stored in the database. To further increase security, a random 32 bit SALT is added along with the password as input to the hash function. This ensures safety against Brute force attacks, dictionary attacks and nullifies the use of rainbow tables.

3) **Other measures:** Login attempts have been limited to 3, failing which the user's IP address is blocked for 24 hours which further increases security. Passwords entered must have at least 1 special character, 1 number and 1 capital alphabet. The minimum character allowed in a password are 8 which is the global standard. A JavaScript password strength meter suggests the strength of the password to the user while registering.

VI. HORTONWORKS HDP

The Hortonworks Sandbox is a single node implementation of the Hortonworks Data Platform(HDP). It is packaged as a virtual machine to make evaluation and experimentation with HDP fast and easy.

For experimentation in a local environment, we used the authentication system along with Hortonworks HDP sandbox.

HDP sandbox provides all the Hadoop tools such as Pig, Hive and HCatalog to name a few.

VII. CONCLUSIONS

This paper proposed a classic approach using dual servers. The proposed authentication protocols enhance Hadoop security as the authentication mechanism is a global standard. As the servers keep the interpreted and distinct form of user credentials, there is very less chance to reveal the user

credentials to the adversary. The generation of 64 bit hexadecimal password improves the security level as they cannot be easily hacked. So the utilization of this protocol will make the Hadoop environment more secure.

ACKNOWLEDGMENT

It is matter of great satisfaction and pleasure to present project report on **"Authentication System for Hadoop in Cloud Environment- a Collaborative Approach "** we wish to express our sincere thanks to **H.O.D. Ms. Prof. Vaishali Londhe** who extended their valuable support during the course of project. We again wish to express our sincere thanks and gratitude to **Ms. Prof. Gayatri Naik** for becoming a guide in this project and for her constant guidance and motivation. We also thank her for her valuable support and encouragement throughout the preparation of project report without which the project report would have not been completed. We also thank our colleagues who have helped in successful completion of the project report. Last but not the least we would like to thank all our friends, who helped us directly and indirectly. Helpful hand rendered by all of them will remain for long time in our memory. Finally we admit the cooperation and hard work are our keywords for success

REFERENCES

[1] Tom White, "Hadoop- Definitive guide", O'Reilly,2009

[2] Owen O'Malley, "Integrating Kerberos into Apache Hadoop", Kerberos Conference 2010, 26-27 October 2010, MIT, USA.

[3] Mark Brunet, "Perfect Password: Selection, Protection, Authentication", Syngress, 2005.

[4] Cloud Security Alliance "Domain 12: Guidance for Identity & Access Management V2.1",April 2010

[5] Her-Tyan Yeh, Hung-Min Sun and Tzonelih Hwang, "Efficient Three- Party Authentication and Key Agreement Protocols Resistant to Password Guessing Attacks", Journal of Information Science and Engineering, vol.19, no.6, pp. 1059-1070, 2003.

[6] L in, C.L., and T. Hwang, " A password authentication scheme with secure password updating", Computer & Security, vol.22, no.1, pp.68–72, 2003.

[7] Eun-Jun Yoon, Eun-Kyung Ryu and Kee-Young Yoo, " Attacks and Solutions of Yang et al.'s Protected Password Changing Scheme", Informatica, vol.16 , no. 2, pp. 285-294, April 2005.

[8] Yanjiang Yang, Feng Bao, "Enabling Use of Single Password Over Multiple Servers in Two-Server Model",2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

[9] Dexin Yang, Bo Yang, "A Novel Two-Server Password Authentication Scheme with Provable Security",2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

[10] Jun Ho Lee and Dong Hoon Lee, "Secure and Efficient Password-based Authenticated Key Exchange Protocol for Two- Server Architecture",2007 International Conference on Convergence Information Technology

[11] Yanjiang Yang, Robert H. Deng, Senior Member, IEEE, and FengBao, "A Practical Password-Based Two-Server Authentication and Key Exchange System",, " international journal of dependable and secure computing, vol. 3, no. 2, April-June 2006

[12] Yanjiang Yang, Feng Bao, "Enabling Use of Single Password Over Multiple Servers in Two-Server Model",2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

[13] LishanKang, Xuejie Zhang(2010), "Identity – Based Authentication in Cloud Storage Sharing",2010 International Conference on Multimedia Information Networking and Security

[14] Zhidong Shen, Qiang Tong(2010), "The Security