Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 2s (2024)

Enhancing Data Security in SPARK Cluster: A Novel Symbol-based Authentication Approach

[1]J.Balaraju, [2]C.Dastagiraiah, [3]P.Ravinder Rao, [4]T.Srikanth, [5]K.Jyothi Goud, [6]V.Subramanyam

[1*]Corresponding author: Associate Professor,Department of CSE,Anurag University,Hyderabad,India.,balarajucse@anurag.edu.in

[2] Assistant Professor, Department of CSE,Anurag University,Hyderabad,India., Dastagiraiah.cse@anurag.edu.in.

[3]Assistant Professor,Department of CSE,Anurag University,Hyderabad,India., ravinderraocse@anurag.edu.in.

[4] Assistant Professor,Department of CSE,Anurag University,Hyderabad,India., jyothi.kurremula@gmail.com.

[5]Assistant Professor,Department of CSE,Anurag University,Hyderabad,India.,srikanthcse@anurag.edu.in.

[6]Associate Professor,Department of CSE,Anurag University,Hyderabad,India., voore.subramanyam206@gmail.com.

Article History:

Received: *01-04-2024*

Revised: *19-05-2024*

Accepted: *02-06-2024*

Abstract:

User authentication is the process of confirming an individual's identity prior to granting them access to a connected device, an online service, or any other valuable resource. Its importance lies in its capability to protect data, applications, and networks for organizations by restricting access to authorized individuals or approved processes. In this study, the widely used Apache Spark technology was employed for storing and analyzing vast amounts of data, and a unique authentication framework was introduced. A dynamic symbol selection authentication offers a promising alternative to traditional alphanumeric passwords, as well as biometric and facial authentications. This authentication method has been thoroughly tested in the highly distributed Apache Spark cluster. The implementation utilizes SHA512 cryptography in various ways and compares the results with existing authentication and machine learning algorithms. The authentication scheme, combined with the powerful Apache Spark distributed system consisting of 10 nodes, yielded exceptional outcomes.

Keywords: Distributed Computing, Authentication SHA512, Spark, Cryptography.

1. INTRODUCTION

Big data security is a critical aspect of Apache Spark cluster security, and it is frequently linked to the human factor. The significance of these factors lies in the development of secure systems, security features, and human-computer interaction. The primary concern in this context is the authentication issue. User authentication plays a pivotal role in computer and big data security scenarios, ensuring that the individual requesting a resource is indeed the person they claim to be. Presently, most authentication solutions incorporate a combination of a username and password. A password serves as a form of confidential authentication information utilized to regulate access to data and resources. It remains undisclosed to unauthorized individuals, and those seeking access are required to provide the correct password in order to be granted or denied entry. However, the drawback of passwords lies in the necessity to keep them secret and for users to remember them. Each authentication method comes with its own set of guidelines and restrictions, such as the inclusion of alphanumeric and special characters, as well as a minimum length for the password. These passwords primarily consist of text-based combinations. Passwords have been employed since ancient times, when soldiers guarding a location exchanged secret codes and only permitted entry to those who were aware of them[1][2][3].

A. Apache Spark

Apache Spark is a framework that is open source and is used for processing large amounts of data by utilizing key Hadoop components. It employs a multi-stage processing method that operates in-memory and is 100 times faster than map-reduce when compared to other big data processing tools like Hadoop and Storm. Spark has the capability to access and process any Hadoop data source that can be executed on Hadoop clusters. At its foundation, Apache Spark encompasses fundamental features pertaining to task scheduling, memory management, error recovery, and communication with storage systems [4].

B. Cryptography

Encryption is a crucial process that utilizes mathematical algorithms and protocols to secure data by encrypting and decrypting it. Its primary purpose is to prevent unauthorized access and maintain the integrity of the data. Encryption plays a vital role in key-based authentication (SBA) as it ensures the confidentiality and integrity of the authentication process. In token-based authentication, cryptography is employed in various ways. Typically, cryptographic hashes are utilized to convert passwords or image patterns into fixed-length hashes, which can be securely stored or transmitted without revealing the original data. Cryptography is also employed to safeguard the communication between the client and the server during the authentication process[5][6].

SHA512 Algorithm.

SHA-512 is a hashing algorithm that operates on input data. It is utilized in various fields such as blockchains, digital certificates, and internet security. Hashing algorithms play a crucial role in digital security and encryption. SHA-512 is part of the SHA-2 family, alongside SHA-256, and is specifically employed in hashing the Bitcoin blockchain. Consequently, SHA-512 executes its functions through a series of steps[7].

2. PROBLEM DEFINITION.

Most data today is stored in widely-used open-source distributed systems like Hadoop and Apache Spark, which do not have built-in security features and instead rely on third-party security protocols. Apache Spark utilizes Hadoop cores for big data storage and processing. Hadoop itself lacks its own security mechanisms and heavily relies on external security protocols like Kerberos and Apache Knox. The current authentication processes are inefficient, leading to risks of unauthorized access not only to personal accounts but also to corporate accounts and sensitive corporate data.

3. LITERATURE REVIEW.

Srinivas et al[8] are devised a symbol-oriented input method that safeguards against shoulder attacks by employing a convex hull approach. To successfully click within the convex hull created by the moving objects, the user needs to identify these objects. If the user desires a more challenging password to guess, a vast array of objects can be utilized. However, this leads to overcrowded images and objects that are nearly impossible to distinguish. On the other hand, using fewer elements reduces the passage space, resulting in a larger convex hull.

Balaraju et al [9] in their study, highlighted the challenging nature of big data storage, safety, and security. Rather than opting for alternative big data technologies, the Hadoop framework emerges as a viable solution to address these concerns. Notably, Hadoop has significantly enhanced its security features over the last decade through continuous improvements. The latter part of the paper delves into the persistent authentication challenges, metadata security issues, and data streaming concerns faced by Hadoop and Spark..

Licheng Ma. et al [10] developed a novel approach for graphical password authentication. In their method, users are required to click on approximate regions of a specific image to gain access to predetermined destinations. Passlogix further enhanced this concept by introducing the verification of legitimacy through the correct sequence of clicking on different items..

Prasad Rao et al [11] in their study, delve into the topic of authentication security measures in Hadoop and Spark. They explore the utilization of third-party security providers like Kerberos and Apache Knox, which come with a notable computational burden in safeguarding data. However, despite the cost, this security approach falls short in establishing a truly trusted environment. Furthermore, the authors highlight the limitation of solely authenticating users without considering their processes. To address these concerns, they propose a novel security method that employs a secure authentication interface to ensure the protection of big data within a Hadoop/Spark cluster.

Yuxin. et al [12] developed a technique that involves utilizing a mouse to sketch a digital signature for authentication purposes. This technique is divided into two main stages: registration and verification. During the registration phase, the user signs their name using the mouse, following which the system erases the signature area. The system takes the user's signature as an input, undergoes a normalization process, and subsequently extracts the signature parameters during the validation phase.

Balaraju and Prasada Rao et al [13] Developed a technique that involves utilizing a mouse to create a digital signature for authentication purposes. This technique is divided into two main stages: registration and verification. During the registration phase, the user signs their name using the mouse, following which the system clears the signature area. The user's signature is taken as input, undergoes a normalization process, and then the signature parameters are extracted during the validation phase.

4. MATERIALS AND METHODS

Spark Multi-node Cluster can support up to 16,000 nodes by segregating master and child services. A significant number of users can operate within the cluster. The basic services enable any user to join the cluster with equal privileges, without the need for user authentication..

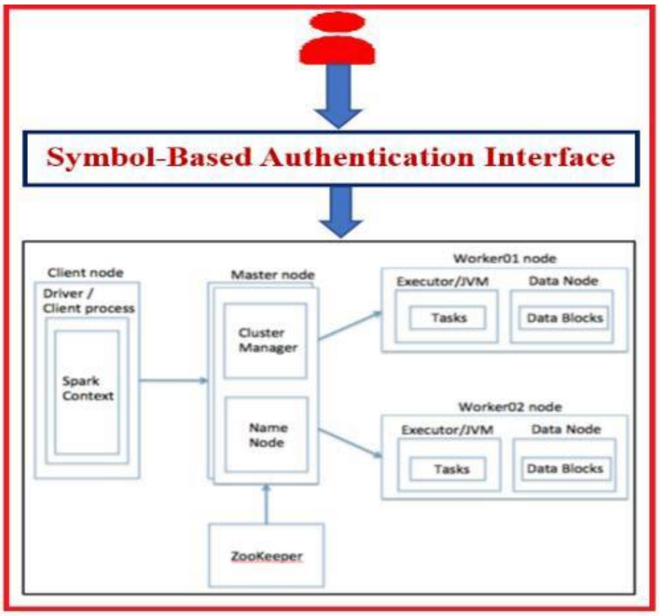
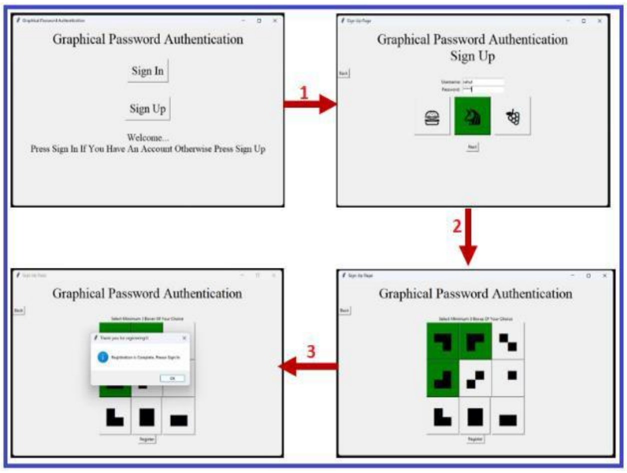


Fig 1: Proposed Symbol-based Authentication for SPARK Cluster.

A Hadoop /Spark cluster configuration consisting of multiple nodes was established with the following specifications: two servers acting as hosts equipped with Xeon processors, 32 GB RAM, and 256 × 4 GB SAS controller hard disk capacity; a Core i3 processor, 4 GB of RAM, and 512 MB storage with 117 numbers; and a Core i5 processor, 8GB RAM, and 1GB hard drive with 126 digits, serving as a slave. The entire cluster, comprising 245 nodes, operates on the open source CentOS Linux operating system and is interconnected through a high-speed gigabit switch. This setup lays a solid foundation for future performance analysis and related endeavors. To enhance security, a graphical user interface has been designed to employ symbol-based authentication. Instead of relying on traditional alphanumeric passwords, users authenticate their identity by selecting a series of images, clicking on specific points, or following a pattern using a mouse or touch screen. Symbol-based authentication offers greater security compared to traditional passwords, as images and patterns are more difficult to guess or brute force. Moreover, it tends to be more memorable for users, facilitating the recollection of their credentials[14][15].


Graphical Password Authentication

Welcome...
Press Sign In If You Have An Account Otherwise Press Sign Up

Graphical Password Authentication
Sign Up

Graphical Password Authentication

Graphical Password Authentication

Fig 2: Process of User registration using Symbols.

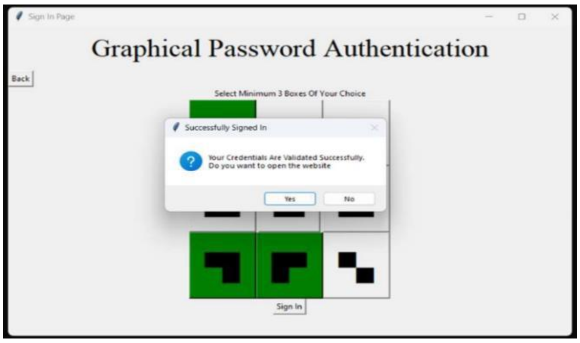
There are various authentication methods based on tokens, such as image-based, gesture-based, and 3D virtual environment-based authentication. Examples of image-based authentication include Pass faces, where a user selects a face from a group of images, and DAS, where a user selects numbers from a grid. On the other hand, 3D virtual environment-based authentication is a newer and more experimental approach that requires users to interact with objects in a virtual environment to verify their identity. In general, token-based authentication presents a promising alternative to traditional alphanumeric passwords, offering enhanced security and a more user-friendly experience [16].

A. *Recognition- Based Technique*

In verification-based strategies, the user is required to authenticate their identity by identifying specific photos chosen during the registration procedure. For instance, in facial recognition systems, the system needs to retain the user's portfolio photos to exhibit them accurately. This information should be securely stored, allowing the system to utilize it in its original format (possibly safeguarded by reversible encryption). Additionally, it should be accessible to authorized individuals, including those who have access to the stored data, like shoulder surfing or potential phishing attempts[17].

B. *HASHLIB Module*

The primary objective of this module is to employ a hash function to encrypt a string in a manner that renders decryption nearly impossible. The length of the encrypted string greatly hinders any attempts to retrieve the original string. This module offers a unified interface for various secure hashing and message digest algorithms, encompassing FIPS-approved secure hashing algorithms as defined in Internet RFC 1321, as well as RSA MD5 technology. The FIPS-protected hash algorithms encompass SHA1, SHA224, SHA256, SHA384, and SHA512. The terms "message digest" and "secure hash connection" can be used interchangeably. In the past, earlier algorithms referred to message digests, but nowadays, Secure Hash has become the prevailing term[18][19].


Sign In Page
Graphical Password Authentication
Back
Select Minimum 3 Boxes Of Your Choice
Successfully Signed In
Yes
No
Sign In

Your Credentials Are Validated Successfully.
Do you want to open the website

Fig 3: Authentication process based on Symbol.

5. RESULTS AND DISCUSSION.

The Apache Spark cluster is currently secured using Kerberos and Apache Knox authentication mechanisms. To enhance security, a new verification method has been developed and implemented on the cluster. This method utilizes reduced computations and offers improved security compared to existing authentication methods. A comparative study has been conducted, which includes token-based authentication as well as other authentication methods like Kerberos, Secure Authentication interface, and DNA Authentication[20][21].

Existing Authentication mechanisms	Computations for Authentication
SAI	5
DNA	4
Kerberos	6
SBA (Proposed)	1

Table 1: Computation for One User Authentication.

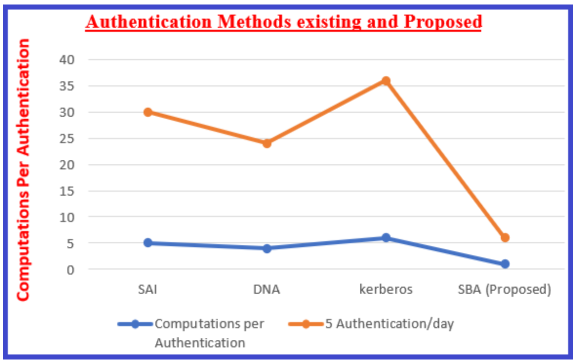


Figure 4: Computation for One User Authentication.

Authentication Methods	Computations / Authentication	5 Authentication / day
SAI	5	25
DNA	4	20
Kerberos	6	30
SBA (Proposed)	1	5

Table 2: Computation for One User Authentication 5 times authentication/day

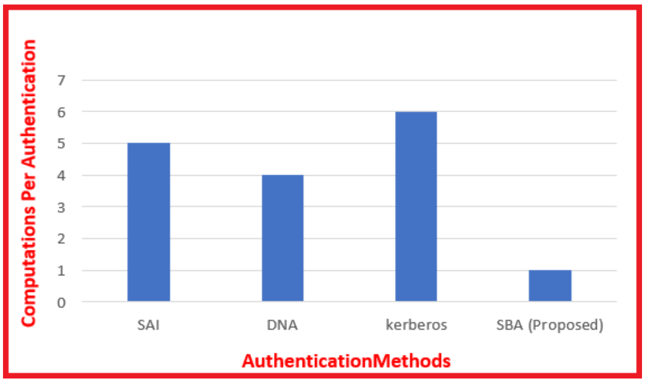


Figure 5.: Computation for One User Authentication 5 times authentication/day.

6. CONCLUSION AND FUTURE SCOPE.

An innovative framework has been proposed for a rapid and effective authentication system designed to safeguard large datasets within an Apache Spark environment. This system presents a robust alternative to current authentication methods. The primary objective behind the successful implementation of this system was to establish a token-based authentication mechanism that could be seamlessly integrated with popular big data technologies like Apache Spark, thereby restricting unauthorized access to data, which can be utilized in various mobile applications or websites. The authentication process incorporates SHA512 encryption in multiple ways. This approach is appealing as words tend to be retained in memory longer than images. By employing a straightforward authentication system alongside a 10-node Apache Spark distributed system, optimal outcomes were achieved. The next phase of this project involves transitioning to a multi-node cluster environment with Hadoop and Fire, expanding the distributed system to encompass over 1000 nodes. This scalability is made possible by the fact that the aforementioned clusters are compatible with commodity hardware and are continuously expanding their node support. Furthermore, this mechanism facilitates the utilization of data for future endeavors by assigning each user a unique identifier and tracking user activities..

REFERENCES

[1]. Balaraju, J., Prasada Rao, PVRD., ―Recent advances in big data storage and security schemas of HDFS: a surve, Special Issue (Emerging Trends in Engineering Technology) ,6,132-138.

[2]. Chellappan S., Ganesan D. (2018) Introduction to Apache Spark and SparkCore. In: Practical Apache Spark.Apress, Berkeley,CA. https://doi.org/10.1007/978-1-4842- 3652-9_3.

[3]. M. M. Shetty and D. H. Manjaiah,(2016), "Data security in Hadoop distributed file system", International Conference on Emerging Technological Trends (ICETT),1-5.

[4]. Salloum, S.; Dautov, R.; Chen, X.; Peng, P.X.; Huang, J.Z. Big data analytics on Apache Spark. Int. J. Data Sci. Anal. **2016**, 1, 145–164.

[5]. Amalraj, A.J. and Jose, J.R., 2016. A survey paper on cryptography techniques. International Journal of Computer Science and mobile computing, 5(8), pp.55-59.

[6]. Revenkar, P. S., Anisa Anjum, and W. Z. Gandhare. "Secure iris authentication using visual cryptography." arXiv preprint arXiv:1004.1748 (2010).

[7]. Gueron, Shay, Simon Johnson, and Jesse Walker. "SHA-512/256." In 2011 Eighth Interna- tional Conference on Information Technology: New Generations, pp. 354-358. IEEE, 2011.

[8]. Kiran, T. Srinivasa Ravi, et al. "A symbol-based graphical schema resistant to peeping at- tack." International Journal of Computer Science Issues (IJCSI) 10.5 (2013): 229.

[9]. Balaraju J., Prasada Rao. PVRD, ―Innovative Secure Authentication Interface for Ha- doop Cluster Using DNA Cryptography: A Practical Study‖. In (eds) Soft Computing and Signal Processing. ICSCSP 2019. Advances in IntelligentSystems and Computing, vol 1118. Springer, Singapore.https://doi.org/10.1007/978-981-15-2475-2_3

[10]. Gao, Haichang, Wei Jia, Fei Ye, and Licheng Ma. "A survey on the use of graphical pass- words in security." J. Softw. 8, no. 7 (2013): 1678-1698.

[11]. Balaraju, J., Prasada Rao. PVRD, "Investigation and Finding A DNA Cryptography Layer for Securing Data in Hadoop Cluster." Int. J. Advance Soft Compu. Appl 12.3 (2020).

[12]. Meng, Yuxin. "Designing click-draw based graphical password scheme for better authenti- cation." In 2012 IEEE Seventh International Conference on Networking, Architecture, and Storage, pp. 39-48. IEEE, 2012.

[13]. Balaraju, J., Prasada Rao. PVRD, ―Designing authentication for Hadoop cluster using DNA algorithm‖. Int. J. Recent. Technol. Eng. (IJRTE) ,8(3), 2019. ISSN: 2277-3878. https://doi.org/ 10.35940/ijrte.C5895.0983.

[14]. Li, Yue. "On Enhancing Security of Password-Based Authentication." (2019).

[15]. Jiya, Gloria Kaka, Ishaq Oyebisi Oyefolahan, and Joseph O. Ojeniyi. "Recognition based graphical password algorithms: A survey." (2021).

[16]. Saranya Ramanan1, Bindhu J S “a survey on different symbol-based authentication schemes” IJIRCEVol. 2, Issue 12, December 2014. [6] Greg E. Blonder (1996). U.S. Patent No.5559961.

[17]. Towseef Akram, Vakeel Ahmad, Israrul Haq, & Monica Nazir. (2017). Symbol-based Authentication.

[18]. Awais, A., Muhammad, A., M., K. H., & Talib, R. (2016). Secure Symbol-based Tech- niques against Shoulder Surfing and Camera-based Attacks.

[19]. Anwar, Muhammad Rehan, Desy Apriani, and Irsa Rizkita Adianita. "Hash Algorithm In Verification Of Certificate Data Integrity And Security." Aptisi Transactions on Techno- preneurship (ATT) 3.2 (2021): 181-188.

[20]. Balaraju, J.,Prasada Rao. PVRD, ―A Novel Node Management in Hadoop Cluster using DNA‖, International Journal of Information Technology ProjectManagement.,12(4), June 2021, ISSN: 1938-0232.

[21]. Balaraju, J., et al. "Dynamic Password to Enforce Secure Authentication Using DNA." *International Journal of Intelligent Systems and Applications in Engineering* 12.1 (2024): 55-61.