Big Data Authentication and Authorization using SRP Protocol

Miti Jhaveri
Student, Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Devang Jhaveri
Student, Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Narendra Shekokar, PhD
HOD, Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

ABSTRACT
The advances in the digital world has rendered the collection of colossal amounts of data called Big Data. This voluminous data if analyzed in an expedient way can help us gain valuable insights. On the other hand, securing Big Data has become an important aspect as it can lead to disastrous results if intercepted by intruders. Through this paper, we describe the management of authentication and access control by providing an overview of the existing protocols. Further, we propose a modified protocol which amalgamates the authentication and authorization processes there by speeding up the entire procedure.

General Terms
Big Data, Security

Keywords
Secure Remote Password Protocol, Access Control, Authentication

1. INTRODUCTION
Google1 defines Big Data as "extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions." In other words, Big Data is a reservoir of voluminous amount of data originating from various sources in several forms such as structured, unstructured and semi-structured data.

The ever increasing usage of smart phones, e-commerce websites, social networking websites, WSNs and GPS generate a large amount of data which if analyzed systematically can give us valuable insights. Big data is indeed Big. Currently, it is measured in zettabyte which is equal to over a trillion gigabytes (1,099,511,627,776 GB, to be exact or 1021 bytes) and this number is expected to increase exponentially in the coming years [2].

The importance of Big data Analytics is far beyond imagination. Big Data when captured, processed and analyzed properly bolsters business companies with precious information which leads to efficiency improvements, increased sales, lower costs, better customer service, and/or improved products and services [3]. For instance, the proliferation of smart phones and other GPS devices offers advertisers an opportunity to target consumers when they are in close proximity of a store, a coffee shop or a restaurant. This opens up new revenue for service providers and offers many businesses a chance to target new customers. On the other hand, television providers have started customizing advertisements based on individual household demographics and viewing patterns.

Big Data revolves around six important characteristic which makes it distinct from the traditional or relational databases [4]. These characteristics are described as:

- Volume: Big data implies enormous amount of data generated by machines, networks and systems like social media etc. Analyzing such large amount of data is not an easy task.
- Variety: The variety characteristic refers to various types of structured and unstructured data originating from miscellanies sources. Traditionally, data was stored structurally in database files or excel sheets, but today data comes in the form of images, audio files, videos, photos etc. This variety of unstructured data causes several problems in storing and analyzing data.
- Velocity: Velocity refers to the enormous rate at which the data is produced. This flow is massive and continuous. analyzing real-time data at such a great speed is again a challenge to big data systems.
- Veracity: The data contains a lot of biases, noise and abnormalities and thus, has to be cleaned before processing. Veracity refers to the resistance provided by Big data to keep the "dirty data" from accumulating.
- Volatility: Validity refers to the point after which the data becomes futile for further analysis and hence, must be truncated.
- Validity: Validity tackles the issue of data trustworthiness and data precision. While analyzing colossal amount of data it is of utmost importance that the data is correct and accurate.

The paper is structured as follows, in section II we provide a comprehensive view about the SRP Authentication protocol and propose our modified version to incorporate the authorization phase. Next, in section III we defend the SRP protocol by discussing various security attacks and how it successfully avoids them. Finally in section IV, we conclude our study and also discuss the future scope of this paper.

2. IMPORTANCE OF SECURITY IN BIG DATA
Given the volume and complexity involved with Big Data, providing security becomes a complicated task. Security is one of the most important parameters to consider in Big Data because if the data is intercepted by a rival business firm, it can gain insights and release policies which will combat the

progress of the current firm [5]. There are three main security issue related to Big Data:

- Data Confidentiality

Data confidentiality is defined as protection of data against unauthorized access. It also refers to distorted data from theft [6]. In other words, data confidentiality aims to implement access control mechanisms in order to authorize the users. Big Data invites data from variety of sources and each source has their own unique access control policy. Merging large number of access control policies under one common policy is a complicated task. Further, to combat the velocity characteristic, Big Data authorization should be done automatically rather than manually (digital signatures).

- Data Integrity

Data trustworthiness should ensure that data is always consistent throughout the big data life-cycle [7]. Implementing Data provenance and data correlation techniques is challenging in big data [8].

- Data Privacy

Data privacy aims to protect the intruders from identifying information from the data base. With the advent of social media, the issue of Data Privacy has become momentous as Big data research is being carried out to create and analyze profiles of us and if the gleaned information is not protected, it can be used for nefarious purposes [9].

In this paper we intend to work on the Data confidentiality and data trustworthiness issue related to big data. We explain the Secure Remote Password (SRP) Protocol, which is a strong authentication protocol and propose a modified version which manages to incorporate attribute based access control within the authentication step, in turn speeding up the process, necessary while dealing with Big Data.

3. SECURE REMOTE PASSWORD PROTOCOL

Authentication is the process of identifying an individual, usually by verifying credentials agreed upon by both the parties. In the case of Big Data, authentication becomes complicated as data originating from multiple sources tries to authenticate itself to a common server. Secure Remote Password Protocol (SRP) protocol is a strong authentication protocol providing resistance to both active and passive attacks [10]. The beauty of the SRP protocol lies in its simplicity. It provides zero knowledge security as the password is never shared over the network. It is always computed using some mathematical formulae and that is sent over the network. Hence, the password remains safe with the client at all times. We take inspiration from SRP protocol and propose a modified version of it to incorporate attribute based access control (ABAC) in the same step. The original version of the SRP is described as follows:

Registration Step: Before Authentication, the client has to register itself to the Big Data server. The client, being a normal human being is expected to remember only two things, username and a simple password.

Authentication Step: After registration, the user sends his ID or username to the server to begin the authentication process. The entire process is shown in Fig 1.

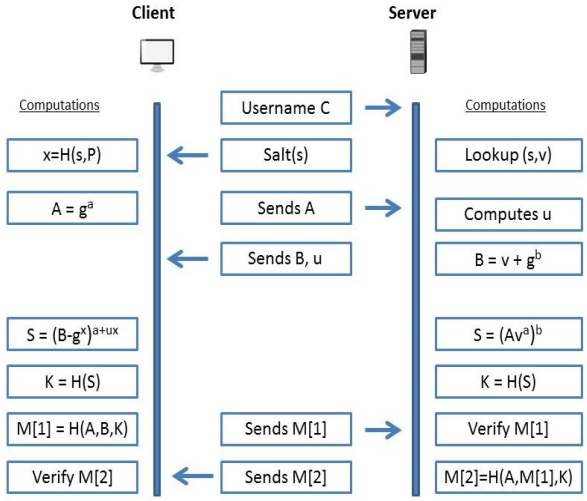

Figure 1 Steps in SRP Protocol

4. LV-SRP PROTOCOL

The modification of SRP protocol, LV-SRP protocol, is proposed to provide attribute based authorization for Big Data servers. In traditionally used protocols, after the authentication is granted authorization is done separately. Authorization or in other words access control is necessary to implement selective restrictions to access resources since, not all users should be given all the access control rights. Some examples of rights or permissions include Read, Write, Modify, Delete, Create, etc. Attribute based access control is defined as a paradigm where access rights are granted to users through the use of policies which combine some attributes together. These attributes can be dynamically changing quantities like time, date, user IP address or company policies like threshold amount withdrawal; threshold permitted transactions, and so on.

The registration step in the LV-SRP protocol is same as the original one. Hence, the user computes the verifier (using v = g^x) and passes the username, salt value and the verifier over an unsecured network to the server. The server stores these values in a table.

The LV-SRP works on the basis of "level" values which are unique values denoting a set of permissions. For instance, Level-1 will have permissions to Read, Level-2 will have permissions to Read, Write and so on.

In order to authenticate itself, the client sends his/her username to the server. Once this is done, the server looks up for each configured attribute and assigns a Level value on the basis of the attribute values. To illustrate this point, we take the example of a banking application. Let us assume it has three constraint attributes: Amount to be withdrawn, Number of Transaction permitted and the Time in MM: SS. The bank sets up a policy and accordingly an if-else ladder is formulated.

if (Amount >= 50,000 || Trans >= 50 || Time >= 18:00)

Assign L = 1; ## only read.

Else if (Amount > = 50,000 && Trans < 50 && Time < 18:00)

Assign L = 2; ## Read, Write.

In such a way, the server calculates the L value for a given user depending on the attribute value at that point of time.

The modified SRP protocol uses a Hash append function to calculate the LV value and fragment the table on the basis of these LV values.

$$LV = A (L, v)$$

Where A () is a function of appending L before the most significant bit of V. This will make the LV values of a specific access control right, say L = 1, fall in a particular range of values. For instance, if L = 1 and v = 24532, the calculated LV value would be 124532. Also, all LV values with L = 1 will fall under the range of 10000 to 19999 (inclusive). Hence, the range of the LV calculated successfully determines the assigned access rights to the user.

Table 1. Table at server end after fragmentation

Labels	Username	Salt	LV
Label 1(R)	.....	.....	.....
Label 2(R,W)	.....	.....	.....
.....	.....	.....	.....
Label n(R,...)	.....	.....	.....

Next, the server sends the salt value as well as the L value to the client. The client generates back 'x' from his secret password. Verifier is again regenerated at the client's end and then appended with the L value to produce LV at the client's end. After this, at every instance instead of V, LV is used.

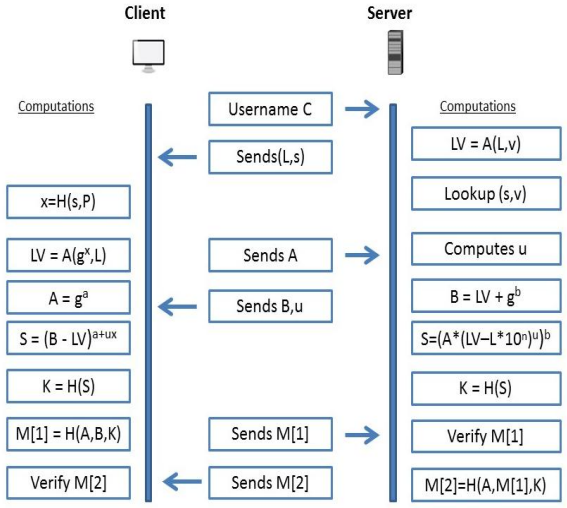

Figure 2. Steps of LV- SRP Protocol

5. ATTACKS AVOIDED BY LV – SRP PROTOCOL

5.1 Brute Force Attack

In this attack, the attacker tries to attain the password or any personal information by a trial-and-error method. The attacker tries all the possible combinations of passwords in order to invade the system and may traverse the entire search space in worst case scenarios in order to invade into the system. However, in case of LV-SRP, the verifier that is computed and sent over the unsecured channel is an exponential value. Therefore, the attack is not only time consuming but is also very costly.

5.2 Dictionary Attack

The dictionary attack is a form of Brute Force attack wherein all the strings listed in the dictionary are tried. The dictionary basically contains those strings that are most likely to succeed [12]. Unfortunately, if the user is made to insert an exponentially hashed value of the password, then the cost associated with this attack surpasses that of what is achieved. Therefore, LV-SRP remains immune to such an attack.

$$V = g^x \qquad \text{.... (i)}$$

where g is the base known publicly. Computing V is known as discrete exponential and its inverse is known as discrete logarithmic. Finding discrete logarithmic is a computationally difficult problem to solve specially for large values of n (512 bits).

5.3 Replay Attacks

A Replay attack occurs when a third party captures a command in transmission and replays it at a later time. By capturing the correct messages, an intruder may be able to gain access to a secure computer or execute commands which are normally encrypted and unreadable. It is often not necessary to decipher the command to use it. Because of this, "Replay attacks are typically simple to perform and require little or no sophistication"[13]. These attacks fail in case of SRP for two reasons. Firstly, the LV-SRP protocol does not involve the transmission of the password through the channel. The values that are transmitted include the verifier and the salt. In case, an eavesdropper obtains the verifier, it is very difficult to obtain the inverse logarithm of such a high exponential number. Let us assume it is a number less than 128-bits. In such a case, finding the inverse logarithm is not that tough. Even if the inverse is found and the 'x' value is obtained, it is still impossible to obtain the password or retransmit this x. The value 'x' is computed using a one-way hash which is practically impossible to reverse. Thus, replay attacks fail in case of LV-SRP.

5.4 Man-In-The-Middle Attack

In Man-In-The-Middle attack an eavesdropper tries to impersonate either the client or server and thus change the message passed or attain valuable information. The attack is formally defined as "Computer security breach in which a malicious user intercepts — and possibly alters — data traveling along a network."[15] In case of LV-SRP, if the eavesdropper attains either the verifier or salt but would still not manage to impersonate either system due to the one way hash function. Although, when he tries to decrypt the message it appears like any normal password but when he uses that as the key, the authentication fails.

The eavesdropper can intercept only A, B, u which is sent over the unsecure network. This will still not lead him in getting the session key, since the LV-SRP Protocol is reducible to Diffie Hellman Protocol [14].

A = g ^ a
B = LV + g ^ b
u = H (A, B)

Table 2. Reduction of LV-SRP to Diffie Hellman

Client	Server
$S=(B - LV)^{a+ux}$	$S = (A(LV-L*10^n)^u)^b$
$S=(LV+g^b-LV)^{a+ux}$	$S = (g^a (g^x)^u)^b$
$S = (g^b)^{a+ux}$	$S = (g^b)^{a+ux}$

As shown in Table 2, the client and server ultimately compute the same values at their respective ends providing no valuable information to the eavesdropper. The eavesdropper has access

only to values such as A, B and u. These values are not enough for the calculation of the session key S, which requires the random variable a ($0 < a < n$) at the client's end and random variable b ($0 < b < n$) at the server's end. The random values are set and known to the client and server respectively and are never shared over the network.

5.5 Denning-Sacco Attack

In the Denning-Sacco attack, the eavesdropper attains 'K' which is the session key which is computed by hashing A, B and K. He then tries to apply brute-force attack and attains the password or tries to impersonate the system. Just by attaining M [1], M [2] or K, the eavesdropper cannot attain any valuable information as explained above. Thus, the SRP is resistant to such attacks.

Thus, this protocol prevents many attacks even if the medium is not completely secure, that is, it resists itself from being compromised. As shown above, the system repels a wide range of attacks thereby protecting each unit and thus, the entire system. Since, the random variables are changed during each authentication cycle, obtaining the past session key would also be of no use. This would prevent any type of past information to enable any compromises in the future.

6. CONCLUSION

Authentication and authorization are done in two separate ways in most systems thus requiring extra time and effort. To improve the efficiency of the system, in this paper, we have tried to coalesce the two pillars of security into a single one. Currently, the most widely used protocol for authentication is Secure Remote Password Protocol. We have proposed a model to provide dynamic attribute-based access control by adopting the research that is done in a static role-based model and integrating it with the SRP Protocol. We have also elaborated the attacks that are repelled by the LV-SRP protocol by mathematically proving the repulsion.

This model is flexible to fit different attribute requirements as per the need of the organization. It strives to provide a stronger and complex access control mechanism due to the use cases that are usually enabled by them. It also makes endeavors to increase the speed at which the amalgamated process is performed in comparison to the traditional methods.

7. REFERENCES

[1] Jainendra Singh, "Real Time BIG Data Analytic: Security Concern and Challenges with Machine Learning Algorithm," IEEE, 2014.

[2] "Why Big Data is important," http://www.navnit.com, May, 2012.

[3] Kevin Normandeu "Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity," InsideBigData, September, 2013.

[4] Elisa Bertino, "Big Data – Security and Privacy," IEEE International Congress on Big Data, 2015.

[5] Akash Gupta, Alok Shukla, S.Venkatesanj, "Big Data: Cryptographically Enforced Access Control and Secure Communication" Proceedings of 6th IRF International Conference, Chennai, India, 10th May, 2014, ISBN: 978-93-84209-16-2I.

[6] L. Xu, D. Sun, and D. Liu, "Study on methods for data confidentiality and data integrity in relational database," 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT '10), vol. 1, pp. 292–295, July, 2010.

[7] Akhil Mittal, "Trustworthiness of Big Data," International Journal of Computer Applications (0975 – 8887), Volume 80 – No.9, October, 2013.

[8] Boris Glavic,"Big Data Provenance: Challenges and Implications for Benchmarking", Specifying Big Data Benchmarks,Volume 8163, Springer Berlin Heidelberg, 2014.

[9] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, "Big Data Privacy Issues in Public Social Media", 6th IEEE International Conference on Digital Ecosystems Technologies (DEST), June, 2012.

[10] Thomas Wu, "The Secure Remote Password Protocol," Internet Society Network and Distributed System Security Symposium, 1998.

[11] Matturdi Bardi, Zhou Xianwei, LI Shuai,Lin Fuhong, "Big Data security and privacy: A review," China Communications Supplement No.2, 2014

[12] Saikat Chakrabarti "Password Based Authentication: Preventing Dictionary Attakcs," IEEE Computer Society, 2007.

[13] Dr. Fred Cohen, http://www.all.net/CID/Attack/Attack76.html

[14] Maryam Ahmed, Baharan Sanjabi, Difo Aldiaz, Amirhossein Rezaei, Habeeb Omotunde, "Diffie-Hellman and Its Application in Security Protocols," IJESIT, November, 2012.

[15] Definition of man-in-the-middle, Webpage 2002-03-26, Retrieved 2002-09, http://www.wordspy.com/words/maninthemiddleattack.asp