

# Securing Big Data over Network using MD5 Algorithm Technique

Bindiya M.K.  
Research Scholar VTU, Belgaum

Ravi Kumar G.K., PhD  
Research Guide, CSE board VTU Belgaum

## ABSTRACT

Big data refers to a collection of information that is too vast and complex to be effectively collected, processed and analyzed using traditional algorithms, tools and approaches. In order to utilize big data, researchers, business and governments are focusing efforts on datasets characterized by three challenges, volume, velocity and variety. These challenges requires research and innovation at all levels of computing, from the physical networks needed for capturing and transporting such data to advanced algorithms for effectively ,securing, organizing, processing and ultimately, making effective use of such data .Now a days Big Data gradually become a hot topic of research and business and has been everywhere used in many industries. Big Data security and privacy has been increasingly concerned. In this paper, first we reviewed some possible methods and techniques to ensure big data security and privacy and then present an MD5 cryptographic algorithm in networking for big data security and privacy.

## Keywords

Big data, security and privacy, RACK

## 1. INTRODUCTION

Big Data is used to describe a massive, volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques [1].Recently, there has been an increasing interest in Big Data. However, the term Big Data remains vague. Big Data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. A widely recognized definition belongs to IDC: “big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, analysis”[2]. The current era is the Age of Big Data” [3]. In the past few years, the total amount of data created by human has exploded [4]. From 2005 to 2020, the amount of data is predicted to increase 300 times, from 130 hexabytes to 40,000 hexabytes [5]. These and business informatics, government, Internet search, social networks, document, photography, audio, video, logs, click streams, mobile phones, sensor networks and so on. Big Data is the result of the dramatic increase of data.

Big Data have significant value. Many Organizations worldwide collect and analyze their own business process data in order to improve their internal decision making. The Obama regime has announced a Big Data research and development initiative based on recognition of the great social and economic value captured in the data in 2012 [6]. Big Data promotes the economy and scientific research, transforms traditional business models and scientific methods and creates new opportunities through data analysis data are generated

from scientific research, finance transforms traditional business models and scientific methods and creates new opportunities through data analysis [7].

However, exploring and using the extraordinary value of Big Data must increase risks of security and privacy. For example, “Amazon monitors our shopping preferences and Google learns our browsing habits, while Twitter knows what’s on our minds. Facebook seems to catch all that information too, along with our social relationships. Mobile operators know not only whom we talk to, but who is nearby. With Big Data promising valuable insights to those who analyzes it, all signs seem to point to a further surge in others’ gathering, storing, and reusing our personal data.” [8]. If the internet age threatened security and privacy, the age of Big Data endanger them even more.

Big Data security usually is to the use of the Big Data to implement solutions increasing security, reliability, and safety of a distributed system. Big Data privacy focuses on the protection of Big Data from unauthorized use and unwanted inference [9].

It is well known that big data are a priceless source of information at the basis of robust and accurate security solutions. However, Big Data often contain sensitive information that needs to be protected from unauthorized access and release. Obviously, there are not any challenges f security and privacy if we do not extract value from Big Data. Thereby, the principles of Big Data security and privacy must be balanced against additional societal value of Big Data.

In this paper, first section discusses some of the possible methods and techniques to ensure big data security and privacy and then present an MD5 cryptographic algorithm in networking for big data security and privacy. And rest of the paper section II reveals big data security and privacy techniques and some possible methods to ensure big data security and privacy. Section III shows proposed method to ensure the big data security and privacy using cryptographic MD5 algorithm in networking. Section IV shows Design Frame work and Section V shows the Results obtained performance and evaluation. Section VI is the conclusion of this paper.

## 2. BIG DATA SECURITY AND POLICY TECHNIQUES

Organizations used various methods of de-identification to ensure security and privacy. The most common solution to ensure security and privacy may be oral and written pledges. However, history has shown that this method is flawed. Passwords, controlled access, and two-factor authentication is low-level, but routinely used, technical solution to enforce security and privacy when sharing and aggregating data across dynamic, distributed data systems. Access permissions such as these can potentially be broken by both the intentional

sharing of permissions and the continuation of permissions after they are no longer required or permitted. More advanced technological solution is cryptography. The famous encryption schemes have AES and RSA. Recent revelations show that the National Security Administration (NSA) may have already found ways to break or circumvent existing Internet encryption schemes [10]. Virtual barriers such as firewalls, secure sockets layer and transport layer security are designed to restrict access to data. Each of these technologies can be broken, however, and thus need to be constantly monitored, with fixes applied as needed. Tracking, monitoring or auditing software is developed to provide a history of data flow and network access by an individual user in order to ensure compliance with security related. The limitation of this technology is that it is difficult and costly to implement on a large scale or with distributed data systems and users because it requires dedicated staff to read and interpret the findings, and the software can be exploited to monitor individual behavior rather than protecting data. All in all, the traditional de-identification techniques are not applicable in the era of Big Data since the de-identification technique widespread uses. The tasks of ensuring Big Data security and privacy become more difficult as information is increased. “Computer scientists have repeatedly shown that even anonymized data can often be re-identified and attributed to specific individuals” [11].

A novel technological named the integrated Rule-Oriented Data (iRODS) is proposed to be the solution to ensure security and privacy in big data [10]. iRODS was architected and designed to address these challenges across a broad spectrum of communities, with differing institutional goals and security and privacy concerns, by providing each adopter community the ability to develop and deploy solutions for data management and sharing that are specific to organizational needs [12][13]. Key technological features of iRODS include: federated data grids or “intelligent clouds”, a distributed rules engine, an “iCAT” metadata catalog, a storage access layer that allows common access, a rich combination of graphical user interface and command-line-based clients and APIs for interaction with an iRODS data grid. iRODS is used in a number of data management applications and has been adopted by numerous institutions around the world. Many publications describe the myriad ways in which iRODS technology has been adapted and applied to solve of variety of challenges in policy based, large-scale data management [12][13][14][15]. The iRODS technology provides improvements in common approaches to securing data and ensuring privacy, including: comprehensive set of security controls, improved control of data access and use through metadata, storage virtualization and data security lifecycle, and persistent identifiers. The big data security techniques data also involve released anonymity protection, social networking anonymity protection and data provenance and so on. For structured data in big data, the data released anonymity protection which is the basic means and key technologies to achieve protection of privacy are still in the stage of continuous development and improvement. The early [16] [17] and optimal [18] [19] [20] k-anonymity protecting scheme focus on static and one-time data released situation. In reality data released often faced continuous and repeatedly situation. In contrast, the data released anonymity protection is more sophisticated in big data scene.

### 3. PROPOSED METHOD

Cryptographically Enforced Access Control and Protected data experiment was carried with a new technique (Read Acknowledgement) RACK, with the utilization of MD5 algorithm in big data, to ensure that the most sensitive private data is end-to-end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. Specific research in this area such as attribute-based encryption has to be made richer, more efficient, and scalable. To ensure authentication agreement & fairness among the distributed entities a cryptographically secure communication frameworks has been implemented. Sensitive data is routinely stored unencrypted in the cloud. The main problem to encrypt data, especially large

Data sets, is the all-or-nothing retrieval policy of encrypted data, disallowing users to easily perform fine grained actions such as sharing records or searches. On the other hand, we have encrypted less sensitive data as well, such as data useful for analytics. Such data has to be communicated in a secure and agreed-upon way using a cryptographically secure communication framework

### 4. DESIGN FRAME WORK

#### Data Structure:

The structure of the data flowing is described below,

#### 1. Request data structure

This is the data or the message sending from the sender to the receiver.

MD5(SenderId:SenderPassword)|  
 Base64(length)|Base64((message|messageID))|MD5(EndStrin)

#### 2. Response data structure

Response data structure refers to response message from the receiver to sender on

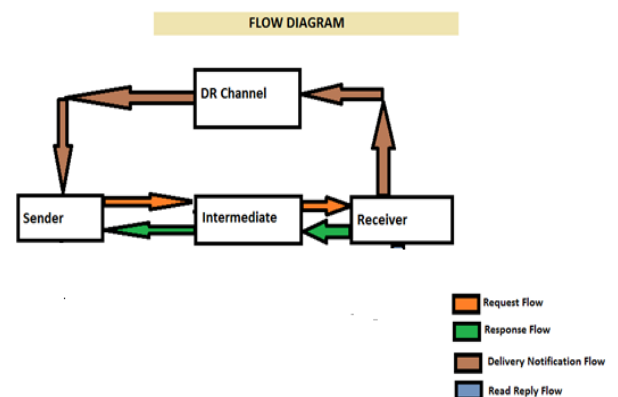
Request message sent, delivery notification message and read reply message.

MD5(user:pwd)|Base64(msgId|Status)|MD5(endStr)

Response: ACK: SUCCESS/ACK: Failure

Delivery Notification: SUCCESS/Failure

ReadReplyNotificaiton: SUCCESS/Failure



## PORTS UTILIZATION

Sl. No	Component	Listener	Port Number
1	Sender	Delivery Notification	17001
		Read Reply Notification	17002
2	Intermediate	Request Listener	7001
3	Receiver	Request Listener	7002
4	Delivery Report Channel	Deliver Report Notification	27001

**Sender:** Is the request initiator. In reference to the mobile technology, Sender will be referenced to Mobile user. Once the sender sends the request, two servers (listeners) will be waiting, one for delivery notification and another for read reply notification.

**Intermediate:** This is the intermediate nodes on the network which carries or forwards the data to the recipient/receiver. There can be huge number of such intermediate routers in the production scenario. However, for the simulating environment we have used only one such intermediate carrier.

**Receiver:** This is the actual final recipient, who is supposed to receive the message. In reference to the mobile technology, receiver refers to either mobile user or the application (short code). Receiver will be waiting for the incoming messages and on once received, will send the delivery notification to the sender to acknowledge the message received.

**Read Reply Channel:** This is another channel who will notify the sender that receiver has read the message/data he sent

**Hash Function:** Hashing techniques available are based on the concept of a hash function that transforms a given input of arbitrary length to a value of a fixed length, called the hash code. The transformation is done in a manner that it is computationally infeasible to transform the hash value to the original value. Hash functions are very efficient as they do not involve heavy computations and hence are applied in the area of security for message authentication and integrity checks.

Base64:Base64 is an encoding scheme that represents binary data in an ASCII string format. It is commonly used when there is a need to encode binary data that needs to be stored and transferred over media that are designed to deal with textual data. Base64 encoding takes three bytes, each consisting of eight bits, and represents them as four printable characters in the ASCII standard

## 5. RESULTS OBTAINED

```

[sushma@localhost SourceCode]$ ./run_sender
Starting the Delivery Notification Server
Waiting for the data
Starting the Read Report Server
Waiting for the data

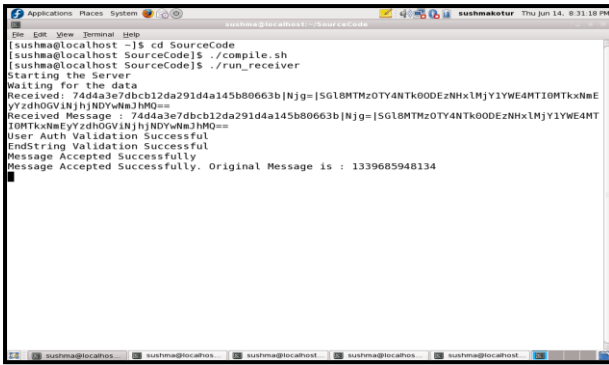
Type the message to send and press ENTER : Hi
    
```

Fig 1: Sender Channel

```

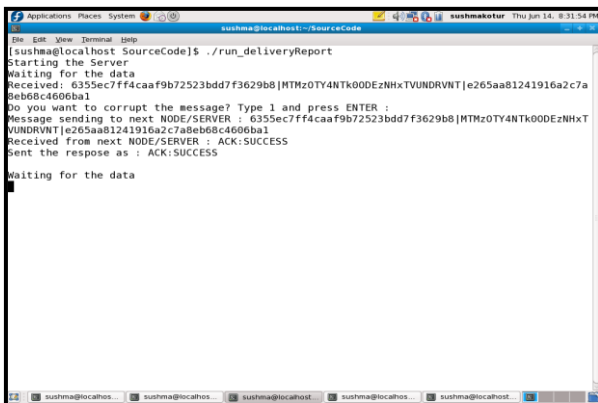
[sushma@localhost SourceCode]$ ./run_intermediate
Starting the Server
Waiting for the data
Received: 74d4a3e7dbcb12da291d4a145b80663b|Njg=|SgI8MTMzOTY4NTk0ODEzNHxLMjY1YWE4MTI0MTkxNmE
yYzdhOGVlNjhjNDYwNmJhMQ==
Do you want to corrupt the message? Type 1 and press ENTER :
Message sending to next NODE/SERVER : 74d4a3e7dbcb12da291d4a145b80663b|Njg=|SgI8MTMzOTY4NTk
00DEzNHxLMjY1YWE4MTI0MTkxNmEyZdhOGVlNjhjNDYwNmJhMQ==
    
```

Fig 2: Intermediate Channel



```
[sushma@localhost ~]$ cd SourceCode
[sushma@localhost SourceCode]$ ./compile.sh
[sushma@localhost SourceCode]$ ./run_receiver
Starting the Server
Waiting for the data
Received: 74d43e7dbc12da291d4a145b80663b|Njg=|SjG18MTMzOTY4NTk0ODUzNHx1MjY1YWE4MTIOMTKxNmE5YzdhOGVlbnJhNDYwNmJHM0==
Received Message : 74d43e7dbc12da291d4a145b80663b|Njg=|SjG18MTMzOTY4NTk0ODUzNHx1MjY1YWE4MTIOMTKxNmE5YzdhOGVlbnJhNDYwNmJHM0==
User Auth Validation Successful
EndString Validation Successful
Message Accepted Successfully
Message Accepted Successfully. Original Message is : 1339605948134
```

Fig 3: Receiver Channel



```
[sushma@localhost SourceCode]$ ./run_deliveryReport
Starting the Server
Waiting for the data
Received: 6355ec7ff4caaf9b72523bd47f3629b8|MTMzOTY4NTk0ODUzNHx1TVUNDRVNT|e265aa81241916a2c7a8eb6c4606ba1
Do you want to corrupt the message? Type 1 and press ENTER :
Message sending to next NODE/SERVER : 6355ec7ff4caaf9b72523bd47f3629b8|MTMzOTY4NTk0ODUzNHx1TVUNDRVNT|e265aa81241916a2c7a8eb6c4606ba1
Received from next NODE/SERVER : ACK:SUCCESS
Sent the response as : ACK:SUCCESS
Waiting for the data
```

Fig 4: Delivery Channel

## 6. CONCLUSION

Security and privacy is essential service for the big data in networks. The characteristics of data that pose both challenges and opportunities in achieving security goals, such as confidentiality, authentication, integrity, availability, and non-repudiation. This paper gives a brief idea about the RACK scheme and discussed different aspects of the RACK scheme for large set of data. The RACK scheme uses cryptographic based technique to find out misbehaving activities during the transmission of data. Our analysis and experiment shows that proposed technique is efficient and practical. The future scope is once the data is received at the receiver side, user can apply map reduce framework to resolve the big data problem.

## 7. ACKNOWLEDGMENTS

The author would like to thank Dr. Ravi Kumar G.K. for his continuous support and guidance for carrying the research work.

## 8. REFERENCES

- [1] Big Data Security and Privacy: Review MATTURDI Bardi1, ZHOU Xianwei, LI Shuai, and LIN Fuhong.2014:135-145.
- [2] Gantz J, Reinsel D. Extracting value from Chaos [J]. IDC iView, 2011: 1-12.
- [3] Lohr S. The age of big data [J]. New York Times, 2012, 11.
- [4] McCune J C. Data, data everywhere [J]. Management Review, 1998, 87(10): 10-12.
- [5] Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and Biggest growth in the Far East [J]. IDC iView: IDC Analyze the Future, 2012.

- [6] Weiss R, Zgorski L. Obama administration Unveils “Big Data” Initiative: Announces \$200 Million in New R&D Investments [J]. Office of Science and Technology Policy, Washington, DC, 2012. Data P. The emergence of a New Asset Class[C]//World Economic Forum Report. 2011.
- [7] Anderson C. The end of theory: the data Deluge makes the scientific method obsolete. Wired Magazine 16.07[J]. 2008.
- [8] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity [J]. 2011
- [9] Mayer-Schönberger V, Cukier K. Big data: Arevolution that will transform how we live, Work, and think [M]. Houghton Mifflin Harcourt, 2013.
- [10] Perlroth N, Larson J, Shane S. NSA able to foil basic safeguards of privacy on web[J]. The New York Times, 2013,
- [11] Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization [J]. UCLA L. Rev., 2009, 57: 1701.
- [12] Rajasekar A, Moore R, Hou C, et al. iRODS Primer: integrated rule-oriented data System [J]. Synthesis Lectures on Information Concepts, Retrieval, and Services, 2010, 2(1): 1-143.
- [13] Rajasekar A, Moore R, Wan M, et al. Applying rules as policies for large-scale data Sharing[C]//Intelligent Systems, Modelling And Simulation (ISMS), 2010 International Conference on. IEEE, 2010: 322-327.
- [14] Barg I, Scott D, Timmermann E. NOAO E2E integrated data cache initiative using iRODS[C]//Astronomical Data Analysis Software and Systems XX. ASP Conference Proceedings. 2011, 442: 497-500.
- [15] Schnase J L, Webster W P, Parnell L A, et al. The NASA Center for Climate Simulation Data Management system[C]//Mass Storage Systems and Technologies (MSST), 2011 IEEE 27th Symposium on. IEEE, 2011: 1-6
- [16] Sweeney L. k-anonymity: A model for pro- tecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 557-570.
- [17] Sweeney L. Achieving k-anonymity privacy protection using generalization and Suppression [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 571-588.
- [18] Bayardo R J, Agrawal R. Data privacy throughoptimal k-anonymization[C]//Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on. IEEE, 2005: 217-228.
- [19] Lefebvre K, DeWitt D J, Ramakrishna R.Incognito: Efficient full-domain k-anonymity[C]//Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005: 49-60.
- [20] Lefebvre K, DeWitt D J, Ramakrishna R.Mondrian multidimensional k-anonymity[C]//Data Engineering, 2006. ICDE'06.Proceedings of the 22nd International Conference on. IEEE, 2006: 25-25.