Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Using Encryption Technique for Effective Data Storage

Sandeep Kaur
Research Scholar, Punjabi University,
Patiala, Punjab, India

Abstract: Today we are living in a digital world. Data is generated from sensors, social networking sites and satellites at an unbelievable speed. In this digital world, data is generated and collected at a rapid rate that is exceeding the boundary range. According to IBM, 2.5 quintillion bytes of data are generated every day. The data with high speed, high variety and high volume is known as Big Data. Integrity, privacy, security, scalability, analysis and storage of data are main challenges of Big Data. Big Data Storage is a challenging issue in Big Data. Various techniques like MongoDB, SimpleDB, BigTable, Dynamo and Apache Hbase are used for Big Data Storage. In this paper, a new storage method is proposed. A client- server framework is generated in which data is stored on the server after getting splitted and encrypted.

Keywords— Big Data, Data Management, Data Storage, Security and Diffie-Hellman

I. INTRODUCTION

Today is the era of cloud based applications and services deployed over the cloud. Cloud computing refers to distributed computing or delivery of computing services over the Internet. Cloud computing services allow individuals and organizations to use software and hardware that are managed by third parties at remote locations. Cloud services include online file storage, webmail, social networking sites, and online business applications. Big Data and Cloud Computing are conjoined with each other. Cloud Computing provides the underlying engine through the use of Hadoop and Big Data provides users the ability to use commodity cluster computing for processing the distributed queries across multiple datasets and return the resultant sets in a timely manner. Rather than using local storage, big data uses distributed storage technology based on cloud computing.

According to McKinsey in [18], Big Data refers to datasets whose size is beyond the capability of typical database software tools to capture, store, manage, and analyse. Big data has four main characteristics as volume, velocity, variety, and value. It is a recent phrase that has come into existence as sheer volume of data is being generated today in almost every aspect of life. Various challenges are associated with big data such as privacy and security, access and sharing information, data management, analytical and technical challenge. Data management further consists of challenges like data storage, data integration, data variety, data processing and resource management.

Data storage is one of the most important challenge that requires much attention as data is growing at unprecedented rates. Traditional storage devices and technologies fail to compensate the needs of storing big data as they have high failure rates and replacement costs and less scalability. Hadoop, Google File System, NoSQL and NewSQL data stores have been used widely for big data storage. As a part of this research, a framework has been designed that incorporates splitting and storage of data based on type of the file for secure storage, splitted data is further encrypted and then stored. The designed framework is a local -client server framework in which users or clients can store or upload their data on servers securely. The proposed framework can be adopted for big data storage.

In this paper, firstly literature survey related with Big Data and its storage and management is discussed. In the next sections, proposed methodology, results and conclusion and future scope are discussed.

II. LITERATURE SURVEY

Many of the previous work has been done in the field of big data and its storage . The study of work done in context of data management and storage is as below-

Big Data is defined in terms of variety, volume, velocity, variability, complexity and value. Two challenges are associated with data in terms of storage and processing. One is to handle or store large amount of data efficiently and effectively. Second is to filter the most important data from all the data collected by the organization. They proposed building up indexes right in the beginning while collecting and storing the data in order to reduce processing time [6]. The implementation of big data should rely on the technologies and methods, i.e., low cost of storage devices, widely used of sensors and data capture technologies, which also combine cloud service, virtual storage equipment and advanced software technology [8]. The architecture used in big data named Lambda specifies a data store that removes the update and delete aspects and only allows creation and reading of records [16].

Chang et al. [1] gave the concept of Bigtable, a sparse, distributed, persistent multidimensional sorted map. The map is indexed by a row key, column key, and a timestamp and each value in the map is an uninterpreted array of bytes.

It is a distributed storage system for managing structured data that is designed to scale to a very large size. Many projects of Google store data in Bigtable, including web indexing, Google Earth, and Google Finance.

Azemovic and Music [2] proposed Hybrid way of storing unstructured data after discussing previously used methods of storing unstructured data i.e. Unstructured data inside relational database environment and Unstructured data outside relational database environment. The proposed method combines the features of both methods and they considered this method an efficient method for storing data.

Ji et al. [3] proposed a general view of Big Data Management (BDM) technologies and applications. They categorized BDM system into three parts namely distributed file system, non-structural and semi-structured data storage and open source cloud platform. Distributed File System include Google File System , Hadoop Distributed File System, Amazon Simple Storage Service, Elastic Storage System . Non-structural and Semi-structured Data Storage include Big Table, PNUTS, Dynamo, Llama.

Zaslavsky et al. [4] suggested different data storage products like Greenplum, IBM DB2 or Netezza, Microsoft SQL Server, MySQL, Oracle, or Teradata as storage is an essential component in big data. In NoSQL technologies, there are four varieties: Key-value, document store, wide column stores, and graph databases. Popular key-value based data storage products are Dynamo, Amazon SimpleDB, and Windows Azure Table Storage. Popular document store technology based products are CouchDB and MongoDB. Popular wide column based products are Apache HBase Hadoop and Cassandra. They categorized big data management challenges into two categories namely engineering and semantic related to perform data management activities such as query and storage efficiently and is to extract the meaning of the information from massive volumes of unstructured data.

Song [5] discussed the challenges of big data storage such as storage area networking and network attached storage solutions adopted in the enterprise environment are not suitable for such big data environments due to their high cost and complexity. They further discussed project named Apache Hadoop that is the most well known industry initiative to develop open-source software for reliable, scalable, distributed storage and computing.

According to the survey report of 2013 done by Philip Russom [7], Hadoop Distributed File System, MapReduce, and other tools are widely used for big data management. Others schemes for big data management include complex event processing (for streaming big data), NoSQL databases (for schema-free big data), in-memory databases (for real-time analytic processing of big data), private clouds, in-database analytics, and grid computing.

Kaisler et al. [9] discussed various issues related to big data like storage and transport issues, management issues and processing issues. In context of big data security, IDC suggested five levels of increasing security: privacy, compliance - driven, custodial, confidential and lockdown.

Zhang and Xu [12] discussed four techniques of distributed file system for big data storage that is storage of small files, load balancing, copy consistency and deduplication after discussing the challenges of big data storage.

Tao et al. [11] proposed a management and analysis system for structured big data called ThumpStorage by integrating the bottom distributed structure of Hadoop distributed file system (HDFS) and the partitioning and scheduling technology of the massive parallel processing database. This system shows high efficiency, low latency and high scalability. It is applicable for managing and analyzing the massive structured data at PB level above.

Zhou et al. [13] analyzed and characterized the performance and energy impact brought by deduplication under various big data environments. There are three sources of redundancy in big data workloads that are deploying more nodes, expanding the dataset and using replication mechanisms. Deduplication is being widely used to reduce cost and save space in data centers. It eliminates redundancy by removing data blocks with identical content. The benefits of this technique are saving disk space, which further leads to saving money on buying storage devices and reducing IO traffic, which results in higher I/O throughput. Since redundancy is inevitable in big data workloads, this technique is valuable for a big data storage environment.

Khan et al. [17] proposed a data life cycle that uses technologies and terminologies of big data for big data management. The proposed data life cycle consist of different stages i.e. collection, filtering & classification, data analysis, storing, sharing & publishing, and data retrieval & discovery. In storing phase, it consists of management plans, content filtering, distributed system, partition tolerance and consistency and use Simple DB, Bigtable, Hadoop, MapReduce, Memcache DB and Voldemort techniques for data storage. They further discussed big data security in terms of privacy, integrity, availability and confidentiality.

M. et al. [15] proposed a two-layer architecture based on a hybrid storage system that is able to support a Platform as a Service (PaaS) federated Cloud scenario. The proposed architecture combines the benefits of SQL and non-SQL database solutions. It allows us to extend SQL-like legacy systems and manages big data through an XML-like, non-SQL distributed storage system according to a Cloud federation approach. It provides high stability and efficiency for the management of information inside a Cloud.

Li et al. [14] illustrated the big data storage and management status, analyzed the problems and finally proposed solutions. The most applicable technologies of storage and management are: distributed file system, distributed database, access interface and query language. Distributed file system store the data in distributed nodes and devices, the nodes are connected by network. Distributed database contains NoSQL and NewSQL. NoSQL system tends to increase performance and scalability. He advised us to use optimize storage techniques i.e. storage systems which are efficient and have reasonable cost that will reduce costs while ensuring the performance. Improving index of the big data and query technology in distributed system and storage and processing of real-time and streaming big data are solutions for big data storage and management.

III. PROPOSED METHODODLOGY

Methodology is the systematic study of methods that are, can be, or have been applied within a discipline. Here, research methodology is the experimental study in which a framework is designed to store data securely on the servers. In this framework, two methods are evaluated- base paper [10] method i.e. Information Dispersal technique with AES encryption is evaluated and then this IDA technique method is joined with Diffie Hellman encryption and then results of both methods are compared. The following steps are to be followed:

a) Set-up of Client - Server Environment
b) Categorization of Files
c) Splitting and Encryption
d) Allocation of files to memory

The description of steps is as -

Set-up of Client - Server Environment: A local client-server environment can be generated using .net framework. For generating this environment, two softwares are required that are Microsoft Visual Studio 2010 and SQL Server Management Studio 2008. Microsoft Visual Studio is an integrated development environment from Microsoft. It is used to develop computer programs for Microsoft Windows, websites, web applications and web services. It supports different programming languages and allows the code editor and debugger to support nearly any programming language. Built-in languages include C, C++ and C++/CLI, VB.NET and C#. Visual Studio 2010 comes with .NET Framework 4 and supports IBM DB2 and Oracle Databases, in addition to Microsoft SQL Server. It includes tools for debugging parallel applications. Download Microsoft Visual Studio 2010 from the link i.e. http://www.visualstudio.com/ and install it into the computer and start generating this environment.

Categorization of Files: In this client - server environment, clients store their data on the server after logging into the system. Here data or files are stored on the basis of data type. Data can be in any form. Image, audio, video, and text are different forms of data. This framework categorizes data into four forms on the basis of data type i.e. txt, jpg, doc and mp3. The data type of file is checked as user attempts to store data on the server. By checking this data type, the server manages the files according to type i.e. text related files are stored in text files allocated memory block and image related files are stored in image files allocated memory block and so on.

Splitting and Encryption : After checking data type, file gets splitted and encrypted. In this process, file gets splitted into many parts. This splitting process is different for each type of file and is based on size of file and splitting process is shown in table 1. After getting splitted, each part gets encrypted separately. For encryption, this environment uses Diffie-Hellman algorithm for proposed method and AES algorithm for integrated method. Diffie-Hellman establishes a shared secret key that can be used for secret communications while exchanging data over a public network. This allows users to exchange keys in a manner that does not allow an eavesdropper to generate the key in a fast manner. Here small data sizes are taken for evaluating these two methods. This local client-server environment can also be tested for larger data sizes.

Table 1. Splitting of data in parts

Type	Size	Parts
Audio	<= 2 MB	4
Audio	2 - 5 MB	5
Audio	> 5 MB	7
Doc	<= 200 KB	5
Doc	200 - 500 KB	8
Doc	> 500 KB	10
Image	<= 200 KB	4
Image	200 - 400 KB	9
Image	> 400 KB	16
Text	<= 40 KB	4
Text	40 - 180 KB	6
Text	> 180 KB	8

Splitting procedure is same for integrated and proposed method for secure data storage.

Allocation of files to memory : As upload button is clicked, after splitting and encryption procedure then all the data gets their space on the memory and stored on the server.

IV. RESULTS AND DISCUSSIONS

In this section, splitting results as well as time analysis and throughput of Integrated Method and Advanced Method is discussed.

a. **Splitting Results**: In splitting process, file gets splitted in different parts based on data type of file and as well as size of file. Splitting results of 9 files is shown in table 2.

Table 2. Splitted data results

Filename	Size	Splitting Parts	Size of each part
audiofile11	1.26 MB	4	324 KB
audiofile21	4.40 MB	5	903KB
audiofile32	8.72 MB	7	1.25 MB
docfile11	132 KB	5	27 KB
docfile21	476 KB	8	60 KB
docfile31	1.44 MB	10	147 KB
textfile22	134 KB	6	23 KB
textfile31	218 KB	8	28 KB
audiofile34	29.6 MB	7	4.24 MB

b. **Time Analysis** : Time can be measured for storing each file on server. Here Time can be defined as a time taken to store data on the server. Also total time can be calculated by adding the time of all files that are stored on the server. Total time can be calculated as :

$$Total\ time\ (T_{Total})\ = \sum_{i=0}^{n} T_i$$

Table 3 shows time analysis of integrated method. Here time is calculated for storing 11 files on integrated method based servers. Total time taken by this server for storing 11 files is 35.12 seconds.

Table 3. Time Analysis (Integrated Method)

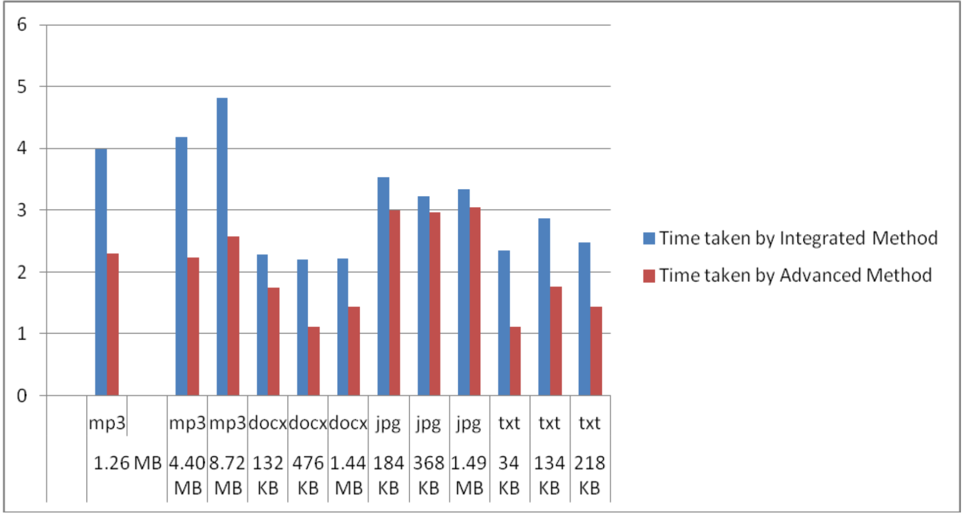
S.No.	Filename	Size	Type	Time(sec.)
1	audiofile11	1.26 MB	mp3	3.99
2	audiofile21	4.40 MB	mp3	4.19
3	audiofile32	8.72 MB	mp3	4.81
4	docfil11	132 KB	docx	2.29
5	docfil21	476 KB	docx	2.20
6	docfil31	1.44 MB	docx	2.22
7	imagefil13	184 KB	jpg	3.54
8	imagefil21	368 KB	jpg	3.22
9	imagfileb31	1.49 MB	jpg	3.33
10	textfileb 22	134 KB	txt	2.86
11	textfileb31	218 KB	txt	2.47

Table 4 shows time analysis of Proposed Method. Here time is calculated for storing 11 files on Proposed Method based servers. Total time taken by this server for storing 11 files is 23.61 seconds.

Table 4. Time Analysis (Proposed Method)

S.No.	Filename	Size	Type	Time(sec)
1	audiofilea11	1.26 MB	mp3	2.30
2	audiofilea21	4.40 MB	mp3	2.23
3	audiofilea32	8.72 MB	mp3	2.57
4	docfilea11	132 KB	docx	1.74
5	docfilea21	476 KB	docx	1.12
6	docfilea31	1.44 MB	docx	1.43
7	imagfilea13	184 KB	jpg	3.00
8	imagefilea21	368 KB	jpg	2.97
9	imagfile31	1.49 MB	jpg	3.05
10	textfilea22	134 KB	txt	1.76
11	textfilea31	218 KB	txt	1.44

Comparative time analysis proofs that Advanced method takes less time as compared to integrated method to store data as shown in graph 1. So, Advanced method is better than integrated method in terms of storage time.



Graph 1. Comparative Time Analysis

c. Throughput Analysis: Framework can also be measured by calculating throughput . Here throughput can be defined as rate of transactions done in a given time. It can be calculated as :

$$Throughput = \frac{Total\ no.of\ files}{Total\ Time} * 100$$

In case of Integrated Framework,

$$Throughput = \frac{11}{35} * 100 = 31\%$$

In case of Integrated Framework,

$$Throughput = \frac{11}{24} * 100 = 46\%$$

Calculation of throughput shows that Advanced method is more efficient than the Integrated method .

Thus, from above results it is proved that Advanced method performs much better than the previous method i.e. Integrated method in terms of time and throughput also.

V. CONCLUSION AND FUTURE WORK

Local client-server environment is generated that stores and manages data properly. The Diffie- Hellman algorithm is used for data encryption. This algorithm is faster and secure than AES algorithm. Future work can be done to test this environment with different forms of data like video and pdf files etc. and check the efficiency of this framework with large data sizes. During splitting process, some data might be lost, work can be done to recover the entire data. Here work is only done for storing or uploading, future research can be done in downloading this stored data. During downloading, splitted parts will be joined. So, joining process will be done in such a manner that quality of data remains same as that of original file.

REFERENCES

[1] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, "*Bigtable: A Distributed Storage System for Structured Data*", Appear in OSDI : Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November 2006

[2] Jasmin Azemovic and Denis Music, "*Comparative analysis of efficient methods for storing unstructured data into database with accent on performance*", 2nd International Conference on Education Technology and Computer (ICETC), Vol:1, pp. V1-403-V1-407, 2010

[3] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, "*Big Data Processing in Cloud Computing Environments*", Published in IEEE journal, pp. 17-23, 2012

[4] Arkady Zaslavsky, Charith Perera, Dimitrios Georgakopoulos "*Sensing as a Service and Big Data*", Proceedings of the International Conference on Advances in Cloud Computing (ACC), Bangalore, India, July 2012

[5] Young - Sae Song, "*Storing Big Data - The Rise of the Storage Cloud* ", seamicro.com, pp. 1-5, December 2012

[6] Avita Katal, Mohammad Wazid, R H Goudar, "*Big Data: Issues, Challenges, Tools and Good Practices*", Published in IEEE journal, pp. 404-409, 2013

[7] Philip Russom, "*Managing Big Data*", tdwi.org, TDWI Best Practices Report , pp. 1-36, 2013

[8] Jinsong Zhang, Yan Chen, Taoying Li, "*Opportunities of Innovation under Challenges of Big Data*", 10th International Conference on Fuzzy Systems and knowledge Discovery, Published in IEEE, 2013

[9] Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money, "*Big Data : Issues and Challenges Moving Forward* ",46th Hawaii International Conference on System Sciences, Published in IEEE, 2013

[10] Ahmad B. Alnafoosi, Theresa Steinbach, " *An Integrated Framework for Evaluating Big-Data Storage Solutions – IDA Case Study* ", Science and Information Conference (SAI), London, UK, pp. 947 - 956, October 2013

[11] Xu Tao , Fu Ge , Tan Huaiyuan , Zhang Hong and Liu Xinran , "*ThumpStorage : A Management and Analysis System for Structured Big Data*", International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC) , pp. 2424 - 2427, 2013

[12] Xiaoxue Zhang and Feng Xu, "*Survey of Research on Big Data Storage*", 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES), pp. 76-80, September 2013

[13] Ruijin Zhou, Ming Liu and Tao Li, "*Characterizing the efficiency of Data Deduplication for Big Data Storage Management*", IEEE International Symposium on Workload Characterization(IISWC), pp. 98-108, September 2013

[14] Jie Li, Zheng Xu, Yayun Jiang, Rui Zhang, "*The Overview of Big Data Storage and Management*", IEEE 13th International Conference on Cognitive Informatics & Cognitive Computing, pp. 510-513, 2014

[15] Fazio M., Celesti A., Villari M. and Puliafito A., "*The Need of a Hybrid Storage Approach for IoT in PaaS Cloud Federation*", 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 779-784, May 2014

[16] Mahendra S. Patil, Jinesh K. Kamdar and Chintan B. Khatri, "*Big Data - An Overview*", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 7, July 2014

[17] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz and Abdullah Gani, "*Big Data: Survey, Technologies, Opportunities, and Challenges*", Hindawi Publishing Corporation, The Scientific World Journal, Article ID- 712826, pp. 1-18, July 2014

[18] Wenhong Tian, Yong Zhao, "*Big Data Technologies and Cloud Computing*", Optimized Cloud Resource Management and Scheduling Theory and Practice, pp. 17-49, October 2014