Secure And Cost-Effective Big-Data Analysis In Cloud Computing

Jayaraj T, J. Abdul Samath

Abstract: Storing and analyzing Big-Data in cloud computing is now gaining popularity. Together, these two technologies provide powerful information and benefits for the business. There are some inevitable challenges when combining these two powerful technologies. The first big-data analysis requires a great deal of computational resources. This is considered a barrier to big-data analysis. Data security becomes a question mark when we store big-data in the cloud. The main objective of this proposed research is to develop a secured and cost effective system for big-data analysis and storage in cloud computing. A good data deduplication algorithm is first developed, and this data deduplication algorithm reduces the unnecessary processing power and storage space that it takes for Big-Data processing and storage. This is very useful for small and medium type of business. The second is to propose a better security model. This multiplies the Cloud Services User's (CSU) trust in the Cloud Service Providers (CSP). Finally the proposed method is evaluated by existing methods. Experimental results make it clear that this proposed method requires minimal storage space and processing power.

Index Terms: Cloud Computing, Big-Data, Data Security, Data Mining, Data Deduplication.

1 INTRODUCTION

Owing to the rapid development of science and technology data are produced consistently. The modern society has entered the era of big data [6][7]. Business industries have set up different types of applications to gather useful information from the big data. However, all these data sources have different operating environment and different data formats. To gather useful information from these is a matter of challenge. Now a day, small, medium and large industries are using cloud services to store and analyze their data. These cloud services provides the Pay-As-You-Go (PAYG) model for business when needed. Cloud services provide many significant benefits for big data, which include: High computational cost, storage, automated tools and reconfigurable verbalized resources. It is very useful for large organizations to develop their business [8][9]. At the same time, most of the data mining algorithms consume large amounts of computational resources to analyze this big data. This means enterprises need to pay huge amounts of money unnecessarily to cloud service providers. Also, analyze big data and get information from it businesses and organizations need to wait longer. Because of this, small and medium types of business are very vulnerable [10][11]. The first objective of this proposed research is to create a better light weight data analysis model. For this we need to develop a data deduplication algorithm. The Deduplication algorithm not only saves storage space, but also reduces the time consume for data processing.

- *Jayaraj T, Research Scholar *, Research and development Centre, Bharathiar University, Coimbatore, India. yoursjayan@gmail.com.*
- *Dr. J. Abdul Samath, Assistant Professor, Chikkana Government Arts and Science College, Tiruppur, India, bdul_samath@yahoo.com*

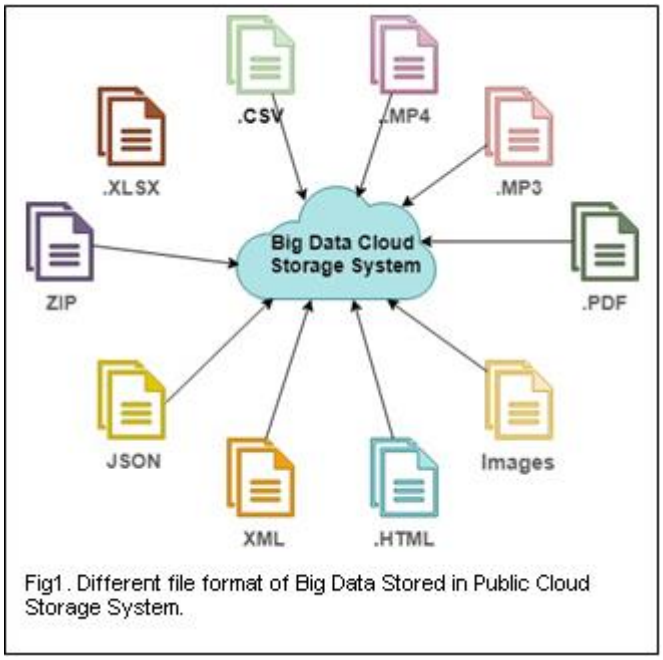

Fig1. Different file format of Big Data Stored in Public Cloud Storage System.

Heterogeneity is one of the most important features of big data. Heterogeneous data contains files with multiple formats. As shown in Fig. 1, how heterogeneous big data is stored in a public cloud. This paper organized as follows section 2 literature survey is summarized. In section 3 HDIA architecture and big data, the analysis module is explained. Section 4 the analysis of the experimental result. Finally, this research is concluded.

2 LITERATURE REVIEW

Zheng Yan et al proposed Deduplication on Encrypted Big Data in Cloud based on based on ownership challenge and proxy re-encryption [2]. This method integrates cloud data deduplication with access control. According to this method only authorized data holders can obtain symmetric keys Used for data encryption. This method contains the following main parts: Encrypted Data Upload, Data Deduplication. Data Deletion, Data Owner Management and Encrypted Data Update. They also proposed a cryptoGPS identification

scheme based on ownership verification protocol. JINBO XIONG et al proposed secure role re-encryption system [3]. This method is based on convergent encryption and the role re-encryption algorithm. Main job of this method is to prevent the privacy data leakage in cloud and this method also supports ownership checking. Three modules of this research is: authorized deduplication, proof of ownership, and role key update. Moreover, they introduce role authorized tree to manage the user's roles and the corresponding role keys. In order to reduce computation cost and management overhead management center is used. Mi Wen et al proposed session-key-based convergent key management scheme, Session-Key-Based Convergent(SKC) and Convergent Key Sharing Scheme(CKS), for cross-user data deduplication in cloud service providers [4]. In SKC, each have the capability to verify the correctness of the convergent key. Besides, encrypted data blocks and session key can be dynamically refreshed when data block changes or other data owner joins in CPSS. Jan Stanek et al developed a Thresholded Data Deduplication Scheme for Cloud Storage. They modify the Stanek et al. [1] to improve its efficiency and emphasize clear functionality. [5]Prevents deduplication of unpopular data and allows their automatic transition to a popular state. This method moves the handling of sensitive decryption shares and popularity state information out of the cloud storage, allowing for simpler security proofs, easier adoption and improved security notion.

3 PROPOSED METHODOLOGY

The proposed system consists of four key components: Format Wise File Separation, File Chunking, Hash Value Finding and Grouping hash values. The architecture of the proposed system is shown in fig2. In this section, we first describe the architecture of the proposed system. Next, we have explained the process flow and algorithm of the proposed method. The pseudo-code is explained in algorithm 1.

3.1 Format Wise File Separation

This is one of the most important components of our proposed method. Heterogeneous Big Data is a large volume of data with a variety of file extensions. We refer to this as Fs = {f1,f2….fn}. In this case, f1, f1, .. fn are all different types of files. ei .txt, .doc, .pdf, .CSV, .HTML etc. We have categorized the files using the K-Means clustering method. By categorizing this, the Heterogeneous data deduplication system can greatly improve the efficiency and the speed of the deduplication process.

3.2 File Chunking and Hash Value Finding

File chunking is the next important part of our proposed system. For this, we first get the file extension. We get file extension through getFileExtension java method. Next, the file size is set 32MB file fixed size. By SplitFile Java function, the input file is cut into pieces of 32MB size. The MD5 algorithm is used to find the hash value in the proposed method. It generates the same hash value for the input value of the same type. This is its special feature. This way we can easily find the data deduplicate file.

3.3 Grouping the Hash Values

The hash value found is stored in the MySql table. A unique identification number is generated to identify each file. Next, grouping hash values through the GroupBy function. If a group's count is more than one, then the first file is stored and everything else is deleted. Finally, this table is updated. Its structure is given in Table 1.

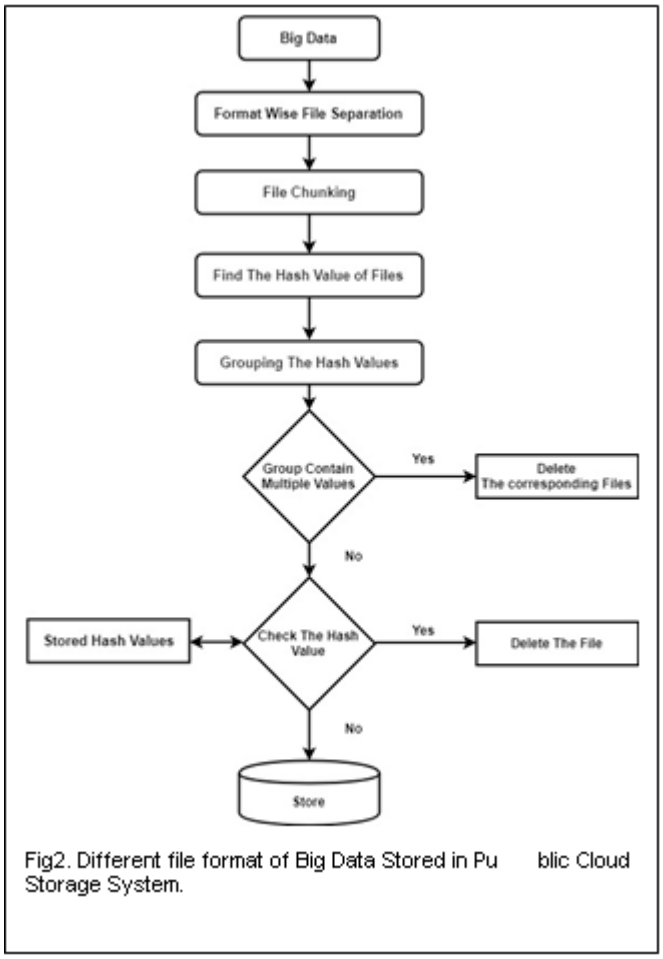

Fig2. Different file format of Big Data Stored in Pu blic Cloud Storage System.

Table 1 *MySql Structure for Storing Hash Values*

File_Id	*Hash Value*	*Unique Name*
1	*f9f07adac05637bc4da1c7ff37e1f052*	*F1*
2	*7ac66c0f148de9519b8bd264312c4d64*	*F2*
3	*e10adc3949ba59abbe56e057f20f883e*	*F3*
4	*403227d744bb5bdae59ea9e5841fcb94*	*F4*
5	*e10adc3949ba59abbe56e057f20f883e*	*F5*
6	*babe6dbe5e71cbd26c2aa48b0a2705a3*	*F6*
7	*f9f07adac05637bc4da1c7ff37e1f052*	*F7*
8	*403227d744bb5bdae59ea9e5841fcb94*	*F8*
9	*66f7ad95f4f39282e1ab8d640a76cc2b*	*F9*
10	*f9f07adac05637bc4da1c7ff37e1f052*	*F10*
11	*babe6dbe5e71cbd26c2aa48b0a2705a3*	*F11*

Algorithm1 Data Deduplication

```
Input: Big Data
Output: Store Unique Data File
Function Data_Deduplication(InputData)
{
File Format Separation Fs={f1, f2, f3, ……..fn}
Fc=FileChunking(Fs)
Hash_Value= GetHashValue(Fc)
G=Grouping(Hash_Value)
Count=G.Count
If (Count>1)
   Keep the single file and delete all other files.
```

```
    Update G
  Else
      If(set.contains(G))
          Delete the corresponding files.
      Else
          Store the file into cloud storage
  }
  Function FileChunking(fileToSplit)
  {
      String chunkNme="chunk"+ Dynamic Value
      String ChunkExtension=getFileExtension(fileToSplit)
      Int maxChunkSize= 32000 //kb
      Split=splitFile(fileToSplit,chunkNme, ChunkExtension,
maxChunkSize)
Return splitFile
  }

      Function GetHashValue(InputFile)
  {
MessageDigest MD_Value =
MessageDigest.getInstance("MD5");
byte[] Digest_Val = MD_Value.digest(InputFile.getBytes());
BigInteger Num_Val = new BigInteger(1, Digest_Val);
String Hash_Val = Num_Val.toString(16);
while (Hash_Val.length() < 32) {
      Hash_Val = "0" + Hash_Val;
return Hash_Val;
      }
```

4 EXPERIMENTAL ANALYSIS

The prototype of this proposed method is developed by JDK 1.8. We have written approximately 7600 lines of codes to implement it. The advantages of this method is evaluated by state-of-the-art algorithms. Which includes: [2] J. Stanek, A. et al, [3] Zheng Yan et al, [4] JINBO XIONG et al, [5] Jan Stanek et al. 4GB of Data set has been downloaded from DATA.GOV to implement the proposed data deduplication method. The eclipse IDE is used for the development environment. The hadoop opensource tool has also been used to implicate the proposed system. The proposed work is run by a computer with Intel Core i7-4310u @ 1.9 GHZ processor, 8GB of RAM, 4TB Hard Disc and Get Force RTX 2080 GPU. This computer is connected to 10MB data transfer speed network connectivity.

Table2 *storage efficiency comparison with state of art methods.*

Input Data	*Proposed Method*	*[1] J. Stanek, A. et al*	*[2] Zheng Yan et al*	*[3] JINBO XIONG et al*	*[5] Jan Stanek et al*
1GB	*0.76GB*	*0.81 GB*	*0.83 GB*	*0.87 GB*	*0.89 GB*
2GB	*1.63 GB*	*1.76 GB*	*1.82 GB*	*1.83 GB*	*1.86 GB*
3GB	*2.23 GB*	*2.4 GB*	*2.61 GB*	*2.71 GB*	*2.86 GB*
4GB	*2.98 GB*	*3.2 GB*	*3.41 GB*	*3.52 GB*	*3.65 GB*

Storage efficiency is shown in figure 3 and summarized in table2. This ensures that the proposed method provides better storage efficiency. [5] Jan Stanek et al and [3] JINBO XIONG et al fail to achieve the highest storage efficiency. The proposed method saves 1.2 GB of storage space when the maximum input data is 4GB.

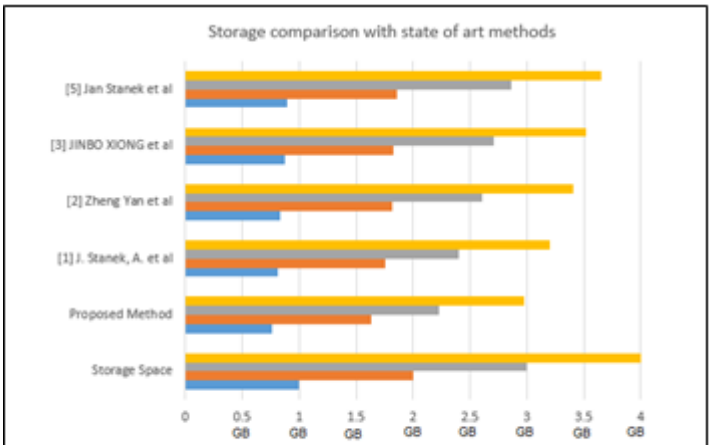

Fig3. Storage efficiency comparisonchart with state of art methods

Here we use the monthly storage cost of four reputed cloud service providers to calculate the cost efficiency of the proposed method: Backblaze, Amazon Web Services, Microsoft Azure and Google Cloud. The calculated cost is summarized in Table 3. Fig. 4 shows the Capricorn chart. Fig. 4 illustrates that the proposed method is a cost effective method.

Table3 *cloud monthly storage cost comparison with state of art algorithms.*

CSP	*Proposed Method*	*[1] J. Stanek, A. et al*	*[2] Zheng Yan et al*	*[3] JINBO XIONG et al*	*[5] Jan Stanek et al*
Backblaze	*23*	*31*	*35*	*41*	*47*
Amazon web services	*28*	*36*	*37*	*47*	*49*
Microsoft azure	*34*	*43*	*41*	*49*	*53*
Google Cloud	*39*	*49*	*46*	*54*	*57*

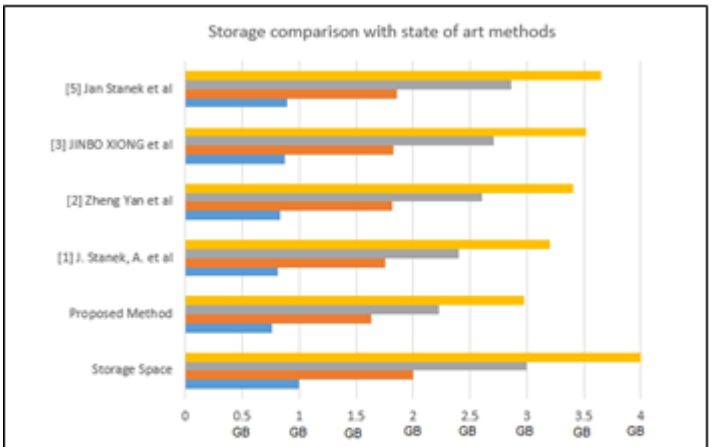

Fig4. Cloud monthly storage cost comparison chart with state of art algorithms.

5 CONCLUSION

Data deduplication contributes significantly to the storage and management of big data. In this paper, an efficient data deduplication method is proposed. The proposed method is more flexible with a Big Data storage system that stores heterogeneous data. This saves storage space multiple times and save the cost of CSP. Experimental results make it clear that its efficiency is excellent.

6 REFERENCES

[1] J. Stanek, A. et al, "A secure data deduplication scheme for cloud storage," in Financial Cryptography and Data Security - 18th International Conference, Christ Church, Barbados, March 3-7, 2014,.

[2] Zheng Yan et al," Deduplication on Encrypted Big Data in Cloud", IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 2, APRIL-JUNE 2016.

[3] JINBO XIONG et al," Secure Encrypted Data With AuthorizedDeduplication in Cloud", Received May 20, 2019, accepted May 31, 2019, date of publication June 5, 2019, date of current version June 21, 2019.

[4] Mi Wen et al, Secure Data Deduplication With Reliable Key Management for Dynamic Updates in CPSS, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS.

[5] Jan Stanek et al, "Enhanced Secure Thresholded Data Deduplication Scheme for Cloud Storage", IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING.

[6] Seref Sagiroglu et al, "Big data: A review", 2013 International Conference on Collaboration Technologies and Systems (CTS).

[7] Dimpal Tomar et al, "Integration of Cloud Computing and Big Data Technology for Smart Generation", 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

[8] Neha Sharma et al, "Big data analytics: Impacting business in big way", 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI).

[9] Nader Mohamed et al, "Real-time big data analytics: Applications and challenges", 2014 International Conference on High Performance Computing & Simulation (HPCS).

[10] Amit Kr. Gupta et al, "Challenges and Issues in Data Analytics", 2018 8th International Conference on Communication Systems and Network Technologies (CSNT).

[11] Anita Gupta, "Challenges of Cloud Computing & Big Data Analytics", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).