iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** III **Month of publication:** March 2018

DOI: http://doi.org/10.22214/ijraset.2018.3704

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Prevention of Data De-Duplication

M. K. Patil[1], U. R. Pawar[2], D. D. Pahade[3]
[1, 2, 3,] *Department of Computer Engineering, MET's Institute of Engineering* Associate *Prof. P. M. YAWALKAR, Dept. of Computer Engineering, MET BKC IOE, NASHIK*

Abstract: ***Data De-duplication is an effective technique to discard duplicate repeating data. Data De-duplication technique is mostly used in cloud computing to decrease the volume of storage space and save bandwidth but not applied on a mobile application for data sharing feature. Current mobile application downloads data i.e. text File, images without checking data duplication, which turn into more data consumption more memory requirement. The solution for this problem is to perform block level de-duplication checking at server-side and client side. Propose system solve this problem of data duplication by performing proactive data checking in mobile device cloud server. In this approach, the shared data is divided into a number of different blocks based on contents and block hash value is calculated for data de-duplication checking.***
Keywords: Project Data De-duplication System, Cloud based system, Network based android application.

I. INTRODUCTION

The world is producing the large number of digital data that is growing rapidly. According to a study, the information producing per year to the digital universe will increase more than 180 Exabyte to 988 Exabyte between 2008 and 2012, growing by 57 of information is imparting a considerable load on storage systems. The terror attacks of the 9/11 and the information lost of various organizations in those attacks Confirmed that information lost is tremendous to modern organization. So it is important to secure the information regularly to a disaster recovery site for data availability and integrity. This section describes the term Data De-duplication and helps in saving storage space and bandwidth. In Cloud computing data de-duplication is a unique data compression method for deleting duplicate same copies of repeating data. Related and somewhat similar terms are important (data) compression and single-occurrence (data) storage.

A. Data De-Duplication App

The goal of data de-duplication is to maximize data storage efficiency by reducing duplicate data from storage on system. We can maximize the volume of data stored at a given cost by eliminating the volume of redundant information in the file system.

II. RELATED WORK

The idea of data de-duplication with protected manner is the main objective of the system, B. Aprana proposed secure data de-duplication technique by differentiating sensitive data and non-sensitive data while uploading into cloud system and apply the cryptographic algorithm for sensitive data by using this data get protected and authenticated.[1] The System proposed by Deepika Singh, Preetika Singh to present various risks induced in cloud storage services due to the use of de-duplication. Taking into consideration the amount of network bandwidth and disk space saved by de-duplication, various methods have been tackled against these risks. In the proposed system solution that not only removes the risk of all the three attacks described but also it helps to establish a trust between the cloud service provider and the user. [2] The System proposed by Nehav kaurav, thoroughly gives idea about various Data De-Duplication methods widely used in storage servers worldwide. This study is also useful for a new researcher who wants to work in field of Data de-duplication and this study can be a start guide for it. We are focusing on the new load balanced algorithms which are scalable. [3]

III. MATHEMATICAL MODEL

1) System defined as
2) S={I,F,T,S,H,D}
3) I=set of Input image
4) $I=\{i_1,i_2,i_3\}$
5) F=setoff feature vector created from Image
6) $F=\{f_1,f_2,f_3\}$
7) T=set of Tag's generated from Image feature vector
8) $T=\{t_1,t_2,t_3\}$

©IJRASET (UGC Approved Journal): All Rights are Reserved

9) S=S is cloud storage services
10) H= Set of Hash value from Images Tag's
11) H= $\{h_1,h_2,h_3\}$
12) D= D is Shared Data on Cloud
13) Function f1
14) F1(I) →$\{f_1,f_2,f_3\}$ → F
15) Function f2
16) F2(F) → $\{h_1,h_2,h_3\}$ → H
17) Function f3
18) F3(H) → $\{t_1,t_2,t_3\}$→ T
19) Function f4
20) F4(T,S) → Upload Data on Cloud Server F5 (T,D) → Check Data Duplication

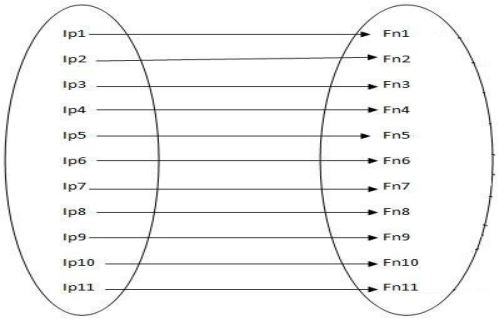


Figure-1: venn diagram

A. File Upload steps
1) Request for file upload
2) Checks for duplication.
3) Verify whether the file already exists or not.
4) If file already exist compare its hash value with Receiver value and download existing file by receiver.
5) . If the file does not exist then send new required file.
6) System divide file into number of blocks (shares).
7) File (shares) store onto the cloud.

B. File Download steps
1) Request for file download.
2) Cloud sends the hash value.
3) Verify if file exists or not.
4) If file already exist then send existing data link file.
5) If file does not exist send request to system for file download.
6) File downloads.

IV. METHODOLOGY

The software development model that was used is the iterative model. This model was used because the iterative model allows for requirement changing.
The System Architecture for the Android based data de-duplication system illustrating all components and sequence of the system is shown below.

©IJRASET (UGC Approved Journal): All Rights are Reserved

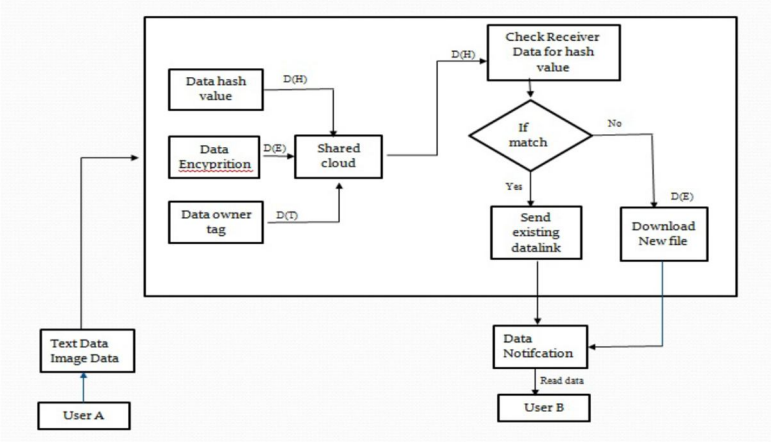


Figure-2: System Architecture

In the current scenario, there are the various operations performed. Operations such as data encryption, data hash value generation, data owner tag. Data hash value is calculated and compared with Cloud server. Comparison is made for forwarded File at cloud server and receiver side. If file is already present then send existing link to receiver. If file is not present save on cloud server and then downloaded by receiver.

User A: Here User A is the sender in the communication (TEXT, IMAGE)

Data Hash Value: Hash value is defined as numeric value of a permanent length that identifies data uniquely. Huge amount of Data is represented by Hash values. Which are smaller numeric values, so they are used with digital signatures. The Function of these Block is to calculate Hash value for the Given Data (TEXT, IMAGE).

Data Encryption: In process of Encryption, Data or information is encoded in such a way that authorized users can only access it and those who are not authorized they cannot get access. In this Block End to End Encryption is used for secure data transfer.

Data owner Tag: Ownership tag schemes allows owner of the data to prove to the cloud storage server that he owns the data in a robust manner. In these block A unique Data owner tag is Generated for every Data (TEXT, IMAGE).

Shared Cloud: Cloud storage is a structure of data storage in which the digital data or information is stored in logical pools, the physical storage contains multiple no. of servers and the real environment is typically owned and handled by a hosting company.

V. RESULTS AND DISCUSSION

The proposed system will generate data hash value, data ownership tag along with cloud computing technique and Duplicate file will be eliminated thus saving storage space and bandwidth.

VI. CONCLUSION AND FUTURE SCOPE

Data de-duplication is the process of eliminating similar data by comparing new data with data already stored and maintaining only one copy. Various vendors are adding capabilities to their products to help maximize its adoption rate and to fight difficulties from alternative solutions like the cloud, tape and even regular disk. As a result, data de-duplication has firmly established itself in the backup market.

VII. ACKNOWLEDGMENT

We would like to thank Prof. Prashant Yawalkar, MET's Institute of Engineering, Nashik for his expert guidance and valuable contribution in project Development.

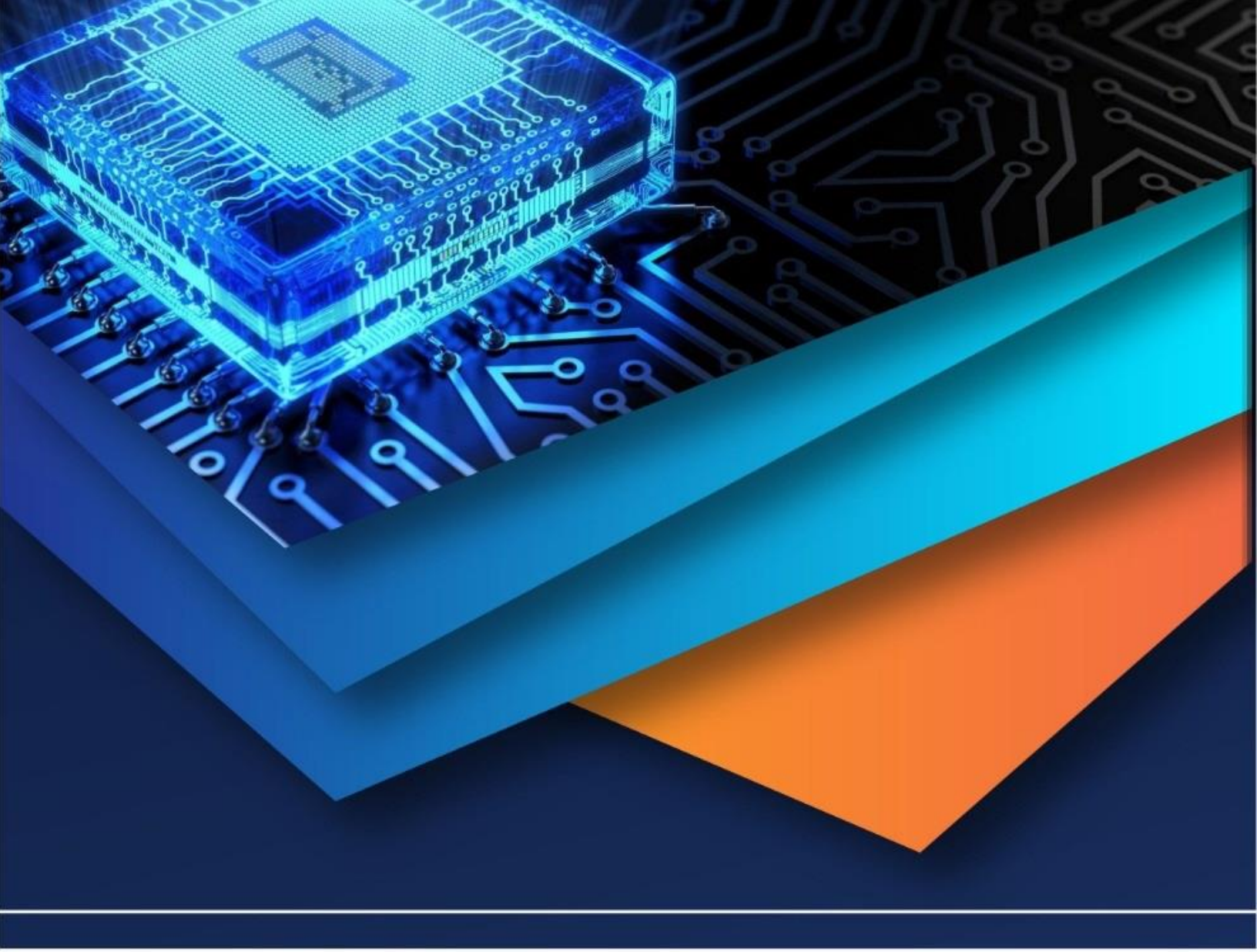
REFERENCES

[1] B. Aparna, "Privacy Preserving and Authorized data de-duplication in Public Cloud Framework", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.5, Issue.10, 2015.

[2] Deepika singh, Preetika singh, "New Challenges for Security against De-duplication", International Journal of Advanced Research in Computer Science and Management studies, Volume 2, Issue 1, January 2014.

[3] Neha kaurav, "An Investigation Data De-duplication Methods And it's Recent Advancment", ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-112.

©IJRASET (UGC Approved Journal): All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)