iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

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

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Improved De-Duplication Methods to Enhance Performance in Cloud Storage

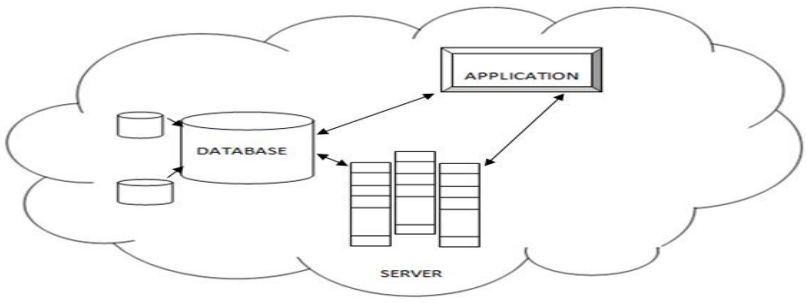
Er. Suman Kumar Mishra[1], Er. Saroj Ranjan2
[1]*Department of Computer Science*
[2]*Sityog Engineering College*

Abstract: ***Cloud computing is the emerging technology that helps in consolidation of resources. Many organizations have their public as well as private clouds. Private clouds can be built from unused resources to store data. But as private cloud has limited ways of storage, so they need to be utilized properly. De-duplication is the means of storing data in effective way over clouds. This paper will discuss about the use of De-duplication method din cloud computing for the storage of the data in effective way. Cloud computing plays a main role in the business domain current as computing processes are delivered as utility on demand to client over the internet. Cloud storage is one of the services offered in cloud computing which has been increasing in famous. The major benefits of using cloud storage their time in buying and maintaining storage environment while only giving for amount of storage requested, which can be scaled-up and done upon demand. With the increasing data size of cloud computing, a reduction in data density could help providers decreasing the costs of running large storage system and convertible energy consumption. De duplication Data is the method which compresses the data by deleting the numerous copies of inner data and it is widely used in cloud storage to save bandwidth and minimum the storage space. To secure the private of sensitive data during De duplication, the hashing method is used to generate the hash for save data before outsourcing.***
Keywords: Cloud computing, De-duplication and Key technique.

I. INTRODUCTION

In the meantime cloud computing is most important thing for our present and future technologies. Cloud computing is most important and famous topic for the researchers it provides a large area for work.

Basically cloud computing provides an in center database for both senders and receivers. Cloud had changed complete scenario for the present technologies and also for the upcoming period. Cloud computing having un-counted benefits. As cloud computing provides multiples of services. The Computing in which the resources like data, storage, various software's are allotted over the network and are managed through the Internet by a service provider is termed as cloud computing. It is also popularly called an Internet based computing because the users interact with the service provides through the Internet and also the customers are given the services via Internet. The advantages and disadvantages related to cloud computing are discussed too. In addition to its benefits and drawbacks we have lot of challenges in the cloud computing.



Cloud is the large pool of resources. In cloud all resources are virtually connected to each other. These resources can be reconfigured according to storage needs. There are mainly five features that cloud has:

A. On-demand self-service

B. Broad network access

C. Resource pooling

D. Measured service

E. Rapid elasticity

Cloud computing contains both hardware and application provided to users. Computing resources are limited and will saturate at some time. This scalability alarms the clouds to structure your data accordingly. It means its giving way to compress the data for further scalability. So here comes cloud storage.

Cloud storage is that model where data can be placed, managed, back up, stored and modified. Cloud storage makes available data to clients in any time, with high storage space [1] and also makes it user friendly so that availability of data increases. There are many companies that help to store data in cloud server online. There is a need of online interface when user wants to store data online. Cloud storage is done mainly to back up the data. Cloud storage is a SLA services i.e. it is Service Level Agreement. Cloud storage is of three types: Public, Private and hybrid.

F. Advantages of Cloud Storage

Availability to access the data from any place/ location increases.

No need to carry physical storage device.

Other trusted people can also allow sharing the data.

Cloud Computing is the novel emerging trends in the novel generation technology. Each client has big amount of data to share to store in a quickly available protected place. The concept of De duplication is reached here to efficiently utilize the bandwidth and circle disc usage on cloud computing. To escape the De duplication copies of the similar data on cloud may cause loss of time, bandwidth consumption and space [1]. Cloud computing is the internet based, a network of remote servers associated over the internet to store, share, change, add and resourcing of data. The benefits of cloud computing: No longer have to pay for someone (or a team of someone's) to do things such as install and update software, install and manage email servers and/or fine servers, run backups – the loveliness of cloud computing is that all of the business of maintaining the service or application is the accountability of the cloud vendor, not yours .No longer have to buy software. Besides the convenience of not consuming to buy software programs and install them on your own servers/computers, using cloud applications instead can be cheaper. One may be able to consolidate your separate application needs into one multi-application cloud computing service. For illustration, Google Apps for Business includes email, a calendar scheduling application, Google Docs for generating documents, presentations and forms and using online file storage and Google Sites for creating websites, all for only $5/month for each person on your account. Now think about the price of, let's say, Microsoft Office including Microsoft Outlook for email Able to cut back on system hardware. File storage, data backup and software packages all take up a lot of space on servers/computers. With cloud computing, you use someone else's servers to store all this data instead, liberation up your in-house [2] computer equipment for other purposes or even letting you get rid of some of it. A cloud computing application may make integration easier. Because many cloud computing applications contain an Application Programming Interface (API) you may be able to find "compatible" applications rather than having to pay to have the submissions you want to be integrated customized for you. Cloud computing applications are habitually updated, so you don't have to spend time and money doing it – and giving you the advantage of always having access to an application's latest features and purposes. Cloud computing allows you and your employee's easy access to applications and data from different computers and devices. "As more consumers and businesses adopt tools such as smart phones and tablets, the ability to cloud data in the cloud and access it from just about anywhere on the planet is quickly becoming vital. Cloud computing lets you start up or grow your small business quickly. It's a lot easier and faster to sign up for a [3] cloud computing application than to buy a server, get it up and running and install software on it. And since you don't need to buy hardware and software, your start up or expansion is cheaper, too.

G. Related Work

Vasilios et.al.2013 [6] presents a migration support network, in which fundamental elements are cost effective system. They proposed a three level framework that satisfies al the necessity in view of cost assumption. They utilized the windows azure policy as a part of creating prototyping model. Besides, the ability to consolidate necessities for numerous administration sorts, e.g., information stockpiling and systems administration, is imagined to be given, encouraging the choice making in relocation sorts past the off-stacking of the application stack on a VM.Haitao et.al. 2011 [7]proposed relocation methods taking into account (dynamic,

receptive and shrewd procedures), albeit basically in light of the present data, can make the mixture cloud-helped VoD organization set aside to 30% transmission capacity cost contrasted and the Clients/Server mode. They can likewise handle unpredicted the glimmer group activity with little cost. It likewise demonstrates that the cloud cost and server transmission capacity picked assume the most essential parts in sparing expense, while the distributed storage size and cloud substance upgrade system assume the key parts in the client experience change .C. Ward et.al. 2010[8]Acquainted the augmentations with a coordinated mechanization capacity called the Darwin structure that empowers workload movement for this situation and talk about the effect that computerized relocation has on the expense and dangers ordinarily connected with relocation to cloud.Kang et.al.2013[9] Proposed the migration algorithm .The VM to its best PM specifically, with the proviso that it has adequate capacity. Then, if the migration constraint is gratified, we transfer another VM from this PM to oblige the new VM. In addition, we study a hybrid scheme where a batch is working to accept upcoming VMs for the on-line development. Evaluation results prove the high efficiency of our algorithms. Xian Xinet.al. 2013[10] proposed a dynamic prototype system termed Cyber Live App to support application sharing and migration on demand among various users. Cyber Live App gives two key administrations: a safe multi-client sharing administration for the virtual desktop of a VM and multi-VM application sharing and movement

II. LITERATURE SURVEY

Yinjin Fu, et.al [2] introduced a technique for de duplication that will optimize the performance of look up. This technique has been used for personal environment and to reduce overhead. Earlier methods are focusing only on removing redundancy. But this method focused on faster time retrieval.

AvaniWildani et. al [3] demonstrated the effectiveness of our approach using a simple neighborhood grouping that requires only timestamp and block number, making it suitable for a wide range of storage systems without the need to modify host file systems.

Don fang Zhao, et.al [4] proposed system that is based on Hy Cache. It provides the transparency to exchange of data, to modification of data. This caching advance shows 29X speedup over the conventional LRU algorithm. De -duplication on primary storage system.

Puzio et.al [5] proposes Clouded up, a secure and well-organized storage service which assures block-level de-duplication and data privacy at the same time. Although based on convergent encryption, Clouded up remains secure thanks to the definition of a component that execute an additional encryption operation and an access control mechanism. Furthermore, as the requirement for de-duplication at block-level raises an issue with respect to key management, we propose including a new component in order to execute the key management for each block together with the actual de-duplication operation. We show that the in the clouds introduced by these new components is minimal and does not impact the overall storage space and computational costs.

Dirk Meister, et.al [6] proposed a method in which earlier backup in sequence is used to predict the future backup. These methods enhance the lookup performance. It is better than BLC approach.

Andrew J. Younger et al"Efficient Resource Management for Cloud Computing Environments,". [7] described a frame work for efficient green enhancement in cloud architecture. It is based on power aware scheduling, variable management and minimal virtual machine design. It has improved overall system efficiency. It is used to evaluate the performance and overall capacity of virtual machine by using power based scheduling of virtual machine.

Dipti Bhansali et al. "An Optimistic Differentiated Job Scheduling System for Cloud Computing," [31] explained the mechanism of an Optimistic Differentiated Job Scheduling System. This algorithm is developed to serve the multiple requests. This method is proposed to handle multiple requests of services like uploading and downloading. Multiple requests are processed by use of non-primitive algorithm. Its main goal is to provide optimistic value of service. The users get the quality of service and service providers gain maximum profit. It exploits the under-utilized resources at non-peak times. Utilization of resources is done in transient way. This paper implements static load balancing based on size of files.

Isam Azawi Mohialdeen "Comparative Study of Scheduling in Cloud Computing Environment," [24] surveyed about the scheduling algorithm used in cloud computing. Scheduling is an important aspect to schedule the jobs on virtual machines. In cloud, single scheduling algorithm is not sufficient because single algorithm does not consider all performance metrics and Maintain quality of service. Many scheduling algorithm have been proposed to enhance the systems performance in terms of throughput and cost. This paper compares 4 types of scheduling algorithms under cloud namely Round Robin (RR), Minimum Completion Time, Random Resource Selection and Opportunistic Load Balancing. These algorithms have been evaluated in terms of their ability to provide quality of service and to maintain fairness amongst all jobs. Each scheduling algorithm have performed superior on some metrics. All measuring characteristics are not provided by each and every algorithm. The selection of good scheduling is based on

characteristics that fulfil the needs of customers as well as service provider.

A. Qura-Tul-Ain Khan et al. , "Usage & Issues of Cloud Computing Techniques in Small & medium Business Organizations," [46] explained about a computing platform that exists in large data centre is cloud computing. Cloud computing is dynamically able to provide servers the ability to address wide range of needs in almost every field. Many problems are involved to deliver cloud computing resources if they were utilities like electricity, privacy issues, security, and access, regulations, reliability and other issues.

III. APPROACHES TO DEDUPLICATION

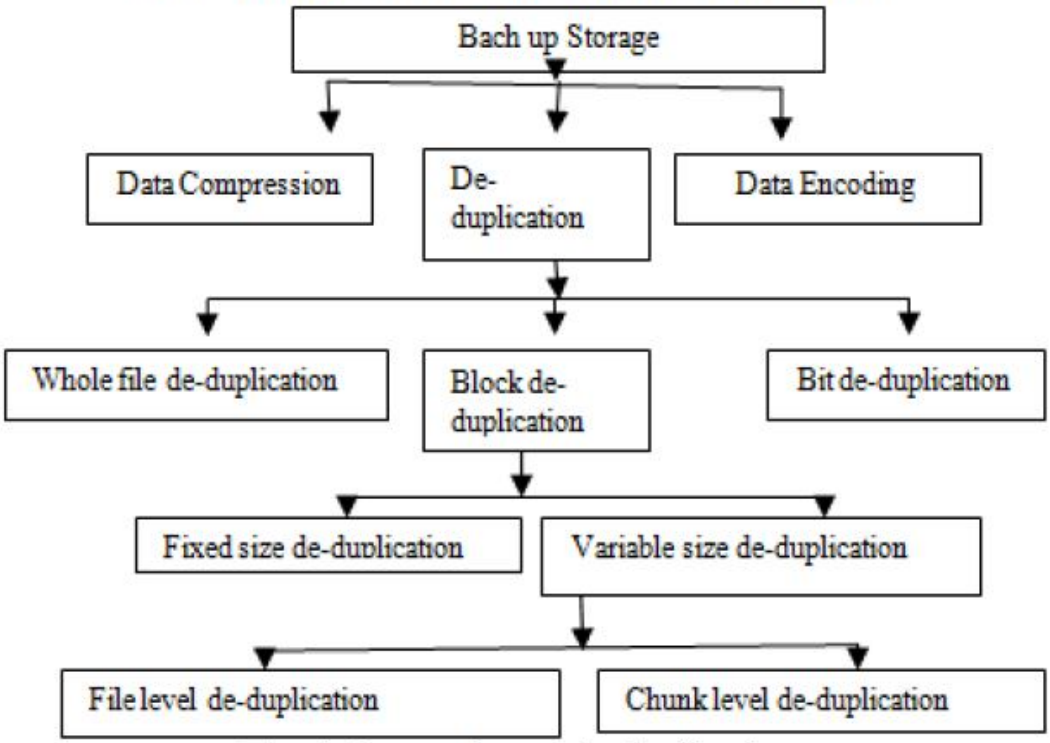


Fig. 1 Approaches to de-duplication

Data de replication describes a class of approaches that diminish the storage capacity needed to store data or the amount of data that has to be transfer over a network. These approaches detect coarse-grained redundancies within a data set, e.g. a file system; Data de duplication not only reduces the storage space requirements by eliminate redundant data but also minimizes the network transmission of duplicate data in the network storage systems. It splits files into numerous chunks that are each uniquely identified by a hash signature called a fingerprint. It removes duplicate chunks by checking their fingerprints, which avoids byte by byte comparisons. Mainly data de duplication listening carefully on different terms like throughput, advance chunking schemes, other type of storage capacity and cluster method and system workload.

III. PURPOSE ALGORITHMS (SHA SECURE HASH ALGORITHM)

A. Various types of SHA

1) SHA 0
2) SHA 1
3) SHA 256
4) SHA 512

The Secure Hash Algorithm is one of a number of cryptographic hash functions. There are currently three generations of Secure Hash Algorithm:

5) SHA-1 is the original 160-bit hash function. The similar to the earlier MD5 algorithm.
6) SHA-2 is a relation of two similar hash functions, with dissimilar block sizes, known as SHA-256 and SHA-512. They differ in

the word size; SHA-256 uses 32-bit words where SHA-512 uses 64-bit words.

7) sHA-3 is a future hash function standard still in development [11].

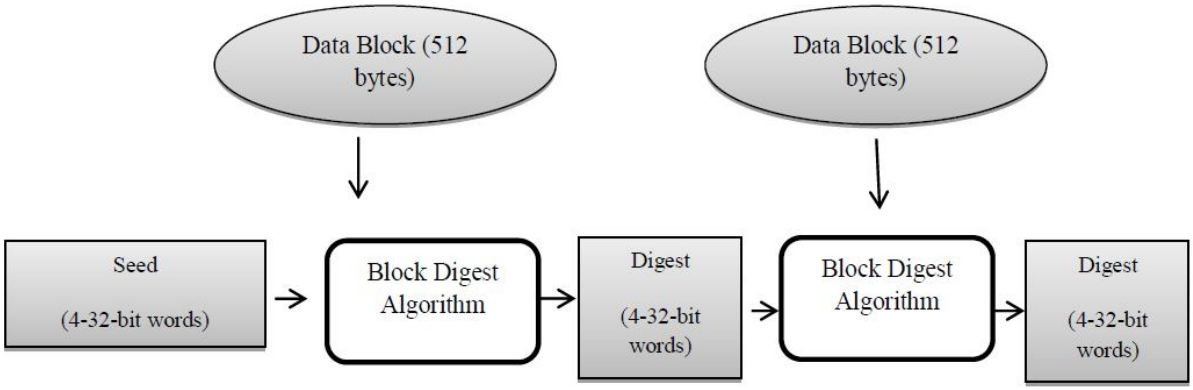
SHA-1 Algorithm: 1-The SHA algorithm uses 5 state variables, each of which is a 32 bit integer (an unsigned long on most systems). These variables are sliced and dice and are (eventually) the message digest. The variables are initialized as follows:

h0 = 0x67452301
h1 = 0xEFCDAB89
h2 = 0x98BADCFE
h3 = 0x10325476
h4 = 0xC3D2E1F0

There are 80 rounds in SHA Algorithm. The hash value created by the SHA hash function.

B. MD-5 Algorithm

MD5 is an algorithm that is used to confirm data integrity through the creation of a 128-bit message summary from data input (which may be a message of any length) that is demanded to be as unique to that specific data as a fingerprint is to the specific separate. MD5, which was developed by Professor Ronald L. Rivest of MIT, is intended for use with digital signature requests, which require that large files must be trampled by a secure method before being encrypted with a secret key, under a public key cryptosystem. MD5 [12] is currently a normal, Internet Engineering Task Force Request for Comments 1321. Allowing to the standard, it is "computationally infeasible" that any two messages that have been input to the MD5 algorithm could have as the output the same communication digest, or that a false message could be shaped through uneasiness of the message digest. MD5 is the third message digest algorithm created by Rivest[12]. All three (the others are MD2 and MD4) have similar constructions, but MD2 was optimized for 8-bit machines, in assessment with the two later formulas, which are optimized for 32-bit machines. The MD5 algorithm is an allowance of MD4, which the critical review found to be fast, but possibly not unconditionally secure. In comparison, MD5 is not quite as fast as the MD4 algorithm, but offers much more declaration of data security.



IV. CONCLUSION AND FUTURE SCOPE

In this paper, I have discussed about storage issues in the cloud computing and how de-duplication method solves the difficulty of storage at cloud. Cloud is a costly storage provider, so the motivation is to use its storage area efficiently. De-duplication has proved to reduce memory consumption by removing the useless duplicate files. So far from the previous studies file level de-duplication is the better method to be used, the focus of the proposed work will be on file level de-duplication. In this work, a dynamic data De duplication scheme for cloud One way of storage space is using hash functions but they are to effective approach so there is an enhancement needed in which use of external hard disk is required.

REFERENCES

[1] Rashid, Fatema, Ali Miri, and Isaac Woungang. "A secure data de-duplication framework for cloud environments." Privacy, Security and Trust (PST), 2012 Tenth Annual International Conference on. IEEE, 2012.
[2] Yinjin Fu, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu,''A De-duplication: An Application-Aware Source De-duplication Approach for Cloud Backup Service "IEEE International Conference on Cluster Computing in the Personal Computing Environment (2011)
[3] A. Wildani, E. L. Miller, and O.Rodeh. HANDS: A heuristically arranged non-backup in-line de-duplication system. Technical Report UCSCSSRC- 12-03,

University of California, Santa Cruz, March 2012.

[4] Dirk Meister, Jorgen Kaiser," Block Locality Caching for Data De-duplication". In Proceedings of the 11th USENIX Conference on File and Storage Technologies (FAST). USENIX, February 2013

[5] Dongfang Zhao, KanQiao, IoanRaic, y, "HyCache: Towards Scalable High-Performance Caching Middleware for Parallel File Systems", Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357, 2014. [6] Pasquale Puzio, Melek O¨ nen, Sergio Flourier, "Clouded up: Secure De-duplication with Encrypted Data for Cloud Storage", IEEE, 201

[6] D. Bhansali, S. Ambike, J. Kshirsagar and J. Bansiwal, "An Optimistic Differentiated Job Scheduling System for Cloud Computing," International Journal of Engineering Research and Applications, vol. 2, no. 2, pp. 1212-1214, 2012.

[7] Q. Khan, S. Nasser, F. Ahmad and M. Khan, "Usage & Issues of Cloud Computing Techniques in Small & medium Business Organizations," International Journal of Scientific & Engineering Research, vol. 3, May 2012.

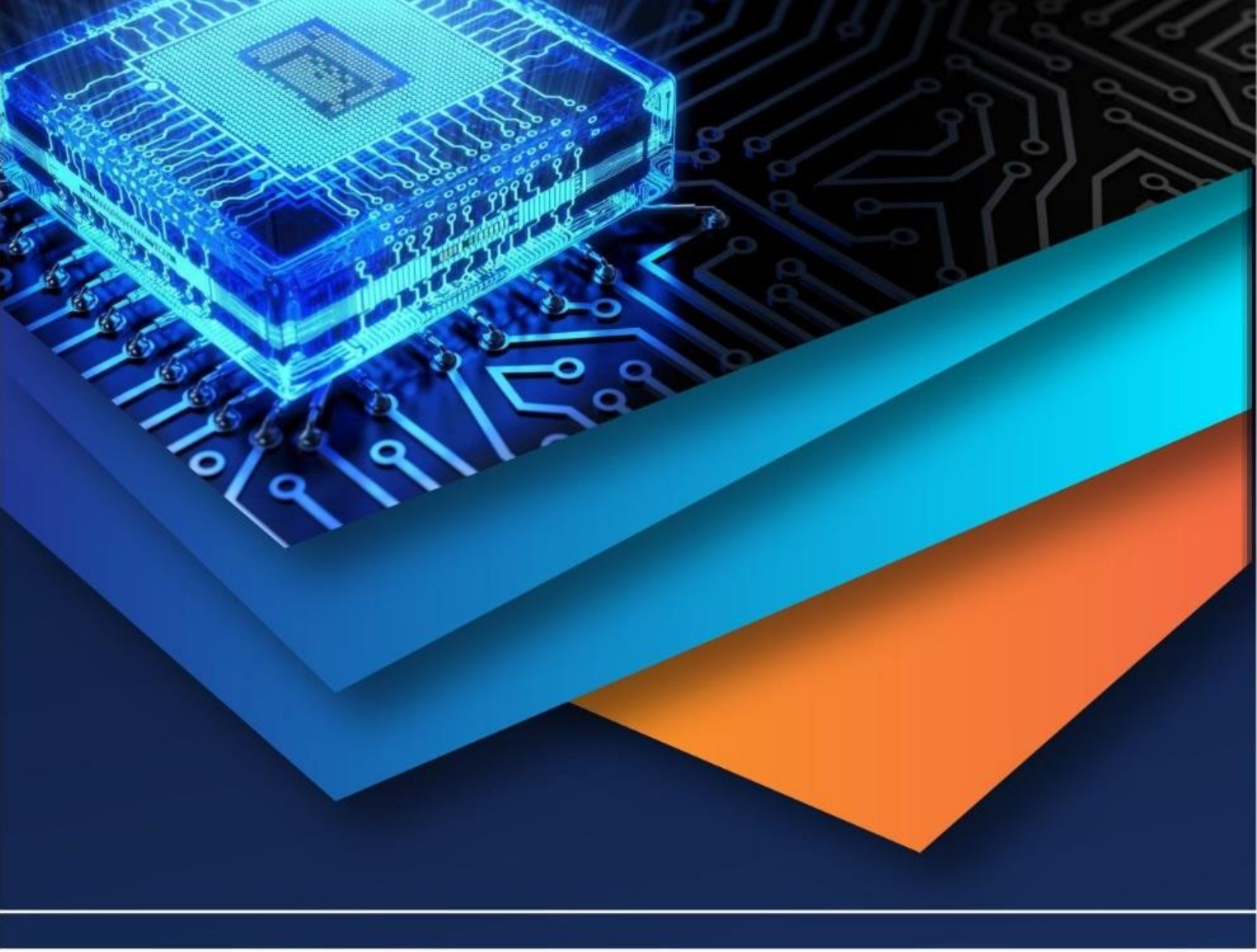
[8] C. Ward, N. Aravamudan, K. Bhattacharya, K. Cheng, R. Filepp, R. Kearney, B. Peterson, L. Shwartz, C. C. Young, "Workload Migration into Clouds – Challenges, Experiences, Opportunities", 2010 IEEE 3rd International Conference on Cloud Computing, pp. 164-171, 2010

[9] Kangkang Li, HuanyangZheng, and JieWu ."Migration-based Virtual Machine Placement in Cloud Systems", 2013 IEEE 2nd International Conference on Cloud Networking (CloudNet, IEEE, pp. 83-90, 2013.

[10] Jianxin Li, Yu Jia a, Lu Liub, TianyuWoa, " CyberLiveApp: A secure sharing and migration approach for live virtual desktop applications in a cloud environment, Elsevier, Vol. 29, pp.334-340, 2013.

[11] Jung, Ho Min, et al. "Efficient data deduplication system considering file modification pattern." International Journal of Security and Its Applications6.2 (2012): 421-426.

[12] Zhang, Yang, Yongwei Wu, and Guangwen Yang. "Droplet: a distributed solution of data deduplication." Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing. IEEE Computer Society, 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY