© 2016 IJSRSET | Volume 2 | Issue 4 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

# Efficient Data Distribution Technique for Hadoop in Cloud Environment of Cluster Node

**Prof. Shashvat Sanadhya, Shashank Mishra**
Shri Ram Institute of Technology, Jabalpur, Madhya Pradesh, India

## ABSTRACT

There has been a quick progress in cloud, with the growing amounts of associations turning number of associations relying upon use resources in the cloud; there is a requirement for securing the data of various customers using concentrated resource. Circulated capacity organizations avoid the cost stockpiling organizations dodges the cost exorbitant on programming, staff keeps up and gives better execution less limit cost and flexibility, cloud advantages through web which construct their presentation to limit security vulnerabilities however security is one of the critical weaknesses that balancing incomprehensible relationship to go into appropriated processing environment. The Proposed wear down HADOOP stockpiling strategies, Map reduces approach with synchronization between tasks and this purpose of interest and its impediments.

**Keywords:** Cloud Computing, Data Storage, Data Security, Map Reduce HADOOP.

## I. INTRODUCTION

Cloud computing is a delivered computer services over the network. Cloud computer kind of computing where by resources and it related capabilities are provided as services to the outer customer using Internet technique. Cloud computer is an environment for providing information resources that are delivered as services to the end user over the internet on demand cloud is defined with file essential characteristics [on-demand self-services, Broad network access, resource pooling, rapid elasticity, measured services].SPI service models [Saas, Paas, Iaas]deployment models [Public, Private, Hybrid, Community]. Reason for moving into cloud is simply because of cloud allows user to access application from anywhere at any time to internet. Cloud provides benefits such as, flexibility, distracter, recovery, software updates automatically, pay- per-use model and cost reduction. Cloud also includes the major risk such as security, data integrity, network dependency and centralization. The pioneer of cloud computing vendors, amzon simple storage service (AS3) and Amazon electric compute cloud are both well-known examples. While this internet based online service provide huge amount of storage space and customizable computing resources, this responsibility of local machines for data maintains at same time.

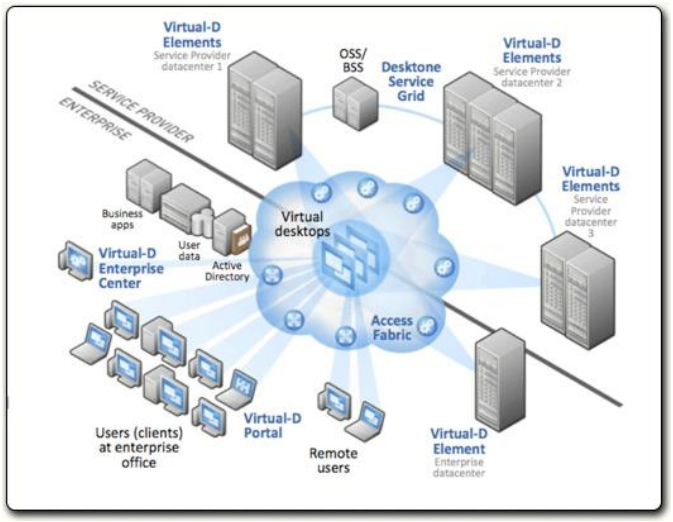


**Figure 1.** Cloud Computing Environment

However security concern has become the biggest obstacles to adaption of cloud becomes all information and data are completely under the control of cloud service provider for such as reason. CSA, ENISA and NIST published general security guidance and recommendations, for cloud usage in order to provide

some level of protections ranging from physical security to network/System/application security.

This paper discusses the methodologies and various techniques to effectively store data. Also, analyzed the advantages, drawbacks of those techniques and provides some direction for features research work.

## II. METHODS AND MATERIAL

### Related Work for Storage Techniques in Cloud Computing

In this section various existing techniques has been discussed. Cloud storage is regarded as a system of disseminated data centres that generally Utilizes virtualization technology and supplies interface for data storage.

### 1. Implicit Storage Security to Data in Online

Providing implicit storage security to data in online is more beneficial in a cloud computing. The use of a data partitioning scheme for implementing such security involving the roots of a polynomial in finite field.

In this scheme data is partitioned in such way that each portion is implicitly secure and does not to be encrypted. These portions are stored on different servers on the network which are known only to the user.

Reconstruction of the data requires access to each server and the knowledge as to which servers the data portions are stored. Several versions of this scheme are described, which include the implicit storage of encryption keys rather than the data and where a subset of the partition may be brought together to recreate the data.

Cloud computing products, also called cloud service delivery models which are often roughly classified into a hierarchy of -as a service terms, presented here in order of increasing specialization:

**Infrastructure-as-a-service (IaaS):** where cloud providers deliver computation resources, storage and network as an internet-based services. This service model is based on the virtualization technology. Amazon EC2 is the most IaaS provider.

**Platform-as-a-service (PaaS):** where cloud providers deliver platforms, tools and other business services that enable customers to develop, deploy, and manage their own applications, without installing any of these platforms or support tools on their local be hosted on top of IaaS model or on top of the cloud infrastructures directly. Google Apps and Microsoft Windows Azure are the most known.

**Software-as-a-service (SaaS)**: applications hosted on the cloud infrastructure as internet based service for end users, applications on the customers' computers be hosted on top of PaaS, IaaS or directly hosted on cloud infrastructure. Salesforce CRM is an example of the provider.

### 2. Identify –Based Authentication

An identify based encryption (IBE) and decryption and identity based signature IBS schemes for IBHMCC. Resources and services are distributed across numerous consumers. So there is a chance of various security risks. Therefore authentication of users as well as services is an important requirement for cloud security. When SSH Authentication protocol (SAP) was employed to cloud, it becomes very complex. As an alternative to SAP, proposed a new authentication protocol based on identity which is based on hierarchical model with corresponding signature and encryption scheme.

Identify based authentication protocol constrains sequence of steps. In step (1) the client C sends the servers a client Hello message. The message contains a fresh random number Cn. session identifier ID and c specification. In step (2) the server S responds with a server Hello message which contains new fresh random number Sn.

### 3. Public Auditing with Complete Data Dynamic Support

Verification of data integrity at unreliable servers is the major concern in cloud storage with public audit ability trusted entity with expertise and capabilities data owners do not possess can be delegated as an external audit party to access the risk of outsourced data when needed. It also provides a transparent yet cost effective method for data owners to gain trust in the cloud. To accomplish, dynamic data support, the existent proof read of PDF (or) POR scheme is improved by spoofing the basic Markel Hash tree (MHT).

**4. Efficient Third Party Auditing (TPA)**

Cloud consumers save data in cloud server so that security as well as data storage correctness is primary concern. The data owners having huge amount of outsourced data and the tak of auditing the data correctness in a cloud environment can be difficult and expensive for data owners. To support third party auditing where user safely delegate in integrity checking tasks to third party auditors(TPA)this scheme can almost guarantee the simultaneous localization of data error(ie the identification of misbehaving servers). A novel and homogeneous structure is introduced to provide security to different cloud types. To achieve data storage security, BLS (Bonch-Lynn-Sachems) algorithm is used to signing the data blocks before outsourcing data into cloud. Reed Solomon technique is used for error correction and to ensure data storage correction.

**5. Way of Dynamically Store Data in Cloud**

Data storage is cloud may not be completely trustable because the clients did not have local copy of data stored in cloud. To address these issues proposed a new protocol system using the data reading protocol algorithm to check the data integrity services providers help the clients to chenk the data security by the proposed effective automatic data reading algorithm. A flexible distributed storage integrity auditing mechanism (FDSIAM), these mechanisms utilizes the homomorphism tokens, blocking erasure and unblocking factors and distributed erasure coded data.

**6. Effective and Secure Storage Protocol**

Current trend is users outsourcing data into service provider who have enough area for storage with lower storage cost. A secure and efficient storage protocol is proposed that guarantees the data storage confidentiality and integrity. This protocol is invented by using the construction of elliptic curve cryptography and sober sequence is used to confirm the data integrity. Data and software process protocol step executed by cloud customers to add the privacy enforcement structure to the software and data before transferring them to the cloud.

Challenge response protocol is protocol is credential so that it will not exposes the contents of the data to outsiders. Data dynamic operations are also used keep the same security assurance and also provide relief to users from the difficult of data leakage and corruptions problems.

**7. Storage Security of Data**

The data is secured in server based on user's choice of security method so that data is given high secure priority resources are being shared across server trouble to data security in cloud. Transmitting data over internet is dangerous due to the intruder attacks data encryption plays an important role in cloud environment. Introduced a consistent and novel structure for providing security to cloud types and implemented a secure cross platform. [14]The proposed effective and flexible distribution scheme two-way handshakes based on token management by utilizing the homomorptic token with distributed verification of erasure coded data, our scheme achieves the integration of storage correctness insurance and data error location that is the identification of misbehaving server.

**8. Secure and Dependable Storage Service**

Storage service of permits consumers to the data in cloud as well as allowed to utilize the available well qualified application with no worry data storage maintance. Although cloud providers benefits, such a service gives up the self control of user's data that introduced fresh valuability hazards to cloud data correctness [8]. The proposed a flexible distributed storage integrity auditing mechanism, utilizing the homomorphism token and distributed [6] coded-data. The proposed design further support secure and efficient dynamic operation on outsource data including block modification, deletion and append.

**9. Optimal cloud storage systems:**

Cloud data storage which requires no effort is acquiring more popularity for individual,enteprise and institutions data backup and synchrozation.[15]the proposed system describe, at a high level, a possible architecture for a cryptographic storage service. At is core, the

architecture consists of their components, a data processor(DP)that processes data before it is sent to the cloud a data verifier(DV)that checks whether the data in the cloud has been tampers with, and a token generator(TG)that generator tokens which enables the cloud storage providers to retrieve segments of consumer data.

### 10. Process of access and store small files with storage

To support services extensively, Hadoop distributed file system server reasons are examined for small file trouble of native Hadoop distributed file system. Burden on Nane Node of HADOOP distributed file system is enforced by large amount of small files, for data placement correction are not considered prefetching mechanism is not also presented. In order to overcome these small size problems, propsed an approach that these small size problem, proposed an approach. That improves the small file efficiency on Hadoop distributed file system [10], in a large cluster, thousands of servers both host directly attached storage and execute user application task. By distributing storage and computation across many servers the resource a grow with demand while remaining economical at every size.

### 11. File storage security management:

To assure the security of stored data in cloud, presented systems which utilize distributed scheme [11].proposed system consists of a master server and a set of slave server. These are not direct commutation link between clients and slave servers in the proposed model. Master server is responsible to process the client's request and at slave server chunking operation in order to provide data backup for file recovery in future. Clients file is stored in the form of tokens on main server and files were chunked on slave server for file recovery.

## III. PROPOSED METHODLOGY AND RESULTS

### Wave Model:

Our previous analysis gives us some hint to propose a computation model to estimate ex-ecution time for a MapReduce job. In order to facilitate our description we define a term wave and this wave model applied on word count process of map reduce mechanism. Definition: A wave is a group of parallel tasks, the length of which is the number of parallel nodes n. There is map wave in map phase and reduce wave in reduce phase. The number of waves is computed by:

$$Nwave = \frac{[Ntask\ ] + 1}{n}$$

Ntask where Ntask represents the number of tasks. The operation [Ntask/n] means obtaining the n integer part of Ntask/n . The number of waves $N_{wave}$ indicates that roughly $N_{wave}$ tasks will be allocated to one node. Figure 2 shows that three tasks form a wave, which are represented by three vertical grids in the figure. Furthermore, in order to facilitate our description, we name the series of tasks executed by an individual node as a task chain. Figure 2 shows that three nodes have three task chain.

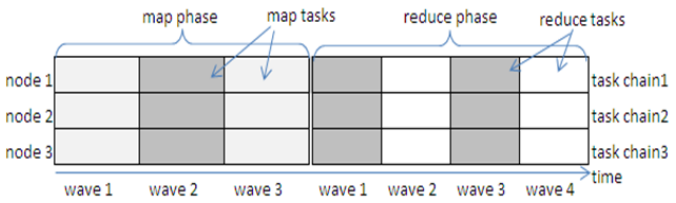


**Figure 2.** The illustration of concept "wave"

For evaluation and testing of the proposed system Ubuntu 14.04 system has been used for as a Hadoop testbed.

This speedup behavior is roughly shown in Figure 3. Even though copy speedup doesn't affect map speedup and other phases' speedup, as a part of the whole job does affect the final speedup.

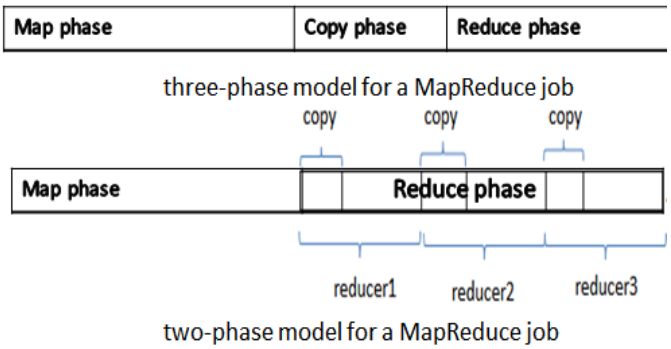


**Figure 3.** Models to describe a MapReduce job

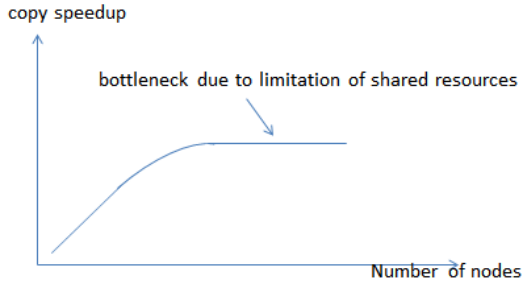


**Figure 4.** Bottleneck due to limitation on shared resources

We mainly did experiments on WordCount.We have two types of experiments: traning experiments and test experiments. Training experiments are used to accumulate experience, while test experiments are used to verify the correctness of our model. We obtained the start time and end time of every task. The duration is represented by their gap. We gathered them to form task chains and traced their execution time.

Instead of executing a job several times to compute its average execution time, we measure it by computing the expectancy from task chains. The advantage of this is we can reduce the number of experiment time.

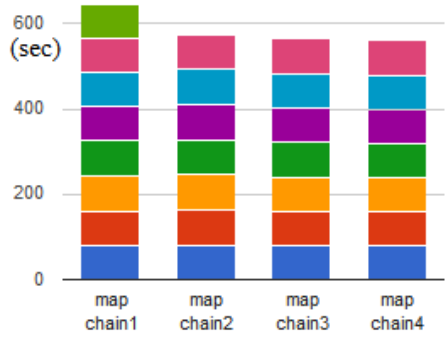


**Figure 4.** Map wave of WordCount.

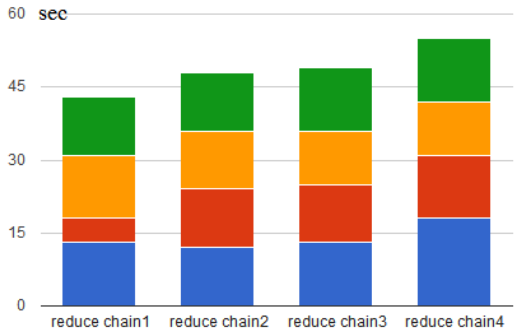


**Figure 5.** Reduce wave of WordCount.

## IV. CONCLUSION

Cloud computing is a new computational paradigm that offers on innovative business model for organization to adopt It without upfront investment. Cloud computing moves the application software and database to the large data centre where the data management and service may not be fully worthy. The security is an important aspect of quality of service. Cloud storage is much more beneficial and advantages then the earlier tradinitional storage system especially in scalability, cost reduction, portability and functionality requirement. This paper presented a storage technique in cloud computing several storage technique that provide integrity to data in cloud have been discussed in details.

## V. REFERENCES

[1] Parakh A, and Kak s (2014).online data storage using implicit security, Information Sciences, vol 179(19), 3323-3331.

[2] Li H,Dai Y et al.(2012)Identity-based authentication for cloud Computing, M. G. Jaatun, G. Zhao, and C, Rong(Eds): Cloud Computing, Lecture Notes in Computer Science, Vol 5931,157-166

[3] Wang Q,Wang C et al(2013).Enabling Public auditability And Data Dynamics For Storage Security in Cloud Computing, IEEE Transactions On Parallel and Distributed Systems,Vol22(5), 847-859.

[4] Balakrishnan S, Saranya G, et al. (2013).Introducing effective Third Party Auditing (TPA)For Data Storage Security in Cloud, International Journal of Computer Science And Technology ,vol 2,(2) 397-400.

[5] Dinesh C (2011).Data Integrity and Dynamic storage Way in Cloud Computing, Distributed Parallel, and cluster Computing.

[6] Kumar S P,Subramanian R (2014),An efficient and secure proctol for ensuring data storage security in Cloud Computing ,Internation Journal of computing Science ,vol8(6),No1,261-274.

[7] Sajithabanu S, Raj E G (2015).Data Storge Security in Cloud, International Journal of Computer science and technology, vol 2(4), 436-440

[8] Wang C, Wang Q et al. (2012),Towards secure And dependable Storage Services in Cloud Computing , IEEE Transactions on Services Computing ,vol5(2),220-232.

[9] Spillner J, Muller J et al (2012).Creating Optimal Cloud Storage System,future Generation Computer Systems ,vol29(4),1062-1072

[10] Dong B, Zheng Q et al. (2012). An optimized Approach for storing and accessing small files on cloud storage, Journal of Network and Computer Applications, 35(6), 1847-1862

[11] Deahmukh P M, Gughane A S et al. (2012).Maintaining Files Storage Security in Cloud Computing International Journal of Emerging Technology and Advance Engineering, vol2 (10), 2250-2459.

[12] Tang Y, Lee P P C et al (2010).FADE: A Secure overlay Cloud Storage System with File assured Deletion, 6th International ICST Conference, Secure Comm.

[13] Wang W,Li Z et al(2009).Secure and efficient Access to outsource Data, CCSW '09 Proceedings of the 2009 ACM workshop on Cloud Computing Security,55-66.

[14] Ensuring Data Storage Security in Cloud Computing .IOSR Journal of engineering –vol 2(12)-(2012) 2250-3021.