ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

A Survey on Cauchy Coding Approach to Optimize Fault Tolerance on Cloud Storage Systems Using hadoop: A Survey

Sunil Emekar, Prof. Vaishali Nandedkar

Dept. of Computer Engineering, PVPIT College of Engineering, Bavdhan Pune, India

ABSTRACT: In recent days bigdata integrity maintenance is the major objective in distributed systems storage. It includes audition using middle ware party for unauthorized access. This work implements protecting the data and regeneration of data if someone mishandles it. This job will be assigned to a Proxy server. The data of the users will be stored in public and private area of the cloud. So that only public cloud data will be accessed by user and private cloud will remain more secured. Once any unauthorized modification is made, the original data in the private cloud will be retrieved by the Proxy server and will be returned to the user. Cloud storage generally provides different redundancy configuration to users in order to maintain the desired balance between performance and fault tolerance. Data availability is critical in distributed storage systems, especially when node failures are prevalent in real life. This research work explores recovery solutions based on regenerating codes, which are shown to provide fault-tolerant storage and minimum recovery bandwidth. It presents specification of algorithm for write operation and how to implement it. In the work system describe First, CaCo uses Cauchy matrix heuristics to produce a matrix set. Second, for each matrix in this set, CaCo uses XOR schedule heuristics to generate a series of schedules. Finally, CaCo selects the shortest one from all the produced schedules. In such a way, CaCo has the ability to identify an optimal coding scheme, within the capability of the current state of the art, for an arbitrary given redundancy configuration. The load balancing is another feature of system, it can also maintain the secure access control mechanism for authenticated user.

I. INTRODUCTION

One of the biggest challenges in designing cloud storage systems is providing the reliability and availability that users expect. Once their data is stored, users expect it to be persistent forever, and perpetually available. Unfortunately, in practice there are a number of problems that, if not dealt with, can cause data loss in storage systems. So, the failure protection offered by the standard RAID levels has been no longer sufficient in many cases, and storage designers are considering how to tolerate larger numbers of failures [1] Technology shifts and market forces are changing the composition and design of storage systems. Topics for this diverse issue include the emergence of nonvolatile storage technologies, virtualization technologies that reduce the distinction between storage and computing platforms, advances in tape densities, the growing use of commodity and distributed storage, and the increasing importance of error and disaster recovery, autonomic storage management, peta scale file and archival storage, and long-term data preservation.[2]Cloud services inevitably fail: machines lose power, networksbecome disconnected, pesky software bugs cause sporadic crashes, and so on. Unfortunately, failure recovery itself is often faulty; e.g. recovery can accidentally recursively replicate small failures to other machines until the entire cloud service fails in a catastrophic outage, amplifying a small cold into a contagious deadly plague.

II. LITERATURE SURVEY

1] Ibrahim Adel Ibrahim at. al. proposed Intelligent Data Placement Mechanism for Replicas Distribution in Cloud Storage Systems in 2016 IEEE International Conference on Smart Cloud.
This system proposed a new replica placement policy for HDFS. The issue of load balancing is addressed in this work by evenly distributing replicas to cluster nodes. Therefore, there is no more need for any load balancing utility. IDPM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

can generate replica distributions that are perfectly even and satisfy all HDFS replica placement rules as confirmed by the simulation results. IDPM is designed for cluster environments in which all cluster nodes have the same computing capabilities. there is an exciting future work for the proposed policy. HDFS replica placement policy the replicas of data blocks cannot be evenly distribute across cluster nodes, so the current HDFS has to rely on load balancing utility to balance replica distributions which results in more time and resources consuming. These challenges drive the need for intelligent methods that solve the data placement problem to achieve high performance without the need for load balancing utility.

Advantages

1: The system provide effective approach for data storage into distributed databases in cloud as well as hadoop environment.

2: minimum time required for multiple chunk creation.

Disadvantages

1: No hadoop base slot configuration approach consider for execution.

2: Multiple replica's generate very high database cost.

2] Son Hoang Dau at. al. proposed Constructions of MDS Codes via Random Vandermonde and Cauchy Matrices over Small Fields Fifty-third Annual Allerton Conference in 2015.

System proposed a useful technique to tackle the constrained coding problems that includes random selection of the evaluation points of a Vander monde or a Cauchy matrix. This system proposed solutions require small finite fields whose sizes are polynomial in the dimensions of the generator matrices. It believe that this technique will be useful for solving a broad range of coding problems. System also discuss a technique of constructing MDS codes with some additional constraints. The key idea is that instead of using a totally random matrix where all entries are randomly selected from a finite field, it start from a Vandermonde or a Cauchy matrix and then randomly select their evaluation points. In this way, system can exploit not only the rich structure of Vandermonde and Cauchy matrices but also the benefits of random codes.

Advantages

1: System proposed new approach MDS code that can provide minimum time time complexity as well as efficiency tan all existing approaches.

2: System can work on multi cloud as well as signle cloud environment without any configuration changing.

Disadvantages

1: Cant discussed any scenario for distribution file systes.

2: System also work on single cloud that can be generate very high complexity.

3] Si Wu, Yinlong Xu at. Al. proposed Enhancing Scalability in Distributed Storage Systems with Cauchy Reed-Solomon Codes in IEEE 2014

This system studies the scaling of distributed storage systems with CRS codes. In particular, system formulate the scaling problem with an optimization model in which both the post scaling encoding matrix and the data migration policy are assumed to be unknown in advance. To minimize the I/O overhead for CRS scaling, it first derive the optimal post scaling encoding matrix under a given data migration policy, and then optimize the data migration process using the selected post scaling encoding matrix. Proposed scaling scheme requires the minimal data movement while achieving uniform data distribution. To validate the efficiency of our scheme, system also implement it atop a networked file system. It can also address the scaling problem for CRS codes from both the theoretical and the practical perspectives. Theoretically, system first present a method to design the post-scaling parity matrices which can improve the scaling efficiency, and then propose a searching algorithm to generate efficient data migration schemes with very low computational complexity.

Advantages

1: The system follow ACID properties that can be eliminate the data keakage and inconsistency issue.

2: System use SMC protocol for multi party authentication that can enhance the security of system

Disadvantages

1: Some time hard to generate Solomon code as matrix generation because it can take high time when data is big.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

2: Space complexity is another issue of multiple chunks.

4]: Brad Calder at al. proposed Windows Azure Storage: A Highly Available Cloud Storage Service with Strong Consistency in October 2011.

The system introduced Windows Azure Storage (WAS) is a cloud storage system that provides customers the ability to store seemingly limitless amounts of data for any duration of time. WAS customers have access to their data from anywhere at any time and only pay for what they use and store. In WAS, data is stored durably using both local and geographic replication to facilitate disaster recovery. Currently, WAS storage comes in the form of Blobs (files), Tables (structured storage), and Queues (message delivery). In this paper, system describe the WAS architecture, global namespace, and data model, as well as its resource provisioning, load balancing, and replication systems.

The Windows Azure Storage platform implements essential services for developers of cloud based solutions. The combination of strong consistency, global partitioned namespace, and disaster recovery has been important customer features in WAS's multi tenancy environment. WAS runs a disparate set of workloads with various peak usage profiles from many customers on the same set of hardware. This significantly reduces storage cost since the amount of resources to be provisioned is significantly less than the sum of the peak resources required to run all of these workloads on dedicated hardware.

Advantages

1: System demonstrate storage abstractions, Blobs, Tables, and Queues, provide mechanisms for storage and workflow control for a wide range of applications that can be enhance the flexibility of system.

2: Faster deployment times and client adoption.

3: Greater pool of development resources**.**

Disadvantages

1: costs and competition are going to be a huge struggle, especially as pricing models change.

2: Increased self-hosting and integration costs.

5] Liping Xiang and at al. proposed Optimal Recovery of Single Disk Failure in RDP Code Storage Systems in June 2010.

The system proposed an optimal and efficient disk recovery scheme, Row- Diagonal Optimal Recovery (RDOR), for single disk failure of RDP code that has the following properties:

(1) It is read optimal in the sense that it issues the smallest number of disk reads to recover the failed disk;

(2) It has the load balancing property that all surviving disks will be subjected to the same amount of additional workload in rebuilding the failed disk.

The system also explores the design state space and theoretically shows the optimality of RDOR. It also carry out performance evaluation to quantify the merits of RDOR on some widely used disks. The aim of this work is on designing an optimal recovery scheme for single disk failure in a storage system that uses double fault tolerant array codes. System considers the single failure recovery problem for storage systems which use the RDP codes. System show that one can exploit both parity disks to reduce the number of disk reads for the recovery of single disk failure and propose a hybrid recovery scheme RDOR. Using RDOR, the disk reads for recovery can be reduced by approximately 25% compared with the conventional recovery strategies.

Advantages

1: System proposed optimal recovery scheme for single disk failure which can reduce disk reads and therefore improves system recovery performance.

2: RDOR has the load-balancing property that all surviving disks will experience the same amount of workload to recover the failed disk, and the recovery scheme matches the lower bound on the number of disk reads.

Disadvantages

1: By using RDOR, the access time of an individual disk for recovery is increased by 15.00% to 22.00% than conventional scheme.

2: It can take a very high time for total data recovery than available approaches.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

6] Osama Khan and at. al. proposed Rethinking Erasure Codes for Cloud File Systems: Minimizing I/O for Recovery and Degraded Reads.

The system proposed a algorithm that finds the optimal number of codeword symbols needed for recovery for any XOR-based erasure code and produces recovery schedules that use a minimum amount of data. It also provide differentiate popular erasure codes based on this criterion and demonstrate that the differences improve I/O performance in practice for the large block sizes used in cloud file systems. In this system when a data disk fails in an erasure coded disk array, it is natural to reconstruct it simply using the P drive. Each failed symbol is equal to the XOR of corresponding symbols on each of the other data disks, and the parity symbol on the P disk. System call this methodology “Reading from the P drive.” It requires k symbols to be read from disk for each decoded symbol. Several open problems remain with respect to optimal recovery and degraded reads. While our algorithm can determine the minimum number of symbols needed for recovery for any given code, it remains unknown how to generate recovery-optimal erasure codes. System are pursuing this problem both analytically and through a programmatic search of feasible generator matrixes. Rotated RS codes are a first result in lowering degraded read costs.

Lower bounds for the number of symbols needed for degraded reads have not been determined.

Advantages

1: System first present an algorithm that finds the optimal number of symbols needed for recovering data from an arbitrary number of disk failures, which also minimizes the amount of data read during recovery.

2: System also demonstrates that minimizing recovery data translates directly into improved I/O performance for cloud file systems.

Disadvantages

1: System introduce minimum symbol required for data recovery, so any attack can be possible at the time data recovery.

2: Required hugh space, that be generate heating problem on database servers.

7] Kevin M. Greenan and at. Al. proposed Flat XOR-based erasure codes in storage systems: Constructions, efficient recovery, and tradeoffs in May 2010.

System described a number of two- and three-disk fault tolerant non-MDS flat XOR-code constructions: Chain codes, flattened parity-check array codes, Stepped Combination codes, and HD-Combination codes. Chain codes are based on previously known constructions, and flattening is a technique that applies to previously known codes. The combination codes are novel constructions. System expect additional flat XORcode constructions will be discovered that fill in more of the tradeoff space between replication and other MDS codes..The algorithm for flat XOR-codes that enumerates recovery equations, i.e., sets of disks that can recover a failed disk. System also describe two algorithms for flat XOR-codes that generate recovery schedules, i.e., sets of recovery equations that can be used in concert to achieve efficient recovery. Finally, system analyze the key storage properties of many flat XOR-codes and of MDS codes such as replication and RAID 6 to show the cost-benefit tradeoff gap that flat XOR-codes can fill. As the scale of storage systems increases, two- and three disk fault tolerant protection, or more, is needed. There is a pronounced gap between the cost-benefit propositions of replication and other MDS codes such as RAID 6 in multidisc fault tolerant systems. In such systems, flat XOR-codes offer cost-benefit tradeoffs that cover a large portion of the gap between replication and other MDS codes.

Advantages

1: System introduce the shortage data recovery equations that can be take the minimum time for execution.

2: System codes achieve optimal storage overhead and small write costs.

Disadvantages

1: Its hard to recovery for logical backups because no any provision introducing for periodical backup for system.

2: Back system is manual execution.

8] Daniel Ford and et. Al. proposed Availability in Globally Distributed Storage Systems in 2010.

System characterize the availability properties of cloud storage systems based on an extensive one year study of Google’s main storage infrastructure and present statistical models that enable further insight into the impact of multiple design choices, such as data placement and replication strategies. With these models its compare data availability under a variety of system parameters given the real patterns of failures observed in our fleet. In particular,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

though disks failures can result in permanent data loss, the multitude of transitory node failures account for most unavailability. System present a simple time-window-based method to group failure events into failure bursts which, despite its simplicity, successfully identifies bursts with a common cause. System develop analytical models to reason about past and future availability in our cells, including the effects of different choices of replication, data placement and system parameters. Inside Google, the analysis described in this paper has provided a picture of data availability at a finer granularity than previously measured. Using this framework, system provide feedback and recommendations to the development and operational engineering teams on different replication and encoding schemes, and the primary causes of data unavailability in our existing cells.

Advantages

1: Determining the acceptable rate of successful transfers to battery power for individual machines up on a power outage.

2: Focusing on reducing reboot times, because planned kernel upgrades are a major source of correlated failures.

Disadvantages

1: High server as well as maintenance cost.

2: Required high IAAS base environment for hosting and storage.

III. CONCLUSION

Many efforts have been devoted to achieve this goal. Initially, people discover that the density of a Cauchy matrix dictates the number of XORs. For this reason, an amount of work has sought to design codes with low density. Moreover, some lower bounds have been derived on the density of MDS Cauchy matrices. In the current state of the art, the only way to discover lowest-density Cauchy matrices is to enumerate all the matrices and select the best one. In the proposed work, system focus on CaCo, a new approach that incorporates all existing matrix and schedule heuristics, and thus is able to identify an optimal coding scheme within the capability of the current state of the art for a given redundancy configuration. The selection process of CaCo has an acceptable complexity and can be accelerated by parallel computing. It should also be noticed that the selection process is once for all. The experimental results demonstrate that CaCo outperforms in hadoop 2.7 with data node as mongoDB.

IV. FUTURE WORK

For the future enhancement system can consider as load rebalancing in hybrid cloud environment with hadoop. The thermal management and energy saving approach with resource virtualization is another interesting area for such concepts. Cauchy's rule can be incubated with bandwidth of performance with respect to time for series computation of delay in network node retiring. The delay time reduction under narrow network is still a bottle neck situation. Either of this can be improvised with technical reduction of data sets and its indexing

REFERENCES

[1] Ibrahim Adel Ibrahim at. al. proposed Intelligent Data Placement Mechanism for Replicas Distribution in Cloud Storage Systems in 2016 IEEE International Conference on Smart Cloud.

[2] Son Hoang Dau at. al. proposed Constructions of MDS Codes via Random Vandermonde and Cauchy Matrices over Small Fields Fifty-third Annual Allerton Conference in 2015.

[3] Si Wu, Yinlong Xu at. Al. proposed Enhancing Scalability in Distributed Storage Systems with Cauchy Reed-Solomon Codes in IEEE 2014

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, „Above the clouds: A berkeley view of cloud computing‟, EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28, February 2009.

[5] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, „Reclaiming space from duplicate files in a serverless distributed file system‟, Distributed Computing Systems, International Conference on, Vol. 0, pp. 617, 2002.

[6] D. Russell, Data Deduplication Will Be Even Bigger in 2010, Gartner, February 2010.

[7] H. S. Gunawi, N. Agrawal, A. C. Arpaci-Dusseau, R. H. Arpaci- Dusseau, and J. Schindler, „Deconstructing commodity storage clusters‟, in ISCA ‟05: Proceedings of the 32nd annual international symposium on Computer Architecture. Washington, DC, USA: IEEE Computer Society, pp. 60–71, 2005.

[8] S. Quinlan and S. Dorward, „Venti: a new approach to archival storage‟, in First USENIX conference on File and Storage Technologies, Monterey, CA, 2002.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

[9] L. L. You, K. T. Pollack, and D. D. E. Long, „Deep store: An archival storage system architecture", In Proceedings of the 21st International Conference on Data Engineering IEEE, pp. 804–815, 2005.

[10] Y. Fu, H. Jiang, N. Xiao, L. Tian, and F. Liu, „AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment", in Proceedings 13th IEEE International Conference on Cluster Computer, pp. 112-120, 2011.