

SURVEY ON DATA REPLICATION BASED ON HADOOP FILE SYSTEM

¹A.Roshan Kumar Reddy, ²V. Banu Priya¹UG scholar, ²Assistant professor

Computer Science and Engineering,

Saveetha School of Engineering, Saveetha University, India

¹roshankumarreddy007@gmail.com, ²banupriyav.sse@saveetha.com

Abstract: Replication means maintaining the same copies of data at more than one system. Replication is one of the important design requirement of any distributed file system(DFS). Now a days most of the academia and industries are using Hadoop DFS for storing and processing their data.Hadoop distributed file system (HDFS) is designed to store huge data sets reliably and used for processing massive data in parallel. In HDFS inefficiency of Replication is one of the problem that causes the performance decrement of a file system. To solve this problem, we propose an efficient data replication scheme in Hadoop framework which improve the availability of the DFS. To solve the data locality problem, we propose an efficient data replication scheme based on access count prediction in a Hadoop framework. By the previous data access count, the existing data replication scheme predicts the next access count of data files using Lagrange's interpolation. Then, the proposed data replication scheme determines the replication factor with the predicted data access count, whether it generates a new replica or it uses the loaded data as cache selectively. Finally, the proposed scheme provides improvement of data locality. By performance evaluation, proposed efficient data replication scheme is compared with default data replication setting of Hadoop that shows proposed scheme reduces averagely of the task completion time in the map phase. Regarding the data locality, proposed scheme provides the increase of node locality by and the decrease of rack and rack-off locality

Keywords: .Hadoop, Data locality, Access Prediction, Data Replication, Data Placement

1. Introduction

Because of requirement of efficient programming model for large scale data in a distributed computing environment, Map Reduce is developed by Google Hadoop is one of the open sources implementation of the Map Reduce Hadoop Distributed File System (HDFS) is developed by Apache Foundation that provides advantages of the power of high-speed computing clusters and storage and high performance in

big data storage .Using HDFS, high availability and fault-tolerance such as replication are provided. In Hadoop scheduling policy, there is the case of the data locality problem that can occur, when the assigned node load the data block from another node. The main factor of data locality in Hadoop refers to the distance between data and the assigned node. In this paper, we propose an efficient data replication scheme in a Hadoop framework based on access count prediction. Proposed data replication scheme focuses improving the data locality in the map phase, and reducing the overhead of data transfer that increases total execution time. In proposed data replication scheme, we efficiently determined the increasing replication factor and avoiding the unnecessary data replication Proposed scheme optimizes and maintains the replication factor effectively by proposed predictor. Proposed scheme minimizes the data transfer load between racks by proposed data replication algorithm. The rest of this paper is organized as follows. In Section, we discuss previous works on data locality in a Map Reduce framework and introduce problems of data locality. The proposed efficient data replication scheme based on access count prediction is presented in Section. Section shows the performance evaluation of proposed scheme. Finally, Section gives our conclusions.

Replication means maintaining the same copies of data at more than one system. Replication is one of the important design requirements of any distributed file system (DFS).Now a days most of the academia and industries are using Hadoop DFS for storing and processing their data. Hadoop distributed file system (HDFS) is designed to store huge data sets reliably and used for processing massive data in parallel. In HDFS inefficiency of Replication is one of the problems that cause the performance decrement of a file system. To solve this problem, we propose an efficient data replication scheme in Hadoop framework which improve the availability of the DFS.Data replication is required, it replicates the data block in the rack that does not have the data block Hence, it tries to increase a probability that a data block is located in the rack in which the node reside the authors proposed a delay scheduling method for both data locality and fairness If the job cannot launch a local task due to the data

locality reason, it waits for a while to allow other jobs to launch tasks. Although the delay scheduling method is designed to improve data locality. Let the jobs wait for a small amount of time, resulting in violating the fairness for jobs. Data replication is required, it replicates the data block in the rack that does not have the data block. Hence, it tries to increase a probability that a data block is located in the rack in which the node reside the authors proposed a delay scheduling strategy for the two information territory and decency. On the off chance that the activity can't dispatch a nearby undertaking because of the information territory reason, it sits tight for some time to enable different occupations to dispatch errands. In spite of the fact that the defer booking strategy is intended to improve information region. Give the employments a chance to sit tight for a little measure of time, bringing about disregarding the reasonableness for occupations. To take care of the information territory issue, we propose an effective information replication plot in light of access include expectation a Hadoop structure. By the past information get totally, the current information replication plot predicts the following access check of information records utilizing Lagrange's interjection. At that point, the proposed information replication conspire decides the replication factor with the anticipated information get to check, regardless of whether it produces another imitation or it utilizes the stacked information as reserve specifically. At last, the proposed conspire gives change of information area. By execution assessment, proposed productive information replication conspire is contrasted and default information replication setting of Hadoop that shows proposed plot decreases averagely 8.9% of the errand fruition time in the guide stage. With respect to information territory, proposed plot gives the expansion of hub area by 6.6% and the abatement of rack and rack-off region by 38.9% and 56.5%. In Hadoop booking approach, there is the situation of the information region issue that can happen, when the doled out hub stack the information hinder from another hub. The fundamental factor of information territory in Hadoop alludes to the separation amongst information and the doled out hub. Regarding the information region, the overhead of rack-off region is more noteworthy than overhead of rack area. To keep the information territory issue, we propose a productive information replication plot utilizing forecast by the entrance check of information records and an information replication arrangement calculation lessening instance of rack and rack-off area.

2. Related Works

1 Smita Hegde Web documents. Replication is a key strategy for improving reliability, fault tolerance and availability in distributed systems. This paper gives a brief introduction to replication and the basic principles

used for replication in DFS. The next section discusses the enormous load on the web servers and inability of the caching techniques lead to the demand for replicating web documents for a better, speedy, and efficient way of delivering information. This has lead to a lot of research in the areas of replication in web documents. This paper introduces the latest research conducted by Guillaume Pierre and others and introduces Globule, a platform which automates all aspects of replication. Between the 2 approaches, Optimistic replication approach is usually followed in DFS. this can be a trade off with consistency against convenience, because it permits updates to require place even within the presence of communication failures at a price of typically violating single-copy serializability. though optimistic replication systems are in use for some time, it's seen that there's still an absence of adequate metrics. Client-Server model is usually used however once addressing mobile users Peer-to-peer model is employed. this enables direct communication and synchronization between all the peers however has measurability issues. therefore relying upon the system the replication approach and model is chosen. As each have some trade off. Due to the rise within the net, the only server approach isn't enough to service all the requests. therefore replication or caching techniques area unit 2 solutions for the network congestion and latency time issues caused by single server. Similarly, even with net replication no single replication or caching policy is perfect for all documents. The Adaptive strategy with based-trace performance provides higher results. The bestowed platform, globe for net documents replication is purposed to own dynamic creation and removal of replicas, consistency management and automatic consumer redirection. It is expected to fulfill the ever-increasing quality of service expectations that users have.

2. Dr. M Praveen Kumar Due to the advent of new technologies, devices, Now-a-days the measure of information delivered by humanity is developing quickly consistently. Enormous information incorporates gigantic volume, high speed, and extensible assortment of information. The Big information is as Structured information, unstructured information and Semi Structured information. It is extremely a dull assignment to process such information through a customary database server. Google tackled this issue utilizing a calculation called Map Reduce, utilized as a part of the innovation Hadoop. In this paper we talk about some security issues in Hadoop related with Big information. Huge information applications are particularly valuable to Organizations, all kind of Companies might be little or extensive scale organizations and to the Industries This paper exhibits the tremendous data information and traits used as a piece of around the globe. The issues

are in addition said to give thought with respect to the gigantic data issues continuously. The security issue is coordinated all the more all together toward increase the security in colossal data. We can improve security in tremendous data by using any of the approach or by joining these three strategies in Hadoop Distributed File System which is the base layer in Hadoop, where it contains extensive number of pieces. techniques are familiar with vanquished certain issues occurs in the name center point and moreover in Data center point. In Future these procedures are also completed in various layers of Hadoop Innovation

3. Dr. Sunita Varma Department of Computer Technology and Application Distributed computing gives administrations utilizing ritualized assets through Internet on pay per utilize premise. These administrations are conveyed from a great many server farms which are associated with each other. Cloud framework comprises of ware machines. The customer information is put away on these machines. Likelihood of equipment disappointment and information debasement of these low execution machines are high. For adaptation to non-critical failure and enhancing the unwavering quality of the cloud framework the information is recreated to different frameworks. Hadoop Distributed File System is utilized for conveyed capacity in cloud framework. The information is put away in the type of settled. The information put away in HDFS is duplicated on numerous frameworks for enhancing the unwavering quality of the cloud framework. Square reproduction arrangement calculation is utilized as a part of HDFS for duplicating the information piece. In this calculation, QoS parameter for imitating the information square isn't indicated among st customer and administration supplier as administration level assertion. The situation of copy utilizing QoS-mindful imitation situation calculation has less replication cost as thought about to piece copy arrangement calculation for estimated 3 GB size of documents and 2000 Data Nodes. The replication time of bigger documents in light of QoS-mindful information reproduction situation calculation was enhanced when contrasted with piece reproduction position calculation as appeared in fig. 8. In spite of the fact that capacity furthermore, number of Data Nodes can ascend to petabytes of information furthermore, a huge number of Data Nodes. It will give critical outcome

4. Jacob Leverich, Christos Kozyrakis Conveyed preparing structures, for example, Yahoo's! Hadoop and Google's Map Reduce, have been effective at outfitting far reaching datacenter assets for substantial scale information investigation. Be that as it may, their impact on datacenter vitality productivity has not been investigated. In addition, the file system segment of these systems successfully blocks downsize of bunches sending these structures (i.e. working at diminished

limit). This paper displays our initial work on adjusting Hadoop to permit downsize of operational groups. We locate that running Hadoop bunches in fragmentary arrangements can spare in the vicinity of 9% and half of vitality utilization, and that there is a tradeoff between execution vitality utilization. We likewise plot further look into the vitality proficiency of these structures Hadoop's Map Reduce, take note of that H Base and Big Table expand upon the same file system framework as their kin Map Reduce structures. While they have comparable necessities of the capacity layer as Map Reduce (enormously adaptable, circulated, hearty), the idea of their workload and action is far expelled. Map Reduce calculations are portrayed by long, unsurprising, spilling I/O, enormous palatalization, and non-intelligent execution. Then again, the previously mentioned organized information administrations are frequently utilized progressively information serving situations

5. Nikhil Chaturvedi This paper concentrates on replication in conveyed document frameworks. Replication is a key technique for enhancing dependability, accessibility and execution in conveyed frameworks. This paper gives a short prologue to replication and different calculations have been examined and a nitty gritty examination has been performed. The correlation table shows it unmistakably that a calculation fulfills the replication requirement. Another favorable position of a dispersed record framework is the capacity to impart data to numerous various clients. DFS is a productive, extensible framework. Moreover, the server tracks which customers have stored duplicates of records, decreasing the requirement for the customer to always inquiry the server, and also lessening the system and server stack. the server tracks which customers have stored duplicates of records, lessening the requirement for the customer to always question the server, and in addition diminishing the system and server load. This is an exchange off with consistency against accessibility, as it enables updates to occur even within the sight of correspondence disappointments at a cost of a few times abusing single-duplicate serializability. Albeit hopeful replicate particle frameworks have been being used for quite a while, it is seen that there is as yet an absence of sufficient measurements. Customer Server demonstrate is for the most part utilized yet when managing portable clients Peer-to-peer show is utilized. This permits coordinate convey particle and synchronization between every one of the associates however has adaptability issues. Along these lines relying on the framework the replication approach and model is chosen. As both have some exchange off.

6. Neha M Patel Narendra M Patel Mosin I Hasan Mayur M Patel A circulated record framework is a customer/server-based application that enables customers to access and process information put away

on the server. At the end of the day, we can state Distributed record framework comprises of programming living on organize servers and customers that straightforwardly interfaces shared envelopes situated on various document servers into a solitary namespace for enhanced load sharing and information accessibility. At the point when a customer gadget recovers a record from the server, a document shows up as an ordinary document on the customer machine, and the client can work with the document in an indistinguishable courses from in the event that it were put away locally on the workstation. At the point when the client completes the process of working with the record, it is come back to the server, which stores the now-adjusted document for recovery at a later time. The stream of this examination papers is: area two gives diagram of various Distributed File System and related work, segment three explains outlining issues of appropriated document framework and correlation between various record framework on the premise of different factors, and segment 4 portrays Hadoop DFS read compose operation., and toward the end we propose a proof of idea for composing copies on Hadoop DFS utilizing parallel approach. At long last, finish of this paper is illustrated. Information replication is a procedure regularly used to enhance information accessibility. In HDFS each piece is imitated on various hubs. In our approach imitation factor is three. In this paper we introduce new approach for productive reproduction arrangement on HDFS that can build throughput by imitating information hinders in parallel and furthermore expansive volume of information are exchange which can expand information exchange rate.

7. Dejene Boru · Dzmitry Kliazovich · Fabrizio Granelli · Pascal Bouvry · Albert Y. Zomaya Cloud computing is an emerging paradigm that gives computing, verbal exchange and storage resources a provider over a community. Conversation resources frequently end up a bottleneck in provider provisioning for plenty cloud packages. Therefore, statistics replication which brings records towards facts customers is visible as a promising solution. It lets in minimizing community delays and bandwidth utilization. on this paper we observe records replication in cloud computing facts facilities. not like different processes available inside the literature, we consider both electricity efficiency and bandwidth consumption of the gadget. that is further to the stepped forward nice of provider QoS acquired as a result of the reduced conversation delays. The assessment results, received from each mathematical model and sizable simulations, help to unveil overall performance and electricity efficiency tradeoffs in addition to guide the layout of destiny statistics replication answers. Information replication in geographically allotted cloud computing facts centers and proposes a novel replication solution which in addition to traditional overall performance

Metrics, inclusive of availability of network bandwidth ,Optimizes energy efficiency of the gadget. Similarly, optimization of verbal exchange delays leads to improvements in first-rate of consumer experience of cloud programs

8. B. Saraladevia, N. Pazhanirajaa, P. Victor Paula, M.S. Saleem Bashab, P. Dhavachelvanc Hadoop (Highly Archived Distributed Object Oriented Programming) was made by Gouge Cutting and Mike Cafarella in 2005 for supporting a circulated web crawler Project. It is an Open source Java Framework innovation stores, access and increase substantial assets from huge information in a conveyed form at less cost, high level of adaptation to non-critical failure and high adaptability. Hadoop [5] handles huge number of information from variousframework like Images, videos, Audios, Folders, Files, Software, Sensor Records, Communication information, Structured Query, unstructured information, Email& discussions, and anything which we can't think in any configuration. Every one of these assets can be put away in a Hadoop bunch with no pattern portrayal as opposed to gathering from various frameworks. There are numerous segments associated with Hadoop like Avro, Chukwa, Flume, Base, Hive, Lucene, Oozie, Pig, Snoop and Zookeeper. The Hadoop Package additionally gives Documentation, source code, area mindfulness, Work planning. A Hadoop group contains one Master hub and Many Slave hubs. The ace hub comprises of Data node, Name hub, Job Tracker and Task Tracker where slave hub goes about as both a Task Tracker and Data hub which holds register just and information just laborer hub. it is used to access a data blocks correctly and also only by unauthorized user. Here Ticket Granting Ticket and Service Ticket playing a major role in providing a security in name node. The Second approach is based on Bull Eye Algorithm Approach explains about the security method from node to node and also scan the nodes in all the angles to prevent from attacks

9. Sharmila Krishnamoorthy The Master hubs regulate putting away of information and running parallel calculations on every one of that information utilizing Map Reduce. The Name Node regulates and facilitates the information stockpiling work in HDFS, while the Job Tracker directs and organizes the parallel handling of data utilizing Map Reduce. Slave Nodes are most by far of machines and do all the cloudy work of putting away the information and running the calculations. Each slave runs a Data Node and Task Tracker daemon that speaks with and gets guidelines from their lord nodes. The Task Tracker daemon is a slave to the Job Tracker in like manner the Data Node daemon to the Name Node. Data Replication in Hadoop system to research the information region issue was tested and demonstrated that replication enhances execution and furthermore diminishes it as far as

possible is crossed. Facilitate the replication factor was bolstered by an entrance check expectation calculation for information documents utilizing Lagrange's introduction which advances the replication factor per information record. Execution assessment demonstrated that our information replication conspire decreases the errand fulfillment time. The errand finishing time for Pi esteem figuring began with 680 s and boiled down to 8.12 s and comparably the Word Count application starts with consummation time of 1384300 s and boiled down to 1300430 s. In any case, once as far as possible is achieved the finish time shoots up to 1608600 s. A circulated framework is a pool of self-sufficient register hubs associated by quick systems that show up as a solitary workstation. As a general rule, taking care of complex issues includes division of issue into sub undertakings and each of which is comprehended by at least one register hubs which speak with each other by message passing. The present slant towards Big Data investigation has prompt such process serious undertakings.

10. Cristina L. Abad, Yi Lu, Roy H. Campbell Putting information as close as conceivable to calculation is a typical routine with regards to information concentrated frameworks, normally alluded to as the information region issue. By dissecting existing generation frameworks, we affirm the advantage of information area also, find that information have diverse prevalence and changing relationship of gets to. We propose DARE, an appropriated versatile information replication calculation that guides the scheduler to accomplish better information region. Set out takes care of two issues, how numerous copies to allot for each record and where to put them, utilizing probabilistic examining and a focused maturing calculation freely at every hub. It exploits existing remote information gets to in the framework and causes no additional system utilization. Utilizing two blended workload follows from Facebook, we demonstrate that DARE enhances information territory by more than 7 times with the FIFO scheduler in Hadoop and accomplishes more than 85% information territory for the FAIR scheduler with defer booking Turnaround time and occupation stoppage are lessened by 19% and 25%, respectively. Hadoop is utilized on numerous little and medium measured organizations that much of the time utilize outsider virtual bunches of in-house Hadoop server farms. Regardless of whether these administration suppliers move up to higher-end organize textures, dispute still exists for the mutual system asset. Also, for all intents and purposes distributed hubs may not be on a similar rack, which additionally focuses on the system and adds idleness to organize communications. Mechanism that can enhance information territory by more than 7- times for a FIFO scheduler and 70% for the Fair scheduler, without bringing about in additional

systems administration overhead. Turnaround time and lull are enhanced by 19% and 25%, separately. Moreover, our plan is scheduler-freethinker also, can be utilized as a part of parallel with different plans that mean to enhance area

3. Conclusion

To tackle the information region issue, we proposed a proficient information replication plot in a Hadoop structure. Proposed effective information replication conspire goes for enhancing the information territory in the guide assignment stage, and decreasing the aggregate handling time. By forecast of access tallies of information document, we upgrade the replication factor per every datum record. Proposed information replication conspire decides producing new imitation or utilizing the stacked information as store. By execution assessment, we demonstrate three noteworthy focal points of proposed plot. As a matter of first importance, proposed plot upgrades and keeps up the replication factor successfully. What's more, proposed plot limits the information exchange stack between racks. At long last, proposed conspire diminishes the preparing time of Map Reduce occupations.

References

- [1] Dejene Boru ,Dzmitry Kliazovich, Fabrizio Granelli ,Pascal Bouvry · Albert Y. Zomaya Received: 18 March 2014 / Revised: 4 July 2014 / Accepted: 24 September 2014 © Springer Science Business Media New York 2015
- [2] B.SaraladevN.Pazhaniraja,P.Victor Paul, M.S.SaleemBasha, P.Dhavachelvan 2nd International Symposium on Big Data and Cloud Computing Procedia Computer Science 50 (2015) 596– 601
- [3] JunghaLee, JaehwaChung and Daewon Lee. International Journal of Software Engineering and Its ApplicationsInternational Journal of Software Engineering and Its ApplicationsVol.9, No. 12 (2015), pp. 177-186
- [4] Replication in Distributed File Systems By Smita Hegde ACM Computing Survey, Vol. 22, No 4, December 2012
- [5] Analysis of Replication and Replication Algorithms in Distributed System Kevin Chetty, Graeme E. Smith and Karl Woodbridge. Volume2,Issue 5, May 2012 ISSN: 2277 128XInternational Journal of Advanced Research in Computer Science and Software Engineering
- [6] Adaptive Data Replication for Efficient Cluster Scheduling 2011 IEEE International Conference on Cluster Computing

- [7] Improving Data Transfer Rate and Throughput of HDFS using Efficient Replica Placement International Journal of Computer Applications (0975 – 8887) Volume 86 – No 2, January 2014
- [8] Security Issues in Hadoop Associated With Big Data IOSR Journal of Computer Engineering (IOSR-JCE)e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. VI (May - June 2017), PP 80-85 www.iosrjournals.org
- [9] QoS-Aware Data Replication in Hadoop Distributed File System Int. J. Advanced Networking and Applications Volume: 07 Issue: 03 Pages: 2741-2751 (2015) ISSN: 0975-0290
- [10] On the Energy (In) efficiency of Hadoop Clusters Jacob Leverich, Christos Kozyrakis. 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
- [11] Abad, L. Cristina, Y. Lu, and R. H. Campbell, "DARE: Adaptive data replication for efficient cluster scheduling", Cluster Computing (CLUSTER), 2011 IEEE International Conference on. Ieee, (2011).
- [12] Khanli, L. Mohammad, A. Isazadeh, and T. N. Shishavan, "PHFS: A dynamic replication method, to decrease access latency in the multi-tier data grid", Future Generation Computer Systems, vol. 27, no. 3, (2011), pp. 233-244.
- [13] S. Seo, "HPMR: Prefetching and pre-shuffling in shared Map Reduce computation environment", Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on. IEEE, (2009).
- [14] X. Zhang, "An effective data locality aware task scheduling method for Map Reduce framework in heterogeneous environments." Cloud and Service Computing (CSC), 2011 International Conference on. IEEE, (2011).
- [15] M. Zaharia, "Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling", Proceedings of the 5th European conference on Computer systems. ACM, (2010).
- [16] Dong, X., El-Gorashi, T., Elmirghani, J.M.H.: Green IP over WDM networks with data centers. J. Lightwave Technol. 29(12), 1861–1880 (2011)
- [17] Ping, F., Li, X., McConnell, C., Vabbalareddy, R., Hwang, J.H.: Towards optimal data replication across data centers. In: International Conference on Distributed Computing Systems Workshops (ICDCSW), pp. 66–71. (2011).
- [18] Li, W., Yang, Y., Yuan, D.: A novel cost-effective dynamic data replication strategy for reliability in cloud data centers. In: International Conference on Dependable, Autonomic and Secure Computing(DASC), pp. 496–502. (2011).
- [19] Kliazovich,D., Bouvry, P.,Khan, S.U.: GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. J. super compute. 62(3), 1263–1283 (2012)
- [20] Chernicoff, D.: The shortcut guide to data center energy efficiency. Real time publishers, New York (2009)
- [21] S.V.Manikanthan and K.Baskaran "Low Cost VLSI Design Implementation of Sorting Network for ACSFD in Wireless Sensor Network", CiiT International Journal of Programmable Device Circuits and Systems,Print: ISSN 0974 – 973X & Online: ISSN 0974 – 9624, Issue : November 2011, PDCS112011008.
- [22] T.Padmapriya, Ms. N. Dhivya, Ms U. Udhayamathi, "Minimizing Communication Cost In Wireless Sensor Networks To Avoid Packet Retransmission", International Innovative Research Journal of Engineering and Technology, Vol. 2, Special Issue, pp. 38-42.

