A Survey on Meliorate Data Distribution in Big Data

S.Shefali[1] S. Guna Sundari[2]
[1]M.E. Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Velammal Engineering College

Abstract—The data distribution techniques hold an important role by having an influence on increasing the rate by which data can be processed as well as transferred in a cloud. A study has been made on various data distribution approaches one can look up on either in single cloud or multi cloud processing or storage systems. Centralized data distribution approach is the easiest way to set up in a client-server model but falls short of being an optimum model by the virtue of single point failure. A slight improvement over centralized is semi centralized approach. Peer to Peer approach stands firm by overcoming the single point failure, decreasing the latency in the data distribution. In this paper, various data distribution techniques along with P2P have been discussed in detail to uplift the data distribution in big data. A holistic study has been done on the hadoop ecosystem and the impact on load balancing which would eventually be taken as a parameter for testing its efficacy. Through all the required survey on various interlinked topics of data distribution mechanisms, we have inferred that P2P is still not an utopian scheme for data distribution and more work has to be contributed in coming days to remove all or few of its effecting demerits on data distribution.

Key words: Peer to Peer, Map Reduce, Hadoop, Centralized, Semi Centralized, Load Balancing

I. INTRODUCTION

Big data is a popularly used to describe exponential growth and availability of data both structured and unstructured . Big data maybe as important to business- and society as the Internet has become. Usually, the main technique for data crunching was to move the data to the computational nodes, which were shared.

In this, the system assumes that data is available at the machine that will process it, as data is stored in a distributed file system such as HDFS. Newly provisioned VMs need to contain the data the will be processed. In this paper we solve lack of high data transfer rationalizing load distribution , latency factor and low bandwidth and updating of peer VM.

A. Effective Data Distribution technique for Multi Cloud Storage in cloud computing

Cloud data storage redefines the outsource data as well as the input data given from various services, service provider and user. [1]Cloud computing holds good for a wide range of data, reliability of the data, security of the data and the processing the data. The single point of failure affects the data availability which would lead to crash in cloud service provider and henceforth it becomes difficult to retrieve the stored data from the cloud. That scenario is mainly observed in non distributed or centralized distributed network. In this scheme of cloud storage an enormous amount of time, data, security are breached. In order to stop the service provider to observe the data, data can be fragmented and distributed to several service providers. Data is fragmented so that applications work with views rather entire relation, better efficiency and gets stored close to where its more frequently used. There are four types of fragmentation , those are horizontal, vertical, mixed and derived. An implementation over distributed cloud storage is hierarchical approach where there is no single point failure.

B. Distributed Data Clustering in Multi-Dimensional P2P Network

P2P network requires costly reorganization of data which includes distribution of data on a priority basis. In P2P computing, network data is distributed among the nodes (peers) to perform a critical function in decentralized manner. All nodes are users and provider of resources in a given or a selective cloud. Data clustering matches well with the features of the P2P network. Since these model uses local information and therefore clustering is effective for handling dynamic topological network. Multi-dimensional P2P system is used to communicate between the peers, distribute the data between the peer and eventually the load is distributed across the cloud making it firmly balanced. [2] Data clustering in multi-dimensional P2P network does not require a specific re organization of the network and alternating or compromising the service of P2P system.

C. Optimal power allocation and load distribution for multiple Heterogeneous Multicore Server Processors across Cloud and Data Centers

For multi-heterogeneous servers the performance of the cloud depends on three important factors[3] :

- Power allocation
- Load-distribution
- Energy-efficiency

For an increase in cloud site, power allocation becomes an important factor to attain optimization. For an increase in no.of VM's load balancing should be taken care of, to attain an hierarchical approach. The performance can be optimized by Load distribution. In a scenario, where companies are computing for better cloud performance , Load balancing will play an important role in retrieving the data from a specific server.

D. Semi Distributed Cloud Computing System With Load Balancing Algorithm.

Cloud computing involve delivery of hosted service through virtualized machine that run of efficacy on balanced node. In a cloud computing scenario computers are connected with different physical machine over the data structure to smoothly balance the load and implement the data distribution either in centralized or non-centralized approach. There are two ways to perform dynamic load balancing they are distributed and non-distributed approach. In a distributed one dynamic load balancing is done by nodes present in a system and load balancing is shared among the nodes. In non-distributed form, each node work independently towards the local goal fixed by the nod.eg: the response time of the local task.[6] Dynamic load

balancing of distributed nature generate more messages to the non-distributed one because every node in system interacts with every other node of a system. Therefore an evident benefit of a distributed dynamic load balancing is even if one more node of system fails it will not cause the load balancing process to halt. Distributed dynamic load balancing can introduce extreme stress on a system which needs to introduce status information with the every other node in a system. In a non-distributed type either one node or group of node do the task of load balancing.

E. Hierarchical Peer-To-Peer System

Peer to Peer systems are popular for performing computing in a cost efficient manner. A central problem of p2p systems is to assign and locate resources among peers through a p2p look up services. Look up service essentially performs the basic function of determining the peer that is responsible for a given key. Efficiency of a look-up services is generally measured in terms of no .of peer hopes needed to route a message, to responsible peer as well as the size of routing table maintained by each peer. Implementation of distributed look up service are often referred as distributed hash table (DHT)[5] .DHT provides improvement to an application in terms of showcasing fault of the system. Hierarchical design offers highest ability by using more reliable peers at the top level.

F. PeerDB: A P2P system for Distributed Data Sharing

Peer to peer is an emerging paradigm used for distributed data system. This nature gives opportunity for new applications to develop. P2P based system participates in a fully fledged data distribution so as to attain a wide bandwidth in optimum load balancing. P2P data distribution supports mechanisms which are dynamically close to hierarchical approach in which the data chooses a path so as to reduce the loss due to error[6] .P2P distributed data management is designed in support distributed data base system in data base technology. There are many features of distributed data base system which differ from nominal P2P system. In P2P system nodes can join and leave any time where as in distributed data base system nodes are added and removed on demand i.e. when there is need for growth or retirement. In P2P system there is no predetermination of node where as in distributed data base system the nodes are static which follows regular path.

G. In Comparison between centralized and distributed cloud storage data center topology

Distributed data center topology has an enhanced effect on average through put compared to centralized topology. Dropbox, skydrive use a centralized topology which connects client over data center. These data center are synonym to machine which are connected virtually in globe of cloud. The efficiency of centralized and distributed cloud storage are determined on the basis of throughput, load balancing, load scheduling[7]. The mode of data distribution differentiates different topology that are necessary to personal and corporate world. The advantage of centralized data center topology is low operation expenses at the same time the disadvantage is high risk of single point failure. The distributed cloud storage specialized over centralized cloud by removal of single point failure but includes the higher expenses to implement. Performance is calculated from connection set up till the end of the file to as to measure the large delays while uploading files on the server.

H. An optimal task selection scheme for Hadoop Scheduling

Hadoop is an open source implementation map reduce and widely used by vast amount of users. Map reduce is the popular programming model used to solve wide range of big data application in cloud computing environment. Hadoop uses master-slave architecture for running application concurrently. Fifo lacks in performance in terms of response time and fair sharing among users[8]. To overcome this hadoop fair scheduler is used to share a cluster effectively among users. Hadoop scheduler chooses a job to assign a slot accordingly to the scheduling policy where a node becomes free. After deciding the job, the scheduler starts searching to find the local task from a list of unscheduled task of the job, to launch a free node.

I. A novel approach for replica synchronization in a Hadoop distributed File Systems.

Map reduce framework provides a scalable data intensive computing in a cloud. The performance and system scalability in a dfs are attained by replica synchronization (RS). Replica synchronization needs to make sure the changes in the data send to one replica and received by other relevant replicas only. But in RS, updates are made to all the relevant replicas as many times as the number of write request leading to increase in bottlenecks. Adaptive RS is used over normal RS to improve the I/O throughput, communication bandwidth and performance of DFS which split the large data into chunks[9]. The chunk list consist list of all the information about the replicas which belong to the same chunk file. Adaptive replica synchronization aims to improve the feasibility, efficiency and applicability when compared to the algorithms.

J. Research on Scheduling Scheme For Hadoop clusters.

Map reduce framework has been employed to develop a wide variety of data intensive applications in large scale systems. Three factors makes predictive scheduling regardable and possible

- The underutilization of CPU processors.
- The growing importance of Map reduce performance
- The Hadoop file distribution model offers data storage information.
- The interaction between master node and slave node.

Predictive scheduling mechanism increases the performance of MR by arranging the task in the system[10]

II. DESIGN ISSUES

S.no	Paper name	Highlight
1	Data Distribution technique for multi cloud storage in cloud computing	Single point failure
2	Distributed Data clustering in multi dimensional P2P network	Critical data distribution between the peers.

3	Optimal power allocation and load distribution for multiple heterogeneous multi-core server processor	Load distribution.
4	Semi distributed cloud computing system with load balancing algorithm	Extreme Stress on the system
5	Hierarchical peer-to-peer systems	To obtain optimal service.
6	PeerDB: A P2P system for distributed data sharing	Nodes are static
7	In Comparison between centralized and distributed cloud storage data center topology	Lack of through put
8	An optimal task selection scheme for Hadoop Scheduling	To reduce the response time
9	A novel approach for replica synchronization in hadoop system	Delay in processing speed.
10	Research on scheduling scheme for Hadoop Clusters	Predictive scheduling

Table 1: Design Issues

III. COMPARISON OF VARIOUS DISTRIBUTION TECHNIQUES

S.no	Centralized Approach	Semi-centralized Approach	Hierarchy	Peer to Peer
1.	Data-distribution	Data-distribution	Data-distributionas	Data distribution in
	Sequentially	simultaneously	per priority	a random way.
2.	Single point failure	Single point failure	Parent node failure stops data flowing to child note	No fault tolerance
3.	Denial of service	Flash crowd effect	Firewall issue	No firewall issue

Table 2: Comparison

IV. CONCLUSION

Data distribution techniques, processing of data and storage in Hadoop Distributed File System are reviewed. Even though different data distribution techniques were discussed they still carry demerits along with them. Hierarchical data distribution is considered over centralized and semi centralized approach but still it introduces latency and might even break the connection between parent and child node if any one of the VMs in the cloud fails. Hence forth an improvement over hierarchical approach is Peer to Peer (P2P) approach which has no latency and requires no special full time administrator. In spite of good features of P2P, setting up a P2P network is a tough task as the entire P2P is decentralized, security can also be breached because of wide network, data recovery is also cumbersome. Apart from the above stated points, P2P is the most widely used data distribution technique in corporate world due to its dynamic approach in processing of data. Finally, there is still a scope for further research on securing an optimal data distribution technique which will provide better security and also provide an ease to recover the crucial data in coming days.

REFERENCES

[1] B.Amarnath reddy, T.Rajashekar reddy "Effective data distribution technique for multi cloud storage in cloud computing". Internal journal of engineering research and application (IGERA) Vol. 2 september- October 2012

[2] Stepano lodi, Gian luca moro "Distributed data clustering in multi dimensional peer to peer networks" proceeding 21st Australian data base conference (ADC) 2010

[3] J.Cao,K.Li, I.Stojmenovic "Optimal power allocation and load distribution for multiple heterogenous multicore server processor", IEEE Transactions On Computers, Vol. 63, No. 1, JANUARY 2014

[4] Payal A Pawad, Prof. VT Gaikwad "Semi distributed cloud computing system with load balancing algorithm" International journal of computer science and information technology Vol.5 (3) 2014

[5] L. Garces-Erice, E. Biersack, K. W. Ross, P. A. Felber, and G. Urvoy-Keller, "Hierarchical peer-to-peer systems," ser. Euro-Par 2003 Parallel Processing Lecture Notes in Computer Science, New York, NY, USA: Springer, vol. 2790, 2003, pp. 1230–1239

[6] Wee siong ng, Beng chin ooi "PeerDB: A P2P system for distributed data sharing" National university of Singapore, Conference on data Mining systems, Vol 3 2011.

[7] Maurice, Bolhuis "In Comparison between centralized and distributed cloud storage data center topology" University of twente, 2012.

[8] S.Suresh,N.P.Gopalan,"An Optimal Task Selection for Hadoop Scheduling", 2014 International Conference on future Information Engineering, IERI Procedia 10 (2014) 70 – 75

[9] J. Vini, Rachel nallathambi, C R Rene robin "A novel approach for replica synchronization in Hadoop system", 2nd Internationa l symposium on Big data and cloud computing at Elsevier

[10] J.Xie, F.Meng, H.Wang, H.Pan, J.Cheng, X.Qin, "Research on scheduling scheme for Hadoop clusters", International Conference on Computational Science, ICCS 2013, Procedia Computer Science 18 (2013) 2468 – 24.

