International Journal of Pure and Applied Mathematics
Volume 115 No. 6 2017, 629-635
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

TRAFFIC AWARE PARTITIONING AND AGGREGATION USING MAP REDUCE WITH DATA SECURITY

H.Sowmya, Dr P.Perumal
[1,2]Department Of Computer science, Sri Ramakrishna Engineering College,Coimbatore
[1]haldoraihn@gmail.com
[2]perumalp@srec.ac.in

Abstract: Cloud computing which is one of the new and an emerging concepts that can provide a lot of resources for processing the big data, the most important problem is to handle the data from various locations and also some sort of security are needed if the data are very much sensitive, the solution which are proposed for the problem is to handle all the data using map reduce and thus the security is provided by using blow fish algorithm

Keywords: MapReduce,Hadoop,Virtualization

1. Introduction

MapReduce which has an two major tasks Map Task and Reduce Task which executes as an two phases, these are generally used for managing an large scale datasets in an distributed clusters, in this project the Exisisting system uses an hash based functions which are generally used for partioning all the intermediate tasks, but the proposed system deals with the distributed algorithm for dealing with the large scale problems and thus the security are provided for sensitive data, sensitive data's are those that the datasets from GPS location of an user and the datasets from various CCTv's which needs an security, the security is provided by using blowfish algorithm. MapReduce has a recent framework for handling a large a large amount of data because they have simple programming model and also they consists of automatic management which are used for parallel execution. MapReduce consists of open source implementations which are well liked by various social web chat media like twitter, Facebook. The big data has various applications, such as an cloud computing, data mining and bioinformatics, in map phase all the map tasks are tend to be launched in parallel which can convert the original input data into an intermediate data which has an form of key and value pairs, these pairs are stored in the machines which has an multiple data partitions, in the reduce phase each data partitions are allotted to the reduce phase once all the map tasks are to be fetched into an reduce phase then the desired result are generated when an intermediate data are shuffled according to the hash function can lead to an large network traffic because these can ignore the network topology and the size of the data that are associated with each key. The word Big Data usually tells about the various methodologies that are used for storing, distributing capturing, managing and analyzing large size datasets with which it has high velocity, high volume and also different structures. Big data can have different forms like structured, unstructured or semi structured data which can result in incapability methods. The Data which are originated from various web sources can reach the system at various processing rates, in order to process all those data its very expensive and it also takes more time, the most efficient way to process the data parallelism concepts are used, big data has its data in the form of complex structure it requires more and more new technologies and architecture every day, Hadoop is one of the core platform which are used for structuring and analyzing big data, hadoop generally solves the problem by making analytics, hadoop which is one of the open source software that allows the distributed processing on various large data sets across various clusters on an commodity server. The major characteristics of big data are volume, variety and velocity which means the size of the data and rate of growth of data makes them very difficult to be captured, managed and analyze the by using the conventional technologies, virtualization and cloud computing is one of the best solutions for managing large amount of data? The cloud which is a new technology can offer different level of scalability, flexibility and availability, this can offer access to the data very easily from various location, these contains user defined and system designed meta data, Sensors are those that can gather information about the physical environment but the resources gathered by them requires some form of storage, cloud computing provides an good storage and data's that are stored are not missed, storing, processing, accessing of data can be done efficiently by using hadoop and virtualization.

The big data strategy was implemented by ICT strategy, which provides data analytic capability for improving the policy and service delivery, these strategies are used by the senior executives and the business program managers

2. Related Works

A. *Cloud Blast*

Cloud blast is nothing but a combination of mapreduce and virtualization which are used for bioinformatics application [2], these contain distributed resources, this approach uses a mapreduce paradigm for parallelizing the tools and also used for managing the execution. These Implementation consists of hadoop , mapreduce and virtualization for deploying commonly used bio informatics tool called NCBI BLAST, these can take unacceptably long times that if they are implemented on small systems

B. *Iterative Mapreduce For Largescale Machine Learning*

Big data is the one which operates at large clouds; the mapreduce doesn't provide any iteration but some program can be done to provide the iteration [3], this can be implemented by looping an rst-class construct which in turn propose the extension of an mapreduce paradigm called as the iterative map reduce

C. *Mapreuce Meets Wider Variety Of Application*

Most of the industries are building data centre scale like computer to overcome the high storage and process the demands that are generated by data intensive and the compute intensive applications [4], the reason for building data centres is because the data that are generated by the digital media has been doubling every year.

D. *Introducing Map-Reduce To High End Computing*

Petabytes of data sets are the frontiers of High End computing, these can process and generate peta bytes of data [5], these application comes from an variety of application such as an cosmology, bioinformatics

E. *Mapreduce-Simplified Data Processing On Clusters*

In this technology the map phase takes input in the form of key value pair process them into an intermediate value pairs of output [6] , the function of the same key values are merged

F. *Optimizing The Data Shuffling In An Data Parallel Mode*

A data- parallel computation generally involves multiple parallel computation phases which are defined by using a user defined functions [7]

G. *Scheduling The Maptask In Map Reduce By Using Data Locality*

Generally a mapreduce cluster consist of ten thousands of machines, these kind of data are organized by using an distributed file system [8], like hadoop distributed file systems which divides the large datasets into data chunks and store them as replicas on each chunk

3. Exisisting System

The Exisisting system consists of MapReduce algorithms for solving the entity relationship problem which consists of huge collection of entities with multiple keys, this system consists of combination based locking and load balanced matching strategies. The combination based locking strategies utilizes multiple keys and sort out the entity pairs that are used for future matching, the load balanced strategies evenly distributes all the similar computation for each reducer in an matching step, the effectiveness and scalability for MapReduce based implementation on an complex intensive task are calculated by even distribution of map tasks and the reduce tasks, Redistribution of data can be used for achieving load balancing among all the reduce tasks that are to be2 executed in parallel. The reason for using the distribution of similarity computation in a reducer step is to remove the bottle neck of skewed matching computations. In order to solve the complex problem generally there are two approaches such as skew handling and load balancing are used, these approaches uses an blocking techniques in order to Reduce the search space, the evaluation that are done on an real cloud infrastructure can display the value and effectiveness of an proposed load balancing strategies, the disadvantages of the Exisisting work focuses only on load balancing which occurs at the reduce tasks, it actually ignores the network traffic that has been occurred during an shuffle phase ,these can be only applied to the network topology with an server directly linked to other server directly linked to other servers, which has limited use in the current system scenario, these approaches are traffic efficient, hash based works can give unstable work load

4. Proposed System

The proposed system uses an distributed algorithm that are used for Big data Applications, they decompose the original Large Scale Problems into an several Sub problems, that are solved using parallelizing techniques, the system design has an online Algorithm the basic idea is to Post Pone the operation, when the cumulative traffic cost exceeds the threshold, the network topology generally consists of three tier Architecture, an access tier aggregation , the system that stakes out the network traffic reduction with an mapreduce job by using traffic aware, intermediate data partioning and data aggregation along multiple tasks.

It offers computers in the form of physical and virtual machines. For implementing the parallelization concepts virtual cluster are used , virtual machine configuration, virtual machine placement, virtual machine consolidation can be used for achieving high throughput, fast response, balancing the load and low cost.

The network topology for an virtual cluster has its impact on the application that are running because of the physical nodes that are located can be linked in different ways, one of the most special architecture called as hierarchical network used where two or more physical nodes can lie at different local area networks, the distributed algorithm can solve problem on multiple machines in an parallel manner.

The basic idea is to break down the original larger problem into a several sub problems these sub problems are coordinated by an high level master problem, these sub problems can run concurrently on different parts of the independent processors.

These algorithms are traffic efficient, the data partioning and aggregation can be performed in a dynamic manner, and the network traffic costs are highly reduced. This is one of the new schemes that can exploit both aggregator placement and also the traffic aware, each and every reduce task can aggregate the related data partition that belong to it and store them in a distributed file system.

Map Reduce has two major phases Map Phase and Reduce Phase, the sub problems that are generated by using distributed algorithm are given has an input to the map phase to generate the key value pair the same key value pair are fetched to the same reducer phase to obtain the output

Table 1. Existing Vs Proposed

The below Table I: describes about the various advantages over the Exisisting and proposed system

MAP REDUCE	MILP
These can reduce cost of data traffic under the constraint of with and without network	These can effectively reduce the data traffic only under the constraint of network
Map Reduce resource allocation system, to enhance the performance of Map Reduce jobs in the cloud by locating intermediate data to the local machines or close-by physical machines. Sup problem may be raised.	It solves the problem by distributing the data on various system through parallelization techniques
Map Reduce is a programming model based on two Primitives: map function and reduce function. The input is divided into chunks that are Assigned to map tasks.	It develops a combination based locking, load balancing strategies, these uses it ignores network traffic during shuffling of hash based inputs
The input data are divided into independent chunks that are processed by map tasks in parallel.	It Dynamically adjusts data partition and aggregation during the execution of map and reduces tasks.i.e chunks are processed dynamically.

Blowfish Algorithm is a form of an encryption algorithm which provides the network security, this algorithm can be used as an replacement for data encryption standard or any IDEA algorithm, it is one of the symmetric key block cipher .which uses an Variable length key from 32 bits to 448 bits which it makes it useful for domestic and commercial use. This Algorithm was designed in the year 1993 , this algorithm was designed in such an way to provide an alternative to Exisisting algorithms such as AES, DES, 3DES, the major advantage of blow fish algorithm are fast, compact, simple, and highly secure, its Fast because it can encrypt data that are present on an 32bit Microprocessor which can process at an rate of 26 clock cycles per byte this algorithm can run in less than 5k memory, these are suitable for application where the key doesn't change often, it encrypts block of data of 64 bits at an time, it generally follows an fiestel network, it consists of two different parts namely key expansion and Data Encryption. It can convert a key of 448 bits into several Sub Key which totally consists of 4168 bytes, it uses large number of sub keys, The P- array consist of 18 to 32 bits sub keys from p1…P18 four 32bit boxes consists of 256 Entries.

Initialize the first P-array to one of the four S-boxes these strings can consists of hexa decimal digits. XOR the P array Pn with the first 32 bits, XOR the Second P array Pr with the second 32 bits, repeat the cycle until an entire P-array has been XOR using key bits, there are total of 521 iterations are required for generating the required sub keys

```
STEP 1: Divide X into 32 bit halves XL, XR
STEP 2: For i=1 to 16
STEP 3: XL= XL XOR Pi
STEP 4: XR= F (XL) XOR XR
STEP 5: Swap XL and XR
STEP 6: XR= XR XOR P17
STEP 7: XL= XL XOR P18
STEP 8: Recombine XL and XR
```

The above describes about various steps that are used for performing the blow fish Algorithm

A. Client

The Client which acts like a station that can receive data from the server its operation is to send the request to another computer these may be located nearer another or they may be far from one another

B. Server

The servers are generally used for sending the request to the client, services generally provides many of the functionalities called as services, these are used for sharing the data or resources among various clients or they are used for performing the computation and calculation on the client, an single server can act on an multiple client similarly single client can act on multiple servers.

C. Data Uploading

The Master acts by considering the data locality and assigning each task to individual workers, data's with same key value pair are assigned to same partition

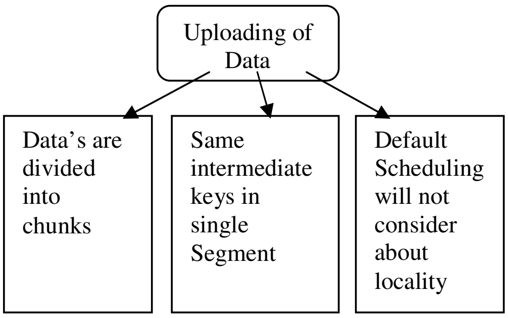
Uploading of Data
Data's are divided into chunks
Same intermediate keys in single Segment
Default Scheduling will not consider about locality

Figure 1. Uploading Data Sets to the Workers

D. Segmentation:

Consider a System which consists of typical Map Reduce job on a Large Clusters which consists of N set of machines. Let XY be the two distances between Two distances for the two machines x and y which generally represents the cost for delivering an unit of data, when the job tend to be executed, mainly two types of tasks are created, namely map tasks and reduce tasks, the map task and reduce tasks are denoted by M and R. These are already placed on the machines, the input that are present on the machines are divided into independent chunks these are processed by map Tasks in an parallel fashion, the results that are obtained are in the form of an key value pair which are sorted in the necessary framework which are fetched into the reduce tasks to produce the final output.

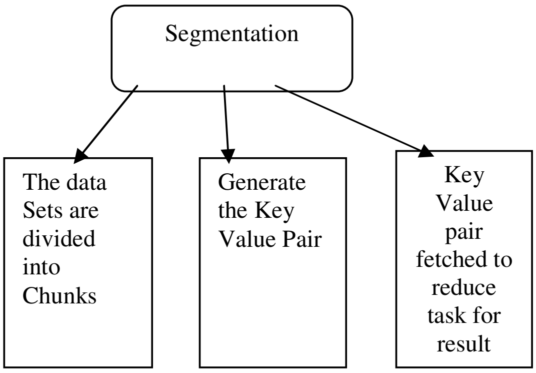
Segmentation
The data Sets are divided into Chunks
Generate the Key Value Pair
Key Value pair fetched to reduce task for result

Figure 2. Segmentation of Data Sets

E. Task Assignment

The Access Link tier is made up of Ethernet Switches Connecting a Set of Virtual Machines; the Switches are connected to Ethernet through Aggregation switches which in turn connected to a layer of core switches

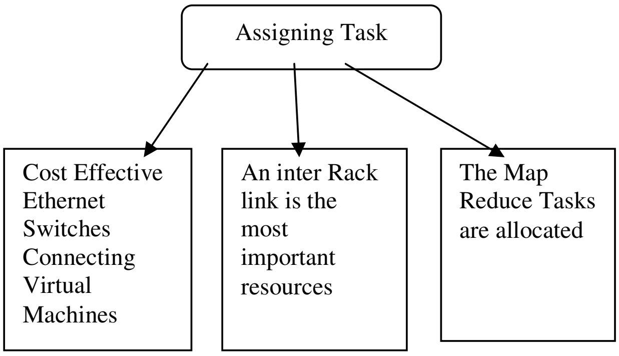
Assigning Task
Cost Effective Ethernet Switches Connecting Virtual Machines
An inter Rack link is the most important resources
The Map Reduce Tasks are allocated

Figure 3. Assigning Of data Sets

The System will divide the execution of a MapReduce job Into Several Minutes or in a length of an hour

F.Evaluation Process

The System will evaluate the performance of proposed algorithm under the online cases by comparing with them to other two algorithms OHRA and OHNA.

G. Secured Data Storing

Cryptography which is one of the valuable technique for protecting data in the data center. In this project the blowfish algorithm is used, Blow fish algorithm which is one of the incredibly fast cipher, it has relatively simple structure and it also very effective, blowfish algorithm generates key which are very robust like cereal box decoder ring which is the benefit for security which can increase the speed of computer processing. Blowfish can be able to create a key, which are most difficult to hack the data which are stored at the data center.

5. Conclusion And Future Work

This Paper deals with the two major problems of Big Data, one issue is to handle large data sets and the other problem is to provide the security to those data, handling of large data sets are done by MapReduce and Hadoop and the security is provided by using blowfish algorithm.

In this paper the Blow Fish Algorithm are used for protecting the data, there are many other techniques that can be used as an enhancement and can produce effective results than Blow fish Algorithm. Some of the algorithms are RSA, Toe Fish Algorithms. Even Biometric Techniques can also be used. Biometric Techniques like palm print; thumb print can also be used.

6. Results And Analysis

This paper has discussed about the mapreduce technology in handling large data sets which has its size in petabytes and an graph has been drawn according to the analysis that has been got during the reduce phase.

The below Table named Table 1 describes about how an Mapper and an Reducer works before aggregation and after aggregation here the nodes are represented by using the terminal “N” and the mappers are represented using terminal “M

Mapper	N1		N2	N3	N4	N5
size Of data	M1	M2	_	M2	M3	M4
Data Before Aggregation	174.51 M	177.92 M	_	176.21 M	177.17 M	176.19 M
Data After Aggregation	139.66 M		_	176.21 M	98.23 M	94.7 M
Reducer (R1)	81.17M			96.17 M	98.23 M	82.02M
Reducer(R2)	58.49 M			80.04 M	78.94 M	
PRACTICAL COST	2690.48M					
SIMULATED COST	2690.48M					

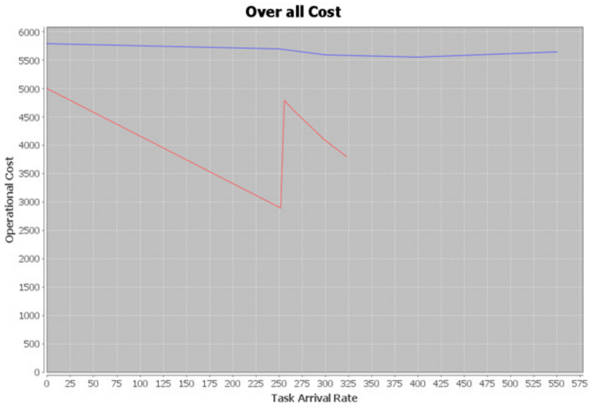

Figure 4. Task Arrival

The above Figure Describes about the task arrival rate and operational cost calculated based on the Joint and Non Joint Optimization

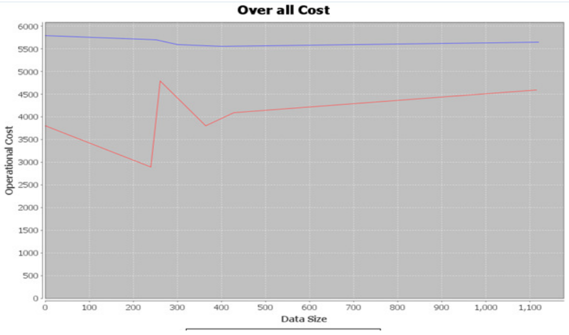

Figure 5. Data Size

The above Figure describes about the data Size and operational cost calculated based on joint and non joint optimization

References

[1] Huan kee, Peng Li Song Guo IEEE and Minyui Gyo "On traffic Aware Partioning and Aggregation using MapReduce", IEEE Transactions on Parallel and Distributed Systems Vol 27, No 3 March 2016.

[2] A. Matsunaga, M. Tsugawa, and J. Fortes, "Cloud blast: Combining mapreduce and Virtualization on Distributed resources for Bioinformatics applications," in science, 2008.EScience'08.IEEE Fourth International Conference, 2008, pp. 222–229.

[3] J. Rosen, N. Polyzotis, V. Borkar, Y. Bu, M. J. Carey, M. Weimer, T. Condie, and R.Ramakrishnan, "Iterative mapreduce for large scale machine learning," arXiv preprint ArXiv: 1303.3517, 2013.

[4] S. Chen and S. W. Schlosser, "Map-reduce Meet wider varieties of applications," Intel Research Pittsburgh, Tech. Rep. IRP-TR-08-05, 2008.

[5] Introducing Map-Reduce to High End Computing Grant Mackey, Saba Sehrish, John Bent, Julio Lopez, Salman Habib, Jun Wang University of Central Florida

[6] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters, "Communications of the ACM, vol. 51, no. 1, pp.107–113, 2008.

[7] MapReduce: Simplified Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemawat

[8] "Optimizing Data Shuffling in Data-Parallel Computation by Understanding User-Defined Functions" Jiaxing Zhang, Hucheng Zhou, Rishan Chen, Xuepeng Fan, Zhenyu Guo, Haoxiang Lin, Jack Y. Li, Wei Lin Jingren Zhou, Lidong Zhou University of Science and Technology Georgia Institute of Technology

[9] Wang, K. Zhu, L. Ying, J. Tan, and L. Zhang, "Map task scheduling in map reduce with data Locality: Throughput and heavy traffic Optimality," in Proc. IEEE INFOCOM, 2013, pp. 1609–1617.

[10] H. Lv and H. Tang, "Machine learning Methods and their application research," IEEE Int IEEE Int. Symp. Intel. Info. Process. Trusted Comput. (IPTC), pp. 108–110, Oct. 2011

[11] Y. Wang, W. Wang, C. Ma, and D. Meng, "Z Put: A speedy data uploading, approach for The hadoop distributed file system," in Proc. IEEE Int. Conf. Cluster

[12] J. Wang, D. Crawl, I. Altintas, K. Tzoumas, And V. Markl, "Comparison Case study," in Proceedings of the Fourth International Workshop on Data Intensive Computing in The Clouds (Data Cloud), 2013.

[13] S.Venkataraman, E. Bodzsar, I. Roy, A. Young, and R. S. Schreiber, "Presto: Distributed Machine learning and graph Processing with sparse Matrices," in Proceedings of the 8th ACM European Conference on Computer Systems. ACM, 2013, pp. 197– 210.

[14] T. White, Hadoop: the definitive guide: the Definitive guide. " O'Reilly Media, Inc." 2009.

[15] K. Meena , S. Vinothini , T. Pallavi , R.Vasugi, "Surveillance Based Gcm Home Security System Using Object Motion Detection", International Innovative Research Journal of Engineering and Technology vol. 2, no. 1, pp. 44-47.