ISSN 2395-1621

Implementation of Cost Optimization Approach for Big Data Stream Processing in Hadoop Framework

[#1]Suhas Sunil Mohite, [#2]Vaibhav Rajendra Patil, [#3]Shital Sunil Pacharne
[#4]Anuja Kalidas Upasani

[1]suhasmohite33@gmail.com
[2]patil.v.r94@gmail.com
[3]shitalpacharne26@gmail.com
[4]upasani.anu30@gmail.com

[#1234]Deparment of Computer Engineering
JSPM's, ICOER, Wagholi, Pune.

ABSTRACT

Big Data contains large-volume, complex and growing data sets with multiple, autonomous sources. Big data processing is the explosive growth of demands on computation, storage, and communication in data centers, which hence incurs considerable operational expenditure to data center providers. Therefore, to minimize the cost is one of the issue for the upcoming big data. Using these three factors, i.e., task assignment, data placement and data routing, deeply influenced by the operational expenditure of geo distributed data centers. In this paper, we are ambitious to study the cost minimization for big data processing in distributed data centers.

Keyword: **Hadoop, Big data, distributed data center, Minimize cost**

ARTICLE INFO

Article History

Received: 26th April 2017

Received in revised form :

26th April 2017

Accepted: 30th April 2017

Published online :

30th April 2017

I. INTRODUCTION

Big data is an one of the emerging hot research topic because its mostly used in data center application in human society, such as government, climate, finance, and science. Currently, most research work on big data falls in data mining, machine learning, and data analysis. The name itself contains the meaning of data will be so big in large volume of both structured and unstructured data present [1]. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain [8]. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration. fig number 1 shows the definition of the big data and the important characteristics of big data.

Fig 1 Big Data Definition[7]

- Big Data is large amount and growing dataset with multiple sources.
- To minimize the time for processing is one of the issue for upcoming big Data.
- In this project, we are ambitious to study the time minimization for Big Data processing in Data centers.

II. BIG DATA

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data. Figure No. 2 gives Layered Architecture of Big Data System. It can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to bottom.

Figure 2: Layered Architecture of Big Data System [8]

III. HADOOP SYSTEM

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper.

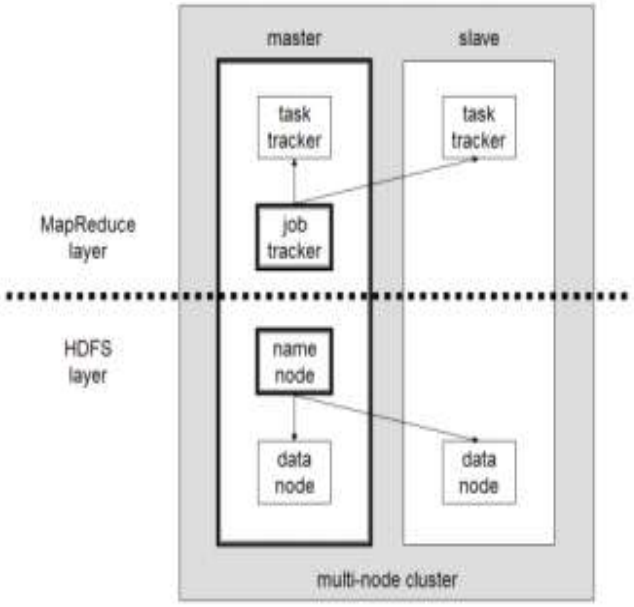

Fig 3. Hadoop Architecture[8]

A. HDFS Architecture

Hadoop includes a fault tolerant storage system called the Hadoop Distributed File System, or HDFS [4]. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them.

Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.

B. MapReduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel.

From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied.

IV. PROPOSED SYSTEM

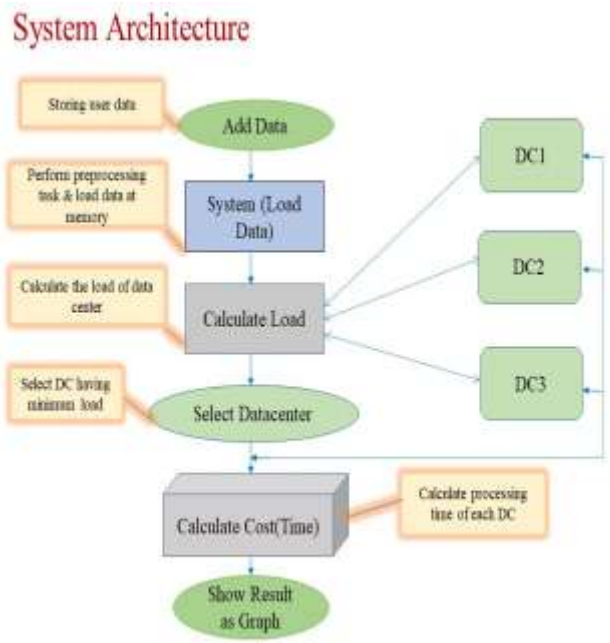

Fig. 4. System architecture

We are the first to consider the cost minimization problem of big data processing with joint consideration of data placement, task assignment and data routing. To describe the rate-constrained computation and transmission in big data processing process resize data centers to achieve the operation cost minimization goal [6].

To deal with the high computational complexity of solving MINLP, we linearize it as a mixed-integer linear programming (MILP) problem, which can be solved using commercial solver. Through extensive numerical studies, we show the high efficiency of our proposed joint-optimization based algorithm. We first present the constraints of data and task placement, remote data loading, and QoS. Then, we give the complete formulation of the cost minimization problem in a mixed-integer nonlinear programming form.

The two dimensional Markov chain process describe the rate-constrained computation and transmission in big data processing. And derive the expected task completion time in closed form. Formulate the cost minimization problem based on the closed-form expression in a form of mixed integer nonlinear programming (MINLP). Linearize it as a mixed-integer linear programming (MILP) problem for solving the complexity of MINLP.

Module:

Date Upload:

- Select the big data and stored into the hadoop environment for the for performing map reduce on hadoop.
- The data should be loaded into the VM server location.
- After Uploading the file the data segmentation is performed for further process.

Segmentation:

- Packet segmentation improves network performance by splitting the packets in received Ethernet frames into separate buffers.
- Packet segmentation may be responsible for splitting one into multiple so that reliable transmission of each one can be performed individually.
- The packet processing system is specifically designed for dealing with the network traffic.

Task Assignment:

- The Data Center should be selected according to computation and storage capacity of servers reside in the data center.
- Identification of Data Center is important matter for minimizing operational expenditure of servers reside in the each data centers.
- Task is assigned to data center according to Memory requirement for effectively processing of data.

Data Loading:

- A Data Placement on the servers and the amount of load capacity assigned to each file copy so as to minimize the communication cost while ensuring the user experience.
- Joint optimization scheme that simultaneously optimizes virtual machine (VM) placement and network flow routing to maximize energy savings.

Processing of Task:

- The high computational server should not process the low population of data chunk.
- Because it increases the operational expenditure of server, wastage of storage and transmission cost.

V. SYSTEM ANALYSIS

We have created system in java. Data is processing on hadoop environment windows system. We have created a java application with local server.java application that communicates with local server and Trustee Server using REST API. We have uploaded text document on hadoop. We have evaluated time, speedup, automatically processing required for uploaded dataset. Here we also calculate the load each datacenter shown analysis purpose.

Advantages:

- Cost for high computational data is minimized.
- Reduce the system operation increases system reliability.

- Energy Consumption is minimized.
- We can Jointly Optimize the problem of Task Assignment, Data Placement and Data Movement.
- The Task Completion time is minimized.

VI. RESULT

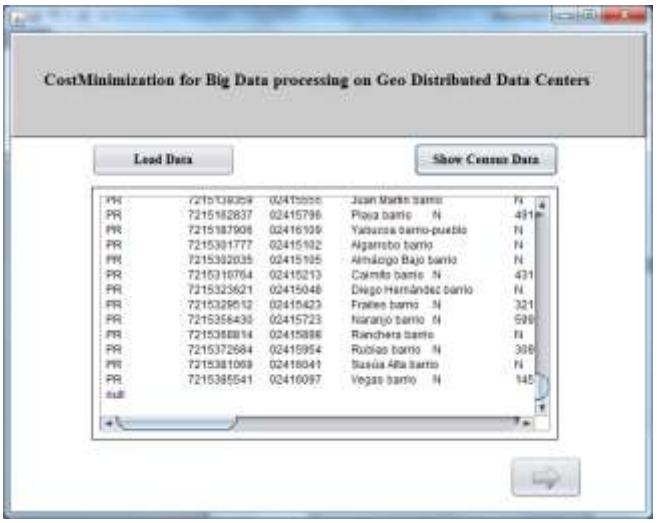

Fig 5. Load dataset on hadoop

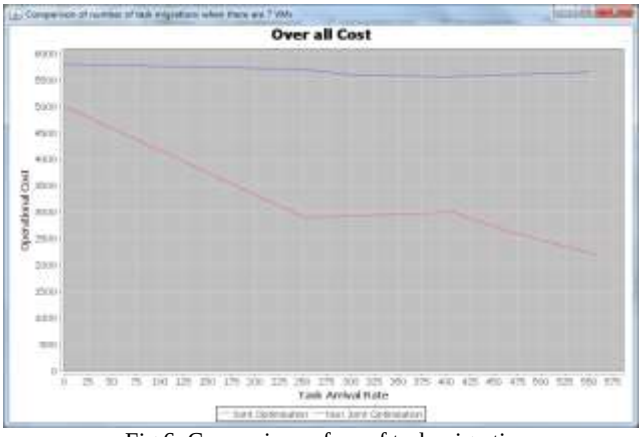

Fig 6. Comparison of no of task migration

VII. CONCLUSION

In this paper we study the distributed data centers issues. We jointly study the data placement, data center resizing and data routing to reduce the operational cost in geo distributed datacenters for big data processing. And we minimize the cost of data center.

ACKNOWLEDGEMENT

We wish to express my profound thanks to all who helped us directly or indirectly in making this paper. Finally we wish to thank to all our friends and well-wishers who supported us in completing this paper successfully we especially grateful to our guide for him time to time, very much needed, valuable guidance. Without the full support and cheerful encouragement of my guide, the paper would not have been completed on time.

REFERENCES

[1] Lin Gu, Student Member, IEEE, Deze Zeng, Member, IEEE, Song Guo, Senior Member, IEEE, Yong Xiang, Senior Member, IEEE, Jiankun Hu, Member, IEEE A General Communication Cost OptimizationFramework for Big Data Stream Processing in Geo-distributed Data Centers

[2] G. Lee, J. Lin, C. Liu, A. Lorek, and D. Ryaboy, "The Unified Logging Infrastructure for Data Analytics at Twitter," in Proceedings of Very Large Data Base Endow., vol. 5, no. 12, pp. 1771–1780, 2012.

[3] G. Mishne, J. Dalton, Z. Li, A. Sharma, and J. Lin, "Fast data in the era of big data: Twitter's real-time related query suggestion architecture," in Proceedings of the 2013 International Conference on Management of Data, ACM, pp. 1147–1158, 2013.

[4] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley,M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster com-puting," in Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, USENIX Association, pp. 2–2, 2012.

[5] Z. Zhang, M. Zhang, A. G. Greenberg, Y. C. Hu, R. Mahajan, and B. Christian, "Optimizing Cost and Performance in Online Service Provider Networks," in Proceedings of the USENIX Network System Design and Implementation, USENIX Association, pp. 33–48, 2010.

[6] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The Cost of a Cloud: Research Problems in Data Center Networks,"SIGCOMM Comput. Commun. Rev., vol. 39, no. 1, pp. 68–73, 2008.

[7] Sarannia, N. Padmapriya, Asst prof, "Survey on Big Data Processing in Geo Distributed Data Centers" Volume 4, Issue 11, November 2014.

[8] Harshawardhan S. Bhosale "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-315.

[9] H. C. Zhao, C. H. Xia, Z. Liu, and D. Towsley, "A Unified Modeling Framework for Distributed Resource Allocation of General Fork and Join Processing Networks," in Proceedings of International Conference on Measurement and Modeling of Computer Systems, ACM, pp. 299–310, 2010.

[10] K. LaCurts, S. Deng, A. Goyal, and H. Balakrishnan, "Choreo: network-aware task placement for cloud applications," in Proceed-ings of Internet Measurement Conference, ACM, pp. 191–204, 2013.

[11] B. Chinoy and H.-W. Braun, "The National Science Foundation Network," Technical Report GA-A21029, SDSC, 1992.

[12] N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, "Inter-datacenter Bulk Transfers with Netstitcher," in Proceedings of Conference of the Special Interest Group on Data Communication, ACM, pp. 74–85, 2011.

[13] Y. Feng, B. Li, and B. Li, "Postcard: Minimizing costs on inter-datacenter traffic with store-and-forward," in proceedings of In-ternational Conference on Distributed Computing Systems Workshops, IEEE, pp. 43–50, 2012.

[14] L. Liu, H. Wang, X. Liu, X. Jin, W. B. He, Q. B. Wang, and Y. Chen, "GreenCloud: a New Architecture for Green Data Center," in Proceedings of the 6th International Conference Industry Session on Autonomic Computing and Communications Industry Session, ACM, pp. 29–38, 2009.

[15] R. Cohen, L. Lewin-Eytan, J. Naor, and D. Raz, "Almost Optimal Virtual Machine Placement for Traffic Intense Data Centers," in Proceedings of International Conference on Computer Communications, IEEE, pp. 355–359, 2013.

[16] X. Meng, V. Pappas, and L. Zhang, "Improving the scalabil-ity of data center networks with traffic-aware virtual machine placement," in IProceedings of International Conference on Computer Communications, IEEE, pp. 1–9, 2010.

