ISSN 2395-1621

A Survey 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 geo-distributed data centers.

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

ARTICLE INFO

Article History

Received: 14th November 2016

Received in revised form :

14th November 2016

Accepted: 16th November 2016

Published online :

16th November 2016

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. 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. 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

© 2016, IERJ All Rights Reserved

issue for upcoming big Data.

- In this project, we are ambitious to study the time minimization for Big Data processing in Data centers.

II. LITERATURE SURVEY

G. Lee, J. Lin, C. Liu, A. Lorek, and D. Ryaboy, "The Unified Log-ging Infrastructure for Data Analytics at Twitter," in Proceedings of Very Large Data Base Endow., vol. 5, no. 12, pp. 1771–1780, 2012.
Analysis: Analysis Twitter's production logging infrastructure and its evolution from application-specific logging to a unified "client events" log format, where messages are captured in common, well-formatted, flexible Thrift messages.
Finding: This approach afforded significant flexibility and allowed for very fast application logging development.

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.
Analysis: He present the architecture behind Twitter's real-time related query suggestion and spelling correction service.
Finding: Build a generic data processing platform capable of handling both "big data" and "fast data.

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.
Analysis: He present Resilient Distributed Datasets (RDDs), a distributed memory abstraction that lets programmers perform in-memory computations on large clusters in a fault-tolerant manner.
Finding: RDD benefits distributed memory abstraction, we compare them against distributed shared memory (DSM)

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.
Analysis: He present a method to jointly optimize the cost and the performance of delivering traffic from an online service provider (OSP) network to its users.
Finding: He find that by OSP can reduce its traffic cost by 40% without any increase in path latency and with acceptably low overheads.

P. Bod´ ık, I. Menache, M. Chowdhury, P. Mani, D. A. Maltz, and I. Stoica, "Surviving Failures in Bandwidth-constrained Datacen-ters," in Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, ACM, pp. 431–442, 2012.
Analysis: He propose and evaluate a novel optimization framework that achieves both high fault tolerance and significantly reduces bandwidth usage in the network core by exploiting the skewness in the observed communication patterns.
Finding: explore the tradeoff between improving fault tolerance and reducing bandwidth usage.

K. yin Chen, Y. Xu, K. Xi, and H. Chao, "Intelligent virtual machine placement for cost efficiency in geo-distributed cloud systems," in Proceedings of International Conference on Communications, IEEE, pp. 3498–3503, 2013.
Analysis: In the CAVP problem formulation, we capture the intrinsic trade-off between electricity cost and WAN communication cost, as well as the electricity price diversity at different geographic locations.
Finding: The results show that the potential of performance improvement is significant and partial-optimizing heuristics.

III. 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.

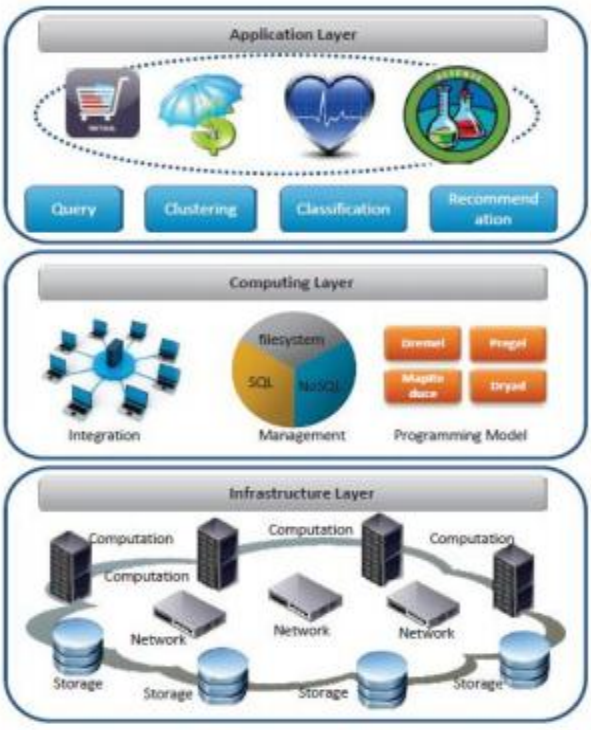

Figure 2: Layered Architecture of Big Data System [8]

© 2016, IERJ All Rights Reserved

IV. 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.

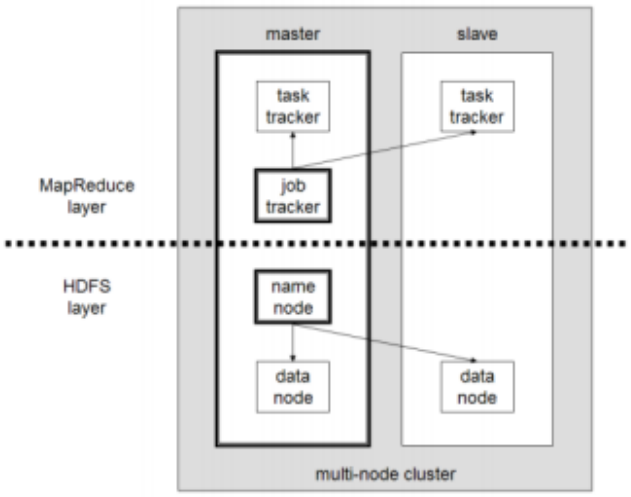

Fig 3. Hadoop Architecture[8]

A. HDFS Architecture

Hadoop includes a fault - tolerant storage system called the Hadoop Distributed File System, or HDFS. 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.

V. CONCLUSION

In this paper we study the geo 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.

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 Log-ging 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.

© 2016, IERJ All Rights Reserved