Cost Reduction Approach in Geo-Distributed Data Center for processing Big Data

K.Anuradha*, L.Devi, M.V.D.V.Prasad
Faculty, Dept. of Computers, Sir C R Reddy College (Autonomous), ELURU, India.
*mail id: anu.komati@gmail.com

ABSTRACT

Now a day's large amount of data storage is a heavy burden on data computation, storage and communication. Due to the demand in data storage, the data can be partitioned and geographically distributed known as geo distributed data center. Operational expenditure is a big challenge for data centers for providing Communication. Cost reduction has become an emergent issue for the upcoming big data era. One of the main features of big data services is the tight coupling between data and computation as computation tasks can be conducted only when the corresponding data are available. Task assignment, Data placement, and Data movement are the main factors- deeply influence the operational expenditure of data centers. In this paper, we proposed an approach to study the cost reduction by optimizing these three factors for big data services in geo-distributed data center. We provide an algorithm for the optimization of cost in the movement of the big data from one data center to another.

Keywords: *Big data, Cost reduction, Data center, Data storage, Geo-distributed data center*

I.INTRODUCTION

Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process data within a tolerable elapsed time. For example, the volume of data Facebook or Youtube need require it to collect and manage on a daily basis, can fall under the category of Big Data. However, Big Data is not only about scale and volume, it also involves one or more of the following aspects − Velocity, Variety, Volume, and Complexity.

Fig.1 Big Data Definition

Geo-distributed data centers are operated by many organizations such as Google and Amazon are the powerhouses behind many Internet-scale services.They are deployed across the Internet to provide better

latency and redundancy. These data centers run hundreds of or thousands of servers, so it consumes megawatts of power with massive carbon footprint, and also incur electricity bills of millions of dollars.

II. RELATED WORK

Raghavendra.P,Z.Wang[2] proposed the key challenges in data center environments are Power delivery, power consumption, and heat management. Propose using different power management strategy such as virtual machine controller and efficiency controller. Using these strategy to validate the power in data centers. A.Sivasubramanian,B.Urgaonkar et al[7]Proposed the Datacenter power consumption has one of the a significant impact on both its recurring electricity bill (Op-ex) and one-time construction costs (Cap-ex). They develop peak reduction algorithms that combine the UPS battery knob with existing throttling based techniques for minimizing power costs in datacenter. Sarannia, N. Padmapriya[8] proposed to to study the cost minimization problem via a joint optimization of these three factors for big data processing in geo-distributed data centers. Proposed using n-dimensional markov chain and procure average task completion time. A. Dhineshkumar, M.Sakthivel[9] proposed to to study the cost minimization problem and optimization of these factors for big data services in geo-distributed data centers. To describe the task completion time with the consideration of both data transmission and computation.

III. SYSTEM MODEL

Geo distributed data center means many data centers are geo graphically distributed and connected through the WAN environment. In recently many organizations move to this geo distributed data center. Because they stored large or massive volume of data. If they are using our own data center means only limited storage will be there so only many of them used this geo distributed data centers.

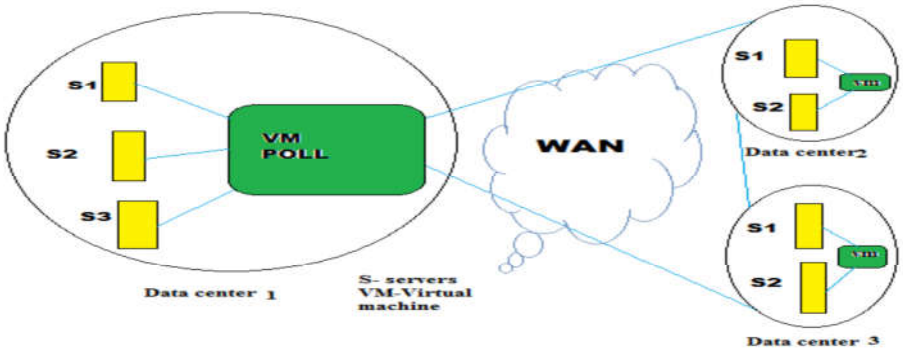


Fig.2 Geo distributed datacenters

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. Because of large explosion of data it is necessary to distribute data between data centers. Many large companies can use the data centers because of storage to maintain all the data that they process tera or peta bytes of data.

IV.EXISTING SYSTEM

The existing routing strategy among data centers fails to exploit the link diversity ofdata center networks. Due to the storage and computation capacity constraints, not all tasks can be placed onto the same server, on which their corresponding data reside. The Quality-of-Service (QoS) of big data tasks has not been considered in existing work. Existing work on data center cost optimization, big data management or data placement mainly focuses on one or two factors. To deal with big data processing in geo-distributed data centers, we argue that it is essential to jointly consider data placement, task assignment and data flow routing in a systematical way. However, the existing routing strategy among data centers fails to exploit the link diversity of data center networks. Due to the storage and computation Cost Minimization for Big Data Processing in Geo Distributed Data Centers capacity constraints, not all tasks can be placed onto the same server, on which their corresponding data reside.

Disadvantages of Existing System:

- It does not have more flexibility.
- Resizing the data centers is also very difficult in cloud environment.
- Wastage of resources is occurred because of data locality.

V.PROPOSED SYSTEM

Map Reduce is a programming paradigm that runs in the background of Hadoop to provide scalability and easy data-processing solutions. Map Reduce is a programming model for writing applications that can process Big Data in parallel on multiple nodes. Map Reduce provides analytical capabilities for analyzing huge volumes of complex data. Traditional Enterprise Systems normally have a centralized server to store and process data. Traditional model is certainly not suitable to process huge volumes of scalable data and cannot be accommodated by standard database servers.

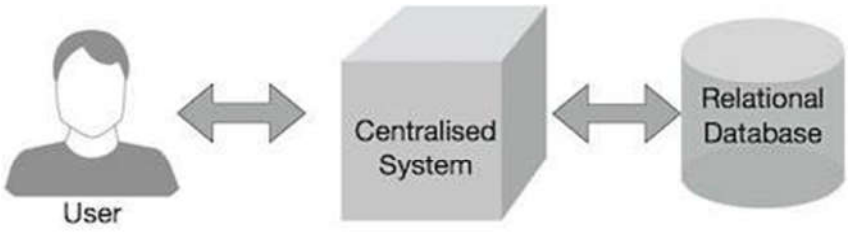


Fig.3 Centralised System

Google solved this bottleneck issue using an algorithm called Map Reduce. Map Reduce divides a task into small parts and assigns them to many computers. Later, the results are collected at one place and integrated to form the result dataset.

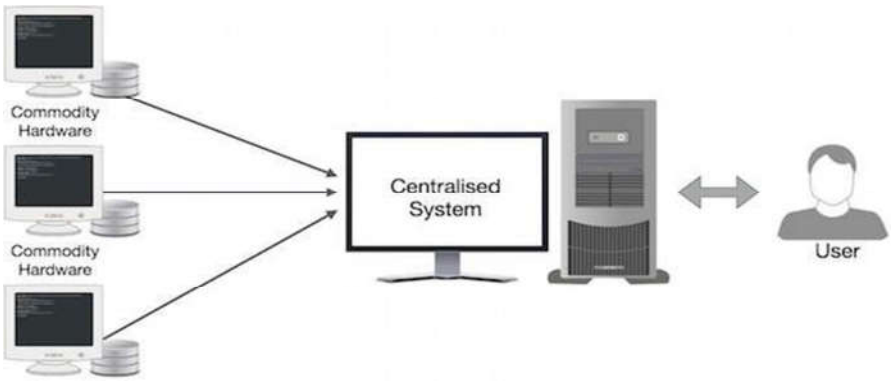


Fig.4 Integrating the Centralised System

The Map Reduce algorithm contains two important tasks, namely Map and Reduce. MapReduce divides a task into small parts and assigns them to many computers. Later, the results are collected at one place and integrated to form the result dataset.

The Map Reduce algorithm contains two important tasks, namely Map and Reduce.

- The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).
- The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.

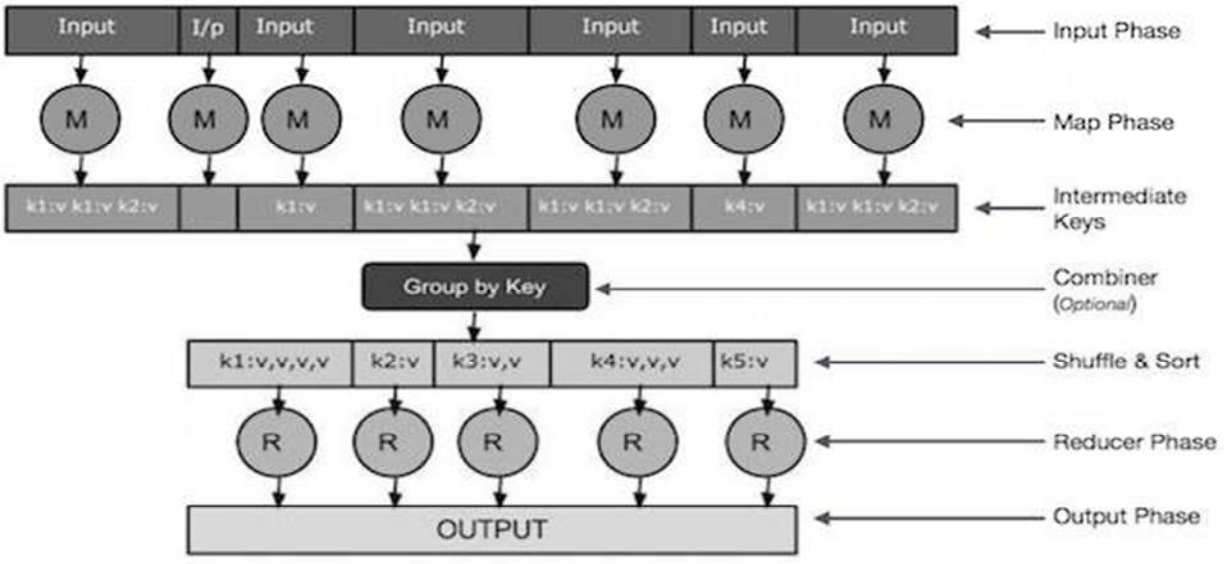
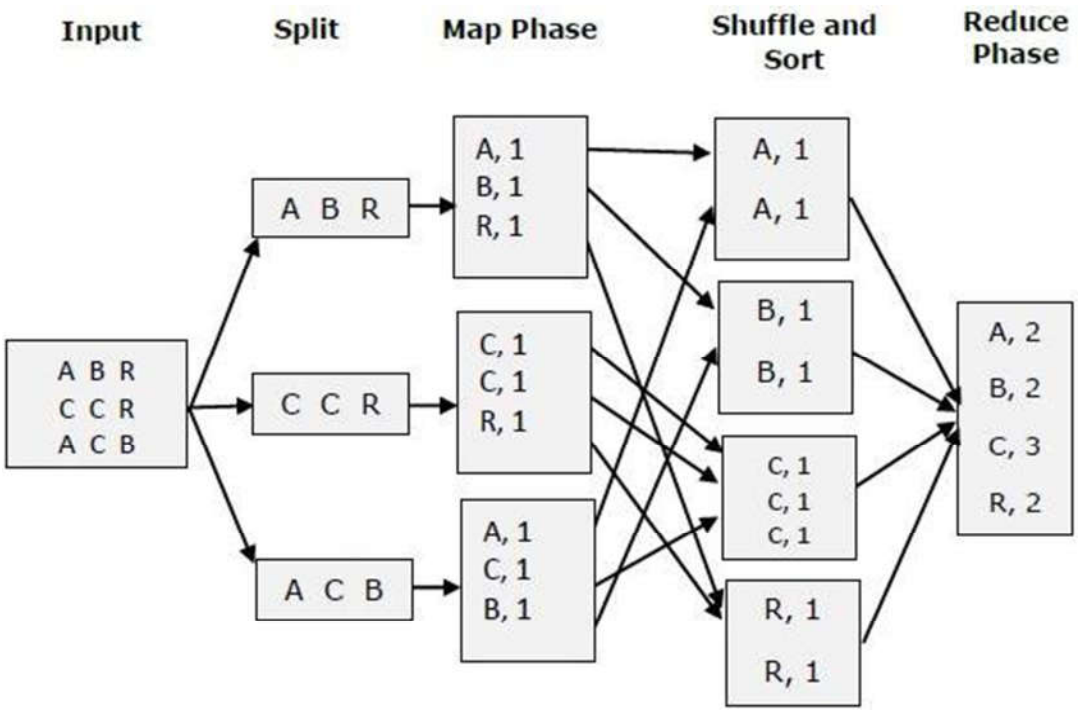


Fig.5 Phases in Map Reduce

- ***Input Phase*** − Here we have a Record Reader that translates each record in an input file and sends the parsed data to the mapper in the form of key-value pairs.
- ***Map*** − Map is a user-defined function, which takes a series of key-value pairs and proccesses each one of them to generate zero or more key-value pairs.

- ***Intermediate Keys*** − They key-value pairs generated by the mapper are known as intermediate keys.
- ***Combiner*** − A combiner is a type of local Reducer that groups similar data from the map phase into identifiable sets. It takes the intermediate keys from the mapper as input and applies a user-defined code to aggregate the values in a small scope of one mapper. It is not a part of the main MapReduce algorithm; it is optional.
- ***Shuffle and Sort*** − The Reducer task starts with the Shuffle and Sort step. It downloads the grouped key-value pairs onto the local machine, where the Reducer is running. The individual key-value pairs are sorted by key into a larger data list. The data list groups the equivalent keys together so that their values can be iterated easily in the Reducer task.
- ***Reducer*** − The Reducer takes the grouped key-value paired data as input and runs a Reducer function on each one of them. Here, the data can be aggregated, filtered, and combined in a number of ways, and it requires a wide range of processing. Once the execution is over, it gives zero or more key-value pairs to the final step.
- ***Output Phase*** − In the output phase, we have an output formatter that translates the final key-value pairs from the Reducer function and writes them onto a file using a record writer.

Let us try to understand the two tasks Map & Reduce with the help of a small diagram −



VI.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. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. It divides a task into small parts and assigns them to many computers then the results are collected at one place and integrated to form the result dataset.For this process the algorithm uses key & values in some phases to get the final phase.

REFERENCES

[1] "Data Center Locations," http://www.google.com/about/data centers/inside/locations/index.html.
[2] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu, "No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center," in *Proceedings of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 2008, pp. 48–59.
[3] www.google.com/bigdata /images/definition of big data
[4] J. Dean and S. Ghemawat. MapReduce*: simplified data processing on large clusters.OSDI*, 2004.
[5] B. Hu, N. Carvalho, L. Laera, and T. Matsutsuka, ―*Towards big linked data: a large-scale, distributed semantic data storage,"* in Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, ser. IIWAS '12. ACM, 2012, pp. 167–176.
[6] Cost Minimization for Big Data Processing in Geo-Distributed Data Centers Lin Gu, Student Member, IEEE, Deze Zeng, Member, IEEE, PengLi, Member, IEEEand Song Guo, Senior Member,IEEEDOI10.1109/TETC.2014.2310456, IEEE Transactions on Emerging Topics in Computing2014.
[7] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, ―Benefits and Limitations of Tapping Into Stored Energy for Datacenters,"in Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA). ACM, 2011, pp. 341–352.
[8] *Survey on Big Data Processing in Geo Distributed Data Centers* - Sarannia, N. Padmapriya, Asst prof Department of Computer Science & Engg., IFET College of Engineering, Villupuram, India Volume 4, Issue 11, November 2014
[9] *Big Data Processing of Data Services in Geo Distributed Data Centers Using Cost Minimization Implementation*-A. Dhineshkumar, M.Sakthivel Final Year MCA Student, VelTech HighTech Engineering College, Chennai, India Assistant Professor, Department of MCA, VelTech HighTech Engineering College, Chennai, India Vol. 3, Issue 3, March 2015