International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 02 | Feb-2016 www.irjet.net p-ISSN: 2395-0072

Efficient Cost Minimization for Big Data Processing

Pooja Gawale[1], Rutuja Jadhav[2], Shubhangi Kumavat[3], Pooja[4]

[1234] *Student, Department of Computer Engineering, MET's Bhujbal Knowledge City, Maharashtra, India*

Abstract - Big data is a collection of data sets which are large and complex and difficult to process by conventional tools. The growth in big data is found to be at a very high pace each day. As the rising trend of big data, it brings new challenge to the infrastructure and service providers because of its volume, velocity and variety. This heavy importunity to meet the restraints like storage, computation and communication in the data centers incurs high expenses in order to meet their needs. Hence, cost reduction is extremely necessary and major fact in this period of big data. The operational expense of the data centers is profoundly driven by three main factors, i.e., data loading, task assignment and data migration. In this paper, big data processing is characterized using two-dimensional Markov chain and the expected completion time is evaluated and further we formulate the problem as mixed non-linear programming (MNLP) based on closed-form expression to deal with high operational complexity. To reduce communication traffic and search effective solution we present solution which is based on weighted bloom filter (WBF), known as Distributed incomplete pattern matching (DI-matching). A bloom filter is a simple randomized organization of data that answers membership question with no false negative and small false positive probability Traditional bloom filter is generalized to weighted bloom filter incorporates the information on query frequencies and membership likelihood of elements into its most effective design.

Key Words: *Big data, Data center, Cost minimization, WBF, Data movement, Communication cost*

1. INTRODUCTION

The continuous increase in volume and detail of data captured by organization, for instance growth of social media, Internet of Things, as well as multimedia, has created an uncontrollable wave of data in either structured or unstructured format and it is referred as Big data. Big are characterized by three aspects: (a) data are existing in large numbers, (b) data cannot be classified into regular relational data bases and (c) data are produced, captured and processed rapidly. Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variation) and process of expanding (velocity) make them hard to confine, conduct, analyze by usual technology and tools. Therefore efficient management of processing and transmitting ids needed to avoid overwhelming data centers and networks with large volumes of data.

Data centers are spread at different locations. The oversized amount of data transmission leads to high communication cost. The main objective is to place the data chunks in the server, to distribute task onto servers, to move data between data centers. Two dimensional Markov chain model is used to process in the first data center and the output is passed as an input to the next data center. Likewise all the data chunks are processed in differently located distributed data center [3]. In this paper we present a solution to find target patterns over distributed environment which is based on well design weighted bloom filter (WBF) called, Distributed Incomplete pattern matching (DI-matching). Specifically, to reduce communication cost and ensure pattern matching in distributed incomplete patterns, use of WBF is done to encode query pattern and disseminate the encoded data to each data center [7]. We study the cost minimization problem of big data computation through combine optimization of data assignment, data placement and data migration in distributed data centers. To deal with high computational complexity of solving mixed non-linear programming (MNLP), linearize it [1].

2. RELATED WORK

To assume the challenges of successfully handling big data, many opinions have been suggested to recover the repository and processing cost. The advantage in manipulating big data is to secure and efficient data loading. It is reported that use of flexibility in data loading policy to increase efficiency in data centers[6]. Data center resizing and data placement are those techniques that have catch lots of attention. From then there has been considerable amount of work being carried on to minimize energy expenditure in the server. As per Rao, investigation on how to reduce the cost by routing user request to data centers located at different places with updated sizes that matched the user requests [5]. Big data service frameworks, comprises a distributed file system, which

distributes data chunks and their replica across the data centers for fine grained load balancing and very similar data access efficiency. To reduce the communication expenditure, few current studies shows to get better data locality by placing jobs on the servers where input data reside to avoid remote data loading . As per Jin, the transmission expenditure is directly proportional to the number of network connection is used. The additional connection used, the higher cost will liable [4].

2.1 Conventional System

In Existing scenario, general focus is on operation and repository suppression, while ignoring the suppression of transmission expenditure. The existing routing strategy between distributed data centers fails to exploit the join diversity of data center networks. Due to storage and computation cost minimization for Big data processing in data centers capacity constraints, all tasks cannot be resides onto the identical server, on which there correspondent data reside.

Drawbacks of conventional system are wastage of resources due to data locality, does not support flexibility and operational task, communication cost optimization has not been achieved and data center resizing difficulty.

2.2 Data center resizing

Cost minimization Data center resizing has been proposed to optimize cost by adjusting the numerous active servers through placing jobs [5]. Data center resizing and data placement are usually jointly consider to match the computing requirement. Existing data center resizing methods only focus on the control of switching on or off servers. The critical problem is non linear relationship between the processing speed and the energy consumption.

2.3 Big data processing management

In recent years, there is big demand for obtaining knowledge from big data to create business values or make society efficient. Among big data processing, streams are becoming conventional and expands, e.g., extroverted medial streams, sensor data streams, lock streams and stock exchanges streams.

3. SYSTEM OVERVIEW

We have considered the reduction of cost problem of big data processing with together consideration of data loading, task assignment and data migration. To elaborate the rate constraint computation and transmission in big data processing we put forward a two dimensional Markov chain and derive the expected task completion time in closed form. Based on the generated closed form expression, we describe the cost minimization problem in a structure of mixed integer nonlinear programming (MINLP) to satisfy the following queries: 1) how to place data modules in the server 2) how to circulate task onto the server without resource constraint violation and 3) how to resize data centers to achieve a goal of operation cost reduction

3.1 Network Model

The distributed data centers is a system that spans many data centers at many locations generally considered for storage. For storing and managing and retrieving data in data center, every data center has many servers. Each data center contains and support task assignment, data loading and data migration. The data centers A, B and C in the figure are connected with each other.

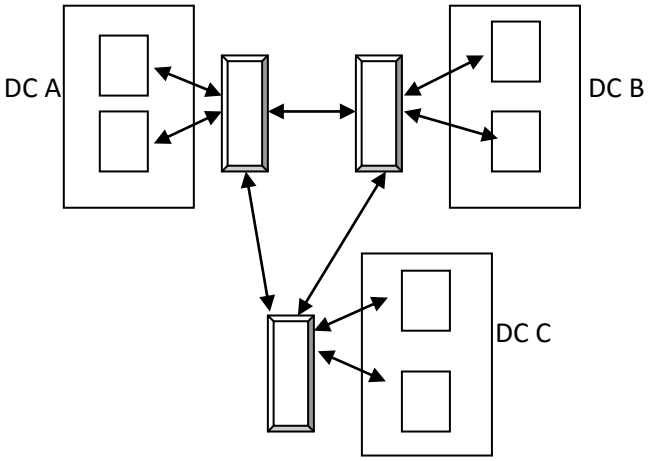

Fig -1: Data Center Network Model

3.2 Markov Chain

A Markov chain is a process that consist finite number of states and some known probabilities having the property that, given the present, the future is conditionally independent of the past. A simple random walk is an example of Markov chain. A series of independent events satisfies the definition of Markov chain. To describe the rate constrained figures and transmission in big data process, a two-dimensional Markov chain is applied and expected task completion time is calculated. The total cost can be calculated by summing up the cost on each server across all distributed data centers and this can be formulated as mixed integer non-linear programming problem and further it is linearized to deal with high operational complexity.

3.3 Data Center Formation

A promising and prominent way is to adjust number of active servers in different data centers to adapt to the user request is known as data center resizing. In this following problem should be studied: how to distribute workload or user request in available data centers. The Quality of Service may not be guaranteed due to the congestion, the data center resizing and request scheduling using weighted bloom filter algorithm is together considered.

The measures on the data loading include data flow from source data centers to target data center, where data modules is required. In order to reduce data migration inside the data center and also to save communication expense task assignment is used. It includes assigning tasks on that number of servers, which are enough to serve the recent request and turn off remaining servers [2] .

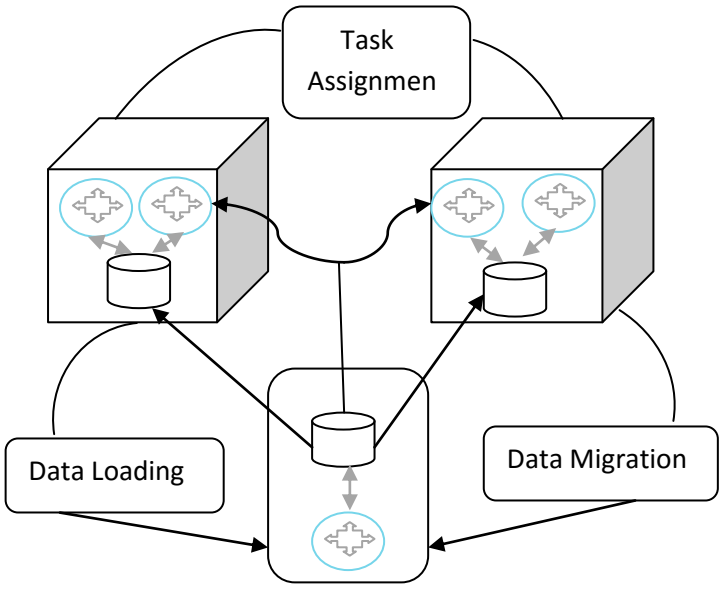

Fig -2: Data Center Formation

3.4 Weighted Bloom Filter

A bloom filter is a compact randomized organization of data for representing a set in order to support membership. Bloom filter is space efficient and hence is very appealing in network application. Weighted bloom filter is conceptualized as: the given pattern is represented and encoded by weighted bloom filter (WBF), where the weight encodes the relation of local patterns and a corresponding global pattern along the time. Then the represented patterns stored in data center could be sampled and hashed into WBF to check whether it is the pattern of interest. Matched patterns together with the corresponding weights are submitted to data center. Finally, the data center aggregates the weights of the patterns and ranks them by weight values. To study the pattern matching in terms of incomplete, dynamic and distributed data *incomplete pattern matching* is used. To create communication efficient and search effective framework, DI matching is used to address incomplete pattern matching.

3. CONCLUSIONS

In this paper, we reviewed important aspects of big data handling in distributed data center. We study how to minimize the cost that occurs during the big data handling by combine consideration of three main factors i.e., data loading, task assignment and data migration by two dimensional Markov chain and formulating problem as MINLP. We proposed weighted bloom filter to reduce communication cost and search effective mechanism. It reduces processing time and increases performance and to ensure pattern matching DI-matching concept is used.

REFERENCES

[1] Lin Gu, Deze Zeng, Peng Li and song Guo, "Cost Minimization for Big Data processing in Geo-Distributed Data Centers,", 2014.

[2] S. A. Yazad, S. Venkatesan, and N. Mittal, "Boosting energy Efficiency with Mirrored data block replication policy and energy Schedular," SIGOPSOper.Syst.Rev.,Vol.47,no.2,pp.33-40,2013.

[3] S.Agarwal,J. Dunagan,N.Jaim,S.Saroju,A.Wolman,and H.Bhogan "Volley:Automated Data Placement for Geo-Distributed Cloud Services,"in the 7th USENIX Symposium on Networked Systems Design and Implementation(NSDI),2010,pp.17-32

[4] H.Jin,T.Cheocherngngam,D.Levy,A.Smith,D.pan,J.Liu,and N.Pissinou," Joint Host-Network Optimization for Energy-Efficient Data Center Networking," in Proceedings of the 27th International Symposium on Parallel Distributed Processing(IPDPS),2013,pp.623-634.

[5] L.Rao, X.Liu, L.Xie and W.Liu, "Minimizing Electricity cost: Optimization of Distributed Internet Data Centers in a Multi-Electricity-Market Environment," in Proceedings of the 29th International Conference on Computer Communications (INFOCOM), IEEE, 2010, pp.1-9

[6] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, "Mad skills: new analysis practices for Big data, " Proc. VLDB Endow., vol.2, no.2,pp. 1481-1492, 2009.

[7] Siyaun Liu, Lei Kang, Lei Chen, Lionel M. Ni, fellow, "How to Conduct Distributed Incomplete Pattern Matching, " vol. 25, no.4, April 2014.