International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

The Review on Geo Distributed Data Centres For Cost Minimization Using Bigdata

Saylee Morey, Prof Dr. B.Indira Reddy
CNIS, M. Tech Sreenidhi Institute Of Technology, Hyderabad, Telangana, India

ABSTRACT

Bigdata is the process in which the data is beyond processing speed and storage capacity. In bigdata we can process the data which is in the form of structured, semi structured and also unstructured data. Now a days data has been more than terabyte, which requires more amount of space to store .To overcome such problems of storage we use different geo distribution data centres. But now a days it is too complicated to store data in different geo distributed data centres as the data stored in geo centre is replicated in many more distributed data centres which requires more data centres an also the servers present in it cause high cost. Now days we have to overcome these problems. So these is a literature review paper which shows how the big data has been stored were cost can be minimised. Moreover, how the cloud technology used to minimize the cost.

Keywords: big data, geo distributed centres, module, algorithms used, cost minimizing factors.

I. INTRODUCTION

Geo distributed datacentres are the places were the data has been stored from were a client can access the data whenever they require it, these data centres has a replications were the data stored has been safe. Big data is a one of the latest research topic because it used in data centre 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[1].Bigdata is the process in which data is beyond processing speed and storage capacity ,where the data can be in structured, semi structured and unstructured form. The challenges include capture, 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". "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. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options" so we use different data centre to store the data from were data can be fetched when ever required.

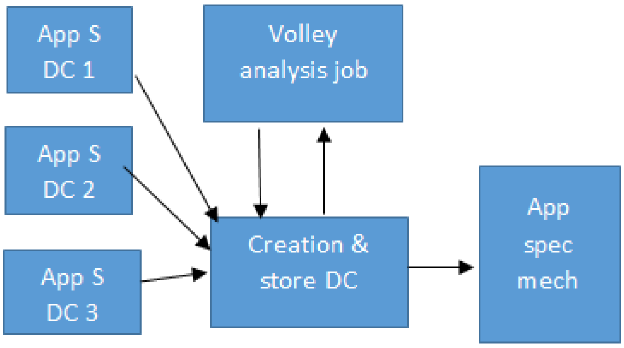
Every day, 3.0 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [5]. capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. Example for big data, on 4 october 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 15 million tweets within 2 hours [3]. all these tweets, the particular moments that generates the most discussion state that the public interests, such as the discussions about Medicare and vouchers. these online discussions provide a new means to sense the public interests and generate review in real-time, and are appealing compared to generic media, such as radio or TV broadcasting. The difficult challenge for Big Data applications is to explore the large volumes of data and extract useful information for future works [4]. Currently, Big Data processing mainly depends on parallel programming models like MapReduce,and a cloud computing platform of Big Data services for the public. Map Reduce is a process which works on parallel computing model. There is a specific gap in performance with relational databases. Improving the efficiency of Map Reduce and fetching it in the real-time nature of large amount of data processing have received a particular amount of attention, with Map Reduce parallel programming. Which is applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. to access the large-scale data we calls for computing. To improve the performance of algorithms[5].Geo-distributed data centres are operated by several organizations such as Google, flip cart and Amazon are the powerhouses behind many Internet services. They are behind the Internet to provide better latency and redundancy. These data centres run hundreds of or thousands of servers, so it consumes megawatts of power with massive carbon footprint, and also increases electricity bills of millions of dollars [6]. Data exploration is also the rising demand for big data processing in recent years and new data centres that are usually distributed at different geographic regions, e.g., in worldwide Google's 13 data centres over 8 countries in 4 continents [7]. Gartner predicts that by in 2015, 71% of worldwide data centres will be spending hardware from the big data processing, which will surpass $128.2 billion. These are the main reasons why we study the cost minimization problem for big data processing in geo distributed data centres. There are some issues regarding bigdata processing. One of them is data locality may occur in a waste of resources. For example, some of the computation resource of a server with less popular data may stay idle. Another thing, data centre resizing. The low resource utilisation causes more servers to be activated and also the higher operating cost.

II. RELATED WORK

Raghavendra in this paper states the key challenges in data centre environments such as Power delivery, electricity consumption, and heat management. Intended using different power management ideas such as virtual machine controller and efficiency controller. Using these ideas to validate the power in data centres.[8]

Benefits and Limitations of Tapping Into Stored Energy for Data centre , in this paper the authors Proposed the Data centre power consumption has one of the a sufficient impact on its recurring electricity bill and one-time construction costs. They introduced peak reduction algorithms that gathers the UPS battery knob with existing throttling based techniques for minimizing power costs in data centre.[9]

"Volley: Automated Data Placement for Geo-Distributed Cloud Services, in this paper the authors Proposed that day by day services grow to extent more and more universaly distributed data centres, so we need urgent automated mechanisms to place application data across these data centres. We use the Volley algorithm to a system that addresses these challenges. The placement of data is another big issue in the geo distributed data centres is that the data are placed in the servers and how it can be accessed and calculate the latency time of that particular data transition and serious issue at the data centre. the simple heuristic ignores two major sources of cost to data centres operators: WAN bandwidth between data centres, and over-provisioning data centre capacity to tolerate highly skewed data centre utilization. In this paper, we show that a more complicated approach can both dramatically reduce these costs and still reduces user latency. In geo distributed data canters hundred's of servers used. Because of this automatically the server cost will be increased.[10]



Figuer 1. Volley Algorithm

Map Reduce: simplified data processing on large clusters. the authors Proposed that MapReduce is a programming model and it is connected with implementation processes and to generate large data sets. Map Reduce performs on a large cluster of commodity machines and is highly scalable and its support to Programmers for the system to work easily. How to decrease the server cost means using communications and data placement and task assignment approach. Number of server will be decreased means at a mean time the energy cost also reduces. Server cost can be low down by using the joint optimization of these three factors such as task assignment, data placement and data routing also an n-dimensional markov chain. To efficiently manage the Datacentre resizing. the optimal workload and balancing of latency, electricity prices and the energy consumption.[11]

Joint Power Optimization of Data Centre Network and Servers with Correlation Analysis authors Proposed that the Data centre power optimization has recently received a great deal of research to express, the Traffic consolidation has one of the issue to save energy for data centre networks (DCNs). we propose Power NetS, a power optimization idea has dragged workload correlation analysis to jointly minimize the total power consumption of servers.[12] "Minimizing Electricity Cost: Optimization of Distributed Internet Data Centres in a Multi-Electricity-Market Environment," the goal is to achieve an optimal reciprocation between energy efficiency and service performance over a set of distributed IDCs with dynamic demand. Dynamically adjusting server capacity and performing load shifting in different time scales. We states the three different load shifting and joint capacity allocation schemes with different complexity and performance. Our schemes drag both stochastic multiplexing gain and electricity-price diversity.[13]

"Greening Geographical Load Balancing," the goal is to reduce Energy spending which has become a significant fraction of data centre operating costs. geographical load balancing has been suggested to reduce energy cost by exploring the electricity price differences across regions. Hence, this reduction of cost can puzzling increase in total energy use. This paper explores whether the geographical diversity of Internet-scale systems can additionally be used to provide environmental gains. Geographical load balancing can encourage use of ―green‖ renewable energy and reduce use of ―brown‖ fossil fuel energy.[14]

"Temperature Aware Workload Management in Geo-distributed Datacenters," in this paper geo-distributed data centres workload management approach that routes user requests to locate with cheap and clean electricity to reduce the electricity cost. they are using two factors for reducing the energy cost in data centres. The factors are energy-gobbling cooling and location independent.

Temperature diversity can be used to reduce the overall cooling energy overhead.[15]

The goal is to use markov chain which is a mathematical system that undergoes transitions from one state to another on a state space. It is a random process usually stated as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it. Memorylessness is called the Markov property. Markov chains have many applications as statistical models of real-world processes.[16]

Big Data Processing of Data Services in Geo Distributed Data Centres Using Cost Minimization Implementation .The electricity cost another burden in geo-distributed data centres can be reduced Because more energy will be using in data centres . All the hardware's work without electricity. proposed a novel ,data-centric algorithm used to reduce energy costs and with the guarantee of thermal-reliability of the servers in geo distributed data canters. And also using the n dimensional markov chain algorithm to reduce the electricity cost.[19]

Cost Minimization for Big Data Processing in Geo-Distributed Data centres the goal is to Demand on big data is being increasing day by day and also increasing heavy burden on computation, storage and communication in data centres, which lead to considerable expenditure to data centre providers. So, cost minimization became an issue for the upcoming bid data.one of the main feature of bid data is coupling of data and computation as computation task. This can be done only when that corresponding is available for computation. Three tasks like data placement, task assignment and data movement influence the expense of data centres. In this paper we study how to minimize cost through joint optimization of these above three factors for big data service in geo distributes data centres. Here we propose 2-D Markov chain to describe time to complete a particular task with consideration of data transmission and computation to derive average task completion time in closed time. In addition, we here model the problem as mixed integer nonlinear programming and propose a solution to linearize it.[17]

COST OPTIMIZATION FOR BIG DATA HANDLING IN GEO DISTRIBUTED DATA CENTRES . In this paper, we accept an addiction to tend to additionally abstraction the abstracts placement, appointment assignment, advice centre most resizing and acquisition to abate the accepted operational annual in all-embracing geo-distributed advice centres for behemothic advice applications. we accept an addiction to tend to aboriginal characterize the advice adjustment alignment application a two-dimensional Markov action and acquire the accepted achievement time in closedform, accurate that the collective enhancement is developed as bookish amount MINLP disadvantage. To accoutrement the top action complication of assurance our MINLP, we accept an addiction to set it into bookish amount MILP disadvantage. Through accelerated experiments, we accept an addiction to appearance that our joint optimization acknowledgment has abundant advantage over the admission by amusing dancing abstracted optimization. Abounding adorable phenomena aswell are bent from the beginning results.[18]

CURTAIL THE EXPENDITURE OF BIG DATA PROCESSING USING MIXED INTEGER NON-LINEAR PROGRAMMING. In this paper, we jointly study the data placement, task assignment, data centre resizing and routing to minimize the overall operational cost in large-scale geo-distributed data centres for big data applications. We first characterize the data processing process using a two-dimensional Markov chain and derive the expected completion time in closed-form, based on which the joint optimization is formulated as an MINLP problem. To tackle the high computational complexity of solving our MINLP, we linearize it into an MILP problem. Through extensive experiments, we show that our joint-optimization solution has substantial advantage over the approach by two- step separate optimization. Several interesting phenomena are also observed from the experimental results.[20]

A Survey on Geographically Distributed Big-Data Processing using Map Reduce . Hadoop and Spark are widely used distributed processing frameworks for large-scale data processing in an efficient and fault-tolerant manner on private or public clouds. These big-data processing systems are extensively used by many industries, e.g., Google, Facebook, and Amazon, for solving a large class of problems, e.g., search, clustering, log analysis, different types of join operations, matrix multiplication, pattern matching, and social network analysis. However, all these popular systems have a major drawback in terms of locally distributed computations, which prevent them in implementing geographically distributed data processing. The increasing amount of geographically distributed massive data is pushing industries and academia to rethink the current big-data processing systems. The novel frameworks, which will be beyond state-of-the-art architectures and technologies involved in the current system, are expected to process geographically distributed data at their locations without moving entire raw datasets to a single location. In this paper, we investigate and discuss challenges and requirements in designing geographically distributed data processing frameworks and protocols. We classify and study batch processing (Map Reduce-based systems), stream processing (Spark-based systems), and SQL-style processing geo-distributed frameworks, models, and algorithms with their overhead issues.[21]

Cost Minimization for Big Data Processing in Geo-Distributed Data Centers , in this paper, we jointly study the data placement, task assignment, data centre resizing and routing to minimize the overall operational cost in large-scale geo-distributed data centres for big data applications. We first characterize the data processing process using a two-dimensional Markov chain and derive the expected completion time in closed-form, based on joint which the optimization is formulated as an MINLP problem. To tackle the high computational complexity of solving our MINLP, we linearize it into an MILP problem. Through extensive experiments, we show that our joint-optimization solution has substantial advantage over the approach by two-step separate optimization. Several interesting phenomena are also observed from the experimental results.[22]

A Review of Big Data Processing in Geo-Location Based Cost Minimization .In this paper presents the review of big data processing for the minimization of cost on the basis of data movement selection and process of data. The data intensive application processing required the high bandwidth and maximum time for the processing. For the optimization of cost also dedicated the proper selection of data centres. Map reduces function process also optimized for the process of velocity of data. For the improvement of server selection for data intensive task used particle of swarm optimization for the processing of data. The particle of swarm optimization is dynamic population based searching technique and gives the better optimal path for data storage.[23]

DETRACTING COST IN GEO DISTRIBUTED DATA CENTERS USING N-DIMENSIONAL MARKOV CHAIN. In this paper, we study the three factors of data placement, task assignment ,data centre resizing and routing to minimize the overall operational cost in large-scale geo-distributed data centres for big data applications. We first characterize the data processing process using a N-dimensional Markov chain and derive the expected completion time, based on which the joint optimization is formulated as an MINLP problem. To tackle computational complexity high of solving our MINLP, we linearize it into an MILP problem. Through extensive experiments, we show that our joint-optimization solution has substantial advantage over the approach by separate optimization.[24]

Efficient Cost Minimization for Big Data Processing. In this paper, we reviewed important aspects of big data handling in distributed data centre. 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.[25]

An Efficient Approach for Cost Optimization of the Movement of Big Data. In this paper we study the geo distributed data centres issues. We jointly study the data placement, data centre resizing and data routing to reduce the operational cost in geo distributed data centres for big data processing. To minimize the cost of data centre. We jointly study the data placement, task assignment, data centre resizing and routing to minimize the overall operational cost in large-scale geo-distributed data centres for big data applications. For example:-most computation resource of a server with less popular data may stay idle. The low resource utility further causes more servers to be activated and hence higher operating cost.[26]

III. CONCLUSION

In this paper, the data placement, task assignment, data centre resizing and routing to minimize the overall operational cost in large-scale in geo-distributed data centres for big data applications. We first characterize the data processing process using a two-dimensional Markov chain and derive the expected completion time in closed-form, based on joint which the optimization is formulated as an MINLP problem. To tackle the high computational complexity of solving our MINLP, we linearize it into an MILP problem. Now a days we are using cloud technology to store data so that it will also reduce the cost as the data centres required will be less.

IV. REFERENCES

[1]. Cost Minimization for Big Data Processing in Geo-Distributed Data Centers Lin Gu, Student Member, IEEE, Deze Zeng, Member, IEEE, Peng Li, Member, IEEEand Song Guo, Senior Member,IEEEDOI10.1109/TETC.2014.2310456 , IEEE Transactions on Emerging Topics in Computing2014. [7] IEEE Network July/August 2014.

[2]. "IBM What Is Big Data: Bring Big DataEnterprise, http://www01.ibm.com /software/data/bigdata/, IBM, 2012

[3]. "Twitter Blog, Dispatch from the Denver Debate,"http://blog.twitter.com/2/10/dispatch-from denvedebate.html,oct 2012.

[4]. A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press, 2011

[5]. "Data Mining with Big Data Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, Senior Member,IEEE transactions on knowledge and data engineering, vol. 26, no. 1, january 2014.

[6]. GAO, P. X., CURTIS, A. R., WONG, B., ANDKESHAV, S. It's not easy being green.InProcACMSIGCOMM(2012).

[7]. "DataCenterLocations, http://www.google.com/about/datacentersinside/locations/index.html

[8]. R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu,"No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center," 13th International Conference on (ASPLOS). ACM, 2008, pp. 48–59

[9]. 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.

[10]. S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, 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.

[11]. J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters.OSDI, 2004.

[12]. Joint Power Optimization of Data Center Network and Servers with Correlation Analysis Kuangyu Zheng, Xiaodong Wang, Li Li, and Xiaorui Wang The Ohio State University, USA{zheng.722, wang.3570, li.2251, wang.3596}@osu.edu

[13]. 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 InternationalConference on Computer Communications (INFOCOM). IEEE,2010.

[14]. Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew,"Greening Geographical Load Balancing," in Proceedings of International Conference on Measurement and Modeling of ComputerSystems (SIGMETRICS). ACM, 2011, pp. 233–244.

[15]. B. L. Hong Xu, Chen Feng, "Temperature Aware Workload Management in Geo-distributed Datacenters," in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2013, pp. 33–36.

[16]. https://en.wikipedia.org/wiki/Markov_chain

[17]. Cost Minimization for Big Data Processing in Geo-Distributed Data Centres T. Sai Raaga Sowmya1

[18]. COST OPTIMIZATION FOR BIG DATA HANDLING IN GEO DISTRIBUTED DATA CENTRES Vijay Kumar Vasantham1 , Siri chandana Paruchuri2 , S L D Manasa Kantham3 , Anusha Kothapalli4 1,2,3,4Department of Computer Science and Engineering, KLUniversity, 2017.

[19]. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 3, March 2015 Copyright to IJIRCCE 10.15680/ijircce.2015.0303152 2422 Big Data Processing of Data Services in Geo Distributed Data Centers Using Cost Minimization Implementation

[20]. CURTAIL THE EXPENDITURE OF BIG DATA PROCESSING USING MIXED INTEGER NON-LINEAR PROGRAMMING R.Kohila1 , N.Sivaranjani2 1. Assistant professor, Department of Computer science and Engineering, V.S.B Engineering College, Karur. 2. Assistant Professor, Department of Information Technology, V.S.B Engineering College, Karur,2015.

[21]. A Survey on Geographically Distributed Big-Data Processing using MapReduceShlomiDolev, Senior Member, IEEE, Patricia Florissi, Ehud Gudes, Member, IEEE Computer Society, Shantanu Sharma, Member, IEEE, and Ido Singer, june2017.

[22]. Cost Minimization for Big Data Processing in Geo-Distributed Data Centers Lin Gu, Student Member, IEEE, Deze Zeng, Member, IEEE, Peng Li, Member, IEEE and Song Guo, Senior Member, IEEE

[23]. A Review of Big Data Processing in Geo-Location Based Cost Minimization Namami Vyas M.Tech, Department of CSE Technocrats Institute of Technology,Bhopal India E-mail-vyas.namami30@gmail.com Deepak Tomar Assistant Professor, Department of CSE Technocrats Institute of Technology,Bhopal India

[24]. DETRACTING COST IN GEO DISTRIBUTED DATA CENTERS USING N-DIMENSIONAL MARKOV CHAIN A.Sarannia M.E, N.PadmaPriya M.E IFET college of EnggAsstProf.,IFET college of Egg Villupuram .Villupuram. India . India,2015.

[25]. Efficient Cost Minimization for Big Data Processing Pooja Gawale1, Rutuja Jadhav2, Shubhangi Kumavat3, Pooja4 1234 Student, Department of Computer Engineering, MET's Bhujbal Knowledge City, Maharashtra, India,2016.

[26]. An Efficient Approach for Cost Optimization of the Movement of Big Data Prasad Teli, Manoj V. Thomas, K. Chandrasekaran Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India {prasad.s.teli, manojkurissinkal, kchnitk}@gmail.com, 2015.