International Journal of Research
Available at
https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

A Cost Minimization Implementation Data Processing for BigData Application

Divya Byri
Department of CSE

ABSTRACT: Currentexertions have advanced a promising method, calleddata analytic, which uses statistical and cloud computing toreduce using size of Big Data to a controllable size to extractinformation, build a knowledge base using the derived data,and eventually develop a nonparametric model for the BigData.Diverse from conventionalcloud services, one of the main structures of big data services is the tight coupling between data and computation ascomputation tasks can be conducted only when the corresponding data are available. As a result, three factors, i.e., taskassignment, data placement, and data movement, deeply influence the operational expenditure of data centers. In thispaper, we are interested to study the cost minimization problem and optimization of these factors for big data servicesin geo-distributed data centers.

KEYWORDS-big data , cost minimization, data placement.

I. INTRODUCTION

Cloud computing has been driven fundamentally by theneed to process an exploding quantity of data in terms ofexabytes as we are approaching the Zetta Byte Era.One critical trend shines through the cloud is Big Data.Indeed, it's the core driver in cloud computing and willdefine the future of IT. When a company needed to storeand access more data they had one of two choices. Oneoption would be to buy a bigger machine with more CPU,RAM, disk space, etc. This is known as scaling vertically.Of course, there is a limit to how big of a machine we canactually buy and this does not work when you start talkingabout internet scale. The other option would be to scalehorizontally. This usually meant contacting some databasevendor to buy a bigger solution. These solutions do notcome cheap and therefore required a significant investment.Today, the source of data generated not only by the usersand applications but also “machine- generated,” and suchdata is exponentially leading the change in the Big Dataspace.

Big Data processing is performed through aprogramming paradigm known as MapReduce. Typically,implementation of the MapReduce paradigm requiresnetworked attached storage and parallel processing. Thecomputing needs of MapReduce programming are oftenbeyond what small and medium sized business are able tocommit.

Cloud computing is on-demand network access toComputing resources, provided by an outside entity.Common deployment models for cloud computing includeplatform as a service (PaaS), software as a service (SaaS),infrastructure as a service (IaaS), and hardware as a service(HaaS).Platform as a Service (PaaS) is the use of cloudcomputing to provide platforms for the development anduse of custom applications. Software as a service (SaaS)provides businesses with applications that are stored andrun on virtual servers – in the cloud. In the IaaS model, aclient business will pay on a per-use basis for use ofequipment to support computing operations includingstorage, hardware, servers, and networking equipment.

HaaS is a cloud service based upon the model of timesharing on minicomputers and mainframes.The three types of cloud computing are the public cloud,the private cloud, and the hybrid cloud. A public cloud isthe pay- as-you-go services. A private cloud is internal datacenter of a business not available to the general public butbased on cloud structure. The hybrid cloud is a combinationof the public cloud and private cloud.Three major reasons for small to medium sizedbusinesses to use cloud computing for big datatechnology implementation are hardware cost reduction,processing cost reduction, and ability to test the value of bigdata

Big data is a collection of data sets so large andcomplex which is also exceeds the processing capacity ofconventional database systems. The data is too big, movestoo fast, or doesn't fit the structures of our current databasearchitectures. Big Data is typically large volume of unstructured (or semi structured) and structured data that getscreated from various organized and unorganizedapplications, activities and channels such as emails,tweeter, web logs, Facebook, etc. The main difficulties withBig Data include capture, storage, search, sharing, analysis,and visualization. The core of Big Data is Hadoop which isa platform for distributing computing problems across anumber of servers. It is first developed and released as opensource by Yahoo!, it implements the MapReduce approachpioneered by Google in compiling its search indexes.Hadoop's MapReduce involves distributing a datasetamong multiple servers and operating on the data: the"map" stage. The partial results are then recombined: the"reduce" stage. To store data, Hadoop utilizes its owndistributed file system, HDFS, which makes data availableto multiple computing nodes. Big data explosion, a resultnot only of increasing Internet usage by people around theworld, but also the connection of billions of devices to theInternet. Eight years ago, for example, there were onlyaround 5 exabytes of data online. Just two years ago, thatamount of data passed over the Internet over the courseof a single month.

II. RELATED WORKS

A.Sivasubramanian,B.Urgaonkar et al proposed the Data center power consumption has one of the a significantimpact on both its recurring electricity bill (Op-ex) and one-time construction costs (Cap-ex). They develop peakreduction algorithms that combine the UPS battery knob with existing throttling based techniques for minimizingpower costs in datacenter .

SharadAgarwal, John Dunagan et al proposed the Nowadays services grow to span moreand more globally distributed datacenters, so we need urgent automated mechanisms to place application data acrossthese datacenters. Proposed the MapReduce is a programming model and its associated with implementation forprocessing and to generating large data sets. MapReduce runs on a large cluster of commodity machines and is highlyscalable and its support to Programmers for the system easy to use.

Kuangyu Zheng, Xiaodong Wang et al proposed theData center power optimization has recently received a great deal of research attention .Traffic consolidation has one torecently proposed to save energy for data center networks (DCNs). we propose PowerNetS, a power optimizationstrategy that leverages workload correlation analysis to jointly minimize the total power consumption of servers. DanXu XinLiu , Bin Fan, The goal is to achieve an optimal tradeoff between energy efficiency and service performanceover a set of distributed IDCs with dynamic demand. Dynamically adjusting server capacity and performing loadshifting in different time scales. We propose three different load shifting and joint capacity allocation schemes withdifferent complexity and performance. Our schemes leverage both stochastic multiplexing gain and electricity-pricediversity.

Zhenhua Liu, Minghong Lin, Energy expenditure has become a significant fraction of data center operatingcosts. Recently, geographical load balancing has been suggested to reduce energy cost by exploiting the electricityprice differences across regions. However, this reduction of cost can paradoxically increase total energy use. This paperexplores whether the geographical diversity of Internet-scale systems can additionally be used to provide environmentalgains. Geographical load balancing can encourage use of greenrenewable energy and reduce use of brown fossilfuel energy.

Hong Xu, Chen Feng, Baochun Li, For geo-distributed datacenters workload management approach thatroutes user requests to locations with cheaper and cleaner electricity to reduce the electricity cost.

III. SUGGESTED METHODOLOGIES

International Journal of Research
Available at
https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

Big data refers to exponentially growing based orunstructured data. Production of big data is created viacompanies, the Internet, society and cyber physical structures.Another viable definition of massive statistics refers to those data units which might be complex and massive which makes it tough toprocedure to be had control gear or traditional paradigms.One of the maximum promising paradigms to manipulate huge data has been data analytic [9]. Data analytic refers back to the evaluation through inspection, cleaning, transformation, modelsand verification working towards creation of conclusionsand decision making at the actual that means of the data. Theshowing Fig.1 depicts the concepts of data analytic.

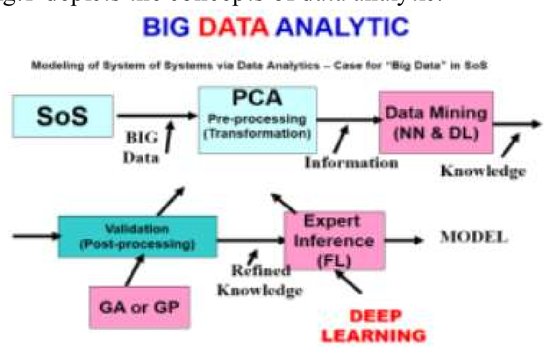

Fig. 1. Data Analytic Tools for Big Data Management

Hadoop is an open-source programming structure for processing and storing big data information in an appropriated style on enormous groups of item equipment. Basically, itachieves two assignments: enormous evidencestoringand quickermanagement. Open-source software: Open sourceprogramming varies from business programming because ofthe expansive and open system of designer that make anddeal with the projects. Habitually, it's permissible to download,utilize and add to, however more business adaptations ofHadoop are getting to be accessible.

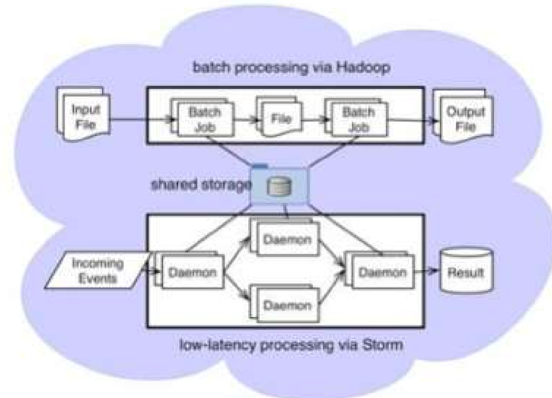

Fig. 2. Structure of Hadoop

• **Framework:**In this case, it suggests all that we haveto generate and run our product applications is givenprograms, device sets, associations, and so. [Fig.2]TheHadoop brand comprises a extensive range of tools. Two ofthem are center parts of Hadoop;

• Hadoop Distributed File System (HDFS) is a virtual document framework that look like some other data framework with the exemption of than when we movea data on HDFS, this document is part into numerouslittle documents, each of those data is reproducedand put away on (ordinarily, might be altered) threeservers for adaptation to internal failure requirements.

• Hadoop MapReduce is anmethod to part every solicitation into slighter solicitations which are sent tonumerous little servers, permitting a really adaptableutilization of CPU power.

For this portion, we try to explain the relationships of thetheoretical and implementation parts of Cloud Computing.They can be explained as following;

• Understand what defines Cloud Computing and be ableto explain the nature and make up of typical cloudscenarios

• Understand how to use MPI programming with Python

• Understand NoSQL database structure and theory, Map/Reduce algorithm, and implementations such as HadoopThis approach would be reflecting several skills, suchas programming, implementing a program for a particularproblem, data gathering, project management, and someother workflows.

On the other hand, establishing the study was challengingwhile performing the feasibility studies. We have faced withthese constraints. We were able to manage our plan, andstarted to handle these obstacles in time. Furthermore, wewere gathering the data from the City of Austin (https://data.austintexas.gov) that includes several important datasheets, such as water quality samplings, restaurant samplingrecords, APD crime summaries, etc. For this experiment, wepicked the historical crime data that entered by the officials.This data sheets includes several data fields and differentattributes as can be seen on Table-I. Moreover, for buildingthe best approach we were back through 2008 to 2011, andadded the most recent entries to make a precise comparisonas described year-to-date for 2014. The data files can beobtained from the following links.

- https://data.austintexas.gov/api/views/r6sgxka2/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/ei2nfehk/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/4c6htv2y/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/gr59-ids7/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/b4y9-5x39/rows.csv?accessType=DOWNLOAD

Cost minimization using mapreduce algorithm

Numerous algorithms were defined earlier in the analysis of large data set. Will go through the different work done tohandle Big Data. In the beginning different algorithm was used earlier to analyze the big data. In work done by Hall. etal. there is defined an approach for forming the rules of the large set of training data. The approach is to have a singledecision system generated from a large and independent n subset of data. Here we use cost minimization using mapreduce algorithm as follows

Cost Minimization using MapReduce Algorithms. Denote by S the set of input objects for the underlying problem.

Let n, the problem cardinality, be the number of objects in S, and t be the number of machines used in the system.Define m = n/t, namely, m is the number of objects per machine when S is evenly distributed across the machines.

Consider an algorithm for solving a problem on S.

We say that the algorithm is minimal cost if it has all of the following properties.

- **Minimum footprint:** at all times, each machine uses only O(m) space of storage.
- **Bounded net-traffic:** in each round, every machine sends and receives at most O(m) words of information over thenetwork.

Constant round: the algorithm must terminate after aconstant number of rounds.

- **Optimal computation:** every machine performs only O(Tseq/t) amount of computation in total (i.e., summing over allrounds), where Tseq is the time needed to solve the same problem on a single sequential machine. Namely, thealgorithm should achieve a speedup of t by using t machines in parallel.

Each machine M has at most 2 groups remaining, i.e., with keys kmin (M) and kmax (M), respectively. Hence, thereare at most 2t such groups on all machines. To handle them, we ask each machine to send at most 4 values toM1 (i.e.,to just a single machine). The following elaborates how:

Map-shuffle (on each Mi, $1 \le i \le t$):

Step 1. Obtain the total weight Wmin(Mi) of group kmin(Mi), i.e., by considering only objects in Mi.

Step 2. Send pair (kmin(Mi),Wmin(Mi)) toM1.

Step 3. If kmin (Mi) 6= kmax (Mi), send pair (kmax (Mi),Wmax (Mi)) to M1, where the definition of kmax (Mi) issimilar to kmin (Mi).

Reduce (only on M1):

Let (k1,w1), ..., (kx,wx) be the pairs received in the previous phase where x is some value between t and 2t. For eachgroup whose key k is in one of the x pairs, output its final aggregate Pj|kj=k wj . The minimality of our group-byalgorithm is easy to verify. It suffices to point out that the reduce phase of the last round takes O(t log t) = O(nt log n)time (since t ≤ m = n/t)

IV. CONCLUSION

By means of proposed approaches, cloud environments can be securedfor compound business operations. Using big data tools to study the enormous amount of threat data received daily,and correlating the different components of an attack, allowsa security vendor to uninterruptedly update their global threat intelligence and equates to improved threat data andunderstanding.Through big data analytics fraud can be identified themoment it happens and appropriate measures can be takento constrain the harm.Wetogether study the data placement, data centre resizingand data routing to reduce the operational cost in geo distributed data centres for big data processing..Todiminish thecost of data centre. We together study the data placement, task assignment, data centre resizing and routing to minimizethe overall operational cost in large-scale geo-distributed data centres for big data applications.

REFERENCES

[1] Cost Minimization for Big Data Processing in Geo-Distributed Data Centers Lin Gu, Student Member, IEEE, DezeZeng, Member, IEEE, PengLi, Member, IEEEand Song Guo, Senior Member,IEEEDOI10.1109/TETC.2014.2310456, IEEE Transactions on Emerging Topics inComputing2014.

[2] A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press, 2011[3] M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A. G. Dimakis,R. Vadali, S. Chen, and D. Borthakur, "Xoring elephants: novel erasure codesfor big data,in Proceedings of the 39th international conference on Very Large Data Bases, ser. PVLDB'13. VLDB Endowment, 2013, pp. 325–336.

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

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

[6] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing Electricity Cost:Optimization of Distributed Internet Data Centers in a Multi-Electricity-MarketEnvironment," in Proceedings of the 29th InternationalConference on Computer Communications (INFOCOM). IEEE,2010.

[7] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloudcomputing reference architecture: Cloud service management perspective. Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.

[8] Gczy, P., Izumi, N., &Hasida, K. (2012). Cloudsourcing: Managingcloud adoption. Global Journal of Business Research, 6(2), 57-70.

[9] Tannahill, B. K., Maute, C. E., Yetis, Y., Ezell, M. N., Jaimes, A.,Rosas, R., . &Jamshidi, M. (2013, June). Modeling of system ofsystems via data analyticsCase for Big Data in SoS. In System ofSystems Engineering (SoSE), 2013 8th International Conference on(pp. 177-183).IEEE.