JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Cost Effective PSO Model for MapReduce in Cloud Environment

Vidhyasagar B S[1*], Ajithkumar M[2], Shaik Sajid[3,] Syed Khadeer [4] , Rahul P [5,] J. Arunnehru[6],

[1] Dept. of CSE, SRM Institute of Science and Technology, Chennai, India
[2] Dept. of CSE, SRM Institute of Science and Technology, Chennai, India
[3] Dept. of CSE, SRM Institute of Science and Technology, Chennai, India
[4] Dept. of CSE, SRM Institute of Science and Technology, Chennai, India
[5] Dept. of CSE, SRM Institute of Science and Technology, Chennai, India
[6] Dept. of CSE, SRM Institute of Science and Technology, Chennai, India

[*]*Corresponding Author: vidhyasagar.s@vdp.srmuniv.ac.in, Tel.: +91-9940301968*

Available online at: www.ijcseonline.org

Received: 20/Mar/2018, Revised: 28/Mar/2018, Accepted: 19/Apr/2018, Published: 30/Apr/2018

Abstract Cloud service provides everything as a service over the Internet or Intranet. Provisioning and allocation of virtual resource over the network requests based on used demand (pay-as-you-go). Big Data, which has large set of data that are so voluminous and complex that traditional method is not enough to process the data, Hadoop MapReduce framework is used to process the large set of data in a distributed manner. Efficient slave nodes selection is difficult to setup Hadoop cluster in cloud environment which led to more cost. We have proposed an algorithm called Particle Swarm Optimization(PSO) that determines the optimal number of nodes in the Hadoop cluster utilizes based on the data sets which provides efficient job execution on minimal set of DataNodes in cloud environment.

Keywords*— Hadoop, MapReduce, Virtualization, PSO,YARN, HDFS*

I. Introduction

Cloud Computing is a technology which provides various kinds of virtual service to the user on demand basis from the basic stack of cloud such as Software-as-a-service, platform-as-a-service and infrastructure-as-a-service. Virtual Applications such as e-mail (electronic mail),Web conferencing, Customer Relationship Management(CRM) etc., are available in cloud platform. Public cloud[1-2]can be accessible by everyone in the organization on pay as you use model. In private cloud[3] offers service within the organization over the intranet. Community cloud meant for community users such as education, Medical, automobile etc. The combination of public and private cloud services together called Hybrid cloud[4].Infrastructure-as-a-Service (IAAS) is the delivery of service on demand basis such as virtual machines, server and storage etc. Platform as a Service (PAAS) provides development environment to the user as a service Example such as Amazon web services, Azure, Salesforce.com and Rackspace. Software-as-a-Service (SAAS) provides virtual applications as a service to the user such as Document, E-mail, One drive etc.

Big data used for processing large sets of data. Hadoop[5] is an open source software used for storing and processing huge amount of data. Hadoop has three basic core components in Hadoop architecture namely HDFS (Hadoop Distributed File System), MapReduce[6-9] and YARN. HDFS[10] is specially designed file system for storing huge datasets for storing and processing in distributed manner. MapReduce is a technique in Hadoop to process the structured and unstructured data in HDFS YARN provides resource management service for allocating resource to process the job. Big data are large datasets of voluminous and complex traditional data processing paradigm to deal with them.

The volume of the big data is tremendously large day by day it is going to increase and now a day's dealing with just 4.4 zeta bytes of data and by 2020 it arises to 44 zeta bytes of data. Variety: Different kinds of data were generated from different sources. Structured or tabular: Having proper schema for data and using RDBMS and process how to store and process the data by writing queries.

Semi-Structured: Log files (JSON, XML,CSV,TSV,E-mail).Un-Structured: videos, images, text, audio. Velocity is speed of fetching and transferring the data. Value mining the useful data from the data set and then after perform analysis and whatever analysis done it is of some value which is useful. Velocity: Big data will have a lot of inconsistency because while dumping huge data some data package are bound to loose in the process so need to fill up the data.

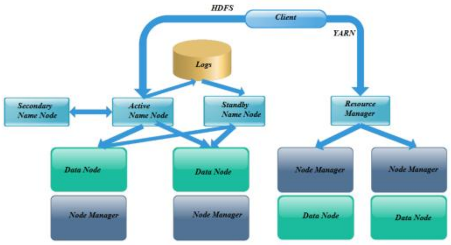


FIGURE.1.HADOOP ARCHITECTURE

Hadoop architecture and it's components shown in figure [1], components of HDFS: Namenode is also called as Master daemon (daemon is processing in the background internally but no physical appearance). NameNode is the responsible for storing the Hadoop distributed file system metadata, the metadata keeps track of all the files that are present in the HDFS it stores information related to files and blocks of the file system map to each files that is present in HDFS. DataNode is also called as Slave daemon, it is responsible for storing and retrieves the data as instructed by the NameNode and DataNode always keeps reporting to the NameNode and DataNode keeps multiple files for each file which is useful for backup that is present in the HDFS. Secondary NameNode: Every transaction was record in edit log file and edit log file was large. NameNode fails because the cluster cannot be operational so now secondary NameNode instructs to record transaction in new edit log file. Now the secondary node copies fsimage and edit log file to its check point directory and if once these files are copied then secondary NameNode loads fsimage and applies all the transaction from the editlog file and stores the information to an new and compacted fsimage file and then after copies to NameNode and renames as a edit log file. Components of YARN: Resource Manager: Resource per cluster to manage use of resources across a cluster. Node Manager is in on all nodes to launch different containers. NameNode stores the metadata of Hadoop distributed file system. NameNode execute various file system operations such as opening, closing and renaming files and directions. All details are kept in Name node, this metadata is available in active name node,Active Namenode is the primary NameNode works and runs in the cluster. Standby NameNode, which has similar metadata as active NameNode. When the active NameNode fails then the Standby NameNode takes place in the cluster. Hence, the cluster is never without a NameNode and so it never fails.

There are soft-wares that can install on the top of the Hadoop and all are open source. Hive is a data warehouse tool to process structured data. Pig is a language developed to execute queries on large datasets that are stored in Hadoop HDFS. Chukwa is a system that collects the data to manage large distributed systems. Avro is the technique to translating data structures into a format that can be stored.Spark is real time data processing with the huge data. Zookeeper is a service that maintains configuration information, deliver distributed synchronization services, and provide group services. Sqoop is the interface of the command-line application for transferring data between both Hadoop and relational databases. Phoenix is the process of converting HBase in to SQL database. Flume can move large amounts of streaming data from one to another place. Example from web servers to Hadoop cluster. Storm is a system to processing streams of data and then reliable and distributed.

The rest of the paper is organized as follows. Section II describes the various algorithms and related works for scheduling and provisioning of resources. Section III described a brief background of the proposed PSO model. Experimental setup, results and conclusions are discussed in section IV , V and section VI respectively.

II Related Work

VM method does not require dedicated server. Resource management plays biggest role in resource utilization in cloud computing. Resource utilization can be improved with VM allocation. Here author proposed that by combining different multiple resources utilization of server was improved [13]. The author states that service provider and its customers have [14] service -level –agreement for types of services available in the cloud. Scalability, reliability and cost-effectiveness of are based on a code-based framework. The novelties of project therein the following aspects. Dynamic [15] parameterization increases efficiency of existing resources. Specified error correction with localized property this minimize cost during encoding and decoding. Artificial neural network learning combined with evolutionary paradigm with help of neuro-evolution [16] approach. The work has divided as two parts namely, Job placement and then after VM placement. Job placement by the cloud broker uses the best-fit heuristic approach and then, author proposed that the VM placement is by the cloud broker that follows the worst fit of heuristic approach. Virtualization [17] technology provides instant server resources allocation. This paper describes reduction of skew value and VM cost, Effective load balancing and future load prediction prevents overload and improve server resource utilization with minimum energy[18]consumption. Hypervisor uses algorithm to choose host resources, this prevents virtual machines from running out of resources. Clustering approach made easy decision to map virtual machine to physical [19] machines. The new advent of digitized has lead to rise of volume, velocity and variety of data by passing every hour. In reality, most of the data were stored on Cloud for managing the storage system. Most of the organizations were moving their data into the Cloud. Thus, Cloud Computing will play a vital role in today's world. Cloud Computing provides remote access to the Client on basis of Pay-What-Use manner. In this case, instead of buying any physical machine's user can rent the machine for their requirements in the Cloud computing environment, provisioning resources is a demanding task, so framework of resource provisioning in Cloud for big data applications. In this framework, job scheduling done through MapReduce technique. The main objective of this job is to reduce cost and execution time. This paper [19] states that VM pool manager has variety of instances small, large and medium sizes and the instance selection as per the user requirements. The author implements nagios in openstack environment to monitor the virtual resource and instance for avoiding in balances in the cluster[20].

III. System Architecture

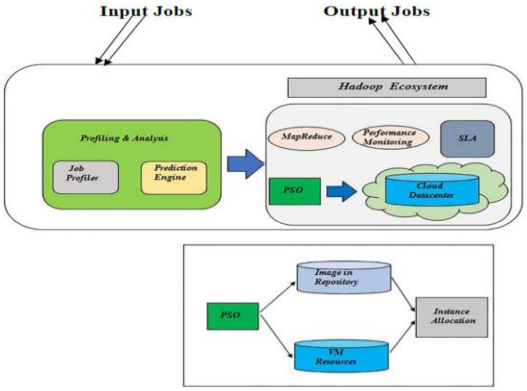


FIGURE 2. CPMC MODEL

Initially job are sent to job profiler for gathering details about the datasets. Then prediction engine predicts the given job for processing. After analysis jobs are sent to scheduler which uses particle swarm optimization (PSO) scheduling algorithm to allocate resource to execute the Job scheduled and data management was taken care by VM pool management system[12]. Finally, PSO sends output as by matching the instance with its equivalent image in repository. It understands the Input dataset maps the available set of resources in the VM pool manager dynamically. VM Pool manager contains the database which has large number of small instances and small number of large instances. Different configuration of CPU resource available in the VM database. Image repository contains various components of Hadoop VM templates available for launching of Hadoop instance in the cloud to provide Hadoop as a service. This whole functions adapted by PSO which brings cost effective to run any jobs in the Hadoop cluster

Particle Swarm Optimization

Particle Swarm Optimization [11] Scheduling algorithm is a method of assuming fitness value to the jobs or datasets are known as PVALUE among those the job with the small pvalue will be considered as pbest value and will be scheduled to the process. From this we can find total execution cost of virtual machine.

S = {R, I, D, TET, TEC}
R = {t1, t2, t3, t4,…, tn}
I = {I1, I2, I3, I4, …, In}

Where

S is Scheduling,
R is Resources,
I is Image in Repository,
t is CPU Instance,
TET is Total Execution Time,
TTEC is Total Execution Cost,
D is Dataset,
J is Type of Job.

The total execution time is directly proportional to the cost, so choosing optimal DataNode in the Hadoop cluster which reduces the cost for executing jobs.

TEC = $\sum_{i=1}^{|R|} \mathrm{CVMI} * [\frac{\mathrm{LET}-\mathrm{LST}}{t}]$ Equation 1

Where

TET ε **TEC**
LET is End time,
LST is starting time,
t is no of minutes.

$x(\mathbf{t+1}) = \mathbf{v_i}(t) + \mathbf{v_i}(t)$ Equation **2**

$\mathbf{v_i}(t+1) = w.\mathbf{v_{i(t)}} + \mathbf{c_1 r_1}(\mathbf{x_i}(\mathrm{t}) - \mathbf{x_i}(\mathrm{t})) + \mathbf{c_2 r_2}(\mathrm{x(t)} - \mathbf{x_i}(\mathrm{t}))$ Equation **3**

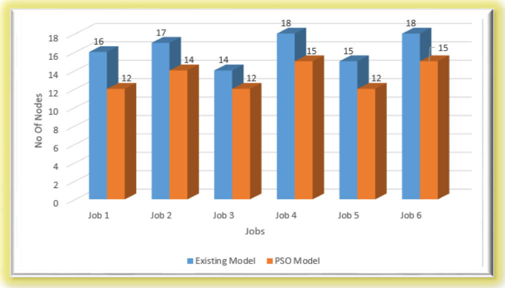
MACHINES	INSTANCES	RAM
M	SMALL	4GB
	MEDIUM	8GB
	LARGE	16GB
	XLARGE	32GB
G	SMALL	4GB
	MEDIUM	8GB
	LARGE	16GB
	XLARGE	32GB
C	SMALL	4GB
	MEDIUM	8GB
	LARGE	16GB
	XLARGE	32GB

Table.2. Instance Cost

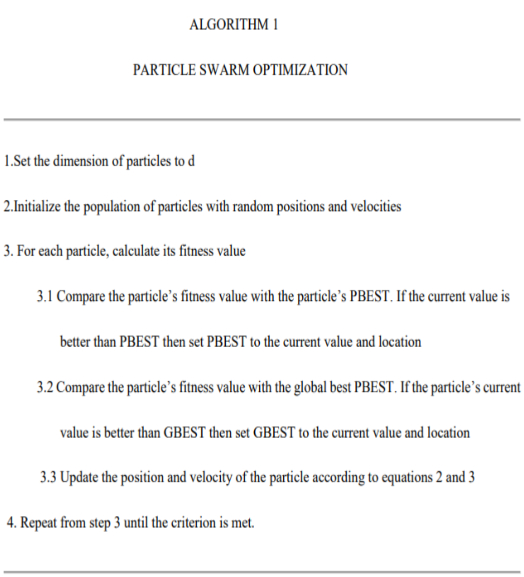
Machines	CPU	Price in Minutes
M.SMALL M.XLARGE	2.4 GHZ 2.5 GHZ	0.166 INR 0.333 INR

G.LARGE	2.7GHZ	0.5 INR
C.XLARGE	2.9GHZ	0.666 INR
C.XLARGE	3.0 GHZ	0.833 INR

Table.3. Instance Types



Graph.1. Nodes Usage in Hadoop Cluster



```
ALGORITHM 1
PARTICLE SWARM OPTIMIZATION

1.Set the dimension of particles to d
2.Initialize the population of particles with random positions and velocities
3. For each particle, calculate its fitness value
    3.1 Compare the particle's fitness value with the particle's PBEST. If the current value is
        better than PBEST then set PBEST to the current value and location
    3.2 Compare the particle's fitness value with the global best PBEST. If the particle's current
        value is better than GBEST then set GBEST to the current value and location
    3.3 Update the position and velocity of the particle according to equations 2 and 3
4. Repeat from step 3 until the criterion is met.
```

IV. Experimental Setup

In our experimental setup, We used 30 nodes of CPU configuration 2.7GHZ, 8GB RAM, I7 processor and 1TB HDD to setup cloud infrastructure for providing Hadoop Cluster as a service. We have implement PSO system in the Master nodes to select the optimal data nodes for cluster formation, Various cloud instance price and types shown on the table [1-2] which has been configured in the master node for billing purpose.

V. Results and Discussions

We have executed different set of jobs in both existing model and PSO model in Hadoop Cluster. The graph[1] shows the DataNodes used by the master node in the Hadoop cluster. The job 1 needs 12 nodes than 15 nodes. In job 2 it needs 16 nodes than 20 nodes for complete its execution. Likewise all other jobs required less number of DataNodes compare to other existing models.

VI. CONCLUSION

We proposed new model for a Hadoop cluster called as CPMC model. Which offers MapReduce as a service at low cost. In previous state of art in which allocates a resources to the Hadoop cluster at varying cost. In our model, we introduce SLA, Which compliance based on elastic characteristic. It keeps the efficient execution of jobs in Hadoop cluster. In our experimental results shows more flexible and adequate for production workloads at reduction of intra cost and also cost effective in cloud environment.

REFERENCE

[1]. Selvaprabhu, “Fragile data Storing in public cloud for hospital administration”2017 14th, VOL 5, NO14, “IEEE International Conference on Services Computing”.

[2]. .AniketMalatpure,”Testing Private Cloud Reliability Using a Public CloudValidation SaaS”,2017,IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW).

[3]. .Joonseok Park, “Pattern-based Cloud Service Recommendation and Integration for Hybrid Cloud”,Research Institute of Logistics Innovation, Volume 41, Number 1, January 2011.

[4]. Mukhtaj Khan, Yong Jin, Maozhen Li, Yang Xiang, and Changjun Jiang, “Hadoop Performance Modeling for Job Estimation and Resource Provisioning”, FEBRUARY 2016, NO.2, VOL. 27, “IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS”.

[5]. Tzu-Chi Huang, Kuo-Chih Chu, Yu-Ruei Rao, “Smart Intermediate Data Transfer for MapReduce on Cloud Computing”, Volume 1 Issue 4, 2011, “International Conference on Cloud Computing and Big Data”.

[6]. R.Thangaselvi, Ananthbabu, Aruna, Jagadeesh,” Improving the efficiency ofMapReduce scheduling algorithm in Hadoop”, Number 1, Volume 19, “IEEE ComputerSociety”.

[7]. J. Dean and S. Ghemawat, “MapReduce: Simplified data process-ing on large clusters,” in Proc. 6th Symp. Operating Syst. Des. Imple-mentation, 2004, p. 10.

[8]. G. Ananthanarayanan, S. Kandula, A. Greenberg, I. Stoica, Y. Lu, B. Saha, and E. Harris, “Reining in the outliers in Map-Reduce clusters using Mantri,” in Proc. 9th USENIX Conf. Operating Syst. Des. Implementation, 2010.

[9]. Dr. (Mrs.) Ananthi Sheshasaayee ,“ A Theoretical Framework for Cloud Resource Provisioning using MapReduce Technique “PG and Research Department of Computer Science&Application.

[10]. K. Kambatla, A. Pathak, and H. Pucha, “Towards optimizing Hadoop provisioning in the cloud,” in Proc. Conf. Hot Topics Cloud Comput., 2009.

[11]. Amol C. Adamuthe,”Solving Resource Provisioning in Cloud using GAs andPSO”,Dept. of CSE, RIT, Rajaramnagar-Islampur, MS, India ,2013 Nirma University International Conference on Engineering (NUiCONE).

[12]. Balaji Palanisamy, “Cost-Effective Resource Provisioning for MapReduce in a Cloud”, IEEETRANSACTIONS ON

PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 26, NO 5, MAY 2015.

[13]. Bramantyo Adrian, “Analysis of K-means Algorithm For VM Allocation in Cloud Computing”, Yogyakarta, Indonesia, International Conference on Data and Software Engineering, 2015.

[14]. Clayton Maciel Costa, “Service Response Time Measurement Model of Service Level Agreements in Cloud environment”, IEEE International Conference on Smart City, SocialCom together with DataCom, 2015.

[15]. Yongmei WEI, “A cost-effective and reliable cloud storage“, Nanyang Polytechnic Singapore, IEEE International Conference on Cloud Computing, 2014.

[16]. P.Varalakshmi, Maheshwari.K, “Cost-Optimized Resource Provisioning in Cloud“, Department of Information Technology, Anna University-MIT, International Conference on Recent Trends in Information Technology (ICRTIT), 2013.

[17]. Mahesh B. Nagpure,”An Efficient Dynamic Resource Allocation Strategy for VM Environment in Cloud”, Student M. tech 2nd year, Dept of Computer Science &Engineering, 2015 International Conference on Pervasive Computing (ICPC).

[18]. Vivek Rajani, “A VM Allocation Strategy for Cluster of Open Host in Cloud Environment”, Research Scholar ITSNS- GTU PG SCHOOL, 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

[19]. Berl, A., Gelenbe,“Energy-Efficient Cloud Computing”, The Computer Journal August 19 2009.

[20]. Vidhyasagar. B. S., S. Aravinda Krishnan, D. Manikkannan, and J. Arunnehru. "An Implementation and Performance Monitoring of Virtual Machines using Ganglia in Eucalyptus Private Cloud." International Journal on Computer Science and Engineering (IJCSE), 2017

Authors Profile

Mr. B.S VIDHYASAGAR pursed Bachelor of Engineering from Anna University, Chennai, India in 2007 and Master of Engineering from Anna University in year 2011. He is currently pursuing Ph.D in Anna University and currently working as Assistant Professor in Department of Computer Science and Engineering,SRM Institute of Science and Technology, Tamilnadu, Chennai since 2016. He is a member of IEEE, IET,ACM, IAENG and Indian Science Congress. He has published more than 8 research papers in reputed international journals and conferences. His main research work focuses on Cloud Computing, Big Data Analytics and IoT, He has 8 years of teaching experience and 2 years of Research Experience.

Mr. M. AJITH KUMAR pursuing Bachelor of Technology in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai, Tamilnadu. His research area of interest includes cloud and Big Data.

Mr. SHAIK SAJID pursuing Bachelor of Technology in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai, Tamilnadu. His research area of interest includes cloud and Big Data.

Mr. SYED KHADEER pursuing Bachelor of Technology in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai, Tamilnadu. His research area of interest includes cloud and Big Data.

Mr. RAHUL P pursuing Bachelor of Technology in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai, Tamilnadu. His research area of interest includes cloud and Big Data.

Dr. J. Arunnehru received his Diploma in Computer Technology from Muthiah Polytechnic College, Annamalai Nagar, in 2004. He obtained his B.E. degree in Computer Science Engineering in 2008, M.E. degree in Computer Science and Engineering in 2010, and the Ph.D. degree in Computer Science and Engineering in 2017 from Annamalai University, Annamalai Nagar. Presently, he is working as a Assistant Professor in the Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamilnadu. He has published papers in 25 internationally reviewed journals and he has presented 8 international conference papers. His area of specialization includes image and video processing, pattern classification and machine learning. He is a life member of the Indian Society of Technical Education. He received the best paper award at the Springer international conference in 2014. He is a reviewer for the Elsevier and other international reputed journals. (E-mail address: arunnehru.aucse@gmail.com)