Study of Elastic Hadoop Cluster on Private Cloud

Pratiksha D. Mandal, Madhuri S. Kadam, Sayali R. Kakade, Monali J. Reddy, Guided By - Prof. Amar More

Department Of Computer Engineering,MITAOE,Pune

Abstract- Whatever kind of industry are you in being able to obtain information based on analysis of data coming from wide variety of sources can help make better decisions. In 2004 Google developed MapReduce, a programming framework for the processing of large datasets across distributed systems. MapReduce got more popularised by open source Apache Hadoop framework. In 2009 Amazon introduced Elastic MapReduce which is used for processing large datasets efficiently using the Apache framework on the fly. It allows customer to write their MapReduce application without dealing with hardware, network and Hadoop configuration. User only needs to submit their Map and Reduce functions along with required number of nodes and in return user will get simplified data as per his specifications mentioned in the application. Issue with Amazon EMR is the usage of computing resources provided only by Amazon datacenter at a certain cost. The goal of this project will be open sourcing of Amazon Elastic MapReduce. It will add the features of Elastic MapReduce to open source private cloud.

Index Terms- MapReduce, Hadoop, Cloud Computing, Amazon EMR, Openstack.

I. INTRODUCTION

Cloud computing has become a hot topic of industry and is emerging as a new computing mechanism. Cloud Computing[1] provides various service models(IaaS, PaaS, SaaS) and various kinds of web applications are deployed to cloud computing environment, leaving petabytes of data to be processed. Various kinds of scientific calculations like data mining and information extraction from massive data facility needed to be performed by cloud computing environment. Hadoop[2] an open source project maintained by Apache Software Foundation, is an implementation of MapReduce framework and is widely used at Yahoo!, Facebook, Amazon[3] etc. Hadoop is a suitable tool to parallelly deal with massive data processing applications. Current hadoop environments are manually deployed on physical servers independently and is time consuming process. Hadoop solves the problem of data storage and retrieval problem but with the limitation that data scientist should be handy with the lots of machine over which the hadoop infrastructure can run.With the growth of cloud computing technology through which computer scientists are able to provision number of the required machines, the required infrastructure problem could be solved.Thus by combining ideas of these two technologies i.e. Cloud computing and Hadoop, the new model is emerging known as Hadoop-as-a-Service,Amazon came up with a service called as Elastic Map Reduce[4].

It solved the problem of the infrastructure requirement but Amazon is a public cloud provider and data scientists having access to private cloud are not able to utilize this service. It would be very much beneficial if we could have a service with which we would be able to run the hadoop service over the private cloud.

Thus this paper describes study and approaches to use Hadoop-as-a-Service over private cloud.We introduce an elastic MapReduce framework, providing the ability for a cluster to be dynamically elastic, i.e. expand or reduce its size on-demand, without requiring for the job to be stopped and restarted but rather paused while reconfiguration occurs. We provide the application programmer the ability to start a given job early before all resources necessary for such a job are available, and then progressively add compute nodes as they become available to the user, thus saving valuable time.

II. LITERATURE SURVEY

A. Hadoop Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. It comprises main two components i.e HDFS[5] and MapReduce[6]. HDFS stands for Hadoop Distributed File System which is a fault tolerant storage system. HDFS stores huge amount of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop MapReduce framework is the processing pillar of the Hadoop ecosystem.

There are two functions in MapReduce as follows:
Map the function takes key/value pairs as input and generates an intermediate set of key/value pairs Reduce the function which merges all the intermediate values associated with the same intermediate key Hadoop works in cluster which is in distributed format. To form a hadoop cluster is a tedious and time consuming job so the idea is to provide hadoop incubated VM's. Hadoop cluster formation [7] involves
Step 1: Change the hostname of each VM node to node0, node1, node2 and so on respectively.
Step 2: Chooses node0 to be the master node of the hadoop cluster and edits the file /etc/hosts to record the mapping of hostname of each slave node to its IP address.
Step 3: Edits the configuration files masters, slaves, core-site.xml,mapred-site.xml,hdfs-site.xml under $HADOOP HOME/conf/ to make a uniform configuration for the new hadoop cluster.
Step 4: Copy the file /etc/hosts and hadoop configuration files masters, slaves, core-site.xml, mapred-site.xml, hdfs- site.xml to the corresponding directory of each slave node by SSH.

Through the above four procedures, a new hadoop cluster is configured and ready to use.

B. Openstack

Openstack[8] is open-source cloud computing platform which provides Infrastructure-as-a-Service (IaaS) and provides tools for creating and managing virtual machines over available resources. It popular and widely used because of its flexible, scalable and open-source in nature. Researchers in [9] describe about deploying openstack over virtual and dedicated hardware. According to their research OpenStack deployed over dedicated hardware always performs better than OpenStack running over virtualized environment that means compared to one-level virtualization two-level virtualization induces significant performance overhead during computational resource usages.

C. Amazon Elastic MapReduce

Amazon Elastic MapReduce is a public cloud provider on pay as you use basis.[10]It allows you to focus on analytics instead of infrastructure as it is scalable and elastic. Amazon Elastic MapReduce (EMR) is one such service that provides fully managed hosted Hadoop framework on top of Amazon Elastic Compute Cloud (EC2). Amazon Elastic MapReduce(Amazon EMR) simplifies running Hadoop and related big data applications on AWS. It removes the cost and complexity of managing the Hadoop installation.

There are many limitations of Amazon EMR as it works on the public cloud and does not allow to use EMR service on the private cloud,due to which client is not able to use their own infrastructure for task execution and there is waste of their available infrastructure.Another limitation is client needs to pay for this service.

III. DESIGN AND IMPLEMENTATION

In this section, we introduce the design and implementation of EMR system. The overall architecture is described as Fig 1. By using virtualization technology, we integrate the underlying physical resources. Authorization head listens to clients request, once authenticated the resource manager interacts with virtualization platform to offer virtual machines from the available resource pool and deploy elastic hadoop cluster for client user.

IV. DETAILED DESIGN

1. If user want to perform mapReduce job, user must login using its userid and password.

2. Authentication and Authorization: Authentication and Fig. 1. Architectural Flow Diagram authorization can be carried out by simple checking of userid and password. But problem with this approach is userid and password can be easily theft. So to avoid this we can encrypt the userid and password. The algorithm in order to achieve this goal is Symmetric Encryption RC-6 Algorithm[11].

3. Once logged in successfully user submits its data, MapReduce function and job credentials to request handler. Request handler further negotiates with the resource manager for allocation of requested resources.

4. Resource Allocation: In above proposed system, multiple user may attempt to access resources, so proper resource scheduling algorithm must be used. Here we are using Priority Based Dynamic Resource Allocation Algorithm in Cloud Computing.This algorithm decides the priority of user at run time and then according to the priority of the user resources is allocated to the user.[12].

5. Hadoop Cluster Deployment(Creating and Destroying VM's): According to client's request, the request handler communicates with resource manager for creating or destroying of the virtual machines. If client want to deploy hadoop cluster then required no of VM's are allocated. After that according to hadoop architecture, master and multiple slave nodes are configured on the VM's. And then return the IP address of the master node ,user login and password of that VM's for accessing the cluster. once the user gets access to the cluster he can perform his mapreduce task. Further VM's can be destroyed on the user request or after the task completion.

6. MapReduce Function: Once the cluster is formed, given MapReduce function runs over the input data and output result is sent back to HDFS. Further this generated output is given back to user.

Fig. 2. Elasticity

V. CHALLENGES

This section describes about the challenges involved in deploying elastic hadoop on cloud.

A. Elasticity

Elasticity is one of the healthiest challenge which our system is going to face as it is not easy to maintain resources and processing. Elasticity is the ability of the system to scale as per the user requirements i.e. in our project, whenever need arises the user should be able to add the nodes as well as user can remove the nodes within the cluster. With whatever changes user does the system should be able to balance the loads on resources as well as processing should be better and efficient.[13][14]

1) Motivation For Elasticity:

Many people using Hadoop for data processing but it requires its cluster to be clearly defined by the time a given job is to be started. Once a job has started, Hadoop MapReduce does not allow for cluster membership changes,Nodes can neither be added nor replaced in the cluster while an application is underway. To remove the node from cluster is not user choice but rather to node failure.

_ Mapreduce cluster is static in nature, so it may result into time loss,especially for long running applications.

_ For such long running applications, node replacement become difficult due to which user is stuck with static configuration of their cluster for the entirety of the job.

_ Also, dead nodes cannot be replaced with new node rather the load gets transferred to another live peers in the cluster.

2) How To Achieve Elasticity:

In EMR system Request Handler negotiates the client request with Resource Manager who keeps track of resource pool as well as application progress running on the cluster.

Reconfiguration of the cluster can be done in two ways as per the client request or according to the size of the job for load balancing purpose. _ Adding the Node:

a) Whenever client wants to add the nodes within the cluster while task is running, needs to send the Add Request to the Request Handler. Request Handler communicates with Resource Manager, if the required resources are available, Resource Manager will process the request.

b) Hadoop cluster works in master slave architecture, Resource Manager send request to slave nodes to pause the running task and send the progress report back to master node.

c) After all slaves sends the progress report to the master node, it send request to the available node to get added in the cluster.

d) After successful addition of the node in the cluster,paused task is evenly distributed over all the nodes.

e) Mapreduce job then continues to run successfully on the cluster. _ Removing the Nodes:

a) At the same time if client want to remove the nodes from the cluster, the request is needed to be sent to Resource manager.

b) Resource manager handles the request, locates the checkpoint and on that checkpoint pauses the task running on all slaves.

c) Further, paused task status is sent back to the master node and then the suitable node gets deleted.

d) MapReduce task then again gets distributed on all the nodes and takes its running status back.

B. Handling Multiple Hadoop Clusters:

As the system works on the Hadoop and cloud concept, there will be many users accessing the EMR system simultaneously. The cluster formation will be done for each user according to the Task which is to be performed. As there are many simultaneous users, the user and the Task information must be maintained within the system and the jobs to be handled correctly by cluster manager. This is tedious task to do as handling of multiple clusters requires perfection.

VI. CONCLUSION

With the rapid development of internet applications, various kinds of web applications are deployed to cloud computing environment, leaving petabytes of data to be processed. In cloud environment, we often need to perform all kinds of scientific computing like matrix multiplication and do data mining and information extraction on massive data. Hadoop, an open-source implementation of MapReduce, is a suitable tool to deal in parallel with these kinds of applications. While current hadoop environments are manually deployed on physical servers independently and lacks flexibility. We are presenting a framework just like Hadoop MapReduce, but capable of supporting a cluster to be dynamically elastic, i.e. expand or reduce its size on-demand, without requiring for the job to be stopped and restarted but rather paused while reconfiguration occurs. It also provides the ability for the application programmer and user to start a given job early before all resources necessary for such job are available, and then progressively add compute nodes as they become available to the user. We are trying to implement Elastic MapReduce service as a new feature that would provide the same functionality as that of Amazon EMR but on the top of private cloud built on Openstack Platform. This will enable users having access to private cloud to run their MapReduce computations task without having to worry about cloud resources management, failures handling, etc that too free of cost.

REFERENCES

[1] ”Benefits and Challenges of Three Cloud Computing Service Models” Joel Gibson, Darren Eveleigh, Robin Rondeau, Qing Tan” 2012 Fourth International Conference on Computational Aspects of Social Networks(CASoN)

[2] ”Hadoop: the definitive guide - White - 2012.”

[3] ”https://en.wikipedia.org/wiki/Amazon Web Services”

[4] ”https://aws.amazon.com/elasticmapreduce/”

[5] ”Data-Intensive Computing with Map-Reduce and Hadoop” Shamil Humbetov 978-1-4673-1740-5 /12/ 31.00

[6] ”An In-depth Study of MapReduce in Cloud Environment” Novia Nurain, Hasan Sarwar, Md.Pervez Sajjad, Moin Mostakim 978-0-7695-4959-0/13 $25.00 2013 IEEEDOI 10.1109/ACSAT.2012.70

[7] ”http://doctuts.readthedocs.org/en/latest/hadoop.html”

[8] ”https://www.openstack.org”

[9] ”Deploying OpenStack: Virtual Infrastructure or Dedicated Hardware” Robayet Nasim, Andreas J. Kassler 2014 IEEE 38th Annual International Computers, Software and Applications Conference Workshops

[10] ”Amazon Elastic MapReduce Developer Guide API Version 2009-03-31”

[11] ”Web Secure Login Design With Symetric Encription RC-6 Algorithm”

[12] ”Priority Based Dynamic resource allocation in Cloud Computing”

[13] ”DELMA: Dynamically ELastic MApReduce Framework for CPUIntensive Applications” Zacharia Fadika, Madhusudhan Govindaraju 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing

[14] ”Towards Deploying Elastic Hadoop in the Cloud” Hong Mao, Zhenzhong Zhang, Bin Zhao, LiminXiao,Li Ruan 22011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery

AUTHORS

First Author – Pratiksha D. Mandale, Department Of Computer Engineering,MITAOE,Pune, p2prati2@gmail.com
Second Author – Madhuri S. Kadam, Department Of Computer Engineering,MITAOE,Pune, madhurikadam300@gmail.com
Third Author – Sayali R. Kakade, Department Of Computer Engineering,MITAOE,Pune, sayalikakade22@gmail.com
Fourth Author – Monali J. Reddy, Department Of Computer Engineering,MITAOE,Pune, monalireddy6@gmail.com
Fifth Author – Prof. Amar More, Department Of Computer Engineering,MITAOE,Pune, amarmore2006@gmail.com