Big Data: Intelligent Scaling of Query Service

HemaMalini B. H, Dr. L. Suresh. Dr. Radhika K. R.

Abstract— Cloud is a good concept of using the existing resources efficiently, but still there is lot of room for improving the efficiency. The present work will be focused on Amazon web service namely EMR service. Today the user has to pay for 1 hour even he/ she uses the service for just 20 minutes. Similarly, if the user uses the service for one hour and ten minutes, he pays for two hours. A company who does extensive data mining, running 1000 instances of data nodes, running couple of thousand jobs for a week, will be paying a huge sum of money. The proposed tool or framework will help to reuse the existing clusters, without shutting down; saving bootstrap time & making sure the jobs are queued properly for execution.

Index Terms— Big data, Cloud computing, IaaS, PaaS, SaaS, query service, private, public hybrid.

1 INTRODUCTION

FOR an executive of a large IT company, the main responsibility is to make sure all the employees have the right software and hardware to perform their duties. Just buying computers will not solve the purpose- the required software or software licences is also to be bought to provide the employees the tools to work. Whenever, new employees are hired, again more software has to be bought or make sure the current software allows another user. This is traditional computing. An alternative to this is cloud computing. Instead of loading a suite of software every time to each computer, it is required to load only one application. That application would allow the workers to log into a Web-based service which hosts all the programs the user would need for his or her job. Some other company would own the remote machines. They run everything from email to word processing and to complex data analysis programs. Thus, cloud computing change the entire computer industry.

2 CLASSIFICATION

Cloud computing can be classified into 3 types based on the deployment:

2.1 Private Cloud

- Private cloud or corporate cloud or internal cloud provides hosted services to a limited number of people behind a firewall. It is a marketing term for a proprietary computing architecture.
- The corporate network and data centre administrators have become service providers due to the advances in virtualization and distributed computing that meet the needs of their "customers" within the corporation.

2.2 Public Cloud

In public cloud, the service provider makes resources, such as storage and applications, available to the general public over the Internet. It may be free or offered on a pay-per-usage model. The advantages of using a public cloud service are:

- Scalability to meet the user requests.
- Resources are not wasted because the user pays for whatever is used.
- Inexpensive and easy set-up since the hardware, bandwidth and application cost is covered by the provider.

2.3 Hybrid Cloud

A hybrid cloud is a cloud computing environment in which an organization provides and manages some resources in-house and has others provided from outside. For example, an organization might use a public cloud service, such as Amazon Simple Storage Service (Amazon S3) for archived data but it may continue to maintain in-house storage for operational customer data.

The hybrid approach allows a business to make use of the cost-effectiveness and scalability that a public cloud computing environment offers without exposing mission-critical applications and data to third-party susceptibilities. This type of hybrid cloud is also referred to as hybrid IT.

To be effective, a management strategy for hybrid cloud deployment should address security, configuration management, fault management, change control, and budgeting. Because a hybrid cloud combines public cloud and private data centre principles, it is possible to plan a hybrid cloud deployment from either of these starting points. If a better starting point is chosen, it will be easier to address business goals.

3 SERVICES

3.1 PaaS (Platform as a Service)

In this category, the development platform itself is provided as a service. The developer uses the platform provided by the

- *HemaMalini B H is currently pursuing doctoral of philosophy program in BMSCE, CSE Research Centre, Bangalore, Visveswaraya Technological University, and is an Associate Professor, BMSIT, Bangalore, India, PH:+91-7204844628 E-mail: bhhemaraj@bmsit.in*
- *Dr. L. Suresh, is the Principal, Cambridge Institute of Technology, Visveswaraya Technological University, India, PH-+91-9686001199. E-mail: suriakls@gmail.com*
- *Dr. Radhika K. R, is a professor in department of Information Science & Engineering, BMSCE, Bangalore, India. , PH-+91-99845387862. E-mail: rkr.ise@bmsce.ac.in*

cloud. The platform will be hosted on the cloud. The platform is accessed by the user using a browser with internet connection and the user creates and runs his application on the platform. Providers: Microsoft's Azure, Google App Engine [2].

3.2 SaaS (Software as a Service)

In traditional computing, users have to buy the licensed version of the software and install it for use. In cloud computing, the software is provided as a service on the basis of pay-per-use model. It provides multi-tenant means at the backend. The same infrastructure is shared amongst multiple users while in the front end each user feels that the software is dedicated to a single user. It supports running multiple instances of the software too. Providers: Google Docs, Zoho [2].

3.3 IaaS (Infrastructure as a Service)

In IaaS, the vendor provides the infrastructure itself as a service. The infrastructure can be provided in the form of technology, data centre or IT service, such as offering CPU time on a hourly basis, considering for storage usage, as well as assessing for data transfers per gigabyte, often with differing rates for uploads against downloads. Amazon's Elastic Computing Cloud (EC2) is a good example. It is similar to traditional "outsourcing" at less cost and effort. The developer can use the infrastructure provided by the vendor for paying for it for a specified period, thus saving the enormous investment on setting up the infrastructure, and also saving maintenance cost. Providers: Amazon's S3, Sun's cloud service [2].
NIST provides on-demand self-service, resource pooling and rapid elasticity. On-demand means that the service is available to turn on or off as needed. Resource pooling means that multiple users share a bank of servers (including storage devices and other computing resources) over the Internet, as a substitute to using dedicated servers. Rapid elasticity means the cloud offering can be dramatically scaled up and down as needed. With as-a-service, you use as much as you want you only pay for whatever you use.

4 PROPOSED WORK

The proposed system will provide effective service mechanisms that are evaluated for performance.
Query Service: In the world of big data, searching or running query on a terabyte of data is a massive problem; Today Google performs search on peta bytes of data. Google stores the data permanently and they are not bothered about the cost, since they already have good operations team to maintain the infrastructure that they want. With most of the other companies, it may not be true. Companies are migrating into cloud, since it is an in-expensive affair and have less overhead of operations involved. If the data growth is unpredicted and if the data is not stored permanently then, the existing query services like Hive will not scale. It may take about days to query the data and to get the results from a 100 terabyte of data sitting in Apache HDFS (Hard Disk File System).
When it comes to cloud, keeping a 100 terabyte is very expensive. Data scientists bring the data they want to analyze only when it is required and query on them. The tools and solutions provided will be used by the Data Scientists. They query the data and extract the information from it. It these data points indicate something, it help the companies to improve their business.
Cloud is a good perception of using the existing resources proficiently, but still there is lot of scope for improving the efficiency. The proposed work will be focused on Amazon web service namely EMR service. Presently, the user has to pay for one hour even he/ she uses the service for only 20 minutes. Likewise, if the user uses the service for one hour and five minutes also, he has to pay for two hours. A software company which does massive data mining, running around thousand instances of data nodes and couple of thousand jobs for a week will be paying enormous money.
The proposed tool or framework would help to reuse the existing clusters, without shutting them down and thus saving the bootstrap time and making sure the jobs are queued properly for execution. For this a model has to be developed which can say this job will run for specific time. Depending on that it will have to choose which cluster to run the job.

5 CONCLUSIONS

Cloud Computing and Big data are buzzwords. More and more software companies are getting into cloud and are providing services for public use, because of the feature of Cloud Systems, i.e. reduced upfront cost, high availability, infinite scalability, incredible fault tolerance capability, expected performance, and so on. The proposed work will be focused on Amazon web service namely EMR service. The performance evaluation also will be done based on the amount of data queried and the number of data nodes connected. The work is planned to improve the efficiency.

6 ACKNOWLEDGMENT

The authors wish to thank honourable Dr. S. C. Sharma, Chief Mentor and Professor of Eminence, BMSIT, Principal Dr. S. Venkateswaran, Head of the department Dr. Thippeswamy G, Dr. Keshav Prasanna, Associate Professor, CSE and Prasad G. J, CEO, Data Integration System who has supported in bringing out this paper.

7 REFERENCES

[1] Minqi Zhou, Rong Zhang, DadanZeng, WeiningQian, "Services in the Cloud Computing Era: A Survey", Software Engineering Institute, East China Normal University, Shanghai 200062, China. National Institute of Information and Communications Technology, Kyoto 619-0289, Japan. IUCS 2010, 978-1-4244-7820-0/10/$26.00 ©2010 IEEE

[2] Mamoun Hirzalla,"Realizing Business Agility Requirements through SOA and Cloud Computing", DePaul University, College of Computing and Digital Media andIBM, Chicago, IL USA, 2010 18th IEEE International Requirements Engineering Conference, 1090-705X/10 $26.00 © 2010 IEEE, DOI

10.1109/RE.2010.70, pp379-380.

[3] Hailong Sun, Xu Wang, Chao Zhou, Zicheng Huang, Xudong Liu, "Early Experience of Building a Cloud Platform for Service Oriented Software Development", School of Computer Science and Engineering, Beihang University, Beijing, China, 978-1-4244-8396-9/10/$26.00 ©2010 IEEE.

[4] Michael Mattess, Christian Vecchiola, and RajkumarBuyya, "Managing Peak Loads by Leasing Cloud Infrastructure Services from a Spot Market", Cloud Computing and Distributed Computing (CLOUDS) Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Australia, 12th IEEE International Conference on High Performance Computing and Communications, 2010, 978-0-7695-4214-0/10 $26.00 © 2010 IEEE, DOI 10.1109/HPCC.2010.77, pp180-188.

[5] Li Guo, YikeGuo_ and XiangchuanTian, "IC Cloud: A Design Space for Composable Cloud Computing", Department of Computing, Imperial College London, UK, IEEE 3rd International Conference on Cloud Computing, 2010, 978-0-7695-4130-3/10 $26.00 © 2010 IEEE, DOI 10.1109/CLOUD.2010.18, pp 394- 401.

[6] SameeraAbdulrahmanAlmulla, Chan YeobYeun, "Cloud Computing Security Management", Khalifa University of Science, Technology and Research (KUSTAR), Shrjah Campus, P.O. Box 573, Sharjah, United Arab Emirates.

[7] Paul Marshall, Kate Keahey1,2 and Tim Freeman1, "Elastic Site: Using Clouds to Elastically Extend Site Resources", Department of Computer Science, University of Colorado at Boulder Boulder, CO USA, 1Computation Institute, University of Chicago, 2Argonne National Laboratory, Chicago, IL USA, 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010m 978-0-7695-4039-9/10 $26.00 © 2010 IEEE, DOI 10.1109/CCGRID.2010.80, pp 43-52.

[8] WesamDawoud, Ibrahim Takouna, ChristophMeinel. "Infrastructure as a service, Security: Challenges Solutions", HassoPlattner Institute Potsdam, Germany, Ministry of Education & Higher Education, Palestine.

[9] Ammar Khalid, "Cloud Computing: Applying Issues in Small", Department of Computer Science & IT, The Islamia University of Bahawalpur, Bahawalpur, Pakistan. International Conference on Signal Acquisition and Processing, 2010, 978-0-7695-3960-7/10 $26.00 © 2010 IEEE, DOI 10.1109/ICSAP.2010.78, pp 278-281.