Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems (NESUS PhD 2016)
Timisoara, Romania

Jesus Carretero, Javier Garcia Blas
Dana Petcu
(Editors)

February 8-11, 2016

Distributed Processing in Cloud Computing

Ilias Mavridis
Aristotle University of Thessaloniki, Greece
imavridis@csd.auth.gr

Eleni Karatza
Supervisor
Aristotle University of Thessaloniki, Greece
karatza@csd.auth.gr

Abstract

Cloud computing offers a wide range of resources and services through the Internet that can been used for various purposes. The rapid growth of cloud computing has exempted many companies and institutions from the burden of maintaining expensive hardware and software infrastructure. With characteristics like high scalability, availability and fault tolerance, cloud computing meet the new era needs for massive data processing at an affordable cost. In our doctoral research we intend to study, analyze, evaluate and make proposals in order to further improve the performance of cloud computing.

Keywords Cloud computing

I. Introduction

Cloud computing has evolved into a major computing platform that is used by many companies. By using cloud computing, companies offer their services or process their data without the need of in-house IT infrastructure [1]. The term of cloud computing usually refers to providing computational services as utilities via the Internet [2]. These services may include infrastructure, platform and software. The increasing use of cloud computing can be explained by the fact that cloud offers "on-demand" scalability, high availability, flexible cost policy, ease of customization and other elements that positions it ahead of classic distributed technologies such as the Grid [1].

The aim of this thesis is to address open issues and limitations in cloud computing and propose techniques in order to overcome the potential obstacles and improve the performance of cloud computing. Through the doctoral research we will study the current bibliography and we will conduct several experiments to analyze and evaluate the current cloud computing technologies.

II. Ongoing Study

At the first phase of our research we investigated the use of main memory in cloud computing and we studied how it affects the computation performance. We analyzed and compared the widespread cloud computing framework Hadoop[3] with the relatively new general engine for large-scale data processing Spark[4]. Spark (unlike Hadoop's MapReduce) uses effectively the main memory and claims that can achieve up to one hundred times higher performance for certain applications compared to Hadoop's MapReduce [4].

In order to experimental evaluate the two frameworks we developed and executed log file analysis application in both frameworks. Log file analysis in cloud was proposed and investigated by many papers [5] - [15] for various reasons. Also many big companies like Facebook, Amazon, ebay, etc. use cloud computing solutions to analyze the enormous amount of log data that they produce. However to the best of our knowledge this is the first work that investigates and compares the performance of real log analysis applications in Hadoop and Spark.

In bibliography there are many papers that investigate the performance of cloud computing from different perspectives and explore how various factors affect

it [16] - [25]. To evaluate the performance of the two frameworks we focus on three performance indicators. The execution time, resource utilization and scalability. The experimental results showed that Spark presents almost the same scalability as Hadoop but Spark is significantly faster and makes better resource utilization than Hadoop.

The output of this study is published in the proceedings of the Second International Workshop on Sustainable Ultrascale Computing Systems (NESUS 2015) in Krakow, Poland; paper entitled "Log File Analysis in Cloud with Apache Hadoop and Apache Spark" [26] and an extended version of this work is submitted to an international journal.

III. Related Work

As we mentioned before in order to evaluate the performance of Hadoop and Spark we developed log file analysis applications in both frameworks. After an extensive search in bibliography we found that cloud computing for log analysis has been investigated and proposed by many papers, however the majority of them studied and proposed Hadoop-based algorithms and systems.

In papers [5] - [9] the authors recognized that logs are produced in higher rate than traditional systems can serve. To overcome the bottleneck of massive data processing of traditional relational databases they proposed and implemented log file analysis using Hadoop cluster.

The paper [10] presents a Hadoop-based log analysis system for intrusion detection and in [11] a MapReduce log analysis algorithm was used to identify security threats and problems. In both works they used Hadoop MapReduce in order to improve the response time of large log files analysis applications and as a result to achieve a faster reaction by the system's administrator.

In [12] the authors implemented a MapReduce-based framework for anomaly detection that follows a specific methodology to analyze log files. First, it collects logs from each node of the monitored cluster to the analysis cluster. Then, it applies K-means clustering algorithm to integrate the collected logs. Finally executes a MapReduce-based algorithm to parse these clustered log files.

A Hadoop-based flow logs analyzing system was proposed in paper [13]. This system uses for log analysis a new script language called Log-QL, which is a SQL-like language that was translated and submitted to the MapReduce framework. After experiments the authors concluded that their distributed system is faster and can handle much bigger datasets compared to a centralized system.

Paper [14] presents a scalable platform named Analysis Farm, for network log analysis with fast aggregation and agile query. To achieve storage scale-out, computation scale-out and agile query, OpenStack was used for resource provisioning, and MongoDB for log storage and analysis.

A cloud platform for log data analysis with the combination of Hadoop and Spark was presented in paper [15]. The authors proposed a cloud platform with batch processing and in-memory computing capabilities by using at the same time Hadoop, Spark and Hive/Shark. They claim that the proposed platform managed to analyze logs with higher stability, availability and efficiency than standalone Hadoop-based log analysis tools.

IV. Thesis idea

Cloud computing has been a focused area of research in the last years and there is still a great research interest in cloud computing. In our research we will study the state of the art cloud technologies and we will deal with open issues. As we continue our research we will study current trends in cloud computing and we will identify and try to propose solutions to problems.

V. Conclusion and future work

At the beginning of our research we dealt with the effective use of main memory in cloud computing and we studied how it can significantly improve its performance. We will continue our research in different areas of cloud computing with the goal of further improve the cloud performance.

Acknowledgment

We would like to acknowledge the contribution of the academic cloud service okeanos [27] for giving us the ability to create the necessary virtual machines for the above case study. We would also like to acknowledge the contribution of the COST Action IC1305 NESUS (Network for Sustainable Ultrascale Computing).

References

[1] I.A. Moschakis and H.D. Karatza, "A meta-heuristic optimization approach to the scheduling of Bag-of-Tasks applications on heterogeneous Clouds with multi-level arrivals and critical jobs," *Simulation Modelling Practice and Theory*, Elsevier, vol. 57, pp. 1-25, 2015.

[2] G.L. Stavrinides and H.D. Karatza, "A cost-effective and QoS-aware approach to scheduling real-time workflow applications in PaaS and SaaS clouds," in *3rd International Conference on Future Internet of Things and Cloud (FiCloud'15)*, Rome, Italy, August 2015, pp. 231-239.

[3] `http://hadoop.apache.org/`

[4] `http://spark.apache.org/`

[5] B. Kotiyal, A. Kumar, B. Pant and R. Goudar, "Big Data: Mining of Log File through Hadoop," in *IEEE International Conference on Human Computer Interactions (ICHCI'13)*, Chennai, India, August 2013, pp. 1-7.

[6] C. Wang, C. Tsai, C. Fan and Sh. Yuan, "A Hadoop based Weblog Analysis System," in *7th International Conference on Ubi-Media Computing and Workshops (U-MEDIA 2014)*, Ulaanbaatar, Mongolia, July 2014, pp. 72-77.

[7] S. Narkhede and T. Baraskar, "HMR log analyzer: Analyze web application logs over Hadoop MapReduce," *International Journal of UbiComp (IJU)*, vol.4, no.3, pp. 41-51, 2013.

[8] H. Yu and D.i Wang, "Mass Log Data Processing and Mining Based on Hadoop and Cloud Computing," in *7th International Conference on Computer Science and Education (ICCSE 2012)*,Melbourne, Australia, July 2012, pp. 197.

[9] H. Kathleen and R. Abdelmounaam, "SAFAL: A MapReduce Spatio-temporal Analyzer for UNAVCO FTP Logs," in *IEEE 16th International Conference on Computational Science and Engineering (CSE)*, Sydney, Australia, December 2013, pp. 1083-1090.

[10] M. Kumar and Dr. M. Hanumanthappa, "Scalable Intrusion Detection Systems Log Analysis using Cloud Computing Infrastructure," in *2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Tamilnadu, India, December 2013, pp.1-4.

[11] S. Vernekar and A. Buchade, "MapReduce based Log File Analysis for System Threats and Problem Identification," in *Advance Computing Conference (IACC), 2013 IEEE 3rd Internationa*, Patiala, India, February 2013, pp. 831-835.

[12] Y. Liu, W. Pan, N. Cao and G. Qiao, "System Anomaly Detection in Distributed Systems through MapReduce-Based Log Analysis," in *3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, Chengdu, China, August 2010, pp. V6-410 - V6-413 .

[13] J. Yang, Y. Zhang, S. Zhang and Dazhong He, "Mass flow logs analysis system based on Hadoop," in *5th IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT)*, Guilin, China, November 2013, pp. 115-118.

[14] J. Wei, Y. Zhao, K. Jiang, R. Xie and Y. Jin, "Analysis farm: A cloud-based scalable aggregation and query platform for network log analysis," in *International Conference on Cloud and Service Computing (CSC)*, Hong Kong, China, December 2011, pp. 354-359.

[15] X. LIN, P. WANG and B. WU, "Log analysis in cloud computing environment with Hadoop and Spark," in *5th IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT 2013)*, Guilin, China, November 2013, pp. 273-276.

[16] J.Conejero, B. Caminero and C. Carron, "Analysing Hadoop Performance in a Multi-user IaaS Cloud," in *High Performance Computing and Simulation (HPCS)*, Bologna, Italy, 21-25 July 2014, pp. 399 - 406.

[17] G. Velkoski, M. Simjanoska, S. Ristov and M. Gusev," CPU Utilization in a Multitenant Cloud," in *IEEE EUROCON 2013*, Zagreb, Croatia, 1-4 July 2013, pp. 242-249.

[18] L. Gu and H. Li, "Memory or Time: Performance Evaluation for Iterative Operation on Hadoop and Spark," in *IEEE 10th International Conference on High Performance Computing and Communications and 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC EUC)*, Zhangjiajie, China, 13-15 Nov. 2013, pp. 721-727.

[19] P.R. Magalhaes Vasconcelos and G. Azevedo de Araujo Freitas, "Performance analysis of Hadoop MapReduce on an OpenNebula cloud with KVM and OpenVZ virtualizations," in *9th International Conference for Internet Technology and Secured Transactions (ICITST)*, London, 8-10 Dec. 2014, pp. 471-476.

[20] Eug. Feller, Lav. Ramakrishnan and Chr. Morin, "Performance and energy efficiency of big data applications in cloud environments: A Hadoop case study," *Journal of Parallel and Distributed Computing Special Issue on Scalable Systems for Big Data Management and Analytics*, vol. 79âĂŞ80, pp. 80âĂŞ89, May 2015.

[21] B.G. Batista, J.C. Estrella, M.J. Santana, R.H.C. Santana and S. Reiff-Marganiec, "Performance Evaluation in a Cloud with the Provisioning of Different Resources Configurations," in *2014 IEEE World Congress on Services (SERVICES)*, Anchorage, Alaska, June 27-July 2 2014, pp. 309-316.

[22] B. El Zant and M. Gagnaire, "Performance evaluation of Cloud Service Providers," in *2015 International Conference on Information and Communication Technology Research (ICTRC2015)*, Paris, France, May 17-19 2015, pp. 302-305.

[23] J. Gao, P. Pattabhiraman, B. Xiaoying and W.T. Tsai, "SaaS Performance and Scalability Evaluation in Clouds," in *2011 IEEE 6th International Symposium on Service Oriented System Engineering (SOSE)*, Irvine, USA, 12-14 Dec. 2011, pp. 61-71.

[24] T. Jiang, Q. Zhang, R. Hou, L. Chai, S.A. Mckee, Z. Jia and N. Sun, "Understanding the behavior of in-memory computing workloads," in *2014 IEEE International Symposium on Workload Characterization (IISWC)*, Raleigh, USA, 26-28 Oct. 2014, pp. 22-30.

[25] T.C. Chieu, A. Mohindra and A.A. Karve, "Scalability and Performance of Web Applications in a Compute Cloud," in *2011 IEEE 8th International Conference on e-Business Engineering (ICEBE)*, Beijing, China, 19-21 Oct. 2011, pp. 317-323.

[26] I. Mavridis and E. Karatza, "Log File Analysis in Cloud with Apache Hadoop and Apache Spark," in *Second International Workshop on Sustainable Ultrascale Computing Systems (NESUS 2015)*, Krakow, Poland, 10-11 Sept. 2015, pp. 51-62.

[27] `https://okeanos.grnet.gr`