Resource Usage Characterization for Social Networks Analytics on Spark

[Extended Abstract]

Irene Manotas
University of Delaware
imanotas@udel.edu

Rui Zhang
IBM Research - Almaden
ruiz@us.ibm.com

Min Li
IBM TJ-Watson Research Center
minli@us.ibm.com

Renu Tewari
IBM Research - Almaden
tewarir@us.ibm.com

Dean Hildebrand
IBM Research - Almaden
dhildeb@us.ibm.com

ABSTRACT

Platforms for Big Data Analytics such as Hadoop, Spark, and Storm have gained large attention given their easy-to-use programming model, scalability, and performance characteristics when processing large scale data in parallel. Along with the wide adoption of these big data platforms, Online Social Networks (OSN) have evolved as one of the major sources of information given the large amount of data being generated in a daily basis from different online communities such as Twitter, Facebook, Flickr, etc. Considering the growing demand for social network applications and their considerable data generation characteristics, it is important to understand if there exist special data patterns and resource demands on systems processing data generated by OSN. One of the basic activities to test and analyze the performance of big data solutions is achieved by running benchmarks that allow both to stress cluster resources and to evaluate the underlying platform under various conditions. Currently, different benchmarks such as BigDataBench, BigBench, SparkBench, etc., have been proposed as a way to evaluate big data systems when applications from different domains and purposes are executed. These benchmarks provide a variety of workloads, ranging from micro-benchmarks (e.g., terasort, grep, etc.), passing through domain-specifc systems (e.g., retailer systems), up to workloads that execute different types of algorithms. However, none of these benchmarks consider the evaluation of big data platforms for OSN data. Therefore, the evaluation of resource utilization for systems and platforms doing OSN analytics is missing. Given this lack of OSN-based benchmark evaluation for big data platforms, we have characterized the resource utilization and performance of Spark in the context of OSN analytics by using two popular social networks datasets as input to two workloads.

The selected workloads represent popular graph-based algorithms used in the analysis of OSN data. The first workload consists of the Page Rank algorithm [2]. Page Rank is a link analysis algorithm which computes the importance of nodes of a graph by considering the number of edges between nodes. Page Rank can be used for Social Network analysis to rank the nodes of a social graph. Some benchmarks, such as BigDatabench [8] and SparkBench [6], include Page Rank in their workload suite. Nevertheless, their performance analysis is based on a web graph dataset i.e., Google Web Graph [3], which is not an OSN dataset and is relatively small, that is, its size is about 70MB. We present the resource and data patttern characterization of Page Rank when using two different input data sizes from two different Social Networks (i.e., Live Journal and Twitter).The input data ranges from 1GB to 25GB and represents the social graph of the Live Journal [4] and Twitter [5] social networks, respectively. The second workload is related to text analytics of OSN data. Specifically, we used the Latent Dirichelt Allocation (LDA) algorithm [1] to identify topics in a set of public tweets extracted from Twitter during the end of the NFL season in 2013 and 2014 [7].

Our results include the characterization of these two workloads in terms of CPU, Disk, Memory and Network, and the data access patterns i.e., number of tasks, shuffle reads/writes, etc. In terms of Disk utilization, for Page Rank different I/O disk patterns are observed when the input data size increases; writes are dominant for the smallest dataset, while reads operations are more common for the largest dataset. On the contrary, LDA exhibits a constant read/write pattern representative of the shuffle operations originated by the term-topic distribution updates in each iteration. CPU utilization is not saturated neither for Page Rank nor for LDA. Memory utilization presents a peak in Page Rank when the input size is small, but when the input size increases the memory is saturated and its utilization is stable; For LDA, memory follows an increasing utilization pattern that follows the creation of RDDs to support the analysis of documents in the corpus. Network I/Os are bursty in Page Rank, while for LDA the network I/Os present a more uniform pattern. Finally, we present future directions for the analysis of OSN data on Big Data platforms.

1. REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, pages 107–117. Elsevier Science Publishers B. V., 1998.

[3] SNAP Datasets. Google web graph, 2002.

[4] SNAP Datasets. Livejournal social network, 2006.

[5] Haewoon kwak, Changhyun Lee, and Sue Moon. Twitter social graph, 2009.

[6] Min Li, Jian Tan, Yandong Wang, Li Zhang, and Valentina Salapura. Sparkbench: A comprehensive benchmarking suite for in memory data analytic platform spark. In *Proceedings of the 12th ACM International Conference on Computing Frontiers*, pages 53:1–53:8. ACM, 2015.

[7] Tech Tunk. Super bowl tweets, 2014.

[8] Lei Wang, Jianfeng Zhan, Chunjie Luo, Yuqing Zhu, Qiang Yang, Yongqiang He, Wanling Gao, Zhen Jia, Yingjie Shi, Shujie Zhang, Chen Zheng, Gang Lu, K. Zhan, Xiaona Li, and Bizhu Qiu. Bigdatabench: A big data benchmark suite from internet services. In *High Performance Computer Architecture (HPCA), IEEE 20th International Symposium on*, pages 488–499, Feb 2014.