Reproducibility Report for ACM SIGMOD 2020 Paper: “Automating Incremental and Asynchronous Evaluation for Recursive Aggregate Data Processing”

THOMAS HEINIS, Imperial College London, United Kingdom

The results could by and large be reproduced on a hardware set up provided by the authors of the paper.

1 INTRODUCTION

This document comments on reproducing the paper "Automating Incremental and Asynchronous Evaluation for Recursive Aggregate Data Processing" [1] authored by Qiange Wang, Yanfeng Zhang, Liang Geng from Northeatern University China and Rubao Lee, Xiaodong Zhang, Ge Yu, Hao Wang from The Ohio State University.

The results of the paper have been successfully reproduced on a hardware and software installation provided by the authors.

2 SUBMISSION

The submission primarily contained a GitHub repository https://github.com/Wangqge/PowerLogae.git which contains all relevant bits, that is: scripts, code and data used. The setup also requires installation of a number of software packages such as Postgres — all of which is outlined in detail in the reproducibility submission.

However, setting up Hadoop without root privileges — required to reproduce the results using the scripts — turned to be impossible and so reproduction was carried out on a cluster and a software setup provided by the authors of the paper.

3 HARDWARE AND SOFTWARE ENVIRONMENT

The hardware and software setup proved to be difficult to replicate (given limited permissions as the scripts require sudo access without password) and the authors of the paper therefore provided access to their system.

The set up is therefore exactly the same as described in Section 6.2 of the paper. Specifically, the hardware and distributed setup is as follows:

The distributed cluster has 17 Aliyun ECS nodes. Each node is an "ecs.r5.xlarge" instance that features 4 vCPUs and 32GB memory with Ubuntu 16.04 LTS OS. The network bandwidth between nodes is 1.5 Gbps/s. Each instance is configured with 100GB disk storage. One node is dedicated as the master and the others are configured as workers. Each worker uses 4 parallel threads.

The software setup on the nodes is as follows:

PowerLog runs on Ubuntu 16.04 and Ubuntu 14.04 (but is expected to run correctly on other Linux distributions). The tested MPI version is openmpi-3.0.0, the JDK version is jdk1.8.0$_1$61, $theHadoopversionis$2.6.5.

4 REPRODUCIBILITY EVALUATION

4.1 Process

Reproduction of the results proved to be fairly easy based on the scripts provided by the authors. Some of the scripts, or rather the executable invoked, exited abnormally. However, this could be addressed, as written in the authors notes, by running scripts to clean up the environment.

A script was provided for each of the datasets which had to be run to reproduce a data point.

The results of the scripts were a bit cryptic and it took a while to find the line reporting on the execution time amidst all the debug messages but then again all these messages added a nice realism to the experiments.

4.2 Results

All major results (i.e., Figures 9-11) in the paper were reproduced using the scripts and setup provided. The execution time of each script varied somewhat as is to be expected. By and large timings were off ± 5-10%, however, the relative runtimes (between approaches and datasets tested) and the trends were the same as reported in the paper.

REFERENCES

[1] Qiange Wang, Yanfeng Zhang, Hao Wang, Liang Geng, Rubao Lee, Xiaodong Zhang, and Ge Yu. 2020. Automating Incremental and Asynchronous Evaluation for Recursive Aggregate Data Processing. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD '20).*