Reproducibility Report for ACM SIGMOD 2021 Paper: "ExDRa: Exploratory Data Science on Federated Raw Data"

YANNIS FOUFOULAS, National and Kapodistrian Univ. of Athens and Athena Research Center, Greece
YANNIS IOANNIDIS, National and Kapodistrian Univ. of Athens and Athena Research Center, Greece

The core thesis of the paper was successfully reproduced, as the greatest majority of the experiments and figures presented in the paper were successfully recreated.

1 INTRODUCTION

The current report targets the reproducibility evaluation of ExDRa [1], which is authored by Sebastian Baunsgaard, Matthias Boehm, Ankit Chaudhary, Behrouz Derakhshan,Stefan Geißelsöder, Philipp M. Grulich, Michael Hildebrand, Kevin Innerebner,Volker Markl, Claus Neubauer, Sarah Osterburg, Olga Ovcharenko, Sergey Redyuk, Tobias Rieger, Alireza Rezaei Mahdiraji, Sebastian, Benjamin Wrede, and Steffen Zeuch. The authors are affiliated with the following organizations:

- Siemens AG; Berlin/Erlangen, Germany
- DFKI GmbH; Berlin, Germany
- Graz University of Technology; Graz, Austria
- Technische Universität Berlin; Berlin, Germany

The core thesis of the paper evaluated is successfully reproduced.

2 SUBMISSION

The authors submitted a link to a GitHub repository[1] that contains all the necessary material (code, scripts, data generators, documentation), including a list with the required dependencies (e.g., Java 8, Maven 3.6+, git, rsync, intel MKL). Installation of the required software ("*the requirements*") is done manually by the reproducer, nevertheless, this does not present a major problem as it involves widely used and known packages. The repository provides scripts to install SystemDS on all machines, generate the datasets, and distribute them to the cluster. There is one script per experiment (i.e., LAN and WAN experiments, Local experiments, and experiments with other tools) as well as scripts to synchronize the results, create the plots, and compile the paper. Finally, the submission included an sh file[2], which needs to be edited by the reproducer to setup the experiments.

3 HARDWARE AND SOFTWARE ENVIRONMENT

Reproduction of the experiments took place on a local machine and the same cluster used for the paper, which consists of 8 machines. Their specifications are shown in table 1.

Table 1. Hardware & Software Environment

	Cluster machines	Local machine
CPU	AMD EPYC 7302 16-Core Processor	Intel(R) Xeon(R) CPU E5-2630 v4
cores	16	20
MHz	3000	3100
RAM	132GB	128GB
Storage	SSD	SSD

[1]https://github.com/damslab/reproducibility/tree/master/sigmod2021-exdra-p523
[2]https://github.com/damslab/reproducibility/blob/master/sigmod2021-exdra-p523/experiments/parameters.sh

4 REPRODUCIBILITY EVALUATION

4.1 Process

In this section, we outline the reproducibility process, which consists of several steps:

- Setup hardware/cluster
- Install requirements
- Generate datasets
- Run algorithms
- Produce figures and PDF

Setup hardware/cluster. Since we did not have a cluster that would cover the minimum requirements documented by the authors of the paper, we requested access to their own infrastructure, which we received after fifteen days, as the authors had to resolve firewall and routing issues that had emerged. Receiving the authors' feedback, we produced an ssh config file to obtain access to the cluster and setup the connectivity between the main node and the worker nodes.

Install requirements. The automatic installation of SystemDS produced several errors, which were communicated to the authors. The errors proved to be due to incompatibilities between current and previous Java versions in the cluster, which led to incompatibilities of some Python code of SystemDS with the newest versions, as well as some other minor mishaps. All these issues were resolved by the authors, who updated the repository appropriately.

Generate datasets. Dataset generation failed the first time with "Host key verification failed" errors. In discussions with the corresponding author, it was proposed to first manually launch ssh connections from any potential master node to any potential worker node, as a handshake, so that the keys would be added to the .ssh/known_hosts file. After applying this, the data generation scripts worked successfully.

Run algorithms. The scripts to run the algorithms worked successfully from the beginning.

Produce figures and PDF. This process faced several problems in the following sequence. (i) Figure creation failed the first time with "Command 'python' not found". The solution was to find and replace "python" occurrences with "python3" as this was the new version installed at the time of reproducibility evaluation. (ii) Pandas import from CSV generated a "Specified a sep and a delimiter; you can only specify one." error. After discussions with the authors, we looked at a requirements file in the repository[3], which includes the necessary versions for the python requirements and downgraded Pandas to version 1.1.0. After this, the scripts creating the figures worked successfully. (iii) Some data points were not appearing in the figures produced. We discussed the issue with the corresponding author and it turned out that, for different algorithms, different lines in the parameters.sh file had to be commented out for the scripts to run properly. In particular, each one of the FFN and CNN algorithms had to be treated separately. By rerunning the scripts multiple times with appropriate commenting, the full figures were produced.

4.2 Results

The core thesis of the paper, as captured by the greatest majority of its experimental result figures, was successfully reproduced. However, the authors' guidance was required for two figures:

- Figure 5. Federated execution of L2SVM decelerated when the number of nodes increased. According to the authors, this was due to different versions of SystemDS being required to create and distribute the datasets and a different one being used in the paper's experiments. The solution was to edit the parameters.sh file and update the systemdshash value after

[3]https://github.com/damslab/reproducibility/blob/master/sigmod2021-exdra-p523/experiments/requirements.txt

distributing the datasets, reinstall SystemDS, and rerun the experiment, which resulted in the original figures.

- Figure 7. Tensorflow ran faster than Local at FFN. According to the authors, this difference from the paper was an issue cold vs hot caches. We repeated the experiment with an appropriate setup and the results were, then, reproduced accurately.

5 SUMMARY

Reproduction of the paper's results was largely achieved and supported the paper's core thesis. During the reproducibility process, the authors cooperated well with the reproducers, providing almost always timely responses to every issue that occurred. Their scripts were not tested in other clusters with the appropriate specifications, since we did not have access to one such. Nevertheless, we used successfully two different Ubuntu local machines for the WAN experiments. According to the authors, the system itself is supported on Mac and Windows but their scripts are based on bash, so it was not possible to have the system tested for portability under Windows (this could work only if installing a Linux subsystem). Finally, the documentation did not include automated scripts to test the algorithms with other datasets.

REFERENCES

[1] Sebastian Baunsgaard, Matthias Boehm, Ankit Chaudhary, Behrouz Derakhshan, Stefan Geißelsöder, Philipp M Grulich, Michael Hildebrand, Kevin Innerebner, Volker Markl, Claus Neubauer, et al . 2021. ExDRa: Exploratory Data Science on Federated Raw Data. In Proceedings of the 2021 International Conference on Management of Data. 2450–2463.