Reproducibility Report for ACM SIGMOD 2021 Paper: "At-the-Time and Back-in-Time Persistent Sketches"

ILIA PETROV, Reutlingen University, Germany

In this report we describe the reproducibility effort for Zhao et al. "*At-the-Time and Back-in-Time Persistent Sketches*" [1]. The publicly available codebase and the instructions are detailed and very well-organized. The results of this effort indicate a high level of reproducibility.

1 INTRODUCTION

In this report we describe the reproducibility effort for "*At-the-Time and Back-in-Time Persistent Sketches*" (ATTP/BITP sketches) [1] which is a joint work of *Benwei Shi*, *Zhuoyue Zhao*, *Yanqing Peng*, *Feifei Li* and *Jeff M. Phillips*, University of Utah Salt Lake City, USA.

This work introduces the novel *At-The-Time Persistent (ATTP)* sketches and *Back-In-Time Persistent (BITP)* sketches. The former allows for approximately and efficiently querying the data at a specific point in the past, while the latter enables approximate queries from an older time point to the present state. ATTP/BITP sketches [1] are space-efficient and their accuracy comes with strong theoretical guarantees.

2 SUBMISSION

The reproducibility submission comprises the codebase of ATTP/BITP sketches [1], a set of configuration files and matrix data generators together with download and dataset preparation scripts for the *world cup* dataset. *Vertica 10 Community Edition*, which needs to be downloaded externally, is used as an additional system. To assist the user the reproducibility package contains configuration and import scripts. As a guide and base for comparison, the authors also provide their own result logs and prepared log data. Furthermore, *Jupyter notebooks* for graph generation are also provided.

Codebase. The codebase [2] of ATTP/BITP sketches [1] (available under https://github.com/Yanqing-UTAH/ATTPCode). The code package comprises a set of python scripts and C++ sources. Most of the scripts are located under the `ATTP HOME` directory and are easy to access from there. It requires C++ 17, which is relatively new and the respective libraries are only supported by newer Linux kernels. The use of Ubuntu is recommended. The codebase is well-organized and easy to use.

Readme and Instructions. The ATTP/BITP package [2] with a well-defined and detailed description in a *readme* file (available under https://github.com/Yanqing-UTAH/ATTPCode/blob/master/repro_readme/repro_readme.pdf). It contains a detailed description of the repeatability procedure.

Datasources. The reproducibility relies on two types of datasets. Firstly, the *FIFA World Cup 98* dataset for the ATTP/BITP heavy hitter and ATTP frequent direction experiments. It can be downloaded by invoking the script ./data_proc/world-cup/prepare_data.sh in the `ATTP home` directory. Secondly, generated *big, medium, and small* matrix datasets. Noticeably, the datasets require 300GB-400GB of free space.

3 HARDWARE AND SOFTWARE ENVIRONMENT

The original paper [1] reports that the experiments have been performed on two CPU types. The reproducibility package [3] defines three types of experimental nodes with their precise parameters. The hardware used for the present reproducibility effort is comparable to their *Node Type 2* [3] although not identical. A brief comparison to our hardware is provided in Table 1.

Table 1. Hardware & Software environments

	Repository [3]	Reproducibility Review Setup
CPU	Intel Xeon E5-1650 v3	2× AMD Opteron 4386 (2 sockets/NUMA nodes)
cores	6	4 (per socket)
GHz	3.6	3.8
RAM	128GB (DDR4-2133)	64 GB (DDR3-1333)
Storage	HDD 7200RPM 1TB	HDD 7200RPM 2TB
OS	Ubuntu	Debian
OS Version	18.04.6 LTS	SMP Debian 5.10.92-1

4 REPRODUCIBILITY EVALUATION

4.1 Process

The users are asked to make some preparatory steps prior to running the repeatability experiments, i.e. installing and configuring *Vertica CE*, downloading and preparing the *FIFA World Cup 98* dataset for the ATTP/BITP heavy hitter, and ATTP frequent direction experiments, as well as generating the matrix datasets. Noticeably, this step requires 300GB-400GB of free space.

Running the repeatability experiments is well automated by a `driver` script located in the `ATTP HOME` directory. The authors provide a predefined set of benchmarking configurations located in the `config` directory, that can be adapted to configure different scenarios.

The authors also provide scripts extracting data from the benchmarking logs and preparing it for creating the graphs. These are located under `plot` directory. Furthermore, this directory contains *Jupyter notebooks* for the graph generation. Because of the difference in the operating systems (Table 1, the reproducibility system runs on Debian instead of Ubuntu) some of the scripts run with minor glitches.

4.2 Results

We were able to reproduce the large majority of the experimental results of [1] and this deemed results reproducible. Table 2 shows a detailed view of our results. We attribute the deviations to the difference in the operating systems that were used.

Table 2. Hardware & Software environments

Figure from [1]	reproducible	Figure from [1]	reproducible
3 right	fully	9	partially
5	fully	10 left	fully
6	fully	10 right	partially
7	almost fully	11	fully
8	fully	12(c)	fully
		16	fully

REFERENCES

[1] Benwei Shi, Zhuoyue Zhao, Yanqing Peng, Feifei Li, and Jeff M. Phillips. 2021. At-the-Time and Back-in-Time Persistent Sketches. In *Proceedings of the 2021 International Conference on Management of Data*. Association for Computing Machinery, New York, NY, USA, 1623–1636. https://doi.org/10.1145/3448016.3452802

[2] Zhuoyue Zhao and et al. 2022. ATTP, BITP GitHub Repository. https://github.com/Yanqing-UTAH/ATTPCode.

[3] Zhuoyue Zhao and et al. 2022. ATTP, BITP Readme. https://github.com/Yanqing-UTAH/ATTPCode/blob/master/repro_readme/repro_readme.pdf.