Reproducibility Report for ACM SIGMOD 2021 Paper: "Adaptive Compression for Fast Scans on String Columns"

JOHN PAPARRIZOS, University of Chicago
CHUNWEI LIU, University of Chicago

The work on the reproducibility of this project is praiseworthy. All required dependencies and build steps are carefully noted in the provided git repository. A series of scripts allows to automatically rerun the experiments, reproduce the results, and recreate some of the plots in the paper. The reproduced results are similar to the values reported in the paper and, importantly, all relationships between the compared methods are maintained.

1 INTRODUCTION

This is a reproducibility report for the paper [1]. To summarize, the central results and claims of the paper are supported by the submitted experiments. The key figures have been reproduced accurately enough. The reproducibility scripts are easy to use and well-documented.

2 SUBMISSION

The reproducibility submission consists of detailed instructions on project dependencies and how to rerun the experiments with Makefile and Python scripts acting as a command-line entry point for the reviewer. Several Python scripts are provided for running experiments and recreating results. Paper figures can be generated automatically by figure scripts with detailed log files.

The submission contains:

- Github repository with code and scripts at: https://github.com/johnfouf/SIGMOD_REPRODUCABILITY
- Data sources at: https://drive.google.com/file/d/13voW0OmvHjPhWxPNkQ-GvTCcs41ACf73/view?usp=sharing

3 HARDWARE AND SOFTWARE ENVIRONMENT

Table 1 describes the resources used in the original paper used and our reproducibility effort.

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	Intel i7-4790	Intel(R) Xeon(R) Gold 6240R
Cores	4	24
GHz	3.60	2.40
RAM	16GB	192GB
Platform	Ubuntu 20.04	Ubuntu 20.04
g++	10	10

4 REPRODUCIBILITY EVALUATION

4.1 Process

The experiments are reproduced on the five datasets (EDGAR Log File Dataset, TPC-DS, Microsoft Academic Graph, Yelp, and OpenAIRE) attached to this submission. The scripts run the proposed approaches and other baselines sequentially on those datasets. The plot scripts parse the produced

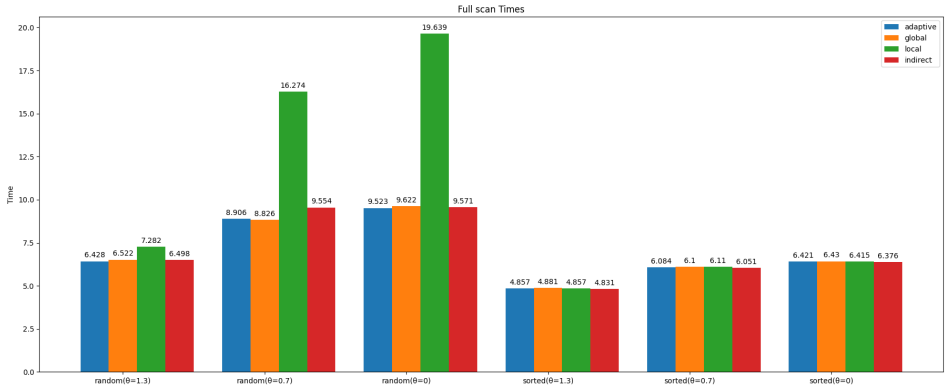


Fig. 1. Full scan (corresponding to original paper Fig 5a top)

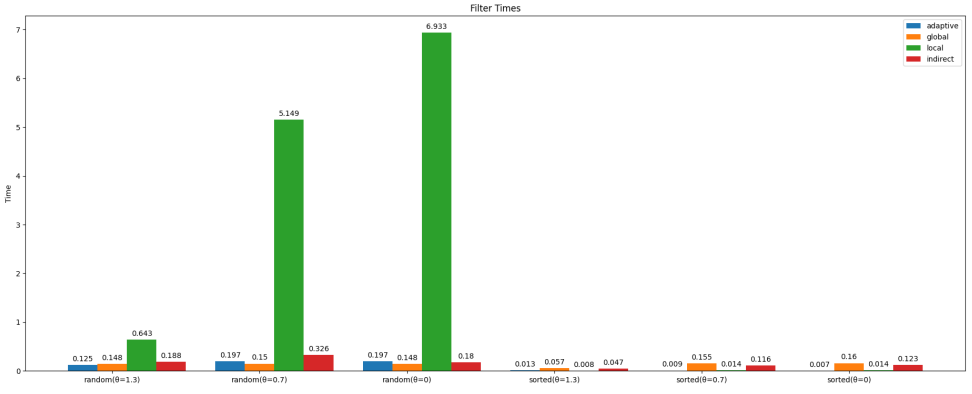


Fig. 2. Random value filtered scan (corresponding to original paper Fig 5a bottom)

files and generate the figures shown in the paper. It was possible to follow the reproducibility instructions without the authors' help.

4.2 Results

The following figures and tables have been reproduced: Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7 and Table 2. The obtained numbers and the visual plots appear to be close enough to the paper's reported values. The deviation is attributed to the differences in hardware. Most importantly, the relationships between different baselines' performances match the ones reported and discussed in the paper.

5 SUMMARY

The major figures have been reproduced on the reproducibility platform. The ideas, claims, and findings supported by these figures are therefore reproduced as well.

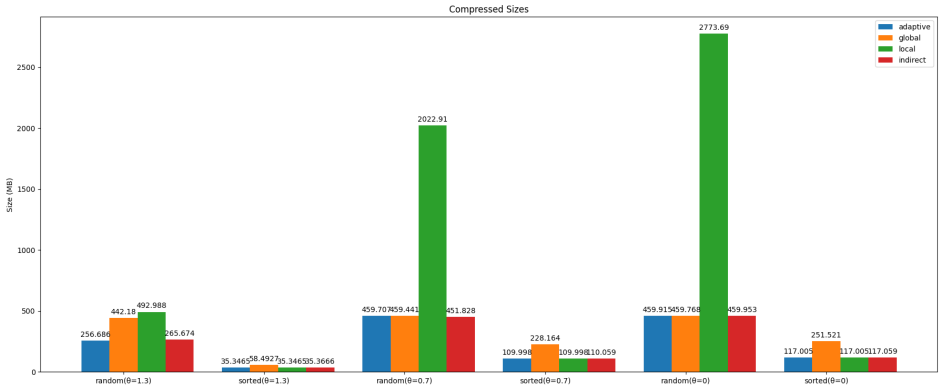


Fig. 3. Compressed size (corresponding to original paper Fig 5b top)

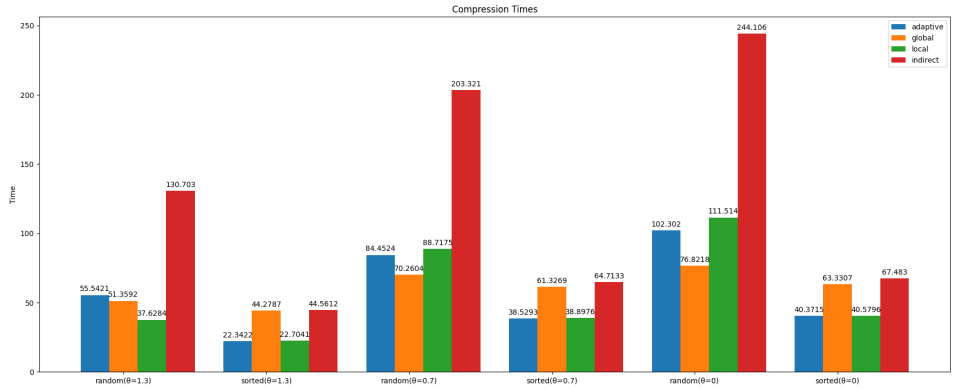


Fig. 4. Encoding time (corresponding to original paper Fig 5b bottom)

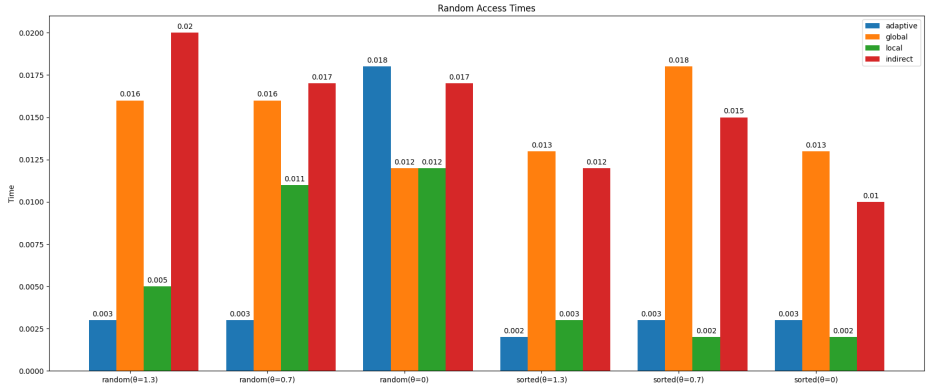
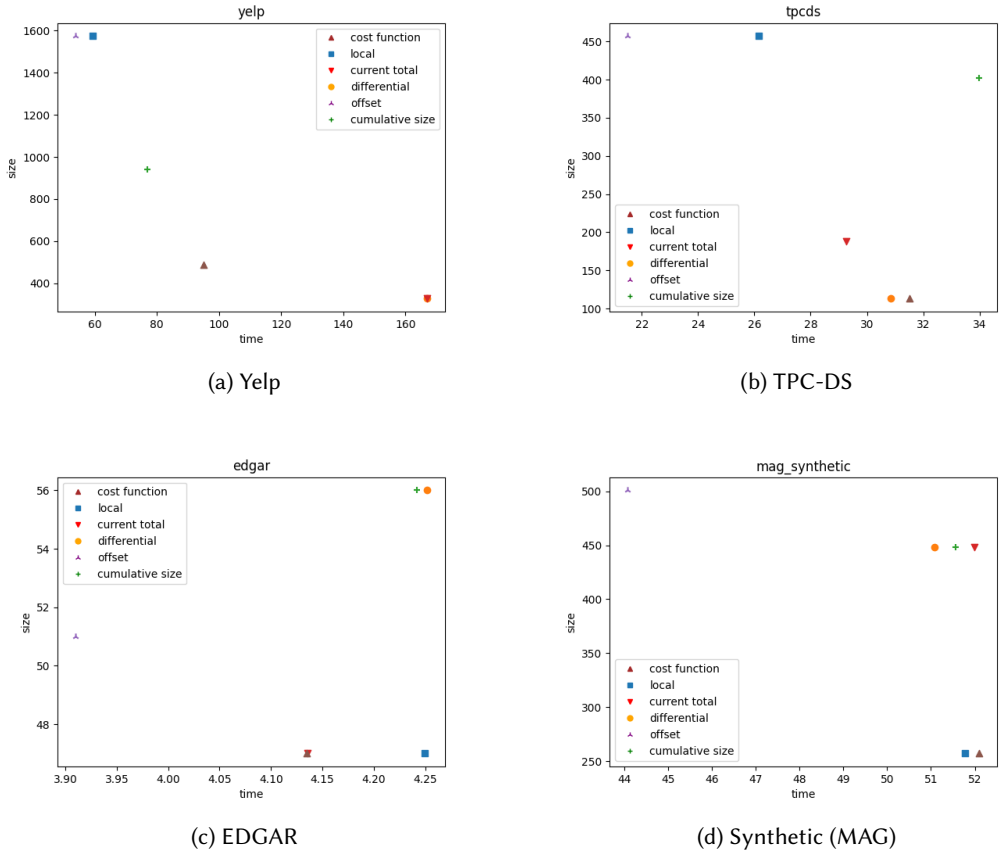


Fig. 5. Random access (corresponding to original paper Fig 6)

value	format	time	I/O	omit pages/offset scan
98.6.98.djf	diff	97187	93481	169/190
98.6.98.djf	diff/no omit	116830	111596	0/359
98.6.98.djf	global	167215	123361	9/0
98.6.98.djf	global/no omit	165973	123929	0/0
98.6.98.djf	local	224132	201061	9/352
98.6.98.djf	local/no omit	172920	151951	0/361
101.81.231.jde	diff	112455	105084	36/293
101.81.231.jde	diff/no omit	117418	109190	0/329
101.81.231.jde	global	165328	122999	0/0
101.81.231.jde	global/no omit	165845	122933	0/0
101.81.231.jde	local	172302	148615	0/336
101.81.231.jde	local/no omit	176167	154239	0/336

Table 2. Page omittance for value (corresponding to original paper Table 2 and 3)



(a) Yelp

(b) TPC-DS

(c) EDGAR

(d) Synthetic (MAG)

Fig. 6. Compression size and encoding time for different encoding strategies (corresponding to original paper Fig 7)

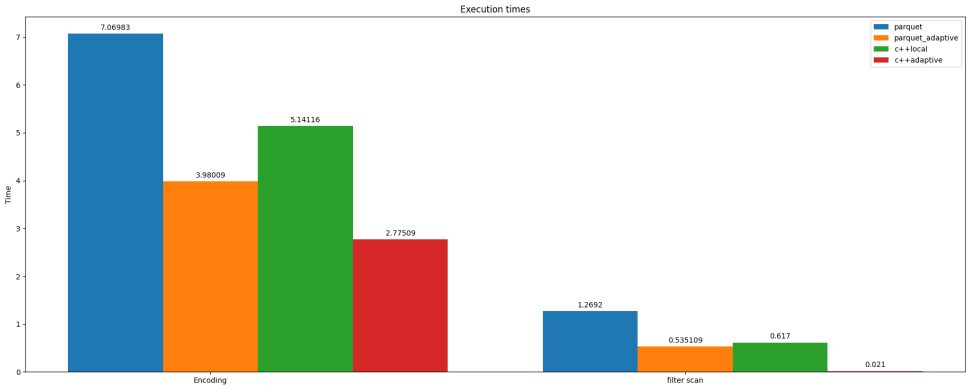


Fig. 7. Parallel Scan Execution time (corresponding to original paper Fig 8)

REFERENCES

[1] Yannis Foufoulas, Lefteris Sidirourgos, Eleftherios Stamatogiannakis, and Yannis Ioannidis. 2021. Adaptive Compression for Fast Scans on String Columns. In *Proceedings of the 2021 International Conference on Management of Data*. 554–562.