DATA ANALYSIS

BY NIMO BEEREN, MARIA POGODAEVA, HENRIQUE DIAS, PANAGIOTIS BANOS, ÇAĞLA SÖZEN, GABRIELA SLAVOVA

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

GHTORRENT DATASET

GHTorrent dataset contains 11 million records of GitHub metadata, such as user id, country, state, company and many more

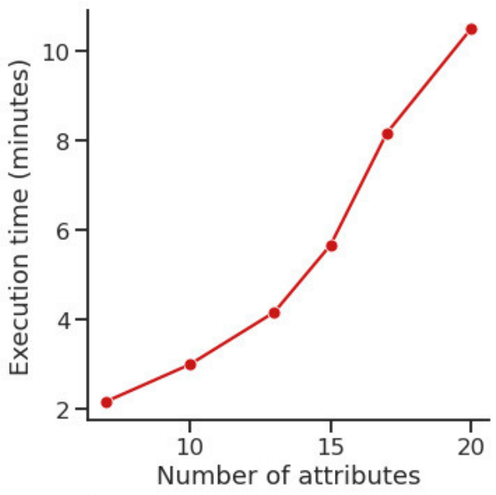
OPTIMIZATIONS

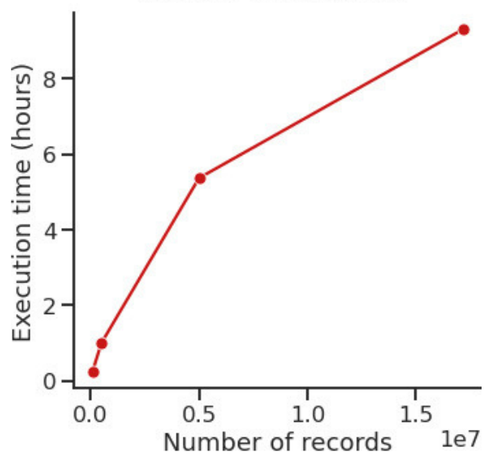
- **Sampling:** Reduce the number of FD candidates that will be checked on the full dataset by filtering out candidates with sampling
- **Batching candidates**: Group and process candidate sets of same size across all machines
- **Purge non minimal Functional Dependencies**
- **Δ-comparisons:** Optimize for cases of numerical and cases of string comparison.

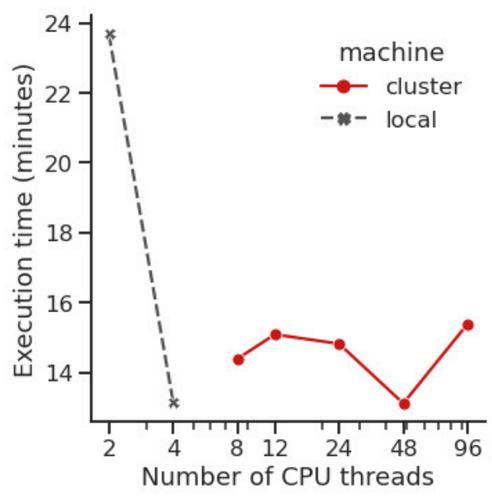
FINDINGS

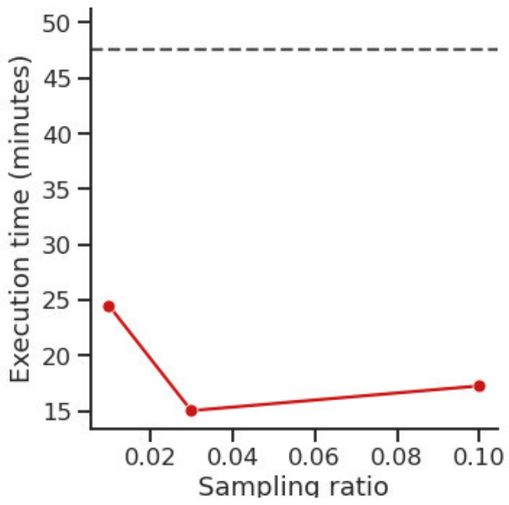
Running time

- grows sub-linearly for the # of rows and exponentially with # of attributes.
- decreases almost by half with sampling and increases as the sampling ratio increases.
- decreases as the # of cores increase locally
- does not change as the # of cores change on the HPC







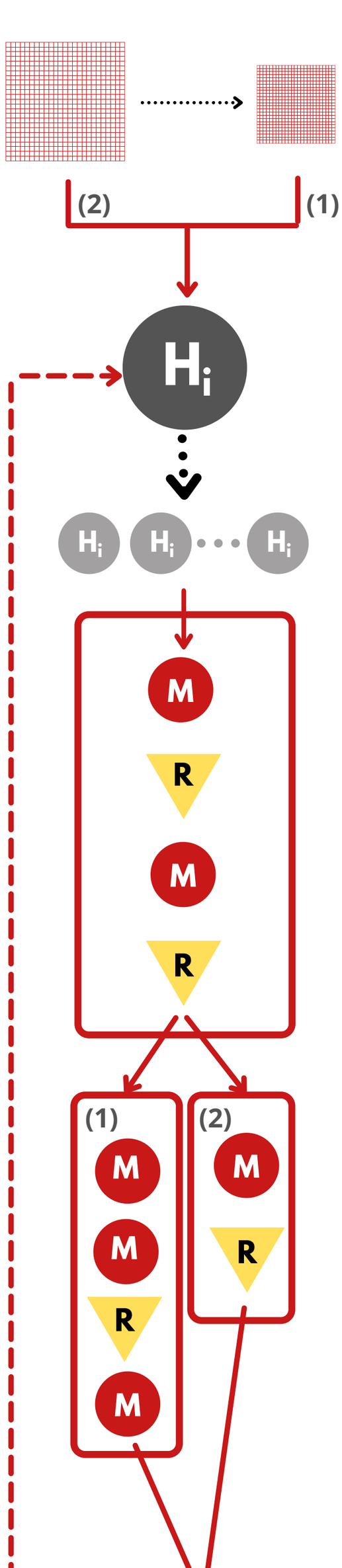


some of the

DETECTED DEPENDENCIES

H *country_code → country*
S *deleted → type*
δ *location → type*
S *location → city*
S *company → state*
S *state → country*

DISTRIBUTION OF COMPUTATION



M Map R Reduce

THE ALGORITHM IS FIRST APPLIED TO SAMPLES AND AFTERWARDS TO THE COMPLETE DATASET

¡ H :SET OF ALL HYPOTHETICAL FUNCTIONAL DEPENDENCIES A→B WITH # OF LHS ATTRIBUTES

SPLIT H_i INTO SMALLER BATCHES AND SEND TO SPARK BATCH BY BATCH

MAP EVERY SINGLE RECORD INTO A ((LHS,RHS),1) PAIR

SUM ALL NON UNIQUE RECORDS TO ((LHS,RHS),C)

MAP EACH PAIR ((LHS, RHS), C) TO A PAIR (LHS, (RHS, C))

REDUCE ALL PAIRS (LHS, (RHS, C)) BY KEY LHS TO A PAIR (LHS, S) WITH S = (RHS_1, C_1), (RHS_2, C_2), ...)

FOR δ - FDS

MAP ALL PAIRS OF RECORDS BASED ON THEIR LHS. IF THEIR Δ-COMPARISON IS WITHIN THE THRESHOLD, WITH A TRUE VALUE, FALSE OTHERWISE

AGGREGATE ALL PAIRS AND CALCULATE THEIR CONJUNCTION. IF THE AGGREGATED VALUE IS TRUE, THEN A δ-FD EXISTS

FOR HARD & SOFT - FDS

MAP ALL PAIRS TO ANOTHER PAIR BASED ON THEIR LHS VALUE AND THE PROBABILITY OF RANDOMLY CHOOSING TWO RECORDS WITH THE SAME LHS AND RHS.

MAP ALL NEWLY CREATED PAIRS TO ANOTHER PAIR WHERE THE PROBABILITY IS WEIGHTED BY THE TOTAL # OF RECORDS WITH THE LHS OF THE PAIR

AGGREGATE ALL PAIRS AND MAP THE AGGREGATED VALUES TO CALCULATE THE SOUGHT PROBABILITY ON THE WHOLE DATA FROM ALL WORKERS.

***Dataset is available at: https://drive.google.com/drive/folders/16xy-5SntyM0atfJNQh3Ognw2XZcrdZHz**
****Video available at: https://drive.google.com/file/d/1LCZ_tnd1CGerpOpIhOdNZGuuilQkImQ6/view?usp=sharing**

Video avaiable at: