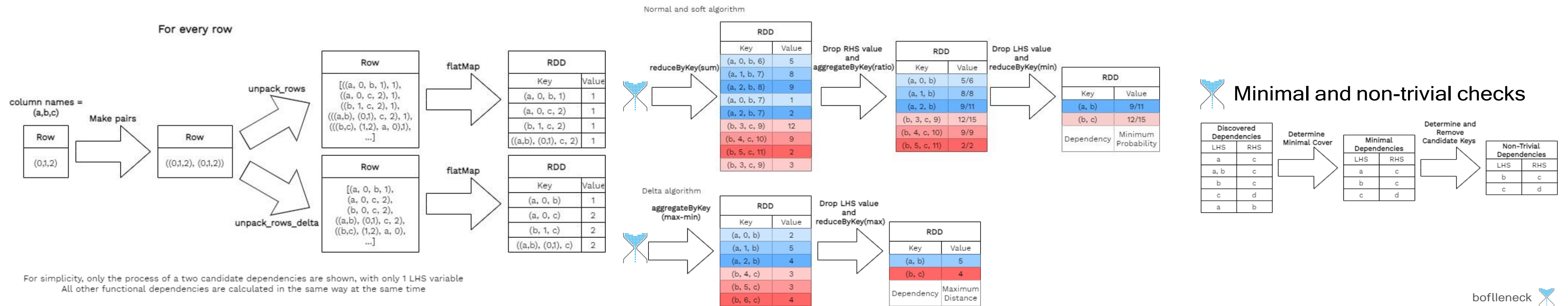
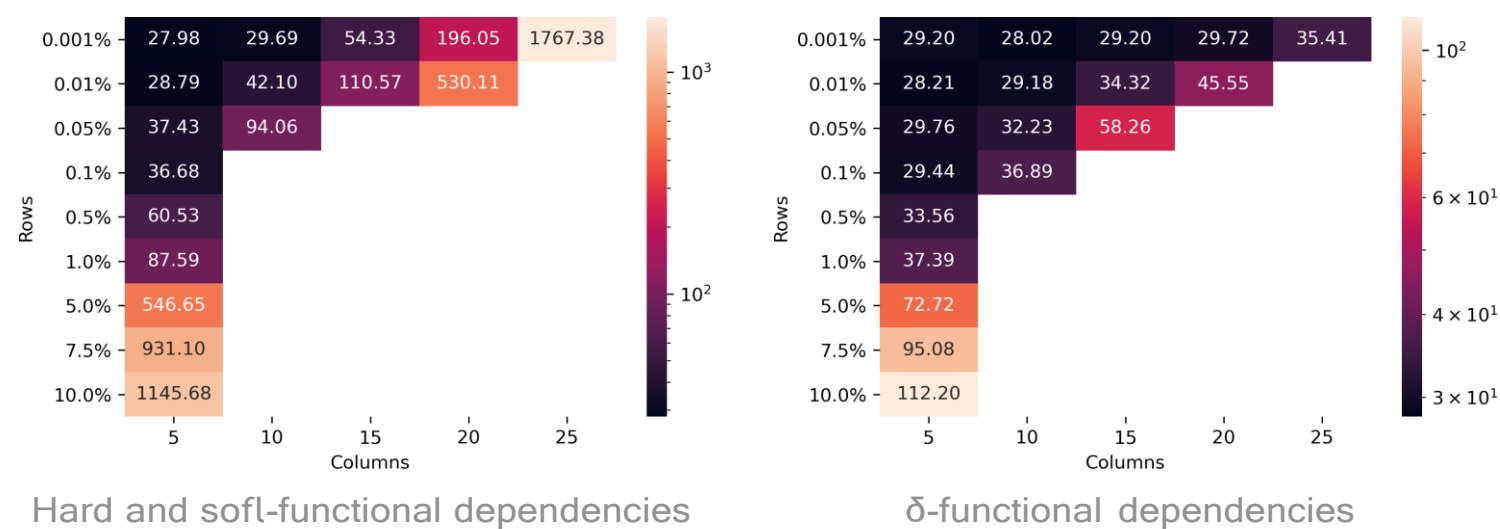


Voters Dataset | Big Data Management | Milestone 2

distribution of computation



experimental performance and scalability



Execution time for different combinations of number of rows and columns. Time in seconds, blank spaces represent a failed run.

CPU: Ryzen 7 3700X 8 Cores / 16 Threads
RAM: 32GB DDR4 3600Mhz
Disk: SSD NVME Western Digital SN550 2400 MB/s

detected dependencies

hard dependencies

- {house_num, last_name} -> {voter_status_reason_desc}
- {first_name} -> {state_cd}
- {first_name, street_name} -> {mail_state}
- {first_name, zip_code} -> {res_city_desc}
- {county_desc} -> {county_id}
- {first_name, street_name} -> {voter_status_desc}
- {midl_name} -> {state_cd}
- {county_id, last_name} -> {county_desc}
- {midl_name, voter_status_reason_desc} -> {voter_status_desc}
- ...

soft dependencies for $\tau=0.8$

- {county_id} -> {county_desc}
- {voter_status_reason_desc} -> {voter_status_desc}
- {state_cd} -> {mail_state}
- {voter_status_desc} -> {mail_state}
- ...

delta dependencies for $\delta=25$

- {age_group} -> {age}
- {sex} -> {salary}
- {age, voter_reg_num} -> {county_id}
- {last_name, paky_desc, street_name} -> {county_id}
- {county_desc, ncid} -> {zip_code}
- {midl_name, res_city_desc, street_name} -> {zip_code}
- {first_name, house_num, race_desc} -> {salary}
- {house_num, last_name, midl_name} -> {county_id}
- {county_id, house_num, last_name} -> {zip_code}
- ...

interesting unexpected dependencies (blue)
expected dependencies (grey)