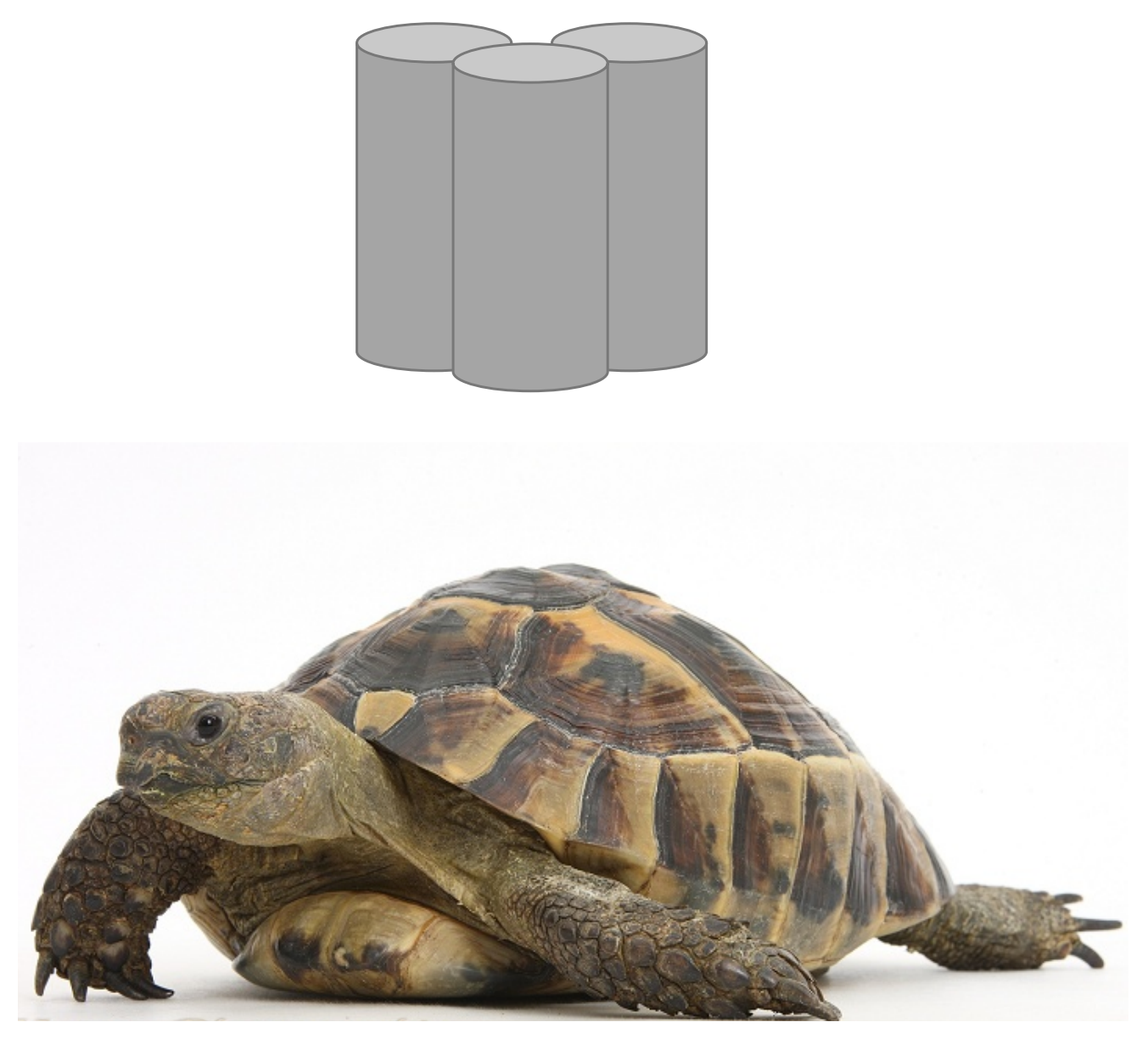
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Adaptive Denormalization

Zezhou (Alex) Liu, Stratos Idreos

Normalization



⊕ less storage/update costs

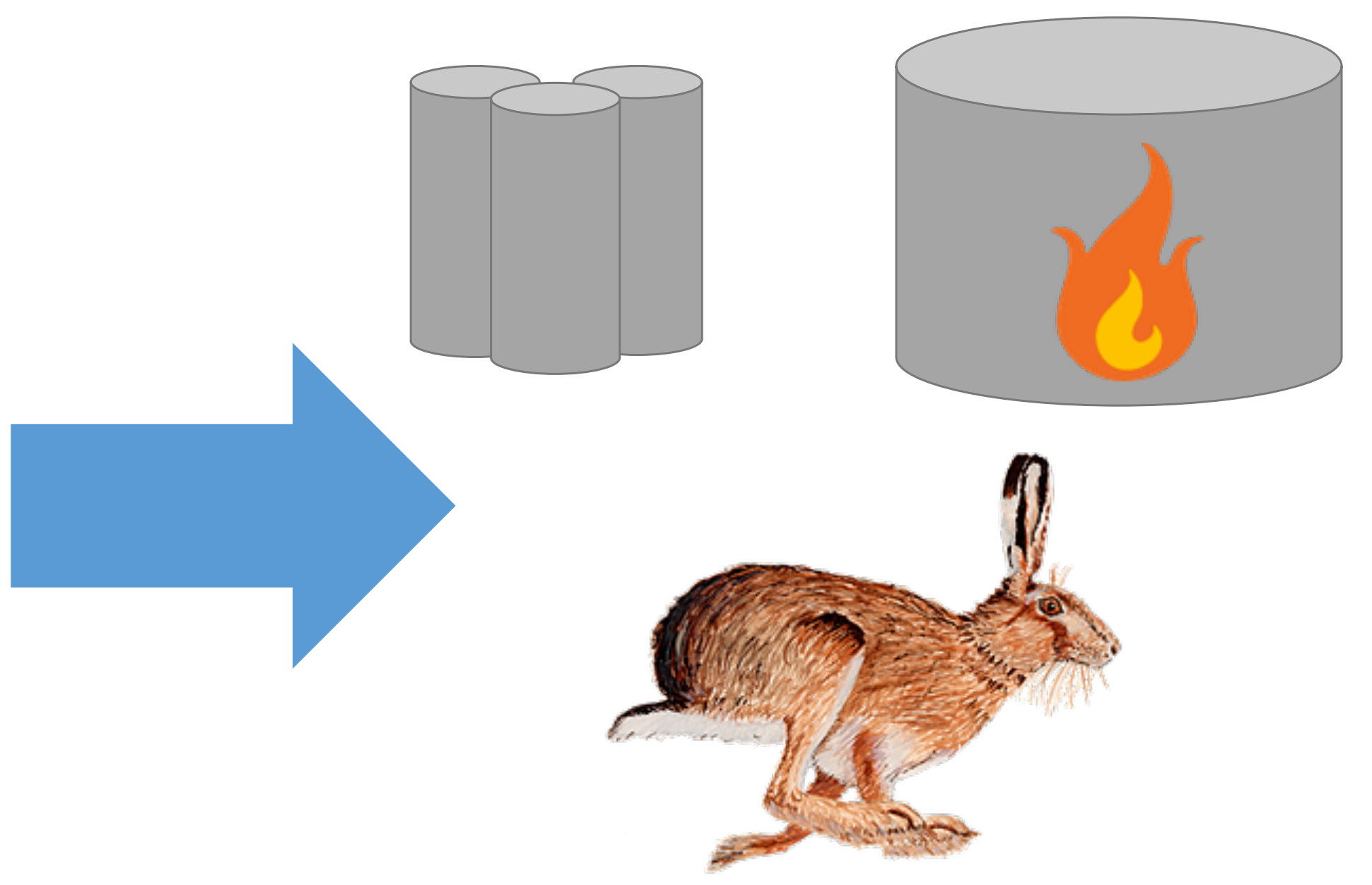
⊖ slow queries (joins)

Denormalization

⊖ more storage/update costs

⊕ fast queries (scans, no joins)

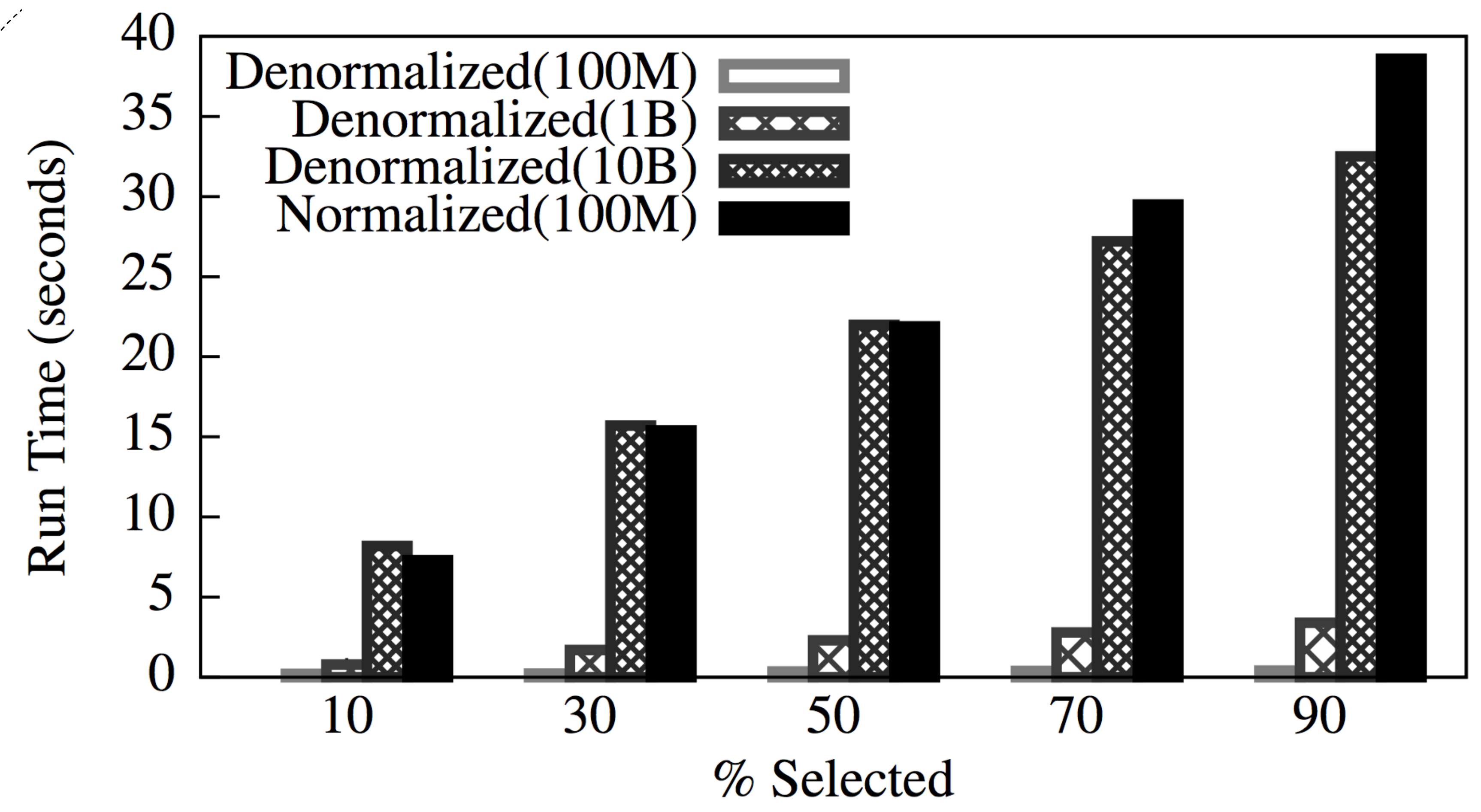
Adaptive Denormalization



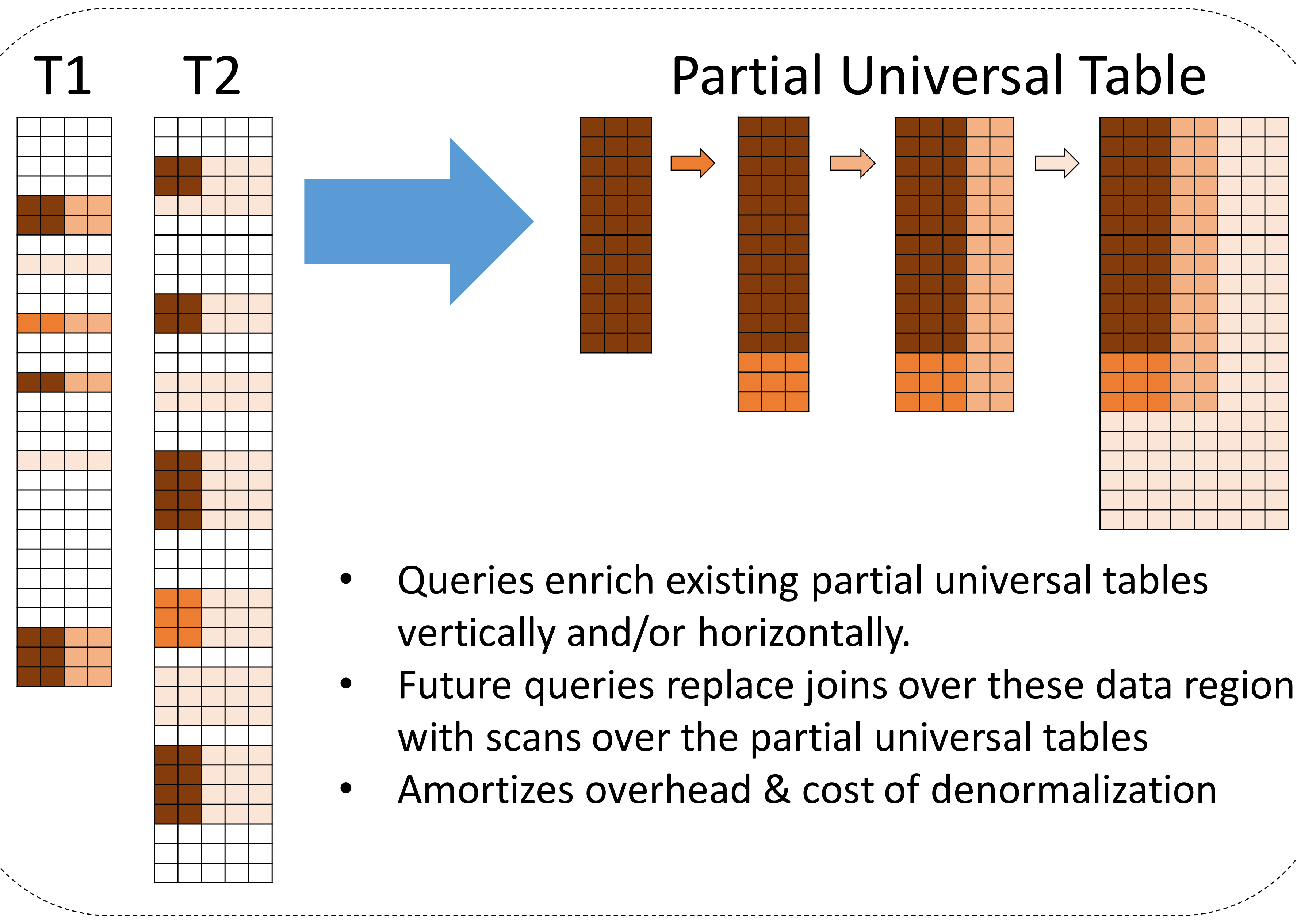
⊕ less storage/update costs

⊕ fast queries (scans)

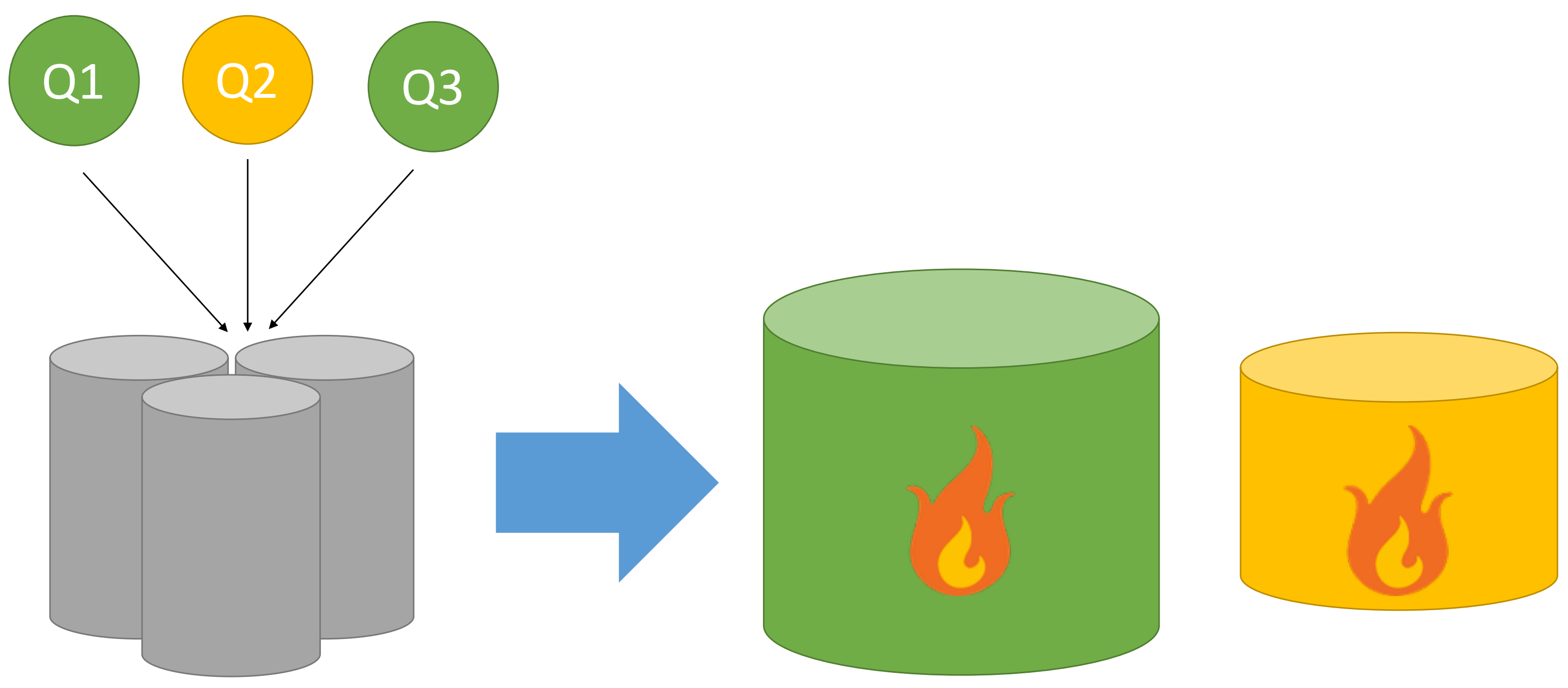
- base data lies in a normalized state
- hot data is adaptively and partially denormalized on-demand
- enables the advantages of both normalization and denormalization
- future queries can benefit from faster query processing over the denormalized data
- still maintains the efficient space utilization, updates, and loading time characteristics found in normalized schemas



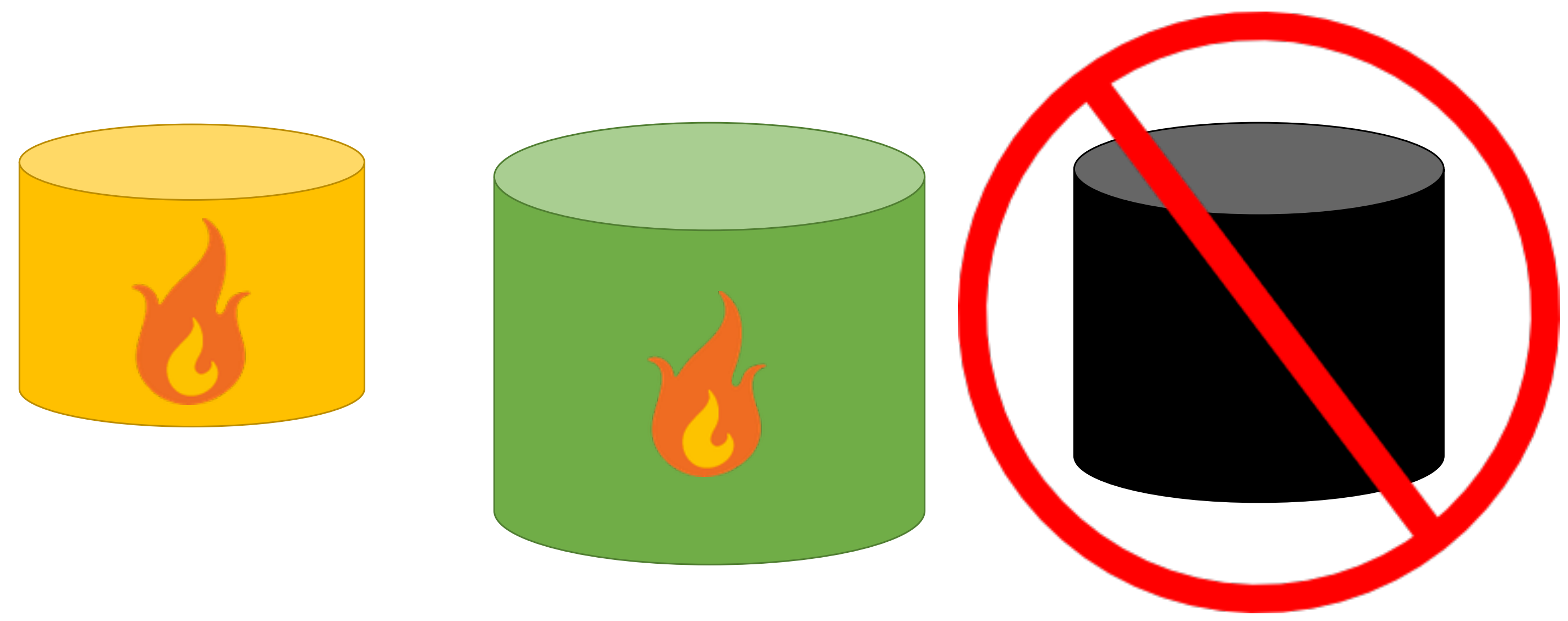
In the time it takes to join inputs of 100 million rows in a normalized schema, we can perform a (logical) join by scanning over 10 billion rows in denormalized schema. The disparity is larger when a higher percentage of rows are selected.



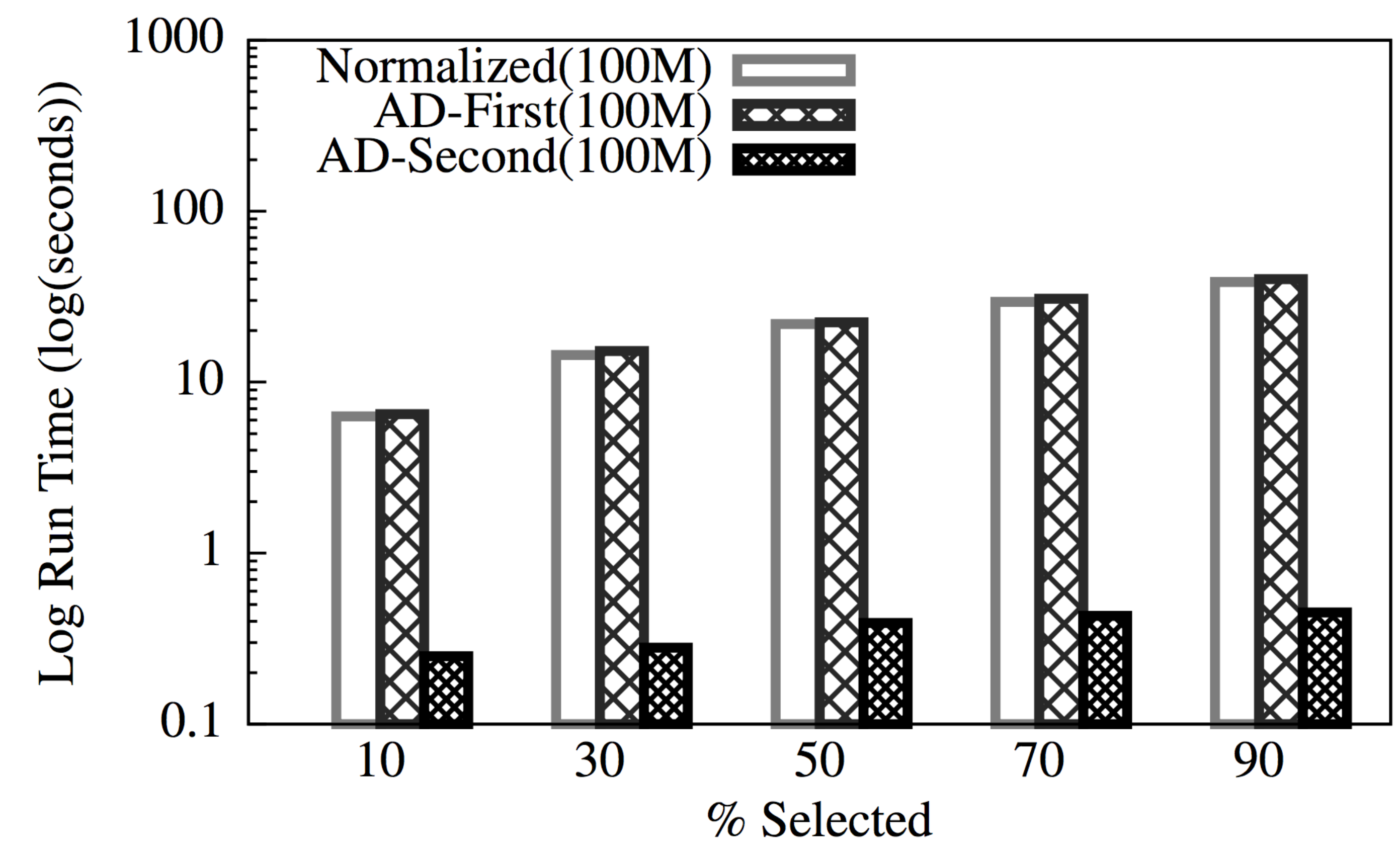
- Queries enrich existing partial universal tables vertically and/or horizontally.
- Future queries replace joins over these data regions with scans over the partial universal tables
- Amortizes overhead & cost of denormalization



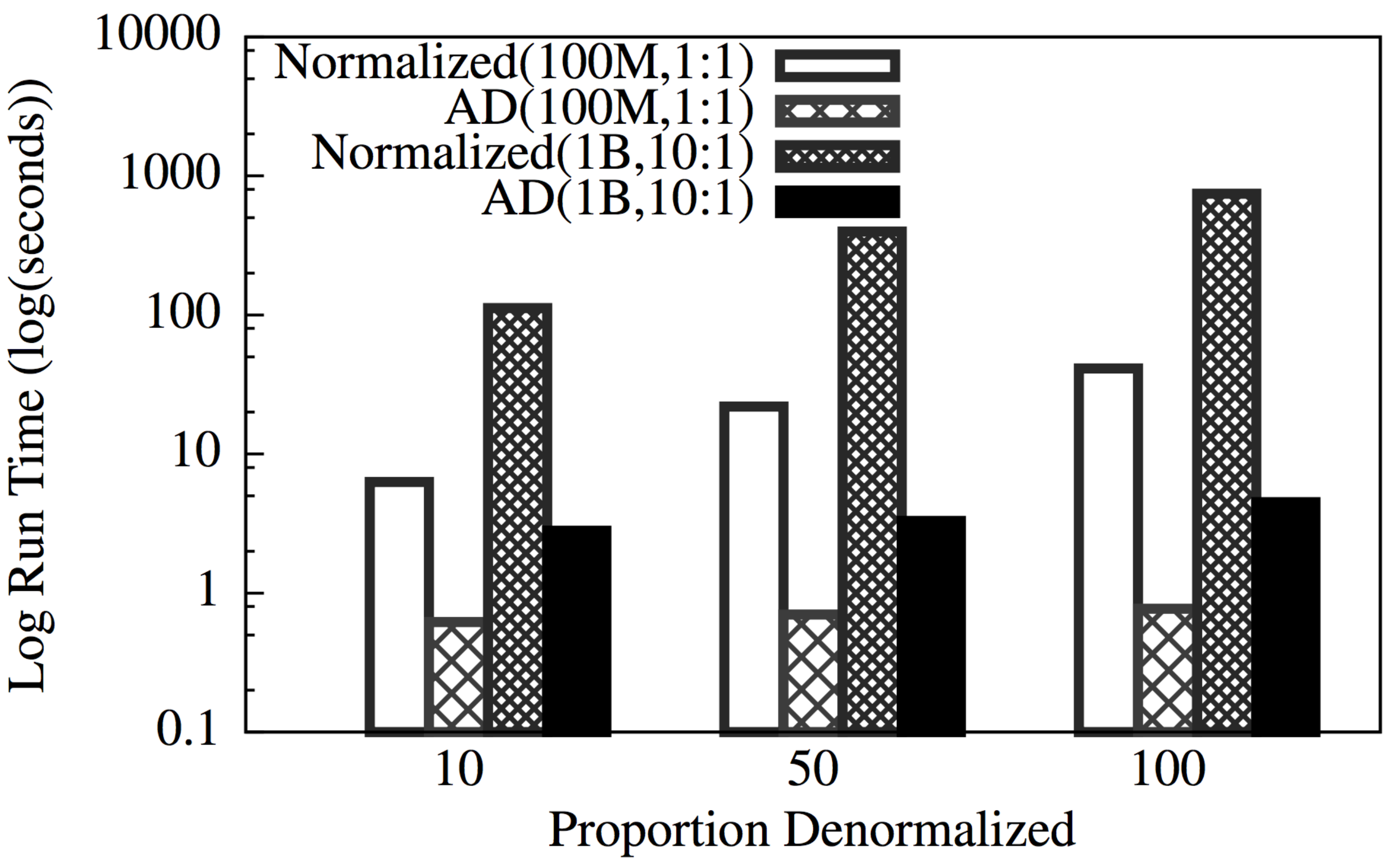
Adaptively denormalizes only regions of the data as they are queried and only data that has not yet been denormalized by previous queries.



Operates within the given memory budget by dropping regions of the partial universal table in response to memory pressures.



Adaptive denormalization (AD) improves significantly over repeated join patterns without penalizing the first join queries.



Adaptive denormalization (AD) achieves significant benefits even when the required data is only partially denormalized.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

daslab.seas.harvard.edu

DASlab