Adaptive Data Skipping in Main-Memory Systems

Wilson Qin, Stratos Idreos
Harvard University
wilsonqin@college.harvard.edu, stratos@seas.harvard.edu

Fast Data Access Patterns

Lightweight indexes such as Zonemaps logically segment data into contiguous zones, keeping metadata for each zone. Metadata informs database operations, indicating which data zones are non-pertinent to a query - this is called data skipping.

Fast Knowledge Extraction

At scale, data skipping enables faster queries without the overhead of maintaining full indexes on the data. Faster data filtering operations during scans, ultimately means a faster knowledge discovery cycle for users.

Zonemaps keep minima, maxima statistics on fixed sized zones.

However, for certain datasets and query workloads, a zonemap layout can even be less effective than a scan.

Are there lightweight methods to avoid zones with useless metadata (A), and create useful zones (B)?

Zones with wide domains, and useless metadata.

Variation in zonemap effectiveness by layout indicates the potential for dynamic structures beyond static layouts that can adapt to a changing query workload over time.

We introduce Adaptive Data Skipping as a robust data skipping framework applicable to a much wider array of data distributions and ad-hoc query workloads than standard data skipping methods.

Incremental

Opportunistically improve relevant zones by gradually restructuring them during query time - as underlying data is being scanned.

On-Demand

Adapt

Increase Skipping

1.4X speed up

Adaptive Zonemap prototype: early results for select operations.

daslab.seas.harvard.edu