Distributed Caching for Processing Raw Arrays

Weijie Zhao, Florin Rusu, Bin Dong, Kesheng Wu, Anna Ho, and Peter Nugent

Distributed Caching Architecture

Raw Array Data Model

Raw Array Chunking

When to split?
- Chunk is sufficiently large.
- Query subarray does not contain any cells.

How to split?
- According to the query subarray boundaries?

Algorithm 1 Chunk Split

Input: Chunk \( \alpha \) with bounding box \( \alpha \), that intersects query subarray \( Q \).
- Minimum number of cells threshold \( \text{MinC} \)

Output: Chunks \( \beta \) and \( \gamma \) after splitting \( \alpha \), if \( (|\text{cells in } \alpha| < \text{MinC}) \) and \( (|\text{cell in } \alpha \in Q|) \) then return

1. \( \min_{\text{vol}} = +\infty \)
2. \( \text{for each boundary } b \in Q \) that intersects with \( \alpha \), do
3. \( \gamma = \text{split cells in } \alpha \text{ into two sets by boundary } b \)
4. \( \text{if vol}(\alpha_b) + \text{vol}(\gamma_b) < \text{vol}_{\text{vol}} \text{ then} \)
5. \( \text{min}_{\text{vol}} = \text{vol}(\alpha_b) + \text{vol}(\gamma_b) \)
6. \( \beta = \text{bounding box}(\alpha_b), \gamma = \text{bounding box}(\gamma_b) \)
7. \( \text{end if} \)
8. \( \text{end for} \)

Cost-Based Caching

Cache Eviction
We must scan a file entirely even if only one accessed chunk isn’t cached.
- We aim to cache all the queried chunks in a file.

\[ \text{cost}(Q, \alpha, (C_j)) = w_\alpha \sum_{f \in (C_j)} \text{size}(f) \]

Cache Placement
We piggyback on the replication induced by query execution.

\[ \text{cost}(C, n, P, W) = \sum_{Q \in W} w_\alpha |C_j \in P \land (C_j, C_k) \in Q| \]

Experiments

PTF catalog: PTF[time=1,153064;ra=1,100000;dec=1,50000]
1 billion objects, 343 GB in CSV, 262 GB in HDF5, and 221 GB in FITS

LinkedGeoData: GEO[long=1,100000;lat=1,50000]
30 billion objects, 1.6 TB in CSV.

3 query patterns: real workload, shifting ranges and alternative queries.

Query execution time:

Similarity join execution time:

Optimization time

Weijie Zhao, Florin Rusu (UC Merced); Bin Dong, Kesheng Wu (LBNL), Anna Ho (CalTech), Peter Nugent (LBNL)