PFLOCK Report

Andres Calderon

University of California, Riverside

May 29, 2020

Re-visiting self-distance join for pair finding

- The goal is to obtaing good local partitioning from the very beginning.
- To find the centers and the points they contain first I have to find the set of pairs of points.
- Applying a partition-based approach we expect to keep those local partitions for subsequents steps.

Solving issues with partition-based Join

- Using GeoSpark Quatree. Keeping fixed number of levels (levels = 5) and size of samples (fraction = 1).
- Variying maximum number of items per node (capacity).
- Finally solving the performance issue in my approach (a really fool mistake).
- Results were validated comparing with the baseline and index-based approach (outputs were identical).

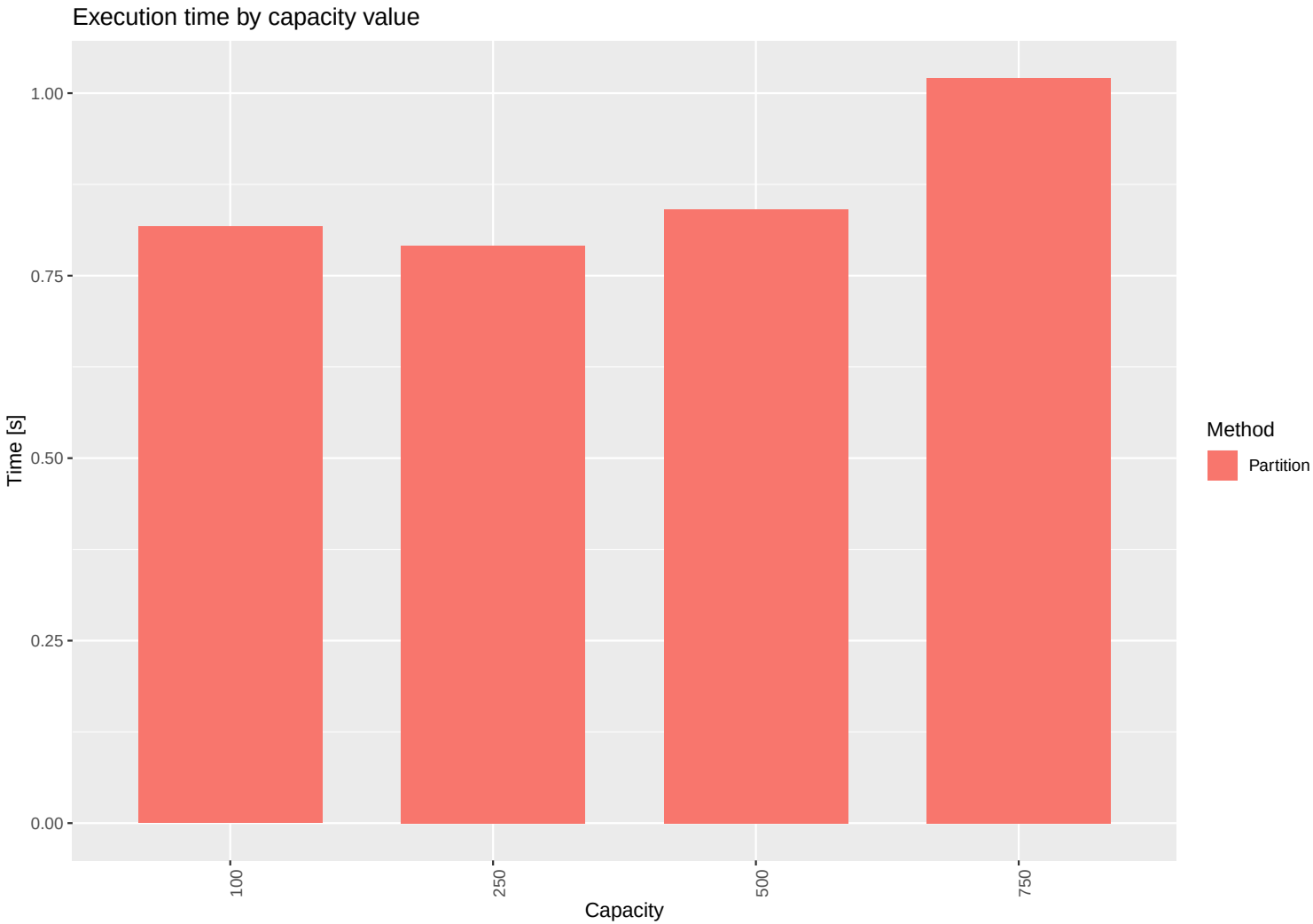
Experiments

Setup

- Assuming a global partion with 10000 points.
- Finding pairs of points which are $\varepsilon = 10m$ each other.
- Runing locally.
- Average of 10 runs.

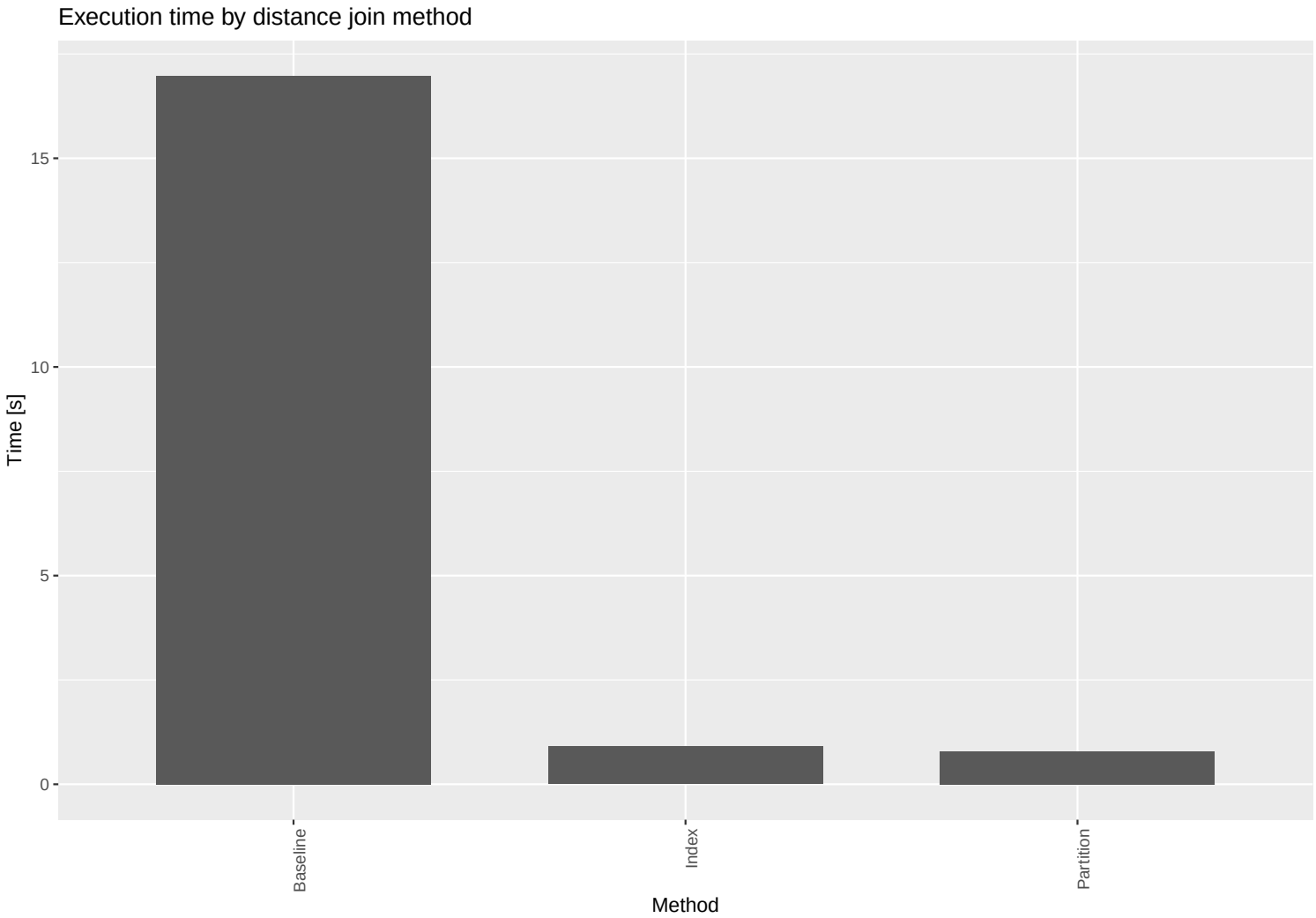
Experiments

Results



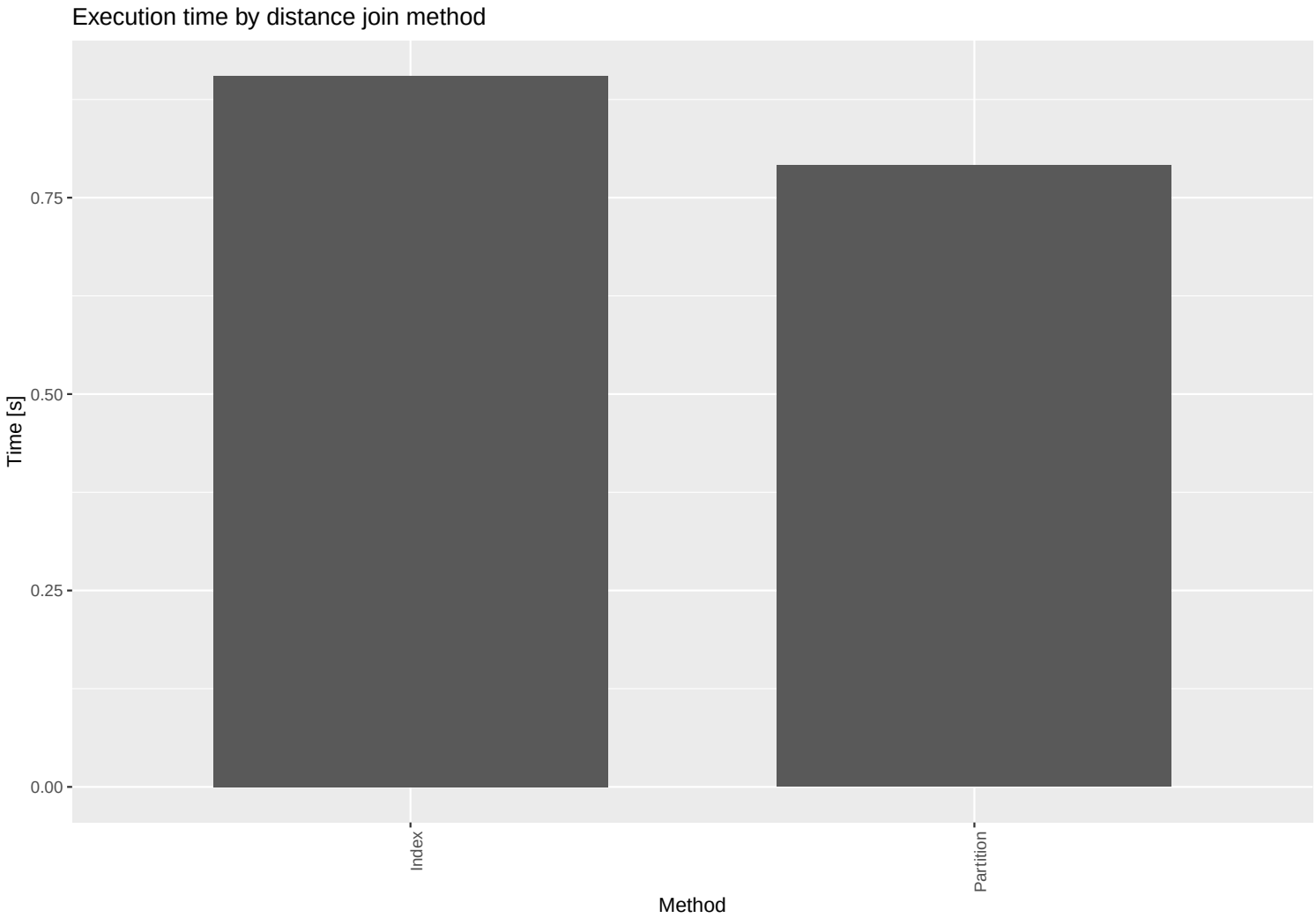
Experiments

Results



Experiments

Results



What's next

- Integrating the aproach to perform points vs centers join (taking advantage of the current local partitioning).
- Validate and test performance.