Similarity Join Database Operator for Multi-Dimensional Data

Faculty: Yasin N. Silva, Ph.D. Student: Spencer Pearson
Division of Mathematical and Natural Sciences, Arizona State University

Why a Database Similarity Join operator?

- Similarity join is useful in many fields to find data which closely match each other
  Applications in:
  - Marketing
  - Multimedia and video applications
  - Sensor Networks
- It is possible to implement a Similarity Join query using standard database operators, but this is inefficient and does not allow a full integration of the operator with the database optimizer.
- The support of Similarity Join for Multi-dimensional data is key because this data type is extensively used to analyze complex objects: images, videos, geographical data, etc.
- We propose a Similarity Join database operator for multidimensional data. This operator has the following benefits:
  - It is fully integrated with the query and optimization engine
  - Better interaction with other database operators
  - Better query execution time and scalability
  - It is easier to construct queries
  - Supports a very useful data type

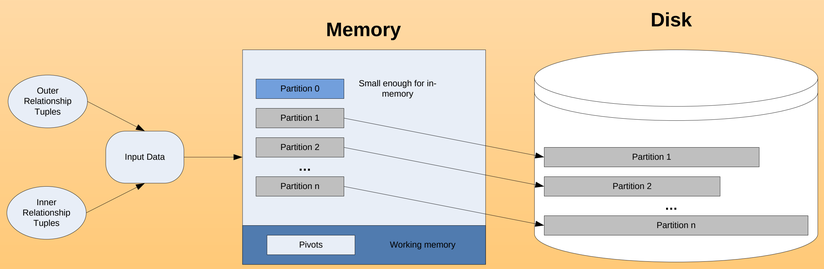
Similarity Join Example

Goal: Find pairs of restaurants and movie theaters that are close to each other.

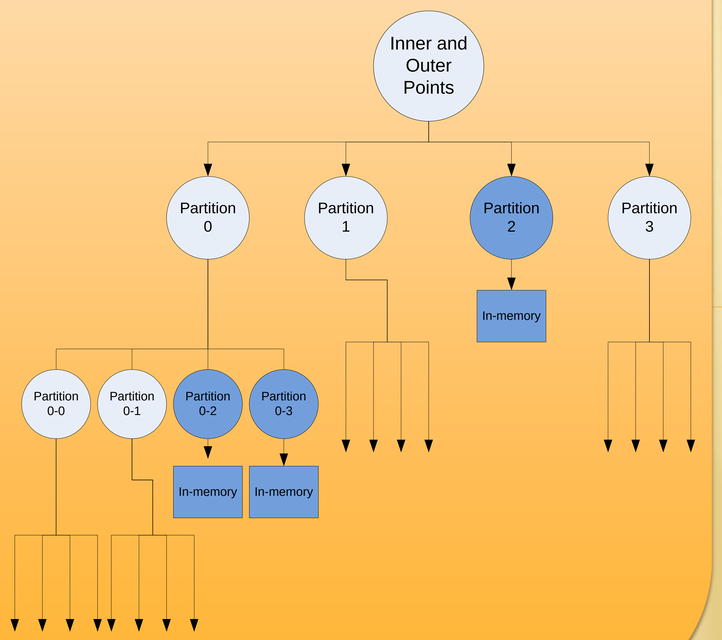
"close" = within ε of one another

Solution Design

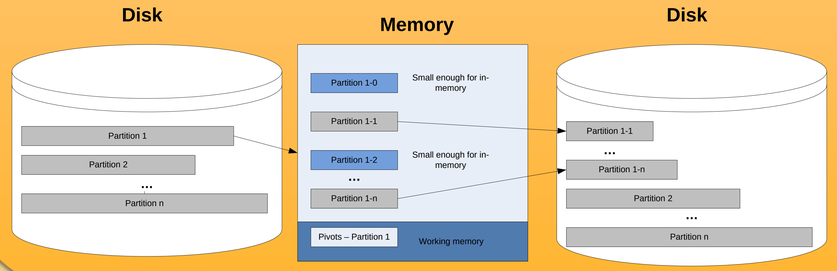
Pass 0



Partition of Data



Pass 1, …, m



Future Work

- Implementation
  - In the query engine of an open source database system
- Testing
  - Determine efficiency/accuracy
  - Make improvements on the algorithm