SIMILARITY JOIN FOR BIG GEOGRAPHIC DATA

YASIN SILVA, JASON REED, LISA TSOSIE, TIMOTHY MATTI, KYLE GERVAIS
ARIZONA STATE UNIVERSITY

Motivation

The Problem

- Cloud-based systems are crucial to processing and analyzing large amounts of data
- Similarity Joins (SJ) are a key data processing and analysis tool
- Very little work on Similarity Joins has been done for big geographic data

Our Contribution

- We propose MRSimJoin –a MapReduce-based algorithm to efficiently solve the SJ problem
- The algorithm is general enough to be used with data that lies in any metric space
- Our focus is on the study of this operation with big geographic data
- Thorough evaluation of performance and scalability with real world and synthetic geographic data sets

Geographic Data

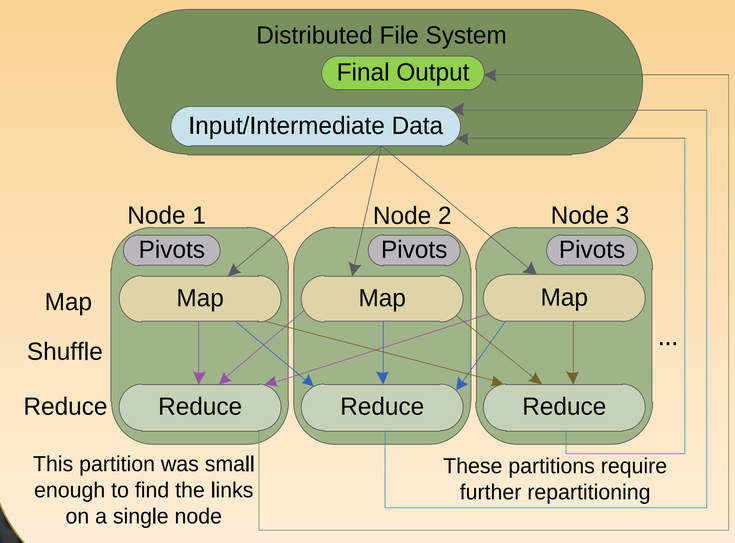
- Geographic data uses latitude (φ) and longitude (λ) coordinates to represent a location on a sphere
- There are several methods of calculating distance between two points
  - Euclidean Distance
  - Great Circle Distance
  - Tunnel Distance
- This presentation considers the case of Euclidean Distance on a plane where a spherical earth is projected using equirectangular projection
- Euclidean Distance is fast to compute and accurate at small distances
- Given two points
  - $r_1 = (\varphi_1, \lambda_1)$
  - $r_2 = (\varphi_2, \lambda_2)$

- The Euclidean Distance between them is as follows:

$$geoDist(r_1, r_2) = R\sqrt{(\Delta_\varphi)^2 + (\cos(\varphi_m)\Delta_\lambda)^2}$$

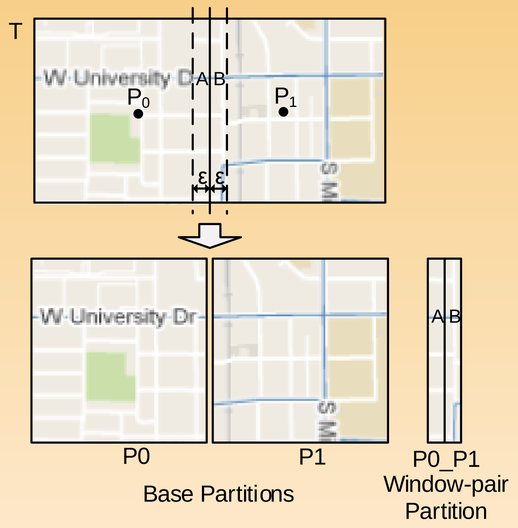
Partitioning

- MRSimJoin iteratively partitions the data into smaller partitions until each partition is small enough to be efficiently processed by a single-node SJ routine
- This process is done in multiple rounds, each corresponding to a MapReduce job
- Each round outputs result links and intermediate data requiring further partitioning

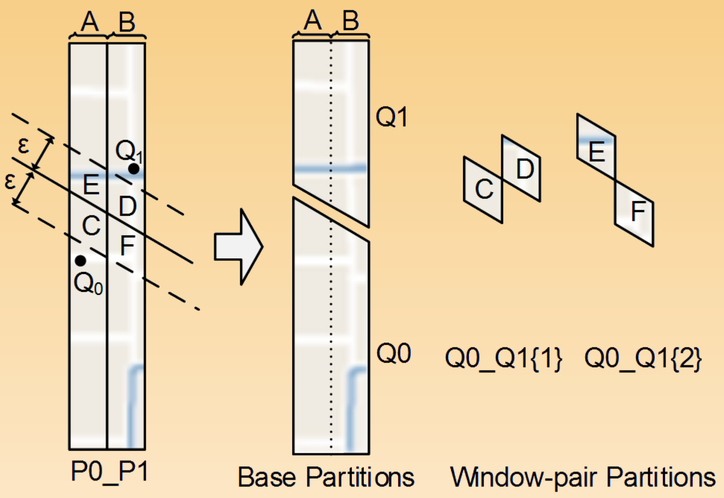


Partitioning in an MRSimJoin Round

- Data partitioning is performed using a set of K pivots (conceptually similar to QuickJoin), which are a random subset of the records to be partitioned
- The process generates two types of partitions: base partitions and window-pair partitions
  1) A base partition contains all the records that are closer to a given pivot than to any other pivot
  2) A window-pair partition contains the records in the boundary between two base partitions



Partitioning a base partition

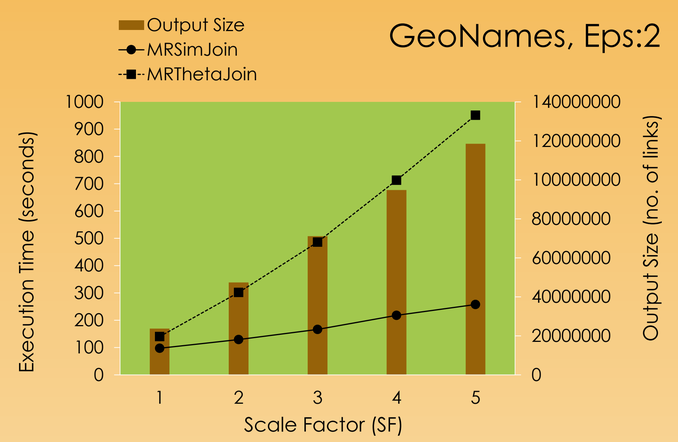


Partitioning a window-pair partition

Performance Evaluation

Tests run over 2 million (SF1) records

Increasing Scale Factor



Increasing Epsilon