Simba: Spatial In-Memory Big data Analysis

Dong Xie1, Feifei Li1, Bin Yao2, Gefei Li1, Zhongpu Chen2, Liang Zhou2, Minyi Guo2

1University of Utah, 2Shanghai Jiao Tong University

{dongx, lifeifei}@cs.utah.edu  {yaobin@cs., oizz01@, chenzhongpu@, nichozl@, guo-my@cs.}sjtu.edu.cn

Big Spatial Data Analysis at Ease

SELECT poi.id, count(*) as c
FROM poi
WHERE poi.distanceJoin(data, Point(poi("lat"), poi("long")),
data("lat"), data("long"), 3.0)
GROUP BY poi.id
ORDER BY poi.id

kNN Join -- RKJSpark

- R-Tree kNN join (RKJSpark)
- For each partition Rᵢ, find Sᵢ ⊂ S, s.t. ∀r ∈ Rᵢ, knn(r, S) = knn(r, Sᵢ)
- Define crᵢ as the centroid of partition Rᵢ
- Take a uniform random sample S’ ⊂ S, and let knn(crᵢ, S’) = {s₁, ..., sₖ}
- For each partition Rᵢ:
  - uᵢ = max{r|crᵢ}
  - γᵢ = 2uᵢ + |crᵢ, sᵢ|
  - Sᵢ = {s ∈ S | crᵢ, s ≤ γᵢ}

Comparison with Existing Systems

Experimental Results

It is now open sourced at:
https://github.com/InitialDLab/Simba