PFLOCK Report

Andres Calderon

University of California, Riverside

March 27, 2020

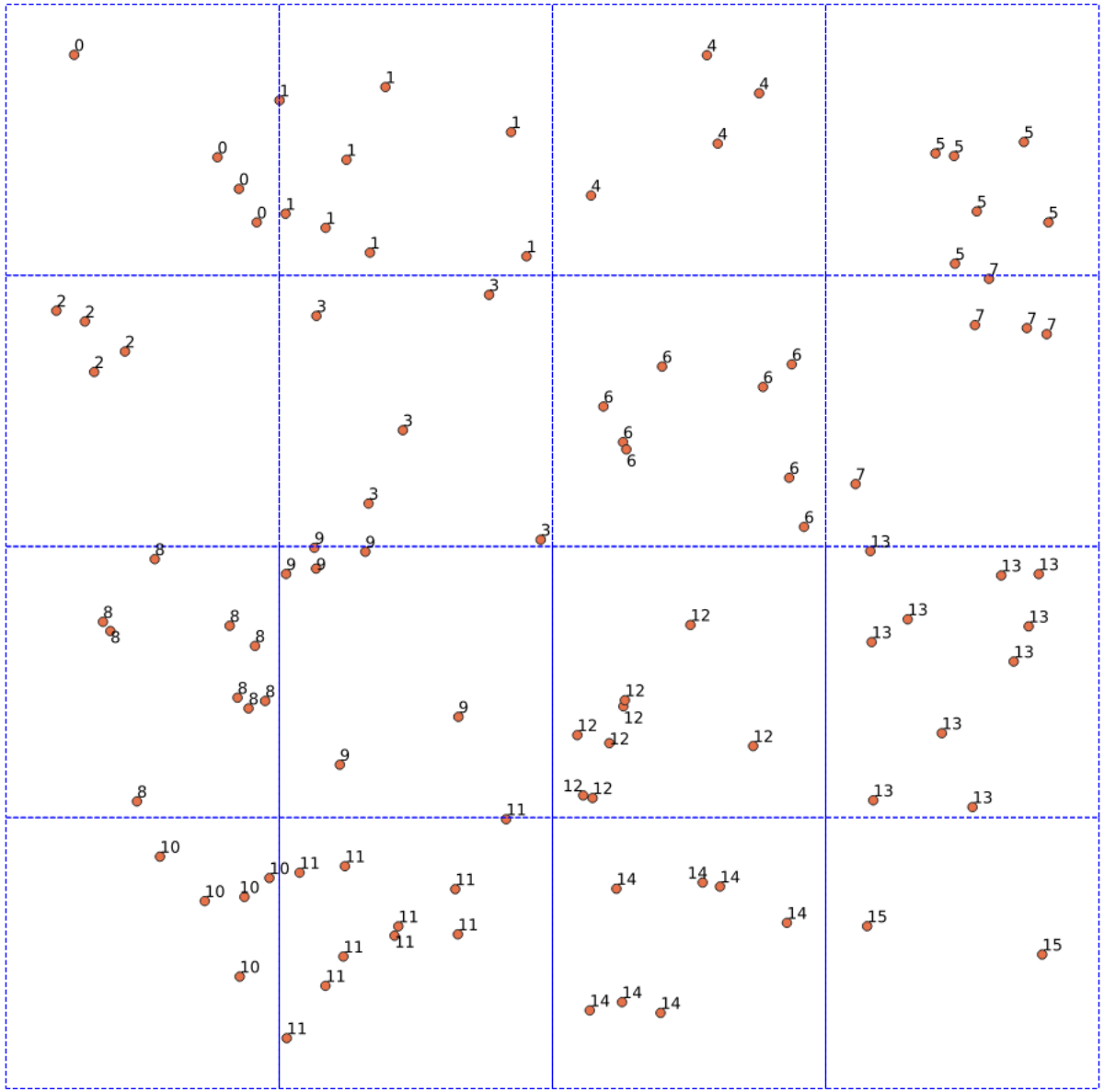
What was the problem?

- Apache Spark do not preserve partitioning by default. It has to be done explicitly in some operation such as MapPartitions, which are usually use in Joins.
- GeoSpark uses a *mapToPair* function which does not allow to set the parameter *preservesPartitioning*[1].
- In the new implementation in Scala it is easier to force this parameter to True during a *MapPartitions* operation.

[1] http://apache-spark-user-list.1001560.n3.nabble.com/preservesPartitioning-td10019.html

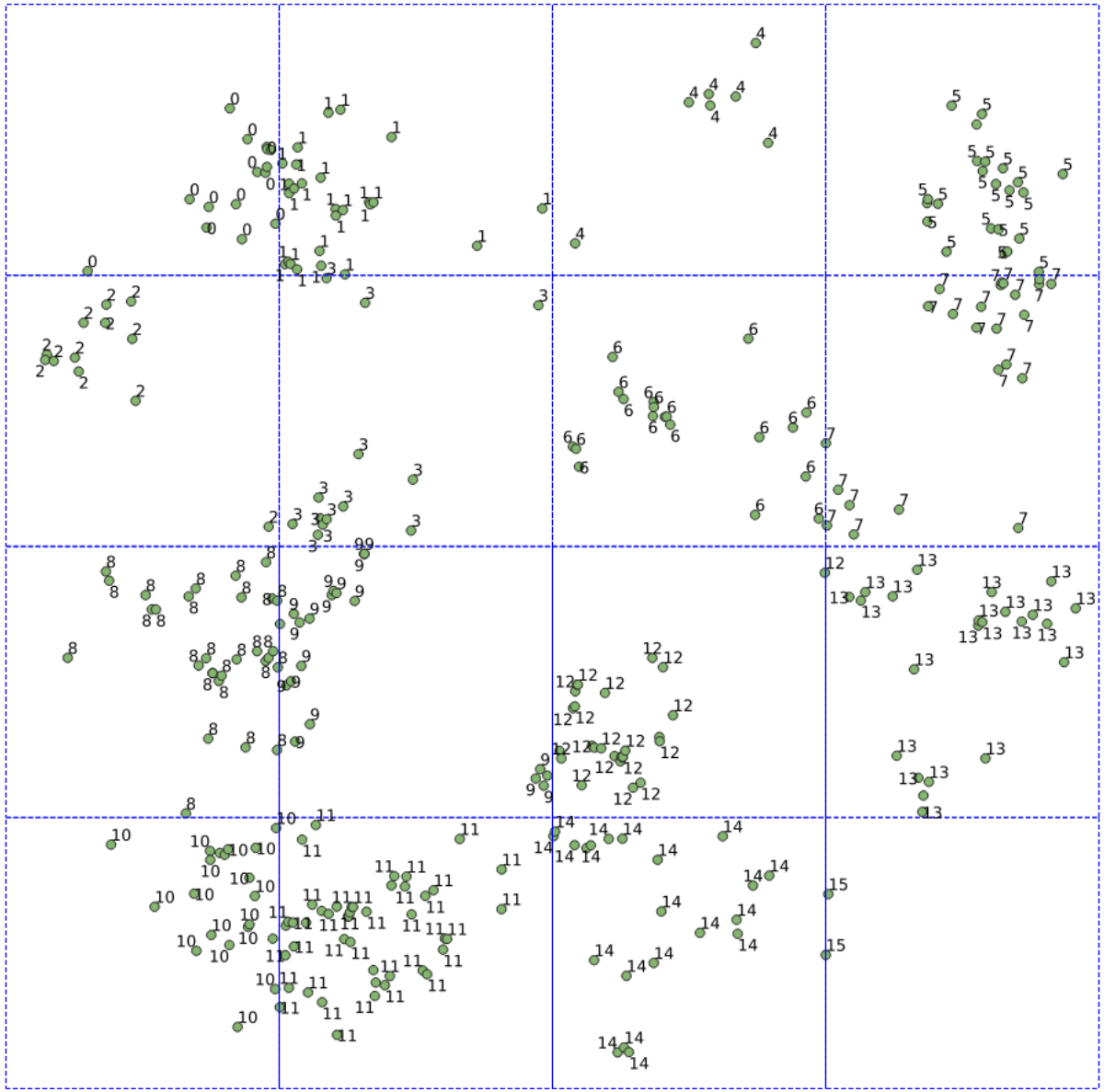
Solving distance join issue in GeoSpark...

Pointset A



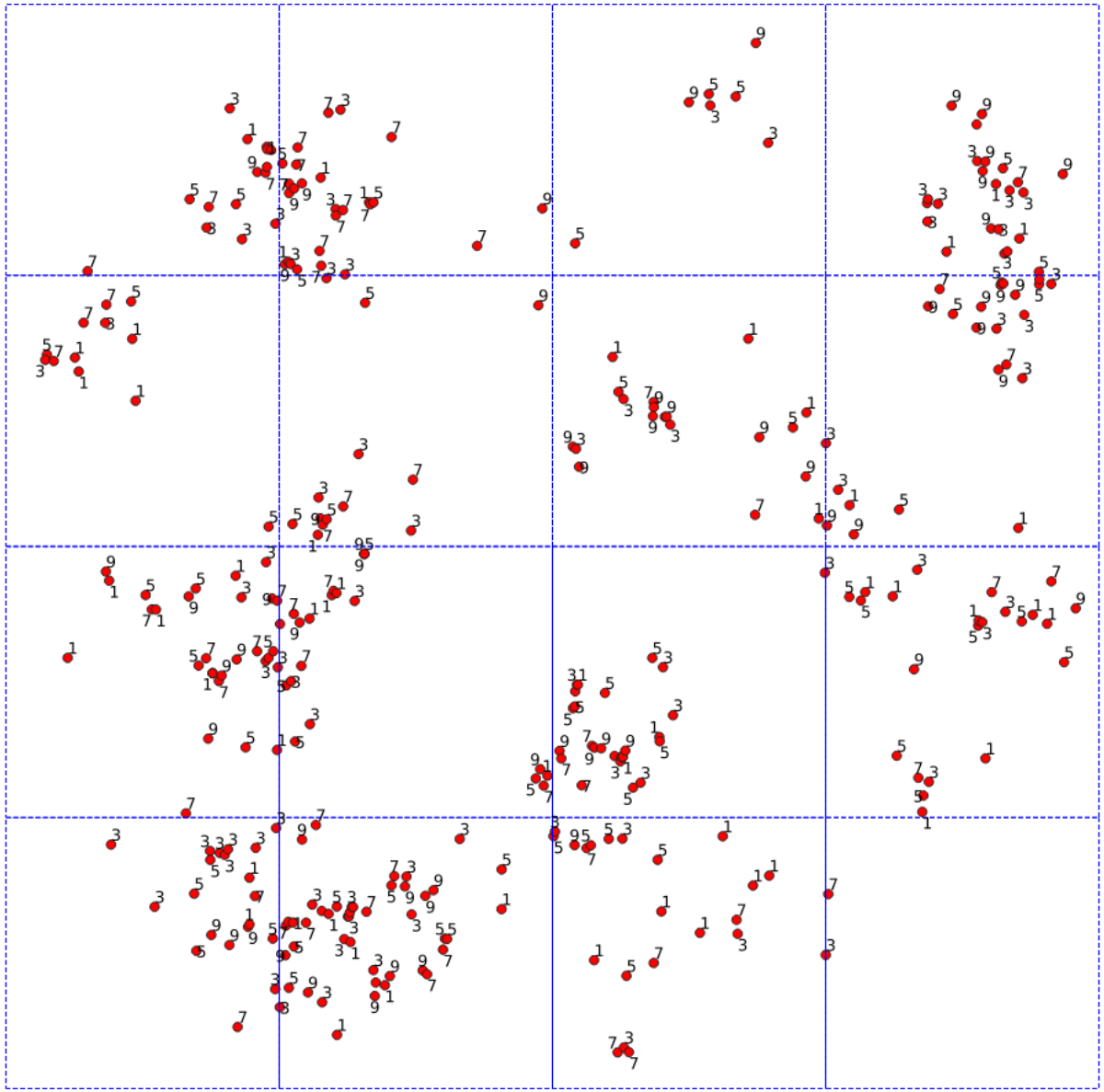
Solving distance join issue in GeoSpark...

Pointset B



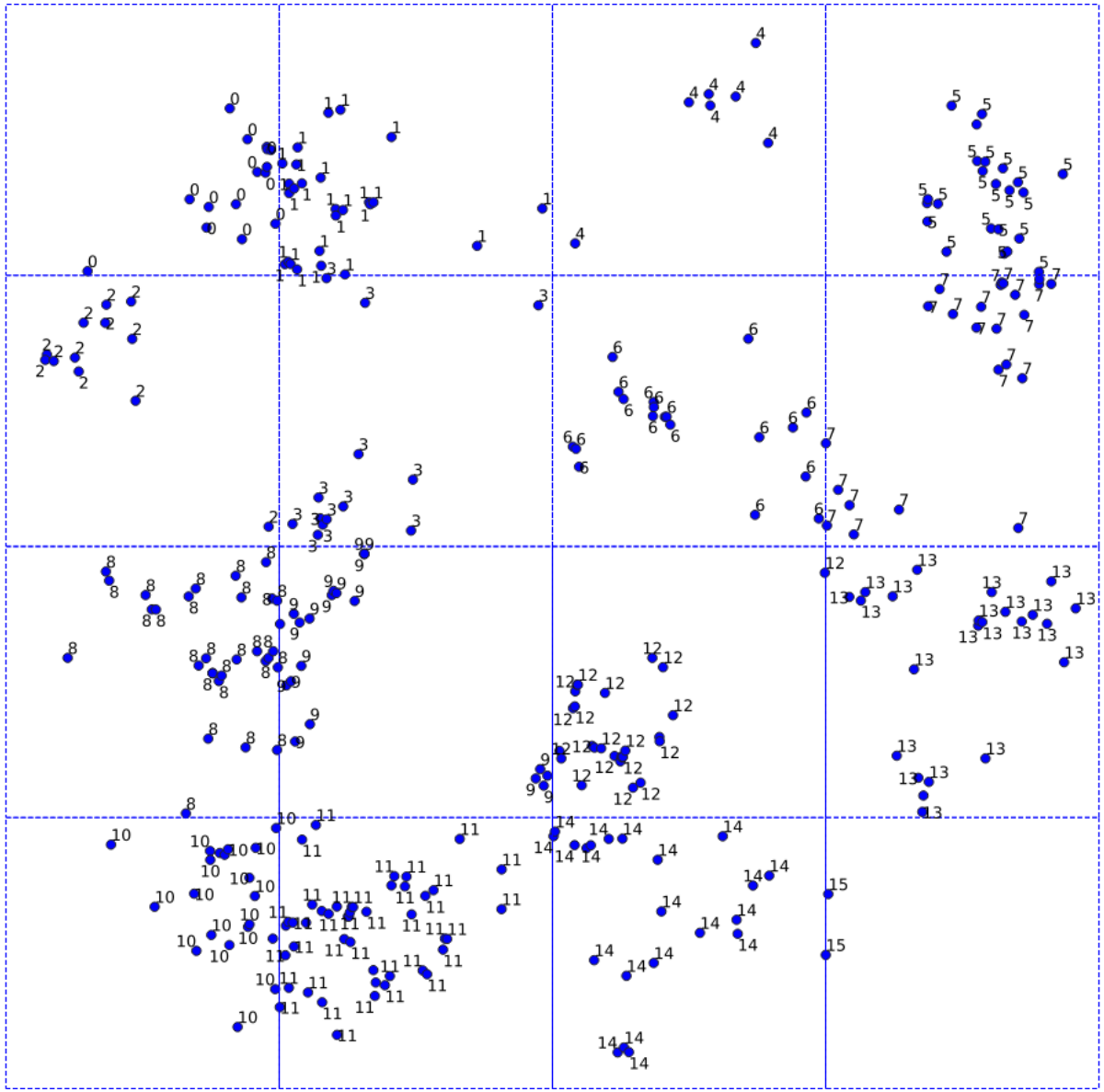
Solving distance join issue in GeoSpark...

Previous



Solving distance join issue in GeoSpark...

Current



What's next?

- Currently working on the new partition-based indexing implementation.
- Implement the new changes in the previous code.