PFLOCK Report

Andres Calderon

University of California, Riverside

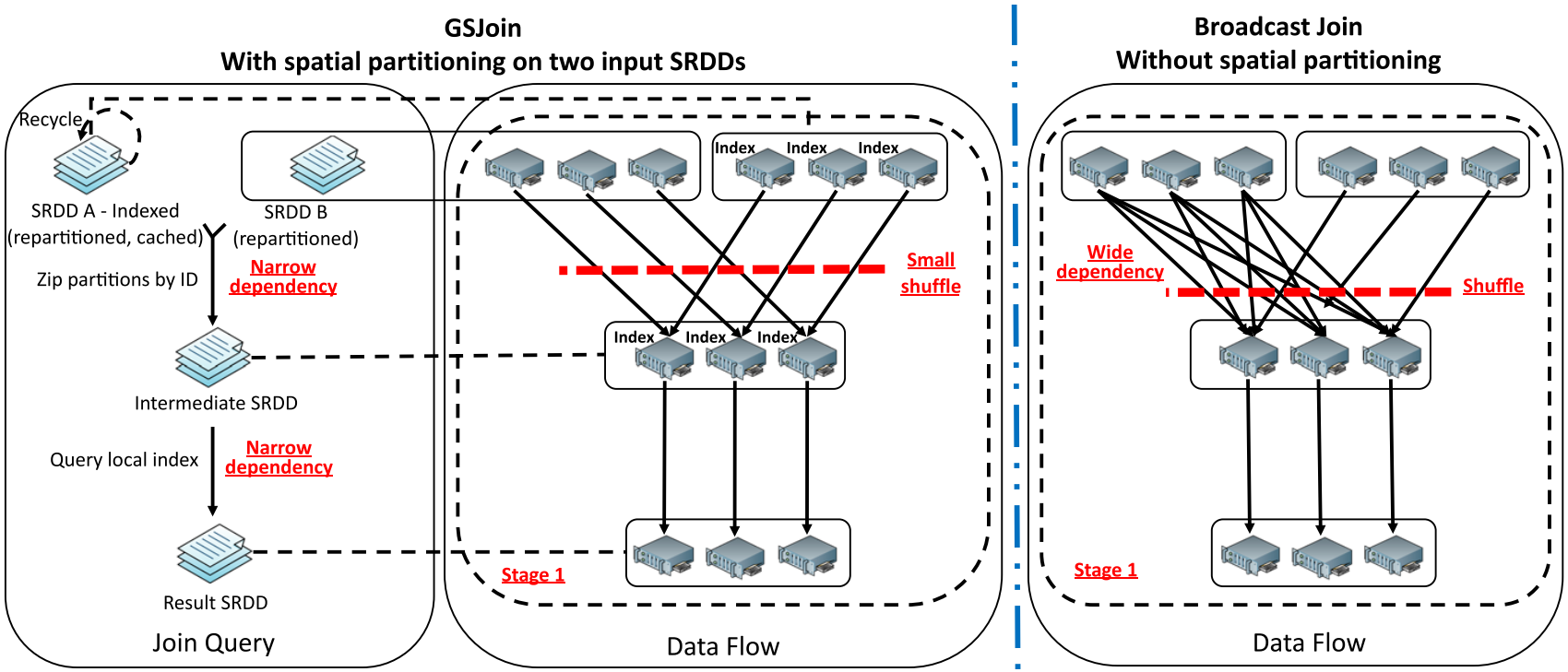
March 16, 2020

About indexing...

- GeoSpark makes partitioning and indexing in a **SpatailRDD** in two diferent stages:
  1. Once a partitioner (Quadtree, KDBtree, ...) is created, a **SpatialRDD** call `partition` method to move data to the corresponding partition (it use the `partitionBy` method of Apache Spark).
  2. Given a **SpatialRDD** already partitioned you can call `buildIndex` method which make a call to the `IndexBuilder` class (it will create a Quadtree or Rtree in the data of each partition).
- Feeding the index at the moment the data is moved would demand modification of Apache Spark code...

About Join...

- GeoSpark provides two types of joins:



About Join...

- GSJoin is actually a combination of Index-based and Nested-loop joins:
  1. If one of the SpatialRDDs is indexed, it will query that index to get a set of candidates and then refine the search.
  2. If no index is present, it will run a nested loop join.
- By default, GSJoin uses left OR right index but not both...

About Join...

- GSJoin performs some verifications which are not needed for distance joins involving Point datasets (CRS and partition matches, statistic collection).
- I have been able to port the code to Scala and remove unnecesary verifications.
- It saves $\approx$ 1.5s but still have to perform more robust tests.
- Currently checking the merge operation to see if they are using just the data in the border of the partition...