CG_Hadoop: computational geometry in MapReduce.

Ahmed Eldawy, Yuan Li, Mohamed F. Mokbel and Ravi Janardan
University of Minnesota.

Andres Calderon

October 16, 2016

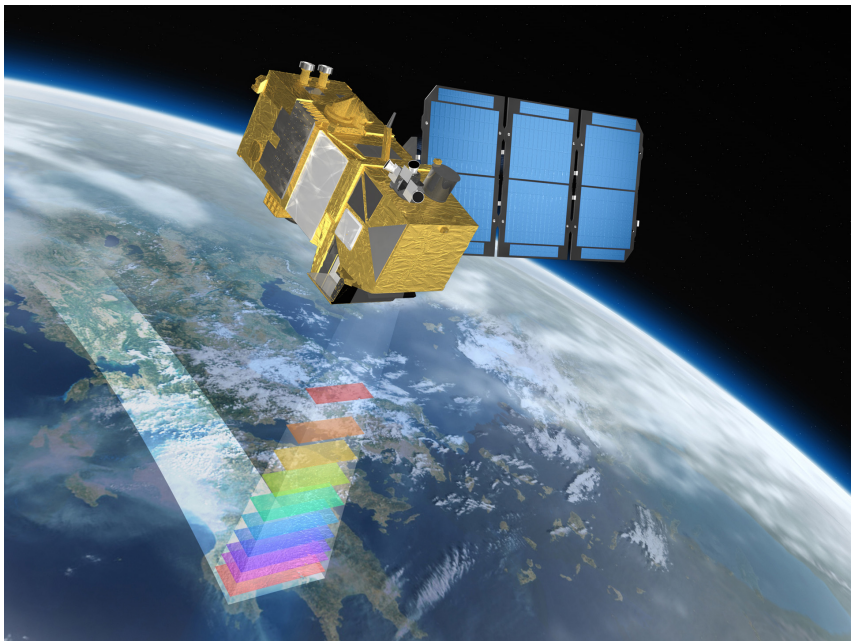
Agenda

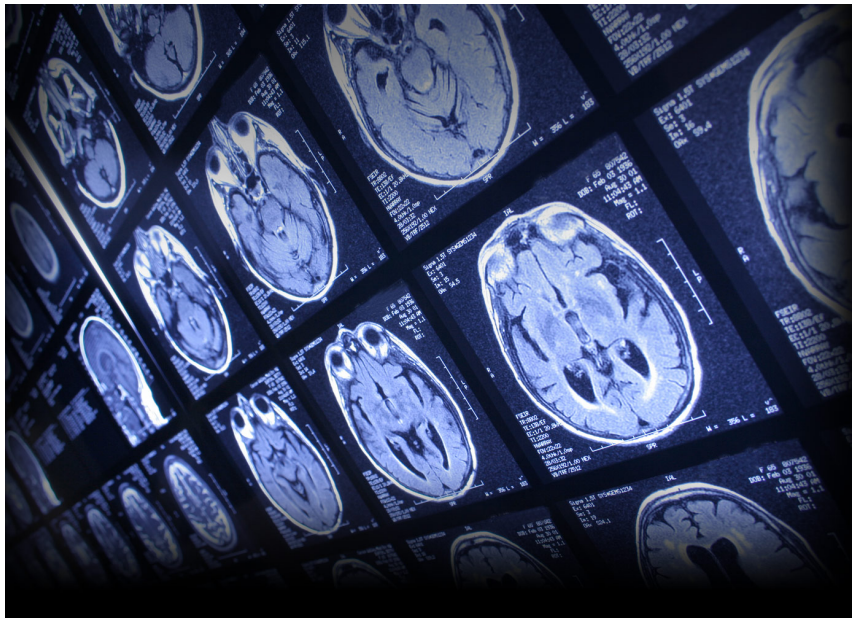
1. Background
2. Computational Geometry Operations
   - Union
   - Skyline
   - Convex Hull
   - Farthest Pair
   - Closest Pair
3. Experiments
4. Conclusions

Agenda

Introduction

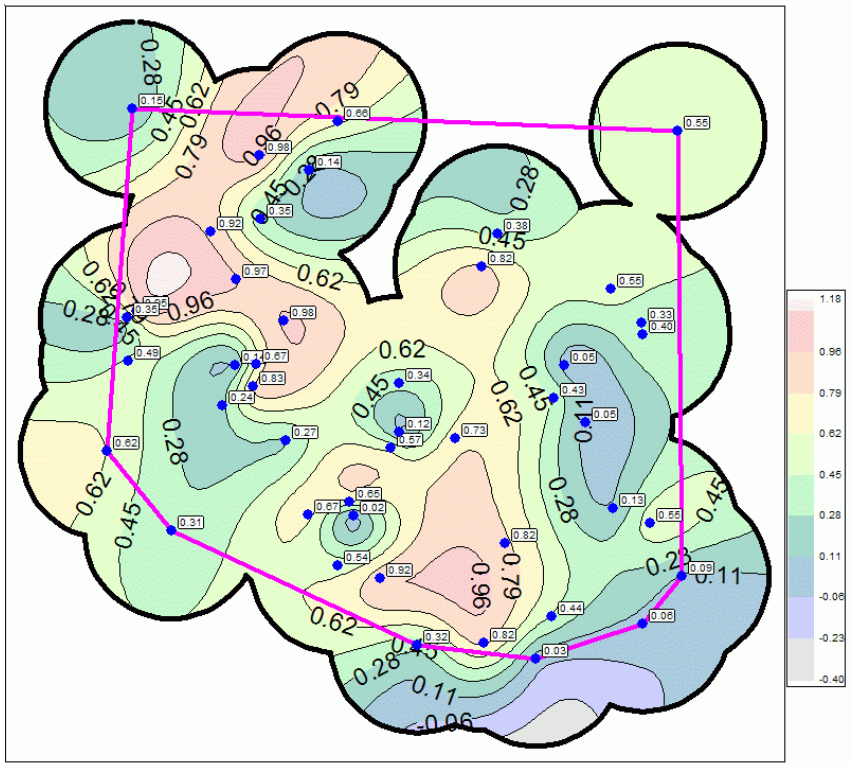
- Hadoop is widely used on many applications...
  - Machine learning, Graph processing, Sorting...
- But we are watching a big [spatial] data explosion...
  - satellites, smart phones, space telescopes, medical devices...

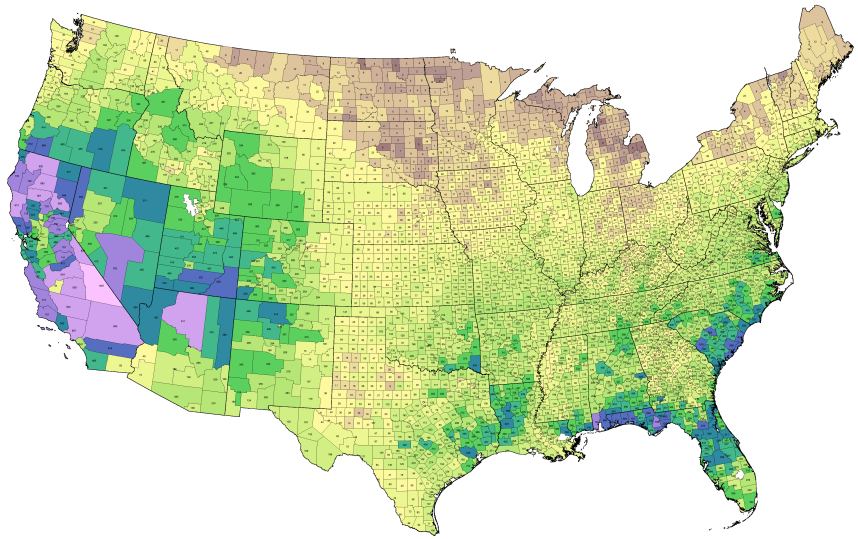




Introduction

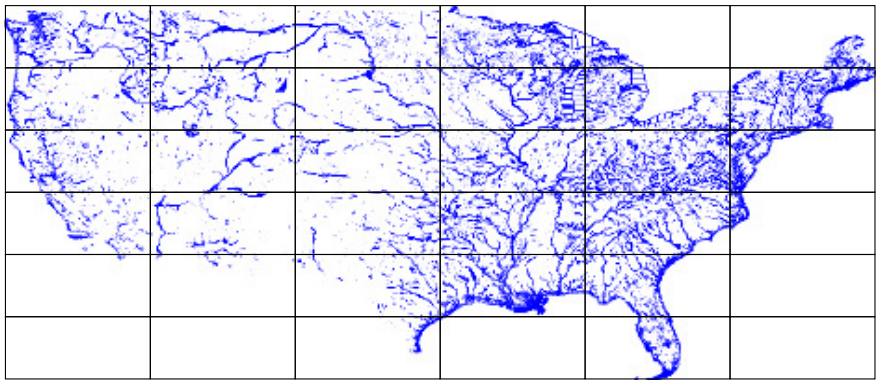
- Computational geometry operations are quite costly.
  - Convex hull - 4 billion points - up to three hours.
  - Polygon union - 5 million polygons - around one hour.

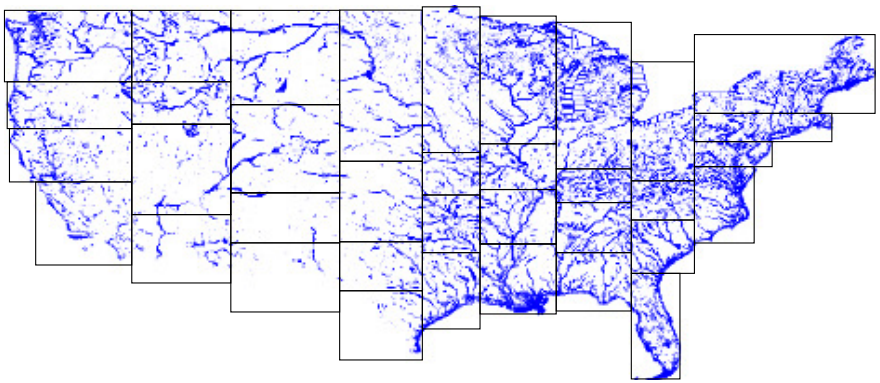




SpatialHadoop

- Extend Hadoop for spatial analysis.
  - Efficient operation for spatial querying.
  - Support for spatial types (point, polygons, rectangle).
  - Spatial partitioning and indexing techniques (Grid, R-tree, R+-tree).
  - Open source (`http://spatialhadoop.cs.umn.edu/`).



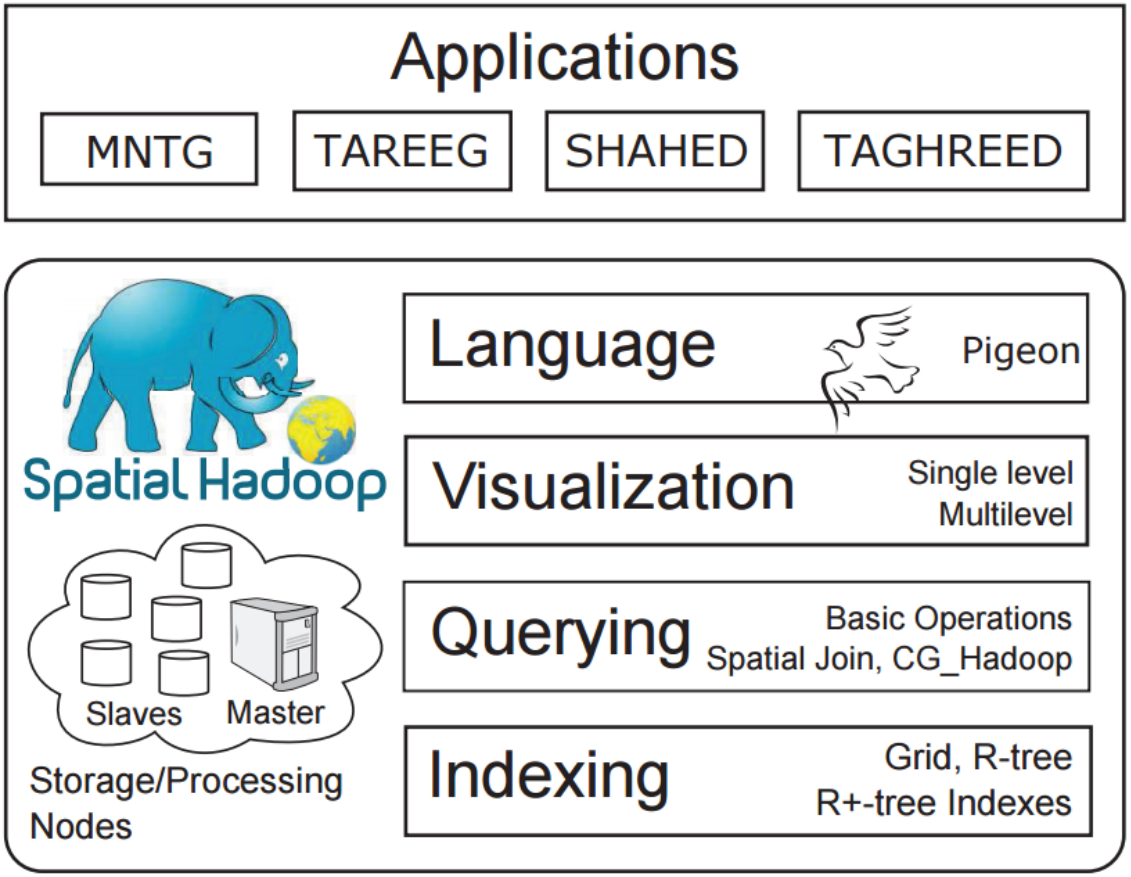


(Eldawy and Mokbel, 2015)

CG_Hadoop

- Apply a divide and conquer approach.
- Exploit spatial indexing and early pruning (SpatialHadoop).
- Main idea:
  1. Partition the input.
  2. Filter partitions that do not contribute to the answer when it is possible.
  3. Apply the algorithm locally for each partition.
  4. Merge the local answers to compute a global result.

CG_Hadoop

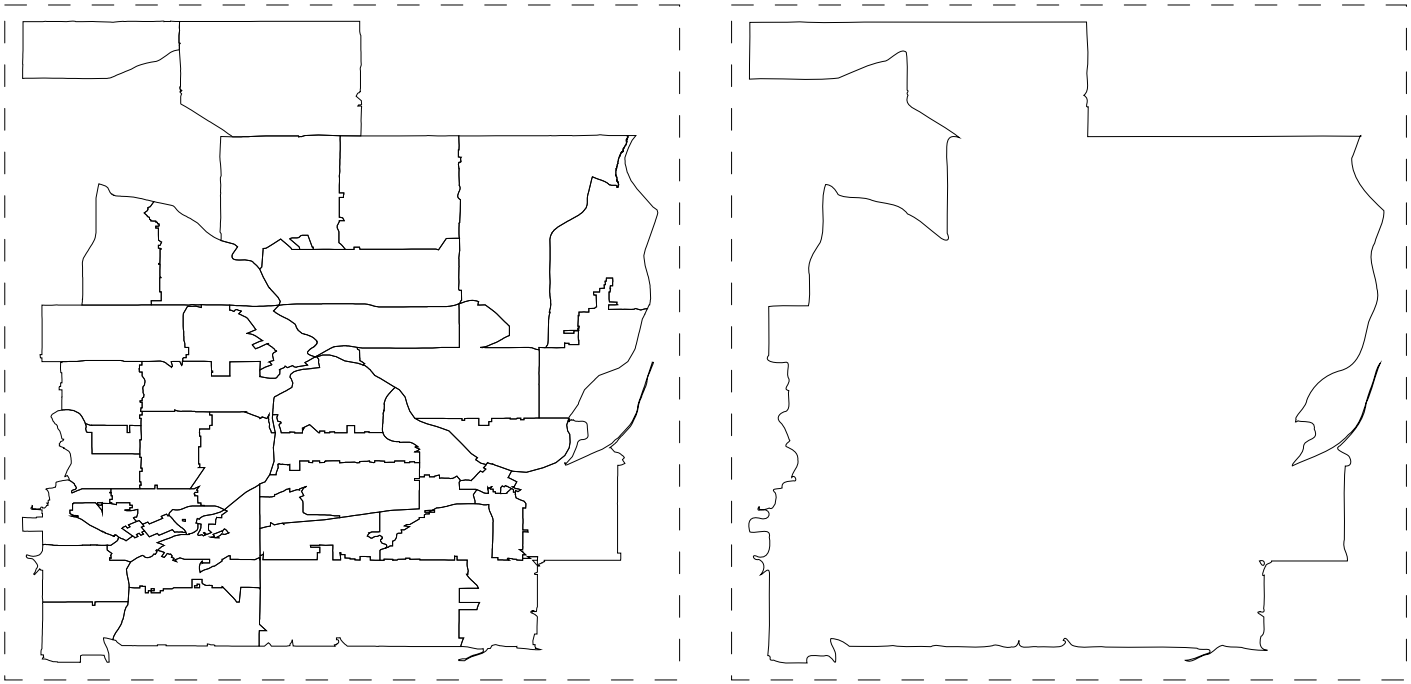

Applications
MNTG
TAREEG
SHAHED
TAGHREED
Spatial Hadoop
Slaves
Master
Storage/Processing Nodes
Language
Pigeon
Visualization
Single level
Multilevel
Querying
Basic Operations
Spatial Join, CG_Hadoop
Indexing
Grid, R-tree
R+-tree Indexes

Agenda

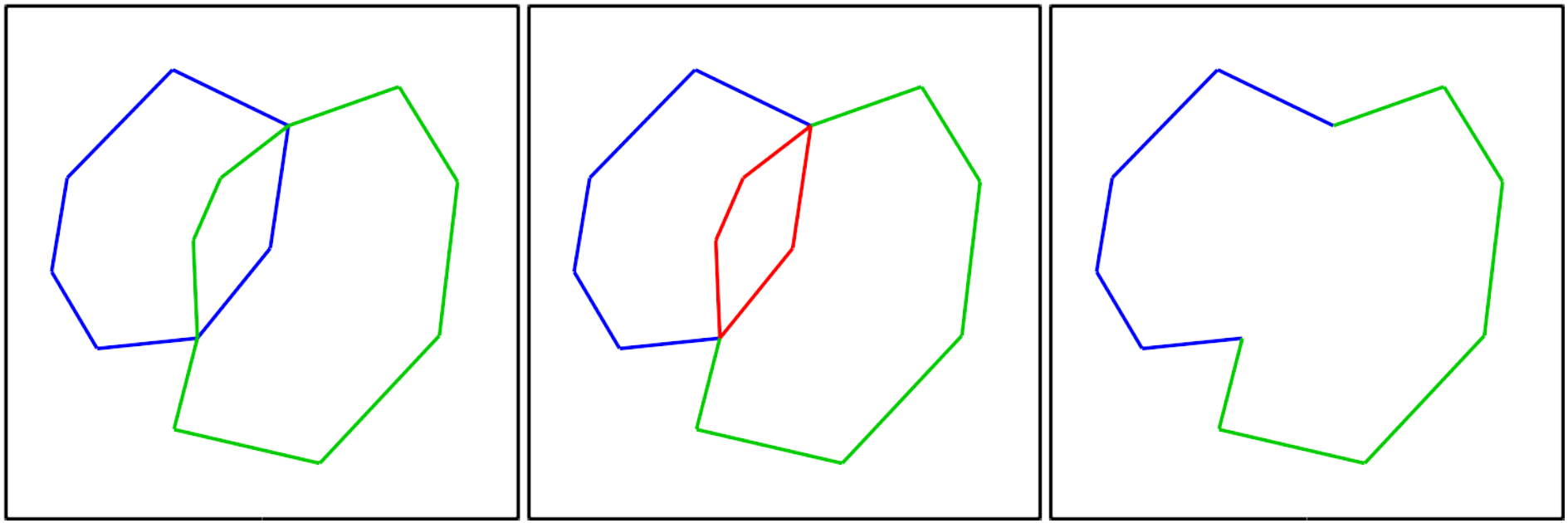
1. Background
2. Computational Geometry Operations
   - Union
   - Skyline
   - Convex Hull
   - Farthest Pair
   - Closest Pair
3. Experiments
4. Conclusions

Agenda

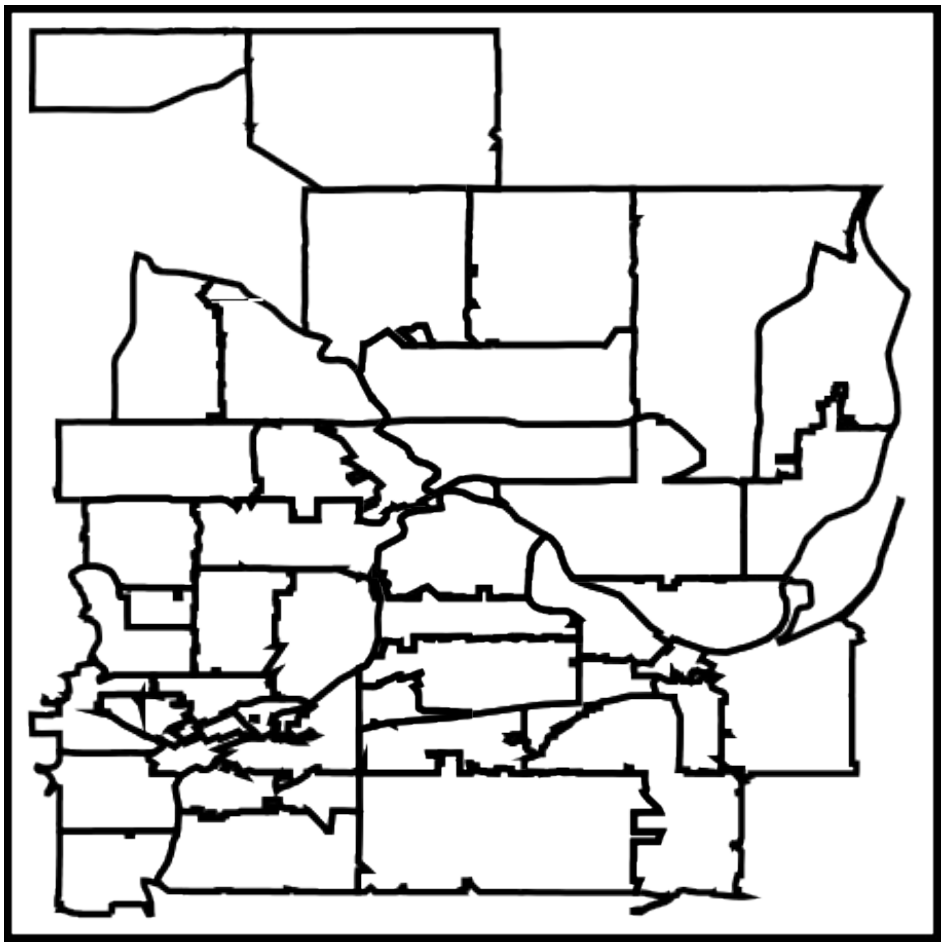
Union



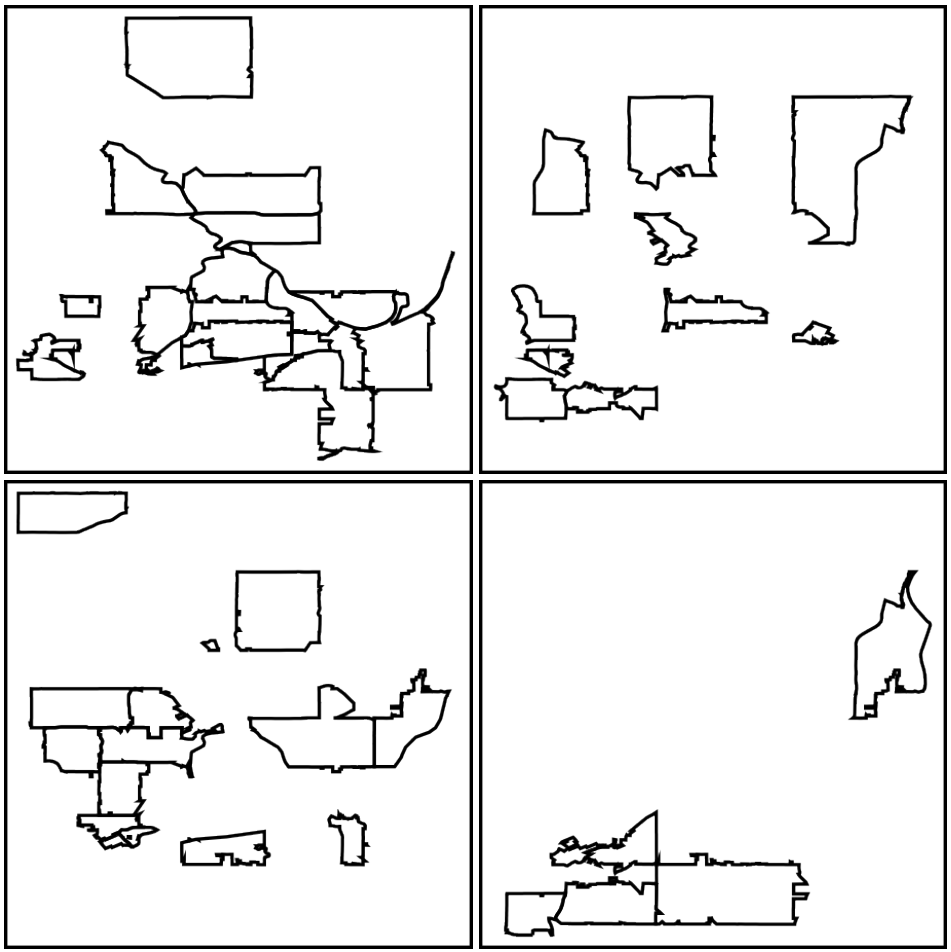
Traditional algorithm



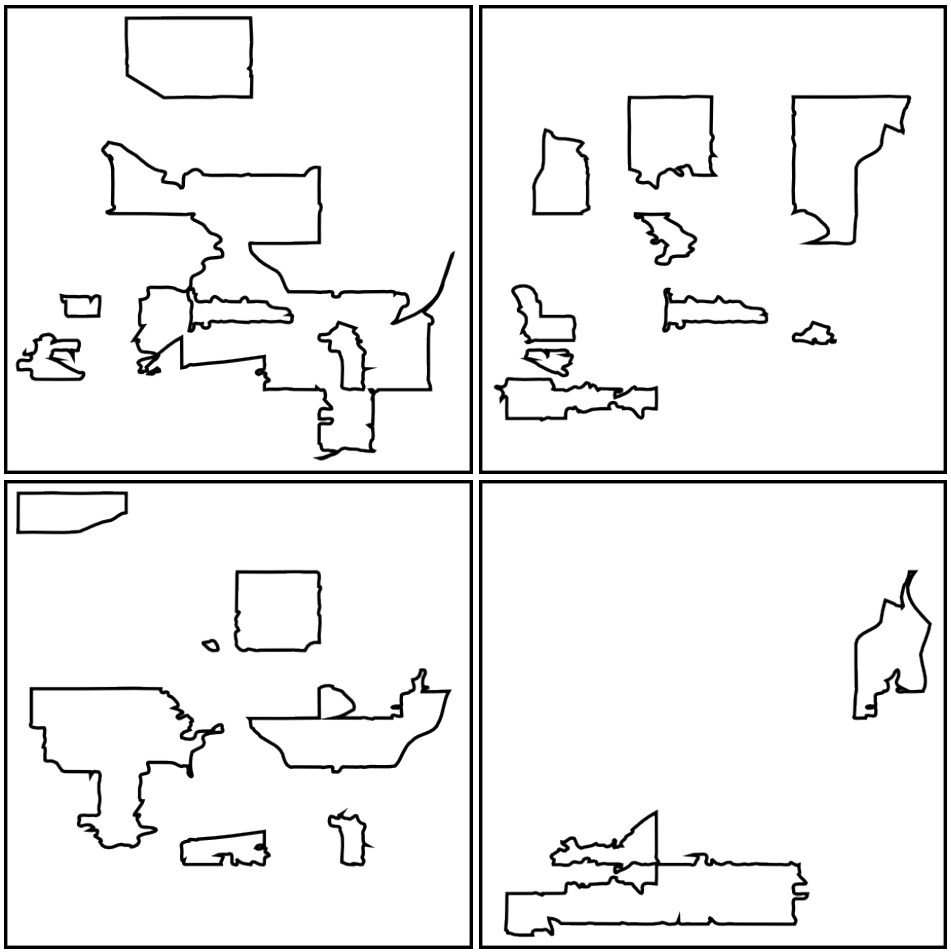
Hadoop - Partition



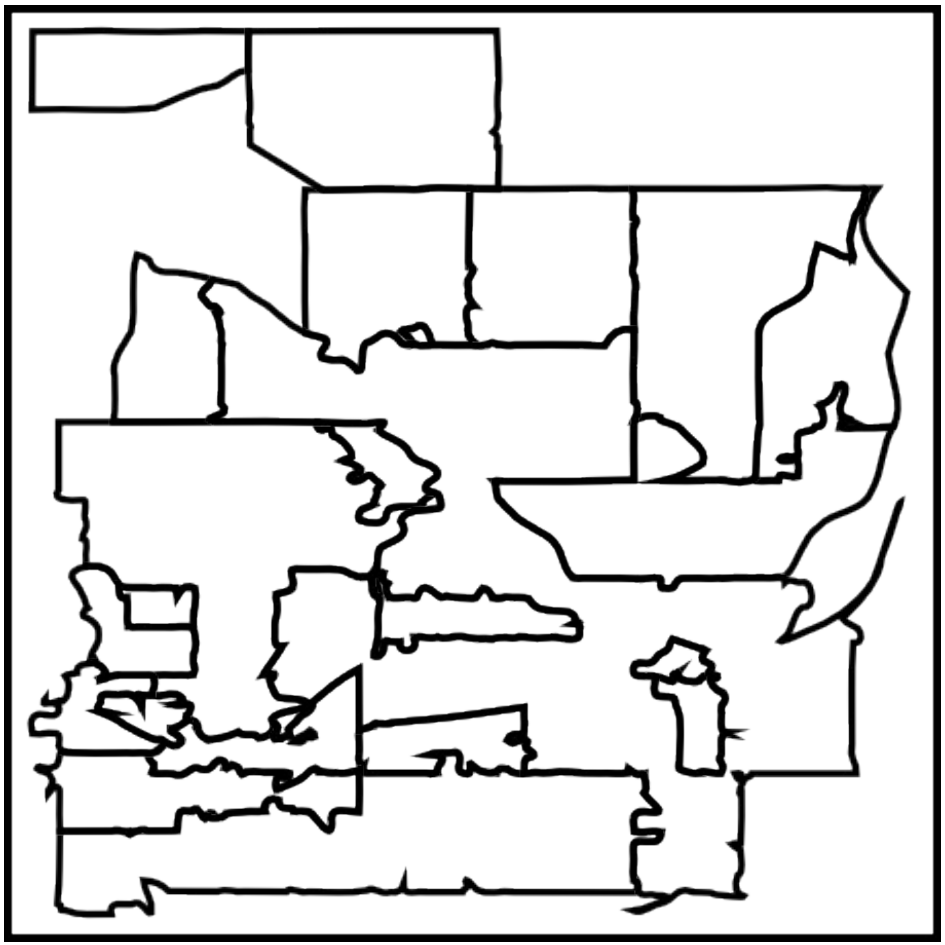
Hadoop - Partition



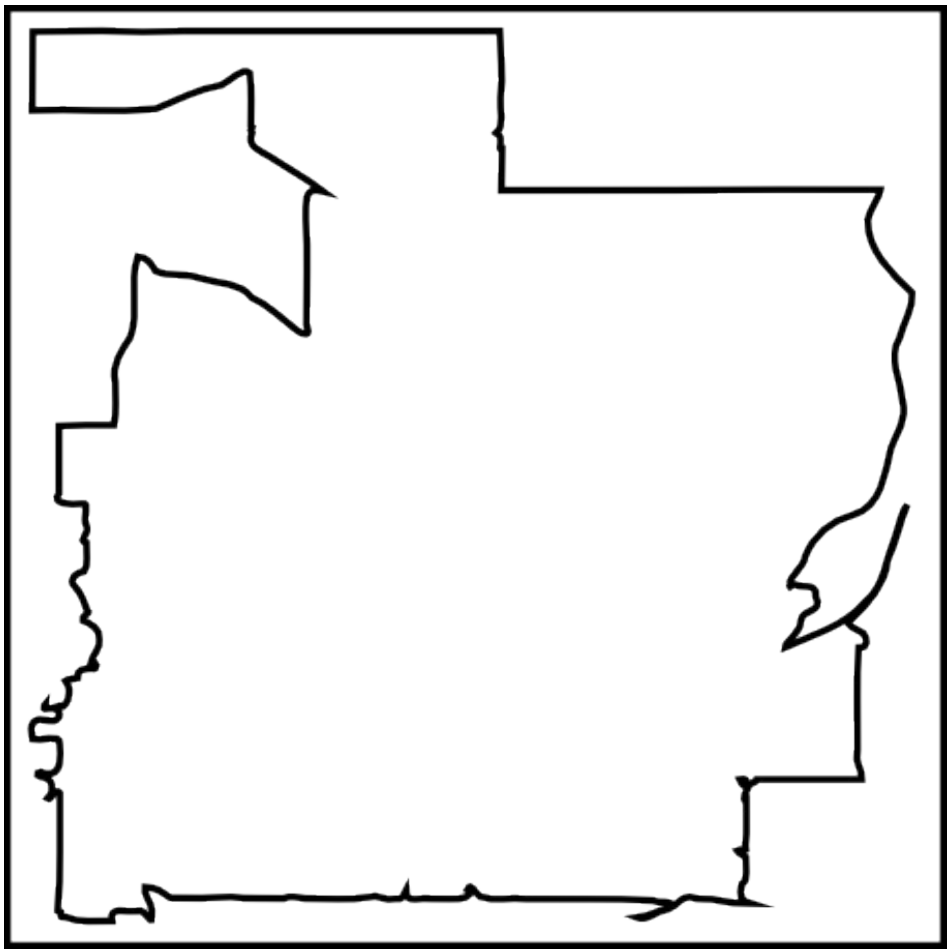
Hadoop - Local union



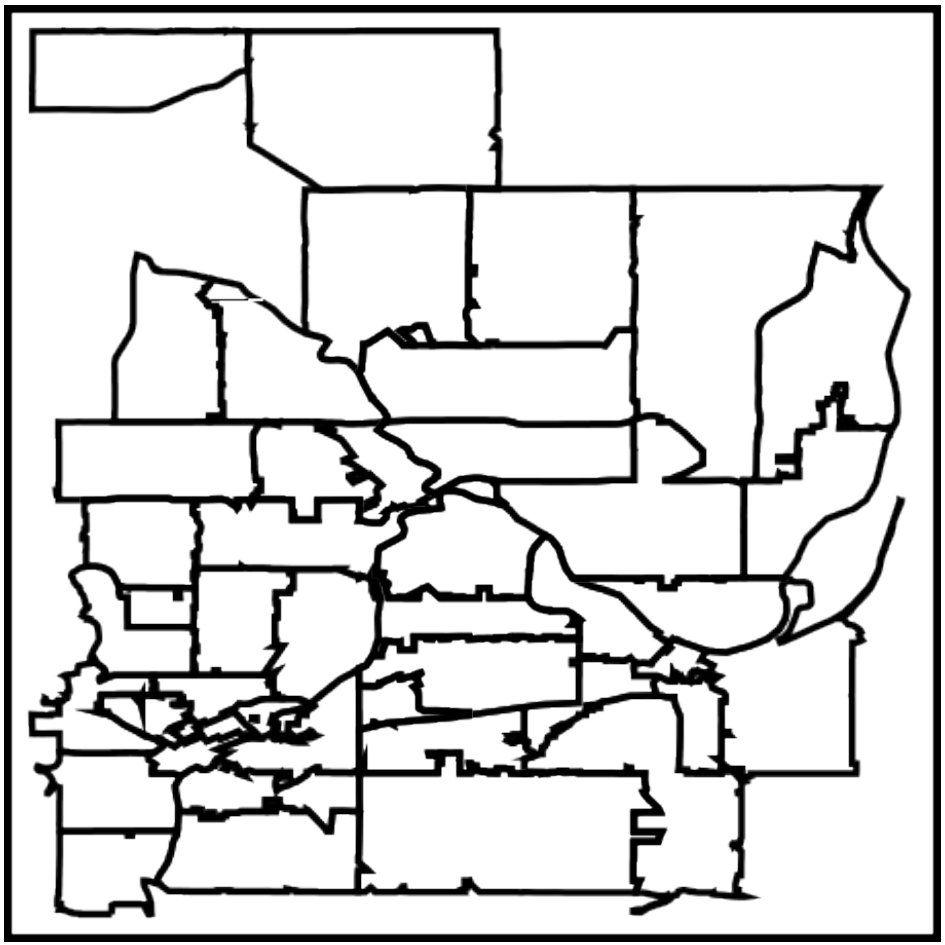
Hadoop - Global union



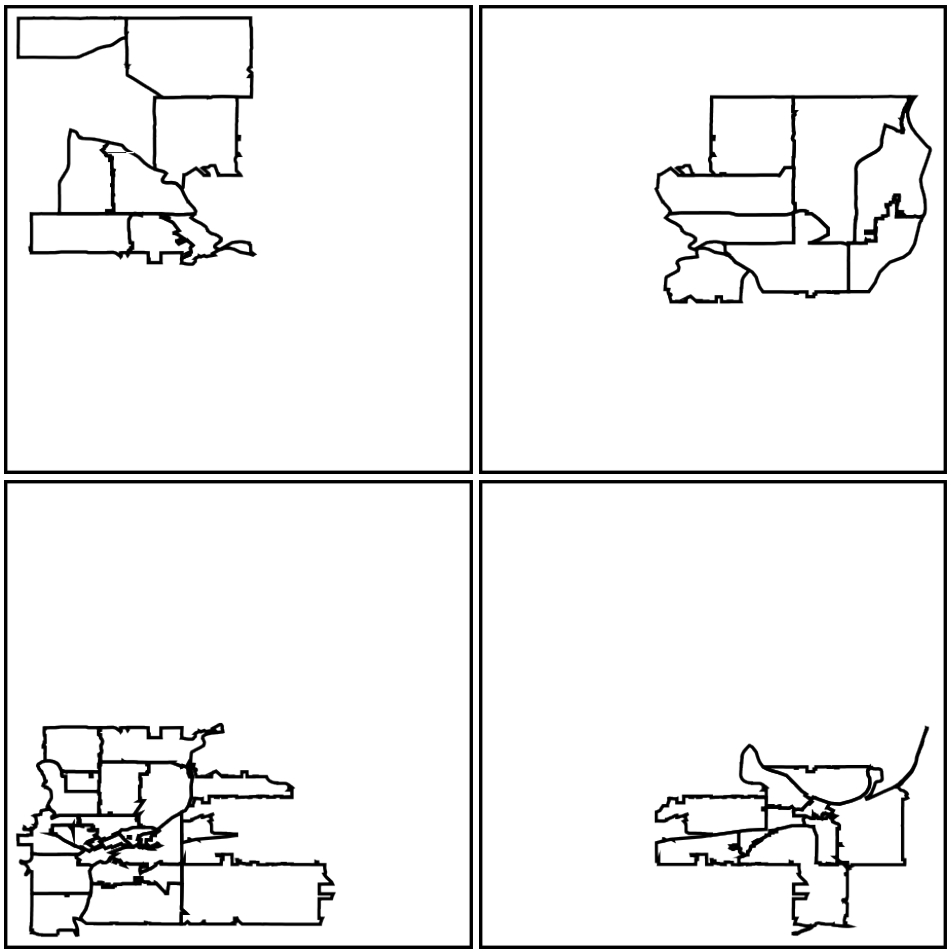
Hadoop - Global union



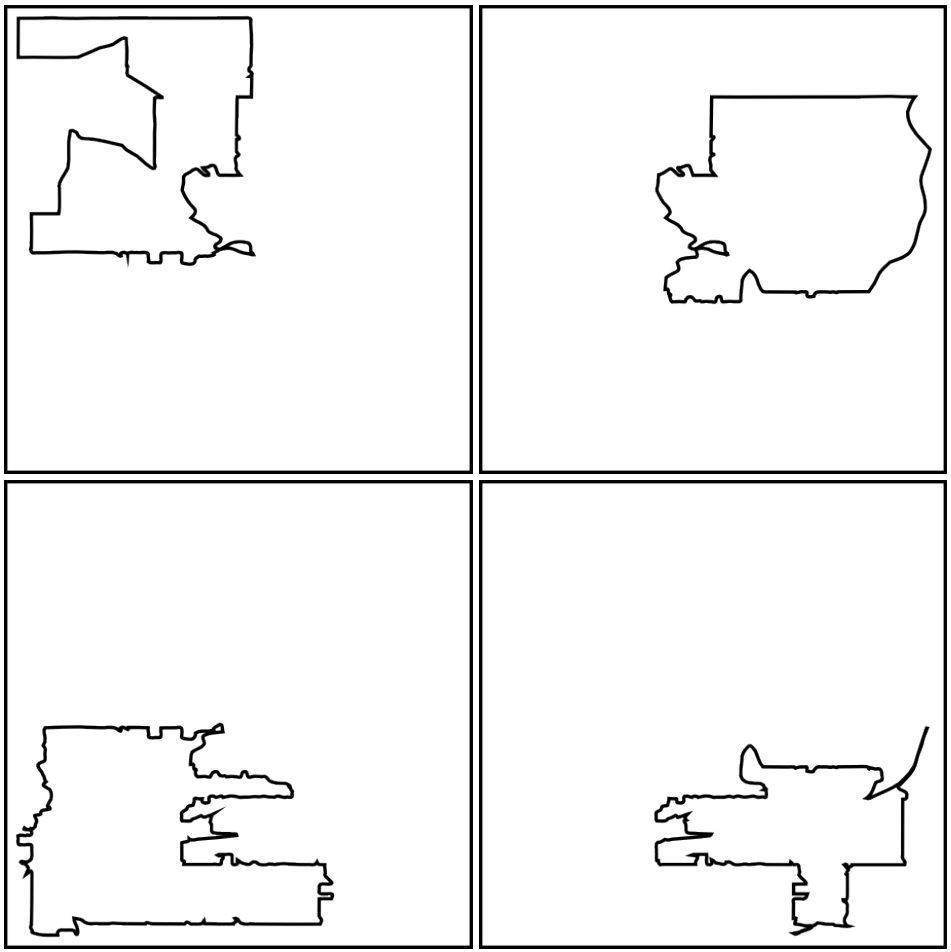
SpatialHadoop - Partition



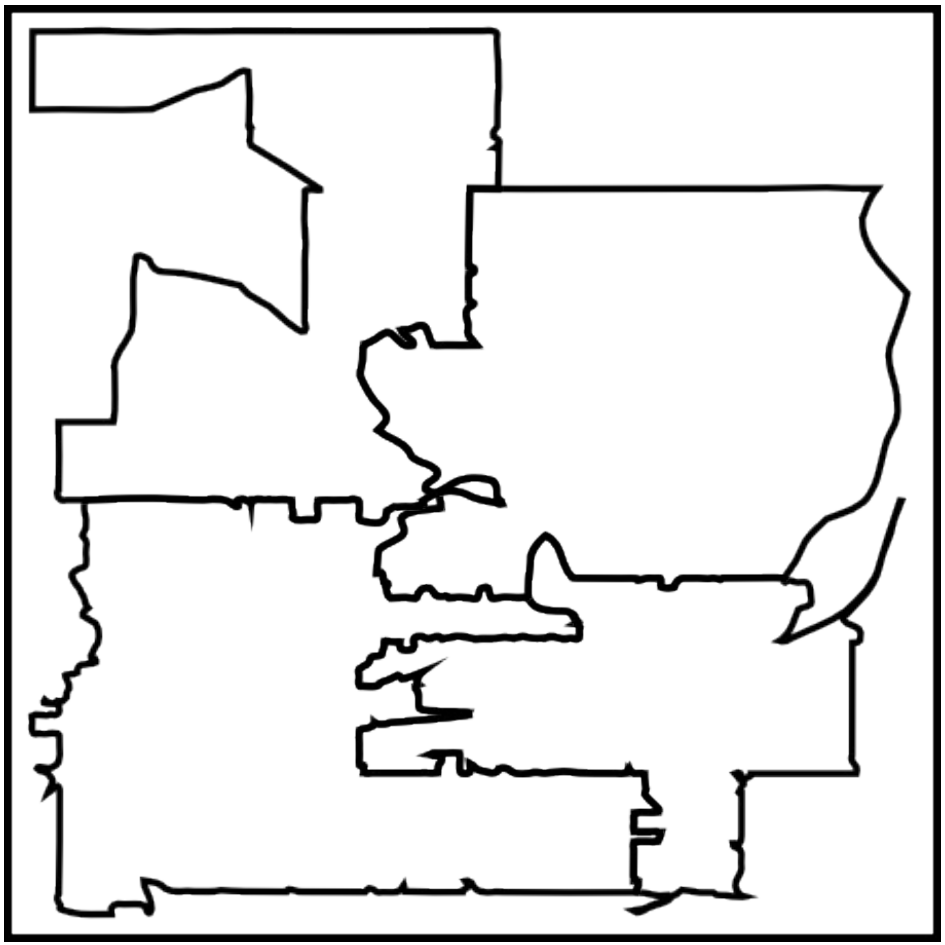
SpatialHadoop - Partition



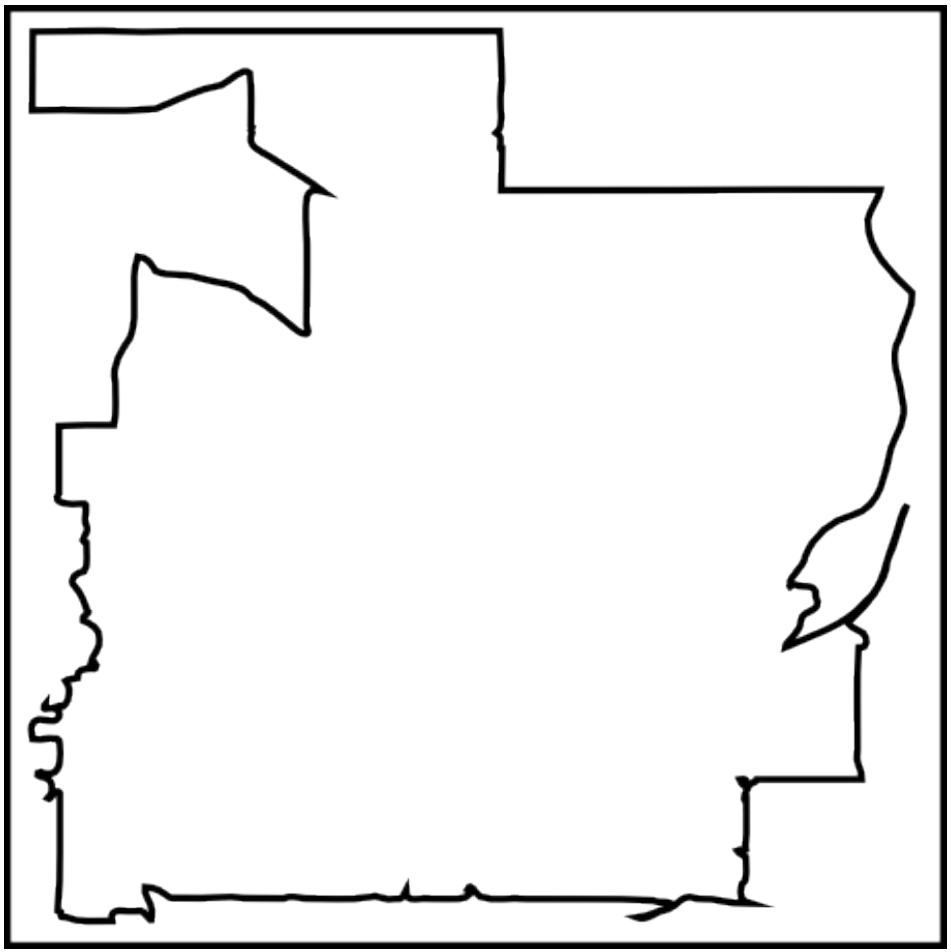
SpatialHadoop - Local union



SpatialHadoop - Global union

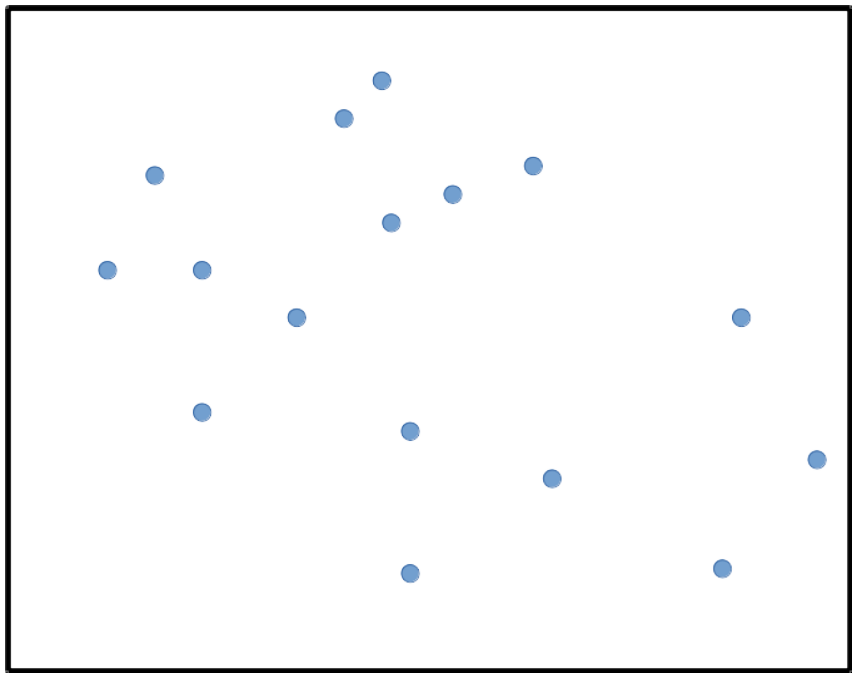


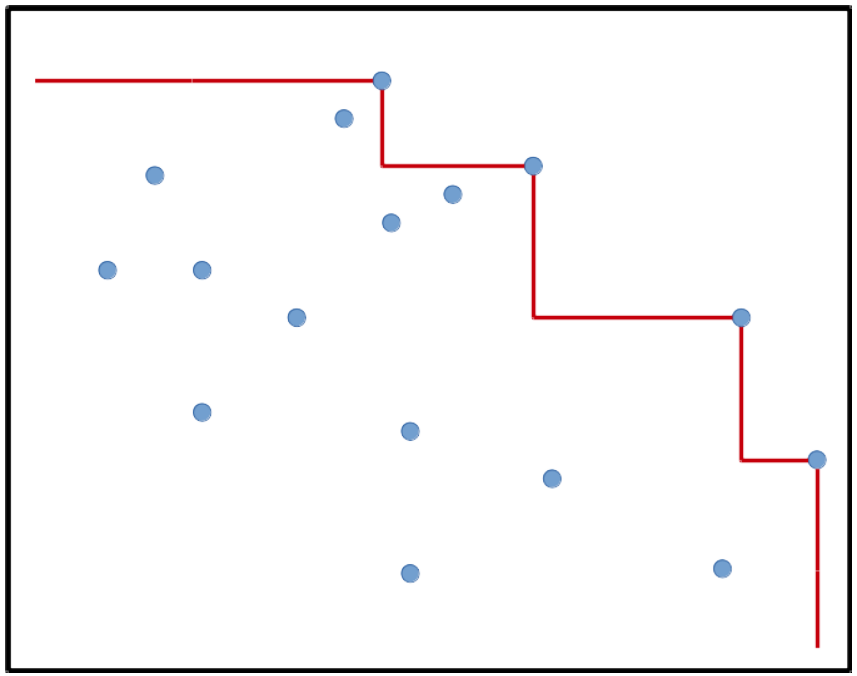
SpatialHadoop - Global union



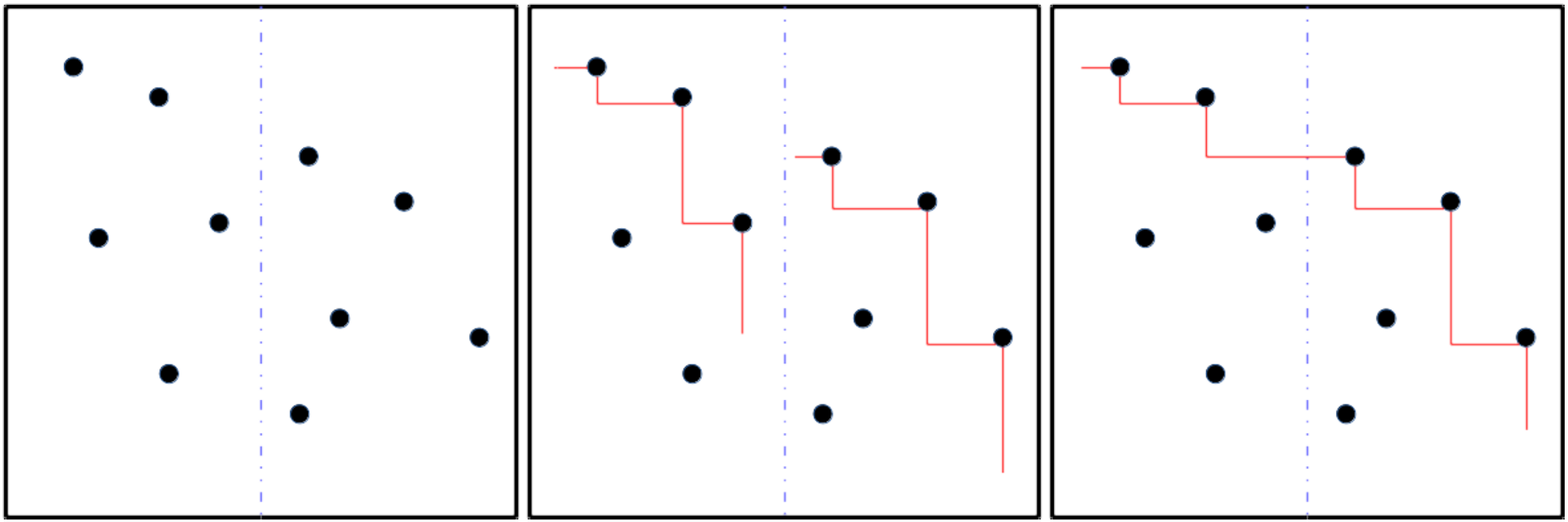
Agenda

Skyline

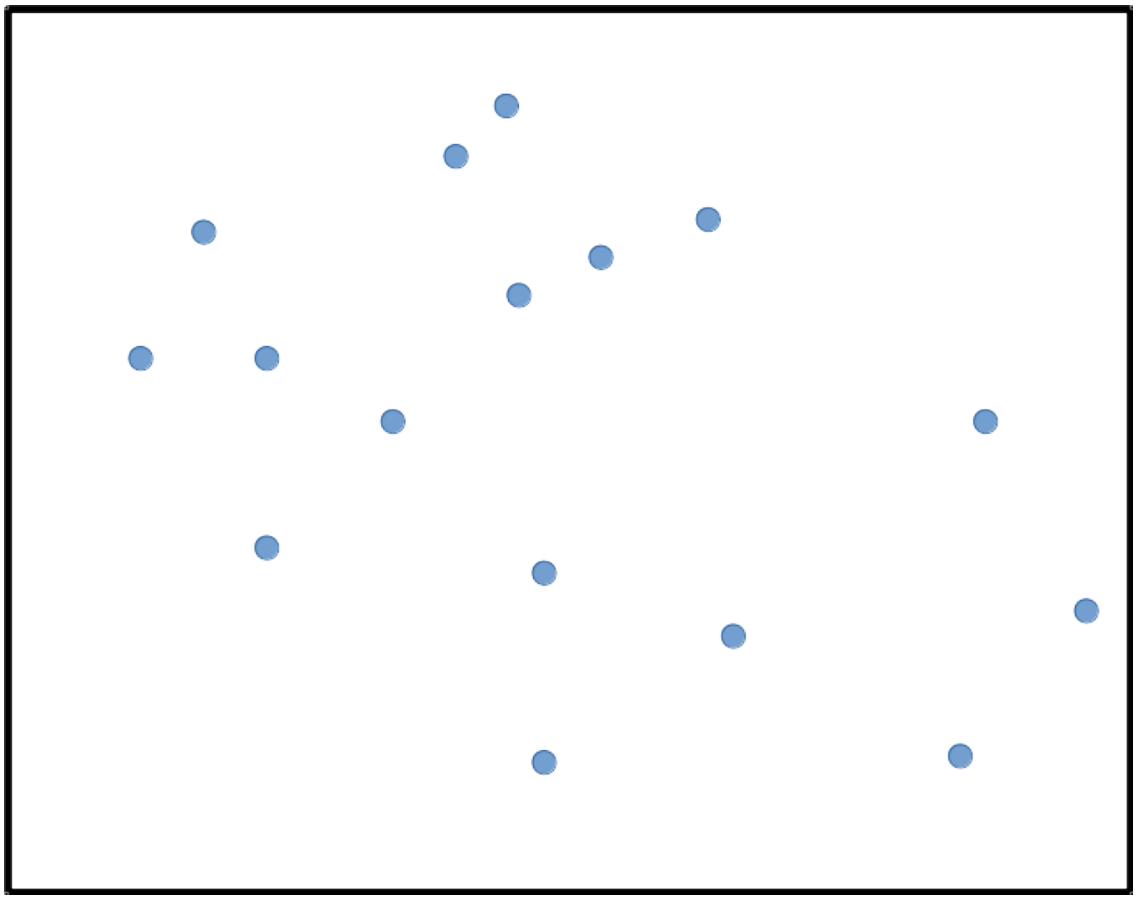




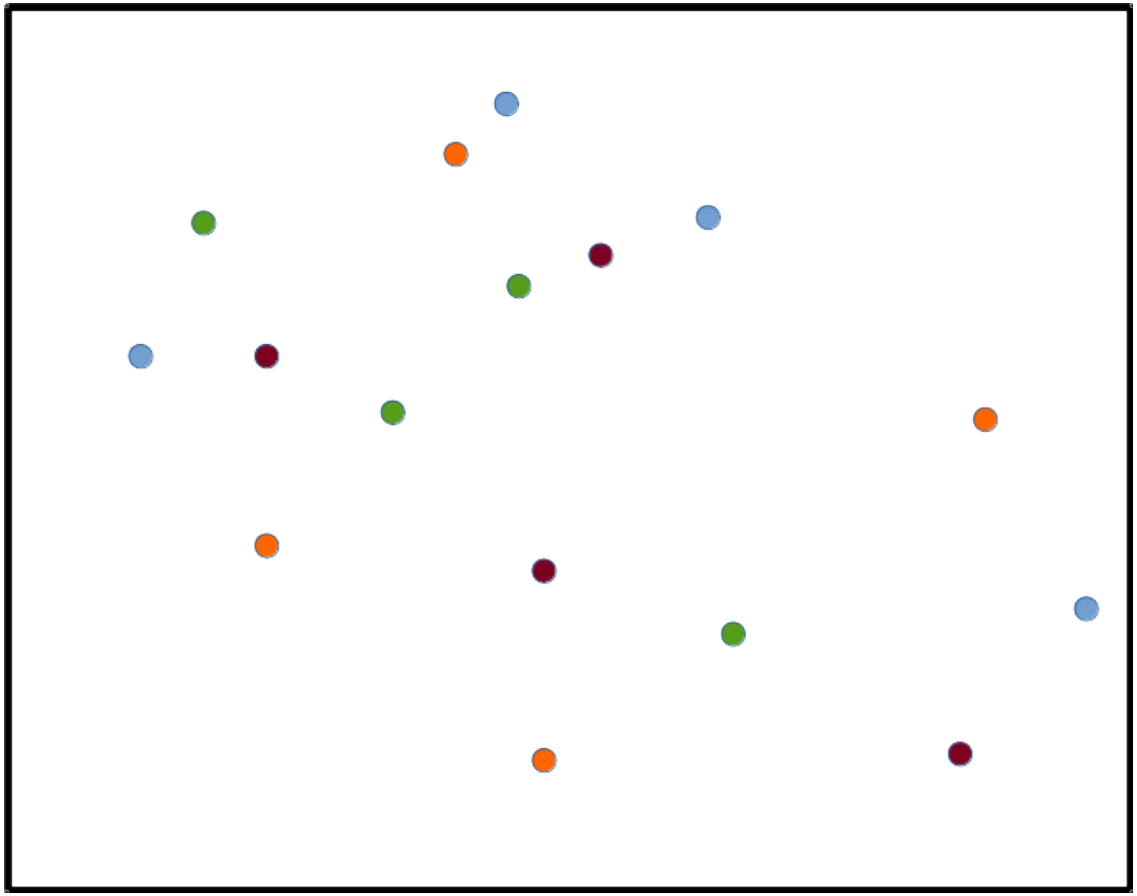
Divide and conquer approach



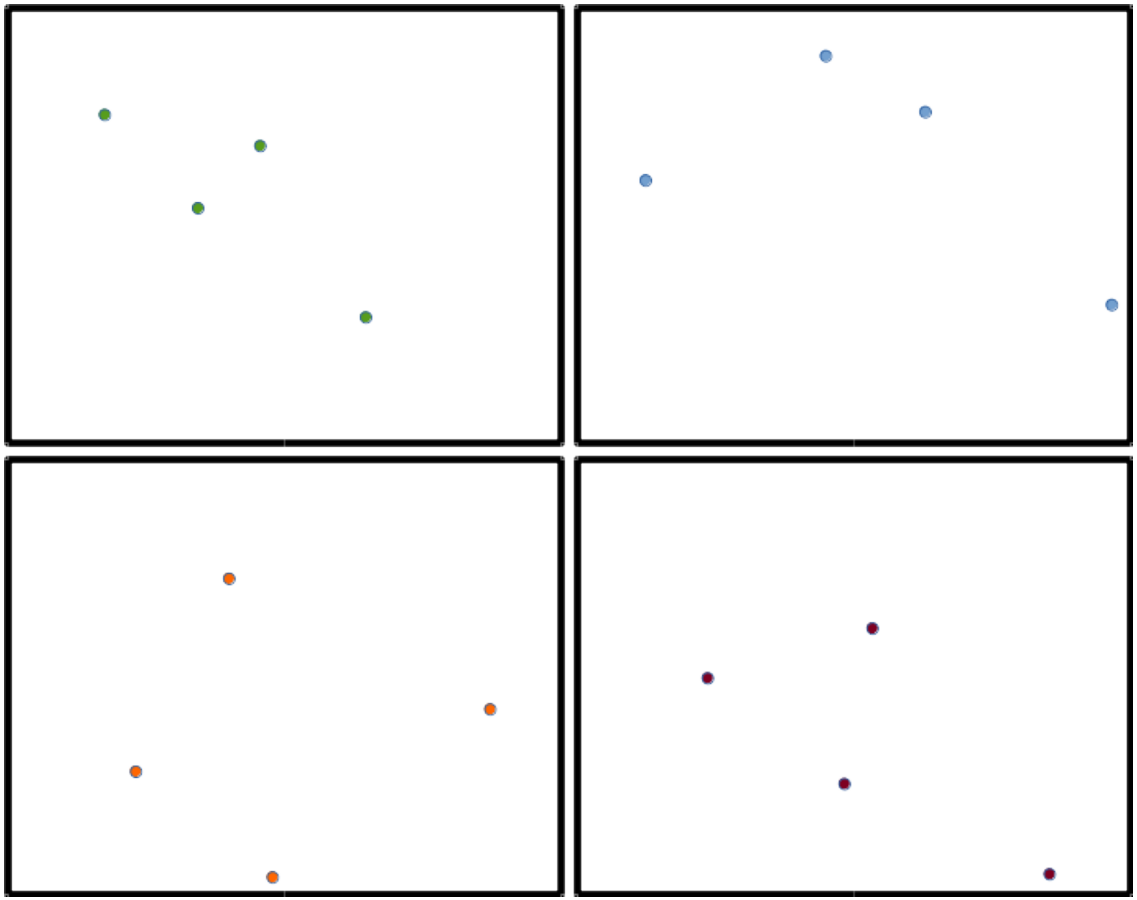
Hadoop - Partition



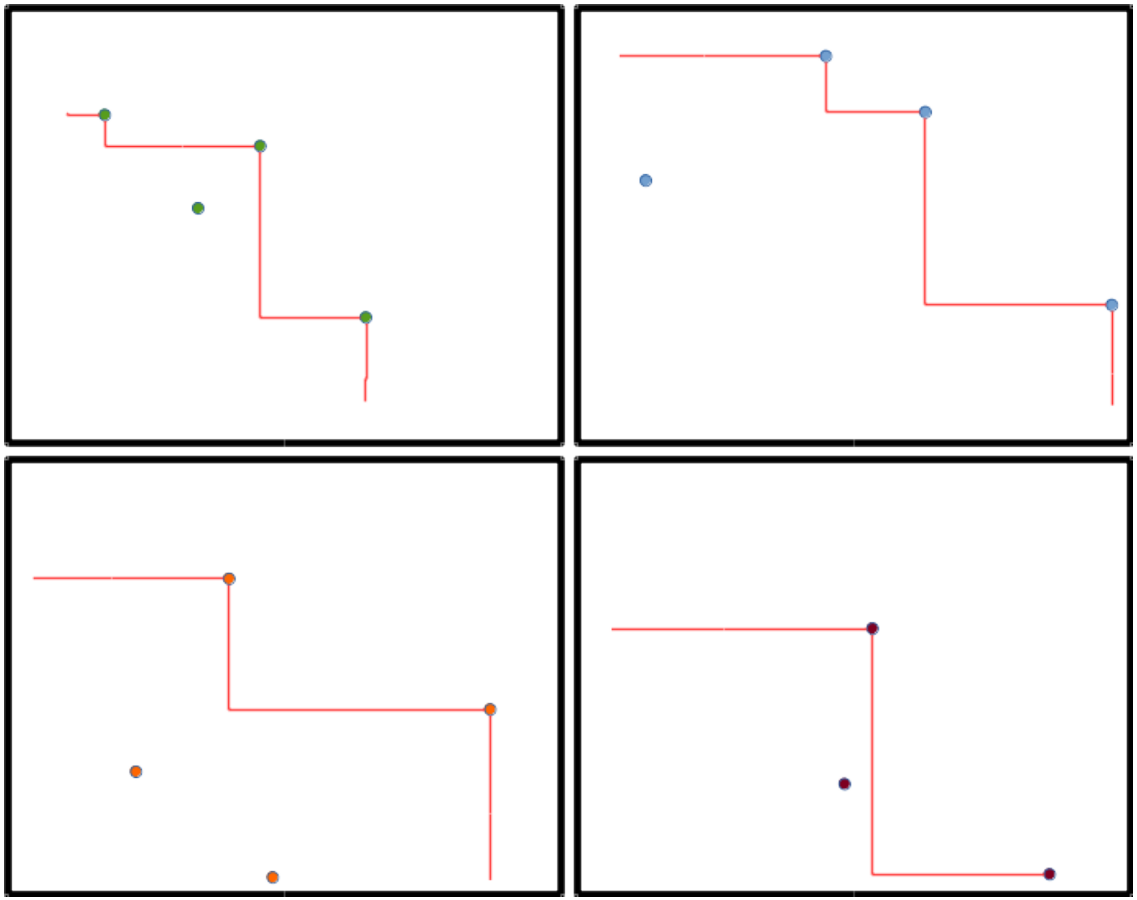
Hadoop - Partition



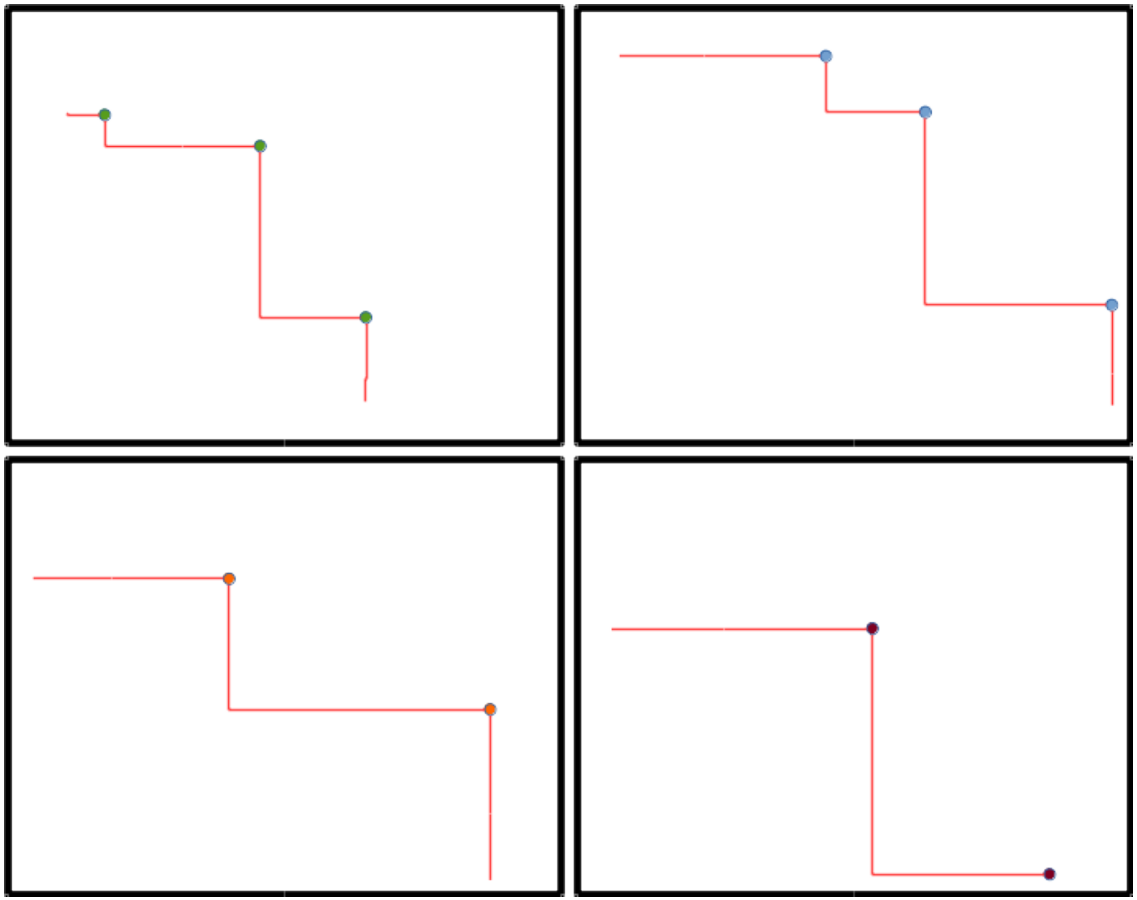
Hadoop - Partition



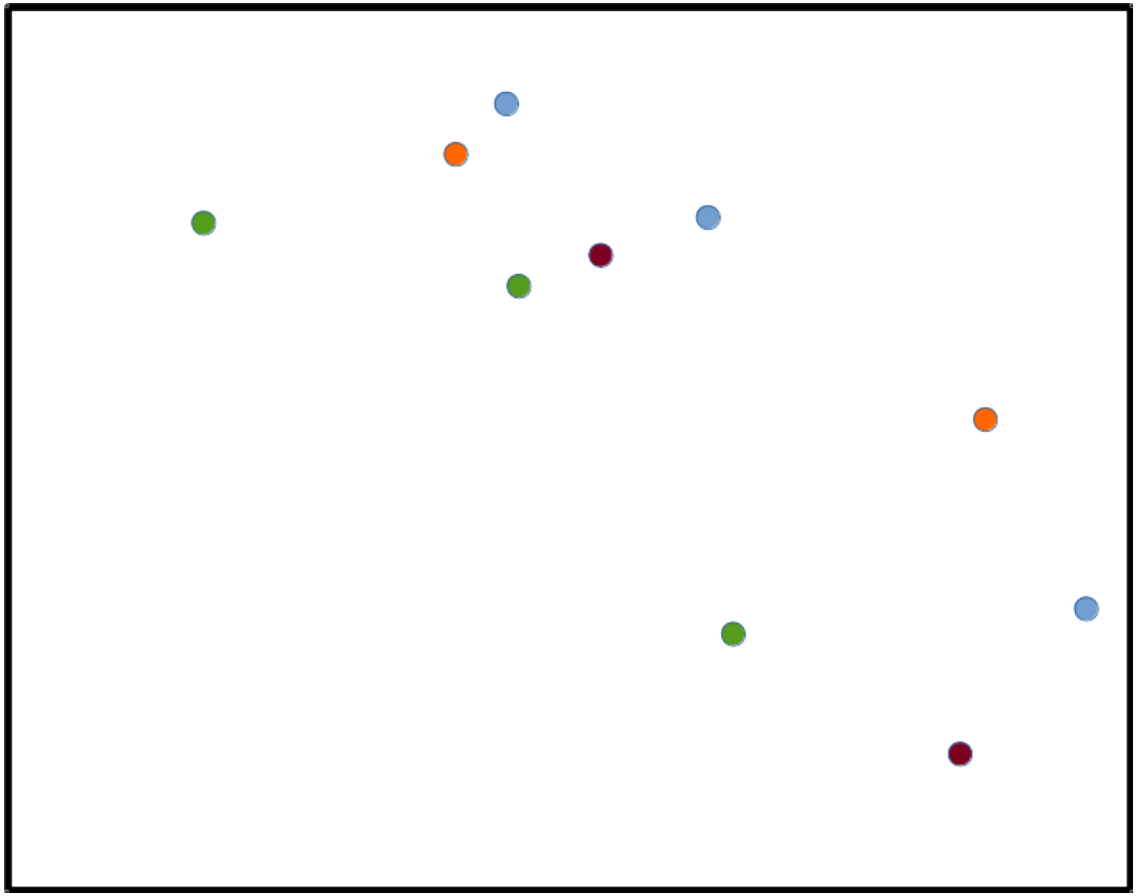
Hadoop - Local



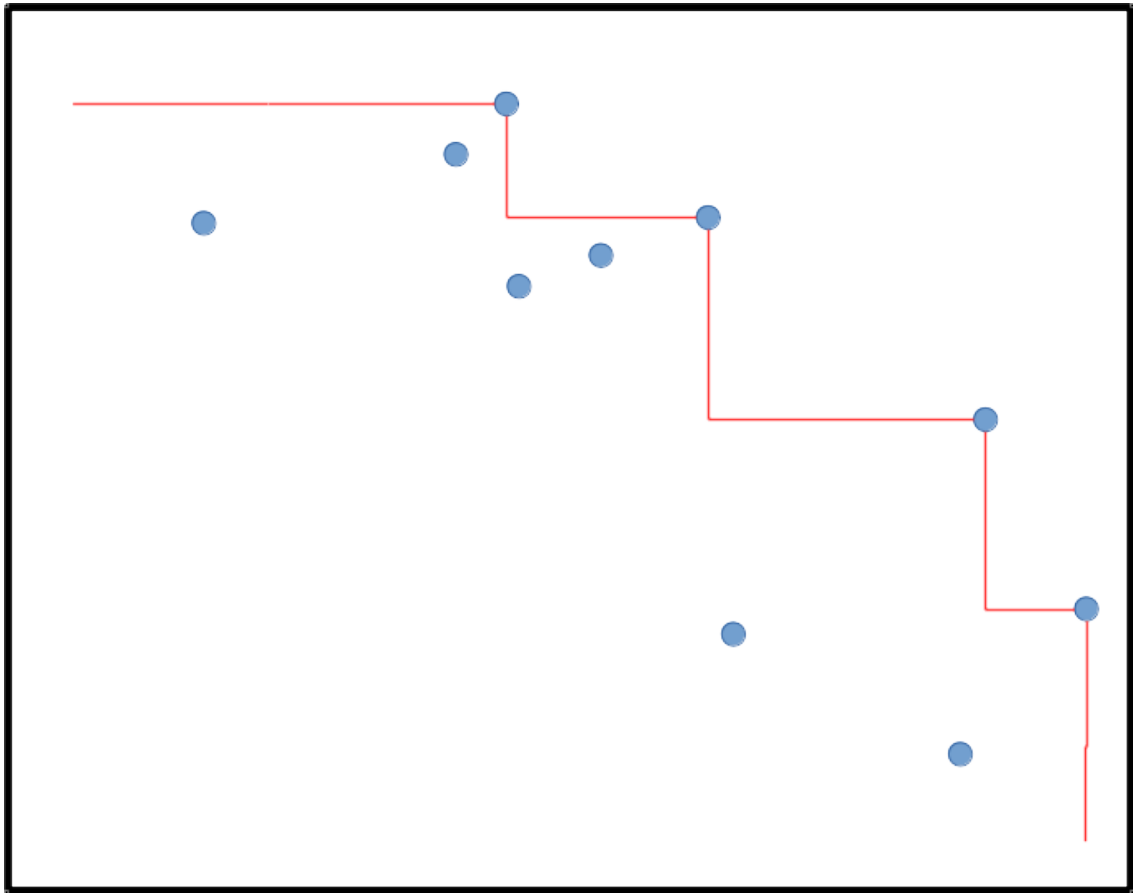
Hadoop - Local



Hadoop - Global



Hadoop - Global



Skyline in SpatialHadoop

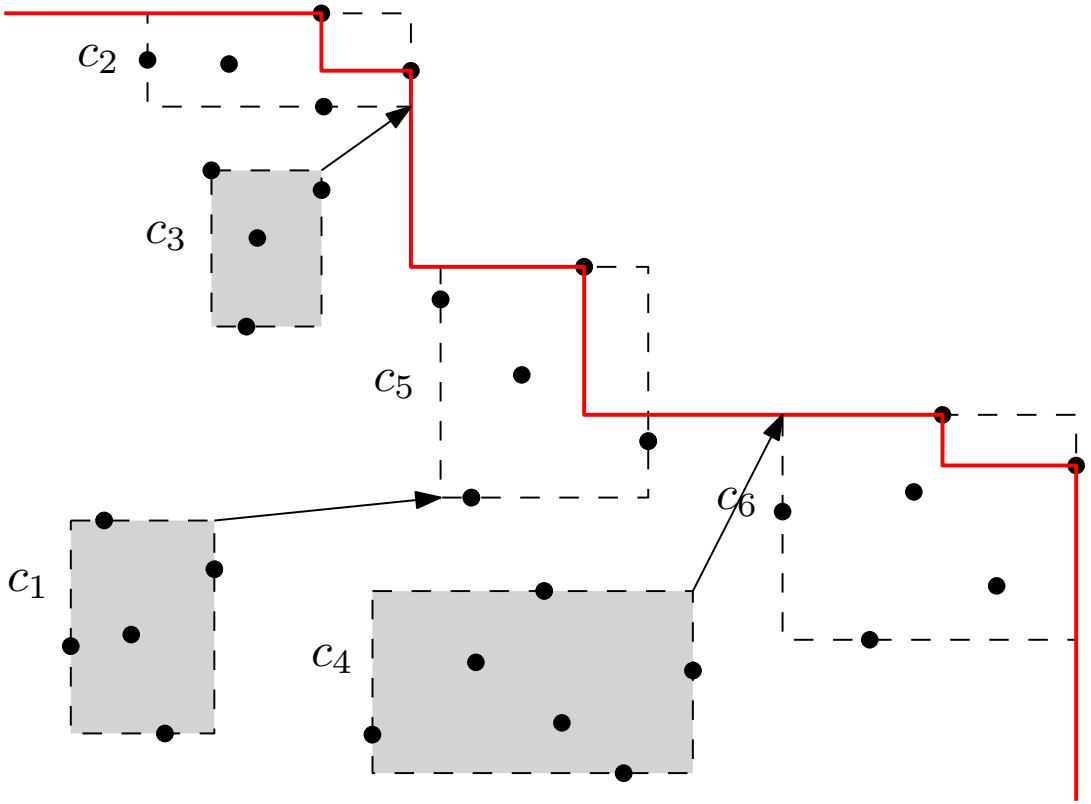
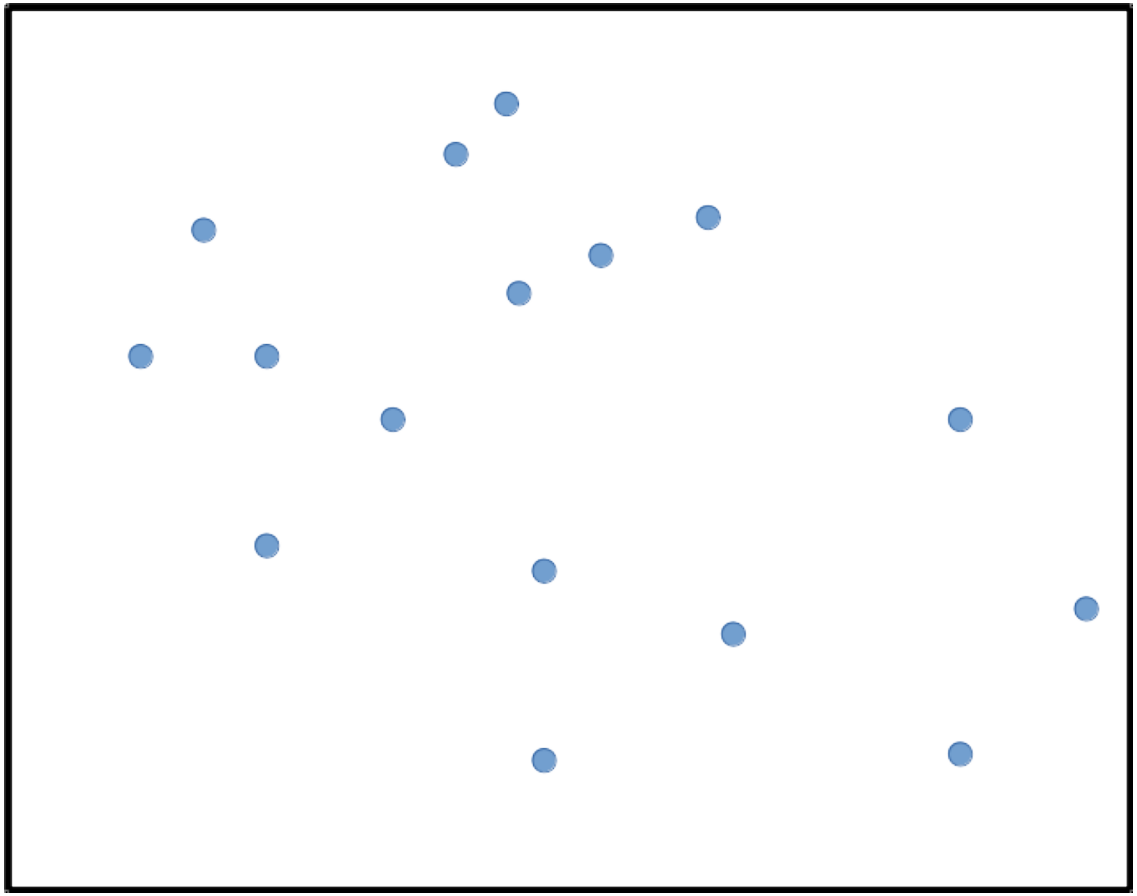
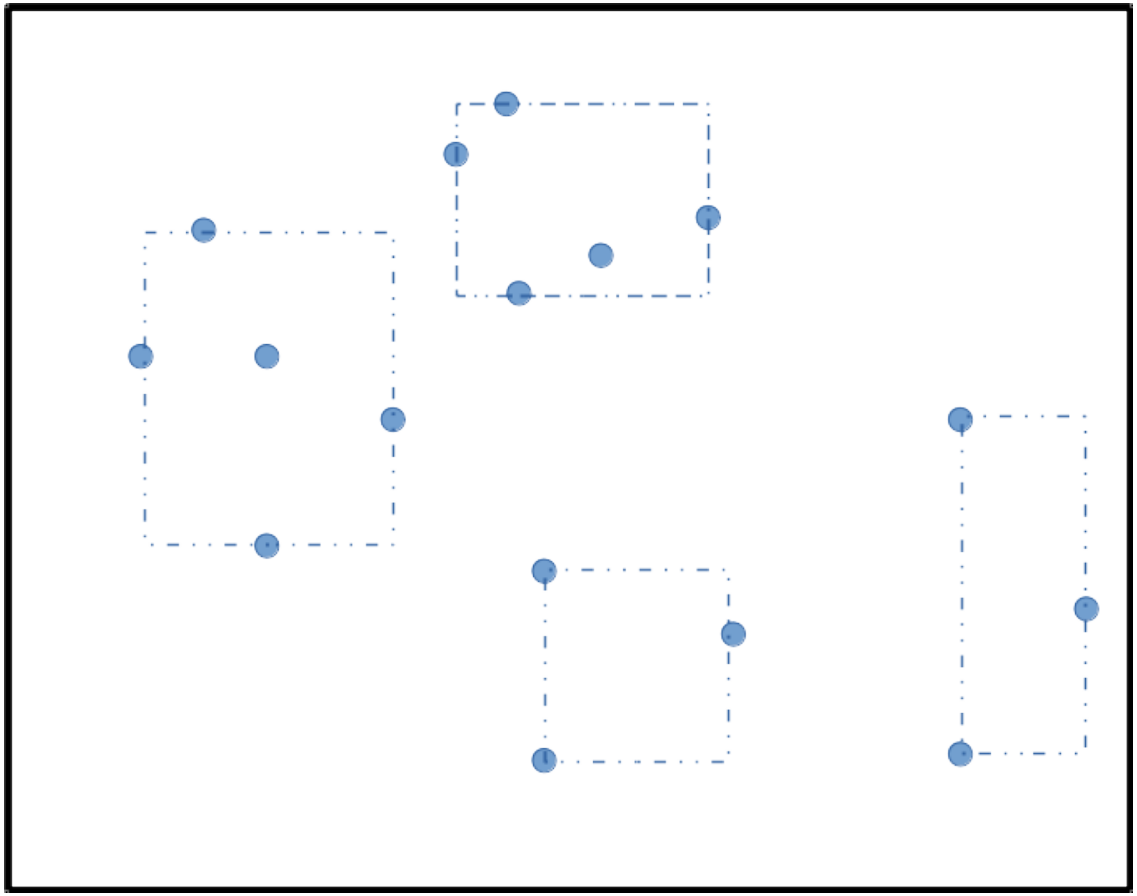


Figure 4: Skyline in SpatialHadoop

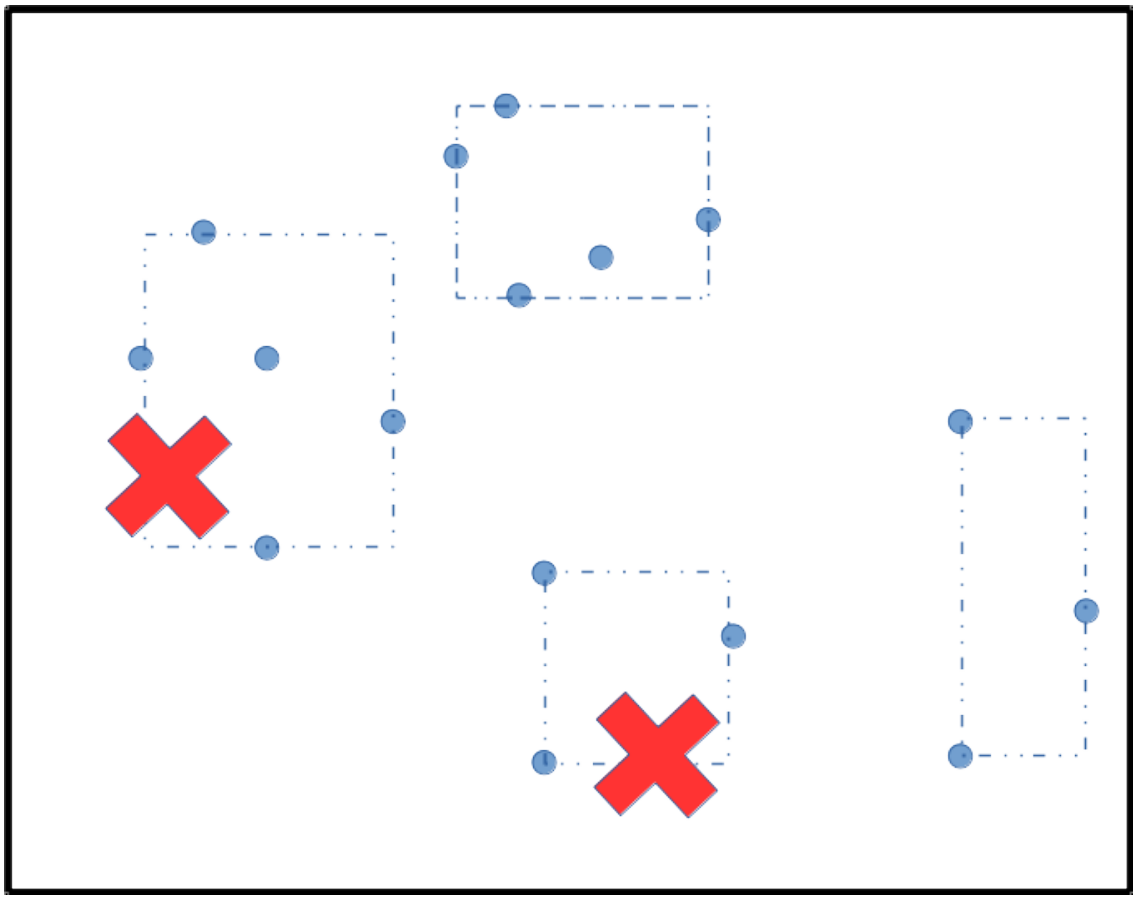
SpatialHadoop - Partition



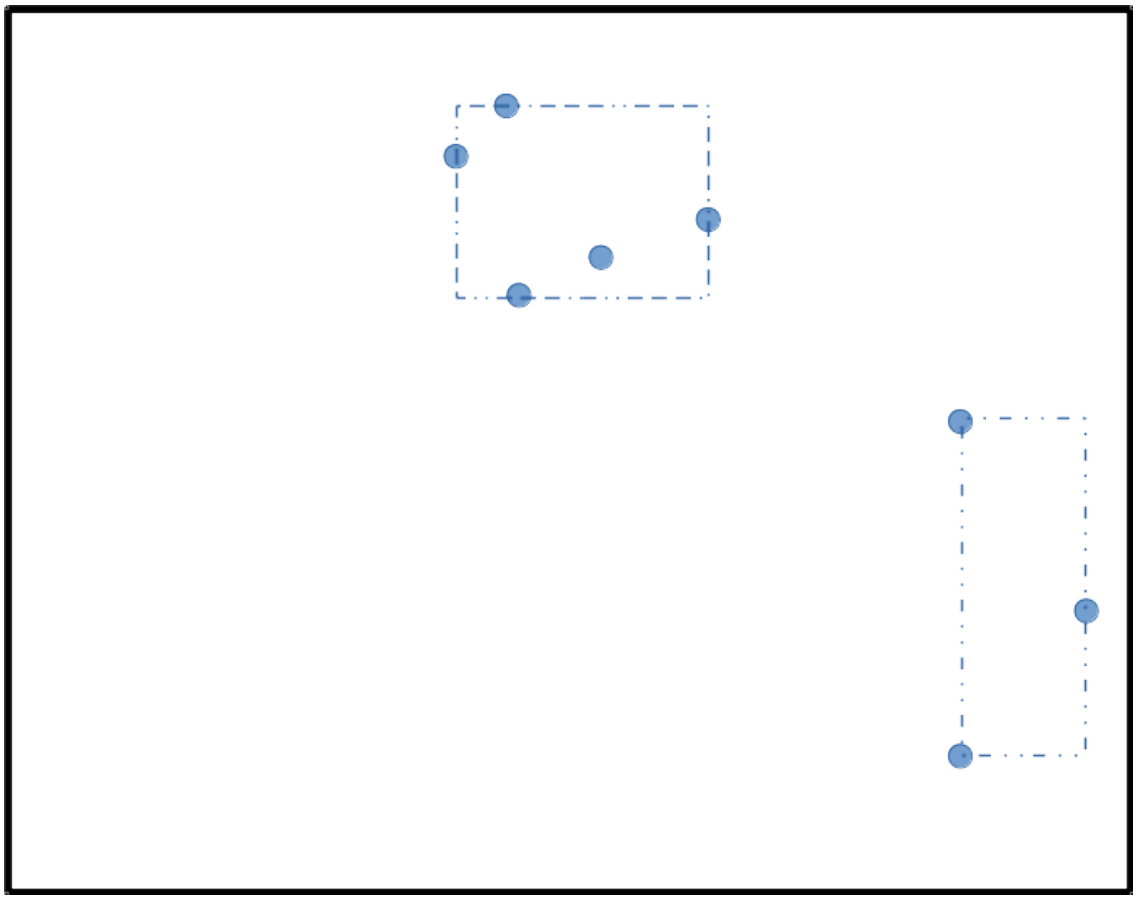
SpatialHadoop - Partition



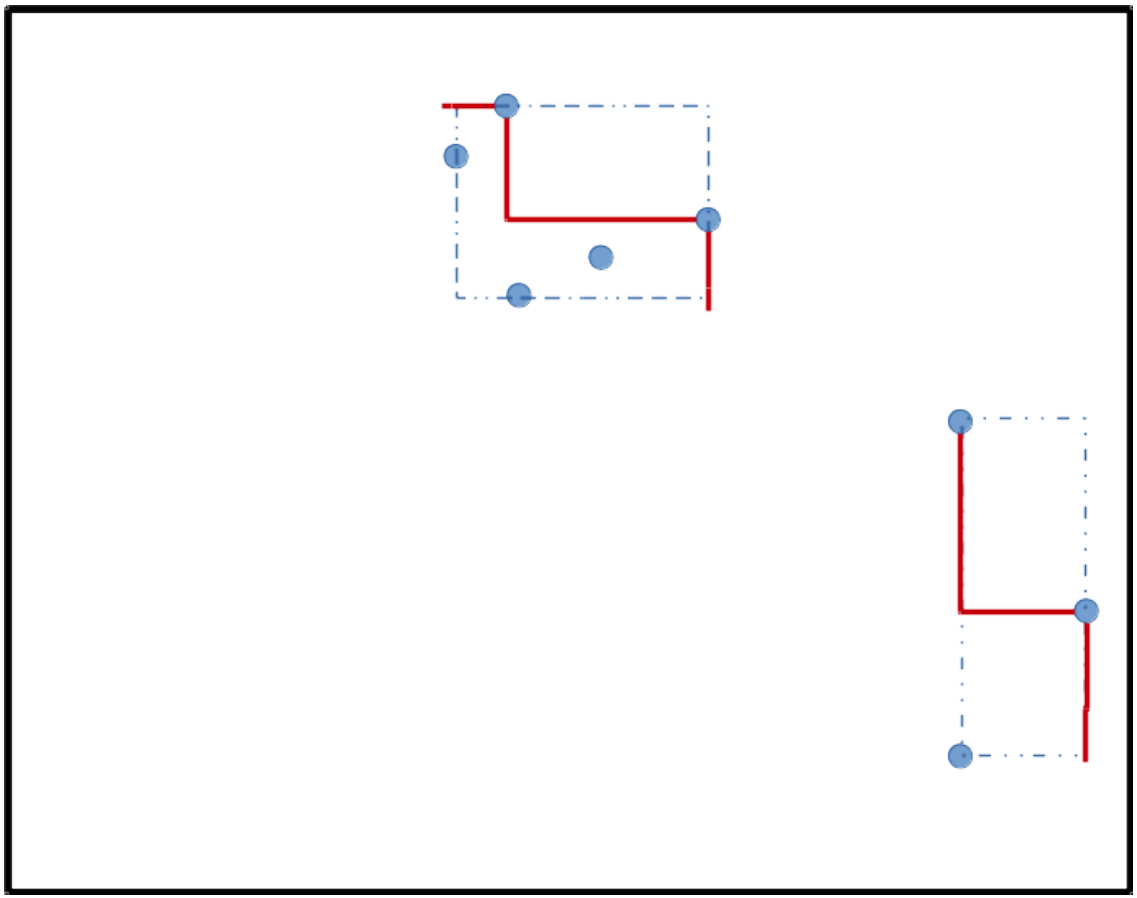
SpatialHadoop - Pruning



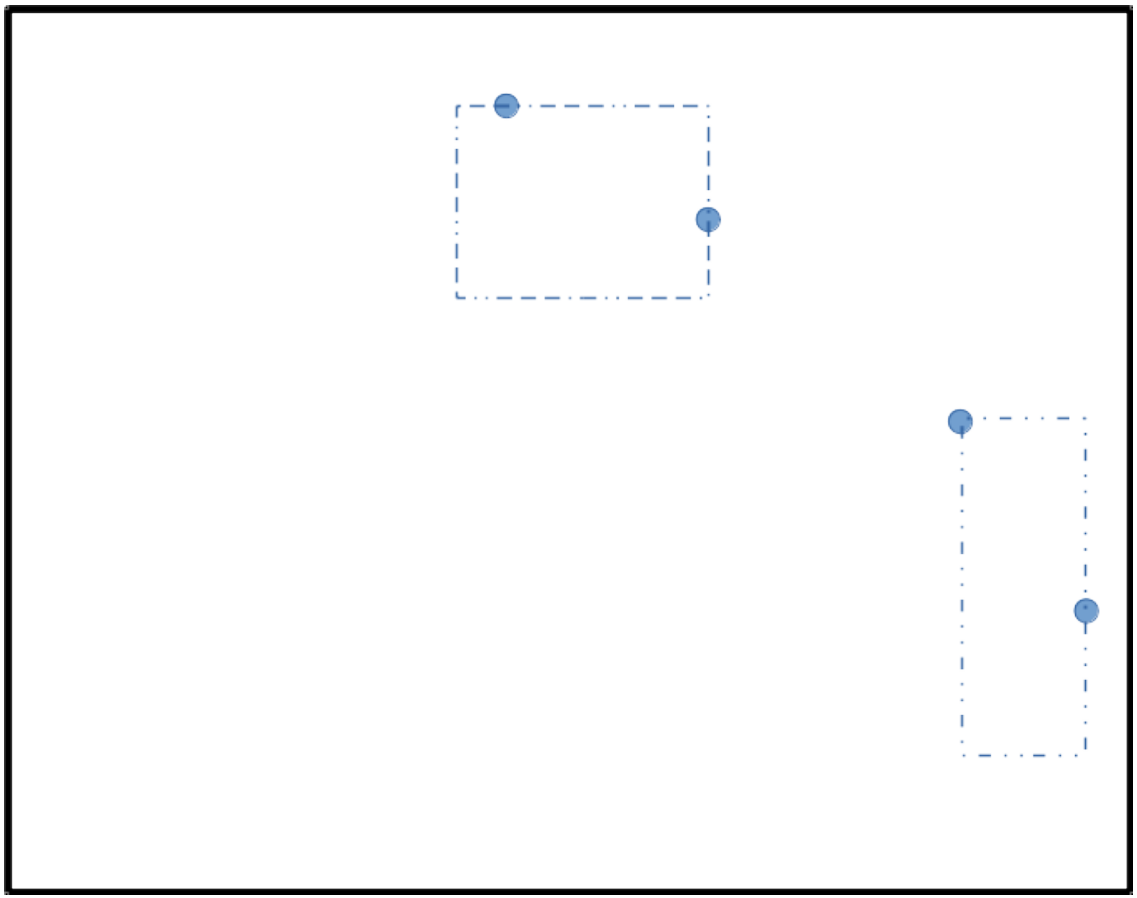
SpatialHadoop - Pruning



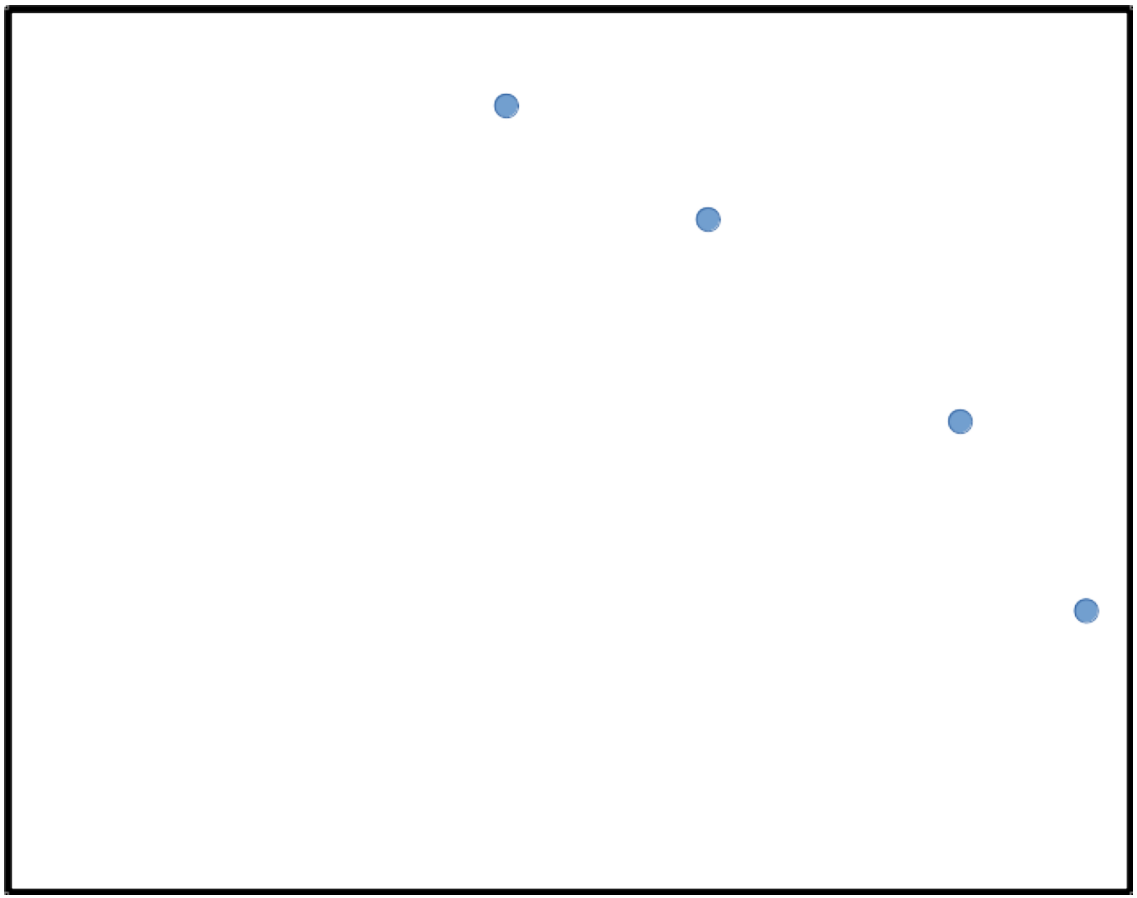
SpatialHadoop - Local



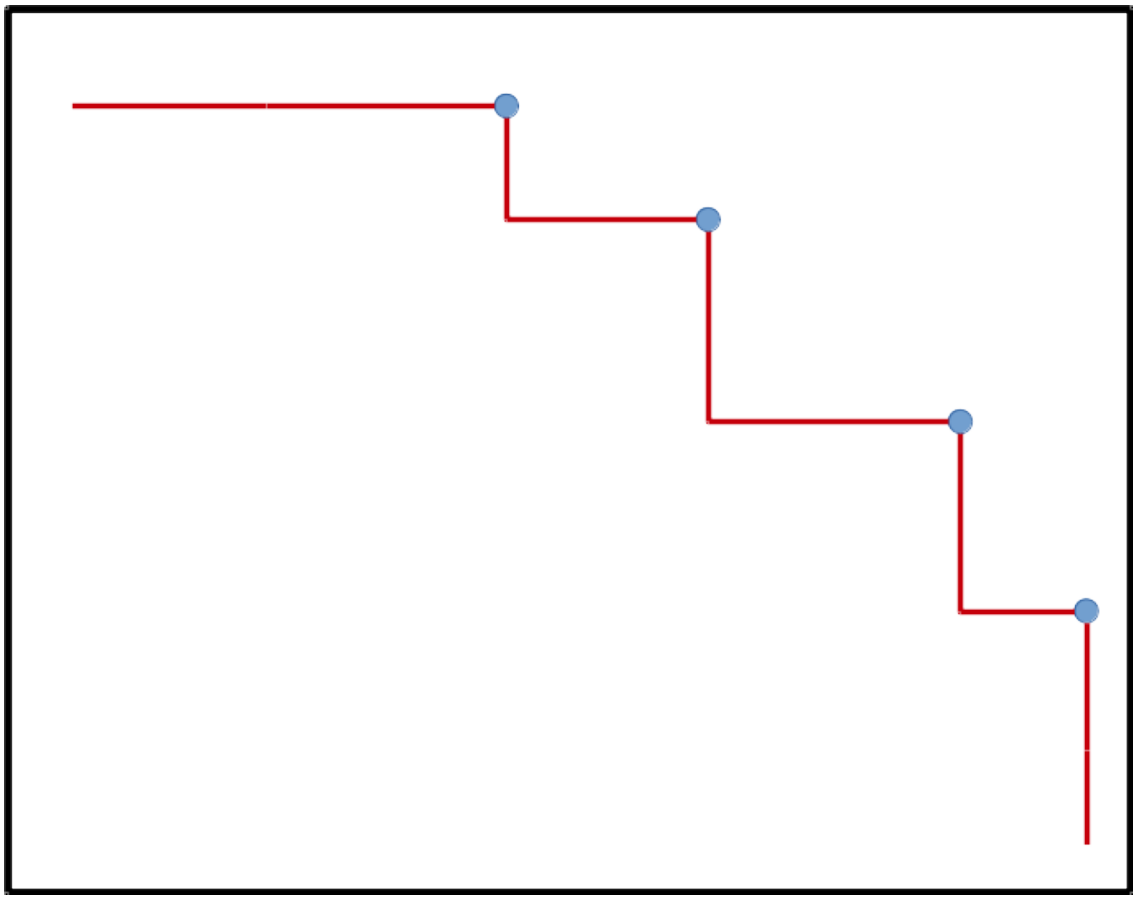
SpatialHadoop - Local



SpatialHadoop - Global

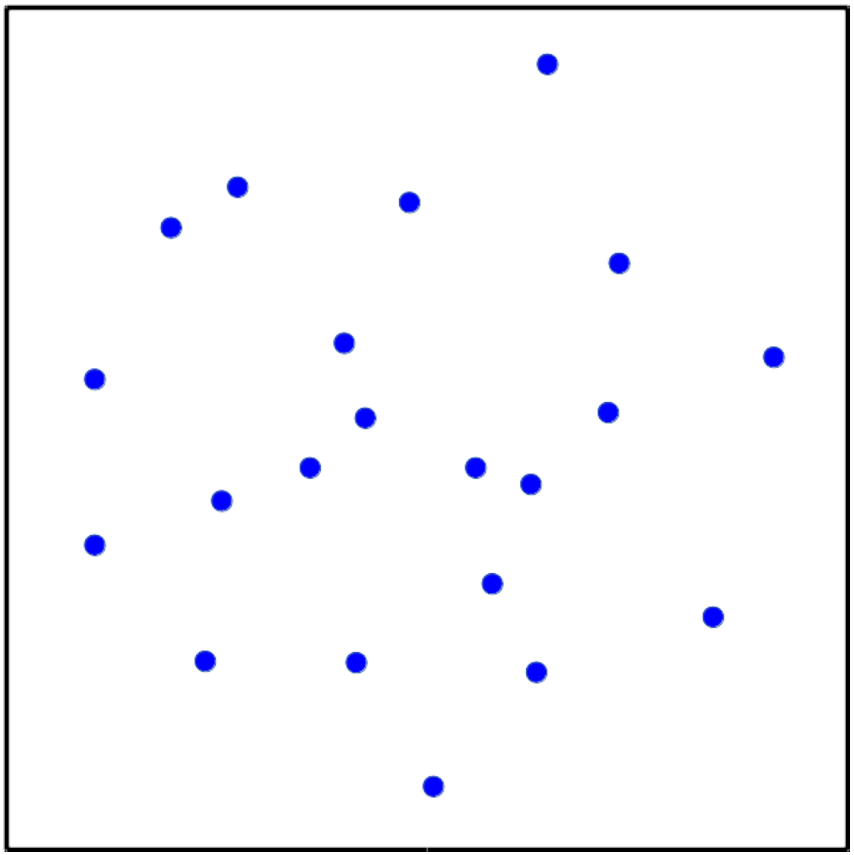


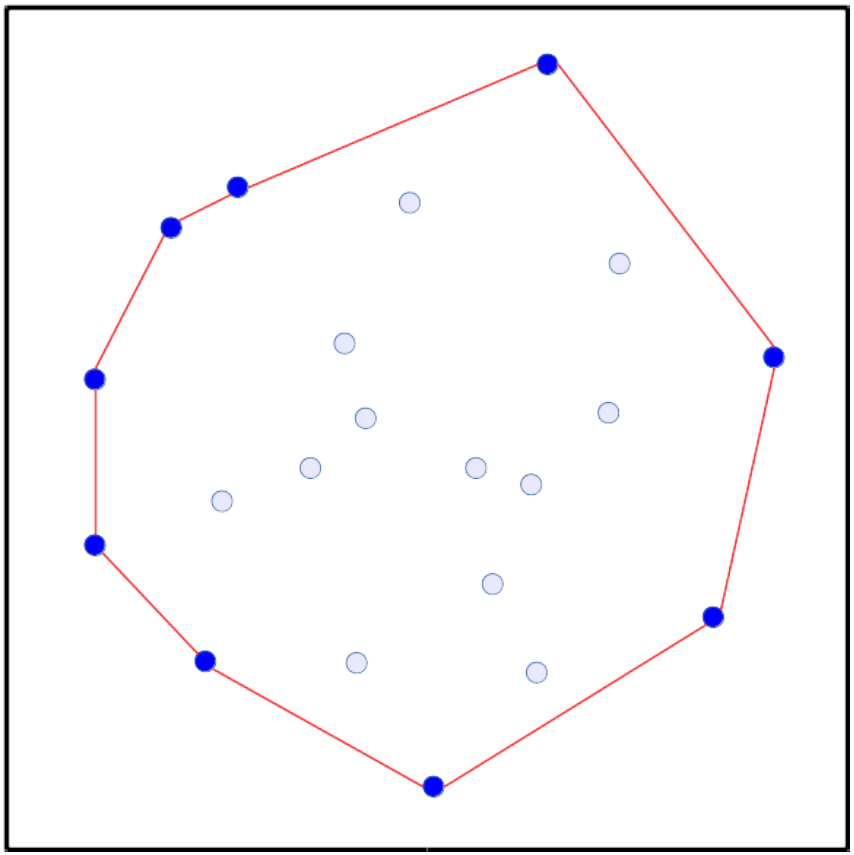
SpatialHadoop - Global



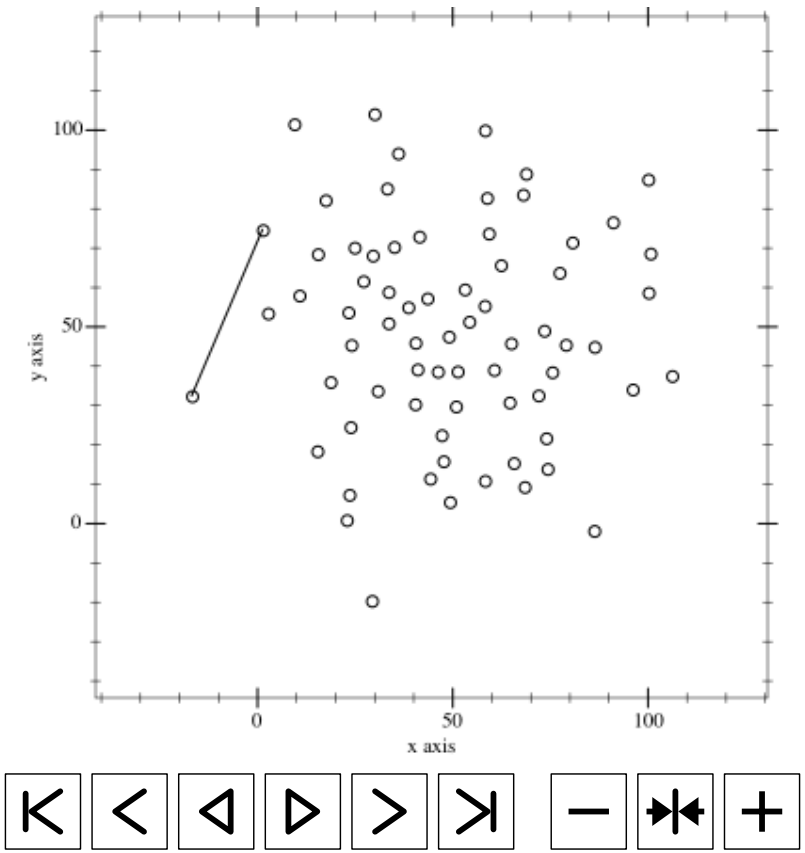
Agenda

Convex Hull

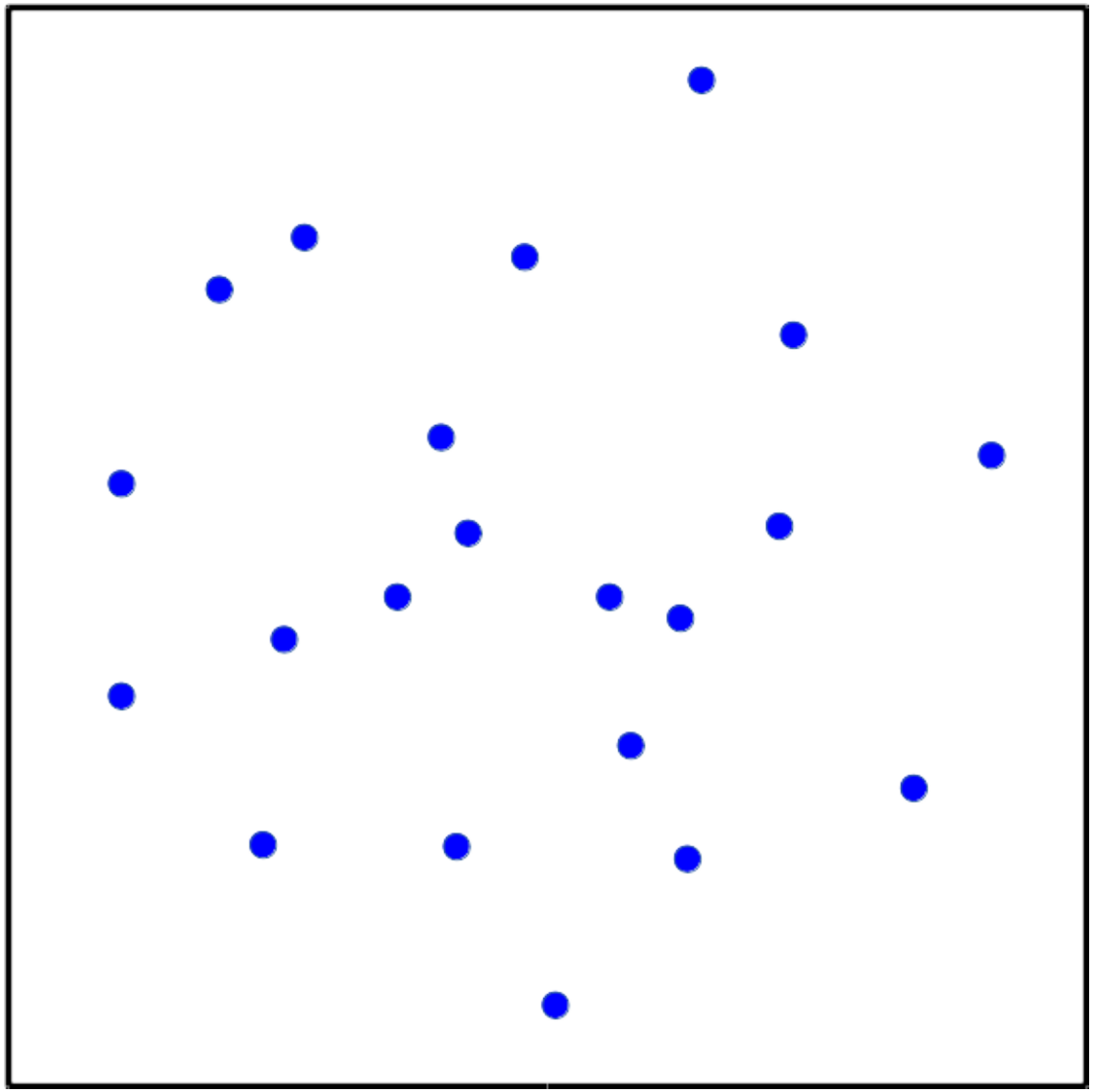




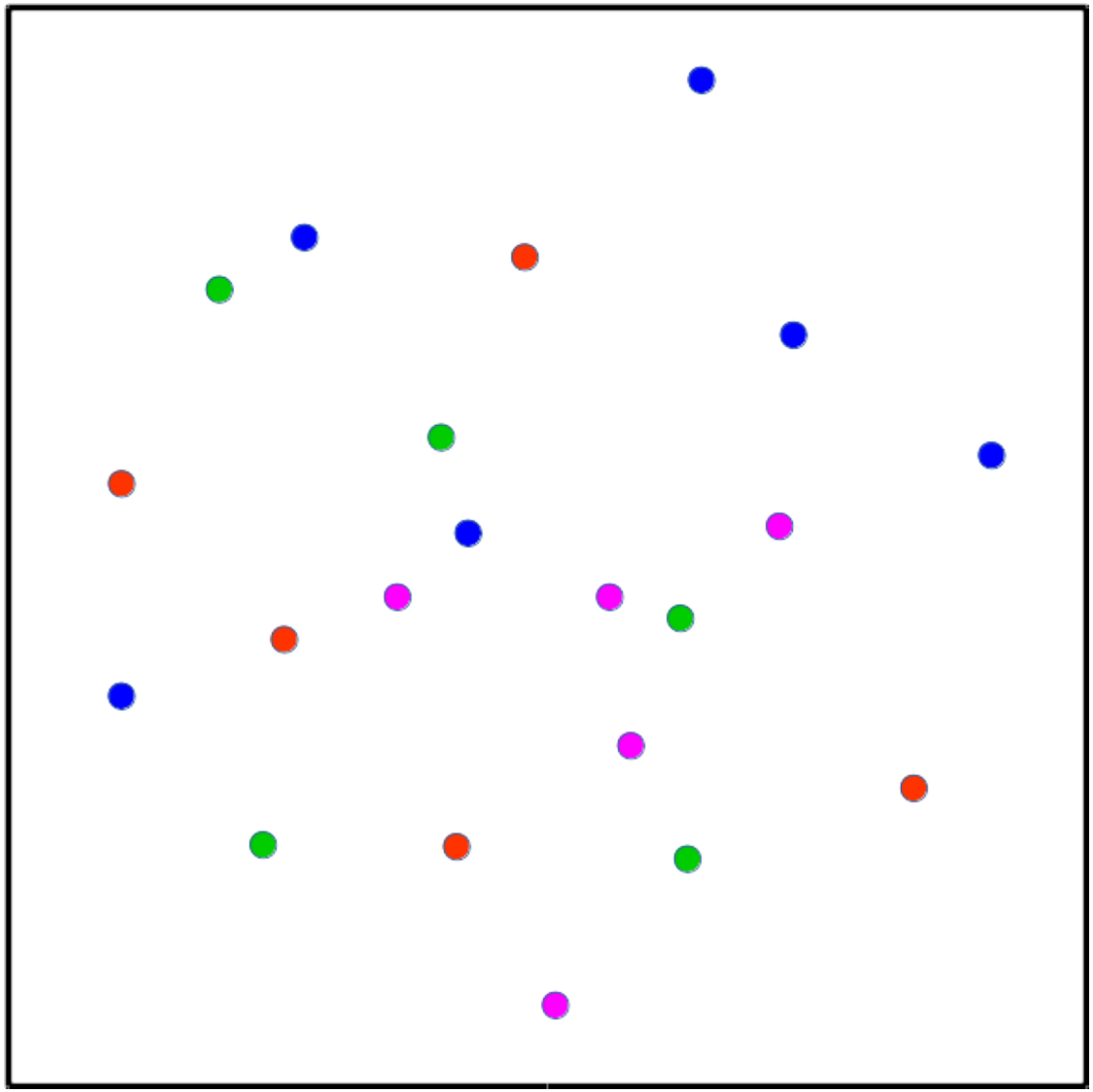
Andrew's Monotone Chain algorithm



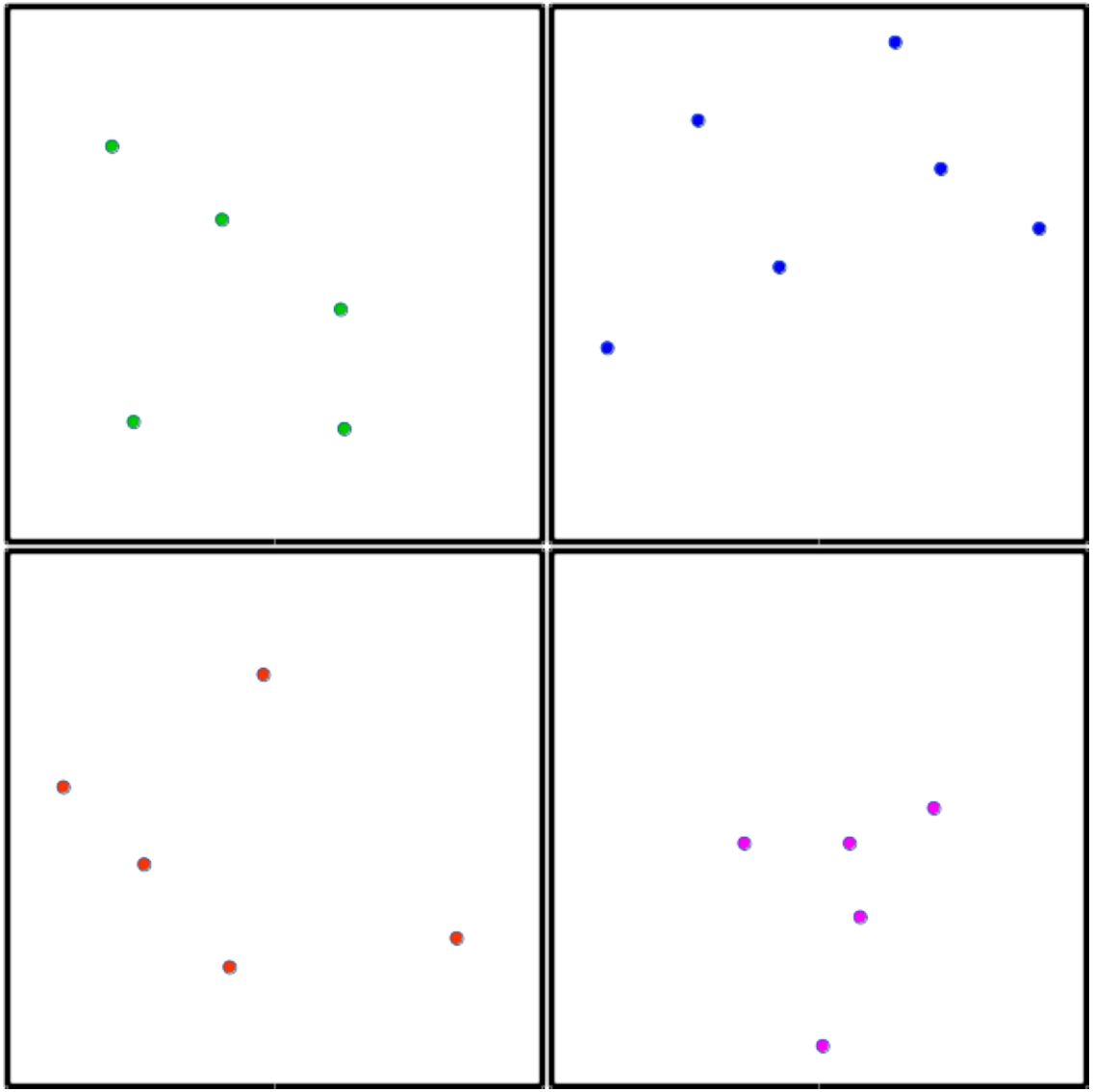
Hadoop - Partition



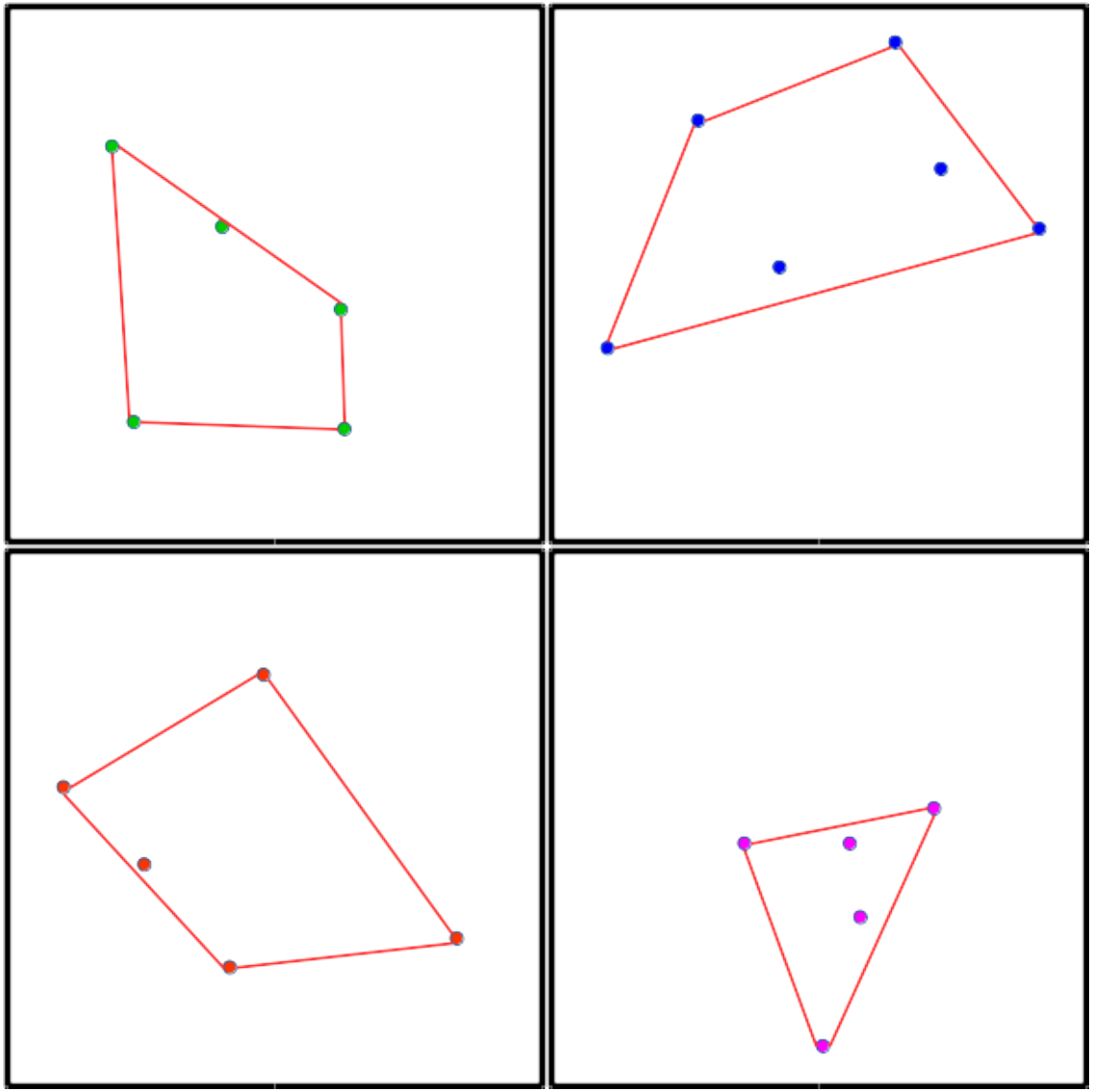
Hadoop - Partition



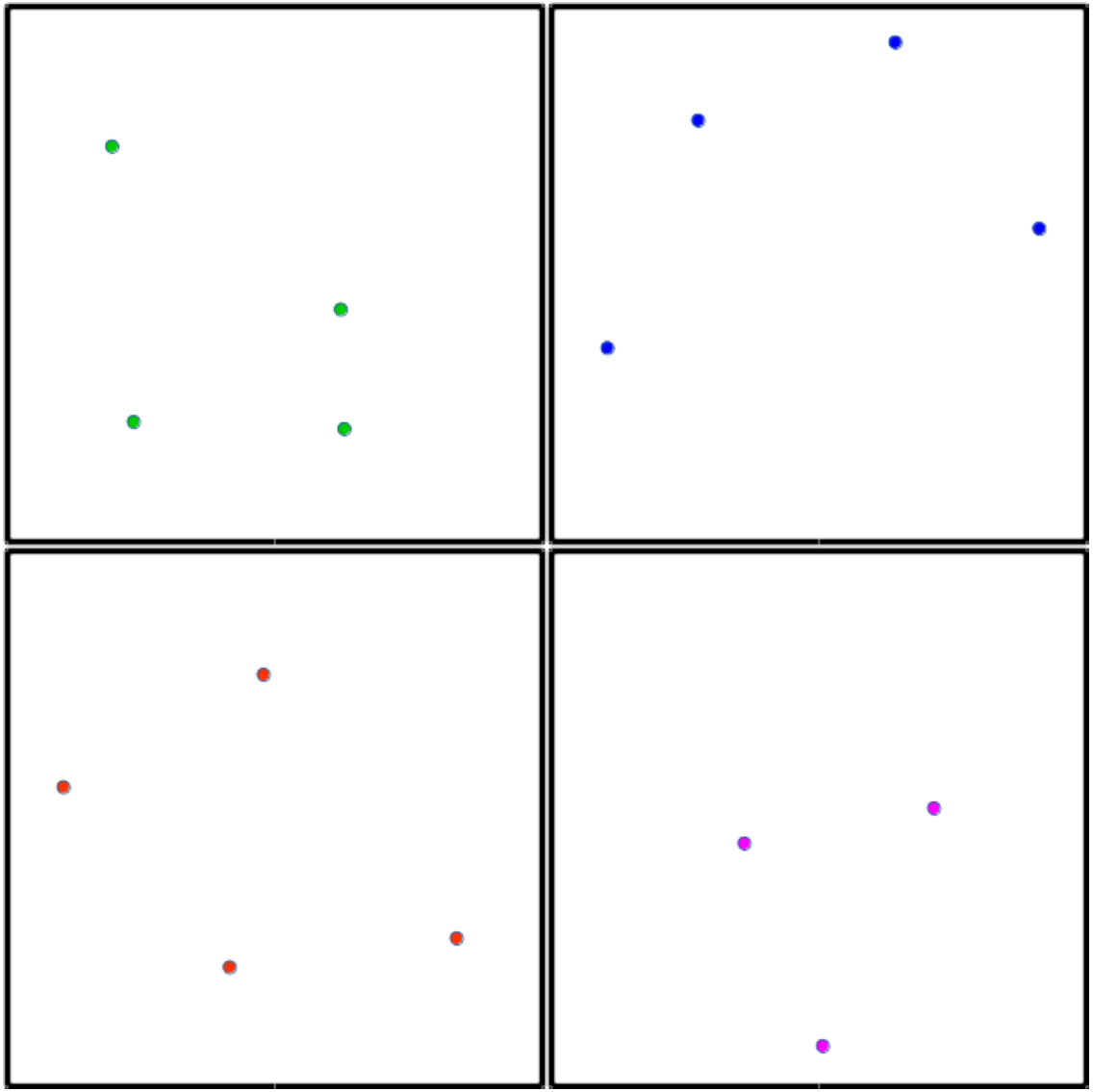
Hadoop - Partition



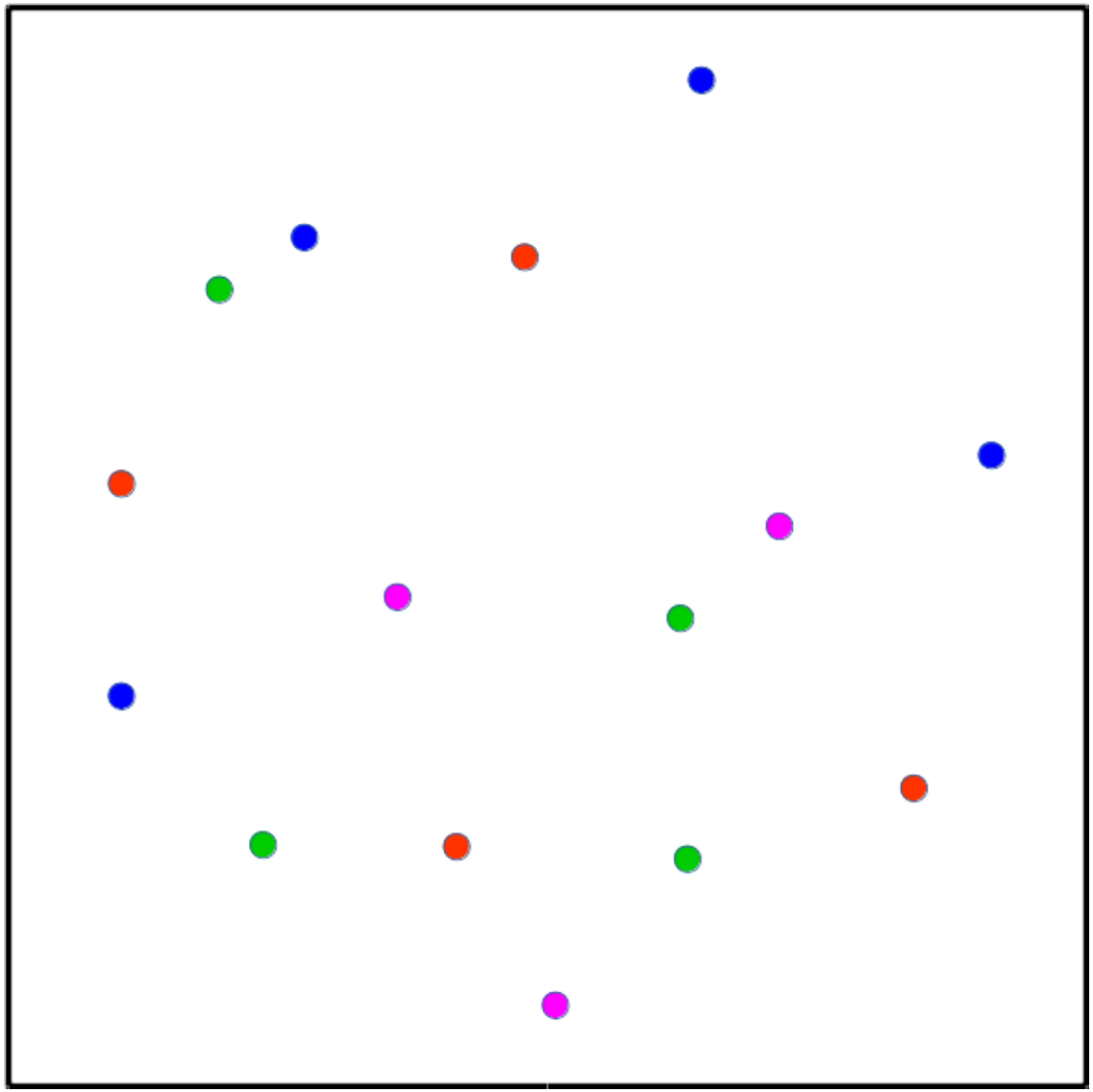
Hadoop - Local



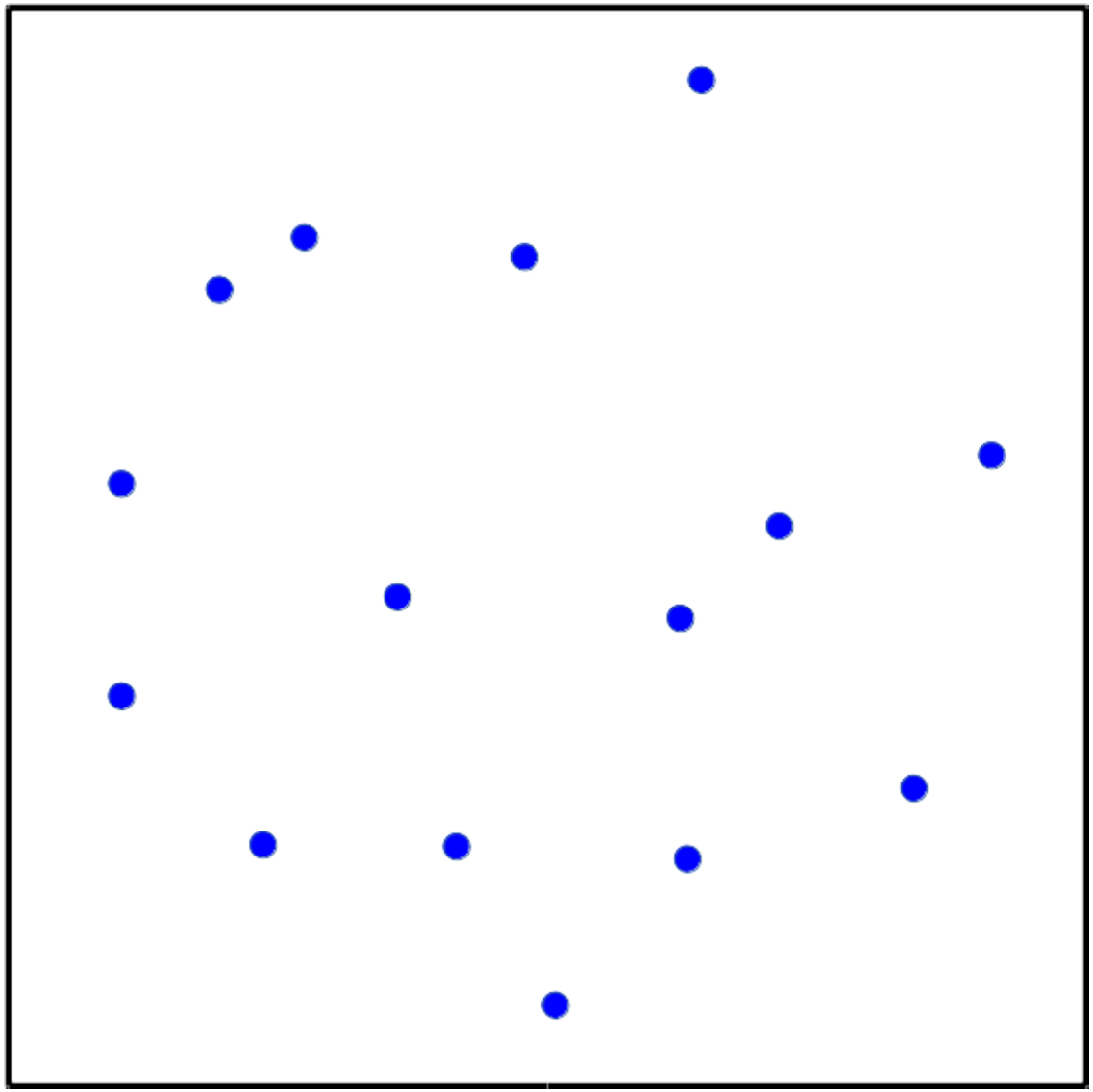
Hadoop - Local



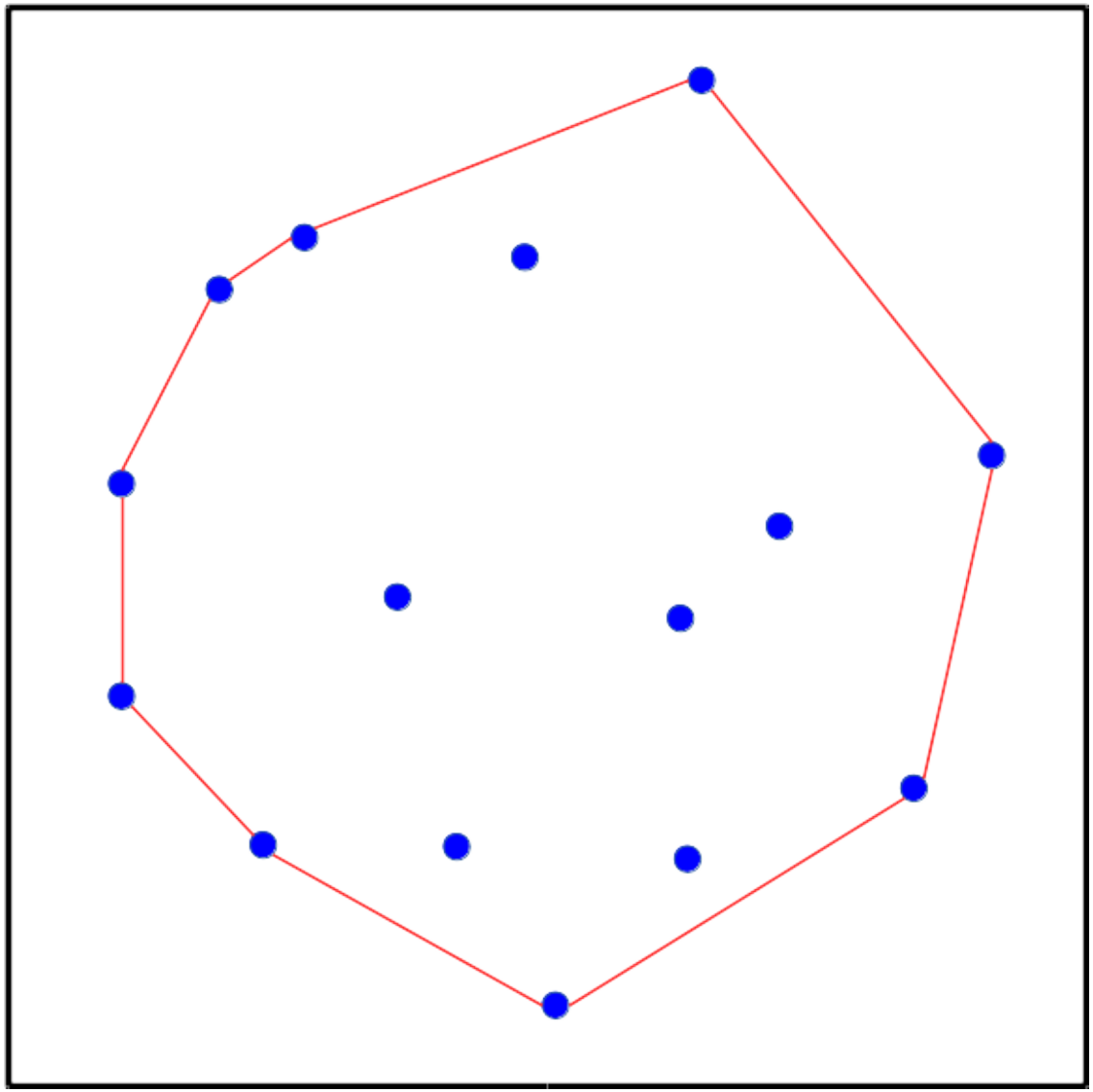
Hadoop - Global



Hadoop - Global



Hadoop - Global



Convex hull in SpatialHadoop

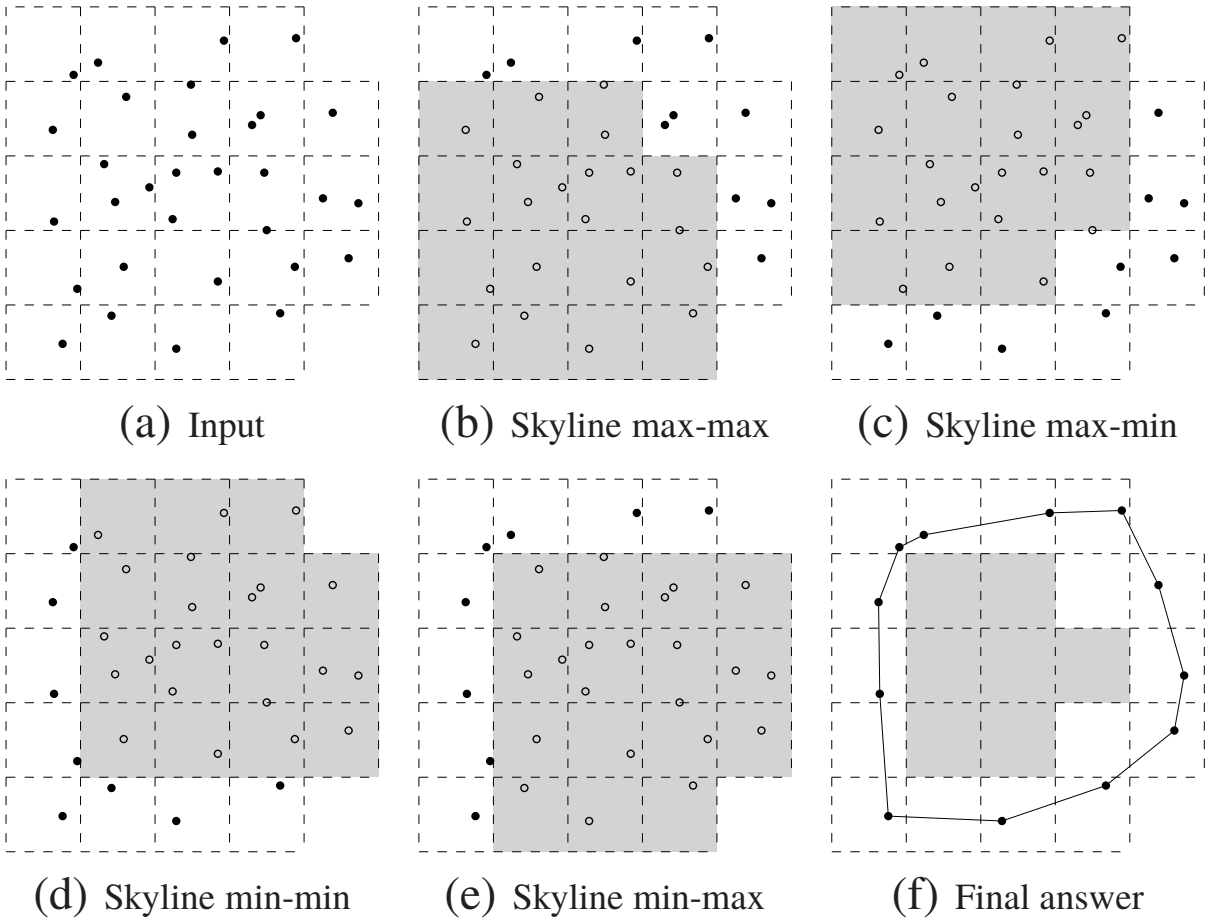
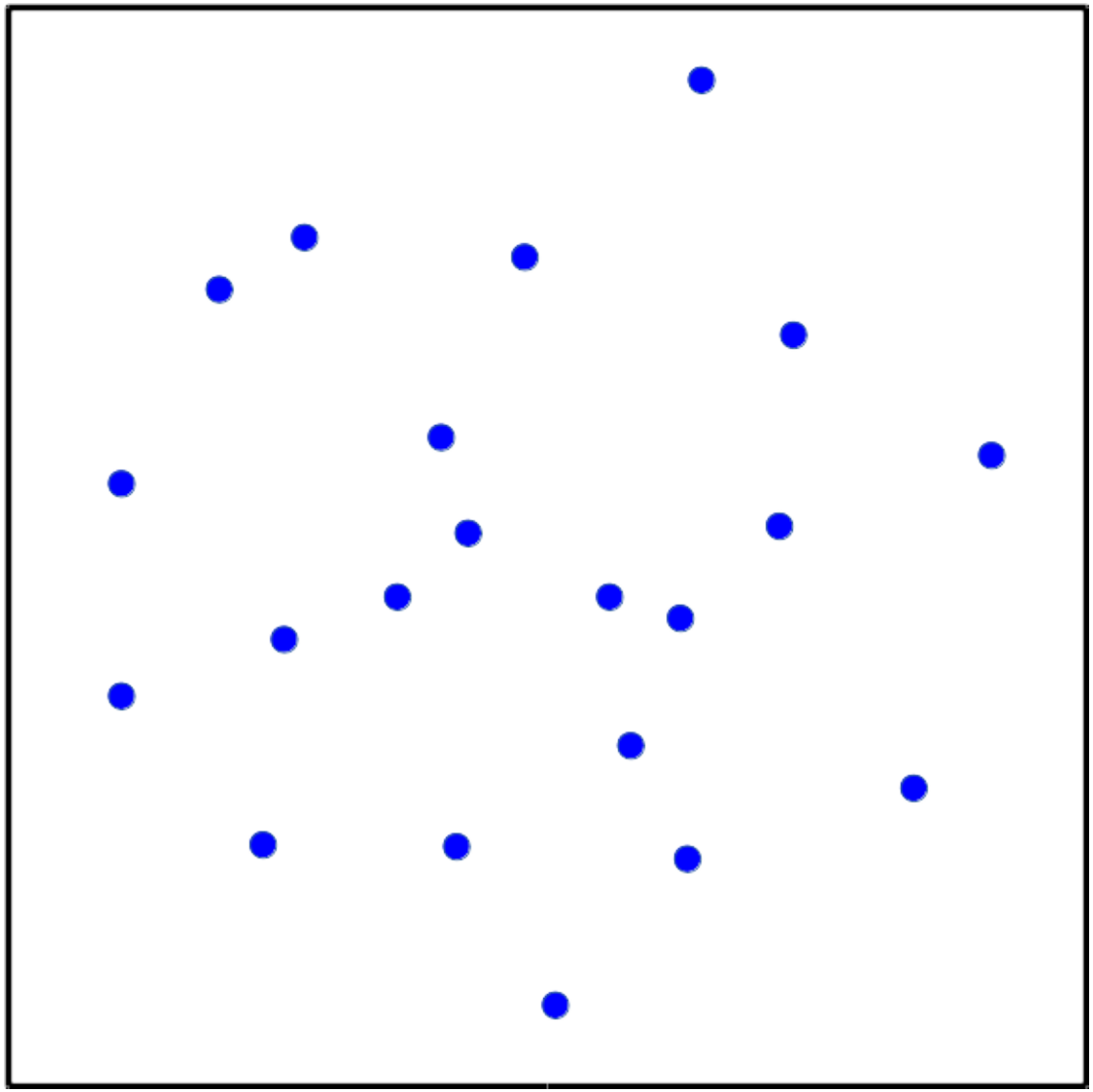
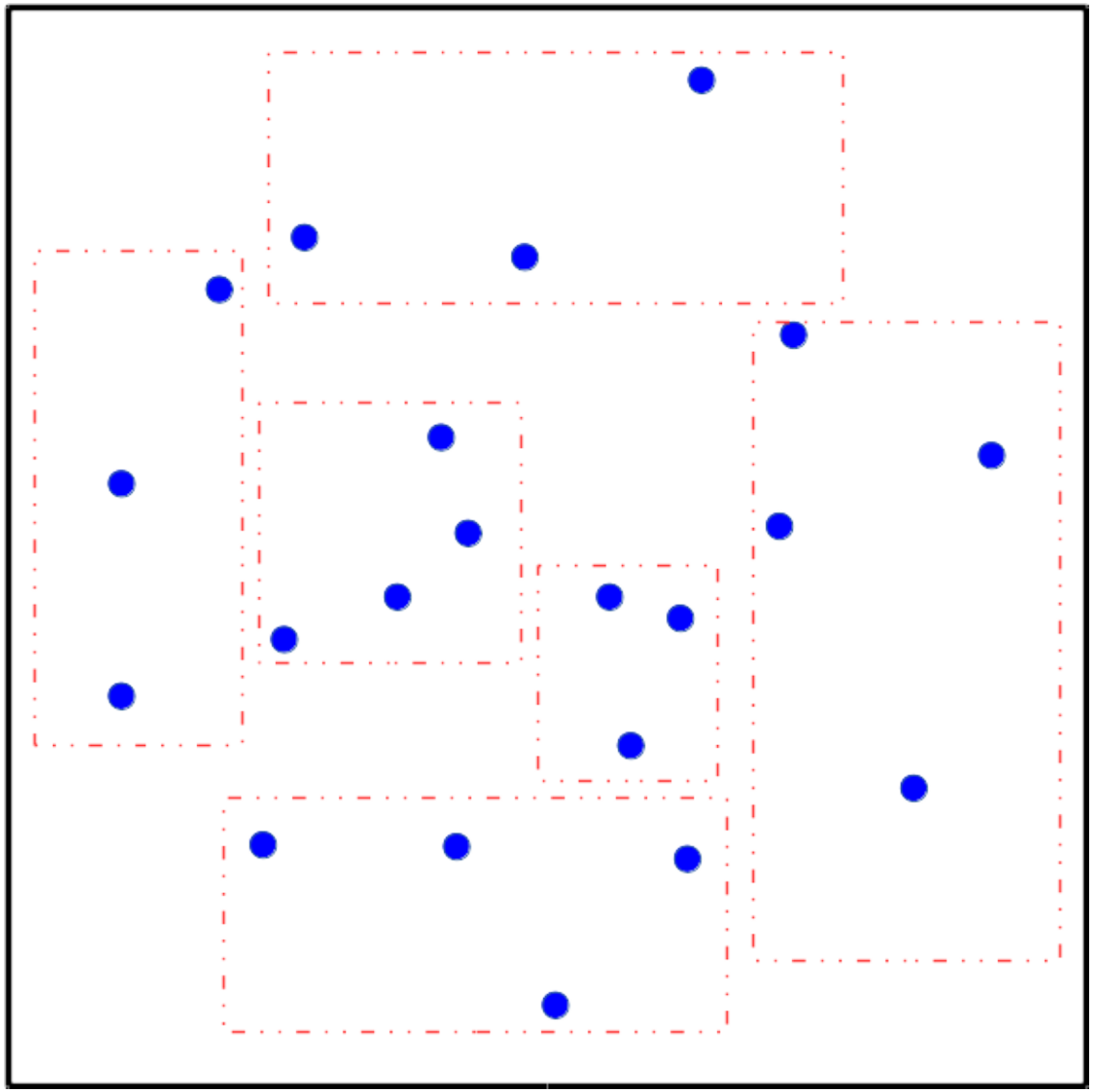


Figure 5: Convex hull in SpatialHadoop

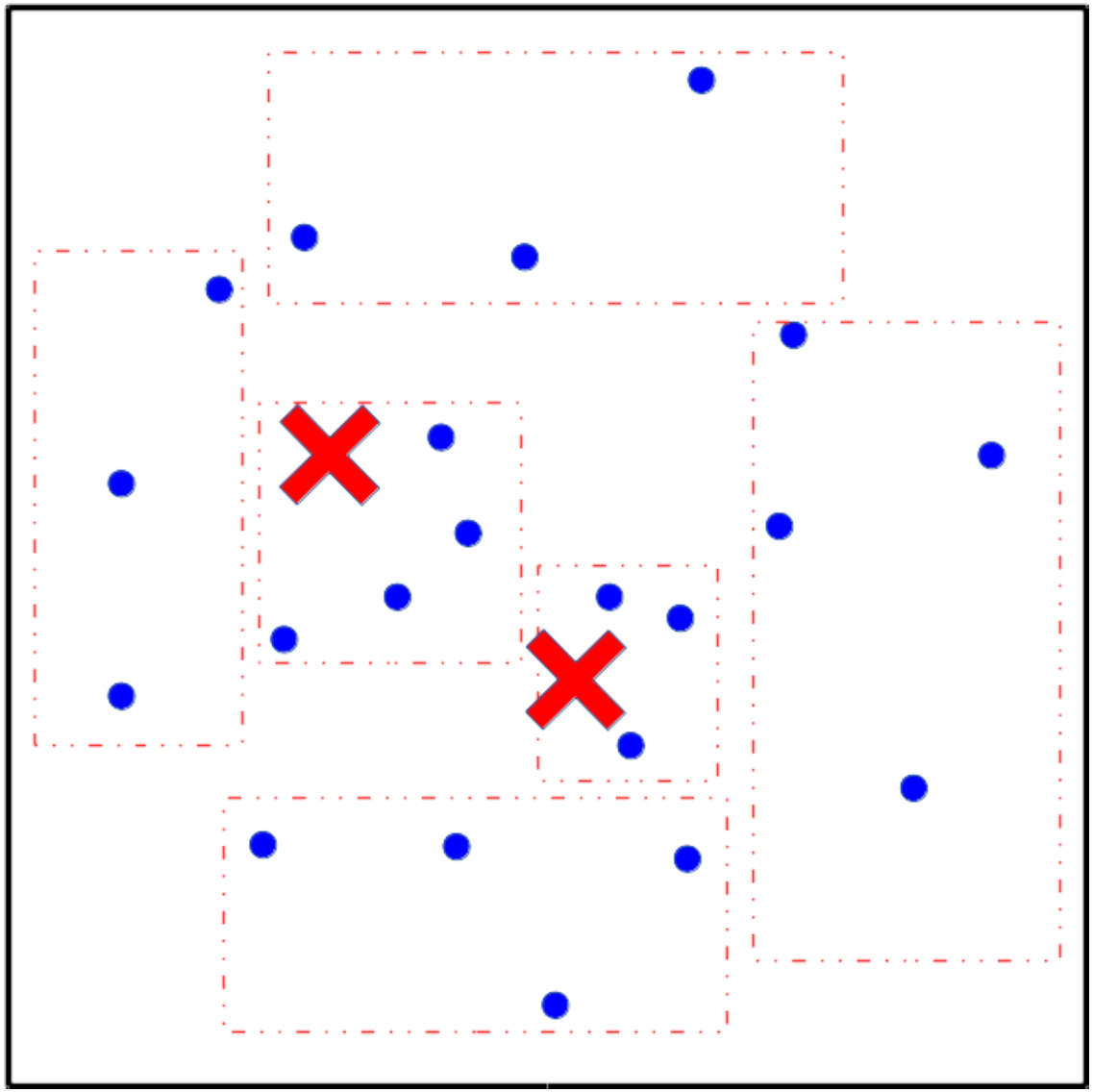
SpatialHadoop - Partition



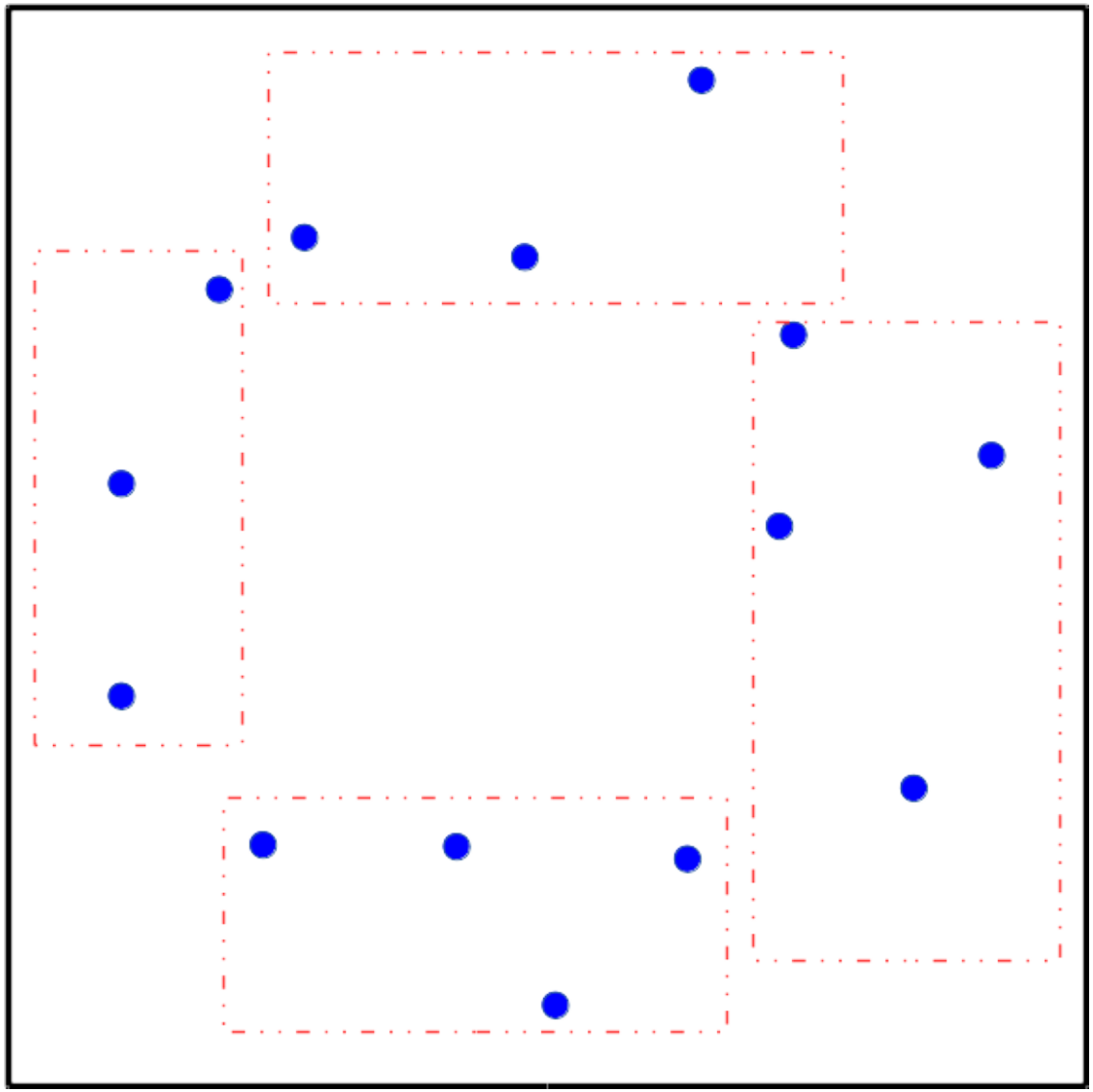
SpatialHadoop - Partition



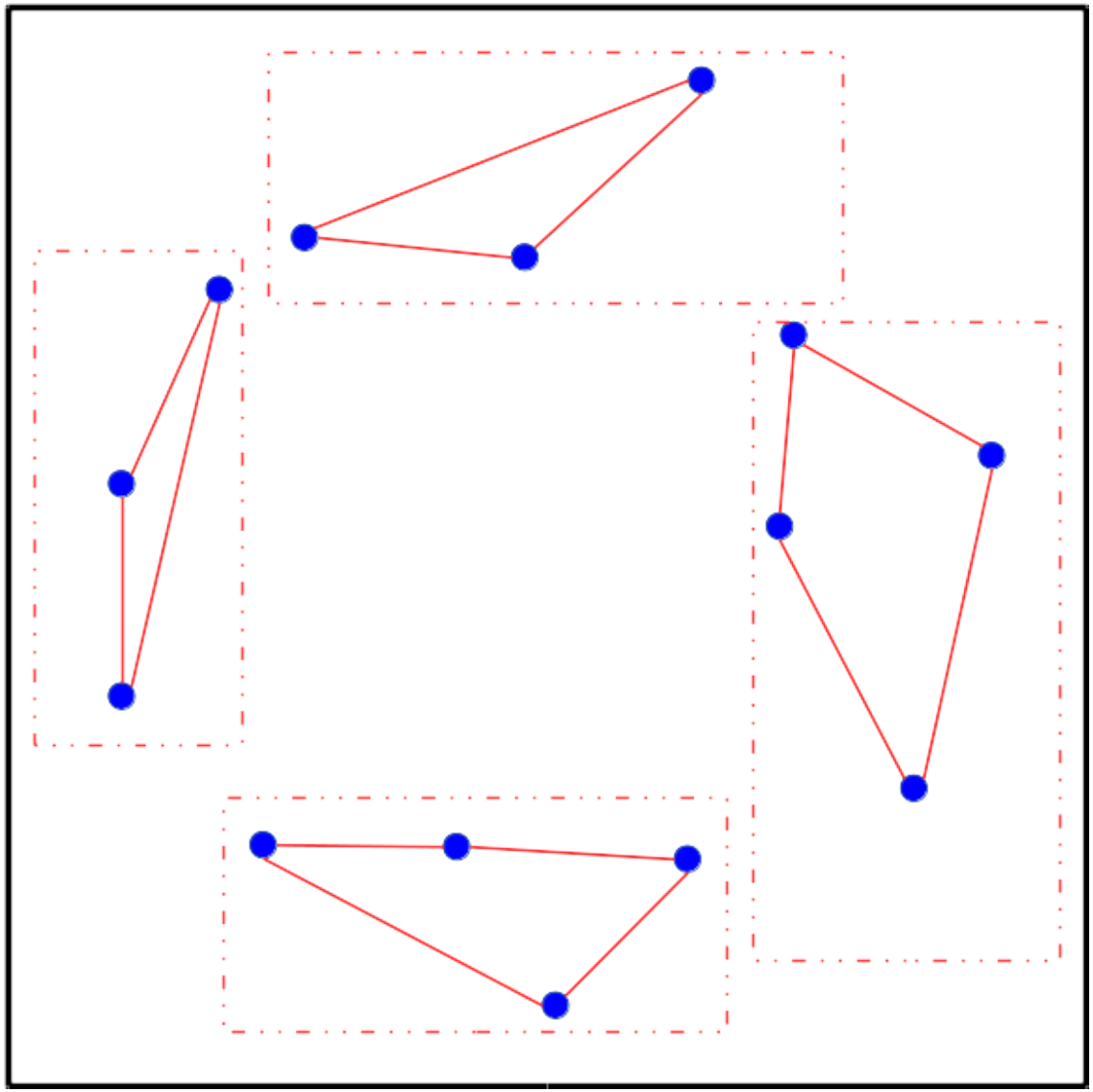
SpatialHadoop - Pruning



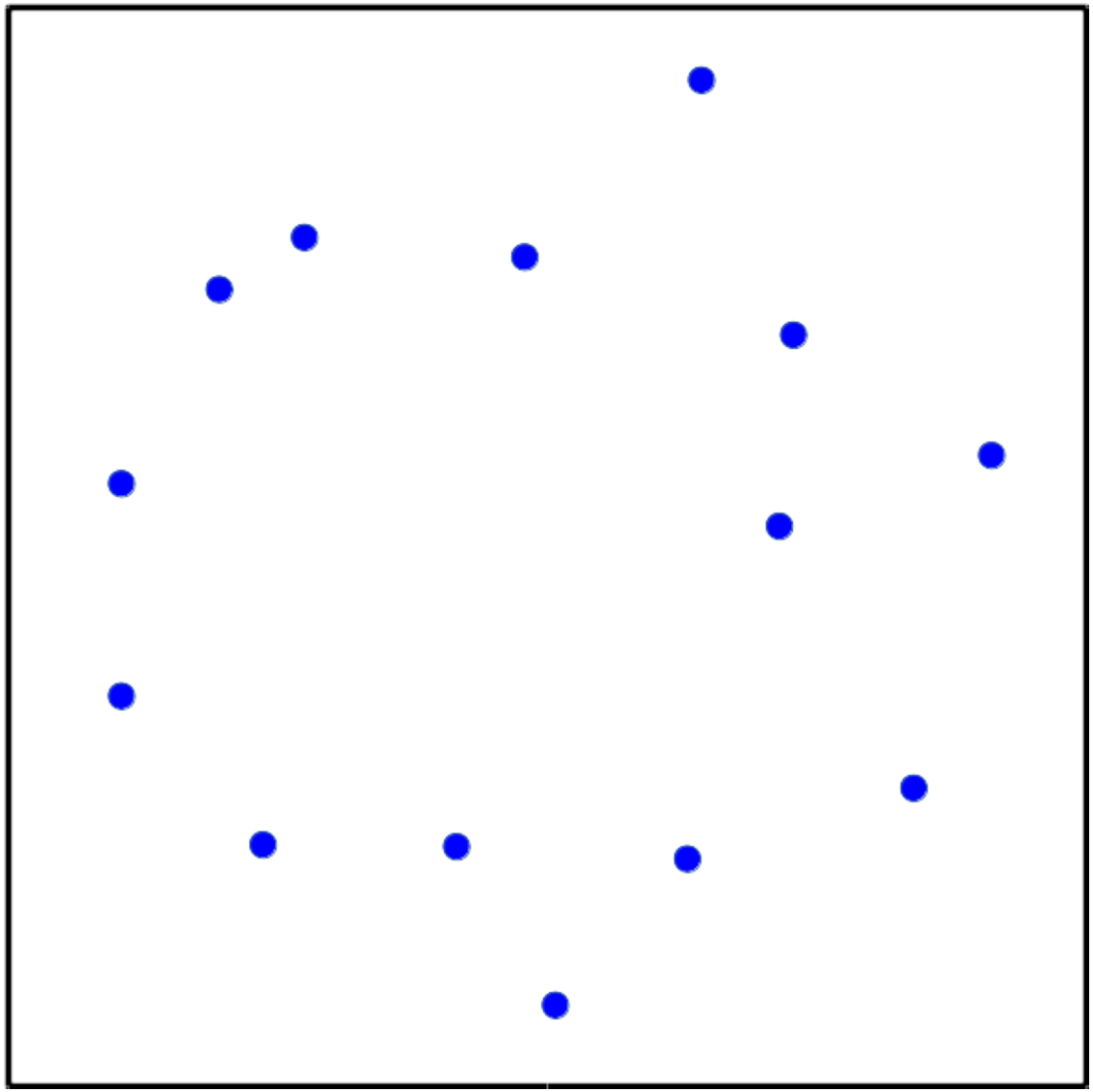
SpatialHadoop - Pruning



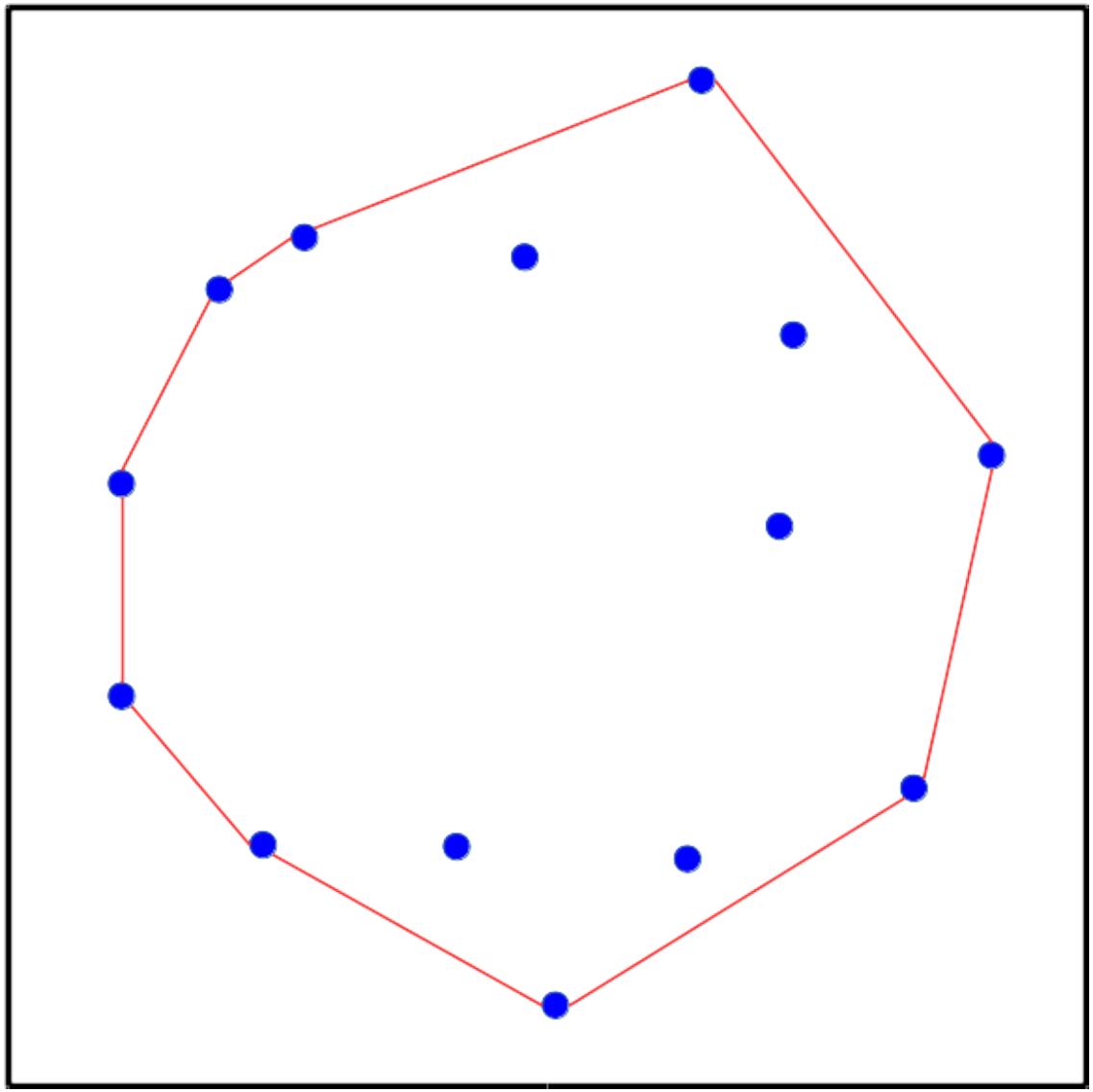
SpatialHadoop - Local



SpatialHadoop - Global

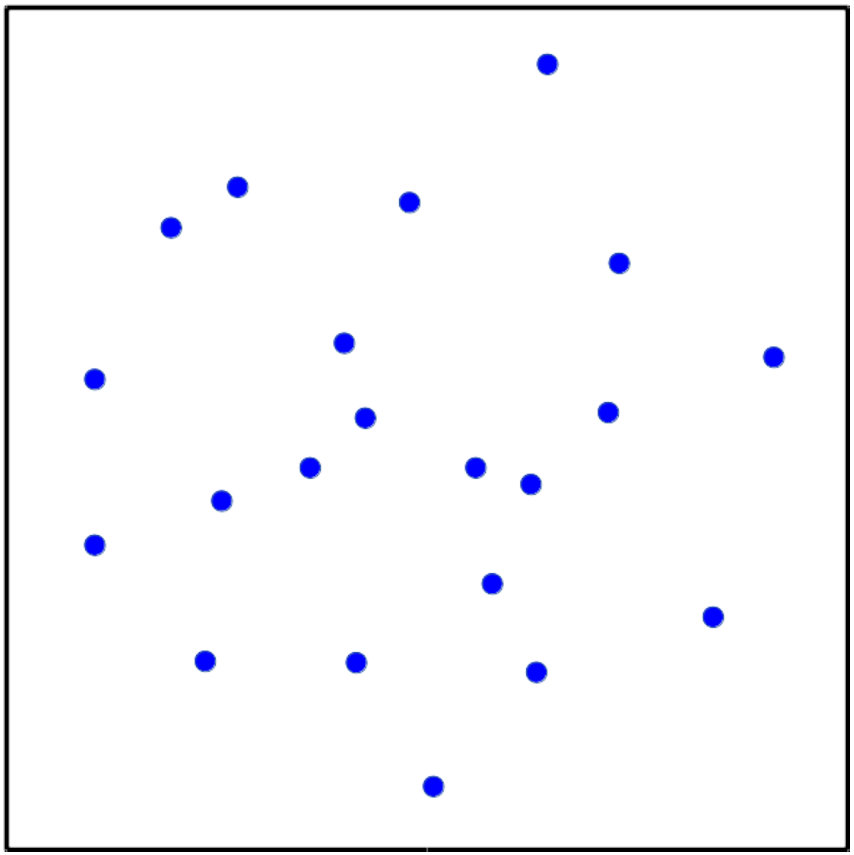


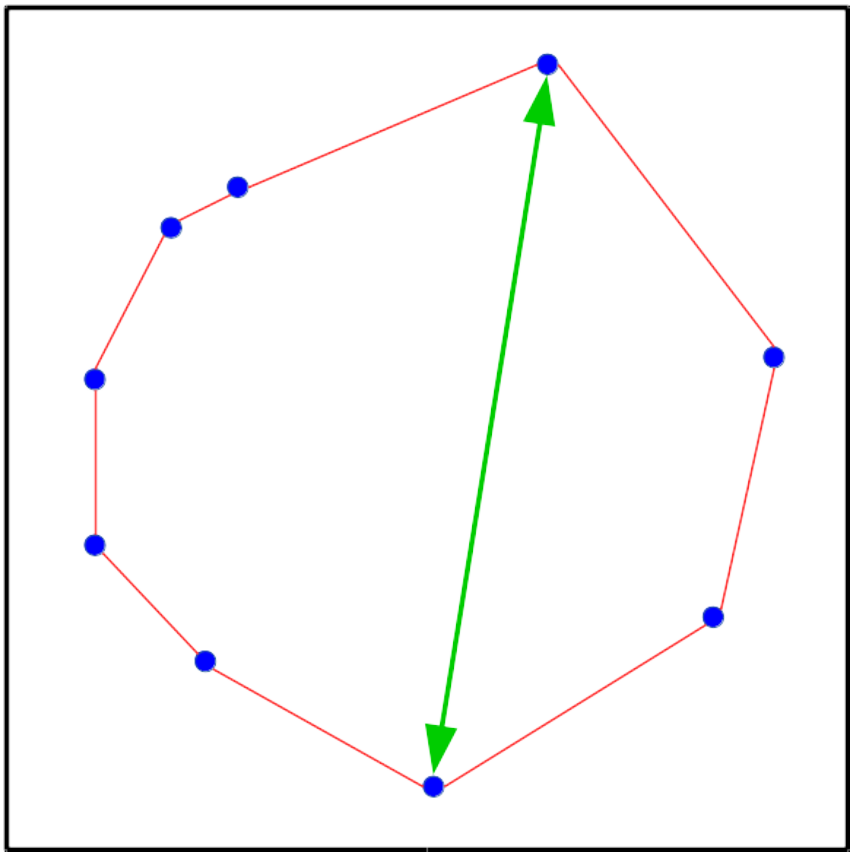
SpatialHadoop - Global



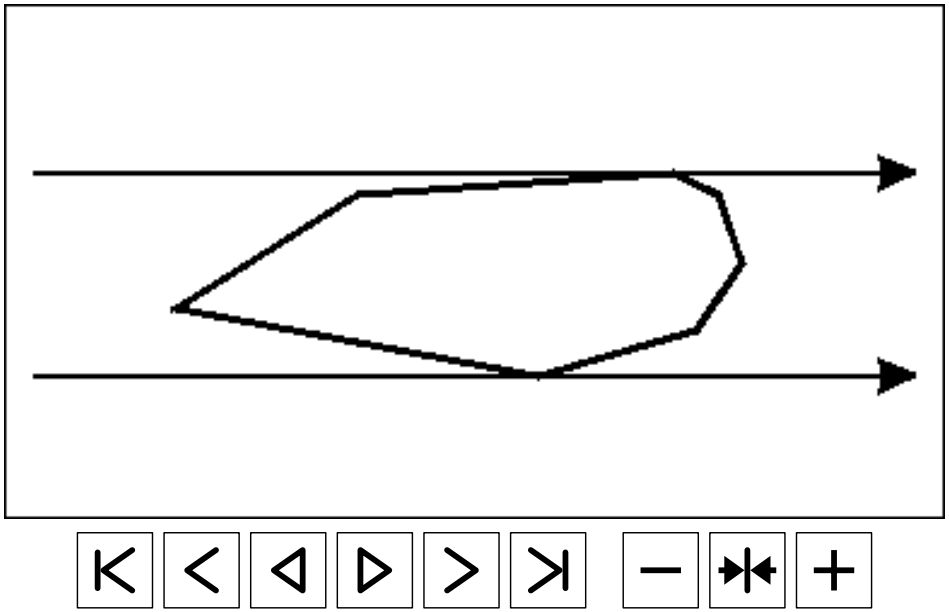
Agenda

Farthest Pair

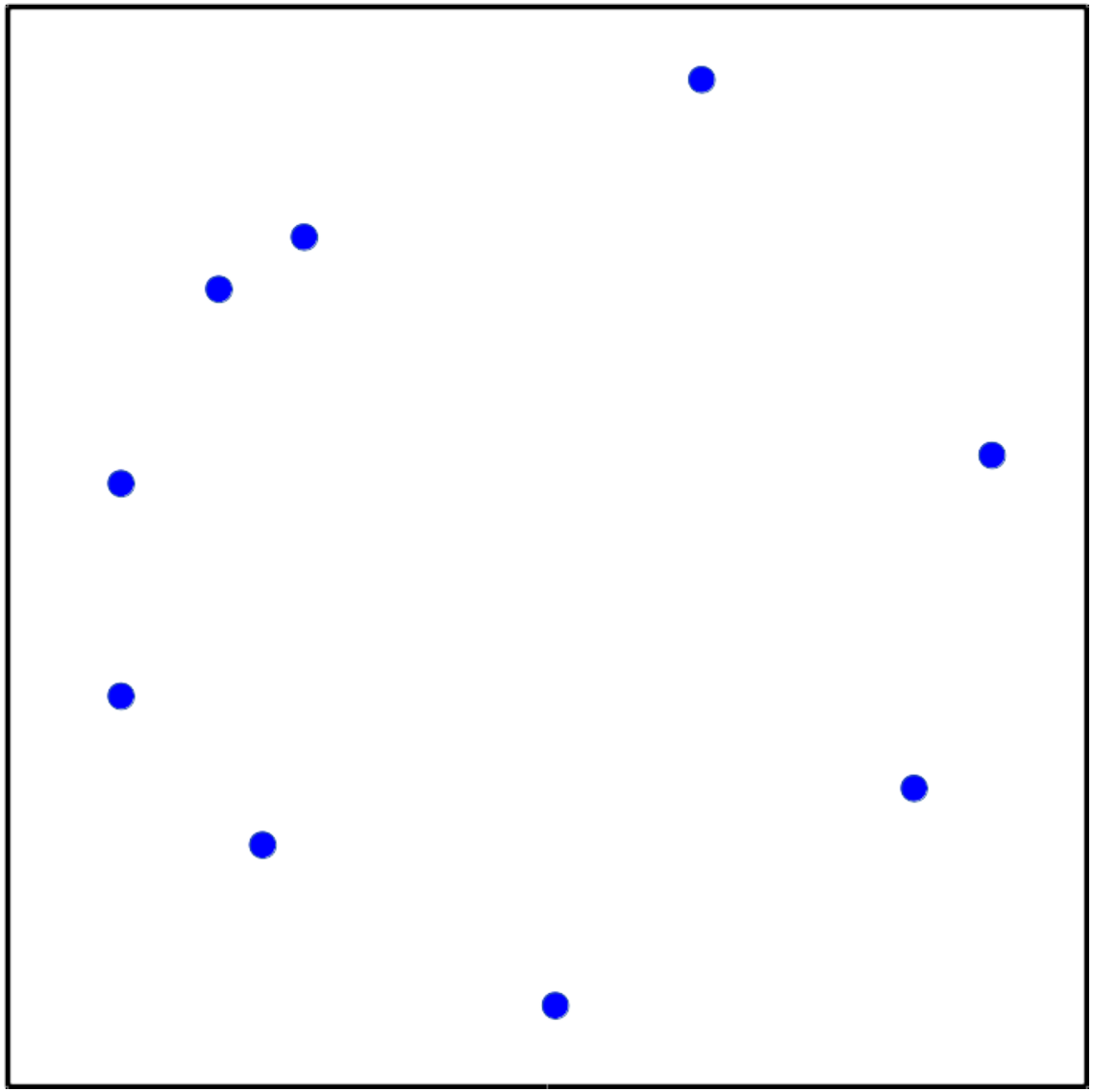




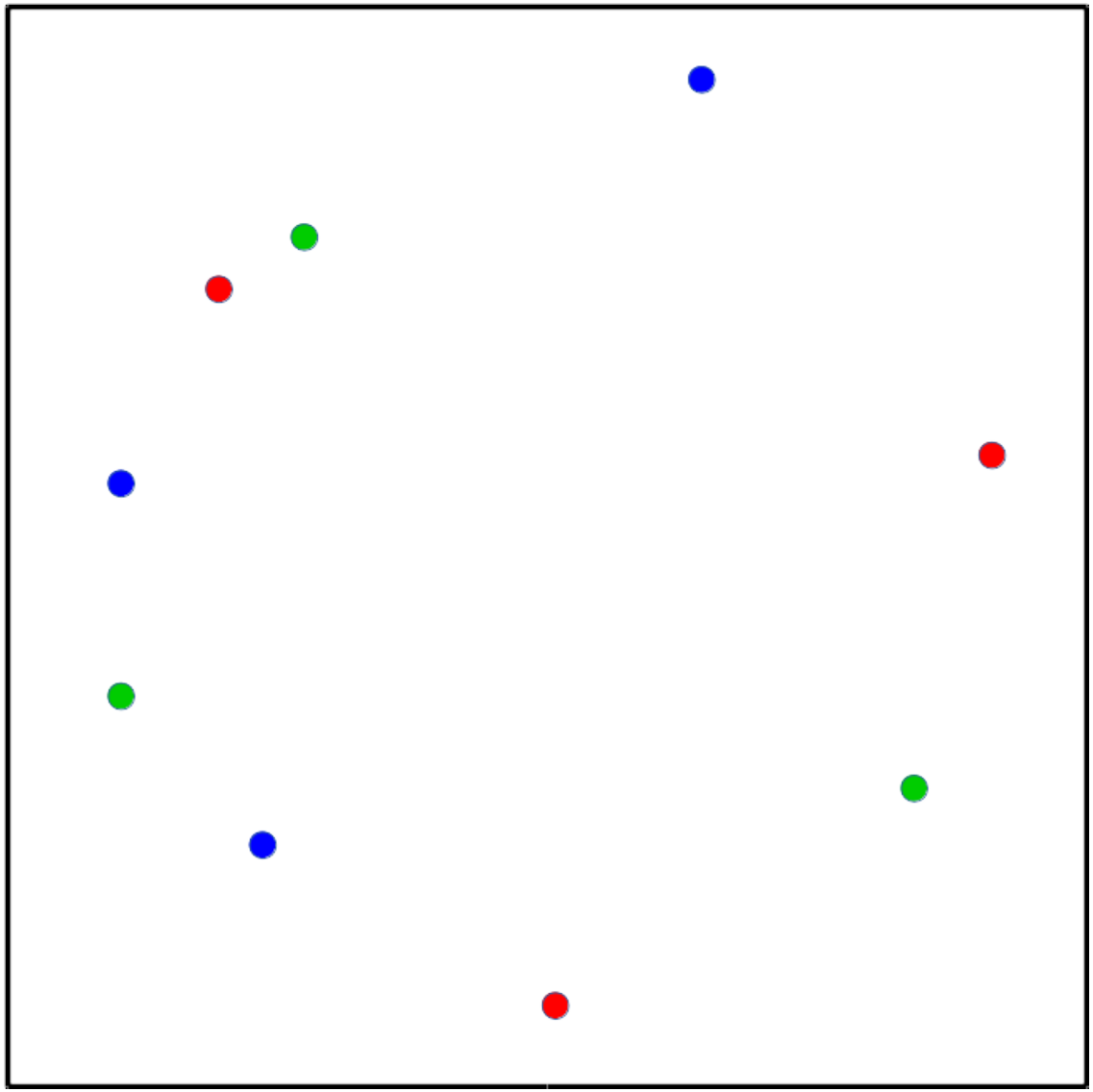
Rotating Calipers method



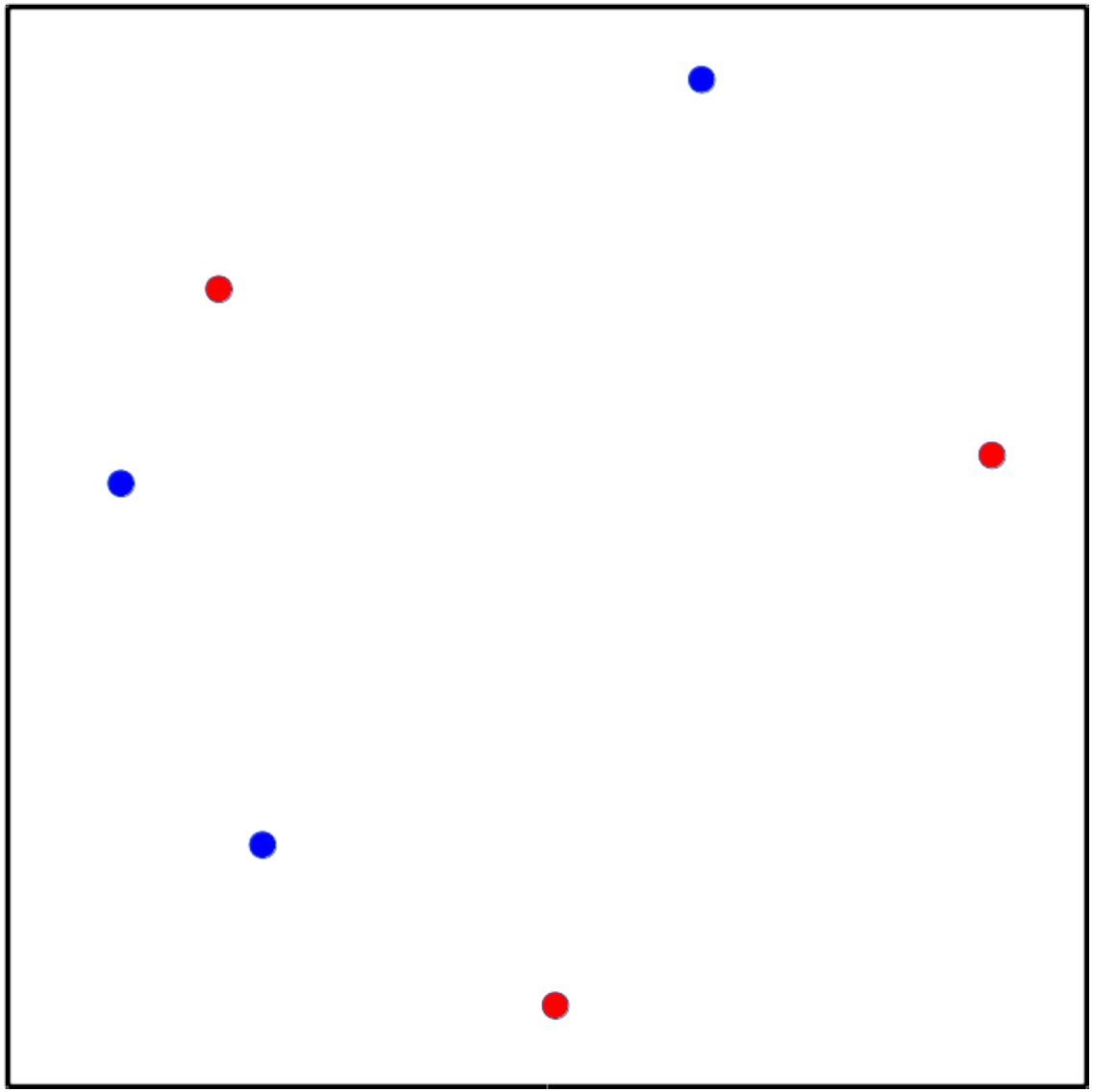
Hadoop - Partition



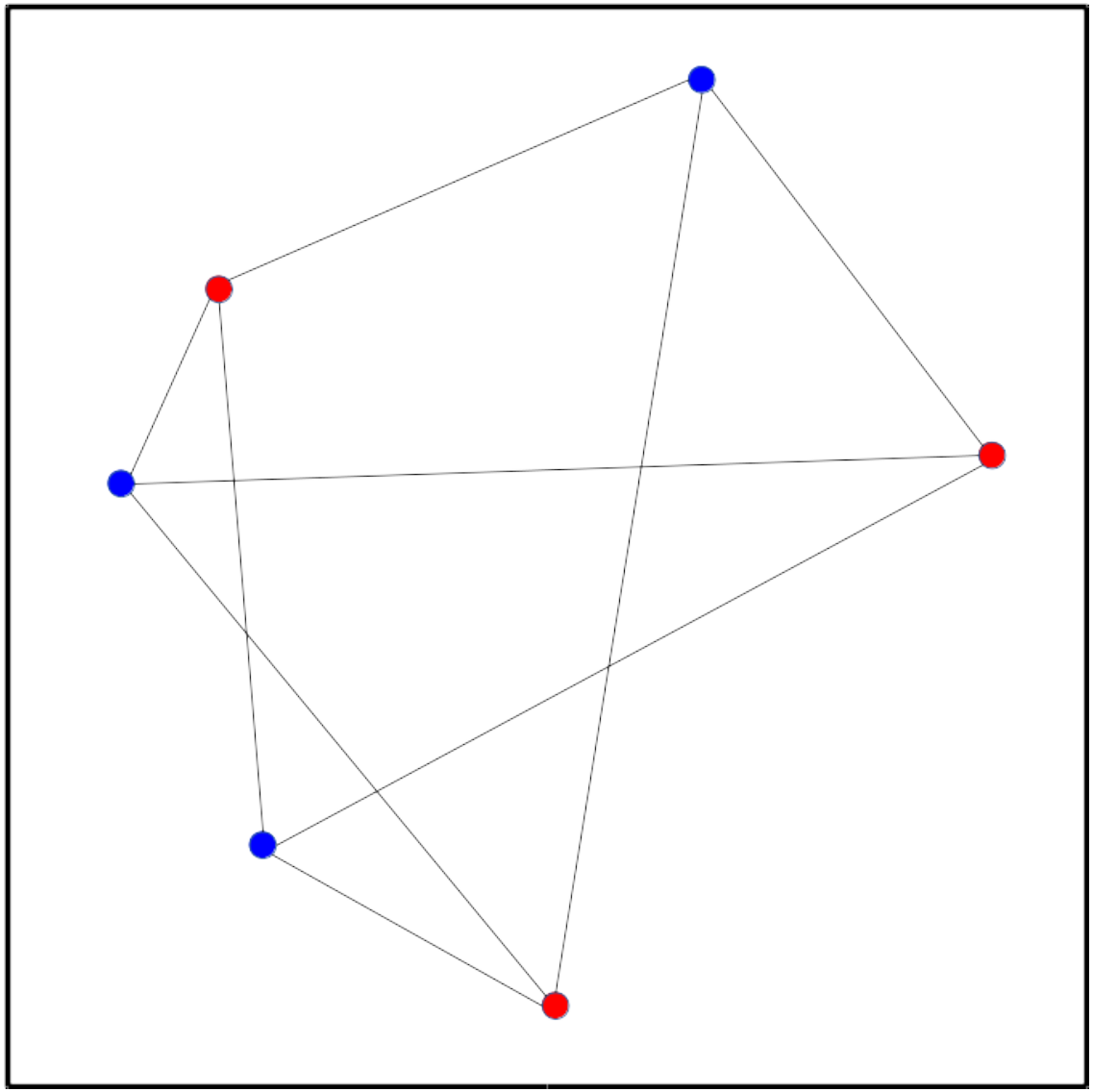
Hadoop - Partition



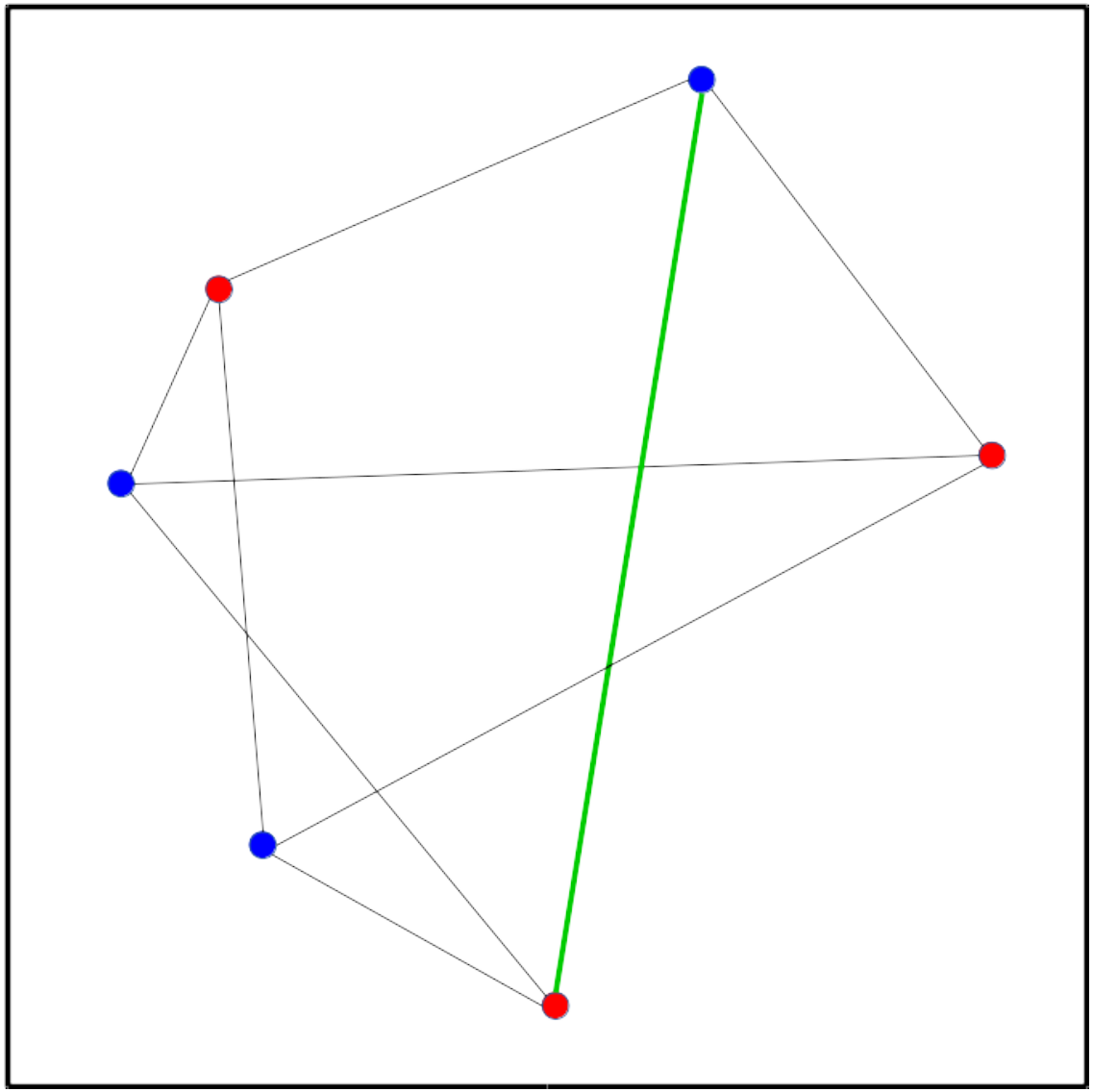
Hadoop - Local



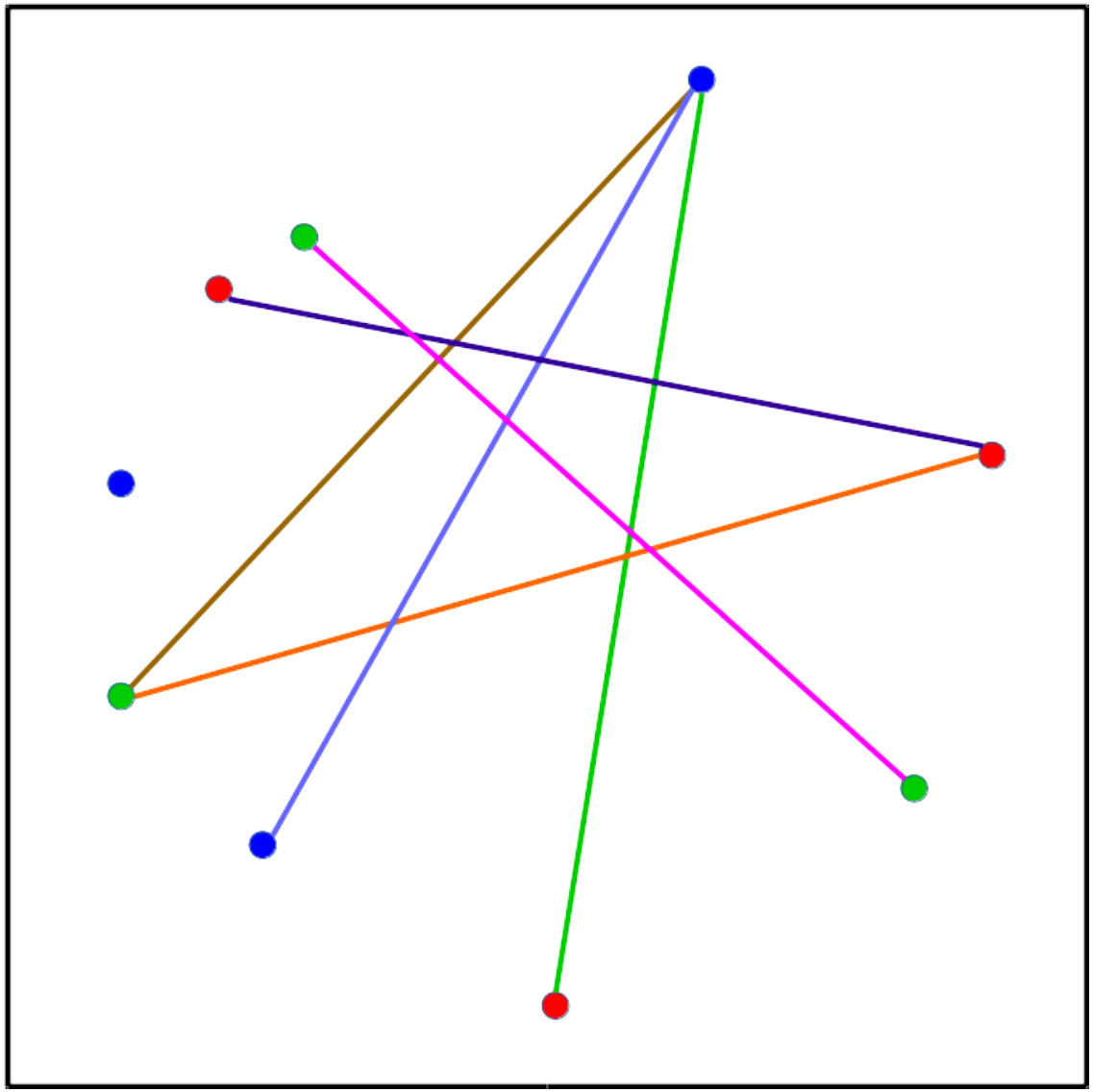
Hadoop - Local



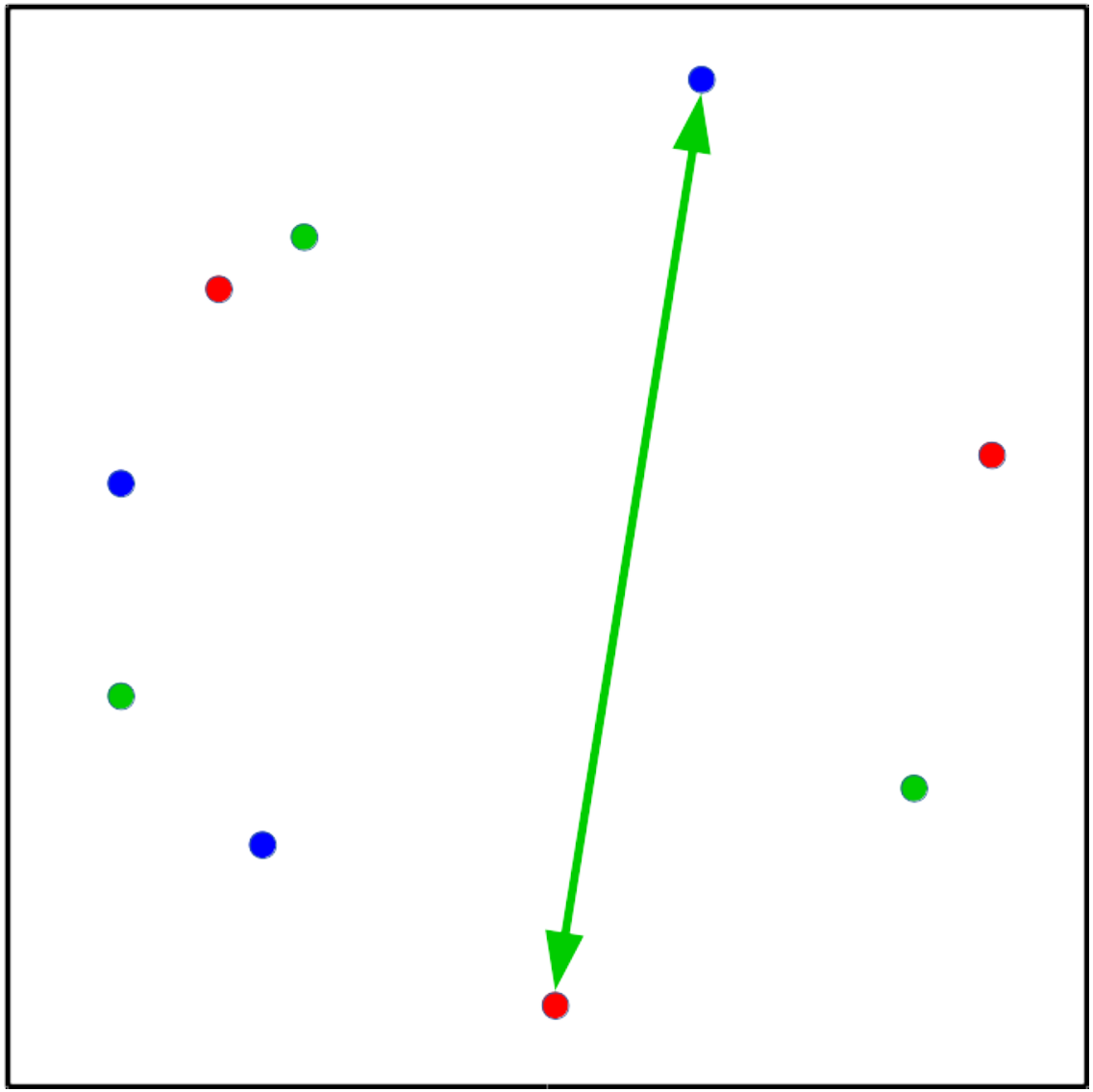
Hadoop - Local



Hadoop - Global



Hadoop - Global



Farthest pair in SpatialHadoop

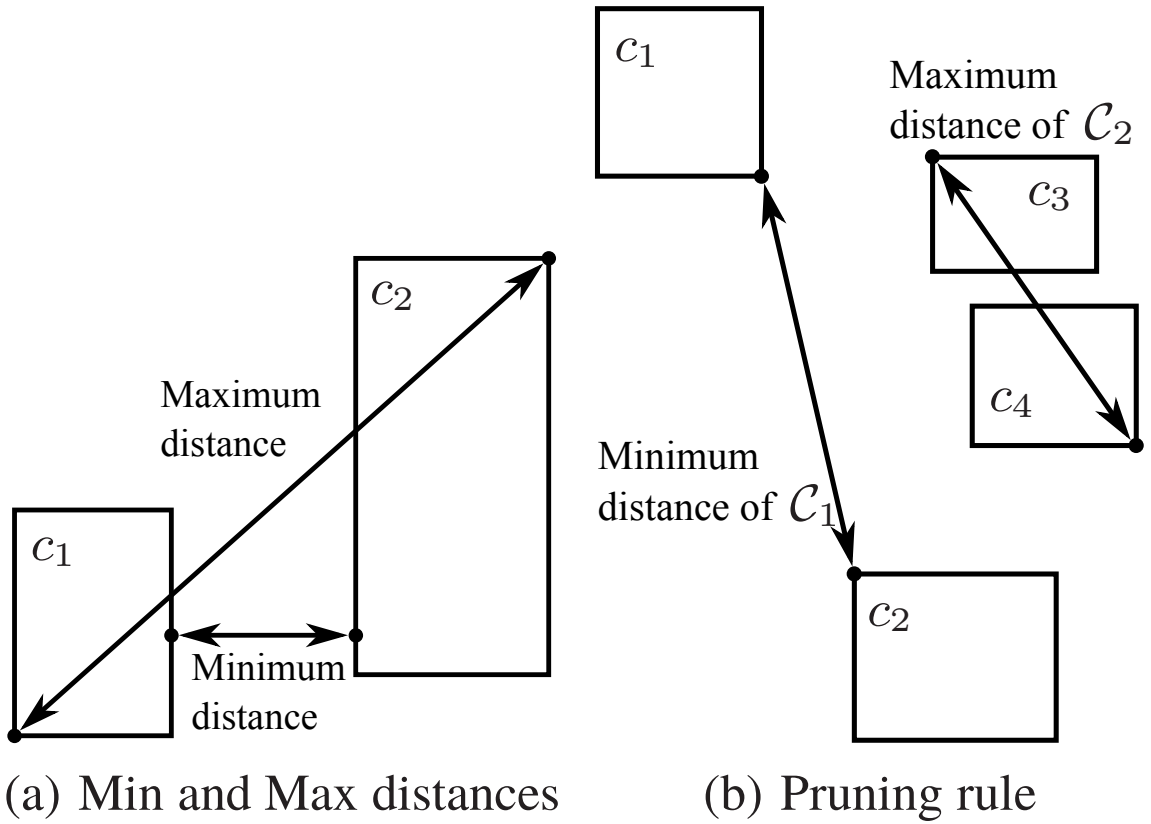
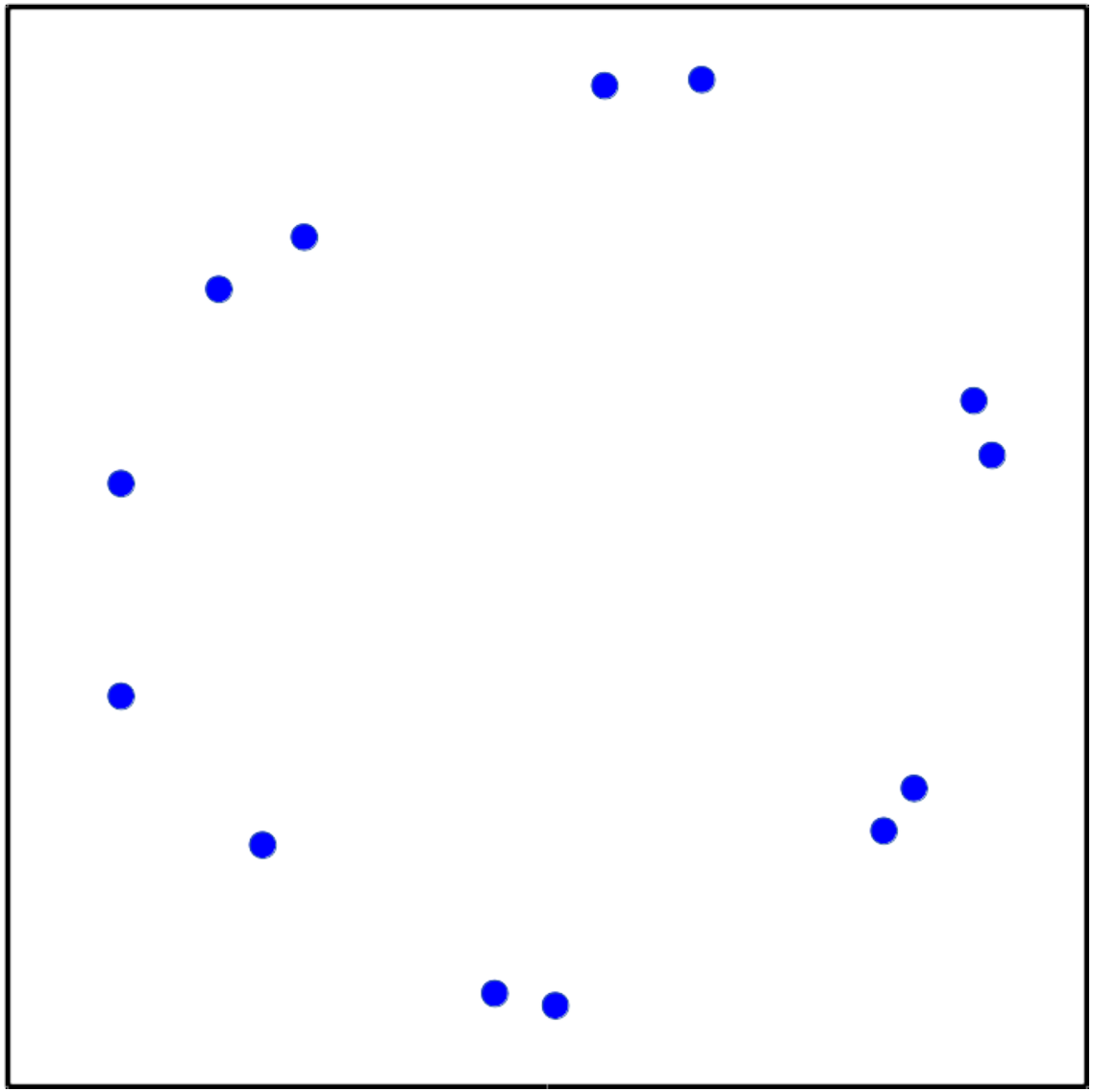
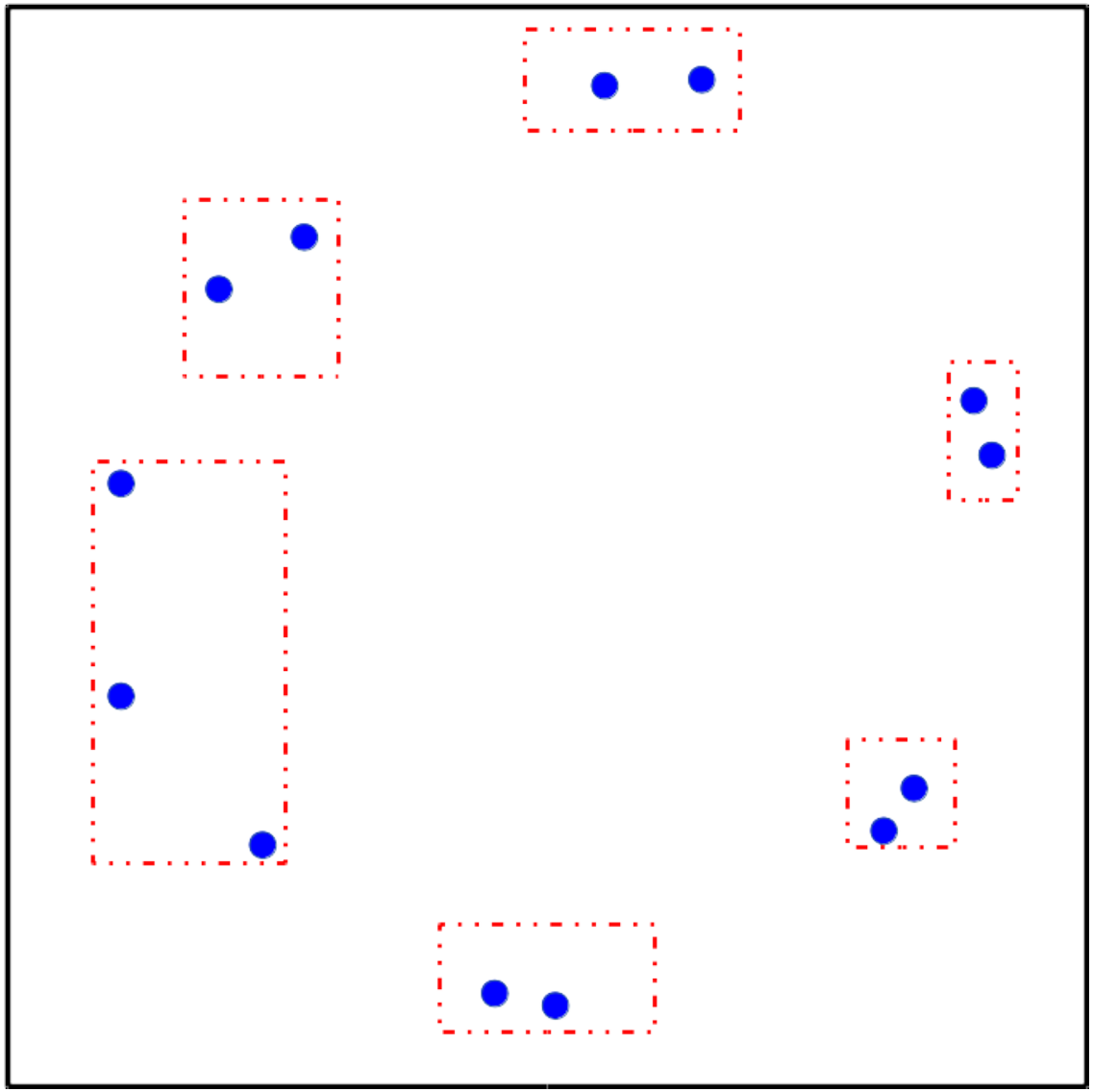


Figure 6: Farthest pair algorithm in SpatialHadoop

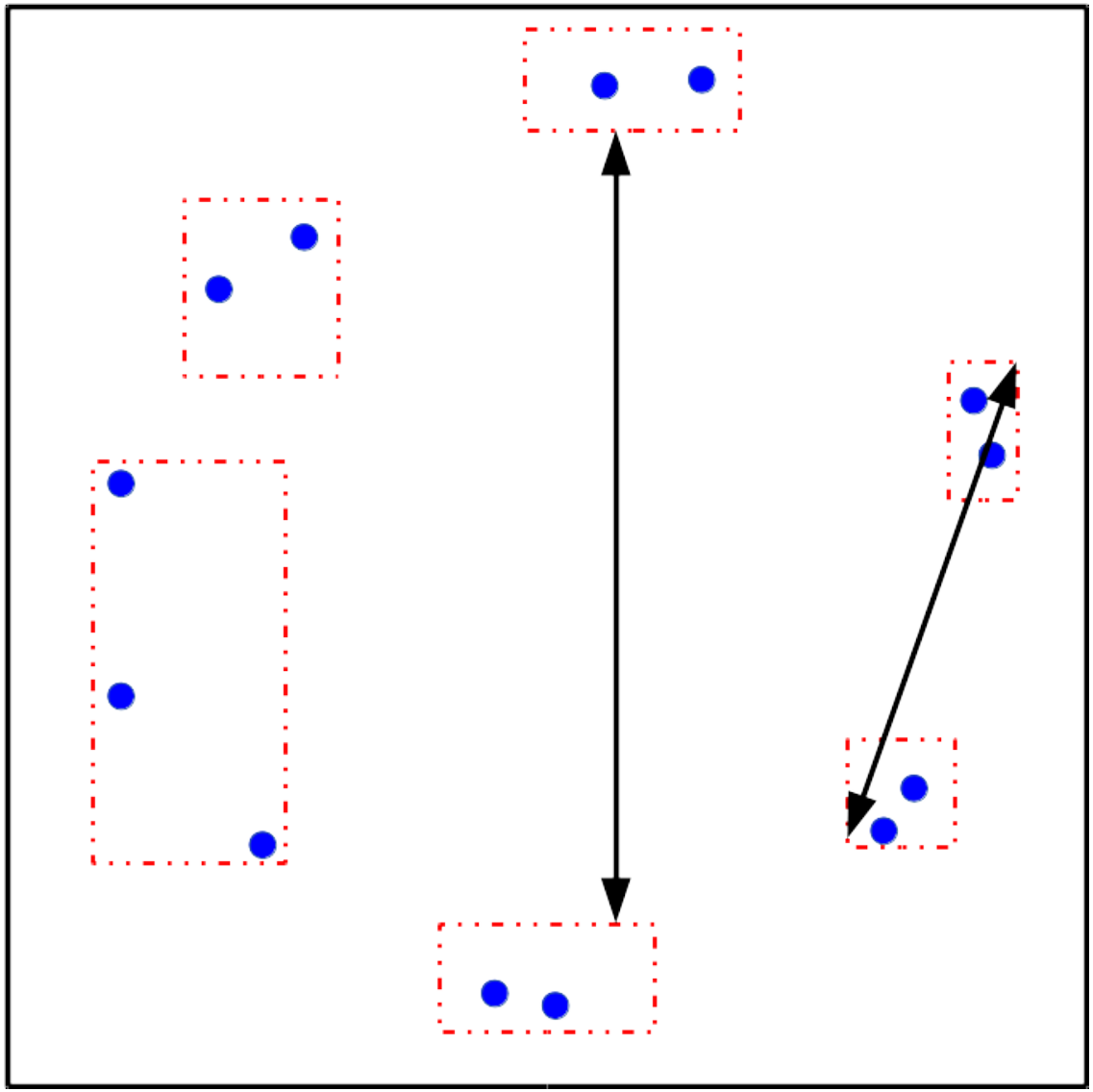
SpatialHadoop - Partition



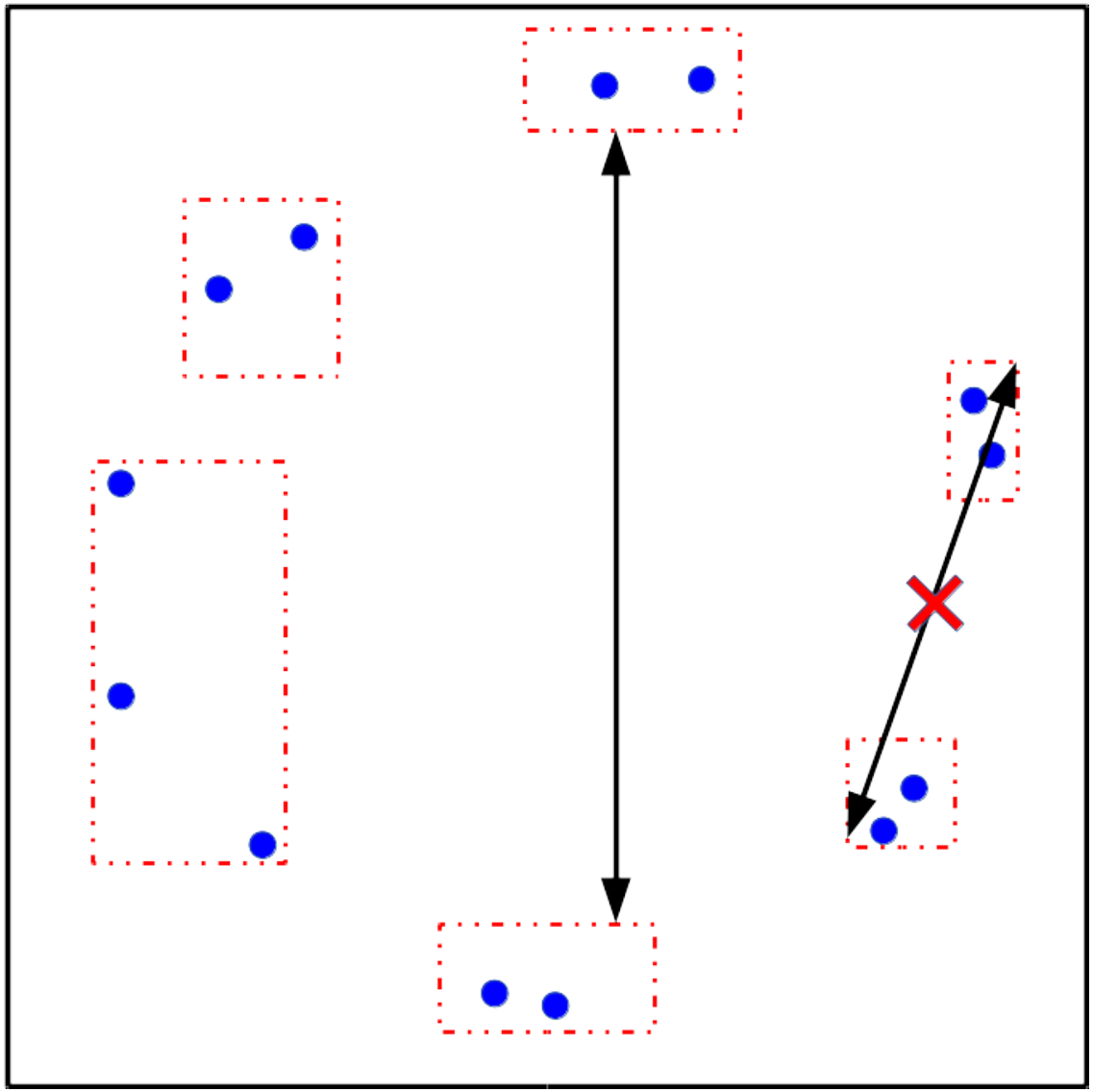
SpatialHadoop - Partition



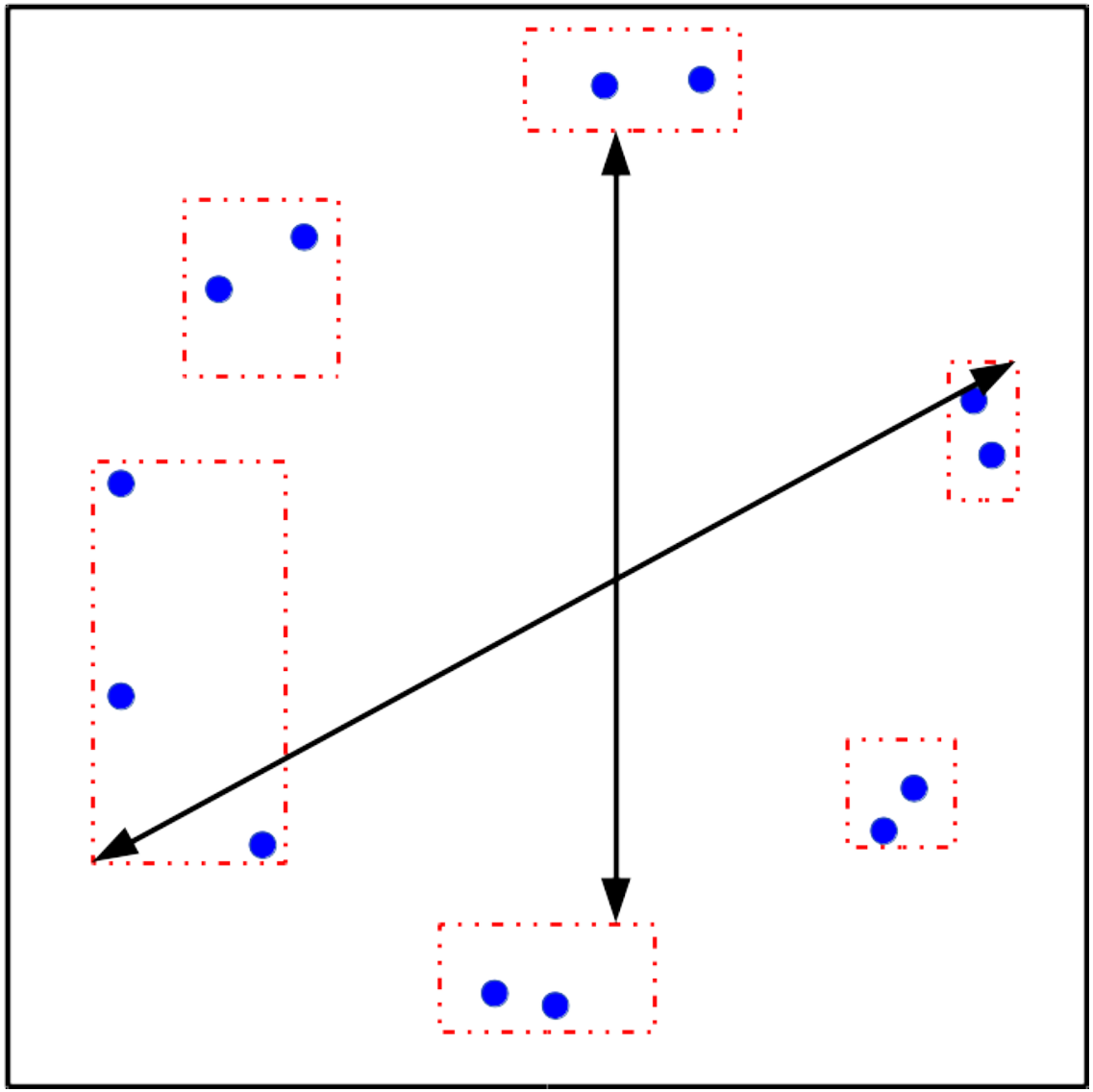
SpatialHadoop - Pruning



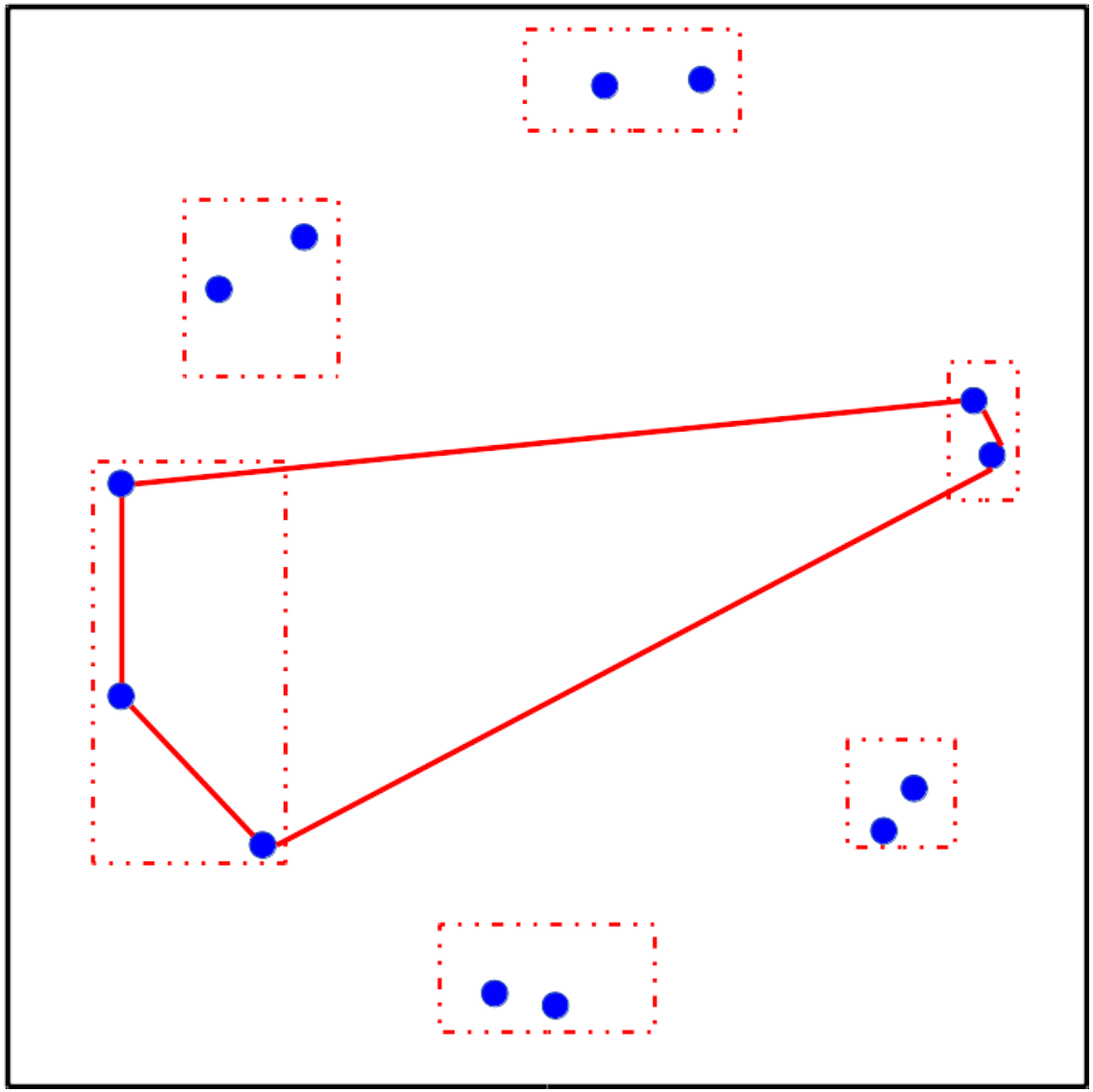
SpatialHadoop - Pruning



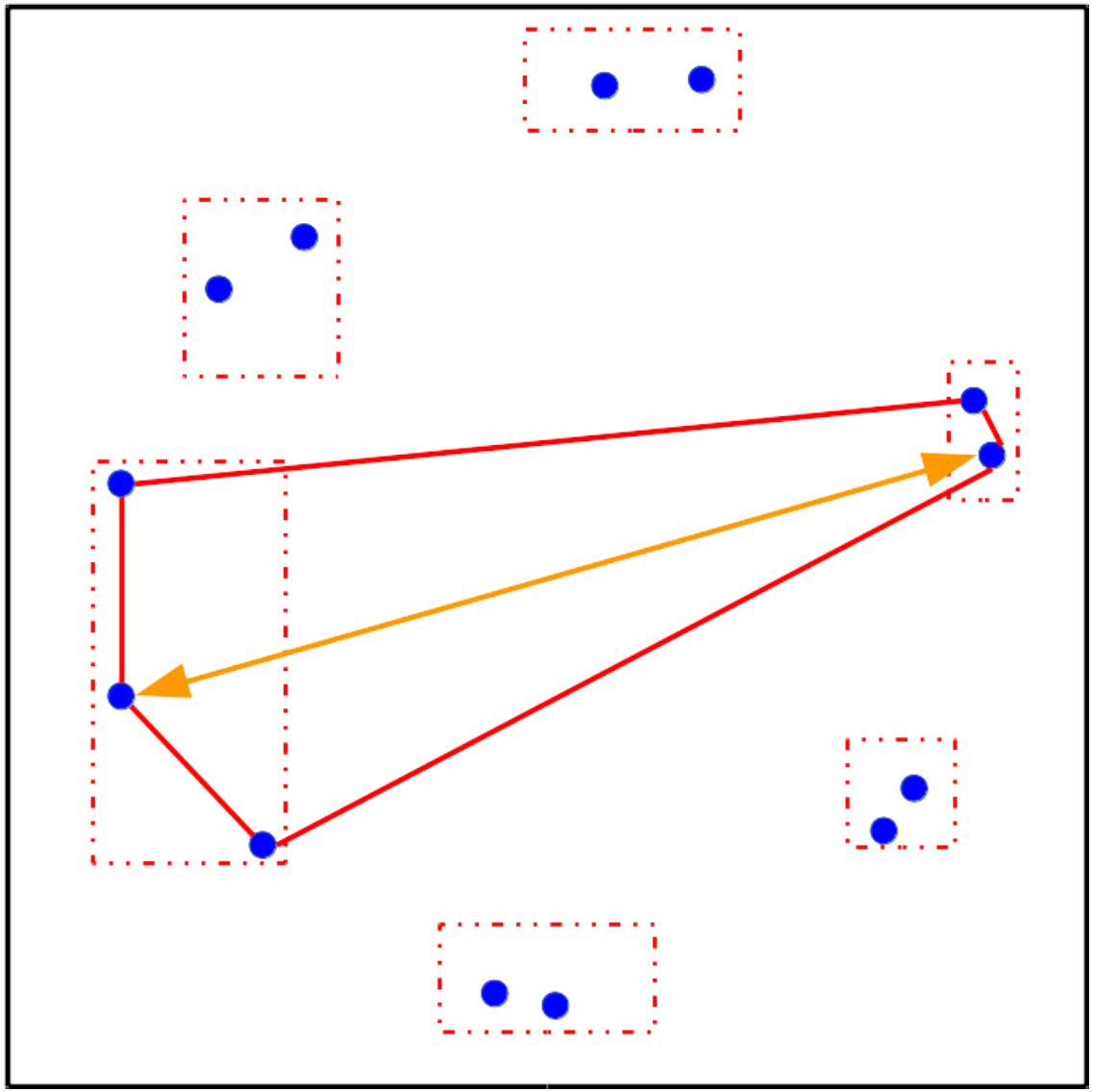
SpatialHadoop - Pruning



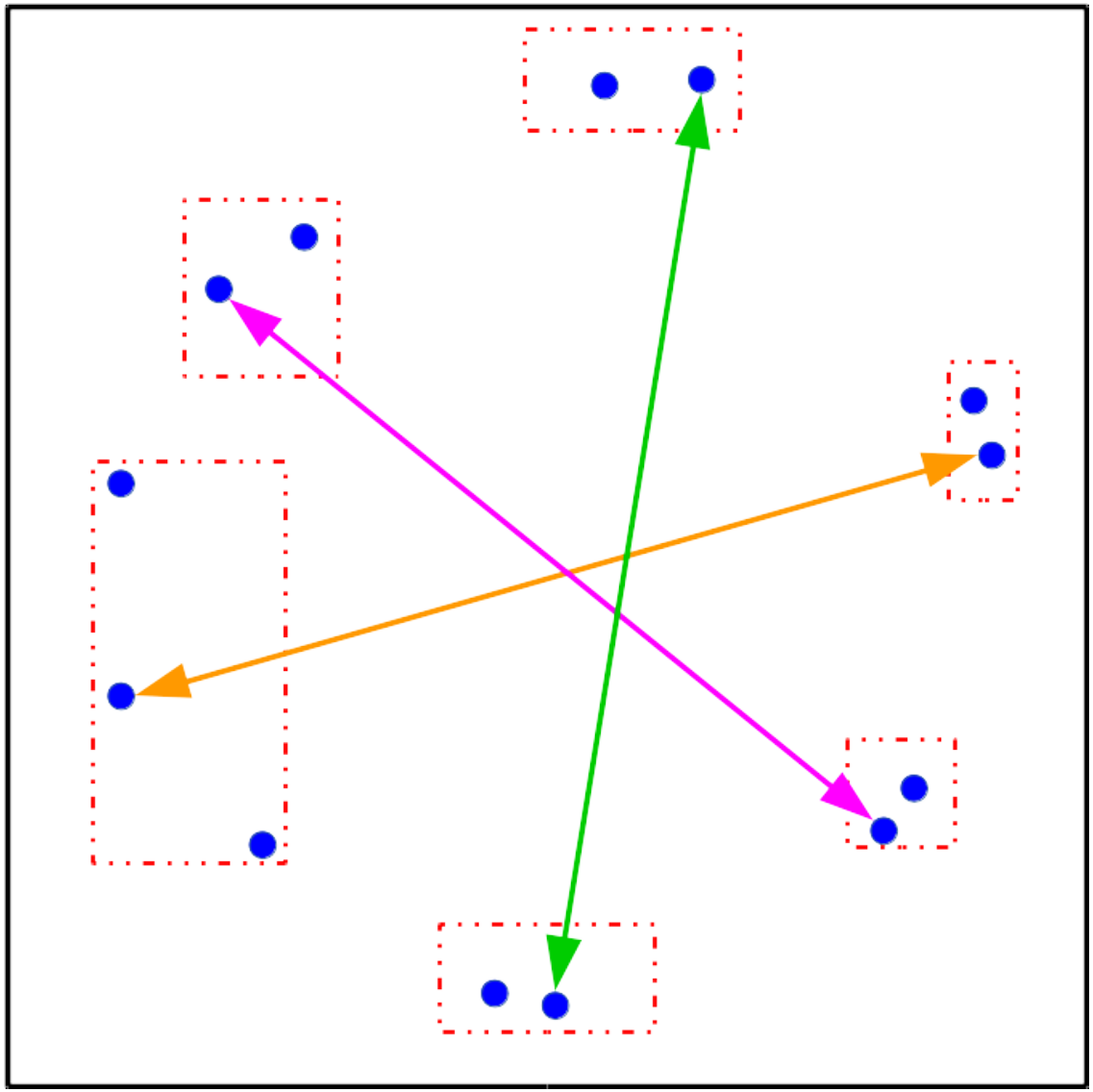
SpatialHadoop - Local



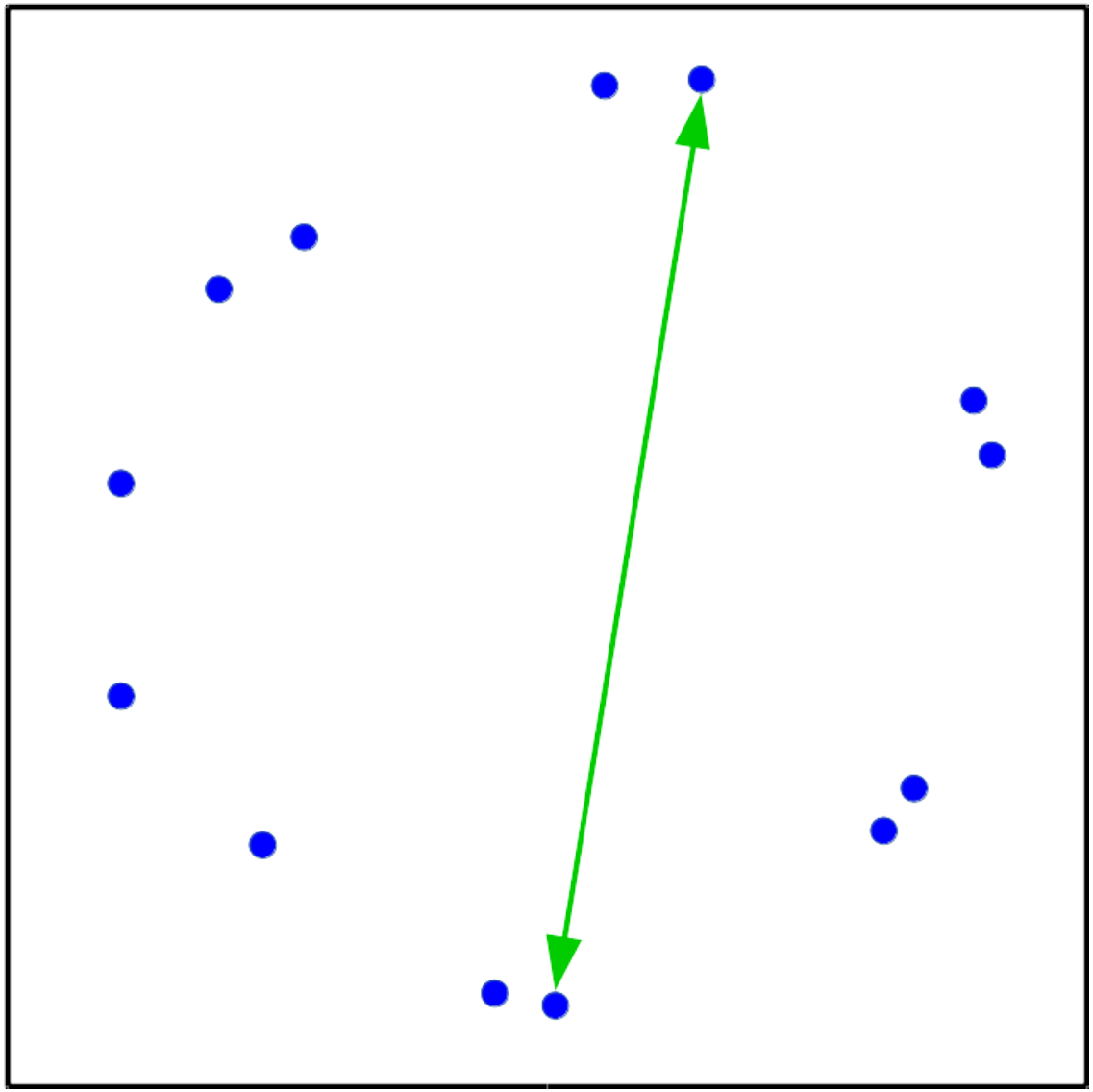
SpatialHadoop - Local



SpatialHadoop - Global

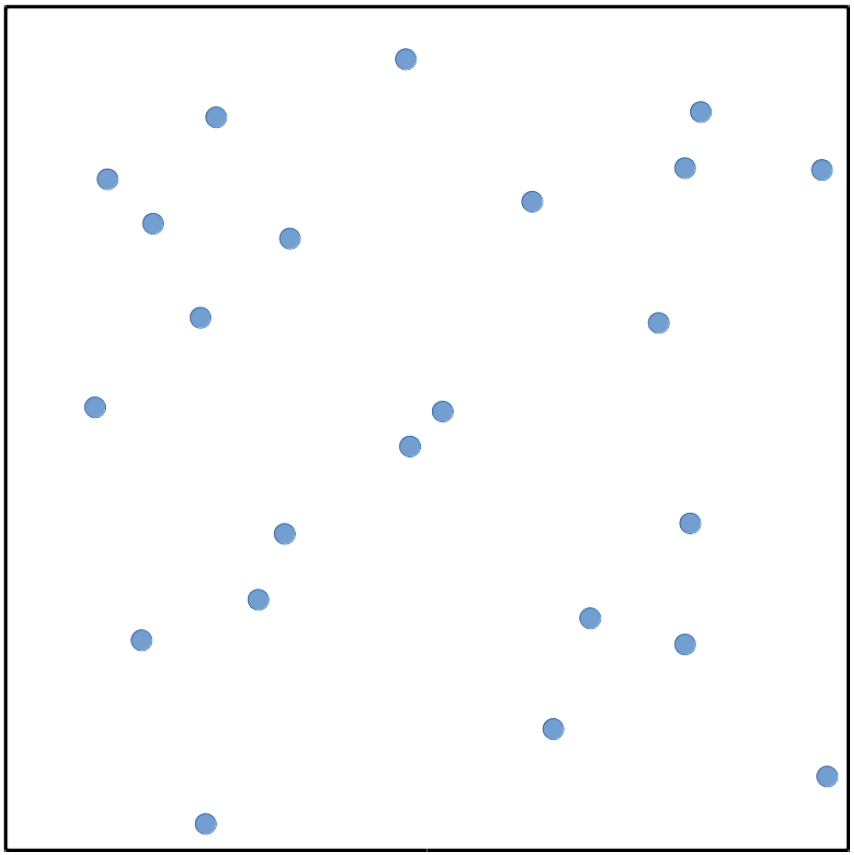


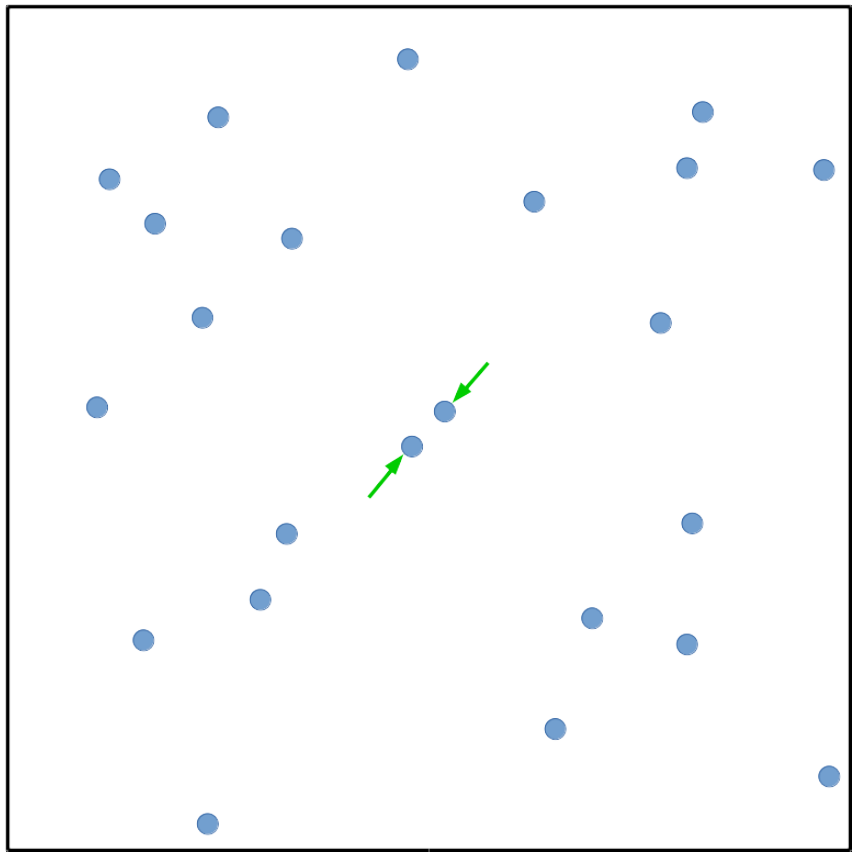
SpatialHadoop - Global



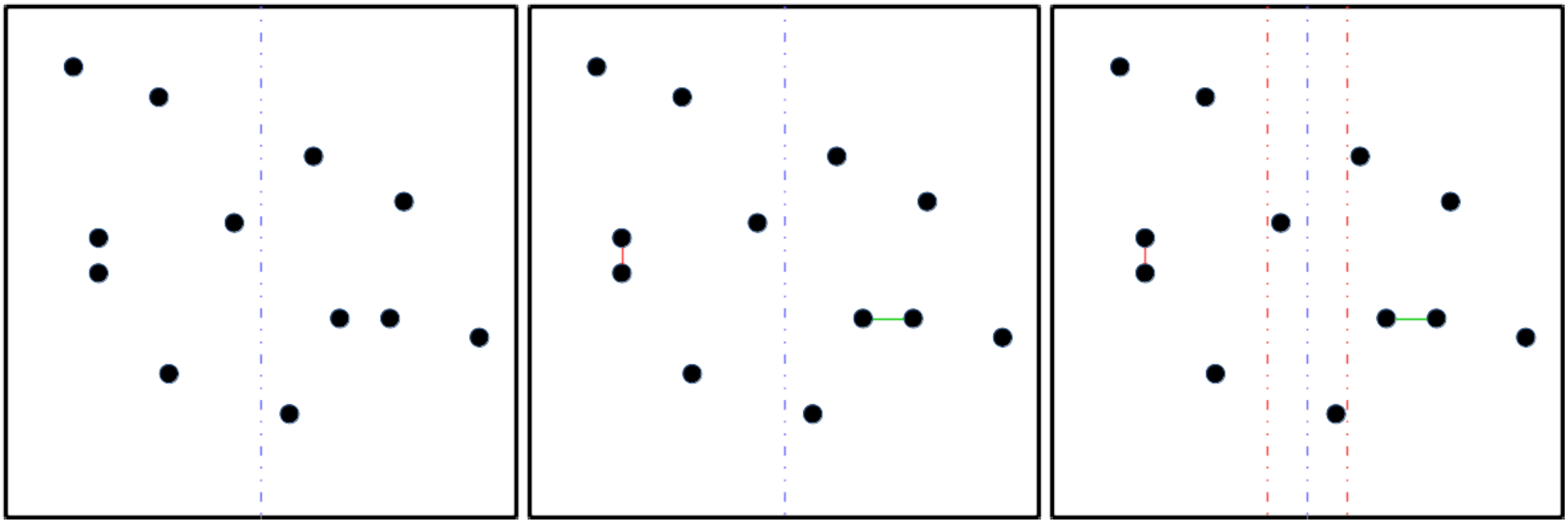
Agenda

Closest Pair

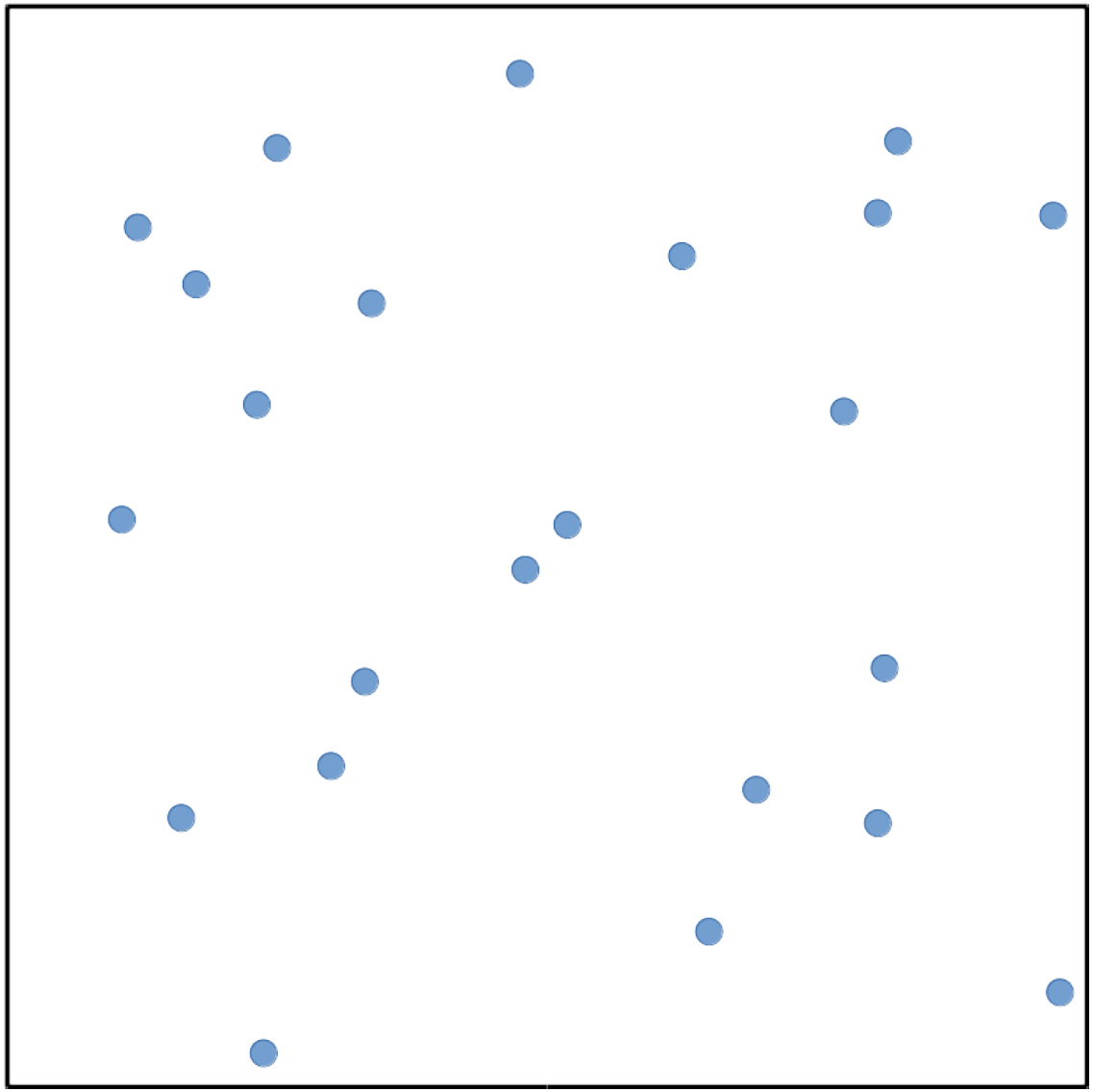




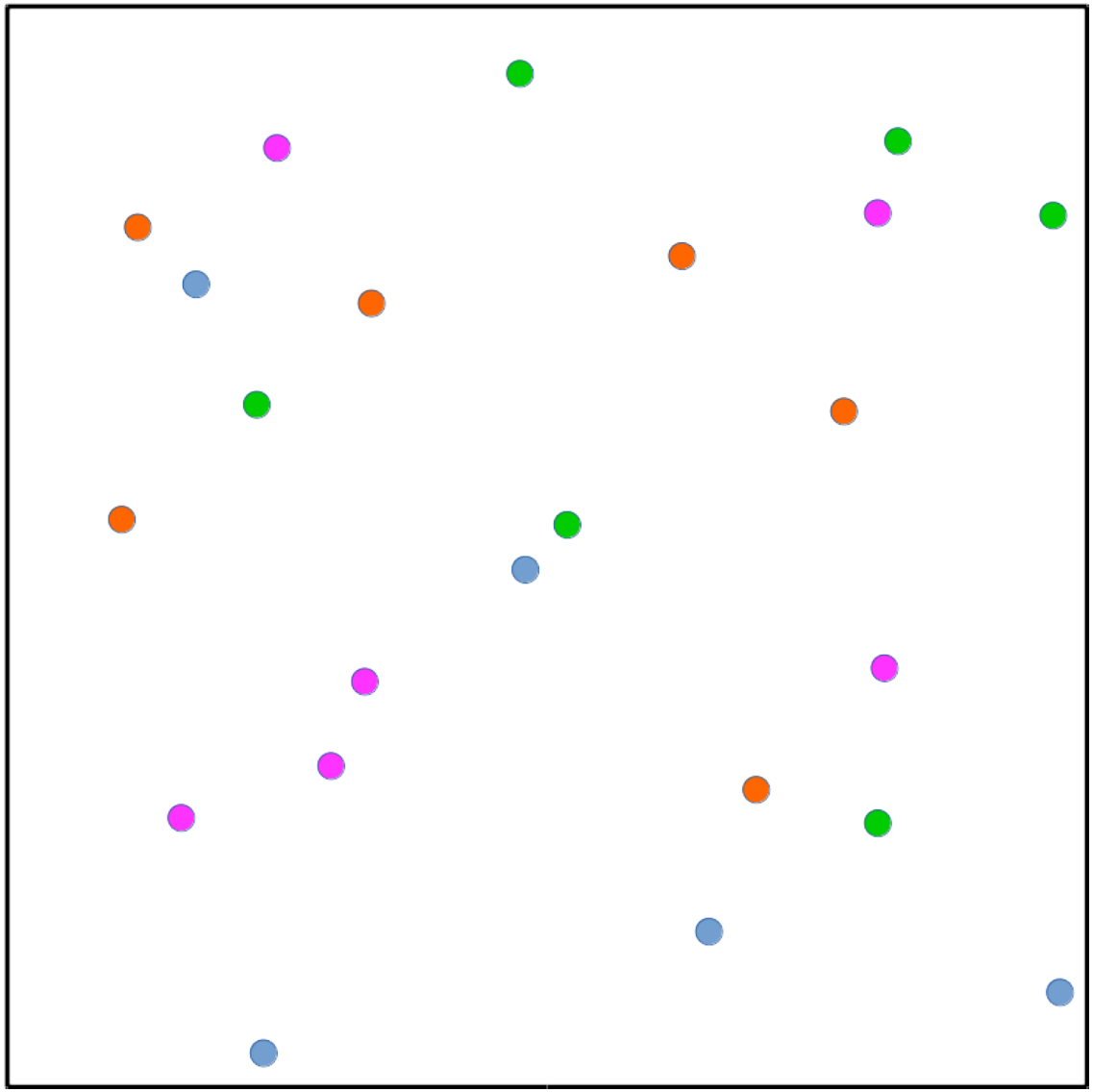
Divide and conquer approach



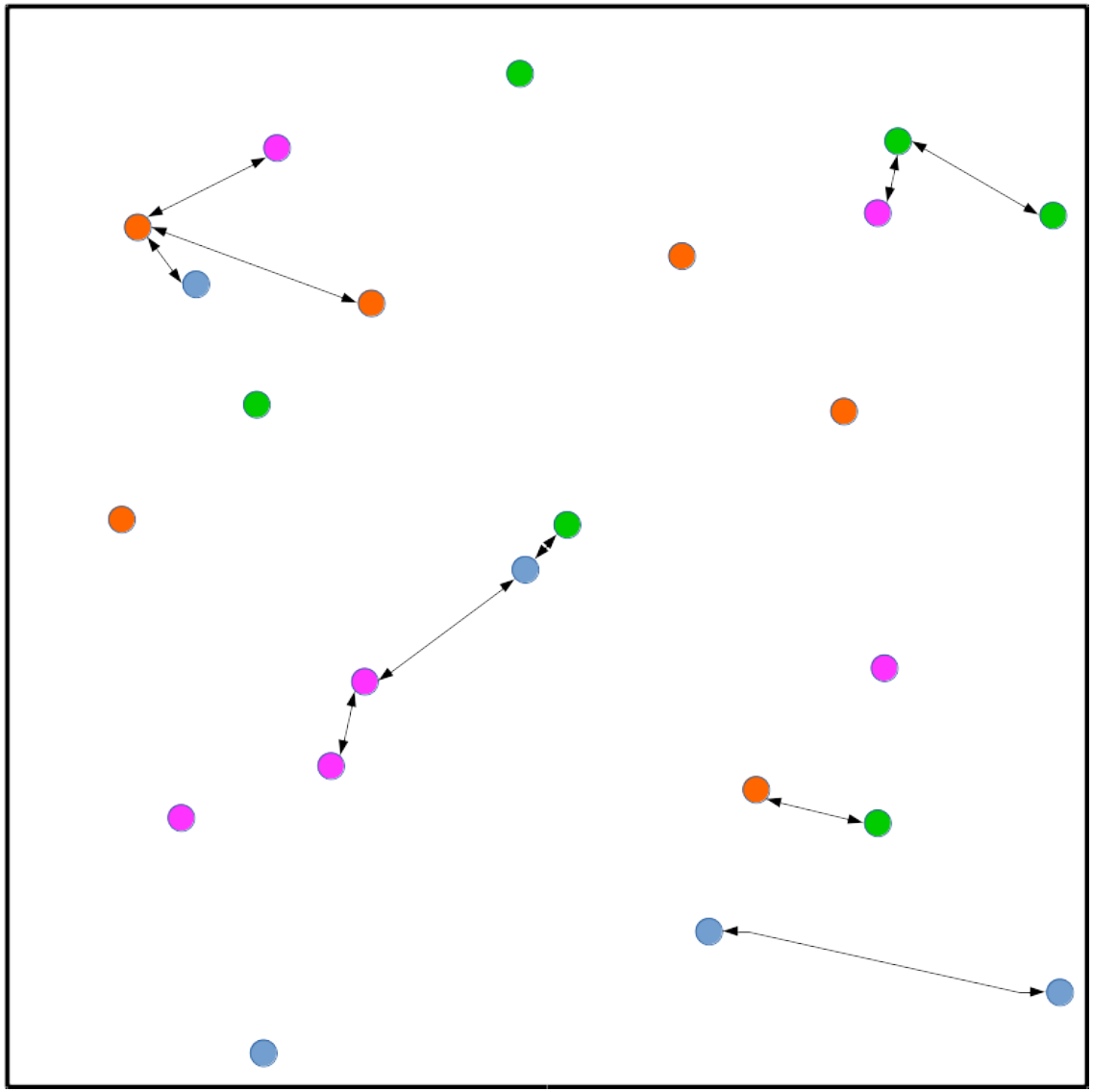
Hadoop - Partition



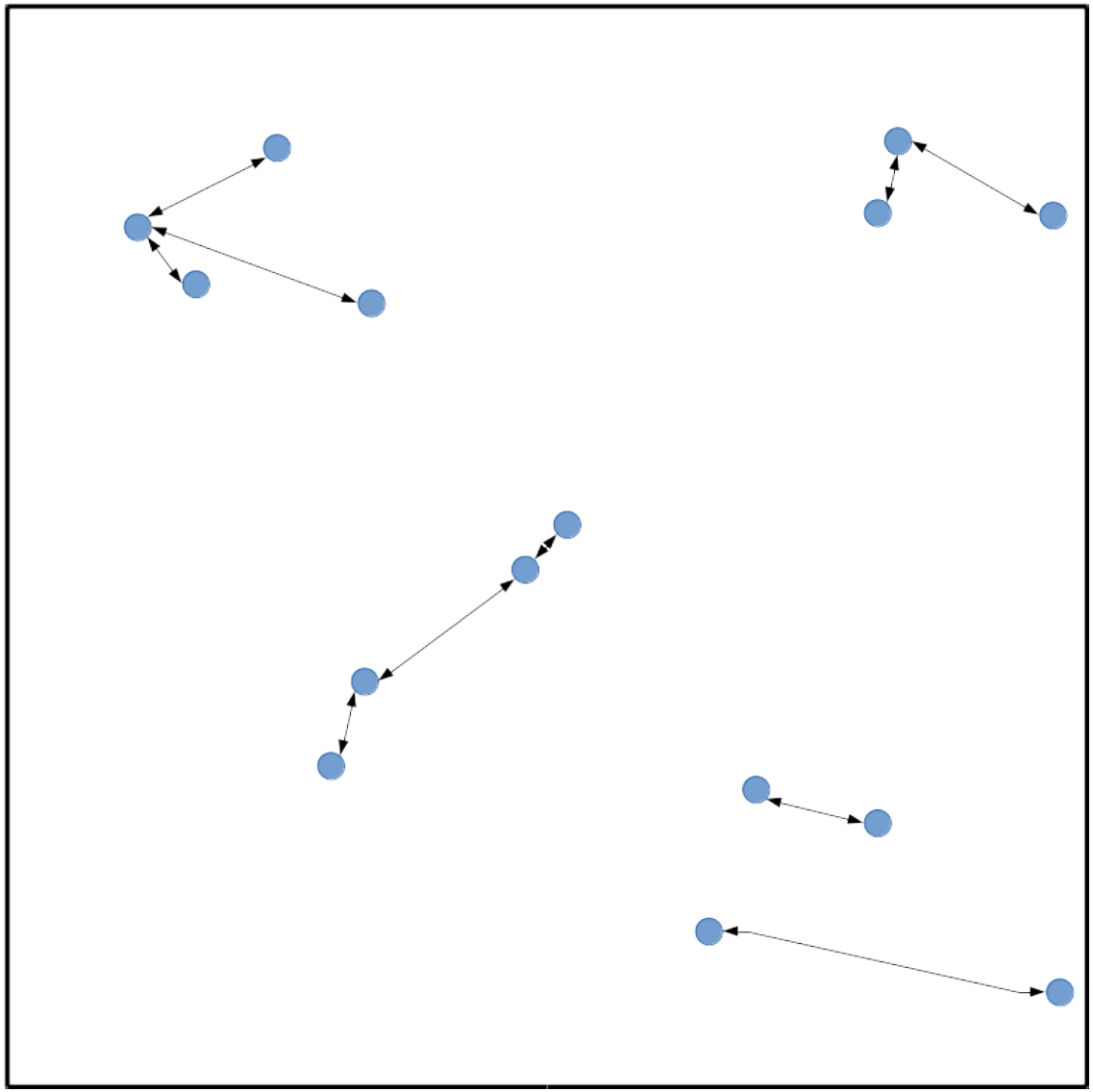
Hadoop - Partition



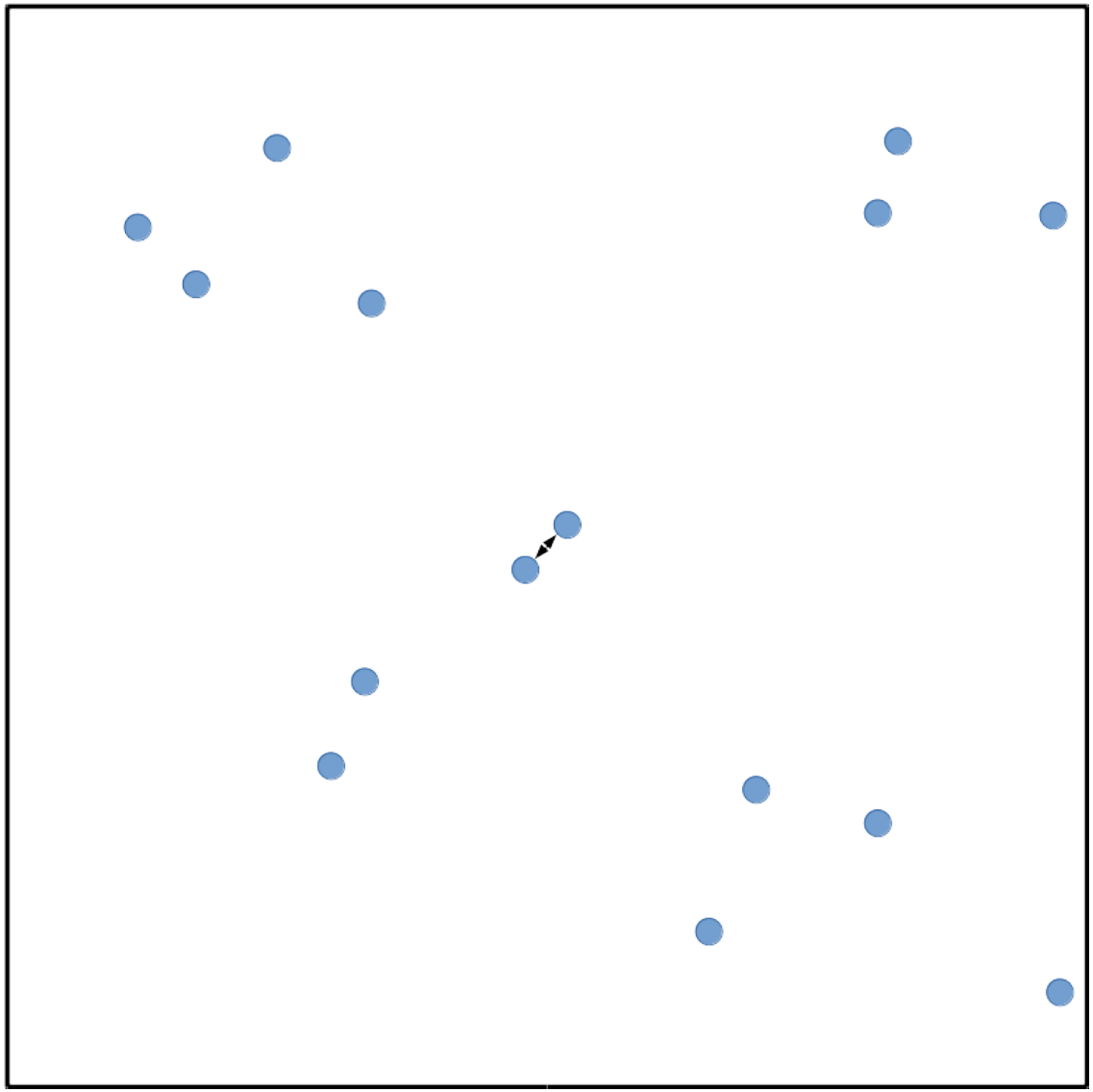
Hadoop - Local



Hadoop - Global



Hadoop - Global



Closest pair in SpatialHadoop

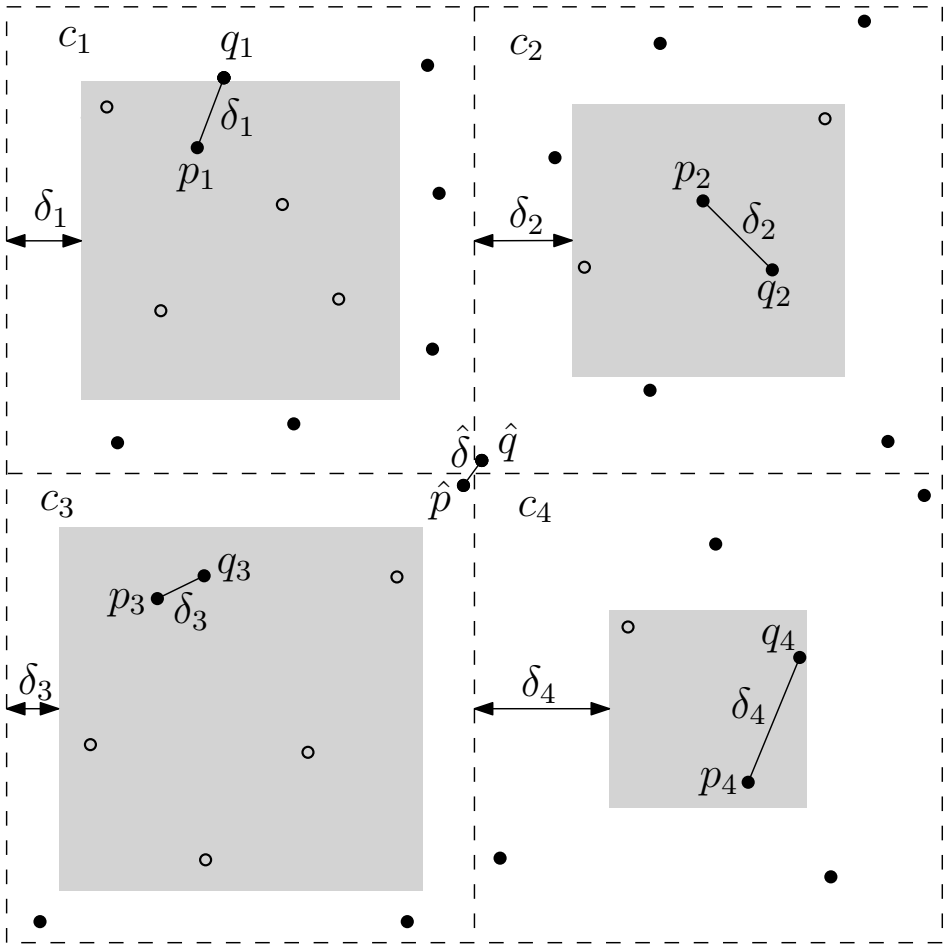
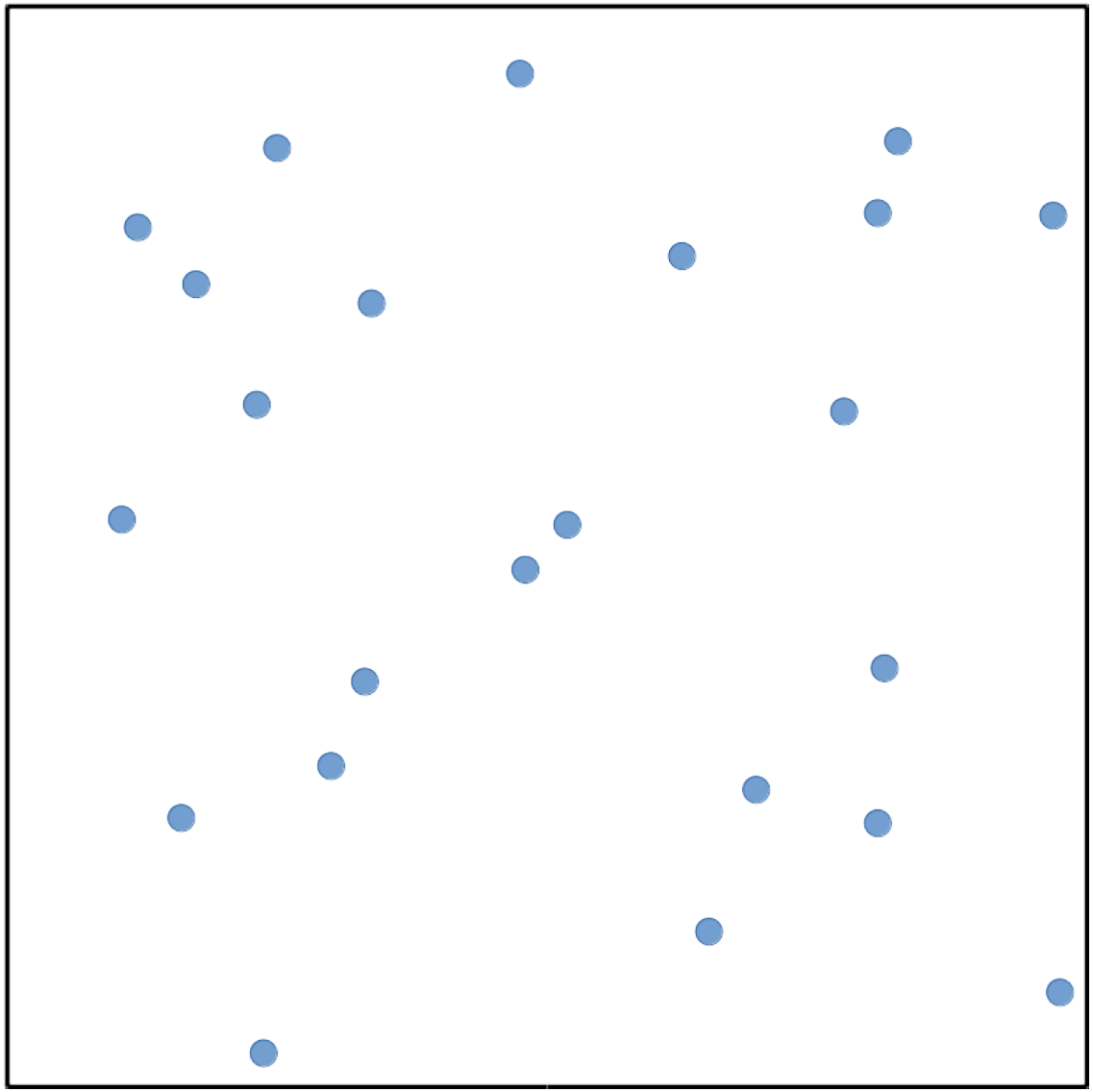
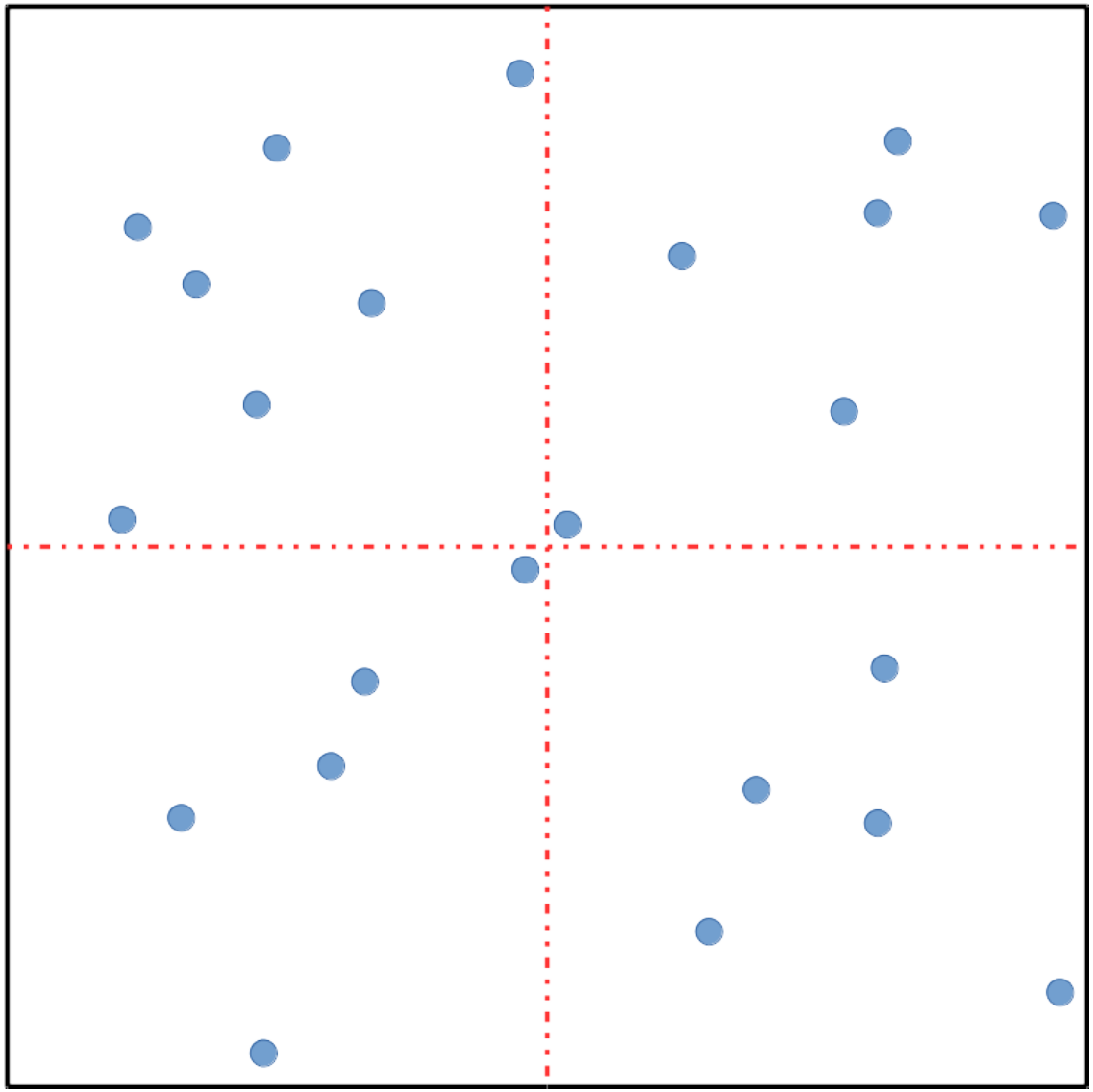


Figure 7: Closest Pair in SpatialHadoop

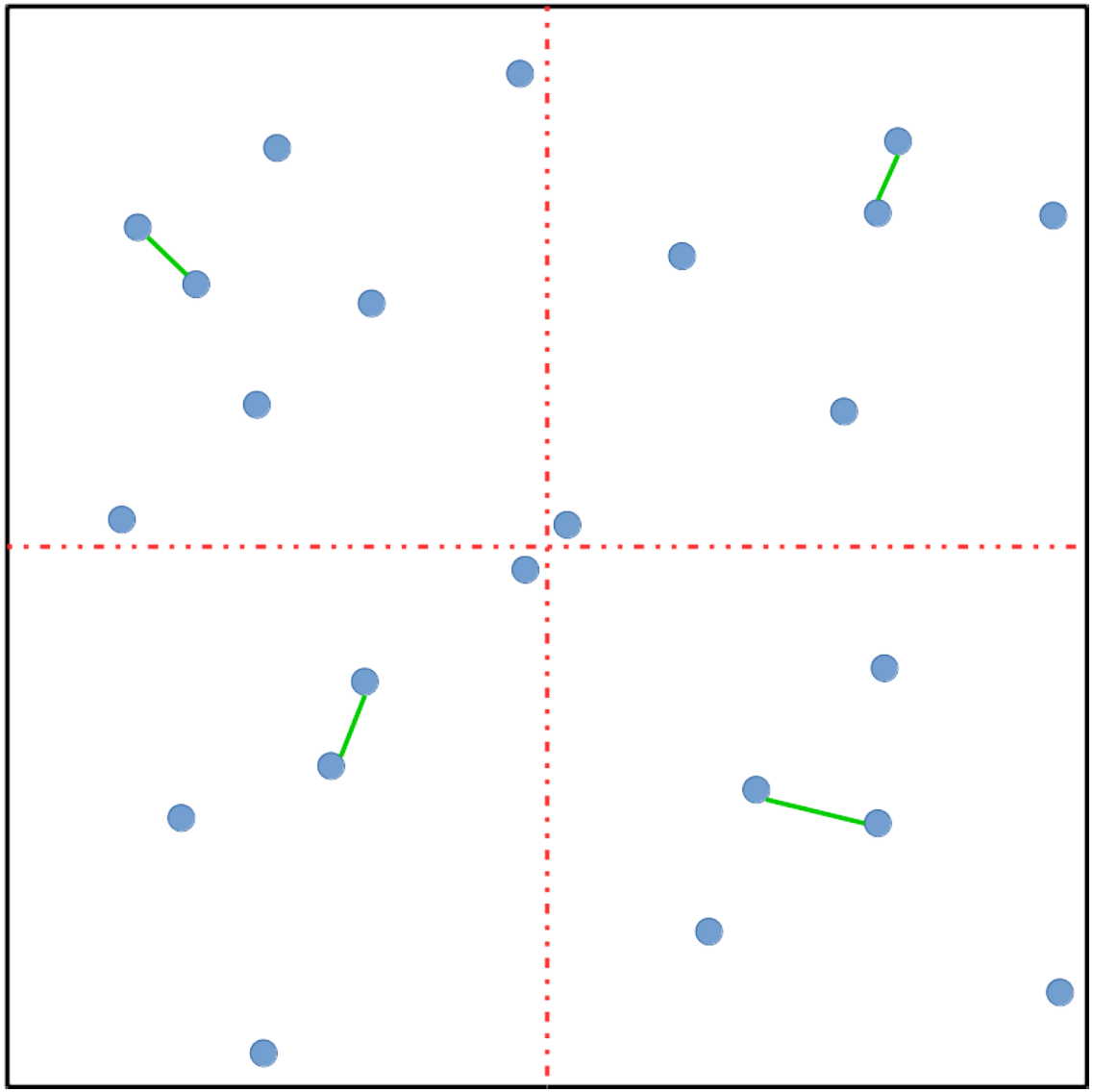
SpatialHadoop - Partition



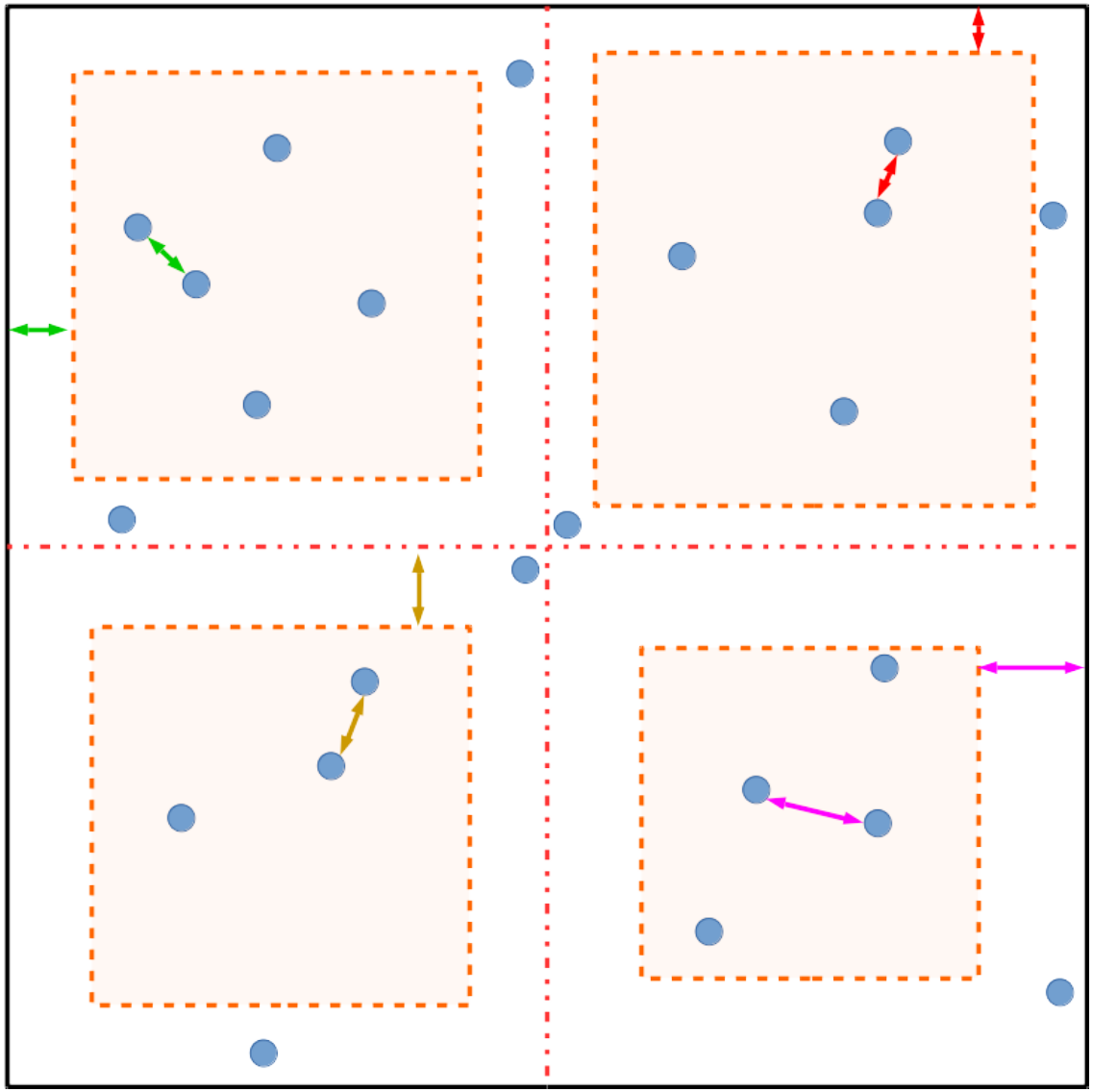
SpatialHadoop - Partition



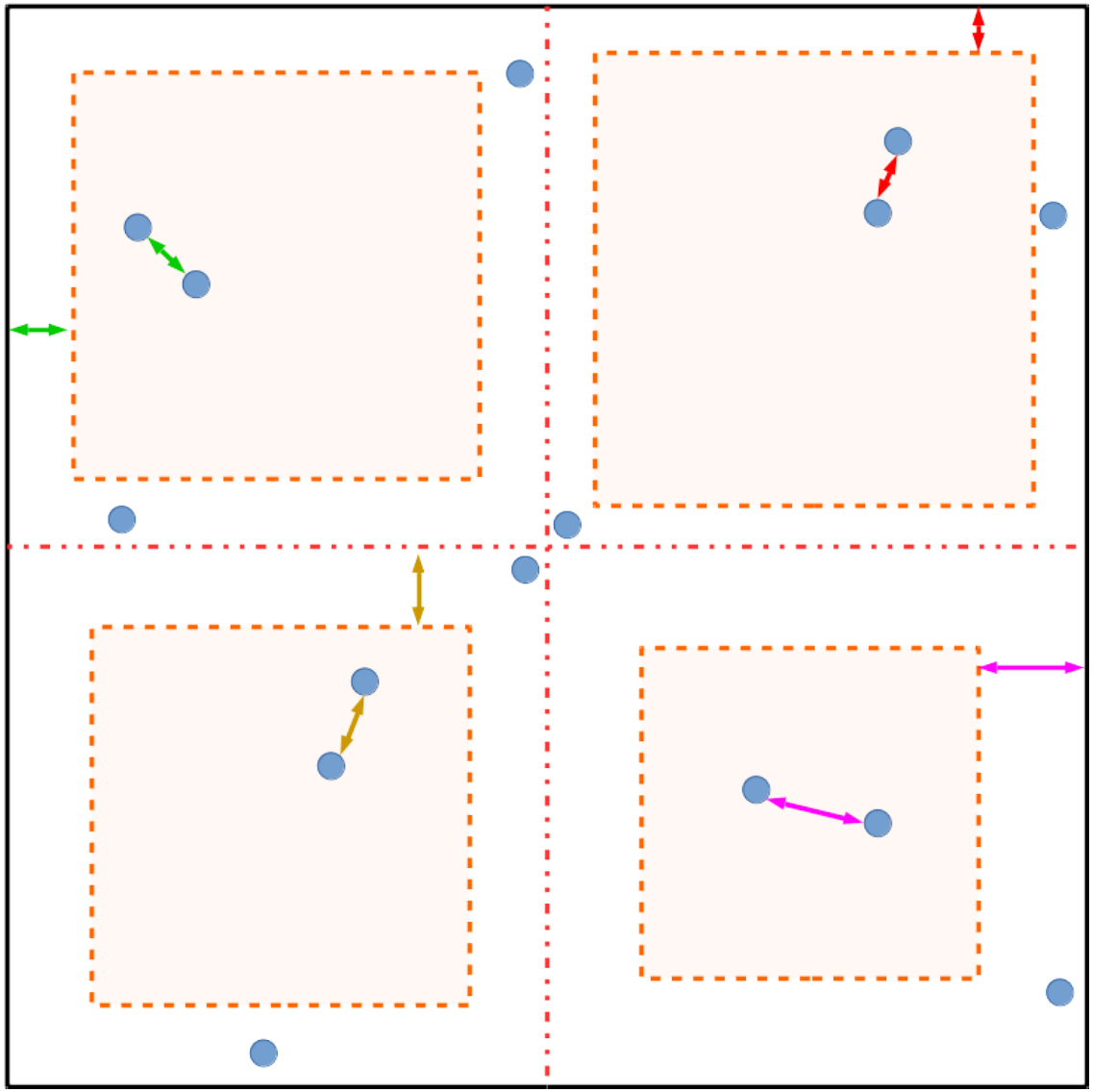
SpatialHadoop - Local



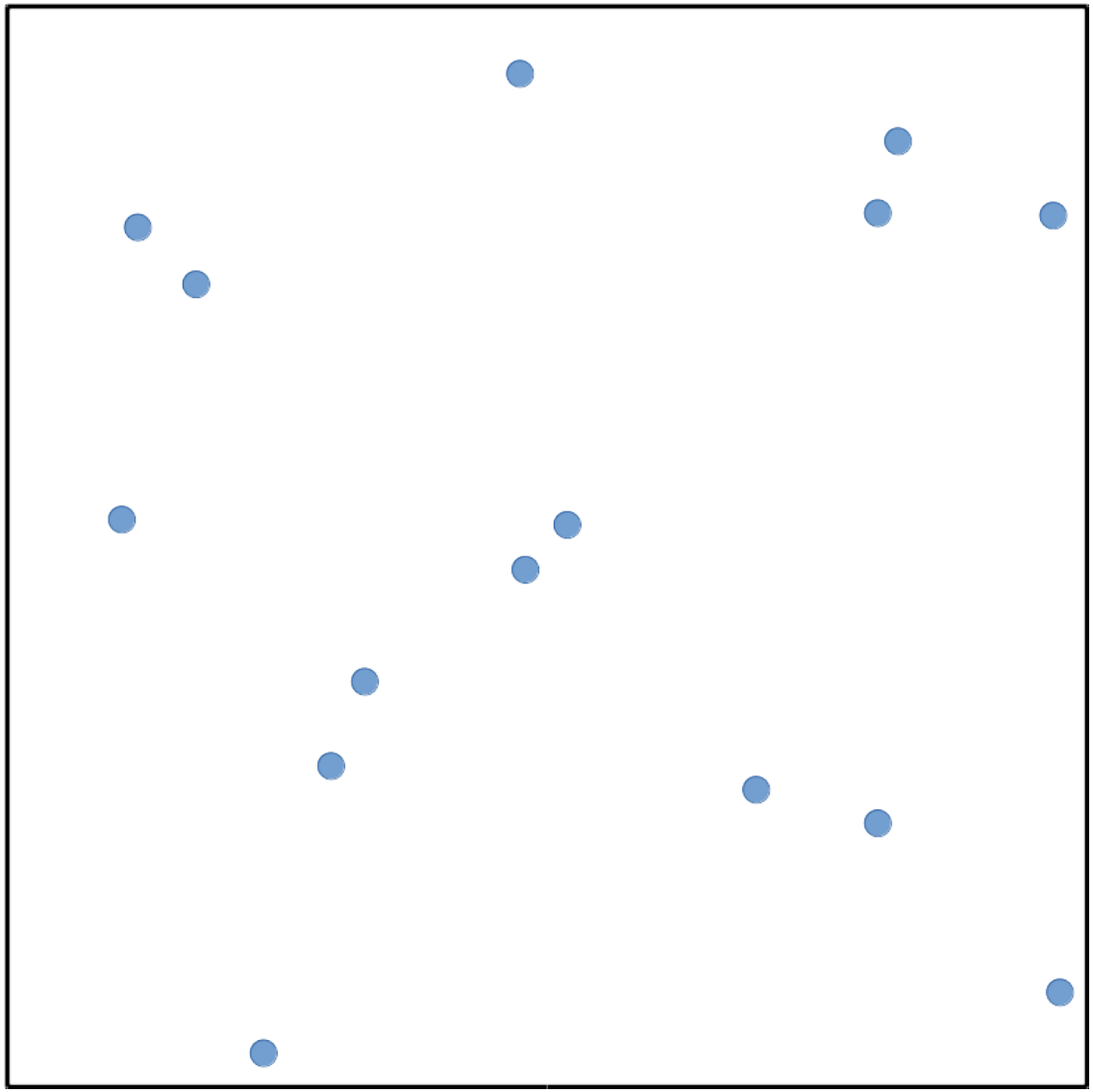
SpatialHadoop - Local



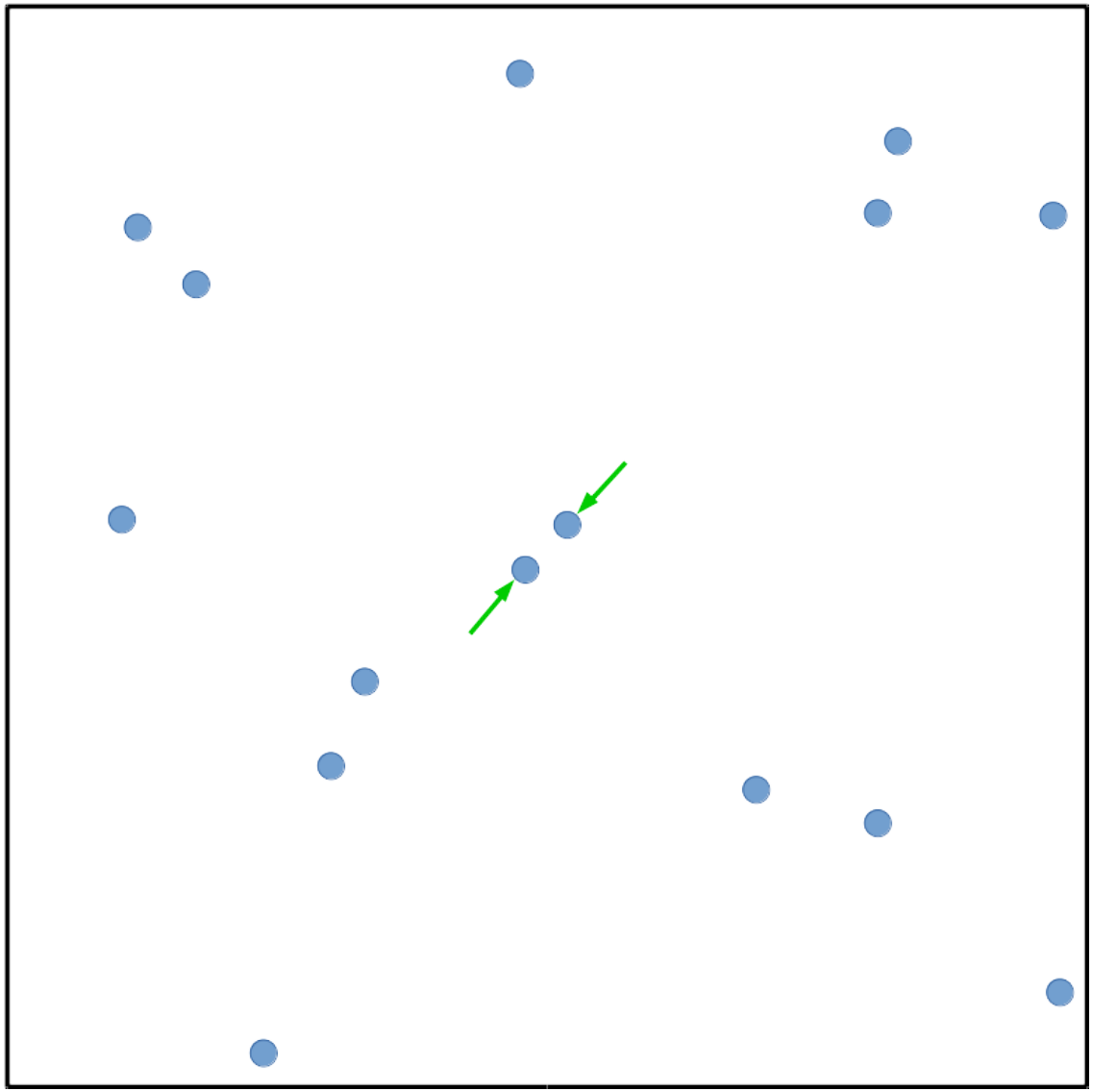
SpatialHadoop - Local



SpatialHadoop - Global



SpatialHadoop - Global



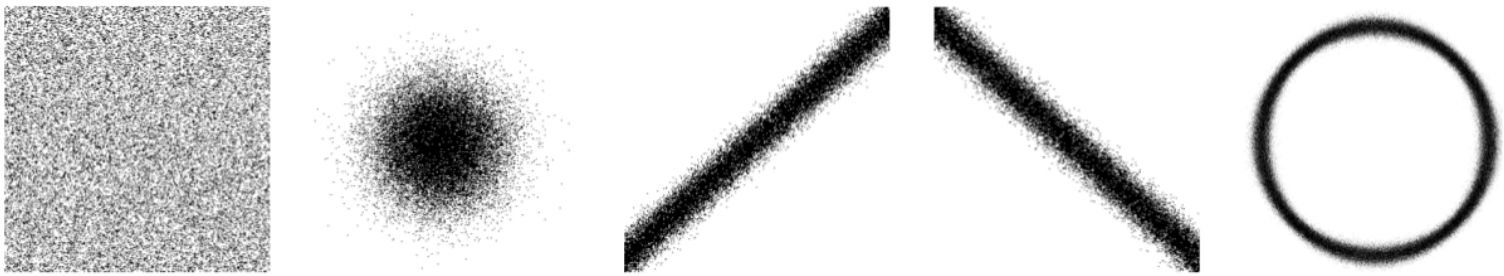
Agenda

Setup

- Cluster of 25 nodes:
  - HDD from 50GB to 200GB.
  - RAM from 2GB to 8GB.
  - Processors 2.2GHz to 3GHz
- Single machine:
  - HDD 2TB.
  - RAM 16GB.
  - Processor 3.4GHz.

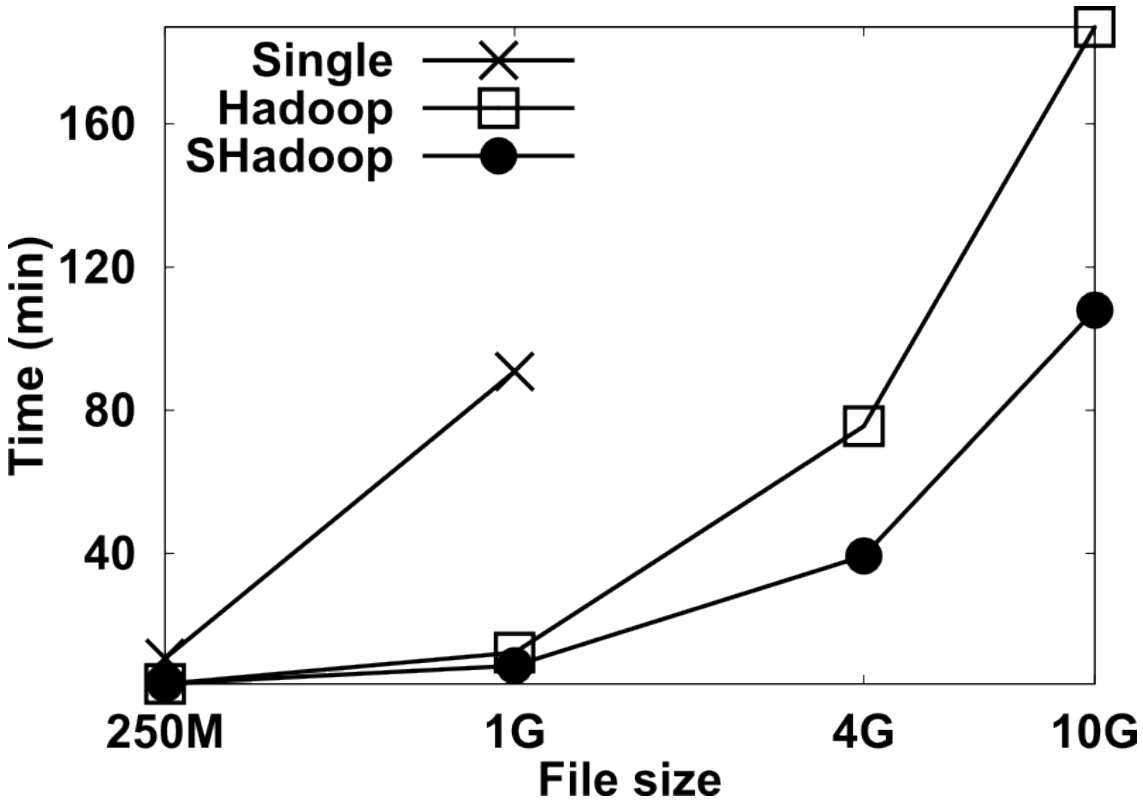
Datasets

- Real datasets (from OpenStreetMap):
  - OSM1: 164M polygons, 80GB.
  - OSM2: 1.7B points, 52GB.
- Synthetic dataset:
  - SYNTH: 3.8B points, 128GB.
  - Five different distributions.



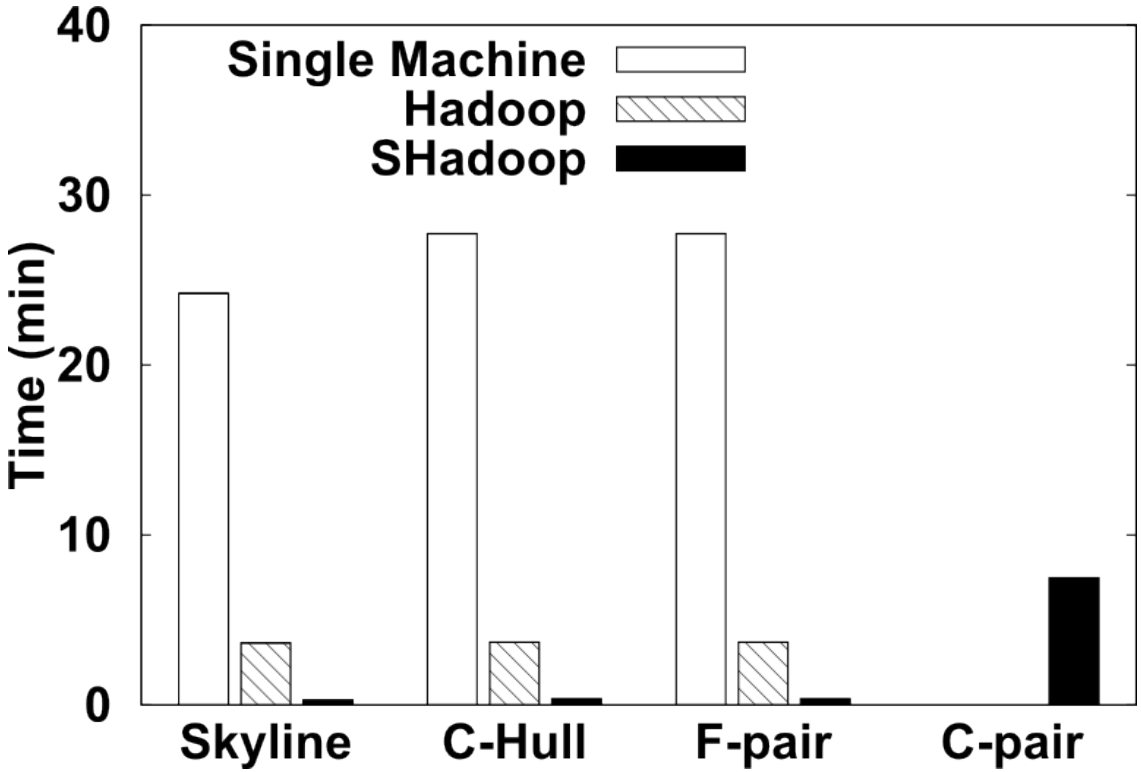
(a) Uniform (b) Gaussian (c) Correlated (d) Anti-correlated (e) Circular

Union


Single
Hadoop
SHadoop
160
120
80
40
Time (min)
250M
1G
4G
10G
File size

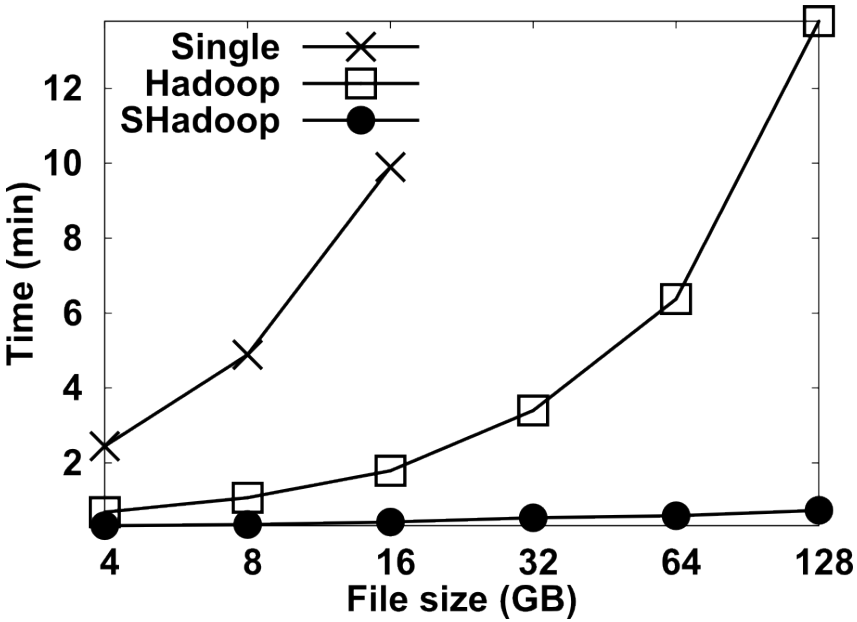
OSM1

Others

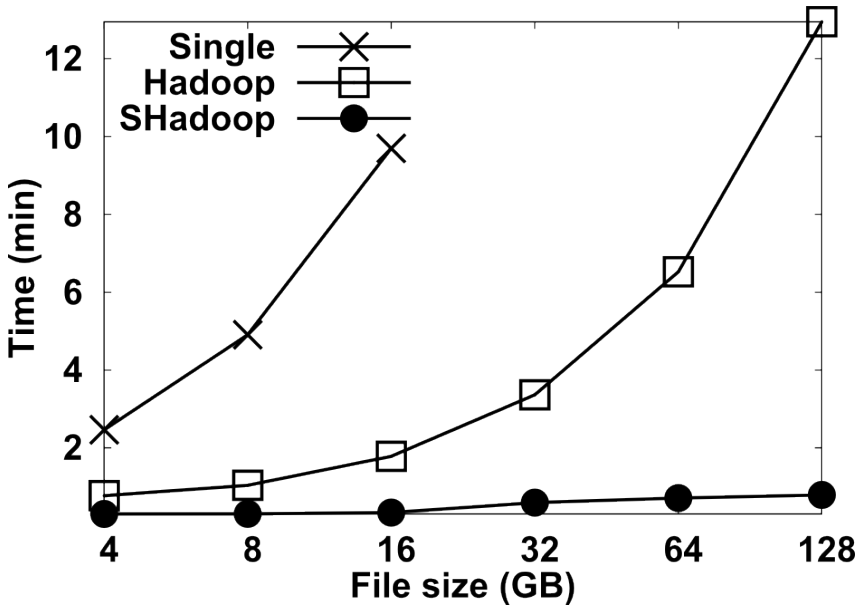

40
30
20
10
0
Time (min)
Single Machine
Hadoop
SHadoop
Skyline
C-Hull
F-pair
C-pair

OSM2

Skyline

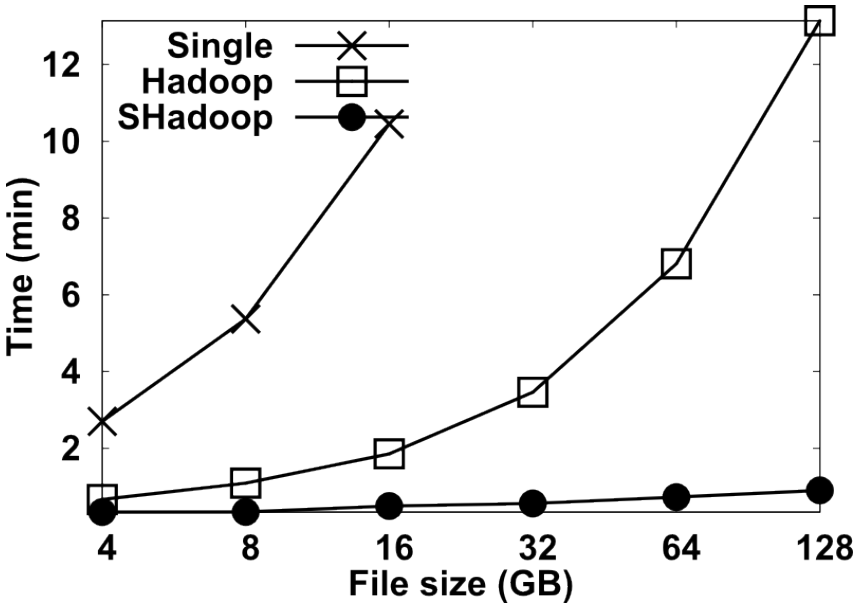


Correlated

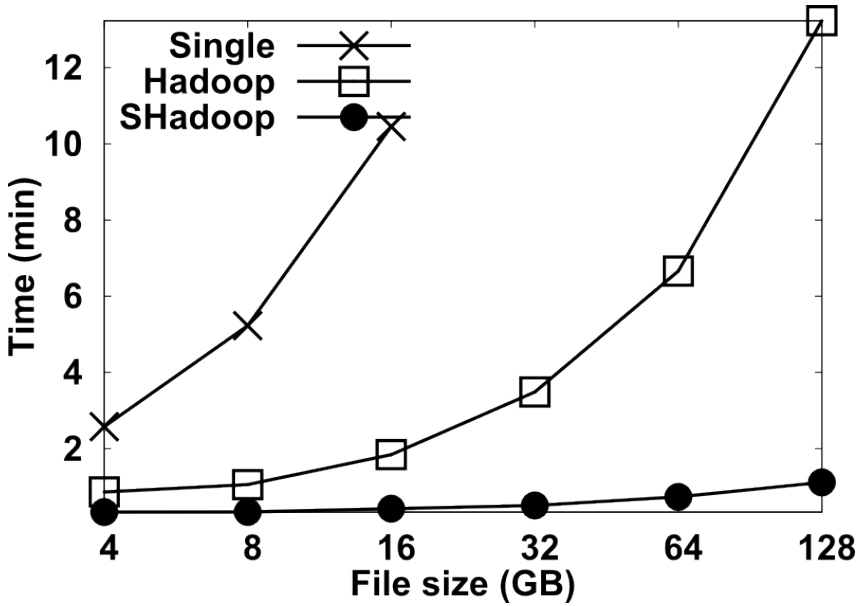


Anti-correlated

Convex hull

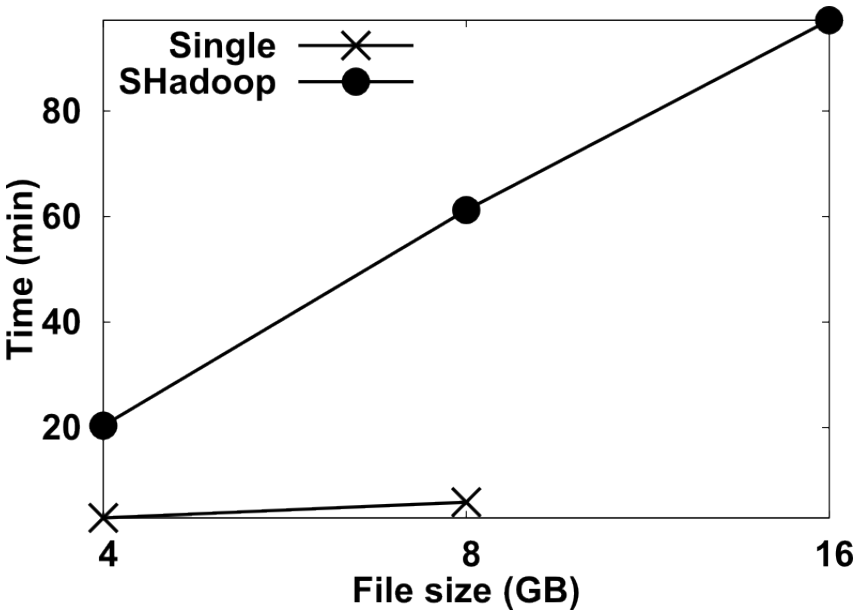


Uniform

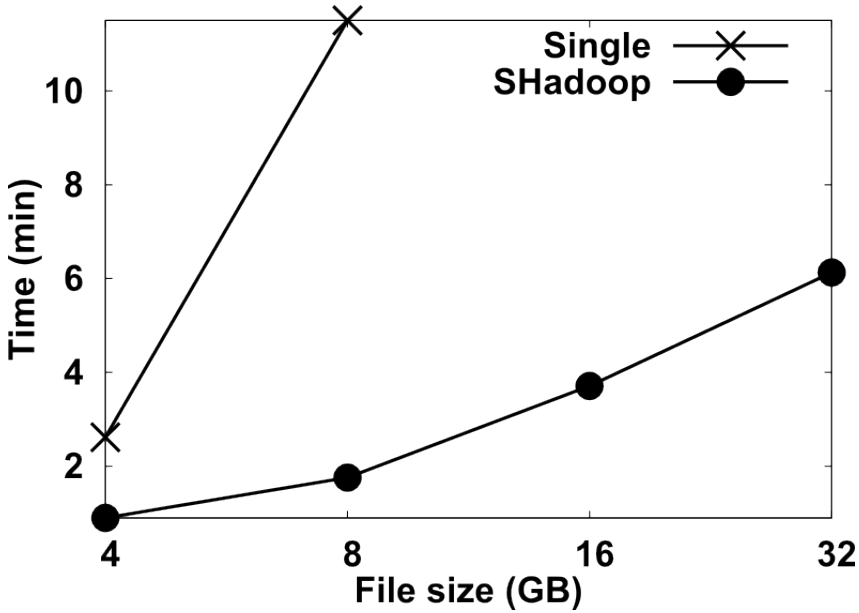


Gaussian

Closest/Farthest pair



Farthest Pair



Closest Pair

Agenda

Conclusions

- This paper introduced CG_Hadoop as a scalable and efficient MapReduce library.
- Focused on 5 fundamental computational geometry problems...
  - Polygon union, Skyline, Convex hull, Farthest and Closest Pairs.
- Provided versions for Apache Hadoop and SpatialHadoop systems.
- Distributed approach speed up performance.
- Spatial partitioning allows early pruning which make it even more efficient.
- Achieve up to 29x and 260x better performance.

Future ideas

- Working on more complex operations, for example motion patterns.
- Explore ports to new distributed platforms such as Spark or Simba.

Thank you!!!

Do you have any question?