Locality-Preserving Spatial Partitioning for Geo Big Data Analytics in Main Memory Frameworks

Isam Mashhour Al Jawarneh, Paolo Bellavista, Antonio Corradi, **Luca Foschini**, Rebecca Montanari

Department of Computer Science and Engineering - DISI
University of Bologna, Italy

{luca.Foschini, isam.aljawarneh3, paolo.bellavista, antonio.corradi,Rebecca.montanari}@unibo.it,

IEEE GLOBECOM 2020
7th - 11th December 2020

Agenda

- *Spatial Partitioning*: Background and Motivations
- *QoS-aware Spatial partitioning for in-memory systems*
  - SCAP (Spatial Co-locality-aware Partitioner)
  - Supported spatial queries: proximity-alike
- Deployment
  - Baselines
  - Testing setup
- Results and Discussion
  - Azure Cloud Spark Test
- Concluding remarks

Smart City and Big Data Context

Smart City
Advanced technological services

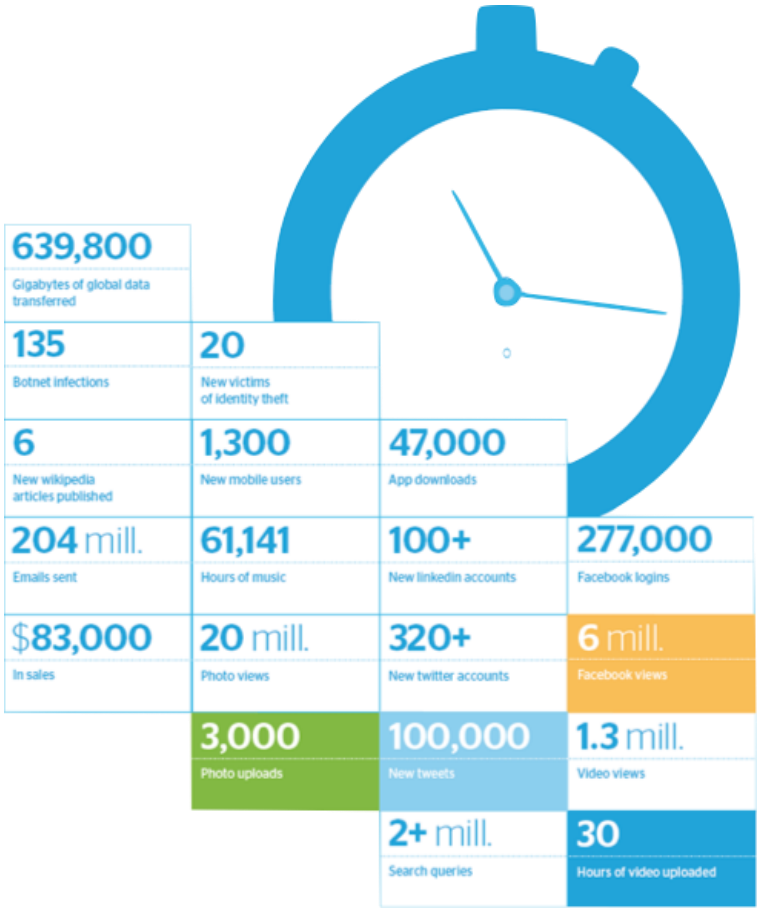
Geographic Big Data
Huge amount of information

Mobile Sensing
Automatic collection of data

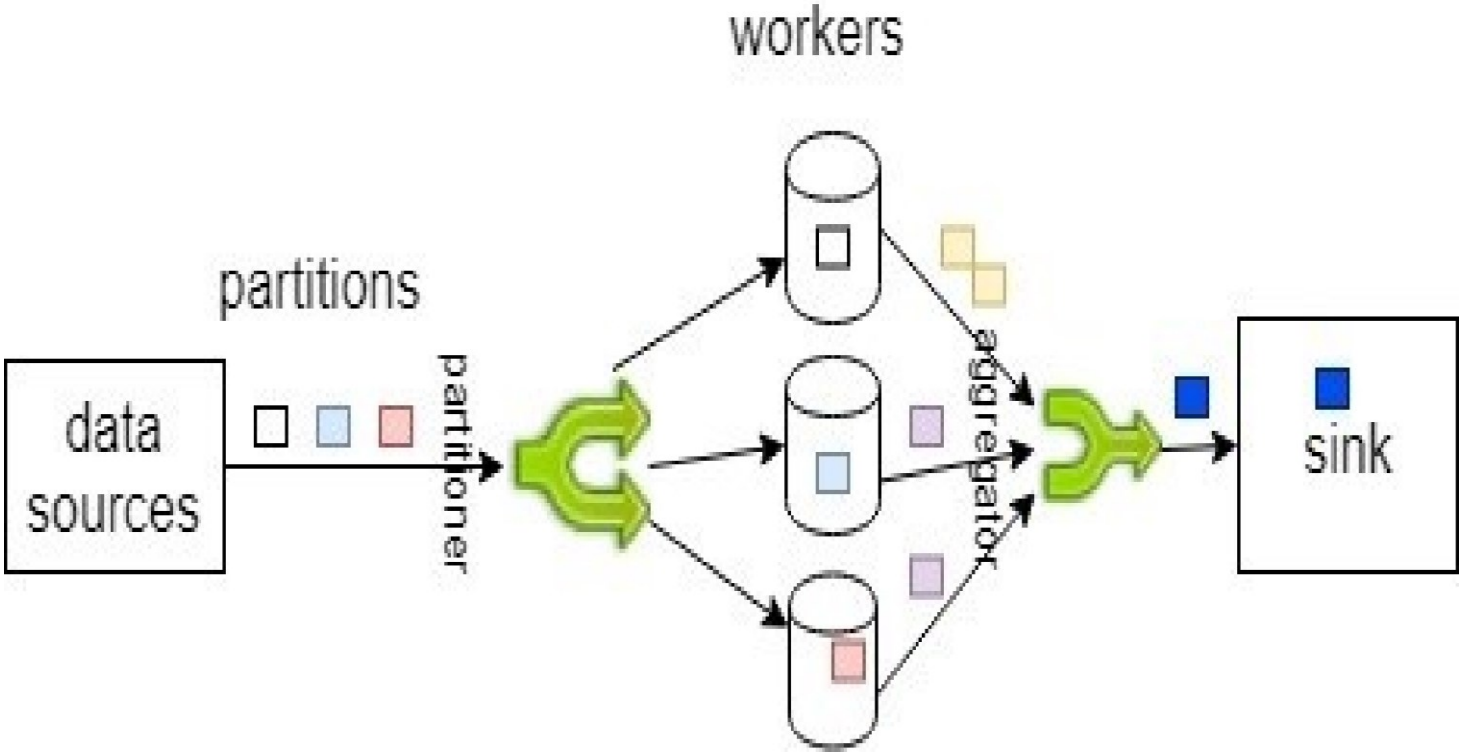
Geo-located Big Data

- Large amounts of geolocated data, **exceeding processing capability** of traditional database management systems
- Charateristics
  - **Volume**: amount of data
  - **Velocity**: streaming data
  - **Variety**: multiple sources, heterogeneous data

Requires distributed data processing systems



Typical data parallelization for IoT in Cloud



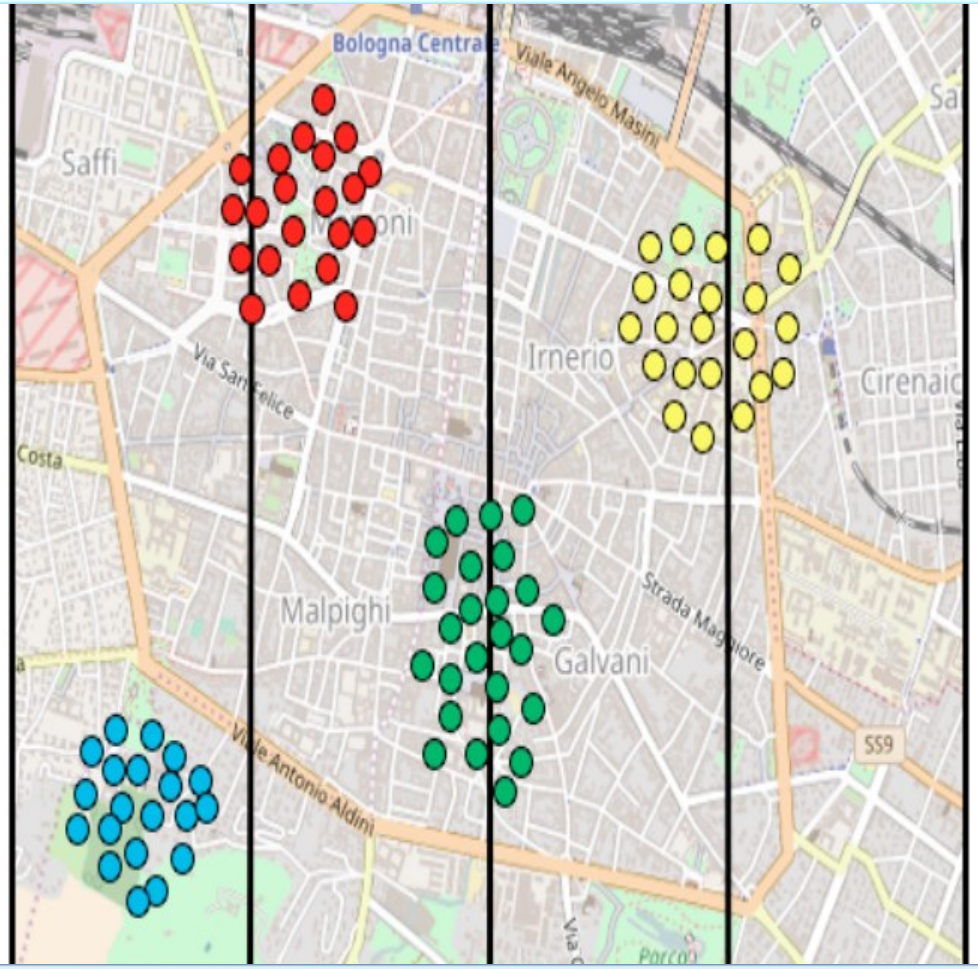
Conflicting Partitioning Challenges

Load Balancing

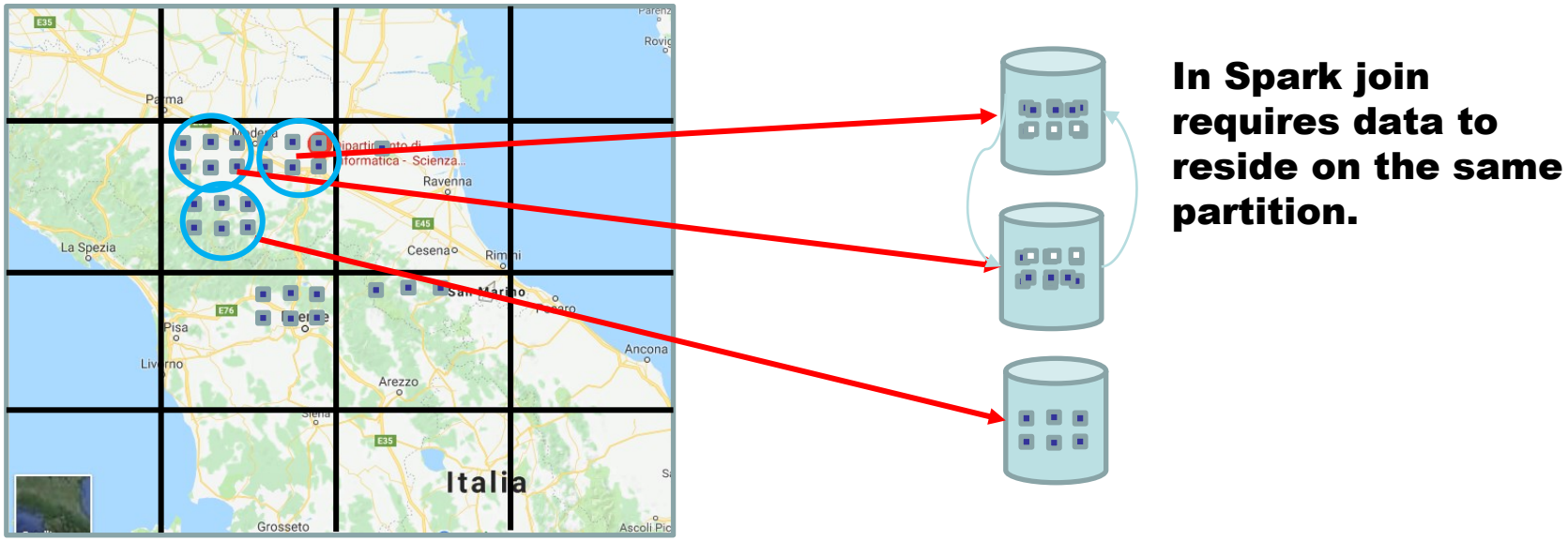
Spatial co-locality

Boundary objects

Example: Boundary Spatial Objects (BSO)



Load Balancing



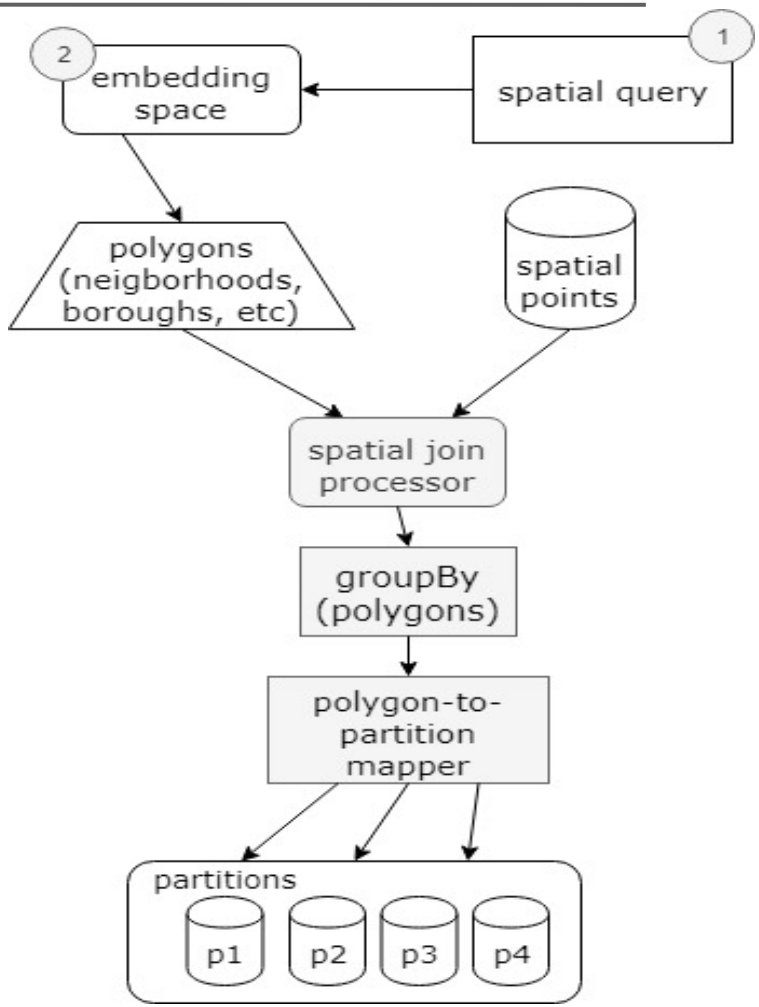
In Spark join requires data to reside on the same partition.

Only load balancing = shuffling (huge toll) for co-location queries.

Spatial Co-locality-aware Partitioner (SCAP)

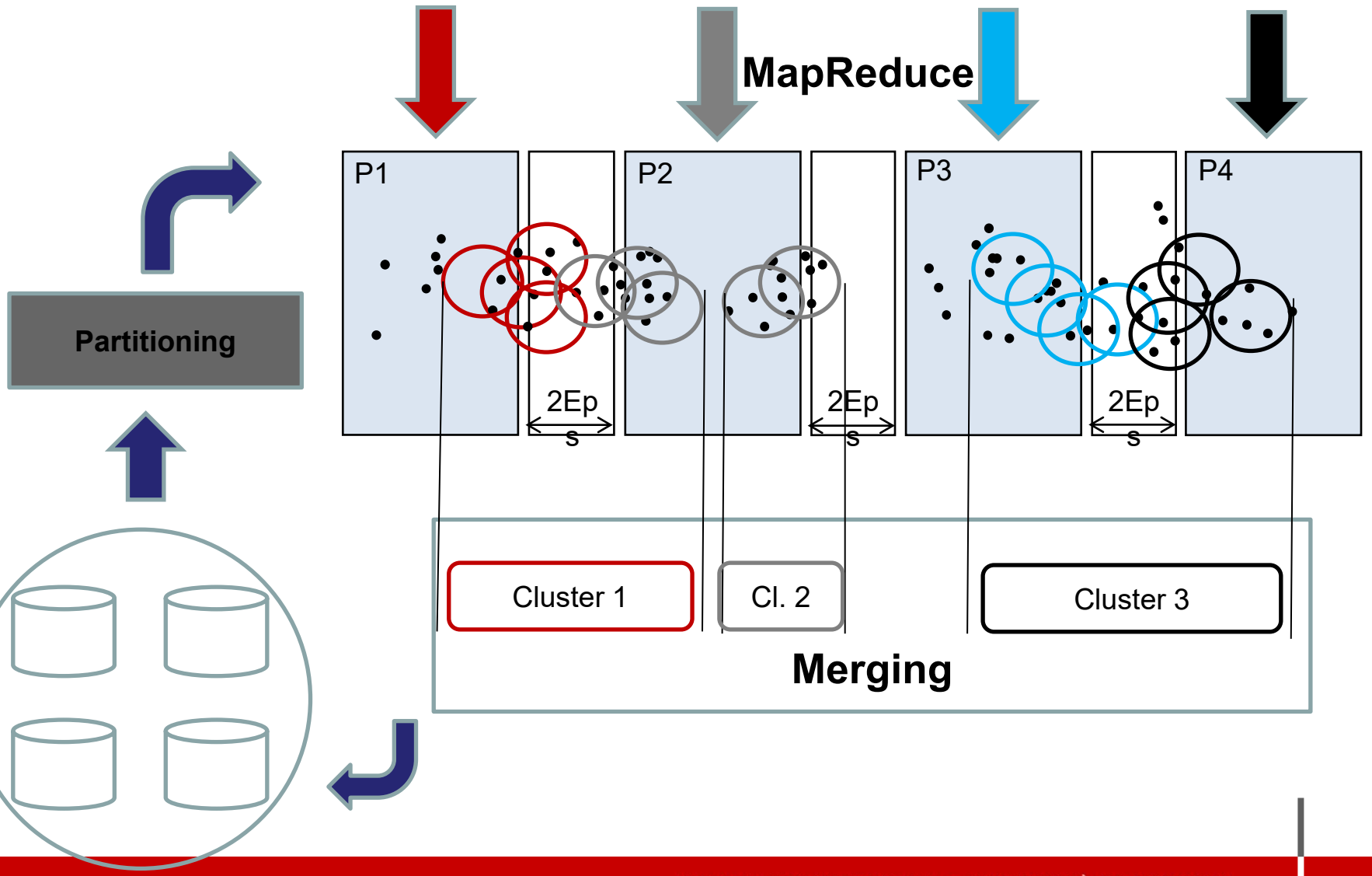
- **SDL** preservation is a priority
- **BSOs** and **load balancing** to a lesser extent

- ***Clump*** geometrically co-located objects into single chunks
- ***Split*** *overloaded* chunks
- ***Map*** chunks *to partitions*



SCAP

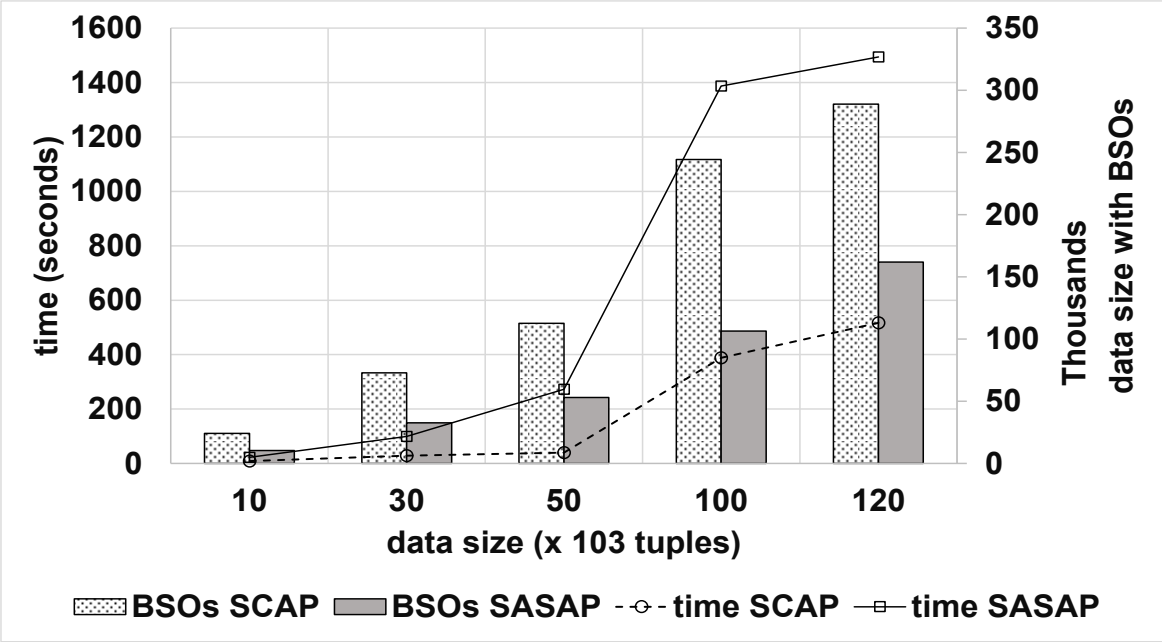
DBSCAN-MR


MapReduce
P1
P2
P3
P4
2Eps
2Eps
2Eps
Partitioning
Cluster 1
Cl. 2
Cluster 3
Merging

Experimental setup

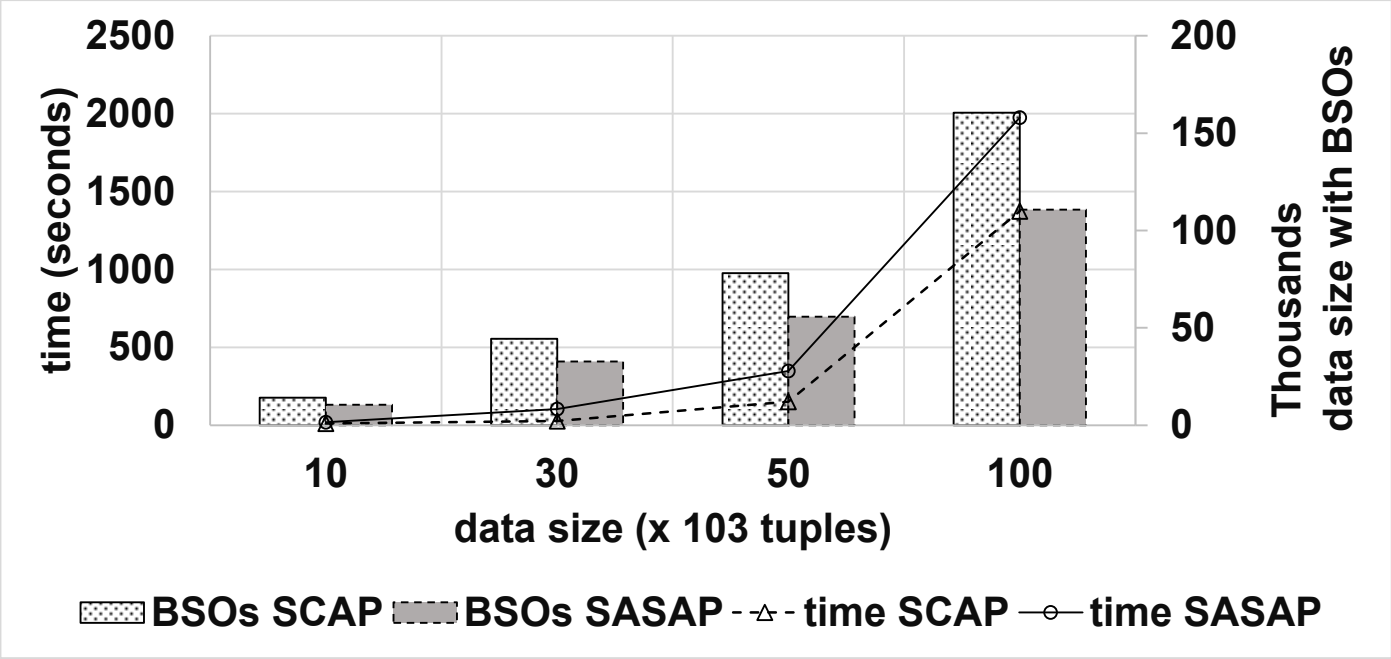
- Evaluation questions
  - Data size vs Accuracy & BSOs count
  - Adaptability effect
- Testbed
  - Cluster: 6 nodes (Microsoft Azure HDInsight Cluster)
  - Datasets:
    - NY City taxicab trips datasets. 150k points representing a portion of data captured through taxi rides for the first half of 2016
    - 150k spatial points collected during the ParticipAct project

Data size vs Accuracy & BSOs count



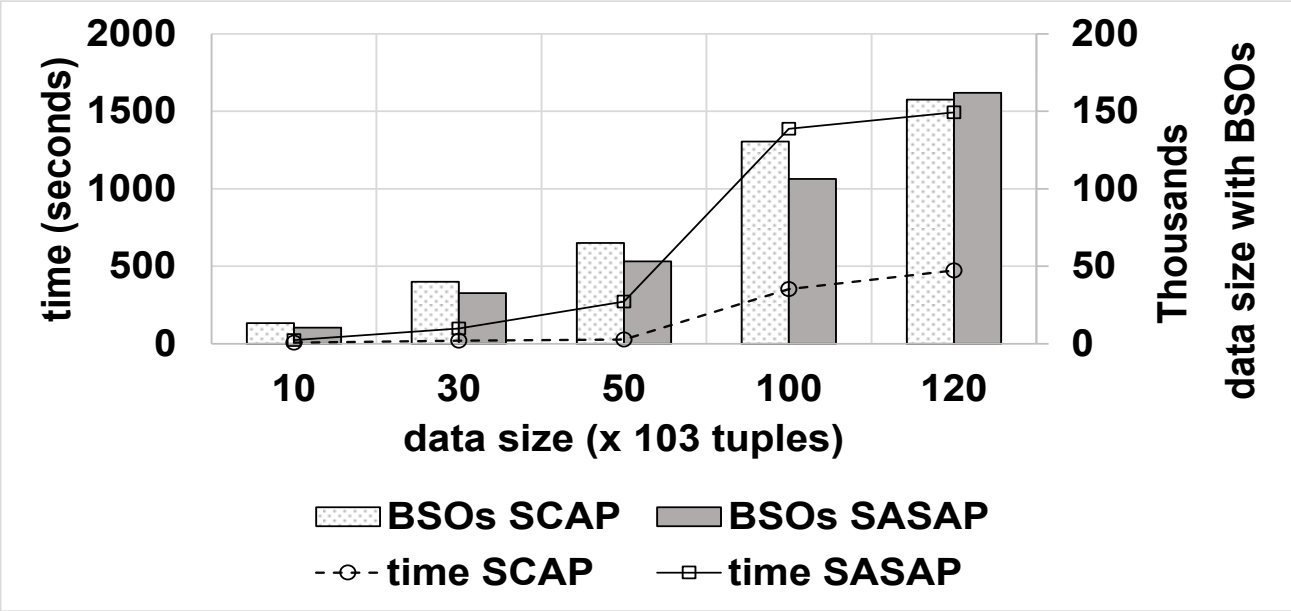
- Running times and number of BSOs of our retrofitted version of DBSCAN-MR over SCAP against SASAP-based version tested on NYC taxicab datasets. Parameters: eps 0.15, minPts 300, geohash 30.
- Better time, but an increased number of bordering replicated points (i.e., BSOs)

Data size vs Accuracy & BSOs count



- Running times and number of BSOs of our retrofitted version of DBSCAN-MR over SCAP against SASAP-based using the ParticipAct dataset. Parameters: eps 0.15, minPts 300, geohash 30.
- Better time, but an increased number of bordering replicated points (i.e., BSOs)

Adaptability effect



- The effect of tweaking geohash precision on the number of BSOs generated by SCAP on NYC taxicab dataset. Parameters: eps 0.15, minPts 300, geohash 35
- Changing the geohash precision has a utility in determining the number of BSOs
- changing the tweakable geohash from 30 to 35 precision yields less BSOs for SCAP
- wider geohash precision implies a smaller size of the cells that this geohash order pass through, which then reduces the overlapping areas between bordering cells, thereby reducing the BSOs count.
- we obtain roughly 32% gain by changing the geohash precision.

Concluding remarks

- Communication service provisioning targets designing communication networks that serve as robust infrastructures for the management of huge amounts of data with QoS guarantees.
- Inappropriate configurations of parallel data processing frameworks may deteriorate the benefits we reap by the elasticity .
- Spatial data partitioning in Cloud computing frameworks is a determinant factor that should be prioritized.
- We have designed SCAP, a locality-preserving spatial partitioning scheme for quality spatial analytics in distributed main memory frameworks.
- Focusing on spatial locality problem to help in minimizing the data shuffling around the network while processing costly proximity-alike queries.
- As a future perspective, we could offload a portion of the data partitioning to IoT devices near the edge.
- Also, we plan to exploit porting parts of the spatial processing to Fog computing near the Edge, and possibly thereby sending only summaries to the Cloud for further processing, thus lowering the overall latency.

Q&A and Contacts

Thanks for your attention!

Questions time...

Isam Al Jawarneh
Email: isam.aljawarneh3@unibo.it

Luca Foschini
Email: luca.foschini@unibo.it

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI INFORMATICA - SCIENZA E INGEGNERIA

IEEE GLOBECOM 2020
7th - 11th December 2020