Interactive and Scalable Exploration of Geospatial Data

Jia Yu

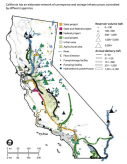
Advisor: Mohamed Sarwat

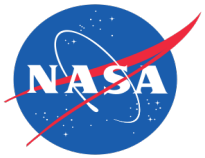
ASU Arizona State University

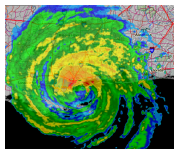
Social Media

1.19 billion monthly active users as of September 30, 2013

Scientific Data





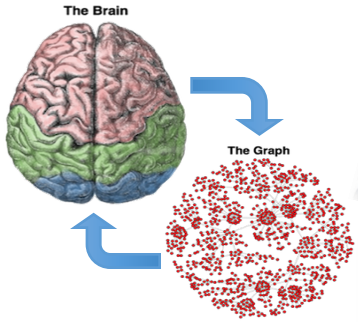


Urban Data

400GB of Road Network Data

UBER

Medical Data



85 Billion neurons in the brain nervous system

Mobile Devices

Big Spatial Data

- Iterative analysis (e.g. Spatial data mining)
- Interactive operation (e.g. Users have no time to wait)

- Two stages M R
- Intermediate data on disk

- DAG scheduler
- Intermediate data in memory

ASU ARIZONA STATE UNIVERSITY

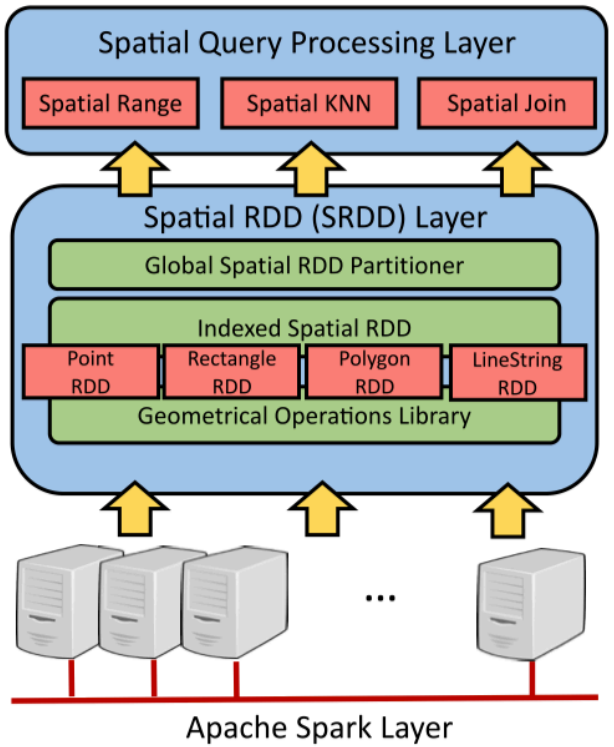
Outline

- Interactive and scalable geospatial data processing
- Interactive and scalable geospatial data visualization

Geospatial data processing

GeoSpark: A cluster computing framework for processing large-scale spatial data

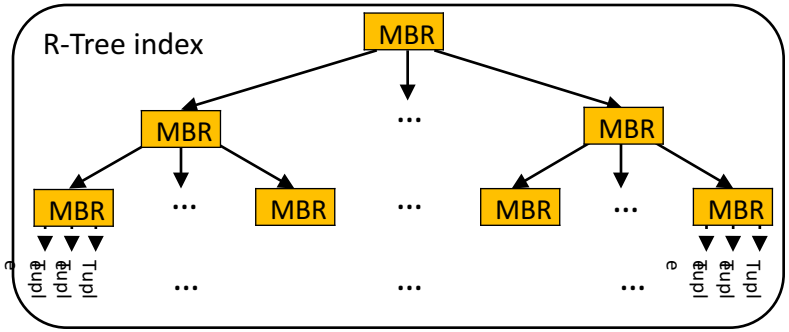
- Spatial RDD
- Spatial Index
- Spatial Data Partitioner
- Spatial Query

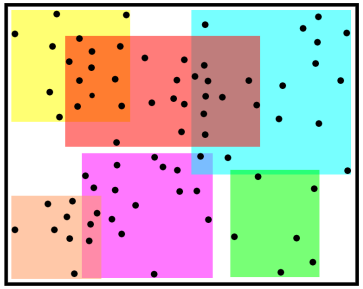


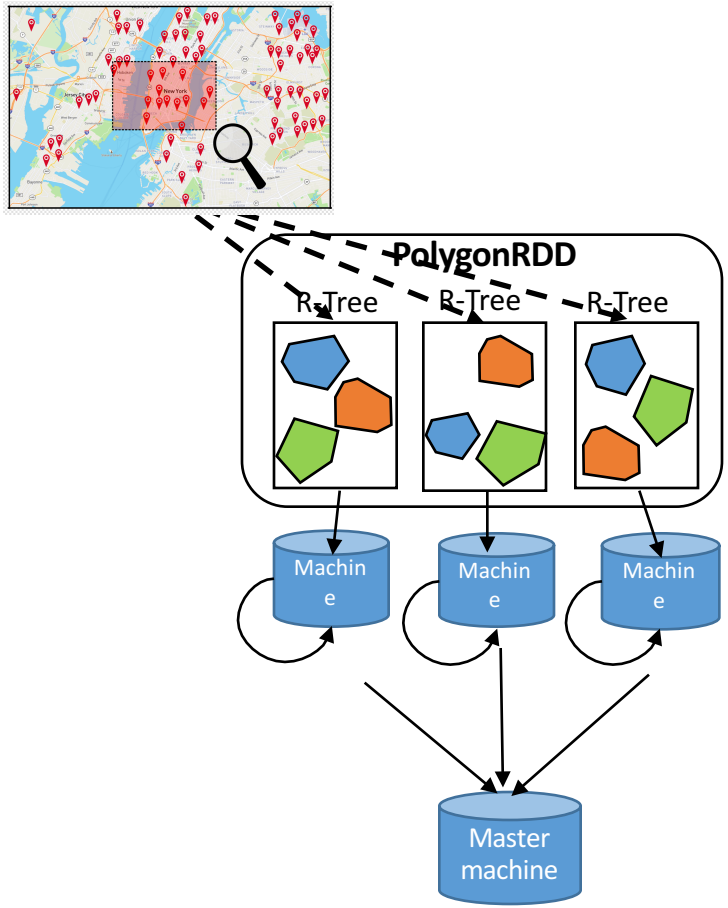
Spatial RDD

- Data source: Local disk, HDFS, Amazon S3…
- Data format: CSV (Comma-Separated Values), TSV, WKT(Well-Known Text), Shapefile, GeoJSON, HDF (NASA Earth data), User Supplied Format…
- Point RDD, Polygon RDD, LineString RDD, Hybrid geometries? Spatial RDD
- Built-in geometrical library: Minimum Bounding Rectangle, Polygon Union, Convex Hull…

Spatial Index

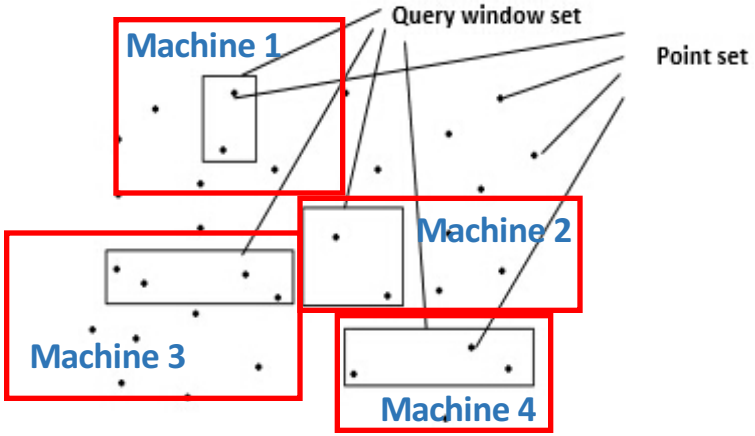






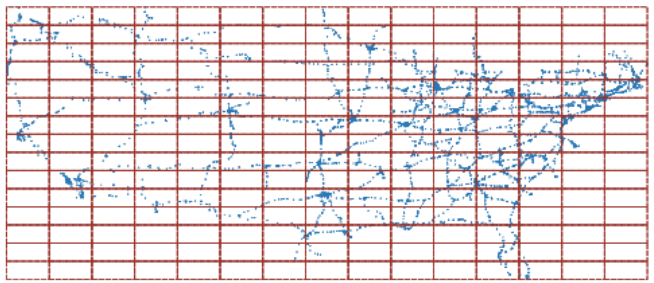
Spatial Data Partitioner

- Regular data partitioner doesn't work: Round-robin, Hash partitioning
- Load balance + spatial proximity



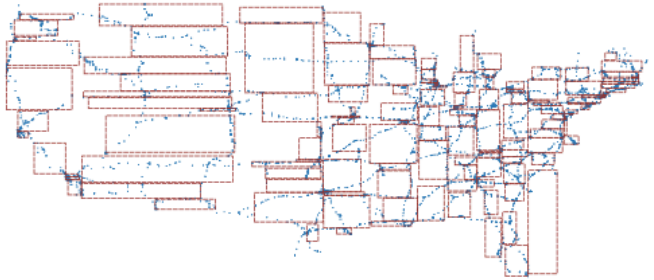
Spatial Data Partitioner (cont.)

- GeoSpark Spatial Partitioning Technique

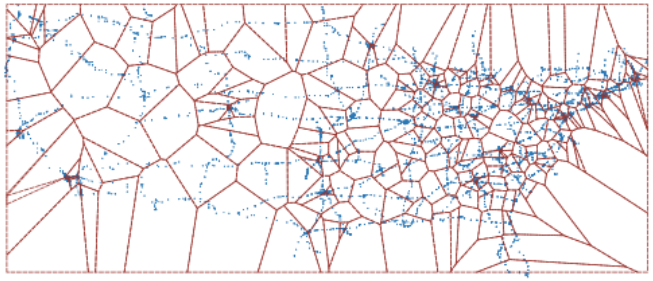


(a) SRDD partitioned by uniform grids

(b) SRDD partitioned by Quad-Tree



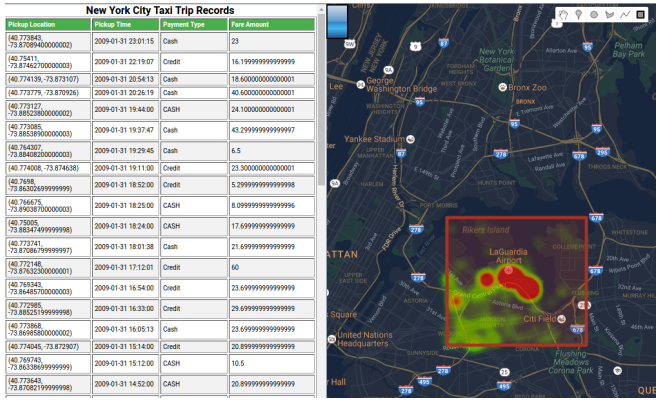
(c) SRDD partitioned by R-Tree



(d) SRDD partitioned by Voronoi diagram

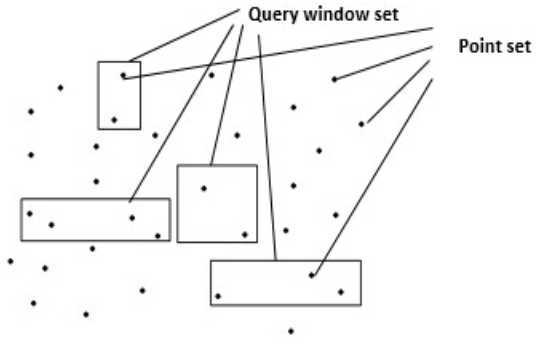
Spatial Queries

Spatial range query



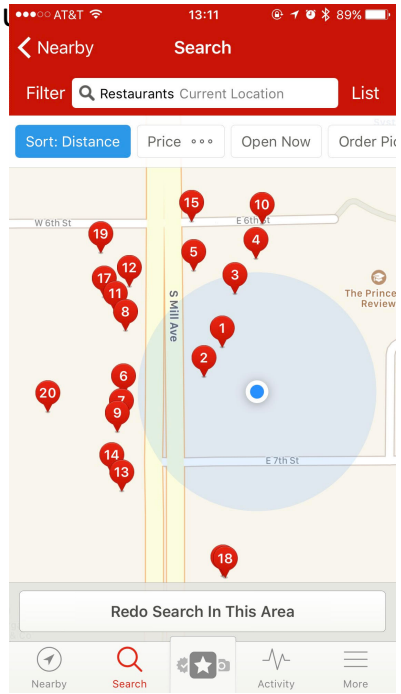
Return Restaurants WITHIN Tempe;

Spatial join query



Return restaurants WITHIN each US city;

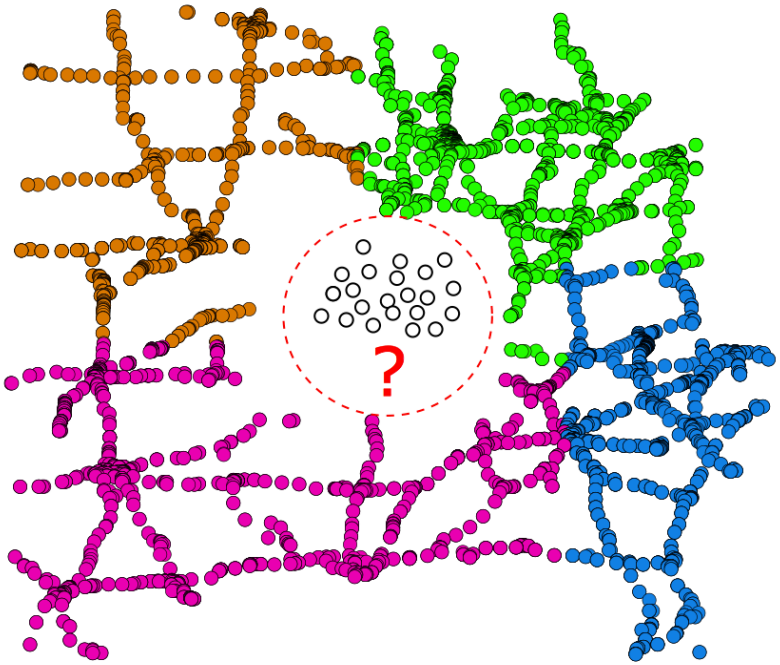
Spatial K nearest neighbor qu



Return 20 nearest restaurants;

ASU Arizona State University

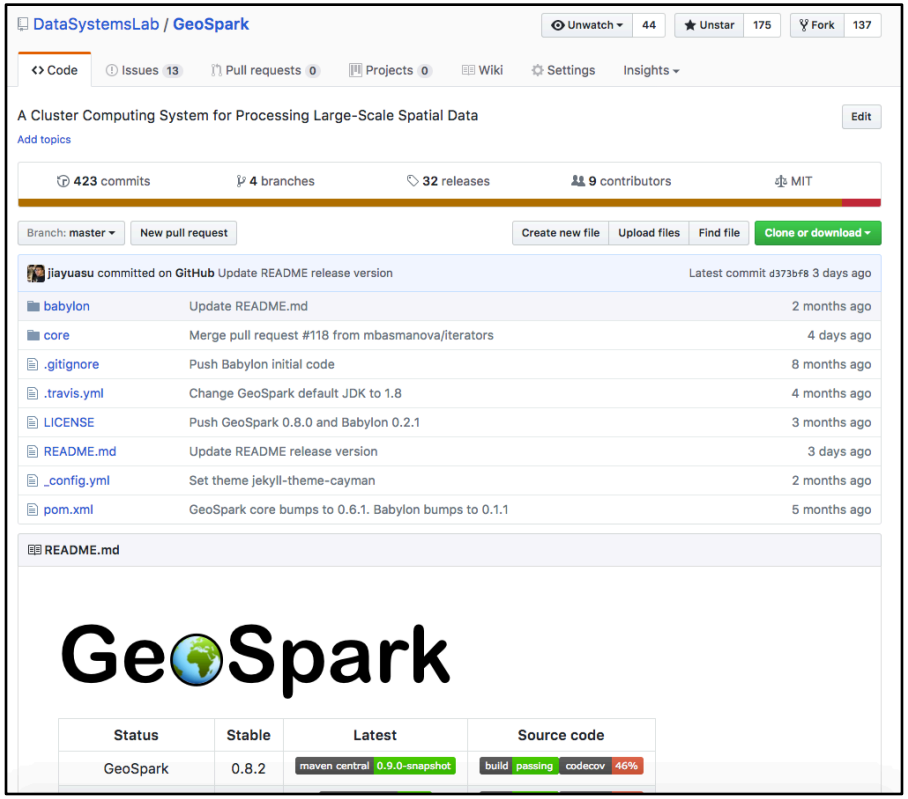
User Cases: Spatial Data Mining



KNN: Spatial Object Classification

Join: African Zebra and Lion Habitats Co-location

GeoSpark is open-sourced



DataSystemsLab / GeoSpark

Unwatch 44 | Unstar 175 | Fork 137

Code | Issues 13 | Pull requests 0 | Projects 0 | Wiki | Settings | Insights

A Cluster Computing System for Processing Large-Scale Spatial Data

Edit

Add topics

423 commits | 4 branches | 32 releases | 9 contributors | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

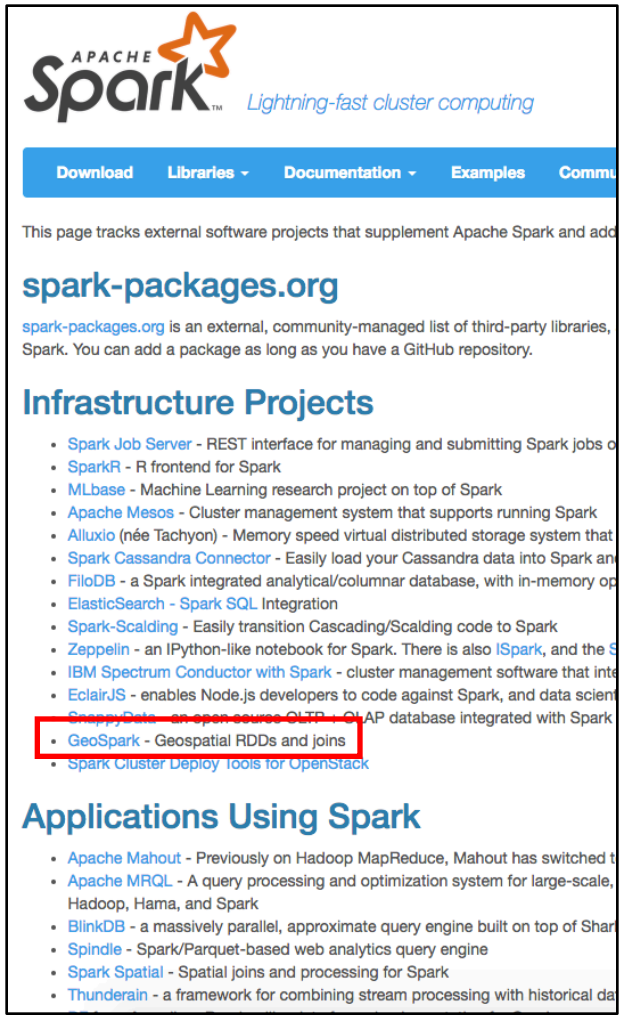
jiayuasu committed on GitHub Update README release version — Latest commit d373bf8 3 days ago

babylon	Update README.md	2 months ago
core	Merge pull request #118 from mbasmanova/iterators	4 days ago
.gitignore	Push Babylon initial code	8 months ago
.travis.yml	Change GeoSpark default JDK to 1.8	4 months ago
LICENSE	Push GeoSpark 0.8.0 and Babylon 0.2.1	3 months ago
README.md	Update README release version	3 days ago
_config.yml	Set theme jekyll-theme-cayman	2 months ago
pom.xml	GeoSpark core bumps to 0.6.1. Babylon bumps to 0.1.1	5 months ago

README.md

GeoSpark

Status	Stable	Latest	Source code
GeoSpark	0.8.2	maven central 0.9.0-snapshot	build passing codecov 46%



APACHE Spark — Lightning-fast cluster computing

Download | Libraries | Documentation | Examples | Commu

This page tracks external software projects that supplement Apache Spark and add

spark-packages.org

spark-packages.org is an external, community-managed list of third-party libraries, Spark. You can add a package as long as you have a GitHub repository.

Infrastructure Projects

- Spark Job Server - REST interface for managing and submitting Spark jobs o
- SparkR - R frontend for Spark
- MLbase - Machine Learning research project on top of Spark
- Apache Mesos - Cluster management system that supports running Spark
- Alluxio (née Tachyon) - Memory speed virtual distributed storage system that
- Spark Cassandra Connector - Easily load your Cassandra data into Spark an
- FiloDB - a Spark integrated analytical/columnar database, with in-memory op
- ElasticSearch - Spark SQL Integration
- Spark-Scalding - Easily transition Cascading/Scalding code to Spark
- Zeppelin - an IPython-like notebook for Spark. There is also ISpark, and the S
- IBM Spectrum Conductor with Spark - cluster management software that inte
- EclairJS - enables Node.js developers to code against Spark, and data scient
- SnappyData - an open source OLTP + OLAP database integrated with Spark
- GeoSpark - Geospatial RDDs and joins
- Spark Cluster Deploy Tools for OpenStack

Applications Using Spark

- Apache Mahout - Previously on Hadoop MapReduce, Mahout has switched t
- Apache MRQL - A query processing and optimization system for large-scale, Hadoop, Hama, and Spark
- BlinkDB - a massively parallel, approximate query engine built on top of Shar
- Spindle - Spark/Parquet-based web analytics query engine
- Spark Spatial - Spatial joins and processing for Spark
- Thunderain - a framework for combining stream processing with historical da

ASU Arizona State University

GeoSpark project is active

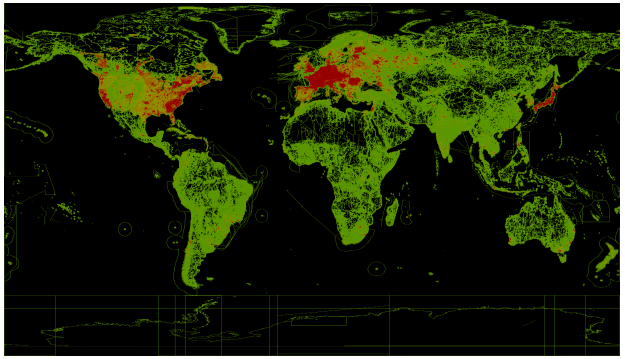
- GeoSpark contributors
  - ASU PhD and Master, Facebook, Europe
- GeoSpark is production-ready

…and many startup companies…

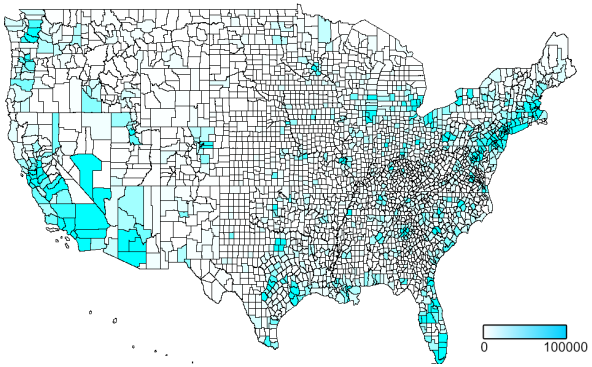
ASU Arizona State University

Geospatial data visualization

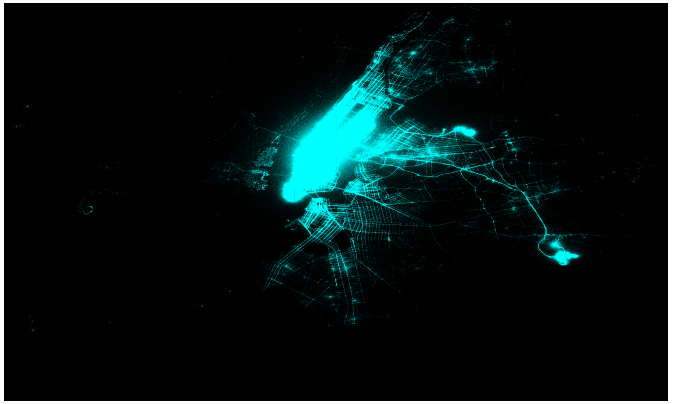
Visualization is useful



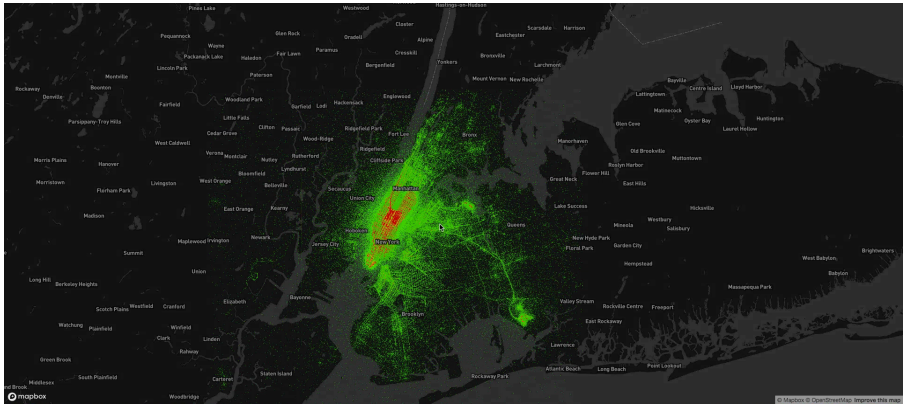
Road network heat map



Tweets per county



New York taxi pickup point dot map



New York taxi pickup point heat map

ASU ARIZONA STATE UNIVERSITY

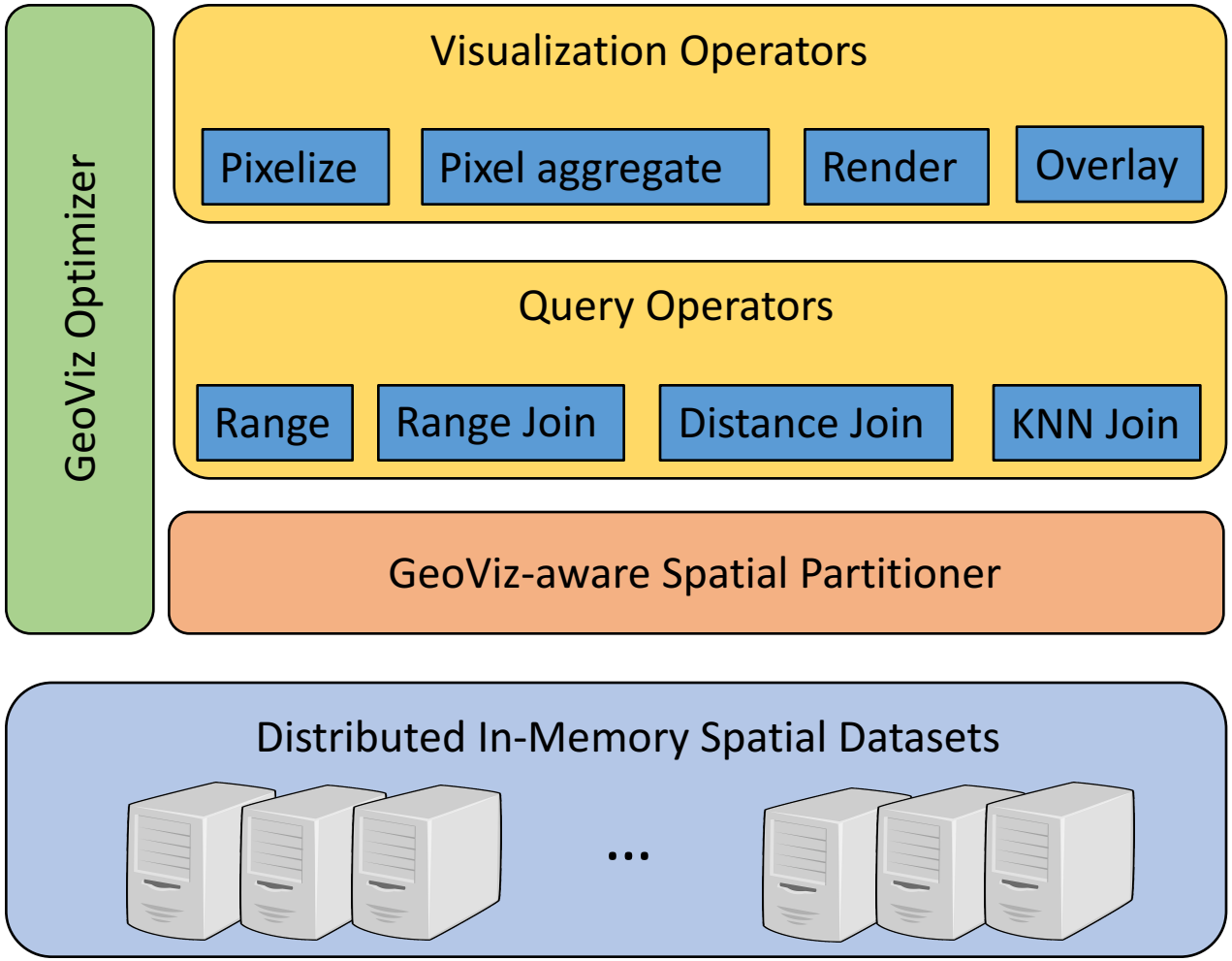
Motivation

- Geospatial visual analytics: two phases
  - Do data preparation FIRST:
    - Load / Parse data: HDFS, S3, PostgreSQL, disk file
    - Spatial Query: PostgreSQL, MySQL, Hadoop, Spark (GeoSpark)
  - Then do data visualization:
    - Google Map, MapBox
    - ArcGIS, QGIS
    - MapD
- Jump between data preparation and visualization
  - Painful, when geospatial data is big
- Always first data preparation, then visualization
  - Is it possible to co-optimize both phases, if I know my workload is geospatial visual analytics?

Babylon: a large-scale geospatial visual analytics platform

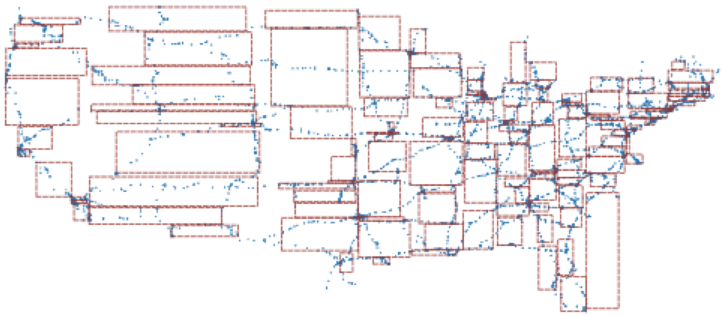
- One stage solution
  - Data preparation + visualization
- Support geospatial visual analytics (GeoViz)
- Co-optimize both phases to produce an efficient execution plan

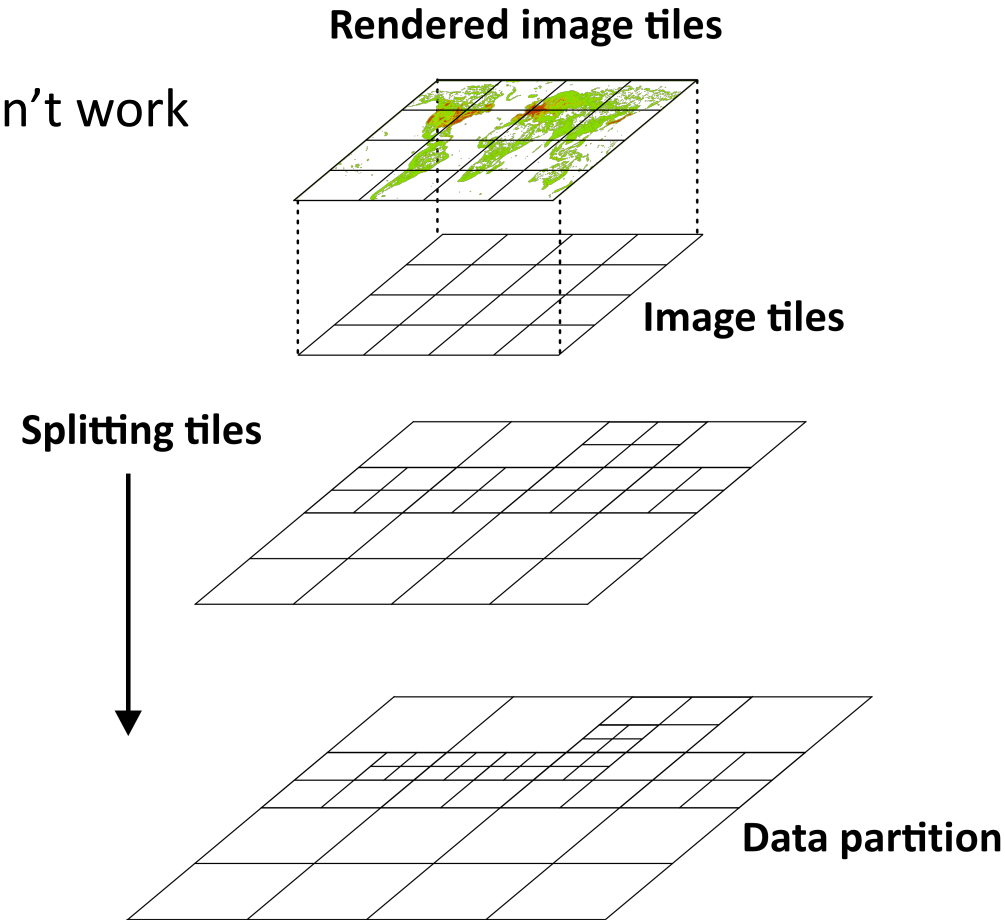
Babylon overview



GeoViz-aware spatial data partitioner

- Spatial data partitioner doesn't work
- Load-balance
- Spatial proximity
- Map tile compatible





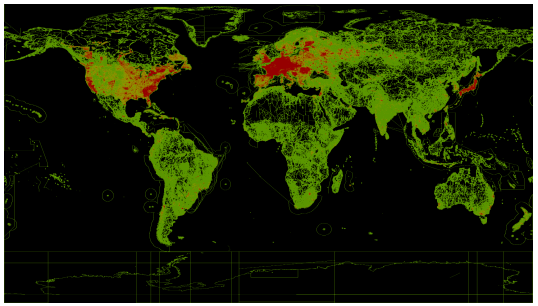
ASU Arizona State University

SQL Interface

- Design your GeoViz visualization effect

```
CREATE GEOVIZ [Name] ([Input] GEOMETRY)
RETURNS MAP

TILE [Tile quantity]
RESOLUTION [Map resolution]
RULE [Pixel aggregation rule]
COLOR [Color expression]
RETURN MAP
( RENDER PIXEL_AGGREGATE(*)
  FROM (PIXELIZE([Input])); )
```



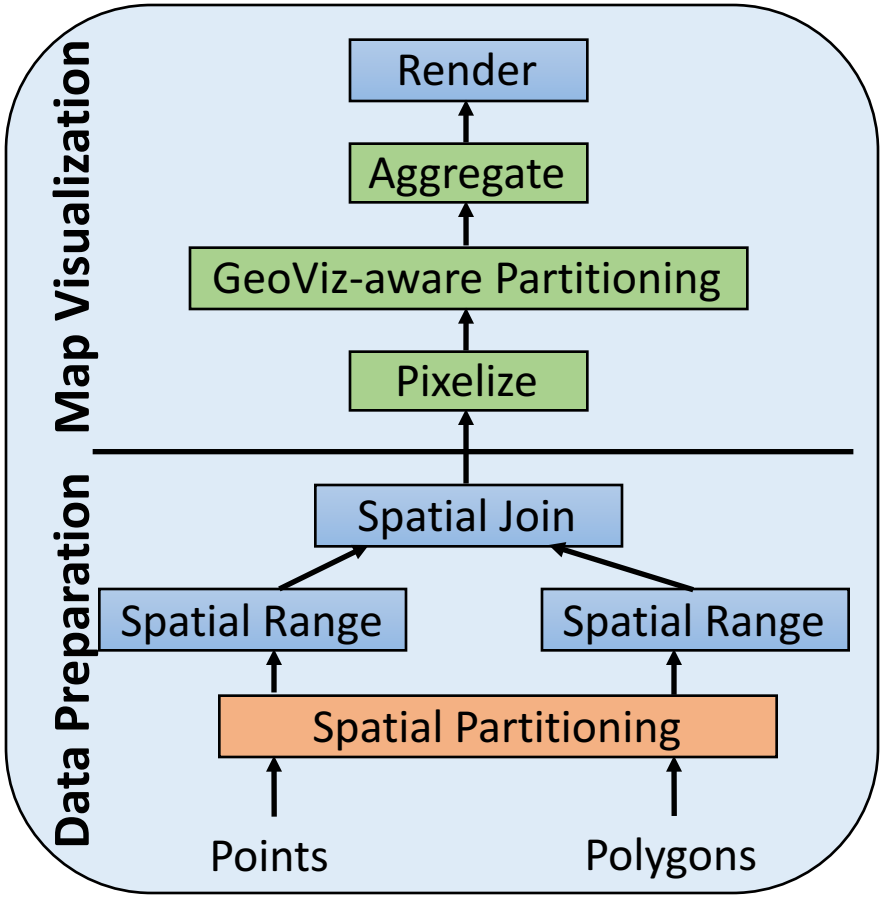
- Pass data to GeoViz function

```
SELECT [GeoViz name]([Table].[Attribute])
FROM [Table]
WHERE [Where clause]
```

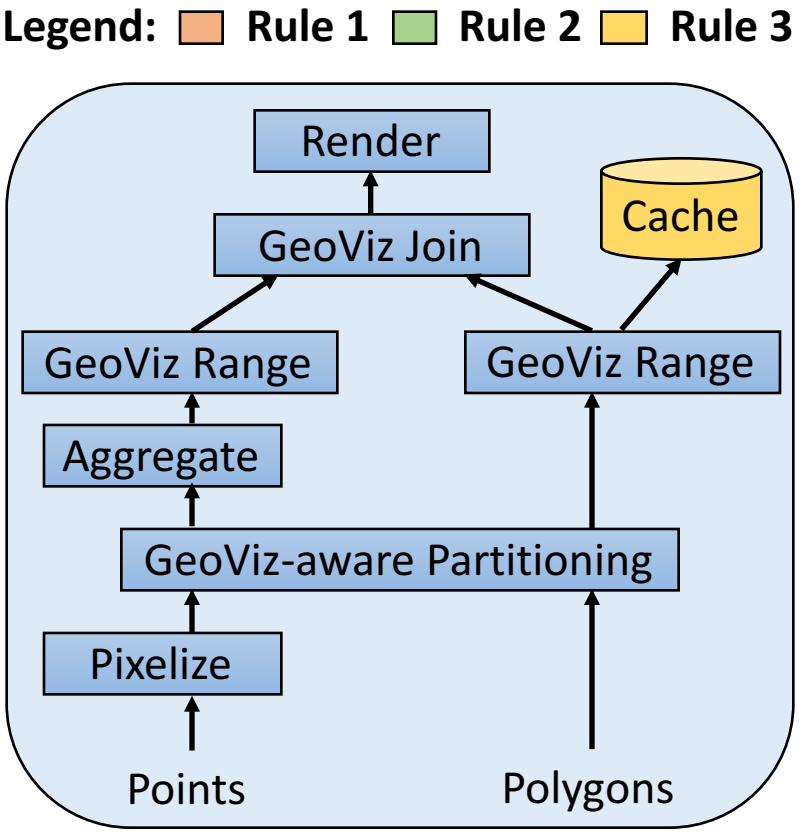
Babylon Optimizer

- Skip unnecessary steps for visualization purpose
- Estimate the cost of each operator
- Re-organize operators
  - Rule 1: Merge repeated operators together
  - Rule 2: Reduce dataset scale in advance
  - Rule 3: Cache frequently accessed datasets

Babylon Optimizer (cont.)



(a) Plan I: Non-optimized

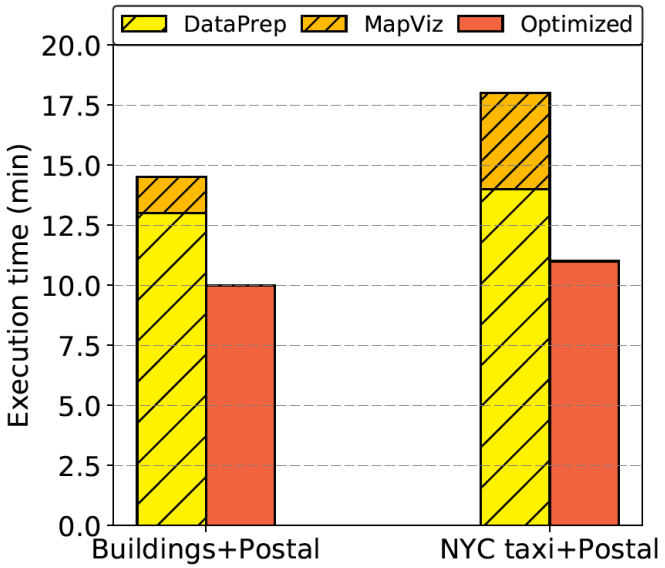


(b) Plan II: Optimized

ASU Arizona State University

Preliminary Result

- Traditional solution
  - GeoSpark + viz extension, no optimization
  - Babylon: co-optimize data preparation + visualization



ASU Arizona State University

Thank you!

ASU Arizona State University