Large Scale Analytics of Vector + Raster Big SPATIAL DATA

CONTENT

- ZONAL STATISTICS PROBLEM
- VECTORIZATION METHOD
- RASTERIZATION METHOD
- SCANLINE METHOD (NOVEL APPROACH)
- EXPERIMENTS AND RESULTS
- CONCLUSION

INTRODUCTION

- NASA EOSDIS provided public access to 17 petabytes (1 TB= 0.001PB) of earth observational data.
- SENTINEL-1A Satellite launched by European Space Agency collected 5 petabytes in 2 years.
- EUROPIAN XFEL projects collects X-ray images of atoms at 10 petabytes /year

- Process raster data and vector data.
- Poor performance when both are combined.

Hadoop-GIS
Spatial Big Data Solutions

GeoSpark

Simba

SciDB

rasdaman
raster data manager

GeoTrellis

- Zonal statistics problem- It aggregates all the values from raster layer that overlap with a set of polygons. Example:

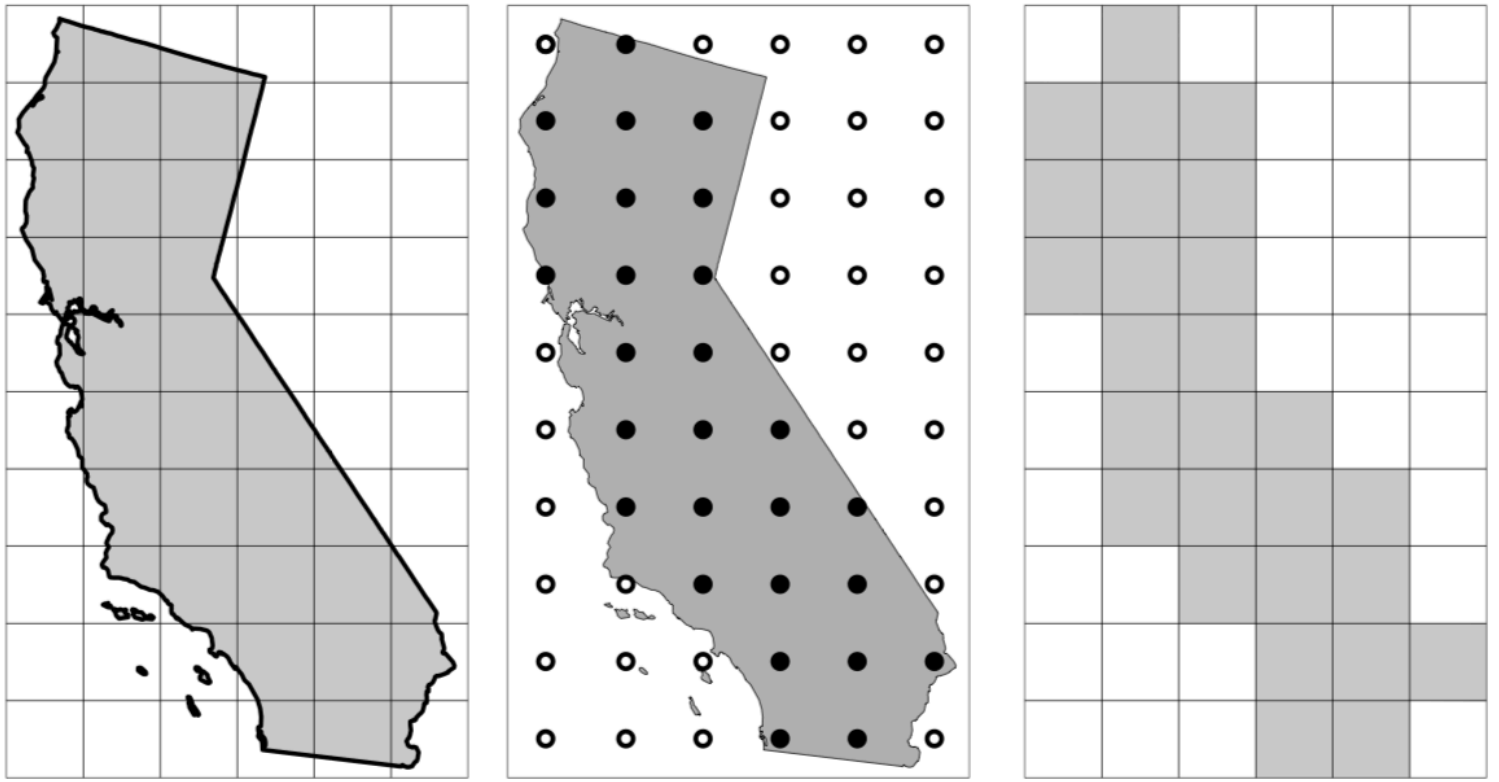
  - Computes the average temperature for each state in the US.
  - Effect of vegetation and temperature on human settlement.
  - Analyzing terabytes of socio-economic and environmental data.

- Input to the problem includes-
  - raster layer ‘r’ - maps each matrix entry into geographic location.
  - vector layer ‘v’ consists set of disjoint polygons represented as points each defined by a geographic location.
  - accumulator ‘acc’ takes pixel values, and computes the statistics of interest.

- Output is a value for the accumulator for each polygon in the vector layer.

EXAMPLE

- r represents the temperature in the world,
- v represents the 50 US states
- acc is an average accumulator
- Output of this problem will be the average temperature for each state in the US.



(a) Input (b) Vectorize (c) Rasterize

VECTORIZATION METHOD

- Vectorize step- raster layer from the raster representation ⟹ vector representation.

  -Converts each pixel to a point and the point location is determined by G2W mapping.

- Spatial join - inner-join operation on polygons and vectorized points

  -produce polygon-point pairs for each point that lies inside a polygon.

- Grouped aggregate- groups all pairs by polygon id

  -runs the user-defined accumulator on each group.

RASTERIZATION METHOD

- Rasterize- polygon filling algorithm

  - mask layer

  - Pixels inside the polygon have values of 1 and outside has 0.

- Clip- removes all pixels in the data layer that corresponds to 0 of the mask layer.

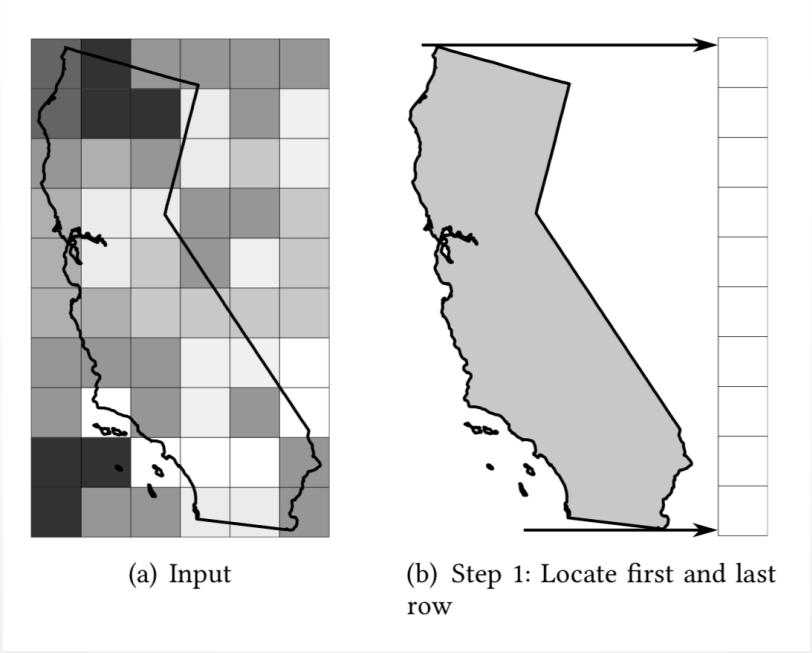
- Aggregate - computes the desired aggregate function, e.g., average.

DRAWBACK

- Vectorize- Resolution of the raster layer ↑
  - ↑ points in vectorization step
  - ↑ point-in-polygon tests will occur on the spatial join step.
  - With billions of pixels, this method takes excessively long time.

- Rastor- Size of the mask layer grows ↑ for high-resolution data
  -requires extra storage and processing.

SCANLINE METHOD



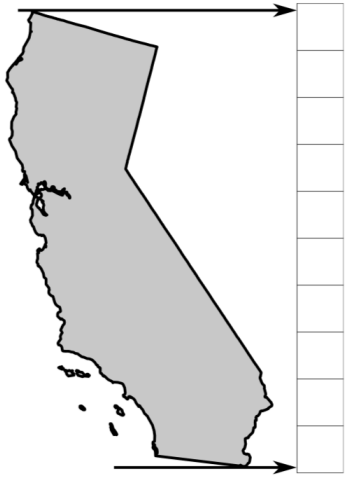
(a) Input

(b) Step 1: Locate first and last row

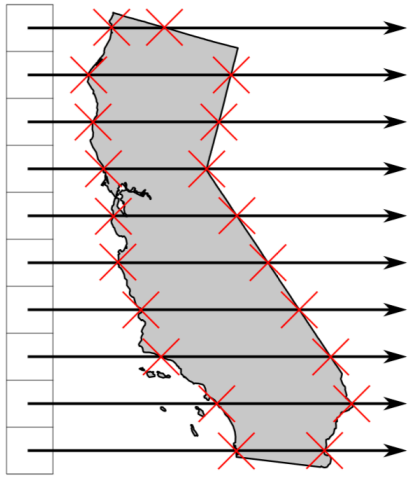
- Step 1- Polygon boundaries are scanned to locate the lowest and highest points.

  -Two points are mapped to the raster layer to locate range of rows.

SCANLINE METHOD



(b) Step 1: Locate first and last row



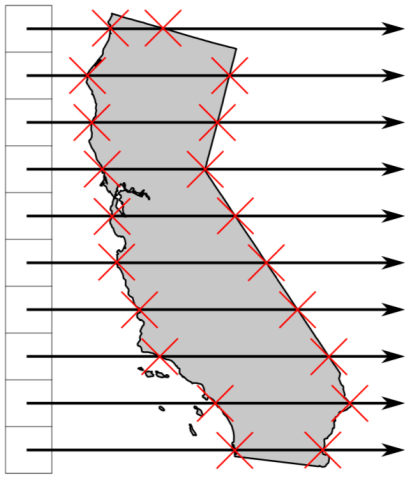
(c) Step 2: Find intersections with scan lines

- Step 2 - runs scanlines from the center of each pixel

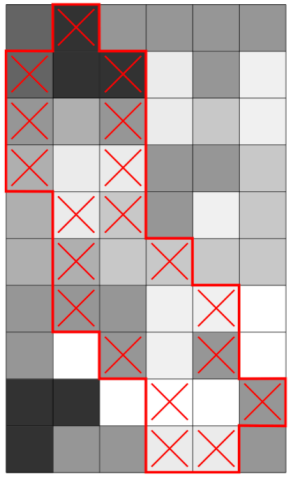
  -finds the intersections of each scanline with the polygon boundary.

  -Intersections then are sorted by their x-coordinates in the increasing order.

SCANLINE METHOD



(c) Step 2: Find intersections with scan lines



(d) Step 3: Process the pixels

- Step 3 - maps each intersection to a pixel in the raster layer.

  - Fill all those pairs of coordinates that are inside polygons

  -ignore alternate pairs.

ADVANTAGES

- Processes inputs in raw format and does not require any conversion from raster to vector making it much faster.
- Preprocessing step can be avoided.
- Requires a minimal intermediate storage for the intersection points.
- Accesses the pixels inside the polygon which improves Disk IO for very large raster layers.

EXPERIMENTS

Raster datasets

Dataset	Resolution in Pixels	File Size
glc2000	40,320×16,353	629 MB
MERIS	129,600×64,800	7.8 GB
US-Aster	208,136×89,662	35 GB
Tree cover	1,296,036×648,018	782 GB

Vector datasets

Dataset	Polygons	Segments	$\frac{\#segments}{\#polygons}$	File Size
Counties	3,108	51,638	17	978 KB
States	49	165,186	3,370	2.6 MB
World	284	3,817,412	13,440	60 MB

- GLC2000 and MERIS 2005 datasets from European Space Agency, US-Aster from Shuttle Radar Topography Mission (SRTM) and Hansen developed the global Tree Cover change dataset which covers the entire globe was used as for the experiments.
- 3000 US continental counties, 49 US continental features and 249 national boundaries are represented using vector dataset.

SETUP

- Intel Xeon E3-1220 v5 3.00ghz Quadcore processor, 64 GB of RAM, and a 2 TB HDD running Ubuntu 16.04.2 and oracle java 1.8.0_102.
- Open source Geotools library 17.0.
- Rasterization method is implemented using PostgreSQL 9.3 and PostGIS 2.3.2.
- Four aggregate values are minimum, maximum, sum, and count

RESULTS

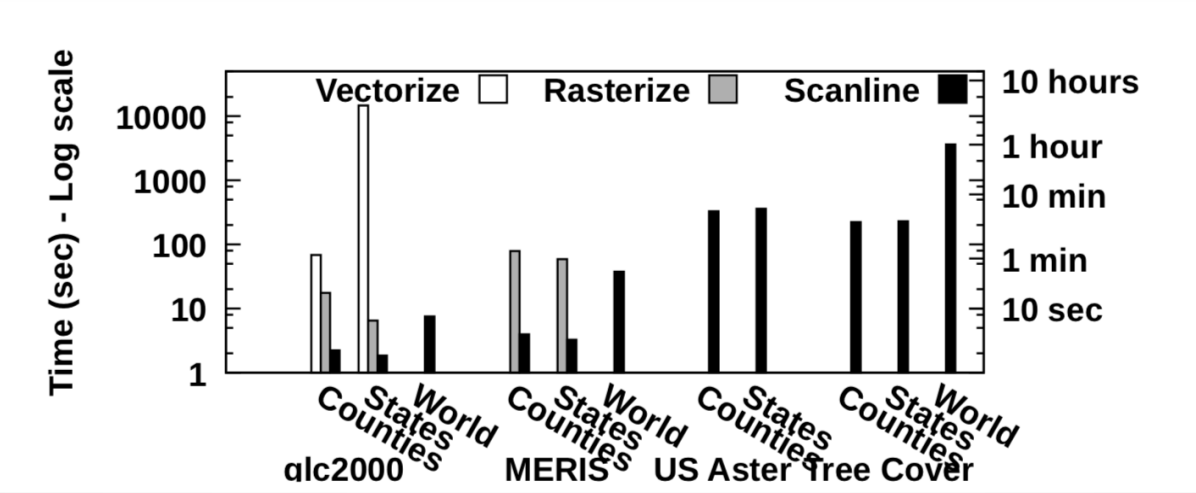


Figure: Overall running time comparison for three methods.

- Faster and process the high-resolution datasets, US Aster and Tree Cover.
- US Counties and States have same performance. The reason is that the running time of the scanline algorithm is dominated by disk IO while the CPU overhead is negligible. This can be improved by
  - Preprocessing the raster file to provide partial aggregates for big chunks that can minimize disk IO.
  - Running in a distributed environment where multiple disks can be accessed simultaneously.

RESULTS

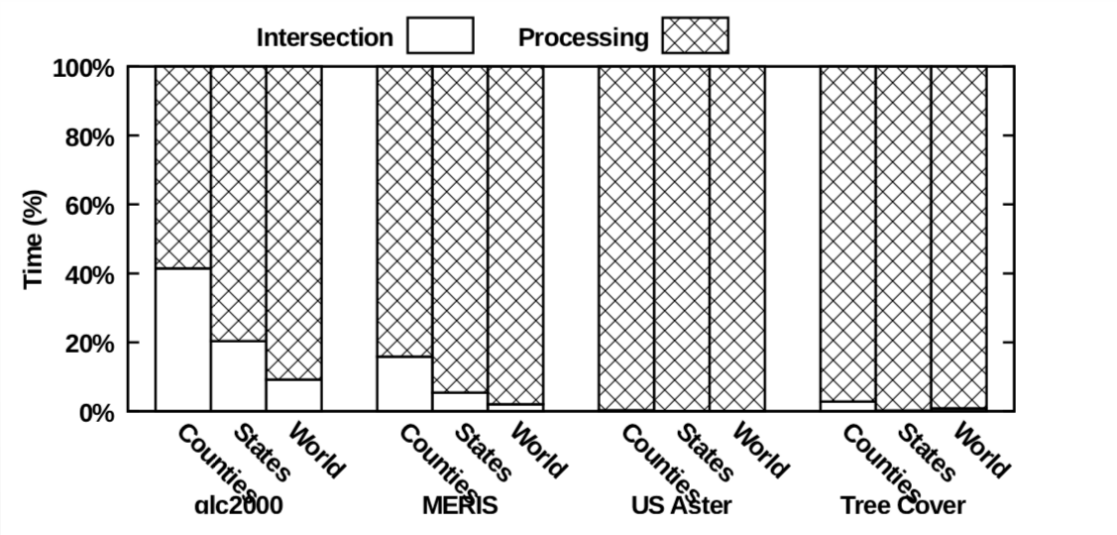


Figure: Scanline method running time breakdown in two phases.

- The intersection phase contains all the processing, except for reading the pixels from the raster file.
- The processing phase reads the pixel values and accumulates them.
- All numbers are normalized to the overall running time for readability.
- Size of the raster layer ⇑, the processing phase dominates the running time. This makes scanline optimal as it is dominated by disk IO for reading only the pixels that need to be processed.

CONCLUSION

- Zonal statistics problem.
- The two baseline methods provide a subpar performance.
- A novel method, termed scanline, which processes inputs in raw format.
- Future work- Using parallel and distributed algorithms.