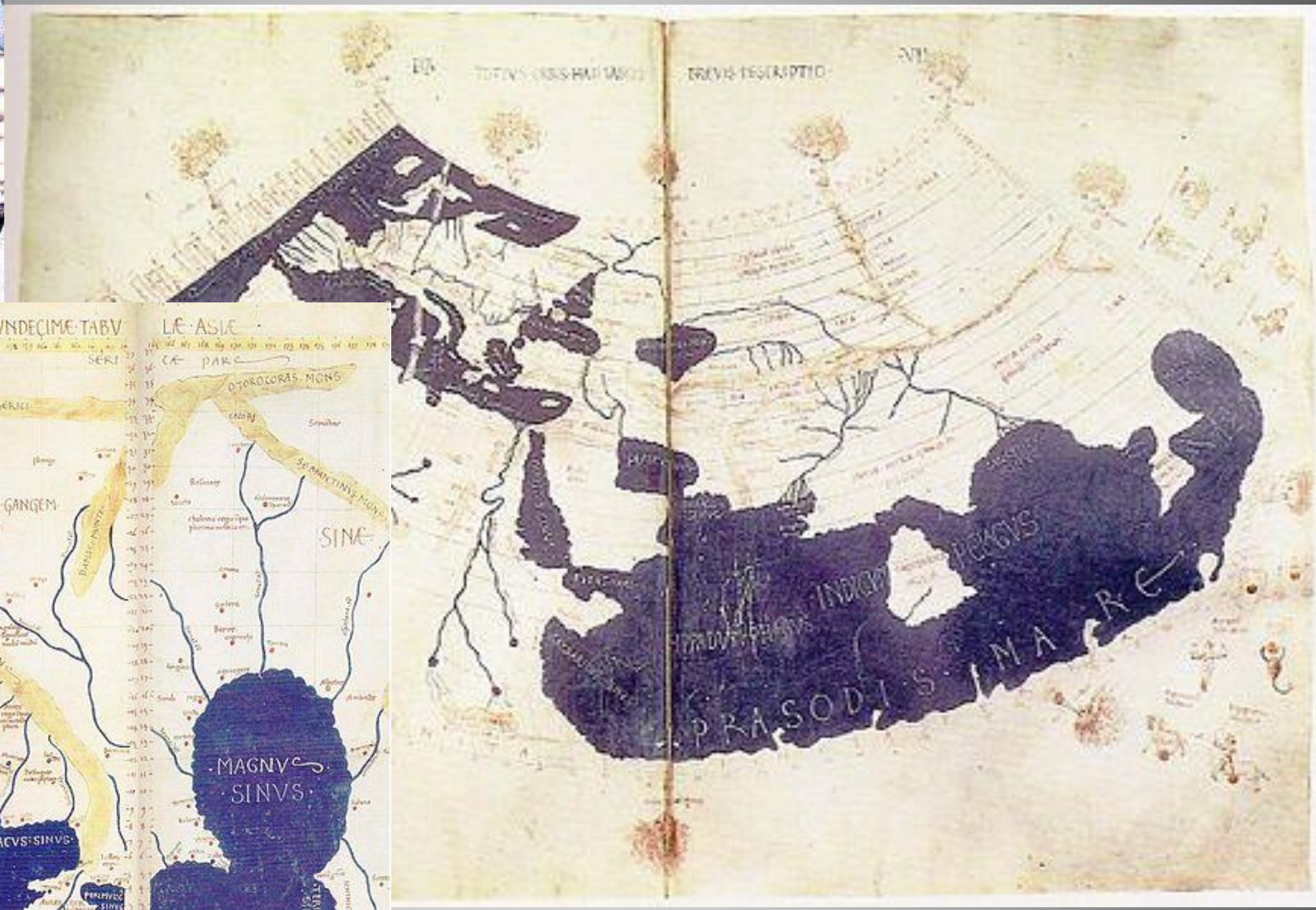
The Era of Big Spatial Data

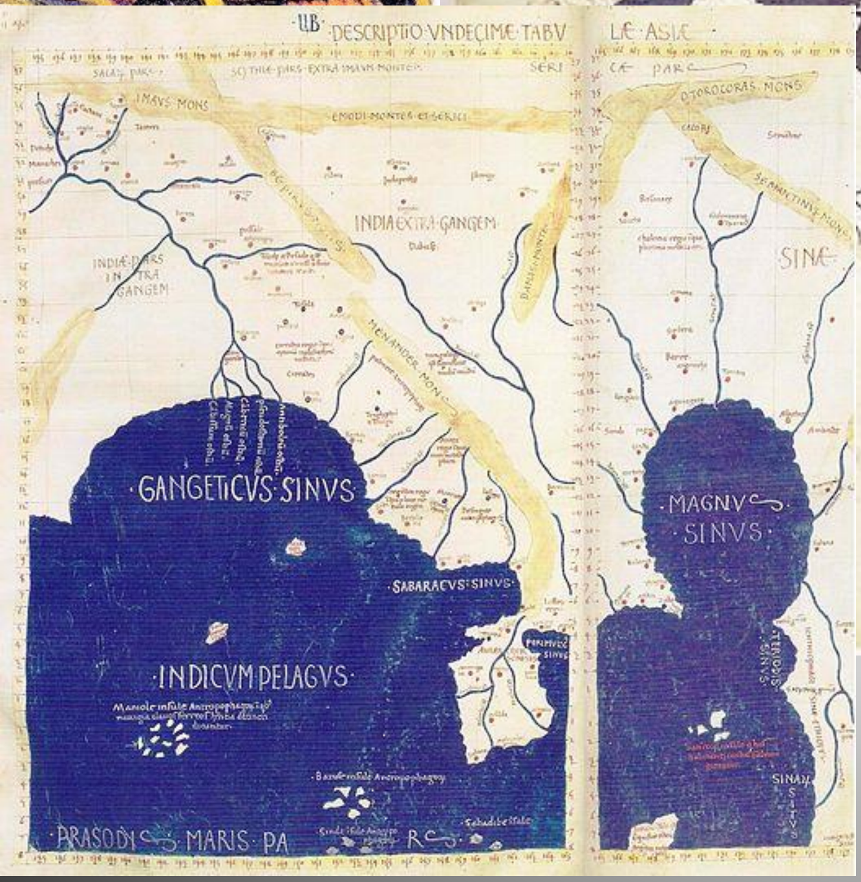
Ahmed Eldawy
University of California, Riverside

Mohamed Mokbel
University of Minnesota

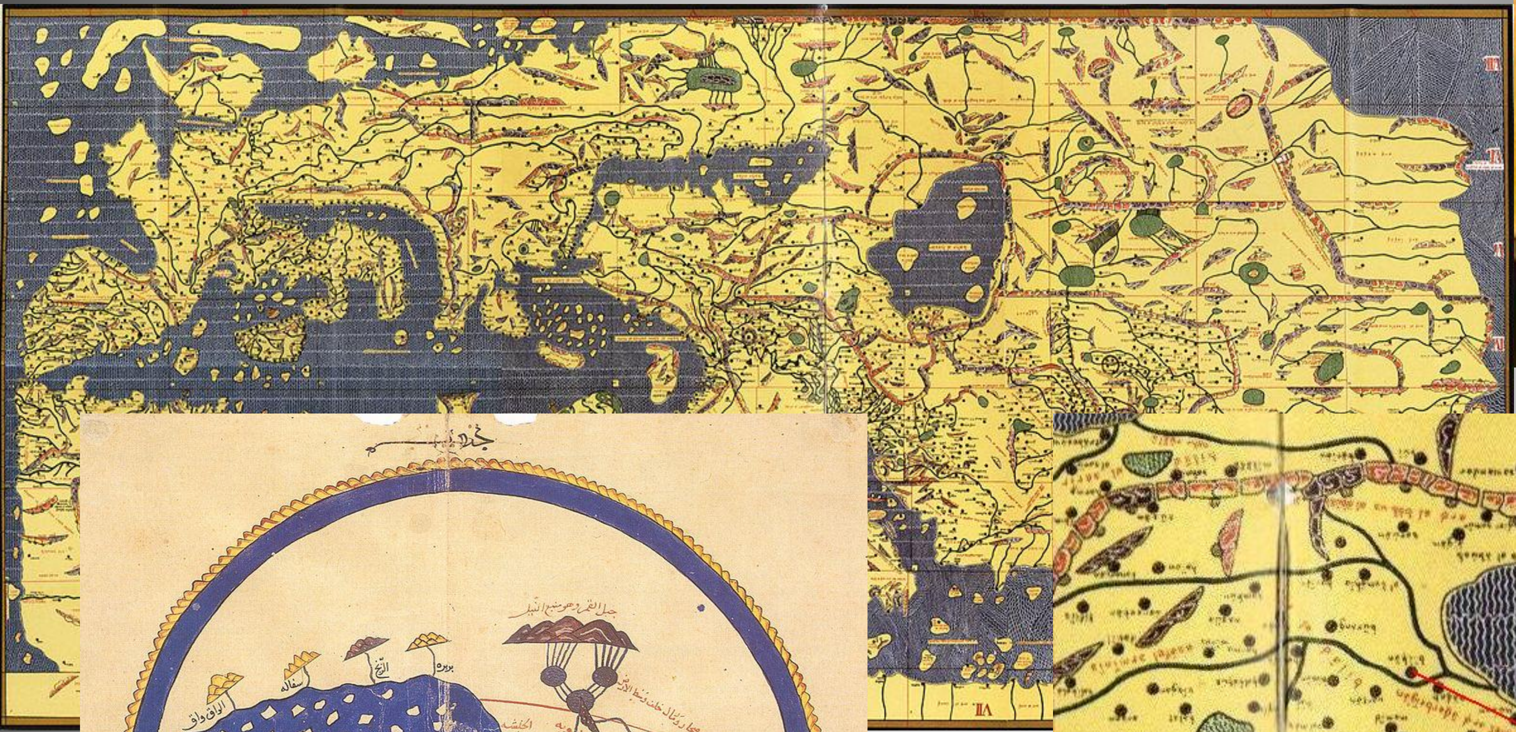

Once upon a time...

Claudius Ptolemy (AD 90 – AD 168)

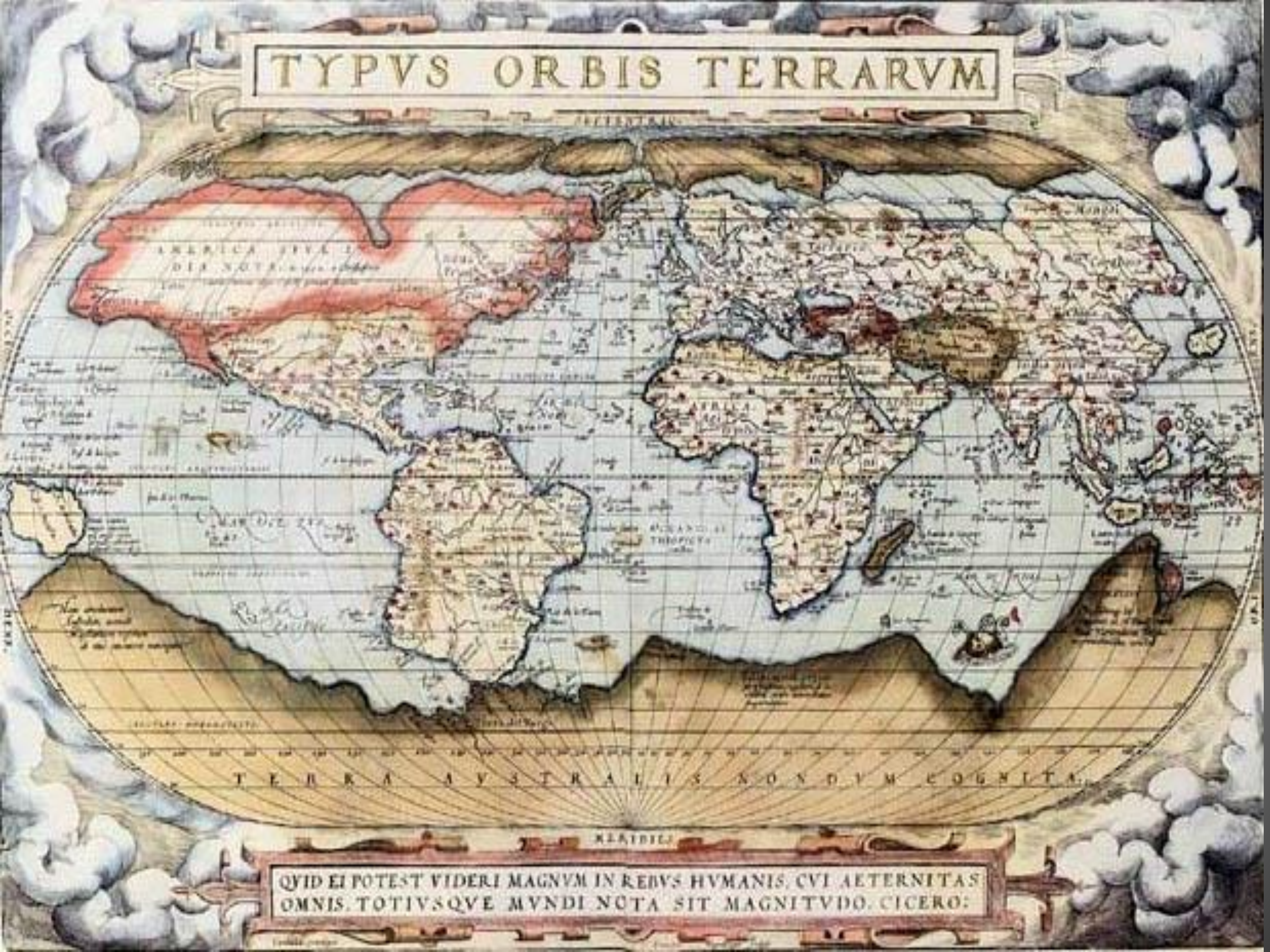




Al Idrisi (1099–1165)

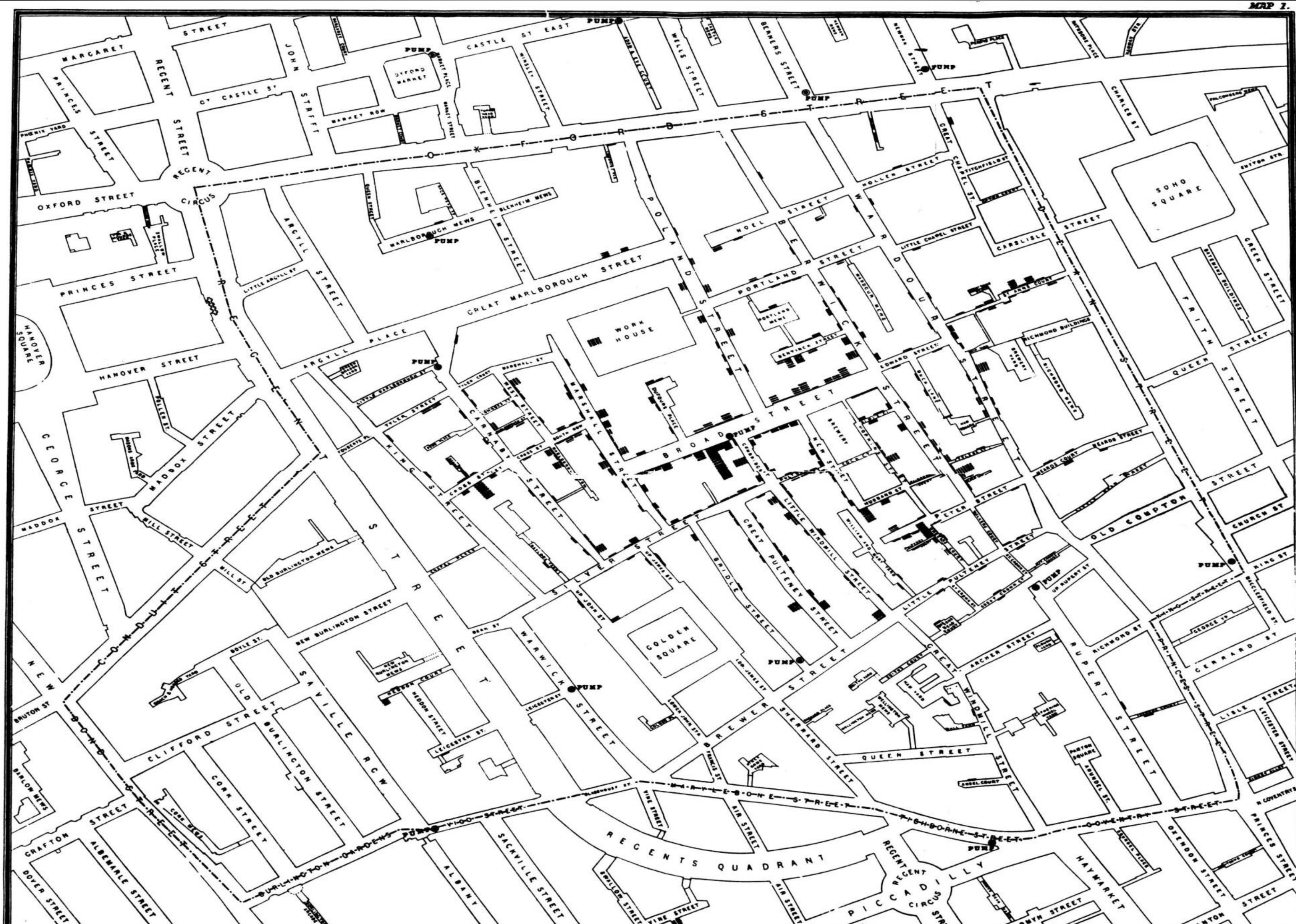


Mar Atlantico
EVROPA
AFRICA
INDIA
Tropicus Cancri
Tropicus Capricorni
Mar Indico
Equinoctialis
Brasil
Mar do Sul
Mar Pacifico
Mar Meridional


TYPVS ORBIS TERRARVM
AMERICA SIVE INDIA NOVA
TERRA AVSTRALIS NONDVM COGNITA
QVID EI POTEST VIDERI MAGNVM IN REBVS HVMANIS, CVI AETERNITAS OMNIS, TOTIVSQVE MVNDI NOTA SIT MAGNITVDO. CICERO:


ARGONAVTICA.
EVROPAE PARS.
SAVROMATAE.
SCYTHAE.
ALANI.
MAEOTIS PALVS.
PARS.
AXENVM Aequor, Iasonio pulsatum remige primum.
CAVCASEVM MARE.
CELTAE.
THRACIA.
CAPPA DOCIA.
ASIAE
SARDOVM PELAGVS.
TYRRHENVM AEQVOR.
IONIVM MARE.
PELOPIS REGIO.
LIBYSTICVM MARE.
Ex conatibus Geographicis Abrah. Ortelij Antuerp.
ILLVSTRISSIMO PRINCIPI CAROLO COMITI ARENBERGIO, BARONI SEPTIMONTII, DOMINO MIRVARTII, EQVITI AVREI VELLERIS, ETC.
Syrtes.
ATLANTICVS Ager.
Hesperides.
LIBYAE PARS.
MERIDIES.
THRACIAE PARS.
PROPONTIS.
BITHYNIA.
Myrtoum pelagus.

Cholera cases in the London epidemic of 1854



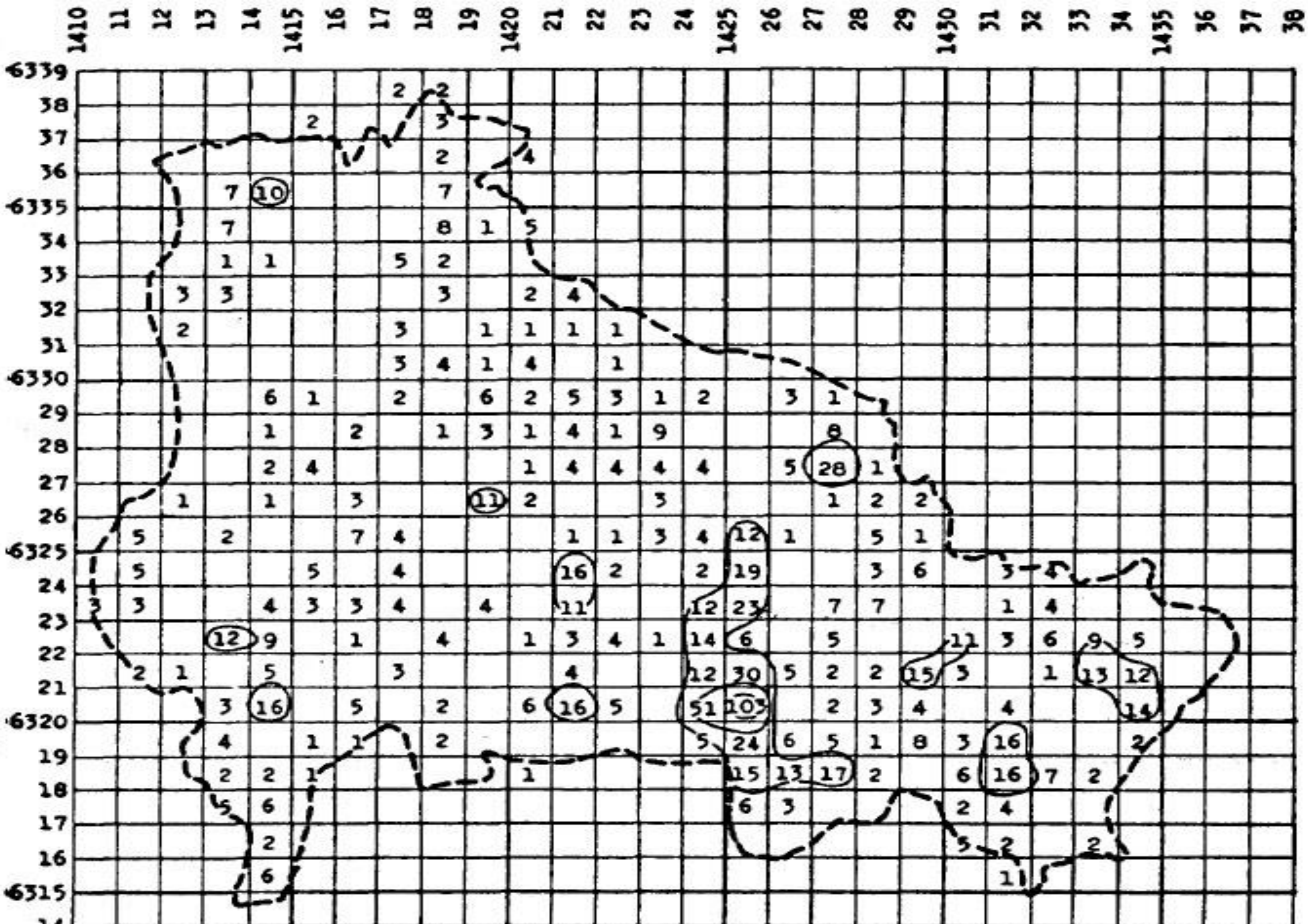


FIGURE 3—Children under 15 years of age in 1940.

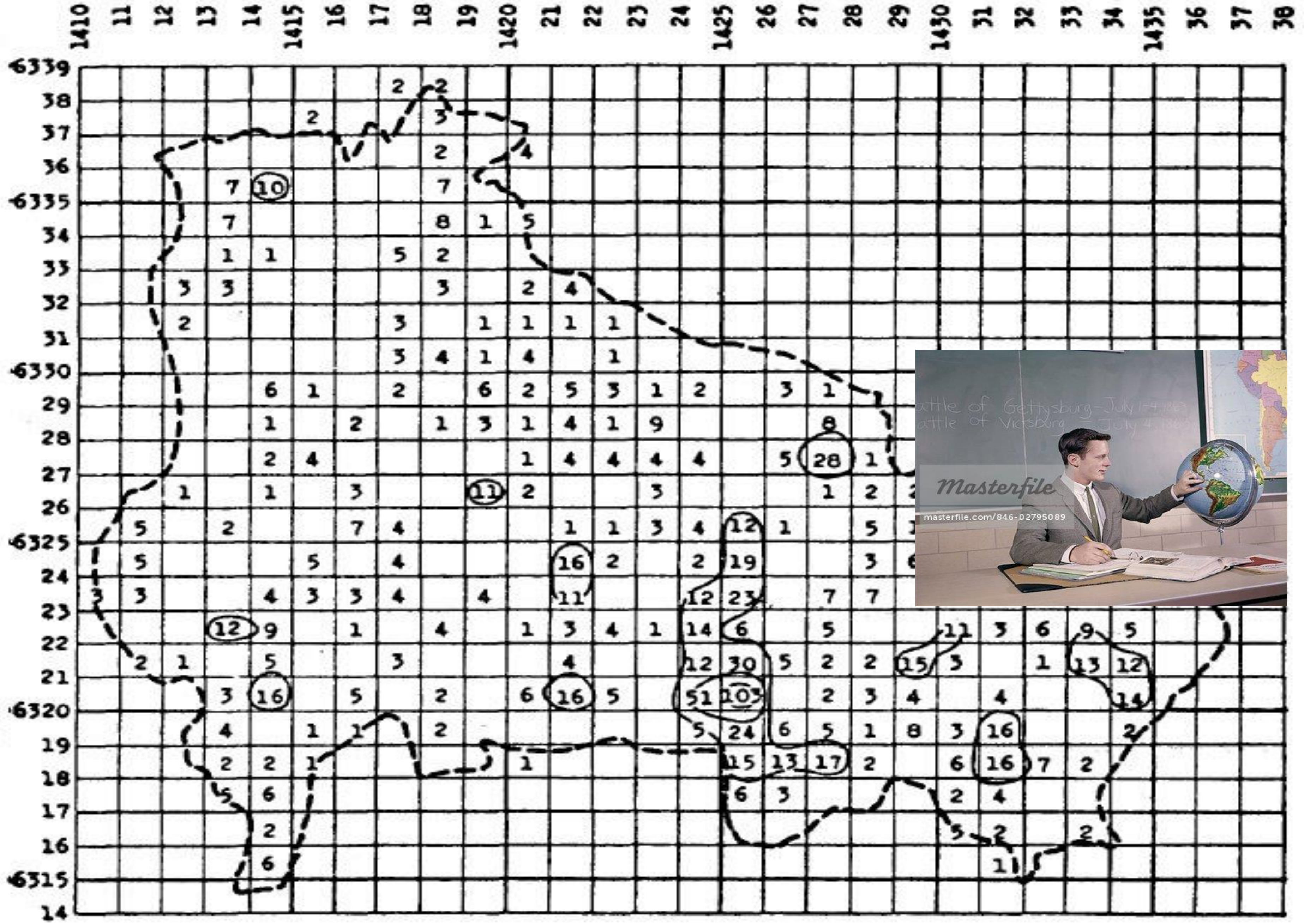
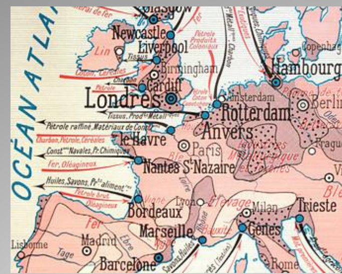
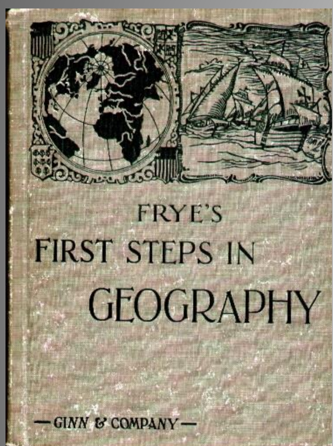
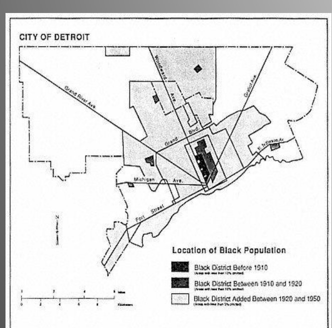


FIGURE 3—Children under 15 years of age in 1940.


UAR
5 M

OCÉAN ATL
Newcastle
Liverpool
Birmingham
Cardiff
Hambourg
Londres
Rotterdam
Anvers
Berlin
le Havre
Paris
Nantes St Nazaire
Lyon
Bordeaux
Marseille
Milan
Gênes
Trieste
Madrid
Barcelone
Rome

FRYE'S
FIRST STEPS IN
GEOGRAPHY
— GINN & COMPANY —

Masterfile
masterfile.com/846-02795089

CITY OF DETROIT
Location of Black Population
Black District Before 1910
Black District Between 1910 and 1920
Black District Added Between 1920 and 1950

Masterfile
masterfile.com/846-02793618

UNIVAC


IBM



Cool computer technology..!! Can I use it in my application
I have BIG data. I need HELP..!!
Oh..!! But, it is not made for me. Can't make use of it as is
My pleasure. Here it is.
CITY OF DETROIT
Location of Black Population
IBM


Masterfile
masterfile.com/846-02795089

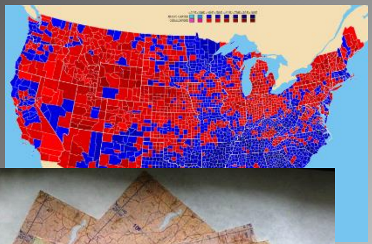

IBM

1969

Kindly let me understand your needs

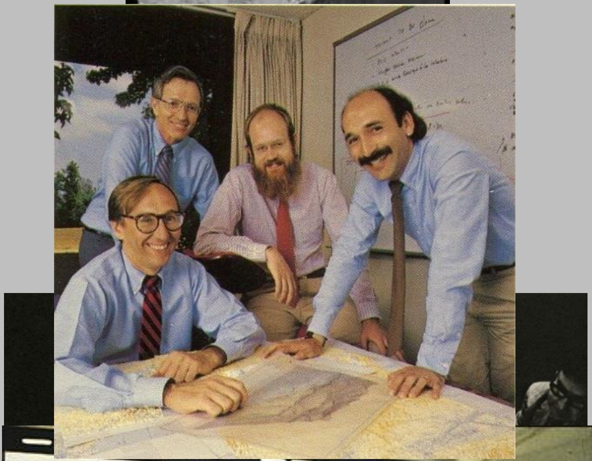
Kindly let me get the technology you have

ESRI

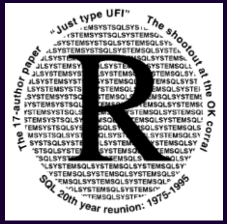









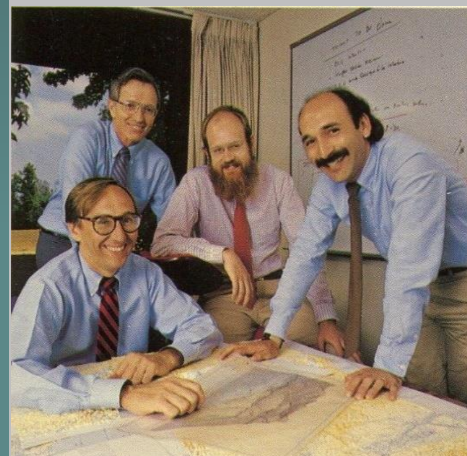
ESRI



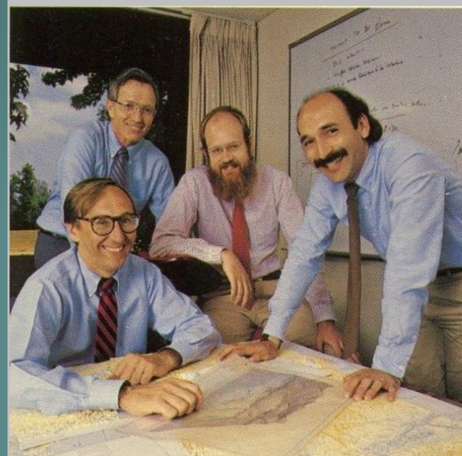

PostgreSQL

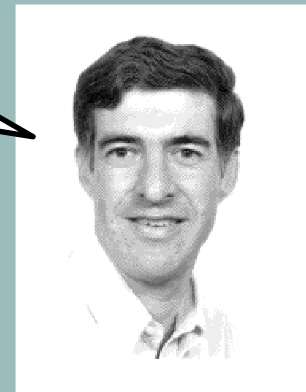
Informix


HELP..!! I have BIG data. Your technology is not helping me
mmm…Let me check with my good friends there.
Cool Database technology..!! Can I use it in my application?
Oh..!! But, it is not made for me. Can't make use of it as is
My pleasure. Here it is.
ESRI
PostgreSQL
Informix
ORACLE
AIRLINE BOOKING
Banking

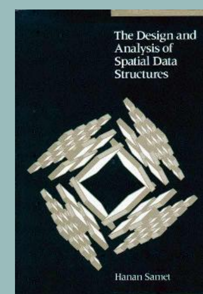


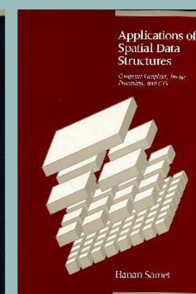
Kindly let me understand your needs



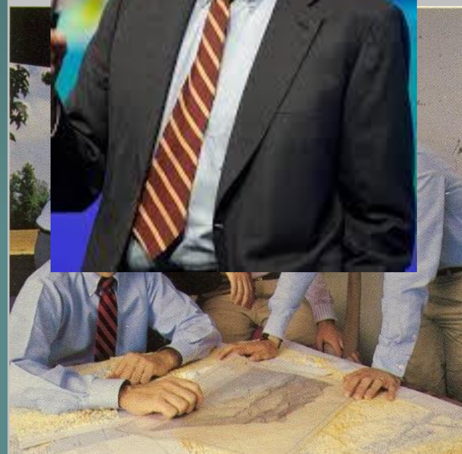


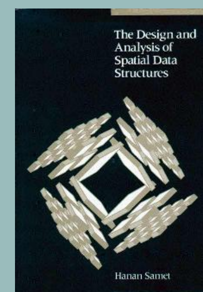
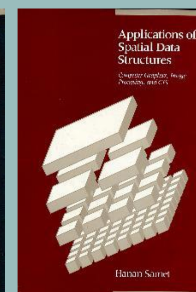
Kindly let me get the technology you have

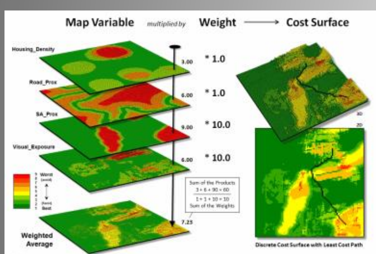


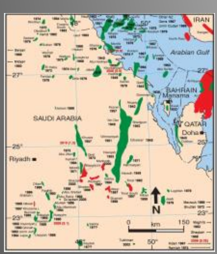


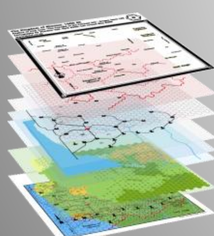


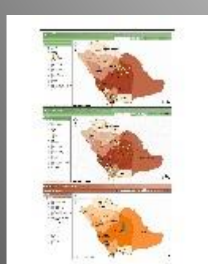




The Design and Analysis of Spatial Data Structures
Hanan Samet

Applications of Spatial Data Structures
Hanan Samet

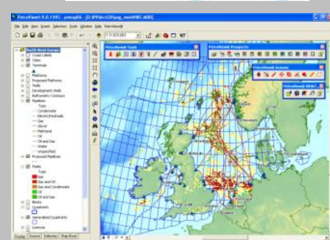


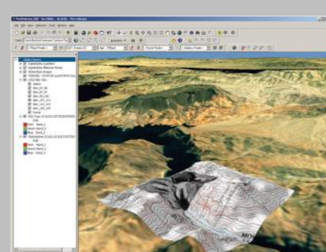






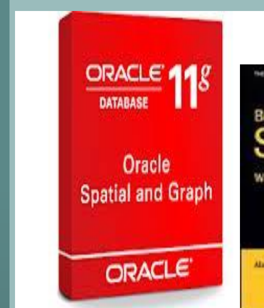
Arc ESRI GIS

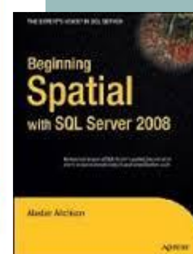


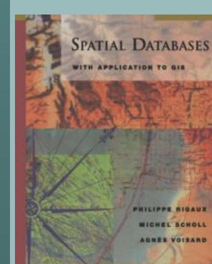


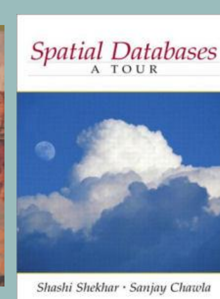
idrisi

PostGIS
Geospatial Objects for PostgreSQL









ORACLE

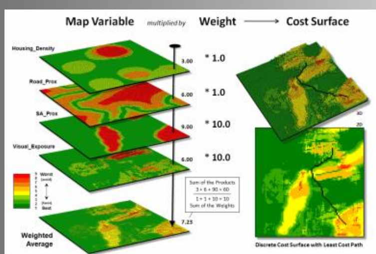
IBM DB2

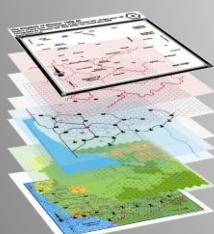
Microsoft SQL Server

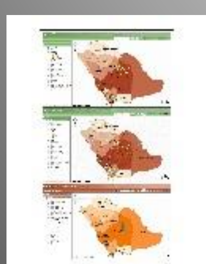
PostgreSQL
the world's most advanced open source database

MySQL

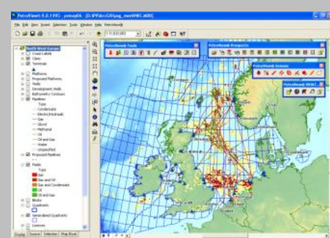
SYBASE | An SAP Company

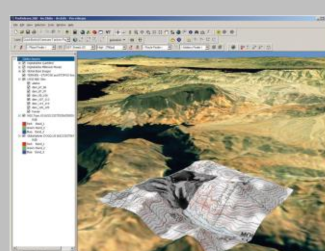






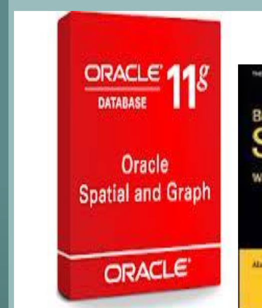
ESRI ArcGIS

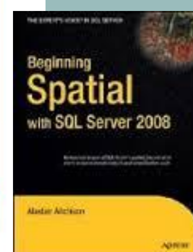


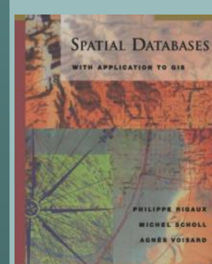


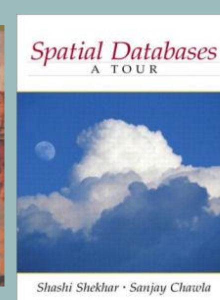
idrisi

PostGIS
Geospatial Objects for PostgreSQL









ORACLE

IBM DB2

Microsoft SQL Server

PostgreSQL
the world's most advanced open source database

MySQL

SYBASE | An SAP Company

Weight
Cost Surface
Weighted Average
ArcGIS
ESRI
idrisi
PostGIS
Geospatial Objects for PostgreSQL
ORACLE 11g
DATABASE
Oracle Spatial and Graph
ORACLE
Beginning Spatial with SQL Server 2008
Spatial Databases
with application to GIS
Philippe Rigaux
Michel Scholl
Agnès Voisard
Spatial Databases
A Tour
Shashi Shekhar · Sanjay Chawla
Google
bing
twitter
facebook
Map Reduce
amazon web services
EC2
hadoop
cloudera
IMPALA
Spark
ORACLE
IBM
DB2
Microsoft SQL Server
PostgreSQL
MySQL
SYBASE
An SAP Company
NASA

Let me check with my other good friends there.
Cool Big Data technology..!! Can I use it in my application?
My pleasure. Here it is.
HELP..!! Again, I have BIG data. Your technology is not helping me
Sorry, seems like the DBMS technology cannot scale more
Oh..!! But, it is not made for me. Can't make use of it as is
Google
bing
facebook
MapReduce
amazon web services
EC2
hadoop
cloudera IMPALA
Spark
PostGIS
Geospatial Objects for PostgreSQL
ESRI GIS
ORACLE 11g DATABASE
Oracle Spatial and Graph
ORACLE
Beginning Spatial with SQL Server 2008
Spatial Databases with application to GIS
Spatial Databases A Tour
Shashi Shekhar · Sanjay Chawla
NASA
idrisi
ORACLE
IBM DB2
Microsoft SQL Server
PostgreSQL
MySQL
SYBASE An SAP Company

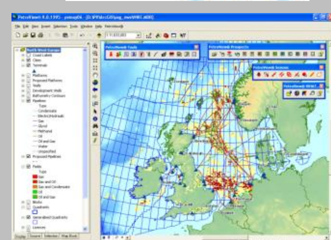

Google
bing
twitter
facebook
Map Reduce
amazon web services
EC2
hadoop
cloudera
IMPALA
Spark
Arc GIS
ESRI
PostGIS
Geospatial Objects for PostgreSQL
ORACLE 11g DATABASE
Oracle Spatial and Graph
ORACLE
Beginning Spatial with SQL Server 2008
NASA
idrisi
Spatial Databases
with application to GIS
Philippe Rigaux
Michel Scholl
Agnès Voisard
Spatial Databases
A Tour
Shashi Shekhar · Sanjay Chawla
ORACLE
IBM
DB2
Microsoft SQL Server
PostgreSQL
MySQL
SYBASE
An SAP Company

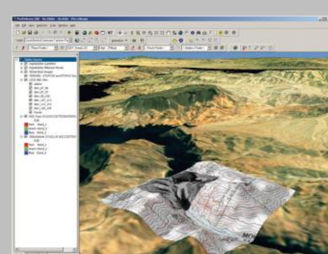

Kindly let me understand your needs
Kindly let me get the technology you have
facebook
MapReduce
amazon web services
EC2
hadoop
cloudera
IMPALA
Spark
ESRI
Arc
GIS
NASA
idrisi
PostGIS
Geospatial Objects for PostgreSQL
ORACLE 11g
DATABASE
Oracle Spatial and Graph
ORACLE
Beginning Spatial with SQL Server 2008
Spatial Databases
with Application to GIS
Spatial Databases
A Tour
Shashi Shekhar · Sanjay Chawla
ORACLE
IBM
DB2
Microsoft SQL Server
PostgreSQL
MySQL
SYBASE
An SAP Company

The Era of Big Spatial Data

NASA

Arc GIS ESRI





idrisi

Google

bing

twitter

facebook

MapReduce

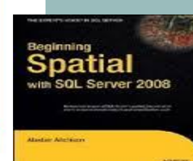
amazon web services EC2

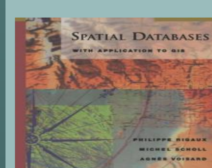
hadoop

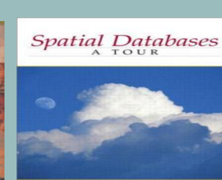
cloudera IMPALA

Spark

PostGIS
Geospatial Objects for PostgreSQL







ORACLE

IBM DB2

Microsoft SQL Server

PostgreSQL

MySQL

SYBASE | An SAP Company

Big ~~Spatial~~ Data Systems

UCR

hadoop

FLUME

HIVE

APACHE HBASE

Spark

APACHE GIRAPH

accumulo™

sqoop

mahout

APACHE STORM™

TEZ

Asterix DB

Hortonworks

OOZIE

The Era of Big Spatial Data

UCR

Recently, a few products have emerged …

Spatial Hadoop
I can play with your spatial data

rasdaman
raster data manager

geomesa

GeoTrellis

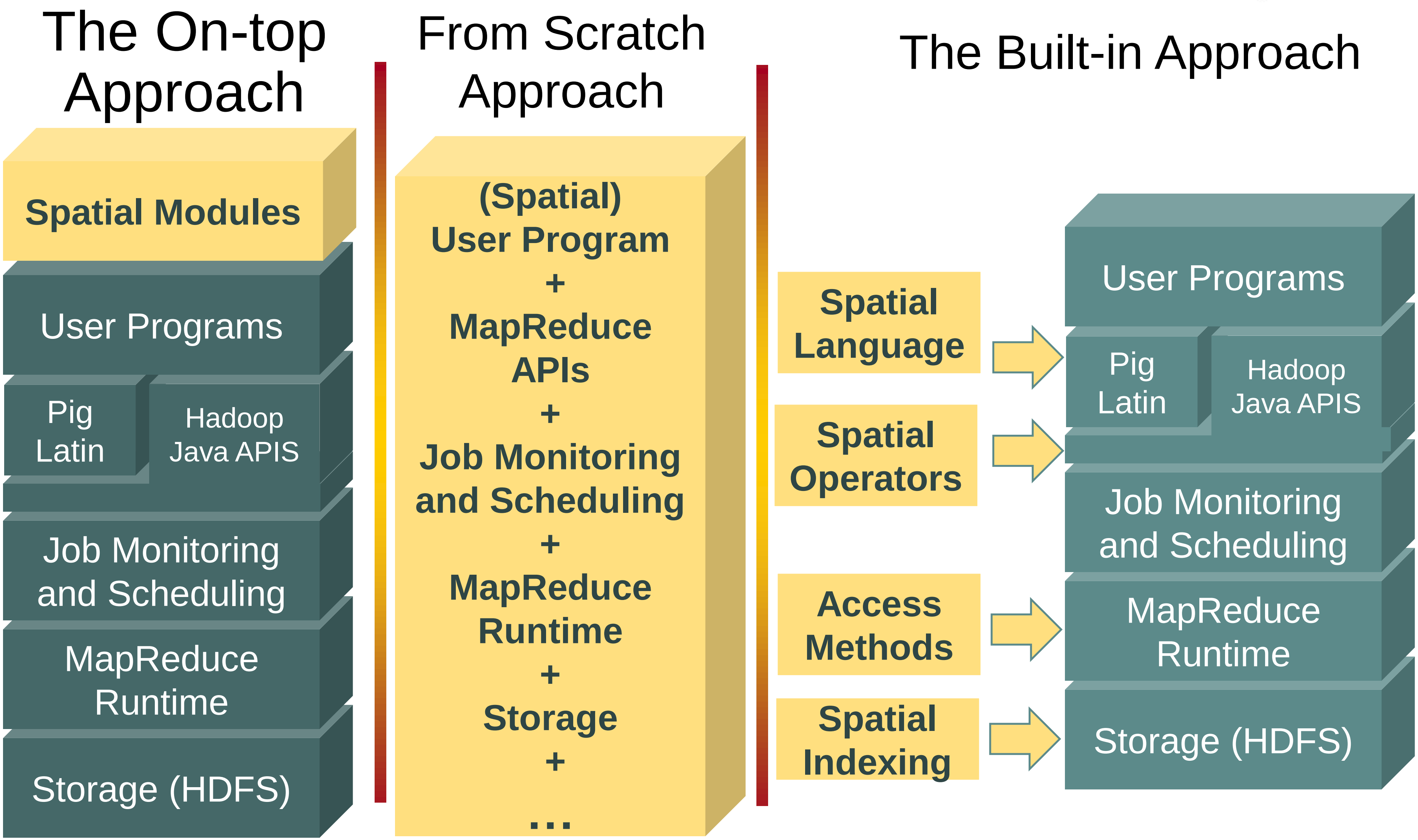
Hadoop-GIS
Spatial Big Data Solutions

CYBER GIS

SpaceCurve

GeoSpark

Approaches for Building A Big Spatial Data System



A Lesson from History: Spatial Database Management Systems

- Database management systems are general purpose for any data types
  - No special treatment for spatial data
  - Spatial data types is defined with traditional types
  - Range and K-NN queries can be supported with traditional SQL
  - B-tree indexing is used for spatial data
- Still can work fine, but there should be better tailored techniques for spatial data
- Spatial database systems are introduced with its own data types, spatial operators and index structures

System Architecture for Big Spatial Data

UCR

Applications
Satellite Imagery, GIS, Microblogs, Medical Imagery, …

Language

Visualization
Single level and multilevel images

Query Processing
Basic Queries, Spatial Join, and Computational Geometry

Indexing
Grid, R-tree, Quad tree, K-d tree, …

Indexing

Applications
Satellite Imagery, GIS, Microblogs, Medical Imagery, …

Language

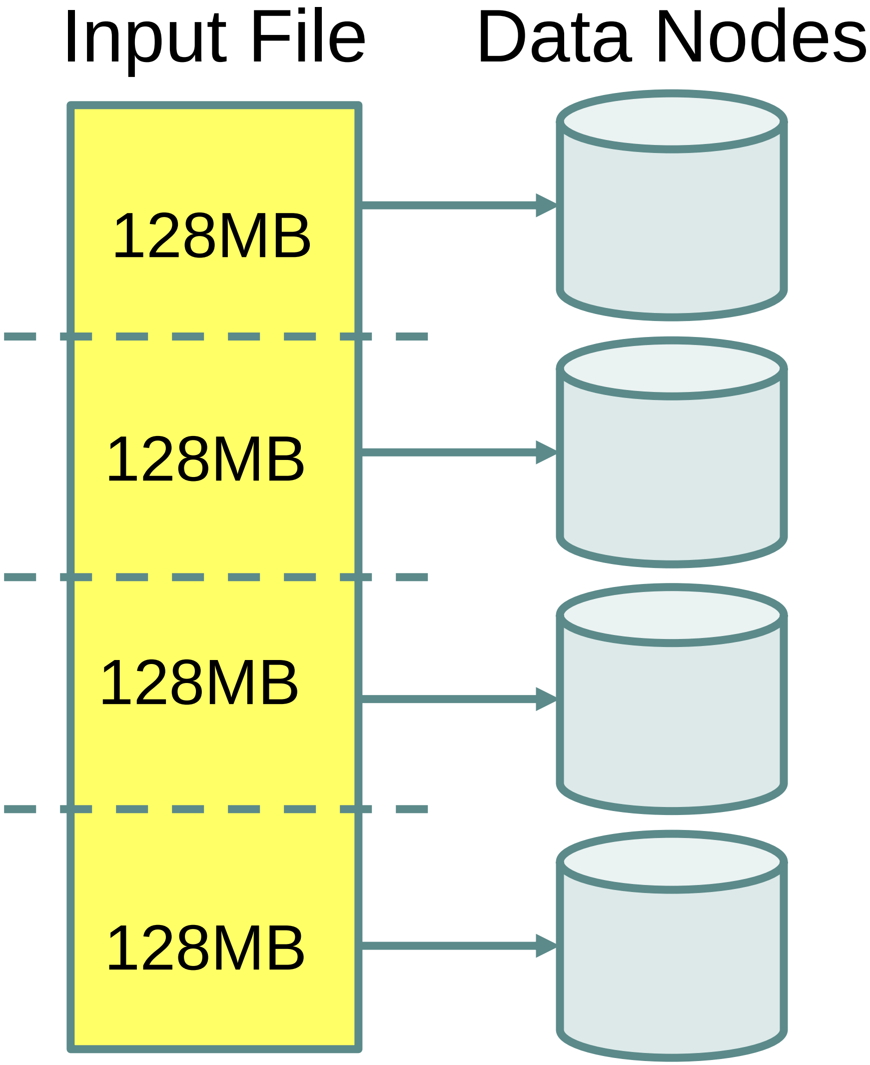
Visualization
Single level and multilevel images

Query Processing
Basic Queries, Spatial Join, and Computational Geometry

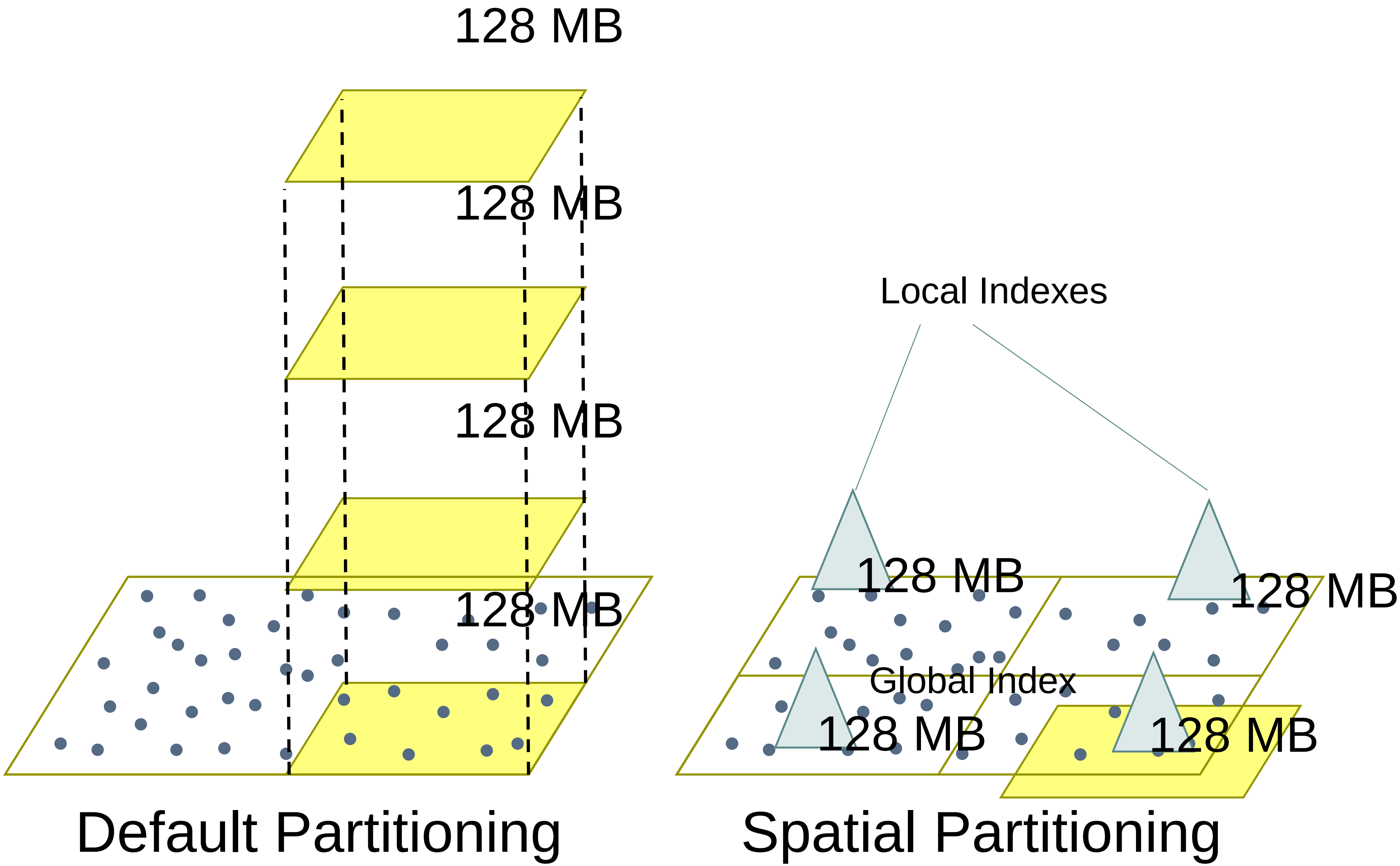
Indexing
Grid, R-tree, Quad tree, K-d tree, …

Data Loading in Hadoop

- Hadoop Distributed File System (HDFS) is widely used.
- HDFS is unaware of spatial data
- Challenges:
  - Big data size
  - HDFS files are sequential and write once



Spatial Indexing



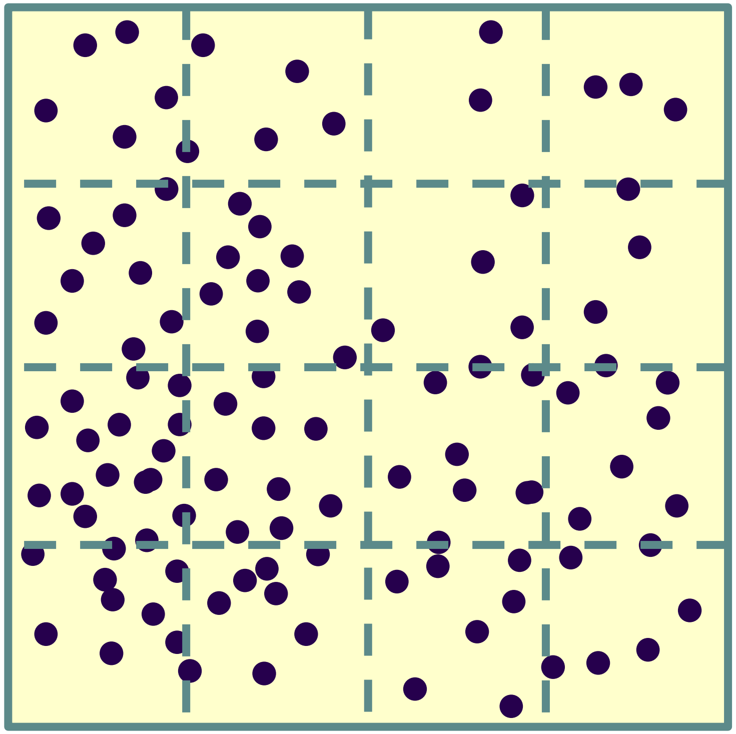
Spatial Indexing Classification

1. How to calculate number of partitions?
2. What is the type of global index?
3. What is the type of local indexes?
4. Is it a clustered or unclustered index?
5. Is it a static or dynamic index?

Uniform Grid Index

- Apply a uniform grid of size $\sqrt{n} \times \sqrt{n}$
- Scan the input and assign each record

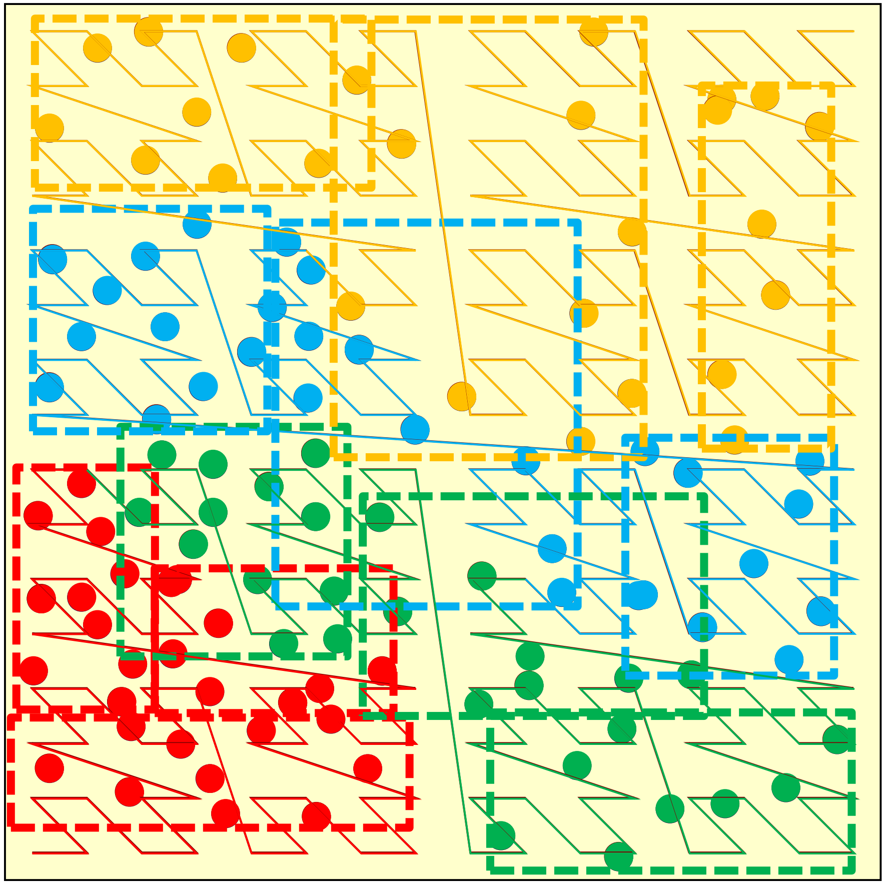
# of Partitions	User-defined [1] # of HDFS blocks [2]
Global	Grid
Local	None
Clustered	
Static	



[1] A. Aji, *et al*.**"Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce"**. In VLDB, 2013
[2] A. Eldawy and M. F. Mokbel. **"SpatialHadoop: A MapReduce Framework for Spatial Data"**. In ICDE, 2015.

R-tree construction

- Sample
- Sort by Z-curve
- Divide into *n* ranges
- Scan input records and partition to the



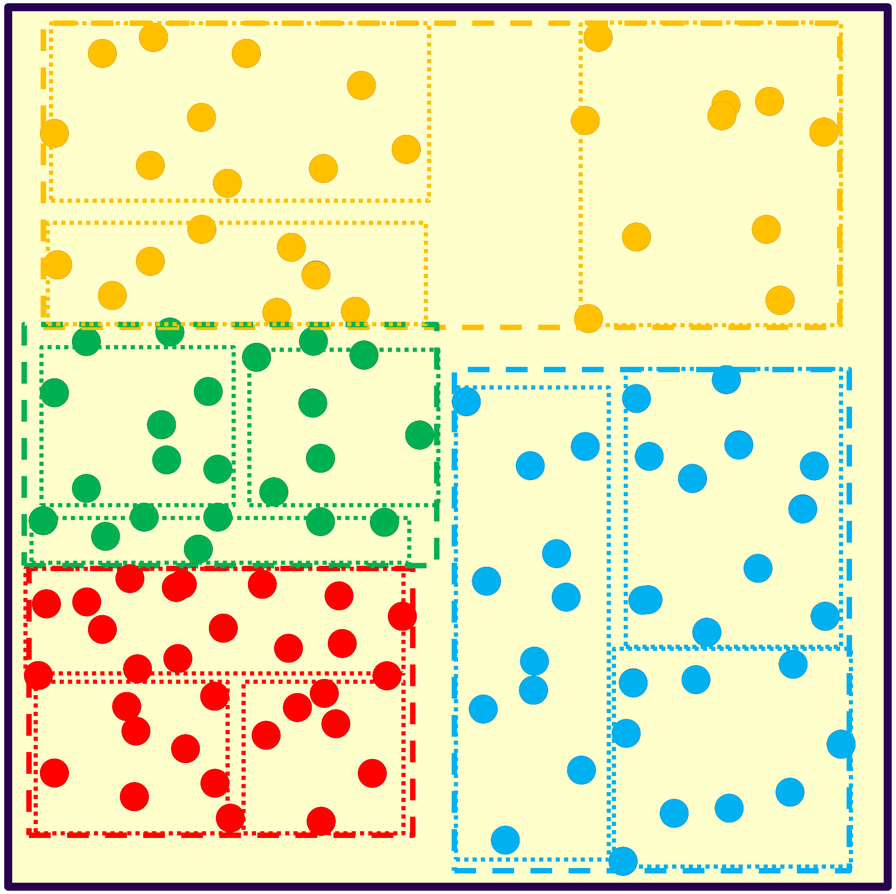
# of Partitions	# of Machines
Global	Z-curve
Local	R-tree
Clustered	
Static	

A. Cary, Z. Sun, V. Hristidis, and N. Rishe. **“Experiences on Processing Spatial Data with MapReduce”**. In SSDBM, 2009

R-tree and R+-tree

- Number of partitions (blocks): $n = \frac{Expected\ size}{block\ capacity}$
- Find partition boundaries
  - Step 1: Sampling
  - Step 2: Bulk load in an R(+)-tree
  - Step 3: Partition boundaries are the MBRs of leaf nodes
- Scan input file, assign each record to its partition(s)

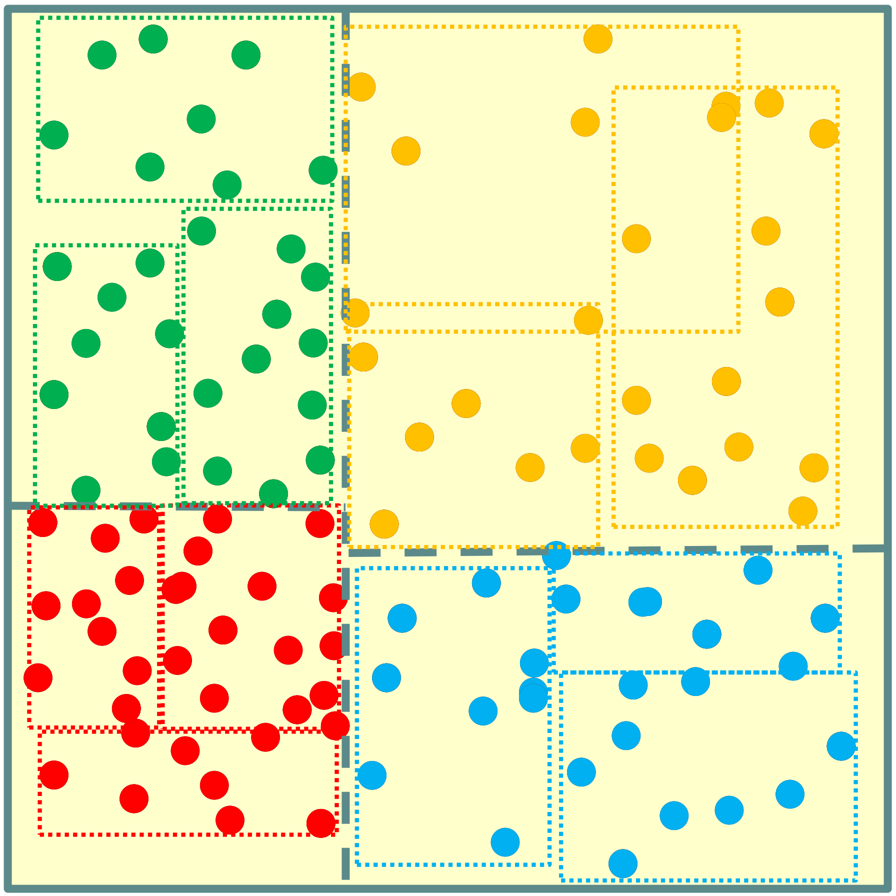
# of Partitions	# of HDFS blocks
Global	R(+)-tree
Local	R(+)-tree
Clustered	
Static	



A. Eldawy and M. F. Mokbel. **“SpatialHadoop: A MapReduce Framework for Spatial Data”**. In ICDE, 2015.

ScalaGiST

- Generalization of the sampling-based indexing for Generalized Search Trees (GiST)
- The sample is partitioned using K-d tree
- Each partition is indexed using a GiST-based index

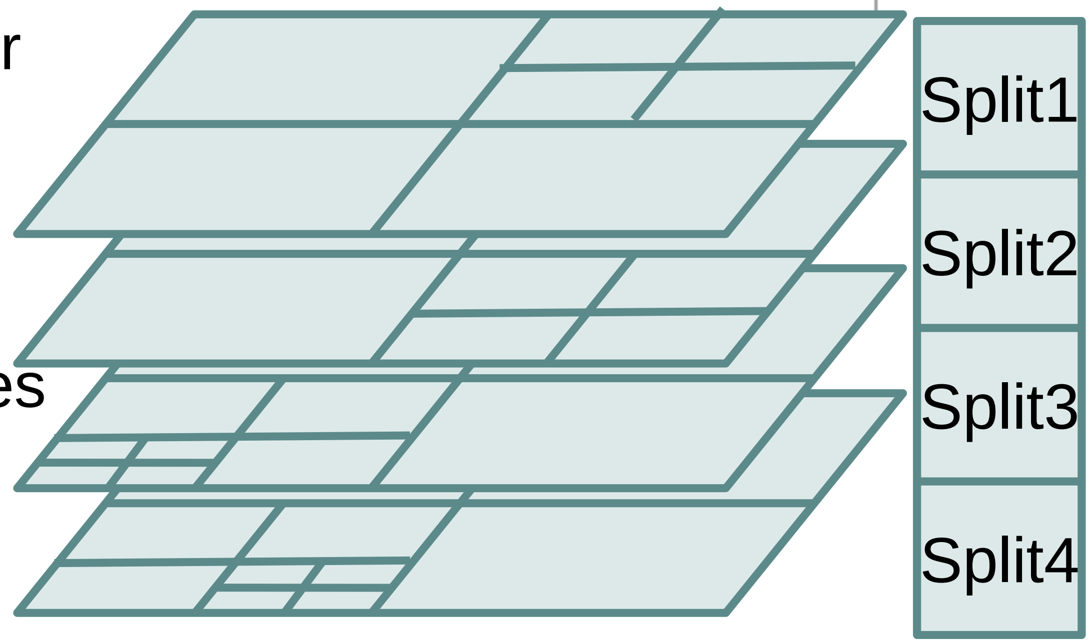


# of Partitions	# of HDFS blocks
Global	K-d tree
Local	GiST-based
Clustered or Unclustered	
Static	

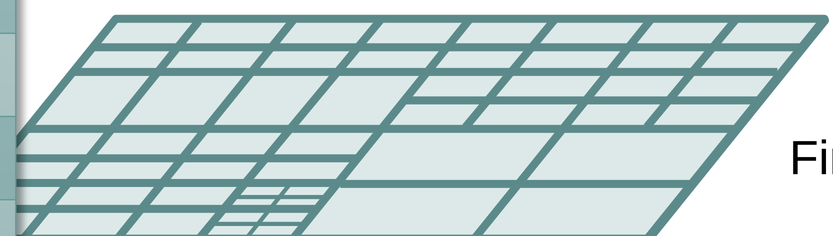
P. Lu. *et al*. **"ScalaGiST: Scalable Generalized Search Trees for MapReduce Systems"**. PVLDB, 7(14), 2014

Quad tree

- Split the input file over machines
- Create a Quad tree for each split
- Partition the leaf nodes across machines [M1-M4]
- Merge leaf nodes to



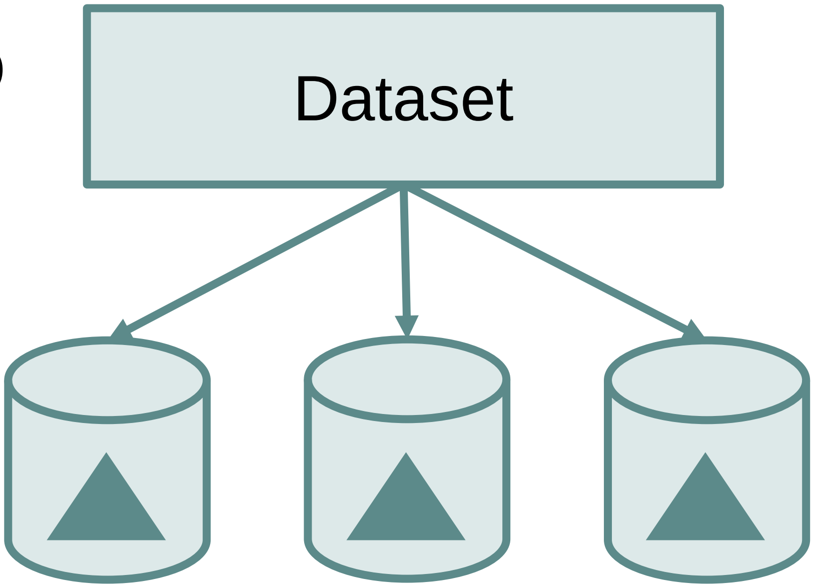
# of Partitions	User-defined
Global	Quad-tree
Local	Quad-tree
Clustered or Unclustered	
Static	



R. T. Whitman, M. B. Park, S. A. Ambrose, and E. G. Hoel. **“Spatial Indexing and Analytics on Hadoop”**. In SIGSPATIAL, 2014

Spark-based Indexes

- Builds on the RDD data model
- Uses the sample-based technique to construct R-tree partitions
- Each partition in the RDD can be further indexed



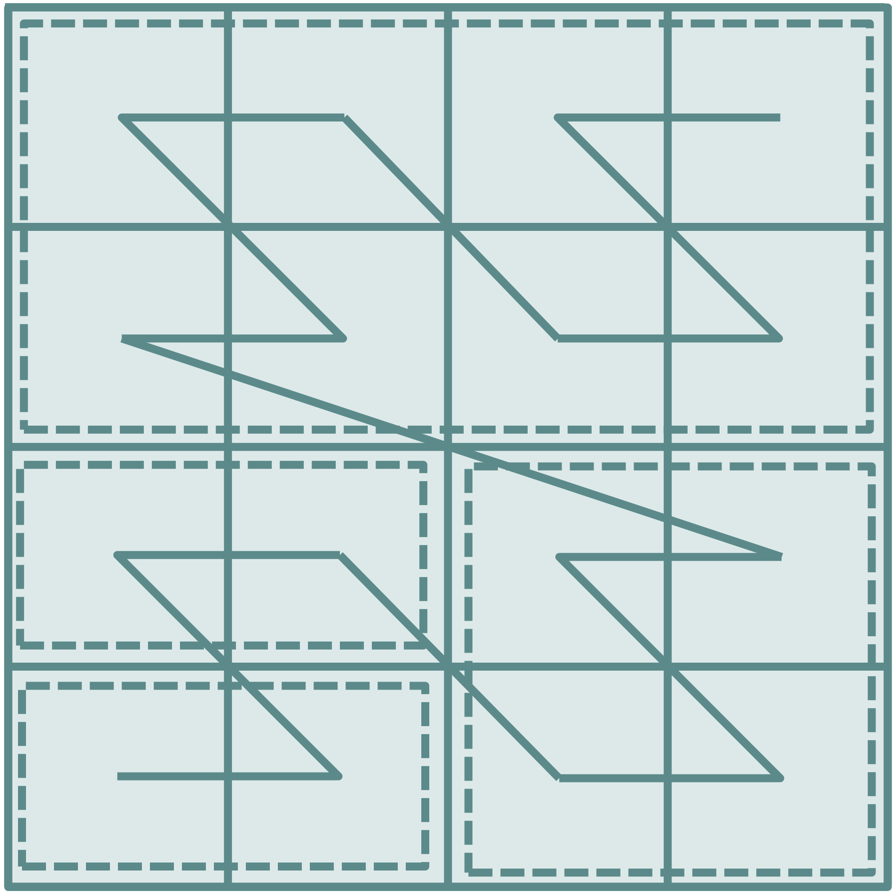
# of Partitions	User-defined
Global	STR-based
Local	R-tree or Quad-tree
Clustered	
Static	

J. Yu *et al*, **"A Demonstration of GeoSpark: A Cluster Computing Framework for Processing Big Spatial Data,"** in ICDE 2016

D. Xie *et al*, **"Simba: Efficient In-Memory Spatial Analytics,"** in SIGMOD 2016

$\mathcal{MD}$-HBase

- Utilizes the linear index in HBase
- Keeps points sorted by Z-curve order
- Builds a virtual Quad-tree or K-d

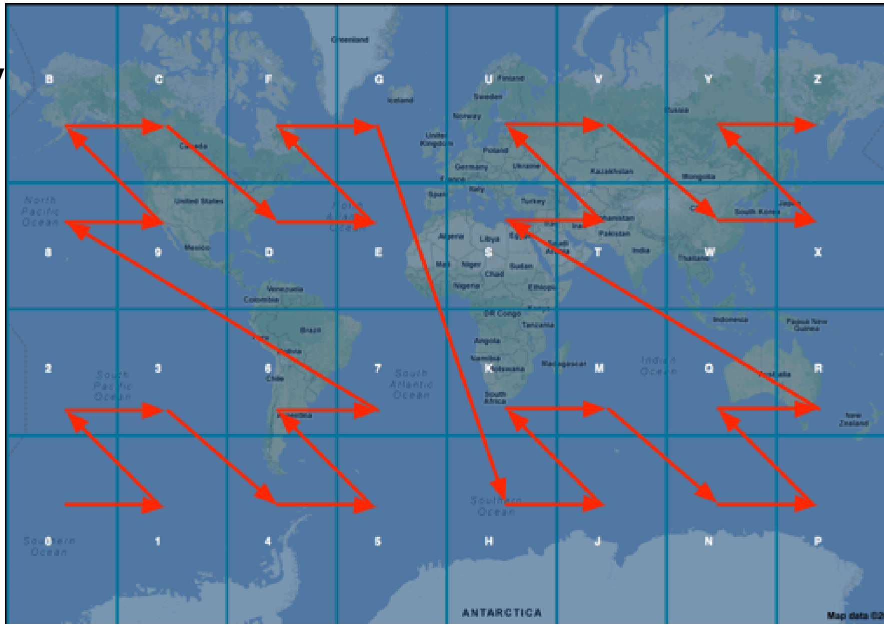


# of Partitions	# of HDFS blocks
Global	K-d tree or Quad-tree
Local	--
Clustered	
Fully dynamic (Insertion and Deletion)	

S. Nishimura *et al*. **"MD-HBase: Design and Implementation of an Elastic Data Infrastructure for Cloud-scale Location Services"**. DAPD, 31(2), 2013

GeoMesa

- Utilizes the sorted index employed in Accumulo
- Keeps records sorted by geohash
- Support spatio-temporal indexing



# of Partitions	# of HDFS blocks
Global	Geo-hash
Local	Geo-hash
Clustered	
Dynamic (Insertion and Deletion)	

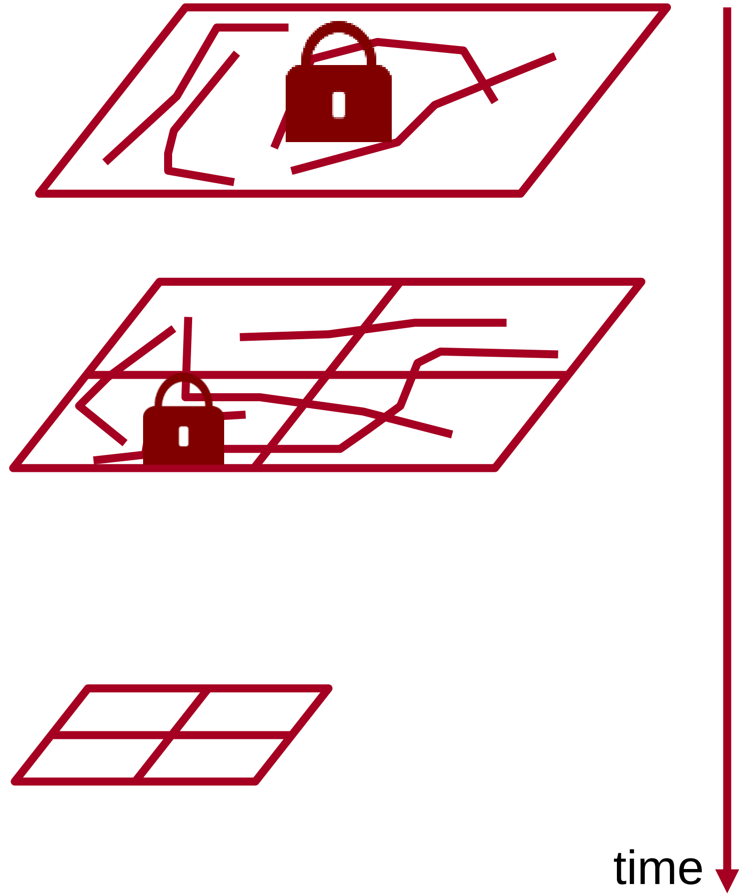
A. Fox *et al*. **"Spatio-temporal Indexing in Non-relational Distributed Databases"**. In International Conference on Big Data, 2013

Quad-tree-based index

- Initially, all trajectories are stored in one partition
- As the partition fills up, new partitions are created for new records
- Each partition is defined

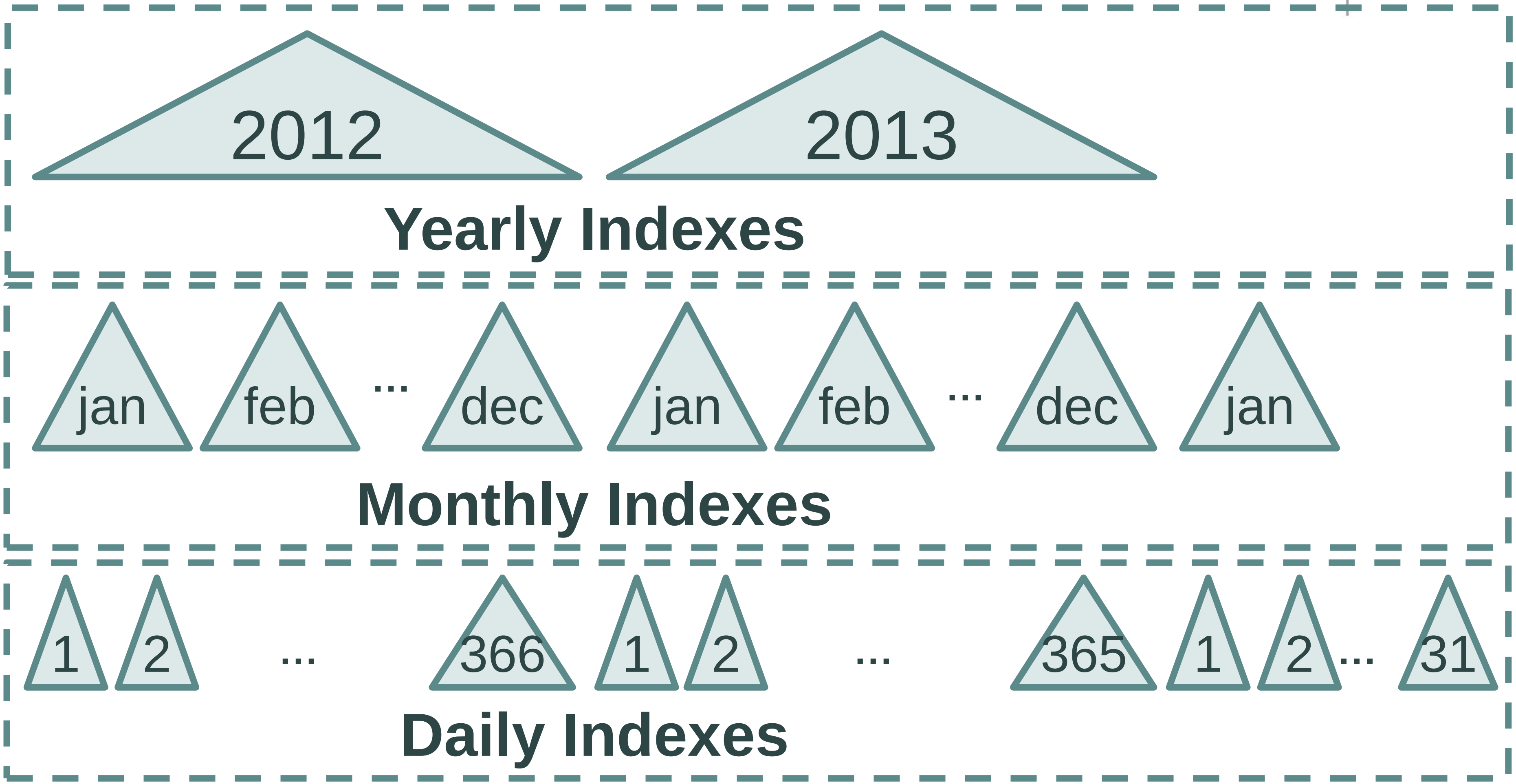
val)

# of Partitions	# of HDFS blocks
Global	Quad-tree
Local	--
Clustered	
Partially dynamic (Insertion only)	



Q. Ma *et al*, **“Query Processing of Massive Trajectory Data Based on MapReduce”**. In CLOUDDB, 2009.

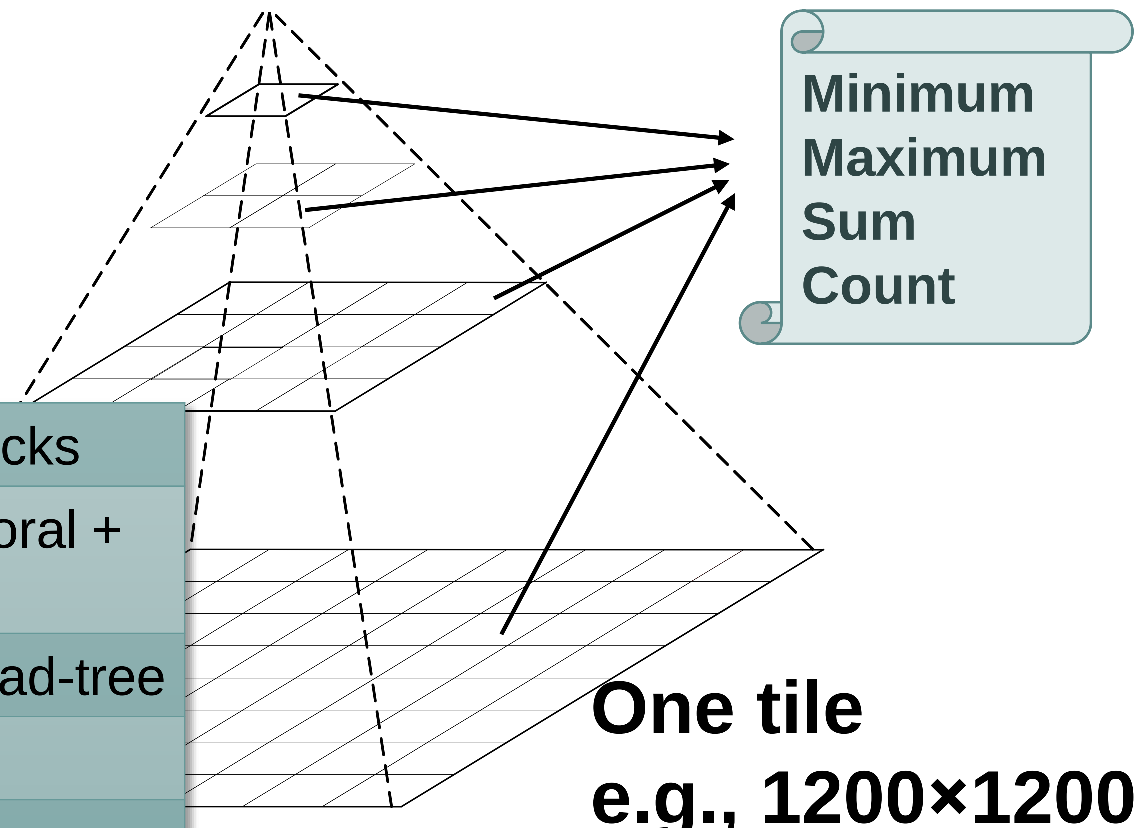
Spatio-temporal Index



A Eldawy *et al*, **"SHAHED: A MapReduce-based System for Querying and Visualizing Spatio-temporal Satellite Data"**, ICDE 2015
L. Aalarbi *et al*, **"A Demonstration of ST-Hadoop: A MapReduce Framework for Big Spatio-temporal Data,"** VLDB 2017

Aggregate Quad Trees

Amend all Quad tree nodes with aggregate values



# of Partitions	# of HDFS blocks
Global	Multires-temporal + Uniform grid
Local	Aggregate Quad-tree
Clustered	
Dynamic (Insertion only)	

A Eldawy *et al*, **"SHAHED: A MapReduce-based System for Querying and Visualizing Spatio-temporal Satellite Data"**, ICDE 2015

Indexes in HDFS

Index	# of Partitions	Global	Local	C	U	Dynamic
Hadoop-GIS	User-defined	Uniform grid	-	✓		
R-tree building	# of machines	Z-curve	R-tree	✓		
SpatialHadoop	# of Blocks	R(+)-tree	R(+)-tree	✓		
ScalaGiST	# of machines	K-d tree	GiST	✓	✓	
ESRI-Hadoop	# of machines	Quad Tree	Quad Tree	✓	✓	
GeoSpark	User-defined	Grid	R&Quad-tree	✓		
MD-HBase	# of Blocks	K-d tree Quad tree	-	✓		✓
GeoMesa	# of Blocks	GeoHash	GeoHash	✓		✓
Trajectory Index	# of Blocks	Quad-tree-based	-	✓		Insertion
SHAHED	# of Blocks	Mulitres temporal + Grid	Aggregate Quad-tree	✓		Insertion

Query Processing

Applications
Satellite Imagery, GIS, Microblogs, Medical Imagery, …

Language

Visualization
Single level and multilevel images

Query Processing
Basic Queries, Spatial Join, and Computational Geometry

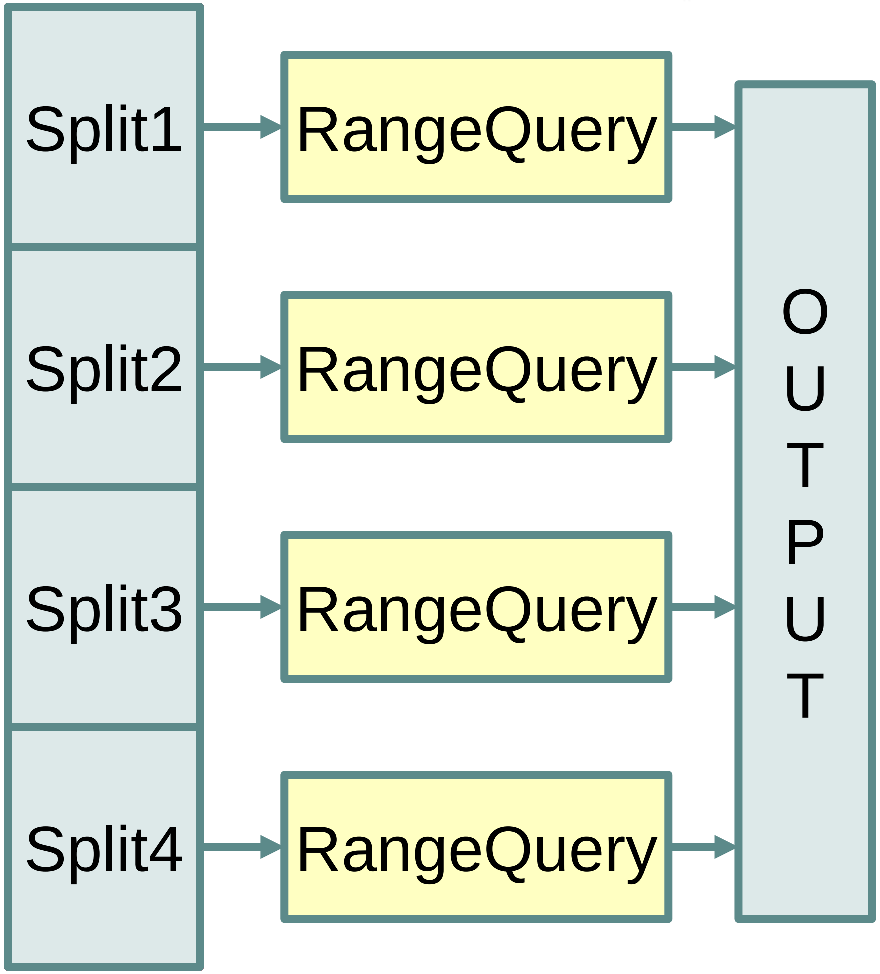
Indexing
Grid, R-tree, Quad tree, K-d tree, …

Basic Spatial Queries

Operation	Approach	Indexes (if any)	Systems
Range Query	Full table scan	-	ZHL+09
	Filter-refine	Grid, R-tree, R+-tree, Quad Tree	SpatialHadoop Hadoop-GIS MD Hbase GeoMesa ESRI Tools PRADASE
KNN	Full table scan	-	ZHL+09 Hadoop-GIS
	Incremental	Grid, R-tree, and Quad Tree	SpatialHadoop MD HBase ESRI Tools

Range Query by Full Scan

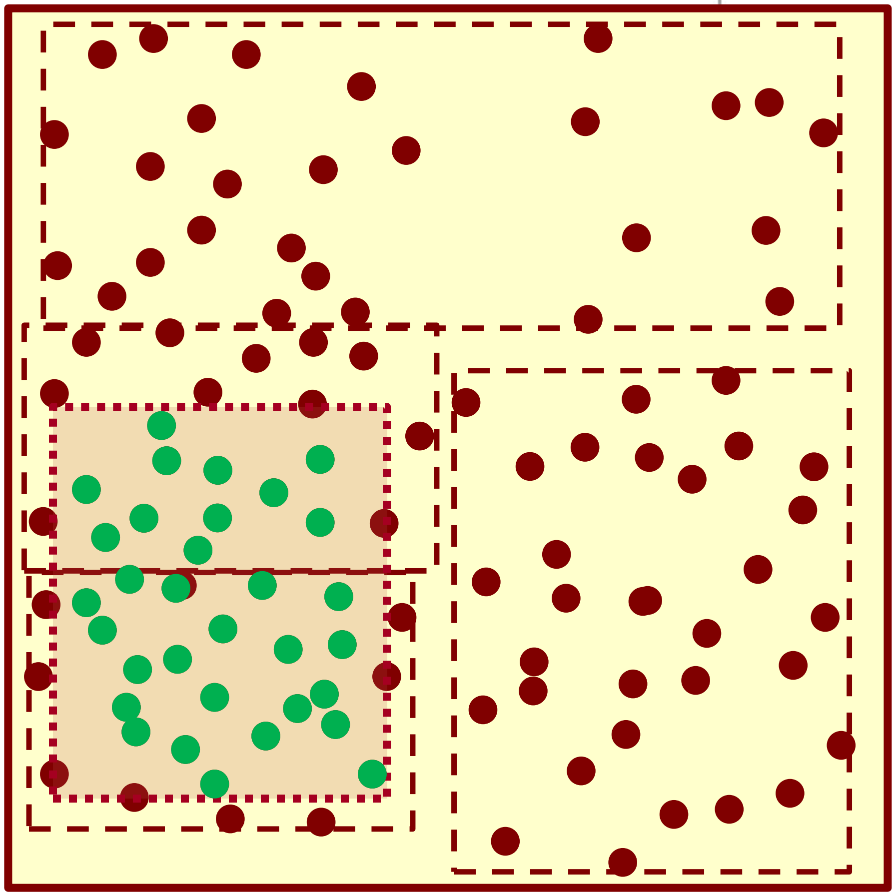
- Split the input file using the default HDFS partitioning
- Each mapper scans records in the assigned split
- Matching records are written to the output
- No reduce phase is required



S. Zhang, J. Han, Z. Liu, K. Wang, and S. Feng. **"Spatial Queries Evaluation with MapReduce"**. In GCC, 2009.

Indexed Range Query

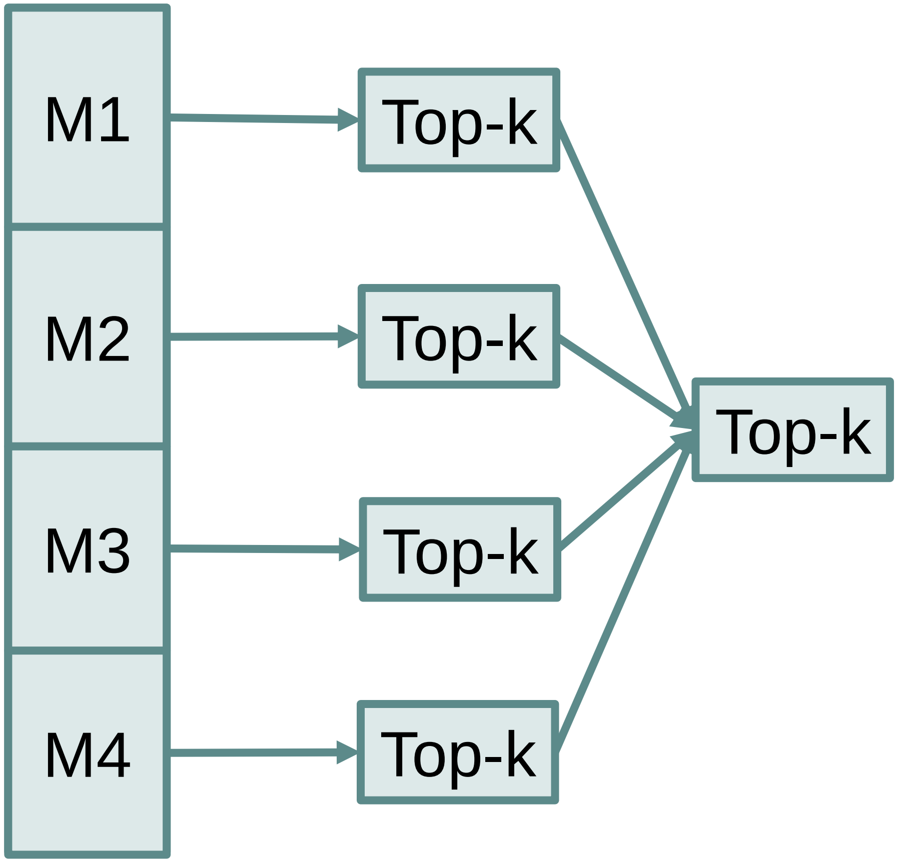
1. Filter: Select overlapping partitions in the global index
2. Refine: Select matching records in each partition
3. Duplicate avoidance: remove duplicates if records are replicated in the index (e.g., R+-tree and Grid)



Full Scan K-Nearest Neighbor

- Straight forward solution, no index required

1. Scan the input. Calculate distance to each point.
2. Select top-k on each machine
3. Combine all matches in one machine and



[1] S. Zhang, *et al*. **"Spatial Queries Evaluation with MapReduce"**. In GCC, 2009.
[2] A. Aji, et al.**"Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce"**. In VLDB, 2013

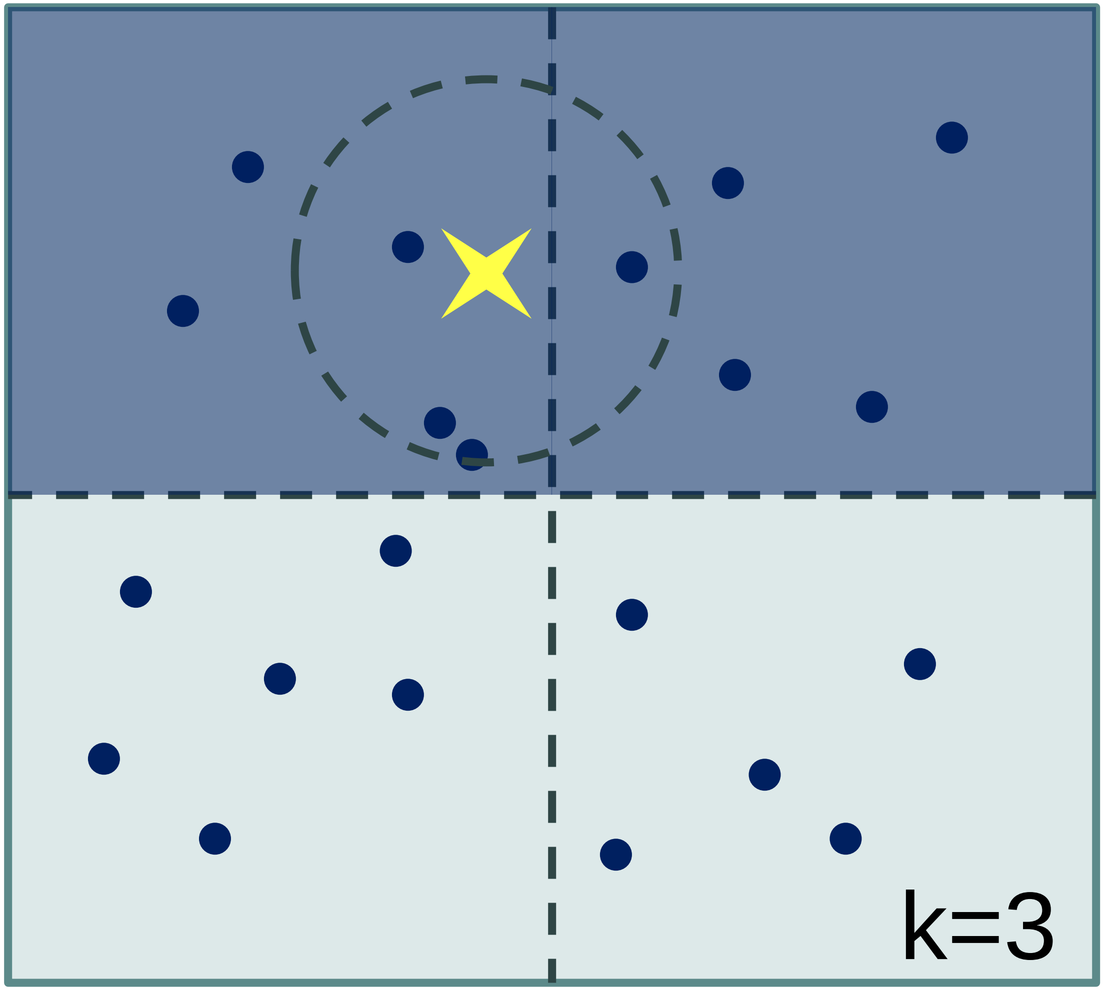
KNN over Indexed Data

First iteration runs as before and result is tested for correctness

- ✗ Answer is incorrect

Second iteration processes other blocks that might contain an answer

- ✓ Answer is correct



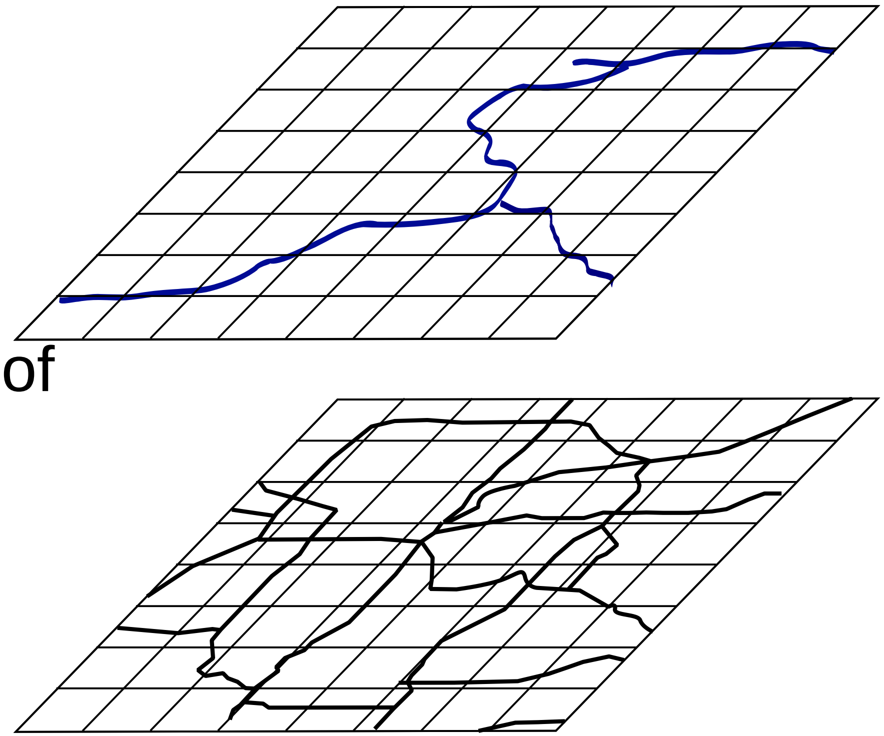
SpatialHadoop, ESRI, ScalaGiST, MD-HBase

Spatial Join Queries

Join Query	Approach	Indexes (if any)	Systems
Self Join	Index-based	Grid	Hadoop-GIS
Binary Join	PBSM	-	SJMR
	Overlap-Join Repartition-Join	Grid, R-tree R+-tree	SpatialHadoop
	Indexed-nested-loop	R-tree	SpatialSpark SpatialImpala
Mulitway Join	PBSM	-	GCN+13
KNN-Join	Brute-force (Exact)	-	ZLJ12
	Z-curve-based (Approximate)	-	ZLJ12
	Voronoi-based (Exact)	-	LSC+12

PBSM Join (No Indexes)

- Partition both inputs using a common grid
- Replicate a shape to all overlapping cells
- Join the contents of each pair of cells separately
- Duplicate elimination
- Ported to MapReduce as SJMR [1]
- Multiway spatial join [2]



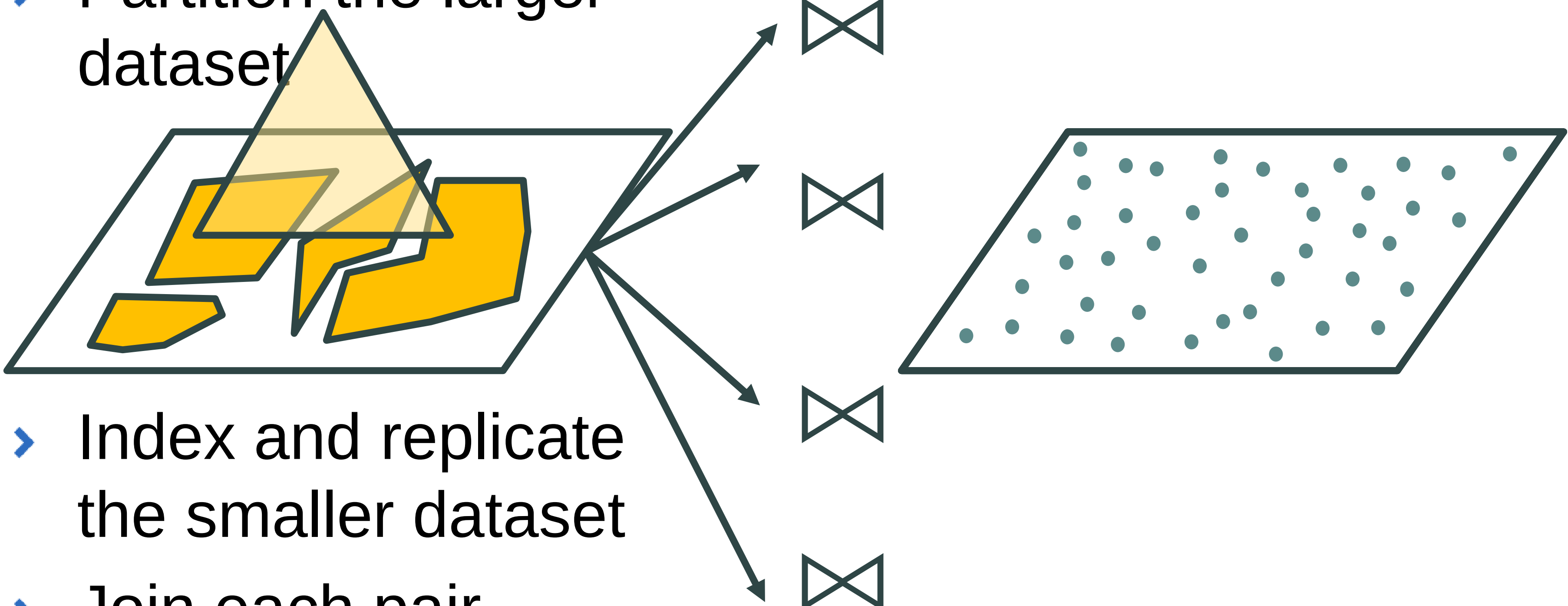
Roads ⋈ Rivers

[1] S. Zhang, et al. **"SJMR: Parallelizing spatial join with MapReduce on clusters"**. In CLUSTER, 2009

[2] H. Gupta, *et al*, **"Processing multi-way spatial joins on map-reduce"**, EDBT 2013

Indexed Nested Loop Join

- Spatial join using point in polygon predicate
- Partition the larger dataset

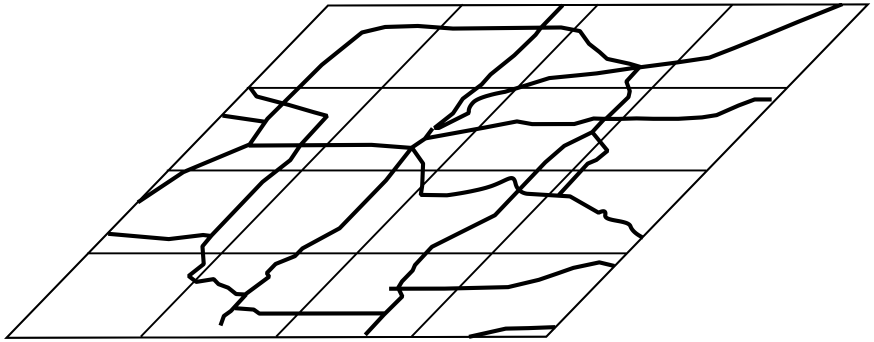


- Index and replicate the smaller dataset
- Join each pair

S. You, J. Zhang, L. Gruenwald, **"Large-Scale Spatial Join Query Processing in Cloud"**, CloudDM, 2015

Self-Join using a grid index

- Utilize the grid index to avoid partitioning the input
- Perform a local self-join on each partition
- Implemented as a map-only job

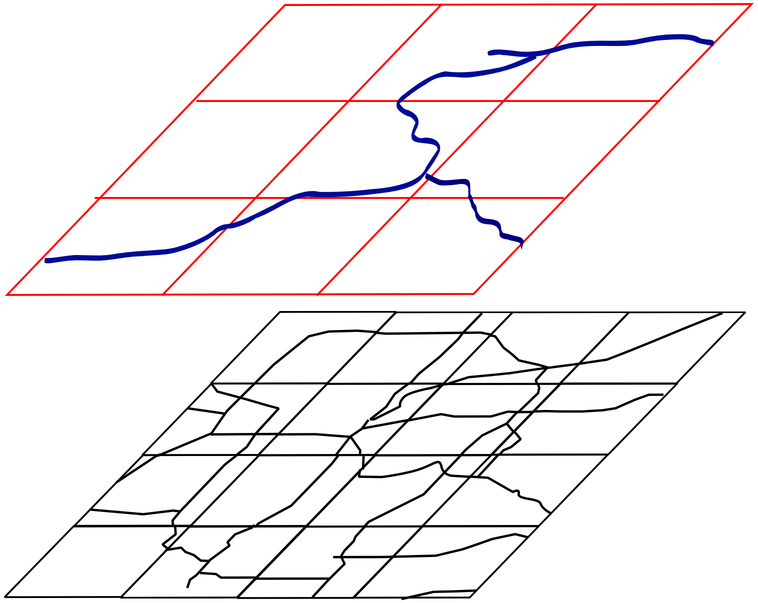


A. Aji, *et al*.**"Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce"**. In VLDB, 2013

Binary Spatial Join

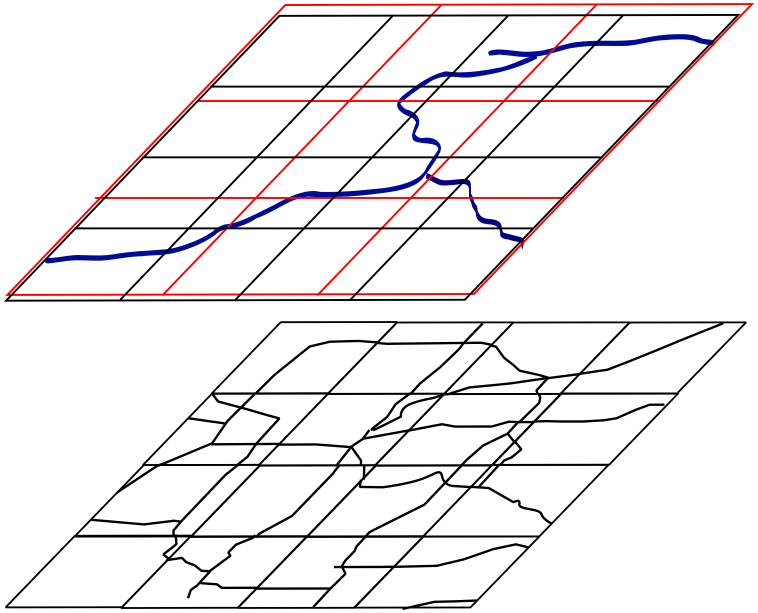
- Two different indexes

Approach 1: Join Directly



Total of 36 overlapping pairs

Approach 2: Partition – Join

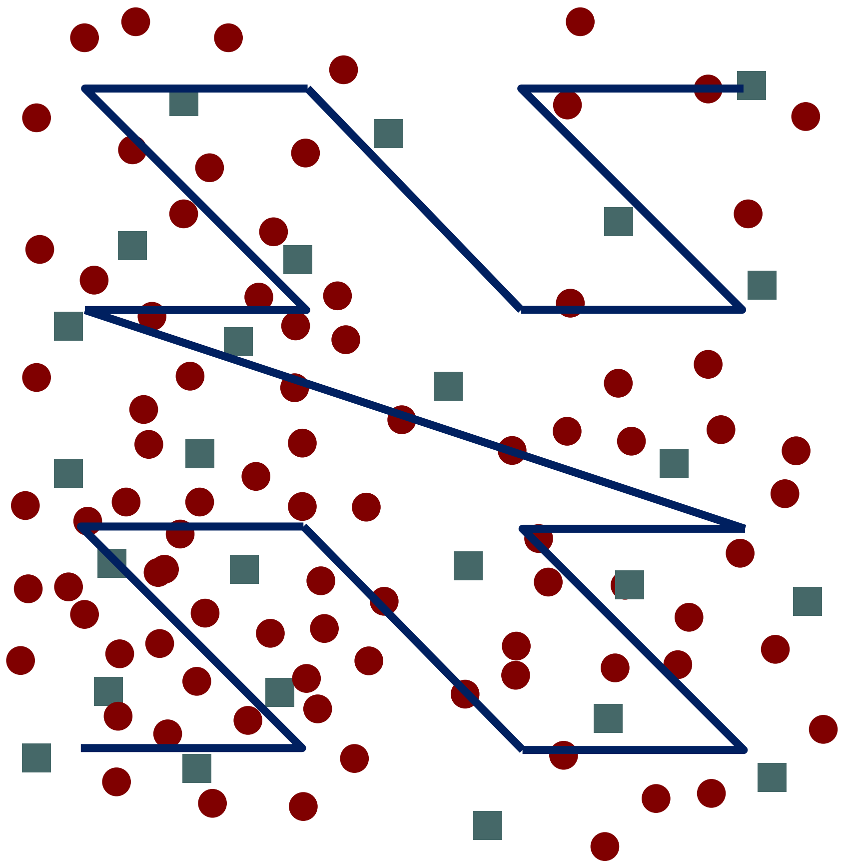


Only 16 overlapping pairs

A. Eldawy and M. F. Mokbel. **“SpatialHadoop: A MapReduce Framework for Spatial Data”**. In ICDE, 2015.

Approximate KNN Join
using Z-curve

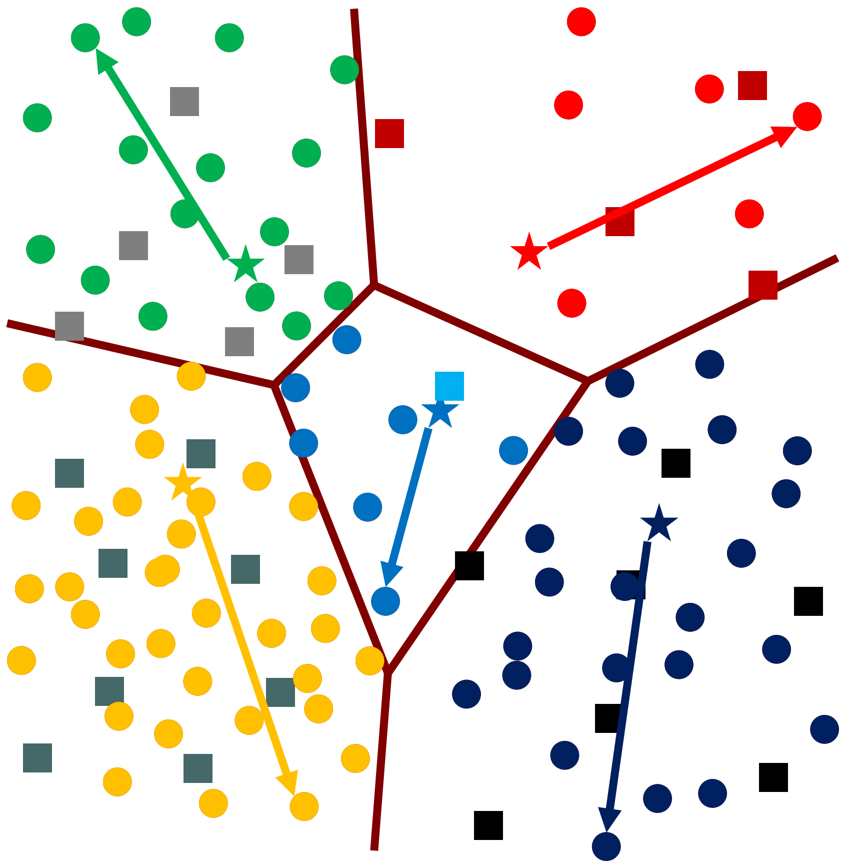
- For each ■, find KNN ●
- Co-partition both R and S using a Z-curve
- Join every pair of corresponding partitions
  - Answer is approximate
- Repeat α-times by shifting the z-values to increase accuracy



C. Zhang, F. Li, and J. Jestes. **"Efficient Parallel kNN Joins for Large Data in MapReduce"**. In EDBT, 2012

Exact KNN Join using Voronoi Diagram (VD)

- Select *n* pivots
- Construct VD for pivots
- Partition R and S into n partitions using VD
- Collect statistics for each partition (e.g., count and maximum distance to pivot)
- Find pairs of partitions (R_i, S_i) that produce answer
- Compute KNN-join between each partition in R and matching partitions in S



W. Lu, Y. Shen, S. Chen, and B. C. Ooi. **"Efficient Processing of k Nearest Neighbor Joins using MapReduce"**. VLDB, 2012

Computational Geometry

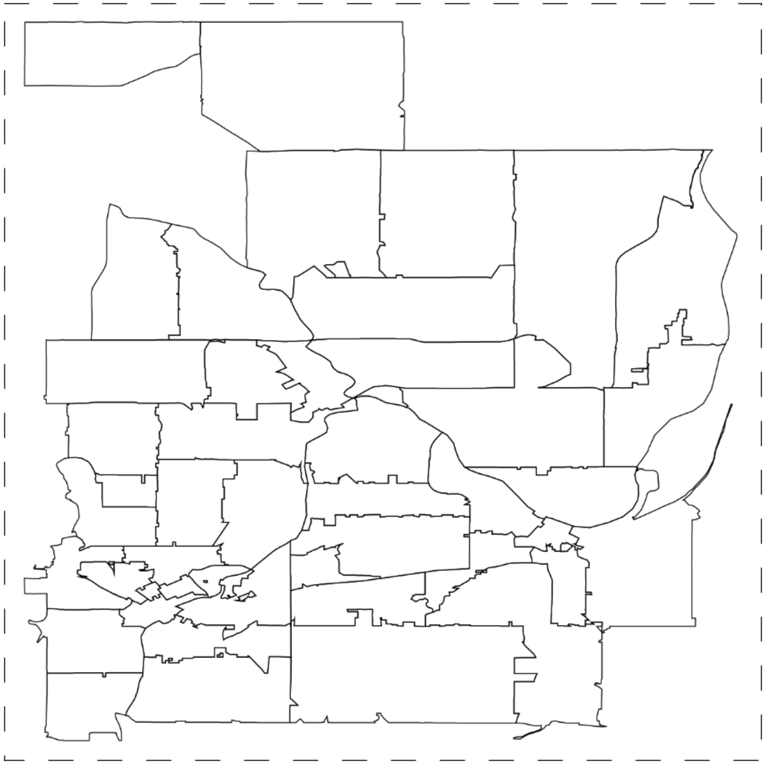
- Divide and conquer approaches can be ported to MapReduce
- General Algorithm
  1. Partition the input using default Hadoop partitioner or SpatialHadoop partitioner
  2. Prune partitions that do not contribute to answer (If spatial partitioner is used)
  3. Apply the algorithm locally in each partition
  4. Combine the partial answers to compute the final result
- Used to implement five CG operations
  - Polygon union, convex hull, skyline, closest/farthest pair

A. Eldawy, *et al*. **"CG Hadoop: Computational Geometry in MapReduce"**.
In SIGSPATIAL, 2013

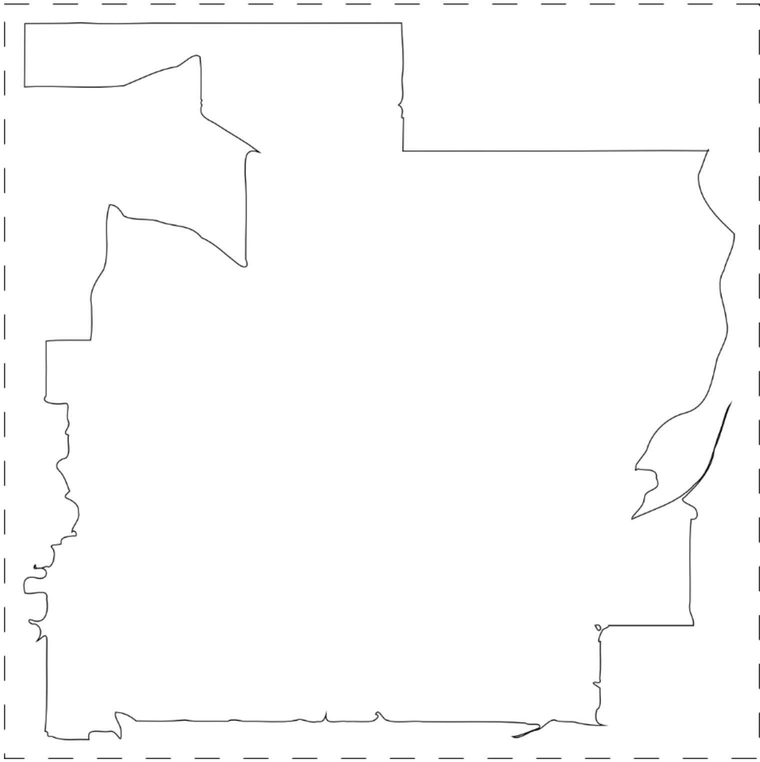
Polygon Union

Computes the union of a set of polygons

Input



Output



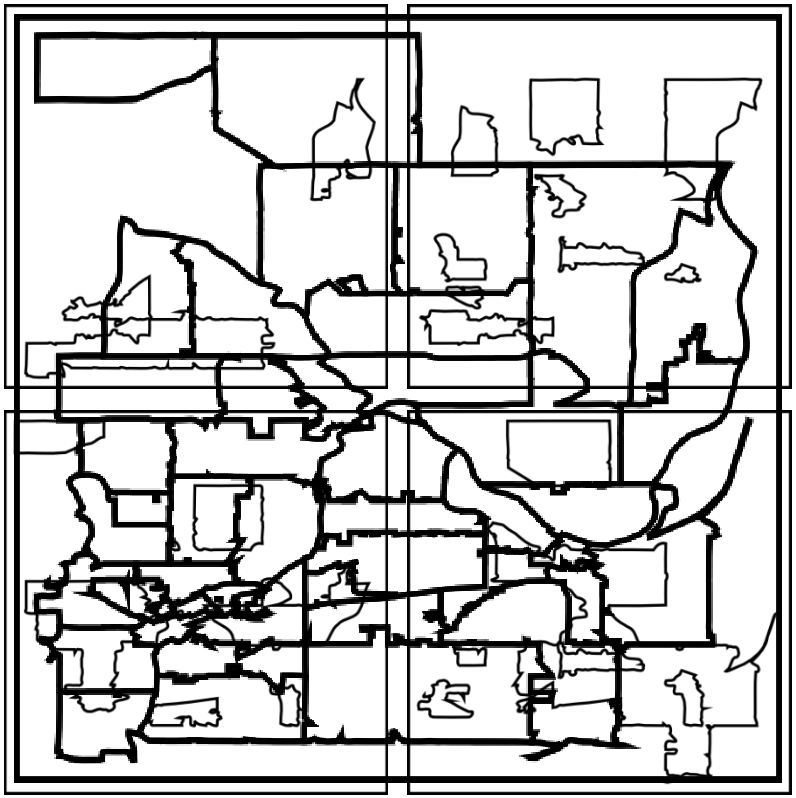
Polygon Union in CG_Hadoop

①Partition

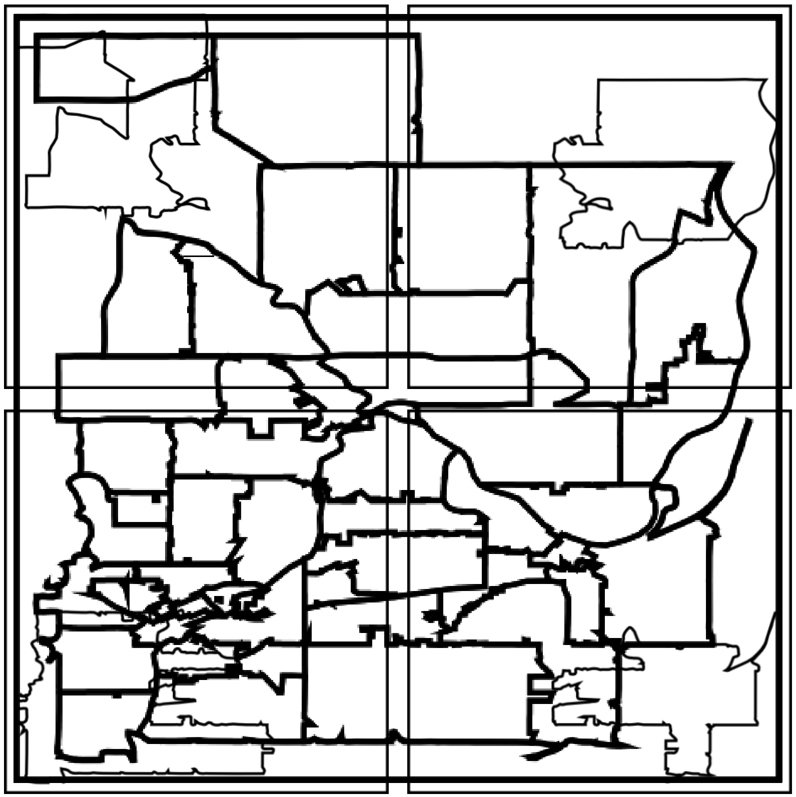
② Local union

③ Global union

Non-spatial partitioning



Spatial partitioning



A. Eldawy, *et al*. **"CG Hadoop: Computational Geometry in MapReduce"**.
In SIGSPATIAL, 2013

Skyline in CG_Hadoop

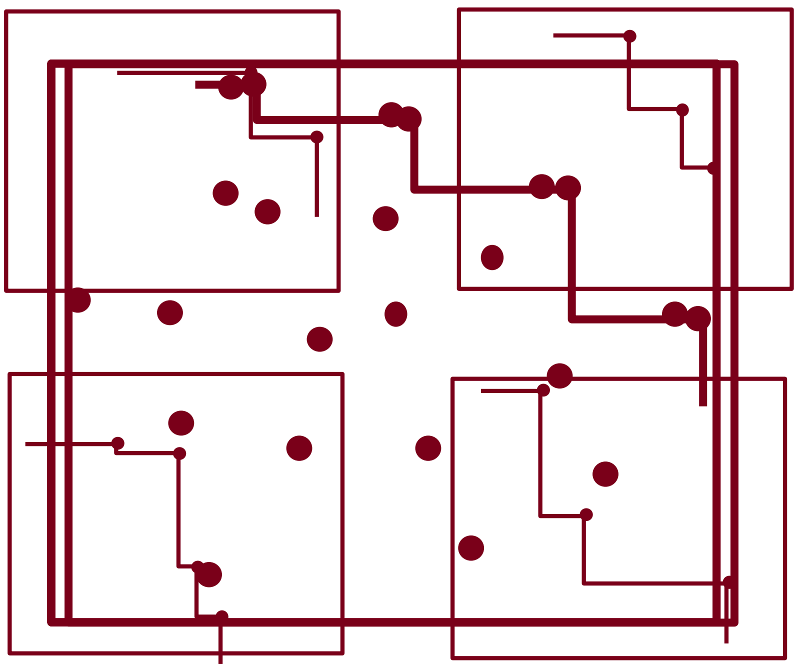
①Partition

②Pruning

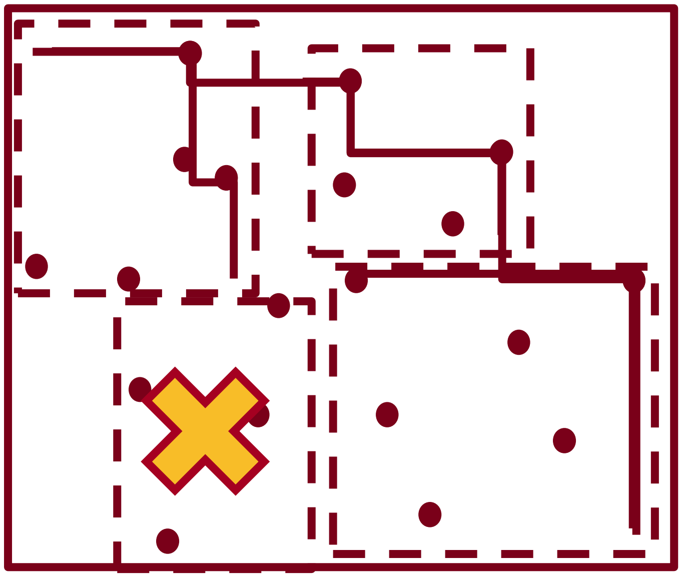
③Local skyline

④Global skyline

Non-spatial partitioning



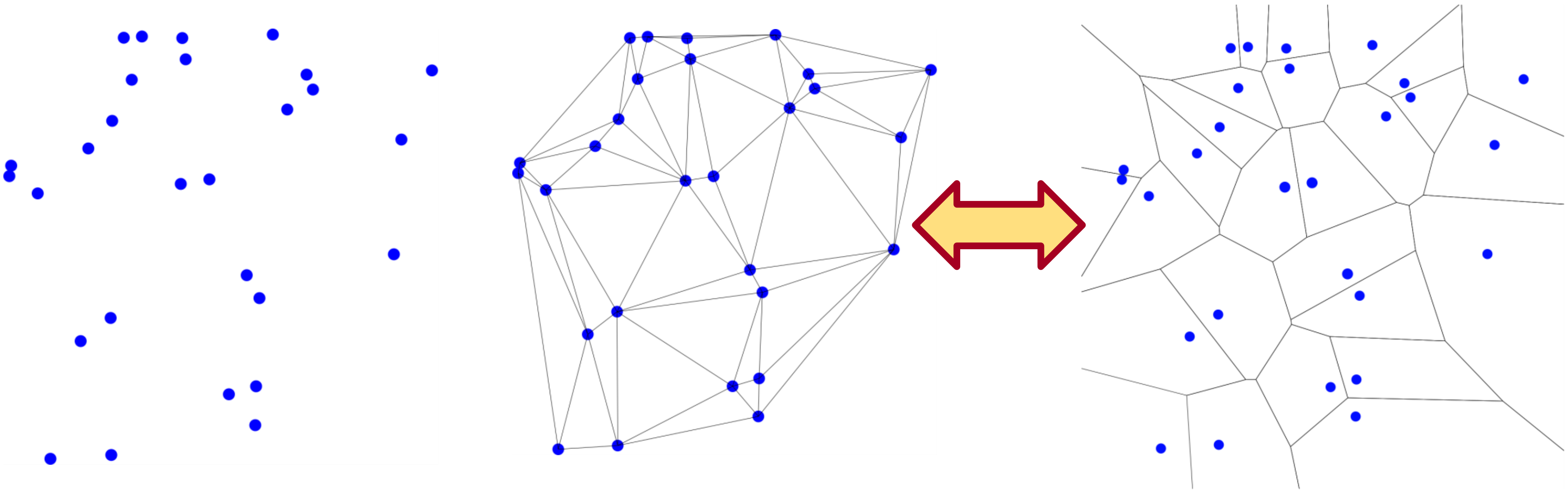
Spatial partitioning



A. Eldawy, *et al*. **"CG Hadoop: Computational Geometry in MapReduce"**.
In SIGSPATIAL, 2013

Voronoi Diagram/ Delaunay Triangulation

Constructs the Voronoi Diagram (VD) or
the Delaunay Triangulation (DT) for a set of points



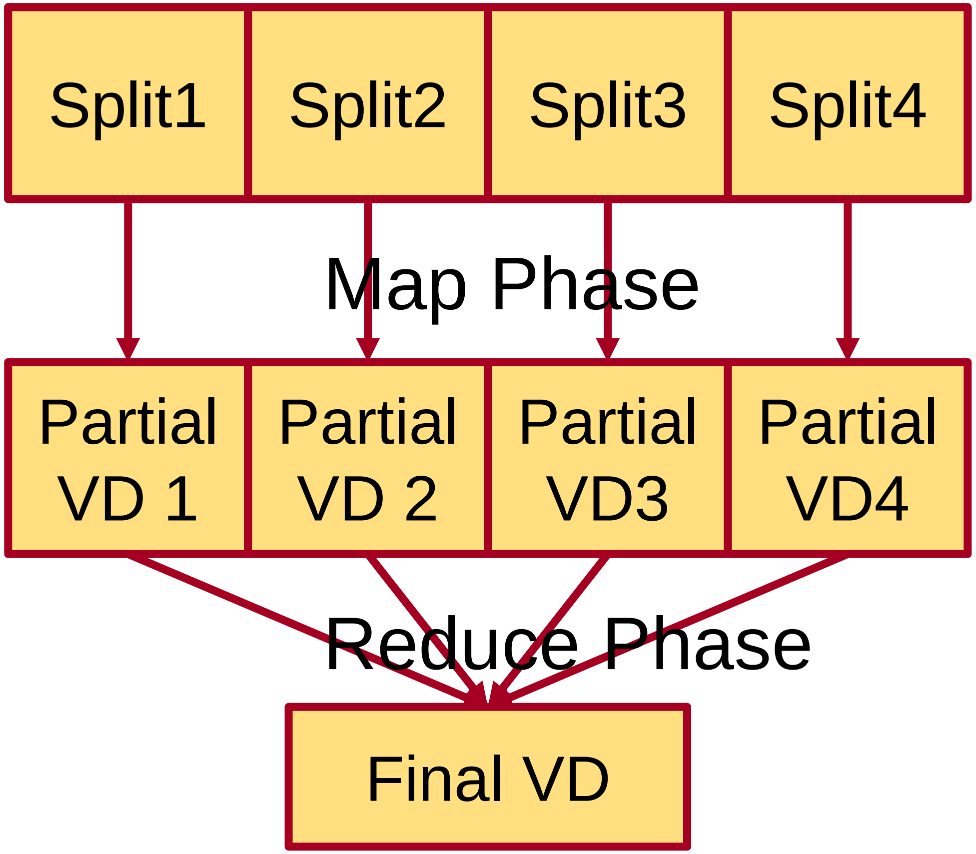
Input

Delaunay Triangulation

Voronoi Diagram

Voronoi Diagram (VD) Construction

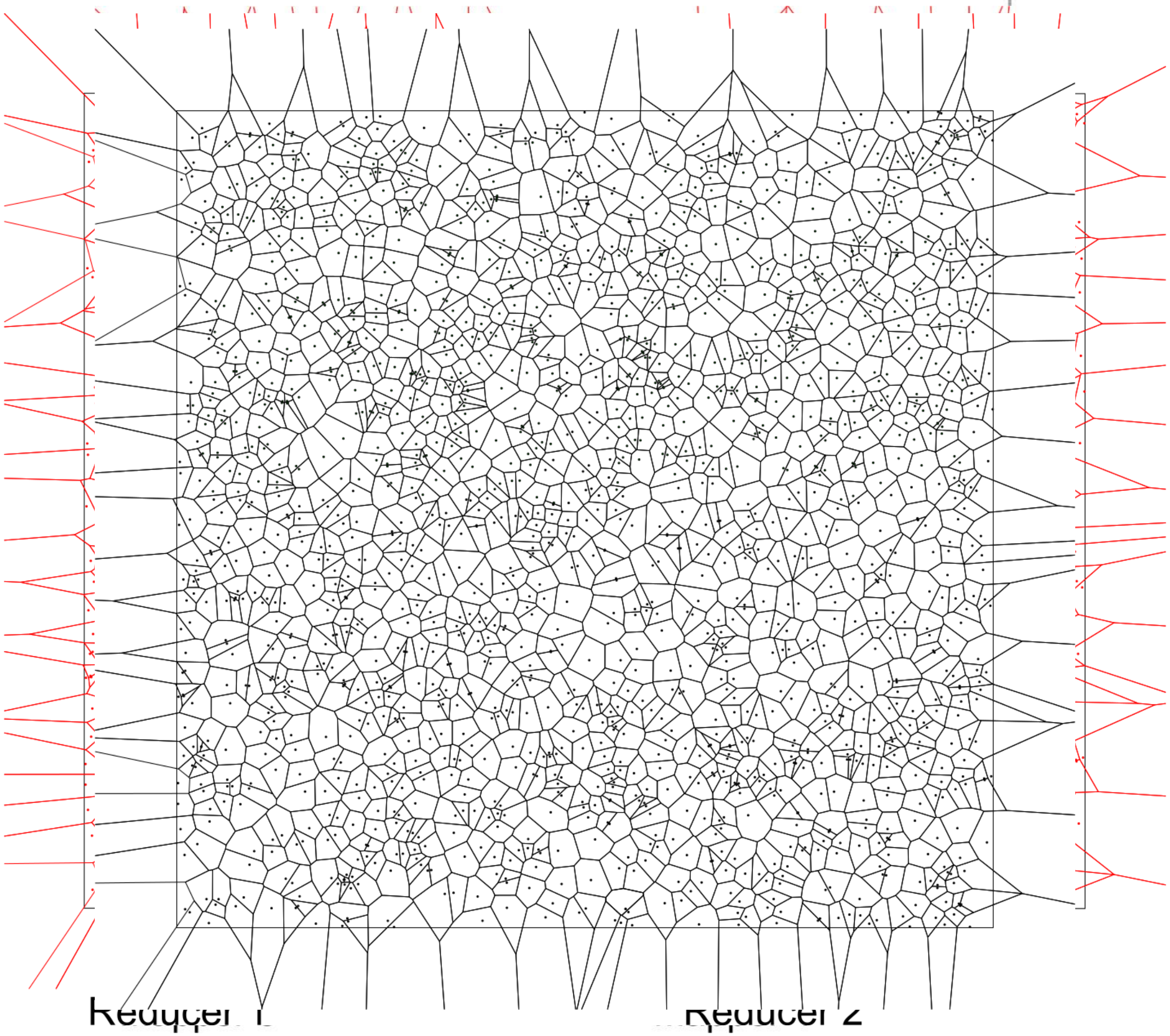
- Assume input points are sorted by one dimension (say x)
- Split the file by the sorted dimension
- Mapper:
  - Construction VD for each split
- Reducer
  - Merge partial VDs into final VD



A. Akdogan *et al,* **"Voronoi-based Geospatial Query Processing with MapReduce"**. In CLOUDCOM, 2010

Voronoi Diagram (VD) Construction

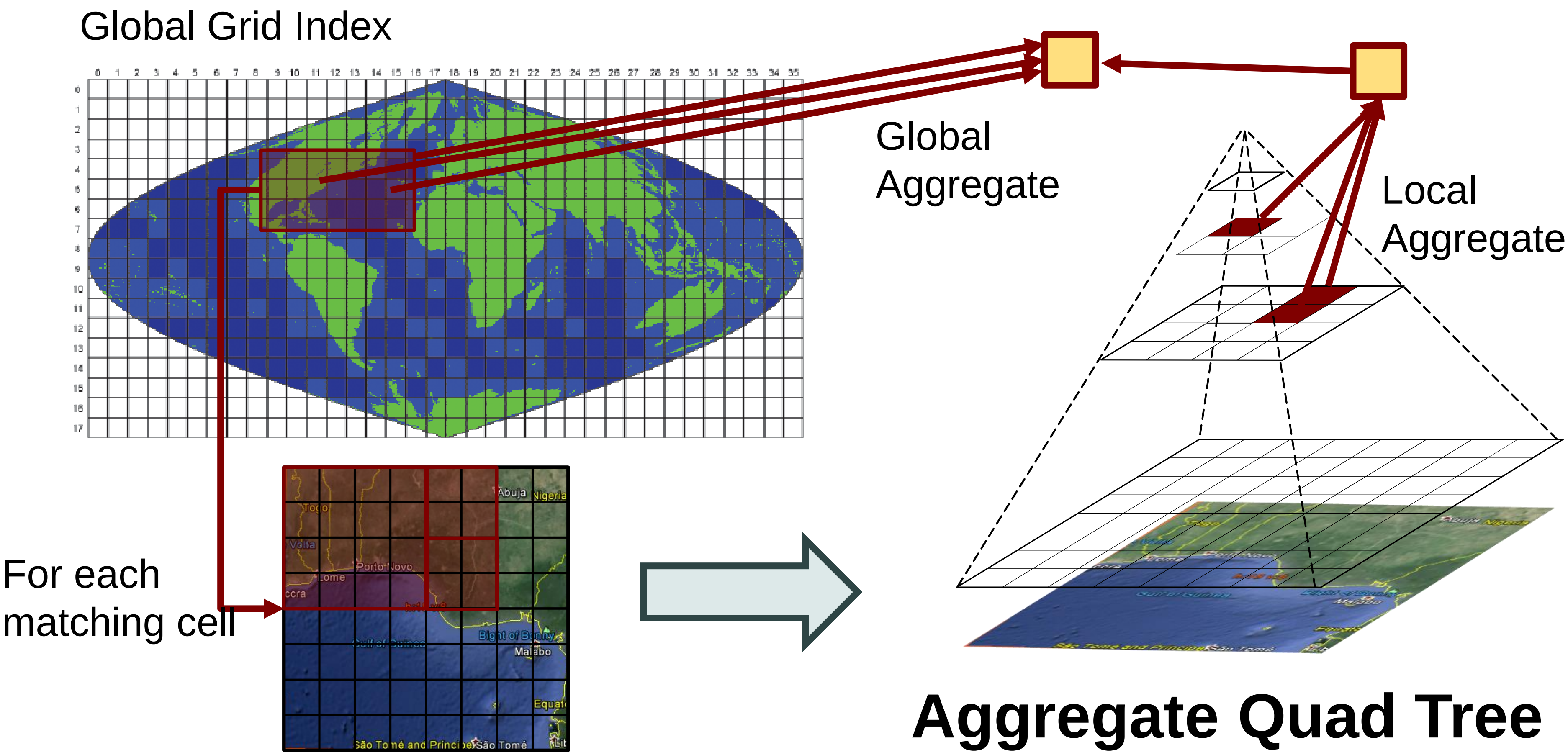
- **Partitioning**
- Local VD
- Pruning
- Vertical Merge
- Pruning
- Horizontal Merge
- Final output



http://aseldawy.blogspot.com/2015/12/voronoi-diagram-and-dealunay.html

Aggregation over Raster Data

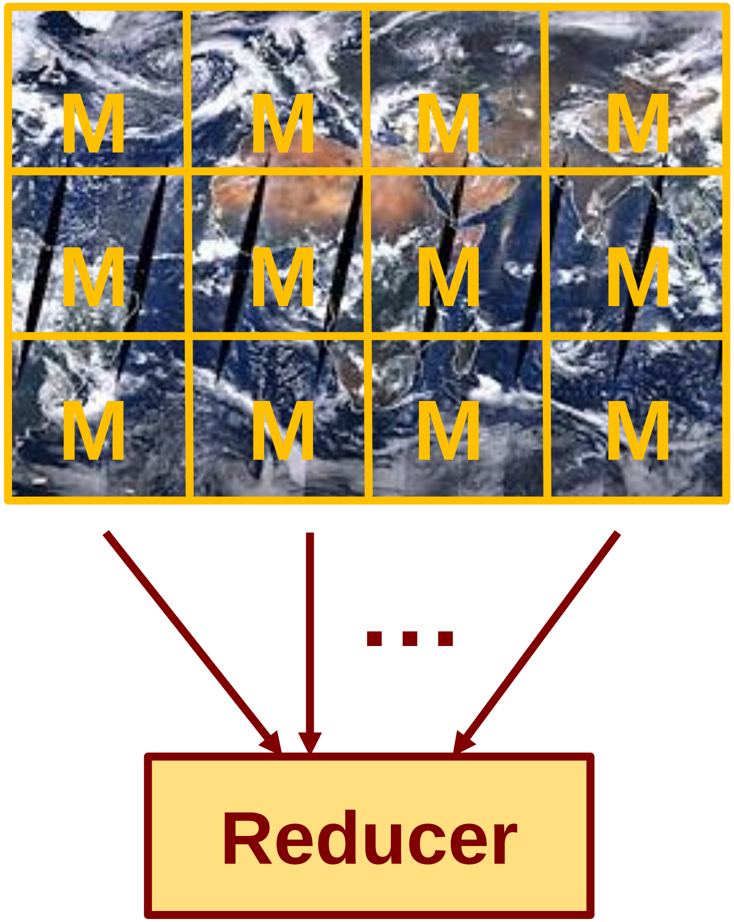
e.g., find the minimum and maximum values in a query range



A Eldawy *et al*, **"SHAHED: A MapReduce-based System for Querying and Visualizing Spatio-temporal Satellite Data"**, ICDE 2015

Image Quality Measurement

- Image quality measurement using MapReduce
- Split the image into tiles
- Map: Assess the quality of each tile
- Reduce: Combine quality measurement of tiles



A. Cary, Z. Sun, V. Hristidis, and N. Rishe. **"Experiences on Processing Spatial Data with MapReduce"**. In SSDBM, 2009

Visualization

Applications
Satellite Imagery, GIS, Microblogs, Medical Imagery, …

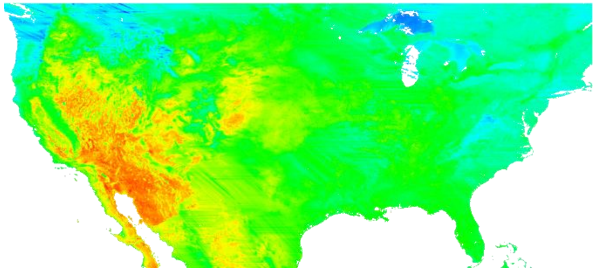
Language

Visualization
Single level and multilevel images

Query Processing
Basic Queries, Spatial Join, and Computational Geometry

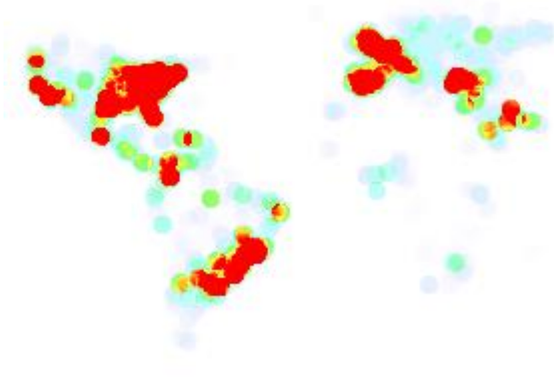
Indexing
Grid, R-tree, Quad tree, K-d tree, …

Spatial Visualization

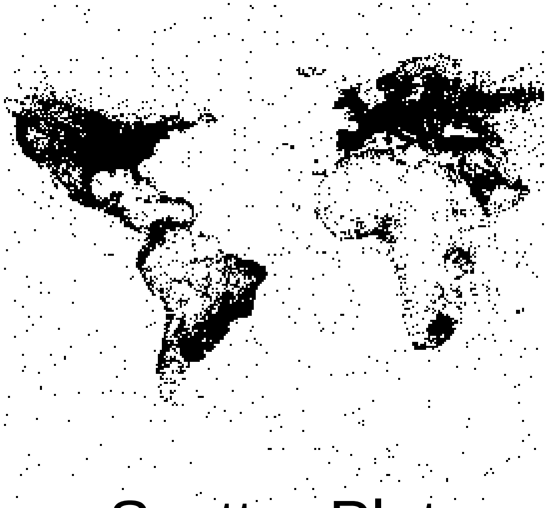


Satellite Data

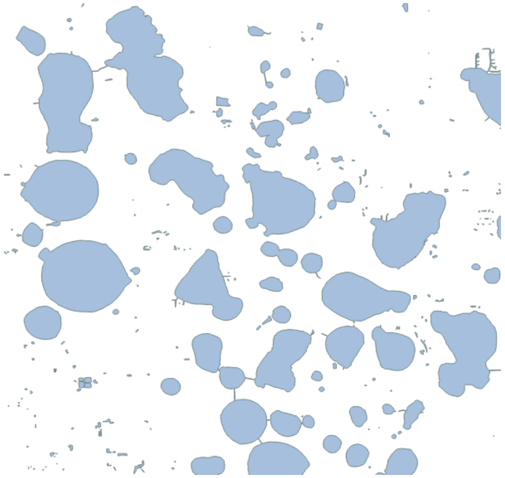
Road Network



Heat Map



Scatter Plot



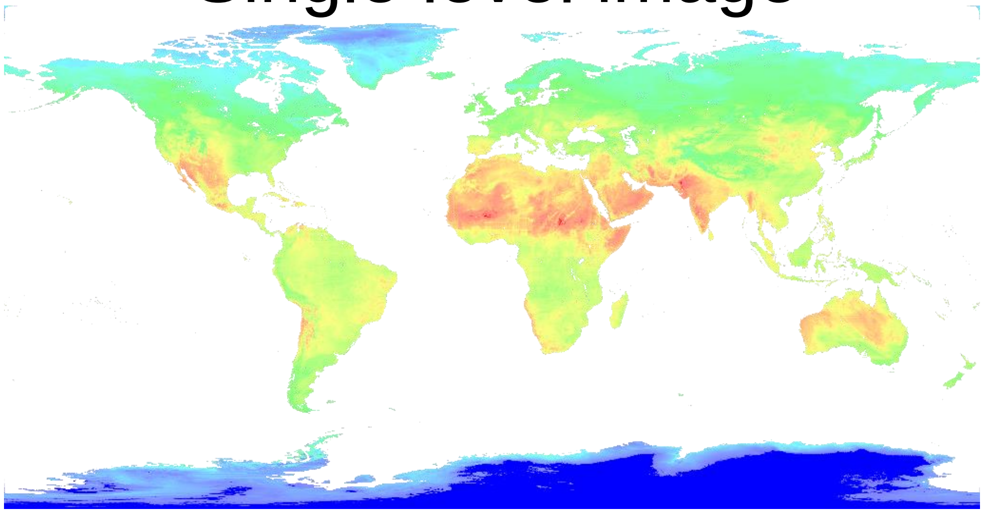
Vector Map

Admin Boundaries

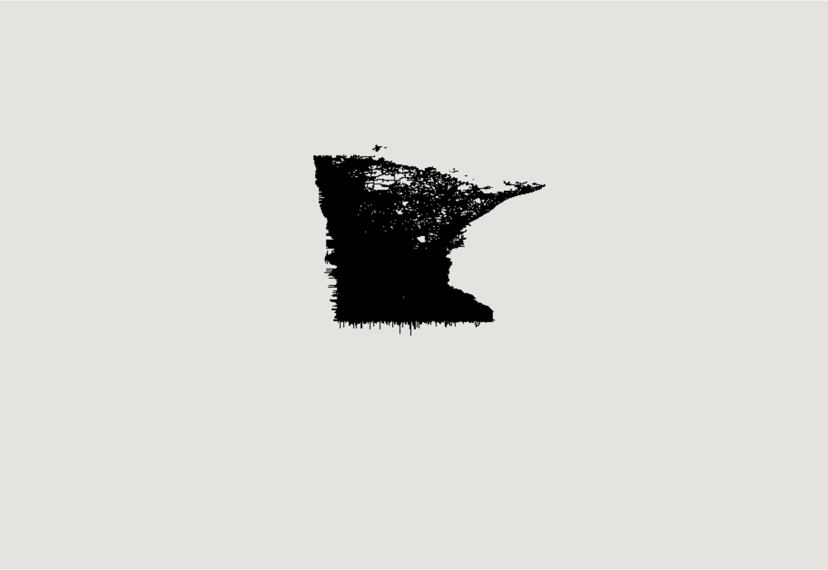
Types of Generated Images

- Single-level image: Fixed resolution
- Multilevel image: Support zoom in/out

Single level image



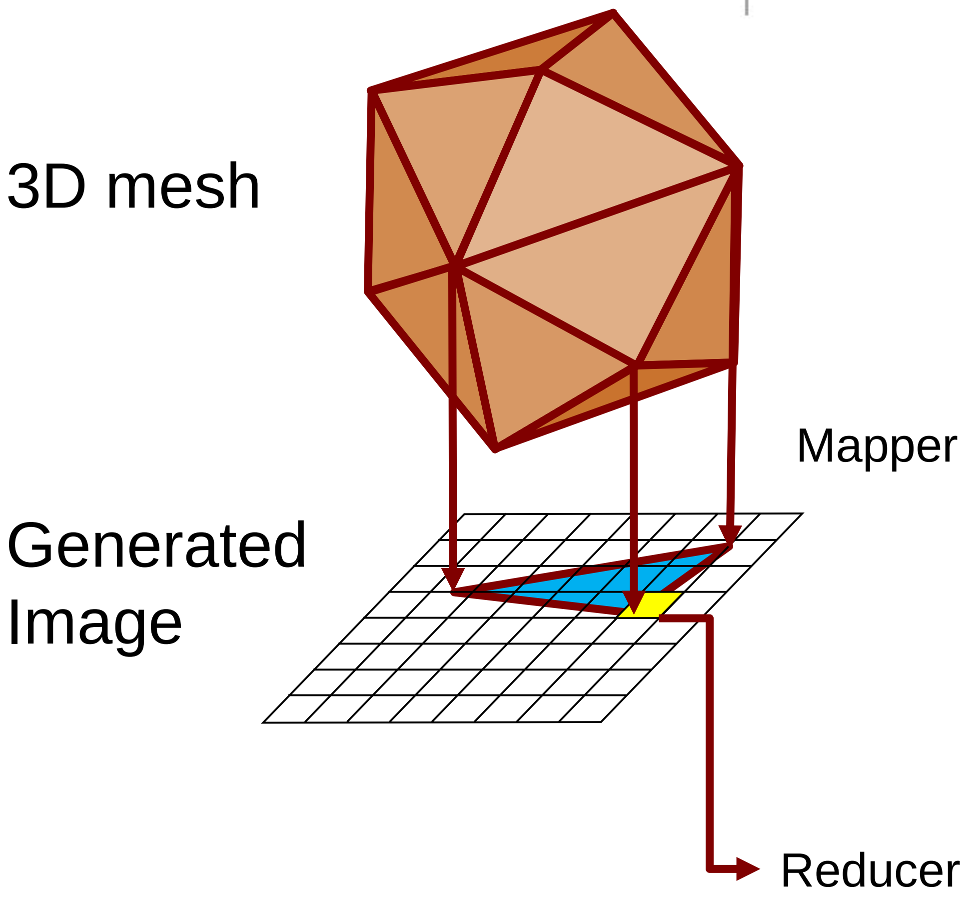
Multi level image



- Challenges
  - Limited resources of one m[illegible] CPU)
  - Generation of giga-pixel images

3D-Mesh Visualization

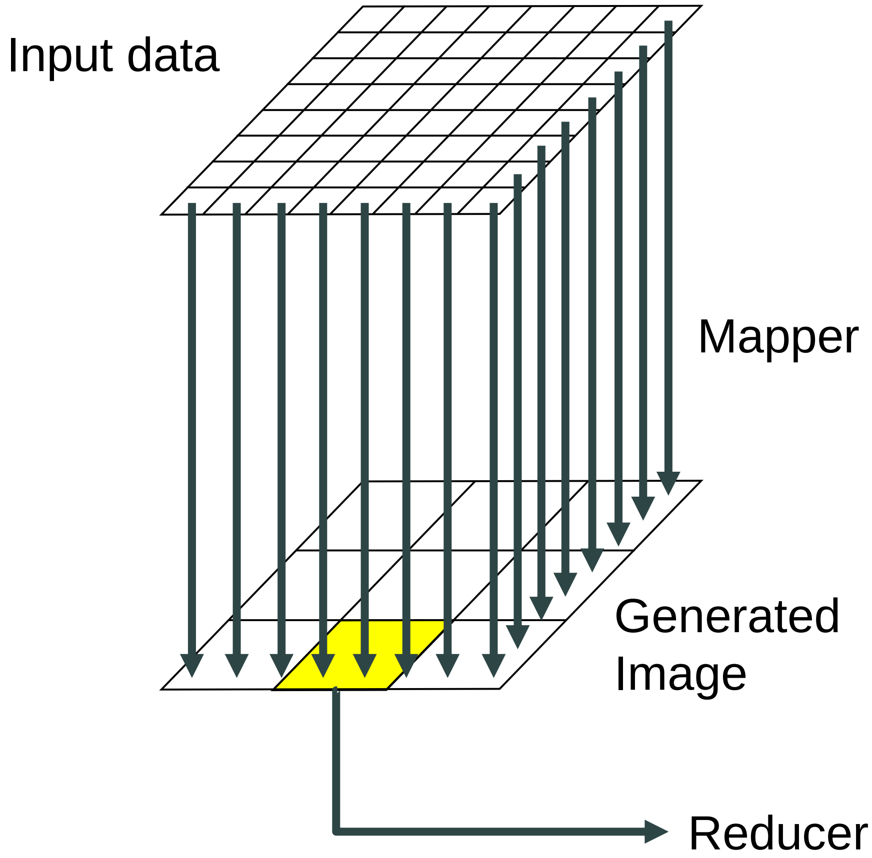
- Mapper:
  - Projects each triangle to the generated image
  - Replicates each triangle to every overlapping pixel
- Reducer:
  - One reducer per pixel
  - Sorts all assigned triangles by z-dimension
  - Generates final color
- Pixel-level partitioning



H. T. Vo. et al. **“Parallel Visualization on Large Clusters using MapReduce”**.
In IEEE Symposium on Large Data Analysis and Visualization, LDAV, 2011

Satellite Heat Maps in SciDB

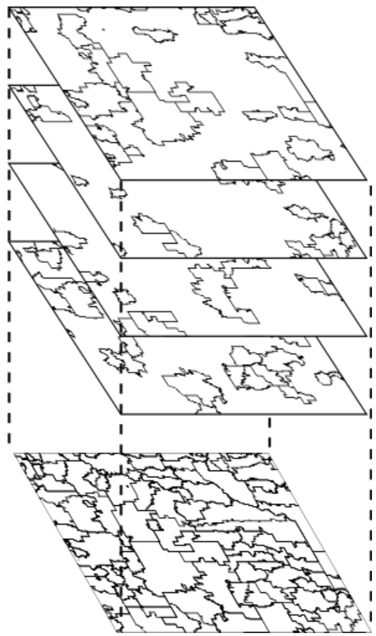
- Mapper
  - Projects each input value to a pixel in the generated image
- Reducer
  - One reducer per pixel
  - Combines all assigned values (e.g., average)
  - Generates a pixel color
- Pixel-level partitioning



G. Planthaber, M. Stonebraker, and J. Frew. **"EarthDB: Scalable Analysis of MODIS Data using SciDB"**. In BIGSPATIAL, 2012

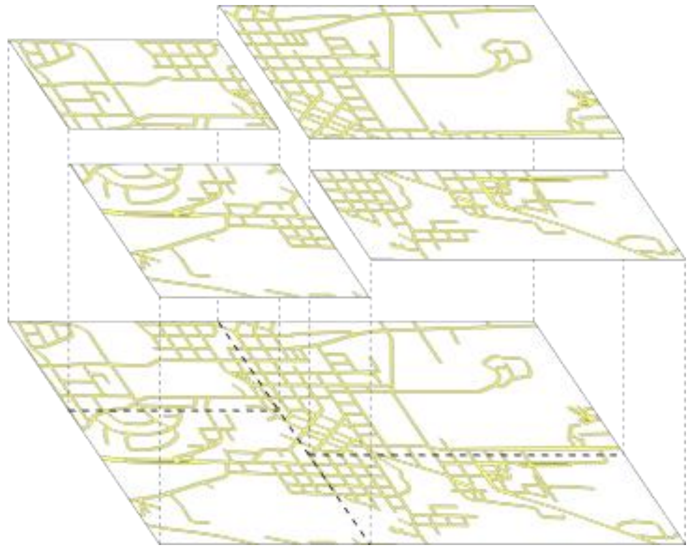
Visualization in HadoopViz

Default Hadoop Partitioning



Overlay

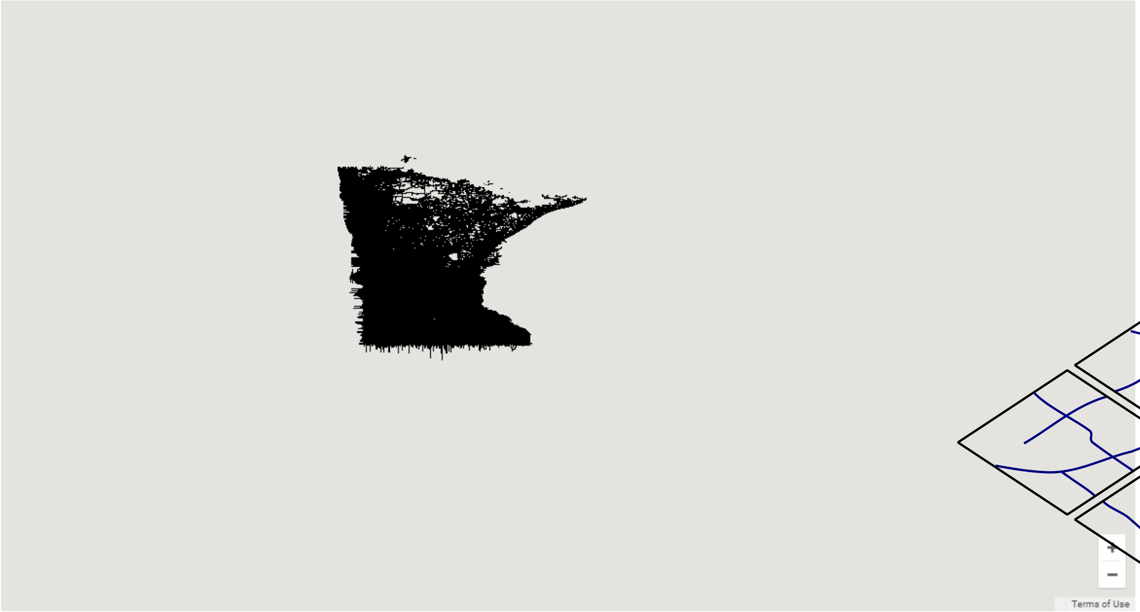
Spatial Partitioning

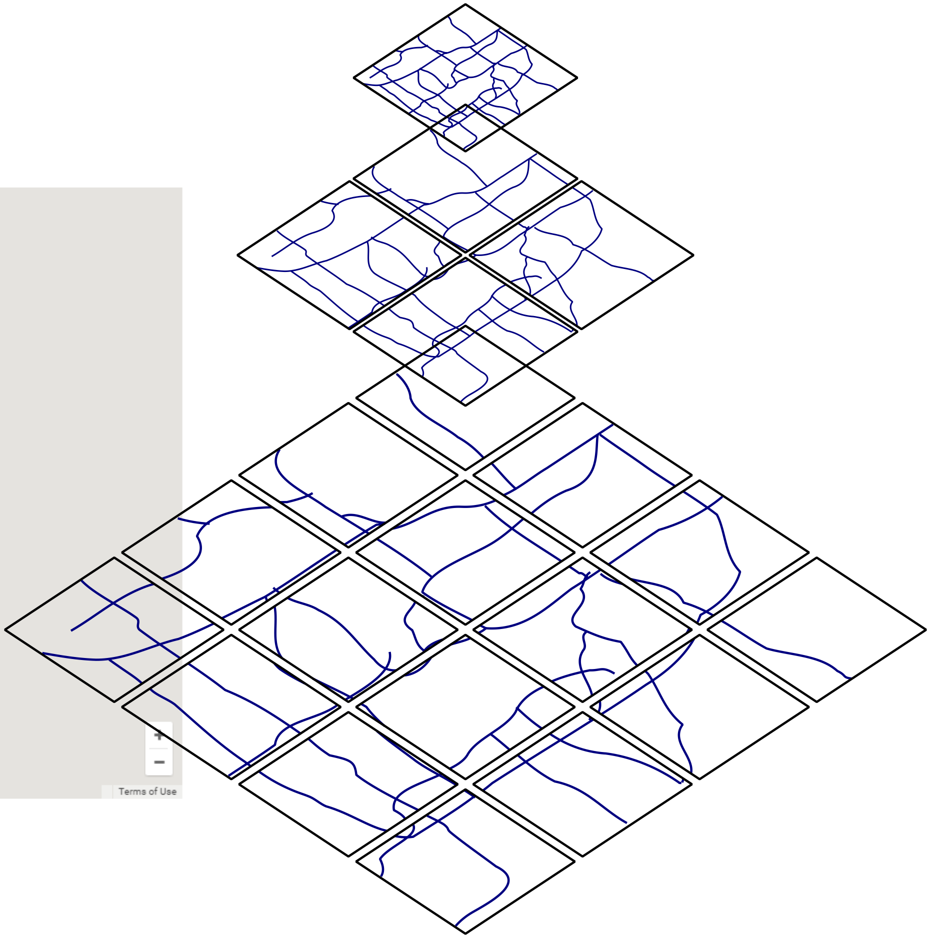


Stitch

A. Eldawy et al, **“HadoopViz: An Extensible MapReduce System for Visualizing Big Spatial Data”**, ICDE 2016

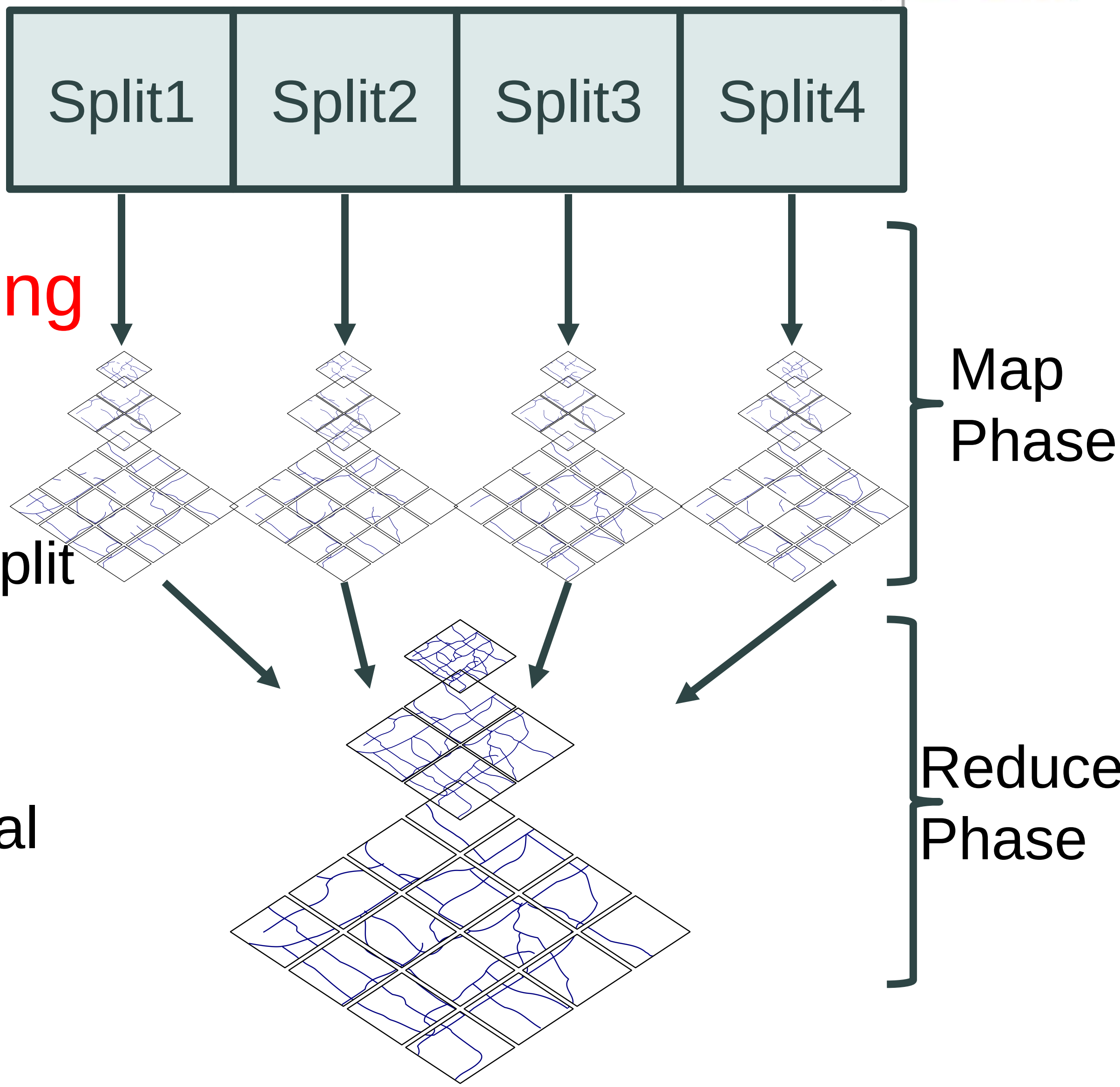
Multilevel Images





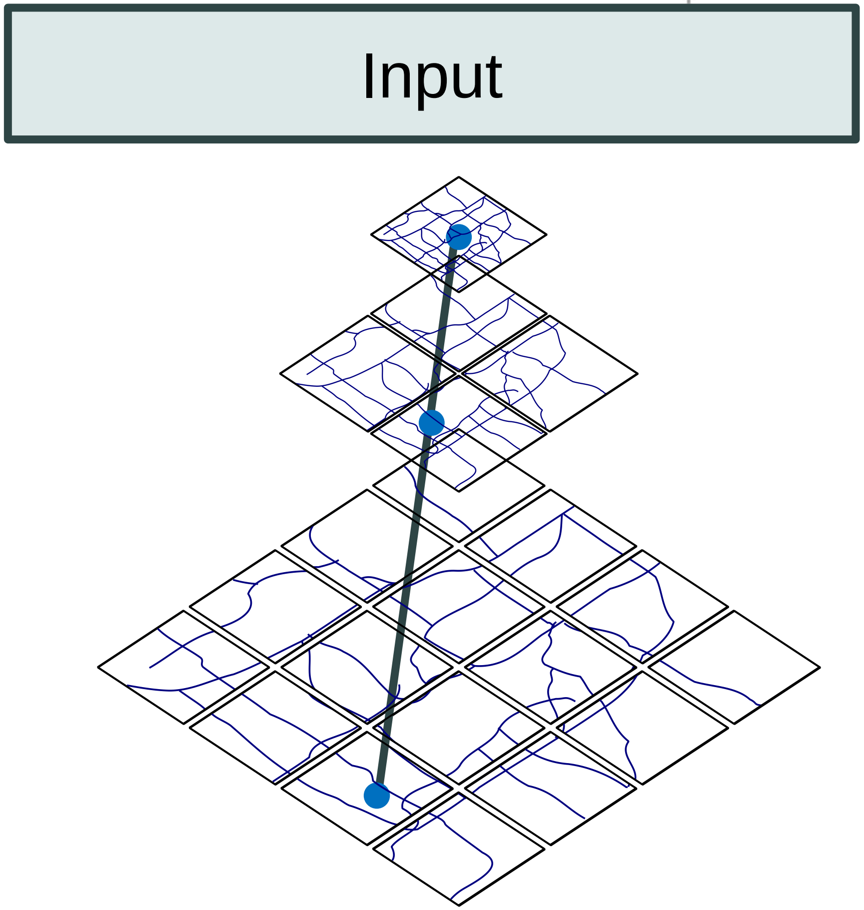
Multilevel Visualization

- Partition using the default Hadoop partitioning
- Mapper:
  - Create a partial pyramid for each split
- Reducer:
  - Merge partial pyramids into a final pyramid



Multilevel Visualization

- Mapper:
  - Multilevel pyramid partitioning
  - Replicate a point to overlapping tiles in each level
- Reducer:
  - Plot an image for each tile
  - Images do not need to be merged



A. Eldawy et al, **"HadoopViz: An Extensible MapReduce System for Visualizing Big Spatial Data"**, ICDE 2016

Language

Applications
Satellite Imagery, GIS, Microblogs, Medical Imagery, …

Language

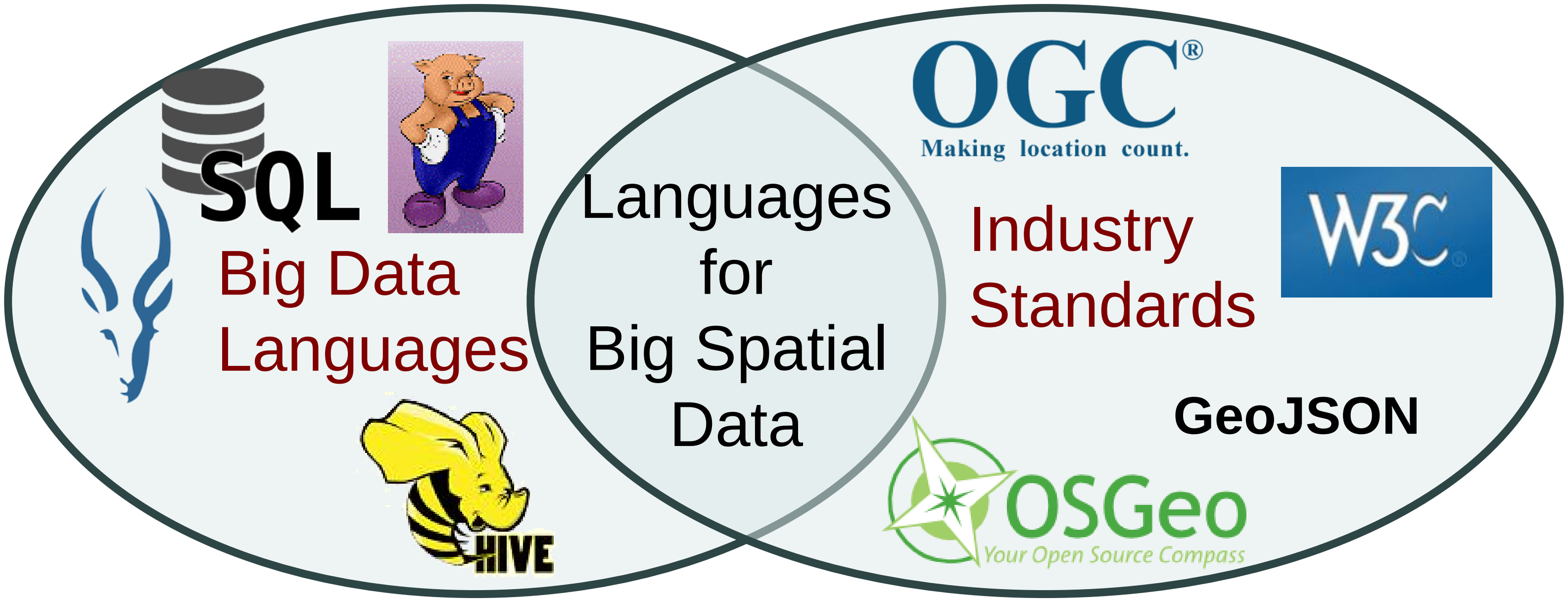
Visualization
Single level and multilevel images

Query Processing
Basic Queries, Spatial Join, and Computational Geometry

Indexing
Grid, R-tree, Quad tree, K-d tree, …

Languages for Big Spatial Data

- Simplifies the system for non-technical user



- Easier to adopt by
  - Existing users of big data systems (e.g., Hadoop, Spark, and Impala)
  - Existing users of traditional systems for spatial data (e.g., PostGIS, Oracle Spatial, and ArcGIS)

Pigeon (by SpatialHadoop)

- Extension to Pig Latin
- OGC-compliant
- Spatial data types
  - E.g., Point, Polygon
- Spatial predicates
- Spatial aggregates

```
FILTER nodes
BY Contains(
  MakeBox(-97.2,43.5,-89.5,49.4),
  MakePoint(node.lon, node.lat));
```

```
zip_codes = LOAD 'zips' AS (zip, city, geom);
zip_by_city = GROUP zip_codes BY city;
zip_union = FOREACH zip_by_city
   GENERATE group AS city, Union(geom);
```

A. Eldawy and M. F. Mokbel. **“Pigeon: A Spatial MapReduce Language”**. ICDE, 2014

GIS Tools for Hadoop (by Esri)

- Extension to Hive QL
- OGC-compliant
- Integrated with ArcMap through plugin tools

```
SELECT counties.name, count(*) cnt FROM counties
JOIN taxi_trips
WHERE ST_Contains(counties.boundaryshape,
  ST_Point(taxi_trips. lon, taxi_trips.lat))
GROUP BY counties.name
ORDER BY cnt desc;
```

http://esri.github.io/gis-tools-for-hadoop/

QL^{SP} (by Hadoop-GIS)

- Extension to Hive QL
- Partial support of OGC-standard operations

```
SELECT ST_Area(ST_Intersection(ta.polygon,tb.polygon))
ST_Area(ST_Union(ta.polygon,tb.polygon)) AS ratio,
ST_Distance(ST_Centroid (tb.polygon),
ST_Centroid(ta.polygon)) AS distance,
FROM markup_polygon ta JOIN markup_polygon tb ON
ST_Intersects(ta.polygon, tb.polygon) = TRUE
WHERE ta.algrithm_uid='A1' AND tb.algrithm_uid='A2' ;
```

A. Aji, *et al*.**"Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce"**. In VLDB, 2013

Applications

Applications

Satellite Imagery, GIS, Microblogs, Medical Imagery, …

Language

Visualization

Single level and multilevel images

Query Processing

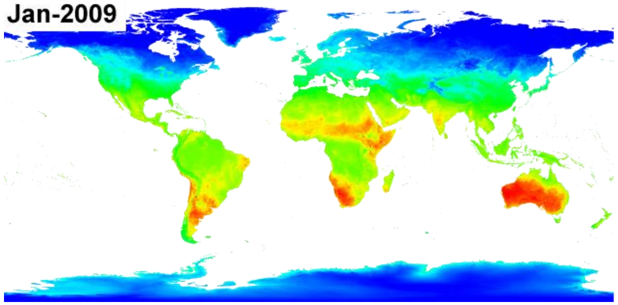
Basic Queries, Spatial Join, and Computational Geometry

Indexing

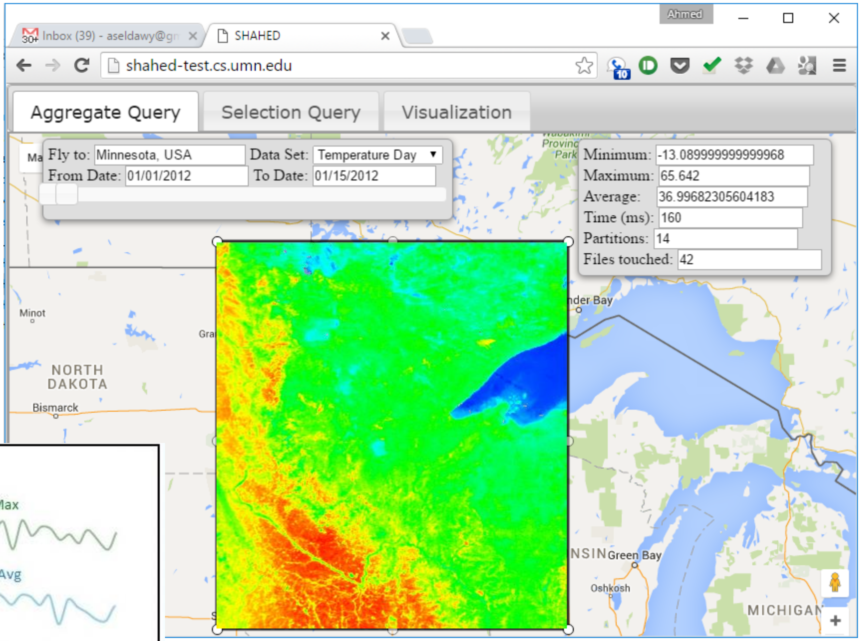
Grid, R-tree, Quad tree, K-d tree, …

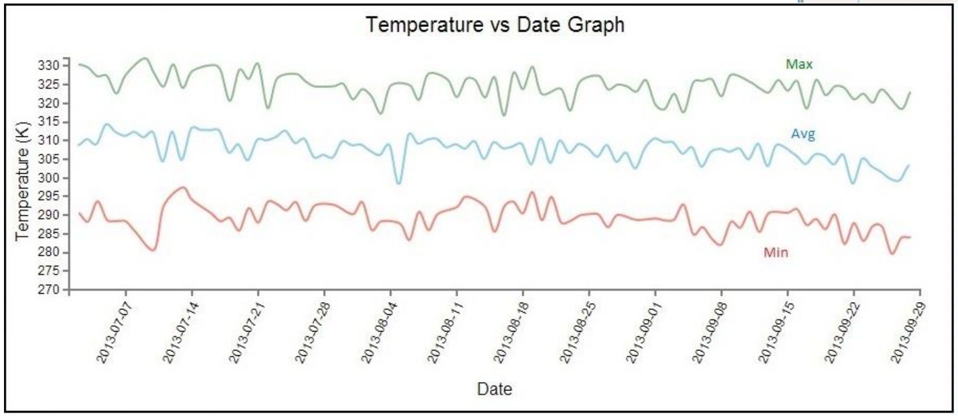
SHAHED – A system for querying and visualizing spatio-temporal satellite data

http://shahed.cs.umn.edu/



Visualize animated heat maps or still images



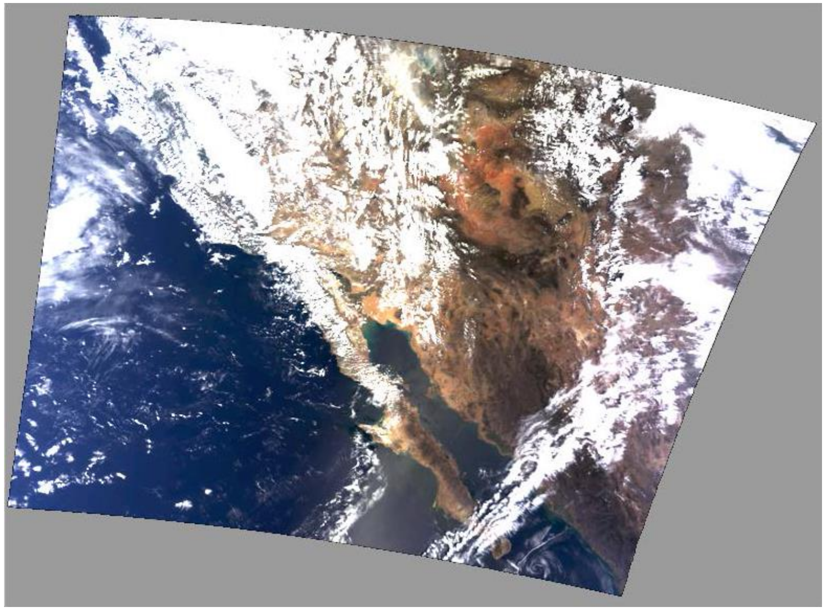


Run spatio-temporal selection and aggregate queries

A. Eldawy *et al*. **"SHAHED: A MapReduce-based System for Querying and Visualizing Spatio-temporal Satellite Data"**, ICDE'15

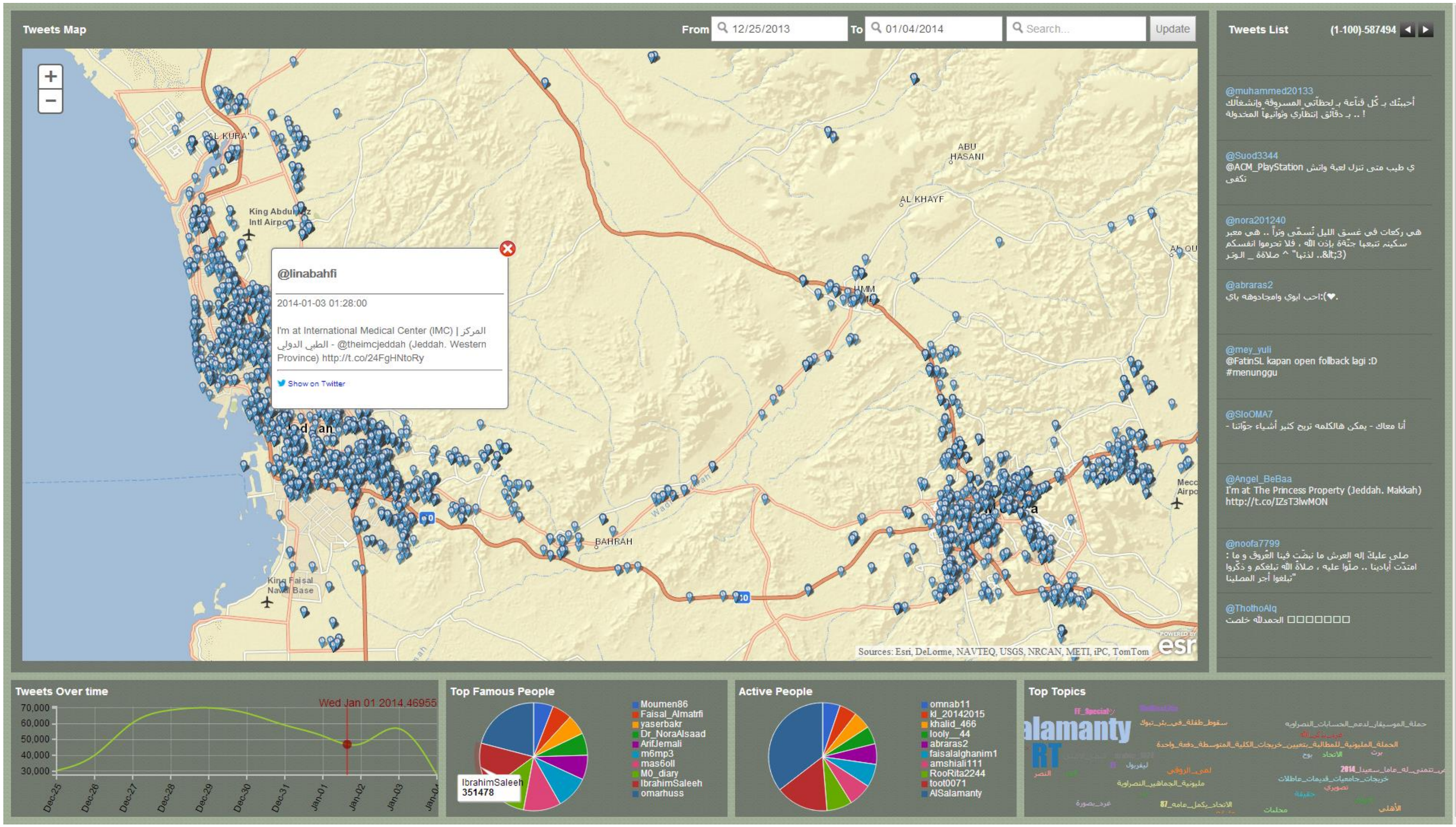
EarthDB: Satellite Data Analysis

- Analyzes and visualizes satellite data using SciDB
- Employs K-d tree partitioning
- Performs analysis queries and visualize the result



G. Planthaber, M. Stonebraker, and J. Frew. **"EarthDB: Scalable Analysis of MODIS Data using SciDB"**. BIGSPATIAL'12.

TAGHREED: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs

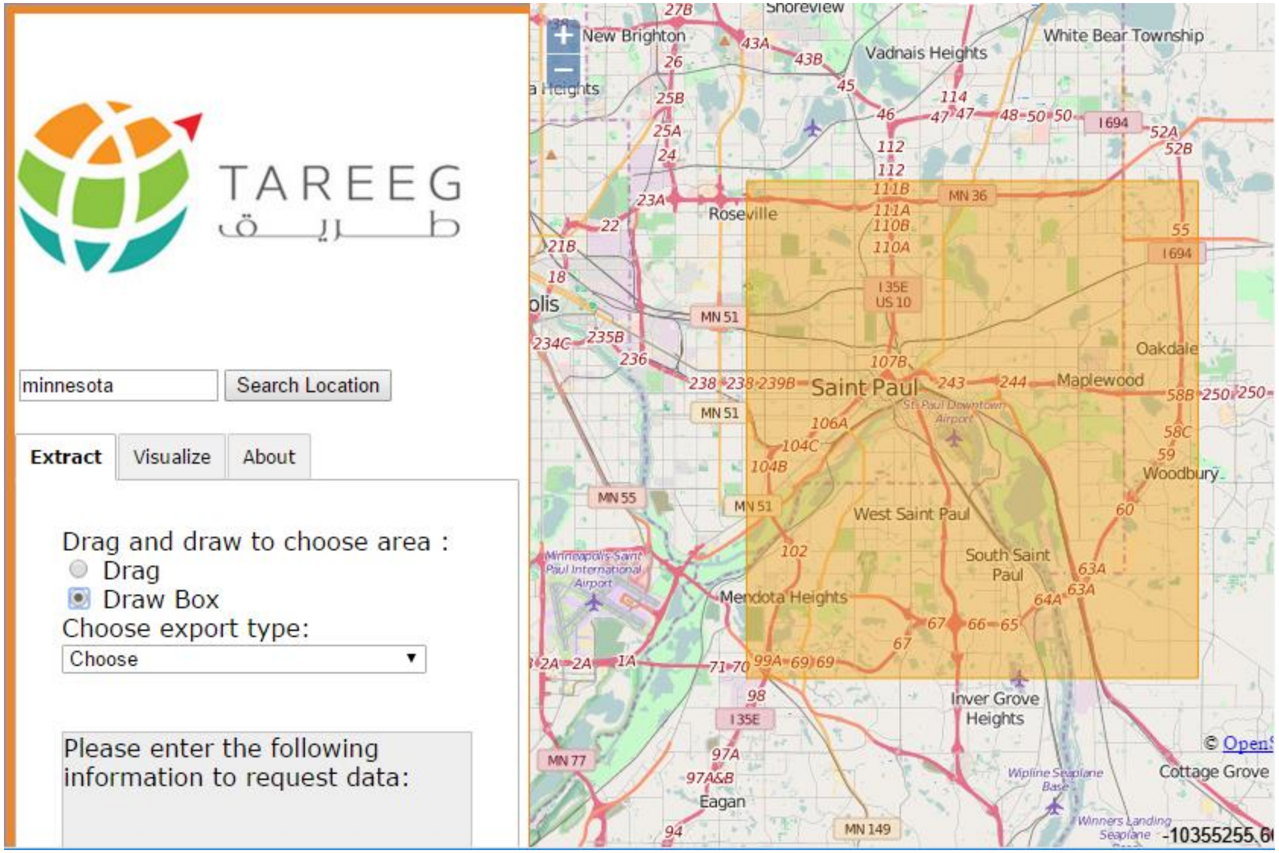


A. Magdy *et al*, **"Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs"**, ICDE 2015

TAREEG – Web-based extractor for OpenStreetMap data using MapReduce

UCR

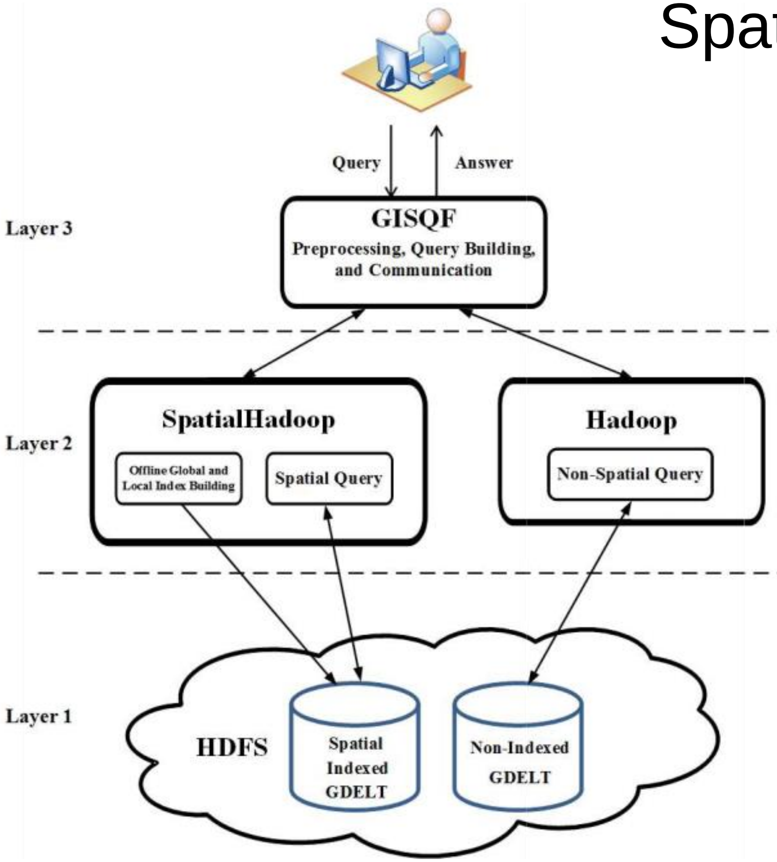
http://tareeg.net/

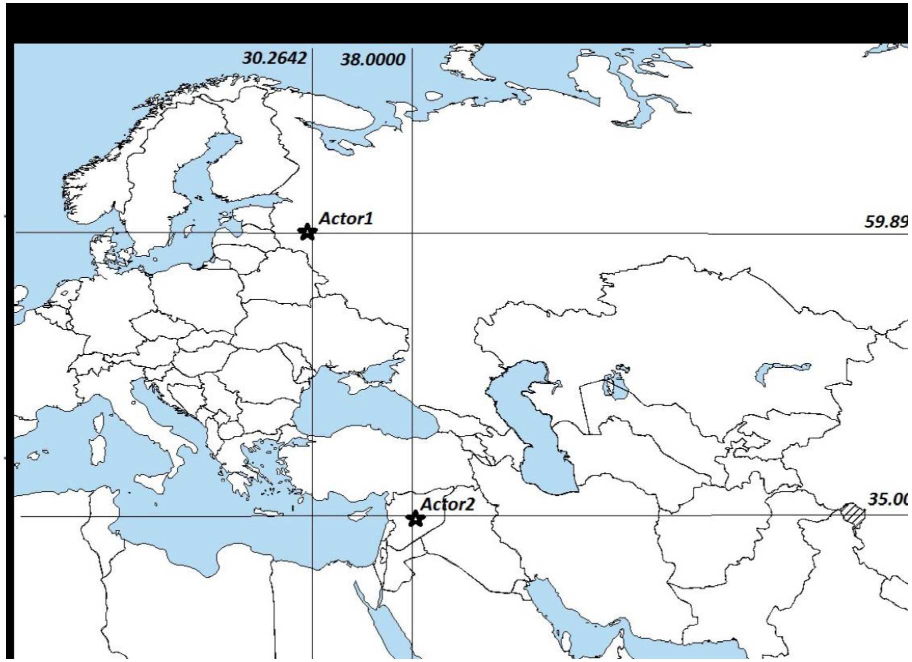


L. Alarabi *et al*, **"TAREEG: A MapReduce-Based Web Service for Extracting Spatial Data from OpenStreetMap"**, SIGMOD'14

GISQF: A SpatialHadoop-based System for Processing Geo-tagged News Events

Spatial selection (point and circle)
Spatial aggregate queries (count)

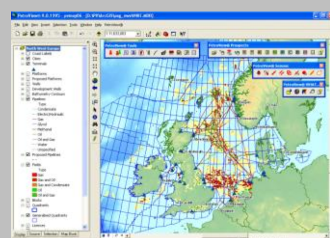


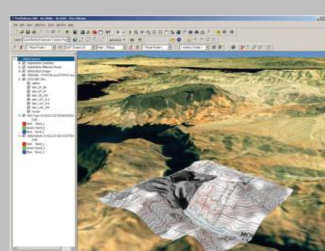


K. Al-Naami et al, **"GISQF: An Efficient Spatial Query Processing System",**
In Proceedings of IEEE Big Data 2014

NASA

Arc GIS
ESRI





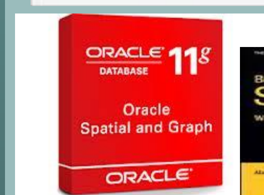
idrisi

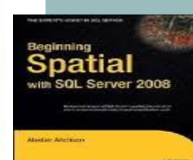
Spatial Hadoop
I can play with your spatial data

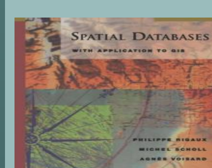
geomesa

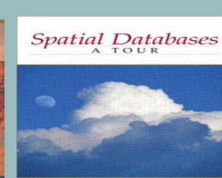
Hadoop-GIS
Spatial Big Data Solutions

PostGIS
Geospatial Objects for PostgreSQL







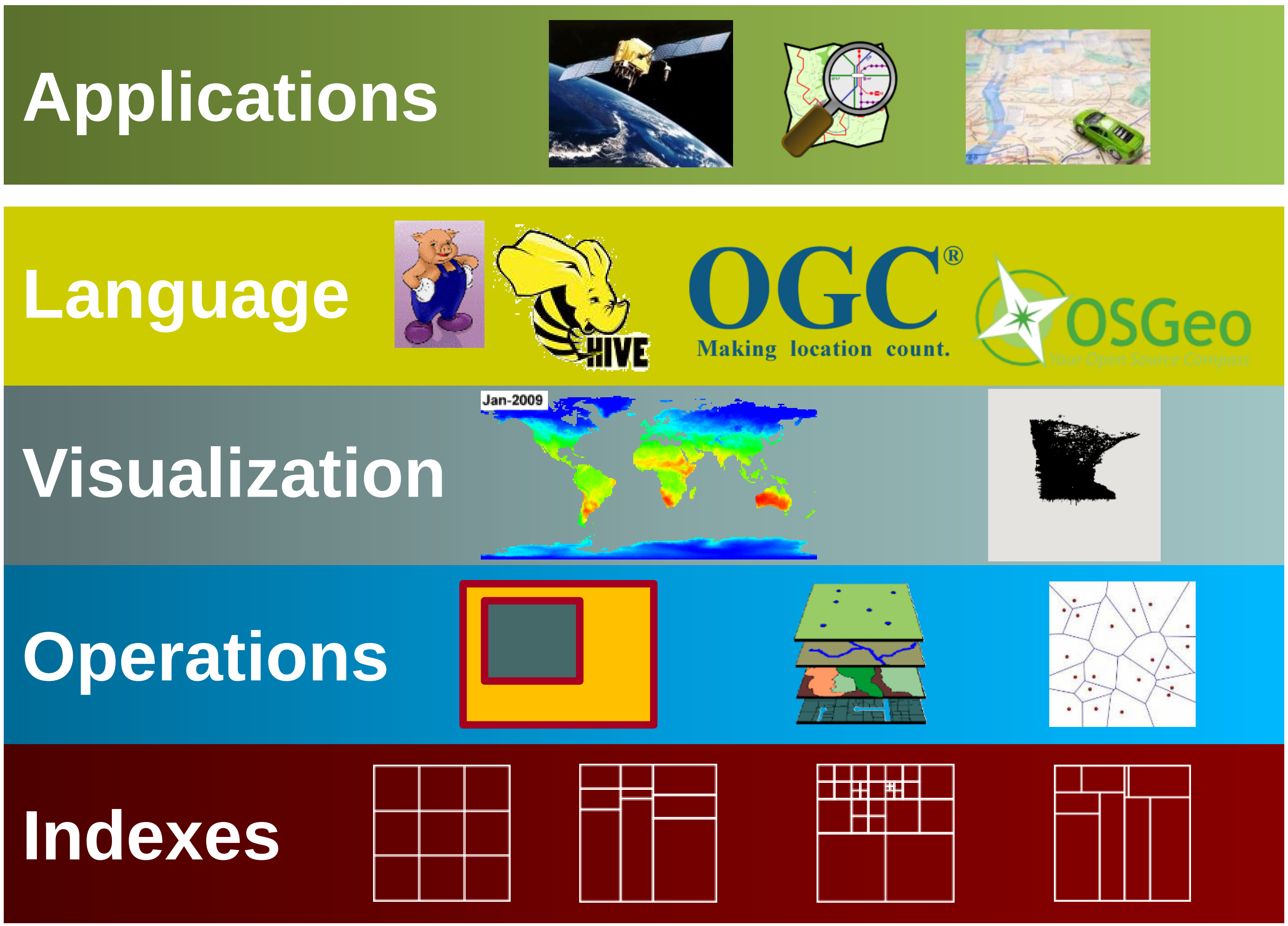


Google
bing
twitter
facebook

MapReduce
amazon web services™ EC2
hadoop
cloudera IMPALA
Spark

ORACLE
IBM DB2
Microsoft SQL Server
PostgreSQL
MySQL
SYBASE | An SAP Company

Summary


Applications
Language
HIVE
OGC®
Making location count.
OSGeo
Visualization
Jan-2009
Operations
Indexes

Thanks You!

Questions?