Spatial Big Data

Joe Niemi

Contents

Spatial Data

- All types of data objects or elements that have geographical information present
- Enables the global finding and locating of individuals or devices
- Also known as geospatial data, spatial information, geographic information

Spatial Data

Raster data

- Geoimages (obtained by satellites for example)
- 3D objects

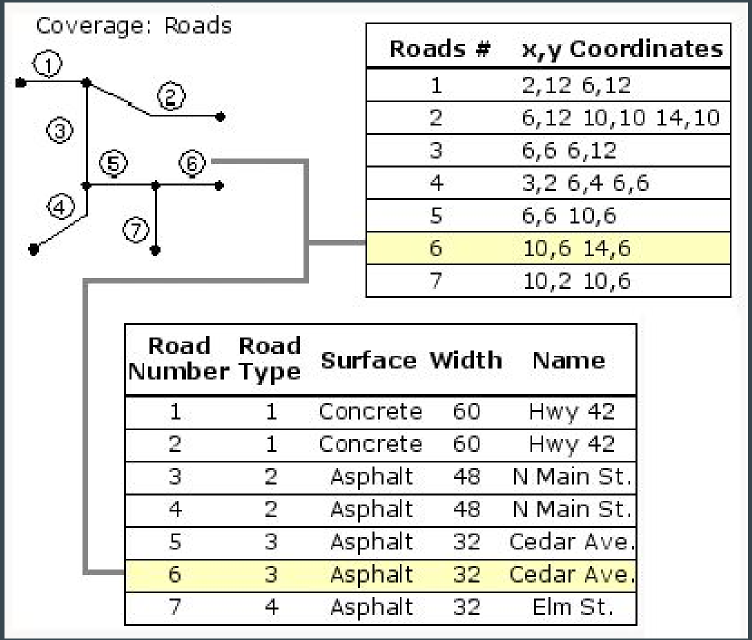
Vector data

- Points, Lines, Polygons

Graph data

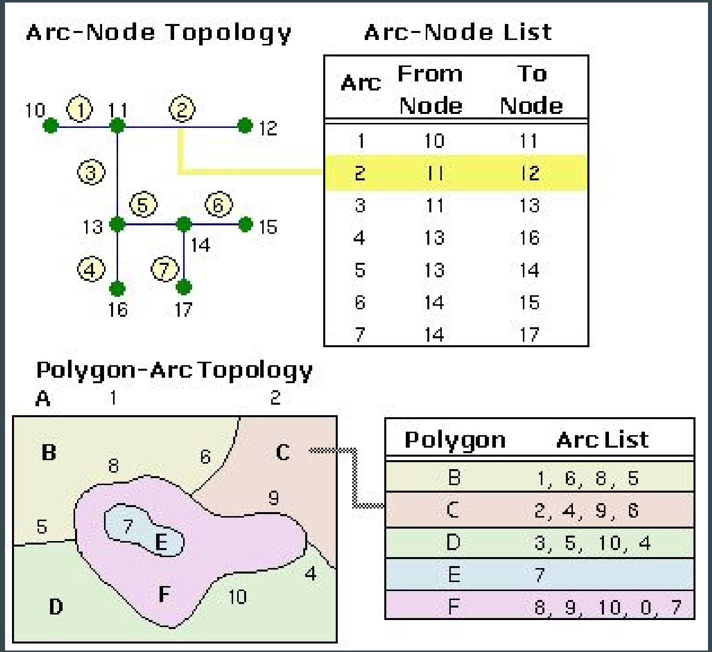
- Road networks (an edge = a road segment and a node = intersection)
- Topological coverage

Topological Coverage



Roads #	x,y Coordinates
1	2,12 6,12
2	6,12 10,10 14,10
3	6,6 6,12
4	3,2 6,4 6,6
5	6,6 10,6
6	10,6 14,6
7	10,2 10,6

Road Number	Road Type	Surface	Width	Name
1	1	Concrete	60	Hwy 42
2	1	Concrete	60	Hwy 42
3	2	Asphalt	48	N Main St.
4	2	Asphalt	48	N Main St.
5	3	Asphalt	32	Cedar Ave.
6	3	Asphalt	32	Cedar Ave.
7	4	Asphalt	32	Elm St.



Arc	From Node	To Node
1	10	11
2	11	12
3	11	13
4	13	16
5	13	14
6	14	15
7	14	17

Polygon	Arc List
B	1, 6, 8, 5
C	2, 4, 9, 6
D	3, 5, 10, 4
E	7
F	8, 9, 10, 0, 7

Contains both the location and attribute data

Spatial Big Data

Spatial Big Data exceeds the capacity of commonly used spatial computing systems

- due to volume, variety and velocity

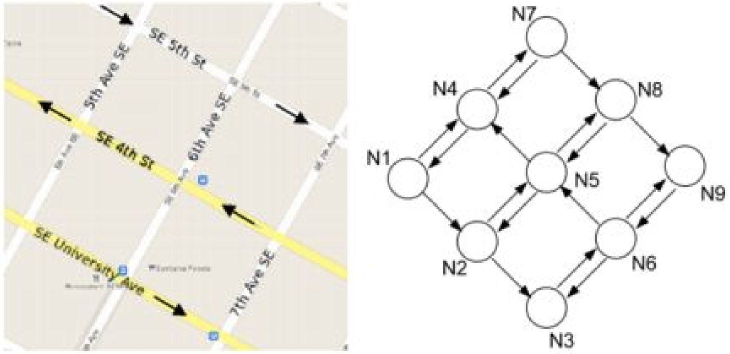
Spatial Big Data comes from many different sources

- satellites, drones, vehicles, geosocial networking services, mobile devices, cameras

A significant portion of big data is in fact spatial big data

Types of Spatial Big Data

- Speed every minute for every road-segment
- GPS trace data from cell-phones
- Engine measurements of fuel consumption (can be estimated from fuel levels, distance travelled and engine idling from engine RPM)
- Greenhouse gas emissions



(a) A Road Map [22] (b) Graph Representation

Nodes

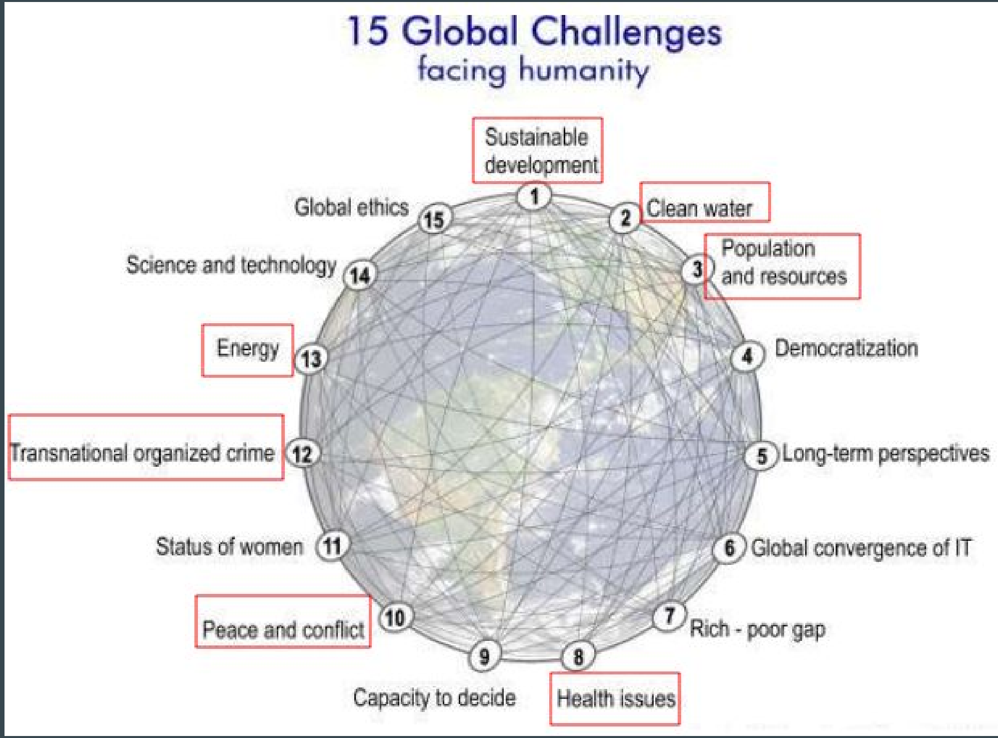
NID
N1
N2
N3
N4
N5
N6
N7
N8
N9

Edges

EID	From	To	Speed	Distance
E1	N1	N2	35mph	0.075mi
E2	N1	N4	30mph	0.075mi
E3	N2	N3	35mph	0.078mi
E4	N2	N5	30mph	0.078mi
E5	N3	N6	30mph	0.077mi
E6	N4	N1	30mph	0.075mi
E7	N4	N7	30mph	0.078mi
E8	N5	N2	30mph	0.078mi
...	...	...	...	...

(c) Tabular Representation of digital road maps

Motivation



15 Global Challenges
facing humanity
Sustainable development
1
Clean water
2
Population and resources
3
Democratization
4
Long-term perspectives
5
Global convergence of IT
6
Rich - poor gap
7
Health issues
8
Capacity to decide
9
Peace and conflict
10
Status of women
11
Transnational organized crime
12
Energy
13
Science and technology
14
Global ethics
15

Motivation

SBD or GIS (Geographic Information System) helps with

- Better decision making
- Saves cost from greater efficiency

From esri 's ArcGIS: "Just about every problem and situation has a location aspect."

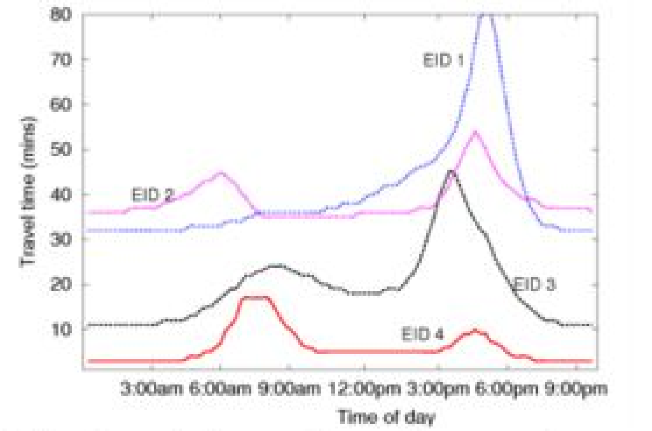
- analyze spatial connections
- get information in real time
- spot location-related patterns that might previously have been undetected

Use cases for Spatial Big Data

1) Eco routing
2) Tracking Endangered Species
3) Better crop production, reducing costs
4) Detecting extreme events

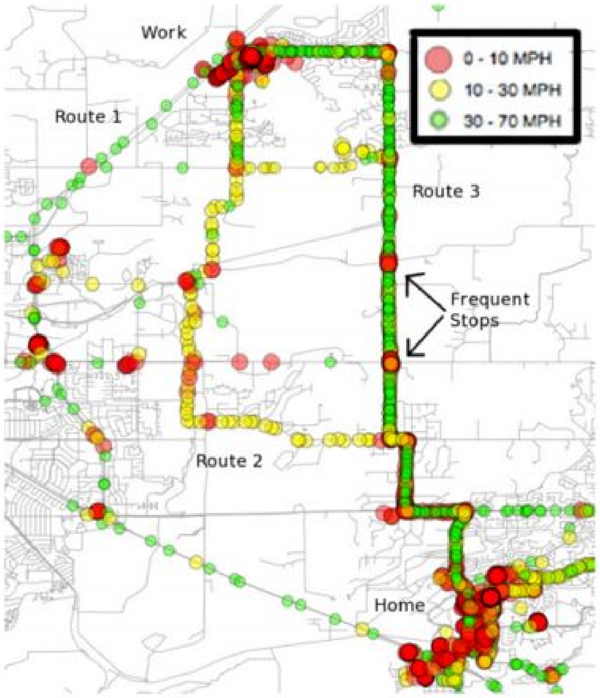
Eco routing

- Next generation routing service
  - avoids congestion
  - reduces idling at red lights
  - avoids left turns
- Estimation: in 2020 about $600 billion is saved an
- Takes into account various datasets
  - real-time and historic traffic data of engine measurements
  - speed-limits
  - road types
  - “rush hour vs non-rush hour”



(a) Travel time along four road segments over a day.

Eco routing

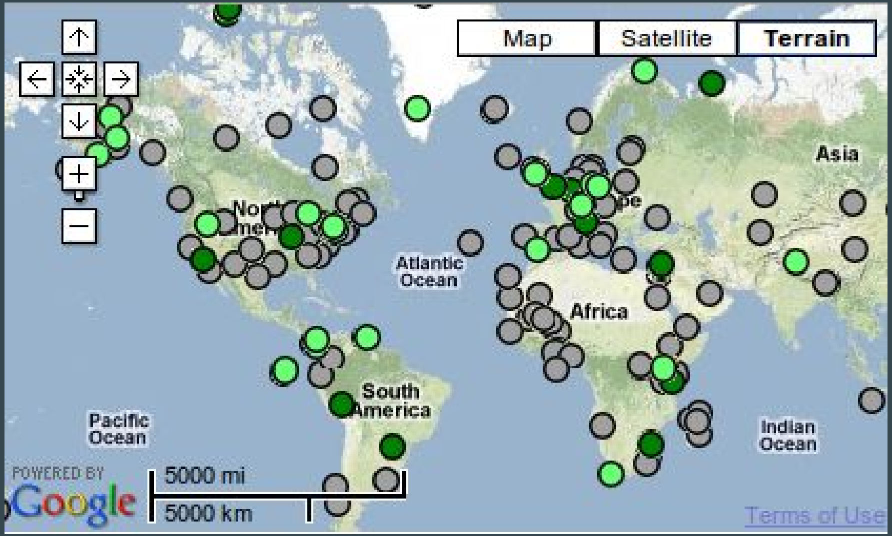


A commuter's GPS tracks over three months

Tracking endangered species

2013: 970 studies over 250 contributors, 41,170 tracks and 61 million locations

Movebank: a free online database of animal tracking data



Better crop production

“If you can grow crop fast in these circumstances, query for similiar places”

Detecting extreme events

- Earthquakes
- Wildfires
- Flooding
- Other calamities

How to detect

- Built-in motion detectors in mobile phones
- Using unstructured data sets can be used such as **tweets**

Future

- New Datasets -> need to rapily integrate new datasets and algorithms

- Computational cost increases as the diversity of Spatial Big Data grows

- Easy to collect, sensors (or sensor networks) are becoming more and more common (Internet of things)

Features of Spatial Big Data

- Access of data depends on the daytime of where it is used
- Changes dynamically
- Recent Spatial Big Data is usually being generated at a very high speed

Challenges of Spatial Big Data

1) Retaining computational efficiency
2) Storing Spatial Big Data into the cloud
3) Applying new data when Spatial Big Data or change old data => repartitioning is needed

Contents

1) Introduction
   - what is Spatial Big Data?
   - motivation
   - use cases
2) **Cloud partitioning**
3) PAIRS (A scalable Spatial Big Data analytics platform)
4) AQWA (Adaptive Query-Workload-Aware partitioning of Spatial Big Data)
5) Summary

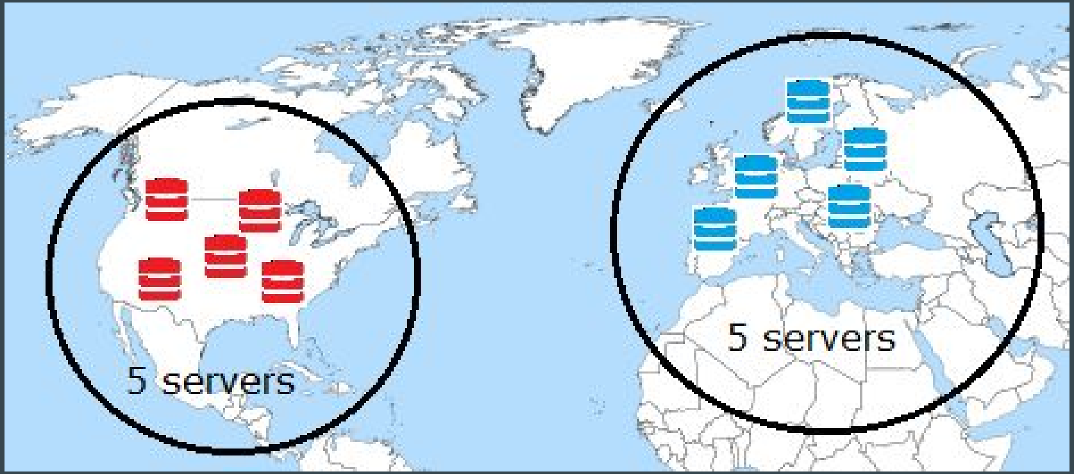
Cloud partitioning of Spatial Big Data

- If partitions are not being accessed, servers remain idle and the user is still charged.
- Most of the existing partitioning approaches co-locate frequently accessed data together to minimize distributed transactions
- Cloud providers often offer time-based pricing models -> users are getting charged even when servers idle or have low CPU usage

Bad example: partitioning of Spatial Big Data

5 servers store data in Europe, 5 servers store data in USA

=> half of the servers are idle for almost a day.

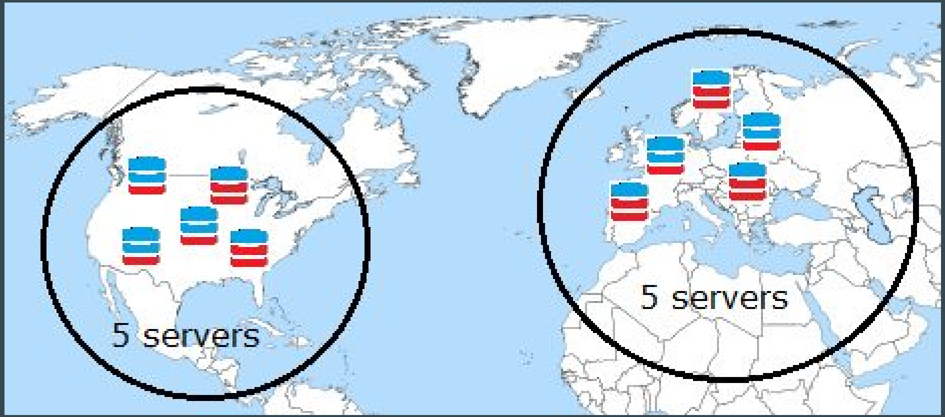


Good example: partitioning of Spatial Big Data

10 servers store data with diverse access patterns to minimize server idle-time

=> Main drawback: Lag or latency problems due to data communication cost

We need a **cache** for servers in Europe to contain frequently accessed data partitions in USA and vise versa

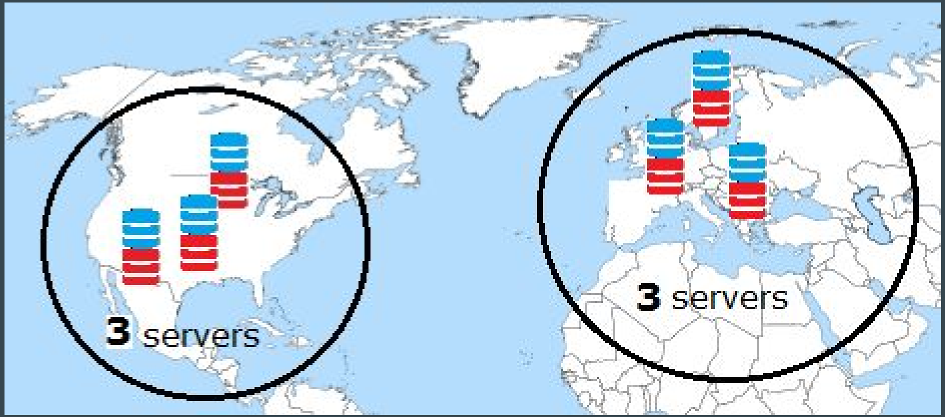


Good example: partitioning of Spatial Big Data

6 servers store data with diverse access patterns to minimize server idle-time

=> Main drawback: Lag or latency problems due to data communication cost

We need a **cache** for servers in Europe to contain frequently accessed data partitions in USA and vise versa



2. Cloud partitioning

Efficient partitioning method

1) Split dataset to partitions based on spatial proximity

- minimizes query throughput

2) Find partitions of diverse access patterns and combine them

- minimizes server idle time and maximizes server utilization

A flatness metric is used to find best possible pair. It shows how diverse access patterns are.

Tabu search algorithm is used that takes into account the history of moves and prevents non-improving moves from happening

Saves up to 40% cost

An easier way to maximize server utilization

In Amazon, based on user defined rules, scale down to a cheaper server if CPU usage is less than 40 percent

- does not take into account server idle-time (they still have to pay for the cheapest server)

Contents

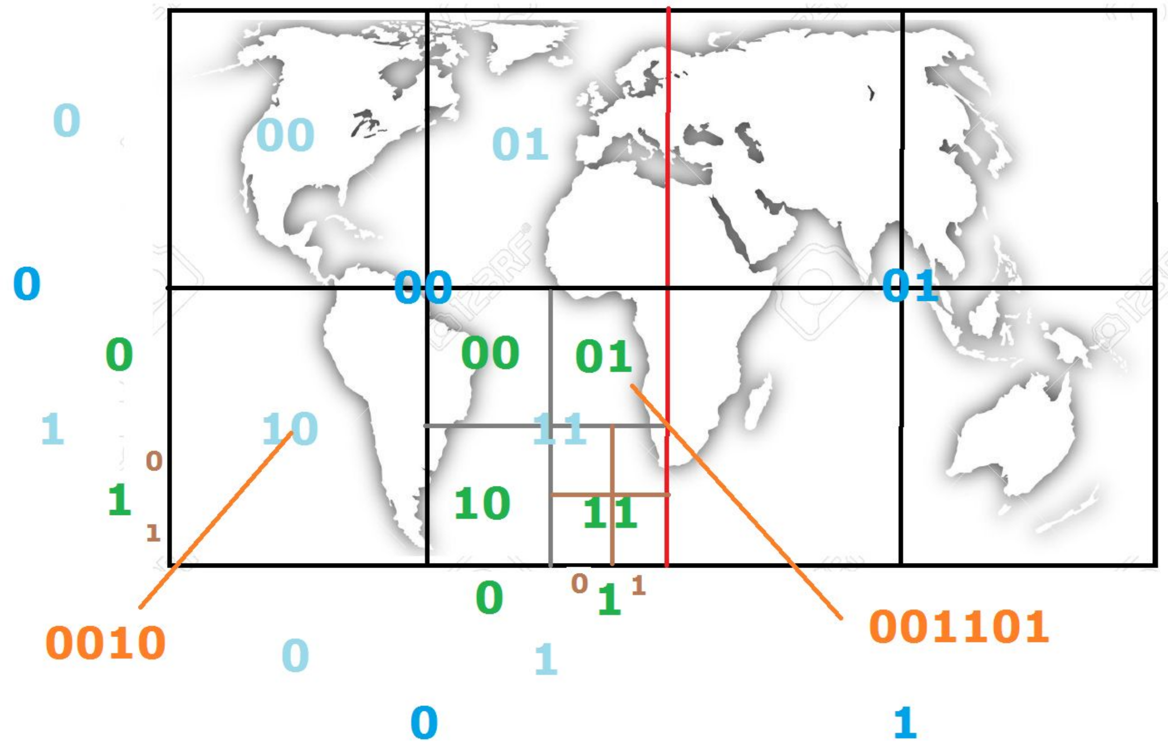
PAIRS

is a cloud service deployed on top of Hadoop and HBase

- PAIRS = Physical Analytics Integrated Repository and Services
- Automatically updates, joins and homogenizes historical and real-time spatial big data that is then available for real-time modeling and analytics
- Data is indexed globally
- Data queries of an area or a single point
  - **parallelized** by **MapReduce**
  - for example a query for a single point (latitude, longitude) for a data layer with daily information for 10 year period, can be retrieved in less than 1 second.

3. PAIRS (A scalable Spatial Big Data analytics platform

Global indexing



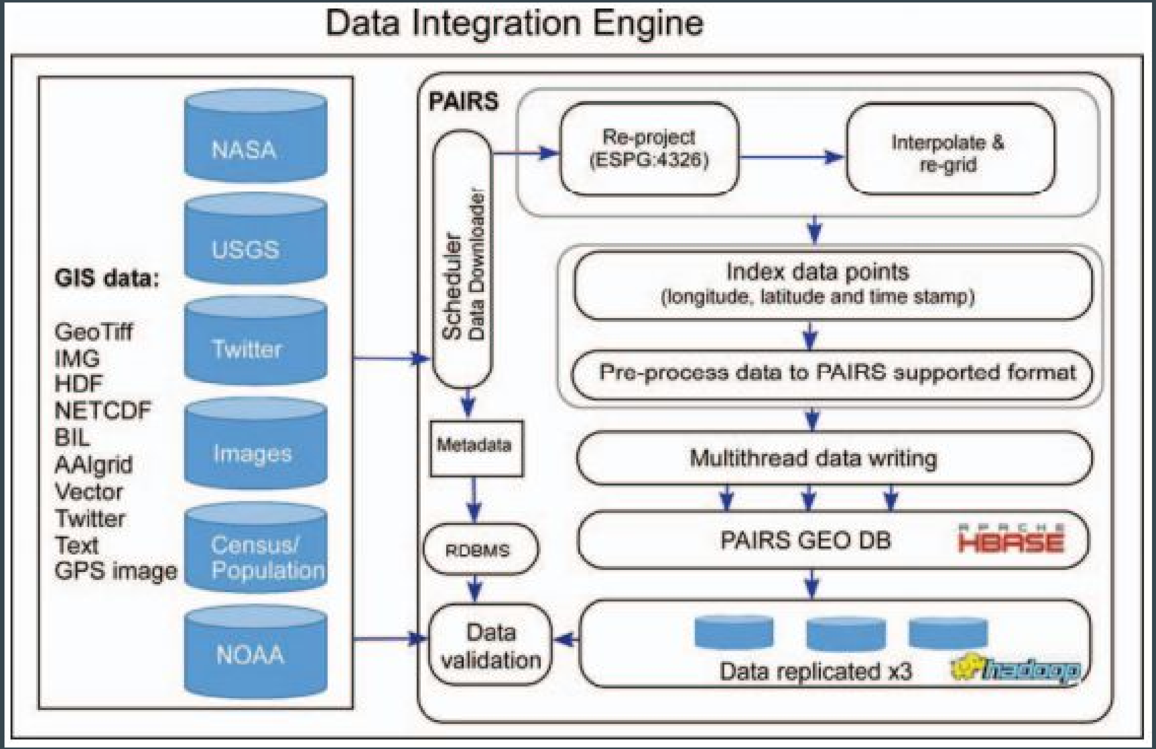
3. PAIRS (A scalable Spatial Big Data analytics platform

PAIRS

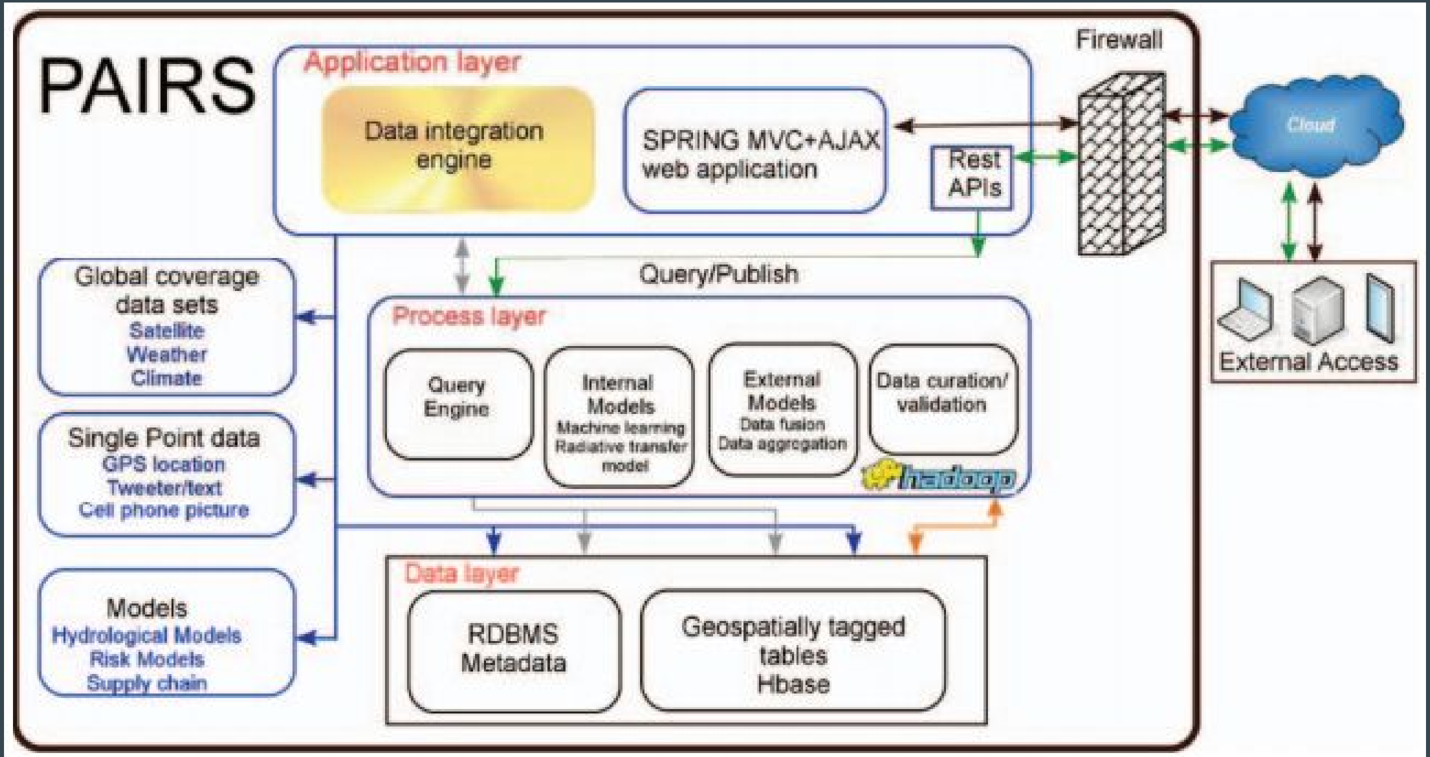
- Eliminates data preprocessing by having all data layers curated and homogenized before being uploaded to the platform
- Data curation means “organization and integration of data collected from various sources so that the value of the data is maintained over time, and the data remains available for reuse and preservation”
- The challenging task is to process unstructured data

3. PAIRS (A scalable Spatial Big Data analytics platform

PAIRS



3. PAIRS (A scalable Spatial Big Data analytics platform



Pairs architecture as a cloud service where a query retrieves metadata from a relational database (PostgreSQL) and pulls spatial data from HBase

3. PAIRS (A scalable Spatial Big Data analytics platform

Contents

AQWA

Adaptive Query-Workload-Aware partitioning of Spatial Big Data

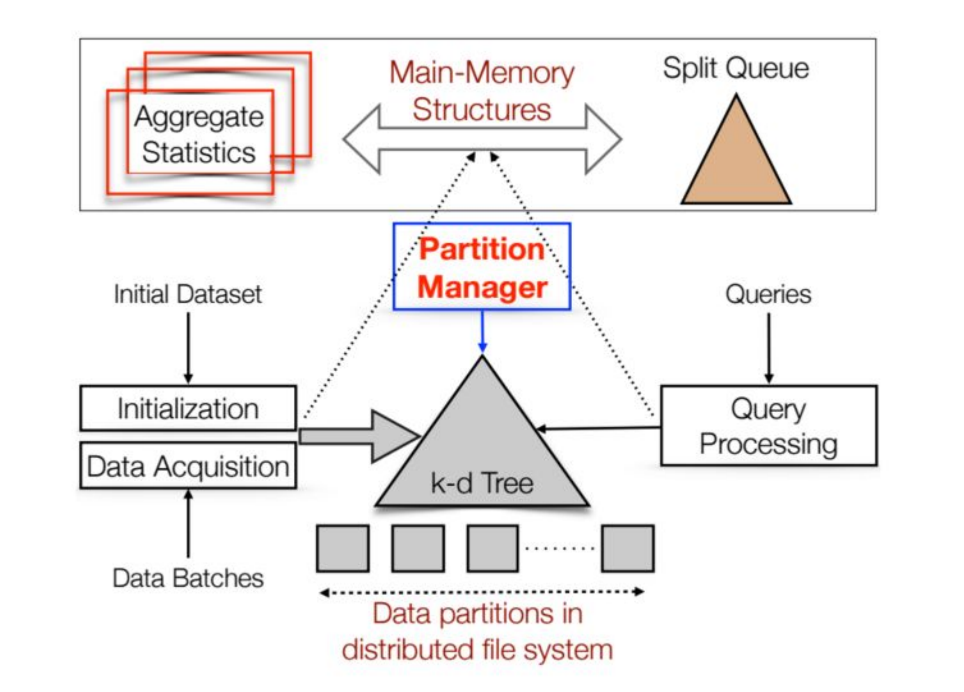
Motivation

Existing cluster-based systems for processing spatial big data

- uses static partitioning methods that cannot efficiently react to data changes
- SpatialHadoop supports static partitioning to handle spatial big data
- Query workload is bad

4. AQWA (Adaptive Query-Workload-Aware partitioning of Spatial Big Data

Overview of AQWA



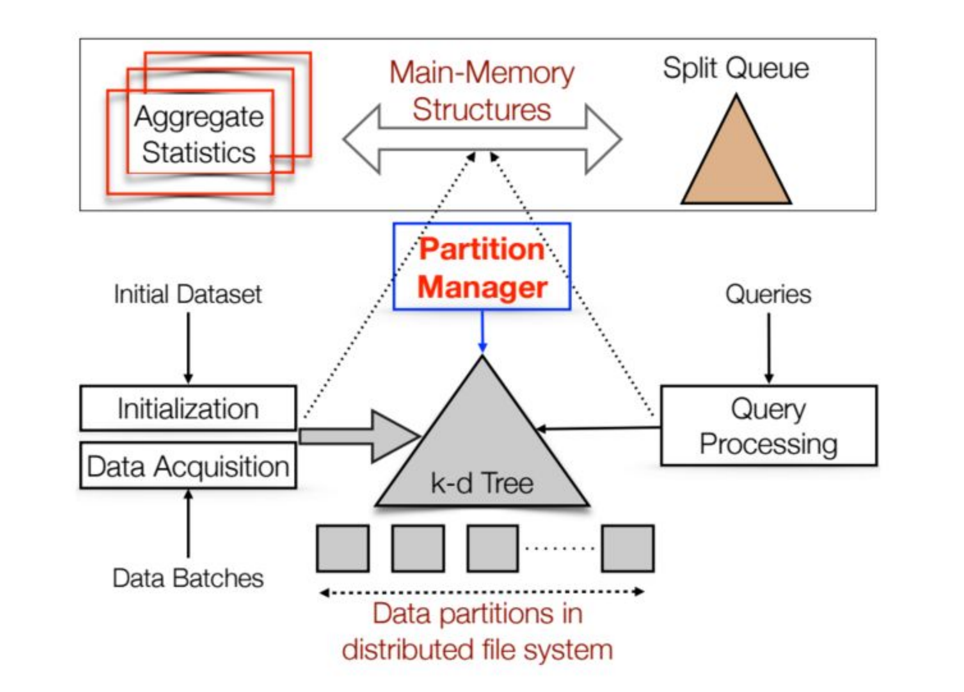
Two main components:

1) a k-d tree of the data

2) a set of Main-Memory structures

- statistics of **data distribution** and **the queries** to data
- flushed to a disk in the case of a system failure

Overview of AQWA



Four processes:

1) Initialization

2) Query Execution

3) Data Acquisition

4) Repartitioning

4. AQWA (Adaptive Query-Workload-Aware partitioning of Spatial Big Data)

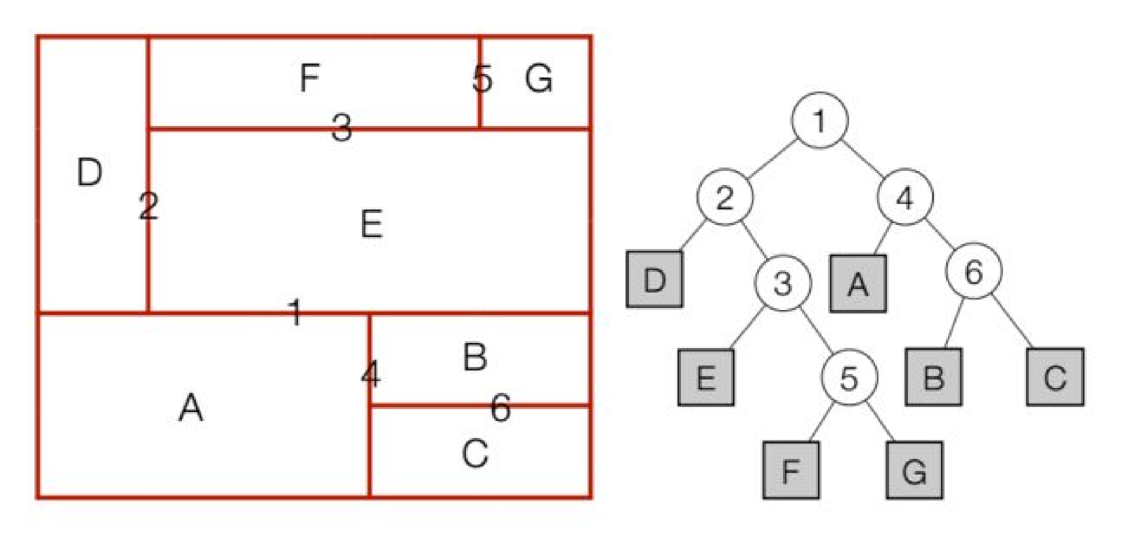
Partitioning of AQWA

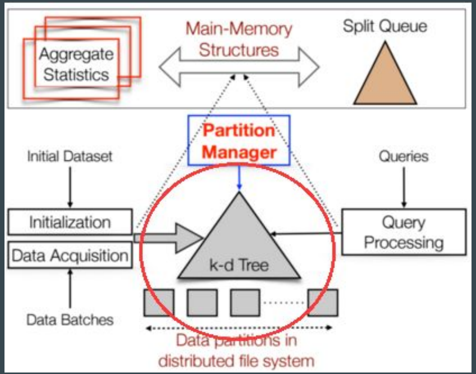
“Partitioned areas that are queried with high frequency need to be partitioned much more often in comparison to other less queried areas”

=> significant savings in query processing time

4. AQWA (Adaptive Query-Workload-Aware partitioning of Spatial Big Data)

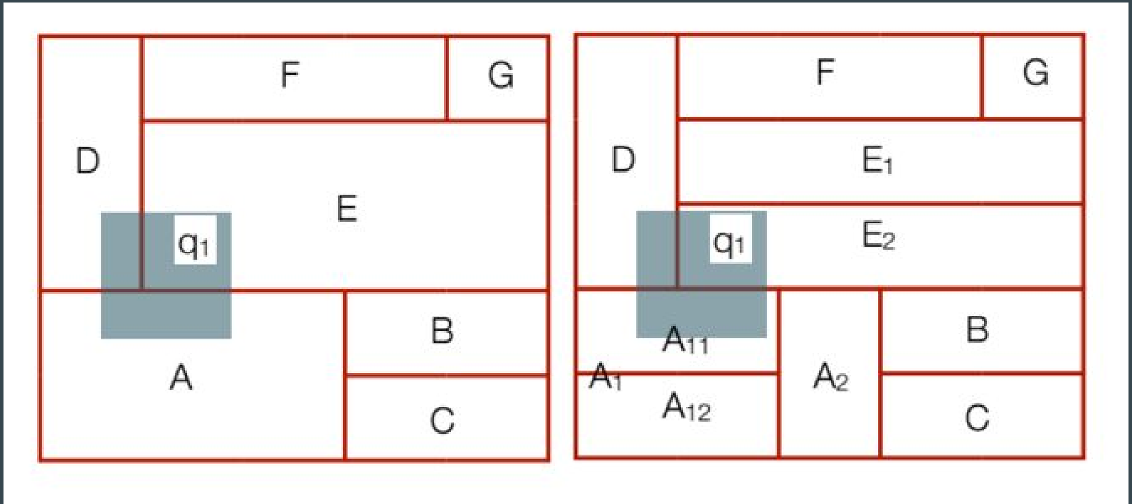
Partitioning of AQWA





An example of a k-d tree with 7 leaf partitions

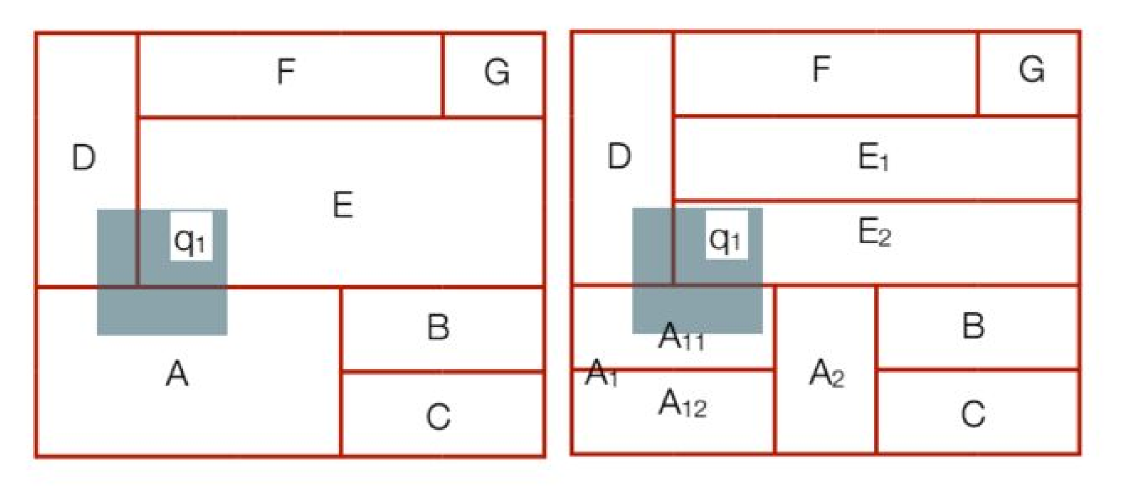
Partitioning of AQWA



Repartitioning of the spatial big data helps with query workload

4. AQWA (Adaptive Query-Workload-Aware partitioning of Spatial Big Data)

Partitioning of AQWA

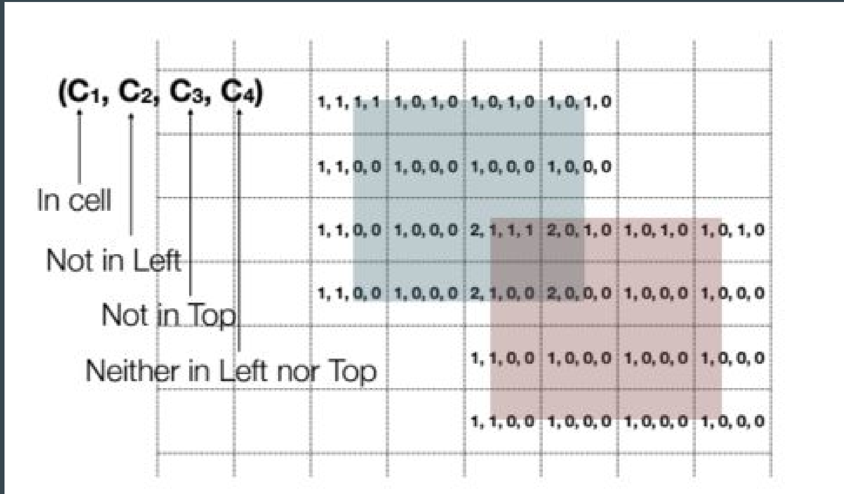


1) How do I know many queries overlap a square?

2) Why not split all of the data into small pieces?

3) How to efficiently determine the best split?

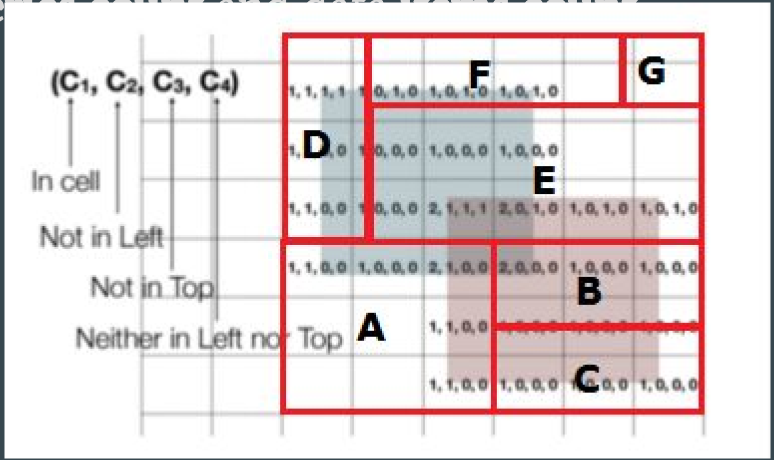
1) How do I know how many queries overlap a square?



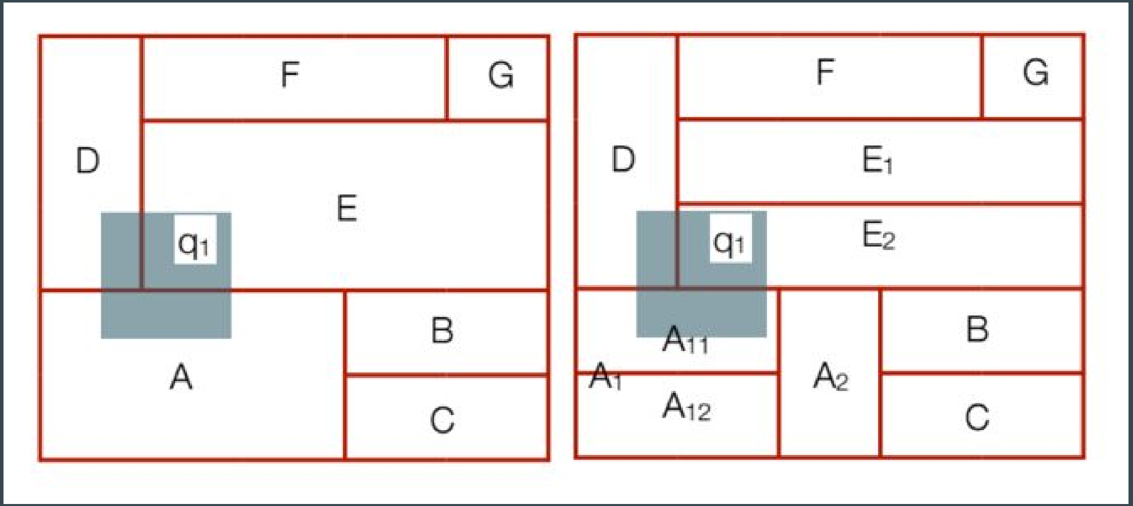
The four counters maintained for each grid cell.

You can get the answer in constant time O (1)

For each **grid**, the main memory has info of **queries count** and data **items count**.



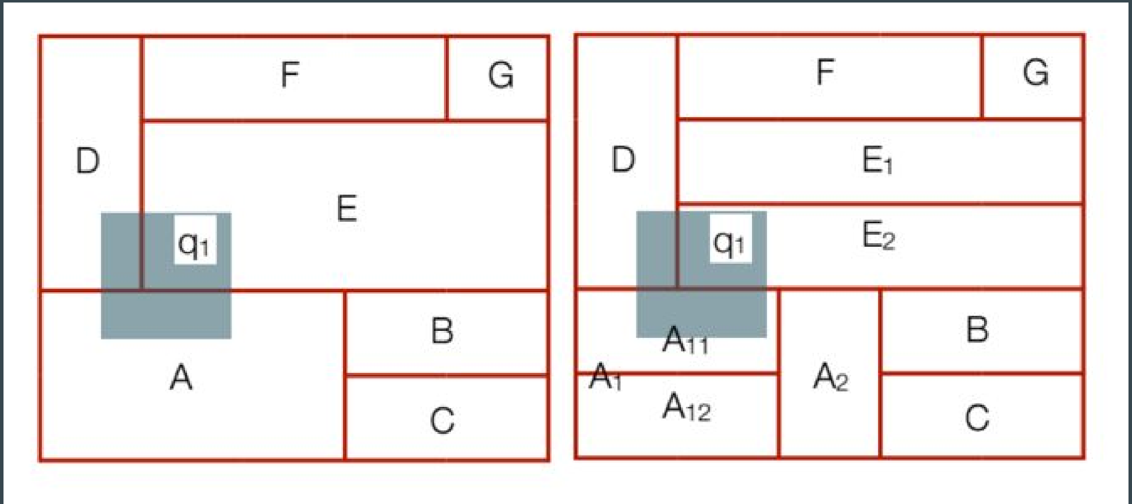
2) Why not just split all of the data into small pieces?



Main memory becomes a performance bottleneck

=> we have **max size** for each **partition** (the block size for example 128MB in HDFS is the minimum size for a partition)

3) How to efficiently determine the best split?



- Priorityqueue
- History of all queries that have been processed
- Time-Fading Weights
  - to avoid unnecessary partitioning
- Cost function
  - integrates the data distribution and the query workload

Summary

Usage of spatial big data depends on

- the location of the user
- the daytime of access

Most of the spatial big data is dynamic

- query workload of spatial big data can change and you should react to it
- new data applied on hourly / daily basis

Spatial big data has many different use cases

Summary

To efficiently handle spatial big data

- the data should have diverse access patterns in each cluster
- it needs to be repartitioned according to query workload changes
    - areas that are queried with high frequency should be partitioned more often in comparison to less queries areas
    - avoid partitioning from a scratch
    - use history of the workload with fading weights

References

Spatial big-data challenges intersecting mobility and cloud computing, Authors: Shekhar, Shashi and Gunturi, Viswanath and Evans, Michael R and Yang, KwangSoo, Year 2012

Geospatial big data: challenges and opportunities, Authors: Lee, Jae-Gil and Kang, Minseo, Year 2015

PAIRS: A scalable geo-spatial data analytics platform, Authors: Klein, Levente J and Marianno, Fernando J and Albrecht, Conrad M and Freitag, Marcus and Lu, Siyuan and Hinds, Nigel and Shao, Xiaoyan and Bermudez Rodriguez, Sergio and Hamann, Hendrik F, Year 2015

Cost-efficient partitioning of spatial data on cloud, Authors: Akdogan, Afsin and Indrakanti, Saratchandra and Demiryurek, Ugur and Shahabi, Cyrus, Year 2015

AQWA: adaptive query workload aware partitioning of big spatial data, Authors: Aly, Ahmed M and Mahmood, Ahmed R and Hassan, Mohamed S and Aref, Walid G and Ouzzani, Mourad and Elmeleegy, Hazem and Qadah, Thamir, Year 2015

Questions?