GeoSpark: A Cluster Computing Framework for Processing Large-Scale Spatial Data

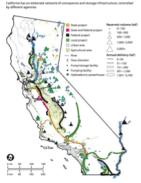
Jia Yu, Jinxuan Wu, Mohamed Sarwat

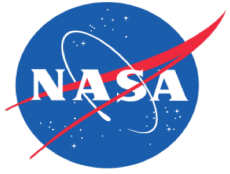
Arizona State University

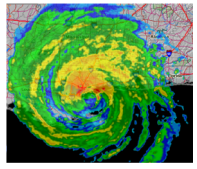
Social Media

1.19 billion monthly active users as of September 30, 2013

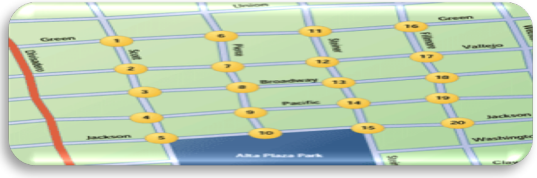
Scientific Data







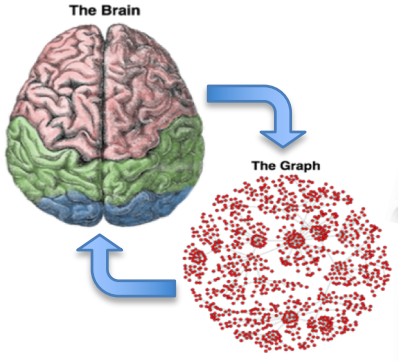
Urban Data



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

GeoSpark Architecture

- *Cluster computing framework!*
- *Easy secondary development!*
- *Spatial data mining compatibility!*

Spatial Query Processing Layer

Spatial RDD Layer

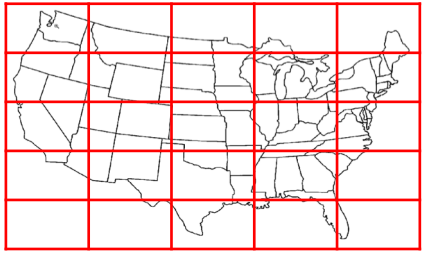
- Spatial RDD (Point, Rectangle, Polygon)
- Data partitioning
- Indexed SRDD

Apache Spark Layer

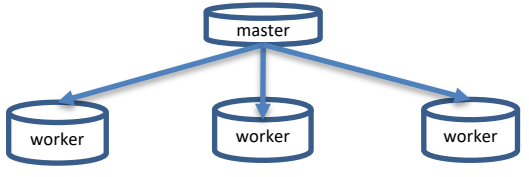
Spatial objects:
PointRDD, RectangleRDD, PolygonRDD

- Geometrical operation library (MBR, Union)

Spatial partitioning: Global grid file



Spatial indexing: (R-Tree and Quad-Tree)



Global grids

Local index
(in each partition)

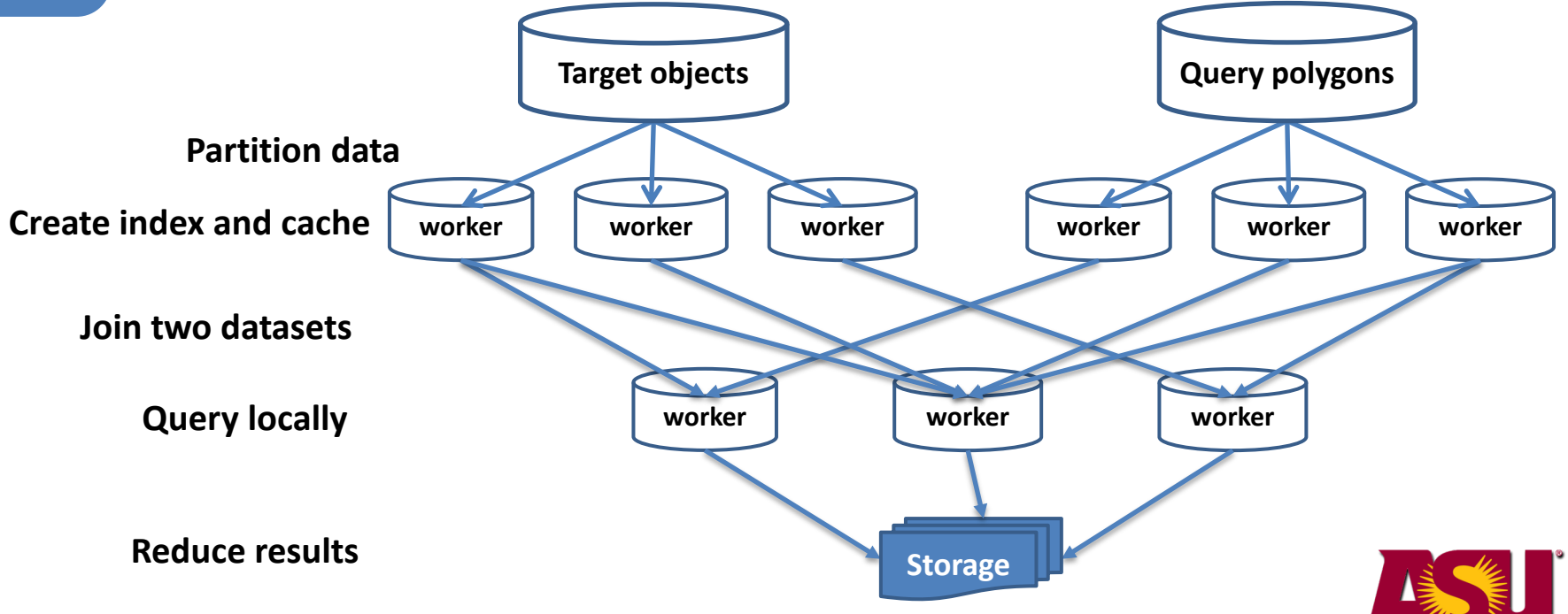
GeoSpark Architecture

- *Cluster computing framework!*
- *Easy secondary development!*
- *Spatial data mining compatibility!*

Spatial Query Processing Layer

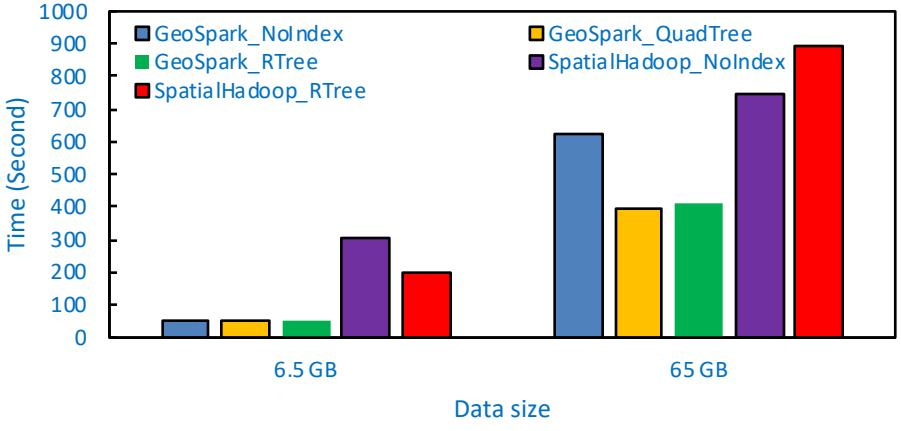
- Spatial Range
- Spatial KNN
- Spatial Join

Execution model (e.g. spatial join)

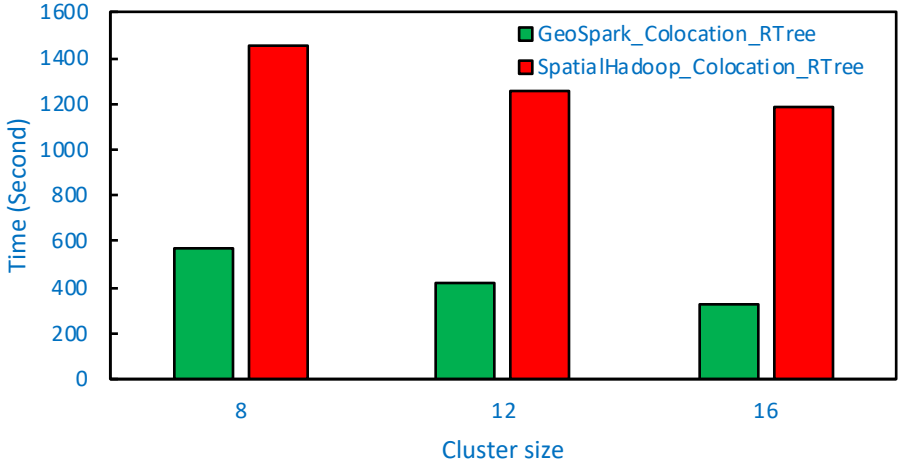


GeoSpark Result

- *Cluster computing framework!*
- *Easy secondary development!*
- *Spatial data mining compatibility!*



Spatial join
(One time analysis)



Spatial Co-Location Pattern Recognition
(Iterative analysis)

See GeoSpark in Poster Session!