Simba: Efficient In-Memory Spatial Analytics.

Dong Xie, Feifei Li, Bin Yao, Gefei Li, Liang Zhou and Minyi Guo

SIGMOD'16.

Andres Calderon

November 8, 2016

Agenda

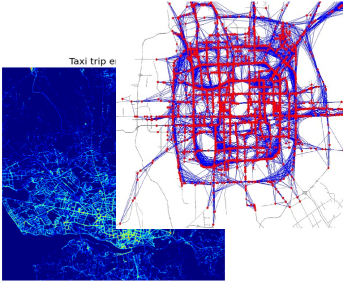
Agenda

Introduction

- There has been an explosion in the amount of spatial data in recent years...

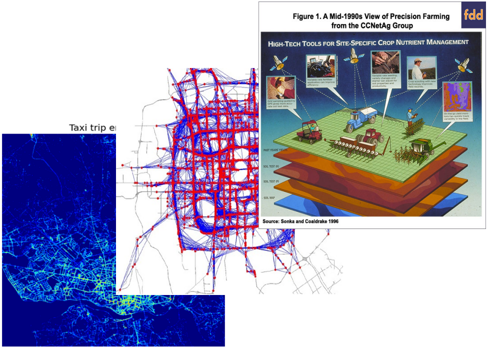
Introduction

- There has been an explosion in the amount of spatial data in recent years...



Introduction

- There has been an explosion in the amount of spatial data in recent years...



Introduction

- But remember that "Spatial is Special"...

PostGIS
Spatial PostgreSQL
ORACLE
SPATIAL
MySQL
Microsoft SQL Server

Spatial Hadoop
I can play with your spatial data
Hadoop-GIS
Spatial Big Data Solutions
MD-Hbase
SECONDO

SpatialSpark
GeoSpark
GeoTrellis
geomesa

Introduction

- Why do we need a new tool???

Introduction

- Simba: **S**patial **I**n **M**emory **B**ig data **A**nalytics.
  - Extends Spark SQL with important spatial operations.
  - Offers simple APIs for both SQL and DataFrame.
  - Support two-layer spatial indexing over RDDs (low latency).
  - Designs a SQL context to run spatial queries in parallel (high throughput).
  - Introduces spatial-aware and cost-based optimizations to select good spatial plans.

Introduction

- Simba: **S**patial **I**n **M**emory **B**ig data **A**nalytics.
    - Extends Spark SQL with important spatial operations.
    - Offers simple APIs for both SQL and DataFrame.
    - Support two-layer spatial indexing over RDDs (low latency).
    - Designs a SQL context to run spatial queries in parallel (high throughput).
    - Introduces spatial-aware and cost-based optimizations to select good spatial plans.

Introduction

- Simba: **S**patial **I**n **M**emory **B**ig data **A**nalytics.
  - Extends Spark SQL with important spatial operations.
  - Offers simple APIs for both SQL and DataFrame.
  - Support two-layer spatial indexing over RDDs (low latency).
  - Designs a SQL context to run spatial queries in parallel (high throughput).
  - Introduces spatial-aware and cost-based optimizations to select good spatial plans.

Introduction

- Simba: **S**patial **I**n **M**emory **B**ig data **A**nalytics.
  - Extends Spark SQL with important spatial operations.
  - Offers simple APIs for both SQL and DataFrame.
  - Support two-layer spatial indexing over RDDs (low latency).
  - Designs a SQL context to run spatial queries in parallel (high throughput).
  - Introduces spatial-aware and cost-based optimizations to select good spatial plans.

Introduction

- Simba: **S**patial **I**n **M**emory **B**ig data **A**nalytics.
    - Extends Spark SQL with important spatial operations.
    - Offers simple APIs for both SQL and DataFrame.
    - Support two-layer spatial indexing over RDDs (low latency).
    - Designs a SQL context to run spatial queries in parallel (high throughput).
    - Introduces spatial-aware and cost-based optimizations to select good spatial plans.

Agenda

Agenda

Agenda

Agenda

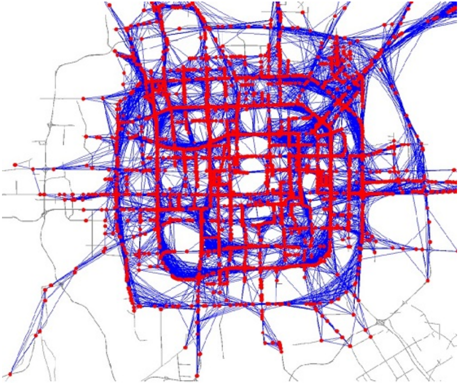
Agenda

Setup

- Cluster of 25 nodes:
    - HDD from 50GB to 200GB.
    - RAM from 2GB to 8GB.
    - Processors 2.2GHz to 3GHz
- Single machine:
    - HDD 2TB.
    - RAM 16GB.
    - Processor 3.4GHz.

Datasets

- Real datasets (from OpenStreetMap):
  - OSM1: 164M polygons, 80GB.
  - OSM2: 1.7B points, 52GB.
- Synthetic dataset:
  - SYNTH: 3.8B points, 128GB.
  - Five different distributions.



Agenda

Conclusions

- This paper introduced CG_Hadoop as a scalable and efficient MapReduce library.
- Focused on 5 fundamental computational geometry problems...
  - Polygon union, Skyline, Convex hull, Farthest and Closest Pairs.
- Provided versions for Apache Hadoop and SpatialHadoop systems.
- Distributed approach speed up performance.
- Spatial partitioning allows early pruning which make it even more efficient.
- Achieve up to 29x and 260x better performance.

Future ideas

- Working on more complex operations, for example motion patterns.
- Explore ports to new distributed platforms such as Spark or Simba.

Thank you!!!

Do you have any question?