SJSU SAN JOSÉ STATE UNIVERSITY

SJSU Undergraduate Research Grants

Macro-benchmarking for NoSQL Database Systems under Geospatial Workloads

Yvonne Hoang and Dr. Suneuy Kim

Department of Computer Science, College of Science

Abstract

Geospatial applications and geospatial data have become increasingly popular in today's technological landscape. Traditional relational databases have their limitations in managing geospatial big data due to the limited scalability and impedance mismatch. NoSQL databases, then, have emerged as an alternative big data management solution. This study is to develop a macro-benchmark suite using GeoYCSB [1] and to conduct macro-benchmarking of MongoDB, the most leading document store. Currently, GeoYCSB only supports micro-benchmarks and lacks macro-benchmarking capabilities. Macro-benchmarks simulate a given application's use cases to create realistic workloads. These use cases may involve multiple data layers and a series of queries performed in sequence, which compose the realistic workload, as compared to micro-benchmarks that test only the primitive spatial functions. The expected outcomes from this project include the developed macro-benchmarks, experimental results, and our analysis.

Project Activities

1. Studying the Design Architecture and Mechanism of GeoYCSB

2. Developing Use Cases: While micro-benchmarks can stress-test the database, they cannot see how the database will behave under realistic query patterns. In order to implement macro-benchmarking, we developed four representative use cases for a geospatial application involving the Graffiti Abatement Incidents dataset. The use cases are listed below:

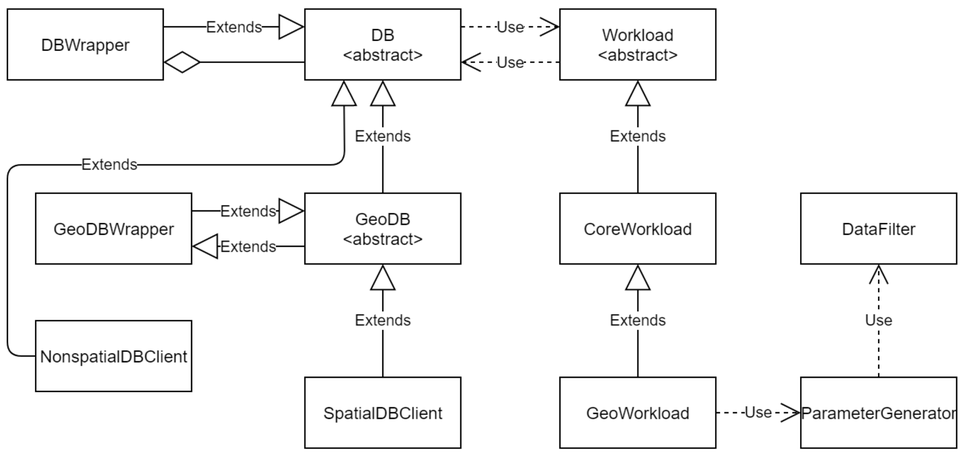
1) A use case performing point-of-interest searches: finding all graffiti near schools
2) A use case performing data analysis: finding neighborhoods with highest graffiti density
3) A use case performing spatial join: finding all graffiti among high traffic areas
4) A use case performing write operations: removing graffiti based on a previous search

3. Setting up a Scalable, Replicated, and Sharded MongoDB Cluster in AWS

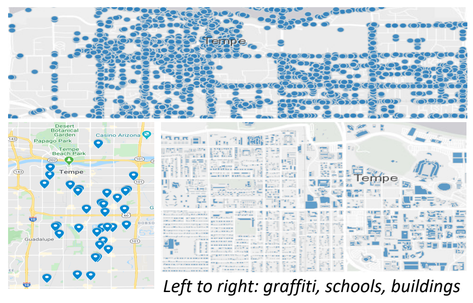
4. Synthesizing Dataset and Populating Database With Geospatial Big Data: We scaled the original dataset with 13,000 documents to over 1 million documents. The method we adopted preserves the geospatial distributions of the original data.

5. Macro-benchmarking and Performance Analysis: The performance factors for our analysis include data size, maximum distance, number of nodes for scalability testing, and write consistency level. The performance metric by which we are measuring are throughput and latency.

GeoYCSB Architecture



Geospatial Datasets



Left to right: graffiti, schools, buildings

Original dataset:
- *13,000 graffiti*
- *50,000 buildings*
- *30 schools*

Expanded dataset totaling over 4GB of data:
- *1,000,000 graffiti*
- *4,000,000 buildings*
- *2,000 schools*

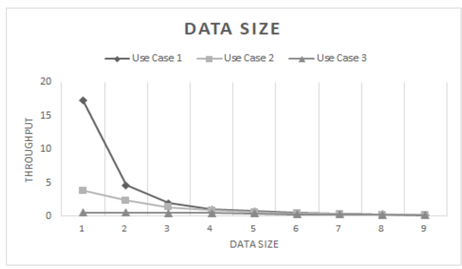
Dataset Sources

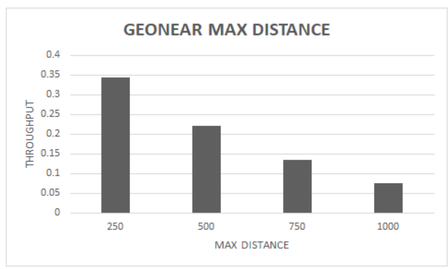
1. "Building Footprints USGS Data Set," https://data-tempegov.opendata.arcgis.com/datasets/building-footprints-usgs.
2. "Graffiti Abatement Incidents Data Set," https://data-tempegov.opendata.arcgis.com/datasets/graffiti-abatement-incidents.
3. "Tempe Union High School District Schools," https://www.tempeunion.org/domain/44.
4. "Tempe Elementary School District Schools," https://www.tempeschools.org/our-schools/.
5. "Kyrene School District Schools," https://www.kyrene.org/Page/1439.

Research Questions

- What use cases are common geospatial application functionalities, and how can we simulate them on our particular dataset?
- How can we synthesize geospatial data to create a realistic, sufficiently large dataset that still preserves the original dataset's distribution?
- How can we develop workloads that properly evaluate a database's performance on this synthetic dataset using these realistic use cases?

Initial Experimental Results





Citations

[1] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan and R. Sears, "Benchmarking Cloud Serving Systems with YCSB", in *Proceedings of the 1st ACM Symposium On Cloud Computing,* Indianapolis, Indiana, USA, pp. 143-154, 2010.

[2] S. Kim and Y. Kanwar, "GeoYCSB: A Benchmark Framework for the Performance and Scalability Evaluation of NoSQL Databases for Geospatial Workloads", In Proceedings of 2019 IEEE International Conference on Big Data (Big Data), pp. 3666 – 3675.