Proyecto 2021-I

Autor: María Lovatón

Descripción teórica del problema

Contexto

A finales de los 80, en la comunidad científica, existía una fuerte necesidad por software especializado en el manejo de data espacial (Haining, 1989). Este tipo de data posee una alta complejidad en el mundo real y requiere de la implementación de variadas estructuras para poder ser representada a nivel computacional. Por lo tanto, crear software para análisis de data espacial no era una tarea fácil. Actualmente, la modernización de hardware y software provee numerosas herramientas que cumplen con ese propósito, las cuáles benefician activamente a múltiples industrias.

Conceptos teóricos

- **Base de datos**: Contenedor que permite almacenar la información de forma ordenada con diferentes propósitos y usos (Anguiano, 2014).
- **Base de datos espacial**: Base de datos que permite describir objectos espaciales a tráves de tres características principales: atributos, localización y topología (Gutiérrez, 2006). Existen dos tipos: georreferenciadas y no georreferenciadas.
- **Sistema de información geográfica**: Mejor conocido como GIS (siglas en inglés), se refiere al conjunto de herramientas que permiten la manipulación y análisis de grandes colecciones de información geográficamente referenciadas.
- **PostGIS**: Extensión para manejo de consultas espaciales de Postgresql (RDBMS).

Problema

Enunciado inicial del proyecto

Se requiere implementar un programa que soporte la ejecución de consultas espaciales sobre un conjunto de datos. En este proyecto, se trabajara con registros de viajes de taxi y datos de la demarcación de barrios de la ciudad de Nueva York.

Consultas

1. ¿Qué viajes tuvieron como origen y destino el mismo barrio?
2. ¿Cuáles son los barrios con mayor cantidad de viajes? Retornar los top 5. (Considerando apenas las coordenadas de partida).

3. Dados dos puntos P1 y P2 que representan una region rectangular, retornar el número de viajes que comenzaron en tal región.
4. Determinar por cuales vecindarios pasó un taxi en un determinado viaje, considerando la ruta como una línea recta.
5. Dado un punto P y una distancia D, encontrar los viajes que empezaron o terminaron a una distancia D como máximo a partir del punto P.

Consideraciones

1. El conjunto de datos principal a utilizar esta compuesto por 36 archivos CSV, que son los registros de 3 proveedores distintos para cada mes del año. Dado que el total del tamaño de los datos esta en el orden de los GBs, se debe considerar esta escala desde el comienzo para el diseño de su propuesta.
2. A pesar que el conjunto de datos completo se conoce a priori, el objetivo es que su programa utilice estructuras de datos dinámicas que ademas puedan permitir las consultas de manera eficiente.
3. Se sugiere se utilice alguna herramienta de terceros para validar el resultado de algunos experimentos tales como PostGIS, MongoDB u Oracle.

Solución

Se describe la solución para cada consulta a continuación:

1. Pre-computar los barrios del origen y el destino en nuevas columnas usando Quadtree (space-driven). Después, se realiza una consulta en base a los valores.

```
SELECT id FROM rides WHERE pickup_neigh == dropoff_neigh
```

2. Usando la pre-computación anterior, se realiza otra consulta.

```
SELECT pickup_neigh, COUNT(id) as cnt FROM rides GROUP BY pickup_neigh ORDER BY c
```

3. Se crearía una función booleana `IS_INSIDE` para verificar si el origen se encuentra dentro del bounding box dado. Esta función utilizaría un Quadtree para realizar la búsqueda en el espacio.

```
SELECT id FROM rides WHERE IS_INSIDE(pickup_point, BBOX(P1, P2))
```

4. Esta vez se correría la consulta en la tabla de barrios. Se crearía un bounding box que representaría la recta entre ambos destinos y se chequearía si existe una intersección entre el polígono del barrio y la recta, usando la función booleana `IS_OVERLAPPED`. Se usaría un Quadtree, similar a lo que se hizo con la consulta 3.

```
SELECT id FROM neighs WHERE IS_OVERLAPPED(BBOX(P1, P2), neigh_polygon)
```

5. Se usaría un KDTree para la función booleana `MAX_DISTANCE` entre 2 puntos. Esto devolvería los puntos que cumplen con la condición.

```
SELECT id FROM rides WHERE MAX_DISTANCE(pickup_point, P, D) OR MAX_DISTANCE(dropo
```

Se utilizará PostGIS para hacer la validación de resultados.

Levantamiento de antecedentes

QUILT: a geographic information system based on quadtrees

Esta investigación describe una nueva herramienta para el manejo de data espacial usando Quadtrees. QUILT es un GIS que usa variantes de Quadtree para representar regiones, líneas y puntos. Esto se logra implementando un Quadtree lineal, organizado en disco usando un B-tree (Shaffer, 1990).

Quadtree and R-tree indexes in oracle spatial: a comparison using GIS data

Esta investigación revisa y resume las variantes de Quadtree y R-tree que se han presentado a lo largo de la literatura científica. Asimismo, se compara el rendimiento de las variantes usando grandes datasets en *Oracle Spatial* y se concluyen en ventajas y desventajas de cada estructura. En sus resultados, los R-trees fueron superiores a los Quadtrees en casi todas las consultas. Sin embargo, en el caso de datasets actualizados en tiempo real, los Quadtrees obtuvieron un mejor rendimiento (Kothuri, 2002).

Planteamiento de objetivos

Objetivo principal

Implementar un programa capaz de realizar las 5 consultas mencionadas.

Objetivos secundarios

1. Implementar un parser para leer e interpretar las consultas. Definir tipos de data y tablas.
2. Implementar funciones básicas de búsqueda por comparación, operaciones booleanas y agrupación. Solo implementar las necesarias: `==`, `OR`, `COUNT`, `GROUP BY`.
3. Implementar la función `BBOX` para representar polígonos.

4. Implementar las funciones `IS_INSIDE` y `IS_OVERLAPPED` usando alguna variante de Quadtree.
5. Implementar la función `MAX_DISTANCE` usando alguna variante de KDtree.

Descripcion de acciones y cronograma previsto

Fecha	Entregable
20/06	Propuesta inicial
27/06	Objetivo secundario 1
04/07	Objetivo secundario 2
11/07	Objetivo secundario 3
18/06	Objetivo secundario 4
25/06	Objetivo secundario 5

Identificacion de limitaciones y riesgos

Limitaciones

Hardware

La limitación de hardware podría limitar la fase de testing del proyecto debido a que solo se podría utilizar una colección limitada de data. Para mejorar esto, se podrían utilizar muestras significativas de data (propensas a errores).

Riesgos

Quadtree

Tal como se menciono en los estudios, el R-tree sobrepasa en rendimiento al Quadtree en la mayoría de casos prácticos. Esto podría resultar en un proyecto ineficiente. Se testeará la implementación del proyecto con la de PostGIS para probar su rendimiento.

Uso de referencias bibliograficas

- Anguiano, J. (2014). Características y tipos de bases de datos. IBM website: https://developer.ibm.com/es/technologies/databases/articles/tipos-bases-de-datos.
- Enunciado del Proyecto de EDA (2021).
- Gutiérrez, M. (2006). El rol de las bases de datos espaciales en una infraestructura de datos. In GSDI-9 Conference Proceedings (pp. 6-10).

- Haining, R. (1989). Geography and spatial statistics: current positions, future developments In Macmillan B (ed) Remodelling Geography. Basil Blackwell, Oxford (pp. 191–203).
- Kothuri, R. K. V., Ravada, S., & Abugov, D. (2002, June). Quadtree and R-tree indexes in oracle spatial: a comparison using GIS data. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (pp. 546-557).
- Shaffer, C. A., Samet, H., & Nelson, R. C. (1990). QUILT: a geographic information system based on quadtrees. International Journal of Geographical Information System, 4(2), 103-131.