MR-DBSCAN — Clustering por Densidad con MapReduce

Reproducción y explicación del algoritmo MR-DBSCAN basado en el paper: "MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm using MapReduce".

Miguel Bolaño, Jose Aguilar y Roger Fuentes

Contexto y Plataforma

Problema

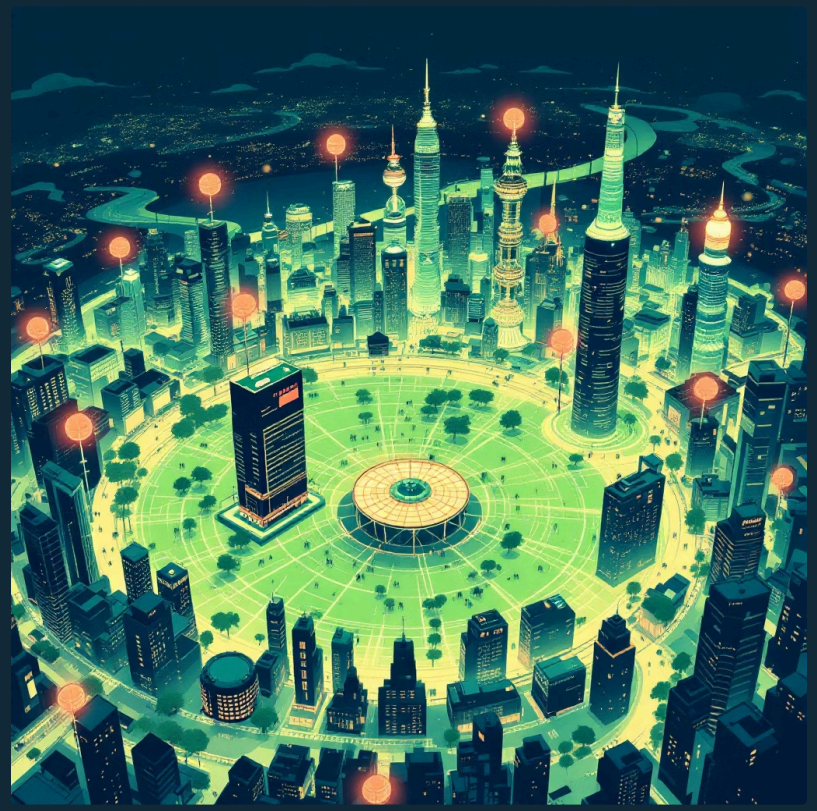
Clustering de datos espaciales a gran escala (trayectorias GPS).

Plataforma

Hadoop + PySpark — procesamiento distribuido con MapReduce.

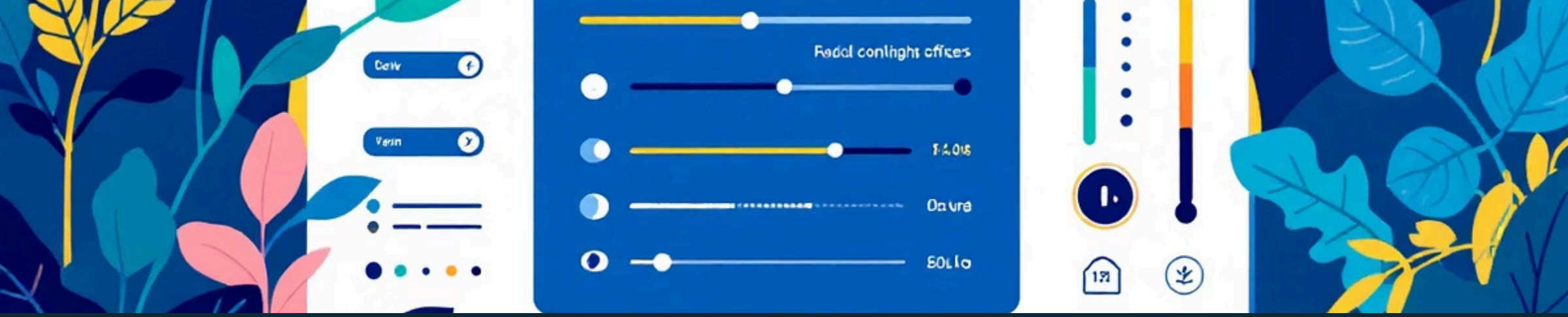
Dataset

GPS real de taxis en Shanghai.



Objetivo del Proyecto

Reproducir MR-DBSCAN implementando las 4 etapas distribuidas por MapReduce para obtener clusters de densidad escalables y comparables al paper original.



Parámetros del Algoritmo

1

Eps (radio)

0.002 y 0.0002 (ajustes según densidad y resolución espacial).

2

MinPts

1000 y 100 (umbral de densidad para clusters locales vs globales).

3

Particionado

Grid en strips: entre 80 y 160 particiones para balancear carga y fronteras.

Fases del MR-DBSCAN

1. Preprocesamiento y particionado

Normalizar coordenadas, eliminar outliers, crear grid de strips.

2. DBSCAN local

Ejecutar DBSCAN en cada partición (Map).

3. Detección de puntos frontera (MC-Sets)

Identificar puntos en la frontera para posibles uniones entre particiones.

4. Fusión global de clusters

Unir clusters conectados vía los MC-Sets (Reduce) y validar etiquetas finales.

Resultados Clave

Puntos procesados

Total de puntos del dataset (GPS taxis).

Clusters

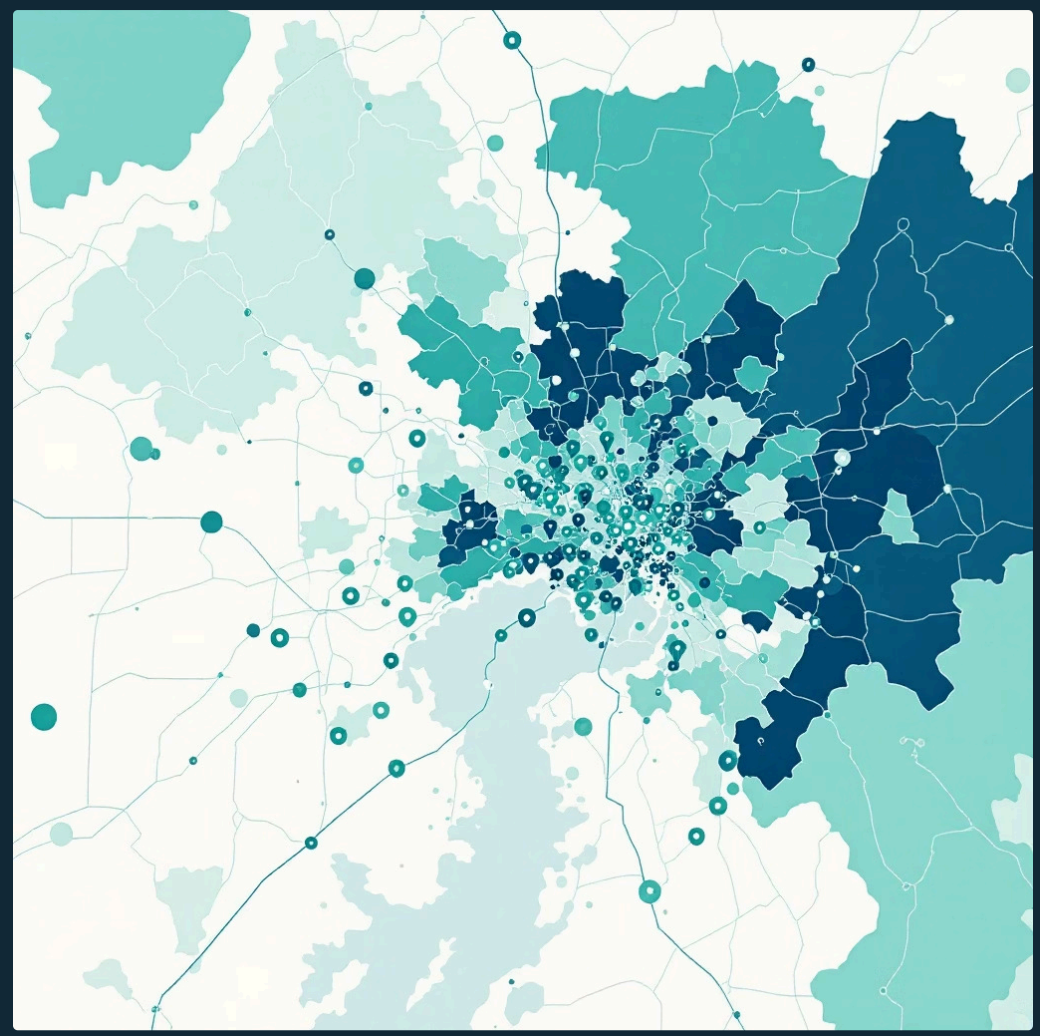
Número de clusters detectados tras fusión global.

Ruido

Porcentaje de puntos etiquetados como ruido.

Tamaños

Distribución final de tamaños de clusters.



Nota: dejar los valores numéricos exactos listos para insertar tras la ejecución reproducida.

Consideraciones de Implementación

Tiempo

El coste depende de particionado y Eps; mayor granularidad → más overhead de comunicación.

Escalabilidad

MapReduce permite escalar horizontalmente, pero la fusión global es sensible a MC-Sets grandes.

Parámetros

Ajustar Eps y MinPts según densidad local; experimentar con 80–160 strips para balancear carga.

Conclusiones y Siguientes Pasos

- MR-DBSCAN facilita clustering distribuido de grandes datos espaciales.
- Ventajas: paralelismo y manejo de grandes volúmenes; desafío: comunicación en fronteras.
- Siguientes pasos: ejecutar experimentos reproducibles, insertar métricas numéricas (tiempos, %ruido, counts), optimizar particionado.