Tercera Entrega
Building Footprint Data Loading and Integration Report

Juan David Castillo Laverde
Juan Pablo Dávila Martinez
Eugenia Victoria Dayoub Barito
Luis Fernando Lee Rodriguez

Pontificia Universidad Javeriana
Facultad de Ingeniería
Administración de Bases de Datos
Bogotá D.C.
Noviembre 2025

Introducción

Metodología de Carga de Datos

Para esta entrega se cargaron dos fuentes de huellas de edificaciones (footprints): Google Open Buildings y Microsoft Building Footprints, ambas reconocidas como datasets abiertos y actuales. El objetivo en esta fase no era aún integrarlos con PDET ni cargarlos en la base NoSQL, sino preparar samples representativos que permitan hacer pruebas posteriores sin necesidad de manejar los millones de registros reales, los cuales exceden la capacidad de procesamiento y almacenamiento práctico a nivel local.

Google Open Buildings

1. Se ingresó al panel de descarga del dataset de Google que se organiza por cuadrículas globales.

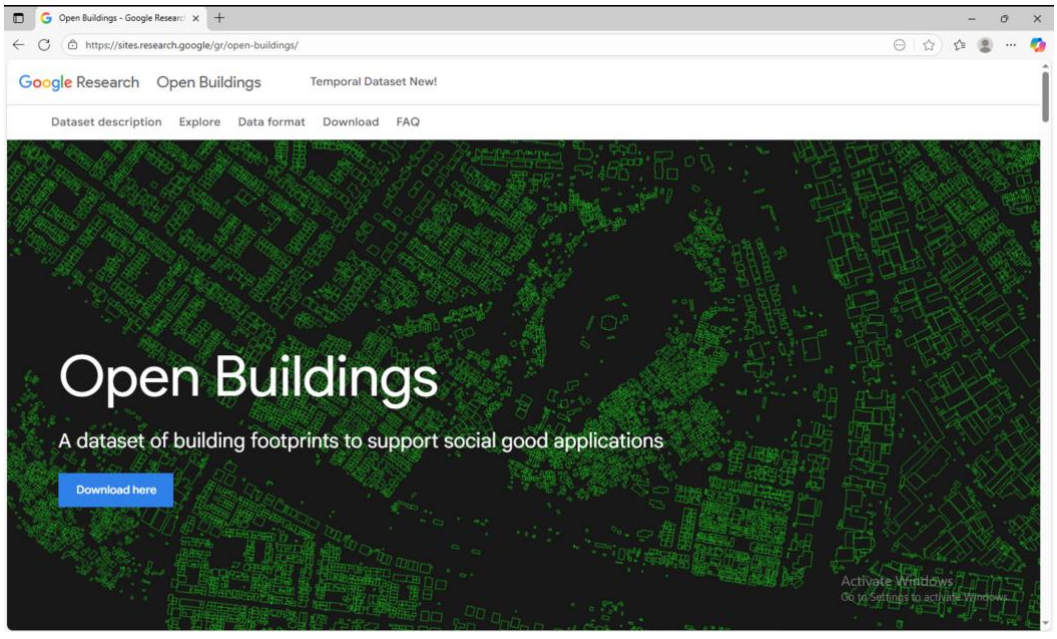

Figura 1: Página de Google Open Buildings

2. Se identificaron 4 tiles correspondientes al territorio colombiano.

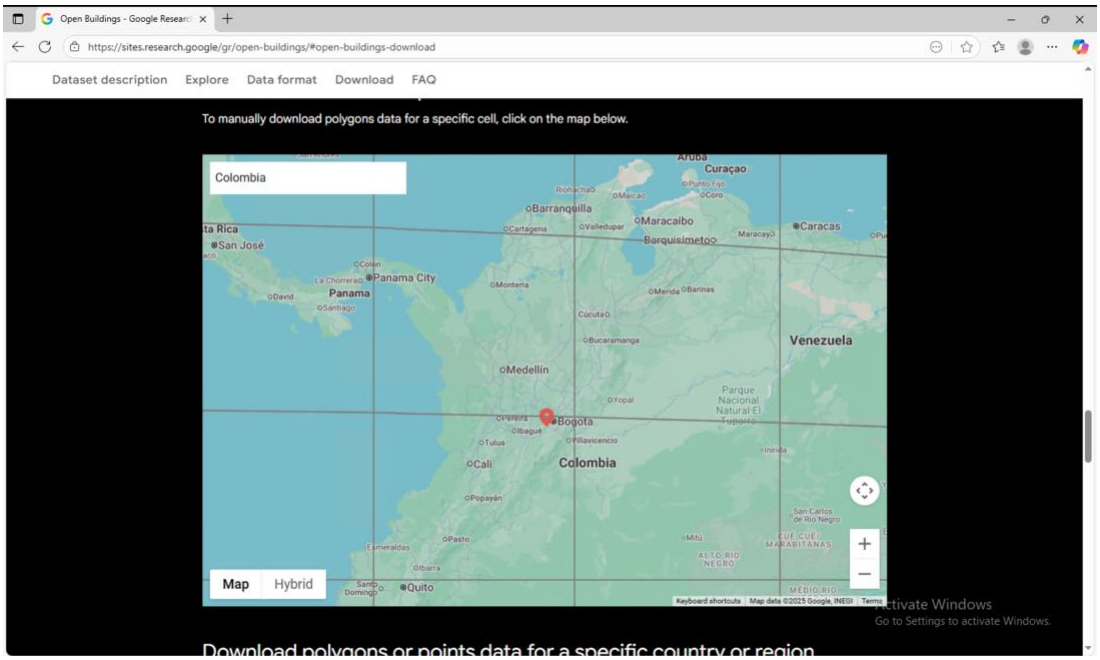

Figura 2: Buscar Colombia en la página

3. Cada tile se descargó en formato .csv.

8e3_buildings	5/11/2025 3:46 p. m.	Comma Separate...	1.721.478 KB
8e3_buildings.csv.gz	5/11/2025 3:46 p. m.	GZ File	667.291 KB
8e7_buildings	5/11/2025 3:46 p. m.	Comma Separate...	1.500.671 KB
8e7_buildings.csv.gz	5/11/2025 3:46 p. m.	GZ File	582.676 KB
8e5_buildings	5/11/2025 3:46 p. m.	Comma Separate...	1.445.771 KB
8e5_buildings.csv.gz	5/11/2025 3:46 p. m.	GZ File	559.530 KB
8e1_buildings	5/11/2025 3:45 p. m.	Comma Separate...	68.899 KB
8e1_buildings.csv.gz	5/11/2025 3:45 p. m.	GZ File	26.609 KB

Figura 3: Los 4 tiles descargados

4. Se agregaron a QGIS como capa de texto delimitado
    a. Delimitador: coma
    b. Tipo de geometría: coordenadas de punto
    c. X = longitude
    d. Y = latitude
    e. Sistema de referencia: EPSG:4326

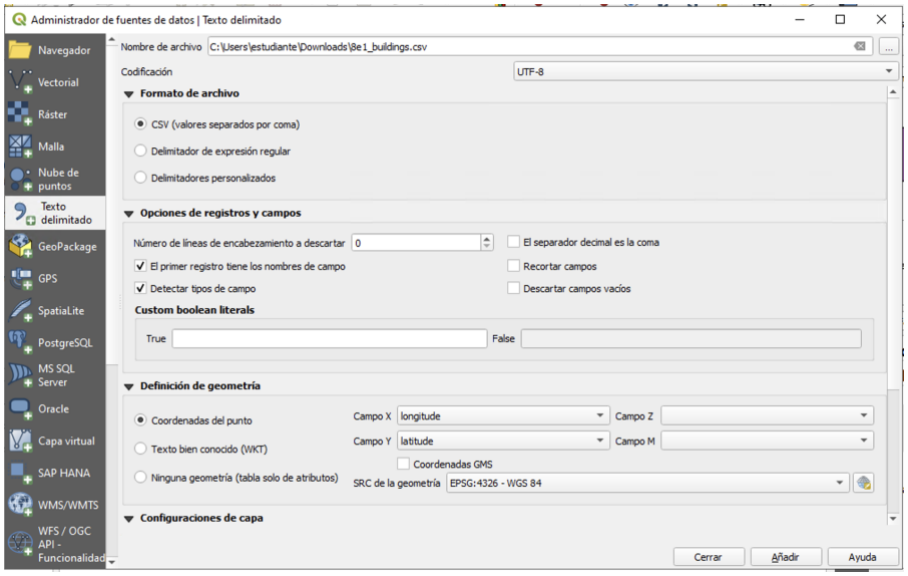

Figura 4: Inserción de cada tile a QGIS

5. Los 4 tiles se visualizan correctamente dentro del territorio nacional.

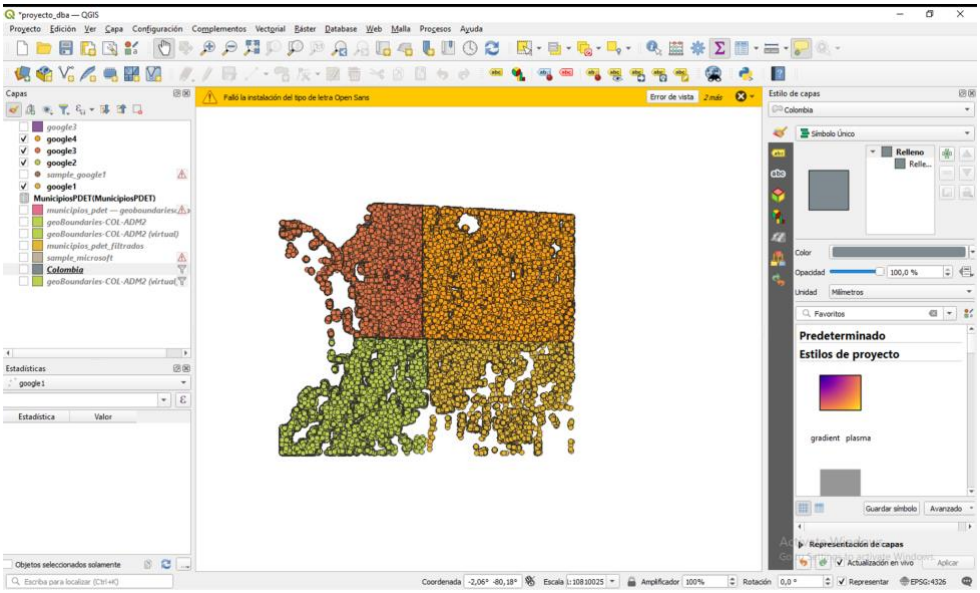

Figura 5: Archivos exportados en QGIS

6. Para obtener un subconjunto manejable se abrió la tabla de atributos del primer archivo cargado (google1), se seleccionaron las primeras 200 filas y se exportaron como entidad nueva.

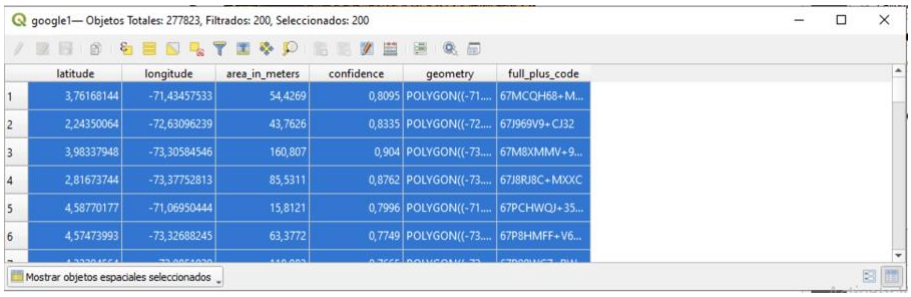

Figura 6: Datos de un archivo filtrados

7. Se generó un archivo GeoJSON con este subconjunto: sample_google1.geojson

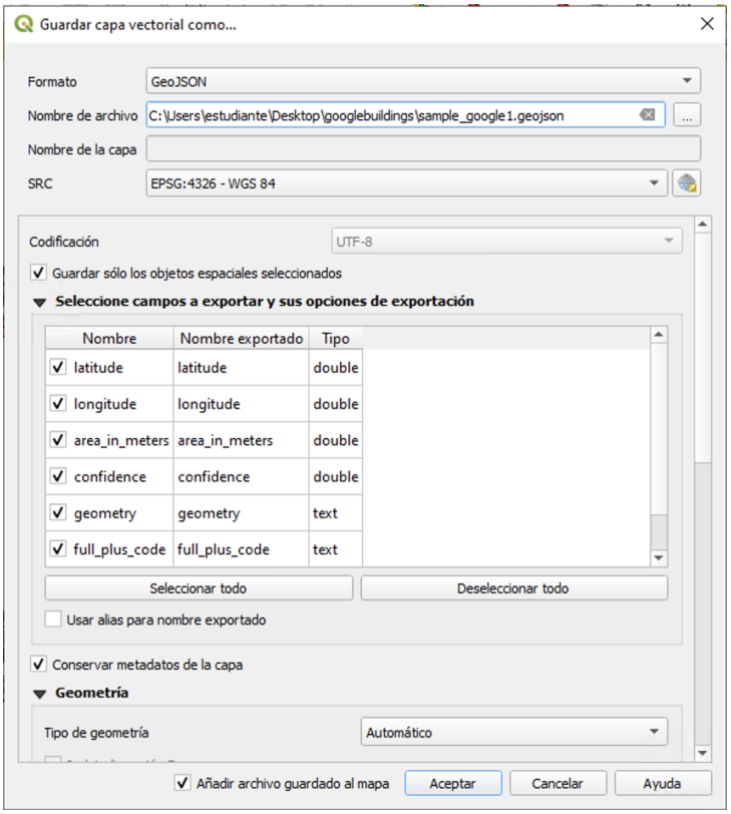

Figura 7: Exportación de archivo con datos filtrados

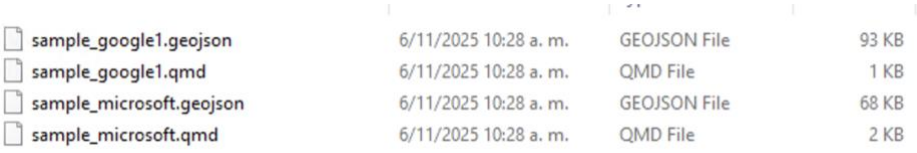

Figura 8: Archivo sample_google1.geojson exportado

Microsoft Building Footprints

1. Se descargó el archivo oficial correspondiente a Colombia en formato .geojsonl.zip. https://minedbuildings.z5.web.core.windows.net/

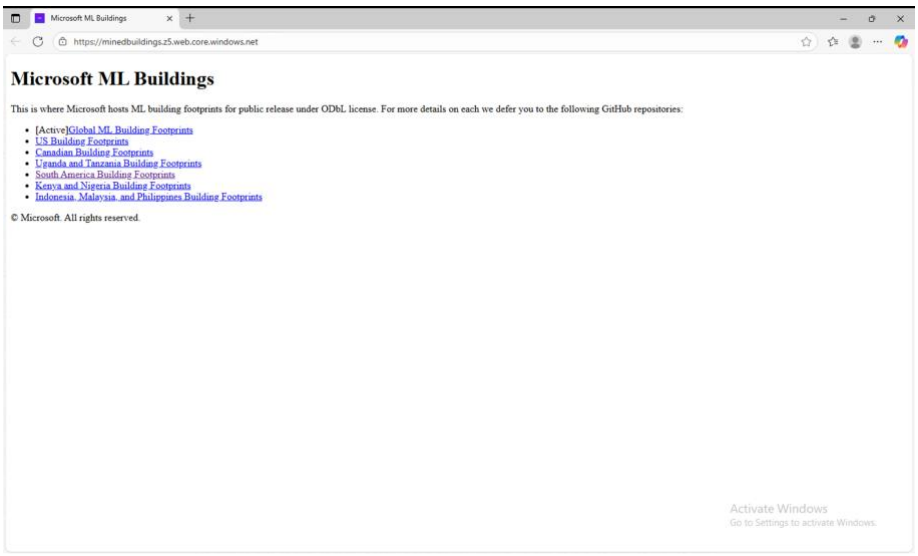

Figura 9: Página de Microsoft ML Buildings

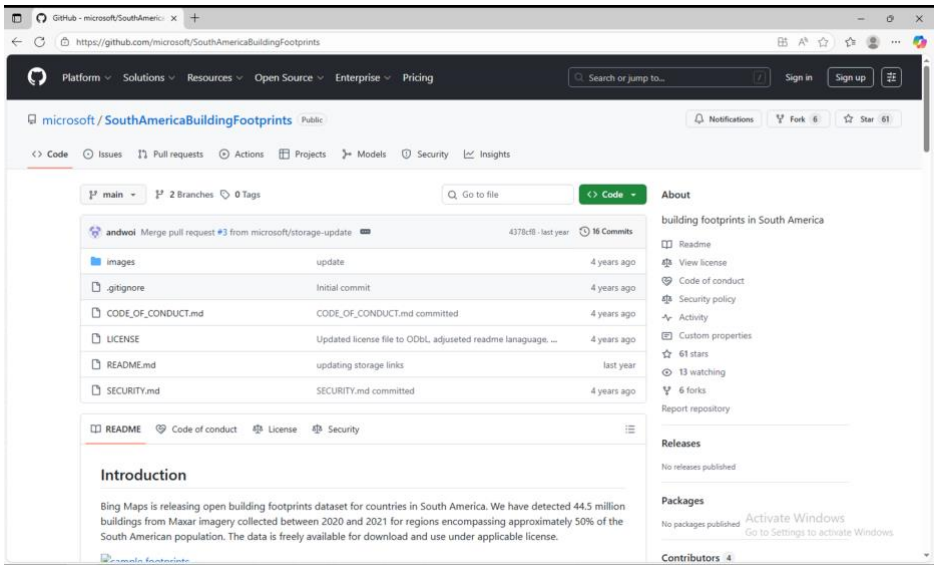

Figura 10: GitHub con datos de footprints

What does the data include?

44,495,865 building footprint polygon geometries located in South America in GeoJSON format. You may download the data in GeoJSON format here:

Location	Count	Link	Size (Compressed)
Continent	44,495,865	SouthAmericaPolygons.zip	15GB
Argentina	3,427,787	Argentina.geojsonl.zip	323MB
Bolivia	1,015,151	Bolivia.geojsonl.zip	82MB
Brazil	18,711,536	Brazil.geojsonl.zip	1.6GB
Chile	2,208,744	Chile.geojsonl.zip	187MB
Colombia	6,083,821	Colombia.geojsonl.zip	482MB
Ecuador	3,674,190	Ecuador.geojsonl.zip	287MB
Guyana	3,339	Guyana.geojsonl.zip	236KB
Paraguay	990,756	Paraguay.geojsonl.zip	73MB
Peru	1,710,431	Peru.geojsonl.zip	144MB
Uruguay	2,656	Uruguay.geojsonl.zip	200KB
Venezuela	6,572,969	Venezuela.geojsonl.zip	497MB

Figura 11: Descarga de datos de Colombia

Figura 12: Archivo descargado

2. Se descomprimió localmente y se cargó el archivo .geojsonl directamente como capa vectorial en QGIS.

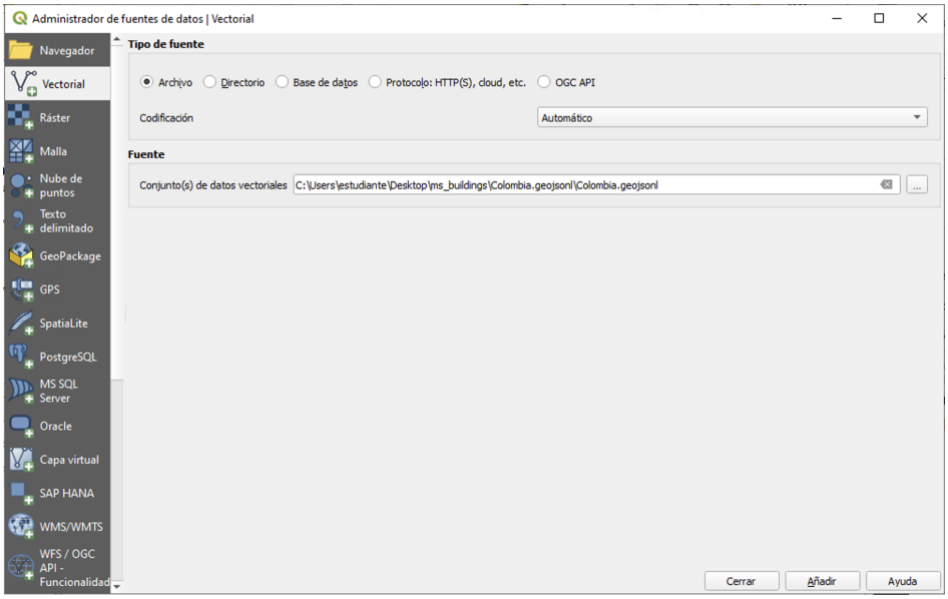

Figura 13: Inserción de archivo en QGIS

3. Aunque la tabla aparecía inicialmente vacía, en el mapa se visualizaron correctamente millones de polígonos, lo cual confirma que la geometría sí fue reconocida.

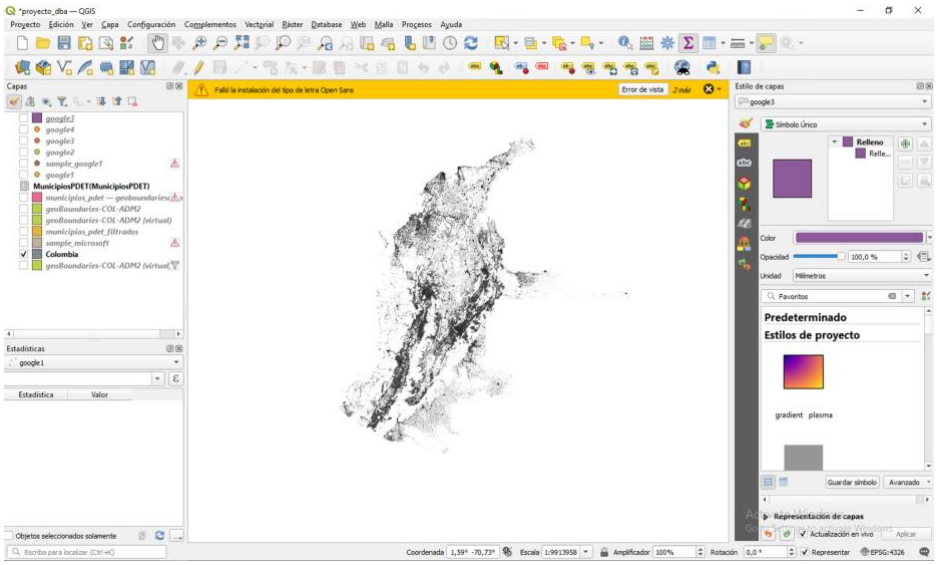

Figura 14: Archivo exportado en QGIS

4. Se visualizaron las primeras 200 entidades aplicando un filtro.

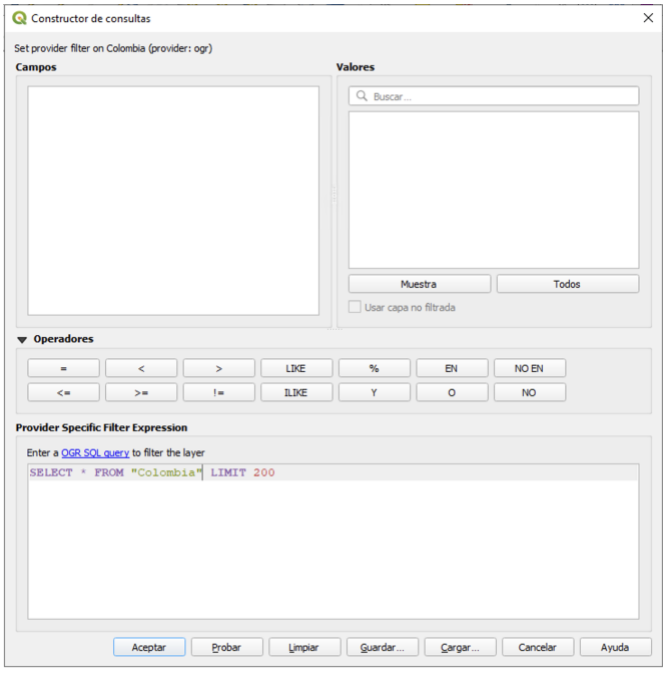

Figura 15: Datos filtrados

5. Ese subconjunto de 200 features se exportó a GeoJSON como: sample_microsoft.geojson

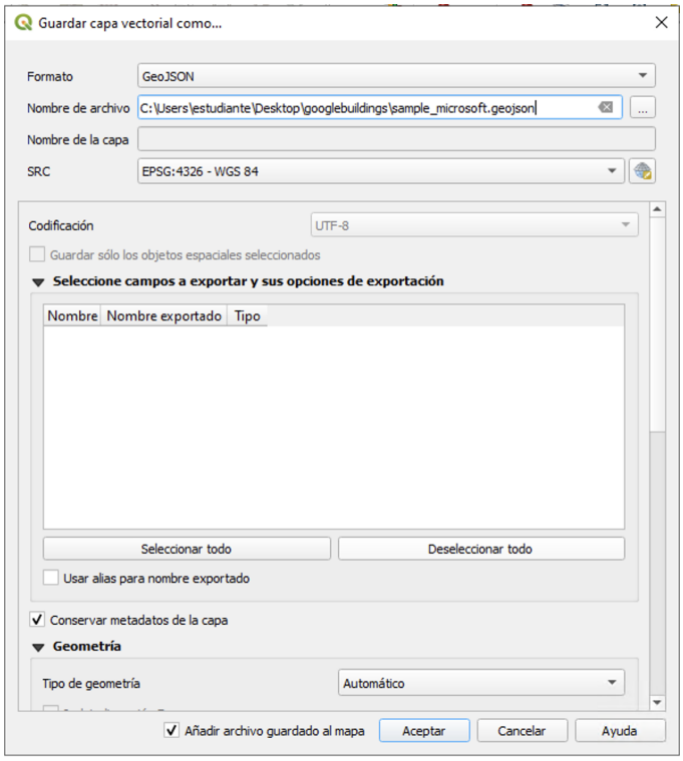

Figura 16: Exportación de archivo con datos filtrados

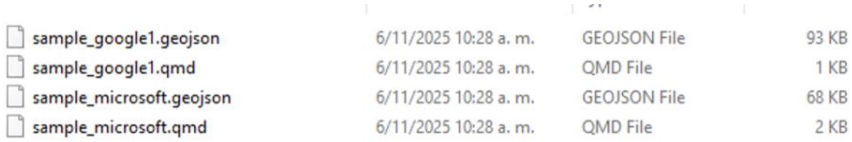

Figura 17: Archivo sample_microsoft.geojson

Estos dos samples representan las dos fuentes requeridas para esta entrega, manteniendo significado estadístico pero un tamaño pequeño que permite continuar con el trabajo sin bloquear carga ni visualización.

Eficiencia de Carga

Los datasets completos de footprints (tanto Google como Microsoft) son extremadamente grandes, con varios millones de polígonos. Trabajar con ellos completos en esta etapa generaría consumo excesivo de memoria, riesgo de fallas al intentar exportar, tiempos de carga muy lentos y dificultad para manipularlos en equipos personales. Debido a esto, se tomó la decisión de aplicar una estrategia de muestreo controlado.

En el caso de Google Open Buildings, el dataset no se descarga por país sino por cuadrículas globales, y la información correspondiente a Colombia está repartida en varios tiles. Por esta razón, se trabajó únicamente con los tiles que cubren el territorio colombiano y, a partir de ellos, se generó una muestra representativa de 200 registros en lugar de intentar manejar el dataset completo.

Por otro lado, el dataset de Microsoft Building Footprints para Colombia se distribuye como un único archivo nacional en formato GeoJSONL, el cual contiene millones de polígonos. Cargarlo completo implicaría tiempos de procesamiento muy altos y posible saturación de memoria. Por este motivo, también se generó una muestra reducida de 200 registros, con el objetivo de facilitar el manejo y permitir pruebas posteriores sin afectar el rendimiento del entorno de trabajo.

Al generar solamente 200 features por dataset, QGIS puede trabajar de manera fluida, y se evita saturar almacenamiento y procesamiento. Esto también permite que el siguiente equipo encargado de NoSQL pueda probar carga, indexación (2dsphere) y consultas sin esperar horas de importación.

Auditoría Inicial (EDA)

Descripción general del proceso realizado

Para llevar a cabo el análisis exploratorio de los datos (EDA), se desarrolló una serie de scripts en Python y Bash que permiten automatizar la carga, inspección y comparación de los datasets de edificaciones.

En particular, se implementaron:

- Dos scripts de Python para cargar los datos del dataset de Google y del dataset de Microsoft.
- Dos scripts de Bash que ejecutan estos procesos de forma automatizada, sin intervención manual.
- Un script de Python para el EDA, encargado de:
    (a) conectarse a MongoDB,
    (b) inspeccionar la estructura de cada colección,
    (c) calcular estadísticas descriptivas,
    (d) evaluar la calidad de los datos,
    (e) y generar un archivo .txt con todo el reporte.

Este flujo garantiza replicabilidad, control y una base sólida para las decisiones posteriores en el proyecto.

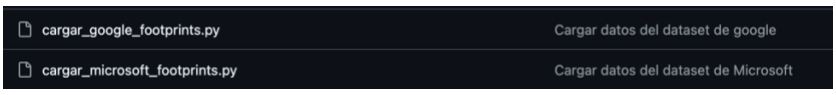

Figura 18. Scripts utilizados para cargar los datasets de Google y Microsoft en MongoDB.

Resumen general de las colecciones

Colección	Cantidad de documentos
Google Open Buildings	**200 edificaciones**
Microsoft Building Footprints	**200 edificaciones**
Municipios PDET	**101 municipios**
Total footprints analizados	**400 edificaciones**

Estructura de los datasets

Google Open Buildings

En el caso de Google Open Buildings, el dataset posee una estructura mucho más completa y rica en metadatos. Cada registro incluye las coordenadas de ubicación (latitud y longitud), el área aproximada de la edificación, el nivel de confianza asignado por el algoritmo de detección, el código Plus Code asociado al lugar y la geometría tanto en formato WKT como en formato GeoJSON. Además, incorpora un campo de fecha que indica el momento exacto en el que la edificación fue cargada a la base de datos.

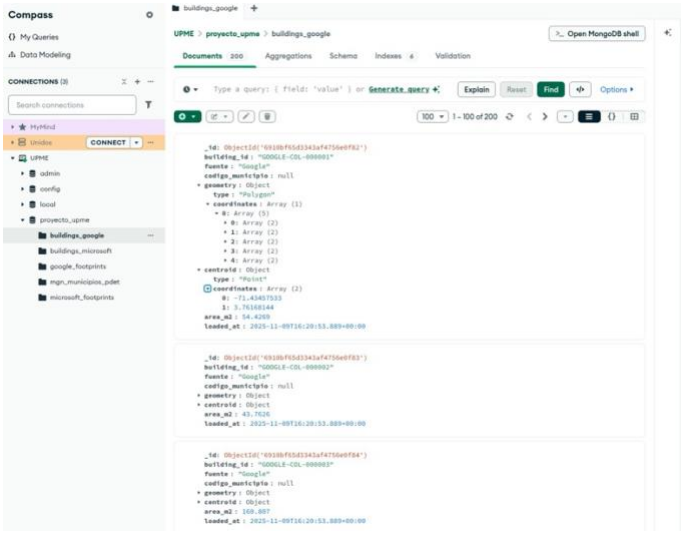

Figura 19. Ejemplo de un registro del dataset Google Open Buildings, mostrando metadatos y geometría.

Microsoft Building Footprints

Por su parte, el dataset de Microsoft Building Footprints se caracteriza por ser significativamente más minimalista. Cada edificación incluye únicamente la geometría en formato Polygon GeoJSON, a la cual se le añadió el cálculo del centroide para facilitar su análisis espacial. A diferencia de Google, este dataset no incorpora información adicional sobre área, niveles de

confianza, plus codes ni otros metadatos complementarios, lo que lo convierte en un conjunto de datos más simple desde el punto de vista descriptivo.

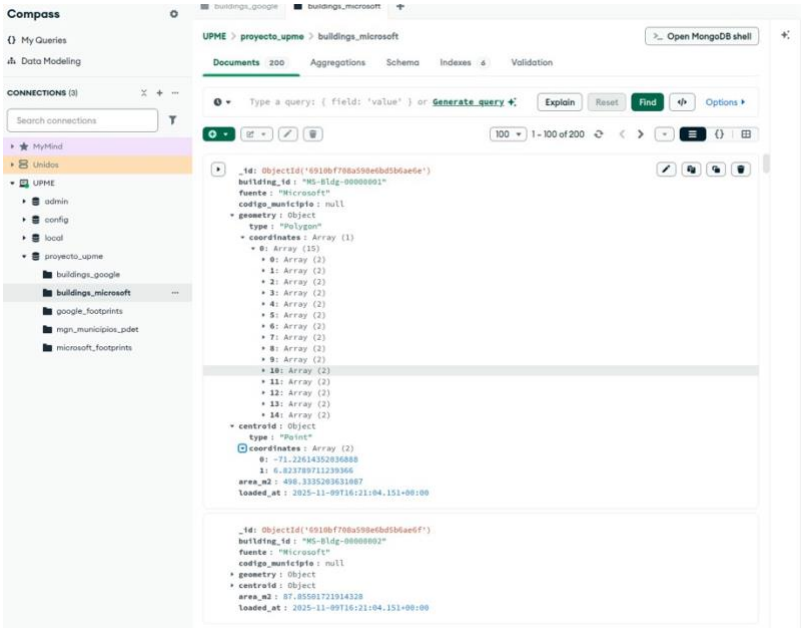

Figura 20. Ejemplo de un registro del dataset Microsoft Building Footprints, representado como un polígono GeoJSON

Resultados del Análisis Exploratorio

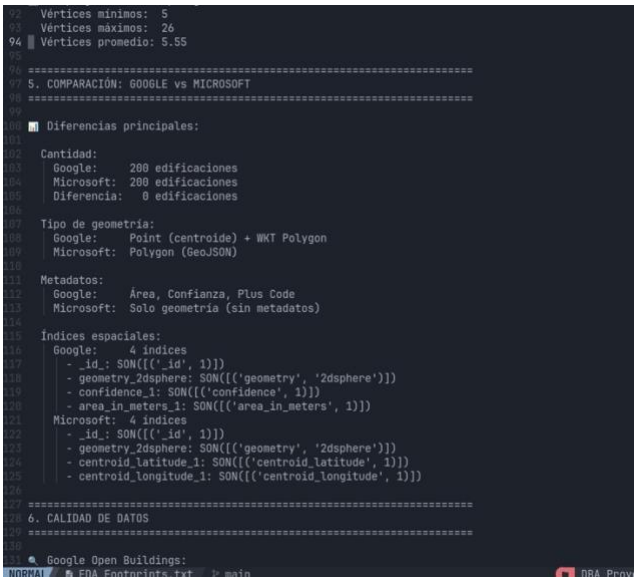

Figura 21. Fragmento del reporte generado automáticamente por el script EDA, mostrando estadísticas de área, confianza y rangos geográficos.

Google Open Buildings

Categoría	Métrica	Valor
Área (m²)	Área mínima	6.56 m²
	Área máxima	2550.45 m²
	Área promedio	97.53 m²
	Área total	19 506.04 m² (1.95 ha)
Distribución por rangos	0–50 m²	66 edificaciones
	50–100 m²	72 edificaciones
	100–200 m²	50 edificaciones
	200–500 m²	11 edificaciones
	>1000 m²	1 outlier
Niveles de confianza	Confianza mínima	0.6512
	Confianza máxima	0.9471
	Confianza promedio	0.7903
Distribución porcentual	Confianza 0.6+	15%
	Confianza 0.7+	37%
	Confianza 0.8+	41.50%
	Confianza 0.9+	6.50%
Rangos geográficos	Latitud	0.261145° a 4.818556°
	Longitud	–73.441730° a –67.558773°

Microsoft Building Footprints

Categoría	Métrica	Valor
Tipo de geometría	Geometrías válidas	100% polígonos (200 geometrías)
Rangos geográficos	Latitud	6.049430° a 12.348873°
	Longitud	–71.497689° a –70.700090°
Complejidad geométrica	Vértices mínimos	5
	Vértices máximos	26
	Vértices promedio	5.55 vértices

Comparación entre Google y Microsoft

Aspecto	Google Open Buildings	Microsoft BF
Cantidad	200	200
Tipo de geometría	Punto + Polygon	Polygon
Metadatos	Área, confianza, plus code	No incluye
Índices espaciales	4	4
Región cubierta	Amazonía	Norte del país

Calidad de los datos

Google Open Buildings: En el caso del dataset de Google Open Buildings, no se identificaron valores nulos en ninguno de los campos analizados (coordenadas, área, confianza y geometría), lo que demuestra una estructura consistente en toda la muestra. Únicamente se encontró un outlier correspondiente a una edificación con un área superior a 1000 m². Más allá de ese caso puntual, todas las geometrías fueron validadas correctamente, presentando un 100% de registros con geometrías completas y sin errores.

Microsoft Building Footprints: Por su parte, el dataset de Microsoft Building Footprints también presentó un comportamiento sólido en términos de calidad: no se encontraron valores nulos en los campos de centroides ni en la geometría. Todos los polígonos resultaron consistentes y estructuralmente correctos, sin geometrías corruptas o incompletas, lo que garantiza la integridad del conjunto de datos para posteriores procesos de análisis espacial.

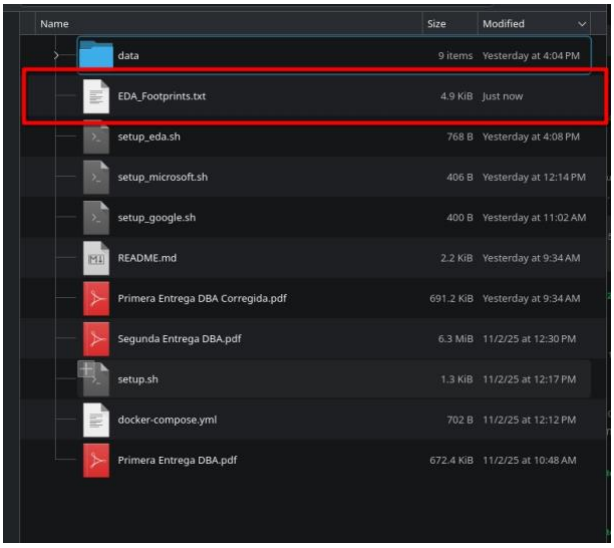

Figura 22. Archivo de resultados exportado por el script EDA, el cual resume la exploración completa de los datasets.

Conclusión

La tercera entrega permitió construir un proceso sólido y reproducible para la preparación y auditoría inicial de los datos de edificaciones. A partir del uso de QGIS, Python, Bash y MongoDB, se generaron muestras representativas de los datasets de Google y Microsoft, evitando el manejo de millones de registros y garantizando eficiencia en la carga y análisis.

El EDA evidenció que ambos datasets presentan buena calidad: no se encontraron valores nulos y todas las geometrías resultaron válidas. Google aporta metadatos adicionales como área y confianza, mientras que Microsoft ofrece polígonos simples pero consistentes. Estas diferencias serán clave para las etapas posteriores de integración e indexación geoespacial.

En conjunto, esta entrega deja preparados datos confiables y un flujo de trabajo automatizado que facilitará las siguientes fases del proyecto, asegurando una base técnica adecuada para continuar con la integración y análisis en la base de datos NoSQL.