Entrega 4
Reproducible Geospatial Analysis Workflow

Juan David Castillo Laverde
Juan Pablo Dávila Martínez
Eugenia Victoria Dayoub Barito
Luis Fernando Lee Rodríguez

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

1. Introducción

El presente informe documenta el desarrollo del Reproducible Geospatial Analysis Workflow correspondiente a la Entrega 4 del proyecto DBA. El objetivo consiste en diseñar e implementar un pipeline geoespacial completamente reproducible que permita:

- Filtrar footprints de edificaciones provenientes de Google Open Buildings y Microsoft Building Footprints usando los límites de los municipios PDET.
- Ejecutar consultas espaciales eficientes mediante índices 2dsphere en MongoDB.
- Calcular el número total de edificaciones y el área total de cubiertas dentro de cada municipio PDET.
- Exportar resultados a archivos CSV para análisis geoespacial y estadístico posterior.
- Asegurar trazabilidad, portabilidad y reproducibilidad en cualquier entorno mediante Docker y scripts estandarizados.

La contenerización con Docker y el uso de scripts automatizados garantizan que cualquier miembro del equipo pueda obtener resultados idénticos y verificables, independientemente del sistema operativo utilizado.

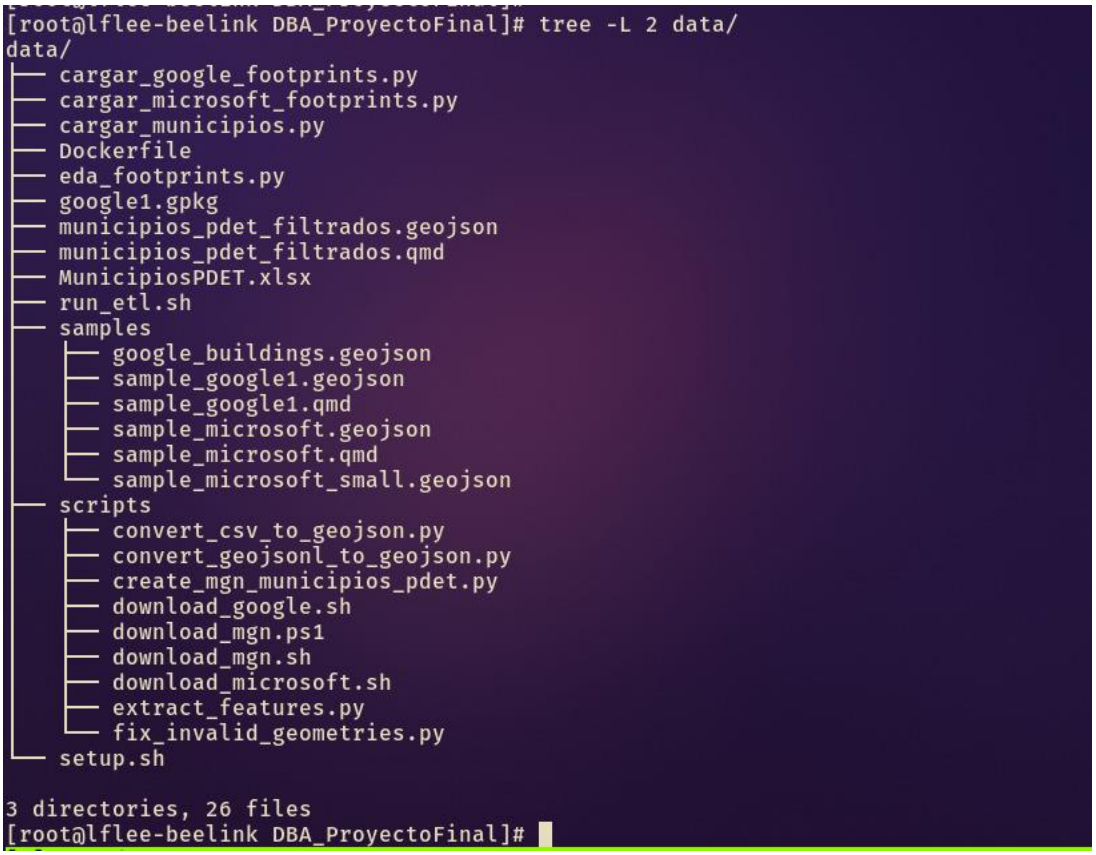


Figura 1. Vista general de los archivos Google/Microsoft o una carpeta del proyecto mostrando la estructura base.

2. Metodología

El workflow se fundamenta en cuatro pilares: automatización, contenerización, scriptabilidad y verificación.

2.1. Automatización

Se desarrolló un script maestro:

pdet_pipeline.sh

Este archivo ejecuta de forma secuencial:

- Validación de conexión a MongoDB.
- Verificación y creación de índices geoespaciales.
- Ejecución del filtrado PDET.
- Cálculo de conteos y áreas totales.
- Exportación de un archivo consolidado en CSV.
- Registro de logs durante todo el proceso.

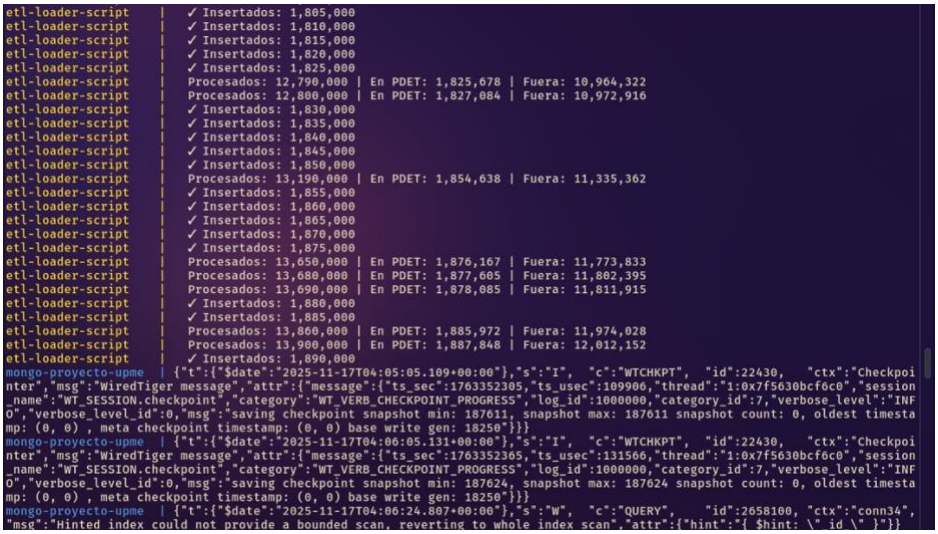


Figura 2. La terminal ejecutando pdet_pipeline.sh.

2.2. Contenerización

Todo el proceso se ejecuta sobre Docker, utilizando el contenedor:

mongo-upme

Ventajas:

- Reproducibilidad entre Windows/macOS/Linux.
- Aislamiento del entorno geoespacial.

- No requiere instalar Python, bibliotecas geoespaciales o mongosh en el equipo del usuario.

Figura 3. Salida de docker ps con el contenedor mongo-upme activo.

2.3. Scriptabilidad

El análisis se realiza exclusivamente mediante scripts:

- cargar_municipios.py
- cargar_google_footprints.py
- cargar_microsoft_footprints.py
- exportar_pdet_google.py
- exportar_pdet_microsoft.py
- run_pdet_pipeline.sh

Esta estructura modular permite reproducir, modificar o extender el workflow sin afectar la coherencia del proceso.


Figura 4. Código de uno de los scripts: ensure_indexes.js, run_pdet_filter.js o los export_*.py.

2.4.Verificación

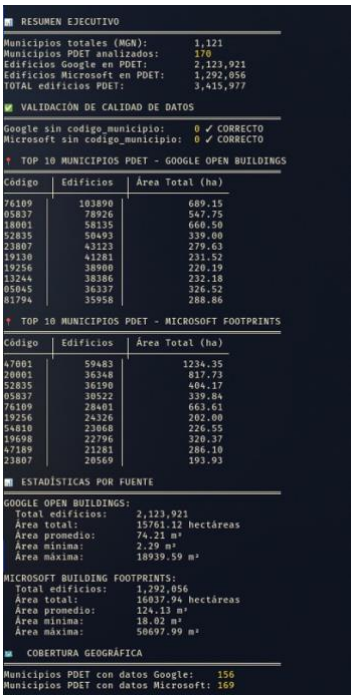
La ejecución genera:

logs_pipeline.txt

Este archivo contiene:

- Timestamps del proceso.
- Validación de índices.
- Conteo de edificios procesados.
- Lectura y escritura de archivos.
- Mensajes de éxito o error.

Este mecanismo garantiza transparencia, trazabilidad y auditoría del pipeline.



```
📊 RESUMEN EJECUTIVO
Municipios totales (MGN):        1,121
Municipios PDET analizados:      170
Edificios Google en PDET:        2,123,921
Edificios Microsoft en PDET:     1,292,056
TOTAL edificios PDET:            3,415,977

✅ VALIDACIÓN DE CALIDAD DE DATOS
Google sin codigo_municipio:     0 ✓ CORRECTO
Microsoft sin codigo_municipio:  0 ✓ CORRECTO

📍 TOP 10 MUNICIPIOS PDET - GOOGLE OPEN BUILDINGS
Código | Edificios | Área Total (ha)
76109  |   103890  |        689.15
05837  |    78926  |        547.75
18001  |    58135  |        660.50
52835  |    50493  |        339.00
23807  |    43123  |        279.63
19130  |    41281  |        231.52
19256  |    38900  |        220.19
13244  |    38386  |        232.18
05045  |    36337  |        326.52
81794  |    35958  |        288.86

📍 TOP 10 MUNICIPIOS PDET - MICROSOFT FOOTPRINTS
Código | Edificios | Área Total (ha)
47001  |    59483  |       1234.35
20001  |    36348  |        817.73
52835  |    36190  |        404.17
05837  |    30522  |        339.84
76109  |    28401  |        663.61
19256  |    24326  |        202.00
54810  |    23868  |        226.55
19698  |    22796  |        320.37
47189  |    21281  |        286.10
23807  |    20569  |        193.93

📊 ESTADÍSTICAS POR FUENTE
GOOGLE OPEN BUILDINGS:
  Total edificios:     2,123,921
  Área total:          15761.12 hectáreas
  Área promedio:       74.21 m²
  Área mínima:         2.29 m²
  Área máxima:         18939.59 m²

MICROSOFT BUILDING FOOTPRINTS:
  Total edificios:     1,292,056
  Área total:          16037.94 hectáreas
  Área promedio:       124.13 m²
  Área mínima:         18.02 m²
  Área máxima:         50697.99 m²

🗺️  COBERTURA GEOGRÁFICA
Municipios PDET con datos Google:    156
Municipios PDET con datos Microsoft: 169
```

Figura 5. Logs del pipeline con la información de Municipios en Google y Microsoft

3. Precisión de las Operaciones Espaciales

3.1. Índices 2dsphere

Para garantizar precisión y eficiencia, se definieron índices en:

municipalities_pdet
buildings_google
buildings_microsoft

Mediante:

```
db.collection.createIndex({ geometry: "2dsphere" });
```

Esto permite:

- Consultas $geoWithin precisas.
- Intersección adecuada de polígonos.
- Procesamiento eficiente incluso con geometrías complejas.

Figura 6. Índices 2dsphere creados

3.2. Filtrado espacial

El filtrado se realizó municipio por municipio usando:

```
{
        geometry: {
        $geoWithin: { $geometry: municipio.geometry }
        }
}
```

Este método es apropiado para esquemas con Polygon y MultiPolygon, asegurando que solo footprints dentro de los límites municipales sean incluidos.

4. Estructura del pipeline

4.1. Flujo general del proceso

- Levantar contenedor MongoDB.
- Ejecutar run_pdet_pipeline.sh.

- Crear índices geoespaciales.
- Filtrar footprints Google por municipio PDET.
- Filtrar footprints Microsoft por municipio PDET.
- Calcular área total (m²).
- Exportar CSV consolidado.
- Registrar logs del proceso.

4.2. Archivos generados

A diferencia de versiones anteriores del pipeline, la versión final genera un único archivo consolidado:

resultados_pdet_por_municipio.csv

Incluye:

- Código DANE del municipio
- Total de edificios Google
- Total de edificios Microsoft
- Área total Google (m²)
- Área total Microsoft (m²)

Además se generó:
- logs_pipeline.txt

5. Resultados

A continuación se presentan las tablas correspondientes al Top 10 de municipios PDET con mayor número de edificaciones según Google Open Buildings y Microsoft Building Footprints.

Estos valores se transformaron a áreas tanto en hectáreas como en metros cuadrados.

Tabla 1: Conteo y área por municipio (Google)

Código municipio	**Edificios**	**Área total (ha)**	**Área total (m²)**
47001	103890	689.15	6891500
20001	78926	547.77	5477700
52835	58135	660.50	6605000
05837	50493	339.00	3390000
76109	43123	279.63	2796300

19256	41281	231.52	2315200
54810	38900	220.19	2201900
19698	38386	232.18	2321800
47189	36337	326.52	3265200
23807	35958	288.86	2888600

Tabla 2: Conteo y área por municipio (Microsoft)

Código municipio	**Edificios**	**Área total (ha)**	**Área total (m²)**
76109	59483	1234.35	12343500
05837	36348	817.73	8177300
18001	36190	404.17	4041700
52835	30522	339.84	3398400
23807	28401	663.61	6636100
19130	24326	202.00	2020000
19256	23068	226.55	2265500
13244	22796	320.37	3203700
05045	21281	286.10	2861000
81794	20569	193.93	1939300

6. Mapas de Resultados Geoespaciales

A continuación, se presentan los mapas generados en QGIS a partir de la unión entre el shapefile oficial del MGN y la tabla resultados_pdet_por_municipio. Estos mapas permiten visualizar de forma clara los municipios PDET para los cuales se obtuvieron resultados válidos desde cada fuente de datos.

6.1.Municipios PDET con datos de Google Open Buildings

En el caso de Google Open Buildings, se obtuvo información válida para 156 municipios PDET.

El siguiente mapa muestra únicamente aquellos municipios donde el pipeline produjo resultados de conteo y área total de edificaciones.

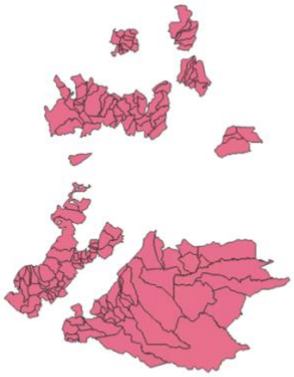


Figura 7. Municipios PDET procesados con Google Open Buildings

6.2. Municipios PDET con datos de Microsoft Building Footprints

Para Microsoft Building Footprints, se obtuvieron resultados válidos para 169 municipios PDET.

El siguiente mapa presenta los municipios donde se identificaron edificaciones mediante esta fuente.

Figura 8. Municipios PDET procesados con Microsoft Building Footprints

6.3. Comparación general

La cobertura espacial difiere entre fuentes:

- Google cubre 156 municipios.
- Microsoft cubre 169 municipios.

Microsoft presenta mayor continuidad en áreas urbanas, mientras que Google detecta mejor edificaciones dispersas.

7. Ejecución del Workflow

Para ejecutar completamente la Entrega 4, el usuario debe correr:

```
cd DBA_ProyectoFinal
docker-compose up -d mongo-upme
chmod +x entrega4/run_pdet_pipeline.sh
./entrega4/pdet_pipeline.sh
```

8. Conclusiones

La Entrega 4 consolidó un flujo de análisis geoespacial totalmente reproducible, que permite estimar de manera consistente y auditable el número de edificaciones y el área total de cubiertas dentro de los municipios PDET. La implementación mediante Docker garantiza independencia del entorno y evita inconsistencias entre sistemas operativos.

El uso de índices 2dsphere permitió lograr precisión en las operaciones espaciales realizadas, especialmente en filtros $geoWithin y cálculos $geoArea. Los archivos CSV generados constituyen la base para análisis comparativos entre Google y Microsoft, y satisfacen los requerimientos técnicos del proyecto.

Finalmente, la estructura modular del workflow facilita su actualización, depuración y ampliación, permitiendo que futuros análisis puedan incorporarse sin modificar la integridad del proceso. Este pipeline cumple los criterios de trazabilidad, reproducibilidad y transparencia exigidos en la entrega.