MODELO DE INTEGRACIÓN DE BIG GEOSPATIAL DATA AL PROCESO DE ACTUALIZACIÓN CATASTRAL: UN ESTUDIO CON LA PLATAFORMA DE CÓDIGO ABIERTO APACHE SEDONA.



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Xiomara Valentina Gutiérrez Mora

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS
FACULTAD DE INGENIERÍA
INGENIERÍA CATASTRAL Y GEODESIA
PROYECTO DE GRADO
Bogotá D.C.
2025

MODELO DE INTEGRACIÓN DE BIG GEOSPATIAL DATA AL PROCESO DE ACTUALIZACIÓN CATASTRAL: UN ESTUDIO CON LA PLATAFORMA DE CÓDIGO ABIERTO APACHE SEDONA.

Autor:

Xiomara Valentina Gutiérrez Mora

Modalidad:

Proyecto de Investigación/Innovación

Presentado para optar por el título de:

Ingeniero(a) Catastral y Geodesta

Director:

Paulo César Coronado Sánchez

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD DE INGENIERÍA

INGENIERÍA CATASTRAL Y GEODESIA

Bogotá D.C.

2025

ÍNDICE GENERAL

1. INTRODUCCIÓN **4**

2. PROBLEMA DE INVESTIGACIÓN **5**

3. OBJETIVOS **7**

3.1. Objetivo general 7

3.2. Objetivos específicos 7

4. MARCO TEÓRICO **7**

4.1. Definición modelo - Norma Iso 42010 7

4.2. Big Geospatial Data 8

4.3. Apache Sedona: Tecnología para el Procesamiento Distribuido de Datos Geoespaciales. 12

4.4. Wherobots Cloud 17

4.5. Catastro multipropósito: Proceso actualización Catastral. 19

4.6. Método ADM – TOGAF 24

5. MÉTODOS Y TÉCNICAS DE ANÁLISIS DE DATOS **27**

5.1. Metodología aplicada en la construcción del modelo. 28

5.2. Metodología para la integración de Big Geospatial Data en la actualización catastral. 39

5.3. Metodología para la adquisición del dominio tecnológico de Apache Sedona. 42

5.4. Metodología de validación técnica del experimento: Entorno tecnológico. ... 48

6. DESCRIPCIÓN DE RESULTADOS **49**

6.1. Dimensiones de Big Data en relación con la actualización catastral. 50

6.2. Proceso de integración de Big Geospatial Data en la actualización catastral. 56

6.3. Resultado desempeño del prototipo en comparación con métodos tradicionales. 57

6.4. Resultado del modelo de integración de Big Geospatial Data en la actualización catastral. 59

7. DISCUSIÓN DE RESULTADOS **69**

7.1. Desafíos actuales en la gestión de datos geoespaciales 69
7.2. Rol de Apache Sedona y el procesamiento distribuido 69
7.3. Más allá del acceso: el reto está en el manejo 70
7.4. Limitaciones y proyecciones del modelo propuesto 71
7.5. Consideraciones éticas y de gobernanza de datos. 72
8. CONCLUSIONES 73
9. RECOMENDACIONES 74
10. BIBLIOGRAFÍA 75
11. ANEXOS 83
Anexo A 83
Guía Técnica para la Instalación y Configuración de un Clúster Standalone de Spark con Apache Sedona. *83*
Anexo C 88
Wherobots Cloud: Aplicación práctica con Apache Sedona. *88*

ÍNDICE DE FIGURAS

Figura. 1. A)Ecosistema de Apache Sedona. B) Arquitectura del sistema Sedona. 14
Figura. 2. Diagrama: Lectura desde PostgreSQL mediante CDC (Change Data Capture). 16
Figura. 3. Wherobots: The Spatial Intelligence Cloud 17
Figura. 4. Estructura ADM. 25
Figura. 5. Desarrollo del modelo empleando método ADM Fase 1. 31
Figura. 6. Desarrollo del modelo empleando método ADM Fase 2. 39
Figura. 6. Proceso de identificación de problemáticas en el uso de datos 41
Figura. 7. Fundamentos conceptuales y preparación técnica; Apache Sedona. 45
Figura. 8. Fases 3 y 4 del modelo propuesto: Implementación práctica inicial e implementación en entornos tecnológicos con Apache Sedona. 48

ÍNDICE DE TABLAS

Tabla 1. Preguntas orientadoras para delimitar alcance de las herramientas 49
Tabla 2. Dimensión de Big Data: Variedad y actualización catastral. 51
Tabla 3. Dimensión de Big Data: Volumen y actualización catastral. 52
Tabla 4. Dimensión de Big Data: Velocidad y actualización catastral. 53
Tabla 5. Dimensión de Big Data: Veracidad y actualización catastral. 55
Tabla 6. Dimensión de Big Data: Valor y actualización catastral. 56
Tabla 7. Integración de Big Geospatial Data en los procesos de actualización catastral. 57
Tabla 8. Comparación del desempeño del prototipo frente a métodos tradicionales en entornos local y nube. 58
Tabla 9. Preguntas genereales formuladas para estructurar el modelo. 60

1. INTRODUCCIÓN

En la actualidad, la cantidad de datos geoespaciales que se generan en los procesos catastrales y en diferentes fuentes de información han aumentado en cuanto a volumen, velocidad y variedad. Esta situación ha hecho necesario adoptar arquitecturas de datos que permitan gestionar, integrar y procesar esta información de manera organizada. En el ámbito de la gestión catastral y catastro multipropósito, donde se manejan grandes volúmenes de datos geoespaciales, la implementación de tecnologías que optimicen su procesamiento es clave para mejorar la toma de decisiones en el territorio.

Según la Resolución 1040 de 2023 el catastro multipropósito no es solo un inventario de bienes inmuebles, es una herramienta que se divide en cuatro procesos: formación, actualización, conservación y difusión catastral, con el fin de apoyar a la gestión y ordenamiento territorial en Colombia. Los gestores y operadores catastrales deben asegurar que la información geográfica y alfanumérica esté disponible, estructurada y cumpla con estándares mínimos para gestionar o reportar la información como el modelo "LADM COL SINIC V1.0", que unifica los elementos catastrales del *Modelo Extendido Catastro Registro LADM_COL* con el fin de mantener coherencia entre los datos gestionados localmente por cada gestor catastral y lo que se consolida en el SINIC según la Resolución 301 de 2025. Sin embargo, los gestores y/o operadores presentan retos técnicos para manejar grandes volúmenes de datos, especialmente en términos de geoprocesamiento, recursos de almacenamiento, escalabilidad, eficiencia e interoperabilidad, lo que ha llevado a buscar soluciones que permitan mejorar esos procesos.

Esta investigación propone un modelo de arquitectura de datos empresarial para integrar Big Geospatial Data en el proceso de actualización catastral. El modelo se desarrolla con base en el marco TOGAF, usando el método ADM representado mediante el lenguaje ArchiMate, lo que permite estructurar la arquitectura en capas y alinear la tecnología con los procesos de negocio.

El paradigma de Big Data comprende un amplio campo de conocimiento que puede aplicarse al Catastro Multipropósito. En este proyecto se aborda específicamente el proceso de Actualización Catastral, el cual se caracteriza e integra dentro del ámbito del Big Geospatial Data para la construcción del modelo. Asimismo, se adopta como caso de estudio una tarea operativa dentro de los subprocesos de la actualización catastral: la captura y procesamiento de datos geoespaciales vectoriales asociados a la información normativa urbanística de los predios de Bogotá D.C.. Este enfoque permite evaluar el desempeño de la herramienta de Big Geospatial Data Apache Sedona en comparación con herramientas tradicionales utilizadas en el tratamiento y análisis de datos geoespaciales.

2. PROBLEMA DE INVESTIGACIÓN

En Colombia, el catastro multipropósito enfrenta desafíos debido a la fragmentación de las fuentes de información y la baja interoperabilidad de datos entre entidades públicas. Como lo señala Cabello (2023), en muchos gobiernos, la interoperabilidad y la estandarización de los datos siguen siendo limitadas y los datos se siguen gestionando en silos, lo cual obstaculiza la construcción de un sector público basado en datos de valor. Esta situación se manifiesta en la coexistencia de múltiples fuentes de datos sin integración entre

ellas, lo que restringe su aprovechamiento para procesos administrativos y técnicos.

Según Henry Rodríguez Sosa (2022), el catastro multipropósito y la ciencia de datos se articulan como elementos esenciales para la toma de decisiones territoriales fundamentadas. Para ello es necesario procesar altos volúmenes de información a altas velocidades, para encontrar patrones y representar los resultados de manera comprensible para todos los niveles de la organización. En el contexto del proceso de actualización catastral con enfoque multipropósito, estos problemas se incrementan debido a fallas significativas de datos en los registros (fichas catastrales y escrituras), ausencias en completitud de la parte cartográfica (mapas) y muy baja integración de la parte gráfica con la parte alfanumérica, lo que limita la eficiencia de los sistemas actuales.

A esto se suma la desactualización de las herramientas disponibles para el procesamiento de Big Data. Reíta y Salinas (2016) concluyeron que la adopción de esta tecnología es viable y beneficiosa, ya que puede optimizar procesos y mejorar los análisis para la gestión estratégica. Sin embargo, también señalaron que enfrenta desafíos importantes, como la falta de cultura informática en el país y los altos costos de infraestructura, lo que dificulta su implementación, especialmente en pequeñas y medianas empresas. De acuerdo con el documento de fortalecimiento institucional de la Unidad Administrativa Especial de Catastro Distrital (2020), los sistemas de gestión integral actual no escalan adecuadamente frente a las exigencias operativas, debido a limitaciones en su arquitectura tecnológica para ejecutar procesos de gestión catastral. Esta situación afecta la capacidad de las entidades públicas para ampliar el alcance y el nivel de detalle en sus procedimientos y en su gestión (SDP, 2020).

3. OBJETIVOS

3.1. Objetivo general

- Definir un modelo para la integración de Big Geospatial Data al proceso de actualización catastral a partir de un estudio de la plataforma de código abierto para análisis espacial distribuido Apache Sedona.

3.2. Objetivos específicos

- Caracterizar el rol del Big Geospatial Data en el proceso de actualización catastral.
- Definir un proceso de integración de Big Geospatial Data en los procesos de actualización catastral.
- Evaluar el desempeño del prototipo en comparación con métodos tradicionales para validar la capa tecnológica del modelo propuesto.

4. MARCO TEÓRICO

4.1. Definición modelo - Norma Iso 42010

En el marco de este proyecto, un modelo se considera como un esquema o representación de un dominio de la realidad a partir de sus propiedades fundamentales encarnados en los elementos que lo constituyen, las relaciones entre ellos, en un contexto específico y los principios que guían su desarrollo y evolución.

4.2. Big Geospatial Data

En el contexto de Big Data, según Jesús García Herrero (2018) los conjuntos de datos que superan las capacidades de los sistemas tradicionales de bases de datos en términos de captura, almacenamiento, gestión y análisis juegan un rol fundamental. Big data no se trata únicamente del tamaño de la información; la *variedad* y la *velocidad* con la que se generan, almacenan y analizan los datos son igualmente críticas. Estas dimensiones, conocidas como las "3V" de Big Data —volumen, velocidad y variedad—, se complementan con una cuarta: el valor, que enfatiza los beneficios derivados de su uso, como la reducción de costos, mejoras en la eficiencia operativa y optimización del negocio.

Además de las cuatro dimensiones tradicionales del big data, se incorpora una quinta: la *veracidad*, la cual se refiere a la calidad y confiabilidad de los datos. Para lograr una integración e interoperabilidad adecuadas entre grandes volúmenes de datos en distintos formatos, es necesario que estos sean precisos, consistentes y convertibles. La veracidad permite asegurar que los datos utilizados en los análisis sean pertinentes para el contexto, facilitando así la extracción de información útil y la toma de decisiones basada en evidencia. (García et al., 2018). En este sentido, Mete (2023) indica que la precisión es un aspecto que debe considerarse en el tratamiento del big data, ya que los datos deben contener información verificable y eliminar aquellos que carezcan de valor analítico.

Según Murazzo et al. (2018), el reto del Big Data es convertir grandes volúmenes de datos en inteligencia de negocio, lo cual implica una gestión diferente basada en el análisis en tiempo real para orientar la toma de decisiones. No se trata únicamente de recopilar datos y realizar análisis detallados, sino de generar interpretaciones rápidas a partir de datos provenientes de diversas fuentes. En este contexto, Becerra et al (2018) mencionan que el

uso de Big Data permite implementar mecanismos automatizados de verificación y control de calidad sobre grandes volúmenes de información. Además, el uso de algoritmos puede aumentar la eficiencia, reducir costos operativos y fortalecer la legitimidad de los procesos institucionales (Becerra et al., 2018).

Por tanto, García et al. (2018) definen a la tecnología big data como el "*conjunto de arquitecturas y herramientas informáticas destinadas a la manipulación, gestión y análisis de grandes volúmenes de datos desde todo tipo de fuentes, diseñadas para extraer valor y beneficio de los mismos, con una amplia variedad en su naturaleza, mediante procesos que permitan capturar, descubrir y analizar información a alta velocidad y con un coste reducido*". (p.15)

En el contexto de las ciudades inteligentes, Mete et al (2023), mencionan que el aumento de los datos geoespaciales ha dado origen al concepto de Big Geospatial Data, el cual representa un reto para las herramientas SIG tradicionales. Este crecimiento ha hecho evidente la necesidad de gestionar de manera más eficiente los datos geoespaciales, ya que almacenar, procesar, analizar y visualizar grandes volúmenes de este tipo de datos resulta muy difícil utilizando solo el software y hardware convencional que normalmente se emplea en los sistemas de información geográfica (GIS).

Anteriormente, acceder a este tipo de información requería altos costos y procedimientos complejos; en la actualidad, es posible obtenerla mediante diversas fuentes como imágenes satelitales, captura de datos mediante plataformas aéreas (aviones, helicópteros y drones) o dispositivos móviles (celulares), programas de visualización geográfica como Street View y Google maps, cartografía, uso de datos de localización

recopilados mediante GPS y metadatos de redes sociales. Asimismo, se dispone de información generada por entidades estatales, que complementa otros conjuntos de datos geoespaciales. (Yu, J., & Sarwat, M., 2021)

De igual manera, en la conferencia "How Apache Sedona is Revolutionizing Geospatial Data Analysis", presentada por Jia Yu, menciona que el *Big Geospatial Data* son grandes volúmenes de información espacial que provienen de fuentes como satélites, datos de movilidad, modelos climáticos, GPS y sensores, entre otros. Estos datos se usan en temas como el monitoreo ambiental y la planificación urbana, pero también generan retos importantes para almacenarlos, procesarlos y analizarlos, debido a su complejidad tanto espacial como temporal (Yu, 2023).

En ese sentido, es útil aclarar la diferencia entre algunos conceptos que suelen confundirse. Como mencionan Tokaj, Yu y Sarwat (2023), "*los términos 'datos espaciales' y 'datos geoespaciales' a menudo se utilizan de manera intercambiable, pero tienen significados ligeramente diferentes. Los datos espaciales se refieren a cualquier dato que tenga un componente espacial o geográfico, y que pueda describir la ubicación, forma y relaciones de objetos en el espacio. Los datos geoespaciales son un subconjunto de los datos espaciales que se relacionan específicamente con la superficie terrestre y sus características*" (Tokaj, Yu, & Sarwat, 2023, p. 3).

Los datos geoespaciales, que contienen información con ubicaciones explícitas (como coordenadas GPS) o implícitas (como secciones o regiones), son fundamentales en este proceso. Su obtención y manejo se realiza a través de los Sistemas de Información Geográfica (SIG), los cuales utilizan sistemas de coordenadas para representar la superficie

terrestre en mapas. Esta proyección es necesaria para que la información tenga utilidad práctica en contextos de análisis espacial. A través de los GIS/SIG es posible organizar, interpretar y visualizar los datos geoespaciales, facilitando su uso en la planificación y gestión territorial. La modernización del Sistema Nacional de Referencia Espacial (NSRS) es crucial para mantener la precisión en la utilización de estos datos (National Geodetic Survey, 2019). Además, es fundamental entender los sistemas de referencia de coordenadas que se utilizan en GIS para asegurar la precisión en la proyección y análisis de estos datos (Earth Data Science, 2019).

Como menciona M. O. Mete (2023) en el artículo "Geospatial Big Data Analytics for Sustainable Smart Cities", las aplicaciones de Big Data geoespacial ofrecen herramientas de ciencia de datos predictiva, como la computación en malla (grid computing) y la computación paralela, para un procesamiento eficiente y rápido, orientado a construir un ecosistema urbano inteligente y sostenible. Plataformas como Hadoop y Spark han sido adaptadas para este tipo de procesamiento, mientras que la inteligencia artificial (IA) y el aprendizaje automático han mejorado su análisis, permitiendo el desarrollo de modelos predictivos y la automatización en la toma de decisiones (Zhang y Li, 2022). Sin embargo, cómo menciona Tuladhar et al. (2005) también surgen preocupaciones éticas, como la privacidad y el acceso equitativo a estos datos.

Barik et al. (2017) identificaron diversas problemáticas en el manejo y visualización de Big Geospatial Data, resaltando limitaciones significativas en la capacidad de herramientas como ArcMap y Geomesa para procesar Big Geospatial Data de manera eficiente. Señalaron desafíos relacionados con la integración de datos heterogéneos, el rendimiento en tiempo real y la escalabilidad de las herramientas actuales. A pesar de estas

limitaciones, destacan que cada herramienta tiene ventajas particulares dependiendo del tipo de datos y del entorno de uso, proporcionando una base para futuras investigaciones en la optimización de estas tecnologías.

4.3. Apache Sedona: Tecnología para el Procesamiento Distribuido de Datos Geoespaciales.

Apache Sedona, anteriormente conocido como GeoSpark, se ha consolidado como una herramienta esencial en el análisis de datos geoespaciales, transformando la manera en que se procesan conjuntos de datos geoespaciales en entornos distribuidos (Yu & Sarwat, 2021). De acuerdo con la página oficial del proyecto, entre sus características clave se destacan su capacidad de procesamiento en memoria, su compatibilidad con lenguajes de programación como Java, Scala y Python, y, en particular, su escalabilidad (Apache Sedona, s.f.).

Aach et al. (2023) y Tokaj, Yu y Sarwat (2025) señalan que la escalabilidad es la capacidad de un sistema para procesar volúmenes crecientes de datos mediante la adición de nodos de procesamiento, de forma que el rendimiento aumente proporcionalmente a la carga de trabajo. Este objetivo se alcanza mediante arquitecturas basadas en clústeres, en las que intervienen factores como la capacidad de almacenamiento, el poder de cómputo y el uso de algoritmos, incluyendo el aprendizaje automático, el procesamiento de grafos y el procesamiento por lotes.

En el artículo de Gousios (2018) este señala que, mediante el uso de pipelines basados en técnicas de optimización para el procesamiento de datos distribuidos, Apache

Spark mejora la escalabilidad y reduce los tiempos de procesamiento en entornos de Big Data. Según los desarrolladores de Apache Sedona este framework fue diseñado para tratar los datos geoespaciales como entidades de primera clase dentro de motores de cómputo distribuido como Apache Spark (procesamiento por lotes), Apache Flink (procesamiento en streaming) y Snowflake (SedonaSnow). Esta integración, como se muestra en la figura 1 sobre el ecosistema y arquitectura de Apache Sedona, permite ejecutar consultas espaciales complejas como uniones, agregaciones y cálculos de distancia, todo de forma paralela y optimizada en clústeres de máquinas. (Apache Sedona, s.f.)

Según Tokaj, Yu y Sarwat (2025), Apache Sedona soporta tipos de formatos vectoriales como Shapefiles, GeoJSON, GeoParquet y archivos CSV con geometría en formato WKT, así como datos ráster en formato GeoTIFF. Sobre estos datos es posible aplicar operaciones como álgebra de mapas, recortes espaciales y análisis combinados entre capas vectoriales y ráster, dentro de flujos de trabajo distribuidos y escalables. Sedona incluye estructuras de datos específicas (Spatial RDDs y Spatial DataFrames), técnicas de indexación espacial como R-trees y Quad trees, y más de 200 funciones espaciales disponibles a través de su extensión Spatial SQL.

Huang et al. (2017), en un estudio sobre procesamiento de datos espaciales en entornos distribuidos con Hadoop y ESRI Spatial Framework, identificaron limitaciones de eficiencia relacionadas con el alto número de operaciones de entrada/salida en disco, las cuales ralentizan el procesamiento. En contraste, GeoSpark SQL (Apache Sedona actualmente), mediante su integración con Hive y Spark, permite ejecutar consultas SQL sobre grandes volúmenes de datos espaciales en entornos distribuidos, mejorando los tiempos de respuesta en consultas que requieren alto cálculo, como uniones espaciales y

consultas KNN (k-vecinos más cercanos).

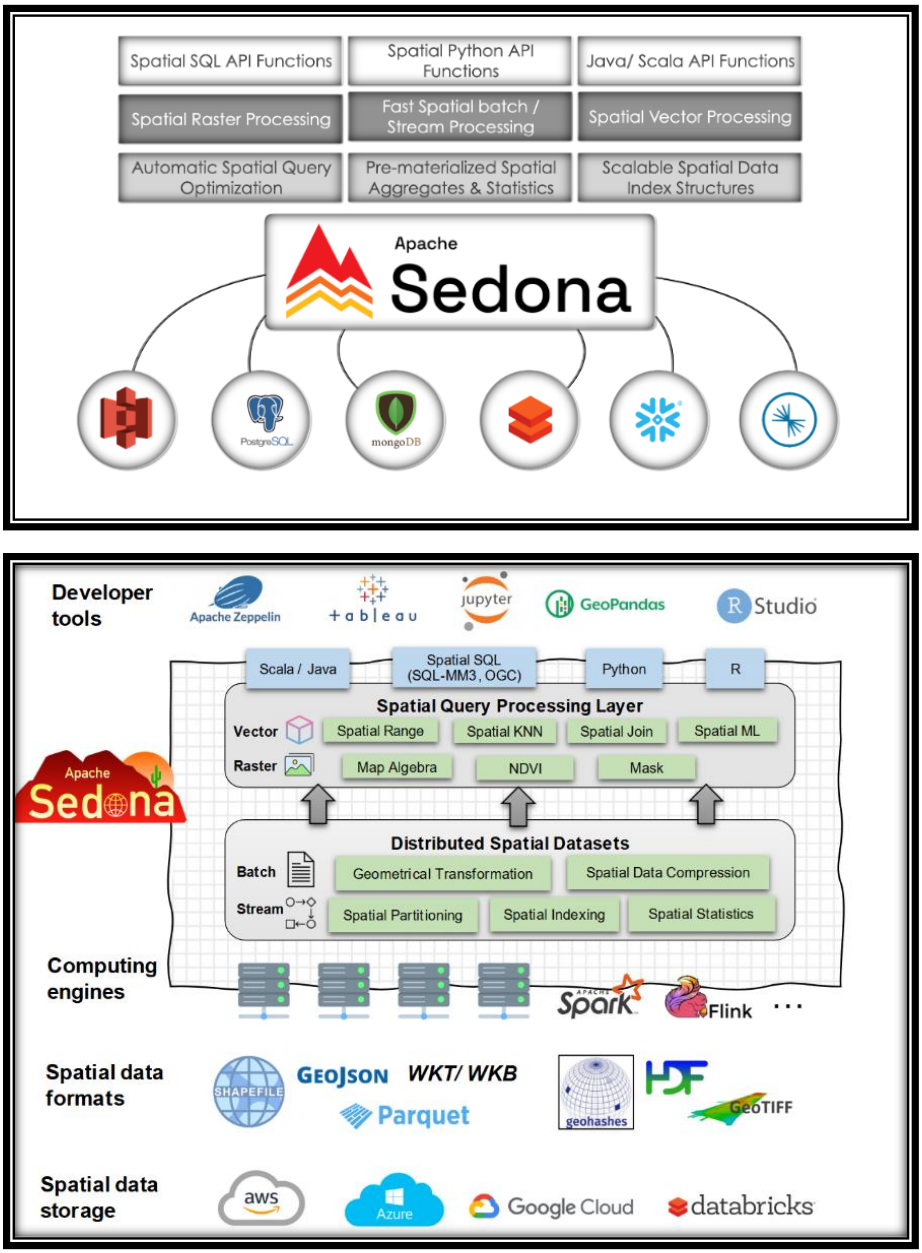


Figura. 1. A)Ecosistema de Apache Sedona. B) Arquitectura del sistema Sedona. Tomado de https://sedona.apache.org/latest/ y Community office hour Sedona.

En la conferencia PyData (2023), Jia Yu señaló que Apache Sedona, anteriormente denominado Apache GeoSpark, gestiona datos distribuidos, indexación y visualización para el análisis de datos geoespaciales. Durante esta conferencia se compararon

GeoPandas y Sedona, concluyendo que ambas herramientas son eficientes con conjuntos de datos pequeños, mientras que Sedona mantiene un mejor rendimiento al escalar a Big Data. De acuerdo con Jia Yu, PyData (2023), estas capacidades permiten trabajar con datos de diversas fuentes y formatos en contextos como análisis urbanístico, logística, monitoreo ambiental o planificación urbana.

Lenka et al. (2016) compararon el rendimiento y la escalabilidad de SpatialHadoop y Sedona, y señalaron que Sedona es adecuado para tareas que requieren alto procesamiento en memoria. Según Huang et al. (2017), Apache Sedona/GeoSpark SQL soporta operadores espaciales como ST_Intersects (determina si dos geometrías se intersectan), ST_Contains (verifica si una geometría contiene a otra) y ST_Within (evalúa si una geometría está completamente dentro de otra), entre otros. También menciona el uso de indexación espacial distribuida como el Rectángulo Mínimo Envolvente (MBR) que permite filtrar datos innecesarios antes de cargarlos en memoria, reduciendo costos de deserialización y cálculo topológico, así como operaciones de entrada/salida, lo que mejora el rendimiento en sistemas distribuidos.

La Figura 2 del libro citado por Tokaj, Yu y Sarwat (2025), O'Reilly Media muestra un modelo de arquitectura en el que se emplea Apache Sedona para integrar datos geoespaciales provenientes de una base de datos PostGIS utilizando un enfoque basado en captura de datos de cambio (CDC). En este proceso, Debezium lee los registros del archivo WAL (Write-Ahead Log) de PostgreSQL, que luego son transmitidos como eventos a través de Apache Kafka. Apache Sedona consume estos eventos desde Kafka, los convierte en un Spatial DataFrame, los geoprocesa, y posteriormente los almacena en formato GeoParquet (Formato de almacenamiento columnar optimizado para consultas

espaciales, con compresión incorporada y soporte de metadatos) sobre un bucket de Amazon S3.

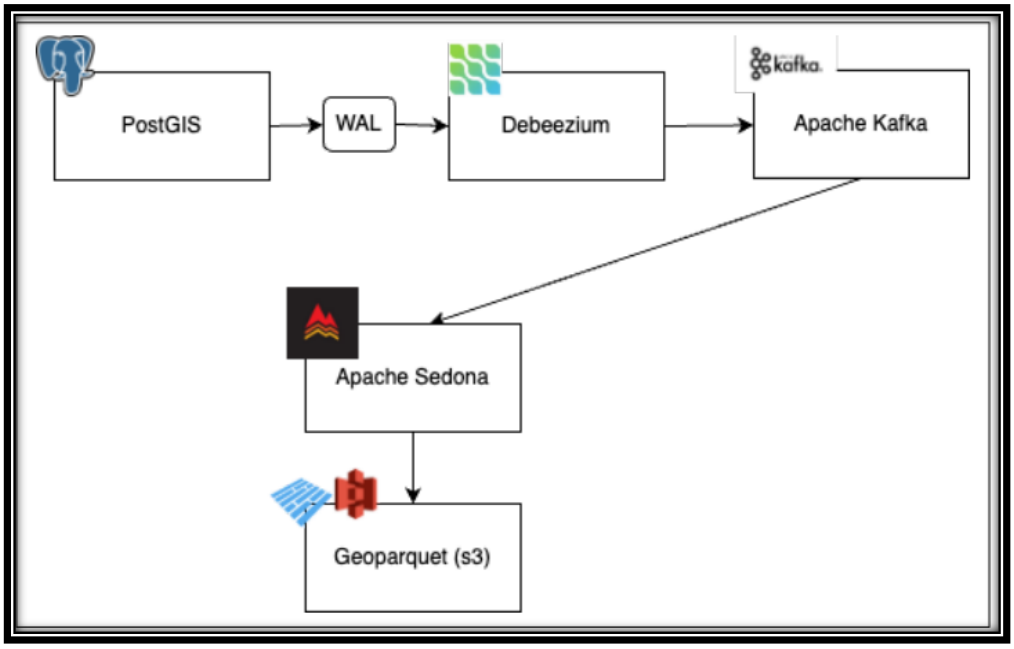


Figura. 2. Diagrama: Lectura desde PostgreSQL mediante CDC (Change Data Capture).
Tomado de Figure 3-7. Diagram showing reading from PostgreSQL CDC, en Cloud Native Geospatial Analytics with Apache Sedona por Tokaj, Yu y Sarwat (2025), O'Reilly Media.

Este flujo de trabajo permite construir un pipeline geoespacial escalable y automatizado, evitando lecturas completas de la base de datos y habilitando la actualización eficiente de los datos a medida que ocurren cambios. Aunque en el ejemplo presentado el enfoque se limita a datos insertados (sin actualizar ni eliminar), el libro sugiere que este modelo puede ampliarse para incluir transformaciones más complejas. (Tokaj, Yu y Sarwat, 2025)

4.4. Wherobots Cloud

Wherobots Cloud es una plataforma de infraestructura de datos geoespaciales diseñada para permitir que las personas y los equipos procesen, analicen y obtengan información de los datos geoespaciales a gran escala. Simplifica el desarrollo de soluciones geoespaciales al encargarse de la arquitectura moderna tipo lakehouse y los servidores, facilitando al usuario el enfoque en el análisis de geoprocesamiento sin preocupación de la infraestructura tecnológica. Wherobots cloud es la plataforma de ETL, análisis e Inteligencia Artificial de datos espaciales creada por los desarrolladores originales de Apache Sedona. (Wherobots Inc, 2025)

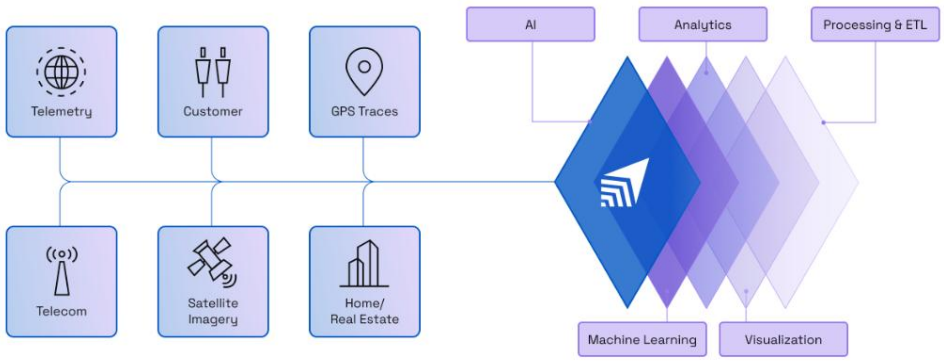


Figura. 3. Wherobots: The Spatial Intelligence Cloud
Tomado de https://docs.wherobots.com/latest/

Entre los servicios que ofrece Wherobots Cloud se incluyen:

Servicios Wherobots Cloud
Compatibilidad con Apache Sedona.
Análisis espacial, rendimiento y escalabilidad.
Análisis unificado de datos vectoriales y raster.
Métodos ETL espaciales nativos de la nube, análisis e inteligencia artificial.
Una arquitectura segura y sin servidor.

Almacenamiento propio en la nube, Amazon S3.
Soporte para compilación con Spatial SQL, Python y Scala.
Herramientas para la productividad de la ingeniería.

Tabla. 1. Servicios Wherobots Cloud.
Fuente: Adaptado de Apache Sedona Office Hour por Wherobots 2025.

Wherobots se basa en Apache Sedona como núcleo de código abierto para implementar una plataforma de datos en la nube orientada al análisis geoespacial y a cargas de trabajo de inteligencia artificial. La plataforma está optimizada para la ejecución de procesos geoespaciales, incorporando tecnología de Apache Sedona. Además de contribuir al desarrollo y mejora de este proyecto, Wherobots implementa optimizaciones propias y funcionalidades adicionales, entre las que se incluye WherobotsDB, un sistema completo de base de datos para análisis espacial. (Wherobots Inc, 2025)

Wherobots Cloud se seleccionó como entorno de prueba debido a que abstrae la complejidad de la configuración de la infraestructura y el software de Spark y Amazon S3, lo que permite centrar los esfuerzos en la validación de las funciones de Sedona. Esta plataforma ofrece la posibilidad de iniciar, con solo unos clics, una instancia de JupyterLab preconfigurada, creando un entorno informático espacial para el desarrollo de análisis geoespaciales y flujos de trabajo de inteligencia artificial (Wherobots Inc., 2025).

Debido a las conclusiones del estudio de Shin et al. (2022), en el que se evaluó el rendimiento de distintos motores de Big Geospatial Data en entornos de computación en la nube utilizando Apache Sedona, y en el que se determinó que HDFS y Amazon S3 presentan mayor escalabilidad y mejor rendimiento que MongoDB, se respaldó la elección de Wherobots Cloud, dado que su arquitectura integra tanto Amazon S3 como Apache Sedona.

No obstante, se debe tener en cuenta que el uso de esta herramienta presenta limitaciones asociadas tanto a la dependencia del proveedor para la gestión de la infraestructura como a los desafíos en la administración de costos en entornos con cargas de trabajo intensivas. Wherobots Cloud incorpora la función Idle Timeout para reducir gastos al pausar el uso de recursos cuando no hay actividad; sin embargo, esta medida no elimina completamente los cargos, ya que se continúa facturando por los recursos mínimos asignados durante el periodo de inactividad. Además, dicha función solo es aplicable a las instancias de cuadernos (notebook instances) y no a las ejecuciones de trabajos (Wherobots Inc., 2025).

Aunque Wherobots reduce automáticamente el nivel de cómputo al mínimo necesario cuando los cuadernos están inactivos, las Unidades Espaciales (Spatial Units) continúan consumiéndose. Por lo tanto, para detener completamente la facturación y liberar todos los recursos asociados, es necesario eliminar manualmente las instancias activas cuando no estén en uso.

En la versión Community Edition, las instancias de cuadernos se apagan automáticamente después de cuatro horas, aunque es posible reiniciarlas si se requiere más tiempo. Esta edición solo permite el uso del runtime de tipo Tiny, mientras que otros entornos de ejecución están disponibles únicamente en versiones superiores. Para los notebooks que requieren más recursos —como el procesamiento raster o el manejo de grandes conjuntos de datos— será necesario activar la suscripción Pro (Wherobots Inc., 2025).

4.5. Catastro multipropósito: Proceso actualización Catastral.

4.5.1. Definición Actualización Catastral.

El Instituto Geográfico Agustín Codazzi, define la actualización catastral cómo un proceso del *Catastro con enfoque Multipropósito* el cuál abarca un conjunto de actividades destinadas a identificar, incorporar y/o rectificar los cambios o inconsistencias en la información catastral durante un período determinado. Para ello, se emplean diversos métodos de intervención en el territorio, como los métodos directos, indirectos, declarativos y colaborativos, además de la integración de múltiples fuentes de información que reflejan los cambios entre la base catastral y la realidad de los inmuebles. La actualización catastral es un proceso técnico-administrativo que busca mantener actualizada la información física, jurídica y económica de los predios. (IGAC, 2022). Este proceso requiere el manejo de grandes volúmenes de datos espaciales y alfanuméricos, provenientes de múltiples fuentes.

4.5.2. Integración Big Data y Actualización Catastral.

Según el artículo 2.2.2.2.6 del Decreto 148 de 2020, el análisis de Big Data se reconoce como un método indirecto para la identificación de información física, jurídica y económica de los bienes inmuebles (DANE, 2020). Este método se integra con lo establecido en el artículo 4.2.4 de la Resolución 1040 de 2023, donde se definen las entradas mínimas requeridas para el proceso de actualización catastral (IGAC, 2023). Dicho artículo contempla, entre otros aspectos, el marco legal y regulatorio vigente, la disponibilidad de personal técnico y administrativo capacitado, y los recursos técnicos, tecnológicos, financieros y logísticos necesarios, todos ellos vinculados al manejo de grandes volúmenes de información espacial y alfanumérica "*4. Datos básicos disponibles (geodesia, insumos cartográficos y cartografía básica y predial) e 5. Información predial*

y relacionada disponible de fuente propia, privada o proveniente o derivada de otros registros administrativos públicos". (IGAC, 2023)

En este marco, el método indirecto comprende además el empleo de modelos estadísticos y econométricos, el uso de imágenes de sensores remotos y la incorporación de fuentes secundarias, como los observatorios inmobiliarios, lo que refleja la aplicación de técnicas de Big Data en la consolidación de datos provenientes de diversas fuentes para su incorporación en la base catastral (DANE, 2020). Un ejemplo de estas fuentes de información espacial contempladas en el Modelo LADM_COL son los croquis de campo, datos crudos, ortofotografías, informes técnicos y registros fotográficos. Estos insumos, que generan grandes volúmenes de datos heterogéneos, son fundamentales para la calidad y actualización de la información catastral, y constituyen un escenario adecuado para la adopción y aplicación de herramientas de Big Data en su gestión y análisis.

En cuanto a la infraestructura tecnológica, el artículo 4.2.1 de la Resolución 1040 de 2023 establece que los procesos catastrales deben contar con herramientas modernas y eficientes para recolectar, procesar, almacenar y gestionar la información, garantizando su interoperabilidad y accesibilidad. Estos lineamientos coinciden con la necesidad de plataformas tecnológicas capaces de soportar arquitecturas de Big Data para la administración y análisis de información. Asimismo, la Resolución 267 de 2023 dispone que, para la ejecución de los proyectos catastrales, se debe contar con una propuesta técnico-económica que defina los recursos y presupuesto necesarios, asegurando que la gestión de la información se realice bajo parámetros técnicos y normativos vigentes.

4.5.3. Servicios que ofrecen Gestor y/o Operador Catastral

De acuerdo con el artículo 4.2.6. de la resolución 1040 de 2023, los servicios que deben ofrecer los gestores y operadores catastrales para cumplir con los productos mínimos para realizar el proceso de actualización catastral comprenden:

Servicios que ofrecen Gestor y/o Operador Catastral: ***Proceso de Actualización Catastral***
Levantamiento de información disponible.
Reconocimiento predial.
Observatorio inmobiliario o bases de datos de ofertas y transacciones del mercado inmobiliario de la región.
Consolidación de información alfanumérica y geográfica actualizada.
Avalúos puntuales y masivos.
Reporte de información al Sistema Nacional de Información Catastral (SINIC).

Tabla. 2. Servicios que ofrecen Gestor y/o Operador Catastral para la actualización catastral. Fuente: Resolución 1040 de 2023

Estos servicios están articulados dentro de un proceso de negocio compuesto que se organiza en tres etapas secuenciales: pre-operativa, operativa y post-operativa. (IGAC, 2023) En cuanto a los plazos legales, los artículos 4.2.3 y 4.2.12 de la resolución 1040 de 2023 establecen que la actualización catastral es obligatoria y debe ejecutarse de forma periódica y sistemática, sin exceder el límite señalado en el artículo 24 de la Ley 1450 de 2011 o su norma modificatoria, en sintonía con la dinámica del mercado inmobiliario y las directrices del IGAC.

En este contexto, la identificación de inconsistencias representa una fase esencial del proceso. Al recolectar la información, se procede a la detección de inconsistencias tanto gráficas como alfanuméricas y registrales, mediante la comparación y cruce de diferentes

bases de datos, con el objetivo final de consolidar y validar los productos de la gestión catastral, garantizando la precisión y fiabilidad de la información (IGAC, 2023). Asimismo, el uso de herramientas de Big Geospatial Data en la gestión eficaz de grandes volúmenes de información, en etapas como el almacenamiento, visualización, análisis y procesamiento, puede, en este caso, para los procesos de gestión catastral, impulsar de manera significativa la construcción verde, la energía verde y los objetivos de carbono neutral en el país (Mete, 2023).

4.5.4. Modelo LADM_COL

Mediante la Resolución Conjunta SNR 05346 – IGAC 662 del 22 de mayo de 2024, se adoptó el Modelo Extendido de Catastro–Registro LADM_COL, derivado del modelo núcleo que establece la base ontológica y semántica de la administración de tierras en Colombia y define un lenguaje común entre las entidades responsables de catastro y registro.

El Modelo Extendido constituye la estructura fundamental para el almacenamiento y gestión de la información predial, organizada en tres paquetes principales: el administrativo, el de interesados y el espacial, este último con un subpaquete de representación geométrica. Cada paquete integra los datos requeridos para caracterizar e identificar los predios conforme a su realidad actual, permitiendo mantener actualizada la información predial y facilitar su uso como insumo para la formulación de políticas públicas y la toma de decisiones.

Este modelo se establece como estándar para la integración e interoperabilidad de

la información catastral y registral con enfoque multipropósito, según lo dispuesto en el repositorio de modelos de la Infraestructura Colombiana de Datos Espaciales (ICDE). A partir de este se derivan modelos de aplicación orientados a la integración, transferencia y reporte de datos entre instituciones. En este marco, la Resolución 301 de 2025 adoptó el Modelo de Aplicación LADM_COL SINIC V1.0, que regula el reporte de información por parte de los gestores catastrales, definiendo las variables mínimas necesarias geográficas y alfanuméricas para consolidar la información en el Sistema Nacional de Información Catastral (SINIC), asegurando su actualización y disponibilidad.

4.6.Método ADM – TOGAF

El método ADM (Architecture Development Method), traducido al español como Método de Desarrollo de Arquitecturas, hace parte del marco de trabajo estandarizado TOGAF (The Open Group Architecture Framework). Este método se centra en los principios de la arquitectura empresarial y proporciona una guía estructurada para el desarrollo, mantenimiento y gestión de arquitecturas organizacionales (Cabrera et al., 2015).

Según The Open Group (2009), ADM se organiza como un ciclo iterativo compuesto por cuatro fases fundamentales: iteración de capacidad de la arquitectura, iteración de desarrollo, iteración de planificación de transición e iteración de gobernanza arquitectónica, cómo se observa en la figura 4. Cada una de estas fases agrupa distintas capas del modelo arquitectónico, lo que facilita una visión integral desde la perspectiva del negocio, las aplicaciones, los datos y la infraestructura tecnológica. Su estructura flexible

permite que el método se adapte a las necesidades particulares de cada organización, promoviendo así la alineación estratégica entre los objetivos del negocio y las soluciones tecnológicas implementadas (Cabrera et al., 2015; The Open Group, 2009).

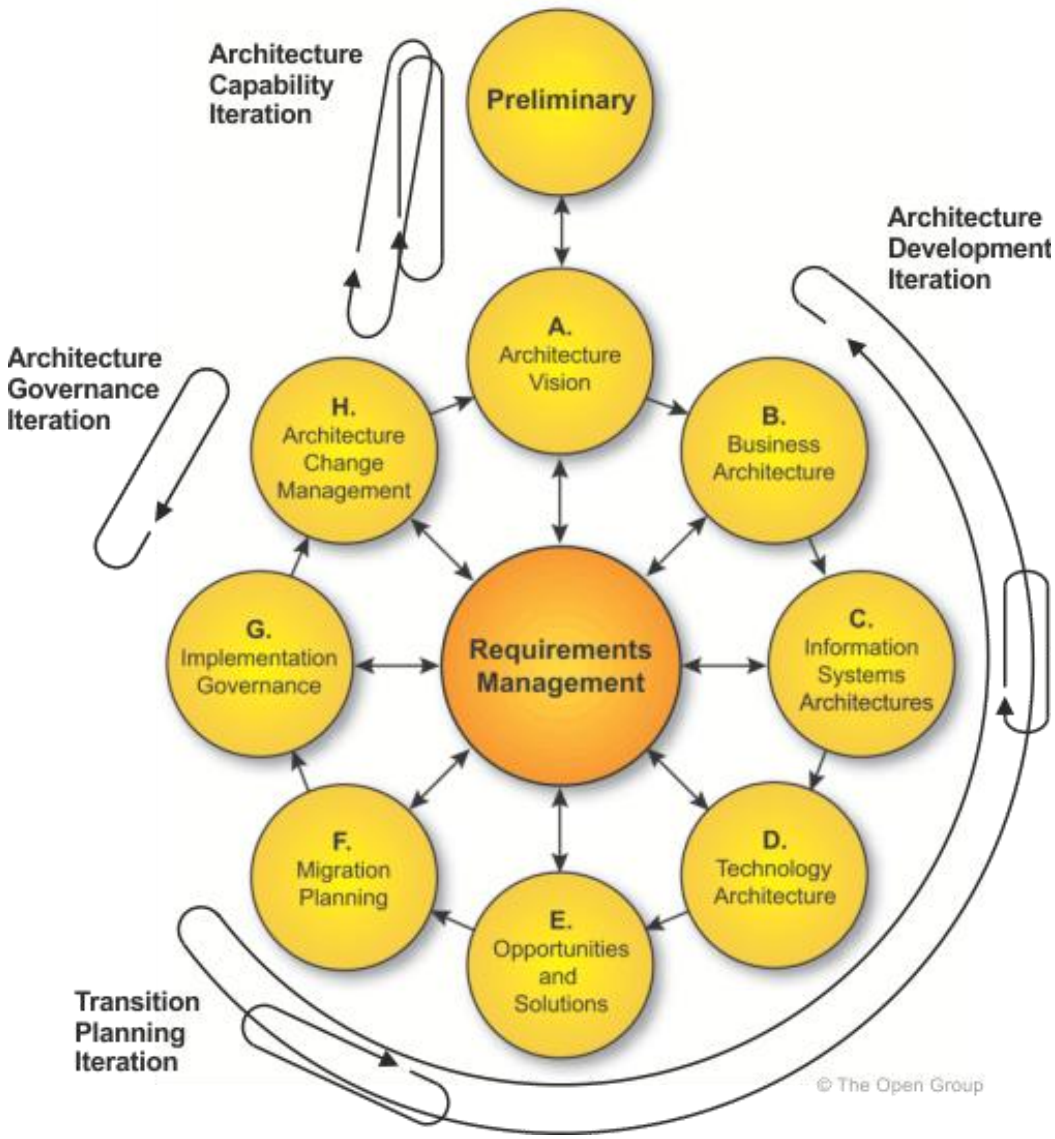


Figura. 4. Estructura ADM.
Tomado de: The Open Group Framework

En este proyecto de investigación se desarrollarán la fase de iteración de capacidad de la arquitectura y la fase de iteración de desarrollo del método ADM para la construcción

del modelo. Según Saboya Ríos et al. (2018) y Ares (2013), la primera fase consiste en identificar la estrategia de la organización, así como lograr el compromiso y la participación de los actores clave del modelo, quienes participan, deciden o se ven afectados por las decisiones de arquitectura en las distintas fases del ADM, esta fase de iteración de capacidad está compuesta por la capa Preliminar y la capa A. Visión de la Arquitectura.

De igual manera, la fase de iteración de desarrollo tiene como propósito identificar y comprender la situación actual que el negocio busca resolver, así como diseñar o proponer el escenario futuro en relación con los dominios de la arquitectura de negocio, los sistemas de información (que abarcan las herramientas y los datos) y la tecnología. Esta segunda fase comprende las capas B. Arquitectura de Negocio, C. Arquitectura de Sistemas de Información y D. Arquitectura Tecnológica.

FASE	*CAPA*	*CONCEPTO*
1. *Iteración de capacidad*	Preliminar	Establece el marco para la arquitectura, definiendo los principios, la gobernanza y el alcance.
	A. Visión de la Arquitectura	Desarrolla un resumen del estado futuro deseado, identificando los impulsores, objetivos y partes interesadas para asegurar que el proyecto sea viable y esté alineado con las metas del negocio.
2. *Iteración de desarrollo de la arquitectura*	B. Arquitectura de Negocio	Es el componente que plasma la estrategia del negocio, los mecanismos de gobernanza y los procesos principales de la entidad.
	C. Arquitectura de Sistema de Información	Detalla los servicios correspondientes a cada sistema de aplicación que se proyecta implementar, así como las interacciones entre ellos y su vinculación con los datos vinculados a los procesos de negocio
	D. Arquitectura de	Detalla la infraestructura tecnológica,

Tecnología	incluyendo el hardware, software y las redes requeridas para respaldar la implementación de arquitectura de sistemas de información.

Tabla. 3. Definición capas Preliminar, A, B, C y D del método ADM TOGAF.
Fuente: Elaboración Propia

5. MÉTODOS Y TÉCNICAS DE ANÁLISIS DE DATOS

La investigación se desarrolló en varias etapas para elaborar un modelo basado en el método ADM, representado en el lenguaje de modelado ArchiMate. Este modelo, según Saboya Ríos et al. (2018), permite la coordinación entre procesos, datos, aplicaciones e infraestructura tecnológica. La concreción del modelo se logra a través de un conjunto de vistas que reflejan los intereses y perspectivas de los actores involucrados.

En cada etapa se identificaron los conceptos relacionados con la integración de Big Geospatial Data en el proceso de actualización catastral, usando como herramienta tecnológica Apache Sedona. Estos conceptos fueron incorporados al modelo según el estándar del método ADM.

Primero, se describe la construcción del modelo conceptual mediante vistas, siguiendo el método ADM del marco TOGAF (sección 5.1). Luego, se expone la metodología empleada para integrar los dominios de Big Geospatial Data al proceso de actualización catastral (sección 5.2), seguidamente se explica cómo se obtuvo el dominio tecnológico asociado a la herramienta Apache Sedona, utilizada como caso de estudio en

la investigación (sección 5.3). Finalmente, se presenta la metodología de validación técnica del experimento, orientada a comprobar la funcionalidad, los retos y la coherencia del entorno tecnológico propuesto (sección 5.4). Las secciones 5.2, 5.3 y 5.4 corresponden a los subprocesos analizados y detallados para la construcción del modelo descrito en la sección 5.1.

5.1. Metodología aplicada en la construcción del modelo.

Para definir el modelo de integración, se optó por utilizar un enfoque de arquitectura, dado que este permite modelar tanto la parte dinámica como estática del sistema. En este caso, se empleó el método ADM (Architecture Development Method), el cual hace parte del marco de trabajo estandarizado TOGAF, propuesto por The Open Group Architecture Framework.

En este proyecto, el modelo contemplará la parte del ciclo correspondiente a las fases 1 y 2: iteración de capacidad e iteración de desarrollo de la arquitectura, que contienen las capas 'Preliminar',' A', 'B', 'C' y 'D', de acuerdo con la *estructura ADM* plasmada en la figura 4. de la sección anterior. Se tuvieron en cuenta los requisitos mínimos establecidos en la normativa vigente para la implementación del proceso de actualización catastral en el desarrollo de las fases 1 y 2.

5.1.1. Desarrollo vista modelo Fase 1. Iteración de capacidad: Capa Preliminar y Capa A. Visión de la Arquitectura (Proceso de Actualización Catastral)

En la fase 1 (iteración de capacidad), que contempla la identificación de la estrategia organizacional y la obtención del compromiso y la participación de los actores clave

(stakeholders) entendidos conceptualmente como los interesados en el modelo, como los gestores y operadores catastrales y el IGAC en su rol de máxima autoridad catastral. Esta fase no implicó contacto directo con los actores, sino que su consideración se realizó de acuerdo con lo establecido en el método ADM y con lo dispuesto en la norma que regula la gestión catastral y el catastro multipropósito, para definir los objetivos y necesidades preliminares del modelo.

La primera vista abarca tanto la etapa preliminar como la capa A. visión de la arquitectura, ambas correspondientes a la fase 1, en la que se representa la motivación para definir la visión general y el alcance del modelo propuesto. Durante esta fase se establece la capa de motivación, en la cual se identifican a los stakeholders o actores clave, sus motivaciones, los objetivos y los requisitos que influyen en el diseño de la arquitectura. En esta fase se reconocen los intereses de los stakeholders involucrados en el proceso de actualización catastral y las condiciones que definen el alcance de la iniciativa. La representación de estos elementos permite establecer la relación entre las necesidades y/o problemáticas que influyen en el diseño de la arquitectura a desarrollar.

De acuerdo con la normativa vigente se tuvo en cuenta para definir las necesidades, objetivos, requisitos, identificar los stakeholders, entre otros, cómo el artículo 81 de la Ley 1955 de 2019 que detalla responsabilidades y sanciones en relación con:

- Suministro de información y consolidación de datos.
- Validación y estandarización de datos catastrales.
- Optimización del proceso de actualización catastral.
- Automatización y mejora en la aplicación de metodologías y estándares.

También para identificar el resultado que se debe obtener en el proceso de actualización catastral y plasmarlo en el modelo se vinculó los productos mínimos que debe entregar el proceso de actualización catastral conforme al artículo 4.2.6 de la Resolución 1040 de 2023. Respecto a la infraestructura tecnológica, el artículo 4.2.1 de la Resolución 1040 de 2023 establece que los procesos catastrales deben contar con herramientas modernas y eficientes para recolectar, procesar, almacenar y gestionar la información, garantizando su interoperabilidad y accesibilidad. Asimismo, según lo establecido en la Resolución 267 de 2023, para la ejecución del proyecto se debe contar con una propuesta técnico-económica que defina los recursos y presupuesto necesarios con el fin de asegurar que el proceso se realice cumpliendo las especificaciones técnicas y normativas vigentes.

En cuanto a los plazos legales, los artículos 4.2.12 y 4.2.3. de la resolución 1040 de 2023 establecen que la actualización catastral es obligatoria y debe ejecutarse de forma periódica y sistemática, sin exceder el límite señalado en el artículo 24 de la Ley 1450 de 2011 o su norma modificatoria, en sintonía con la dinámica del mercado inmobiliario y las directrices del IGAC.

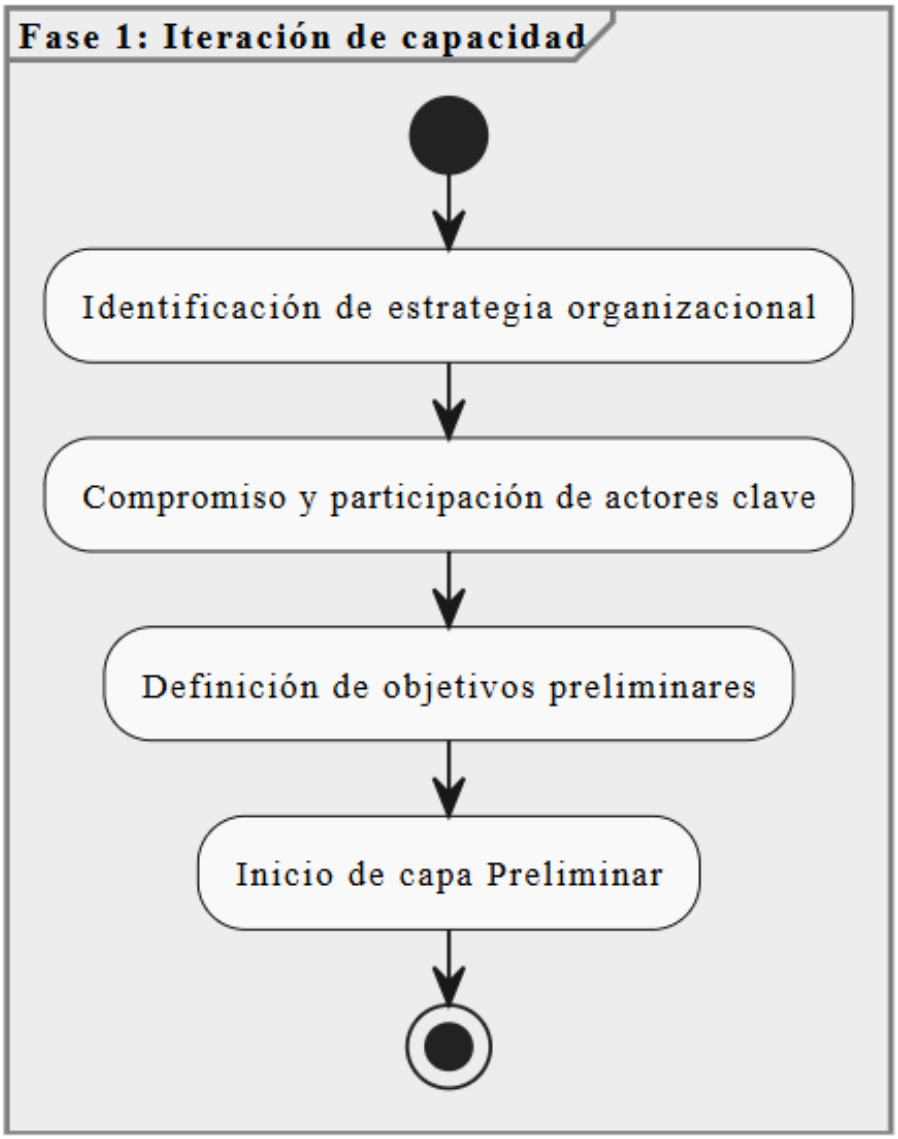


Figura. 5. Desarrollo del modelo empleando método ADM Fase 1.
Elaboración propia.

5.1.2. Desarrollo vista modelo Fase 2. Iteración de desarrollo: Capa B Arquitectura de Negocio. (Proceso de Actualización Catastral)

A su vez, la fase 2 (iteración de desarrollo) tiene como propósito comprender e identificar la situación actual y diseñar la situación futura en los dominios de arquitectura de negocio, sistemas de información y tecnología (Ares, 2013). Lo cual promueve la

investigación del modelo propuesto.

Los servicios que ofrece el gestor y/o operador para cumplir con los productos mínimos de acuerdo con el artículo 4.2.6. de la resolución 1040 de 2023, comprenden:

- Levantamiento de información disponible.
- Reconocimiento predial.
- Observatorio inmobiliario o bases de datos de ofertas y transacciones del mercado inmobiliario de la región.
- Consolidación de información alfanumérica y geográfica actualizada.
- Planos elaborados sobre la base de cartografía actualizada.
- Avalúos puntuales y masivos.
- Reporte de información al Sistema Nacional de Información Catastral (SINIC).

Estos servicios están articulados dentro de un proceso de negocio compuesto que se organiza en tres etapas secuenciales: pre-operativa, operativa y post-operativa. (IGAC, 2023)

Durante la etapa pre-operativa, el stakeholder realiza la identificación de la zona a intervenir utilizando insumos técnicos como ortofotos, levantamientos topográficos, cartografía base, el instrumento de ordenamiento territorial vigente (POT - PBOT - EOT), entre otros, basándose en información y recursos secundarios recopilados previamente que permiten establecer la representación geográfica de los predios del territorio desde la oficina. La actividad se desarrolla conforme al artículo 6 de la Resolución 746 de 2024, que señala los requisitos mínimos de información, recursos y capacidades necesarios para el proceso de actualización catastral. En paralelo, se lleva a cabo la identificación de inconsistencias (alfanuméricas, jurídicas y geográficas) en la información catastral

existente, mediante el análisis de datos registrales —como los derechos jurídicos del predio, el tamaño y la forma de las propiedades—y de bases de datos geográficas y alfanuméricas, entre otras. (IGAC, 2023)

La fase operativa consiste en identificar los aspectos físicos, jurídicos y económicos, de acuerdo con el artículo 32 de la resolución 746 de 2024 y el articulo 2.2.2.2.6. del Decreto 148 de 2020 sobre los métodos de recolección de información catastral, que establece un enfoque flexible para la captura y recolección de información catastral. Para ello, se emplean métodos que incluyen técnicas directas (levantamiento en campo), indirectas (uso de datos existentes o de fuentes externas), colaborativas (aportados por terceros) y declarativas (entregados por los propietarios o poseedores). En esta etapa se validan los datos recogidos, se definen zonas homogéneas físicas y geoeconómicas para efectos valuatorios, se integran los datos en bases de datos estructuradas y se entrega la información. (IGAC, 2023)

Finalmente, la fase post-operativa implica la consolidación de los productos catastrales mediante la liquidación de avalúos, la asignación del Número Predial Nacional (NUPRE) y la generación de documentos técnicos y administrativos asociados a la gestión realizada en el territorio intervenido. Como cierre del proceso, se deben entregar los productos a los municipios y reportar la información al SINIC, cumpliendo con los estándares definidos por las entidades rectoras del sistema catastral. Además, se especifica que este cierre se materializa mediante la entrega de los Registros Catastrales R1 y R2, junto con un archivo XTF que contiene el resultado de la actualización. (IGAC, 2023)

5.1.3. Desarrollo vista modelo Fase 2. Iteración de desarrollo: Capa C Arquitectura

de Sistemas de información. (Rol Big Geospatial Data)

Para el desarrollo de la capa C. Arquitectura de Sistema de Información. se integra el rol de Big Gesopatial Data al proceso de actualización catastral, para lograr una integración efectiva y asegurar la interoperabilidad de grandes volúmenes de información proveniente de distintas fuentes, es esencial que los datos sean precisos, coherentes y transformables. Gracias a los avances tecnológicos permiten que la generación y procesamiento de datos georreferenciados se lleven a cabo con mayor rapidez y precisión.

Como ejemplo práctico y caso de estudio, en el nivel de aplicación de la arquitectura del modelo, se utiliza la herramienta, Apache Sedona, esta herramienta es fundamental para procesos de automatización, optimización y análisis de los datos capturados durante los procesos de actualización catastral.

Sedona anteriormente conocido como GeoSpark, se ha consolidado como una herramienta esencial en el análisis de datos geoespaciales, transformando significativamente la manera en que se procesan conjuntos de datos espaciales en entornos distribuidos (Yu & Sarwat, 2021). Entre las características clave de Apache Sedona se destacan su capacidad de procesamiento en memoria, su compatibilidad con varios lenguajes de programación como Python, Java y Scala (Apache Sedona, s.f.), y, en particular, su escalabilidad. Esta última se refiere a la capacidad de construir sistemas de procesamiento de datos escalables y eficientes, permitiendo manejar grandes volúmenes de datos de manera óptima mediante la adición de más nodos de procesamiento. (Aach et al., 2023).

En la arquitectura de aplicación se define un flujo de procesamiento de Big

Geospatial Data implementado como un pipeline que responde a las tareas del proceso de actualización catastral: identificar, incorporar y rectificar datos catastrales. Un pipeline, o conductor de datos, es un sistema que admite tanto consultas por lotes como interactivas, permitiendo canalizar y cargar los resultados en diferentes destinos para el análisis o la actualización de bases de datos. Este flujo se ejecuta sobre un entorno de procesamiento distribuido, soportado por SparkContext, el punto de entrada para gestionar Big Geospatial Data con Apache Sedona, el cual se integra con la capa de infraestructura tecnológica. En este ambiente, Sedona facilita operaciones sobre grandes volúmenes de datos espaciales mediante funciones de lectura, transformación, particionamiento e indexación geoespacial, entre otras, lo que permite estructurar procesos escalables que cumplen con los requerimientos técnicos para optimizar los procesos de actualización catastral.

Las funcionalidades de Apache Sedona son:

- Procesamiento de Consultas Espaciales:
    - Soporte para Spatial SQL, Python, Scala/Java y R, lo que permite realizar consultas espaciales avanzadas.
    - Algoritmos distribuidos para consultas espaciales, optimizando el análisis de datos en clústeres.
    - Un optimizador de consultas espaciales que mejora la eficiencia en el procesamiento de datos.
- Conjuntos de Datos Espaciales Distribuidos:
    - Particionamiento espacial para organizar datos geoespaciales en clústeres.
    - Indexación espacial para acelerar las búsquedas y consultas.
    - Compresión de datos espaciales, lo que reduce el uso de memoria y mejora el rendimiento.

En este proyecto, en la parte de datos de la arquitectura de sistemas de información, se trabaja en una tarea concreta del proceso de actualización catastral en vez de cubrir todas las operaciones técnicas involucradas y presentadas en la capa de negocio. Se decide concentrar el estudio en el uso del pipeline para consolidar una matriz que vincule la normativa del plan de ordenamiento territorial con la información de cada predio de Bogotá D.C., esto debido a la entrada en vigencia del Plan de Ordenamiento Territorial de Bogotá, según lo establecido en el Decreto 555 de 2021, que modifica la normativa del Decreto 190 de 2004, se hace necesario realizar un análisis integral de todos los predios de la ciudad para actualizar la base catastral.

El enfoque propuesto se centra en evaluar el funcionamiento del pipeline a partir del análisis de una tarea puntual, lo que permitirá explorar la posibilidad de ampliar el modelo de arquitectura a otros aspectos del proceso. La aplicación de Apache Sedona busca facilitar la consulta masiva y consolidación de datos geográficos para su análisis, optimizando así las tareas en tiempo de ejecución. Gracias al uso de herramientas de big data, no será necesario limitar el análisis a un conjunto reducido de datos, ya que se podrán procesar todos los datos objeto de estudio mediante geoprocesamiento distribuido por lotes a cualquier escala. Esto permitirá optimizar procesos como la delimitación de zonas homogéneas físicas, geoeconómicas y la realización de avalúos puntuales.

Entre los datos principales considerados para el estudio se encuentran:

- Datos internos: Predios, definidos como la entidad central, integran componentes de información jurídica, económica y física a través de sus respectivos atributos y relaciones.

- Datos externos: Plan de Ordenamiento Territorial (POT) vigente, que agrupan entidades y atributos relacionados con la clasificación del suelo, tratamientos urbanísticos, áreas de actividad y sectores.

Para la implementación del pipeline, se tiene en cuenta las funcionalidades de Sedona en cuanto al procesamiento de consultas espaciales incluye una serie de operaciones que dependen de las relaciones entre objetos geoespaciales, fundamentales para aplicar en la actualización catastral. En este caso de estudio solo se utilizará formatos geospaciales como shapefile y geojson, tipo de dato vector, Sedona soporta diversos operadores espaciales, entre ellos: ST_Intersects, que determina si dos geometrías se intersectan; ST_Contains, que verifica si una geometría contiene a otra; y ST_Within, que evalúa si una geometría está completamente dentro de otra (Huang et al., 2017). Además, se hace uso de índices espaciales como el Rectángulo Mínimo Envolvente (MBR) para filtrar datos innecesarios antes de cargarlos en memoria.

5.1.4. Desarrollo vista modelo Fase 2. Iteración de desarrollo: Capa D Arquitectura Tecnológica. (Orquestación Big Geospatial Data)

Considerando la importancia del Big Geospatial Data en la actualización catastral y el uso de Apache Sedona en la capa de sistemas de información, los factores clave para gestionar de manera eficiente estos grandes volúmenes de datos en el pipeline son la capacidad de almacenamiento y cómputo. Sedona se respalda en avances en modelos y algoritmos, tales como el aprendizaje automático, el procesamiento de grafos y el procesamiento por lotes (Aach et al., 2023).

Para la creación del clúster se emplearon dos computadoras; un portátil y uno de escritorio, ambas con el sistema operativo Linux Ubuntu. En cada dispositivo se definieron contenedores Docker que permiten desplegar, de manera aislada y escalable, los componentes de Apache Spark y Sedona. Sedona, por defecto, cuenta con una imagen Docker (Sedona Docker Image), disponible en Docker Hub (https://hub.docker.com/r/apache/sedona), que se descarga con el comando docker pull apache/sedona:1.7.1 y se ejecuta mediante docker run -p 8888:8888 -p 8080:8080 -p 8081:8081 -p 4040:4040 apache/sedona:1.7.1. Este esquema facilita la administración de procesos distribuidos, ya que cada contenedor se ejecuta con los recursos y dependencias apropiados para su función, evitando conflictos entre librerías o configuraciones del sistema operativo. (Apache Sedona Project, 2025)

Como complemento, se utilizó el lenguaje ArchiMate, una notación visual derivada de UML dónde se puede representar este tipo de Método de Desarrollo de Arquitecturas. ArchiMate opera como un sistema de símbolos y conectores que estructura y comunica los elementos de la arquitectura, permitiendo su análisis e implementación. Según Hernández León y Lancheros Aguilar (2021), ArchiMate incluye un catálogo que permite identificar los requerimientos desde la perspectiva del negocio y de la implementación, y representar cualquier diagrama requerido por el marco de arquitectura empresarial TOGAF.

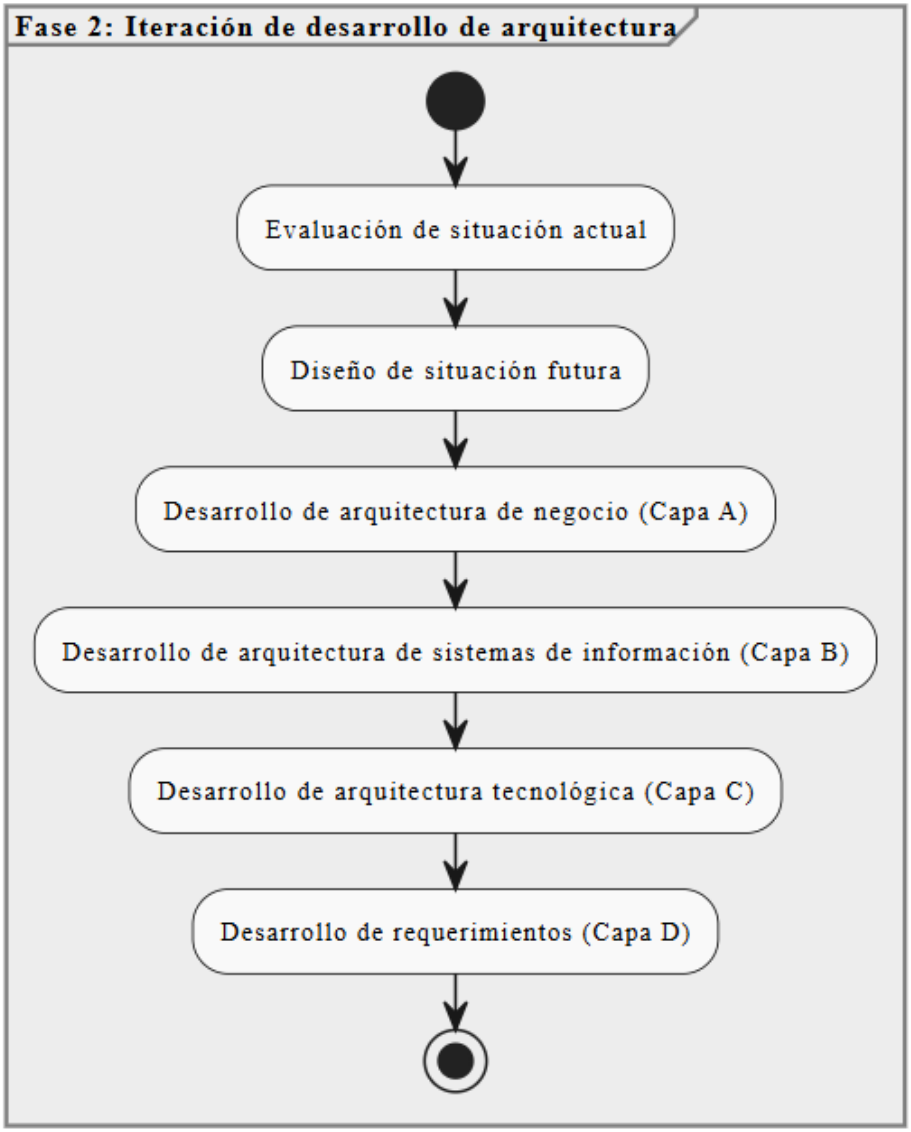


Figura. 6. Desarrollo del modelo empleando método ADM Fase 2.
Elaboración propia.

5.2. Metodología para la integración de Big Geospatial Data en la actualización catastral.

Para abordar el estudio de la integración de Big Geospatial Data con el proceso de actualización catastral, se realizó la caracterización de ambos campos de conocimiento con

el fin de comprender sus conceptos fundamentales y modelar cada fase (1: *iteración de capacidad. 2: iteración de desarrollo*) del modelo objeto de estudio. Se empleó el método estructurado que combinó el análisis teórico-conceptual con la implementación de flujos de trabajo prácticos. Las etapas desarrolladas fueron:

- Revisión de problemáticas actuales: Se analizó el proceso funcional de la actualización catastral con enfoque multipropósito y se identificaron aspectos técnicos, operativos y normativos relacionados. El objetivo fue entender cómo funciona el modelo de negocio según la normativa, la cual establece los requerimientos mínimos para ejecutar este proceso por parte de los actores involucrados. A partir de esto, se identificaron problemáticas relacionadas con el proceso de obtención de información actualizada, especialmente en lo que respecta al uso de análisis de geoprocesamiento de datos geoespaciales masivos y los métodos de extracción y almacenamiento de esta información. Todo esto con el fin de avanzar hacia un objetivo específico en el modelo de integración de Big Data con la actualización catastral, explorando cómo el paradigma de Big Data puede integrarse con el catastro mediante el uso de tecnologías como Apache Sedona y dar una posible solución a las problemáticas identificadas. (*véase Figura 5*)

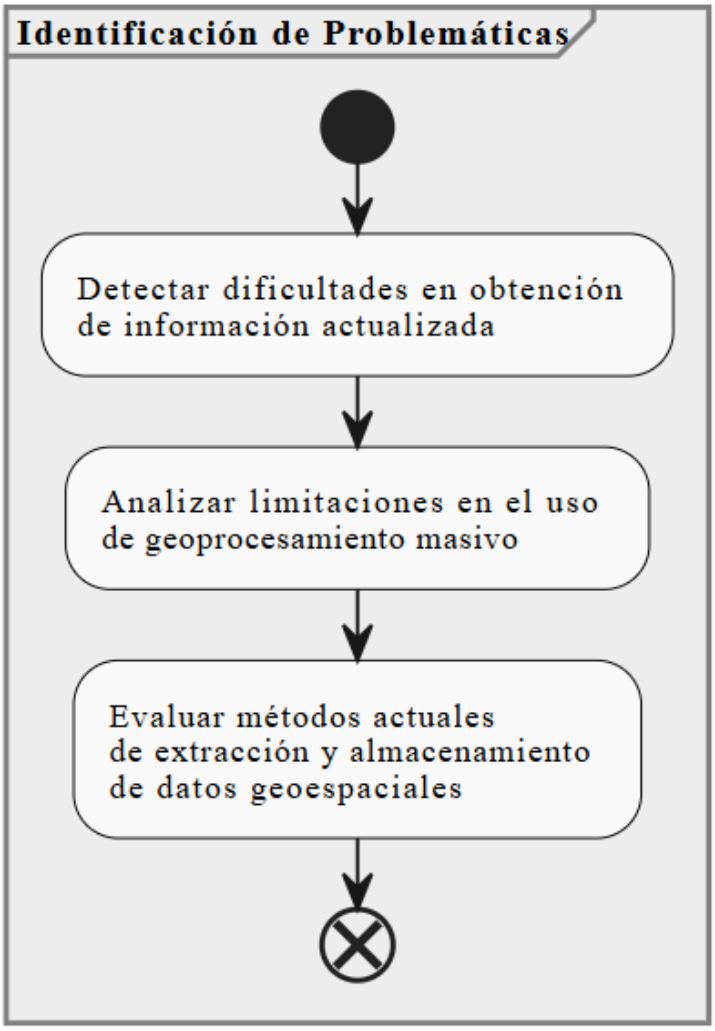


Figura. 7. Proceso de identificación de problemáticas en el uso de datos geoespaciales masivos en el proceso de actualización catastral. Elaboración propia.

- Análisis normativo y de estándares técnicos: revisión de marcos normativos y estándares relacionados con gestión catastral, el proceso de actualización catastral con enfoque multipropósito y el paradigma Big Data para diseñar las respectivas fases del modelo según el método ADM. (véase Anexo A: Vista modelo fase 1: *Etapa Preliminar y Visión de Arquitectura. Fase 2: Capa Arquitectura de Negocio*).

- Análisis del uso de Big Geospatial Data: Se examinó la relación entre los datos geoespaciales masivos y la orquestación de plataformas de Big Data, con el fin de desarrollar las capas o vistas correspondientes al sistema de información y la la de tecnología, de acuerdo con el método ADM. En el marco de este proyecto, el enfoque sobre Big Data se orientó hacia los aspectos técnicos relacionados con el manejo de la información a medida que aumenta su volumen, sin abordar el tratamiento de datos a gran escala. (*véase Anexo A: Capas de Negocio y Sistemas de Información).*

- Implementación técnica: Cómo objeto de estudio se investigó sobre el uso de Apache Sedona como herramienta tecnológica para integrar los datos de Big Geospatial Data y actualización catastral. Para ello, se documentó la configuración e instalación de la herramienta tanto en entorno local como en la nube, así como su uso en el procesamiento distribuido de datos catastrales. (véase *Anexos A y B*).

5.3. Metodología para la adquisición del dominio tecnológico de Apache Sedona.

La metodología se estructuró en cuatro etapas: la investigación de los conceptos fundamentales, el estudio técnico del manejo de la herramienta, la realización de pruebas técnicas y, finalmente, la formulación de la propuesta del clúster.

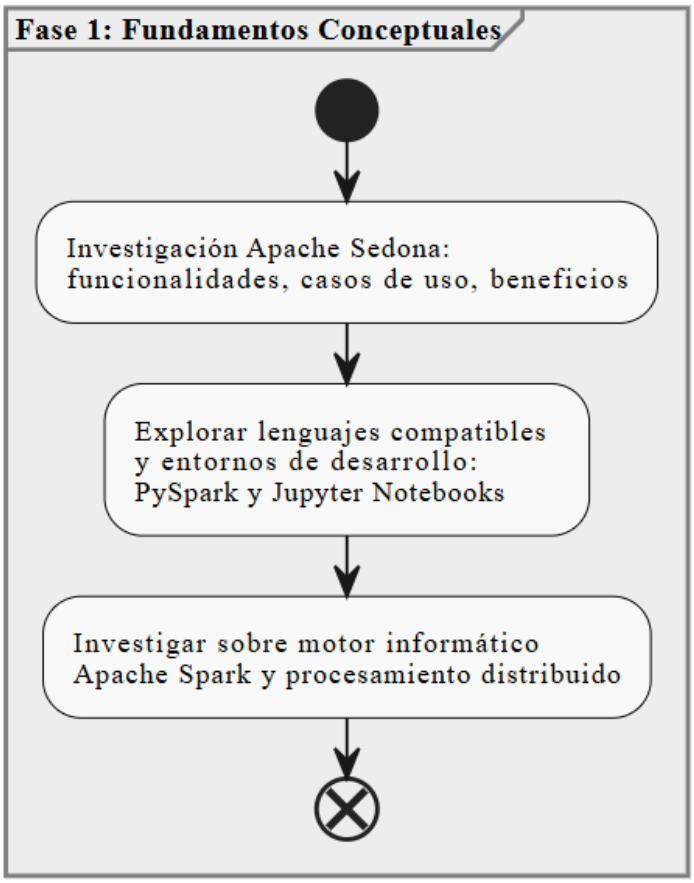
Fase 1: Fundamentos conceptuales

Se abordaron los conceptos fundamentales relacionados con Apache Sedona, incluyendo funcionalidades, casos de uso y beneficios. Además, se identificó la relación estrecha de Sedona con el framework Apache Spark, empleado para el procesamiento distribuido de grandes volúmenes de datos geoespaciales mediante procesamiento paralelo por lotes.

También, se realizó un curso específicamente de la librería de python PySpark y se utilizó Jupyter Notebooks como entorno de desarrollo para desarrollar pruebas, con el propósito de avanzar hacia la fase de preparación técnica requerida para la creación del clúster.

Fase 2: Preparación técnica

Se adquirieron competencias básicas en sistema operativo Linux, Docker Desktop (Docker Engine, Docker Swarm, Docker Images) y Apache Spark para el diseño de la capa tecnológica del modelo, mostrando el funcionamiento del clúster de computadoras en el ambiente local.


Fase 1: Fundamentos Conceptuales
Investigación Apache Sedona:
funcionalidades, casos de uso, beneficios
Explorar lenguajes compatibles
y entornos de desarrollo:
PySpark y Jupyter Notebooks
Investigar sobre motor informático
Apache Spark y procesamiento distribuido

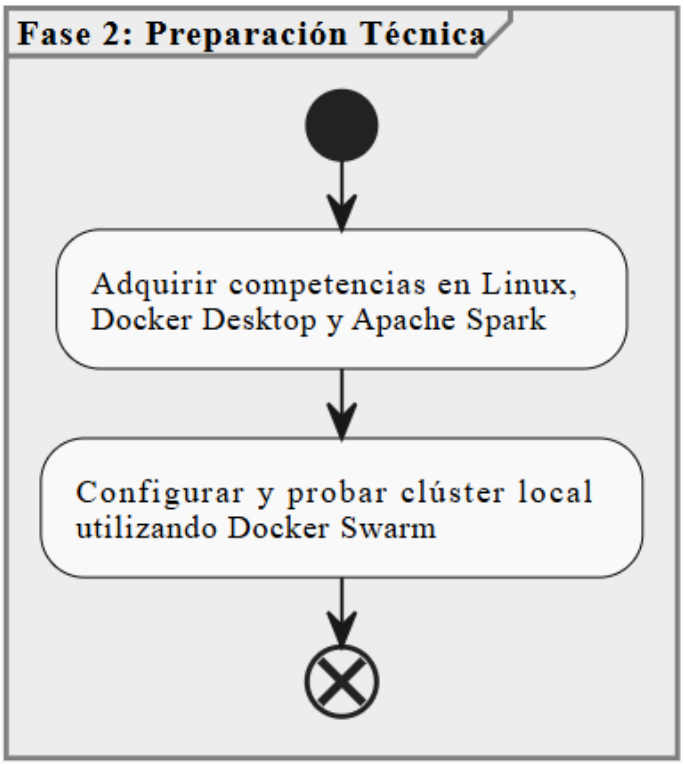


Figura. 8. Fundamentos conceptuales y preparación técnica; Apache Sedona. Elaboración propia.

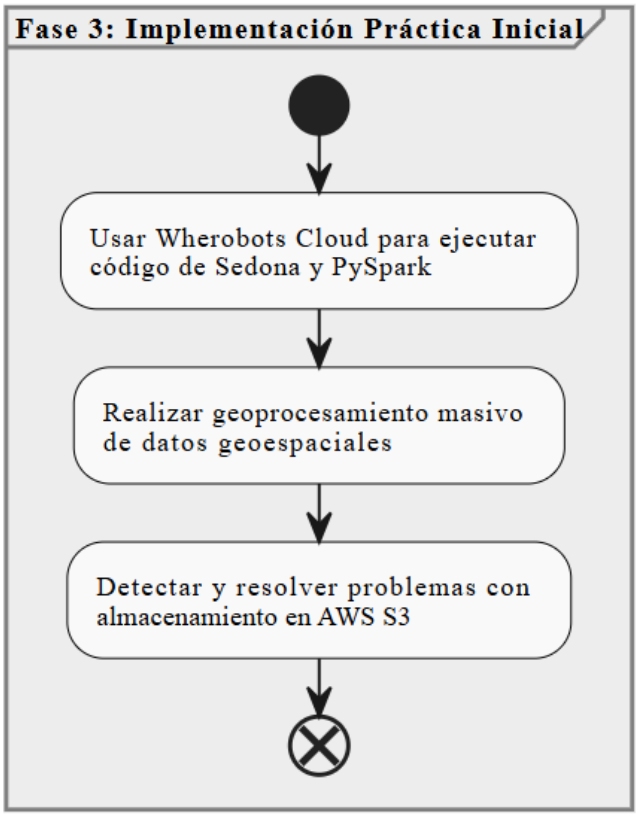
Fase 3: Implementación práctica inicial

Se realizó un acercamiento al funcionamiento de Apache Sedona mediante la plataforma Wherobots Cloud, con el objetivo de comprender la sintaxis necesaria de Sedona junto Pyspark para su ejecución. Se llevaron a cabo geoprocesamientos de grandes volúmenes de datos, obteniendo resultados. Asimismo, se presentaron dificultades relacionadas con la gestión del almacenamiento en AWS S3, debido al desconocimiento de que esta tecnología de Wherobots Cloud continúa ejecutando procesos y genera cobros por los recursos consumidos cuando las sesiones no se cierran correctamente (La duración de la sesión del notebook varía según la versión

del plan de suscripción de Wherobots, véase Anexo C). Estos problemas fueron identificados y resueltos durante el desarrollo de las actividades.

Fase 4: Implementación en entornos tecnológicos

- Entorno local: Se prepararon dos computadoras físicas, una configurada con sistema operativo Linux que actuó como nodo manager y nodo worker, y otra con sistema operativo Windows; utilizando el Subsistema de Windows para Linux (WSL), que funcionó como nodo worker. En ambas máquinas se instaló Docker Desktop y se habilitó Docker Swarm para la creación del clúster. El despliegue del clúster se realizó utilizando la imagen oficial de Apache Sedona (apache/sedona:1.7.1) disponible en Docker Hub.
- Entorno en la nube: Implementación en Wherobots Cloud, gestión de autenticación y credenciales, definición de políticas de acceso y rutas de lectura/escritura para datos catastrales en AWS S3. Se utilizó la plataforma Wherobots Cloud, ya que cuenta con infraestructura preconfigurada para el manejo de Big Geospatial Data. La plataforma se encarga de la arquitectura moderna tipo lakehouse y del funcionamiento de los servidores, lo que facilita la puesta en marcha. El procesamiento espacial se ejecutó mediante Apache Sedona, integrado de forma nativa en Wherobots Cloud, lo que permitió realizar operaciones de geoprocesamiento masivo de forma eficiente y escalable.


Fase 3: Implementación Práctica Inicial
Usar Wherobots Cloud para ejecutar código de Sedona y PySpark
Realizar geoprocesamiento masivo de datos geoespaciales
Detectar y resolver problemas con almacenamiento en AWS S3

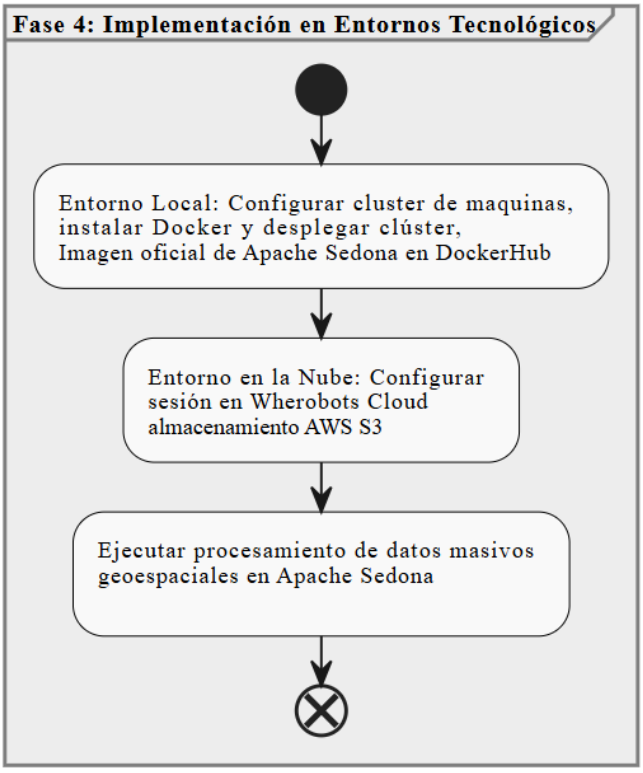

Fase 4: Implementación en Entornos Tecnológicos
Entorno Local: Configurar cluster de maquinas, instalar Docker y desplegar clúster, Imagen oficial de Apache Sedona en DockerHub
Entorno en la Nube: Configurar sesión en Wherobots Cloud almacenamiento AWS S3
Ejecutar procesamiento de datos masivos geoespaciales en Apache Sedona

Figura. 9. Fases 3 y 4 del modelo propuesto: Implementación práctica inicial e implementación en entornos tecnológicos con Apache Sedona.
Elaboración propia.

5.4. Metodología de validación técnica del experimento: Entorno tecnológico.

Para describir las fases 3 y 4, correspondientes a la implementación práctica inicial y a la implementación en entornos tecnológicos orientada a la adquisición de dominio tecnológico mencionadas anteriormente, se definieron preguntas orientadoras que permitieron delimitar el alcance de las actividades y garantizar la coherencia metodológica entre ambas etapas.

Pregunta	Entorno en la Nube: Wherobots Cloud	Entornor local: Apache Sedona y Spark	Entornor local: Postgis
¿Cuál es el objetivo del experimento?	Desarrollar y ejecutar scripts con Apache Sedona integrados con PySpark para realizar geoprocesamiento masivo en un entorno administrado en la nube.	Implementar un clúster de cómputo con Apache Sedona sobre Apache Spark para habilitar y probar geoprocesamientos distribuidos en datos vectoriales.	Ejecutar consultas SQL sobre grandes volúmenes de datos geoespaciales.
¿Cómo se seleccionó la plataforma?	Plataforma con infraestructura preconfigurada y optimizada para Big Geospatial Data, con integración nativa de Apache Sedona y despliegue sin gestión manual de servidores.	Escalabilidad horizontal, compatibilidad con Apache Spark y capacidad para procesamiento distribuido de datos geoespaciales.	Herramienta tradicional y robusta para la gestión de datos espaciales de un solo nodo.
¿Cuál es el proceso de configuración en cada entorno?	Inicialmente entorno gratuito Tiny, posteriormente entorno Medium (plan Professional) con 4 vCPUs y 16 GB de RAM. Datos	Despliegue de imagen Docker de Apache Sedona en un clúster local de tres nodos (1 master y 2 workers) mediante Docker Swarm, con 8 GB de RAM y 4 núcleos cada uno, sobre Apache Spark 3.5.0.	Instalación de PostgreSQL con extensión PostGIS.

	almacenados en bucket de Amazon S3.		
¿Por qué y cómo se eligieron los datos de prueba?	Datos catastrales en formato GeoJSON: i) lotes (946 MB; 933.147 registros), ii) tratamientos urbanísticos según Decreto 555 de 2021 (17,5 MB; 5.636 registros), iii) predios (1,32 GB; 3'768.700 registros).	Mismos datos y formatos utilizados en el entorno en la nube para mantener la comparabilidad.	Mismos datos y formatos que en los otros entornos.
¿Cómo se diseñaron las consultas de prueba?	Creación de vistas temporales y uso de PySpark y SedonaSQL para obtener el tratamiento urbanístico de cada predio en Bogotá.	Igual procedimiento que en el entorno en la nube.	Consulta SQL espacial para el mismo objetivo.
¿Cómo se definieron las métricas de evaluación?	Curva de aprendizaje, tiempo de ejecución por consulta, costo de operación (recursos y Spatial Units) y políticas de cierre de sesión.	Curva de aprendizaje, tiempo total de ejecución, éxito o fallo de la operación, facilidad de implementación y uso máximo de memoria.	Tiempo total de ejecución, éxito o fallo de la operación y uso máximo de memoria.

Tabla 1. Preguntas orientadoras para delimitar alcance de las herramientas.

Nota: La elección de los conjuntos de datos respondió a lo establecido en el Decreto 555 de 2021, que modifica la normatividad vigente del Plan de Ordenamiento Territorial de Bogotá Decreto 190 de 2004. Se empleó cartografía oficial de lotes, normativa urbanística vigente y la base de datos de predios. El tamaño y volumen de estos conjuntos evidencian la necesidad de herramientas de Big Data, ya que, por ejemplo, el archivo CSV de predios excede el límite de registros que puede manejar una hoja de cálculo convencional.

6. DESCRIPCIÓN DE RESULTADOS

La descripción de los resultados obtenidos en esta investigación responde directamente al objetivo general y objetivos específicos planteados. El objetivo general

consistió en definir un modelo para la integración de Big Geospatial Data al proceso de actualización catastral, basado en el estudio y uso práctico de la plataforma de código abierto Apache Sedona.

6.1. Dimensiones de Big Data en relación con la actualización catastral.

En cumplimiento del primer objetivo específico, se realizó la caracterización del rol del Big Geospatial Data dentro del proceso de actualización catastral. En esta etapa se identificaron los conceptos fundamentales de Big Data aplicados al ámbito geoespacial, destacando los aspectos relacionados a Big Data; volumen, variedad, velocidad, valor y veracidad de los datos, y qué relación tiene con el proceso de actualización catastral. Esto permitió establecer una base conceptual que fundamenta el modelo propuesto.

6.1.1 Variedad

En el proceso de actualización catastral existen varios subprocesos, cada uno con métodos distintos para la obtención de datos. Uno de los principales retos de Big Data, en cuanto a la variedad, radica en el tratamiento de conjuntos de datos provenientes de múltiples fuentes y con distintos formatos, lo que exige procesos más complejos de integración y análisis. En este contexto, es común encontrar datos estructurados, semiestructurados y no estructurados, lo cual requiere enfoques técnicos diferenciados para su procesamiento.

Dimensión de Big Data: Variedad y Actualización Catastral

Identificación de la dimensión en la actualización catastral	Integración de datos en distintos formatos: vectores (shapefiles, GeoJSON), raster, imágenes, tablas alfanuméricas con información jurídica, física y económica de los predios, y datos geográficos relacionados con el POT (cartografía oficial, visores, geoservicios), entre otros.
Uso de la herramienta de big data.	Las herramientas de Big Data permiten estandarizar e integrar estos datos heterogéneos para facilitar su procesamiento, almacenamiento y análisis masivos dentro de flujos automatizados.
Ejemplo y/o caso de estudio	Para el desarrollo práctico de la investigación se recopilaron datos geoespaciales de la normativa del POT de Bogotá, establecidos en el Decreto 555 de 2021, así como información sobre construcciones a partir de servicios ofrecidos por Google y datos abiertos proporcionados por entidades públicas, mediante el uso de geoservicios y visores SIG. Se obtuvieron datos geográficos en diferentes tipos de formatos como GeoJSON, WKT y Shapefile, además de tablas estructuradas en Excel. Adicionalmente, se incorporaron datos recolectados manualmente mediante herramientas como Google Street View, el visor SINUPOT y la plataforma Mapas Bogotá.

Tabla 2. Dimensión de Big Data: Variedad y actualización catastral.

Volumen

A medida que avanza la tecnología así mismo avanzan los métodos para la recolección y creación de datos, en el proceso de actualización Catastral se evidencia que existe variedad de fuentes de datos de distintas organizaciones dónde se pueden extraer datos para la consolidación de toda la información y ponerla al servicio del negocio del gestor y/o operador catastral.

En este contexto, el aumento en el acceso a diversas fuentes de datos abiertos facilita, por ejemplo, la identificación de nuevas construcciones mediante herramientas comunes como Street View de Google Maps o a partir de datos generados por entidades públicas. Estas fuentes, al igual que el uso habitual del teléfono celular para la captura de imágenes en campo, contribuyen activamente a la generación de Big Data en los procesos catastrales.

Dimensión de Big Data: Volumen y Actualización Catastral	
Identificación de la dimensión en la actualización catastral	El volumen de datos crece constantemente debido a la generación masiva y continua de información geoespacial y alfanumérica por parte de múltiples entidades tanto públicas como privadas. Este incremento incluye datos como cartografía base, levantamientos topográficos, bases de datos geoespaciales, imágenes satelitales, información registral y bases externas.
Uso de la herramienta de big data.	Se utilizan infraestructuras escalables, como los data lakehouse, para almacenar y organizar grandes volúmenes de datos. Tecnologías como Apache Sedona y Apache Spark permiten el procesamiento distribuido de datos geoespaciales y alfanuméricos a gran escala.
Ejemplo y/o Caso de estudio	En el caso de Bogotá D.C., la ciudad cuenta con aproximadamente 2.890.251 lotes vigencia 2025, según datos oficiales de la Unidad Administrativa de Catastro Distrital (UAECD,2025). Esta magnitud implica gestionar una gran cantidad de información espacial y alfanumérica asociada a cada predio, entre esos datos físicos, económicos y jurídicos.

Tabla 3. Dimensión de Big Data: Volumen y actualización catastral.

6.1.2. Velocidad

De acuerdo con lo mencionado anteriormente en la dimensión de volumen, en la actualidad los datos se generan y actualizan con mayor rapidez y variedad. Anteriormente, la captura de información se realizaba de forma manual, por ejemplo, en papel o hojas de Excel. Hoy en día, esa información puede ser registrada tanto en papel y archivos Excel como a través de encuestas y/o formularios digitales que se integran directamente con tableros de visualización, lo que permite obtener análisis casi en tiempo real.

En este contexto, los procesos de obtención y transformación de datos para la actualización catastral, que antes podían tomar días o incluso meses, se ven optimizados gracias a las funciones y capacidades de las nuevas herramientas tecnológicas.

Dimensión de Big Data: Velocidad y Actualización Catastral

Identificación de la dimensión en la actualización catastral	La actualización catastral enfrenta el desafío de adaptarse a la alta velocidad con la que cambian las condiciones del territorio, impulsada por desarrollos urbanos, cambios normativos y nuevas construcciones.
Uso de la herramienta de big data.	Permite integrar y procesar flujos de datos en tiempo real o casi real mediante automatización y procesamiento distribuido. Esto reduce el tiempo entre la generación de datos y su incorporación al sistema catastral.
Ejemplo y/o Caso de estudio	Durante el desarrollo práctico de la investigación se evidenció que herramientas como Apache Spark y Apache Sedona permiten construir flujos de procesamiento de datos (pipelines) capaces de gestionar múltiples fuentes de datos en paralelo. Se ejecutó operaciones espaciales como intersección, contención y filtrado sobre millones de registros geoespaciales en tiempos optimizados respecto a métodos tradicionales.

Tabla 4. Dimensión de Big Data: Velocidad y actualización catastral.

6.1.3. Veracidad

En el contexto de Big Data, la dimensión de veracidad hace referencia a la integridad, precisión y calidad de los datos utilizados. En procesos como la actualización catastral, esta dimensión cobra relevancia, ya que los análisis técnicos, la planeación territorial y la toma de decisiones debe estar soportada de información verificada y fuentes confiables. Durante la investigación, se identificaron estrategias para asegurar la calidad de la información, como la validación visual, la depuración de registros y la aplicación de reglas de control en el manejo de los datos.

Tal como lo mencionan Becerra, J. et al, ignorar la veracidad de los datos o de los procedimientos asociados puede derivar en consecuencias graves, los autores citan al

Parlamento Europeo (2017), el cuál advierte que "*el uso de datos imprecisos puede llevar al diseño de algoritmos sesgados, producir correlaciones erróneas o subestimar los impactos éticos, sociales y legales de las decisiones automatizadas*", los autores señalan que el big data solo será tan confiable como lo sean los datos que los sustentan. En conclusión, estos autores mencionan que los errores producidos pueden traducirse en procedimientos negativos que pueden afectar la vida de las personas, especialmente en poblaciones vulnerables, y también que comprometen la legitimidad de las instituciones (Becerra, J. et al, 2018).

Dimensión de Big Data: Veracidad y Actualización Catastral	
Identificación de la dimensión en la actualización catastral	En este proceso participan distintos roles de talento humano que cumplen distintas funciones: desde la obtención de información, la generación de datos confiables que respalden decisiones estratégicas, hasta la revisión técnica y jurídica, orientada a garantizar la precisión y validez de los datos, los cuales son transformados en información útil para la actualización catastral.
Uso de la herramienta de big data.	El uso de herramientas de Big Data como Apache Spark y Apache Sedona, integradas en una arquitectura Lakehouse, permite implementar mecanismos automáticos de trazabilidad, control de calidad y validación de datos. Como lo señala Vasconcelos (2024), los Lakehouses "implementan capas que proporcionan propiedades ACID, validación de esquemas y otras funcionalidades", superando así "la falta de procesos de gobernanza y gestión de datos". Esta arquitectura puede materializar la dimensión de veracidad en el proceso de actualización catastral, al garantizar que los datos utilizados sean consistentes, auditables y técnicamente verificables. (Vasconcelos,2024).

Ejemplo y/o Caso de estudio	Como parte de la práctica, se investigó el funcionamiento de pipelines y triggers en bases de datos para la gestión de datos geoespaciales. Un ejemplo de esta implementación fue el uso de QGIS, conectado a una base de datos PostGIS, para corregir datos vectoriales. A través de un script automatizado, el cambio se reflejó directamente en la base de datos, activando un trigger que actualizó el valor del dato y registró su estado anterior para fines de auditoría y trazabilidad. Cabe mencionar que PostgreSQL con PostGIS es compatible con Apache Sedona, lo que permite integrar datos espaciales en pipelines de procesamiento distribuido con Apache Spark.

Tabla 5. Dimensión de Big Data: Veracidad y actualización catastral.

6.1.4. Valor

Becerra et al. (2018) retoman la definición de Hopkins (2011), quien plantea que el aspecto de valor en Big Data se refiere a la utilidad que se obtiene al transformar grandes volúmenes de datos en información significativa para la toma de decisiones y el valor económico, haciendo que las soluciones sean asequibles y ayuden a la inversión de los negocios. (Hopkins, 2011; Becerra, J. et al, 2018) En el caso de la actualización catastral, este valor se materializa al consolidar, integrar y procesar datos provenientes de diversas fuentes, lo que permite generar productos útiles como insumos para el análisis masivo del territorio.

Dimensión de Big Data: Valor y Actualización Catastral	
Identificación de la dimensión en la actualización catastral	Al transformar los datos generados en las fases pre-operativa y operativa en matrices de decisión, se optimizan las actividades del proceso, se identifican inconsistencias y se planifican acciones con mayor precisión. En la fase post-operativa, esto se traduce en la consolidación de productos catastrales como la liquidación de avalúos, la asignación del NUPRE, la generación de los registros R1 y R2 y la creación del archivo XTF.

Uso de la herramienta de big data.	El enfoque Lakehouse, al combinar almacenamiento estructurado con capacidades de análisis masivo, permite organizar y transformar grandes volúmenes de datos espaciales generados en el proceso (Vasconcelos, 2024). Con herramientas de procesamiento distribuido como Apache Sedona junto a Apache Spark, es posible estructurar la información para tareas operativas, técnicas y administrativas asociadas a la toma de decisiones.
Ejemplo y/o Caso de estudio	Durante la investigación se evaluó Apache Sedona tanto en un entorno local mediante contenedores Docker como en la nube, a través de Wherobots Cloud, para realizar operaciones como intersección, agrupamiento y filtrado, con el fin de integrar, transformar y analizar datos espaciales y alfanuméricos masivos. La herramienta se empleó para generar una matriz de decisión con su respectiva base de datos espacial, a partir de información sobre construcciones y normativa urbanística de la ciudad de Bogotá. Esto permitió ejecutar análisis más rápidos y automatizados, identificar la normativa asociada a cada predio, así como detectar datos duplicados, inconsistentes y construcciones no registradas, entre otros.

Tabla 6. Dimensión de Big Data: Valor y actualización catastral.

6.2. Proceso de integración de Big Geospatial Data en la actualización catastral.

Respecto al segundo objetivo específico, se definió y representó un proceso general y estructurado para la integración de Big Geospatial Data en las diferentes fases del proceso de actualización catastral. Este proceso, resumido en la Tabla 7, se construyó con base en lo investigado sobre el ecosistema de Big Data, teniendo en cuenta especialmente el rol de Apache Sedona como caso de estudio y su posible aplicación en cada etapa del proceso catastral.

Fases Actualización Catastral	Proceso Actual	Integración con Big Data	Herramienta técnica.
Pre-Operativa	Identificación de la zona a intervenir mediante insumos como ortofotos,	Automatización de tareas operativas y	Automatización de recolección de datos

	topografía, cartografía base e instrumentos de ordenamiento territorial (POT, PBOT, EOT), con base en recursos secundarios recopilados previamente. (IGAC, 2023)	consolidación de datos. Integración de información del mercado inmobiliario (renta y venta) para análisis espacial.	mediante web scraping y web crawling. Almacenamiento en data lakehouse (Vasconcelos, 2024). Visualización y validación inicial con herramientas SIG tradicionales.
Operativa	Recolección de información física, jurídica y económica de los predios, siguiendo la Resolución 746 de 2024 y el Decreto 148 de 2020. Se aplican métodos directos, indirectos, colaborativos y declarativos. Los datos se validan, se definen zonas homogéneas y se consolida la información. (IGAC, 2023)	Procesamiento masivo de datos espaciales (vectoriales y ráster) con técnicas de limpieza y transformación. Generación de información confiable para estudios técnicos y de mercado.	Apache Sedona para geoprocesamiento masivo. Almacenamiento en data lakehouse. Bases SQL y NoSQL bajo el modelo LADM-COL SINIC V1.0. Automatización de flujos con pipelines. Validación con SIG tradicionales.
Post-Operativa	Consolidación de productos catastrales (avalúos, NUPRE, documentos técnicos y administrativos), entrega a municipios y reporte al SINIC con los resultados de la actualización (Registros R1, R2, archivo XTF). (IGAC, 2023)	Automatización del reporte y trazabilidad de los datos. Reducción de tiempos de respuesta y eficiencia en la consulta de grandes volúmenes de datos.	Automatización de reportes con pipelines y scripts. Uso de Apache Sedona y arquitectura lakehouse para procesamiento distribuido y almacenamiento estructurado tanto de los datos obtenidos durante el proceso cómo ya de los datos resultantes.

Tabla 7. Integración de Big Geospatial Data en los procesos de actualización catastral. Elaboración propia.

6.3. Resultado desempeño del prototipo en comparación con métodos tradicionales.

Con el fin de evaluar el desempeño del prototipo y validar la capa tecnológica del modelo propuesto, se realizaron pruebas experimentales orientadas al geoprocesamiento de Big Geospatial Data. Los experimentos se enfocaron en medir la capacidad de respuesta, el tiempo de ejecución y la viabilidad operativa de las diferentes herramientas en el manejo de grandes volúmenes de datos vectoriales en formato GeoJSON.

Las pruebas incluyeron tanto un clúster local de Apache Sedona sobre Apache Spark como la solución en la nube de Wherobots (también basada en Apache Sedona), y como referencia comparativa, PostGIS como representante de métodos tradicionales. Los resultados obtenidos se presentan en la siguiente tabla:

Herramienta	Sistema Operativo	Configuración	Conjunto de Datos	Tiempo de ejecución	Resultado
Apache Sedona + Apache Spark (local)	Linux	Cluster: Docker Swarm, red overlay, nodo master y nodos worker. Script: JupyterLab con kernel Python, instalación librerías PySpark y Sedona	Datos vectoriales (GeoJSON)	Consulta: 1 min 40 s Descarga: 5 min	Archivo CSV con 947,453 registros y atributos: código_lote, nombre tratamiento, porcentaje_area, area intersection AB, geometría, orden de resultado discriminado por código lote y tipo de tratamiento.
Wherobots Cloud + Apache Sedona (Nube)	Servidores en nube	Infraestructura preconfigurada; entorno JupyterLab con kernel Python y almacenamiento en Amazon S3	Datos vectoriales (GeoJSON)	Consulta: 2 s Descarga: 11 min	Archivo CSV con 947,453 registros y atributos: código_lote, nombre_tratamiento, porcentaje_area, area_intersection_AB, geometría, orden de resultado discriminado por código lote y tipo de tratamiento.
PostgreSQL + PostGIS (Local)	Linux	Instalación de postgresql y creación de la extensión espacial PostGIS y PostGIS_topology	Datos vectoriales (GeoJSON)	-	SQL Error [54000]: ERROR: memoria agotada. Detail: No se puede agrandar el búfer de cadena que ya tiene 1073741670 bytes en 2559 bytes adicionales.

Tabla 8. Comparación del desempeño del prototipo frente a métodos tradicionales en entornos local y nube.

Durante el experimento, se ejecutó una consulta para identificar los tratamientos urbanísticos aplicables a todos los lotes de Bogotá.

- En Wherobots Cloud, la consulta se completó en 2 segundos, requiriendo 11 minutos adicionales para la descarga de los resultados, gracias a su infraestructura preconfigurada.
- En el clúster local de Apache Sedona sobre Apache Spark, la misma consulta tomó 1 minuto y 40 segundos, con una descarga de 5 minutos
- En PostGIS, la consulta no finalizó debido a un error por insuficiencia de memoria, generado por la imposibilidad de ampliar el búfer de cadena ya existente.

Estos resultados confirman que las herramientas tradicionales de software GIS, como PostGIS, enfrentan limitaciones en el procesamiento de datos geoespaciales de gran escala, tanto en consumo de memoria como en tiempos de ejecución. Por el contrario, las soluciones distribuidas basadas en Apache Sedona tanto en despliegue local como en la nube logran completar el procesamiento sin errores y con un uso más eficiente de los recursos.

No obstante, se evidenció que la configuración de un clúster local con Apache Sedona y Spark requiere mayor conocimiento técnico, una curva de aprendizaje más pronunciada y una preparación de infraestructura más compleja en comparación con soluciones en la nube como Wherobots, que abstraen gran parte de la configuración y optimizan automáticamente las consultas espaciales paralelizadas.

En síntesis, los resultados validan que el procesamiento distribuido es más adecuado para manejar grandes volúmenes de datos geoespaciales en contextos como la actualización catastral masiva, respaldando la pertinencia técnica del modelo propuesto.

6.4. Resultado del modelo de integración de Big Geospatial Data en la actualización catastral.

En este sentido, se elaboró un modelo basado en el método ADM-TOGAF y representado mediante el lenguaje ArchiMate, las capas que se desarrollaron del modelo incluyen la capa Preliminar, la visión de la arquitectura, la arquitectura de negocio, la de sistemas de información y la de arquitectura tecnológica, cada una abordada desde su

aplicación práctica con Apache Sedona y con datos relacionados al proceso de actualización catastral.

Capa	Preguntas orientadoras	Resultado del modelo propuesto
Preliminar	¿Cuál es el propósito general del modelo de integración propuesto?	Formular un modelo técnico viable para la integración de Big Geospatial Data en el proceso de actualización catastral, usando tecnologías de código abierto y orientado al cumplimiento normativo vigente.
A. Visión de la Arquitectura	¿Cuál es la meta principal que se busca alcanzar mediante la implementación del modelo?	Demostrar que es posible mejorar el proceso de actualización catastral mediante herramientas distribuidas que permiten procesar grandes volúmenes de datos espaciales y favorezca la eficiencia y ahorro de tiempo y costentes en el proceso y cumplimiento normativo.
B. Arquitectura de Negocio	¿Cómo se estructuran actualmente los procesos de gestión de información catastral por parte de los operadores?	Se identificaron y documentaron los servicios y procesos esenciales en las fases pre-operativa, operativa y post-operativa. Se analizaron requerimientos técnicos de información usados actualmente, como insumos topográficos y normativos.
C. Arquitectura de Datos	¿Qué tipos de datos se utilizaron y cómo se estructuraron en el modelo?	Se trabajó con datos vectoriales (GeoJSON, WKT y Shapefile) y datos alfanuméricos. Se construyó una matriz de decisión para predios no edificados en Bogotá, integrando fuentes como SINUPOT y visores normativos. No se utilizó información ráster.
C. Arquitectura de Sistema de información	¿Cuál es el rol de Apache Sedona dentro del sistema? ¿Qué funcionalidades aporta al proceso de análisis?	Se diseñó un proceso técnico-práctico basado en Apache Sedona sobre Spark, para realizar tareas como filtrado, intersección y análisis espacial masivo. Esta implementación demostró su utilidad en la validación normativa de predios.
D. Architectura tecnologica	¿Cómo se implementa la infraestructura técnica para el procesamiento distribuido? ¿Qué ventajas aporta este enfoque?	Se configuró un clúster local de Apache Spark y contenedores Docker para ejecutar Apache Sedona. Esta arquitectura tecnológica permitió validar la viabilidad técnica del modelo en escenarios prácticos de procesamiento geoespacial distribuido.

Tabla 9. Preguntas genereales formuladas para estructurar el modelo.

Vista modelo fase 1: Etapa preliminar y visión de arquitectura:

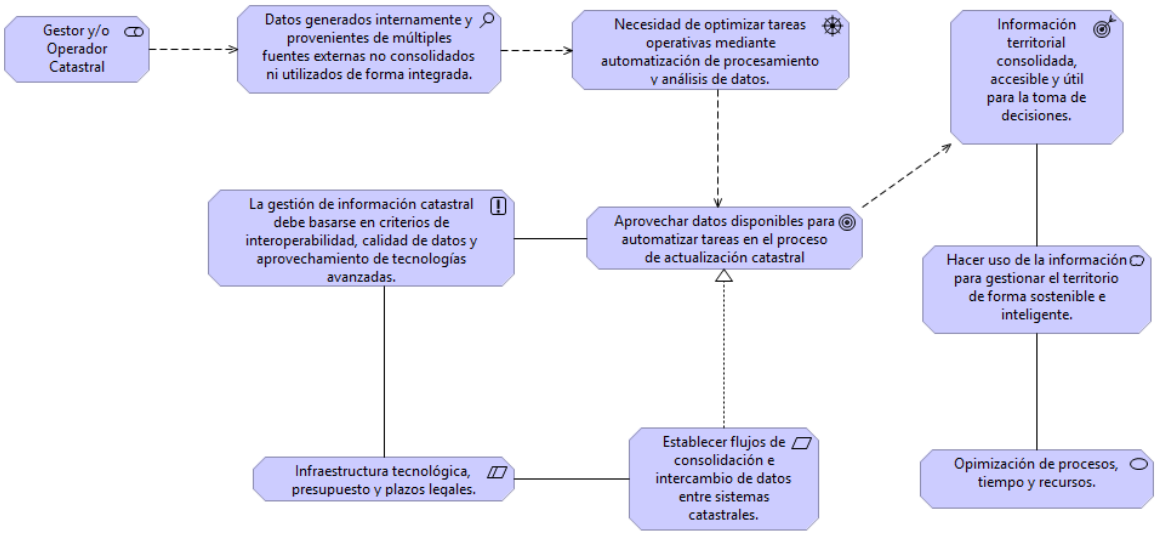


Figura1. Vista modelo fase 1. Elaboración propia.

En el resultado de la vista correspondiente a la capa preliminar y a la visión de la arquitectura, como se observa en la Figura 1, se representa la situación actual en la que el **Gestor y/o Operador Catastral**, en su rol como parte interesada (stakeholder), se enfrenta a una situación en la que, durante los procesos de trabajo, se observa que **los datos generados internamente y aquellos provenientes de diversas fuentes no se han consolidado ni integrado, lo que impide aprovechar de forma eficiente estos recursos en las operaciones técnicas de cada proceso de la gestión catastral**. Esta condición es identificada como un assessment, el cual influye directamente en la aparición de una necesidad (driver): **optimizar tareas operativas mediante la automatización del procesamiento y análisis de datos**.

A partir de esta necesidad y conforme al artículo 81 de la Ley 1955 de 2019, que establece las responsabilidades y sanciones para gestores y operadores del servicio catastral, se define un objetivo (goal) que consiste en **aprovechar los datos disponibles para automatizar y/o optimizar las tareas en el proceso de actualización catastral**. Para alcanzar dicho objetivo, se identifican dos elementos clave:

- Un requisito (requirement) técnico y funcional: **Establecer flujos para la consolidación e intercambio de datos que integren tanto la información proveniente de los sistemas catastrales como la de fuentes externas.**
- Un resultado esperado (outcome): **Información territorial consolidada, accesible y útil para la toma de decisiones, facilitando la optimización de las tareas operativas.**

El resultado esperado (vinculado a los productos mínimos que debe entregar el proceso de actualización catastral conforme al artículo 4.2.6 de la Resolución 1040 de 2023) se asocia

con un significado (meaning): **hacer uso de la información para gestionar el territorio de forma sostenible e inteligente**. El uso integral de la información no se limita a la optimización de tareas operativas, sino que también permite la gestión del territorio con enfoque multipropósito. Este significado se vincula con la generación de valor (Value) para el stakeholder, expresado como **optimización de procesos, tiempo y recursos en la gestión catastral**.

De manera transversal, se establece un principio (principle) que orienta el modelo: **la gestión de información catastral debe basarse en criterios de interoperabilidad, calidad de datos y aprovechamiento de tecnologías avanzadas**. Este principio se relaciona con las limitaciones (constraint), que comprenden la **falta de infraestructura tecnológica, recursos y plazos legales para entregar la información**, lo cual, a su vez, promueve la investigación del modelo propuesto.

Vista modelo fase 2: Arquitectura de negocio (Proceso de actualización catastral):

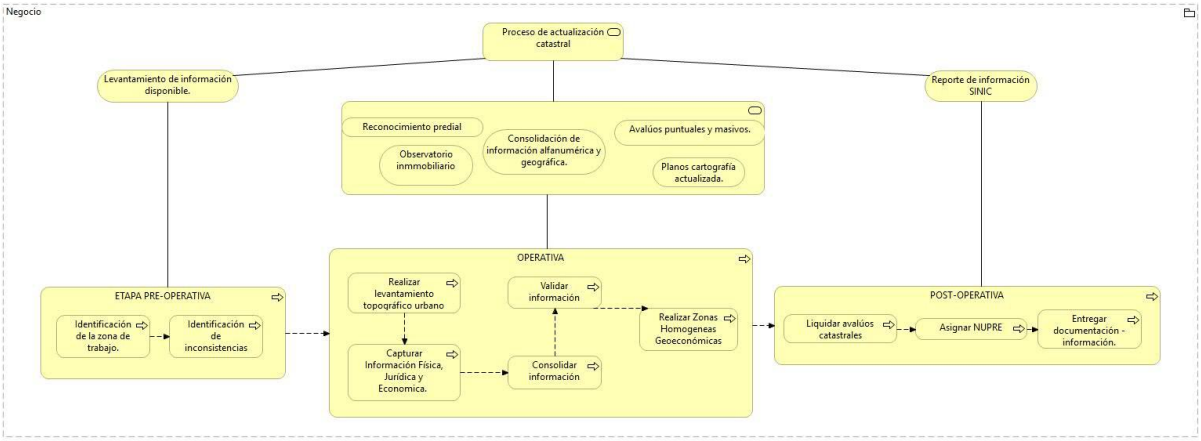


Figura 1-1. Vista modelo fase 2 - Capa arquitectura de negocio. Elaboración propia.

En la Figura 2 se muestra la representación gráfica de los servicios funcionales y procesos operativos que desarrolla el Gestor y/o Operador en el Proceso de Actualización Catastral. Esta vista facilita la comprensión del negocio al identificar los servicios que presta, detallar la estructura operativa y explicar cómo se organizan y coordinan las actividades que conforman el proceso. En la capa de arquitectura de negocio, se detallaron los servicios y procesos operativos esenciales para la actualización catastral, estructurados en etapas pre-operativa, operativa y post-operativa. Se identificaron los métodos y requerimientos técnicos de información necesarios con el fin de crear el modelo.

Vista modelo fase 2: Arquitectura de Sistemas de información (Big Geospatial Data)

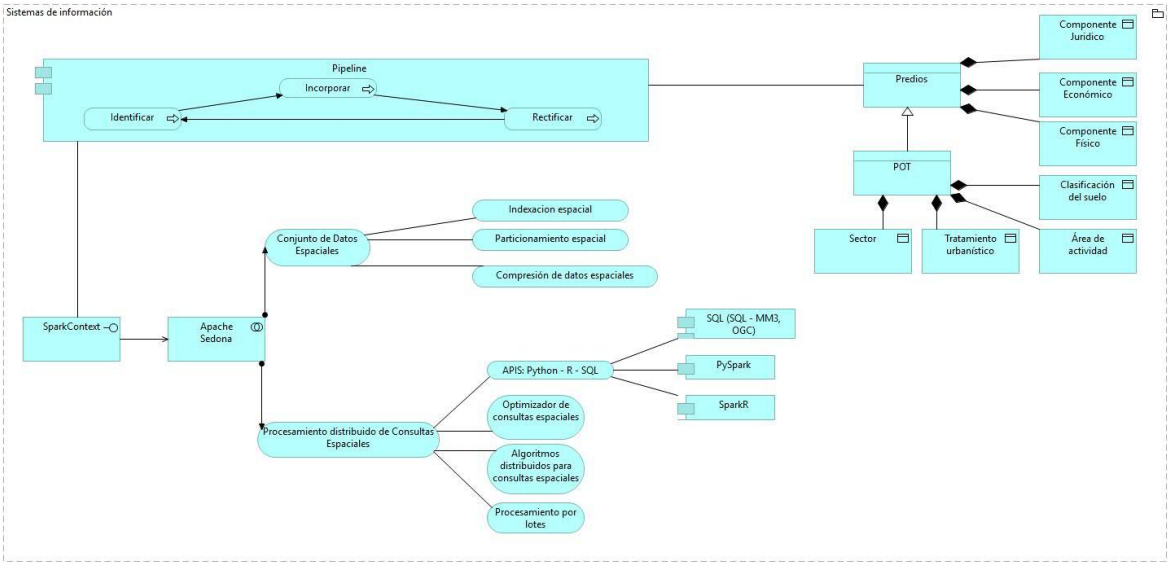


Figura 1-2. Vista modelo fase 2: Capa arquitectura de sistemas de información. Elaboración propia.

La capa de arquitectura de sistemas de información, mostrada en la figura 3, vincula la capa de aplicación y la de datos del modelo para respaldar la visión de la arquitectura siguiendo los parámetros del método ADM. Esta arquitectura surge ante la necesidad de optimizar los procesos operativos que se llevan a cabo en la actualización catastral, plasmados en la capa de negocio, integrando el concepto de Big Geospatial Data para consolidar y aprovechar los datos disponibles. De esta forma, se busca optimizar y/o automatizar las tareas técnicas operativas y reducir costos en términos de tiempos de ejecución y recursos.

La capa de sistemas de información definió un proceso técnico-práctico mediante la implementación de un pipeline utilizando Apache Sedona para la integración y análisis distribuido de grandes volúmenes de datos geoespaciales. Esta implementación práctica,

detallada en los anexos, verificó la eficacia del proceso propuesto, demostrando capacidad para consultas espaciales masivas y análisis relacionados con la normativa territorial vigente.

Específicamente, en la parte de datos, se realizó un análisis de cómo obtener una matriz de decisión para los predios considerados no edificados en Bogotá, en cumplimiento del Decreto 555 de 2021. Este análisis implicó la identificación de reglas para la clasificación de predios según criterios normativos, área de terreno, zona homogénea física y destino económico. La integración de datos alfanuméricos y geográficos se realizó mediante la construcción de una matriz que permitió validar, depurar y caracterizar lotes con base en normativa urbanística vigente. Se utilizó información proveniente de distintas fuentes, incluyendo geoservicios como SINUPOT y visores cartográficos específicos. La digitalización manual complementó la información faltante para asegurar la precisión en la matriz final, esencial para los análisis siguientes.

Vista modelo fase 2: Arquitectura Tecnológica (Clúster de computadores ambiente local)

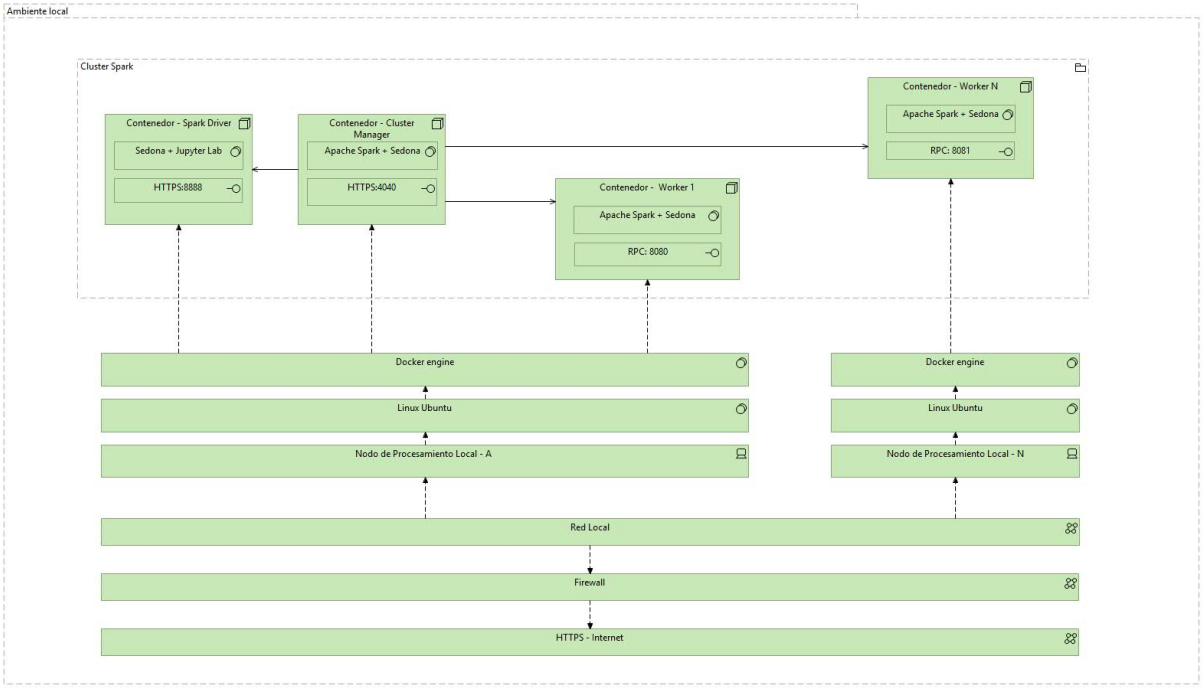


Figura 1-3. Vista del modelo de la fase 2: capa arquitectura tecnológica. Elaboración propia.

Finalmente, en la capa tecnológica, también representada en los anexos, se documentó la implementación técnica del modelo utilizando un clúster de computadores en un entorno local. Para el despliegue distribuido de Apache Sedona se utilizaron tecnologías como Apache Spark, encargado de la configuración del clúster de cómputo, y contenedores Docker para la conexión y gestión de los nodos. Gracias a la escalabilidad horizontal que ofrece esta arquitectura, es posible añadir múltiples nodos worker según los requerimientos del sistema. De esta manera, se validó la viabilidad técnica del modelo propuesto y su adaptabilidad a escenarios prácticos de procesamiento geoespacial.

En el caso de estudio se implementará un clúster de computadoras, el cual se muestra en la vista de la capa de arquitectura tecnológica, tal como se observa en la figura 4, empleando Apache Spark como sistema de cómputo, el cual utiliza un motor informático para distribuir

los datos y las tareas de procesamiento entre el clúster de máquinas. Para configurar el entorno, se establecen parámetros en Spark que reconocen a Sedona como extensión espacial.

Al usar contenedores y marcos de trabajo como Spark para entornos de procesamientos distribuidos, se facilita la escalabilidad horizontal: si se requiere mayor capacidad de procesamiento, es suficiente con añadir más nodos o contenedores en el clúster de Spark. En el modelo de la capa tecnológica, el clúster está formado por dos equipos. El primero aloja tres contenedores: uno para el Spark Driver (puerto 8888), otro para el Cluster Manager (puerto 4040) y un tercero para un Cluster Worker (puerto 8080). El segundo equipo incluye un contenedor para un nodo adicional del clúster. Ambos equipos se interconectan mediante una red local, un firewall e Internet.

En la vista de arquitectura, la segmentación entre la red local e Internet y el firewall refuerzan la seguridad y el control del tráfico de datos. Los puertos expuestos se delimitan y la comunicación entrante y saliente de cada contenedor se regula para evitar accesos no autorizados. El acceso a JupyterLab se realiza vía HTTPS para proteger la conexión con el Spark Driver. A través de esa interfaz, el Spark Driver solicita recursos y coordina la ejecución de Apache Sedona sobre el motor de Spark, delegando la asignación de tareas al Administrador de Clúster. Este recibe los detalles del trabajo y reparte las tareas a los nodos trabajadores, garantizando aislamiento y disponibilidad de CPU y memoria para la ejecución efectiva. El clúster en el ambiente local simula escenarios de producción a menor escala, permite realizar pruebas, validar configuraciones y ajustar el rendimiento antes de la puesta en marcha. Así, el modelo de la arquitectura satisface los requisitos de ejecución y análisis y proporciona un entorno que se adapta a las necesidades de la aplicación.

7. DISCUSIÓN DE RESULTADOS

7.1. Desafíos actuales en la gestión de datos geoespaciales

Los sistemas tradicionales de procesamiento y almacenamiento de información geoespacial se han vuelto insuficientes frente a la magnitud, complejidad y velocidad con la que se producen los datos hoy en día. En el contexto del Catastro Multipropósito, cada actor institucional genera insumos que, sin una adecuada interoperabilidad, terminan fragmentados o incluso desperdiciados. Incluso dentro de una misma organización, las distintas áreas suelen generar datos de forma descentralizada y sin mecanismos de articulación.

La acumulación de archivos en formatos incompatibles, la duplicidad de registros y la falta de trazabilidad dificultan la actualización eficiente del catastro, especialmente en entornos urbanos dinámicos como los que presentan muchas capitales del país. Esta situación se agrava cuando los sistemas no están preparados para integrar información proveniente de distintas fuentes, lo que limita la capacidad de respuesta ante los cambios constantes del territorio.

7.2. Rol de Apache Sedona y el procesamiento distribuido

La incorporación de Apache Sedona al flujo de trabajo catastral representa una respuesta directa a estos retos. Su capacidad para ejecutar geoprocesos a gran escala, especialmente con datos vectoriales, permite automatizar tareas antes manuales, como la detección de cambios, la validación espacial o la generación de estadísticas territoriales. Aunque no busca reemplazar a los SIG tradicionales, Sedona complementa sus funciones al facilitar la integración de grandes volúmenes de información en entornos distribuidos.

El análisis masivo no solo mejora la calidad de los datos, también impacta directamente el rol de los profesionales involucrados. Al liberar tiempo de tareas operativas repetitivas, los equipos técnicos pueden enfocarse en diseñar modelos analíticos, formular hipótesis territoriales o evaluar escenarios futuros. En lugar de ejecutar tareas mecánicas, el ingeniero puede realmente “ingeniar”.

Durante la evaluación de normativas urbanísticas en Bogotá, se identificaron predios con formas irregulares que no coincidían con la normativa asignada. Este hallazgo, difícil de detectar con métodos convencionales, evidencia el potencial del procesamiento masivo para mejorar la toma de decisiones, incluso en aspectos normativos.

7.3. Más allá del acceso: el reto está en el manejo

El acceso a datos ya no es el principal obstáculo; por el contrario, el reto radica en su adecuada gestión. Imágenes satelitales gratuitas, datos climáticos, normativas urbanísticas y registros levantados en campo están disponibles, pero frecuentemente se pierden entre archivos no estructurados o inaccesibles. Cuando la información no es interoperable ni reutilizable, pierde valor. Es aquí donde el Big Geospatial Data aporta: no solo en volumen, sino en estructuración, estandarización y análisis.

Durante la implementación del modelo, se evidenció que dividir los procesos de actualización en bloques operativos más pequeños permite obtener resultados parciales valiosos sin necesidad de esperar a la culminación del flujo completo. Este enfoque de procesamiento progresivo mejora la eficiencia, reduce errores acumulativos y optimiza la asignación de recursos.

Además, la experiencia en este estudio confirma que la limpieza, validación y estructuración de los datos es una fase crítica. La presencia de inconsistencias en etapas tempranas puede contaminar todo el análisis, especialmente cuando se aplican modelos de inteligencia artificial. La fragmentación en subconjuntos de análisis permite detectar errores más fácilmente y aumentar la precisión del procesamiento.

7.4.Limitaciones y proyecciones del modelo propuesto

Es importante señalar que el modelo desarrollado en esta investigación se concentró en la etapa de geoprocesamiento de datos geoespaciales tipo vector, y no abarca la totalidad del flujo de datos del catastro. No contempla, por ejemplo, la implementación de una base de datos espacial transaccional a nivel nacional ni la sistematización completa del intercambio interinstitucional.

Sin embargo, se plantea como una base técnica válida para avanzar hacia flujos de datos más robustos, reproducibles y adaptables dentro del marco del Catastro Multipropósito. La evidencia obtenida sugiere que herramientas como Apache Sedona, en combinación con Apache Spark y otras soluciones escalables, pueden aportar significativamente a la construcción de un sistema más inteligente y sostenible.

No obstante, la aplicación de tecnologías como Apache Sedona requiere más que software: demanda conocimiento técnico en infraestructura distribuida, arquitectura de datos y análisis espacial. Esta brecha se convierte en una limitación recurrente. Para que los beneficios del modelo propuesto sean sostenibles, es imprescindible fortalecer las

capacidades institucionales mediante formación continua dirigida a técnicos, estudiantes y funcionarios públicos.

También se identificó la necesidad de arquitecturas más completas, como los datalakehouses, que centralicen la ingesta, almacenamiento y análisis de datos espaciales. Esta estrategia no solo facilita la escalabilidad del sistema, sino que también mejora la trazabilidad y articulación entre entidades.

7.5.Consideraciones éticas y de gobernanza de datos.

La incorporación de herramientas de Big Data también plantea cuestionamientos sobre el uso responsable de la tecnología. Lo más grande no siempre es lo mejor, especialmente cuando se carece de criterios sólidos para la toma de decisiones. Malas decisiones automatizadas pueden afectar negativamente a comunidades vulnerables y profundizar brechas sociales.

Por ello, se destaca la necesidad de una gobernanza de los datos: instituciones articuladas, normas claras y liderazgo técnico que garantice el uso ético y transparente de la información. Sin estos componentes, cualquier esfuerzo tecnológico puede ser contraproducente.

8. CONCLUSIONES

La integración de Big Geospatial Data mediante Apache Sedona en el proceso de actualización catastral permite modernizar el análisis de datos geoespaciales a gran escala. Esta herramienta no reemplaza los SIG tradicionales, sino que los complementa al ofrecer procesamiento distribuido con escalabilidad horizontal, lo que permite añadir nuevos nodos al clúster a medida que aumenta el volumen de datos cuando se requiera identificar, rectificar o actualizar datos en el proceso de actualización catastral.

Se concluyó que tecnologías orientadas al procesamiento por lotes con escalabilidad horizontal optimizan aspectos clave como la infraestructura tecnológica, el uso eficiente de recursos de cómputo y almacenamiento, y la sostenibilidad de la información geoespacial. Estas soluciones representan una alternativa viable y financieramente sostenible, especialmente para operadores que pueden prestar el servicio a municipios pequeños o rurales.

Al desarrollar un proyecto basado en Big Data, es importante trabajar en paralelo con los sistemas actuales para mantener la continuidad operativa mientras se adoptan nuevas arquitecturas. En el caso del catastro multipropósito, que está trabajando hacia un modelo de transferencia, las herramientas de Big Geospatial Data resultan una opción viable para manejar y analizar grandes volúmenes de información, apoyando la integración y el fortalecimiento del proceso.

El uso de tecnologías Big Data en el proceso de actualización catastral permite que las tareas operativas sean más ágiles, especialmente en el geoprocesamiento de grandes volúmenes de datos. Gracias a su capacidad de procesamiento masivo y distribuido, es

posible detectar con mayor rapidez las inconsistencias en la información catastral y mejorar la calidad de los datos disponibles para la toma de decisiones.

Finalmente, el modelo aquí planteado no pretende resolver todos los desafíos del proceso de Actualización Catastral con enfoque multipropósito, pero sí establece una arquitectura funcional que puede escalarse, integrarse y mejorarse mediante nuevas líneas de investigación técnica e institucional.

Tener en cuenta que se están desarrollando aplicaciones de inteligencia artificial en los procesos catastrales, el uso de Big Data es necesario para garantizar la veracidad e integridad de la información. Los *pipelines* de procesamiento que se plasmó en el modelo permiten mantener una canalización de datos estable, de modo que los modelos de inteligencia artificial utilicen información consistente y actualizada.

9. RECOMENDACIONES

Al tratar de implementar un proyecto de Big Data se debe ser flexible con la metodología y las herramientas; esto se debe a que las dos anteriores son recientes y pueden llegar a presentar problemas al implementarlas. Los grandes volúmenes de datos, o Big Data, requieren grandes cambios en el servidor, la infraestructura de almacenamiento y la arquitectura de administración de la información en la mayoría de las empresas.

Resulta pertinente avanzar hacia la implementación de pipelines de procesamiento de datos y arquitecturas tipo datalakehouse como mecanismos para soportar de manera robusta el modelo de transferencia de información orientado a la prestación del servicio público

catastral. En futuras líneas de investigación, se sugiere desarrollar un datalakehouse como repositorio unificado que permita integrar los procesos de ingesta, transformación, almacenamiento y análisis de datos geoespaciales, cómo es su funcionamiento y que ventajas trae consigo al sistema caatastral. Esta estrategia favorecería la escalabilidad del sistema, así como la coordinación entre las distintas etapas del flujo de información, fortaleciendo el modelamiento de los procesos de transferencia e intercambio de información predial entre gestores y operadores catastrales. Esta línea de trabajo se alinea con las necesidades actuales relacionadas con la interoperabilidad y eficiencia en los modelos de intercambio de información entre las entidades responsables del servicio catastral.

10. BIBLIOGRAFÍA

Aach, M., Inanc, E., Sarma, R., et al. (2023). Large scale performance analysis of distributed Deep learning frameworks for convolutional neural networks. Journal of Big Data, 10, 96. https://doi.org/10.1186/s40537-023-00765-w

Ares, E. (2013). Introducción a TOGAF para el desarrollo de Enterprise Architecture [Diapositivas de SlideShare]. https://es.slideshare.net/netmindit/introduccin-a-togaf-para-el-desarrollo-de-enterprise-architecture

Apache Sedona Project. (2025, 1 de abril). Community Office Hour: An open-source cluster computation engine for processing large-scale geospatial data [PDF].

Barik, R.,Lenka, R., Ali, S., Gupta, N., Satpathy, A., Raj, A. (2017). Investigation into the efficacy of geospatial big data visualization tools. doi: 10.1109/CCAA.2017.8229777 Recuperado de: https://www.researchgate.net/publication/316860265_Investigation_into_the_efficacy_of_geospatial_big_data_visualization_tools

Becerra, J., Cotino-Hueso, L., León, I. P., Sánchez-Acevedo, M. E., Torres-Ávila, J., & Velandia-Vega, J. (2018). Derecho y big data. Bogotá: Editorial Universidad Católica de Colombia Disponible en: https://hdl.handle.net/10983/22964b

Cabello, S. (2023). Análisis de los modelos de gobernanza de datos en el sector público: una mirada desde Bogotá, Buenos Aires, Ciudad de México y São Paulo. Recuperado de: https://repositorio.cepal.org/server/api/core/bitstreams/9e5b987c-9168-4f88- 9503-ca05c6d353ce/content

Cabrera, A., Carrillo, J.M., Abad, M., Jaramillo, D., & Romero, F.R. (2015). Diseño y Validación de Arquitecturas de Aplicaciones Empresariales. RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação, 79-91. Recuperado de: https://www.semanticscholar.org/paper/Dise%C3%B1o-y-Validaci%C3%B3n-de-Arquitecturas-de-Cabrera-Carrillo/029ee612f9ccd1155c7d5ca5cd791a9a718f0ba2

Congreso de la República de Colombia. (25 de mayo de 2019). Ley 1955 de 2019, por el cual se expide el Plan Nacional de Desarrollo 2018-2022 “Pacto por Colombia, pacto por la equidad” [Artículo 81]. Recuperado de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=93970

DANE. (4 de febrero de 2020). Decreto 148 de 2020, [Artículo 2.2.2.2.6]. Recuperado de https://www.igac.gov.co/transparencia-y-acceso-a-la-informacion-publica/normograma/decreto-148-de-2020

Earth Data Science. (2019). Coordinate Reference System and Spatial Projection. Recuperado de https://www.earthdatascience.org

Hernández León, J. A., & Lancheros Aguilar, D. L. (2021). Análisis de alternativas de diseño de una plataforma digital para compartir eventos culturales. Universidad Distrital Francisco José de Caldas. Recuperado de https://repository.udistrital.edu.co/server/api/core/bitstreams/7e966fd1-ddf0-4ff8-91cc-4a86bda9cd9a/content

Huang, Z., Chen, Y., Wan, L., & Peng, X. (2017). GeoSpark SQL: An Effective Framework Enabling Spatial Queries on Spark. International Journal of Geo-Information, 6(285). doi:10.3390/IJGI6090285 Recuperado de: https://www.mdpi.com/2220-9964/6/9/285

García. J, López, J. M. M., De Jesús, A. B., Guisado, M. Á. P., Bustamante, Á. L., & R, W. P. (2018). Ciencia de datos: técnicas analíticas y aprendizaje estadístico. Un enfoque práctico.

Gousios, G. (2018). Big data software analytics with Apache Spark. Proceedings of the 15th International Conference on Mining Software Repositories, 7. Recuperado de: https://doi.org/10.1145/3183440.3183458

IGAC. (11 de junio de 2024). Resolución 746 de 2024, por medio de la cual se modifica parcialmente la Resolución N.° 1040 del 08 de agosto de 2023, única de la gestión

catastral multipropósito [Artículo 6 - 32]. Recuperado de https://www.igac.gov.co/transparencia-y-acceso-a-la-informacion-publica/normograma/resolucion-746-de-2024

IGAC. (18 de octubre de 2023). Procedimiento de formación y/o actualización catastral con enfoque multipropósito (PC-GCT-01, versión 1). Recuperado de https://www.igac.gov.co/sites/default/files/listadomaestro/PC-GCT-01%20V1%20Formacion%20y%20Actualizacion%20Catastral%20CEM.pdf

IGAC. (8 de agosto de 2023). Resolución 1040 de 2023, por medio de la cual se expide la Resolución Única de la Gestión Catastral Multipropósito [Artículo 4.2.1 - 4.2.3 - 4.2.6 - 4.2.12] Anexos: Anexo 1. Glosario de términos para la gestión catastral multipropósito. Recuperado de https://antiguo.igac.gov.co/es/contenido/resolucion-1040-de-2023

IGAC. (21 de febrero de 2023). Resolución 267 de 2023, por la cual se adopta la guía para la elaboración del plan de calidad para el proceso de formación catastral y/o actualización de la formación catastral con enfoque multipropósito. Anexo Guía Plan Calidad Actualización Catastral. *Instituto Geográfico Agustín Codazzi.* Recuperado de https://antiguo.igac.gov.co/es/contenido/resolucion-267-de-2023

IGAC (2023). Actualización catastral - Procesos del Catastro Multipropósito. Recuperado de https://igac.gov.co/catastro-multiproposito/actualizacion-catastral

IGAC (2023). LADM-COL. IGAC. https://igac.gov.co/catastro-multiproposito/ladmcol

IGAC (04 de marzo de 2025). Resolución 301 de 2025. https://www.igac.gov.co/transparencia-y-acceso-a-la-informacion-publica/normograma/resolucion-301-de-2025

IGAC (2025). Instructivo para la estructuración de la información catastral. Modelo de aplicación LADM_COL SINIC V1.0. IGAC. https://www.igac.gov.co/sites/default/files/transparencia/normograma/Anexo%202%20Instructivo%20para%20la%20estructuracion%20LADM_COL%20SINIC%20%20V1.0_0.pdf

Lenka, R. K., Barik, R. K., Gupta, N., Ali, S. M., Rath, A., & Dubey, H. (2016). Comparative análisis of SpatialHadoop and GeoSpark for geospatial big data analytics. IEEE. Recuperado de: https://doi.org/10.1109/IC3I.2016.7918013

National Geodetic Survey. (2019). *NGS Updates Guidance on Modernized National Spatial Reference System*. Recuperado de https://www.ngs.noaa.gov/web/news/modernized-nsrs-updatedguidance.shtml

Mete, M. O. (2023). Geospatial big data analytics for sustainable smart cities. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-4/W7, 141±146. Recuperado de https://isprs-archives.copernicus.org/articles/XLVIII-4-W7-2023/141/2023/isprs-archives-XLVIII-4-W7-2023-141-2023.pdf

Murazzo, M. A., Guevara, M. J., Tello, M., Rodríguez, N. R., Piccoli, F., & Giménez, M. (2018). Orquestación de servicios para el desarrollo de aplicaciones para big data [Objeto de conferencia, VI Jornadas de Cloud Computing & Big Data]. Universidad Nacional de La Plata. https://sedici.unlp.edu.ar/handle/10915/69999

PyData. (2023). Jia Yu - How Apache Sedona is Revolutionizing Geospatial Data Analysis. YouTube. https://www.youtube.com/watch?v=AtgjdK2Gglg&list=LL&index=3.

Revista Geodata. (2022). El catastro multipropósito y la ciencia de datos: un matrimonio necesario. Edición 2. Recuperado de https://revistageodata.icde.gov.co/edicion-2/el-catastromultiproposito-y-la-ciencia-de-datos-un-matrimonio-necesario

Saboya Ríos, N., Loaiza Jara, O. L., & Lévano Rodríguez, D. (2018). Diseño de un modelo de arquitectura empresarial para publicaciones científicas basado en ADM - TOGAF 9.0. Recuperado de https://apuntesuniversitarios.upeu.edu.pe/index.php/revapuntes/article/view/123

SDP (2020). Fortalecimiento institucional de la Unidad Administrativa Especial de Catastro Distrital - UAECD Bogotá. Gobierno Territorial. Recuperado de: https://www.sdp.gov.co/sites/default/files/226_uaecd_2020110010172_7840_0.pdf

Shin, H., Lee, K., & Kwon, H. (2022). A comparative experimental study of distributed storage engines for big spatial data processing using GeoSpark. The Journal of Supercomputing, 78(5), 2556-2579. https://doi.org/10.1007/s11227-021-03946-7. Recuperado de: https://link.springer.com/content/pdf/10.1007/s11227-021-03946-7.pdf

Tang, S., He, B., Yu, C., Li, Y., & Li, K. (2018). A Survey on Spark Ecosystem for Big Data Processing. arXiv: Distributed, Parallel, and Cluster Computing. https://doi.org/10.1109/TKDE.2020.2975652

The Apache Software Foundation. (2025). Spatial SQL app. Apache Sedona. Recuperado de: https://sedona.apache.org/latest/tutorial/sql/

Tokaj, P., Yu, J., & Sarwat, M. (2025). Cloud Native Geospatial Analytics with Apache Sedona: A hands-on guide for working with large-scale spatial data (Early Release ed.).

O'Reilly Media. https://resources.wherobots.com/hubfs/OReilly-Cloud-Native-Geospatial-Analytics-with-Apache%20Sedona.pdf

Tuladhar, A., Radwan, M. M., Abdel Kader, F., & El-Ruby, S. (2005). Federated data model to improve accessibility of distributed cadastral databases in land administration. Recuperado de https://www.fig.net/resources/proceedings/fig_proceedings/cairo/papers/ts_01/ts01_03_tulad har_etal.pdf

Vasconcelos, F. F. (2024). *Uso de Lakehouses para um ambiente analítico espacial Big Data* (Trabalho de conclusão de curso, Universidade Federal de Alagoas). Universidade Federal de Alagoas.

Wherobots. (2025). Create account. Wherobots Documentation. Recuperado el 20 de marzo de 2025, de https://docs.wherobots.com/latest/get-started/create-account/

Wherobots. (2024). Working with files – Getting started with Wherobots Cloud Part 3. Wherobots. Recuperado de: https://wherobots.com/blog/working-with-files-getting-started-with-wherobots-cloud-sedonadb-part-3/

Wherobots Inc. (2025). Organizations and accounts. Wherobots documentation. Recuperado de: https://docs.wherobots.com/latest/get-started/organization-info/organizations-and-accounts/

Yu, J., & Sarwat, M. (2021). Big Geospatial Data Processing Made Easy: A Working Guide to GeoSpark. En M. Werner & Y.-Y. Chiang (Eds.), Handbook of Big Geospatial Data (pp. 35 - 54). Springer. https://doi.org/10.1007/978-3-030-55462-0_2

Zhang, C., & Li, X. (2022). Land use and land cover mapping in the era of big data. Land, 11(10), 1692. https://doi.org/10.3390/land11101692 Recuperado de: https://www.mdpi.com/1859644

11. ANEXOS

Anexo A

Guía Técnica para la Instalación y Configuración de un Clúster Standalone de Spark con Apache Sedona.

Este apartado describe las instrucciones para configurar un clúster standalone de Apache Spark con Apache Sedona utilizando dos máquinas como nodos (Un nodo Master y un nodo worker). Se abordarán los pasos necesarios para desplegar el entorno, asegurando una configuración detallada para el procesamiento de datos espaciales.

1. Requisitos previos

- Instalar Docker Engine en ambas máquinas.
- Para configurar en Windows, Docker Desktop configurado sobre WSL2.
- Conectividad de red entre los nodos (el manager debe ser accesible en el puerto 2377 para uniones de Swarm).
- Conocimiento de la IP de la interfaz del manager (en este ejemplo 192.168.0.7).
- Imagen oficial apache/sedona:1.7.1 disponible en Docker Hub.

Instalación paso a paso:

1. Instalar Docker .

   1.1. Instalación para máquina con sistema operativo Linux Ubuntu.

   Versión Ubuntu 24.04.1 LTS, una distribución basada en Debian (ID_LIKE=debian). La arquitectura es salida es x86_64, entonces se cuenta con un sistema de 64 bits.

   Instalar paquetes requeridos:

   ```
   sudo apt-get install apt-transport-https ca-
   certificates curl software-properties-common
   ```

   Añadir la clave GPG oficial de Docker:

```
curl -
fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-
key add -
```

Añadir el repositorio de Docker:

```
sudo add-apt-
repository "deb [arch=amd64] https://download.docker.com/linux/ubu
ntu $(lsb_release -cs) stable"
```

Actualizar el índice de paquetes:

```
sudo apt-get update
```

Instalar Docker:

```
sudo apt-get install docker-ce
```

Verificar red Docker

```
docker network ls
```

Verificar Version Docker

```
docker --version
```

Verificar la instalación:

```
sudo systemctl status docker
```

Crear o verificar usuario

```
sudo groupadd docker
sudo usermod -aG docker $USER
$USER
newgrp docker
```

1.2. Instalación para máquina con sistema operativo Windows

1.2.1. Instalar Ubuntu en Microsoft Store.

1.2.2. Instalar Docker Desktop desde la página oficial.

https://www.docker.com/products/docker-desktop/

Al momento de ejecutar el .exe marcar WSL (Windows subssystem for linux)

2. Inicialización del Swarm en el nodo Manager

2.1. En la máquina con sistema Linux que actuará de manager, ejecutar:

```
docker swarm init --advertise-addr 192.168.0.7
```

2.2. Salida esperada

```
Swarm initialized: current node (XXXXXXXXXXXX) is now a manager.
To add a worker to this swarm, run the following command:
    docker swarm join --token SWMTKN-... 192.168.0.7:2377
To add a manager to this swarm, run 'docker swarm join-token manager'...
```

3. Unión de nodos Worker

3.1. Copiar el comando docker swarm join … que mostró el manager.

3.2. En cada nodo maquina con sistema Windows (o Linux) que se quiera unir como worker, se debe abrir el PowerShell o terminal WSL y ejecutar:

```
docker swarm join --token SWMTKN-1-0tnht410cx4luqer6mu3runohoudh3ipmdfygz1s8rx5nkmers-anvlnzq1cufcfbzeewws4xgin 192.168.0.7:2377
```

3.3. Verificación:

```
This node joined a swarm as a worker.
```

3.4. Verificar el estado de los nodos

```
docker node ls
```

Observar al menos dos entradas:

- Una con MANAGER STATUS Leader
- Otra con ROLE worker

4. Crear la red overlay

Para que los contenedores se comuniquen a través de todos los nodos:

```
docker network create -d overlay sedona-net
```

5. Descargar la imagen de Apache Sedona en los dos nodos.

```
docker pull apache/sedona:1.7.1
```

6. Desplegar el servicio spark-master

   En el **manager** crea el servicio con los parámetros adecuados de recursos y arranque:

```
        docker service create \
          --name spark-master \
          --constraint 'node.role==manager' \
          --hostname spark-master \
          --network sedona-net \
          --publish 8080:8080 \     # Spark Master UI
          --publish 8888:8888 \     # JupyterLab UI
          --publish 4040:4040 \     # Spark Application UI
          --publish 8085:8085 \     # Sedona Web UI (si aplica)
          --env SPARK_MASTER_HOST=spark-master \
          --env START_MASTER=true \
          --env DRIVER_MEM=2g \
          --env EXECUTOR_MEM=3g \
apache/sedona:1.7.1
```

   - --constraint 'node.role==manager' garantiza que sólo el manager ejecute esta tarea.
   - START_MASTER=true arranca el proceso de master.
   - Ajusta DRIVER_MEM y EXECUTOR_MEM según la memoria disponible (en este ejemplo 2 GB y 3 GB).

7. (Opcional) Actualizar configuración de spark-master

Si se necesita añadir variables o cambiar recursos tras la creación:

```
docker service update spark-master \
  --env-add NEW_ENV_VAR=value \
--limit-memory 4g
```

8. Desplegar el servicio spark-worker

   En el manager, registra un servicio que lanzará procesos workers sólo en nodos con rol worker:

```
docker service create \
  --name spark-worker \
  --constraint 'node.role==worker' \
  --network sedona-net \
  --entrypoint "/opt/spark/sbin/start-worker.sh" \
  apache/sedona:1.7.1 \
    spark://spark-master:7077 \
    --cores 6 \
--memory 2g
```

   - --cores y --memory asignan recursos al worker.
   - El entrypoint inicia el proceso de worker apuntando al master.

9. Verificar servicios y tareas.

```
docker service ls
docker service ps spark-master
docker service ps spark-worker
```

10. Acceso a las interfaces.

    - Spark Master UI: http://192.168.0.7:8080
    - JupyterLab: http://192.168.0.7:8888
    - Spark App UI: http://192.168.0.7:4040
    - (Opcional) Sedona UI: según tu configuración en el contenedor.

Anexo C

Wherobots Cloud: Aplicación práctica con Apache Sedona.

La plataforma de infraestructura de datos geoespaciales Wherobots Cloud está diseñada para permitir que las personas y los equipos procesen, analicen y obtengan información de los datos espaciales a gran escala, simplificando el desarrollo de soluciones geoespaciales. (Wherobots Inc, 2025)

El esquema presentado en la figura 0-1 ilustra el flujo completo de registro, autenticación y utilización de la plataforma Wherobots para la realización de análisis de geoprocesamiento.

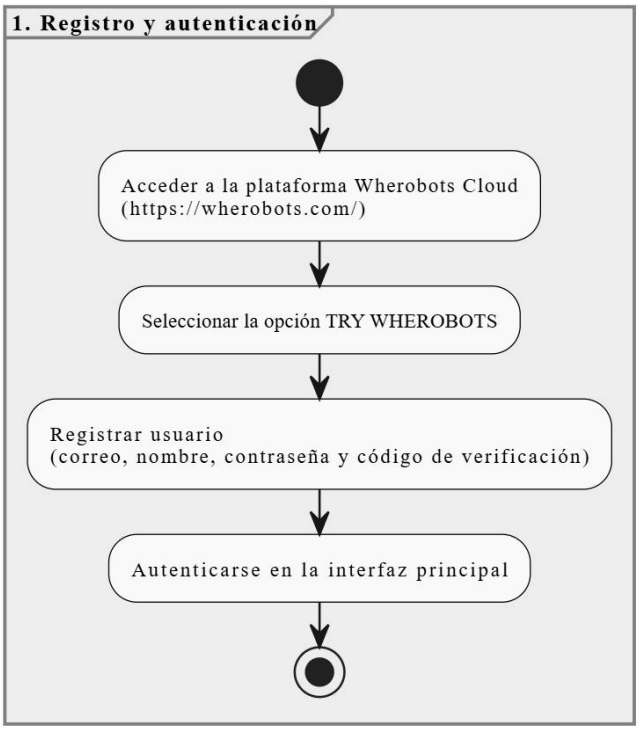


Figura A. Flujo de uso de Wherobots Cloud. 1. Registro y autenticación. Elaboración propia.

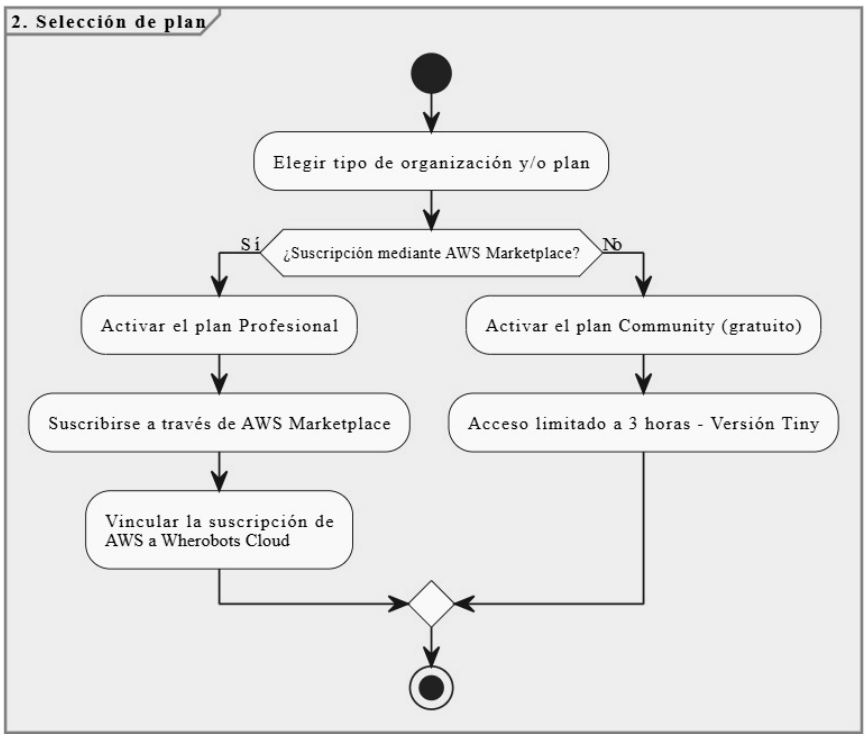


Figura B. Flujo de uso de Wherobots Cloud. 2. Selección de plan suscripción. Elaboración Propia.

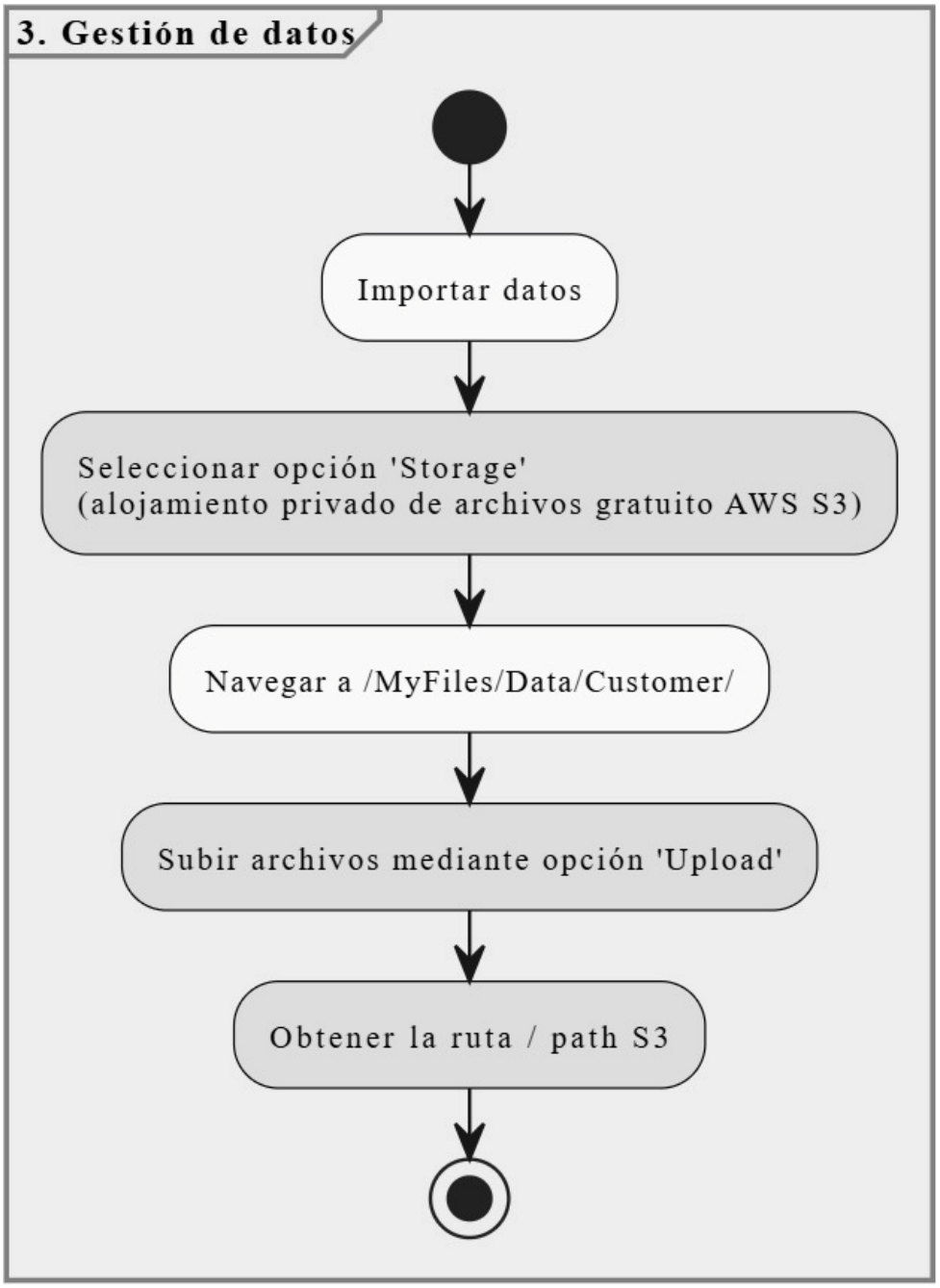


Figura C. Flujo de uso de Wherobots Cloud. 3. Gestión de datos. Elaboración propia.

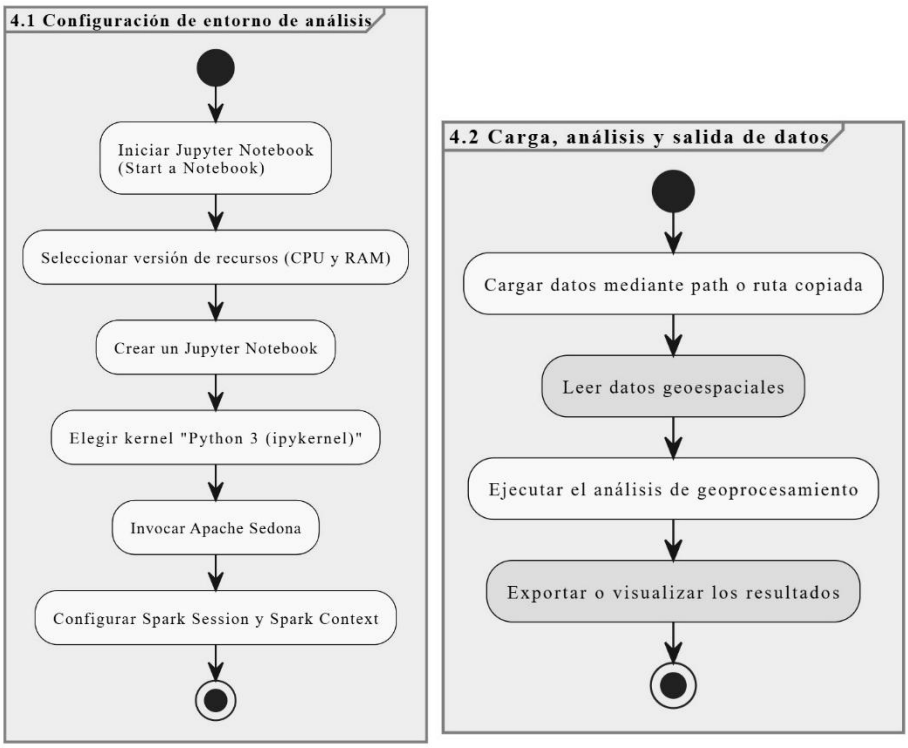


Figura D. Flujo de uso de Wherobots Cloud. 4. Geoprocesamiento de datos a gran escala. A la izquierda 4.1. Configuración de entorno de análisis y a la derecha 4.2. Carga, análisis y salida de datos.

A continuación, se presentará una breve explicación sobre cómo registrarse en Wherobots. Este manual documentará únicamente el proceso de registro para los tipos de organización '**Community'** o '**Professional'** y cómo hacer uso de JupyterLab para aplicar Apache Sedona con Python en un ejemplo práctico.

1. **¿Cómo suscribirse a Wherobots?**

   Para suscribirse a Wherobots, se debe ingresar a la página principal, seleccionar la opción "TRY WHEROBOTS", completar el formulario de registro y verificar el correo electrónico para activar la cuenta. Wherobots ofrece un plan gratuito denominado Community, que permite el acceso a la versión Tiny de Jupyter Notebook. Para acceder al plan profesional, es necesario contar con una cuenta en AWS, suscribirse al producto "Wherobots Spatial Intelligence Cloud Professional Edition" en AWS Marketplace y vincular dicha suscripción con la organización en Wherobots.

2. Trabajar con Wherobots y JupyterLab.

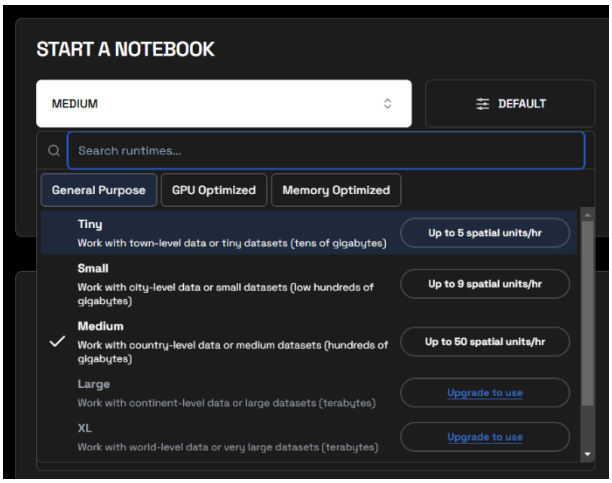
Una vez activada y vinculada la suscripción de Amazon con Wherobots, se procede a iniciar un notebook. Es importante destacar que, aunque no se tenga una suscripción vinculada a Amazon, siempre se puede acceder a la versión Tiny, que es gratuita y está disponible de forma permanente.

a. La versión Tiny es la opción gratuita que siempre estará habilitada.

b. El plan profesional gratuito por un mes de $400 permite acceder a los notebooks Small y Medium, ofreciendo mayor capacidad de procesamiento y recursos adicionales durante ese período.

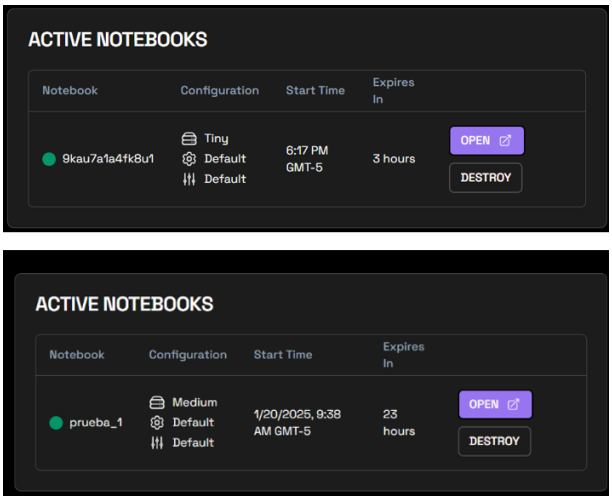
2.1. Iniciar Notebook.

a. Se elige el tipo de notebook según el plan y se procede a dar clic en el botón Start.



Imágen B-1. Iniciar Notebook en Wherobots.

Teniendo en cuenta que se realizó la suscripción al plan profesional, se hará uso de la versión del notebook Medium.



Imágen B-2. Comparación Notebooks Tiny - Medium.

Nota: La duración de la sesión del notebook varía según la versión. Los notebooks tipo Medium tienen una expiración de 23 horas, mientras que los de tipo Tiny expiran aproximadamente a las 3 horas.

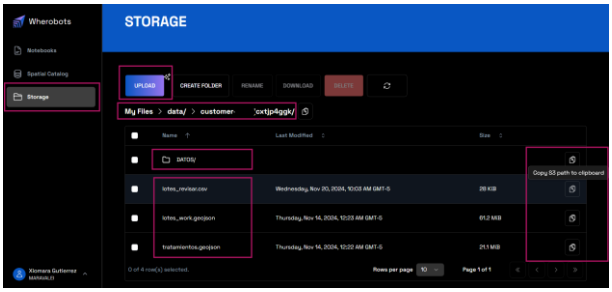
2.2. Importar datos a través de AWS S3.

Un flujo de trabajo común con Wherobots Cloud consiste en cargar varios archivos, realizar análisis geoespaciales y guardar los resultados como parte de un flujo de datos más amplio.

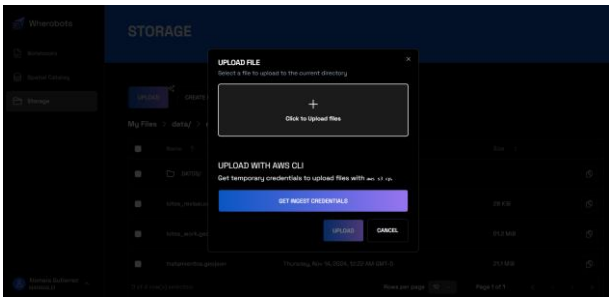
WherobotsDB permite importar datos de prácticamente cualquier formato de datos espaciales, incluyendo archivos planos como CSV y TXT, Shapefiles, GeoJSON, GeoParquet y bases de datos PostGIS. La cuenta de Wherobots Cloud incluye alojamiento privado de archivos gratuito a través de AWS S3.

a. Subir y administrar archivos a través de la pestaña **Storage** de Wherobots Cloud.

Los datos se importan ingresando a la carpeta **Data**, luego haciendo clic en la carpeta **Customer**, y finalmente clic en **Upload** para subir los archivos de manera privada en AWS S3.



Imágen B-3. Interfaz sección Storage – Wherobots.

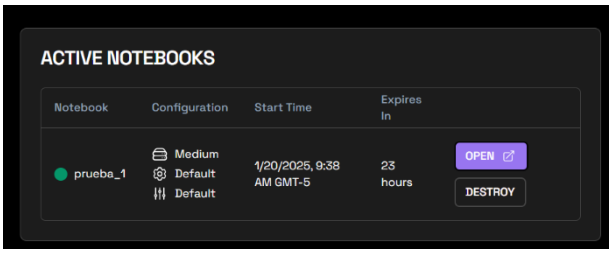


Imágen B-4. Importar datos mediante opción Upload.

Nota: Para cargar esos datos en Jupyter, simplemente copie la ruta dando clic en el archivo y péguela en su código de JupyterLab.

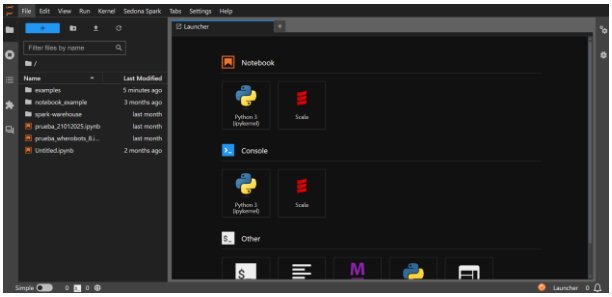
3. Aplicación práctica con Apache Sedona.

Una vez iniciado el entorno de notebook, haga clic en Open para iniciar JupyterLab.



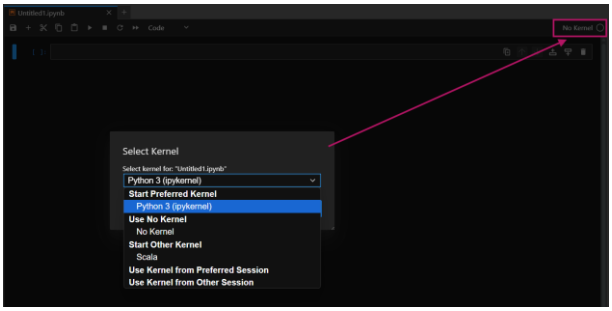
Imágen B-5. Activar y abrir Notebook - Wherobots

El entorno de notebook de Jupyter será la interfaz principal para trabajar con WherobotsDB mediante Python o Scala.



Imágen B-6. Interfaz JupyterLab - Wherobots

a. Inicio del notebook: Selecciona "Python 3 (ipykernel)" en el entorno de JupyterLab. Esto creará un nuevo notebook extensión ".ipynb". Luego, en la esquina superior derecha, seleccionar el kernel de Python 3 para garantizar la compatibilidad con el código.



Imágen B-7. Seleccionar kernel - Python 3.

b. Crear sesión de Spark e inicializar SparkContext: Se establece la conexión con el clúster Spark.

   Si queremos trabajar con pyspark, es necesaio crea una SparkSession, ya que permite ejecutar y gestiona las tareas de Spark.

   Sedonacontext es el punto de entrada para trabajar con SRDD, estos son una estructura de datos básicas en Apache Spark que nos permite manejar Geo Big data.

```
import os
from sedona.spark import *
from sedona.core.formatMapper.shapefileParser import ShapefileReader
from sedona.utils.adapter import Adapter
from sedona.spark import *

config = SedonaContext.builder().\
        appName('prueba_1').\
        getOrCreate()

sedona = SedonaContext.create(config)

sc = sedona.sparkContext

sc.setSystemProperty("sedona.global.charset", "utf8")
```

Imágen B-8. Creación SedonaContext.

c. Lectura de archivos:

   i. Se cargan los datos geoespaciales copiando el path de los archivos que se importaron de manera privada en AWS S3.

```
# Ruta archivos GeoJSON y csv
lotes_path = "s3://wbts-wbc-o2l64ln0km/n2tr4bwgwm/data/customer-b5kdjcxtjp4ggk/DATOS/lotes.geojson"

lotes_revisar_path = "s3://wbts-wbc-o2l64ln0km/n2tr4bwgwm/data/customer-b5kdjcxtjp4ggk/DATOS/lotes_revisar.csv"

tratamiento_path = "s3://wbts-wbc-o2l64ln0km/n2tr4bwgwm/data/customer-b5kdjcxtjp4ggk/DATOS/tratamientos_urbanisticos.geojson"
```

Imágen B- 9. Ejemplo ruta archivo importado Storage AWS S3.

   ii. Se lee el tipo de archivo y se crea una vista temporal.

```
#Codigo para poder leer un CSV
lotes_revisar = sedona.read.format('csv').\
        option('header','true').\
        option('delimiter', ',').\
        option('interSchema', 'true').\
        load(lotes_revisar_path)

lotes_revisar.createOrReplaceTempView("lotes_revisar")

#=============================================================

#Codigo para poder leer Cargar geojson

lotes = sedona.read.format("geojson").\
        option("multiline", "true").\
        option("geometry", "geometry"). \
        load(lotes_path).\
        selectExpr("explode(features) as features").\
        select("features.*")

lotes.createOrReplaceTempView("lotes")
```

Imágen B-10. Ejemplo modo lectura archivo csv y geojson.

iii. Se verifica el esquema de los datos importados, con la función .printSchema()

```
#####################################################
Esquema lotes_revisar
root
 |-- lotlote_id: string (nullable = true)

#####################################################
Esquema lotes
root
 |-- geometry: geometry (nullable = true)
 |-- properties: struct (nullable = true)
 |    |-- LOTCODIGO: string (nullable = true)
 |    |-- LOTDISPERS: string (nullable = true)
 |    |-- LOTDISTRIT: long (nullable = true)
 |    |-- LOTILDISPE: string (nullable = true)
 |    |-- LOTUPREDIA: long (nullable = true)
 |    |-- MANZCODIGO: string (nullable = true)
 |    |-- SHAPE_Area: double (nullable = true)
 |    |-- SHAPE_Leng: double (nullable = true)
 |-- type: string (nullable = true)
```

```
################################################
Esquema tratamientos
root
 |-- geometry: geometry (nullable = true)
 |-- properties: struct (nullable = true)
 |    |-- ACTO_ADMIN: string (nullable = true)
 |    |-- ALTURA_MAX: string (nullable = true)
 |    |-- CODIGO_ID: string (nullable = true)
 |    |-- CODIGO_TRA: string (nullable = true)
 |    |-- CÓDIGO_SU: string (nullable = true)
 |    |-- ESCALA_CAP: string (nullable = true)
 |    |-- FECHA_ACTO: string (nullable = true)
 |    |-- FECHA_CAPT: string (nullable = true)
 |    |-- NOMBRE_TRA: string (nullable = true)
 |    |-- NORMATIVA: string (nullable = true)
 |    |-- NUMERO_ACT: string (nullable = true)
 |    |-- OBSERVACIO: string (nullable = true)
 |    |-- RESPONSABL: string (nullable = true)
 |    |-- SHAPE_Area: double (nullable = true)
 |    |-- SHAPE_Leng: double (nullable = true)
 |    |-- TIPOLOGIA: string (nullable = true)
 |-- type: string (nullable = true)
```

Imágen B-11. Ejemplo vista esquema.

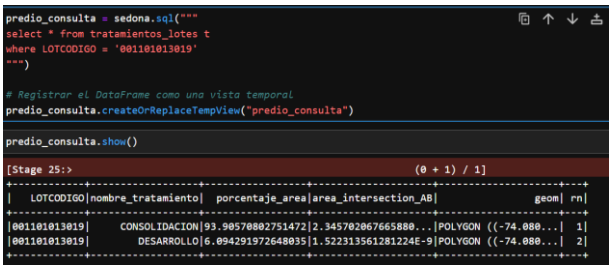
d. Ejecución de consultas: Se realizan consultas geoespaciales utilizando Apache Sedona y su extensión sql.

El propósito de la consulta de ejemplo, implementada utilizando Apache Sedona y WherobotsDB, es identificar los tratamientos urbanísticos correspondientes a todos los lotes de la ciudad de Bogotá y observar si su ejecución es exitosa.

i. Realizar consulta

```
bogota = sedona.sql("""
SELECT p.properties.LOTCODIGO, p.geometry
FROM lotes p
""")
# Registrar el DataFrame como una vista temporal
bogota.createOrReplaceTempView("bogota")

tratamientos_lotes = sedona.sql("""
WITH intersecciones AS (
    SELECT
        a.LOTCODIGO,
        COALESCE(b.properties.NOMBRE_TRA, 'NA') AS nombre_tratamiento,
        ST_Area(ST_Intersection(a.geometry, b.geometry)) / ST_Area(a.geometry) * 100 AS porcentaje_area,
        ST_Area(ST_Intersection(a.geometry, b.geometry)) AS area_interseccion_AB,
        ST_Intersection(a.geometry, b.geometry) AS geom
    FROM
        bogota a
    JOIN
        tratamientos b
    ON
        ST_Intersects(a.geometry, b.geometry)
)
SELECT
    *,
    ROW_NUMBER() OVER (PARTITION BY LOTCODIGO ORDER BY area_interseccion_AB DESC) AS rn
FROM
    intersecciones
""")
```



```
predio_consulta = sedona.sql("""
select * from tratamientos_lotes t
where LOTCODIGO = '001101013019'
""")

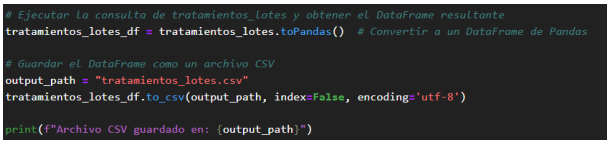
# Registrar el DataFrame como una vista temporal
predio_consulta.createOrReplaceTempView("predio_consulta")

predio_consulta.show()
```

```
[Stage 25:>                                                 (0 + 1) / 1]
+-----------+-----------------+-----------------+--------------------+--------------------+---+
|  LOTCODIGO|nombre_tratamiento|  porcentaje_area|  area_intersection_AB|                geom| rn|
+-----------+-----------------+-----------------+--------------------+--------------------+---+
|001101013019|     CONSOLIDACION|93.90570802751472|2.345702067665880...|POLYGON ((-74.080...|  1|
|001101013019|       DESARROLLO|6.094291972648035|1.522313561281224E-9|POLYGON ((-74.080...|  2|
+-----------+-----------------+-----------------+--------------------+--------------------+---+
```

Imágen B-13. Ejemplo consulta API SQL – Apache Sedona.

ii. Exportar consulta.



```
# Ejecutar la consulta de tratamientos_lotes y obtener el DataFrame resultante
tratamientos_lotes_df = tratamientos_lotes.toPandas()  # Convertir a un DataFrame de Pandas

# Guardar el DataFrame como un archivo CSV
output_path = "tratamientos_lotes.csv"
tratamientos_lotes_df.to_csv(output_path, index=False, encoding='utf-8')

print(f"Archivo CSV guardado en: {output_path}")
```

Imágen B-14. Ejemplo exportar consulta a archivo csv.

Nota: Para la visualización de datos geoespaciales puede realizarse mediante diversas herramientas. Se sugiere considerar Kepler.gl debido a su óptima integración con WherobotsDB, aunque otras opciones pueden ser igualmente válidas según las necesidades del usuario.