DOCUMENTACIÓN – METODOLOGÍA, PROCESOS Y ARQUITECTURA DE DATOS

Sistema de Monitoreo y Detección de Anomalías en Transacciones Inmobiliarias (2015–2025)

Equipo:

William Pabon - Científico de datos.

Nilson de la Ossa - Analista de datos.

Mariana Jaramillo - Desarrolladora .

Kevin Vargas - Ingeniero de datos.

Plataforma: Databricks Lakehouse Platform

---

1. Introducción

Colombia cuenta con uno de los repositorios más grandes de transacciones inmobiliarias de América Latina, con más de 34 millones de registros provenientes de 1.105 municipios. Esta información es crítica para el país, pues soporta procesos como:

- Planeación territorial
- Análisis del mercado inmobiliario
- Supervisión operativa
- Evaluación de riesgos patrimoniales y financieros
- Identificación de fraude documental y jurídico

Sin embargo, la calidad heterogénea, errores de captura y el enorme volumen representan un reto para cualquier sistema tradicional.

Nuestro equipo diseñó e implementó un sistema automatizado que integra, depura, estandariza, monitorea y analiza anomalías, usando Databricks, PySpark, Machine Learning, reglas jurídicas de la Superintendencia de Notariado y Registro (SNR), y fuentes complementarias como los códigos municiapales que provee el DANE en el portal de datos abiertos.

1.1 Propósito del documento

Describir la arquitectura, procesos, metodología y lineamientos técnicos utilizados para integrar, estandarizar, analizar y detectar anomalías en el dataset nacional de transacciones inmobiliarias

de Colombia (más de 30 millones de registros).

1.2 Alcance

Incluye:

- Arquitectura de datos implementada en el proyecto.
- Procesos de ingesta, transformación y estandarización.
- Reglas de calidad y validación.
- Modelos de detección de anomalías.
- Tablero de monitoreo.
- Lineamientos de operación y mantenimiento del sistema.

1.3 Objetivo General

Construir un sistema automático de monitoreo, análisis y detección de anomalías en la dinámica inmobiliaria del país, permitiendo identificar de forma temprana:

- Riesgos operativos
- Riesgos financieros
- Inconsistencias territoriales
- Indicadores de posible fraude
- Problemas de calidad de datos
- Riesgo de lavado de activos o evasión de impuestos
- Errores de anotación, duplicidad y secuencia
- Actos sin soporte o con justificación incompleta

1.3 Objetivos específicos

OE1. Integrar y estandarizar los registros de transacciones inmobiliarias.

Se implementó un proceso de ingesta y homogenización de más de 30 millones de registros usando PySpark.

OE2. Construir modelos de detección de anomalías basados en reglas y machine learning.

Se aplicaron tanto reglas heurísticas basadas en estadística inferencial como modelos de aislamiento estadístico (Isolation Forest).

OE3. Identificar patrones de fraude, valores atípicos, duplicidades y errores de anotación.

Se identificaron como transacciones anómalas aquellas transacciones cuyo valor está muy por debajo o muy por encima de los valores comunes a nivel territorial. También se encontraron transacciones anómalas de acuerdo con la cantidad de participantes en la misma, donde algunas transacciones se dieron entre una cantidad anormal de personas. Adicionalmente, se observaron municipios donde la cantidad de transacciones anómalas está por encima de lo normal.

OE4. Desarrollar un tablero de monitoreo con KPIs de riesgo y calidad.

Se aplicaron tanto reglas jurídicas (SNR) como modelos de aislamiento estadístico (Isolation Forest).

OE5. Integrar fuentes públicas complementarias (DIVIPOLA, ORIPs oficiales, catálogo jurídico).

Se integraron las siguientes fuentes:

Directorio ORIP

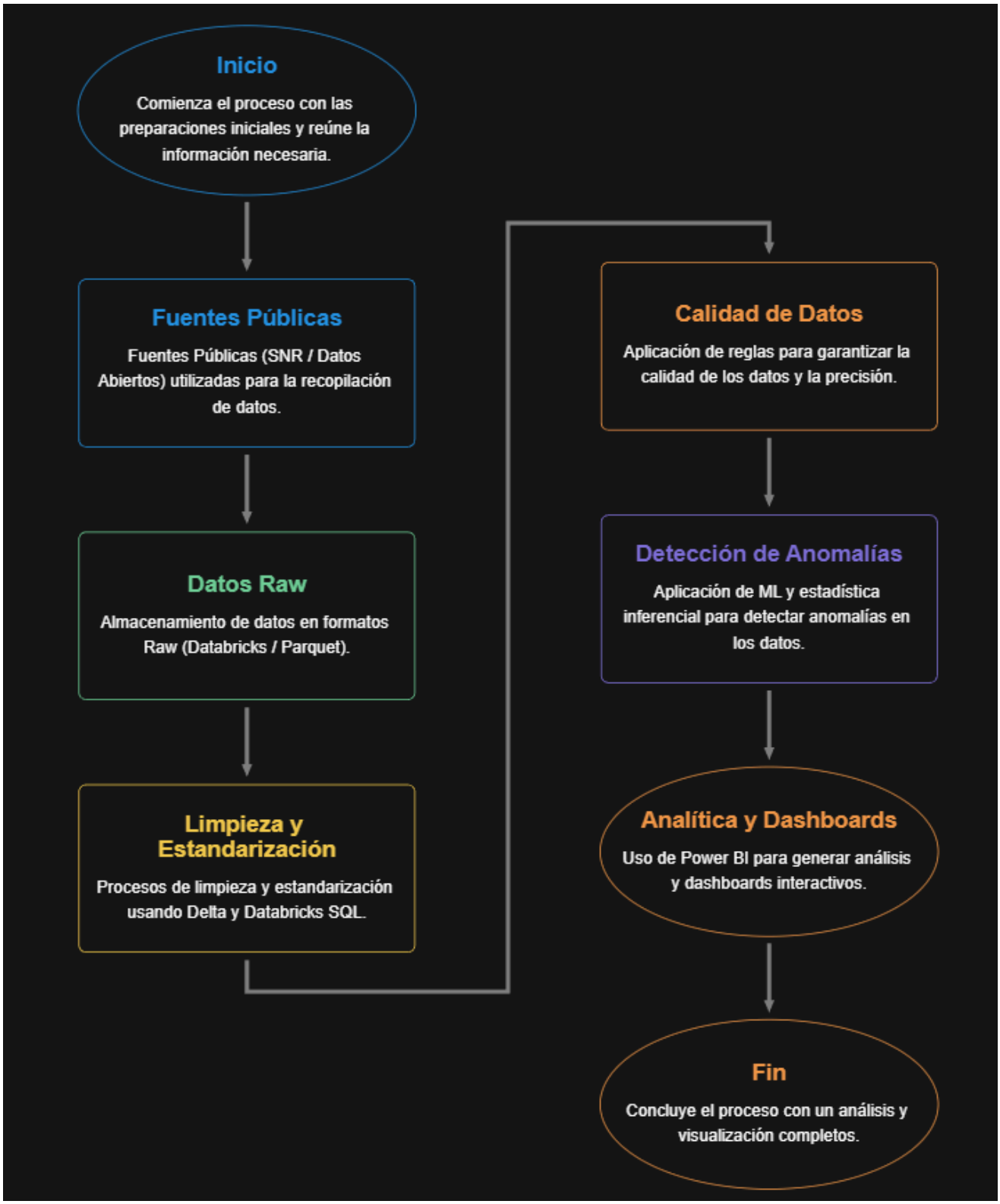
Directorio DIVIPOLA CODIGOS MUNICIPIOS

OE6. Documentar la arquitectura, procesos y metodología. (Este documento)

2. Arquitectura General del Sistema

2.1 Visión General

Representación conceptual del flujo:

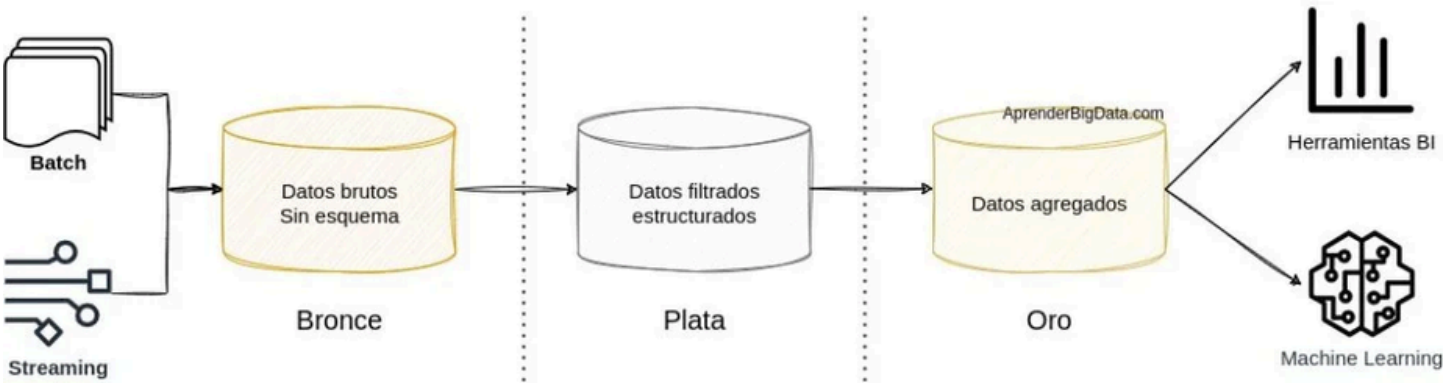


2.2 Componentes Principales

- Databricks Free Lakehouse como entorno unificado
- Parquet/Delta Lake como formato de almacenamiento

- Notebooks para ETL, validación y ML
- PySpark para procesamiento distribuido
- Github para versionamiento
- Power BI

3. Arquitectura de Datos Detallada



3.1 Raw Layer (Bronze) – Ingesta

- Contiene archivos parquet originales sin cambios.
- Dataset cargado directamente desde datos.gov.co.
- Garantiza trazabilidad y reproducibilidad.

3.2 Cleaning Layer (Silver) – Estandarización

Procesos principales:

- Conversión de formatos (fechas, numéricos, strings).
- Normalización de texto (trim, upper/lower, regex).
- Validación de integridad referencial (ORIP ↔ DIVIPOLA ↔ municipio).
- Uso de fuentes de datos abiertos alternativas, como la información de las ORIP y los códigos municipales.

3.3 Quality Layer (Gold) – Reglas y banderas de

calidad

Cada regla produce una columna:

- Cumple_regla (0/1)

3.3.1 Reglas aplicadas.

3.3.1.1 PK.

Regla: Unicidad obligatoria.

Detecciones:

- Duplicados → posible corrupción o reproceso.

3.3.1.2. MATRÍCULA

- Unicidad obligatoria: un predio = una matrícula.
- Validación contra ORIP (primeros 3 dígitos).
- No nula.
- Detección:
  - Matrículas que aparecen en municipios donde no deberían localizarse.
  - Matrículas compartidas por múltiples predios (fraude fuerte).

3.3.1.3. FECHA_RADICA_TEXTO

Validaciones:

- No nula.
- No puede estar en el futuro.
- Múltiples radicaciones el mismo día para el mismo predio → alerta.
- Formato estandarizado a yyyy-MM-dd.
- Si existe, debe existir DOCUMENTO_JUSTIFICATIVO.

Detecta:

- Actos sospechosos registrados el mismo día.
- Manipulación de fechas.
- Registros ilegales sin soporte.

3.3.1.4. FECHA_APERTURA_TEXTO

- Debe existir un único valor por matrícula.
- No puede ser posterior a la radicación.
- No nula.
- No puede estar en el futuro.

Detecta:

- Folios creados después de que ya existían transacciones — dato corrupto.

3.3.1.5. YEAR_RADICA

- Debe coincidir con el año de FECHA_RADICA.
- No puede ser menor que el año de apertura.

3.3.1.6. ORIP

- Validación contra catálogo oficial (198 ORIPs).
- Coherente con matrícula.
- Cambios poco frecuentes.

Detección:

- Jurisdicciones inconsistentes.
- Posible fraude al mover predios entre ORIPs.

3.3.1.7. DIVIPOLA

- Validación con lista oficial.
- No nulo.
- Coherencia territorial: Municipio ↔ Departamento ↔ Matrícula ↔ ORIP.

Detecta:

- Inmuebles ubicados “fuera de lugar” (intento de ocultar información).

3.3.1.8. MUNICIPIO y DEPARTAMENTO

- Ambos deben coincidir con DIVIPOLA.
- No nulos.
- Nombres válidos y estandarizados.

3.3.1.9. NUM_ANOTACION

- Par (MATRÍCULA, NUM_ANOTACIÓN) debe ser único.
- Secuencia creciente.
- Posibles fraudes detectados:
  - Saltos enormes.
  - Regresiones numéricas.
  - Duplicidad.

3.3.1.10. ESTADO_FOLIO

- Folios cerrados no pueden tener anotaciones posteriores.
- Formato estandarizado.

- Solo un estado por folio.

3.3.1.11. FOLIOS_DERIVADOS

- Si existe → debe existir DOCUMENTO_JUSTIFICATIVO.
- Se verifica que los folios derivados existan realmente en el dataset.

3.3.1.12. Dinámica_Inmobiliaria

- No nula.
- 0/1 estandarizado.
- Correspondencia con el tipo de acto jurídico.

3.3.1.13. COD_NATUJUR / NOMBRE_NATUJUR

- Validación con catálogo SNR.
- No nulos.
- Limpieza de texto, estandarización.

3.3.1.14. NUMERO_CATASTRAL / NUMERO_CATASTRAL_ANTIGUO

- Columna venía completamente nula.
- Se reporta como inconsistencia oficial.

3.3.1.15. DOCUMENTO_JUSTIFICATIVO

- No nulo cuando se requiere.
- Formato estandarizado.
- Duplicados revisados.

3.3.1.16. COUNT_A / COUNT_DE

- Valores numéricos válidos.
- Posibles valores extremos.

3.3.1.17. PREDIOS_NUEVOS

- Solo un "1" por matrícula.
- Coherencia con NUM_ANOTACION.

3.3.1.18. TIENE_VALOR / TIENE_MAS_DE_UN_VALOR / VALOR

Validaciones:

- 0/1 estandarizado.
- Si TIENE_VALOR = 1 → VALOR no nulo.
- Valores extremos o improbables.

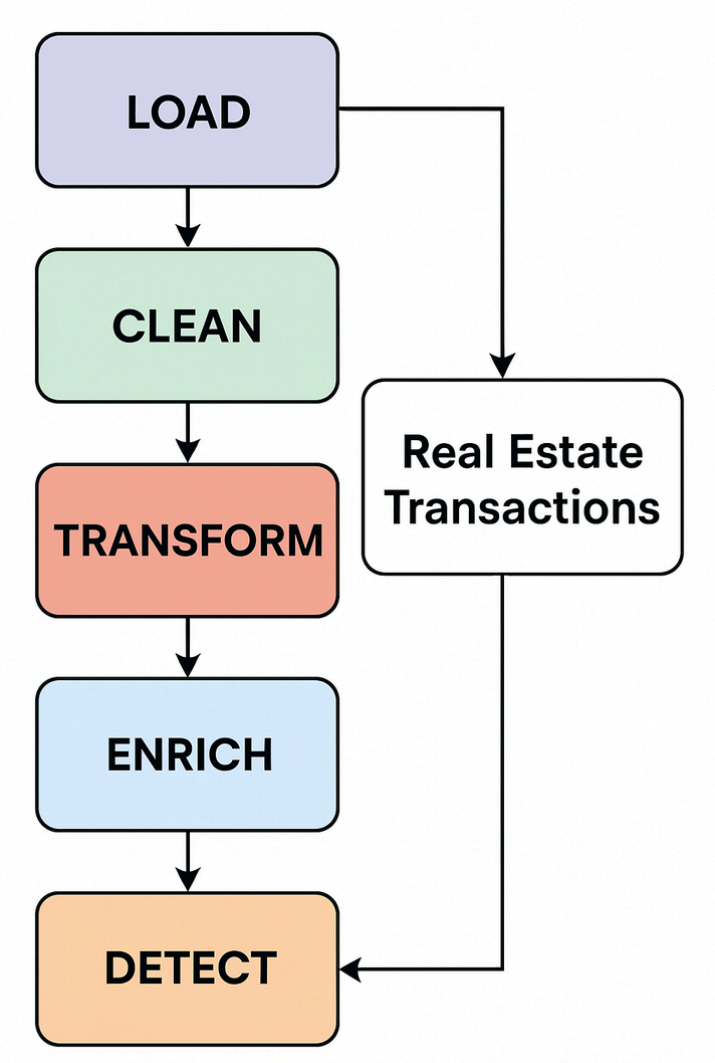
- Coherencia con tipo de acto jurídico.

3.4 Anomaly Detection Layer

Incluye:

- Estadística descriptiva e inferencial
- Reglas heurísticas de negocio basadas en estadística
- Machine Learning (Isolation Forest)
- Indicadores agregados por municipio/ORIP y tipo de anomalía.

4. Metodología de Procesamiento



4.1 Ingesta

- Uso de Databricks para almacenamiento de información en datalake.
- Uso de notebooks en Databricks para lectura de parquets.
- Validación inicial de esquema: tipos, nulos, longitud de columnas.

- Transformación y Limpieza
- Reglas de Identificación
  - Reglas Geográficas
  - Reglas de Tiempo
  - Reglas de Operación

5. Modelos y Métodos Aplicados

5.1 Reglas Estadísticas

Identificación de:

- Outliers por valor a nivel municipal.
- Anomalías de secuencia.
- Registros duplicados.
- Tiempos anómalos.

5.2 Reglas heurísticas

- Reglas heurísticas basadas en estadística inferencial teniendo en cuenta los valores de las transacciones a nivel municipal.

5.3 Machine Learning

Propuesta:

- Isolation Forest para detectar valores atípicos a nivel estructural en los registros.

6. Pipeline de Procesos en Databricks

6.1 Detalle por Notebook

6.1.1. Notebook - Data Cleaning

El notebook “Data Cleaning” implementa el proceso sistemático de limpieza, validación y estandarización de los más de 30 millones de registros de transacciones inmobiliarias contenidos en la base de datos suministrada, siguiendo las reglas operativas definidas por la Superintendencia de Notariado y Registro (SNR), el IGAC y los estándares técnicos de calidad de datos.

1. Se realiza una visualización exploratoria de las columnas y conteo base, verificando la correcta disponibilidad del dataset.
2. Preparación del entorno de transformación: Esta sección configura el ambiente para procesar datos a escala.

3. Implementación de reglas de calidad (Data Quality Rules) Para cada columna crítica del esquema se implementan reglas de calidad mediante **columnas bandera (cumple_regla)**, lo que permite cuantificar: registros válidos, registros sospechosos, registros anómalos, registros incompletos.
4. Generación de métricas de calidad: Este enfoque facilita exportar un reporte final de auditoría de calidad.
5. Estandarización de formatos: Estas transformaciones permiten preparar el dataset para las fases posteriores de detección de anomalías.
6. Preparación para modelos y reglas analíticas

6.1.2. Notebook - Estandarización y Enriquecimiento de los Datos

Este notebook corresponde a la segunda fase del proceso de preparación y saneamiento de datos dentro del sistema de monitoreo y detección de anomalías inmobiliarias. Su objetivo es transformar, normalizar y enriquecer el dataset, asegurando consistencia semántica y técnica entre las variables clave del modelo.

1. Carga del Dataset Limpio Inicial: El notebook inicia cargando el DataFrame resultante de la fase previa ("Data Cleaning"). Este DataFrame ya cuenta con validaciones básicas aplicadas, por lo que esta fase se enfoca en estandarizar formatos, homogenizar valores, crear columnas derivadas, integrar informacion externa.
2. Estandarización de Campos Textuales: Se ejecuta un conjunto de transformaciones orientadas a garantizar consistencia en las columnas categóricas, especialmente aquellas provenientes de diferentes fuentes o con variabilidad en captura manual.
3. Estandarización y Conversión de Fechas: Estas transformaciones son esenciales para los modelos estadísticos, detección de anomalías temporales y el análisis operativo de acumulados por día/mes.
4. Integración de Catálogos Oficiales (Enriquecimiento)
5. Creación de Indicadores Derivados: Estas banderas alimentan el **motor de reglas del sistema final**.
6. Integración y Validación de Valores Numéricos
7. Preparación para Visualización y Modelos

6.1.3. Notebook - Anomalías en predios

Este notebook realiza un proceso de extracción, limpieza, estandarización y almacenamiento de datos relacionados con transacciones inmobiliarias, con el objetivo de preparar la información para un modelo de detección de anomalías en predios.

1. El notebook inicia recuperando la información de transacciones inmobiliarias desde una fuente centralizada, lo que permite acceder a un volumen amplio y heterogéneo de registros.
2. Se realiza una depuración inicial en la que solo se conservan las variables necesarias para el análisis.
3. El notebook transforma las fechas que vienen en múltiples formatos hacia un estándar comprensible y uniforme.
4. Finalmente, la información ya depurada y estandarizada se almacena en una tabla confiable lista para ser consumida por los modelos de análisis y detección de anomalías.

6.1.3. Notebook - Anomalías en predios

Este notebook realiza la preparación integral de los datos inmobiliarios necesarios para construir un modelo de detección de anomalías. A partir de los registros originales, se seleccionan las variables más relevantes, se normalizan montos y fechas, se corrigen valores faltantes y se transforman múltiples campos al formato adecuado para el análisis.

1. Selección de variables para el modelo.
2. Posteriormente, se transforman los valores asociados a las transacciones para asegurar que puedan ser interpretados y comparados correctamente.
3. El notebook reemplaza ausencias de información por valores neutros que permitan continuar el análisis sin distorsionar los resultados.
4. Se transforman variables textuales y categóricas en formatos numéricos y ordenados, permitiendo que el modelo interprete correctamente factores como la existencia de valor registrado, el tipo de predio o el comportamiento de radicación
5. Las fechas asociadas a las transacciones se analizan para detectar errores cronológicos, abrir espacios para señales de riesgo o anomalías en el proceso registral. Identificar incoherencias en años, secuencias o aperturas permite fortalecer el control operativo y mejorar la calidad del dato.
6. El notebook aplica técnicas como escalamiento y transformaciones logarítmicas para reducir la influencia de valores extremos y facilitar la identificación de transacciones sospechosas.
7. Finalmente, se eliminan columnas que no aportan valor al análisis y se consolida un dataset limpio, estandarizado y optimizado para ser consumido por los modelos de detección de anomalías.

7. Dashboard de Monitoreo

7.1 Dasbhoard transacciones con anomalias.

7.1.1. KPIS Propuestos

- Transacciones anómalas.
- Valor total transado
- Valor promedio por transaccion anómala
- Predios con transacciones anómalas.
- Participantes en transacciones anómalas
- Valor por dinamica inmobiliaria
- Valor por tipo de predio
- Transacciones anómalas por detalle de anomalía
- Informacion de las transacciones
- Transacciones anómalas por fecha de radicación
- Valor total transado por fecha de radicación
- Z-score municipal vs valor de la transacción.

7.2 Dasbhoard transacciones inmobiliarias.

7.2.1. KPIS Propuestos

- Transacciones
- Valor total transado
- Valor promedio por transaccion
- Predios con transacciones
- Total de participantes
- Valor por acto juridico
- Estadisticas por municipio

- Estadisticas por orip
- Estadisticas por departamento
- Valor por dinamica inmobiliaria
- Valor por categoria de ruralidad
- Valor total transado por fecha de radicacion
- Transacciones por fecha de radicacion
- Valor por estado de folio
- Valor por tipo de predio
- Transacciones con valor
- Transacciones con multiples valores

7.2 Visualizaciones

- Series de tiempo
- Mapas geográficos
- Indicadores de riesgo
- Drilldown ORIP → municipio → matrícula

9. Integración con Fuentes Externas

Usamos:

- Códigos municipales - DIVIPOLA oficial (DANE)
- Listado oficial de Oficinas de Registro de Instrumentos Públicos - ORIP
- Catálogo jurídico de la SNR

10. Conclusiones Generales

- Se logró construir una arquitectura escalable en Databricks que soporta más de 30 millones de registros.
- El sistema automatiza la identificación de riesgos que antes requerían meses de revisión manual.
- Se detectaron cientos de miles de inconsistencias operativas, territoriales y jurídicas.
- Se sentaron las bases para un observatorio inmobiliario nacional.

11. Recomendaciones Finales

- Implementar el sistema como pipeline programado en Databricks cada 24 horas.
- Ampliar el modelo usando grafos (Neo4j) para detectar redes de fraude inmobiliario.

12. Anexos

- Tabla de reglas aplicadas
- Lista de ORIPs oficiales

- Listado de DIVIPOLA
- Listado de transacciones inmobiliarias.