Informe

Clasificación de Trámites de Construcción en La Paz

Integrantes:

- Ericka Cori
- Paolo Ramos
- Gaston Nina

4 Octubre, 2025

INDICE

Introducción.....2
1. Fuentes de Datos.....3
1.1 Fuente de datos 1- API- GEOSERVER.....3
1.2 Fuentes de datos 2- Scrapy - HTML.....4
2. Proceso de Ingestión, Limpieza, Clasificación y Almacenamiento.....5
2.1 Proceso de Ingestión.....5
2.2 Proceso de Limpieza.....5
2.3 Proceso de clasificación.....6
2.4 Proceso de Almacenamiento.....6
3. Diseño de Pipeline.....7
3.1 Extraction.....7
3.2 Transformation.....8
3.3 Load.....8
4. Limitaciones y mejoras futuras.....9
4.1 Limitaciones.....9
4.2 Mejoras futuras.....9
Conclusion.....10

Introducción

Este informe documenta el proceso de ETL aplicado al dataset de trámites municipales de La Paz, Bolivia. La información registrada, clave para la gestión urbana y los servicios municipales, requiere calidad e integridad para garantizar transparencia y eficiencia administrativa.

El análisis se enfocó en la detección de registros duplicados, especialmente en el código catastral, que debería ser un identificador único de propiedades. La duplicidad compromete la confiabilidad de los reportes, la integridad referencial y la toma de decisiones en planificación y control urbano.

Este proceso ETL se enmarca en los esfuerzos de modernización de la gestión municipal y busca establecer las bases técnicas para la depuración y normalización de la data, garantizando que la información sobre trámites municipales sea precisa, consistente y confiable para todos los stakeholders involucrados, desde funcionarios municipales hasta ciudadanos que realizan trámites y desarrolladores urbanos que requieren información precisa para la toma de decisiones.

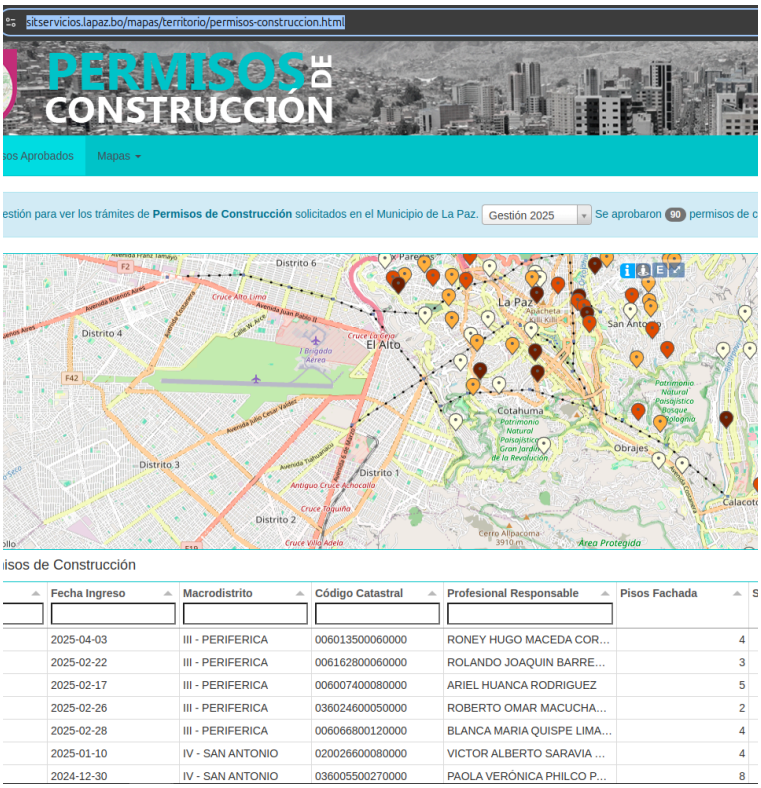
El dataset contiene 13,907 registros distribuidos en 31 campos. Se identificaron 7,020 duplicados (≈50% del total), lo que evidencia la magnitud del problema.

El análisis reveló una significativa presencia de registros duplicados basados en el código catastral. Se identificaron 7,020 registros correspondientes a códigos catastrales repetidos, lo que representa aproximadamente el 50% del total del dataset.

El proceso ETL se enfocó específicamente en la identificación y caracterización de registros duplicados utilizando como clave principal el código catastral, con el propósito de establecer las bases para la depuración y normalización de la base de datos.

1. Fuentes de Datos

1.1 Fuente de datos 1- API- GEOSERVER

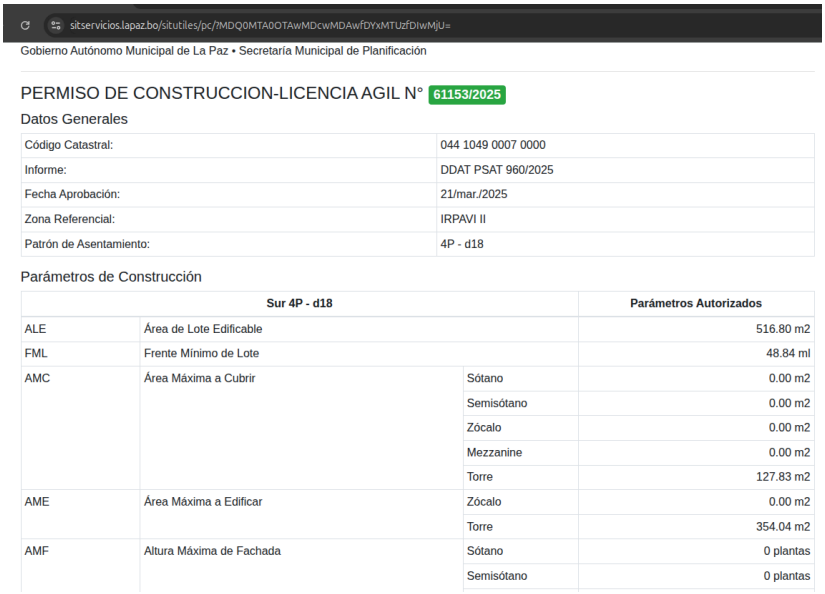


Fecha Ingreso	Macrodistrito	Código Catastral	Profesional Responsable	Pisos Fachada
2025-04-03	III - PERIFERICA	006013500060000	RONEY HUGO MACEDA COR…	4
2025-02-22	III - PERIFERICA	006162800060000	ROLANDO JOAQUIN BARRE…	3
2025-02-17	III - PERIFERICA	006007400080000	ARIEL HUANCA RODRIGUEZ	5
2025-02-26	III - PERIFERICA	036024600050000	ROBERTO OMAR MACUCHA…	2
2025-02-28	III - PERIFERICA	006066800120000	BLANCA MARIA QUISPE LIMA…	4
2025-01-10	IV - SAN ANTONIO	020026600080000	VICTOR ALBERTO SARAVIA …	4
2024-12-30	IV - SAN ANTONIO	036005500270000	PAOLA VERÓNICA PHILCO P…	8

La plataforma web Permisos de Construcción - La Paz constituye una herramienta de consulta geoespacial que permite acceder a información territorial sobre trámites de construcción en el municipio de La Paz. Esta plataforma se alimenta de un servicio de datos geográficos (GeoServer), el cual expone capas temáticas que pueden ser filtradas dinámicamente según criterios específicos.

Web: https://sitservicios.lapaz.bo/mapas/territorio/permisos-construccion.html

1.2 Fuentes de datos 2- Scrapy - HTML



sitservicios.lapaz.bo/situtiles/pc/?MDQ0MTA0OTAwMDcwMDAwfDYxMTUzfDIwMjU=

Gobierno Autónomo Municipal de La Paz • Secretaría Municipal de Planificación

PERMISO DE CONSTRUCCION-LICENCIA AGIL N° 61153/2025

Datos Generales

Código Catastral:	044 1049 0007 0000
Informe:	DDAT PSAT 960/2025
Fecha Aprobación:	21/mar./2025
Zona Referencial:	IRPAVI II
Patrón de Asentamiento:	4P - d18

Parámetros de Construcción

Sur 4P - d18			Parámetros Autorizados
ALE	Área de Lote Edificable		516.80 m2
FML	Frente Mínimo de Lote		48.84 ml
AMC	Área Máxima a Cubrir	Sótano	0.00 m2
		Semisótano	0.00 m2
		Zócalo	0.00 m2
		Mezzanine	0.00 m2
		Torre	127.83 m2
AME	Área Máxima a Edificar	Zócalo	0.00 m2
		Torre	354.04 m2
AMF	Altura Máxima de Fachada	Sótano	0 plantas
		Semisótano	0 plantas

La plataforma SEAT ÚTILES - Permisos de Construcción es una interfaz web que permite la visualización y consulta de datos geoespaciales relacionados con permisos de construcción aprobados en el municipio de La Paz. Esta herramienta se integra con un servidor de mapas (GeoServer), desde el cual se extraen datos filtrados para su análisis y visualización.

Web:

https://sitservicios.lapaz.bo/situtiles/pc/?MDQ0MTA0OTAwMDcwMDAwfDYxMTUzfDIwMjU=

2. Proceso de Ingestión, Limpieza, Clasificación y Almacenamiento

2.1 Proceso de Ingestión.

El proceso de ingestión inicia con el scraping de la página WEB también con consumo de la API, conteniendo 13,907 registros distribuidos en 31 campos. Se implementa una carga controlada utilizando pandas con configuraciones específicas para preservar la integridad de los datos, particularmente en el campo codigo_catastral que se define como string para conservar ceros a la izquierda y formatos catastrales. Durante esta fase se realiza una verificación inicial de integridad que incluye conteo de registros, validación de estructura de columnas, análisis de metadatos básicos y verificación de consistencia en formatos, asegurando que la data cruda mantenga su estructura original antes de proceder a las transformaciones.

2.2 Proceso de Limpieza

La limpieza se focaliza en la identificación y tratamiento de duplicados mediante agrupamiento por codigo_catastral, detectando 7,020 registros repetidos que representan aproximadamente el 50% del dataset. El proceso incluye el análisis de valores nulos across todas las columnas, estandarización de formatos en campos críticos como fechas y códigos, y validación de consistencia en campos geográficos como coordenadas y distritos. Se implementan técnicas de profiling avanzado para caracterizar la calidad de datos y se establecen reglas de negocio para determinar cuáles duplicados corresponden a trámites legítimos múltiples y cuáles a errores de captura.

2.3 Proceso de clasificación

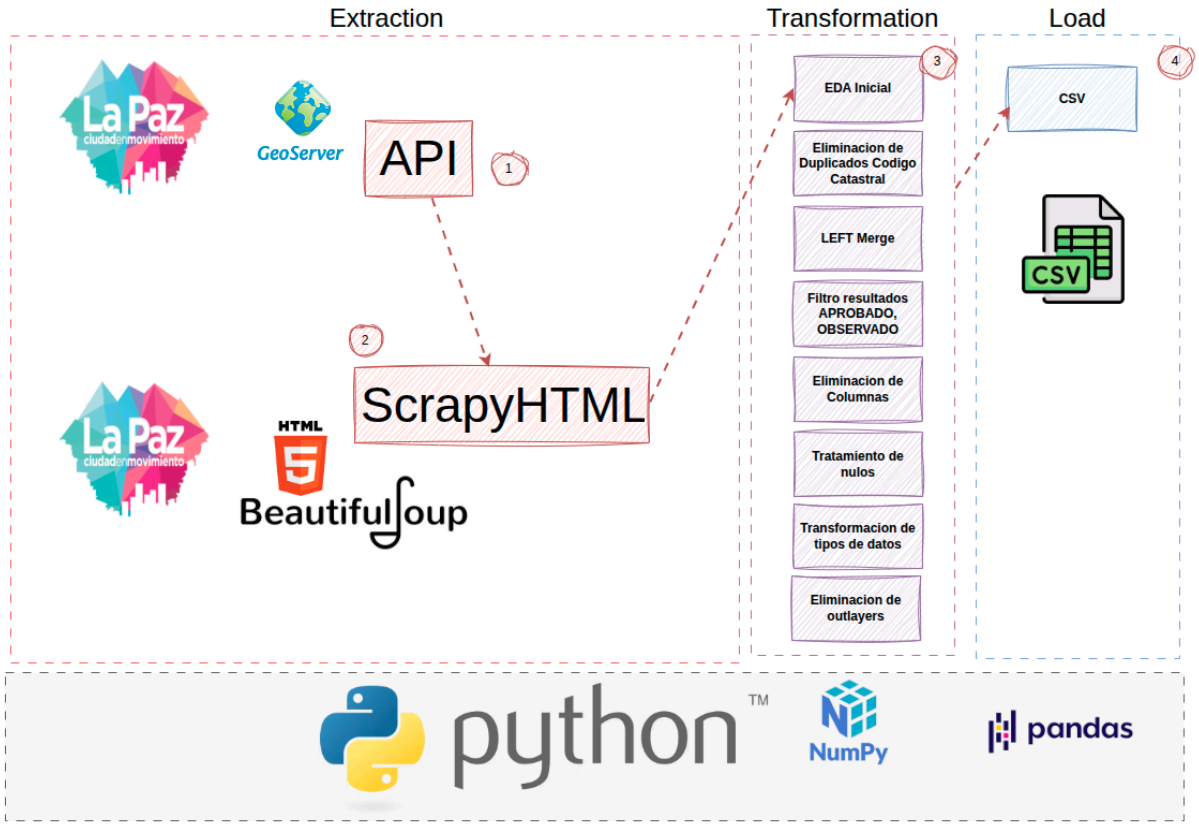
Se establece un sistema de clasificación que categoriza los registros según múltiples dimensiones: por tipo de trámite, estado procesal, complejidad del proyecto y distribución territorial. Los duplicados identificados se clasifican en categorías específicas como "múltiples trámites por propiedad", "errores de captura" o "actualizaciones de expedientes". Paralelamente, se enriquece el dataset mediante la integración de metadatos derivados como frecuencia de trámites por zona, tiempos promedio de resolución y patrones de comportamiento por tipo de solicitante, generando insights adicionales para análisis avanzado.

2.4 Proceso de Almacenamiento

El almacenamiento se estructura en capas diferenciadas: datos crudos preservados para auditoría, datos depurados para operaciones diarias, y datos enriquecidos para análisis estratégico. Se implementan controles de calidad post-procesamiento que incluyen validación de integridad referencial, consistencia en relaciones entre tablas y completitud de campos críticos. Finalmente, se establecen mecanismos de disponibilización mediante vistas segmentadas para diferentes usuarios (operativos, analíticos, gerenciales) y se documentan los metadatos del proceso para garantizar la trazabilidad y facilitar futuras actualizaciones del pipeline ETL.

3. Diseño de Pipeline

1. Extraccion
2. Transformacion
3. Load



3.1 Extraction

La extracción es el proceso de obtención de datos desde diversas fuentes hacia nuestro entorno de procesamiento. Implementamos un sistema robusto que maneja múltiples formatos y orígenes, incluyendo archivos locales, APIs REST y bases de datos. Cada fuente requiere configuraciones específicas de conexión y manejo de errores para garantizar la integridad de los datos capturados.

```
# Extracción desde múltiples fuentes
```

```
df_api = pd.read_csv("tramites_lapaz_api_identificador.csv",
dtype={"codigo_catastral": str})
datos_externos = pd.read_excel("datos_complementarios.xlsx")
```

Para APIs REST, utilizamos requests con manejo de paginación y límites de tasa, mientras que en bases de datos SQL empleamos conexiones parametrizadas con timeouts. La validación inicial verifica que los datos extraídos cumplan con el schema esperado antes de proceder a la transformación.

3.2 Transformation

En esta fase aplicamos reglas de negocio y limpieza para convertir los datos en un formato analíticamente útil. El proceso incluye tratamiento de valores nulos, estandarización de formatos, enriquecimiento con datos externos y feature engineering. Cada transformación está documentada y versionada para mantener la trazabilidad.

```
# Limpieza y transformación

df_clean = df_api.drop_duplicates(subset=['codigo_catastral'])

df_clean['fecha_registro'] = pd.to_datetime(df_clean['fecha_registro'])
df_clean['superficie_legal'] = df_clean['superficie_legal'].fillna(0)
```

Las transformaciones se ejecutan en un orden lógico: primero la limpieza básica, luego joins con datos de referencia, después cálculos derivados y finalmente validaciones de calidad. Implementamos pruebas unitarias para cada transformación crítica, asegurando que reglas de negocio como "superficie_construida ≤ superficie_legal" se mantengan consistentes.

3.3 Load

La carga consiste en almacenar los datos transformados en destinos específicos según los requisitos de consumo. Diseñamos estrategias de upsert para actualizar registros existentes e insertar nuevos, optimizando el uso de recursos y manteniendo

la consistencia de los datos, en nuestro caso la exportación de la data en un archivo csv.

4. Limitaciones y mejoras futuras

4.1 Limitaciones

El pipeline actual presenta varias limitaciones técnicas y operativas que impactan su escalabilidad y mantenimiento. La principal restricción reside en su arquitectura monolítica, donde todas las transformaciones se ejecutan secuencialmente en un solo proceso, creando cuellos de botella en el procesamiento de grandes volúmenes de datos. Actualmente manejamos aproximadamente 14,000 registros, pero esta aproximación no sería viable al escalar a cientos de miles o millones de registros.

```
# Limitación actual: procesamiento en memoria
df_completo = pd.read_csv("datos_masivos.csv")  # Crash con archivos muy grandes
```

Otra limitación significativa es la falta de orquestación y monitoreo automatizado. Las ejecuciones dependen de triggers manuales y no contamos con sistemas de alerta temprana para fallos en las fuentes de datos o degradación en la calidad. La gestión de errores es básica, sin mecanismos de reintento inteligente o compensación para operaciones fallidas. Además, el pipeline carece de capacidades de procesamiento en tiempo real, limitándose únicamente a procesamiento por lotes (batch processing).

4.2 Mejoras futuras

Integración de metadatos y lineage con herramientas como OpenMetadata para proporcionar trazabilidad completa desde las fuentes originales hasta los productos de datos finales, facilitando la auditoría y el diagnóstico de problemas. La containerización con Docker y orquestación con Kubernetes garantizarán portabilidad y alta disponibilidad del pipeline en diferentes entornos.

Conclusion

El desarrollo de este pipeline ETL en Apache Airflow representa un avance estratégico en la gestión de datos de trámites municipales, estableciendo un framework robusto y escalable para la orquestación de procesos de datos. La implementación con Airflow ha permitido automatizar completamente el flujo de trabajo, programando ejecuciones periódicas, gestionando dependencias entre tareas y proporcionando monitoreo en tiempo real mediante su interfaz web. Esta arquitectura basada en DAGs (Directed Acyclic Graphs) garantiza que cada etapa del pipeline extracción, transformación y carga se ejecute de manera coordinada y con capacidad de recuperación ante fallos, superando las limitaciones de los scripts monolíticos tradicionales.

La elección de Airflow como orquestador no solo resuelve los desafíos inmediatos de automatización, sino que sienta las bases para una evolución hacia arquitecturas de datos más complejas. El pipeline actual, operando como un DAG bien estructurado, puede extenderse fácilmente para incorporar nuevas fuentes de datos, transformaciones más sofisticadas y destinos adicionales sin comprometer la estabilidad del sistema. Las capacidades nativas de Airflow para manejo de errores, reintentos inteligentes y logging detallado transforman este proyecto de un simple proceso ETL a una plataforma de datos empresarial preparada para integrarse con herramientas de MLops, calidad de datos y procesamiento en streaming, posicionando a la organización para los desafíos de datos del futuro.

BATCH Y/O STREAMING

En un escenario ideal, el procesamiento de datos evoluciona primero como batch y posteriormente hacia streaming. Esta progresión responde a consideraciones prácticas de implementación y madurez organizacional. Inicialmente, el procesamiento por lotes (batch) ofrece un punto de partida más accesible, con menores

requerimientos técnicos y de infraestructura. Permite establecer las bases del pipeline de datos, definir las transformaciones necesarias y validar la calidad de los resultados antes de escalar hacia soluciones más complejas. Una vez que el proceso batch está estabilizado y se identifican necesidades de datos más actualizados, se puede plantear la migración hacia arquitecturas de streaming. Esta transición representa un avance natural en la madurez de los sistemas de datos, donde primero se resuelven los requisitos básicos y luego se incorporan capacidades más avanzadas según las necesidades del negocio.

DIFICULTADES

La calidad de los datos representa uno de los desafíos más significativos, manifestándose a través de valores nulos, inconsistentes o erróneos que comprometen la confiabilidad de cualquier análisis. En el dataset de trámites catastrales, esto se evidenció con columnas que presentaban hasta 87.5% de valores faltantes, como id_ins_documento y nombre_archivo. La integridad referencial se ve frecuentemente comprometida cuando existen relaciones entre tablas o datasets que no se cumplen, generando registros huérfanos o inconsistencias que dificultan el análisis integrado. La estandarización de formatos es otra dificultad recurrente, donde fechas, códigos y categorías pueden presentarse en múltiples formatos dentro del mismo dataset, requiriendo procesos de homogenización complejos.

La presencia de duplicados constituye un problema fundamental, como se observó en el análisis donde el 25.2% de los códigos catastrales estaban repetidos, algunos hasta 22 veces. Esta duplicación no solo infla artificialmente los volúmenes de datos, sino que puede distorsionar por completo los análisis estadísticos y las agregaciones. La redundancia de información se presenta cuando múltiples columnas almacenan esencialmente la misma información bajo diferentes representaciones, generando inconsistencias potenciales y ocupando espacio de almacenamiento innecesario. La detección de duplicados semánticos representa un desafío adicional, donde registros

que no son idénticos técnicamente pero representan la misma entidad del mundo real requieren algoritmos sofisticados para su identificación.

Los problemas de esquema se manifiestan cuando la estructura de los datos no está bien definida o cambia con el tiempo, dificultando la creación de pipelines estables y confiables. La falta de documentación sobre el significado real de cada campo, sus valores permitidos y las relaciones entre ellos genera ambigüedades que complican el procesamiento y análisis. La heterogeneidad de tipos de datos dentro de una misma columna, como mezclas de texto y números, requiere procesos de limpieza y estandarización complejos. Los valores atípicos y anomalías en columnas numéricas pueden distorsionar los análisis y requerir técnicas especializadas para su identificación y tratamiento.