ETL

ETAPA etl - sir 2023

La guía de usuario está enfocada en caracterizar y definir los artefactos y procesos necesario para implementar un método de extracción, transformación y carga (ETL-Extract, Transform, Load) para el proyecto de calidad en el Sistema de Información Registral, de la Superintendencia de Notaria y Registro.

Contenido

[INTRODUCCIÓN 1](#_Toc125891176)

[DEFINICIÓN DEL AMBIENTE DE TRABAJO 2](#_Toc125891177)

[DISEÑO E IMPLEMENTACIÓN DEL PROCESO ETL 3](#_Toc125891178)

[Notas de informes financieros 7](#_Toc125891179)

[Informe de auditor independiente 8](#_Toc125891180)

[Información de contacto 9](#_Toc125891181)

[Información de la compañía 9](#_Toc125891182)

# INTRODUCCIÓN

La etapa de ETL (Extract, Transform, Load) tiene como objetivo proveer y mejorar el proceso de recopilación, limpieza y carga de datos de diferentes fuentes del Sistema de Información Registral – SIR de la Superintendencia de Notaría y Registro en un sistema de almacenamiento centralizado, mediante el diseño, construcción e implementación del modelo de la base de datos y los flujos de calidad. Es determinado el ambiente de trabajo que permita la extracción de los datos de diferentes formatos y estructuras, la limpieza y transformación de los mismos según las necesidades establecidas, y garantice la carga de datos de manera eficaz. La ejecución de la etapa permitirá mejorar la calidad de los datos, así como reducir los tiempos de procesamiento, contribuyendo así de manera positiva a la operación de la superintendencia.

# DEFINICIÓN DEL AMBIENTE DE TRABAJO

El ambiente de trabajo implementado se determinó a través del análisis y estudio de diferentes plataformas enfocadas en el ‘*Business Intelligence - BI’* y el análisis de datos de código abierto. Los criterios de selección se determinaron según los objetivos del modelo de negocio del proyecto y estipularon en los siguientes conceptos:

* Amplia gama de herramientas para integrar, transformar y analizar datos de diferentes fuentes.
* Ofrezca una interfaz gráfica intuitiva para crear y gestionar procesos ETL.
* Licenciamiento de código abierto.

Acorde a la definición del ambiente de trabajo, se determinó el uso de las siguientes herramientas que permiten el despliegue funcional de la etapa:

## ORACLE DATABASE

Base de datos de Oracle permite la recopilación organizada de información o datos estructurados a través del almacenamiento electrónico en un sistema informático, a través de un sistema de gestión de base de datos.

## ORACLE SQL DEVELOPER

Entorno de desarrollo integrado y regido bajo licencia gratuita que simplifica la gestión y desarrollo de las bases de datos de Oracle en implementaciones tradicionales y en la nube.

Versión: 22.2.1.234.1810 - 12 de septiembre de 2022.

## SPOON - PENTAHO DATA INEGRATION

Herramienta de integración de datos de código abierto escrita en Java que forma parte de la plataforma Pentaho Data Integration. Spoon despliega una interfaz gráfica de usuario que permite a los usuarios crear, probar y ejecutar flujos de trabajo de integración de datos, así como gestionar la integración de datos en tiempo real.

Versión: 9.3.0.0.428 - 12 de abril de 2022.

Requisitos: Preinstalación del JDK 11.

## jdk 11 – java development ki

Kit de desarrollo de software para la plataforma Java, que ofrece un conjunto de herramientas, bibliotecas y documentación referente a aplicaciones Java. Incluye el compilador Java, la máquina virtual Java (JVM), y bibliotecas Java estándar.

Versión: 11.0.13.

# DISEÑO E IMPLEMENTACIÓN DEL PROCESO ETL

## FASE I

Definición de los requisitos que definen el ETL y la base de datos según lo establecido en la versión 2.7 del modelo de negocio del proyecto. Los campos de información sobre las cuales se debe garantizar la calidad de los datos son:

* Dirección del Inmueble
* Nombre del Propietario
* Identificación del Propietario
* Sexo del Propietario
* Código de Departamento y Municipio (Divipola),
* Área del Predio

## DIRECCIÓN DEL INMUEBLE

Figura 1. Estructura del objeto.

Tabla . Campos de información de la entidad dirección del inmueble.

### NOMBRE DEL PROPIETARIO



Tabla 2. Campos de información Entidad dirección del inmueble.

**IDENTIFICACIÓN DEL PROPIETARIO**



Tabla 3. Campos de información Entidad dirección del inmueble.

**SEXO DEL PROPIETARIO**



Tabla 4. Campos de información Entidad dirección del inmueble.

**CÓDIGO DE DEPARTAMENTO Y MUNICIPIO (DIVIPOLA)** 

Tabla 5. Campos de información Entidad dirección del inmueble.

**AREA DEL PREDIO**



Tabla 6. Campos de información Entidad dirección del inmueble.

* Área privada: char [(10)] – Datos de longitud de cadena de 10 bytes de tamaño fijo.
* Área construida: char [(10)] – Datos de longitud de cadena de 10 bytes de tamaño fijo.
* Área total: char [(10)] – Datos de longitud de cadena de 10 bytes de tamaño fijo.
* Coeficiente: numeric (3,0) – Datos numéricos que tiene tres dígitos y cero decimales.

## fase II

Es diseñada la arquitectura de la base de datos a implementar, a partir del planteamiento y desarrollo de los modelos conceptual, lógico y físico, según lo establecido en los criterios de calidad especificados en la fase anterior.

### MODELO CONCEPTUAL

El modelo conceptual aterriza los objetos de información en diferentes entidades de datos y flujos de interacciones. El modelo conceptual diseñado establece las siguientes entidades como artefactos fundamentales de la base de datos a diseñar: 1) Ciudadano, 2) Anotación Ciudadano, 3) Folio, 4) Zona registral, 5) Dirección, 6) Oficina de Origen, 7) Documento, 8) Municipio, 9) Departamento y 10) Vereda.

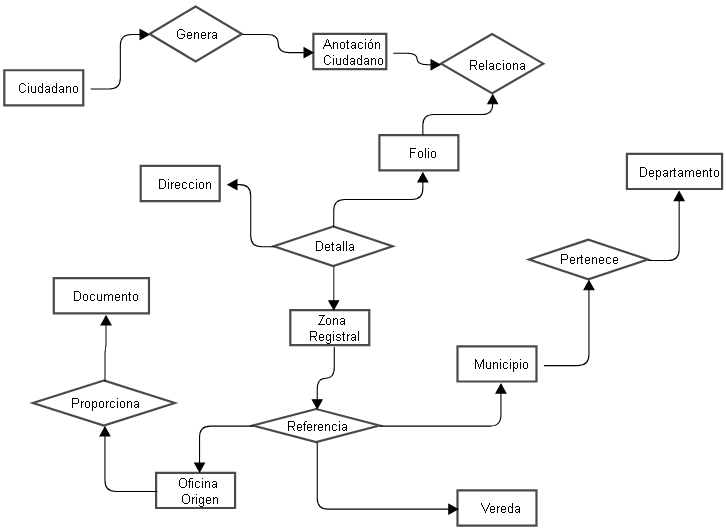


Figura 1. Modelo conceptual

### MODELO LOGICO

A través del modelo son relacionadas y normalizadas las entidades y flujos de interacción, como lo ilustra la Figura 2. A continuación son los especificados los atributos que conforman las entidades del modelo.

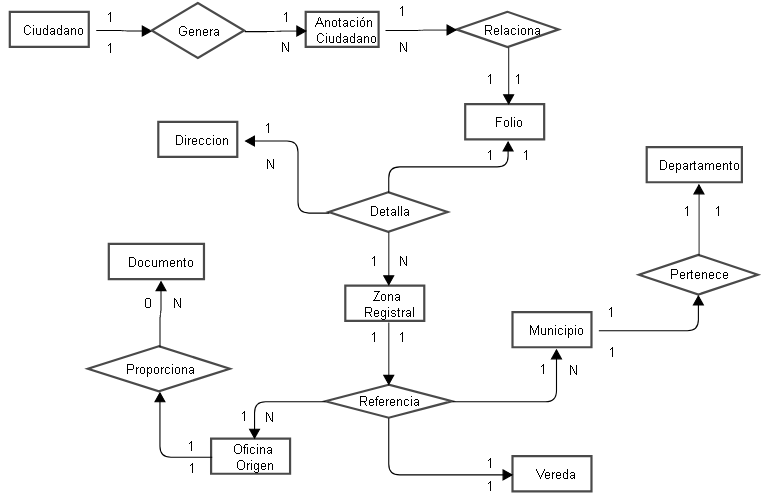


Figura 2. Modelo lógico.

#### Entidad Ciudadano

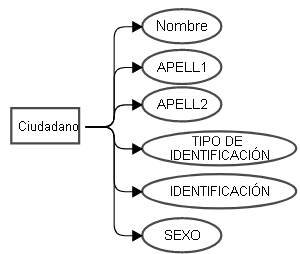


Figura 3. Entidad Ciudadano.

* Nombre: char [(65)] – Datos de longitud de cadena de 65 bytes de tamaño fijo.
* Apell1: char [(30)] – Datos de longitud de cadena de 30 bytes de tamaño fijo.
* Apell2: char [(30)] – Datos de longitud de cadena de 30 bytes de tamaño fijo.
* Tipo de identificación: char [(3)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Identificación: char [(12)] – Datos de longitud de cadena de 12 bytes de tamaño fijo.
* Sexo: char [(1)] – Datos de longitud de cadena de 1 bytes de tamaño fijo.

#### Entidad Anotación - Ciudadano

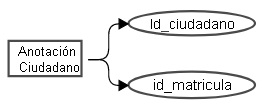


Figura 4. Entidad Anotación - Ciudadano.

* id\_ciudadano: char [(12)] – Datos de longitud de cadena de 12 bytes de tamaño fijo.
* id\_matricula: char [(10)] – Datos de longitud de cadena de 10 bytes de tamaño fijo.

#### Entidad Folio

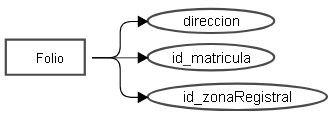


Figura 5. Entidad Folio.

* direccion: char [(n)] – Datos de longitud de cadena de n bytes de tamaño fijo.
* id\_matricula: char [(10)] – Datos de longitud de cadena de 10 bytes de tamaño fijo.
* id\_zonaRegistral: char [(n)] – Datos de longitud de cadena de n bytes de tamaño fijo.

#### Entidad Dirección

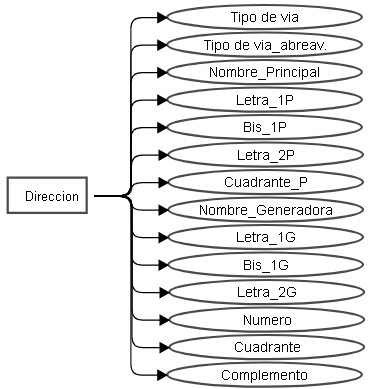


Figura 6. Entidad Dirección.

* TipoDeVia: char [(100)] – Datos de longitud de cadena de 100 bytes de tamaño fijo.
* TipoDeViaAbrev: char [(2)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Nombre\_Principal: char [(45)] – Datos de longitud de cadena de 45 bytes de tamaño fijo.
* Letra\_1P: char [(2)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Bis: char [(3)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Letra\_2P: char [(2)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Cuadrante\_P: char [(5)] – Datos de longitud de cadena de 5 bytes de tamaño fijo.
* Nombre\_generadora: char [(45)] – Datos de longitud de cadena de 45 bytes de tamaño fijo.
* Letra\_1G: char [(2)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Bis\_1G: char [(3)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Letra\_2G: char [(2)] – Datos de longitud de cadena de 2 bytes de tamaño fijo.
* Numero: numeric (4,0) – Datos numéricos que tiene cuatro dígitos y cero decimales.
* Cuadrante: char [(5)] – Datos de longitud de cadena de 5 bytes de tamaño fijo.
* Complemento: char [(45)] – Datos de longitud de cadena de 45 bytes de tamaño fijo.

#### Entidad Zona Registral

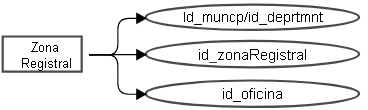


Figura 7. Zona Registral

* Id\_muncp/id\_deprtmnt: numeric (5,0) – Datos numéricos que tiene cinco dígitos y cero decimales.
* id\_zonaRegistral: char [(n)] – Datos de longitud de cadena de n bytes de tamaño fijo.
* id\_oficina: char [(n)] – Datos de longitud de cadena de n bytes de tamaño fijo.

#### Entidad Zona Registral



Figura 7. Zona Registral.

* ID\_Circulo: char [(3)] – Datos de longitud de cadena de tres bytes de tamaño fijo.

#### Entidad Documento

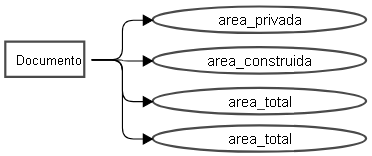
DIRE

Figura 8. Documento.

* Area\_Privada: numeric (10,0) – Dato numérico de diez dígitos y cero decimales.
* Area\_Construida: numeric (10,0) – Dato numérico de diez dígitos y cero decimales.
* Area\_Total: numeric (10,0) – Dato numérico de diez dígitos y cero decimales.
* Coeficiente: numeric (3,0) – Dato numérico de tres dígitos y cero decimales.

#### Entidad Municipio



Figura 8. Documento.

* cplobla: numeric (3,0) – Datos numéricos que tiene tres dígitos y cero decimales.
* Lmunic: char [(60)] – Datos de longitud de cadena de 60 bytes de tamaño fijo.

#### Entidad Departamento



Figura 8. Documento.

* cplobla: numeric (2,0) – Datos numéricos que tiene tres dígitos y cero decimales.
* ldepar: char [(40)] – Datos de longitud de cadena de 40 bytes de tamaño fijo.

### MODELO FISICO

El modelo de datos físicos representa cómo se construirá el modelo en la base de datos. En el cual se identifican todas las estructuras de tabla, las entidades (tablas), el tipo de datos de atributo de las entidades, la clave principal, la clave externa y las relaciones entre las tablas.

### ETL

Para el desarrollo del proceso ETL se definen las reglas de validación acorde al modelo de negocio del proyecto, el modelo de flujo de transformación y representación gráfica de las estructuras desarrolladas.

### Reglas de validación

Los caracteres de control identificados como prioritarios a eliminar son los siguientes:

* Nueva línea (LF o "\n").
* Retorno de carro (CR o "\r").
* Tabulador (HT o "\t").
* Avance de página (FF o "\f").
* Sustitución (SUB o "\0").
* Retroceso (BS o "\b").
* Alerta (BEL o "\a").

El conjunto de caracteres especiales identificados son los siguientes:

* [A-Z] Alfabeto en mayúsculas.
* [a-z] Alfabeto en minúsculas.
* [0-9] Números naturales.
* [Ñ] Letra eñe en mayúsculas.
* [ñ] Letra eñe en minúsculas.
* [áéíóú] Vocales tildadas en minúscula.
* [ÁÉÍÓÚ] Vocales tildadas en mayúsculas.

### Modelo secuencial del proceso ETL

|  |  |  |  |
| --- | --- | --- | --- |
| ORIGEN | PROCESO | OBJETIVO | IMPLEMENTACIÓN |
| CSV | EX00 | Extraer los campos de información normalizados del repositorio. | Selecciona todos los atributos de las entidades y realiza la extracción de los campos de información. |
| EXOO | TR00 | Establecer los atributos que conforman el modelo y eliminar atributos redundantes. | Selecciona solo los atributos requeridos por el modelo físico y verifica la correcta normalización (nombre, tipo y longitud). |
| TR00 | LD00 | Construir el repositorio de trabajo que contendrá todos los registros de información de los atributos del modelo. | Carga los registros de información en una base de datos denominada “REPOSITORIO”. |
| EX00 | TR00 | Excluir los ciudadanos del círculo de pruebas. | Establece un filtro de condición para el circulo “999”. |
| TR00 | TR01 | Establecer y formalizar Los atributos de las entidades. | Normaliza los nombres, tipo y longitud de los atributos del modelo. |
| TR01 | TR02 | Seleccionar el círculo de trabajo. | Establecer un filtro de condición para un círculo especifico. |
| TR02 | TR03 | Establecer un identificador de integración de datos. | Asignar un id de secuencia incremental al objeto de información. |
| TR03 | TR04 | Establecer el inicio de los flujos de trabajo para cada entidad. | Establece el orden de los atributos según la prioridad relacional de las entidades. |
| TR04 | TR05 | Inicializar la entidad Ciudadano. | Selecciona los atributos que componen la entidad Ciudadano. |
| TR05 | TR06 | Determinar los criterios de diferenciación de atributos y eliminación de caracteres de control. | Establece el formato “Mayúsculas “y parametrizan los caracteres de control a eliminar. |
| TR06 | TR07 | Establecer el punto de control de caracteres especiales, errores tipográficos y estandarización de nomenclaturas por atributo. | Valida errores de tipografía relacionados a con caracteres “**/, \ y %”** y la letra Ñ.  Parametriza el conjunto de caracteres no especiales que deben ser eliminados. |
| TR07 | TR08 | Preparar los atributos del Ciudadano a cargar. | Elimina el id integrador. |
| TR08 | LD00 | Cargar la información del Ciudadano. | Genera un archivo de salida SQL. |
| TR04 | TR09 | Inicializar la entidad Anotación Ciudadano. | Selecciona los atributos que componen la entidad Anotación Ciudadano. |
| TR09 | TR10 | Preparar los atributos del Ciudadano a cargar. | Elimina el id integrador. |
| TR10 | LD00 | Cargar la información de la Anotación Ciudadano. | Genera un archivo de salida SQL. |
| TR10 | TR11 | Inicializar la entidad Divipola. | Selecciona los atributos que componen la entidad. |
| TR11 | TR12 | Preparar los atributos de la Divipola a cargar. | Elimina el id integrador. |
| TR12 | LD02 | Cargar la información de la Divipola | Genera un archivo de salida SQL. |
| TR11 | TR13 | Inicializar la entidad Municipio. | Selecciona los atributos que componen la entidad. |
| TR13 | TR14 | Preparar los atributos del Municipio a cargar. | Elimina el id integrador. |
| TR14 | LD03 | Cargar la información de la Divipola | Genera un archivo de salida SQL. |
| TR11 | TR15 | Inicializar la entidad Departamento. | Selecciona los atributos que componen la entidad. |
| TR15 | TR16 | Preparar los atributos del Departamento a cargar. | Elimina el id integrador. |
| TR17 | LD04 | Cargar la información de la Entidad Departamento. | Genera un archivo de salida SQL. |

Tabla 1. Flujo de transformación.

### Modelo grafico del proceso ETL

#### Selector de carga

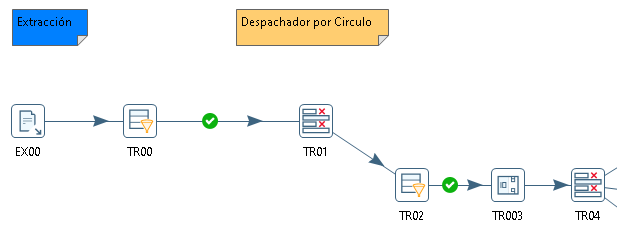


Figura 9. Proceso de extracción y despacho de información.

#### Entidad Ciudadano

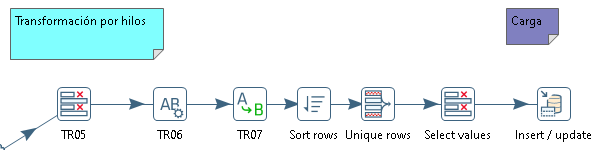


Figura 10. Hilo Entidad Ciudadano

#### Entidad Anotación - Ciudadano



Figura 11. Hilo Entidad Anotación - Ciudadano

#### Entidad Municipio - Departamento

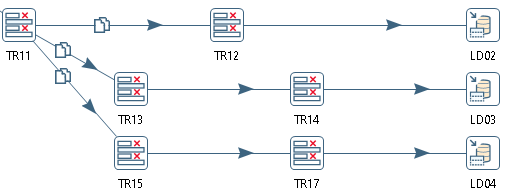


Figura 12. Hilo Entidad Municipio - Departamento