Requerimientos Funcionales - Proceso de Limpieza y Enriquecimiento de Datos

Created by	Juan Alejandro Carrillo Jaimes
Created time	@October 18, 2025 8:32 AM
Tags	AWS, AWS Glue, Cloud, ETL, Fundamentals

1. Objetivo

Establecer las reglas de negocio para la transformación, limpieza y enriquecimiento de datos empresariales, garantizando la calidad y consistencia de la información procesada.

2. Alcance

Este documento define las reglas de negocio aplicables a los datos de registro empresarial, incluyendo la estandarización de campos, validación de información y creación de variables derivadas.

3. Catálogos de Referencia

3.1 Catálogo de Tipos de Identificación

Clase de Identificación	Código	Acción
NIT	11	Mapear al código correspondiente
CEDULA DE CIUDADANIA	33	Mapear al código correspondiente
CEDULA DE EXTRANJERIA	44	Mapear al código correspondiente
PASAPORTE	55	Mapear al código correspondiente
TARJETA DE IDENTIDAD	66	Mapear al código correspondiente
PERMISO POR PROTECCION TEMPORAL	77	Mapear al código correspondiente
PERMISO ESPECIAL DE PERMANENCIA	88	Mapear al código correspondiente
REGISTRO CIVIL DE NACIMIENTO	99	Mapear al código correspondiente
DOCUMENTO EXTRANJERO	101	Mapear al código correspondiente
SIN IDENTIFICACION	-	Registrar en log de errores
NULL	-	Registrar en log de errores

Nota: El catálogo completo se encuentra en el archivo `catalogo_clases_identificaciones.csv`

3.2 Catálogo CIIU - Actividades Económicas

Código	Actividad
0111	Cultivo de cereales (excepto arroz), legumbres y semillas oleaginosas
0112	Cultivo de arroz
4511	Comercio de vehículos automotores nuevos
6910	Actividades jurídicas
6920	Actividades de contabilidad, teneduría de libros, auditoría financiera y asesoría tributaria

Nota: El catálogo completo se encuentra en el archivo `catalogo_codigos_ciuu.csv`

4. Reglas de Negocio

Nº	Regla de Negocio	Descripción	Detalles de Implementación
RN-001	**Estandarización de Fechas**	Convertir todas las fechas (`fecha_matricula` , `fecha_renovacion` , `fecha_actualizacion`) al formato YYYY-MM-DD	• Formato objetivo: ISO 8601 • Valores no convertibles → Log de errores
RN-002	**Limpieza de Registros Duplicados**	Identificar duplicados basados en `matricula` y `nit` , conservando solo el más reciente	• Ordenar por `fecha_actualizacion` descendente • Eliminar duplicados manteniendo el primero
RN-003	**Normalización de Estados**	Unificar valores inconsistentes en `estado_matricula`	• Convertir a mayúsculas • Eliminar espacios al inicio y final • Ejemplo: "activa", "ACTIVA", "Activa" → "ACTIVA"
RN-004	**Creación de Variables Derivadas**	Calcular `antiguedad_empresa` basado en la fecha de matrícula	• Fórmula: `año_actual - año(fecha_matricula)` • Si fecha es NULL → Log de errores
RN-005	**Enriquecimiento de Actividad Económica**	Crear columna `actividad_economica` mapeando `cod_ciuu_act_econ_pri` con el catálogo CIIU	• Join con catálogo por código • Códigos no encontrados → Log de errores
RN-006	**Creación de Llave Única**	Generar campo `id_unico` para identificación inequívoca	• Concatenar: `codigo_camara` + `matricula` + `razon_social` • Si falta algún valor → Log de errores
RN-007	**Gestión de Registros Inválidos**	No eliminar registros, solo reportar inconsistencias	• Exportar errores a CSV • Columnas del log: `columna` , `mensaje_error` , `valor`
RN-008	**Mapeo de Clase de Identificación**	Asignar código numérico según `clase_identificacion`	• Aplicar mapeo del catálogo 2.1 • "SIN IDENTIFICACION" o NULL → Log de errores
RN-009	**Determinación de Tipo de Persona**	Crear campo `tipo_persona` según el tipo de identificación	• Si `clase_identificacion` = "NIT" → tipo_persona = 2 (jurídica) • Otros casos → tipo_persona = 1 (natural)
RN-010	**Estandarización de Nombres de Columnas**	Convertir nombres de columnas a formato snake_case	• Convertir a minúsculas • Reemplazar espacios por guiones bajos • Ejemplo: "Fecha Matricula" → "fecha_matricula"

5. Estructura del Log de Errores

Formato del archivo: `errors_YYYYMMDD.csv`

Campo	Descripción	Ejemplo
columna	Nombre del campo con error	"fecha_matricula"
mensaje_error	Descripción del problema	"Formato de fecha inválido"
valor	Valor original encontrado	"31/13/2023"

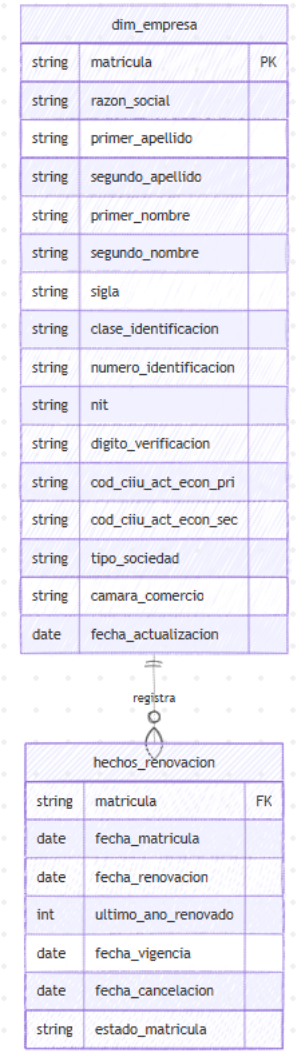
6. Validaciones Críticas

- **Campos obligatorios** que no pueden ser NULL:
  - `matricula`
  - `codigo_camara`
  - `clase_identificacion`
- **Validaciones de consistencia**:
  - Las fechas deben ser lógicamente coherentes
  - Los códigos CIIU deben existir en el catálogo
  - La antigüedad no puede ser negativa

8. Modelo Analítico - Lineamientos de Arquitectura

7.1 Estructura del Modelo Dimensional

El modelo analítico debe construirse siguiendo un esquema estrella (Star Schema) con dos componentes principales:



7.2 Tabla de Hechos: `fact_renovacion`

Propósito: Registrar eventos y actividades que cambian en el tiempo relacionados con la matrícula, renovación y vigencia de las empresas.

Tipo de Datos	Campos	Fuente Origina
Identificador	`matricula`	Campo base

Fechas Clave	`fecha_matricula`
	 `fecha_renovacion`
	 `fecha_vigencia`
	 `fecha_cancelacion`
	 `fecha_actualizacion`
		Columnas temporales

Estado	`estado_matricula`	Indicador del estado
Métricas	`ultimo_ano_renovado`	Extraído del dataset origina

7.3 Dimensión Empresa: `dim_empresa`

Propósito: Almacenar atributos descriptivos y relativamente estáticos de cada empresa.

Tipo de Datos	Campos	Fuente Original	Descripción
Identificador	`matricula`	FK hacia hechos	Llave primaria de la dimensión

Identificación	`numero_identificacion`
	 `nit`
	 `digito_verificacion`
	 `clase_identificacion`
		Columnas base	Datos de identificación legal

Nombre/Razón	`razon_social`
	 `primer_nombre`
	 `primer_apellido`
	 `sigla`
		Columnas descriptivas	Identificación nominal

Clasificación	`tipo_sociedad`
	 `cod_ciiu_act_econ_pri`
	 `cod_ciiu_act_econ_sec`
		Columnas categóricas	Clasificación empresarial y económica

Ubicación	`camara_comercio` `codigo_camara`	Columnas geográficas	Jurisdicción y ubicación

7.4 Consideraciones de Implementación

1. **Integridad Referencial**: La columna `matricula` debe mantener consistencia entre ambas tablas
2. **Granularidad**: Cada registro en `hechos_renovacion` representa un evento de cambio de estado
3. **Historización**: La dimensión empresa debe manejar cambios tipo SCD Tipo 1 (sobrescribir) para mantener simplicidad
4. **Indexación**: Crear índices en:
   - `matricula` (ambas tablas)
   - Campos de fecha en tabla de hechos
   - `cod_ciiu_act_econ_pri` en dimensión

7.5 Campos Derivados Adicionales

Para el modelo analítico, agregar los siguientes campos calculados:

En tabla de hechos:

- `dias_vigencia`: Diferencia entre `fecha_vigencia` y fecha actual
- `flag_vencido`: Indicador si `fecha_vigencia` < fecha actual

En dimensión empresa:

- `tipo_persona`: Derivado según RN-009
- `actividad_economica_descripcion`: Enriquecido desde catálogo CIIU
- `antiguedad_empresa`: Calculado según RN-004

8. Salida Esperada

El proceso debe generar:

1. **Dataset limpio y enriquecido** con todas las transformaciones aplicadas
2. **Archivo de log de errores** con todos los problemas detectados
3. **Estadísticas de procesamiento**:
   - Total de registros procesados
   - Cantidad de duplicados eliminados
   - Cantidad de errores por tipo

Sources 🔗

https://aws.amazon.com/es/glue/features/databrew/

https://docs.aws.amazon.com/glue/latest/dg/machine-learning-teaching.html