Historia de Usuario ETL batería de indicadores. (Release 1).

Nota: Esta historia de usuarios detalla el primer release a implementar del ETL.

Como analista del POT, quiero disponer periódicamente de las variables extraídas desde las fuentes oficiales, validadas y almacenadas en el datalake del distrito; de cada variable necesito contar con la trazabilidad completa del proceso de carga, para poder construir indicadores en etapas posteriores.

Requisitos Funcionales Detallados

1. Fuentes y Extracción

Requisito	Descripción
Disponibilidad de Variables en las fuentes	Se contará con dos fuentes de datos: • **El servidor Windows del DAP** que contendrá una estructura de directorios organizada por sistemas y subsistemas en donde se almacenarán los conjuntos de datos con las variables que requieren los 431 indicadores definidos. • **La Geodata del distrito** que está almacenada en un servidor postgres con su extensión PostGIS. **Características de las variables** • La información de cada conjunto de datos estará representada acorde a lo especificado en el catálogo de datos definido para la batería de indicadores del distrito. • Para cada uno de los 431 indicadores deberá proveerse una o varias variables en donde se especifique la línea base para la medición del indicador.
Extracción Periódica	La extracción debe ocurrir con periodicidad definida por el DAP, para asegurar que los indicadores del POT reflejen la realidad reciente. Esta periodicidad es una decisión administrativa y el proceso ETL se puede lanzar una vez se cumplan los procedimientos internos establecidos por el DAP.

2. Calidad y Validación

Requisito	Variables Afectadas

Validación alfanumérica (Rango)	Toda variable alfanumérica se validará en varios aspectos: ● ¿Permite nulos? ● ¿Permite datos NA (no aplica) y no disponible (ND)? ● Que sus valores estén en un dominio determinado. ● Que sus valores estén en rangos lógicos predefinidos. ● PENDIENTE: Queda por definir en qué casos las variables necesitan algún tipo de transformación.Por ejemplo llevar los valores a la misma escala de medida. ● En caso de unidades de medida la variable deberá estar acompañada de esta; ejemplos: mt, km, ha, etc.
Validación Geográfica (Integridad)	Cada variable que requiera geolocalizarse debe tener un **ID** de la unidad espacial asociada (polígono, punto o multi línea) válida que exista en el Catálogo Geográfico Maestro de la Geodata del distrito.
Bitácora de Errores	Los registros procesados en el ETL que no superen las validaciones deben ser separados en una zona de **ERROR/RECHAZADOS** para su posterior revisión, **sin contaminar** los datos listos para el análisis. Deberá quedar un registro del motivo del rechazo.

3. Trazabilidad y Destino (Data Lake)

Requisito	Propósito
Metadato de Trazabilidad	Cada registro, una vez cargado, debe incluir metadatos de control como el **Fecha_Carga_DL** (Fecha en que ingresó al Data Lake) y el **Origen_FTE** (Fuente de la cual provino), periodo de análisis al que corresponde, asegurando la trazabilidad.
Datos Cargados (RAW)	Debe existir una copia inmutable de las variables originales cargadas. Para ello la estructura del container /POT/ del Datalake consiste de varios folder: */POT/raw/* → los datos de la ingesta tal cual vienen del origen */POT/curated/* → Zona estandarizada y validada acorde al catálogo de datos definido para la batería de indicadores y las transformaciones requeridas. */POT/gold/* → Zona consumible para usar en los cálculos de los indicadores. /POT/error/ → zona con datos rechazados durante la carga.

Datos Listos para Uso	El resultado final (datos validados y consolidados) debe estar disponible en la zona gold del Data Lake, siendo el punto de acceso para el Analista POT que desee generar indicadores.

Criterios de aceptación:

- El sistema debe permitir extraer datos de ambas fuentes de forma periódica.
- Las variables deben ser validadas antes de almacenarse.
- Debe existir trazabilidad sobre qué datos fueron tomados, cuándo y de dónde.
- Los datos deben estar disponibles en el datalake en tres estados: cargados, procesados y listos para uso.
- El analista debe poder consultar las variables sin necesidad de acceder directamente a las fuentes originales.

Revisado y aceptado por:

_____
Nombre usuario:
Cargo en el distrito:
Fecha: