1. INFORMACIÓN BÁSICA

Nombre del Módulo	ETL de las Variables para la Batería de Indicadores POT
Proyecto-Tema-Necesidad	Historia de Usuario para especificar el ETL que carga las variables usadas por los indicadores del POT desde las fuentes al datalake del distrito.
Versión Requisito	1.0
Fecha Requisito	27/11/2025
Tipo de solicitud	[X] Requerimiento [] Proyecto [ ] Incidente

2. DESCRIPCIÓN GENERAL

Requerimiento Funcional

Dentro del proceso de seguimiento y evaluación del POT se requiere la elaboración de una batería de indicadores que permitan medir el cumplimiento de sus metas y objetivos. Estos indicadores se construyen a partir de información suministrada por las dependencias internas y externas que son requeridas.

La información suministrada por las dependencias se define como un conjunto de variables asociadas a la ejecución del POT. Las variables requeridas son suministradas por las dependencias acorde a los parámetros establecidos por el distrito y especificados en el catálogo de datos para la batería de indicadores.

El requerimiento funcional que se describe a continuación detalla la manera como periódicamente se recopilan estas variables desde varias fuentes y su almacenamiento en el data lake del distrito.

Precondiciones:

- Para que el proceso ETL se pueda llevar a cabo las variables requeridas deben ser suministradas por las dependencias relacionadas acorde a los parámetros del distrito.
- Se requiere que el servidor Windows del DAP esté **siempre encendido, conectado y mantenido** por el equipo de TI del DAP para que no falle el proceso inicial de recopilación de datos.

Normatividad:
El sistema de seguimiento y evaluación se realiza acorde a lo establecido en el acuerdo 48 del 2014.

Cómo funciona en la actualidad.

En la actualidad las dependencias directamente entregan los indicadores ya calculados. A partir de este nuevo diseño, las dependencias entregarán las variables requeridas y previa carga en el datalake del distrito.

Estas historias de usuario hacen referencia a la carga periódica de información suministrada por las dependencias al datalake del distrito.

El cargue inicial de la información histórica no está detallado en este documento. Será una historia de usuario que debe especificar el proceso realizado por una única vez en el momento en que el sistema entre en operación.

Historia de usuario 01

Como analista del POT, quiero disponer periódicamente de las variables extraídas desde las fuentes oficiales, que previamente han sido 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.

Requiero:

1. *La disponibilidad de las variables en los servidores definidos.*
   Para disponer la información 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 la información de las variables** que requieren los 431 indicadores definidos.
   - **La Geodata del distrito** que está almacenada en un servidor postgres con su extensión PostGIS.

2. *Tener como referencia para la carga el catálogo de datos del distrito.*
   - 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. El catálogo debe incluir los metadatos necesarios para la validación y homologación (ej: Tipo_Dato_Esperado, Unidad_Homologada, Factor_Conversion, Rango_Logico_Min/Max, etc.)
   - Para garantizar el anterior requerimiento se contará con un instrumento para la recolección de información.
   - Es de anotar que un mismo conjunto de datos puede alimentar información de varios sistemas y subsistemas y por ende varios indicadores.

- En las cargas de inicio de periodo del POT, para cada uno de los indicadores deberá proveerse una o varias variables en donde se especifique la línea base para la medición del indicador.

3. *La extracción periódica de datos desde los servidores.*
   - 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 de manera automática o de manera manual una vez se cumplan los procedimientos internos establecidos por el DAP.

4. *La validación de las variables alfanuméricas***.**
   Toda variable numérica (tipos int32, decimal y float64) se validará y, si es aceptada, se transformará de acuerdo con los metadatos definidos en el catálogo.
   - ¿Permite valores nulos?
   - ¿Permite datos NA (no aplica) y no disponible (ND)?
   - ¿Sus valores están en el dominio establecido en el catálogo de datos?
   - ¿Sus valores están en los rangos lógicos predefinidos?
   - Para las variables que necesiten algún tipo de transformación, este se realice adecuadamente. Por ejemplo, llevar los valores a la misma escala de medida;
   - En el caso de variables que representen unidades de medida estas deberán estar acompañada de la respectiva unidad; ejemplos: mt, km, ha, etc.
   - Estandarización de Precisión: El valor resultante se redondeará a la Precisión Requerida.

5. *La validación de la información geográfica.*
   Cada variable que requiera geolocalizarse debe tener un ID válido de la unidad espacial asociada (polígono, punto o multi línea) y de la capa origen. Por válido se entiende que el ID exista en el catálogo geográfico maestro de la geodata del distrito.
   Adicionalmente, la información de cada variable se debe suministrar bajo los parámetros definidos en el instrumento diseñado para la recolección de información.

6. *Una bitácora de errores que garantice la trazabilidad de la información.*
   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. (ej. "Error: Formato Numérico Inválido", "Error: Valor Fuera de Rango Lógico"). Por otro lado se registran en la

bitácora aquellas fuentes de las cuales no se encontró información para el periodo que se está cargando.

7. *Trazabilidad del periodo al que pertenecen los datos que se cargan.*
    - Cada registro, una vez cargado, debe incluir metadatos de control como el (**Fecha_DL**) fecha en que ingresó al Data Lake, el **Origen_FTE** (Fuente de la cual provino) y periodo de análisis al que corresponde. Para las variables transformadas, se deben registrar la **Unidad_Homologada** y el **Factor_Conversion** aplicado.
    - Si una variable pasa por múltiples transformaciones (*ej. de m a km, y luego se eleva al cuadrado*), la bitácora debe ser lo suficientemente detallada para rastrear todas las modificaciones y no solo el factor de conversión final, asegurando la trazabilidad completa del proceso de carga.

8. *Mantener una copia espejo de los datos originales en el Data Lake del Distrito*.
   Debe existir una copia inmutable de las variables originales cargadas. Para ello la estructura del container /POT/ del Data Lake consiste de varios folder:
    - */POT/raw/* → contiene los datos de la ingesta tal cual vienen del origen
    - */POT/curated/* → Zona con datos estandarizados y validados acorde al catálogo de datos definido para la batería de indicadores.
    - */POT/gold/* → Zona consumible para usar en los cálculos de los indicadores.
    - /POT/error/ → zona con datos rechazados durante la carga; es decir, allí se almacenará el log de trazabilidad del proceso.

9. *Las variables listas para el uso posterior*.
   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 de la Historia de Usuario:

- El sistema debe permitir extraer datos de las dos fuentes definidas en forma periódica.
- Las variables deben ser validadas antes de almacenarse en el data lake
- Debe existir trazabilidad sobre qué datos fueron tomados, cuándo y de dónde.
- Los datos deben estar disponibles en el data lake en tres estados: cargados, procesados y listos para uso.

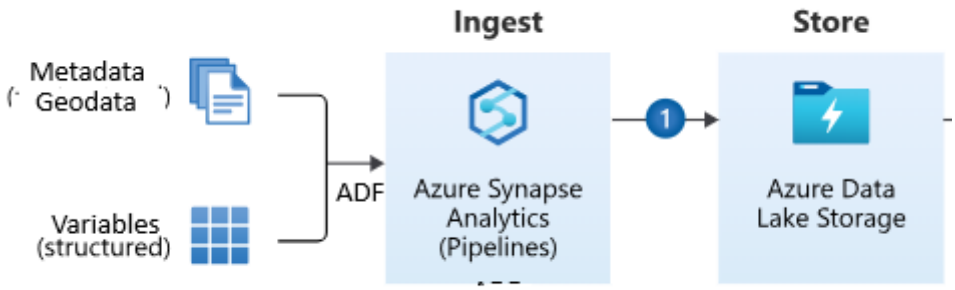
- El analista debe poder consultar las variables sin necesidad de acceder directamente a las fuentes originales.

Historia de usuario 02

Como Ingeniero de Datos quiero configurar un pipeline ETL orquestado por Azure Data Factory/Synapse Pipelines que ingiera datos de dos fuentes, aplica las reglas de calidad y validación y almacene los resultados en las zonas RAW, CURATED, GOLD y ERROR de ADLS Gen2 del distrito.

Requiero:

1. *Un pipeline que incorpore los elementos tecnológicos que se describen en la siguiente figura*.



2. *La conexión al servidor del DAP y la extracción de las variables alfanuméricas.*
   Para habilitar la conexión se instalará un SHIR (Self-Hosted Integration Runtime) en el Servidor Windows del DAP; de esta manera ADF (Azure Data Flow) puede acceder a la estructura de directorios de este servidor y extraer los conjuntos de datos allí almacenados.

3. *La extracción de las variables tabulares y alfanuméricas*.
   - Mediante la opción ADF copy se crea una copia de los archivos que contienen las variables desde el servidor Windows a la zona /POT/raw/ de ADLS Gen2. Se evaluará el mejor almacenamiento para datos tabulares. Esta será una copia inmutable de los datos tal como vienen de las fuentes.
   - El *pipeline* de ingesta debe implementar una **política de reintentos** (máx. 3 reintentos, 5 min de espera) al fallar la conexión SHIR. Si falla definitivamente, se debe enviar una alerta al equipo de TI del DAP.

4. *La extracción de variables geográficas.*
   Usando ADF y Azure Database for PostgreSQL se configura un Linked Service enADF para conectarse a la geodata del distrito (PostgreSQL con PostGIS) y extraer los datos de las

variables, incluyendo la geometría, hacia la zona /POT/raw/. Esta será una copia inmutable de los datos tal como vienen de las fuentes.

5. *Que la extracción sea periódica y programable*.
   Usando la opción ADF Pipeline trigger, se configurará un *Trigger* (manual o programado) para que el proceso se lance acorde a la decisión administrativa del distrito.

6. ***Gestión Centralizada del Catálogo de Metadatos.***
   *Se estudiará la alternativa de implementar **Azure Purview o Microsoft Fabric** como fuente de verdad para el Catálogo de Datos, asegurando que el Data Flow de ADF consuma los metadatos de validación y homologación (Tipo_Dato_Esperado, Rango_Logico_Min/Max, Factor_Conversion, etc.) de manera **programática** y **automatizada**, eliminando la dependencia de archivos manuales.*

7. *La validación de las variables alfanuméricas*.
   Se implementa un Data Flow que consume datos de la zona /POT/RAW y aplica las validaciones de rango alfanumérico (nulos, NA, rangos lógicos) antes de mover los datos a la zona /POT/curated.

8. *La validación de la información georreferenciada*.
   En el Data Flow se implementará lógica para validar que el ID de la unidad espacial asociado a la variable exista en el Catálogo Geográfico Maestro.

9. *La bitácora de errores*.
   El Data Flow de ADF incluirá una división condicional. Los registros que no superen las validaciones serán dirigidos al *sink* de /POT/error/ junto con un campo indicando el motivo del rechazo.
   Se requiere modificar el *Data Flow* para que, antes de la división condicional al *sink* /POT/error/, se implementa una lógica (usando una **columna derivada** o una **expresión de concatenación condicional**) que **acumule y concatena** todos los mensajes de error de validación (ej., 'Error: Formato Numérico Inválido', 'Error: Valor Fuera de Rango Lógico') en un único campo de la bitácora. Esto asegurará la trazabilidad completa de todos los incumplimientos de calidad por cada registro rechazado

10. *La trazabilidad de los datos*.
    En la primera fase de carga (o en la validación), se añadirán los metadatos Fecha_Carga_DL, Origen_FTE y el periodo de análisis a cada registro cargado.

11. *Los datos validados y limpios*.
    Los registros que superen las validaciones serán almacenados en la zona /POT/curated/ en formato optimizado (Ej: Parquet) y estandarizado según el Catálogo de Datos

12. *Los datos listos para consumo de otros sistemas*.

Una vez que se completen las transformaciones y agregaciones, el resultado final y consolidado se almacenará en la zona /POT/gold/ , que será el punto de acceso para el Analista del POT. Su formato de almacenamiento será como Parquet File.

Criterios de Aceptación para la Historia de Usuario 02:

1. El pipeline de ADF se ejecuta correctamente, extrayendo datos tanto del directorio Windows (vía SHIR) como del servidor PostGIS.
2. La data cruda de cada fuente se almacena en la zona /POT/raw/ junto con los metadatos de trazabilidad (Fecha_Carga_DL, Origen_FTE).
3. Los registros que fallan en las validaciones alfanuméricas o geográficas son separados y almacenados en la zona /POT/error/.
4. Los datos validados y transformados para la línea base están disponibles y consolidados en la zona /POT/gold/, listos para ser consumidos por el Analista del POT.
5. Se puede auditar cualquier registro y rastrear su origen y fecha de carga al datalake.
6. La verificación de calidad de los datos se realizará contra lo establecido en el catálogo de datos del distrito.