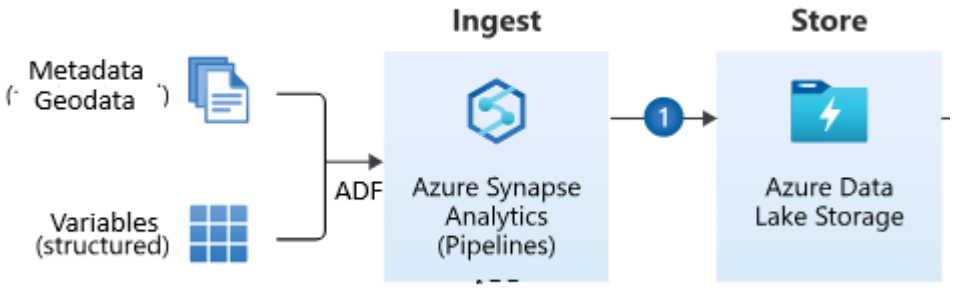
C2.E2-4. Pipeline para el ETL de variables para indicadores (Release 1).

Historia de Usuario:

Nota: Esta historia de usuarios detalla el primer release a implementar del ETL. En caso de ser necesario, en posteriores release se ampliarán las funcionalidades descritas a continuación.



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

1. Conexión de Fuentes y Extracción (ADF y Conectores)

Tarea de Implementación	Componente Tecnológico	Cómo Cumple el Requisito
Conexión a Directorio Windows	ADF / Self-Hosted Integration Runtime (SHIR)	Se instalará un SHIR en el Servidor Windows del DAP para que ADF pueda acceder a la estructura de directorios y extraer los conjuntos de datos.
Extracción de Variables Tabulares y Alfanuméricas.	ADF Copy Activity	Se usarán actividades de copia para mover los archivos de variables desde el servidor Windows a la zona /POT/raw/ de ADLS Gen2. Se evaluará el mejor almacenamiento para datos tabulares.

Extracción de Variables Geográficas	ADF y Azure Database for PostgreSQL	Se configurará un *Linked Service* en ADF 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/.
Periodicidad de la Extracción	ADF Pipeline Trigger	Se configurará un *Trigger* (manual o programado) para que el proceso se lance acorde a la decisión administrativa del DAP.

2. Calidad, Validación y Bitácora (ADF/Pipelines)

Tarea de Implementación	Componente Tecnológico	Cómo Cumple el Requisito
Proceso de Validación	ADF / Flujo de Datos (Data Flow) de Mapping	Se implementará un Data Flow que consume datos de la zona RAW y aplica las validaciones de rango alfanumérico (nulos, NA, rangos lógicos) antes de mover los datos.
Validación Geográfica	Data Flow / Scripts de Transformación	Se implementará lógica para validar que el ID de la unidad espacial asociado a la variable exista en el Catálogo Geográfico Maestro.
Bitácora de Errores	ADLS Gen2 / Zona /POT/error/	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.

Datos Validados y Limpios	ADLS Gen2 / Zona /POT/curated/	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.

3. Trazabilidad y Destino Final (ADLS Gen2)

Tarea de Implementación	Componente Tecnológico	Cómo Cumple el Requisito
Copia Inmutable	ADLS Gen2 / Zona /POT/raw/	Los datos se cargan "tal cual vienen del origen" en la zona RAW, asegurando una copia inmutable del dato fuente.
Metadatos de Trazabilidad	ADF / Actividad de Transformación de 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.
Datos Listos para Consumo	ADLS Gen2 / Zona /POT/gold/	Una vez que se completen las transformaciones y agregaciones para la línea base, el resultado final y consolidado se almacenará en la zona /POT/gold/ , que será el punto de acceso para el Analista del POT.

Criterios de Aceptación para la Implementación:

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.

Revisado y aceptado por:

_____
Nombre usuario:
Cargo en el distrito:
Fecha: