**SOLUCIÓN TECNOLÓGICA PARA LA OPTIMIZACIÓN EN LA GESTIÓN DE RECURSOS HÍDRICOS EN EL PERÚ APLICANDO BUSINESS INTELLIGENCE**

**Documento del Desarrollo del Proceso ETL**

Versión 1.0

**Control de Versiones**

**Lima, 13 de abril del 2025**

**ÍNDICE**

[**1. Introducción 3**](#_heading=h.t4kqk1ckkdqo)

[**2. Objetivo 3**](#_heading=h.6qevqlvf6kjo)

[**3. Descripción General del Proceso ETL 3**](#_heading=h.kile72boe01g)

[3.1. Extracción de Datos 3](#_heading=h.9ru8y11hzxg9)

[3.2. Transformación de Datos 4](#_heading=h.ka5sasiyn91f)

[3.3. Carga de Datos 4](#_heading=h.12mufa276gj)

[**4. Integration Services (SSIS) 5**](#_heading=h.oy5f0leg0epz)

[4.1. Lógica del Proceso ETL 5](#_heading=h.46r0rbt506lj)

[4.2. Configuración de Conexiones 6](#_heading=h.y0pcie1b5kv0)

[4.3. Configuración de Componentes 7](#_heading=h.xo4socvk5iv2)

[4.4. Paquetes 8](#_heading=h.36yvrwa7yx0a)

# **Introducción**

El presente documento define el proceso de ETL (Extract, Transform, Load), un componente vital en el desarrollo de un sistema de Business Intelligence (BI) eficaz y dinámico. Entre las herramientas más utilizadas para este propósito se encuentra Integration Services (SSIS) de Microsoft, que proporciona un entorno robusto para diseñar, implementar y administrar flujos de trabajo de ETL.

# **Objetivo**

Describir de manera detallada el proceso de ETL (Extracción, Transformación y Carga) implementado dentro del entorno de Business Intelligence, especificando las etapas, herramientas utilizadas, fuentes de datos, reglas de transformación y procedimientos de carga hacia el data warehouse.

# **Descripción General del Proceso ETL**

## **Extracción de Datos**

La etapa de extracción consiste en obtener datos desde diversas fuentes, como bases de datos transaccionales, archivos planos o sistemas externos. En este proyecto, se utilizan procedimientos almacenados y conexiones OLE DB/Ado.Net para recuperar los datos necesarios, replicando su estructura en una base intermedia denominada STAGE.

## **Transformación de Datos**

Durante la transformación, los datos extraídos se someten a procesos de limpieza, validación, enriquecimiento y adaptación para garantizar su calidad e integridad. Se aplican reglas de negocio específicas, conversiones de tipo de datos y cálculos necesarios, además de auditorías que registran cada paso mediante claves únicas y recuentos de registros.

## **Carga de Datos**

Finalmente, los datos transformados se cargan en el DataWarehouse, que representa el almacén de datos final. Este proceso incluye la inserción de datos en tablas dimensionales y de hechos, respetando la integridad referencial y utilizando componentes SSIS para automatizar y controlar el flujo completo.

# **Integration Services (SSIS)**

En esta sección, se explora el uso de Integration Services (SSIS) como herramienta principal para la creación de procesos de ETL. Este ofrece una variedad de componentes y funcionalidades que permiten la extracción de datos desde diversas fuentes, la transformación según las necesidades del negocio, y la carga eficiente de los datos en el almacén de datos. Crear una solución de ETL (Extract, Transform, Load) con Integration Services (SSIS) implica una serie de pasos clave que cubren desde el diseño hasta la implementación y monitoreo de los paquetes de flujo de datos.

## **Lógica del Proceso ETL**

Permite establecer conexiones a diversas fuentes de datos como bases de datos SQL Server, archivos planos, fuentes OLE DB, conexiones ODBC, entre otras. El cual, contiene múltiples tareas de flujo de control, como tareas de extracción de datos, transformaciones y cargas en el almacén de datos.

Para el desarrollo de los paqueteques, su lógica se divide en dos:

* **Auditoria:** Se enfoca en registrar y rastrear información relevante sobre el proceso de carga de datos.
  + ***Recuento de datos inicial/final:*** Antes y después de la transformación, se realizan recuentos de datos para asegurar que no se pierdan registros durante el proceso de ETL.
  + ***Consulta del GUID para su creación:*** Se consulta un identificador único (GUID) para asociarlo con cada registro nuevo que ingresa al sistema, lo que permite un seguimiento preciso de los datos a lo largo del tiempo.
  + I***nserción y/o actualización del registro:*** Los datos y sus metadatos asociados, como la fecha de carga y el GUID, se insertan o actualizan en las tablas de auditoría para su posterior análisis y seguimiento.
* **Proceso:** Se centra en las tareas de extracción, transformación y carga de datos en el almacén de datos o el destino final.

## **Configuración de Conexiones**

En el apartado de administración de conexiones, se realizará las creaciones de 2 tipos: Ado.Net y OleDB.

* **Ado.Net:** Proveedor de datos optimizado para trabajar directamente con objetos .NET. Ofrece mayor flexibilidad para tareas como ejecuciones de procedimientos almacenados y manejo de parámetros.
* **OleDB:** Proveedor de acceso genérico a diversas fuentes de datos, especialmente útil para conectarse con bases de datos como SQL Server.

Asimismo,al realizar la creación de estas conexiones se recomienda renombrarlas. Teniendo en cuenta el tipo de conexión Ado.Net/OleDB, Base de datos, Usuario

## **Configuración de Componentes**

En el entorno de desarrollo de Microsoft Visual Studio, se crean las variables necesarias para cada ciclo haciendo uso de los Tasks o Tareas. Las tareas son bloques funcionales dentro de los paquetes SSIS que ejecutan acciones específicas, realizar transformaciones de datos o controlar el flujo de ejecución.

Se necesitan distintos tipos de tareas en un ciclo. De manera general, se dividen de la siguiente forma:

* **Tarea de Recuento de filas anteriores:** Estos son configurados para obtener los datos de una tabla en específico mediante una consulta de SQL. Asimismo, se agrega una variable interna que alojara el resultado, siendo un valor por defecto igual a cero; el cual servirá para contar las filas cargadas.
* **Tarea de Auditoría:** Esta tarea se encarga de insertar un registro en la tabla de auditoría al inicio de cada ciclo ETL. Registra información como el nombre del paquete, GUID, hora de inicio, número de filas iniciales, y un identificador único de ejecución. Esto permite hacer seguimiento de cada ejecución, facilitando la trazabilidad, monitoreo y control de errores en los procesos de carga.
* **Tarea de Recuento de filas posteriores:** Similar a la tarea de recuento inicial, esta se ejecuta al finalizar el proceso de carga para contar nuevamente los registros de la tabla destino. El resultado se almacena en la misma variable usada anteriormente y se utiliza para actualizar el registro de auditoría, permitiendo verificar los datos previamente fueron insertados/ actualizados/ eliminados dependiendo de la logia del proceso.
* **Tarea de Actualización de Auditoría:** Esta tarea se ejecuta al finalizar el ciclo ETL para actualizar el registro correspondiente en la tabla de Auditoría. Utiliza el identificador generado previamente para ubicar el registro exacto, y actualiza campos como la fecha y hora de finalización, el número de filas procesadas (posteriores) y el estado del proceso (éxito o fallo). Esta acción cierra formalmente la auditoría del paquete, asegurando que cada ejecución quede documentada de forma completa y precisa.

## **Paquetes**

Los paquetes representan unidades modulares de trabajo que encapsulan la lógica de extracción, transformación y carga de datos. Cada paquete contiene una secuencia de tareas organizadas mediante flujos de control y flujos de datos.

* **Stage:** Los paquetes cumplen una función de integración inicial: extraer los datos desde las fuentes originales (como sistemas operacionales o archivos externos) y almacenarlos en tablas intermedias que replican fielmente su estructura original. Esta capa actúa como una "zona de aterrizaje" segura, permitiendo validar y preparar los datos antes de que ingresen al almacén principal.

Dentro de este proceso, Stage, hace uso de una tarea adicional. La cual realiza una limpieza antes de cada nueva carga, evitando duplicidad de datos.

* **Data Warehouse:** Los paquetes de este proceso, representan la segunda etapa del flujo ETL. Aquí, los datos provenientes del STAGE se transforman y cargan en estructuras analíticas como dimensiones y hechos. Estos paquetes están orientados a dar formato analítico, aplicar reglas de negocio y garantizar la integridad referencial.   
  Además, implementan una lógica más compleja como MERGE, inserciones condicionales o uso de vistas para unificar múltiples fuentes.   
  Por último, se hace uso del componente Lookup se usa para relacionar los datos con las dimensiones existentes en el DWH, y la correcta configuración de la opción Full Cache para mejorar el rendimiento durante la carga masiva.