Documentación interna

Manual de uso TeseoETL

v2.0.1 - Actualizado 07/02/2025

¿Qué es TeseoETL?

TeseoETL es una herramienta de extracción, transformación y carga de datos (ETL, por sus siglas en inglés) diseñada para facilitar el procesamiento y la gestión de grandes volúmenes de información de manera eficiente y automatizada. Esta plataforma permite conectar diversas fuentes de datos, transformar la información según las necesidades específicas de cada proyecto, y cargar los resultados en sistemas de almacenamiento o visualización para su análisis y toma de decisiones.

TeseoETL está orientado a usuarios que necesitan integrar datos provenientes de diferentes fuentes, como bases de datos, APIs, hojas de cálculo, y archivos planos, entre otros. Su flexibilidad y capacidad de personalización lo convierten en una herramienta ideal para proyectos que requieren soluciones adaptadas a contextos específicos, como el monitoreo electoral, la transparencia gubernamental, y la eficiencia en contrataciones públicas.

Entre sus principales características se incluyen:

- **Automatización de procesos:** Permite programar tareas recurrentes para la extracción y transformación de datos.
- **Conectividad flexible:** Soporta múltiples formatos y protocolos de conexión.

- **Transformación de datos:** Herramientas integradas para limpiar, normalizar y enriquecer datos.
- **Escalabilidad:** Diseñado para manejar tanto pequeñas como grandes cantidades de datos.
- **Interfaz amigable:** Facilita la configuración de procesos ETL sin necesidad de conocimientos avanzados en programación.
- **Bitácora de consultas:** Incluye un registro de las consultas realizadas por los usuarios, facilitando el seguimiento y la auditoría.
- **Consulta de información vía API:** Permite acceder y consultar los datos procesados a través de interfaces de programación de aplicaciones (API), facilitando la integración con otras herramientas y sistemas.

TeseoETL no solo optimiza la gestión de datos, sino que también potencia la generación de conocimiento estratégico, apoyando la toma de decisiones basadas en evidencia en diversos contextos.

¿Cómo funciona TeseoETL?

El trabajo con datos en el mundo real es un conjunto de problemas complejos y simultáneos. TeseoETL es un sistema completo de procesamiento de datos que implementa una solución para cada uno de los siguientes problemas:

- Consumo automatizado de fuentes de datos diversas
- Estandarización y consolidación de información, incluyendo extracción de texto (de imágenes o documentos)
- Actualización continua
- Visualización de grandes cantidades de datos
- Consumo de datos vía API

En este documento se describe en detalle la arquitectura técnica de TeseoETL, incluyendo los requerimientos de hardware y software, y la implementación basada en contenedores. Además, se exploran las funciones y responsabilidades de cada componente dentro de este motor de análisis de datos.

Este documento está destinado a desarrolladores, ingenieros y arquitectos de sistemas que buscan una comprensión básica de la infraestructura que sustenta esta solución.

Componentes

El sistema TeseoETL tiene cuatro componentes principales e independientes que desempeñan las funciones principales del sistema interactuando entre sí. En esta sección explicaremos cada uno de estos componentes, sus funciones principales y las responsabilidades de cada uno dentro del sistema completo.

Orquestador de Pipelines

El orquestador es el motor de los procesos ETL definidos dentro de Teseo. Cada proceso ETL (también llamado pipeline o flujo) se compone por una serie de tareas ejecutadas en serie que consumen una fuente de datos y producen un dataset transformado.

El ambiente en el cual se definen y ejecutan estos pipelines es Apache NiFi. TeseoETL aprovecha este software para implementar una estrategia de ETL basada en la experiencia acumulada tras importar cientos de fuentes de datos.

Las tareas que componen los flujos (también llamadas procesadores) son unidades atómicas de procesamiento de datos que se interconectan para llevar los datos de un lugar a otro en el sistema. Estos procesadores pueden ejecutar tareas sencillas como una descarga de un archivo o una conversión entre formatos, pero también se pueden utilizar para interactuar con los demás componentes del sistema.

Manejo secuencial de datos

Los procesos de ETL realizan un manejo secuencial de los datos; es decir, que los datos fluyen de principio a fin del pipeline en un orden determinado. El sistema cuenta con una serie de scripts en Node.js para consumir, transformar o cargar datos siguiendo esta lógica.

TeseoETL incluye varios scripts listos para su uso directo desde el orquestador. Estos scripts ejecutan funciones básicas como la carga de datos transformados al data warehouse o transformaciones y limpieza de datos. Además, el formato de scripts permite que el sistema sea fácilmente extensible para realizar otros procesos sobre los datos.

Funciones y responsabilidades

- Definición de pipelines de procesamiento de datos mediante la interconexión de procesadores
- Interfaz gráfica de creación y edición de pipelines
- Automatización y ejecución periódica
- Monitoreo de flujo de los datos
- Consumo directo de fuentes de datos
- Transformación, estandarización y consolidación de fuentes
- Alimentación del data warehouse

Procesador de Documentos

El proceso de extracción de texto a partir de documentos no estructurados se realiza en su propio componente, basado en el software Apache Tika. En términos de infraestructura, este componente vive al mismo nivel que el orquestador. Sin embargo, conceptualmente se puede considerar al procesador de documentos como una funcionalidad más del orquestador, habilitando la inclusión de este proceso dentro de los pipelines.

El procesador de documentos ejecuta una tarea compleja: recibe documentos elaborados por humanos para consumo por humanos, y realiza una extracción de texto sin importar el formato del documento. Posee la capacidad de extraer texto de formatos diversos de documentos (documentos de texto, hojas de cálculo) e incluso permite realizar OCR (Optical Character Recognition) sobre documentos escaneados e imágenes.

Funciones y responsabilidades

- Extraer texto de casi cualquier tipo de documento
- Realizar OCR sobre documentos escaneados e imágenes
- Habilitar el procesamiento de documentos dentro del orquestador

Data Warehouse

El data warehouse de TeseoETL está implementado sobre Opensearch, un motor de base de datos no estructurados con gran flexibilidad para almacenar cualquier tipo de datos y alta velocidad de búsqueda.

Funciones y responsabilidades

- Almacenar de manera persistente todos los datos transformados en un repositorio centralizado
- Capacidad para realizar consultas rápidas y eficientes sobre los datos
- Poner los datos a disposición del sistema de visualización y análisis de datos
- Permitir la integración de sistemas tercerizados mediante una API

Business Intelligence

La capa de Business Intelligence, basada en OpenSearch-Dashboards, facilita el análisis de datos mediante una poderosa interfaz de búsqueda sobre los conjuntos de datos y otra interfaz intuitiva de visualización exploratoria y creación de reportes. Este es el componente mediante el cual la mayoría de usuarios finales pueden interactuar con TeseoETL, y permite implementar una capa de seguridad basada en permisos y roles para definir tipos de usuarios y granularidad de acceso a los datos.

Análisis de Datos

A través de la interfaz de búsqueda, el usuario puede realizar consultas avanzadas sobre cualquier conjunto de datos integrado utilizando filtros simples de texto que se ejecutan sobre millones de documentos en una fracción de segundo. Además de búsqueda de texto también es posible la creación de filtros complejos con cientos de condiciones, visualizar los resultados en un formato tabular con los campos preferidos por el usuario, y exportación directa de los resultados en formato CSV.

Visualización y Reportes

Para el análisis avanzado de datos, el componente posee una interfaz de visualización exploratoria y creación de reportes a través de la cual los usuarios

pueden explorar los conjuntos de datos mediante decenas de visualizaciones predefinidas y una interfaz sencilla de drag and drop.

Funciones y responsabilidades

- Realizar consultas simples o complejas sobre datasets completos, desde búsquedas de texto simples hasta combinaciones complejas de filtros
- Ejecutar visualizaciones de datos exploratorias para permitir análisis complejos
- Creación de reportes y dashboards de manera sencilla utilizando una interfaz visual de drag and drop
- Implementar el control de acceso a los datos mediante un sistema de espacios, roles y permisos