

ETL: Tecnologías para procesos de extracción, transformación y carga de datos como solución para integrar y gestionar datos de múltiples fuentes

ETL: Technologies for extracting, transforming, and loading data as a solution for integrating and managing data from multiple sources

Oscar Alejandro López Clímaco ^{1,2}

Correspondencia: lc16024@ues.edu.sv

1 Facultad Multidisciplinaria Paracentral de la Universidad de El Salvador 

2  ORCID: <https://orcid.org/0009-0009-1652-1090>

RESUMEN

Mostrar el impacto de las tecnologías ETL en el entorno empresarial, se incluirá una investigación de la documentación oficial sobre la tecnología de extracción, transformación y carga de datos. Además, se realizará una investigación documental sobre la herramienta Pentaho, que se utilizará para ejecutar procesos (inserción, extracción, actualización y transformación de datos, ejecución de scripts de programación dentro del proceso ETL, etc.) en un ambiente de prueba, con el fin de comprender mejor el funcionamiento de esta herramienta y de las tecnologías ETL.

Palabras claves: ETL, Pentaho, extracción, transformación, carga, scripts.

ABSTRACT

To demonstrate the impact of ETL technologies within the business environment, this study includes a review of official documentation regarding Extraction, Transformation, and Loading (ETL) technologies. Additionally, documentary research will be conducted on the Pentaho tool, which will be used to execute processes (such as data insertion, extraction, updating, and transformation, as well as the execution of programming scripts within the ETL workflow) in a test environment. The goal is to gain a deeper understanding of the functionality of both this tool and ETL technologies in general.

Keywords: ETL, Pentaho, extraction, transformation, loading, scripts.

INTRODUCCIÓN

Uno de los temas de creciente relevancia a nivel internacional es el avance digital, especialmente en el área de la inteligencia artificial. Este fenómeno ha suscitado un amplio debate, ya que la sociedad parece depender de la tecnología en su vida cotidiana. Si dichas herramientas se detuvieran, se generaría un caos general. El ser humano se ha vuelto tan dependiente de la innovación que hasta las tareas más simples se realizan con su apoyo.

El campo de la tecnología es amplio y no puede abarcarse en su totalidad en un solo documento. Existen múltiples tecnologías que requieren un estudio dedicado para cada ámbito. Estas están diseñadas para automatizar actividades humanas y optimizar el tiempo; sin embargo, muchas de ellas han sido creadas específicamente para el ámbito empresarial, un pilar fundamental de la subsistencia humana. Hoy en día, incluso las pequeñas empresas utilizan tecnologías para agilizar sus procesos y enfrentar la creciente competitividad.

Es crucial mencionar que esta demanda empresarial fue lo que generó la necesidad de desarrollar la tecnología ETL, la cual se discutirá a lo largo de este documento. Se presentarán ejemplos prácticos y casos simulados para ilustrar cómo esta tecnología mejora la eficiencia de los procesos empresariales.

METODOLOGÍA

Como punto de partida, se responderán preguntas relacionadas con esta tecnología: ¿qué es ETL?, ¿quién lo creó?, ¿cómo surge?, ¿qué beneficios tiene?, ¿cómo funciona?, ¿quién lo utiliza?, ¿qué herramientas existen?

¿Qué es ETL?

Las definiciones de ETL varían entre diferentes fuentes, pero todas coinciden en que no es una tecnología centralizada para una empresa específica. Es una tecnología aplicable a diversas organizaciones. A continuación, se presentan definiciones de ETL de empre-

sas reconocidas:

- **Google:** El proceso de extracción, transformación y carga (ETL) es el método tradicionalmente utilizado en las organizaciones para combinar datos de varios sistemas en una base de datos, un almacén o un lago de datos. Este proceso permite almacenar datos antiguos o agregar nuevos datos para analizarlos y fundamentar decisiones empresariales.
- **Oracle:** La extracción, transformación y carga (ETL) es el proceso que las organizaciones impulsadas por datos utilizan para recopilar datos de distintas fuentes y reunirlos con el fin de facilitar el descubrimiento, la generación de informes, el análisis y la toma de decisiones. Los orígenes de los datos pueden variar en tipo, formato, volumen y fiabilidad, por lo que es necesario procesarlos para que sean útiles.
- **IBM:** ETL (extraer, transformar, cargar) es un proceso de integración de datos que combina, limpia y organiza datos de múltiples fuentes en un conjunto único y consistente para su almacenamiento en un almacén de datos o lago de datos.
- **Microsoft:** ETL es una canalización de datos utilizada para recopilar datos de varios orígenes, transformarlos según las reglas de negocio y cargarlos en un almacén de datos de destino.

DESARROLLO

Definición de Ralph Kimball

El libro *The Data Warehouse Toolkit* de Ralph Kimball (Kimball & Ross, 2002) menciona brevemente las tecnologías ETL en varios de sus capítulos, ya que están ligadas a los procesos que alimentan los almacenes de datos. En estas menciones, el autor define cada proceso de ETL de manera que se puede comprender mejor en qué consiste esta tecnología. A continuación, se presentan sus definiciones.

La extracción es el primer paso en el proceso de introducir datos en el entorno del almacén de datos. Extraer significa leer y comprender la fuente de datos, así como copiar los datos necesarios para el almacén en

el área de preparación para una mayor manipulación.

Una vez que los datos se extraen al área de preparación, se llevan a cabo numerosas transformaciones, como limpiar los datos (corregir errores ortográficos, resolver conflictos de dominio, tratar elementos faltantes o estandarizar formatos), combinar datos de múltiples fuentes, eliminar duplicados y asignar claves de almacén. Todas estas transformaciones son precursoras de la carga de datos en el área de presentación del almacén de datos.

Independientemente de si se trabaja con archivos planos o estructuras de datos normalizadas en el área de preparación, el paso final del proceso ETL es la carga de datos. La carga en el entorno del almacén de datos se presenta típicamente en forma de tablas dimensionales de calidad garantizada en instalaciones de carga a granel de cada data mart. El data mart de destino debe luego indexar los nuevos datos para optimizar el rendimiento de las consultas. Una vez que cada data mart ha sido recién cargado, indexado y provisto de agregados apropiados, se notifica a la comunidad de usuarios sobre la publicación de los nuevos datos.

La publicación incluye comunicar cualquier cambio ocurrido en las dimensiones subyacentes y en los nuevos supuestos introducidos en los hechos medidos o calculados.

Si consideramos las definiciones anteriores, es evidente que, sin importar el enfoque de cada autor, todos coinciden en que ETL es una tecnología que, independientemente de su aplicación, sigue un proceso similar.

Sin más que agregar sobre la definición de esta tecnología, ahora se abordará la siguiente pregunta:

¿Quién creó la tecnología ETL?

Cuando se habla de una tecnología o herramienta, es fundamental considerar su creador. A menudo, para desarrollar una tecnología se requieren entornos que faciliten la creación de la herramienta y, por supuesto, lenguajes de programación. Al hablar de ETL como tecnología, se puede pensar que detrás de ella hay un creador. Sin embargo, ETL no surgió como una herramienta programada, sino que se originó a partir de

técnicas, metodologías y buenas prácticas relacionadas con el almacenamiento de grandes volúmenes de datos, que consisten en la extracción, transformación y carga de datos.

Dado que ETL no surgió como una herramienta específica, no hay un único creador o empresa que la haya desarrollado. Esta tecnología es el resultado de una evolución que ha ocurrido a lo largo del tiempo, a medida que surgió la necesidad de gestionar grandes volúmenes de datos.

No obstante, se puede atribuir a Ralph Kimball el haber sentado las bases para que las empresas comenzaran a implementar estas metodologías. En las décadas de 1980 y 1990, Kimball trabajó para formalizar sus técnicas de integración de datos, lo que llevó a las empresas a desarrollar aplicaciones y herramientas que sistematizaran estas metodologías. Esto es lo que ha permitido el desarrollo de herramientas específicas para ejecutar procesos ETL.

¿Cómo surgió?

Las empresas modernas enfrentan el desafío de procesar grandes volúmenes de datos generados por sus operaciones diarias. Manejar estos datos utilizando métodos obsoletos o manuales sería extremadamente ineficiente y podría llevar más de una semana, lo que podría estancar el crecimiento de la empresa. Para abordar este desafío, es esencial implementar la tecnología ETL (extracción, transformación y carga).

¿Qué beneficios tiene?

Tomando como referencia algunas de las empresas que antes fueron citadas para conocer la definición de ETL, nos apoyamos en ellas para presentar los beneficios que tiene la implementación de esta tecnología.

ETL sirve como puente que conecta vastas reservas de datos generados por organizaciones con conocimientos procesables. Su importancia radica no solo en el gran volumen de datos que maneja sino también en la precisión y eficiencia con la que los gestiona.

¿Cómo funciona ETL?

ETL es un proceso de tres pasos:

- **Extracción de Datos**

El proceso comienza extrayendo datos sin procesar de fuentes de datos relevantes, incluidas bases de datos, archivos, etc. Los datos extraídos se almacenan en una zona de aterrizaje, también llamada área de preparación. Un área de preparación es un almacenamiento intermedio donde los datos solo se almacenan temporalmente.

- **Carga de datos**

Cargar datos en el sistema de destino es el último paso en ETL. Los datos transformados se mueven desde el área de preparación a un sistema de almacenamiento permanente, como un almacenamiento de datos.

Los datos cargados están bien estructurados, lo que los profesionales de datos y los usuarios comerciales pueden utilizar para sus necesidades de análisis y BI. Dependiendo de los requisitos de su organización, puede cargar datos de diversas formas.

- **Streaming**

En este caso, los datos se cargan casi en tiempo real o en tiempo real a medida que se vuelven disponibles. A menudo se utiliza para transmitir fuentes de datos y es ideal para aplicaciones que requieren datos actualizados para análisis o toma de decisiones. La transmisión de datos de actividad del usuario a un panel de análisis en tiempo real es un ejemplo común.

¿Quién lo utiliza?

Lo utilizan todas aquellas empresas que necesitan extraer, transformar y almacenar datos en sus almacenes de datos para agilizar su negocio y mantenerse en constante crecimiento a medida que los mercados se van diversificando.

¿Qué herramientas existen?

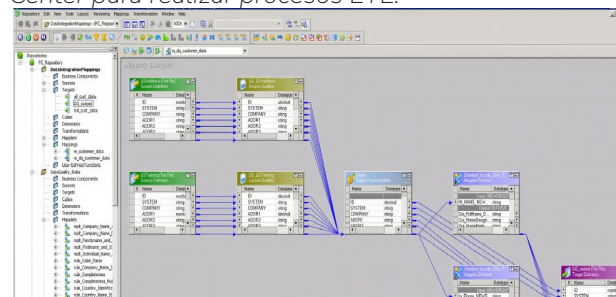
Informática Power center

Informática PowerCenter es una de las mejores herramientas ETL del mercado. Dispone de una am-

plia gama de conectores para lagos y almacenes de datos en la nube, como AWS, Azure, Google Cloud y Salesforce. Sus herramientas con poca o ninguna programación están diseñadas para ahorrar tiempo y simplificar los flujos de trabajo.

Figura1

Representa la herramienta gráfica Informática PowerCenter para realizar procesos ETL.

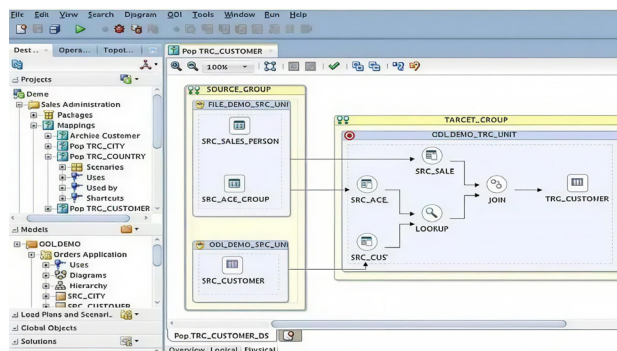


Oracle Data Integrator

Oracle Data Integrator es una herramienta ETL que ayuda a los usuarios a construir, implementar y gestionar almacenes de datos complejos. Viene con conectores listos para usar con muchas bases de datos, como Hadoop, EREP, CRM, XML, JSON, LDAP, JDBC y ODBC.

Figura2

Herramienta gráfica ETL de Oracle Data Integrator



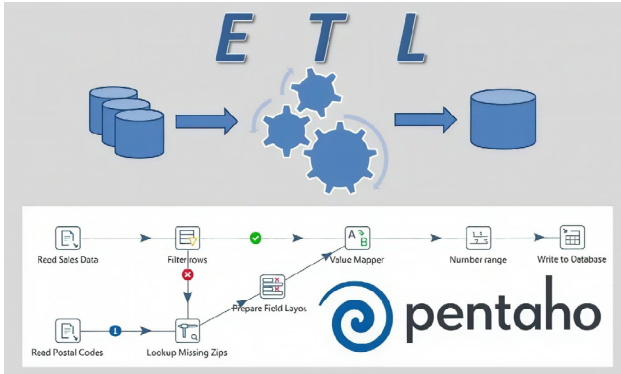
Pentaho Data Integration

Pentaho Data Integration (PDI) es una herramienta ETL ofrecida por Hitachi. Captura datos de diversas fuentes, los limpia y los almacena en un formato uniforme y coherente. Los usuarios pueden diseñar trabajos y transformaciones de datos utilizando el cliente PDI, Spoon, y luego ejecutarlos utilizando Kit-

chen. Por ejemplo, el cliente PDI puede utilizarse para ETL en tiempo real con Pentaho Reporting.

Figura 3

Representación del proceso ETL de la herramienta Pentaho



Hay una infinidad de herramientas más orientadas a la tecnología ETL, pero con las mostradas es más que suficiente para darnos cuenta que no solo una empresa sino muchas se han dedicado a la creación de herramientas para apoyar estos procesos y metodologías de ETL.

RESULTADOS

Se presenta a continuación un ejemplo de cómo funciona ETL con la herramienta Pentaho.

Para el siguiente ejemplo se utilizará una API REST.

Se creará la siguiente estructura y se guardará el Job con el nombre de “job_example4”.

Figura 4

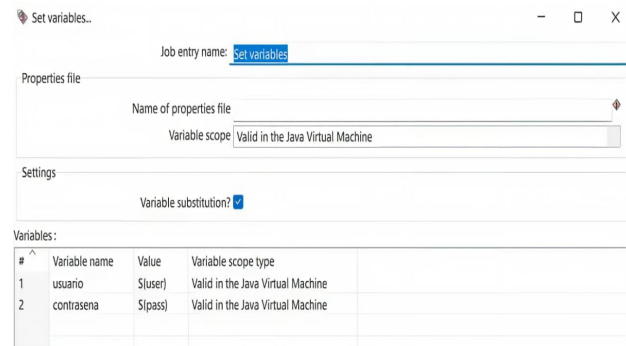
Flujo de ejecución de tareas para la transformación de datos



Se crean las siguientes variables en el paso de “Set variables”:

Figura 5

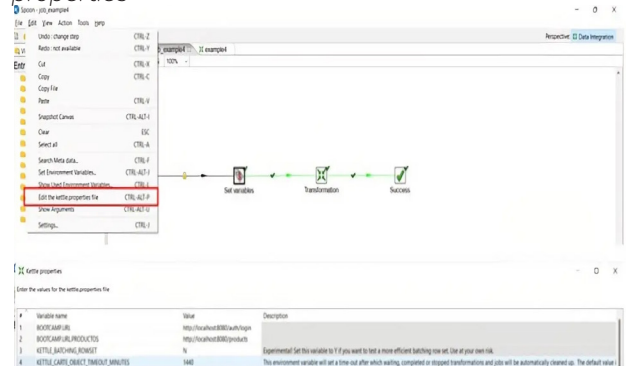
Definición de variables de entorno en el Job Entry



Como siguiente paso definiremos dos variables a utilizar a nivel del kettle.properties.

Figura 6

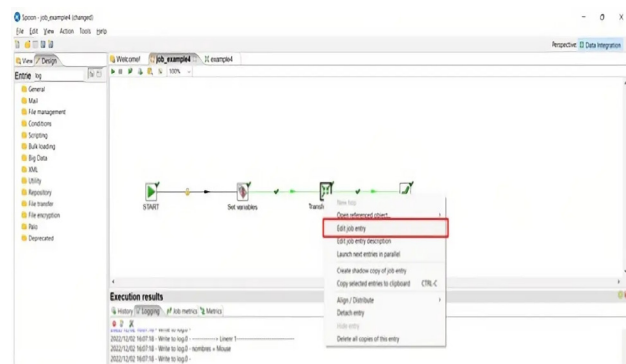
Acceso y edición del archivo de configuración kettle.properties



Se creará una nueva transformación y se guardará con el nombre de “example4”, posteriormente asociar la transformación al Job creado anteriormente.

Figura 7

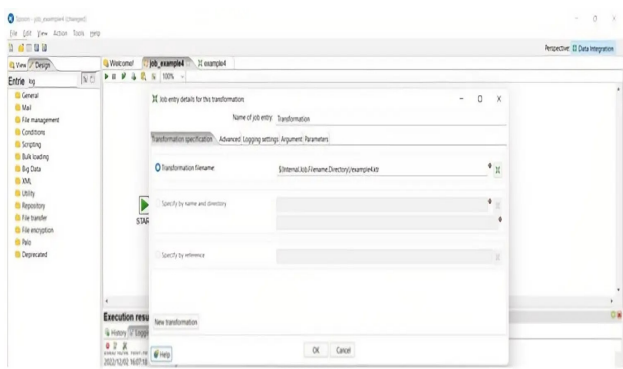
Asociación de la transformación al Job Entry



Asociamos la transformación:

Figura 8

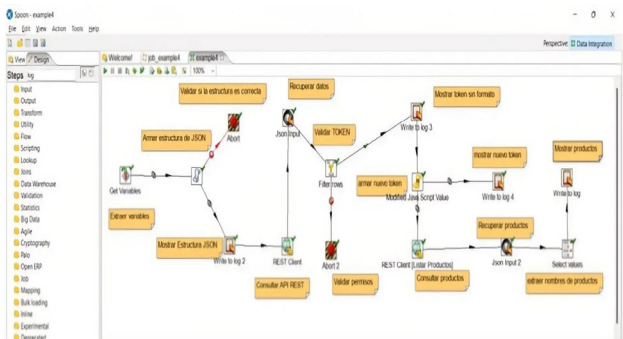
Especificación de la ruta de la transformación en el Job Entry



El siguiente paso es definir la estructura de la transformación, la cual quedará de la siguiente manera:

Figura 9

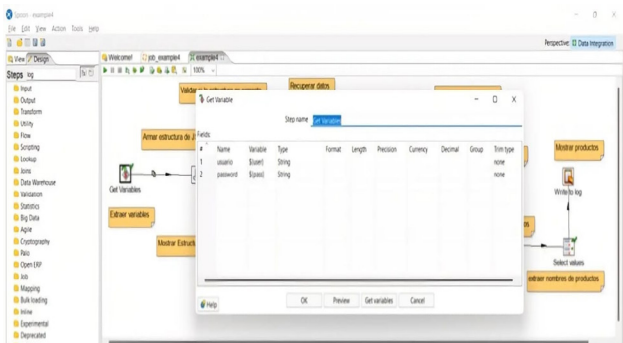
Estructura lógica de la transformación de datos en Spoon



El primer paso será extraer las variables que se definieron en el Job, se debe de extraer el usuario y la contraseña.

Figura 10

Extracción de variables en el paso Get Variables

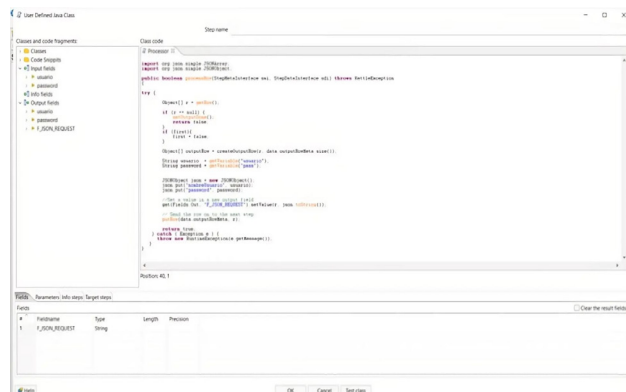


Luego de extraer las variables se debe de armar el for-

mato del archivo JSON que se enviará en la petición.

Figura 11

Construcción del objeto JSON mediante User Defined Java Class



Código Fuente

```
import org.json.simple.JSONArray;
import org.json.simple.JSONObject;

public boolean processRow(StepMetaInterface smi,
StepDataInterface sdi) throws KettleException {

    try {

        Object[] r = getRow();

        if (r == null) {

            setOutputDone();

            return false;

        }

        if (first) {

            first = false;

            Object[] outputRow = createOutputRow(r, data.
outputRowMeta.size());

            String usuario = getVariable("usuario");
            String password = getVariable("pass");
            JSONObject json = new JSONObject();
            json.put("nombreUsuario", usuario);
```

```

        json.put("password", password);

        // Set a value in a new output field
        get(Fields.Out, "F_JSON_REQUEST").setValue(r,
        json.toString());

        // Send the row on to the next step.

        putRow(data.outputRowMeta, r);

        return true;

    } catch (Exception e) {

        throw new RuntimeException(e.getMessage());

    }

}

```

Recordar agregar la variable de salida.

Figura 12

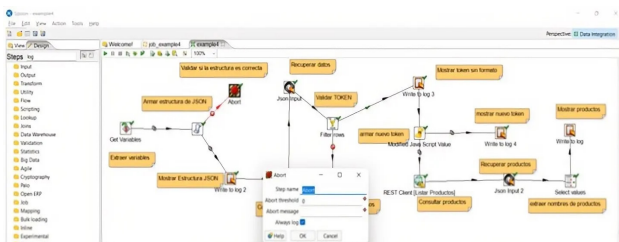
Definición de los campos de salida en el REST Client

Fields				
#	Fieldname	Type	Length	Precision
1	F_JSON_REQUEST	String		

Si el formato no cumple la estructura deseada se abortará el proceso:

Figura 13

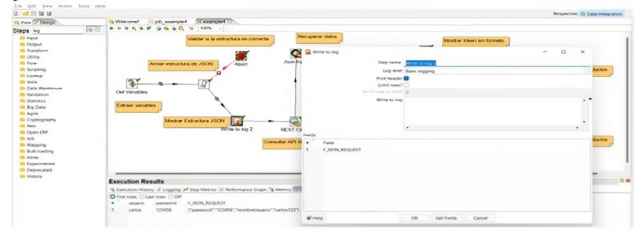
Control de flujo mediante validación de código de estado HTTP



Si el formato es correcto recuperaremos la información y se mostrará en consola de la siguiente forma:

Figura 14

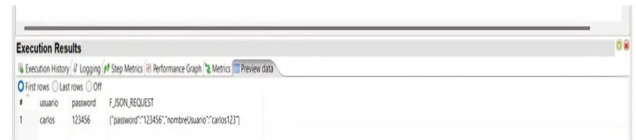
Extracción de campos mediante el componente JSON Input



Hasta este punto se ha formado la estructura del json, se ejecuta la transformación y se valida la estructura.

Figura 15

Generación de registros de confirmación mediante
Write to log

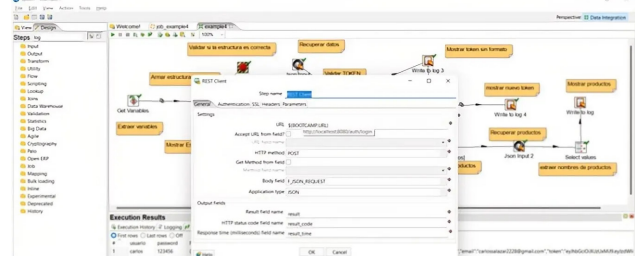


Implementación de API REST:

- Para la URL se utilizará la variable “\${BOOTCAMP_URL}”.
- El método HTTP será tipo POST.
- El “body” será la variable F_JSON_REQUEST definida en los pasos anteriores.
- Las variables serán: “result”, “result_code” y “result_time”.

Figura 16

Ejecución y verificación del flujo de trabajo (Job) en Spoon



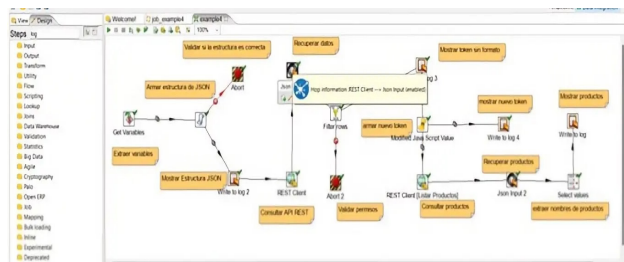
*Solo se definirán parámetros en la pestañan de “General” *

El siguiente paso será recuperar la información alma-

cenada en la variable “result”:

Figura 17

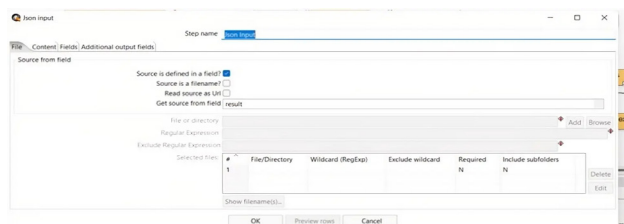
Visualización del registro de eventos (Logging) del proceso



- Definir un nombre al paso.
- Marcar el check de “Source is defined in a field”.
- Agregar la variable donde se extraerá la información.

Figura 18

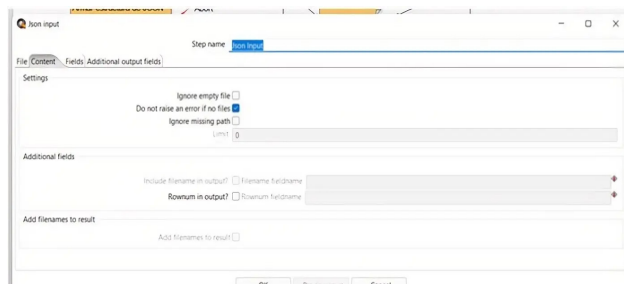
Métricas de rendimiento y flujo de datos por componente



Pestaña Content.

Figura 19

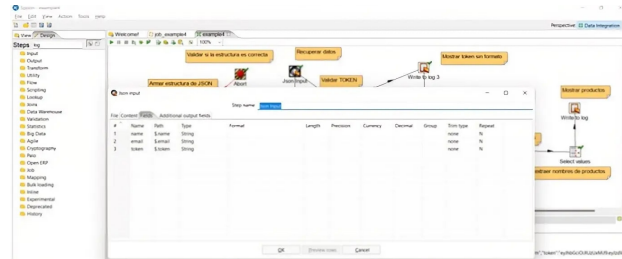
Selección y formateo de campos finales mediante Select values



Definir campos.

Figura 20

Configuración del mensaje de salida final en el componente Write to log



El siguiente paso es validar si el token es válido.

Figura 21

Confirmación de ejecución exitosa de la transformación en Spoon

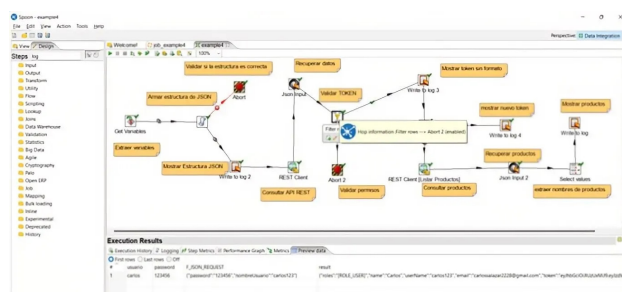


Figura 22

Registro del token obtenido en la consola de logs



Si el resultado del filtro devuelve falso se debe abortar.

Figura 23

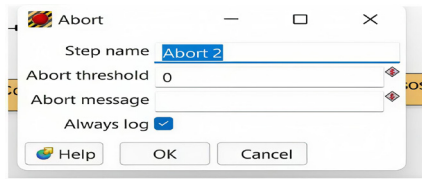
Ventana de confirmación de ejecución exitosa del Job



Si el filtro devuelve verdadero se procede a mostrar los datos en consola:

Figura 24

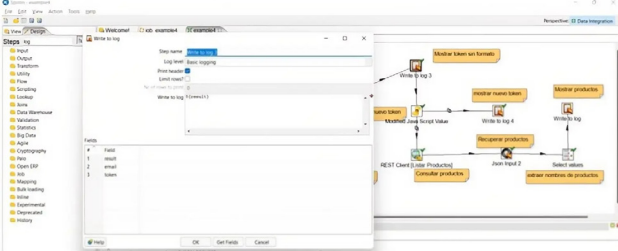
Estado final del Job tras una ejecución satisfactoria



Prueba de ejecución de transformación hasta el momento.

Figura 25

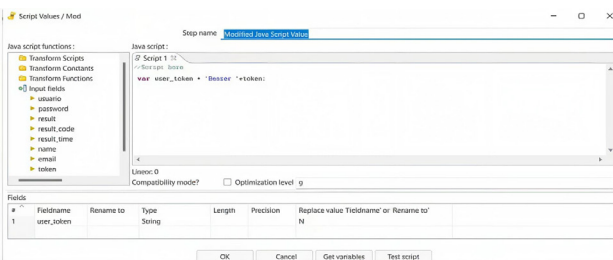
Registro detallado de ejecución del Job principal



El objetivo de esta transformación es listar los productos que se encuentran en la base de datos, pero para listar los productos se debe validar el usuario con un token de jwt, en el siguiente paso se definirá la estructura valida de token que admite jwt.

Figura 26

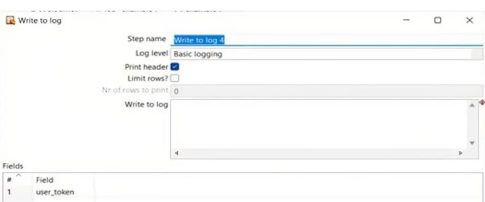
Métricas de rendimiento y tiempos de ejecución del Job



Mostrar el token generado.

Figura 27

Historial de ejecuciones y estados de finalización del Job



Prueba de ejecución hasta el momento:

Figura 28

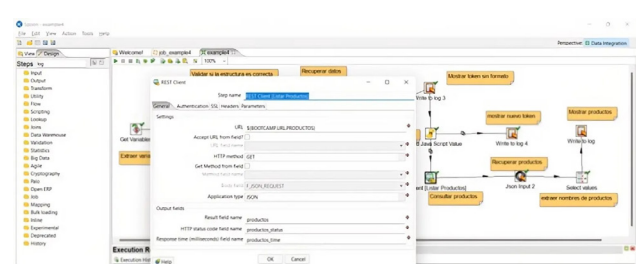
Detalle cronológico de la ejecución y finalización del Job principal



Implementación de API REST para listar productos.

Figura 29

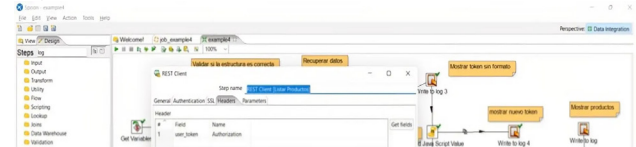
Entorno de diseño y lienzo de trabajo en Pentaho Data Integration



- *definir url: se debe tomar la variable creada en los primeros pasos, método HTTP GET*.
- *tipo de aplicación: JSON*.
- Añadir token en el apartado de "Headers".

Figura 30

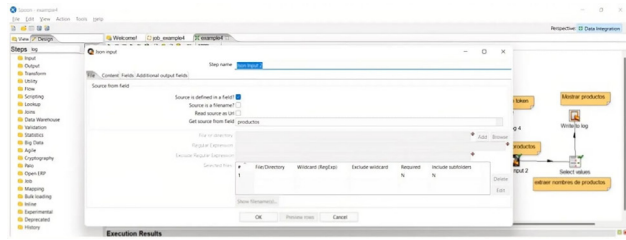
Previsualización de datos en tiempo real durante el flujo



Recuperar respuesta en "Json input".

Figura 31

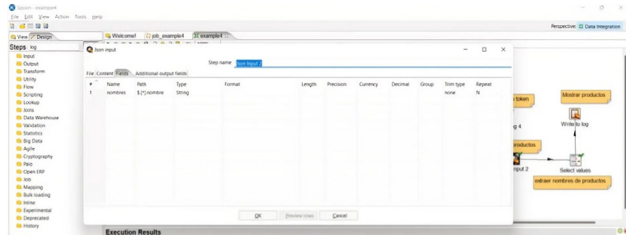
Configuración de la condición lógica en el componente Filter rows



Obtener los nombres de productos:

Figura 32

Interfaz de extracción y salida de datos estructurados (JSON input)



Seleccionar valores.

Figura 33

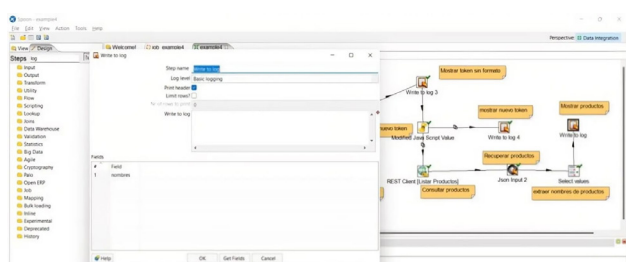
Selección y edición de metadatos de los campos de salida.



Mostrar productos.

Figura 34

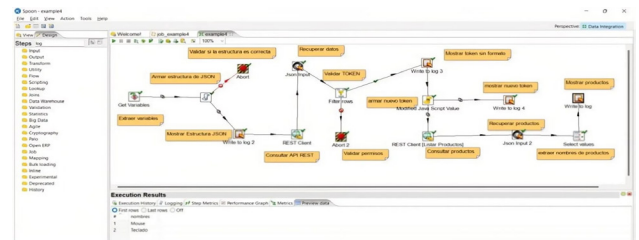
Configuración del registro de salida para la visualización de productos.



Resultado final.

Figura 35

Resultado final de la integración y visualización de datos procesados



CONCLUSIONES

Importancia de ETL: La tecnología ETL facilita el procesamiento y manejo eficiente de grandes volúmenes de datos desde diversas fuentes. Su implementación permite a las empresas integrar datos de manera más efectiva, optimizando la toma de decisiones y mejorando la productividad.

Funcionalidad ETL: ETL se compone de tres fases principales: extracción de datos de múltiples fuentes, transformación de esos datos para cumplir con los requisitos comerciales y, finalmente, carga en un almacén de datos. Esto mejora la capacidad de análisis y permite a las empresas realizar estudios más avanzados, como análisis predictivos y aprendizaje automático.

Beneficios empresariales: La tecnología ETL permite una vista unificada de la información, lo que facilita la toma de decisiones estratégicas. Además, proporciona la capacidad de manejar datos históricos, esenciales para el análisis de tendencias y la eficiencia operativa. Automatiza procesos que de otro modo serían manuales, reduciendo costos y mejorando la calidad de los datos.

Herramientas ETL: Existen múltiples herramientas de software disponibles en el mercado, como Pentaho, AWS Glue, Informatica PowerCenter y Apache Airflow, que permiten implementar procesos ETL. Cada herramienta ofrece funcionalidades específicas, desde conectores con bases de datos hasta integración con servicios en la nube.

Se concluye que las tecnologías ETL son un gran be-

neficio para las empresas ya que les facilita el procesamiento de datos en tiempos cortos permitiendo que la productividad de las empresas se agilice y también se almacenen los datos de forma segura y clara.

REFERENCIAS

Kimball, R., & Ross, M. (2002). The data warehouse toolkit. Wiley Computer Publishing.

Google Cloud. (n.d.). ¿Qué es ETL? Google Cloud. <https://cloud.google.com/learn/what-is-etl?hl=es>

Oracle. (n.d.). ¿Qué es ETL? Oracle. <https://www.oracle.com/lad/integration/what-is-etl/>

IBM. (n.d.). ¿Qué es ETL? IBM. https://www.ibm.com/topics/etl?mhsrc=ibmsearch_a&mhq=etl

Microsoft. (n.d.). ¿Qué es ETL? Microsoft Azure. <https://learn.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl>

Astera. (n.d.). ETL: ¿Qué es y cómo funciona? Astera. <https://www.astera.com/es/type/blog/etl/>

Autor desconocido. (n.d.). The data warehouse toolkit [PDF]. <https://ia801609.us.archive.org/14/items/the-data-warehouse-toolkit-kimball/The%20Data%20Warehouse%20Toolkit%20-%20Kimball.pdf>

DataCamp. (n.d.). A list of the 16 best ETL tools and why to choose them. DataCamp. https://www.datacamp.com/es/blog/a-list-of-the-16-best-etl-tools-and-why-to-choose-them?utm_source=google&utm_medium=paid_search&utm_campaignid=20616617505&utm_adgroupid=154290359917&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=698933344247&utm_targetid=dsa-2220216600867&utm_loc_interest_ms=&utm_loc_physical_ms=9069902&utm_content=DSA~blog~Data-Engineering&utm_campaign=231025_1-sea~dsa~tofu_2-

b2c_3-es-lang_4-prw_5-na_6-na_7-le_8-pdsh-go_9-nb-s_10-na_11-namay24&gad_source=1&gclid=Cj0KCQjw6uWyBhD1ARIsAIMcADosQJsU3_tiret-S6Uhk8sS4hsSnuepGpC18zBqJJ15wnOm5A9cmolaAkVhEALw_wcB

