UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

GRADO EN INGENIERÍA DE SISTEMAS DE LA INFORMACIÓN

Creación de un dashboard para la monitorización de indicadores de desarrollo sostenible

Autor: Guillermo Ramírez Barber.

Tutor: Borja Bordel Sánchez

A mi abuelo Rafael, que con lápiz y papel me transmitió su pasión por las matemáticas y la lógica. En sus enseñanzas comenzó este viaje hacia la informática.

RESUMEN

Gracias al desarrollo que ha ido teniendo el empleo de los datos en estos últimos años, se han podido desarrollar diversos procesos que ayudan a la extracción, tratamiento y carga de los mismos. En este Trabajo de Fin de Grado me he centrado en el desarrollo de un proceso ETL que permita integrar datos abiertos proporcionados por la FAO (Organización de las Naciones Unidas para la Alimentación y Agricultura) en una base de datos PostgreSQL. Para ello, se ha empleado la herramienta Apache Hop, encargada de la extracción, transformación y carga de los mismos, automatizando la limpieza, formateo e integración de distintas fuentes en una estructura relacional.

Una vez cargados los datos en la base de datos, se ha utilizado Power BI como herramienta de análisis y visualización. Gracias a las funcionalidades de las que dispone, ha sido posible representar gráficamente los indicadores clave de emisiones, absorciones y superficie forestal, permitiendo de esta forma un análisis más accesible, visual y orientado a la toma de decisiones.

Durante todo el proyecto se ha trabajado con herramientas empleadas en entornos profesionales, abordando retos técnicos relacionados con la calidad del dato, la gestión de errores en procesos ETL y la visualización efectiva. El resultado es una solución integral que facilita el estudio de fenómenos medioambientales a partir de datos abiertos y actualizables.

Contenido

1. Introducción. 7
1.1. Contexto. 7
1.2. Planteamiento del problema y motivación. 9
1.3. Objetivos. 10
1.4. Organización del documento. 13
2. Estado de la Cuestión. 14
2.1. Introducción. 14
2.2. Procesos ETL. 16
2.3. Fases ETL. 21
Fase de Extracción. 21
Fase de transformación. 26
Fase de carga. 27
2.4. Data Warehouse. 29
2.5. Datos abiertos. 31
3. Herramientas. 32
3.1. Apache Hop. 32
3.2. PostgreSQL. 34
3.3. Power BI. 37
4. Desarrollo del proyecto. 39
4.1. Datos abiertos. 39
4.2. Diseño e Implementación del proceso ETL. 42
4.3. Creación Data Warehouse. 48
4.4. Framework destinado al análisis de datos. 51
5. Pruebas. 65
5.1. Pruebas de carga e integración de datos. 65
5.2. Pruebas con presentación de datos. 71
6. Acceso al fichero PBI. 74
7. Consideraciones Éticas, Legales y de Sostenibilidad. 76
8. Conclusiones. 78
ANEXO. BIBLIOGRAFIA. 81

Tabla de imágenes.

Ilustración 1: Fases del Proceso ETL. 21
Ilustración 2: Apache Hop 32
Ilustración 3: PostgreSQL 34
Ilustración 4: Consulta de versión MySQL 36
Ilustración 5: Connect to Database 36
Ilustración 6: Create DataBase 37
Ilustración 7: Power BI 37
Ilustración 8: Organización de las Naciones Unidas para la Alimentación y la agricultura 40
Ilustración 9: Carga Tabla Element 43
Ilustración 10: Carga Tabla Area 43
Ilustración 11: Carga Tabla Product 43
Ilustración 12: Carga Tabla Source 43
Ilustración 13: Carga Tabla Fire 44
Ilustración 14: Carga Tabla Forest 44
Ilustración 15: Conexión con fichero CSV Incendios 44
Ilustración 16: Conexión con fichero CSV Bosques 44
Ilustración 17: Renaming campos Tabla Element 45
Ilustración 18: Append Streams 45
Ilustración 19: Unique Rows 45
Ilustración 20: Connection to DBTFG 46
Ilustración 21: Insert/Update 46
Ilustración 22: Claves Emissions Land Use Forest 47
Ilustración 23: Connect to DataBase 48
Ilustración 24: Tablas definidas en DB 49
Ilustración 25: Contenido tablas por SQL 49
Ilustración 26: Contenido Tabla Emissions Land Use Fire 50
Ilustración 27: Pestaña Bosques Power BI 52
Ilustración 28: Claves superiores Bosques Power BI 53
Ilustración 29: Gráficos filtrados Bosques Power BI 54
Ilustración 30: Gráfico Superficie forestal Bosques Power BI 54
Ilustración 31: Gráfico Superficie forestal con filtro por país Bosques Power BI 55
Ilustración 32: Gráfico Superficie forestal con filtro por origen Bosques Porwer BI 55
Ilustración 33: Gráfico Emisiones Bosques Power BI 56
Ilustración 34: Gráfico Emisiones con filtros Bosques Power BI 57
Ilustración 35: Gráfico Emisiones por tipo Medición Bosques Power BI 57
Ilustración 36: Gráfico Emisones/Absorciones Bosques Power BI 58
Ilustración 37: Gráfico Emisones/Absorciones por Tipo Medición Bosques Power BI . 58
Ilustración 38: Pestaña Incendios Power BI 59
Ilustración 39: Claves superiores Incendios Power BI 59
Ilustración 40: Gráfico Superficie y biomasa quemada Incendios Power BI 61
Ilustración 41: Gráfico Tipo ecosistema afectado Incendios Power BI 62
Ilustración 42: Gráfico Emisiones CO2 Incendios Power BI 63
Ilustración 43: Gráfico Emisiones CH4 Incendios Power BI 64

Ilustración 44: Conexión relacional tablas Power BI 65
Ilustración 45: Error Carga datos 66
Ilustración 46: Correcta Carga datos 66
Ilustración 47: Error en campo no encontrado en Insert/Update 67
Ilustración 48: Error de campo no tratado en JavaScript 67
Ilustración 49: DataPreview del paso Select Values 68
Ilustración 50: Comparación de KPI Y2021 entre fuente de datos y BBDD 68
Ilustración 51: Ejemplo en el origen de datos decimales y miles 68
Ilustración 52: Código Java aplicado para la resolución del problema 69
Ilustración 53: Error log de campo no encontrado en Select Values 69
Ilustración 54: Renaming de los campos en Select Values 70
Ilustración 55: Aplicación del tipo de dato en Select Values tras renaming 70
Ilustración 56: Ejemplo error de búsqueda de campo en Insert/Update 70
Ilustración 57: Log en pipeline de ejecución 71
Ilustración 58: Total registros cargados en tabla de BBDD 71
Ilustración 59: Datos cargados en Power Bi 72
Ilustración 60: Filtros con múltiples opciones en Power BI 73
Ilustración 61: Acceso a la página web, marcando dónde debe accederse 74
Ilustración 62: Correo electrónico 74
Ilustración 63: Inicio de sesión en UPM 75
Ilustración 64: Importar informe 75

1. Introducción.

1.1. Contexto.

Gracias al auge de la información y al desarrollo de entornos que nos permiten el tratamiento de esta, podemos tomar decisiones más informadas y estratégicas, mejorar la eficiencia operativa y abordar desafíos globales con soluciones basadas en evidencias y análisis precisos.

Esto ha supuesto que, hoy en día, nuestra sociedad viva en una era de la información donde los datos tengan un valor incalculable. Su valor radica en la capacidad de transformar conocimiento en acciones que permiten impulsar el progreso y la sostenibilidad en nuestra sociedad.

El acceso a Internet ha ido experimentando un crecimiento muy significativo a lo largo de las últimas décadas, cambiando la forma que la sociedad tenía de poder trabajar, comunicarse y acceder a información en su día a día.

En la última edición del estudio anual realizado por *We Are Social*, una de las mayores referencias a la hora de analizar la actual realidad del sector digital en nuestra sociedad, se señala que el número de usuarios de internet en el mundo alcanzó los 5.350 millones de personas, suponiendo un 66.2% de la población mundial. El número de personas que tienen acceso a Internet incrementó en 97 millones, lo que supone un 1,8% respecto a 2023.

Con la disponibilidad de internet en nuestros hogares, oficinas e incluso en nuestros dispositivos móviles, un usuario medio pasa siete horas en línea todos los días. Por cada usuario, se crean más de 1.7MB de datos en cada segundo de cada día y no sólo se generan trabajando o acudiendo a fuentes de información disponibles en Internet. Al día se suben más de 500 horas de vídeo a YouTube, se envían más de 41

millones de mensajes de WhatsApp, se gastan más de 1 millón de dólares en compras online y se hacen más de 1 millón de videollamadas.

Toda esta información generada por cada usuario puede suponer un problema si no ha sido almacenada, administrada y mantenida de forma segura. Añadiendo que debido a la naturaleza de los datos éstos se encuentran dispersos. La recolección de estos y su tratamiento puede suponer una propuesta costosa para grandes entidades tecnológicas, así como para las gubernamentales. La disposición de dichos datos procedentes de los usuarios es un valor en activo muy grande.

Este tratamiento es una propuesta costosa debido a la gran cantidad de datos que deben tratarse. Sin embargo, permite organizar, limpiar y estructurar dicha información, transformándolos en información valiosa, accesible y sobre la que se pueden tomar acciones. Al integrar datos de diversas fuentes y asegurar que la calidad de estos es óptima, facilita el análisis profundo y la toma de decisiones con información veraz en campos tan diversos integrados en nuestra sociedad como la gestión medioambiental, la economía…

Por ello, la extracción y gestión de los datos es un proceso que se puede realizar de múltiples formas: el uso de scripts especializados en la captura de datos específicos, el empleo de APIs para acceder a datos de forma estructurada de aplicaciones y servicios, la integración de bases de datos relacionales y no relacionales para la consolidación de dicha información, etc.

A lo largo de este proyecto, nos centraremos en los procesos ETL demostrando cómo estos procesos pueden integrarse para optimizar la recolección, transformación, almacenamiento y tratamiento de datos.

1.2. Planteamiento del problema y motivación.

Tras poder observar que la recolección de los datos y el tratamiento de estos es un gran activo para la sociedad y diferentes entidades, las principales preguntas que nos surgen son

¿Cuándo debo realizar el tratamiento de dichos datos? ¿Qué finalidad tienen? ¿Quién debe tener acceso a ellos? ¿Dónde debo mantenerlos de forma segura y accesible?

Los procesos ETL (Extract, Transform and Load) ofrecen soluciones a dichas preguntas a través de:

1. **Cuando realizar el tratamiento de los datos**: Estos procesos deben llevarse a cabo de forma regular para asegurar que los datos sean los más actuales y relevantes. Dependiendo de la procedencia y naturaleza de estos, además de su organización, el tratamiento puede ser diario, semanal, mensual, anual o incluso en tiempo real.

2. **Finalidad de los datos**: El principal propósito debe definirse previamente para que los datos se encuentren correctamente tratados, con el objetivo de poder obtener dicha información con la mayor fidelidad posible. En función del propósito, los datos se emplearán para realizar análisis de tendencias, generación de informes, mejora de productos y/o servicios, toma de decisiones estratégicas influyentes en la sociedad, etc.

3. **Acceso a los datos**: La gestión de quién tiene acceso a los datos es un aspecto crucial para asegurar la privacidad y seguridad de estos. Solo personas autorizadas deben tener acceso controlado a esta información sensible. Se deben implementar políticas de acceso que estén basadas en roles y niveles de autorización (dividiendo el acceso, tratamiento e incluso la acción que se

puede tener sobre dicha información con el objetivo de evitar que dichos datos acaben proporcionando una información errónea).

4. **Almacenamiento de los datos**: Los datos deben almacenarse en sistemas que sean seguros y confiables. Dependiendo de la volumetría y la sensibilidad de los datos, pueden mantenerse almacenados en bases de datos locales, en la nube, etc. Es esencial que las plataformas de almacenamiento de datos que vayan a emplearse cumplan las principales normativas de protección de datos y nos ofrezcan respaldo en caso de algún peligro.

Los procesos ETL nos permiten estructurar y gestionar estos aspectos mediante el uso de herramientas que automatizan la extracción, transformación y carga de datos, garantizando que dicha información sea accesible, segura y útil en todo momento.

Este proceso, nos permitirá partir de un gran volumen de datos que no disponen de ningún valor aparente a un sistema de datos gestionado de forma correcta los cuáles supondrán un gran valor para el usuario final, permitiendo tomar decisiones a partir de la lectura de estos.

1.3. Objetivos.

Este proyecto tiene como objetivo el desempeño e implementación de un Framework que permita el análisis de una serie de datos predeterminados, así como la creación de un Data Warehouse mediante un proceso ETL.

Dichos datos permitirán el análisis del impacto de los incendios forestales en las superficies forestales de los países que componen la Unión Europea desde 2010 al 2021 así como las emisiones de CO2. Este conjunto de datos se ha generado en diversas fuentes como empresas y gobiernos, redes sociales (observaciones aportadas por usuarios que proporcionan información acerca del impacto de la

degradación forestal en cada país), etc. Esta información ha sido producida o recopilada por administraciones públicas que ponen a disposición de la sociedad para que su acceso se realice de forma libre, así como el empleo de estos.

Empleando datos de la FAOSTAT, el objetivo es evaluar la relación entre los incendios forestales, la pérdida de la superficie forestal y las emisiones de CO2 en diversos países. En análisis permitirá comparar dichos datos con el objetivo de identificación de patrones y proporcionar posibles acciones efectivas de prevención y protección del entorno. Por lo tanto, podremos dividir los objetivos específicos en los siguientes puntos:

1. **Comparación analítica de los datos:** Se podrá examinar las estadísticas de incendios, emisiones/absorciones de CO2 y superficie forestal entre los países que componen la Unión Europea (dicha superficie podrá ser analizada a partir de un porcentaje de devaluación de esta en un rango de 11 años).
2. **Evaluación de los impactos:** A través de este análisis podremos identificar el impacto directo de los incendios en el entorno forestal, así como las emisiones producidas de carbono en función del tipo de medición (desde el carbono producido por los incendios hasta la degradación continua de los bosques desde una o varias etapas de los bosques o el reemplazo de estos).
3. **Propuestas de prevenciones activas:** Sugerencia de medidas para reducir la degradación del entorno y protección de este.
4. **Fomento de la toma de decisiones informada:** Proveer a los responsables gubernamentales de información clave para la planificación de políticas activas medioambientales.

A través de este análisis y de la implementación de futuras decisiones a nivel gubernamental, la sociedad podrá experimentar una serie de beneficios finales a lo largo de los años que permitirá tener un mejor cuidado y desarrollo del medioambiente:

- **Toma de decisiones informada:** Este proyecto ayudará a aportar información clave a los diversos gobiernos en función del clima de cada país, permitiendo diseñar estrategias efectivas de prevención, así como conservación.
- **Monitoreo y evaluación continua:** Facilitará el seguimiento de la efectividad de dichas políticas implementadas a lo largo del tiempo, permitiendo en un futuro poder actuar con antelación.

El proceso ETL se aplica en este proyecto para manejar los datos de manera estructurada y eficiente, permitiendo un análisis detallado. A continuación, se describirán las tres fases clave del proceso:

- **Extracción:** En esta fase, se realiza la extracción de los datos desde las fuentes originales (principalmente FAOSTAT). Estos datos pueden estar en formatos diversos como archivos CSV, Excel o bases de datos estructuradas. La extracción de estos se realizará de manera manual, asegurando que toda la información relevante para el análisis esté disponible y actualizada.
- **Transformación:** Una vez extraídos los datos, se pasará a esta fase donde se limpian, estandarizan y preparan para su análisis. Esta fase también puede implicar la creación de variables derivadas para facilitar el análisis, como porcentajes de devaluación de la superficie forestal.
- **Carga:** Finalmente, los datos transformados se cargan en un sistema de almacenamiento adecuado para el análisis. Esto puede incluir bases de datos SQL o sistemas en la nube. La carga de los datos en este paso es crucial para poder llegar a garantizar su disponibilidad para análisis posteriores. Los datos serán cargados en una única ocasión debido a que no se tratan de registros que experimenten modificaciones y supongan posteriores cargas delta. En este paso, se garantiza que los datos sean fácilmente accesibles para el usuario.

1.4. Organización del documento.

Este documento se encuentra dividido en los siguientes capítulos.

- **Capítulo 1:** Introducción. Donde se muestra la motivación del proyecto y los objetivos que se persiguen en este Trabajo de Fin de Grado.

- **Capítulo 2:** Estado de la cuestión. En este apartado se detallarán aquellos aspectos técnicos aplicados en los procesos ETL, el concepto de Datos Abiertos, así como los sistemas Data Warehouse.

- **Capítulo 3:** Herramientas. Se detallan las herramientas que se han empleado durante la realización del proyecto. Específicamente: pgAdmin, PostgreSQL y Power BI.

- **Capítulo 4:** Desarrollo del proyecto. Se detallan los trabajos que se han realizado durante el diseño y construcción del Framework de análisis de datos.

- **Capítulo 5:** Pruebas. Se realizan una serie de pruebas para asegurar el correcto funcionamiento del sistema, además de errores que han ido surgiendo durante el desarrollo del proyecto.

- **Capítulo 6:** Acceso al fichero PBI. Se ha desarrollado una guía para poder acceder a los gráficos sin necesidad de descarga o instalación de una aplicación.

- **Capítulo 7:** Conclusiones. Se explican las conclusiones, los objetivos logrados, adversidades encontradas, así como conocimientos adquiridos.

2. Estado de la Cuestión.

2.1. Introducción.

Hoy en día convivimos en una sociedad la cual genera millones y millones de datos los cuáles deben ser tratados para un posterior uso correcto. Estos datos pueden provenir de diversas fuentes como pueden ser redes sociales, aplicaciones de negocio, sistemas operacionales o incluso información delicada registrada por aplicaciones de terceros en Internet.
Esta información la cuál a menudo es dispersa y se presenta en formatos heterogéneos, requieren de procesos los cuáles permitan su correcta relación, así como su organización y almacenamiento de una forma eficiente con el objetivo de poder ser aprovechados en futuros análisis estratégicos.
En este contexto es donde adquieren gran importancia los conceptos fundamentales: Proceso ETL y los Data Warehouse.

El proceso ETL se trata de un conjunto de tareas, normalmente ejecutadas mediante aplicaciones software o ejecutadas manualmente mediante operadores del sistema, que se emplean para poder extraer, movilizar y transformar datos desde diversas fuentes hacia una serie de sistemas de almacenamientos como pueden llegar a ser los Data Warehouse.

Este proceso cobró gran popularidad en la década de los 70 debido a que las empresas empezaron a utilizar diversas fuentes de datos para almacenar información con gran valor para los negocios. Por lo tanto, se impuso la necesidad de integración de dicha información.

Debido a esta necesidad que apareció, el proceso ETL acabó convirtiéndose en el principal método empleado para la extracción de

datos de diversas fuentes y permitiendo transformarlos en información valiosa la cuál sería posteriormente cargada en destinos específicos.

Entre finales de la década de 1980 e inicios de la década de 1990, los Data Warehouse fueron creados como una arquitectura de almacenamiento de datos.

Este almacenamiento de datos llamado Data Warehouse (DW), es como un repositorio central. La información viene de una o diversas fuentes de datos donde posteriormente son tratados y transformados. Los usuarios finalmente pueden luego acceder a ellos como herramientas de Business Intelligence. Almacenando toda esta información en un único lugar, permite a las empresas tener una visión conjunta de los datos o de otros elementos cruciales.

Los Data Warehouse proporcionaban acceso integrado a múltiples sistemas como minicomputadoras, computadoras personales o incluso hojas de cálculo. Pero con el tiempo diferentes departamentos de cada empresa elegían las herramientas ETL para utilizarlas con fines de almacenamiento de datos, sin embargo, no se encontraban aplicadas del todo.

Con el paso de los años, las fuentes y sistemas de datos específicos han aumentado enormemente permitiendo tener una gran diversidad a la que poder acudir. El proceso ETL se ha convertido finalmente en uno de varios métodos que se utilizan en diferentes organizaciones empresariales para la extracción, tratamiento y carga de los datos.

Con este proceso se garantizaba que se respeten los estándares de validación y los diferentes tipos de datos con el fin de que se adapten a la estructura precisa para cumplir con los requisitos de salida. Este proceso puede ser empleado también para poder tratar una serie de datos con un formato específico para que, posteriormente, los programadores de diferentes softwares puedan plasmar de una forma

más sencilla dichos datos y el usuario final pueda llegar a tener un mayor entendimiento al analizarlos. Un ejemplo de este caso sería la herramienta SAP BW, el cual permite que se extraigan una serie de registros de diversos orígenes y, mediante el empleo de la plataforma SAP HANA, se transforme dicha información de una forma más visual y fácil para el usuario con el objetivo de poder plasmar información en diferentes aplicaciones como PowerBI o SAP Analytics Cloud.

A causa de la creciente necesidad de poder integrar información en las bases de datos, el proceso ETL se ha establecido como método principal y el más utilizado para extraer datos de diversas fuentes, transformarlos y, posteriormente, cargarlos en los sistemas correspondientes.

El principal objetivo de estos procesos es obtener de diversas fuentes de información esos datos, manipularlos y, posteriormente, transformarlos a un formato más consistente que permita su consulta y análisis eficientemente en el futuro.

Un Data Warehouse permite centralizar y consolidar datos de diversas fuentes, mejorando la calidad y consistencia de la información. Nos facilita el análisis histórico y predictivo, permitiendo una optimización en la toma de decisiones estratégicas y operativas.

2.2. Procesos ETL.

En la actualidad, la aparición de los procesos ETL son fundamentales a nivel de gestión y análisis de datos en las grandes empresas. Esto se debe a que en su totalidad son fundamentales en la extracción de datos de múltiples fuentes, la transformación en formatos adecuados y su carga en uno o varios sistemas de almacenamiento, como Data Warehouse o bases de datos, para su posterior análisis.
En un entorno empresarial el cuál se encuentra cada vez más orientado al tratamiento de datos, los procesos ETL permiten consolidad

información heterogénea, proporcionando una visión más coherente y accesible en los datos para la futura toma de decisiones estratégicas por parte del usuario final.

Su importancia se encuentra en la capacidad para facilitar la integración de los datos que provienen de diversas fuentes heterogéneas y su transformación en información útil y legible. Las empresas actuales, especialmente en sectores orientados al comercio, las finanzas o incluso la tecnología, dependen de estos procesos para obtener informes lo más precisos posibles, análisis predictivos y optimización de sus operaciones.

Por lo tanto, los procesos ETL supusieron y suponen actualmente un cambio significativo en las operaciones empresariales debido a los siguientes puntos:

- **Centralización de los datos:** Permiten que la integración de datos dispersos encontrados en diversos sistemas, plataformas o bases de datos sea mucho más fácil. Esto supone tener una visión única y consolidada de la información de diversas áreas como finanzas, logística e incluso atención al cliente.
- **Mejora de la calidad de la información:** A través de su segunda fase, transformación, los procesos ETL permiten filtrar, normalizar y enriquecer los datos previamente a su carga en un sistema destino, mejorando la calidad de estos.
- **Agilización de la toma de decisiones:** Al encontrarse los datos centralizados y en formatos adecuados para su posterior análisis, las decisiones se pueden tomar de una forma informada, pudiendo acceder a los datos actualizados y coherentes.
- **Reducción de los costes operacionales:** Evitan a las empresas que haya datos duplicados y, con ello, una redundancia de estos. Esto permite que haya una optimización del uso de recursos y mejora la eficiencia en el procesamiento de información.

Los procesos ETL son esenciales para el Business Intelligence (BI), ya que constituyen la base sobre la que deben construirse los informes y análisis que apoyan la toma de decisiones. La capacidad de transformar datos brutos en información consolidada y procesable es crucial para que las herramientas BI puedan extraer *insights* relevantes.

- **Alimentación de los sistemas de BI:** Los procesos ETL alimentan los Data Warehouse o bases de datos dedicadas, que son los repositorios de datos en los cuales se encuentran los datos desde donde las herramientas BI extraen información crucial para la generación de informes, paneles de control y análisis completos.
- **Disponibilidad y análisis de datos históricos:** Permiten almacenar grandes volúmenes de datos históricos de manera estructurada, facilitando la realización de análisis de tendencias, pronósticos o tomas de decisiones en base a datos del pasado.
- **Optimización de la calidad de los informes:** Al tratarse los datos en estos procesos ETL, se mejora la calidad de los informes generados, aumentando la fidelidad de los resultados obtenidos.

Al ser un componente clave en BI, suponiendo más de un 80% del tiempo del BI, los procesos ETL no solo aseguran que los datos estén estructurados y listos para el análisis, sino que también aportan beneficios significativos a nivel de proyecto.

- Mejora de la calidad de datos.
- Centralización y accesibilidad de la información.
- Optimización de los recursos, mejorando la eficiencia operativa y evitando posibles errores.
- Escalabilidad y flexibilidad, permitiendo a las empresas adaptarse a la evolución de sus necesidades de datos, integrando posibles nuevas fuentes o ajustando transformaciones a medida que crecen los volúmenes de datos.

Sin embargo, un mal diseño de procesos ETL puede tener consecuencias graves para una empresa, ya que afecta directamente a la calidad y disponibilidad de los datos utilizados en las tomas de decisiones. Los posibles errores en los procesos pueden ocasiones información incorrecta, retrasos en los análisis y pérdida de confianza en los sistemas establecidos de *Business Intelligence*. Otro punto añadido es que los procesos ETL mal diseñados pueden provocar que se generen cuellos de botella en el sistema, posibles aumentos del tiempo de procesamiento de datos y ocasionar costos innecesarios no tenidos en cuenta en el Budget establecido al inicio de un proyecto debido a la corrección de errores o la reestructuración de los procesos.

Por ello, para evitar un mal diseño de los procesos ETL, deben tenerse en cuenta las siguientes buenas prácticas:

- **Planificación adecuada:** Antes de la implementación de los procesos ETL en un proyecto, es crucial realizar una planificación detallada que defina cuáles son los objetivos del proyecto, las fuentes de datos, los tipos de transformaciones de datos necesarias y los requisitos de almacenamiento.
- **Automatización y monitoreo:** Se deben implementar sistemas de monitoreo que informen del rendimiento y la calidad del proceso ETL para detectar fallos de manera temprana. Esto permite que se identifiquen errores en tiempo real, evitando que se produzca un cuello de botella y afecten a posteriores análisis.
- **Pruebas rigurosas a lo largo de la implementación:** Es crucial que se realicen pruebas a lo largo del proceso ETL antes de su implementación final. Esto incluye que se hagan pruebas de calidad de los datos, rendimiento y capacidad de respuesta ante grandes volúmenes de datos. Estas pruebas deben iniciarse desde entornos de desarrollo hasta entornos réplica de los de productivo.
- **Revisión y optimización continua:** Las necesidades empresariales van evolucionando, así como la propia tecnología, por lo que es fundamental que se realicen ajustes de los procesos

de acuerdo con los cambios que pueden ir sufriendo las bases de datos, así como los posibles aumentos en el volumen de datos o incluso los futuros requisitos en el análisis.

Finalmente, los procesos ETL desempeñan un papel fundamental en el tratamiento de datos desde su origen hasta el desarrollo de informes de análisis de datos, por ejemplo. Gracias a su capacidad para integrar, transformar y cargar datos de diversas fuentes, un buen diseño de estos procesos es clave para garantizar la eficiencia y permitir análisis cada vez más precisos. En una sociedad cada vez más dependiente de los datos, disponer de infraestructuras de información consolidades y optimizadas mediante procesos ETL bien diseñados es esencial para el éxito en la toma de decisiones estratégicas.

2.3. Fases ETL.

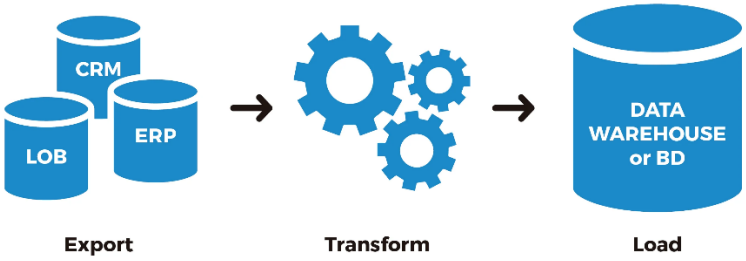


Ilustración 1: Fases del Proceso ETL.

Fase de Extracción.

El proceso ETL consta de las siguientes tres fases:

Se trata de la primera fase en el proceso ETL. Consiste en la obtención de datos desde las diversas fuentes de información, pudiendo tratarse incluso de archivos Excel o bases de datos transaccionales. Estos datos pueden obtenerse de fuentes internas o externas a una organización, los cuales serán tratados posteriormente en la fase de transformación y cargados en un sistema de almacenamiento centralizado.

En esta fase se realiza una primera recopilación de datos, la aplicación correcta del formato correspondiente a cada dato sin alterar su estructura original y se establece la frecuencia con la que se desea extraer dicha información. Este último paso normalmente se basa en la extracción de datos full en una primera instancia para obtener toda la información de la fuente de datos y, posteriormente, ejecuciones delta que permiten la modificación de aquellos registros que presentan cambios en sus datos. A pesar de ello, en función del objetivo que tienen dichos datos pueden realizarse extracciones completas diarias,

extracciones incrementales que hayan presentado modificaciones en el origen o extracciones en tiempo real que permiten la obtención de dichos datos que se generan o actualizan en la fuente en todo momento.

Esta extracción de datos es fundamental porque asegura que los datos relevantes siempre estén disponibles para que sean procesados y analizados posteriormente. Por ello, se trata de una fase que debe ser eficiente, ya que la información debe ser extraída sin interrumpir el funcionamiento normal de las fuentes y sin que llegue a comprometer su calidad y la integridad de esta.

Estas fuentes de datos podemos clasificarlas como heterogéneas u homogéneas.

- **Fuentes heterogéneas:** Se tratan de fuentes de datos que provienen de diversos sistemas, con diferentes formatos, tecnologías o incluso estructuras de almacenamiento. Un ejemplo de este tipo de fuentes se trataría de sistemas no SQL, plataformas relacionadas a redes sociales o aplicaciones en la nube. Estos datos pertenecientes a las fuentes heterogéneas suponen que la fase de transformación siga un proceso de transformación más complejo debido a que estos datos deben ser normalizados y adaptados de una forma más precisa y coherente en relación a las fuentes homogéneas, suponiendo un mayor costo de tiempo dentro de un proyecto.

- **Fuentes homogéneas:** Se tratan de aquellas fuentes cuyos datos provienen de sistemas que disponen de un mismo formato o estructura. Por ejemplo, fuentes que disponen de bases de datos relacionales. La homogeneidad de estas fuentes facilita la extracción y su posterior transformación, ya que los formatos son los adecuados para su tratamiento, lo que reduce el costo de tiempo en un proyecto.

Una vez tratados las diferentes fuentes de datos, un aspecto crucial son los formatos más comunes en los que se pueden encontrar los datos que se precisan extraer, pudiendo variar en función del origen de esos datos.

- **Bases de datos relacionales (SQL):** Se tratan de datos almacenados en bases de datos estructuradas como MySQL u Oracle. Estos datos se disponen en tablas relacionadas entre ellas.
- **Bases de datos NoSQL:** Datos que se encuentran almacenados en bases de datos no estructuradas o semiestructuradas como MongoDB. Estos datos no se disponen en tablas relacionales, por lo que sus fuentes no tienen un esquema fijo y pueden ser más flexibles.
- **JSON:** Se trata de un formato ligero y fácil de leer que se emplea en intercambios de datos entre diferentes aplicaciones.
- **XML:** Formato parecido al JSON que se emplea en intercambios de datos entre diferentes aplicaciones pero que requieren de estructuras jerárquicas de datos.
- **APIs:** Fuente de datos que se emplea en muchas ocasiones y que devuelven la información en formatos como JSON o XML.
- **Archivos planos:** Concebidos como archivos de texto simple, texto sencillo o texto sin formato, se tratan de ficheros que contienen datos simples en un lenguaje entendible por las personas. La simplicidad de estos los hace fáciles de procesar, pero pueden carecer de estructuras complejas.

Tras saberse de qué tipo de fuente de datos se obtendrán los datos, se debe analizar qué tipo de técnica de extracción de datos debe aplicarse, cada cual adecuada para distintos tipos de fuentes y necesidades.

- **Extracciones tipo Full:** Se trata de la extracción completa de los datos de una fuente, sin tener en cuenta cambios previos. Esta técnica es principalmente útil cuando se quiere disponer de una

copia exacta de los datos y se suele aplicar como primer medio de extracción cuando se desarrolla un proyecto empleando un proceso ETL y se quiere extraer toda la información disponible en una fuente de datos, lo que facilitará futuras comparaciones a nivel de año en el análisis que se realizará en cuadros de mando. Puede ser ineficiente si las fuentes contienen grandes volúmenes de datos, teniendo que recurrir a diferentes extracciones full con un específico número de paquetes de datos.

- **Extracciones incrementales:** Permite la obtención de datos que han cambiado desde la última extracción. Este tipo de extracción mejora la eficiencia y reduce el volumen de datos que se van a procesar, lo que es la mejor opción para fuentes de datos con una gran volumetría.
- **Extracciones en tiempo real:** Este tipo de extracciones permiten obtener los datos a medida que se van modificando o generando, proporcionando un acceso directo a los datos más actualizados. Este tipo de extracción es la más útil en sistemas de monitorización o plataformas financieras.
- **Extracciones basadas en eventos:** Estas extracciones se realizan cuando se produce un evento especificado previamente por el desarrollador. Este evento activa la extracción dependiendo de casos como la actualización de una base de datos, otra extracción anterior que se ha programado y tras su finalización provoca que esta extracción se inicie, etc.

Sin embargo, en los entornos empresariales, los datos extraídos pueden proceder de diferentes áreas de una organización y cada tipo de datos dispone de un propósito específico para respaldar la toma de decisiones. Por lo tanto, los tipos más comunes suelen ser:

- **Datos comerciales y/o de marketing:** Información que está relacionada con los clientes, las ventas, productos, campañas publicitarias, etc. Este tipo de datos permite una evaluación de la

efectividad de las campañas estratégicas comerciales y segmentar las audiencias a las que están destinadas.

- **Datos de recursos humanos:** Esta información suele ser de las más sensibles y de las que más se requiere, ya que se trata de datos de empleados, salarios, tiempos de trabajos, formación académica, etc. Son especialmente cruciales para la gestión del talento de una empresa, la evaluación de cumplimiento de políticas internas o gestión de empleados en base al área en el que trabajan.
- **Datos procedentes de redes sociales o plataformas digitales:** Estos datos son valiosos para poder mejorar la imagen de una marca, gestionar su reputación y ajustar las técnicas de comunicación en Internet.
- **Datos financieros:** Estos datos son esenciales para el análisis financiero, gestión de presupuesto y toma de decisiones acerca de inversiones y rentabilidad.
- **Datos logísticos:** Información acerca de pedidos, cadenas de suministro, etc. Permiten mejorar la eficiencia operativa y gestionar el flujo de productos, optimizando las operaciones diarias.

Por lo tanto, la fase de extracción es la fase más crucial en el proceso ETL, ya que permite obtener datos clave de diversas fuentes de información para su posterior transformación y carga. La aplicación de un diseño eficiente en esta fase garantiza que la calidad de los datos sea la adecuad, así como la accesibilidad a los mismos. Una correcta implementación de técnicas de extracción asegura un flujo de datos optimizado y confiable.

Fase de transformación.

En esta fase, los datos extraídos previamente se procesan y transforman para que puedan tener una utilidad y sean adecuados para su posterior análisis. Por ello, en esta fase se producen los siguientes pasos:

- **Conversión de formatos:** Por ejemplo, convertir específicas unidades métricas en otras en función del país de origen del dato.
- **Limpieza de datos:** Si en la fase anterior no se han aplicado una serie de filtros a la información extraída, en la fase de transformación se pueden eliminar o corregir errores en dichos datos o información la cuál no sea necesaria para el objetivo establecido.
- **Enriquecimiento de los datos:** Con la información extraída se pueden añadir nuevos valores o campos a los datos a partir de una serie de funciones o cálculos matemáticos, como por ejemplo el cálculo del porcentaje de índice muscular en base a la masa muscular total en el cuerpo y el peso corporal total.
- **Agregación de la información:** La información extraída puede ser agregada o pueden aplicarse vistas sobre dichos datos para que sean más comprensibles.
- **Normalización:** Ajuste de los valores de los datos para que se encuentren dentro de un rango específico o sigan un patrón especificado por el usuario final.

Esta fase es crucial para poder garantizar que los datos extraídos sean útiles y aptos para su análisis. Mediante los procesos nombrados anteriormente, los datos se ajustan a los requisitos específicos de una empresa y se preparan de la forma más adecuada para tomar decisiones informadas. Un diseño adecuado de esta fase asegura la consistencia de los datos, facilitando su integración en diferentes sistemas de análisis o de *reporting*.

Fase de carga.

La fase de carga en un proceso ETL es el último paso del flujo de datos, en el que la información que ha sido extraída, transformada y procesada, finalmente será insertada en un sistema destino. Esta fase es determinante debido a que establece cómo se almacenan y gestionan los datos para su posterior análisis.

Existen diferentes tipos de cargas que se pueden utilizar dependiendo de las necesidades del negocio y la frecuencia con la que se desea que la información sea actualizada.

- **Carga tipo Full:** Se trata de cargar todos los datos en la base de datos destino, reemplazando los existentes por los nuevos. Para realizar este tipo de carga debe realizarse un borrado previo del volumen de datos en el sistema destino. Es ideal cuando se requiere realizar una actualización completa de los datos o cuando el volumen de estos no sea demasiado grande. Sin embargo, puede ser costoso en tiempo.
- **Carga incremental:** Solo se realiza la carga de los datos que han cambiado desde la última carga. Es más eficiente que una carga full, ya que solo se procesan datos que son nuevos o modificados, reduciendo el tiempo.
- **Carga en tiempo real:** Los datos se cargan en el sistema destino en tiempo prácticamente real, ya que a medida que van generándose o se actualizan en la fuente de datos se disponen a ser cargados. Este tipo de carga es muy común en aplicaciones que requieren de información actualizada constantemente.
- **Carga tipo Delta:** Este tipo de carga es similar a las cargas incrementales. Sin embargo, previamente a iniciar la carga, se rastrean aquellos registros que tienen diferencias con relación a ese mismo registro cargado con anterioridad, modificándose sólo aquellos campos de un dato que ha sufrido una modificación y no todo el registro completo.

Mientras se ejecuta la fase de carga, pueden aplicarse diversas restricciones para garantizar la calidad de los datos asó como la eficiencia del proceso de carga.

- **Restricciones de integridad referencial:** Asegura que los datos cargados mantengan consistencia y coherencia.
- **Restricciones de validación:** Validan que los datos cumplan con una serie de criterios o rangos antes de ser cargados.
- **Restricciones de rendimiento:** Dependiendo del tamaño de los datos, puede que sea necesario la implementación de técnicas de optimización, como la carga en bloques o en lotes, evitando así la sobrecarga de la base de datos destino y garantizar que el proceso de carga se realice de una manera eficiente.
- **Transacciones y bloqueos:** Las transacciones se suelen emplear para asegurar que las cargas se realicen de forma atómica (las cargas deben realizarse correctamente o no se harán). Los bloqueos, sin embargo, se aplican para evitar que otros procesos puedan interferir en la carga mientras se está llevando a cabo.

Esta fase de carga es la etapa final de un proceso ETL, donde los datos procesados se almacenan en la base de datos destino, disponibles para su correcto análisis. Elegir el tipo adecuado de la carga y aplicación de las restricciones que sean necesarias, asegura que los datos sean íntegros y estén eficientemente gestionados. Un proceso de carga que esté bien diseñado es clave para optimizar el rendimiento del sistema y garantizar que los datos estén siempre disponibles para su correcto uso.

2.4. Data Warehouse.

Un Data Warehouse se trata de un sistema centrado en el almacenamiento de datos diseñado principalmente para su consulta y realización de análisis de grandes volúmenes de información procedentes de diversas fuentes. A diferencia de las bases de datos transaccionales, las cuales se encuentran optimizadas para la entrada, modificación y lectura de datos en tiempo real, un Data Warehouse se encuentra optimizado de manera que se puedan realizar diversas consultas más complejas y análisis de nivel histórico. Los procesos ETL son cruciales para alimentar estos sistemas, debido a que permiten extraer datos de diferentes fuentes, transformarlos mediante diversas técnicas aplicadas con el fin de adaptarlos al modelo del Data Warehouse y cargarlos en estos.

En la definición técnica de un Data Warehouse hay varios aspectos fundamentales que deben ser considerados:

- **Orientado al negocio:** Están diseñados para que puedan satisfacer las necesidades específicas de cada tipo de análisis y la toma de decisiones del negocio, pero no para el procesamiento transaccional. Se enfocan en ofrecer una visión más integral de los datos que permita al usuario obtener una información más útil y estratégica.
- **Integrado:** Los datos provienen de diversos sistemas y son procesados e integrados en el Data Warehouse. Esto supone la normalización de los formatos y la creación de esquemas comunes con la finalidad de que los datos de diferentes fuentes puedan ser consultados y analizados de una forma más coherente.
- **Variante en el tiempo:** Los datos que se encuentran almacenados suelen ser normalmente históricos, lo que supone que se conserva y se puede acceder a conjuntos de datos en diferentes puntos del tiempo.
- **No volátil:** Una vez que estos datos han sido cargados en el Data Warehouse, no se modifican ni se eliminan. Esto es crucial para

que la integridad de los datos no se vea afectada y garantiza que el usuario siempre tenga acceso a un conjunto de datos consistente para el análisis.

Los Data Warehouse normalmente se organizan en cubos multidimensionales, en modelos estrella o copo de nieve, donde los datos se almacenan en dimensiones o hechos.

- **Dimensiones:** Contextualizan los hechos como los productos, los clientes, etc. Estas dimensiones permiten que los datos de los hechos se segmenten para que los usuarios puedan realizar consultas más detalladas y específicas.
- **Hechos:** Datos numéricos que se van a analizar por parte del usuario, como ventas, ingresos, *head accounts*, etc. Se tratan de los elementos principales de un análisis.

Para facilitar el análisis de los datos, los usuarios acceden a la información de los Data Warehouse mediante herramientas de *Business Intelligence* (BI) que proporcionan interfaces gráficas, como cubos multidimensionales. Estos cubos permiten a los usuarios realizar consultas y análisis de forma más interactiva, pudiendo visualizar los datos de diferentes dimensiones, filtrando o realizando agregaciones en base a la necesidad del análisis.

En conclusión, un *Data Warehouse* es un elemento clave en la infraestructura de datos de las organizaciones modernas, ya que centralizan la información proveniente de diversas fuentes y la ponen a disposición del usuario para su posterior análisis.

2.5. Datos abiertos.

La Open Data o Datos abiertos se trata de un conjunto de datos producidos o recopilados por las entidades públicas que se encuentran disponibles de forma libre y sin restricciones para su uso, reutilización o distribución. Este acceso abierto promueve la participación de la ciudadanía, permitiendo a cualquier persona o institución hacer uso de la información para realizar investigaciones o tomar decisiones informadas.

Su principal objetivo es fomentar la transparencia y la accesibilidad de información pública, con el fin de que los ciudadanos, empresas u otras organizaciones puedan acceder a los datos para su uso libre en diferentes contextos. Estos datos abiertos permiten mejorar la toma de decisiones y una mejor rendición de cuentas de las administraciones públicas. A través de ellos, se busca crear un ecosistema de información que permita generar diferentes soluciones para posibles problemas sociales, económicos o ambientales, por ejemplo.

Sin embargo, a nivel empresarial, los datos abiertos presentan una gran cantidad de beneficios, debido a que permiten optimizar la toma de decisiones al acceder a información actualizada e importante acerca de tendencias y/o comportamientos. Además, permiten a las empresas mejorar en su eficiencia operativa al acceder a datos de diversas fuentes públicas, reduciendo los costos asociados a investigación o adquisición de datos privados.

3. Herramientas.

3.1. Apache Hop.

Ilustración 2: Apache Hop

Los procesos ETL (Extract, Transform and Load) representan una pieza fundamental para lograr la integración eficiente de datos que provienen de distintas fuentes hacia una base de datos o sistema centralizado. Estos procesos permiten transformar los datos en formatos coherentes, corregir inconsistencias y prepararlos para su análisis por parte de los usuarios finales, garantizando la calidad de la información.

En este contexto existen varias herramientas especializadas que permiten construir flujos ETL de forma visual e intuitiva, destacando entre ellas a Apache Hop. Dicha herramienta, que se ha empleado durante este proyecto, de integración de datos de código abierto está diseñada para desarrollar y planificar pipelines complejos de procesamiento de datos.

Apache Hop proporciona una interfaz gráfica intuitiva y modular que facilita al usuario diseñar procesos ETL sin necesidad de escribir código. A través de los pipelines (procesos de flujo de datos) y workflows (procesos de control o decisión), el usuario puede modelar desde tareas muy simples hasta más complejas. Además, en esta herramienta es posible ejecutar scripts condicionales, transformar y/o limpiar registros, conectarse a múltiples orígenes de datos (CSV, bases de datos, etc.), y exportar los resultados a diferentes destinos.

Su interfaz cuenta con componentes como:

- **Explorador de proyectos:** Espacio donde se gestionan los pipelines, workflows y configuraciones.
- Lienzo de diseño: Donde se arrastran y conectan los distintos pasos del flujo de datos, permitiendo una visualización gráfica del proceso.
- **Configurador de transformaciones:** Cada paso del pipeline se puede parametrizar detalladamente, con opciones como transformación de campos, lectura de archivos, filtros condicionales, etc.
- **Visor de resultados y logs:** Al ejecutar el pipeline, Hop proporciona un panel detallado con las filas leídas, escritas, actualizadas o rechazadas, facilitando así la depuración de errores y el control de calidad.

Durante este proyecto, Apache Hop ha sido la herramienta seleccionada para el desarrollo del proceso ETL, partiendo de la extracción de datos abiertos de la FAO en formato CSV, su transformación y su carga final en una base de datos PostgreSQL. A lo largo del flujo de datos, se han implementado validaciones, conversiones personalizadas mediante scripts JavaScript y estructuras de control que aseguran que los datos insertados en la base de datos cumplen con los requisitos estructurales y de calidad necesarios para su análisis final.

Apache Hop puede descargarse a partir del siguiente link:

https://hop.apache.org/download/

En definitiva, Apache Hop se ha consolidado como una herramienta robusta y visualmente accesible para diseñar procesos ETL, permitiendo a los usuarios manipular datos desde su origen hasta su carga en el sistema final con total control, trazabilidad y monitoreo.

3.2. PostgreSQL.

Ilustración 3: PostgreSQL

Para la implementación del Data WareHouse se ha elegido PostgreSQL, el cuál es un sistema de gestión de bases de datos relacional de código abierto. Está diseñado para manejar grandes volúmenes de datos de manera eficiente, ofreciendo características avanzadas como transacciones ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad), integridad referencial, soportes para procedimientos almacenados, así como una amplia variedad de tipos de datos.

Su relación con pgAdmin4 es clave destacarla debido a que, esta última, se trata de una herramienta gráfica de administración y desarrollo para bases de datos PostgreSQL. Proporciona una interfaz más gráfica que permite gestionar y administrar bases de datos PostgreSQL de manera más intuitiva, facilitando tareas como la creación o modificación de tablas, ejecución de consultas SQL, etc.

PostgreSQL es una opción muy atractiva para la implementación de Data Warehouse y otros sistemas de bases de datos debido a las siguientes características:

- Soporte para SQL avanzado y funciones no estándar.
- Alta escalabilidad: Capaz de manejar bases de datos de gran tamaño.
- Expansión mediante extensiones: Permite extender su funcionalidad mediante el uso de extensiones
- Seguridad avanzada.
- Comunidades activas y una documentación más extensiva.

PostgreSQL utiliza un enfoque relacional para gestionar las bases de datos, lo que supone que almacene la información en tablas organizadas en filas y columnas. Los principales puntos de su funcionamiento son los siguientes:

- **Base de datos:** Contenedor lógico que puede almacenar múltiples esquemas, tablas, vistas, etc. Cada base de datos funciona de manera independiente y se ejecuta en un proceso de servidor separado.
- **Esquemas:** Conjunto de objetos de base de datos dentro de la misma. Proporcionan una forma de organizar los objetos y evitar conflictos de nombres.
- **Tablas y relaciones:** Las tablas son la estructura fundamental en la que se almacenan los datos. Estas tablas están formadas por columnas en la que cada una de ellas puede tener un tipo de dato distinto. Como añadido, PostgreSQL permite establecer relaciones entre tablas mediante claves foráneas, lo que facilita la normalización de los datos y mantiene la integridad referencial.
- **Consultas SQL:** PostgreSQL emplea SQL para interactuar con las bases de datos. Los usuarios pueden emplearlo para realizar consultar de inserción de datos, así como actualización o incluso borrado de los mismos.
- **Índices:** Son estructuras que permiten una búsqueda más rápida de los datos dentro de las tablas. El sistema de gestión de bases de datos crea índices sobre las columnas específicas que se utilizan con frecuencia en las consultas, mejorando el rendimiento de las operaciones de lectura.

PostgreSQL trabaja con bases de datos relacionales, permitiendo que se empleen tablas interconectadas entre sí y que almacenan una información organizada correctamente.

Se ha instalado la siguiente versión de PostgreSQL:

- PostgreSQL 16.2, compiled by Visual C++ build 1937, 64-bit.

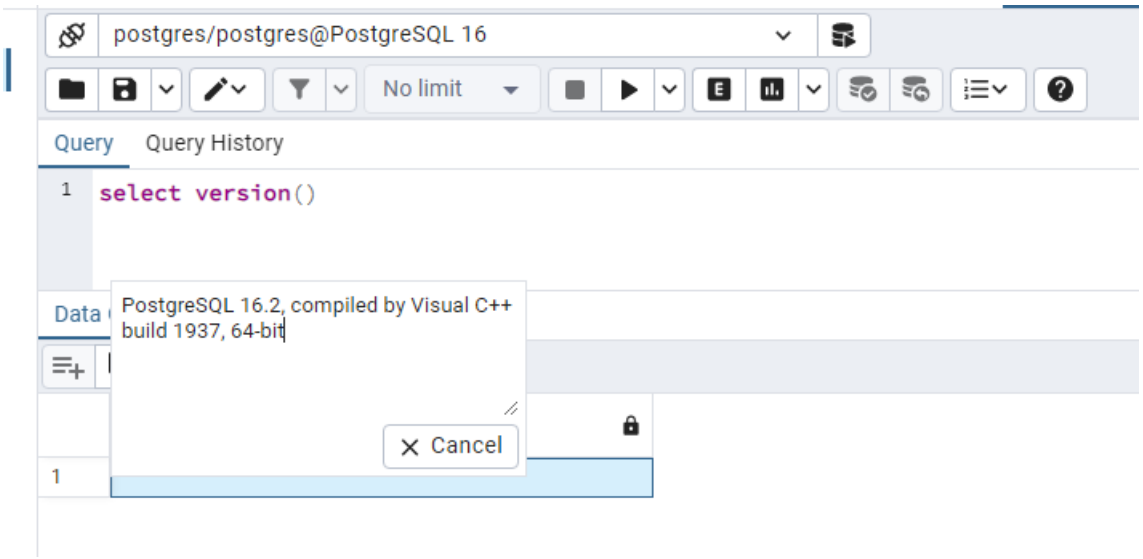


Ilustración 4: Consulta de versión MySQL

En esta base de datos es importante tener en cuenta los siguientes parámetros para una conexión correcta:

Hostname: localhost.

Port: 5432

Username: postgres

Contraseña: RamirezG18

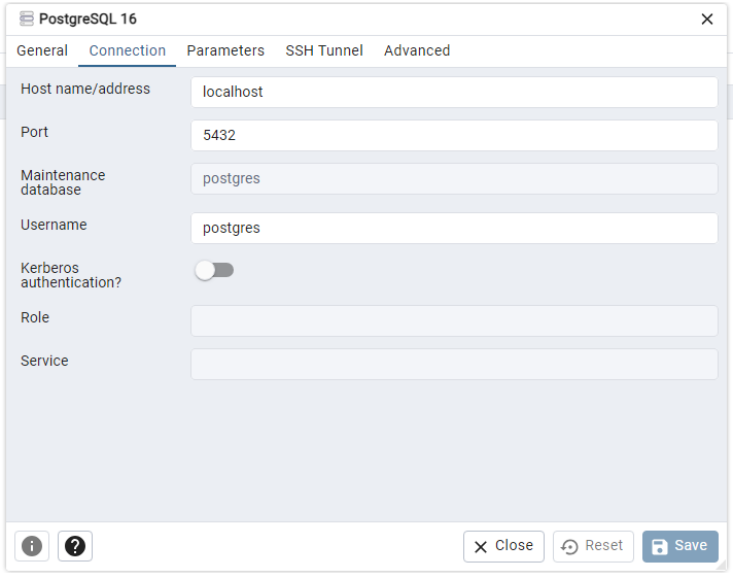


Ilustración 5: Connect to Database

Por último, tras iniciar sesión se crea una nueva base de datos con el siguiente comando.

```
-- Database: TFG_Guillermo

-- DROP DATABASE IF EXISTS "TFG_Guillermo";

CREATE DATABASE "TFG_Guillermo"
    WITH
    OWNER = postgres
    ENCODING = 'UTF8'
    LC_COLLATE = 'Spanish_Spain.1252'
    LC_CTYPE = 'Spanish_Spain.1252'
    LOCALE_PROVIDER = 'libc'
    TABLESPACE = pg_default
```

Ilustración 6: Create DataBase

3.3. Power BI.

Power BI se trata de una herramienta de Microsoft para la visualización y análisis de datos. Permite la transformación de dicha información en informes interactivos y cuadros de mando que facilitan la toma de decisiones.

En este proyecto, se ha empleado Power BI para conectarse a la base de datos de PostgreSQL la cual contiene información relevante que posteriormente será tratada y mostrada en un cuadro de mando. Esta conexión se realiza fácilmente mediante los conectores nativos de Power BI, especificando los parámetros de acceso al servidor.

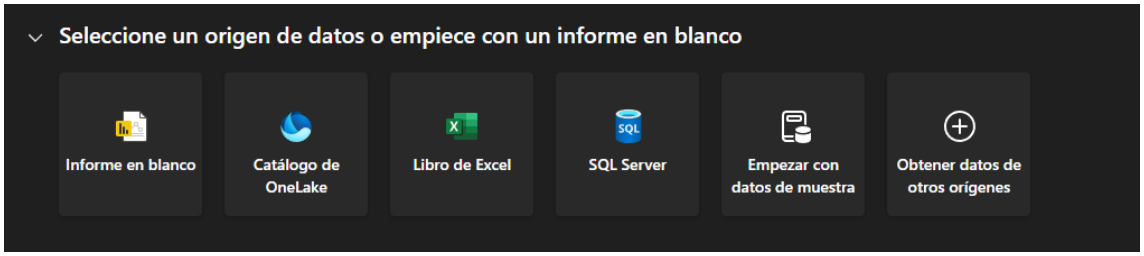


Ilustración 7: Power BI

Una vez conectados, Power BI permite:

- Importar y transformar datos de forma visual.
- Crear relaciones entre tablas y medidas personalizadas.
- Generar gráficos, mapas y paneles interactivos.
- Actualizar automáticamente los datos.

- Compartir los informes de forma online con otros usuarios.

Debido a estas funcionalidades, Power BI ha sido esencial para representar de forma clara, precisa y dinámica los datos extraídos, facilitando el análisis de patrones y repercusiones.

4. Desarrollo del proyecto.

4.1. Datos abiertos.

Los datos que se han empleado tienen como objetivo el análisis del impacto de los incendios en las superficies forestales de los países que componen la Unión Europea desde 2010 al 2021 así como las emisiones de CO2. Este conjunto de datos ha sido generado en diversas fuentes como empresas y gobiernos, redes sociales, etc. Dicha información ha sido producida y recopilada por administraciones públicas que ponen a disposición de la sociedad para que su acceso se realice de forma libre, así como su empleo.

Por ello, la obtención de estos datos se ha realizado a partir de la siguiente página:

Food and Agriculture Organization of United States (FAO)

https://www.fao.org/faostat/es/#data/GF (Emisiones derivadas de los bosques)

https://www.fao.org/faostat/es/#data/GI (Emisiones derivadas de fuegos)

https://www.fao.org/faostat/es/#data/GV (Emisiones derivadas de suelos orgánicos drenados)

Durante el proyecto, se llegó a plantear la posibilidad de empleo de herramientas de generación de datos aleatorios, dando la posibilidad de creación de tablas complementarias.

Sin embargo, se optó directamente con emplear los datos preseleccionados dentro de la página oficial de la Organización de Comida y Agricultura de las Naciones Unidas (FAO).

En su mayoría los datos son compatibles debido a que se trata información relacionada, por ejemplo, en algunas tablas se contemplan medidas de países los cuáles pueden no aparecer en otras tablas.

Estos son los datos tratados:

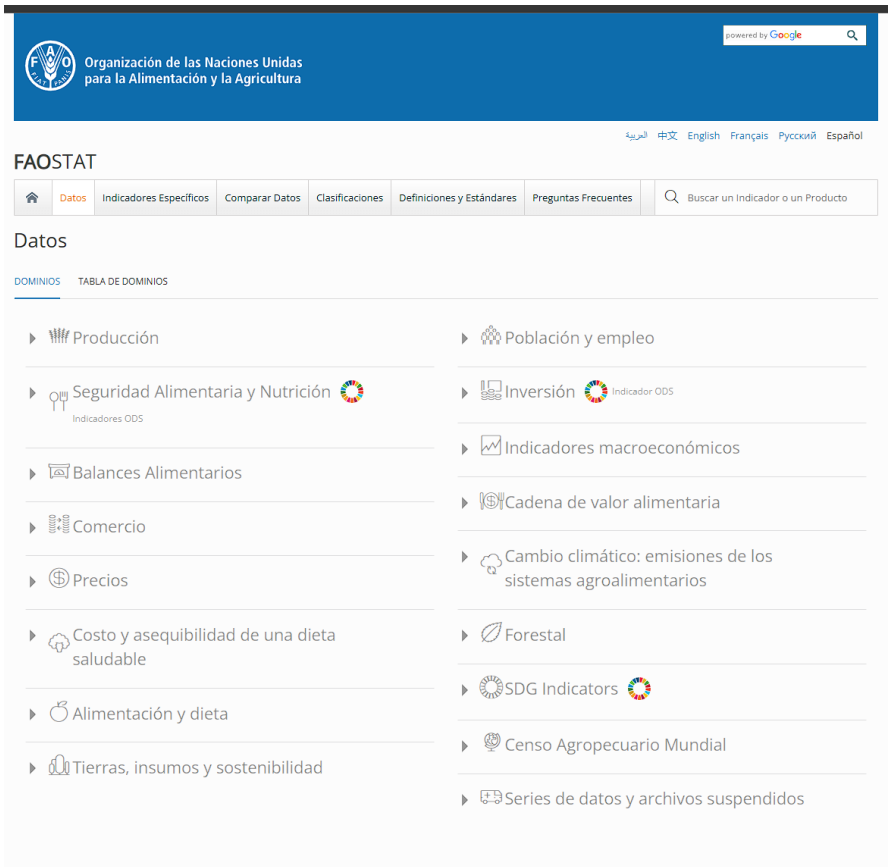


Ilustración 8: Organización de las Naciones Unidas para la Alimentación y la agricultura

Estas son las tablas de datos usadas en el proyecto:

Symbol: Tabla la cual contiene dos registros siendo A o E. El registro A hace referencia a la cifra oficial y el registro E hace referencia al valor estimado. Dicha tabla se emplea meramente para, tras su relación con otras tablas en la base de datos, poder contemplar si los KPIs mostrados se tratan de cifras oficiales o valores estimados.

Emissions Land Use Fire: Se registran estadísticas del impacto ambiental de los incendios en distintos países europeos. Esta tabla contiene información clave relacionada con la superficie afectada por incendios, biomasa quemada, tipo de emisiones, etc. Cada registro en la tabla se representa por el país en el que se ha producido dicho incendio, superficie ambiental afectada, fuente de origen de la información y unidad en la que se representan los indicadores. Las unidades representadas pueden ser toneladas (en el caso de la biomasa

quemada), mil toneladas (en el caso de las emisiones CH4, N2O y CO2) y hectáreas (en el caso de la superficie quemada). Dichos indicadores aparecen representados por años, siendo el inicio del periodo 2010 y el final 2021.

"The FAOSTAT domain Emissions from Fires consists of estimates of methane (CH4), nitrous oxide (N2O) and carbon dioxide (CO2) emissions generated from biomass burning in a range of vegetation types and from fires in organic soils."

Emissions Land Use Forest: Se registran estadísticas acerca de las emisiones y uso de la tierra relacionados con los bosques de la Unión Europea. Esta tabla contiene información clave relacionada con el análisis del suelo forestal y cómo ha ido cambiando durante el tiempo en diferentes países, datos acerca de emisiones y superficies forestales que han sufrido incendios y el estudio del impacto de incendios en términos de pérdida de masa forestal o aumento de emisiones. Por ello, esta tabla representa información acerca de emisiones y absorciones de CO2 correspondiente a los cambios en las reservas de carbono forestal (biomasa de la superficie y biomasa subterránea) aparecen representados por años, siendo el inicio del periodo 2010 y el final 2021.

"The FAOSTAT domain Forests consists of CO2 emissions and removals corresponding to forest carbon stock changes (aboveground and belowground living biomass). Estimates are computed following the 2006 IPCC Guidelines for National greenhouse gas Inventories (IPCC, 2006)."

4.2. Diseño e Implementación del proceso ETL.

Apache Hop se trata de una plataforma de código abierto para orquestar procesos de integración de datos (ETL). Esta herramienta nos permite diseñar, ejecutar y monitorizar flujos de datos visualmente mediante pipelines y workflows, facilitando de esta forma tareas como la transformación y posterior carga de datos.

En este caso, se emplearán pipelines los cuáles son secuencias de pasos (steps) conectados entre sí, que transforman y trasladan datos desde una o varias fuentes a un destino, aplicando operaciones intermedias.

Los componentes principales empleados en los pipelines son los siguientes.

- **CSV Input:** Lee los datos desde un fichero origen CSV y los convierte en un flujo de registros estructurados.
- **Select Values:** Permite seleccionar, renombrar o cambiar el tipo de columnas específicas del flujo de datos.
- **Append Streams:** Combina múltiples flujos de datos en uno solo, siendo especialmente útil para unir resultados de diversas fuentes.
- **Unique Rows:** Filtra registros duplicados según campos clave, dejando pasar sólo filas únicas.
- **Insert/Update:** Inserta nuevos registros o permite actualizar existentes en una base de datos, según una clave definida.

Para la creación del proceso ETL se han creado 7 pipelines, en función del tipo de tabla en la que se deseaba insertar/actualizar datos en la base de datos.

Esos 7 pipelines pueden dividirse en dos grupos:

1. Pipelines destinados a la carga de tablas sin indicadores, estando compuestas por un identificador único y la descripción correspondiente en función del tipo de metadato.

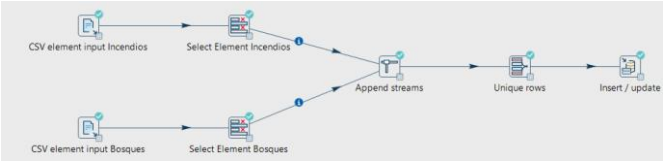


Ilustración 9: Carga Tabla Element

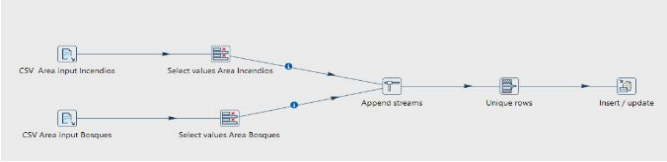


Ilustración 10: Carga Tabla Area

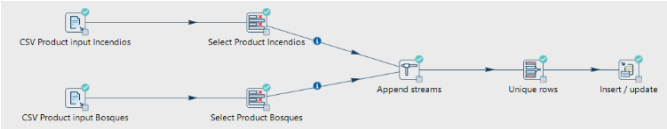


Ilustración 11: Carga Tabla Product

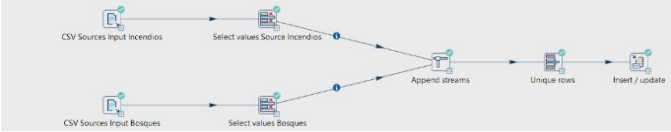


Ilustración 12: Carga Tabla Source

En este caso se han creado 4 pipelines, una por cada tabla, encargadas de transferir los datos desde el origen al destino en función del tipo de campo (elemento, producto, área, source y symbol).

2. Pipelines destinados a carga de tablas con indicadores (KPIs), estando compuestas por los identificadores de las tablas nombradas anteriormente (dimensiones) así como los KPIs relacionados que proporcionan información desde el 2010 al 2021. En este caso se realizaron extracciones directas desde el fichero .CSV a la tabla asignada en la base de datos debido a que los datos extraídos de origen no presentan registros duplicados, así como registros que deban tratarse desde origen.

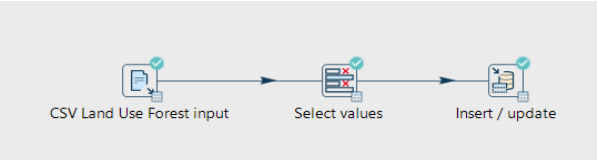


Ilustración 13: Carga Tabla Fire

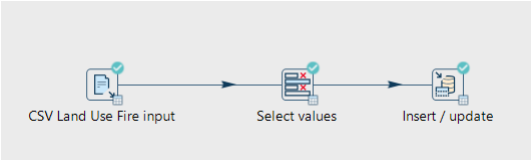


Ilustración 14: Carga Tabla Forest

Carga de datos: Se procede a crear una conexión entre un fichero .CSV y una base de datos. Seguiremos el caso del pipeline "etl_carga_element" encargado de transferir y transformar los datos desde los ficheros origen Emisiones_Uso_de_la_tierra_Bosques y Emisiones_Uso_de_la_tierra_Incendios.

En ambos pasos, se realiza la conexión con los ficheros .CSV y se establece el tipo de dato, la longitud, formato y precisión.

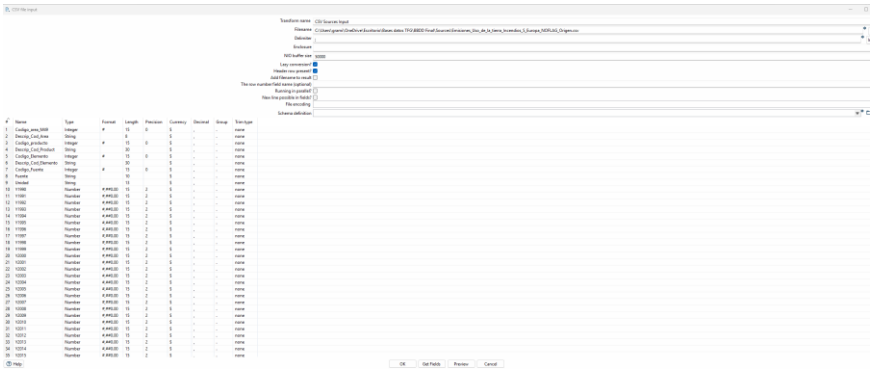


Ilustración 15: Conexión con fichero CSV Incendios

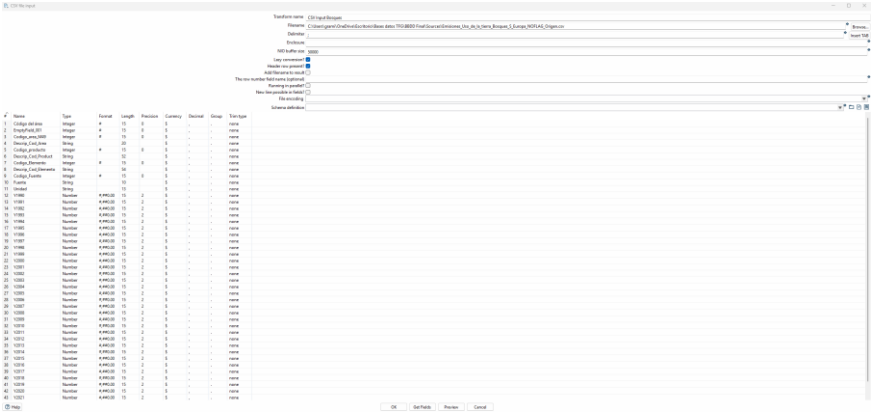


Ilustración 16: Conexión con fichero CSV Bosques

Posteriormente, se descartan las columnas innecesarias como Código del área, código_área_m49, etc. Esto se lleva a cabo en el paso Select Values debido a que en muchas ocasiones no precisamos de esos datos en las tablas y no aportan información real. En este paso se realiza también el renaming del campo, poniéndole el mismo nombre que tendrá en la tabla de la base de datos destino.

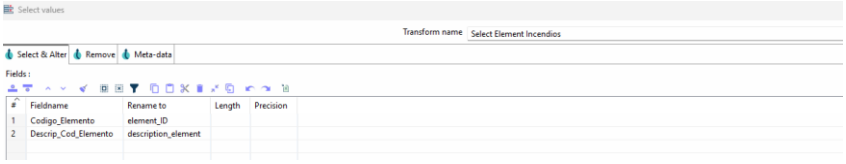


Ilustración 17: Renaming campos Tabla Element

El siguiente paso, Append streams es el encargado de combinar los flujos de datos en uno solo, ya que en el caso de este pipeline se están extrayendo registros de dos fuentes distintas. El objetivo es unir los registros existentes en cada fuente de datos en un único lugar.

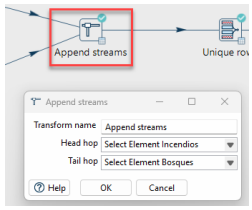


Ilustración 18: Append Streams

El paso Unique rows es el encargado de filtrar los registros duplicados según campos clave, dejando únicamente filas únicas. En el caso de este pipeline, se filtrarán los registros duplicados por el campo element_ID.

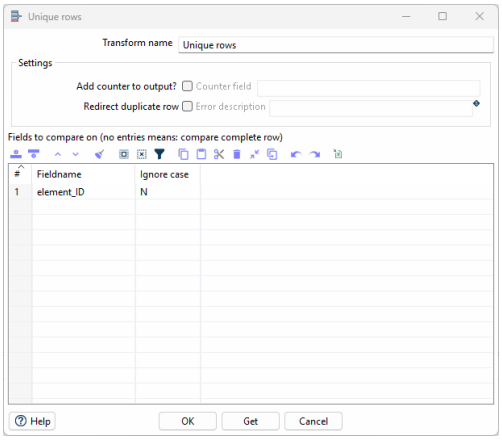


Ilustración 19: Unique Rows

Finalmente, se emplea el componente Insert/Update, en el cual se tiene que elegir la base de datos, el nombre de la tabla en la que se quieren cargar los datos, así como el usuario y la contraseña del servidor. La tabla ha sido creada previamente en la Base de datos, por lo que este paso únicamente se encarga de insertar y/o actualizar datos en dichas tablas.

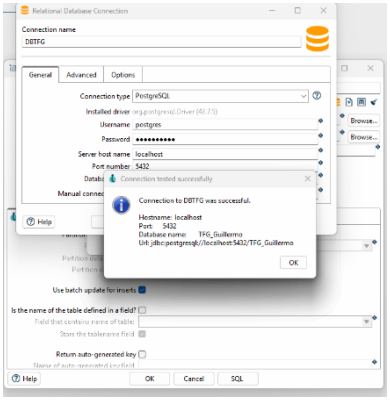


Ilustración 20: Connection to DBTFG

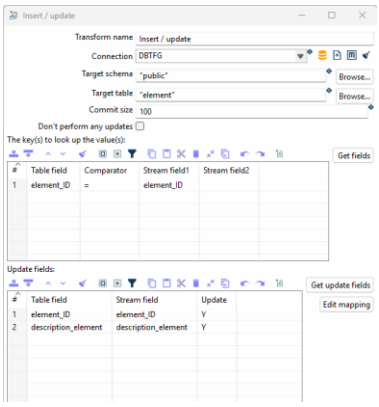


Ilustración 21: Insert/Update

Los datos que van a ser cargados en las tablas deben de tener una Key o Clave especificada previamente para no tener futuros problemas de filas duplicadas, debido a la gran magnitud de datos, es preciso poner claves. Por ello, en el paso del Insert/Update, es preciso seleccionar cuáles serán las claves que se tendrán en cuenta para la carga.

En el caso del pipeline encargado en el proceso ETL del land_use_forest, las claves elegidas han sido las siguientes:

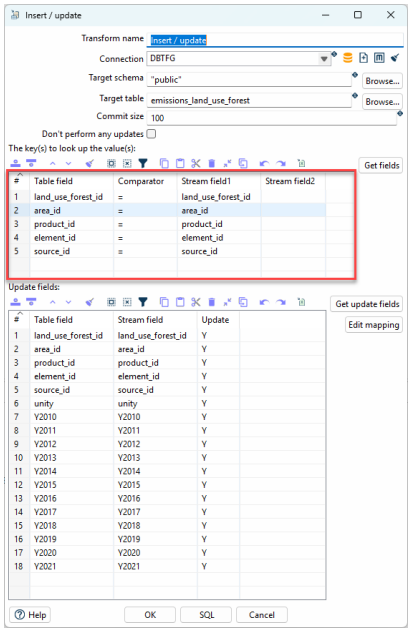


Ilustración 22: Claves Emissions Land Use Forest

4.3. Creación Data Warehouse.

Tras la carga de los datos en las tablas, se puede realizar la comprobación en la aplicación MySQL que se han creado dichas tablas definidas anteriormente.

Para conectarse a la base de datos, deben definirse los siguientes parámetros.

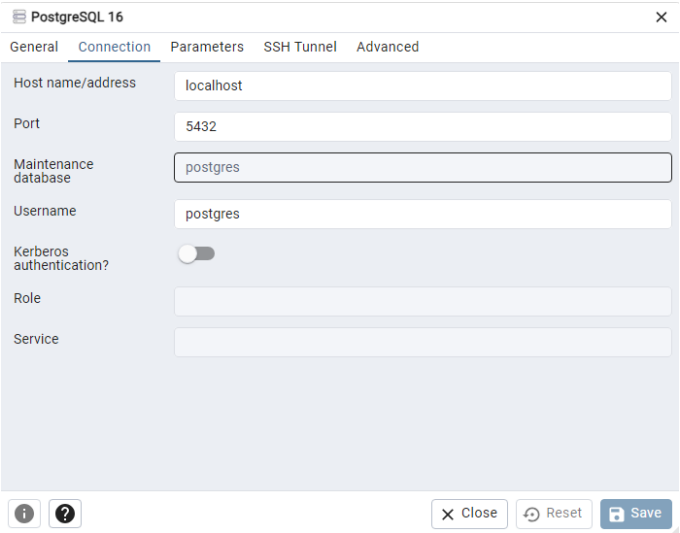


Ilustración 23: Connect to DataBase

Service : localhost:5432

Database : TFG_Guillermo

User : postgres

Password : RamirezG18

Una vez realizadas todas las cargas, se pueden comprobar en la aplicación que se han creado las tablas definidas :

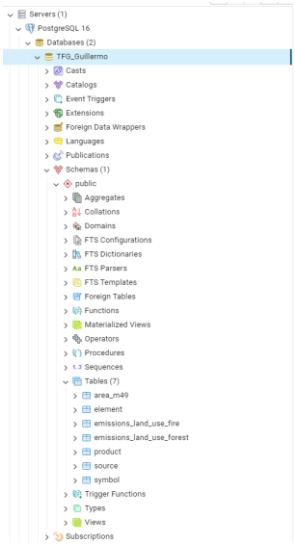


Ilustración 24: Tablas definidas en DB

Y, mediante la ejecución de comandos SQL, se puede observar el contenido de las tablas.

```
1  select * from area_m49
```

Data Output Messages Notifications

	area_id [PK] integer	description_area character varying (50)
1	4	Afganistán
2	2	África
3	17	África central
4	18	África meridional
5	11	África occidental
6	14	África oriental
7	15	África septentrional
8	8	Albania
9	276	Alemania
10	13	América central
11	5	América del Sur
12	21	América septentrional
13	19	Américas
14	20	Andorra

Ilustración 25: Contenido tablas por SQL

```
select * from emissions_land_use_fire
where product_fd = '6992' and element_fd = '7245'
```

Data Output Messages Notifications

	land_use_fire_id [PK] integer	area_id integer	product_id integer	element_id integer	source_id integer	unity character varying (30)	y2010 double precision	y2011 double precision	y2012 double precision	y2013 double precision	y2014 double precision	y2015 double precision	y2016 double precision	y2017 double precision	y2018 double precision	y2019 double precision	y2020 double precision
1	1	8	6992	7245	3050	toneladas	0	33581573000	4332935700	0	0	4334318700	8665407400	22745949400	0	0	0
2	47	276	6992	7245	3050	toneladas	0	10853706300	2170809600	1084937500	34733865500	0	3255159500	28219823300	2170647000	0	0
3	96	20	6992	7245	3050	toneladas	0	0	0	0	0	0	0	0	0	0	0
4	142	40	6992	7245	3050	toneladas	0	1085066300	2169417200	6509453900	15188013600	0	3255370200	6509024800	19526129100	1084892100	2169803200
5	191	112	6992	7245	3050	toneladas	36938911600	16294171200	270460943800	40193151800	507258756100	203109967300	80395062200	32591653500	46953215900	73866565300	296507254600
6	240	56	6992	7245	3050	toneladas	0	2171043200	0	0	3256644800	0	0	0	0	0	0
7	293	58	6992	7245	3050	toneladas	0	0	0	0	0	0	0	0	0	0	0
8	337	70	6992	7245	3050	toneladas	24936135000	216823911200	203802955000	9757772400	26016655600	10839448100	1084034600	52035829600	1084080200	23651556800	149601575400
9	383	100	6992	7245	3050	toneladas	101880528200	66113419700	10836793700	1083974300	11923364700	32518799500	27090168700	23839648600	6502717000	13004048000	16254992600
10	432	200	6992	7245	3050	toneladas	0	0	0	0	0	0	0	0	0	0	0
11	476	203	6992	7245	3050	toneladas	7599317200	2170677400	0	4341840500	43413866000	0	1085515800	15196108300	0	0	0
12	525	191	6992	7245	3050	toneladas	2169034000	182153207200	43371035100	0	6504978000	54187991000	1084140400	53122460100	1084374500	22767726800	111670037500
13	578	208	6992	7245	3050	toneladas	0	0	0	0	0	0	0	0	0	0	0
14	627	703	6992	7245	3050	toneladas	0	0	3255755300	0	7595673500	0	0	2170812900	8681857000	0	0
15	676	705	6992	7245	3050	toneladas	0	65071541700	27115735600	1084589600	3253993200	0	3253931600	0	7592167000	0	0
16	726	724	6992	7245	3050	toneladas	40083993900	169044473300	486349595600	111609869800	86689703400	226489472300	204607223300	56777862700	17333727900	13425813370	47641428200
17	780	233	6992	7245	3050	toneladas	0	2175163800	2175482200	1087639800	0	0	0	4351372600	5437050600	0	0
18	832	643	6992	7245	3050	toneladas	5622646640530	5760205069000	8889325783460	2549043522620	8735177477010	2884860122800	5812125517130	3304782174080	5608776419780	4812619309010	2472716901160
19	881	246	6992	7245	3050	toneladas	8851390300	4426408900	0	885162800	0	0	2655519300	0	1770805400	885182000	0
20	934	250	6992	7245	3050	toneladas	91039031200	47693175600	537565971500	7594275100	197262786700	47693531700	40103820000	577702070200	15181620600	460618339200	99693436800
21	985	292	6992	7245	3050	toneladas	0	0	0	0	0	0	0	0	0	0	0
22	1029	300	6992	7245	3050	toneladas	5412228000	165732036000	94225629700	8662482300	15157568900	9744912000	133190931300	81191535200	21656005000	8967330960	67116563200
23	1082	348	6992	7245	3050	toneladas	0	2170355800	20617598500	0	52084517400	0	0	3255107800	0	3254920000	0
24	1133	372	6992	7245	3050	toneladas	2172109800	0	0	0	0	1086048700	0	14123031600	4345801900	0	6518340600

Ilustración 26: Contenido Tabla Emissions Land Use Fire

La finalidad de esta base de datos es almacenar información de manera organizada y de simple acceso para el futuro.

4.4. Framework destinado al análisis de datos.

Una vez se han completado las fases de extracción, transformación y carga (ETL) de los datos en una base de datos estructurada, es primordial aplicar una tercera capa que nos permita realizar análisis más visuales, intuitivos y dinámicos. Dicha capa facilita la comprensión de los datos, identificar posibles anomalías y tomar decisiones basadas en evidencias visuales de forma mucho más eficiente.

Por ello, se ha empleado una aplicación que convierte los datos recibidos desde la base de datos en representaciones visuales más accesibles para usuarios cuyo perfil no está basado en ser expertos en Big Data o no posee conocimientos informáticos.

Por ello, en este proyecto se ha empleado un framework de Business Intelligence, específicamente la aplicación Power BI, la cual nos permitirá crear paneles de control, visualizaciones interactivas e informes más dinámicos a partir de los datos almacenados en las bases de datos. Esta aplicación facilita al usuario final la posibilidad de poder acceder a la información sin necesidad alguna de ejecutar consultas o poseer conocimientos de SQL o programación.

Estos gráficos o visualizaciones generadas en Power BI permiten interpretar grandes cantidades de información de forma visual e inmediata, lo que resulta importante para:

- La toma de decisiones informadas basadas en evidencias y tendencias.
- Explorar información de forma más interactiva, pudiendo aplicar filtros, jerarquías o segmentos.
- Comparar series temporales, indicadores clave (KPIs) y comportamientos de variables.
- Detectar variaciones o anomalías en los datos.

Esto aporta un valor añadido muy significativo al análisis de datos, especialmente en proyectos en los cuáles los datos pueden ir variando temporalmente.

Por lo tanto, para esta parte del proyecto, se han desarrollado una serie de frameworks de análisis dinámico que utilizan los datos extraídos, transformados y cargados en la base de datos previamente. Cada uno de estos gráficos han sido diseñados con la finalidad de representar diferentes dimensiones de un fenómeno, como la evolución de las dimensiones por país y año, por producto o incluso por fuente de datos. Además, se ha realizado una configuración que permite la conexión entre los dashboards y la base de datos con el fin de que, si los datos van a ser modificados o actualizados, estas operaciones se verán reflejadas en el framework de forma automática. Esto nos garantiza que el análisis realizado sea sobre la información más reciente y evita la necesidad de estar reconstruyendo manualmente los informes tras cada actualización.

Se han diseñado dos pestañas los cuáles proporcionan información acerca de las dos principales fuentes de datos (Emisiones en bosques y emisiones por incendios). Estos son las pestañas y sus diseños:

1. Bosques.

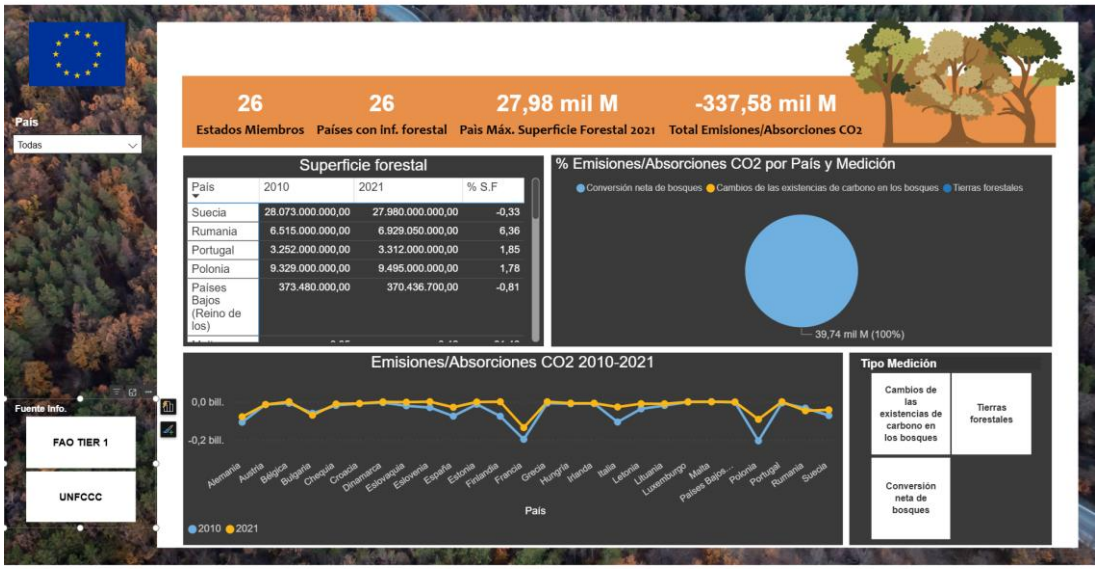


Ilustración 27: Pestaña Bosques Power BI

1.1. Indicadores clave superiores.

Ilustración 28: Claves superiores Bosques Power BI

En esta pestaña, en la parte superior se pueden observar cuatro indicadores clave que resumen la situación forestal y climática en los países analizados. Estos gráficos se encuentran en la parte superior con diferente color para captar la atención del usuario, ofreciendo una visión rápida y comparativa de las variables más relevantes del conjunto de datos que se ha tratado previamente. Por ello, la información clave presentada es la siguiente:

1. El número de países de los estados miembros de los cuáles se puede proporcionar información forestal y/o climática en la base de datos.

2. Número de países de los que se dispone información forestal.

3. El dato del país con mayor superficie forestal en el último año disponible (siendo 2021), expresado en millones de hectáreas y destacando la magnitud de sus recursos forestales.

4. Balance neto de emisiones y absorciones de CO2, expresado en valores absolutos negativos o positivos, permitiendo identificar si en conjunto las tierras forestales están funcionando como emisores o como sumideros de carbono.

Dichos indicadores pueden ser filtrador por el desplegable lateral que hace referencia a la información que quiere ser mostrada por país. En el caso siguiente, como ejemplo, filtraremos para obtener los valores relacionados a Alemania :

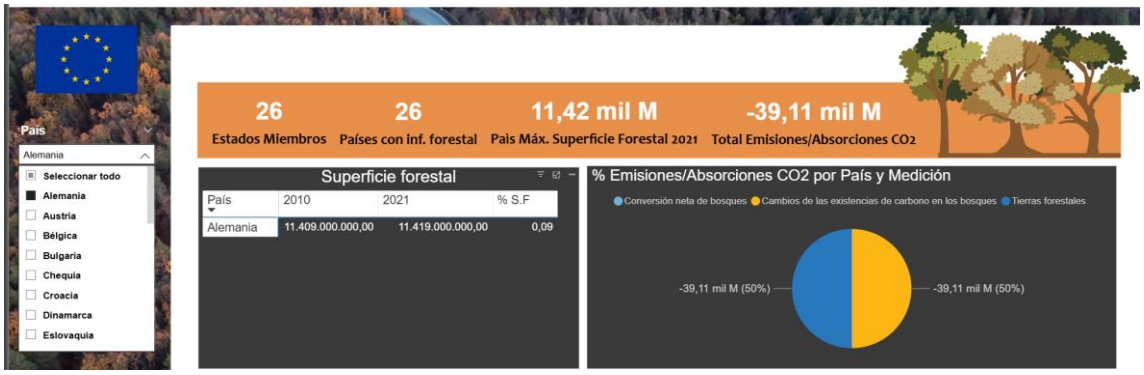


Ilustración 29: Gráficos filtrados Bosques Power BI

1.2. Superficie forestal.

Superficie forestal

País	2010	2021	% S.F
Suecia	28.073.000.000,00	27.980.000.000,00	-0,33
Rumania	6.515.000.000,00	6.929.050.000,00	6,36
Portugal	3.252.000.000,00	3.312.000.000,00	1,85
Polonia	9.329.000.000,00	9.495.000.000,00	1,78
Países Bajos (Reino de los)	373.480.000,00	370.436.700,00	-0,81

Ilustración 30: Gráfico Superficie forestal Bosques Power BI

En este gráfico incluido en el panel, se proporciona una visión comparativa de la evolución de la superficie forestal en distintos países europeos entre los años 2010 y 2021. A través de esta tabla, se puede manejar la siguiente información:

- Superficie forestal en el año 2010 (en hectáreas)
- Superficie forestal en el año 2021(en hectáreas)
- Porcentaje de cambio relativo (%S.F) entre ambos años, permitiendo identificar tendencias de crecimiento o pérdida debido a deforestación, incendios u otros factores.

Este gráfico destaca por su formato numérico, donde los valores aparecen representados con separadores de miles y decimales, permitiendo una lectura rápida y precisa de valores de gran magnitud.
Por otro lado, al estar implementado sobre una visualización dinámica, permite filtrar, ordenar o segmentar los datos en tiempo real agrupando por países o enfocándose en un año completo:

Ilustración 31: Gráfico Superficie forestal con filtro por país Bosques Power BI

O en función de la fuente de información :

Ilustración 32: Gráfico Superficie forestal con filtro por origen Bosques Porwer BI

1.3. % Emisiones/Absorciones CO2 por país y medición.

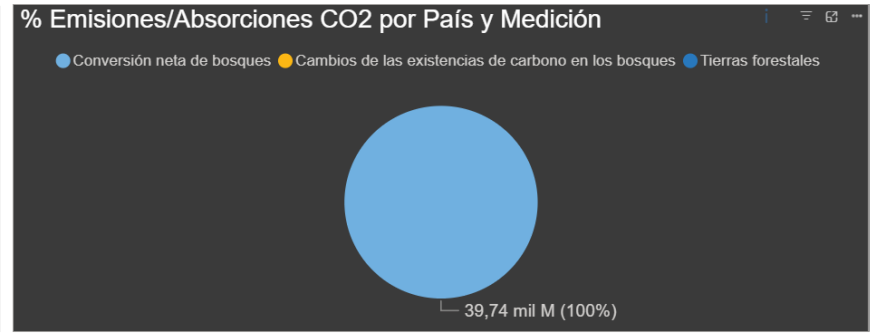


Ilustración 33: Gráfico Emisiones Bosques Power BI

Este gráfico representa de una forma clara y visual cómo se distribuyen las emisiones y absorciones de dióxido de carbono. A través de él, permite observar en temas porcentuales qué proporción de CO2 total es atribuible a cada una de las siguientes categorías de medición definidas:

- **Conversión neta de bosques** : Relacionado a los cambios forestales debido a la transformación de tierras en uso como agricultura, urbanización, etc. Esta medición recoge tanto las pérdidas como las ganancias de superficie forestal, reflejando el impacto sobre las emisiones o absorciones de CO2.

- **Cambios de las existencias de carbono en los bosques** : Variaciones en el carbono almacenado en la biomasa forestal viva o muerta, suelos y la materia orgánica presente en bosques. Pueden deberse estos cambios a la tala, incendios, plagas, etc.

- **Tierras forestales** : Medición referenciada a las emisiones o absorciones derivadas de la gestión o uso de las tierras clasificadas como forestales. Incluye prácticas como la reforestación o la deforestación parcial, representando la contribución directa de este tipo de suelo en las emisiones/absorciones de CO2.

La finalidad de este gráfico tipo donut es proporcionar un análisis visual y rápido de cada tipo de medición sobre el total de las emisiones o absorciones de CO2, ayudando a identificar el mayor contribuyente neto

en términos de carbono y reforzando la comprensión del impacto relativo de cada práctica sobre el balance global de carbono.

Esta visualización vinculada directamente a la base de datos que centraliza los datos transformados permite ser filtrado por país para comprobar cuáles son las principales mediciones que aportan al % de emisiones/absorciones de CO2, así como por la fuente de origen de datos:

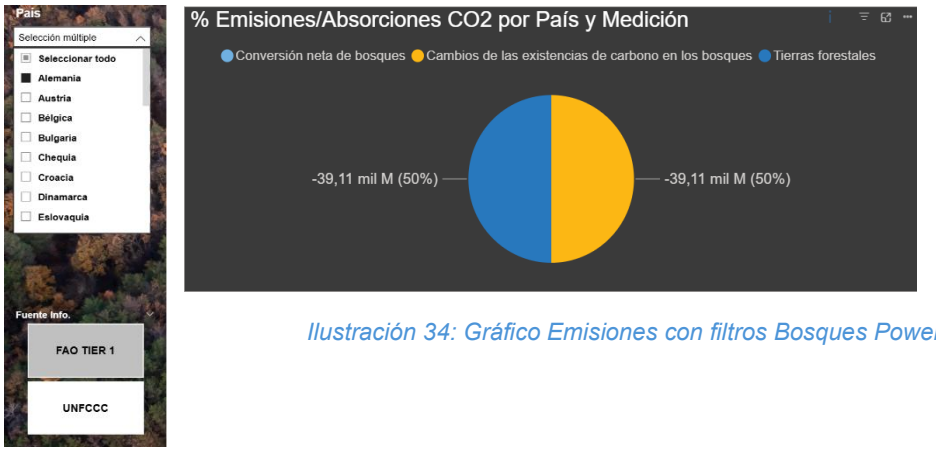


Ilustración 34: Gráfico Emisiones con filtros Bosques Power BI

Por otro lado, también es posible filtrar por tipo de medición directamente a partir del filtro aplicado en la parte inferior del gráfico:

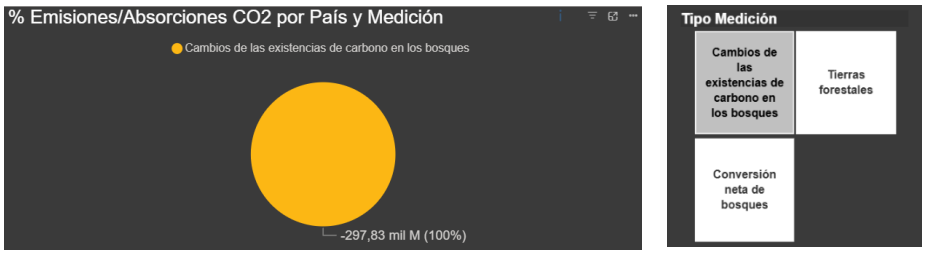


Ilustración 35: Gráfico Emisiones por tipo Medición Bosques Power BI

1.4. Emisiones/Absorciones CO2 2010-2021

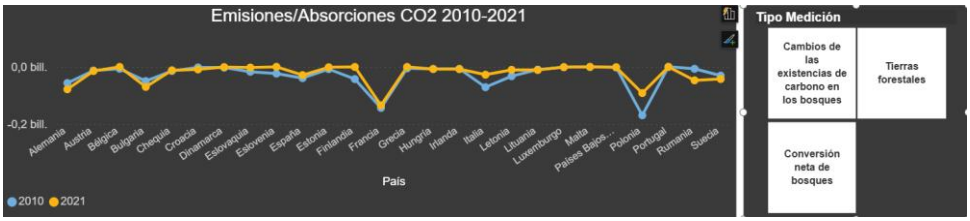


Ilustración 36: Gráfico Emisones/Absorciones Bosques Power BI

Este gráfico de líneas compara las emisiones y absorciones netas de CO2 registradas en los años 2010 y 2021 para cada país miembro. Se representa mediante dos líneas (línea azul para 2010 y línea amarilla para 2021), permitiendo visualizar la evolución del balance de carbono entre el año 2010 y el año 2021.

A través de la información reflejada se puede saber que, si el valor se encuentra por debajo de cero, indica que el país ha absorbido más CO2 del que ha emitido. En caso contrario, implica que hay un exceso de emisiones, contribuyendo al aumento de gases de efecto invernadero.

Este gráfico tiene gran importancia ya que permite comparar entre países y evaluar el impacto relativo en función de sus prácticas de gestión de tierras y bosques, así como identificar tendencias de mejora o posible deterioro en sus políticas forestales y climáticas.

Toda esta información puede ser filtrada por el filtro lateral relacionado con el país, así como el tipo de medición:

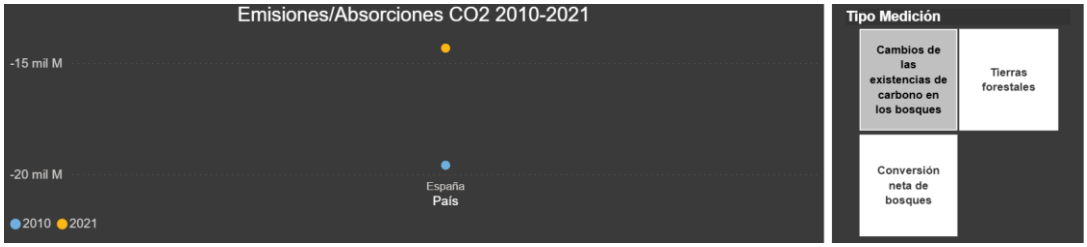


Ilustración 37: Gráfico Emisones/Absorciones por Tipo Medición Bosques Power BI

2. Incendios.

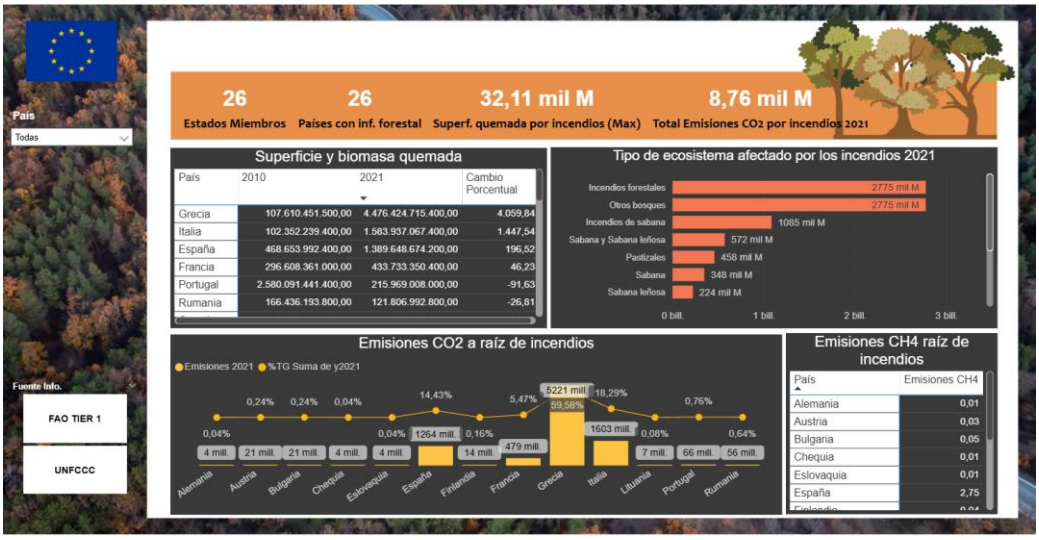


Ilustración 38: Pestaña Incendios Power BI

2.1. Indicadores clave superiores.

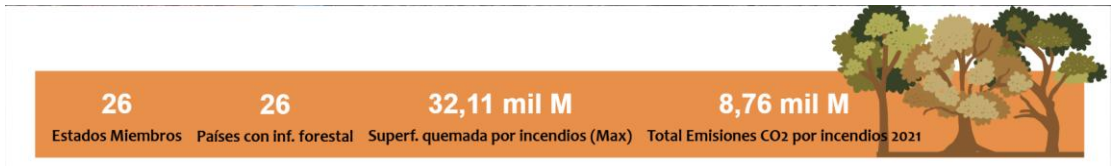


Ilustración 39: Claves superiores Incendios Power BI

Una vez se ha realizado la extracción, transformación y carga de los datos a la base de datos, se han generado paneles interactivos que permiten un análisis más interactivo e intuitivo. En este caso, se tratan de indicadores clave relacionados con los incendios forestales, los cuales resumen la información más relevante para una rápida interpretación por parte del usuario final.

En estos indicadores podemos encontrarnos:

- **Estados miembros:** Indica el número de países considerados en el análisis (disponibles con información en la tabla area_m49), todos pertenecientes a la Unión Europea.
- **Países con información forestal:** Representa cuántos de esos países disponen de información relevante sobre la temática forestal y su impacto en relación con los incendios ocasionados en dichos países.

- **Superficie quemada por incendios:** Muestra el valor máximo registrado en hectáreas de superficie forestal afectada por incendios entre los países que se analizan. Esta información nos permite identificar el nivel de afectación medioambiental más alto durante el año 2021.
- **Total emisiones de CO2 por incendios (2021):** Refleja el volumen total de dióxido de carbono emitido debido a los incendios forestales durante el año 2021. Este indicador es especialmente importante para estudios sobre cambio climático y evaluación de políticas medioambientales.

En el caso de este gráfico, para la superficie quemada máxima, aparece el valor del país con mayor superficie quemada en el año 2021. Sin embargo, puede ser filtrado por el filtro de 'País' el cual permitirá conocer la máxima superficie quemada en el año 2021 para un país específico.
Para el indicador relacionado con las emisiones de CO2 por incendios, dicho filtro también puede ser aplicado.

2.2. Superficie y biomasa quemada.

Superficie y biomasa quemada

País	2010	2021	Cambio Porcentual
Grecia	107.610.451.500,00	4.476.424.715.400,00	4.059,84
Italia	102.352.239.400,00	1.583.937.067.400,00	1.447,54
España	468.653.992.400,00	1.389.648.674.200,00	196,52
Francia	296.608.361.000,00	433.733.350.400,00	46,23
Portugal	2.580.091.441.400,00	215.969.008.000,00	-91,63
Rumania	166.436.193.800,00	121.806.992.800,00	-26,81

Ilustración 40: Gráfico Superficie y biomasa quemada Incendios Power BI

Este gráfico de tabla comparativo permite tener una visión de la superficie y biomasa forestal que se ha quemado en los países europeos de los que se dispone información entre los años 2020 y 2021. Dicha tabla aparece representada con las siguientes columnas:

- Año 2010 y 2021: Proporcionan información del volumen total de superficie y biomasa quemada en cada uno de esos años. Los valores se expresan en unidades grandes, lo que permite y facilita observar la magnitud de los daños entre ambas fechas.

- Cambio porcentual: Calcula la variación relativa entre los años 2010 y 2021, facilitando identificar tendencias de aumento o disminución de los incendios forestales y su impacto.

A través de esta tabla se puede detectar cuáles han sido los países que han experimentado un drástico aumento de su superficie quemada debido a los incendios, así como detectar cuáles son las tendencias preocupantes permitiendo tomar estrategias de prevención más estrictas.

Por ello, esta visualización permite que se pueda poner un contexto al efecto de los incendios en el tiempo y ofrecer una base sólida sobre la que tomas decisiones relacionadas con las políticas medioambientales, la reforestación y la gestión del cambio climático.

2.3. Tipo de ecosistema afectado por los incendios 2021.

Ilustración 41: Gráfico Tipo ecosistema afectado Incendios Power BI

El siguiente gráfico de barras horizontales muestra las principales causas de los incendios forestales ocasionados durante el año 2021. Se encuentran clasificadas por tipo de ecosistema afectado. Estos valores permiten darnos a conocer el impacto en términos de superficie afectada. Dichas categorías son las siguientes:

Incendios forestales
Bosque tropical húmedo
Otros bosques
Incendios de sabana
Matorrales cerrados
Pastizales
Matorrales abiertos
Sabana
Sabana leñosa
Incendios en suelos de turba
Matorrales cerrados y abiertos
Sabana y Sabana leñosa

Este gráfico permite identificar los tipos de vegetación más afectados por el fuego, priorizar técnicas de prevención y reforestación en áreas más dañadas y comprender cómo estos tipos de ecosistema se ven afectados por la presencia humana o por el cambio climático.

2.4. Emisiones de CO2 a raíz de incendios.

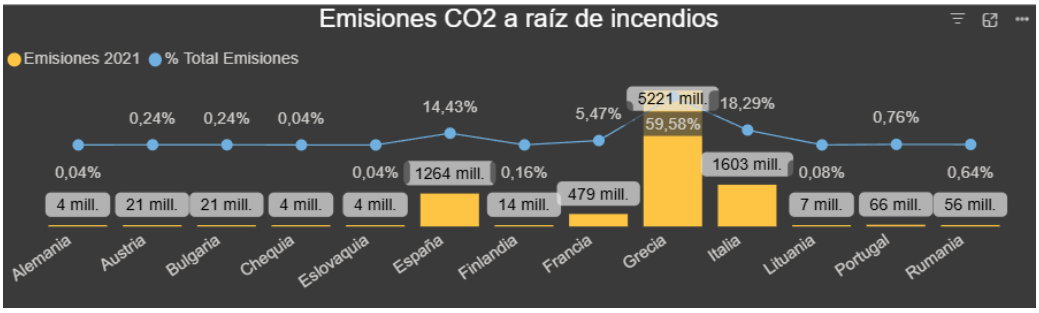


Ilustración 42: Gráfico Emisiones CO2 Incendios Power BI

Este gráfico representa un análisis comparativo de las emisiones de dióxido de carbono derivadas de los incendios forestales en varios países durante el año 2021. Su visualización aparece representada de dos maneras:

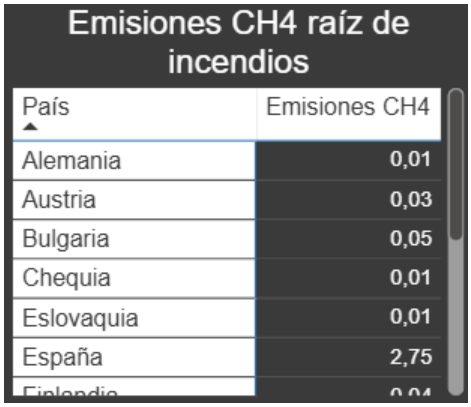
- **Barras amarillas**: Volumen absoluto de emisiones de CO2 por país en el año 2021.
- **Línea azul**: Porcentaje que representa cada país sobre el total de emisiones que se han registrado.

Con estos datos visuales es posible conocer la magnitud que tienen las emisiones por país, siendo en este caso Grecia la principal emisora de dióxido de carbono de los países representados. En su caso, supone aproximadamente un 60% del total de las emisiones totales. Con esta información, se puede llegar a saber el impacto relativo de un país en el conjunto europeo, así como los países que tienen menor impacto y aplicar medidas de prevención de incendios aplicadas en esos países sobre los más afectados.

Con este gráfico se pueden tomar medidas de mitigación y adaptación climática, estudios sobre los ciclos de carbono y sostenibilidad ambiental o políticas de adaptación climática.

Su combinación de datos facilita la toma de decisiones en base al volumen de emisiones a nivel nacional.

2.5. Emisiones de CH4 a raíz de incendios.



Emisiones CH4 raíz de incendios

País	Emisiones CH4
Alemania	0,01
Austria	0,03
Bulgaria	0,05
Chequia	0,01
Eslovaquia	0,01
España	2,75
Finlandia	0,04

Ilustración 43: Gráfico Emisiones CH4 Incendios Power BI

La tabla presenta una comparativa de las emisiones de metano producidas debido a los incendios forestales por país. Este tipo de tablas permiten un análisis más detallado del impacto ambiental que general los incendios más allá del análisis del dióxido de carbono producido, centrándose en otro gas de efecto invernadero que afecta significativamente a corto plazo al calentamiento global.

Esta visualización permite comprobar que países como España aportan valores más elevados de emisiones de metano, lo que afirma que hubo una mayor intensidad de incendios durante el período de estudio (año 2021).

Este análisis está centrado específicamente tanto a nivel científico como a nivel político, debido a que permite priorizar las áreas geográficas para unas medidas de intervención específicas, mejorando los planes de prevención de incendios y evitando aumentar sus efectos sobre el cambio climático.

En esta pestaña, especialmente, se pueden ir aplicando filtros a nivel de país, así como a nivel de fuente de origen de los datos. Esto es debido a la conexión relacional de los campos ID de cada tabla.

Ilustración 44: Conexión relacional tablas Power BI

De esta forma, este modelo relacional permite una estructura flexible, escalable y limpia para poder trabajar con grandes volúmenes de datos, haciendo que el Power BI pueda generar visualizaciones más dinámicas, intuitivas y adaptadas a las necesidades del usuario. Finalmente, el usuario podrá acceder a información de diferentes tablas aplicando una serie de filtros específicos sin que los datos se vean afectados o se representen erróneamente.

5. Pruebas.

5.1. Pruebas de carga e integración de datos.

En cualquier proyecto la realización de pruebas de carga e integración de datos es crucial, debido a que de estas pruebas depende el resultado final y correcto que debe transferirse a la base de datos. Por ello, las pruebas han de ser exhaustivas para garantizar la correcta carga y transformación de la información procedente de fuentes externas.

En consecuencia, como parte de estas pruebas durante el desarrollo de este proyecto, se han realizado en la herramienta de Apache Hop, entorno ETL orientado al diseño visual de pipelines y workflows de integración de datos.

Durante las pruebas, se han validado pipelines cuya estructura sigue la lógica de los procesos ETL, compuestas por los siguientes pasos:

- Lectura de datos desde un archivo CSV (CSV Land Use Fire Input).
- Transformación de los datos mediante un bloque de código JavaScript que ajusta formatos numéricos y realiza validaciones.
- Selección y renombrado de campos con Select Values.
- Carga final de los datos en una base de datos relacional a través del bloque Insert/Update.

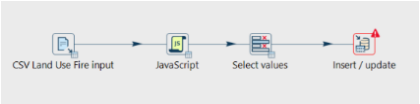


Ilustración 45: Error Carga datos

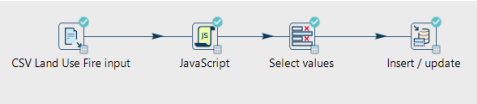


Ilustración 46: Correcta Carga datos

Durante dichas pruebas, se han identificado y prevenido diversos errores que pueden llegar a surgir durante la ejecución de los pasos del pipeline. Dichos errores pueden deberse a las siguientes causas:

- **Errores de formato de datos**: Al convertir cadenas de separadores de miles a tipo numérico sin eliminar los puntos o intentar adaptar correctamente la coma decimal.
- **Valores nulos o especiales no encontrados**: Como la aparición de NOT_FOUND o valores NULL que no estaban tratados anteriormente y que causan errores en tiempo de ejecución. Este error puede producirse especialmente en el paso de JavaScript.
- **Desajustes de nombres de campos**: En pasos como Select Values o Insert/Update, si no se renombran correctamente los campos intermedios, se produce un error por campo no encontrado.
- **Conflictos de tipo de dato en la base de datos**: Al intentar insertar un campo numérico con formato texto o con puntos decimales mal definidos en una columna que esté definida como float o numeric.

- **Errores de clave primaria duplicada**: En el paso de Insert/Update, si no se definen correctamente las claves de búsqueda (lookup fields), pueden generarse inserciones erróneas o duplicados.

Estas pruebas han sido fundamentales para que el proceso de integración funcione de la forma más correcta y sin pérdidas ni alteraciones de datos al transferirlos desde los ficheros fuente a la base de datos final. Este control inicial permite que, posteriormente, herramientas como Power BI trabajen con datos más limpios y estructurados.

Ilustración 47: Error en campo no encontrado en Insert/Update

Ilustración 48: Error de campo no tratado en JavaScript

La transmisión correcta de los registros se puede comprobar a través del Data Preview que contiene cada uno de los pasos, mostrándonos los datos con los que se está trabajando sin tener que ver si el error se ha producido al cargar la información en la base de datos.

Este Data Preview tiene gran importancia debido a que nos muestra toda la información que se ha ido transfiriendo a lo largo de cada uno de los pasos que componen el Pipeline, permitiéndonos ver dónde se está produciendo la carga errónea del dato o su malinterpretación.

Ilustración 49: DataPreview del paso Select Values

En relación con las pruebas de carga e integración de datos, se han ido realizando varias ejecuciones de los pipelines y su comprobación durante o posterior a la ejecución de la misma, pudiendo observar los datos en el DataPreview disponible o en la base de datos tras su carga. Por ello, se han contemplado errores en la carga de datos debido a que se están teniendo en cuenta los puntos como miles y como decimales, debido a ello se cargan erróneamente en la base de datos.

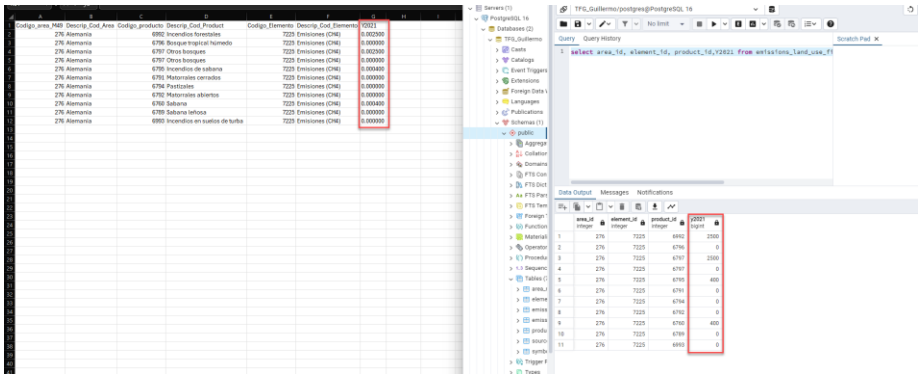


Ilustración 50: Comparación de KPI Y2021 entre fuente de datos y BBDD

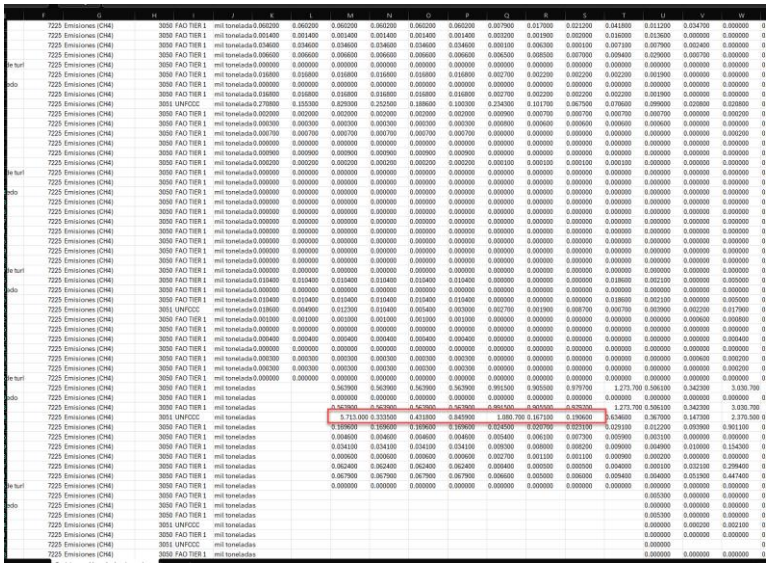


Ilustración 51: Ejemplo en el origen de datos decimales y miles

Debido a ello, se aplica un paso con script java para la interpretación correcta de los KPIs. Como paso principal, se realizó el cambio de los KPIs a tipo string y, posteriormente, en el pasa JavaScript se realiza el cambio de dichos indicadores a números.

Como parte del desarrollo del código, se aplican una serie de ifs los cuáles dejan el valor como está si es decimal o reemplaza el valor de los miles.

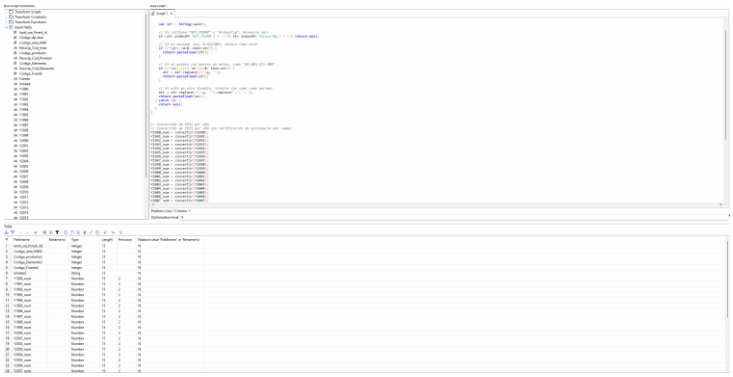


Ilustración 52: Código Java aplicado para la resolución del problema

Cabe destacar, que es imprescindible añadir las columnas de salida que se desean en el apartado 'fields' del paso JavaScript como aparece reflejado en la screenshot anterior. Si no se añaden dichos campos, dará error de falta de columnas de salida. Dichos nombres que saldrán como columnas de salida, deben tener un nombre diferente debido a que no estamos tomando las columnas de origen, si no que estamos aplicando transformaciones en dichos campos.

Si dichos campos de salida del JavaScript no son renombrados en el siguiente paso correctamente, saldrá un error de ejecución debido a que no encuentra dichos nombres de campos en el destino:

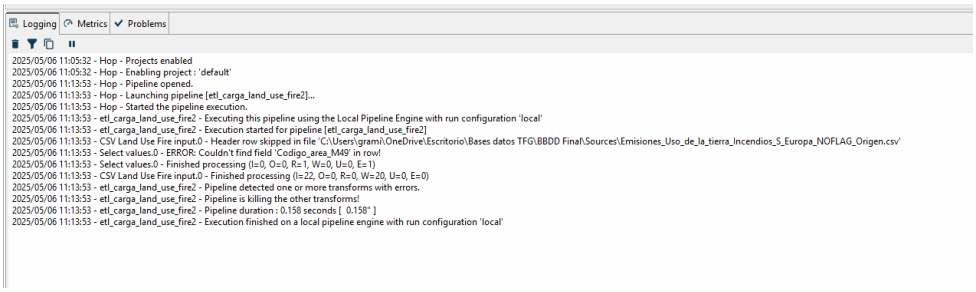


Ilustración 53: Error log de campo no encontrado en Select Values

#	Fieldname	Rename to	Length	Precision
1	land_use_fire_ID2	land_use_fire_id	15	
2	Codigo_area_M49	area_id	15	
3	Codigo_producto2	product_id	15	
4	Codigo_Elemento2	element_id	15	
5	Codigo_Fuente2	source_id	15	
6	Unidad2	unity	15	
7	Y2010_num	Y2010	15	2
8	Y2011_num	Y2011	15	2
9	Y2012_num	Y2012	15	2
10	Y2013_num	Y2013	15	2
11	Y2014_num	Y2014	15	2
12	Y2015_num	Y2015	15	2
13	Y2016_num	Y2016	15	2
14	Y2017_num	Y2017	15	2
15	Y2018_num	Y2018	15	2
16	Y2019_num	Y2019	15	2
17	Y2020_num	Y2020	15	2
18	Y2021_num	Y2021	15	2
19	Y2022_num	Y2022	15	2

Ilustración 54: Renaming de los campos en Select Values

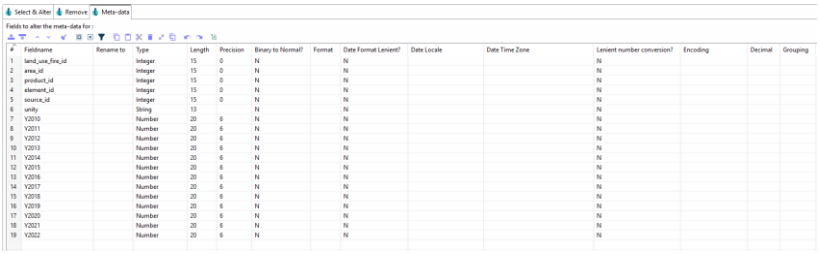


Ilustración 55: Aplicación del tipo de dato en Select Values tras renaming

Si no se realiza una correcta aplicación del renaming, así como del tipo de dato que debe transferirse al Insert/Update, se producirán errores de ejecución en dicho paso debido a que no encuentra el campo:

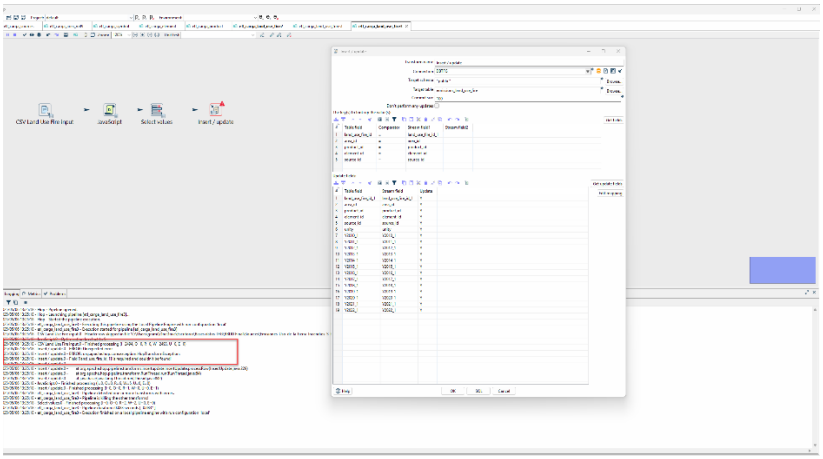


Ilustración 56: Ejemplo error de búsqueda de campo en Insert/Update

Una vez comprobado el estado de las cargas y las actualizaciones de datos en Apache Hop, se comprueba si dichas tablas se encuentran con datos disponibles y bien cargadas en la base de datos.

```
2025/05/14 11:14:01 - Hop - Projects enabled
2025/05/14 11:14:01 - Hop - Enabling project : 'default'
2025/05/14 11:43:38 - Hop - Pipeline opened.
2025/05/14 11:43:38 - Hop - Launching pipeline [etl_carga_land_use_firefinal]...
2025/05/14 11:43:38 - Hop - Started the pipeline execution.
2025/05/14 11:43:38 - etl_carga_land_use_firefinal - Executing this pipeline using the Local Pipeline Engine with run configuration 'local'
2025/05/14 11:43:38 - etl_carga_land_use_firefinal - Execution started for pipeline [etl_carga_land_use_firefinal]
2025/05/14 11:43:38 - CSV Land Use Fire input.0 - Header row skipped in file 'C:\Users\grami\OneDrive\Escritorio\Bases datos TFG\BBDD Final\Sources\Emisiones_Uso_de_la_tierra_Incendios_S_Europa_NOFLAG_Origen.csv'
2025/05/14 11:43:38 - JavaScript.0 - Optimization level set to 9.
2025/05/14 11:43:38 - CSV Land Use Fire input.0 - Finished processing (I=2494, O=0, R=0, W=2493, U=0, E=0)
2025/05/14 11:43:39 - JavaScript.0 - Finished processing (I=0, O=0, R=2493, W=2493, U=0, E=0)
2025/05/14 11:43:39 - Select values.0 - Finished processing (I=0, O=0, R=2493, W=2493, U=0, E=0)
2025/05/14 11:43:39 - Insert / update.0 - Finished processing (I=2493, O=0, R=2493, W=2493, U=0, E=0)
2025/05/14 11:43:39 - etl_carga_land_use_firefinal - Pipeline duration : 1.133 seconds [ 1.133s ]
2025/05/14 11:43:39 - etl_carga_land_use_firefinal - Execution finished on a local pipeline engine with run configuration 'local'
```

Ilustración 57: Log en pipeline de ejecución

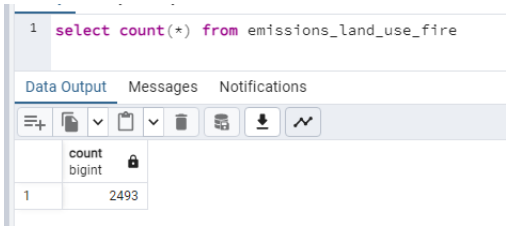


Ilustración 58: Total registros cargados en tabla de BBDD

Además, los datos han de comprobarse visualmente para cerciorarnos de que la información coincide tanto en orden como en valor. Como resultado final disponemos de 7 tablas generadas a partir de los 3 ficheros exportados.

5.2. Pruebas con presentación de datos.

Tras haber completado el proceso ETL de los datos en una base de datos relacional, es imprescindible validar no solo que los pipelines creados sean correctos y cumplan su funcionalidad, sino también la fidelidad con la que los datos son representados en herramientas de visualización y análisis. Este apartado tiene como objetivo evaluar cómo se han presentado los datos transformados que se encuentran disponibles en la base de datos, a través del uso de herramientas visuales, garantizando la calidad de estos para el usuario final.

Un punto esencial que ha de tenerse en cuenta en este punto es la calidad del dato cargado: su limpieza, precisión y coherencia estructural. Esto se suma a la necesidad de asegurar una correcta conexión y estable entre la base de datos y la herramienta de visualización, en este caso el Power BI. Debido a que dicha herramienta puede acceder a la base de datos en tiempo real o a través de actualizaciones programadas mediante modelos, también se comprueba que cualquier modificación posterior en la base de datos se refleje correctamente en los informes,

asegurando que los gráficos estén siempre alineados con el estado actual de los datos.

El primer paso que se ha de realizar tras conectar Power BI con la base de datos y seleccionar las tablas, es que dichas tablas aparezcan en el apartado de Datos dentro de la aplicación para su uso posterior.

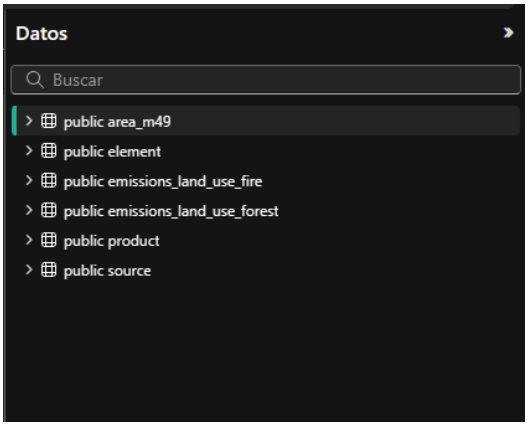


Ilustración 59: Datos cargados en Power Bi

Mientras que una base de datos permite consultar grandes volúmenes de datos mediante lenguaje SQL, Power BI proporciona un entorno visual más intuitivo para verificar la consistencia y utilidad de estos. Su interfaz visual permite detectar errores en formatos o relaciones incorrectas entre tablas con mayor agilidad. Además, esta herramienta facilita la construcción de medidas y KPIs personalizados, facilitando muchos aspectos funcionales respecto a usar consultas en SQL o inspecciones de tablas.

Uno de los componentes más poderosos en Power BI es la “Segmentación de datos”, que permite filtrar de una forma más dinámica los elementos visuales del informe en función de valores seleccionados (como país, fuente de origen, etc.). La segmentación permite al usuario final poder navegar e interactuar con los datos sin necesidad alguna de conocimientos técnicos, facilitando un análisis más específico.

Gracias a estos filtros visuales, es posible validar también que las relaciones entre tablas han sido definidas correctamente y que los gráficos responden de forma coherente a las elecciones del usuario.

Comprobando así que no solo la integridad de los datos, sino también su navegabilidad es correcta.

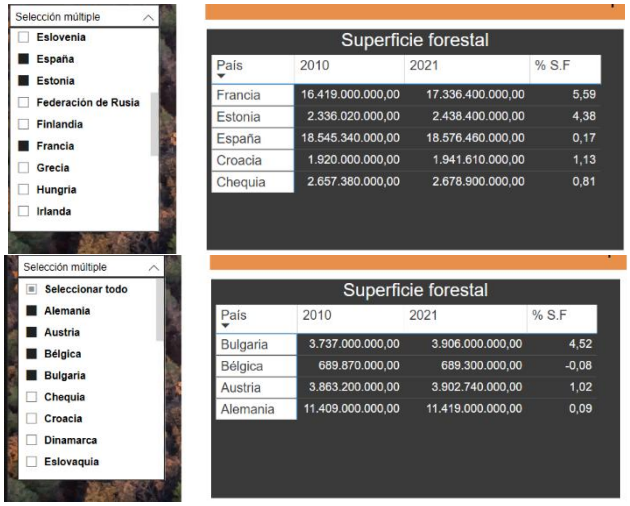


Ilustración 60: Filtros con múltiples opciones en Power BI

A través de la selección establecida, se puede llegar a comprobar que los gráficos varían en función de esta y que la información proporcionada cumple satisfactoriamente. En este caso, se puede llegar a realizar una comparación entre la superficie forestal en los años 2020 y 2021 entre los países seleccionados, así como un análisis del porcentaje de superficie que se ha perdido o ganado entre ambas fechas.

Sin embargo, estos gráficos no pueden compararse todos con todos, puesto que proporcionan información diferente (tipo de mediciones, porcentajes de emisiones y/o absorciones, etc.).

Por ello, es preciso que los gráficos sean usados empleando el sentido común, ya que comparando superficie forestal en hectáreas con emisiones en toneladas no proporcionaría un resultado coherente.

Finalmente, para plasmar correctamente los datos en los diferentes gráficos, ha sido preciso crear las relaciones entre las tablas en la herramienta a partir del apartado ‘Vista de modelo’. Esto se debe a que para poder aplicar los filtros y mostrar los datos correctamente o, incluso, emplear medidas de distintas tablas en los gráficos es preciso que existan relaciones entre tablas.

6. Acceso al fichero PBI.

En el caso de este proyecto, se ha optado por exportar el fichero .pbix para que posteriormente se pueda abrir desde la aplicación Power BI. Por ello, se explicará a lo largo de este punto cómo es posible acceder al archivo desde la propia plataforma web para que el usuario pueda tener una satisfactoria experiencia sin necesidad de instalación de ninguna herramienta.

Esta herramienta está enfocada para que los usuarios puedan acceder a los ficheros .pbix generados siempre y cuando se encuentren en una organización, ya sea académica o corporativa. Por ello, es imprescindible que se disponga de un correo corporativo/académico.

A través del siguiente link, se accederá a la aplicación web:

https://www.microsoft.com/en-au/power-platform/products/power-bi/

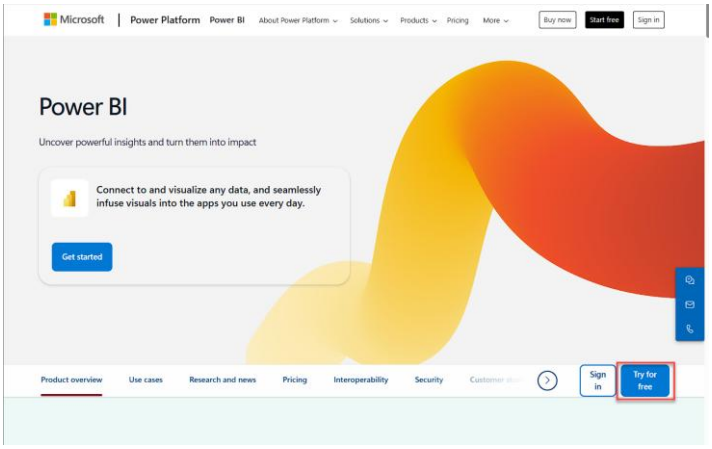


Ilustración 61: Acceso a la página web, marcando dónde debe accederse

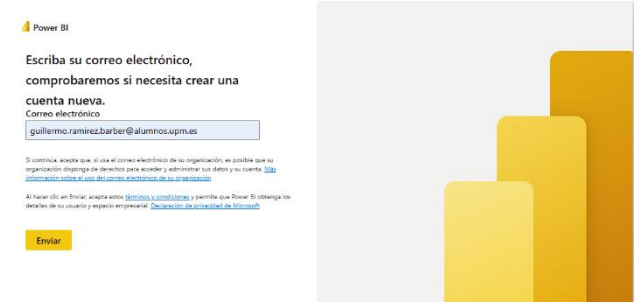


Ilustración 62: Correo electrónico

Ilustración 63: Inicio de sesión en UPM

Se abrirá una ventana externa para el inicio de sesión en la Universidad Politécnica.

Posterior a ello, se procederá a realizar el inicio de sesión en la aplicación donde accederemos a la pestaña “Mi área de trabajo” situada en la parte izquierda de la pantalla y seleccionaremos “importar”. Aparecerá una lista donde se debe seleccionar “Informe, informe paginado o libro” y, posteriormente, “Desde este equipo”.

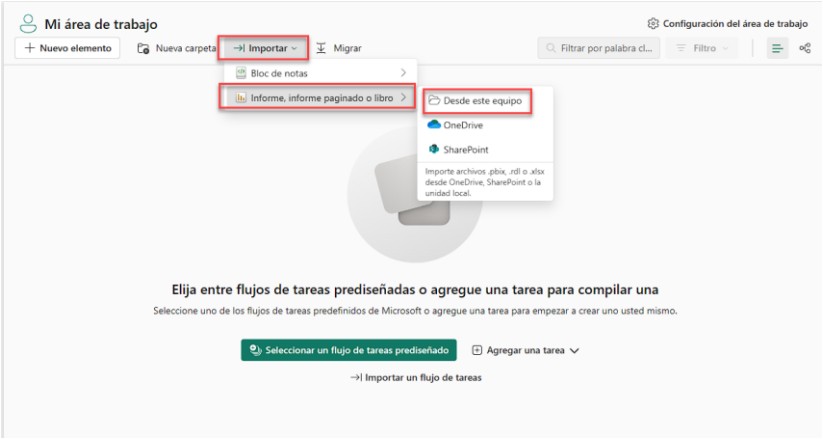


Ilustración 64: Importar informe

De esta forma, es posible acceder a las 5 gráficas diseñadas a través de los datos oficiales de la FAO.

7. Consideraciones Éticas, Legales y de Sostenibilidad.

A lo largo del desarrollo de cualquier proyecto tecnológico no solo ha de contemplarse su viabilidad técnica, sino también su adecuación a principios éticos, normativas legales y a su impacto medioambiental. En este Trabajo de Fin de Grado, se ha trabajado con un proceso ETL que implica la extracción, transformación y carga de datos abiertos proporcionados por la FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura) en una base de datos para su posterior visualización mediante la herramienta Power BI. Esta etapa final tiene como objetivo reflexionar sobre las implicaciones que este tipo de proyectos conllevan, especialmente cuando se trabaja con grandes volúmenes de datos públicos y herramientas de análisis avanzado.

Desde un punto de vista ético, este proyecto se basa exclusivamente en datos abiertos, lo que asegura que no se han vulnerado derechos de privacidad, confidencialidad o consentimiento informado. No se ha empleado información sensible ni personal, por lo que no existe riesgo directo para la integridad de individuos o entidades. Sin embargo, es responsabilidad del desarrollador asegurar la interpretación objetiva y responsable de los datos, evitando sesgos o conclusiones erróneas que puedan llevar a desinformación o decisiones equivocadas por parte de usuarios finales.

Legalmente, el proyecto cumple con los principios de la licencia de datos abiertos, conforme a las políticas de la FAO. Esta fuente de datos permite el uso, distribución y análisis de la información siempre que se reconozca su origen. Se ha respetado la propiedad intelectual de las herramientas utilizadas, como Apache Hop, PostgreSQL y Power BI, las cuales cuentan con sus respectivas licencias de uso. Además, el almacenamiento y gestión de datos en bases de datos locales no vulnera ninguna normativa relacionada con el tratamiento de datos personales, dado que no se ha trabajado con dicho tipo de información.

Aunque el impacto medioambiental de este proyecto sea indirecto, cabe señalar que el análisis de datos orientado a la gestión forestal, emisiones e incendios contribuye a una mayor conciencia ecológica y a la toma de

decisiones más informadas por parte de las instituciones y organismos. El uso eficiente de tecnologías digitales como Power BI permite recurrir la dependencia de recursos físicos y favorece el acceso remoto a la información, promoviendo prácticas más sostenibles en el manejo de datos medioambientales.

En conclusión, este proyecto se ha desarrollado con un gran compromiso hacia la ética, legalidad y sostenibilidad. El empleo de datos abiertos, el cumplimiento de licencias de uso de software y la contribución al análisis medioambiental mediante herramientas tecnológicas, garantizan que el trabajo so solo sea sólido técnicamente, sino también socialmente responsable. Esta reflexión reafirma la importancia de considerar estos tres pilares en cualquier desarrollo tecnológico, especialmente en aquellos con implicaciones en la gestión y análisis de datos públicos.

8. Conclusiones.

Tras la finalización del presente proyecto, se ha logrado cumplir el objetivo principal que se propuso: creación e implementación de un proceso ETL (Extract, Transform and Load) que permite extraer los datos abiertos de la FAO (Emisiones derivadas de fuegos y de bosques), transformarlos y almacenarlos de forma estructurada en una base de datos relacional MySQL. Por otro lado, se ha conseguido establecer una conexión entre la base de datos y la herramienta de visualización Power BI, posibilitando la representación interactiva y dinámica de los datos cargados. Este flujo ha permitido cerrar el ciclo completo del tratamiento y explotación de los datos, desde su origen hasta su presentación final en gráficos para análisis y posterior toma de decisiones.

Durante el proyecto, se han ido adquiriendo una serie de competencias clave relacionadas con las herramientas y los conceptos empleados. Entre los aprendizajes más significativos, se pueden destacar los siguientes:

- Diseño y ejecución de un proceso ETL completo mediante el empleo de la herramienta Apache Hop, comprendiendo la lógica de la integración de datos desde varias fuentes de origen, así como su validación y carga posterior.
- Gestión de las bases de datos en PostgreSQL/MySQL, incluyendo la generación de claves primarias y foráneas y ejecución de consultas SQL que sean eficientes para la validación de la información que se ha transformado.
- Automatización de transformaciones complejas en Apache Hop, como las conversiones de tipo de dato, limpieza de nulos o datos inconsistentes, así como la detección de errores en pipelines.
- Visualización de datos en Power BI, aplicando filtros, segmentaciones y componentes interactivos que permiten al usuario final acceder a la información de una forma más clara e intuitiva.

- Mejora en el procesamiento lógico y estructurado, crucial para planificar, construir y depurar un proceso ETL que garantice la calidad de los datos de principio a fin.

Gracias a la recopilación, transformación y almacenamiento de los datos provenientes de la FAO, el análisis ha permitido:

- Identificar tendencias temporales relacionadas con las emisiones de CO2, uso de la tierra y cambios en la superficie forestal.
- Detectar patrones atípicos que puedan derivarse de eventos climáticos o políticas.
- Proporcionar un soporte visual claro para la toma de decisiones, especialmente en entornos donde se requiere el análisis de datos por usuarios que no poseen conocimientos técnicos de informática. Por ello, Power BI ha supuesto ser fundamental al ofrecer representaciones gráficas intuitivas y segmentadas, permitiendo ajustar visualmente los análisis por país, tipo de medida o fuente de origen de los datos.

Cabe destacar que la realización de este proyecto ha implicado un profundo proceso de aprendizaje en múltiples disciplinas, combinando aspectos de ingeniería de datos, gestión de la información y visualización analítica. A lo largo del proyecto, se han podido evaluar ventajas y limitaciones de las herramientas utilizadas (como la facilidad de integración de Apache Hop, la versatilidad de Power BI o las capacidades relacionales de MySQL).

Por ello, se ha hecho evidente la importancia de la calidad de datos de origen para asegurar resultados fiables. El tratamiento riguroso de esta información durante el proceso ETL ha sido clave para que las visualizaciones en Power BI sean significativas y útiles.

En definitiva, el proyecto me ha permitido consolidar conocimientos técnicos y prácticos entorno a los procesos ETL y su aplicación en entornos reales de datos. Se ha logrado construir un flujo de datos robusto que comienza con la extracción de datos abiertos, pasando por una transformación de estos eficiente y finalizando con su visualización en un entorno dinámico e interactivo como Power BI. Este proyecto no

solo ha cumplido los objetivos planteados al inicio, si no que también ha potenciado mi capacidad analítica frente a herramientas profesionales del ecosistema de datos. La experiencia que he ido adquiriendo sienta una base sólida para abordar futuros proyectos de integración y explotación de información en contextos más profesionales.

ANEXO. BIBLIOGRAFIA.

Arsys. (s.f.). Infografía: ¿Qué ocurre con los datos en Internet en un minuto? Arsys. https://www.arsys.es/blog/infografia-datos-internet

Astera. (s.f.). ¿Qué es la validación de datos? https://www.astera.com/es/knowledge-center/what-is-data-validation/

Computing. (s.f.). La calidad del dato empieza en el origen y es la base del éxito corporativo. https://www.computing.es/encuentros/la-calidad-del-dato-empieza-en-el-origen-y-es-la-base-del-exito-corporativo/

ComputerHoy. (s.f.). La increíble cantidad de datos que se generan en internet cada minuto. https://computerhoy.20minutos.es/internet/cantidad-datos-generan-internet-cada-minuto-1351127

DataCamp. (s.f.). 9 ejemplos de dashboards en Power BI. https://www.datacamp.com/es/blog/9-power-bi-dashboard-examples

DataKitchen. (s.f.). The six types of data quality dashboards. https://datakitchen.io/the-six-types-of-data-quality-dashboards/

DataLadder. (s.f.). La integración de datos explicada: definición, tipos, proceso y herramientas. https://dataladder.com/es/la-integracion-de-datos-explicada-definicion-tipos-proceso-y-herramientas/

DataScientest. (s.f.). Data Warehouse: qué es y cómo utilizarlo. https://datascientest.com/es/data-warehouse-que-es-y-como-utilizarlo

FAO. (s.f.-a). Datos sobre incendios forestales. FAOSTAT. https://www.fao.org/faostat/es/#data/GI

FAO. (s.f.-b). Datos sobre superficies forestales. FAOSTAT. https://www.fao.org/faostat/es/#data/GF

FAO. (s.f.-c). Datos sobre cobertura vegetal. FAOSTAT. https://www.fao.org/faostat/es/#data/GV

GPStrategy. (s.f.). ¿Cómo interpretar correctamente un dashboard? https://gpstrategy.com.co/como-interpretar-correctamente-un-dashboard/

Hop Apache. (s.f.-a). Hop concepts. https://hop.apache.org/manual/latest/getting-started/hop-concepts.html

Hop Apache. (s.f.-b). Create your first pipeline. https://hop.apache.org/manual/latest/pipeline/create-pipeline.html

Marketing4eCommerce. (s.f.). Usuarios de Internet en el mundo. https://marketing4ecommerce.net/usuarios-de-internet-mundo/

Open Data Handbook. (s.f.). ¿Qué son los datos abiertos? https://opendatahandbook.org/guide/es/what-is-open-data/

Qlik. (s.f.). ETL (Extract, Transform, Load). https://www.qlik.com/us/etl

QuestionPro. (s.f.). Métodos cuantitativos: definición, tipos y ejemplos. https://www.questionpro.com/blog/es/metodos-cuantitativos/

SAS Institute. (s.f.). ¿Qué es ETL? https://www.sas.com/es_es/insights/data-management/what-is-etl.html

Unite.AI. (s.f.). Cada segundo de cada día se crean 1,7 MB de datos para cada persona. Así es como podemos controlarlos. https://www.unite.ai/es/cada-segundo-de-cada-d%C3%ADa-se-crean-1-7-MB-de-datos-para-cada-persona.-As%C3%AD-es-como-podemos-controlarlos./

WebsiteRating. (s.f.). Estadísticas y hechos sobre Internet. https://www.websiterating.com/es/blog/research/internet-statistics-facts/

Zeus Vision. (s.f.). Mejores prácticas de visualización de datos y análisis. https://zeus.vision/data-talk/mejores-practicas-de-visualizacion-de-datos-y-analisis/