Portada

2023

UNIVERSIDAD DE ALCALÁ
Escuela Politécnica Superior

Grado en Ingeniería Informática

Trabajo Fin de Grado

Plataforma basada en Python para transformación e integración de datos abiertos (open data) sobre servicios de análisis y visualización

Autor: Marcos Navarro Juan

Tutor/es: Manuel De Buenaga Rodríguez

TRIBUNAL:

Presidente: << Nombre y Apellidos >>

Vocal 1º: << Nombre y Apellidos >>

Vocal 2º: << Nombre y Apellidos >>

FECHA: << Fecha de depósito >>

Agradecimientos.

Son muchas las personas a las que tengo que agradecer que hoy sea mucho mejor persona que cuando empecé este camino, cuando empecé esta bonita travesía de estudiar el grado ingeniería informática.

En primer lugar, quiero dar las gracias a mis padres y a mis dos hermanos, que gracias a su apoyo he podido crecer tanto profesionalmente como persona.

En segundo lugar, ha sido de vital importancia rodearme de mi circulo de amistades, los cuales me han dado apoyo tanto en las mejores situaciones como en la peores. Especialmente a mis compañeros de la universidad, con los cuales hemos sacado multitud de proyectos y exámenes adelante.

Quiero destacar el curso que realice con la beca Erasmus, el año que pase en Varsovia, Polonia. El cual, fue fundamental para entender quién soy a día de hoy y esto es gracias a la Politechnika Warszawska y a las amistades inolvidables que forje en ese increíble viaje.

Por último, quiero agradecer a quienes han hecho posible este recorrido, agradecer a la propia universidad de Alcalá y a los diferentes profesores que me han ido acompañando en cada una de las asignaturas y por supuesto a mi tutor de TFG Manuel de Buenaga.

Resumen.

Abstract.

Resumen Extendido.

Tabla de contenido

Figuras

Tablas

Código.

Acrónimos y Abreviaturas.

Capítulo 1

1. Introducción.

En este primer capítulo del TFG, en primer lugar, se va a exponer diferentes argumentos del porque se ha realizado este proyecto, una serie de objetivos a cumplir, se definirán unas líneas de trabajo a seguir y finalmente se describirá brevemente la estructura del documento.

1.1. Motivación

La información, determinar qué información es la relevante y cual no lo es y exponerla de una forma adecuada en la que se pueda interpretar, tiene un increíble poder en nuestra sociedad. A lo largo de nuestra historia como seres humanos, la información siempre ha jugado un papel determinante y es en la actualidad donde más información recibidos y recabamos. Por ello, todo tipo de gobiernos, empresas, y más organismos invierten tanto tiempo y dinero en poder entender y descifrar la mayor cantidad de información posible y actuar en consecuencia.

En la actual, el papel de internet juega un papel vital. Gracias a internet y la era de las tecnologías en la que vivimos recabar esta información o datos es un proceso que se produce en cada momento que estamos conectados. No importa en que dispositivo o aplicación o en que lugar estemos conectados que permanente se está enviando información relevante.

Al visualizar la información que se considera de valor utilizando las diferentes fuentes de datos, provoca en una mejor toma de decisiones. Al tener una plataforma que facilite la transformación y la integración de datos abiertos, los diferentes agentes realizaran un análisis mas consistente de la situación, por tanto, aumentará la calidad de las decisiones, que esta soportadas por evidencias solidas.

Los datos abiertos contienen una gran cantidad de información de gran valor que ayuda a comprender mejor diversos fenómenos, esta información esta presentada por diferentes organismos públicos o privados que garantizan la consistencia del dato. El acceso a estos datos abiertos es una manera libre, por lo que, provoca un impulso en la innovación y una colaboración ciudadana.

El concepto de Open Data es un concepto que nació en el siglo XXI el cual tiene una tendencia evolutiva exponencial, donde son cada vez más los organismos públicos o privados que ponen a disposición de aquel que lo desee información sobre infinidad de cuestiones.

La utilización del lenguaje Python y su amplia gama de bibliotecas que se van incrementando de forma exponencial supone una eficiencia en el procesamiento de datos y en el análisis de datos. Utilizando e investigando las bibliotecas de Python puede permitir la automatización de tareas tediosas y repetitivas, ahorrando tiempo y recursos en el manejo de grandes volúmenes de datos

Debido a este conjunto de ideas y de la importancia que siempre ha tenido y mas en la actualidad el extraer información de valor sobre un conjunto de datos nace la motivación para la creación de este proyecto. Además, el proyecto en cuestión afectaría a infinidad de campos, donde

algunos de estos suponen de una importancia notable para la sociedad. Por ejemplo, se puede destacar el campo de la medicina, el de la salud y medioambiente, campos financieros, deportivos.

1.2. Objetivos

El objetivo del proyecto es trabajar sobre un conjunto de fuentes de datos que cumplan la característica de ser open data y que pertenezcan a diferentes campos de la sociedad actual. Algunas de estas fuentes de datos estarán relacionadas con proyectos de investigación en los que ha participado el departamento (Eugloh o Smacite). Se aplicarán nuevas funcionalidades a través de Python para la extracción, transformación y carga de los datos (ETL) y para el análisis a los diferentes conjuntos de datos. El resultado de este análisis se presentará de la manera más visual apoyándose en el software de Microsoft PowerBi. Este objetivo principal está compuesto por subobjetivos más concretos. Estos subjetivos son los siguientes:

- La utilización de fuentes de datos abiertas (Open Data), cuyos dominios están relacionados con las ciudades inteligentes (Smacite) y con la salud y el medioambiente. Las administraciones públicas proporcionan una serie de bases de datos abiertas de múltiples campos de información. Algunos de estos campos pueden ser de información relacionada con accidentes de tráfico, la distribución de mobiliario urbano, valores climatológicos, … Donde cualquier tipo de agente individual u organismo puede acceder a esta información y realizar proyectos que fomentan la innovación y el desarrollo.

- Relación al proyecto internacional “Eugloh”, realizado por universidades europeas, incluida la Universidad de Alcalá de Henares. Eugloh es un proyecto relacionado con la salud global y Smart Cities.

- Realización de procesos de extracción, transformación y carga de datos a través de Python para las fuentes de datos abiertas seleccionadas, para que el dato sea consistente y heterogéneo a la hora de aplicar las diferentes técnicas de análisis. Donde incluso las fuentes de datos utilizadas tengan diferente formato, pero sean validas a la hora de integrarse en la plataforma.

- La implementación de técnicas avanzadas basadas en analítica de datos, desarrolladas en Python, sobre las bases de datos abiertas anteriormente descritas y su integración en PowerBi. A partir de las soluciones obtenidas en las diferentes técnicas de análisis se aporta una valoración de los resultados que aportan funcionalidad.

- Visualizar la información recabada tras la implementación de técnicas avanzadas basadas en analítica de datos, de una forma intuitiva y compresible para el usuario. Tambien se visualizará las soluciones obtenidas y la interpretación de valor de cada una de ellas.

1.3. Líneas de Trabajo.

Las líneas de trabajo se componen de 7 etapas, donde pueden cada etapa se desarrolla en un espacio de tiempo determinado.

1. Definición de objetivos y alcance del proyecto. Se definen los objetivos que se quieren alcanzar y se determina si son viables o no, en función del tiempo y de los recursos. También se define el alcance del proyecto teniendo en cuenta las funcionalidades especificas a desarrollar.

2. Investigación y selección de fuentes de datos abiertos. Se investigan y se seleccionan las fuentes de datos abiertas a utilizar en el proyecto. La investigación se realiza en diferentes paginas gubernamentales donde se exponen las diferentes bases de datos y sus características. Se seleccionan las fuentes de datos en función de la temática (en relación con "Eulogh", ciudades inteligentes, salud y medioambiente …) y teniendo en cuenta la cantidad de información y la consistencia de los datos.

3. Implementación de la extracción y transformación de datos. Se realizan los diferentes programas a través de Python para realizar las operaciones ETL para cada una de las fuentes de datos.

4. Implementación de las técnicas avanzadas basadas en analítica de datos. Se realizan las diferentes técnicas de analítica de datos para las diferentes fuentes de datos seleccionadas. Dependiendo de los objetivos esperados de cada fuente de datos se realiza una técnica de u otra. Se realiza el análisis de datos utilizando Python y diferentes bibliotecas que provee el lenguaje.

5. Desarrollo de las herramientas de visualización en PowerBi. En función de los resultados obtenidos en el paso anterior y de las fuentes de datos se utiliza el software PowerBi para visualizar diferentes aspectos relevantes al proyecto.

6. Pruebas y evaluación de la plataforma. A lo largo de todo el proyecto se realizan las pruebas y las evaluaciones correspondientes a cada etapa del proyecto.

7. Documentación. A medida que se va a realizando cada una de las etapas, se va explicando en el documento los pasos realizados y a aportación teórica necesaria para comprender de una forma sencilla el funcionamiento del proyecto.

8. Despliegue. Una vez realizada cada una de las etapas se prepara la plataforma para su despliegue y su complemento funcionamiento.

La línea de trabajo tiene el siguiente flujo de estados:

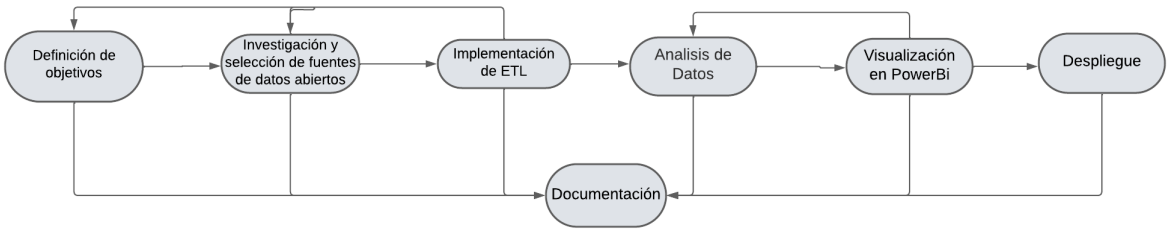

Figura 1. Flujo de Línea de Trabajo 1.

Como se puede observar puede ser necesario complementar una etapa para poder avanzar a la siguiente, también durante el proyecto dependiendo de la etapa se puede volver a la etapa anterior. En cada una de las etapas se realiza de forma paralela la documentación del proyecto.

1.4. Estructura de la Memoria.

La estructura de la memoria está compuesta por capítulos, la estructura es la siguiente:

- Capítulo 1. Se realiza una breve introducción del proyecto, se explica las motivaciones que se han producido para llevar a cabo el proyecto, se define un flujo de trabajo y se explica como va a ser la estructura de la memoria.

- Capítulo 2. Se explican teóricamente los conceptos necesarios para entender el proyecto. Los conceptos que destacan son los de: Open Data, ETL, Análisis de datos, Python y Visualización con PowerBy.

- Capítulo 3. Se explican las herramientas que se han llevado a cabo para realizar el proyecto y el porqué de la utilización de cada una.

- Capítulo 4. Se expone la implementación llevada a cabo. Se el porque cada una de las fuentes de datos escogidas y el porqué de la elección, se explica todo lo correspondiente con el proceso de ETL, todo el proceso del análisis de datos y de la visualización de resultados y la estructura de la plataforma.

- Capítulo 5. Se exponen todos los resultados obtenidos y se explica detalladamente que significado y que información se pueden obtener.

- Capitulo 6. Se comentan las conclusiones obtenidas al finalizar el proyecto, la posible escalabilidad del mismo y temas de consideración como posibles aplicaciones, mejoras y análisis.

- Capitulo 7. Se indican todas las referencias utilizadas para la realización del proyecto.

Capítulo 2

2. Estado del Arte.

A continuación, se explican los conceptos de mayor relevancia del proyecto y las razones por las que estos conceptos son utilizados en el proyecto. Para la compresión del proyecto es necesario tener claro una serie de conceptos, entender cómo funcionan y cuáles son sus puntos fuertes y débiles. Por ello, en este punto se presentan los siguientes conceptos desde una perspectiva teórica y se explican las diferentes ventajas por las que se ha decidido que forman parte del proyecto.

2.1. Open Data

El concepto de "Open Data" se basa en la premisa proporcionar al público ciertos conjuntos de dato de una manera libre y sin restricciones excesivas. Existen diferentes organizaciones que generan o poseen datos que ponen a disposición del publico en general.

El objetivo del concepto es de promover la transferencia y la apertura de la información, haciéndola accesible a todo el que lo desee. Al liberar la información, se busca fomentar la participación ciudadana para colaborar y desarrollar soluciones innovadoras en todo tipo de ámbitos. Tambien puede provocar un aumento de la calidad de la información presentada, ya que, al permitir que los ciudadanos y investigadores analicen y examinen la información pueden detectar errores o mejorías.

Las fuentes de datos que se utilizan en este proyecto son todas de fuentes de datos abierto, ya que, nos brindan las siguientes ventajas en función a los objetivos definidos del proyecto:

- Libre acceso sin costo alguno. Debido a que la característica principal de las fuentes de datos abiertas es que tienen un acceso libre y sin restricciones provoca que el uso de las bases de datos de diferentes ámbitos no tenga coste alguno. Lo cual, es una ventaja significativa a tener en cuenta en todos tipos de proyecto y más cuando son de índole académica.

- Acceso a fuentes de datos de diversos indoles. Al utilizar open data, se tiene un acceso a una amplia variedad de bases de datos de diferentes dominios. Organizaciones como puede ser el gobierno de España o la unión europea ponen a disposición multitud de fuentes de datos abiertas de diferentes categorías como demografía, empleo, medio ambiente, economía, …

- Calidad y confiabilidad de los datos. Como detrás de estas fuentes de datos abiertas se encuentran instituciones de reconocidas provoca un mayor grado de confiabilidad y calidad en comparación con otras fuentes de datos no verificadas. Lo que permite trabajar sabiendo que la información con la se trabaja es precisa y confiable, lo cual, es un aspecto fundamental para el análisis y visualizaciones confiables.

- Innovación. Al utilizar open data se fomenta la innovación y el desarrollo de soluciones creativas. Gracias estos conjuntos de datos se pueden descubrir nuevas conclusiones, patrones y tendencias que pueden impulsar la creación de herramientas analíticas y visuales únicas.

- Enfoque en problemas sociales. El contenido de las fuentes de datos abiertas se corresponde con temas sociales, por lo que, se pueden abordar problemas sociales o deficiencias del

sistema y proporcionarle una solución de calidad que tenga un impacto real en la calidad de vida.

Cabe indicar que España se encuentra entre los lideres del fomento del Open Data a nivel europeo. Todos los gobiernos de los países pertenecientes a la unión europea tienen un sistema donde proporcionan fuentes de datos abiertas que abarcan todos los ámbitos de la vida social. El "Open Data Maturity Report" recoge el desempeño de los países europeos en términos de aportaciones y lo califica con un sistema de puntación de cuatro dimensiones (política, social, impacto y calidad). En los últimos años, los países miembros de la unión europea han mejorado notablemente sus aportaciones, por tanto, sus puntuaciones, en el 2015 la puntuación media era del 44% y en el 2022 se ha incrementado hasta el 81%. En la figura 2 se puede observar las puntuaciones de cada país miembro de la unión europea en función de la política, social, impacto y calidad.

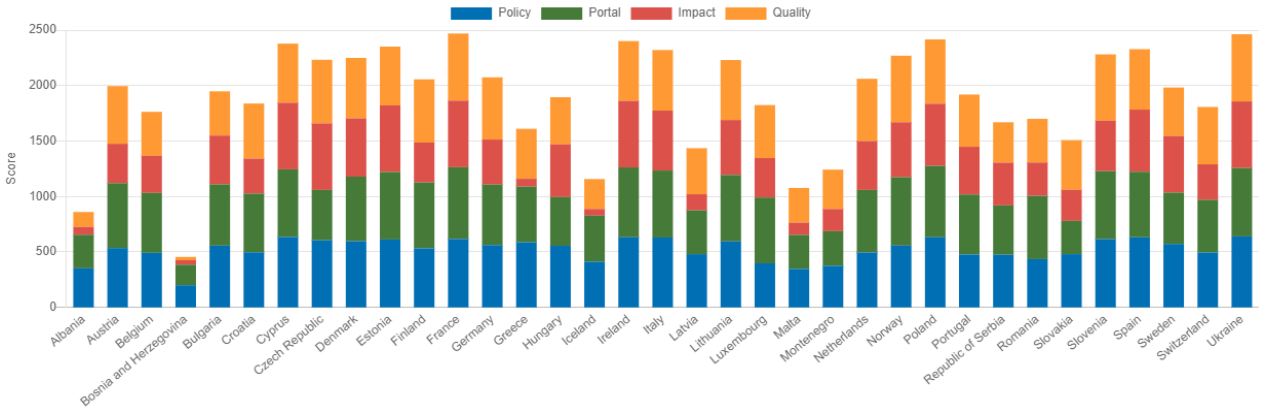

Figura 2. Tabla de puntaciones de los países miembros en el año 2022.

España se encuentra en la tercera posición con mas aportaciones y se encuentra por encima de la media europea como se puede observar en la figura 2.

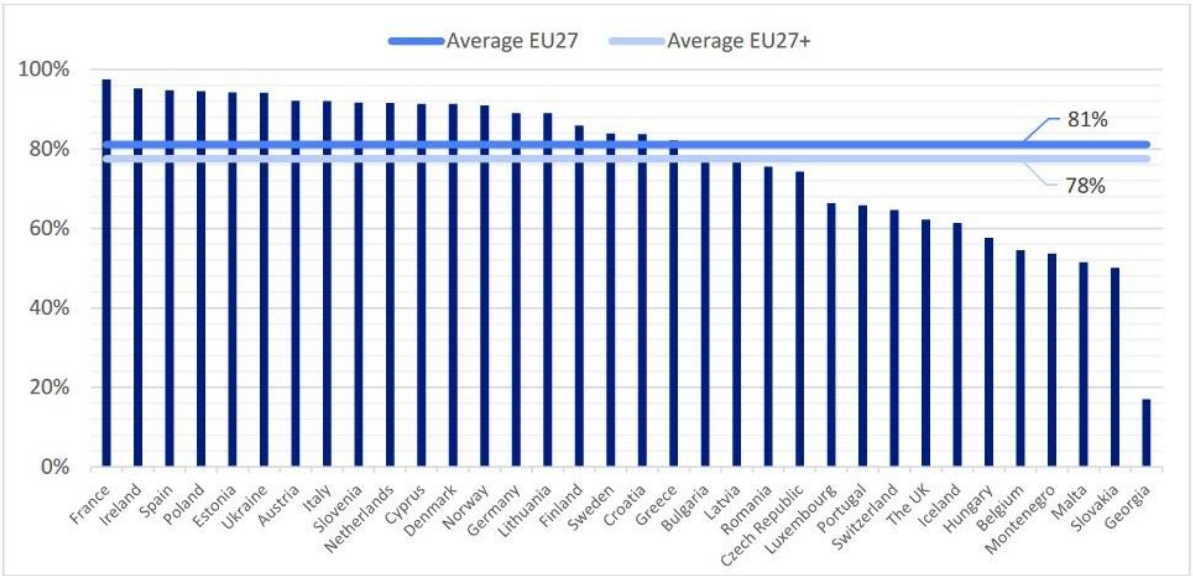

Figura 3. Tabla de puntaciones de la aportación de fuentes de datos abiertas a nivel europeo.

Como se puede observar la utilización de fuentes de datos abiertas es una practica que ha ido aumentado de forma notable y para el futuro se prevé que evolucione siguiendo la misma tendencia exponencial.

2.2. Análisis de datos

El análisis de datos es el proceso por el cual se examinan, se limpian, se transforman e interpretan una serie de conjuntos de datos. Tiene como objetivo encontrar patrones, relaciones para extraer conocimientos que aporten valor para tomar decisiones con información relevante.

Este concepto es fundamental a la hora de tomar decisiones relevantes. Ya que un correcto análisis de datos otorga cierta información de valor, como pueden ser tendencias, patrones o relaciones, que ayudan a comprender situaciones, evaluar opciones y respaldar la toma de decisiones. La información obtenida puede servir para optimizar procesos, al analizar los datos de diferentes procesos, sirve para identificar ineficiencias, cuellos de botella o áreas de mejora, lo cual puede aumentar la eficiencia y reducir costos.

También la información obtenida a partir del análisis de datos sirve para identificar oportunidad y riesgos. Se pueden detectar oportunidades de negocio, nichos de mercado o tendencias emergentes, pero a su vez, también se puede detectar riegos potenciales, como anomalías o desviaciones significativas en los datos.

Existen diferentes tipos de análisis de datos que se pueden realizar sobre un conjunto de datos, cada uno con enfoques y técnicas diferentes para conseguir una serie de objetivos. A continuación, se describen los más significativos:

- Análisis Descriptivo. Consiste en resumir y describir los datos utilizando medidas de estadística básica como pueden ser promedios, medianas, desviaciones. Tiene como objetivo proporcionar una compresión inicial de los datos.

- Análisis Exploratorio. Consiste en descubrir patrones y relaciones ocultas en los datos a partir de técnicas gráficas y procesos estadísticos más avanzados como pueden ser análisis de regresión, análisis de conglomerados (clustering) y análisis de componentes principales (PCA). Tiene como objetivo ayudar a generar una hipótesis sobre los datos y orientar investigaciones futuras.

- Análisis Predictivo. Consiste en identificar patrones históricos y tendencias para posteriormente realizar una predicción que se asemeje con la realidad, toda predicción tiene un error característico. Utiliza técnicas estadísticas y algoritmos de aprendizaje automático (Machine Learning) para construir modelos predictivos.
  Cabe destacar que cuantos más datos se estén procesando y mejor sea el algoritmo o el proceso con el que se esté trabajando, más patrones históricos y tendencias serán identificadas, por lo que, significa que se obtendrán predicciones más reales y con menos error. Este tipo de análisis de datos es fundamental para ámbitos como la salud, finanzas, marketing o logística.

- Análisis Prescriptivo. Tiene como objetivo proporcionar recomendaciones y soluciones optimas basadas en datos a través de combinar técnicas de análisis predictivo con optimización matemática con restricciones y objetivo específicos. Se utilizar para argumentar toma de decisiones y optimización de recursos.

- Análisis de texto y Minería de Datos: Se aplica a conjunto de datos no estructurados, como pueden ser textos, reseñas, comentarios en redes sociales. Combina las técnicas de procesamiento del leguaje natural (NLP) y la extracción de grandes conjuntos de datos. Tiene como objetivo entender patrones de comportamientos, de pensamiento o opiniones.

En el proyecto se realizan diferentes tipos de análisis de datos, desde el que nos proporcionar soluciones más simples hasta el que nos muestra resultados más eficientes, con el objetivo de abarcar

un marco amplio de resultados. El tipo de análisis que predomina en el proyecto es el de análisis predictivo y la utilización de algoritmos de aprendizaje automático o Machine Learning.

El proyecto estará compuesto de tres modelos estadísticos, donde cada uno está enfocado a una fuente de datos abierta diferentes. Dependiendo de cómo este compuesta la fuente de dato abierta y de los objetivos sobre el análisis se utiliza una técnica de análisis u otra.

Los tipos de análisis realizados son, el análisis descriptivo, el análisis exploratorio donde destaca el análisis de regresión y por último el análisis predictivo con diferentes técnicas de Machine Learning.

Cabe destacar que los procesos de extracción, transformación y carga (ETL) como los procesos de extracción, carga y transformación (ELT) forman parte del concepto de análisis de datos. El proceso ETL tiene una gran importancia en el proyecto.

2.3 Proceso ETL.

El proceso ETL (Extract, Transform y Load) es una metodología utilizada en el análisis de datos para la integración y la preparación de datos de diferentes fuentes de datos o data warehouse. Estos procesos se utilizan para garantizar la calidad, coherencia y disponibilidad de los datos antes de su análisis.

El proceso de ETL está compuesto por diferentes fases, a continuación, se explica cada una ellas:

1. Extracción. En esta fase se extraen los datos de las fuentes de datos, que en el caso del proyecto serán fuentes de datos abiertas. Dependiendo de cómo este presentada la fuente de datos se utiliza unas técnicas de extracción u otras, por ejemplo, la extracción se puede realizar en fuentes de datos SQL, o de ficheros Excel o incluso de documentos PDF.
   Se utiliza Python para la realización de las técnicas de extracción, ya que, Python ofrece infinidad de bibliotecas y herramientas que facilitan la extracción de datos, como puede ser Pandas o Beautiful Shop.

2. Transformación. Una vez que ya se han extraído todos los datos, los datos se someten a una serie de procesos de limpieza, normalización, enriquecimiento y consolidación para garantizar su calidad y coherencia, que será fundamental a la hora de realizar el análisis de datos y la visualización de la información pertinente. Las operaciones mas comunes que se suelen realizar en esta fase son operaciones como la eliminación de duplicados, la conversión de formatos, la agregación de datos y la creación de nuevas variables. Para realizar la transformación también se utiliza Python, debido a sus amplias bibliotecas que facilitan la transformación de la información, algunas de estas son Pandas, NumPy o SciPy.

3. Carga. Una vez que ya se ha producido la fase de extracción y la de transformación se vuelven a cargar estos datos en data warehouse. Un data warehouse es un sistema de almacenamiento de datos que esta creado para un posterior análisis y generación de informes.
   El objetivo de esta fase es cargar los datos ya transformados en una estructura de almacenamiento diseñada específicamente para su análisis, para su visualización o su acceso.
   A través de Python se producirá a la carga de datos en los diferentes sistemas data warehouse.

A continuación, se presenta un esquema del funcionamiento global de un proceso ETL:

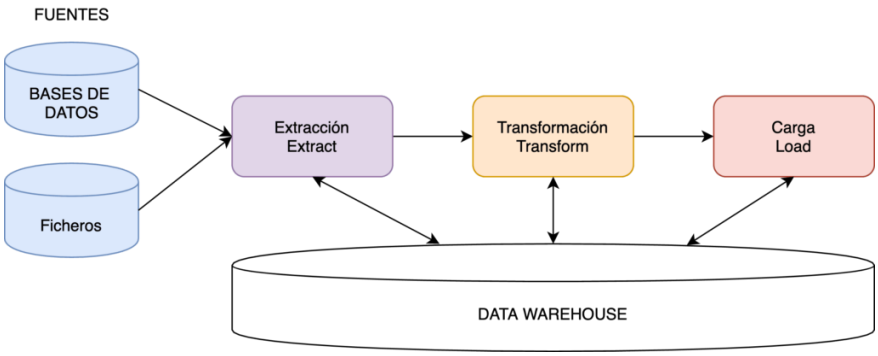

Figura 4. Esquema del funcionamiento de un proceso ETL.

2.4 Python.

Los de procesos de ETL de las diferentes fuentes de datos abiertas, tanto los diferentes tipos de análisis aplicados a las fuentes de datos se realiza a través del lenguaje de programación Python.

Python es un lenguaje de programación de alto nivel y de propósito general. Se caracteriza por ofrecer una amplia gama de bibliotecas y herramientas que hacen más sencillo multitud de procesos. Relacionados con el proyecto, los procesos de manipulación, procesamiento y análisis de datos.

Python destaca por su evolución en los últimos años, ya que, ha experimentado un crecimiento significativo comparado con los otros lenguajes de programación. La utilización de Python frente a otros lenguajes se debe a las ventajas que propone.

En la figura cinco se presenta la evolución en función del tiempo, desde antes del año 2006 hasta el 2022, de los lenguajes de programación mas utilizados globalmente. La información esta extraída de PYPL, donde realizan la gráfica analizando la frecuencia con la que se buscan tutoriales sobre los lenguajes de programación en Google. Como se observa Python se convierte en la más utilizada actualmente con un incide de búsquedas del 27,43% según PYPL.

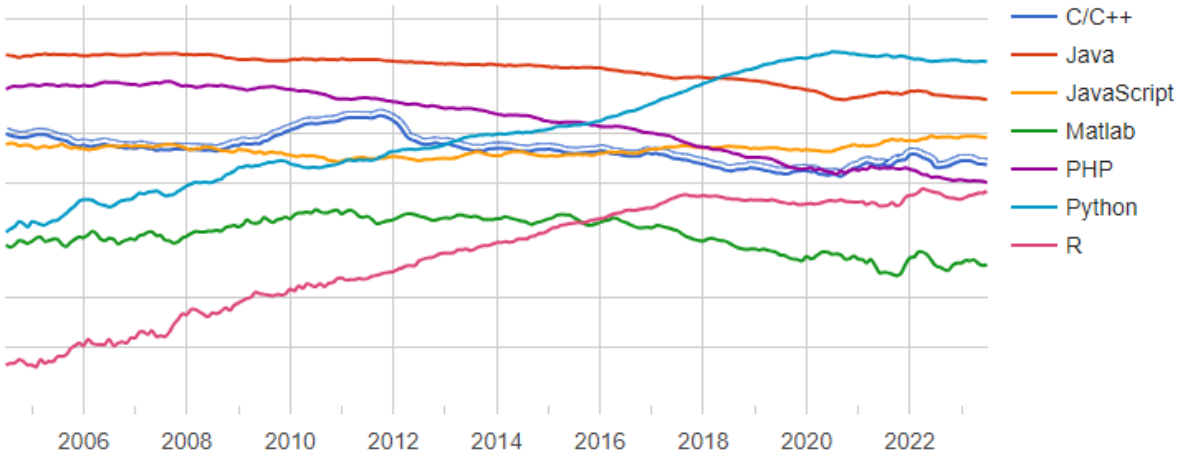

Figura 5. Evolución de los lenguajes de Programación en función del tiempo.

Se escoge este lenguaje de programación y no otro, debido a las ventajas notables que proporciona para un proyecto de este tipo, las cuales son las siguientes:

- Facilidad de uso. Python desataca por tener una síntesis clara y legible, lo cual, a la hora de manipular y transformar los diferentes datos es fundamental tener un código fácilmente comprensible.

- Amplio Ecosistema de Bibliotecas. Python tiene un amplio ecosistema de bibliotecas, que no para de incrementar. Estas bibliotecas son fundamentales a la hora de proporcionar herramientas para realizar tareas de extracción, transformación y carga (ETL) y las diferentes técnicas de análisis de datos.

- Compatibilidad con las fuentes de datos abiertos. Python puede trabajar con diferentes formatos de datos, lo cual es fundamental para el manejo de datos abiertos. No importa que los datos abiertos se encuentren almacenados en archivos de diferentes tipos como pueden ser CSV, JSON, XML que Python ofrece bibliotecas y herramientas que facilitan el proceso de extracción.

- Integración de ostras tecnologías. Python se integra de forma sencilla con otras tecnologías y software. Se puede combinar con bases de datos, servicios web y otras herramientas. En concreto, es compatible con el software PowerBi, por lo que, se puede realizar la visualización de la información de una forma concisa.

- Soporte de la comunidad. Cuenta con una comunidad inmensa de documentación y de desarrolladores, por lo que, facilita el aprendizaje y desarrollo de la plataforma.

2.5 Visualización en PowerBi.

PowerBi es una herramienta de Business Intelligence desarrollada por Microsoft, la cual permite transformar y visualizar datos en informes interactivos o en paneles de control dinámicos.

Se utiliza el software PowerBi para llevar a cabo la visualización de los resultados obtenidos a través de Python y las fuentes de datos abiertas utilizadas.

Capítulo 3

3. Herramientas a Utilizar.

Para lograr los objetivos del proyecto es necesario la utilización de diferentes tecnologías y herramientas a lo largo del proceso. Cada herramienta está relacionada con las diferentes fases del proyecto. Las fases del proyecto se pueden visualizar en la figura 1.

Para la fase de ETL de las diferentes fuentes de datos y posterior análisis de datos se utiliza Python, la versión 3.7, como lenguaje de programación y Visual Studio Code como entorno de desarrollo. Durante el análisis de datos se van utilizando las siguientes bibliotecas:

-

Para el almacenamiento de datos durante el análisis de datos se utiliza un almacén de datos (Data Warehouse) utilizando la aplicación Excel desarrollada por Microsoft y que forma parte de la suite de Microsoft Office.

Para la fase de visualización se utiliza el software PowerBi.

Para la fase de documentación, se utiliza la aplicación de procesamiento de texto Microsoft Word, que forma parte de la suite de Microsoft Office.

Se utiliza el software en línea Lucidchart para la realización de diferentes tipos de diagramas, que posteriormente son implementados en Microsoft Word como figuras.

Para alojar, gestionar el proyecto se utiliza la plataforma en la nube GitHub, que es una plataforma de desarrollo colaborativa para desarrolladores que permite alojar, gestionar y colaborar con proyectos almacenados y sus posibles versiones del proyecto. El proyecto esta almacenado en un repositorio donde se van gestionando las posibles versiones y las actualizaciones, el repositorio es el siguiente:

Capítulo 4

4. Implementación.

Capítulo 7

7. Bibliografía.

1. The official portal for European Data, “Open Data Maturity”, data.europa.eu. Available: https://data.europa.eu/en/publications/open-data-maturity/2022
2. Aprende Big Data, “ETL vs ELT”, aprendebigdata. Available: https://aprenderbigdata.com/etl-vs-elt/
3. Real Python, “Beautiful Soup: Build a Web Scraper with Python” realpython. Available: https://realpython.com/beautiful-soup-web-scraper-python/
4. Pypi “Beautiful Soup4 4.12.2” pypi.org. Available: https://pypi.org/project/beautifulsoup4/
5. Pypi “Pandas 2.0.3” pypi.org. Available: https://pypi.org/project/pandas/
6. Pandas Documentation, “Guide Pandas”, pandas.pydata.org. Available: https://pandas.pydata.org/docs/user_guide/index.html#user-guide
7. A
8. A
9. A
10. A
11. Características de Python, “¿Python el lenguaje del futuro?”, paradigma.digital. Available: https://www.paradigmadigital.com/dev/es-python-el-lenguaje-del-futuro/
12. evolución de los lenguajes de Programación, “Popularity of Programming Lenguage”, pypl.github.io. Available: https://pypl.github.io/PYPL.html

El documento debe igualmente incluir los siguientes apartados obligatorios:

- Índice.
- Resumen en español del trabajo en un máximo de cien (100) palabras.
- Resumen en inglés del trabajo en un máximo de cien (100) palabras.
- Palabras clave con un máximo de cinco.
- Introducción en la que se indique el planteamiento del trabajo y los objetivos a conseguir.

- Descripción del trabajo desarrollado, estructurado como proceda según su tipo, y con los contenidos que se consideren publicables si hay un acuerdo de confidencialidad.
- Conclusiones.
- Bibliografía, que incluirá el conjunto de referencias. Se recomienda el estilo de citación del IEEE.

Se recomienda que la estructura se complete de acuerdo con el siguiente esquema:

- Índice.
- Resumen en español del trabajo en un máximo de cien (100) palabras.
- Resumen en inglés del trabajo en un máximo de cien (100) palabras.
- Resumen extendido del trabajo en un máximo de 4 páginas.
- Glosario de acrónimos y abreviaturas.
- Introducción en la que se indique el planteamiento del trabajo y los objetivos a conseguir.
- Descripción del trabajo desarrollado estructurado con el siguiente esquema:
    - Base teórica en la que se expongan los conceptos teóricos utilizados para la realización del trabajo, así como los cálculos realizados.
    - Descripción experimental, cuando sea necesario, descripción del diseño, resultados, etc.
- Conclusiones y, en su caso, trabajo futuro.
- Bibliografía, que incluirá el conjunto de referencias. Se recomienda el estilo de citación del IEEE.
- Anexos/apéndices:
    - Planos y diagramas, entendiendo por tales los generales, diagramas de bloques, esquemas de detalle y planos, ajustados a la normativa existente sobre el análisis y el diseño de sistemas hardware y software. En caso de restricciones de confidencialidad se excluirá este apartado.
    - Pliego de condiciones (en su caso), que incluya las condiciones generales (normativas), condiciones de materiales y de equipos (características técnicas) y condiciones de ejecución. En caso de restricciones de confidencialidad se excluirá este apartado.
    - Presupuesto, que incluya: ejecución material (materiales y mano de obra), gastos generales y beneficio industrial, honorarios de dirección y redacción (tarifas del Colegio, en su caso), coste de ejecución por contrata y presupuesto total. En caso de confidencialidad se excluirá este apartado.
    - Manual de usuario y/o de instalación y/o de mantenimiento (en su caso), en todos aquellos equipos o programas generados en el trabajo y que vayan a utilizarse posteriormente.