



Hito programación

DAVID MILLÁN, IVÁN ENCINAS Y PAULA CUBERO

CAMPUSFP | Programación

Índice

FASE 1 (Iván y Paula)	2
1. Hablamos de fuentes de datos. De grandes volúmenes de datos. Por ejemplo, data Lake o similar. También es importante tratar la diferencia entre datos estructurados y no estructurados en relación al bigData (Paula)	2
2. Entre las herramientas más interesantes a la hora de gestionar grandes volúmenes de datos nos encontramos con Hadoop y Spark. Habría que tratar sus características y finalidad. (Paula)	5
3. Existen lenguajes de programación “recomendables” para gestionar datos. Entre ellos, están Python y Scala. Sería explicar brevemente por qué. (Iván)	6
PYTHON	6
SCALA	6
4. En la parte de visualización de datos, de mostrar dashboards nos encontramos con PowerBI y Tableau entre otros. Debemos explicar qué son. (Iván)	7
Power BI	7
Tableau	7
FASE 2. Implementación de código (David)	8
Histograma	14
Gráfico de línea	15
FASE 3. Evaluación fase 2 y evolución (David)	16
Evaluación fase 2	16
Evolución	16
Webgrafía	18

FASE 1 (Iván y Paula)

1. Hablamos de fuentes de datos. De grandes volúmenes de datos. Por ejemplo, data Lake o similar. También es importante tratar la diferencia entre datos estructurados y no estructurados en relación al bigData (Paula)

Las **fuentes de datos** son el lugar donde se originan los datos utilizados. Puede ser el lugar donde se crearon los datos o donde se digitalizaron la información física.

Un **data lake** es un repositorio centralizado diseñado para almacenar, procesar y proteger grandes cantidades de datos estructurados, semiestructurados o no estructurados. Puede almacenar datos en su formato nativo y procesar cualquier variedad de datos, ignorando los límites de tamaño.

Un data lake proporciona una plataforma escalable y segura que permite a las empresas realizar las siguientes tareas: transferir cualquier dato desde cualquier sistema y a cualquier velocidad (incluso si los datos provienen de sistemas que son locales, de la nube o de procesamiento perimetral); almacenar cualquier tipo o volumen de datos con fidelidad absoluta; procesar datos en tiempo real o en modo por lotes; y analizar datos mediante SQL, Python, R o cualquier otro lenguaje, datos de terceros o aplicaciones de estadísticas.

A cada elemento de un data lake se le asigna un identificador único y se etiqueta con un conjunto de etiquetas de metadatos extendidas. Cuando se presenta una cuestión de negocios que debe ser resuelta, podemos solicitarle al data lake los datos que estén relacionados con esa cuestión. Una vez obtenidos podemos analizar ese conjunto de datos más pequeño para ayudar a obtener una respuesta.

El principal beneficio de un data lake es la centralización de fuentes de contenido dispares. Una vez reunidas (de sus "silos de información"), estas fuentes pueden ser combinadas y procesadas utilizando big data, búsquedas y análisis que de otro modo hubieran sido imposibles.

El data lake se asocia a menudo con el almacenamiento de objetos orientado a **Hadoop**. En este escenario, los datos de una organización se cargan primero en la plataforma Hadoop y, a continuación, se aplican las herramientas de análisis y de minería de datos a los datos que residen en los nodos clúster de Hadoop.

*Hadoop -> es un entorno de trabajo para software, bajo licencia libre, para programar aplicaciones distribuidas que manejen grandes volúmenes de datos.

*Clúster -> es un conjunto de nodos que trabajan de forma coordinada para almacenar la información y/o realizar el procesamiento.

Otra fuente que gestiona un gran volumen de datos es un **Data warehouse**, es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa u organización. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso. Los datos de un data

warehouse deben almacenarse de forma segura, fiable, fácil de recuperar y fácil de administrar.

Data Warehouse es una arquitectura de almacenamiento de datos que permite a los ejecutivos de negocios organizar, comprender y utilizar sus datos para tomar decisiones estratégicas.

Podemos diferenciar 3 categorías principales de Data Warehouses:

- Data Warehouse de Empresas (EDW), son depósitos de datos centralizados que permiten orientar las decisiones de la empresa. Los datos son organizados y presentados de manera uniforme. Los EDW también permiten clasificar los datos en según su tema.
- Data Stores Operacionales (ODS), los datos se actualizan en tiempo real, lo que los hace muy útiles para actividades cotidianas como el registro de informes y de empleados.
- Data Mart que es una subcategoría de Data Warehouse. Está concebida para empresas de sectores de la venta o las finanzas. Los datos pueden ser recolectados desde diversas fuentes.

Diferencia entre datos estructurados y no estructurados en relación al bigData

Los **datos estructurados** son datos que están en un formato estandarizado, tienen una estructura bien definida, cumplen con un modelo de datos, siguen un orden persistente y son de fácil acceso para humanos y programas. Este tipo de datos generalmente se almacena en una base de datos.

Los datos estructurados están muy organizados y se comprenden fácilmente mediante el lenguaje de máquina. Mediante las bases de datos relacionales se puede ingresar, buscar y manipular datos estructurados con relativa rapidez. Esta es la característica más atractiva de los datos estructurados.

Se ha creado un formato de etiquetado o marcado, dentro del propio lenguaje HTML de las páginas, que permite identificar y describir explícitamente diferentes tipos de información: marcado de datos estructurados que Google utiliza para las búsquedas.

Los **datos no estructurados** son un conjunto de datos que no se almacenan en un formato de base de datos estructurado. Los datos no tienen un formato u organización predefinidos, lo que hace que sea mucho más difícil de recopilar, procesar y analizar.

Los ejemplos de datos no estructurados incluyen texto, vídeo, audio, actividad móvil, actividad en redes sociales, imágenes satelitales, imágenes de vigilancia...

Los datos no estructurados son difíciles de deconstruir porque no tienen un modelo predefinido, lo que significa que las bases de datos no relacionales o NoSQL son las más adecuadas para administrar datos no estructurados.

*Bases de datos no relacionales -> no trabajan con estructuras definidas. Es decir, los datos no se almacenan en tablas, y la información tampoco se organiza en registros o campos.

Tienen una gran escalabilidad y están pensadas para la gestión de grandes volúmenes de datos.

Otra forma de administrar datos no estructurados es hacer que fluyan a un lago de datos, lo que les permite estar en su formato sin formato y no estructurado.

Más del 80% de todos los datos generados en la actualidad se consideran no estructurados, y este número seguirá aumentando con la prominencia del Internet de las cosas.

Las técnicas de minería de datos aplicadas a datos no estructurados pueden ayudar a las empresas a aprender hábitos de compra, patrones en las compras, sentimiento hacia un producto específico y mucho más.

2. Entre las herramientas más interesantes a la hora de gestionar grandes volúmenes de datos nos encontramos con Hadoop y Spark. Habría que tratar sus características y finalidad. (Paula)

De la herramienta **Hadoop** ya he hablado ligeramente en el apartado anterior pero ahora voy a tratar sus características y finalidad.

Hadoop es una estructura de software de código abierto para almacenar datos y ejecutar aplicaciones en clústeres de hardware comercial. Proporciona almacenamiento masivo para cualquier tipo de datos, enorme poder de procesamiento y la capacidad de procesar tareas o trabajos concurrentes virtualmente ilimitados.

Características:

- Capacidad de almacenar y procesar enormes cantidades de cualquier tipo de datos, al instante.
- Poder de cómputo. El modelo de cómputo distribuido de Hadoop procesa big data a gran velocidad.
- Tolerancia a fallos. El procesamiento de datos y aplicaciones está protegido contra fallos del hardware. Si falla un nodo, los trabajos son redirigidos automáticamente a otros nodos para asegurarse de que no falle el procesamiento distribuido. Se almacenan múltiples copias de todos los datos de manera automática.
- Flexibilidad. No tiene que procesar previamente los datos antes de almacenarlos. Puede almacenar tantos datos como desee y decidir cómo utilizarlos más tarde.
- Bajo costo. La estructura de código abierto es gratuita y emplea hardware comercial para almacenar grandes cantidades de datos.
- Escalabilidad. Puede hacer crecer fácilmente su sistema para que procese más datos con sólo agregar nodos.

Las empresas buscan que Hadoop sea su próxima gran plataforma de datos. Sus usos más populares de hoy en día son:

- Almacenamiento y archivo de datos de bajo coste.
- Entorno de pruebas para descubrimiento y análisis. Ofrece una oportunidad para innovar con una inversión mínima.
- Data lake. Los data lake permiten almacenar datos en su formato original o exacto, tanto estructurados como sin estructurar, y sin ningún tipo de procesamiento, con el objetivo de ofrecer una visión sin modificar o sin refinar de los datos a los analistas de datos para que puedan utilizarlos para descubrir y analizar. Ayudando a hacer preguntas nuevas o difíciles sin restricciones.

En cuanto a **Spark** es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos. Spark se puede ejecutar de forma independiente o en Apache Hadoop, Apache Mesos, Kubernetes, la nube y distintas fuentes de datos.

Apache Spark está especialmente diseñado para su implementación en big data y machine learning. Pues su potencia de procesamiento agiliza la detección de patrones en los datos, la clasificación organizada de la información, la ejecución de cómputo intensivo sobre los datos y el procesamiento paralelo en clústers.

Esta herramienta cuenta con la comunidad de código abierto más grande que existe a nivel mundial en cuanto a big data.

Existen 4 componentes que integran y potencian lo que es Spark. Ellos son:

- Spark SQL: permite acceder a los datos de manera estructurada. También facilita la integración de Spark con Hive, ODBC, JDBC y herramientas de business intelligence.
- Spark Streaming: brinda soporte para el procesamiento de datos en tiempo real. Esto mediante un sistema de empaquetamiento de pequeños lotes.
- MLlib – Machine Learning Library: ofrece una biblioteca de algoritmos muy potentes de machine learning.
- GraphX: proporciona una API de procesamiento gráfico para la computación paralela de grafos.

Spark destaca por cumplir las siguientes tres características:

- Es una herramienta rápida ya que ejecuta las cargas de trabajo 100 veces más rápido que con Hadoop MapReduce. Con Spark, disfrutas de alto rendimiento con los datos por lotes y de streaming gracias al programador de grafos acíclicos dirigidos de última generación, al optimizador de consultas y al motor físico de ejecución.
- Su uso es sencillo ya que Spark cuenta con más de 80 operadores generales que facilitan el desarrollo de aplicaciones en paralelo. Puedes utilizarlo de forma interactiva desde el shell de Scala, Python, R y SQL para escribir aplicaciones rápidamente.
- Spark permite usar una gran cantidad de bibliotecas que incluye SQL, DataFrame, MLlib para aprendizaje automático, GraphX y Spark Streaming. Además, puedes combinarlas sin problemas en la misma aplicación.

3. Existen lenguajes de programación “recomendables” para gestionar datos. Entre ellos, están Python y Scala. Sería explicar brevemente por qué. (Iván)

Los conocimientos de programación son fundamentales sea cual sea la dirección que tomes en la ciencia de datos. Mientras que lenguajes como Python, R y SQL sirven de base para muchas funciones de ciencia de datos o analítica, otros son útiles para las trayectorias profesionales en áreas como el desarrollo de sistemas de datos o son más adecuados específicamente para los aspirantes a científicos de datos.

-¿Cómo se utiliza la programación en la ciencia de datos?

El campo de la ciencia de datos se basa en la programación en todas las funciones del trabajo, desde la automatización de la limpieza y la organización de conjuntos de datos brutos hasta el diseño de bases de datos y el ajuste de algoritmos de aprendizaje automático.

PYTHON

Python es un lenguaje de programación popular de propósito general además es un lenguaje de programación orientado a objetos de código abierto, que agrupa datos y funciones para lograr flexibilidad.

En la ciencia de datos, Python suele utilizarse para el procesamiento de datos, la implementación de algoritmos de análisis de datos y el entrenamiento de algoritmos de aprendizaje automático y aprendizaje profundo. Python admite múltiples estructuras de datos y utiliza una sintaxis en inglés sencillo, lo que lo convierte en un gran lenguaje para los programadores principiantes.

SCALA

Scala es una extensión de Java, un lenguaje fuertemente asociado a la ingeniería de datos, con interoperabilidad gracias a la compilación del bytecode de Java y su ejecución en la máquina virtual de Java. Construido como respuesta a los problemas percibidos en Java, es un lenguaje más nuevo y elegante.

Scala permite crear marcos de trabajo de alto rendimiento para el manejo de datos en silos, perfectos para la ciencia de datos a nivel empresarial.

Con vastas bibliotecas y soporte en entornos de desarrollo integrados (IDE) comunes, es funcional y escalable. Scala también admite el procesamiento concurrente y sincronizado.

4. En la parte de visualización de datos, de mostrar dashboards nos encontramos con PowerBI y Tableau entre otros. Debemos explicar qué son. (Iván)

Power BI

Power BI es un conjunto de herramientas que pone el conocimiento al alcance de todos y nos brinda acceso a nuestros datos de forma segura y rápida, generando grandes beneficios para nosotros y para nuestra empresa. Es un sistema predictivo, inteligente y de gran apoyo, capaz de traducir los datos (simples o complejos) en gráficas, paneles o informes por sus cualidades como la capacidad gráfica de presentación de la información, o la integración de Power Query: el motor de extracción, transformación y carga (ETL) incluido en Excel.

Power BI se conforma fundamentalmente de estos componentes:

-Power BI Desktop: aplicación gratuita de escritorio para transformar, visualizar datos y crear informes de los mismos.

-Power BI Service: servicio online (SaaS) con funcionalidad similar a la aplicación desktop y permite publicar informes y configurar la actualización de datos automáticamente para que el personal de la organización tenga los datos actualizados.

-Power BI Mobile: aplicación móvil disponible para Windows, iOS y Android para visualizar informes y que se actualiza automáticamente con los cambios de los datos.

Funcionalidad

Power BI permite conectar a cientos de orígenes de datos en la nube o entorno local, creando informes con objetos integrados o creando objetos personalizados.

El acceso a los datos puede ser desde una tabla Excel, Salesforce, Dynamic CRM, Google Analytics, hasta complejas bases de datos (on-premise o en la nube), información de servicios de Azure, etc., lo cual facilita tener toda la información en una única visualización.

Tableau

¿Qué es?

La plataforma de Tableau es la opción de inteligencia de negocios moderna líder en el mercado. Hace que sea más fácil explorar y administrar los datos. Asimismo, permite descubrir y compartir información más rápidamente a fin de generar grandes cambios en los negocios y en el mundo.

¿Para qué sirve?

La esencia de Tableau es simple y a la vez muy relevante: ayudar a las personas y empresas a ver y comprender todos sus datos. Y esto lo consigue ofreciendo a los usuarios toda una selección de herramientas útiles e intuitivas de inteligencia de negocios.

A través de funciones simples como la de arrastrar y soltar, cualquier persona puede acceder y analizar de forma sencilla datos, e incluso, crear informes y compartir esta información con otros usuarios.

Principales características de Tableau.


- Su facilidad de uso, apto para usuarios que no tienen conocimiento alguno de programación.
- Su principal ventaja, además de la anterior, es por supuesto su planteamiento visual y sencillo de todo tipo de datos complejos. Así, se adapta a organizaciones de sectores muy diversos y con características muy diferentes.
- La sencilla conexión con fuentes de datos para realizar análisis.
- Es una herramienta perfecta para trabajar en equipo, ya que permite el sencillo acceso a los datos por parte de diversas personas.
- La posibilidad de integración en aplicaciones propias de los usuarios.

FASE 2. Implementación de código (David)

En esta fase estaremos implementando la información recopilada de la fase 1.

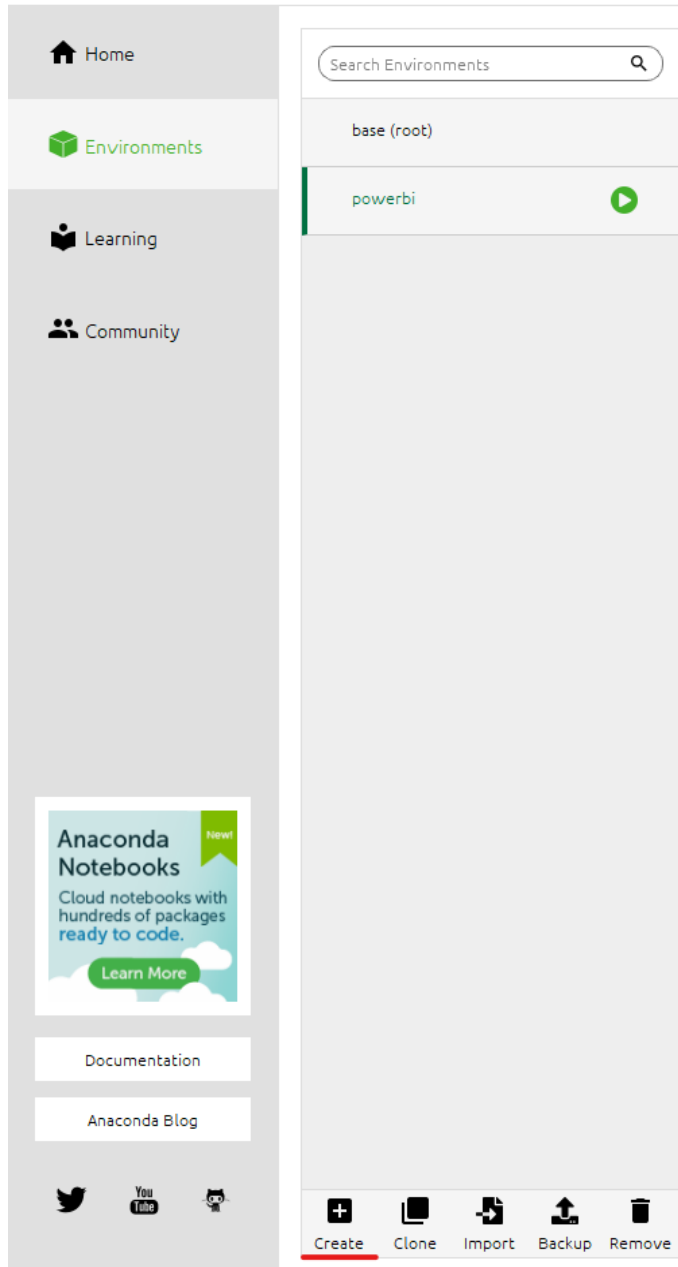
Estaremos usando Python en PowerBi. Pero para ello, hay que preparar el entorno.

En anaconda iremos a “Environments” y crearemos uno nuevo.

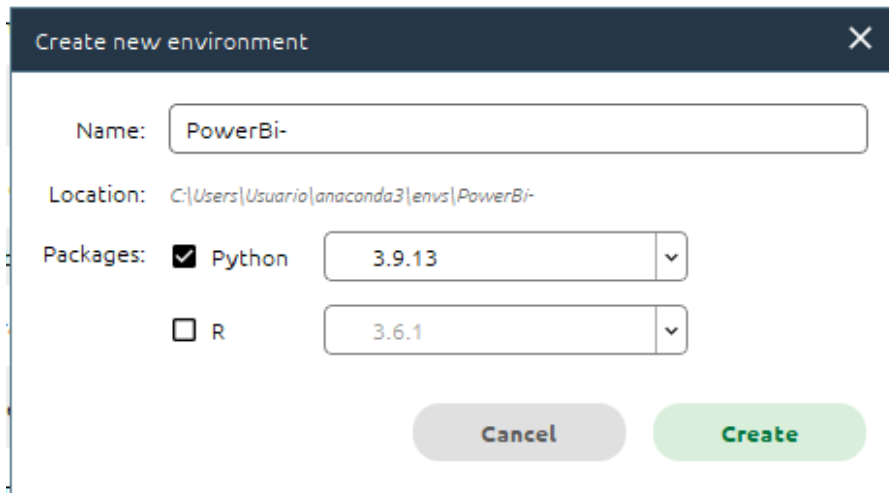
 Anaconda Navigator

File Help

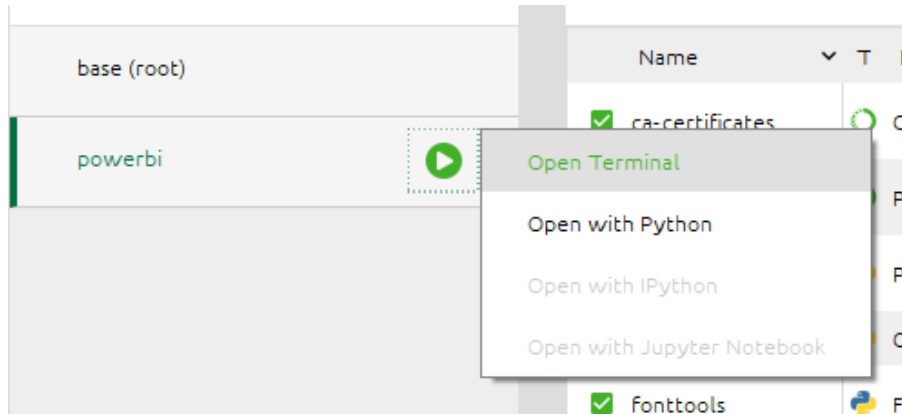
 ANACONDA.NAVIGATOR



Pondremos el nombre que queramos y la versión de Python.



Cuando esté creado el entorno, iremos a él y abriremos su terminal.



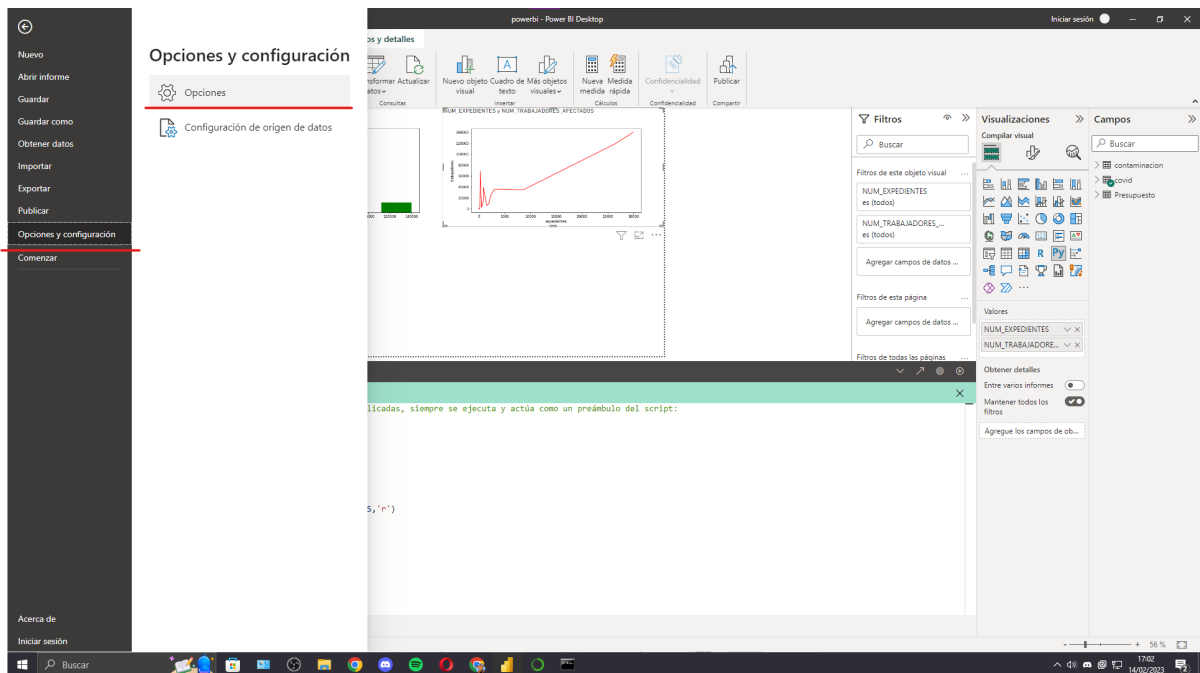
Instalamos pandas y matplotlib.

```
C:\Windows\system32\cmd.exe

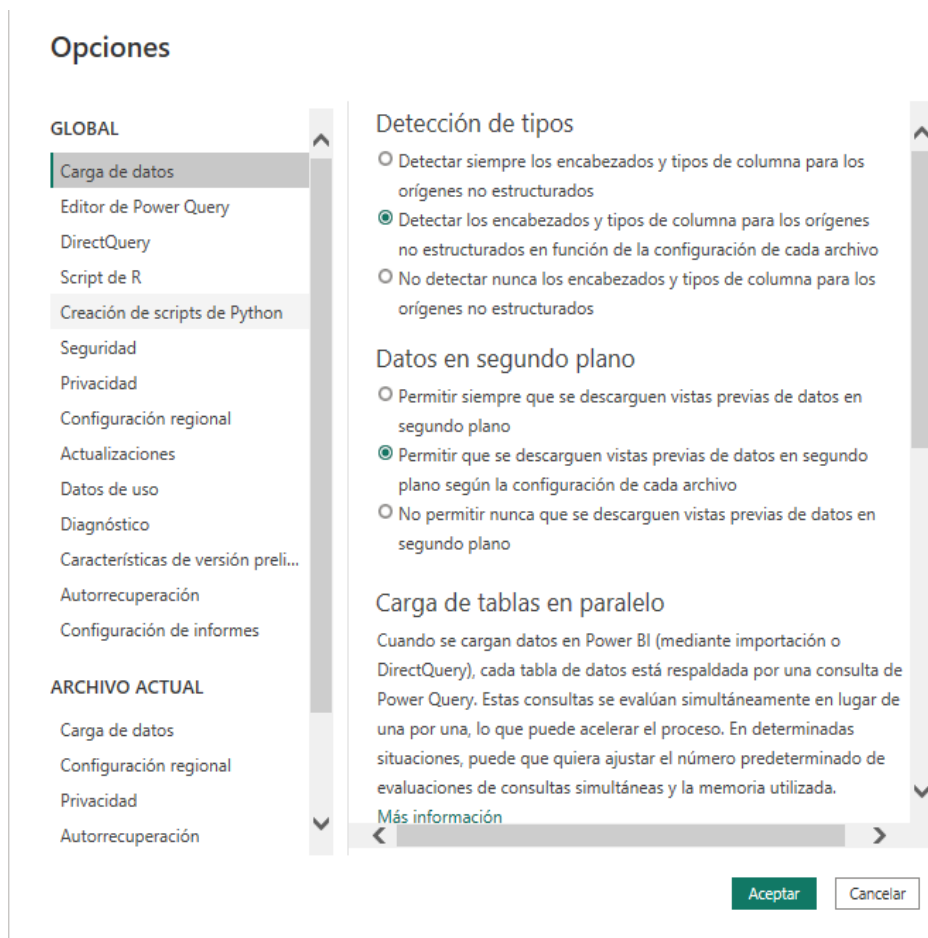
(powerbi) C:\Users\Usuario>pip install pandas
Collecting pandas
  Downloading pandas-1.5.3-cp39-cp39-win_amd64.whl (10.9 MB)
----- 10.9/10.9 MB 4.5 MB/s eta 0:00:00
Collecting pytz>=2020.1
  Downloading pytz-2022.7.1-py2.py3-none-any.whl (499 kB)
----- 499.4/499.4 kB 5.3 MB/s eta 0:00:00
Collecting numpy>=1.20.3
  Downloading numpy-1.24.2-cp39-cp39-win_amd64.whl (14.9 MB)
----- 14.9/14.9 MB 7.1 MB/s eta 0:00:00
Collecting python-dateutil>=2.8.1
  Using cached python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
Collecting six>=1.5
  Using cached six-1.16.0-py2.py3-none-any.whl (11 kB)
Installing collected packages: pytz, six, numpy, python-dateutil, pandas
Successfully installed numpy-1.24.2 pandas-1.5.3 python-dateutil-2.8.2 pytz-2022.7.1 six-1.16.0

(powerbi) C:\Users\Usuario>pip install matplotlib
Collecting matplotlib
  Downloading matplotlib-3.7.0-cp39-cp39-win_amd64.whl (7.6 MB)
----- 7.6/7.6 MB 8.6 MB/s eta 0:00:00
Collecting cycler>=0.10
  Using cached cycler-0.11.0-py3-none-any.whl (6.4 kB)
Collecting pyparsing>=2.3.1
```

Una vez hecho lo anterior, iremos a Powerbi y abriremos “Opciones”.



En opciones buscaremos “Creación de scripts de Python”.



Dentro, cambiaremos los directorios Raíz de Python a donde se nos haya guardado el entorno de anaconda.

Directorios raíz de Python detectados:

Otros

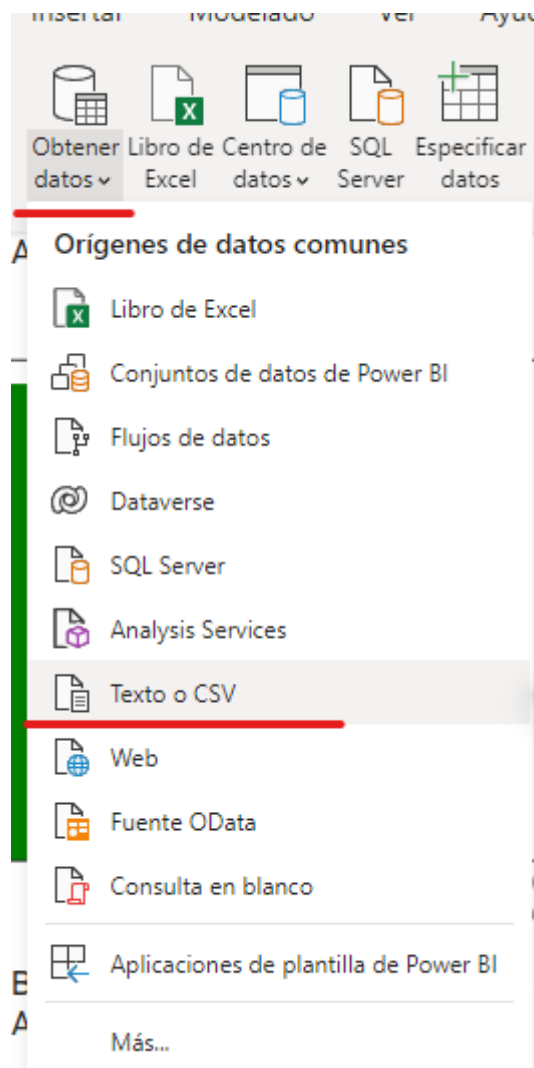
Establezca un directorio raíz para Python:

C:\Users\Usuario\anaconda3\envs\powerbi

Examinar

[Cómo instalar Python](#)

Para insertar un archivo CSV, lo único que tendremos que hacer será ir a “Obtener datos” – “Texto o CSV” y buscaremos el archivo que queremos insertar.



Cuando lo tengamos nos aparecerá de la siguiente manera y le daremos a cargar.

Gastos_Presupuesto.csv

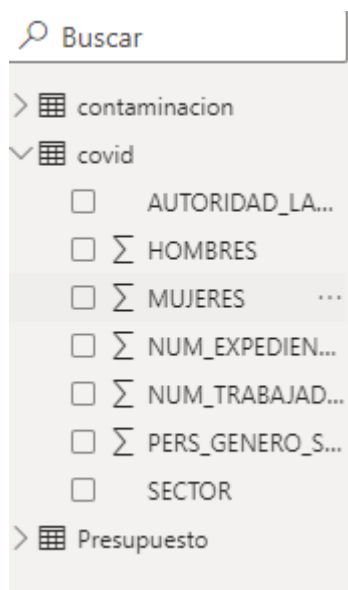
Origen de archivo: 1232: Europeo occidental (Windows) | Delimitador: Punto y coma | Detección del tipo de datos: Basado en las primeras 200 filas

Centro	Descripción Centro	Sección	Descripción Sección	Programa	Descripción Programa	Capítulo	Descripción Capítulo	Económico	Descripción Económico	Fondo	Descripción Fondo	Presupuesto
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	20000	RETRIBUCIONES BÁSICAS			100354
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	21000	RETRIBUCIONES BÁSICAS			16094
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	22000	RETRIBUCIONES COMPLEMENTARIAS			62359
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	23000	SUELDOS DEL GRUPO C2			20400
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	24000	TREINOS			4202
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	25000	COMPLEMENTO DE DESTINO			21999
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	26000	COMPLEMENTO ESPECÍFICO			29195
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	27000	OTROS COMPLEMENTOS			1815
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	28000	PRODUCTIVIDAD			8808
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	1	GASTOS DE PERSONAL	29000	SEGURIDAD SOCIAL			51841
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	2	GASTOS CORRIENTES EN BIENES Y SERVICIOS	22801	ATENCIÓNES PROTOCOLARIAS Y REPRESENTATIVAS			4905
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	2	GASTOS CORRIENTES EN BIENES Y SERVICIOS	22802	PUBLICIDAD Y PROPAGANDA			10000
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	2	GASTOS CORRIENTES EN BIENES Y SERVICIOS	22808	REUNIONES, CONFERENCIAS Y CURSOS			1000
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	2	GASTOS CORRIENTES EN BIENES Y SERVICIOS	23000	DE LOS MIEMBROS DE LOS ÓRGANOS DE GOBIERNO			1000
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	2	GASTOS CORRIENTES EN BIENES Y SERVICIOS	23020	DEL PERSONAL NO DIRECTIVO			1000
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	2	GASTOS CORRIENTES EN BIENES Y SERVICIOS	23100	GASTOS DE VIAJE MIEMBROS DE LOS ORG. DE GOBIERNO			2000
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91216	PRESIDENCIA DEL PLENO	2	GASTOS CORRIENTES EN BIENES Y SERVICIOS	23120	GASTOS DE VIAJE DEL PERSONAL NO DIRECTIVO			2000
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91230	SECRETARÍA GENERAL DEL PLENO	1	GASTOS DE PERSONAL	21000	SUELDOS DEL GRUPO A2			24919
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91230	SECRETARÍA GENERAL DEL PLENO	1	GASTOS DE PERSONAL	22001	SUELDOS DEL GRUPO A2			49182
1	AYUNTAMIENTO DE MADRID	100	PRESIDENCIA DEL PLENO	91230	SECRETARÍA GENERAL DEL PLENO	1	GASTOS DE PERSONAL	22003	SUELDOS DEL GRUPO C1			340439

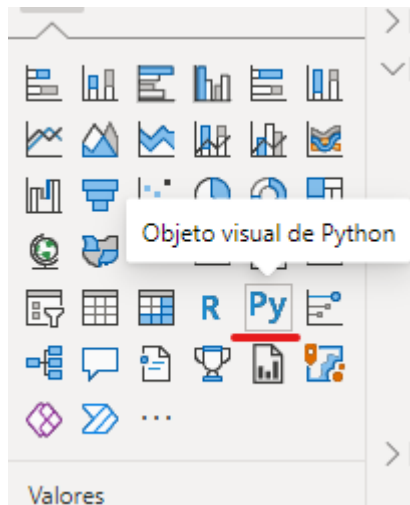
Los datos de la vista previa se han truncado debido a límites de tamaño.

Extraer tabla mediante ejemplos | Cargar | Transformar datos | Cancelar

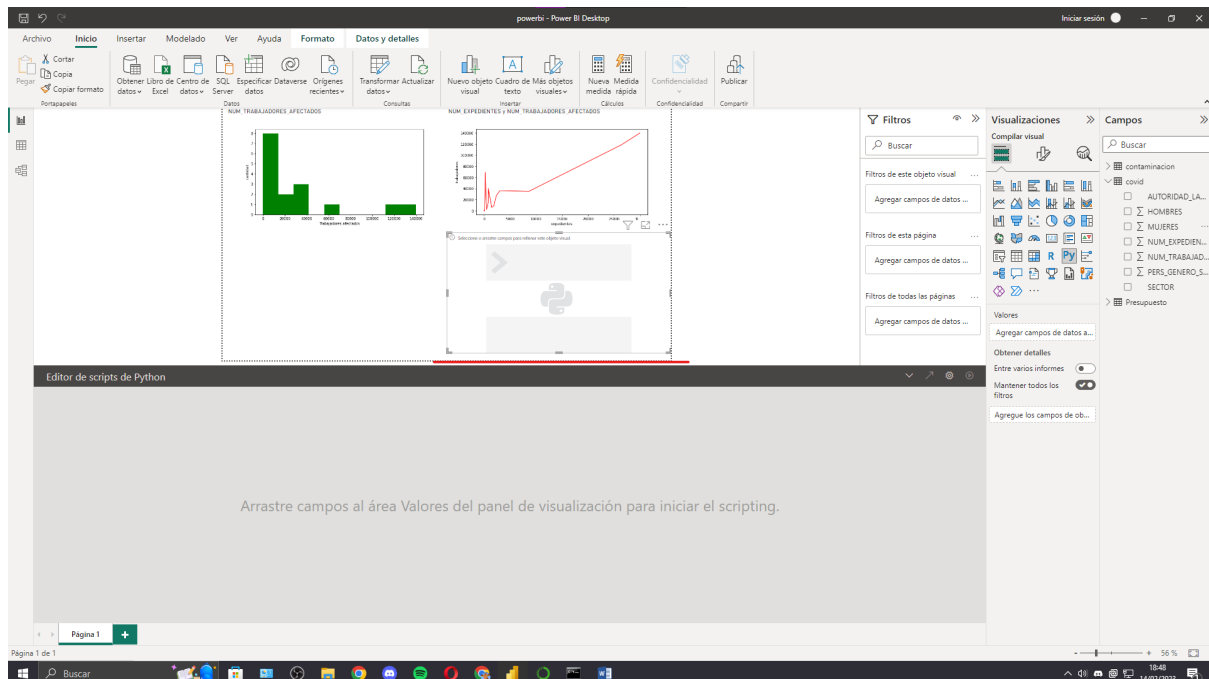
Los archivos que hemos insertado se muestran así, a la derecha del programa.



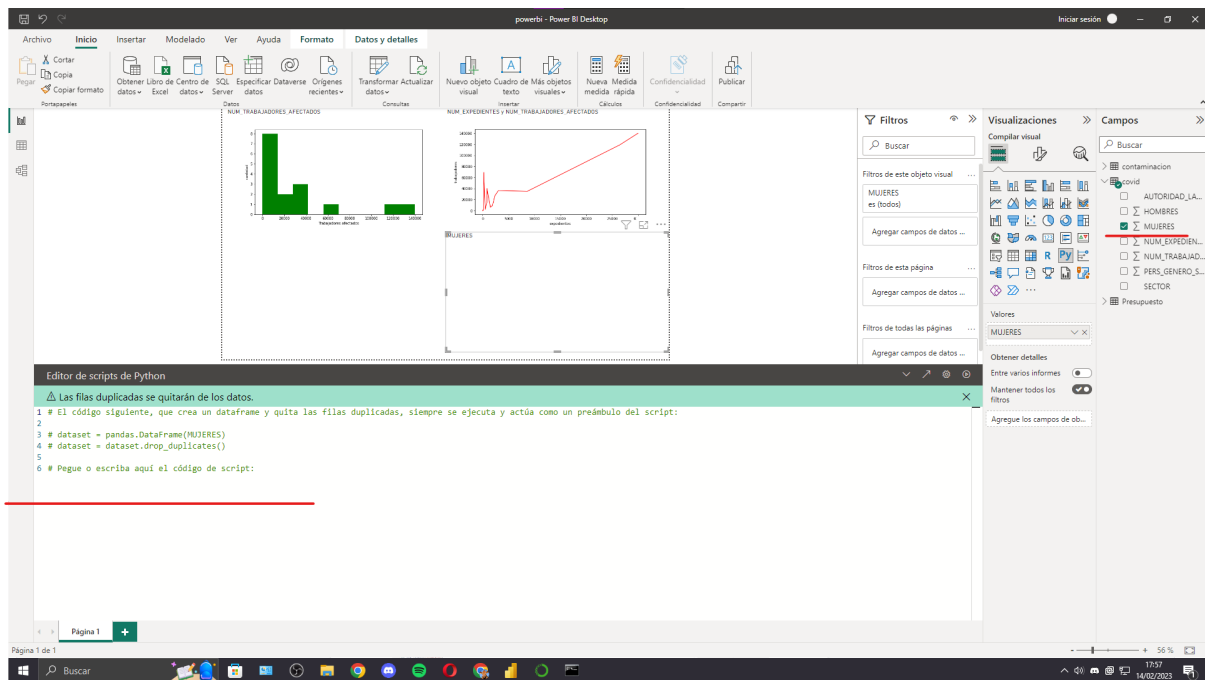
Para agregar un objeto y hacer nuestro gráfico lo que haremos será buscar el icono “Py” entre los diferentes elementos que nos aparecen.



Podemos ver cómo queda. Ahora mismo está desactivado, para activarlo tenemos que seleccionar alguna tabla del archivo que hemos insertado.



Una vez activado ya podremos escribir nuestro código



Histograma

En este gráfico vemos como es un histograma, esto lo he hecho importando pandas y matplotlib, seguido de eso he definido el histograma con los datos y he nombrado los ejes.

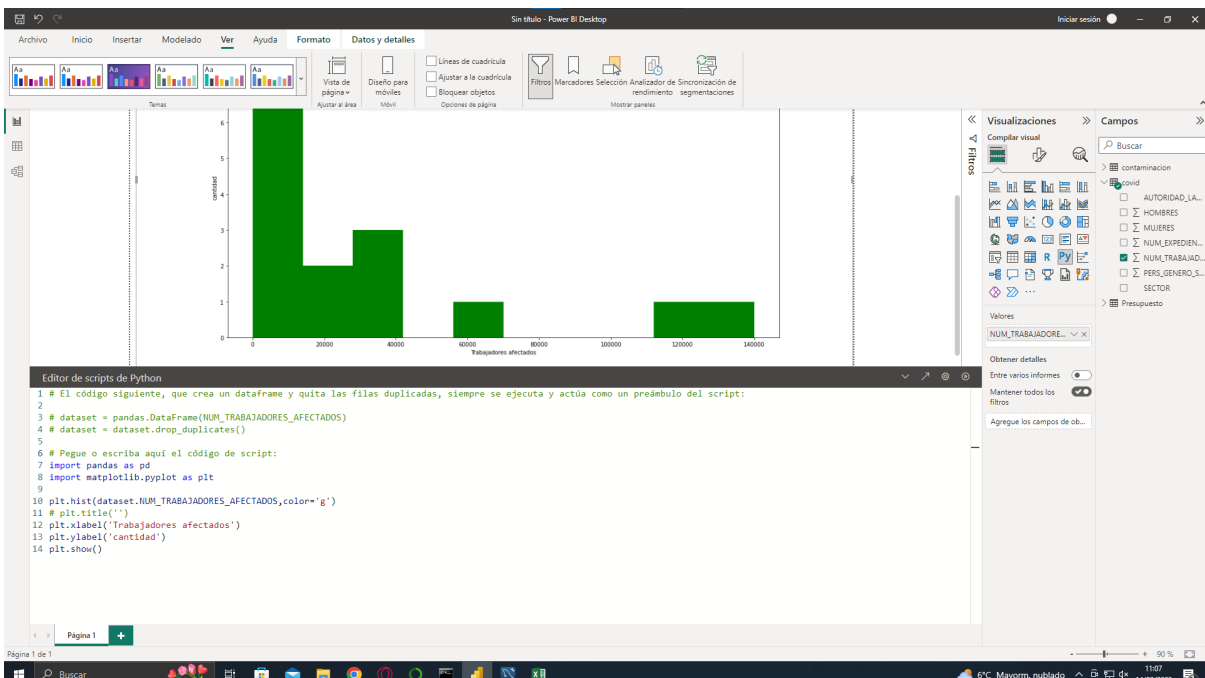
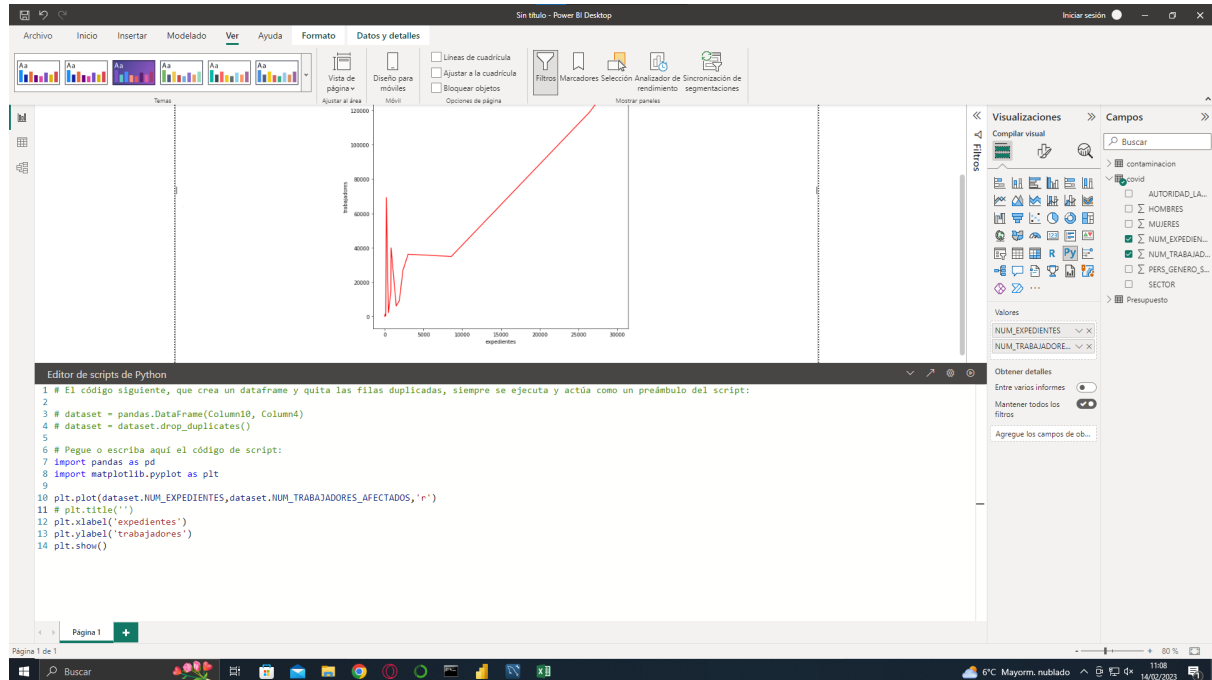


Gráfico de línea

Para hacer este gráfico de línea, lo he hecho importando pandas y matplotlib, seguido de eso he definido el gráfico con los datos y he nombrado los ejes.



FASE 3. Evaluación fase 2 y evolución (David)

Evaluación fase 2

PowerBi: he usado PoweBi porque me ha parecido una aplicación muy fácil de usar y cómoda. Permite insertar archivos de datos y hacer gráficos de forma muy rápida e intuitiva.

Python: he usado Python porque además de que me lo permitía la aplicación, ya se usarlo para poder realizar el trabajo. He usado las librerías de Pandas y Matplotlib, estas sirven para manejo de estructuras de datos y la creación de gráficos, respectivamente.

Evolución

Internet ha cambiado por completo en unos años: los negocios, las redes y el volumen del tráfico ha crecido exponencialmente. Esto ha permitido a los usuarios acceder, almacenar y procesar información en grandes cantidades.

Empezando por el acceso a archivos. Antiguamente la información se guardaba físicamente y manualmente en sistemas de carpetas y ficheros, esto era una tarea extremadamente lenta. Hasta que llegaron las primeras aplicaciones que manejan datos desde un ordenador, estos se encargan de organizar archivos y directorios. Sin embargo lo que se lleva hoy en día servicios de almacenamiento en la nube, como GoogleDrive y Dropbox. Estos te permiten acceder a los datos desde cualquier sitio siempre y cuando tengas conexión a internet. Los servicios de almacenamiento en la nube también tienen mayor seguridad y privacidad de los datos al encontrarse en servidores protegidos.

Las bases de datos por otra parte empiezan a ser más conocidas en los años setenta gracias al modelo relacional, a finales de los setenta aparece SQL y empieza a ser más normalizado. A finales de los 90 se empieza a investigar sobre las bases de datos orientadas a objetos y aparecen sistemas como Excel y Access.

Ahora hay 3 compañías que dominan el ámbito de las bases de datos:

- IBM
- Microsoft
- Oracle

Las API son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos.

API significa “interfaz de programación de aplicaciones”. En el contexto de las API, la palabra aplicación se refiere a cualquier software con una función distinta. La interfaz puede considerarse como un contrato de servicio entre dos aplicaciones.

A finales del año 2000 eBay lanza su API con la intención de impulsar el área de las soluciones de comercio electrónico, pero también fomentaron el uso de las APIs.

En verano de 2002, Amazon sacó también su API. Con la plataforma Amazon Web Services permitieron a los desarrolladores incorporar contenidos de Amazon a sus propias páginas web.

En 2006 fue cuando Facebook sacó su primera API con la que se podía acceder a información de usuario. Un mes más tarde lo hizo twitter.

Después de que Google se uniera a la moda de las APIs, Amazon quiso llevar las suyas propias a otro nivel y saltaron de solo usarlas para comercio electrónico a también usarlas para almacenar datos en la nube y para servidores virtuales en sus “data center”.

Webgrafía

1.1

<https://www.powerdata.es/data-lake>

<https://cloud.google.com/learn/what-is-a-data-lake?hl=es-419>

<https://datascientest.com/es/data-warehouse-que-es-y-como-utilizarlo>

<https://www.tibco.com/es/reference-center/what-is-structured-data#:~:text=Se%20llama%20datos%20estructurados%20cuando,en%20una%20base%20de%20datos.>

<https://ayudaleyprotecciondatos.es/bases-de-datos/diferencias-entre-datos-estructurados-y-no-estructurados/#:~:text=Los%20datos%20estructurados%20est%C3%A1n%20altamente,de%20recopilar%2C%20procesar%20y%20analizar.>

1.2

https://www.sas.com/es_es/insights/big-data/hadoop.html

<https://blog.mdcloud.es/que-es-spark-big-data-y-machine-learning/>

<https://cloud.google.com/learn/what-is-apache-spark?hl=es>

1.3

<https://blog.edx.org/es/9-principales-lenguajes-de-programacion-para-la-ciencia-de-datos>

1.4

<https://www2.deloitte.com/es/es/pages/technology/articles/que-es-power-bi.html>

<https://www.tableau.com/es-es/why-tableau/what-is-tableau>

<https://www.arimetrics.com/glosario-digital/tableau>

2.

<https://www.youtube.com/watch?v=t1ugYWnQZHI>

3.

<https://www.telefonica.com/es/sala-comunicacion/blog/el-drastico-cambio-de-internet-en-los-ultimos-10-anos/>

https://ikastaroak.birt.eus/edu/argitalpen/backupa/20200331/1920k/es/ASIR/GBD/GBD01/es_ASIR_GBD01_Contenidos/website_11_evolucin_de_los_sistemas_de_almacenamiento_de_la_informacin.html

<https://muytecnologicos.com/historia/historia-de-las-bases-de-datos>

<https://aws.amazon.com/es/what-is/api/#:~:text=Las%20API%20son%20mecanismos%20que,meteorología%20contiene%20datos%20meteorológicos%20diarios.>

<https://www.bbvaapimarket.com/es/mundo-api/breve-historia-de-las-apis-del-comercio-electronico-la-era-movil/>