HITO GRUPAL DE PROGRAMACIÓN

ÍNDICE

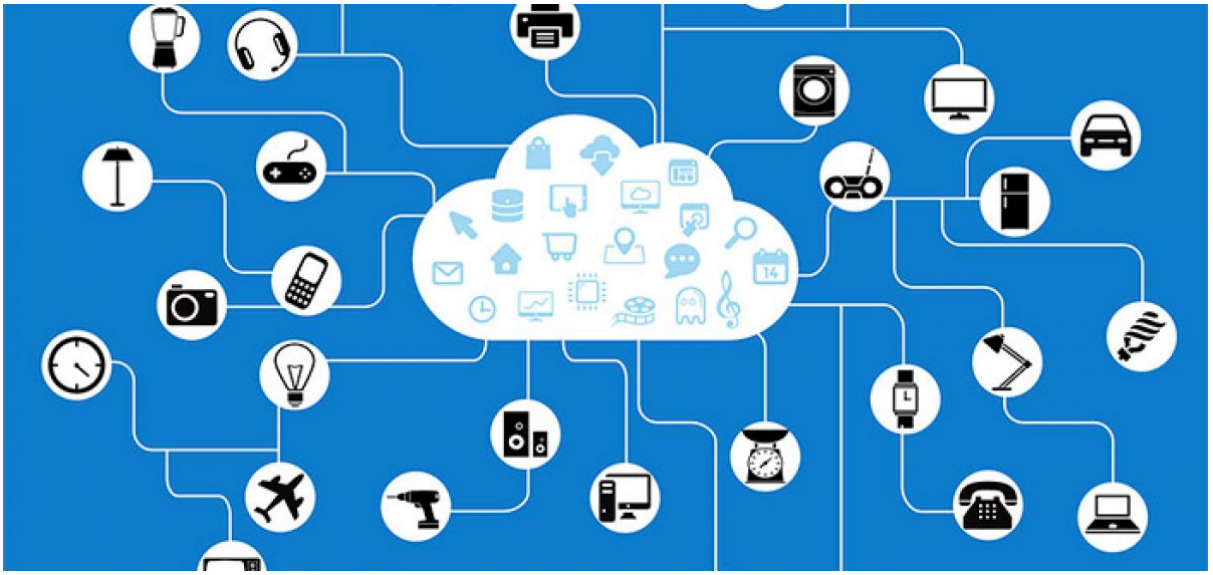
CUESTIÓN 1

1. Fuentes de datos.

Las fuentes de datos son los orígenes desde donde se recopilan y almacenan los datos que se van a utilizar para el análisis. Un ejemplo de fuente de datos es un data lake, que es un sistema de almacenamiento de datos no estructurados, semi-estructurados y estructurados en un solo lugar.

La diferencia entre datos estructurados y no estructurados radica en su formato y organización. Los datos estructurados están organizados en una forma específica y se pueden almacenar en una base de datos relacional o en una tabla, mientras que los datos no estructurados no están organizados y pueden incluir texto, imágenes, audio y vídeo.

En relación a Big Data, los datos estructurados son más fáciles de procesar y analizar, ya que están organizados y tienen un formato definido, mientras que los datos no estructurados requieren más esfuerzo para su procesamiento y análisis debido a su formato no estructurado. Sin embargo, los datos no estructurados pueden tener más información y proporcionar una visión más completa del problema a analizar. Por lo tanto, es importante tener en cuenta ambos tipos de datos en el análisis de Big Data.



2. Características y finalidad de Hadoop y Spark.

Hadoop y Spark son dos tecnologías que se utilizan para manejar grandes cantidades de datos en un entorno distribuido. Estas herramientas permiten almacenar y procesar datos a gran escala de manera eficiente y escalable.

Hadoop es un marco de software de código abierto que se utiliza para almacenar y procesar grandes cantidades de datos distribuidos. Hadoop permite dividir grandes cantidades de datos en pequeños fragmentos y distribuirlos en un cluster de servidores, lo que permite un procesamiento más rápido y eficiente. Además, Hadoop incluye un sistema de procesamiento de datos distribuidos llamado MapReduce que permite ejecutar tareas de procesamiento de datos de manera paralela en varios servidores.

Por otro lado, Spark es un motor de procesamiento de datos en cluster que permite procesar grandes volúmenes de datos de manera rápida y eficiente. Spark es más rápido que Hadoop debido a su capacidad para procesar datos en memoria en lugar de leerlos desde el disco. Esto hace que Spark sea adecuado para aplicaciones que requieren una respuesta rápida. Además, Spark incluye una serie de librerías para el análisis de datos, como SQL y DataFrames, que facilitan la gestión y el procesamiento de los datos.

En términos simples, Hadoop es un sistema que permite almacenar y procesar grandes cantidades de datos en varios servidores, mientras que Spark es un motor que permite procesar esos datos de manera rápida y eficiente.

Es importante tener en cuenta que Hadoop y Spark son complementarios y que muchas empresas utilizan ambos para lograr una solución completa para la gestión de datos. Por ejemplo, se puede utilizar Hadoop para almacenar y distribuir grandes cantidades de datos, y Spark para procesarlos de manera rápida y eficiente.

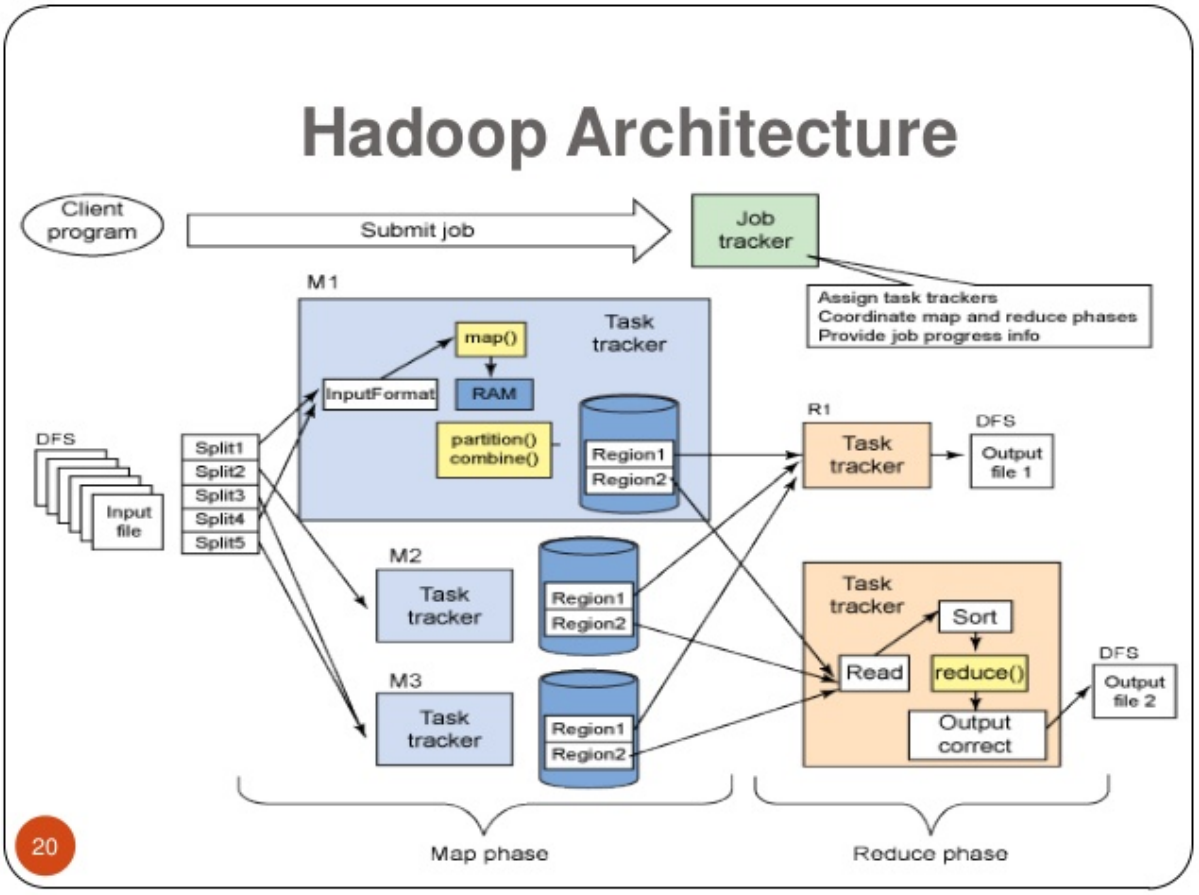
Hadoop y Spark son herramientas esenciales para cualquier empresa que desee gestionar y procesar grandes cantidades de datos de manera eficiente. Ofrecen una solución escalable y flexible para el almacenamiento y procesamiento de datos a gran escala, lo que las hace ideales para aplicaciones en diferentes industrias, como la banca, la salud y la investigación científica. Además, ambas herramientas están en constante evolución y se están actualizando para ofrecer más funciones y mejorar su rendimiento.

Otra ventaja de Hadoop y Spark es su compatibilidad con diferentes sistemas operativos, lo que permite a las empresas utilizarlos en su plataforma existente. Además, ambas tecnologías son compatibles con diferentes lenguajes de programación, lo que las hace aún más accesibles y flexibles para los desarrolladores.

En cuanto a la seguridad, Hadoop y Spark incluyen medidas de seguridad incorporadas, como la autenticación y la autorización, para proteger los datos sensibles. Sin embargo, es importante tener en cuenta que la seguridad depende en gran medida de la implementación y configuración correctas de las herramientas, por lo que es necesario tener cuidado con el acceso a los datos y los permisos de los usuarios.

Hadoop y Spark son herramientas poderosas y esenciales para la gestión y procesamiento de grandes cantidades de datos. Ofrecen soluciones escalables y flexibles para el almacenamiento y procesamiento de datos a gran escala, lo que las hace ideales para diferentes industrias y aplicaciones.

Además, su compatibilidad con diferentes sistemas operativos y lenguajes de programación, así como sus medidas de seguridad incorporadas, hacen de Hadoop y Spark herramientas valiosas para cualquier empresa que desee gestionar y procesar grandes cantidades de datos de manera eficiente y segura.



3. Explicación de Python y Scala.

Python y Scala son dos de los lenguajes de programación más populares y ampliamente utilizados para la gestión de datos. Ambas tienen sus fortalezas y debilidades, y la elección dependerá de las necesidades específicas de cada proyecto.

Python es un lenguaje de programación de alto nivel, fácil de aprender y con una amplia gama de librerías y herramientas para el análisis de datos. Algunas de las librerías más populares son Pandas, Numpy y Matplotlib. La comunidad de Python es muy activa y está constantemente desarrollando nuevas herramientas para mejorar la gestión de datos. Python es una excelente opción para la exploración y análisis de datos, y es utilizado por una gran variedad de industrias, desde la ciencia de datos hasta la finanzas.

Scala, por otro lado, es un lenguaje de programación de propósito general que se ejecuta en JVM (Java Virtual Machine) y tiene una gran integración con el ecosistema de Big Data de Apache Hadoop. Esto lo hace ideal para el procesamiento de grandes cantidades de datos y es ampliamente utilizado en entornos empresariales. Además, Scala permite una fácil integración con Java y otras tecnologías empresariales, lo que lo hace ideal para proyectos en los que se requiere una integración fluida con otras tecnologías empresariales.

Ambos lenguajes tienen una amplia comunidad de desarrolladores y una gran cantidad de recursos disponibles en línea, lo que los hace ideales para cualquier persona que quiera aprender a gestionar datos.

Además, tanto Python como Scala son lenguajes de programación muy versátiles y pueden ser utilizados para una amplia variedad de tareas, desde el análisis de datos hasta la programación de aplicaciones web y móviles.

Python y Scala son dos excelentes opciones para la gestión de datos y la elección dependerá de las necesidades específicas de cada proyecto. Ambas tienen una amplia comunidad de desarrolladores y una gran cantidad de recursos disponibles, lo que las hace ideales para cualquier persona que quiera aprender a gestionar datos.

4. Explicación de PowerBI y Tableau.

PowerBI y Tableau son herramientas de visualización de datos que se utilizan para crear dashboards y representaciones visuales de información empresarial. Estos dashboards permiten a los usuarios comprender rápidamente los patrones y tendencias en los datos, lo que ayuda a tomar decisiones informadas.

PowerBI es una herramienta de visualización de datos desarrollada por Microsoft que permite a los usuarios conectar y visualizar sus datos en una plataforma intuitiva.

PowerBI ofrece una amplia gama de gráficos y visualizaciones, lo que permite a los usuarios explorar y analizar sus datos de manera fácil. Además, PowerBI es compatible con una amplia gama de fuentes de datos, incluyendo bases de datos, aplicaciones en la nube y archivos.

Por otro lado, Tableau es una herramienta de visualización de datos más avanzada que permite a los usuarios crear visualizaciones interactivas y animaciones. Ofrece una amplia gama de herramientas y recursos para ayudar a los usuarios a comprender y explorar sus datos de manera más profunda.
Además, Tableau permite a los usuarios compartir sus visualizaciones con otros en la organización, lo que permite una colaboración más efectiva y una toma de decisiones más informada.

Ambas herramientas son fáciles de usar y ofrecen soluciones accesibles para la visualización de datos.

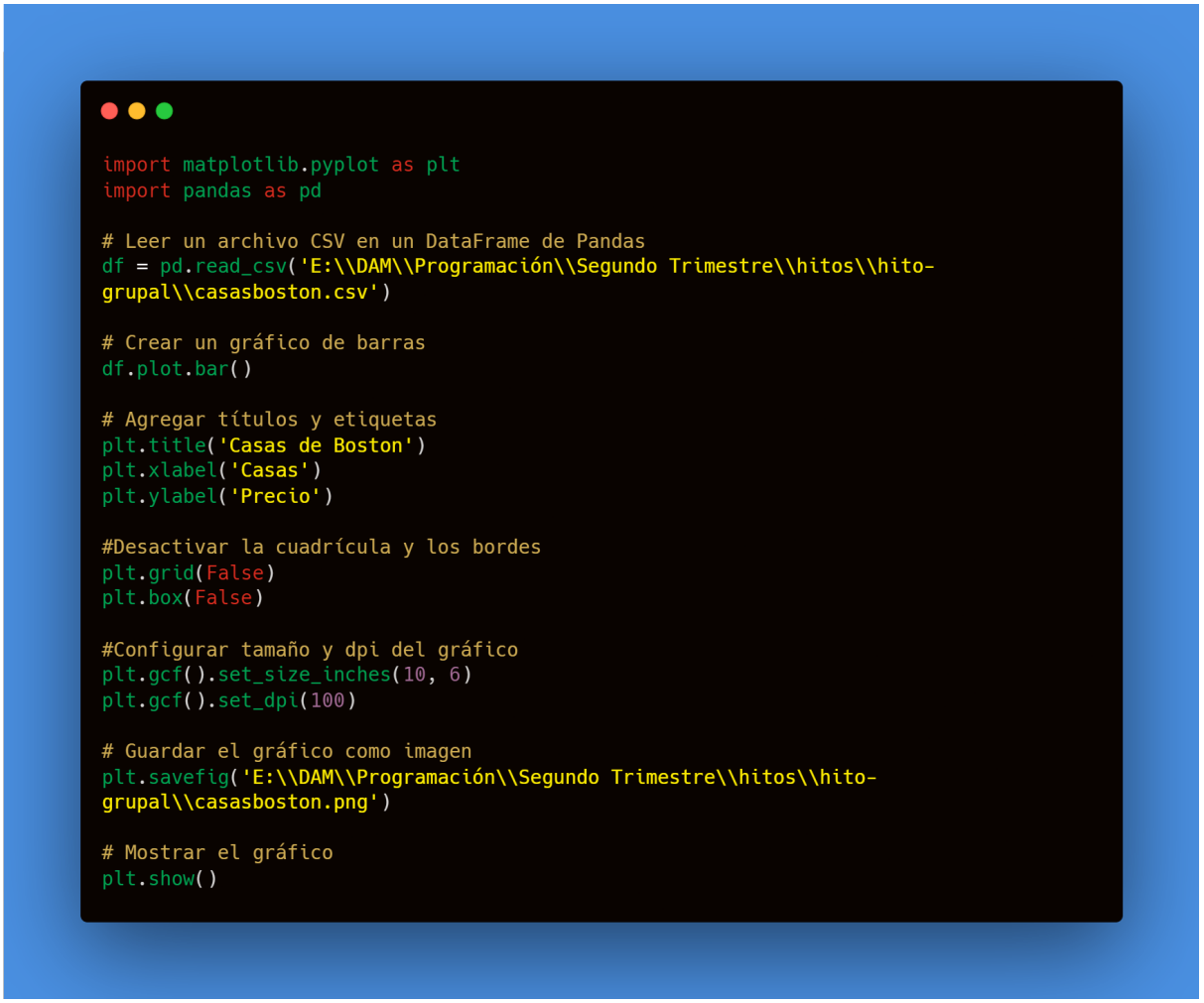
PowerBI es una opción ideal para usuarios que buscan una solución sencilla e intuitiva, mientras que Tableau es una opción más avanzada para usuarios que buscan una solución más completa y profunda.

PowerBI y Tableau son herramientas esenciales para cualquier organización que busque comprender y visualizar sus datos de manera efectiva. Ofrecen soluciones flexibles y accesibles para la visualización de datos, lo que ayuda a las empresas a tomar decisiones informadas y mejorar sus operaciones.

CUESTIÓN 2

En esta segunda fase se realiza la implementación de la investigación. En concreto sería acceder a un volumen de datos y mostrarlo. Podríamos utilizar Scala o Python y mostrar el resultado en PowerBI o Tableau. La idea es que sea algo muy impactante por la calidad de contenido tratado, velocidad de acceso, volumen de datos...

Para abordar esta segunda fase de implementación, hemos desarrollado un script en python para mostrar el gráfico:



```python
import matplotlib.pyplot as plt
import pandas as pd

# Leer un archivo CSV en un DataFrame de Pandas
df = pd.read_csv('E:\\DAM\\Programación\\Segundo Trimestre\\hitos\\hito-grupal\\casasboston.csv')

# Crear un gráfico de barras
df.plot.bar()

# Agregar títulos y etiquetas
plt.title('Casas de Boston')
plt.xlabel('Casas')
plt.ylabel('Precio')

#Desactivar la cuadrícula y los bordes
plt.grid(False)
plt.box(False)

#Configurar tamaño y dpi del gráfico
plt.gcf().set_size_inches(10, 6)
plt.gcf().set_dpi(100)

# Guardar el gráfico como imagen
plt.savefig('E:\\DAM\\Programación\\Segundo Trimestre\\hitos\\hito-grupal\\casasboston.png')

# Mostrar el gráfico
plt.show()
```

El script comienza importando las librerías Matplotlib y Pandas. Luego, se usa Pandas para leer un archivo CSV (ubicado en la ruta especificada) en un DataFrame de Pandas.

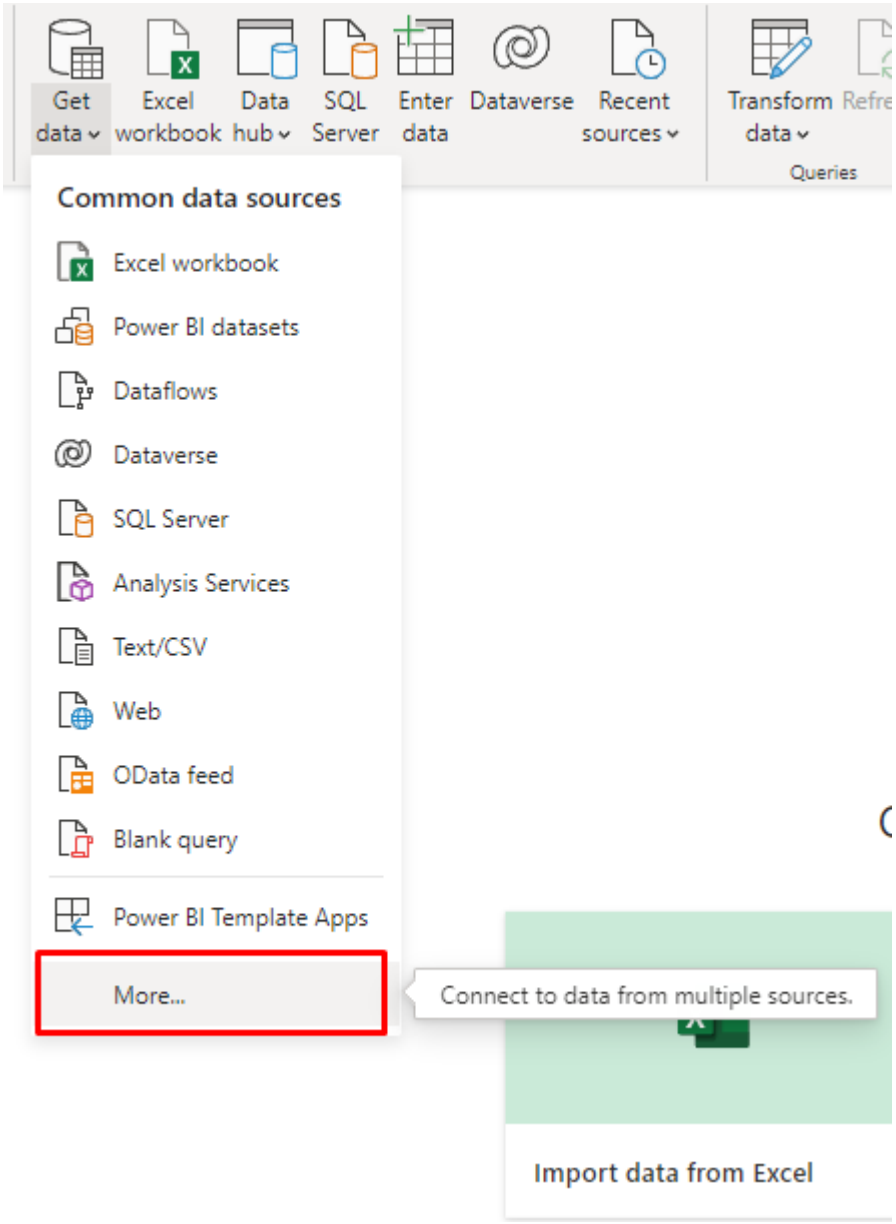
Después, se crea un gráfico de barras a partir de los datos en el DataFrame utilizando el método "plot.bar()".

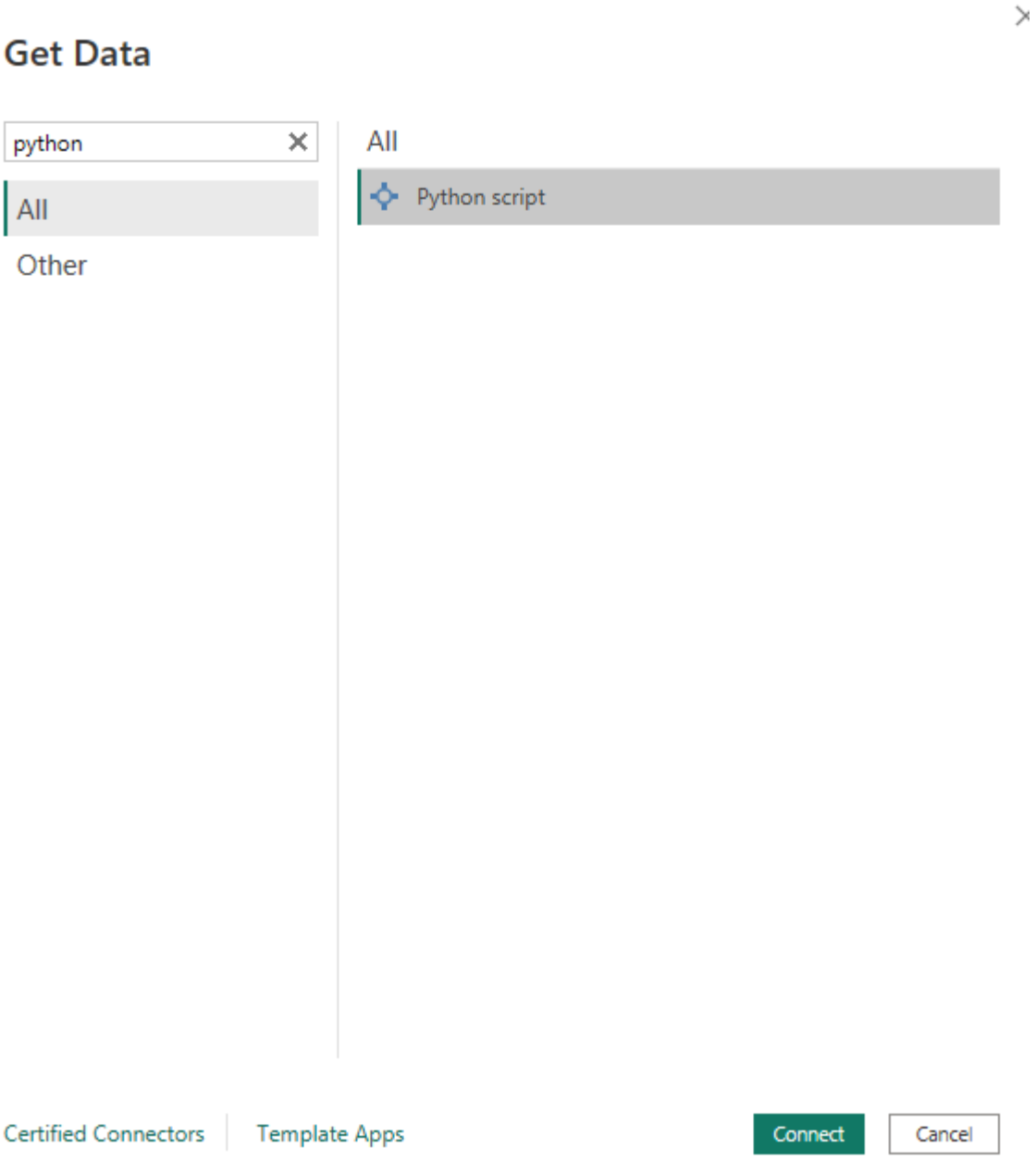
Luego, se agrega un título, etiquetas para el eje x e y, y se desactiva la cuadrícula y los bordes. También se configura el tamaño y la resolución (dpi) del gráfico.

Finalmente, se guarda el gráfico como una imagen (en la ruta especificada) y se muestra en la pantalla.

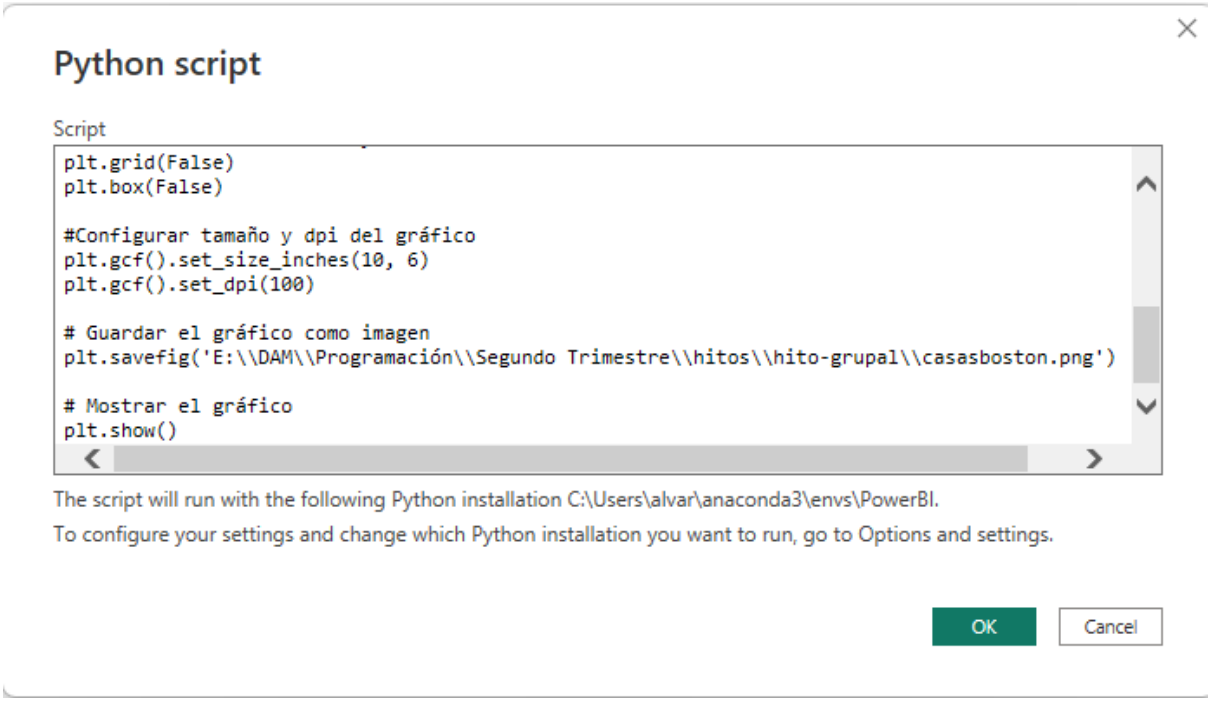
1º Una vez creado el script en Python accedemos a PowerBI y creamos un nuevo informe.

2º Después hay que pulsar la opción de “Obtener datos” y seleccionar “Python”.

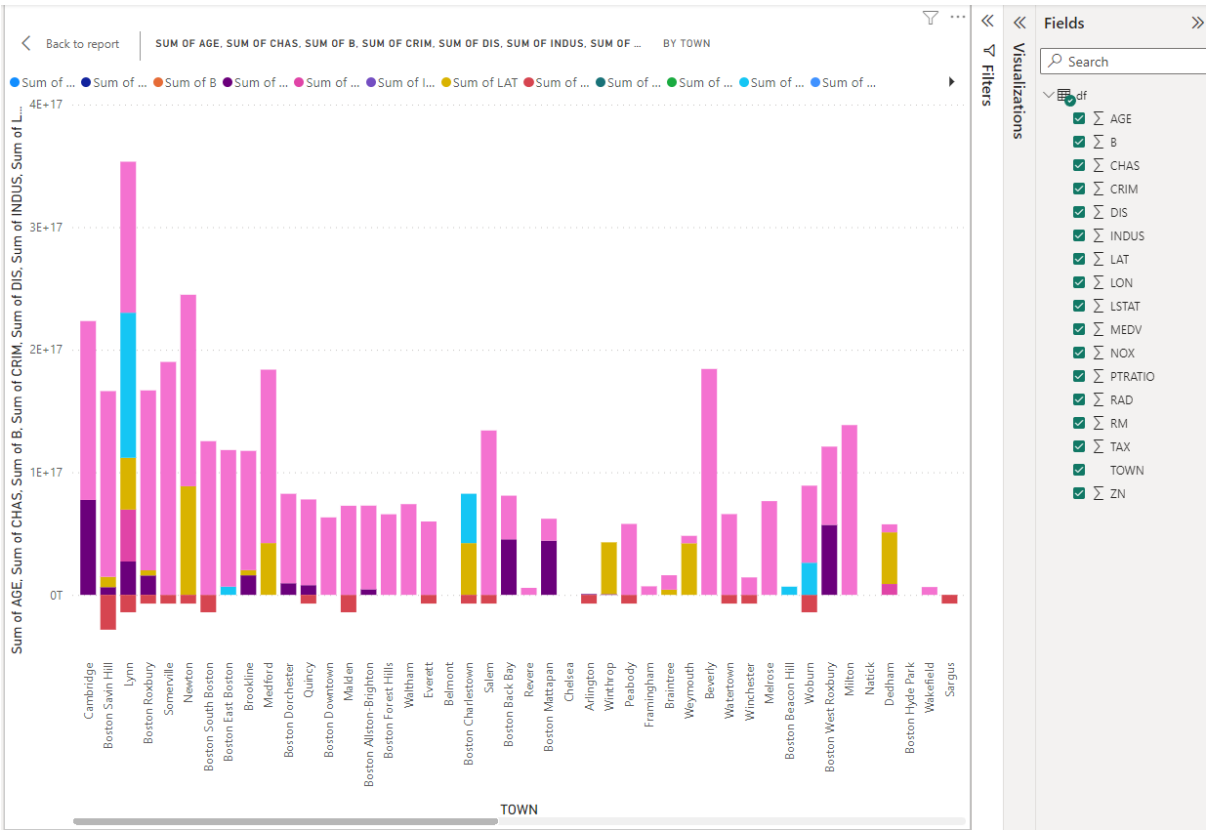



Get Data
python
All
Other
All
Python script
Certified Connectors
Template Apps
Connect
Cancel

3º Una vez que hemos accedido a la obtención de datos en Python, pegamos el script.



4º Una vez se han importado los datos del csv y ejecutado el script, creamos un gráfico con los datos importados.



CUESTIÓN 3

Realizar una evaluación o consideraciones de cómo ha evolucionado el acceso a datos en los últimos años. Desde acceso a ficheros, pasando por base de datos y consumiendo APIs.

En esta fase, evaluaremos y consideraremos la evolución del acceso a los datos a lo largo de los años. Desde el acceso a los archivos hasta la era del Big Data, veremos cómo ha evolucionado la gestión y el análisis de datos.

El acceso a los datos ha evolucionado significativamente en los últimos años. En el pasado, los datos se almacenaban en archivos y se accedía a ellos de manera manual. Con el tiempo, se desarrollaron bases de datos que permitieron una gestión más eficiente de los datos y un acceso más rápido y fácil a ellos.

En la actualidad, con el auge de la tecnología y la necesidad de acceder a grandes cantidades de datos, se han desarrollado APIs que permiten acceder a los datos de una manera más sencilla y eficiente. Además, las APIs también permiten integrar los datos de diferentes fuentes y aprovecharlos de manera más efectiva.

En conclusión, la evolución del acceso a los datos ha sido de una gestión manual de archivos a un acceso rápido y eficiente a través de bases de datos y APIs. Esto ha permitido una gestión más eficiente de los datos y una toma de decisiones más informada.

Tecnologías y servicios para la realización del acceso a datos:

- Base de datos: SQL, NoSQL, NewSQL, etc.
- Data Warehousing: Amazon Redshift, Google BigQuery, etc.
- Cloud Computing: Amazon Web Services, Microsoft Azure, Google Cloud Platform, etc.
- Big Data: Apache Hadoop, Apache Spark, etc.
- Análisis de datos: Python, Scala, R, etc.

Herramientas y software para la gestión de datos:

- Data Lakes: Amazon S3, Microsoft Azure Data Lake, Google Cloud Storage, etc.
- Apache Hadoop: HDFS, MapReduce, etc.
- Apache Spark: Spark SQL, Spark Streaming, etc.
- Visualización de datos: PowerBI, Tableau, QlikView, etc.
- Herramientas de análisis de datos: Jupyter, Zeppelin, R Studio, etc.

Evaluación de herramientas y conceptos de análisis de datos:

- La evolución de la gestión de datos ha sido impresionante, con la aparición de nuevas tecnologías y servicios para el acceso y análisis de datos en tiempo real.
- La tendencia actual es hacia la gestión de grandes volúmenes de datos y la importancia que tienen en la toma de decisiones empresariales.
- Aunque se han presentado grandes avances en la gestión de datos, aún existen desafíos como la seguridad y privacidad de la información.

Recomendaciones para la mejora del análisis de datos:

- Mantenerse al día con las nuevas tecnologías y herramientas para la gestión de datos.
- Implementar estrategias para garantizar la seguridad y privacidad de los datos.
- Desarrollar una cultura de datos en la organización, donde los datos sean vistos como un recurso valioso y estratégico.
- Invertir en la formación y capacitación de los empleados para mejorar su competencia en análisis de datos.
- Integrar el análisis de datos en la toma de decisiones empresariales para obtener mejores resultados.

Bibliografía

- Fuentes de datos
- Hadoop y Spark: ¿qué tecnologías son?¿para qué sirven?...
- ¿Para qué sirve PowerBI?
- Comparaciones Hadoop y Spark