SEGUNDO HITO GRUPAL – PROGRAMACIÓN

CARLOS HERNÁNDEZ ZABALGO

ÁLVARO ACOSTA ESTEBAN

DIEGO GARCÍA LUQUE

ESTA FASE ES UNA EXPLICACIÓN "TEÓRICA" DE CUATRO PUNTOS FUNDAMENTALES:

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 Big Data.

   Las fuentes de datos de grandes volúmenes son sistemas de almacenamiento y gestión de datos diseñados para manejar una cantidad masiva y diversa de datos.
   Ejemplos: data lakes, Hadoop, y sistemas de almacenamiento en la nube.

   Estos sistemas permiten almacenar, procesar y analizar datos en tiempo real para obtener información valiosa y tomar decisiones informadas.
   Una data lake es un almacenamiento de datos sin estructurar que permite almacenar una gran cantidad de datos en su formato original. Permite una variedad de fuentes de datos para ser almacenados de manera organizada para analizarse posteriormente.
   Algunos ejemplos de fuentes de datos incluyen transacciones comerciales, clics en un sitio web, información de redes sociales, entre otros.

   Diferencia de datos estructurados y no estructurados en relación al big Data:
   Los datos estructurados tienen una forma definida, como una tabla de bases de datos. Estos datos son fáciles de analizar con herramientas convencionales de bases de datos y análisis.
   Por otro lado, los datos no estructurados son aquellos que no tienen una forma definida y organizada, como textos, imágenes, vídeos, audio, entre otros. Estos datos son más difíciles de procesar y analizar, ya que no se pueden representar en una tabla.
   En el contexto del Big Data, los datos no estructurados representan un gran desafío para su procesamiento debido a su volumen y diversidad.
   Sin embargo, estos datos contienen información valiosa y es necesario encontrar maneras de procesarlos y analizarlos de manera efectiva. Por eso, se han desarrollado herramientas para manejar grandes volúmenes de datos no estructurados en el marco del Big Data.

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.*

Hadoop y Spark son herramientas importantes para la gestión de grandes volúmenes de datos y están diseñadas para resolver los desafíos asociados con el almacenamiento y el procesamiento de datos en grandes cantidades.

Hadoop es un marco de trabajo de código abierto, que permite el almacenamiento y el procesamiento de grandes cantidades de datos distribuidos en clústeres de computadoras.

Características de Hadoop:

- *Almacenamiento distribuido*: distribuye los datos a través de nodos en un cluster, lo que permite almacenar grandes cantidades de datos sin un sistema de almacenamiento centralizado.

- *Procesamiento en paralelo*: permite el procesamiento de datos en paralelo a través de nodos, lo que aumenta la velocidad y la eficiencia del procesamiento.

- *Escalabilidad*: es altamente escalable y se puede agregar más nodos al cluster para manejar mayores cantidades de datos.

- *Tolerante a fallos*: los datos se replican en varios nodos para garantizar la disponibilidad y la seguridad de los datos.

- *Integración con otros sistemas*: se integra fácilmente con otros sistemas y tecnologías, como Spark, Hive, entre otros.

- *Abierto y flexible*: es de código abierto y ofrece muchas opciones para personalizar y extender su funcionalidad.

La finalidad de Hadoop es permitir el almacenamiento y procesamiento de grandes cantidades de datos distribuidos en clústeres de computadoras. Gracias a esto las empresas pueden analizar y extraer toda la información que necesiten para mejorar su eficiencia y eficacia.

Spark es un marco de trabajo de código abierto que permite el procesamiento de grandes volúmenes de datos en tiempo real y batch.

Este es más rápido que Hadoop, ya que procesa los datos en la memoria en lugar de en el disco duro.

Características de Spark:

- Es compatible con diferentes lenguajes de programación.
- Amplia gama de herramientas para el análisis de datos.
- Amplia gama de herramientas para la inteligencia artificial.
- Procesamiento distribuido: permite distribuir el procesamiento de datos a través de múltiples nodos en un cluster, lo que aumenta la velocidad y la eficiencia.
- Alta velocidad: es capaz de procesar amplias cantidades de datos, lo que permite realizar análisis en tiempo real.
- Interfaz de programación sencilla: ofrece una interfaz de programación de alto nivel que permite a los usuarios escribir código en lenguajes como Scala, Python y R.
- Integración con otras tecnologías: se integra con otros sistemas y tecnologías, como Apache Hadoop, Apache Hive, entre otros.

La finalidad de Spark es proporcionar una plataforma eficiente para procesar grandes cantidades de datos y realizar análisis de datos y ciencia de datos de forma rápida y sencilla.

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

Python y Scala son lenguajes de programación "recomendables" para gestionar datos debido a las siguientes razones:

- Python es un lenguaje de programación de alto nivel que ofrece una gran cantidad de bibliotecas y herramientas para el análisis y la manipulación de datos. Estas incluyen Pandas, Numpy, Matplotlib, entre otras.

Además, tiene una sintaxis sencilla y clara, lo que lo hace ideal para los profesionales que trabajan con datos. Es un lenguaje muy versátil que puede ser utilizado en una amplia gama de aplicaciones, desde ciencia de datos hasta automatización y desarrollo web.

- Scala es recomendable por ser un lenguaje de programación muy escalable y eficiente que se ejecuta en la plataforma de Java.

Es una muy buena opción para el procesamiento de amplios volúmenes de datos gracias a su integración con Apache Spark, un motor de procesamiento de datos distribuido de código abierto. Además, es un lenguaje de programación funcional y orientado a objetos que ofrece una sintaxis clara y potente para la manipulación de datos.

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

   PowerBi y Tableau son herramientas de visualización de datos que son capaces de crear dashboards interactivos y representaciones gráficas de datos.

   - PowerBI permite conectar, analizar y visualizar datos de diferentes fuentes en tiempo real. Ofrece múltiples opciones de gráficos, tablas y mapas para representar los datos.

   - Tableau es una plataforma de análisis y visualización de datos que permite crear visualizaciones avanzadas, con una interfaz de arrastrar y soltar. También permite conectar y analizar datos de diferentes fuentes y compartir los resultados en tiempo real a través de la web y dispositivos móviles.

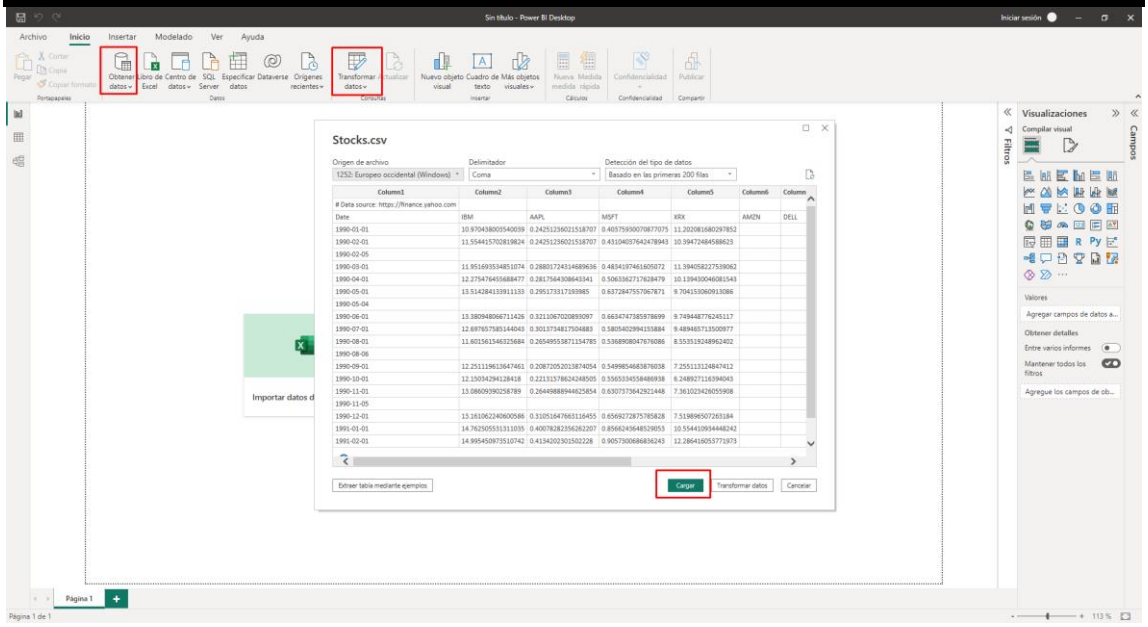
   Ambos son herramientas muy populares en el mundo de la visualización de datos y ofrecen una amplia variedad de opciones para crear dashboards y presentar datos.

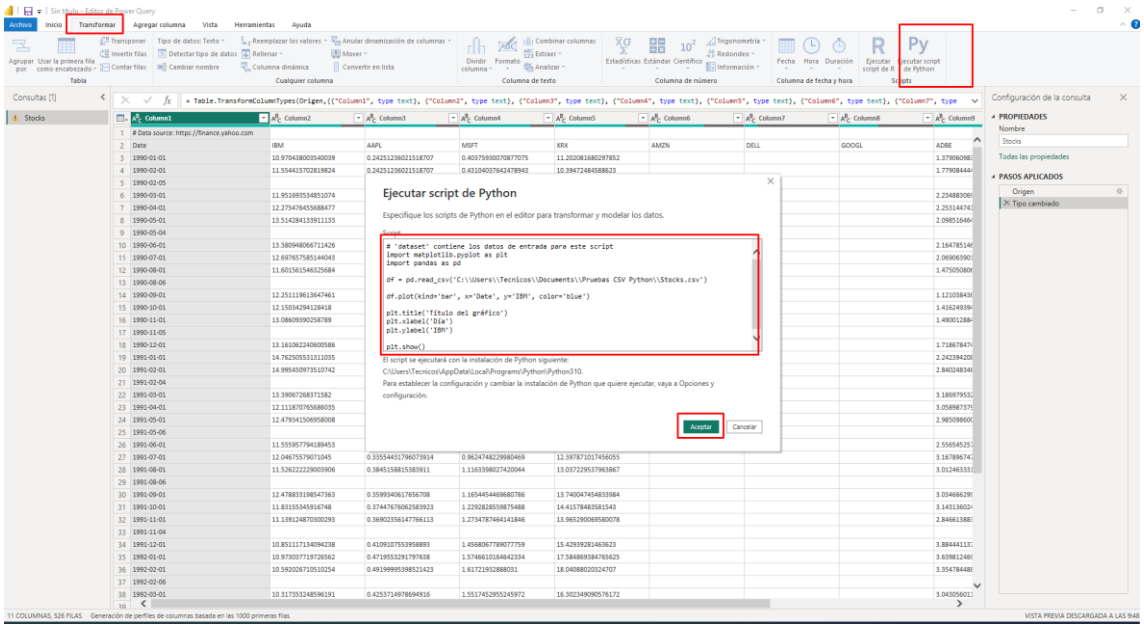
   - PowerBi crea dashboards en tiempo real y es capaz de compartirlos con otros usuarios.
   - Tableau está más enfocada a la visualización de datos avanzados, que permite conectar y visualizar datos de diferentes fuentes. También puede realizar análisis complejos y crear dashboards.

EN ESTA SEGUNDA FASE SE REALIZA LA IMPLEMENTACIÓN DE LA INVESTIGACIÓN. EN CONCRETO SERÍA ACCEDER A UN VOLUMEN DE DATOS Y MOSTRARLO.

Para acceder a una fuente de datos, hemos decidido utilizar el lenguaje de programación de Python. Para ello, hemos decido conectarnos a un archivo CSV de la siguiente manera:

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

# Leer el archivo CSV en un dataframe de pandas
df = pd.read_csv('C:\\Users\\Tecnicos\\Documents\\Pruebas CSV Python\\Stocks.csv')

# Crear un gráfico de barras
df.plot(kind='bar', x='Date', y='IBM', color='blue')

# Agregar títulos y etiquetas
plt.title('Título del gráfico')
plt.xlabel('Día')
plt.ylabel('IBM')

# Mostrar el gráfico
plt.show()
```





Para la realización de este apartado, me he guiado del siguiente video →

https://www.youtube.com/watch?v=t1ugYWnQZHI&t

PARA FINALIZAR, REALIZAMOS UNA EVALUACIÓN O CONSIDERACIONES DE CÓMO HAN EVOLUCIONADO EL ACCESO A DATOS EN LOS ÚLTIMOS AÑOS. DESDE ACCESO A FICHEROS, PASANDO POR BASE DE DATOS Y CONSUMIENDO APIS.

El acceso a datos ha evolucionado mucho en los últimos años, se han producido cambios en:

- Acceso a ficheros: Antes los archivos eran almacenados localmente y se accedía a ellos mediante un ordenador, hoy en día es posible acceder a ellos desde cualquier lugar a través de servicios en la nube.
- Base de datos: Las bases de datos han pasado de ser un recurso centralizado y controlado por un departamento de TI, a ser distribuidas y accesibles a través de la nube.
- Consumiendo APIs: Con el crecimiento de la economía de la API, los datos pueden ser accedidos y utilizados por aplicaciones y sistemas externos, lo que permite una mayor integración y colaboración.

En general, estos cambios han permitido un acceso más fácil y seguro a los datos.

INDICA QUÉ RECOMENDACIONES EN EL ANÁLISIS DE DATOS PROPONES PARA SU MEJORA.

Nuestras recomendaciones para mejorar en el análisis de datos serían:

- Limpiar y validar los datos antes de analizarlos.
- Utilizar gráficos y tablas para representar visualmente los datos y para identificar tendencias y patrones.
- Aplicar modelos estadísticos apropiados para su conjunto de datos y problemas.
- Validar los resultados utilizando técnicas de validación cruzada.
- Automatizar tareas repetitivas y procesos para ahorrar tiempo y mejorar la precisión.
- Colaborar y trabajar con otros expertos en el campo y compartir sus hallazgos y resultados.
- Mantener su conocimiento actualizado y aplicar las últimas técnicas y herramientas en el análisis de datos.

EVALÚA LAS HERRAMIENTAS Y CONCEPTOS DE LA ANALÍTICA DE DATOS ANALIZADA.

- *Almacenamiento de datos* → Herramientas para almacenar y gestionar grandes volúmenes de datos, como bases de datos NoSQL, Hadoop, etc.
- *Aprendizaje profundo* → Algoritmos de aprendizaje automático basados en redes neuronales.
- *Seguridad de datos* → Conjunto de políticas, procesos y tecnologías para proteger la privacidad y seguridad de los datos.
- *Procesamiento de datos en tiempo real* → Herramientas para procesar y analizar datos en tiempo real, como Apache Spark.
- *Integración de herramientas* → Integración de diferentes herramientas de análisis de datos para obtener una visión completa y holística de los datos.
- *Análisis predictivo* → Técnicas para predecir resultados futuros basados en patrones y tendencias en los datos.
- *Análisis de sentimientos* → Análisis de opiniones y percepciones en medios sociales y otros medios.
- *Dashboarding*: → Herramientas para crear paneles de control interactivos para visualizar y monitorear los datos.

JUSTIFICA LAS TECNOLOGÍAS, SERVICIOS, HERRAMIENTAS Y SOFTWARE ELEGIDOS PARA LA REALIZACIÓN DEL ACCESO A DATOS.

La elección de tecnologías, servicios, herramientas y software para el acceso a datos depende de muchos factores, como los requisitos de capacidad de procesamiento, la complejidad de los datos, las necesidades de seguridad y privacidad, el presupuesto, etc.

Algunas justificaciones para elegir ciertas tecnologías incluyen:

- *Almacenamiento de datos* → Un sistema de almacenamiento de datos robusto es necesario para garantizar la disponibilidad, escalabilidad y rendimiento de los datos.
- *Procesamiento de datos en tiempo real* → Si se requiere una respuesta rápida y precisa a las consultas de los datos, es necesario elegir un sistema de procesamiento de datos en tiempo real.
- *Integración de herramientas* → La integración de herramientas de análisis de datos diferentes permite una visión completa y holística de los datos.
- *Aprendizaje automático* → El aprendizaje automático permite identificar patrones y tendencias en los datos de manera más eficiente que con técnicas manuales.
- *Seguridad de datos* → La selección de tecnologías y servicios que cumplen con los estándares de seguridad y privacidad de los datos es esencial para proteger la confidencialidad de la información.

En conclusión, la elección de tecnologías, servicios, herramientas y software para el acceso a datos depende de los requisitos y objetivos específicos del proyecto y de la organización. Es importante evaluar cuidadosamente cada opción antes de tomar una decisión.