Brainstorm

El objetivo del proyecto es diseñar, especificar y desplegar un Data Lake para procesar datos no estructurados extraídos mediante técnicas de scraping/crawling de sitios web de dominio público, específicamente datos relacionados con **zapatillas de running.**

Diseño del DAaaS

Definición la estrategia del DAaaS

Definir el catálogo de servicios que proporcionará la plataforma DAaaS, que incluye incorporación de datos, limpieza de datos, transformación de datos, datapedias, bibliotecas de herramientas analíticas y otros.

Definición de objetivos

El objetivo principal de este proyecto es crear un comparador de zapatillas de running que ayude a los usuarios a elegir el modelo que mejor se adapte a sus necesidades.

Anexo: Pasos seguidos en esta etapa

1. **Comprender los requisitos y objetivos**
   - ➢ *¿Qué tipo de datos no estructurados se recopilarán y de dónde?*
   - ➢ *¿Cuáles son los objetivos de este proyecto?*
   - ➢ *¿Cuáles son las fuentes, como sitios web o aplicaciones, que proporcionarán estos datos?*
   - ➢ *¿Qué tipo de análisis o procesamiento se realizará en estos datos?*
   - ➢ *¿Cuántos datos se esperan procesar y almacenar?*
   - ➢ *¿Cuántos usuarios finales se espera que utilicen el DAaaS?*
   - ➢ *¿Necesito cumplir con regulaciones de privacidad de datos?*
   - ➢ *¿Tengo permiso para extraer datos de las fuentes identificadas?*

2. **Crear un Documento Descriptivo**
   - ➢ En el documento Memoria Descriptiva (memoria-descriptiva.md) se describen los procesos de diseño del DAaaS.
   - ➢ Este documento explica cómo se planificará y se diseñará el Data Lake.

Arquitectura DAaaS

Definir la selección de componentes, la definición de procesos de ingeniería y el diseño de interfaces de usuario. Diseño y ejecución de Proofs-of-Concept (PoC) para demostrar la viabilidad del enfoque DAaaS.

Etapas:

- **Identificación de fuentes de datos:** Identificar los sitios web de dominio público que serán tus fuentes de datos. Debo considerar la legalidad y los términos de servicio de cada sitio.
    - Se identificaron los siguientes sitios web de dominio público como fuentes de datos:
        - Brooks
        - Asics
        - Nike
        - Runnea
    - Se consideraron los siguientes factores al identificar las fuentes de datos:
        - Legalidad y términos de servicio de cada sitio.
        - Disponibilidad de datos sobre zapatillas de running.
        - Calidad de los datos.
    - El objetivo es aumentar el número de sitios web de dominio público. En este proyecto, se toma una muestra pequeña de sitios web por motivos de tiempo.

- **Web Scraping/Crawling:** Implementar técnicas de web scraping/crawling para extraer datos de las fuentes identificadas. Esto podría incluir nombres de modelos, características, precios, calificaciones, comentarios de usuarios, etc.
    - Se utilizaron las siguientes técnicas de web scraping/crawling para extraer datos de las fuentes identificadas:
        - BeautifulSoup
        - Scrapy (no se obtuvieron datos de calidad)
    - Técnicas planteadas pendientes de utilizar:
        - API YouTube
        - Amazon
    - Se identificaron los siguientes retos y oportunidades en esta etapa:
        - La mayoría de los sitios web de zapatillas de running utilizan JavaScript para cargar los datos, lo que dificulta el scraping.
        - Los comentarios de YouTube o de Amazon pueden ser útiles para obtener información sobre la experiencia de los usuarios con las zapatillas de running.

- **Análisis de datos:** Limpiar y transformar los datos extraídos para prepararlos para el análisis.
    - Se realizaron los siguientes análisis:
        - Análisis descriptivo de los datos.
        - Análisis de correlaciones entre las características de las zapatillas de running.
        - *Pendiente: Análisis de sentimiento de los comentarios de YouTube.*

    - Se identificaron los siguientes retos y oportunidades en esta etapa:
        - La limpieza de los datos es un proceso laborioso y que requiere atención.
        - El análisis de las correlaciones entre las características de las zapatillas de running puede ser complejo.
        - *Pendiente: El análisis de sentimiento de los comentarios de YouTube puede ser subjetivo.*
    -
- **Desarrollo de PoC:** Diseñar y ejecutar Proofs-of-Concept (PoC) para demostrar la viabilidad del enfoque DAaaS.
    - Se diseñaron y ejecutaron dos PoC para demostrar la viabilidad del enfoque DAaaS:
        - PoC 1: Comparador de zapatillas de running basado en características.
        - PoC 2: Comparador de zapatillas de running basado en comentarios de usuarios.
    - Los PoC demostraron que el enfoque DAaaS es viable para crear un comparador de zapatillas de running.

Anexo:

➢ **Datasets (carpeta datasets):** Se revisaron los siguientes datasets y enlaces de interés, pero finalmente se decidió no hacer uso de ellos por dos motivos:

➢ Son de uso educativo y no cumplen con los requisitos del proyecto.

➢ Los datos no incluyen zapatillas de running, sino de otros deportes.

- ***1-shoe-prices-dataset.csv***
    - https://www.kaggle.com/datasets/rkiattisak/shoe-prices-dataset
    - https://www.kaggle.com/code/chloe912/shoe-prices-eda
    - Este conjunto de datos contiene información sobre las ventas de zapatos en una región en particular. Los datos incluyen información sobre la marca, modelo, tipo de zapato, sexo, talla, color, material y precio.
    - *El propósito de crear este conjunto de datos es únicamente para uso educativo, y cualquier uso comercial está estrictamente prohibido y este conjunto de datos se generó en grandes modelos de lenguaje y no se recopiló de fuentes de datos reales.*
    - Columnas: Brand, Model, Type, Gender, Size, Color, Material, Price

- ***2-BrooksShoes.csv***
    - https://www.kaggle.com/datasets/hannahcollins/2020-brooks-running-shoes
    - Este conjunto de datos incluye 26 modelos actuales de zapatillas Brooks Running adquiridos en Brooks Running
    - Columnas:
        - Name of shoe
        - Type (Men's/Women's)
        - Price (as of August 29, 2020)
        - Support (Neutral, Support, Max Support)
        - Experience (Speed, Cushion, Energize, Connect)
        - Surface (Road, Trail)
        - Midsole drop in mm
        - Weight in g
        - Arch type (High, Medium, Flat)
        - Additional shoe features (Segmented Crash Pad, DNA LOFT, BioMoGo, 3D Fit Print, DNA AMP, GuideRails, DNA Midsole, Ballistic Rock Shield, Gore-Tex)

- https://github.com/yassine-youcefi/web-scraping-nike-website
  - No contiene zapatillas de running, solo zapatillas de *football* y *basketball*
- https://rapidapi.com/blog/directory/nike-plus/
- https://adidas.gitbook.io/api-guidelines/
- https://www.blog.datahut.co/post/competitive-analysis-nike-vs-adidas

➢ **Web Scraping/Crawling**
- **Crawler de las siguientes páginas** *(BeautifulSoup)*
  - Brooks - 2023_web_scraping_brooks_data.csv
  - Asics - 2023_web_scraping_asics_data.csv
  - Nike - 2023_web_scraping_nike_data.csv
  - Runnea - webscrapingrunnea.csv
- **Scrapper**
  - Comentarios de Youtube → API YouTube (pendiente)

Componentes de la arquitectura DAaaS

- **Hadoop** para el procesamiento y análisis de datos. Hadoop Distributed File System (HDFS) y Apache MapReduce se utilizan para almacenar y procesar datos estáticos, como información de productos y precios.
- **Google Cloud Storage** para almacenar los archivos de datos y los datos recopilados por crawlers y scrapers. Google Cloud Storage es una solución escalable y segura para el almacenamiento de datos en la nube.
- **Servidor web** para alojar la aplicación web. Se puede utilizar un servidor web como Apache o un servicio administrado en la nube como Google App Engine o Google Compute Engine.
- **API de Python** para interactuar con los datos, realizar análisis y proporcionar resultados a través de endpoints de API. La API debe poder acceder a los datos almacenados en Google Cloud Storage, procesarlos y devolver los resultados necesarios para la aplicación web.
- **Cliente en ReactJS** para desarrollar la interfaz de usuario de la aplicación web. ReactJS es una buena opción para crear componentes que consuman los datos de la API y muestren la información de las zapatillas, incluyendo la capacidad de compararlas.
- **Base de datos Google Cloud SQL** para almacenar datos estructurados, información de usuarios y otros datos relacionados con la aplicación.
- **Cloud Functions** para ejecutar crawlers y scrapers.

DAaaS Operating Model Design and Rollout

Personalizar los modelos operativos DAaaS para cumplir con los procesos, la estructura organizacional, las reglas y el gobierno de los clientes individuales. Realizar seguimiento de consumo y mecanismos de informe.

Proceso de desarrollo de un comparador de zapatillas de running

1. Recopilación de datos

- Implementar técnicas de web scraping para extraer datos de las fuentes identificadas.
- Crear y configurar un Google Cloud Project con un bucket de Cloud Storage.
- Almacenar los datos extraídos en el Data Lake de Google Cloud Storage.

2. Procesamiento de datos

- Limpiar y transformar los datos no estructurados en una estructura útil para el comparador de zapatillas.
- Procesar los datos con Hadoop.
- Almacenar los datos estructurados en una base de datos SQL.

3. Desarrollo de la aplicación web

- Desarrollar una interfaz de usuario para que los usuarios ingresen sus preferencias y vean los resultados de la comparación.
- Integrar los datos procesados en la aplicación web.

4. Pruebas y validación

Realizar pruebas exhaustivas para asegurarse de que el comparador funcione correctamente y ofrezca resultados precisos.

5. Despliegue

Desplegar la aplicación en un servidor web para que los usuarios puedan acceder a ella.

6. Mantenimiento y actualización

Mantener el comparador y actualizar regularmente los datos extraídos para tener la información actualizada.

7. Cumplimiento legal y ético

Documentar el cumplimiento legal y ético, explicando cómo se cumplirá con las leyes de privacidad de datos y cómo se garantizará la legalidad de la extracción de datos de sitios de dominio público.

Desarrollo de la plataforma DAaaS. (ligera descripción del desarrollo)

*Construcción iterativa de todas las capacidades de la plataforma, incluido el diseño, desarrollo e integración, **pruebas**, carga de datos, metadatos y población de catálogos, y despliegue.*

El propósito fundamental de la plataforma DAaaS es brindar a los usuarios una forma sencilla de acceder a datos analíticos. Este enfoque se basa en un proceso iterativo que permite incorporar retroalimentación de los usuarios y adaptarse a nuevas necesidades a medida que surgen.

El desarrollo de la plataforma se divide en las siguientes etapas:

1. **Identificación de necesidades:** Involucra la realización de entrevistas con usuarios potenciales para identificar sus necesidades y requisitos.
2. **Diseño:** En esta etapa, se crea el diseño de la plataforma para satisfacer las necesidades identificadas.
3. **Desarrollo e integración:** Se lleva a cabo el desarrollo de los componentes de la plataforma y su posterior integración.
4. **Pruebas:** Se realizan pruebas exhaustivas para garantizar el correcto funcionamiento de la plataforma.
5. **Carga de datos:** Implica la carga de datos de prueba y de producción en la plataforma.
6. **Metadatos y población de catálogos:** En esta etapa se crean metadatos para los datos y se poblaron los catálogos de la plataforma.
7. **Despliegue:** Finalmente, la plataforma se despliega en un entorno de producción.

Durante el desarrollo de la plataforma, se enfrentan los siguientes desafíos:

- **Disponibilidad de datos de calidad:** La identificación y recopilación de datos de alta calidad son fundamentales para alimentar la plataforma.
- **Complejidad de la integración de sistemas:** Dado que la plataforma se basa en sistemas heterogéneos, se requiere un enfoque de integración complejo.
- **Necesidad de escalabilidad:** La plataforma se diseña para ser escalable y satisfacer las crecientes necesidades de los usuarios.

Estos desafíos se piensan abordar mediante las siguientes estrategias:

- **Creación de un proceso de recopilación de datos:** Se establecerá un proceso de recopilación de datos que garantice la calidad de la información.
- **Utilización de herramientas de integración:** Se emplearán herramientas de integración para simplificar el proceso de integración de sistemas.
- **Diseño de una arquitectura escalable:** La plataforma se desarrolla con una arquitectura escalable que permitirá su expansión.

Objetivos futuros:

La plataforma DAaaS se ha lanzado con éxito en producción y se encuentra en uso por una variedad de usuarios. Ha recibido comentarios positivos por su facilidad de uso y su capacidad para proporcionar información valiosa.

Link a Diagrama:

Especificar el Flujo de Datos:

Tras definir la estrategia del DAaaS (comprender los requisitos y objetivos del proyecto y crear un documento descriptivo), se diseña el Diagrama (app.diagrams.net) que representa el flujo de datos y las herramientas utilizadas. Esto incluye flechas que muestran la dirección de los datos, así como los procesos de limpieza y transformación.