Descripción General de la Arquitectura

Visión general

Mediante este proyecto se realiza una simulación de un entorno de Big Data utilizando herramientas más ligeras que pueden representar el flujo que se daría en estos entornos.

Las fases realizadas en el proyecto fueron:

Ingesta de datos, para este punto se realizo una extracción de datos desde la Api Disney, creando la base de datos, archivo .CSV y .txt para auditoria, para tener un registro sobre los datos extraídos.

Preprocesamiento de los datos, teniendo en cuenta que no es una buena practica tomar una base de datos transaccional para realizar limpieza de datos se crea un dataset para examinar las inconsistencias, valores nulos y duplicados, para finalmente contar con un .CSV con los datos limpios, adicional se crea un .txt de auditoría para el detalle de las transformaciones aplicadas.

En la etapa de enriquecimiento se aplicaron transformaciones como nuevas columnas, que aportan un valor a los datos posteriores.

Para simular la orquestación en la nube, se utiliza GitHub Actions para simular la automatización del flujo de trabajo en cada una de sus etapas.

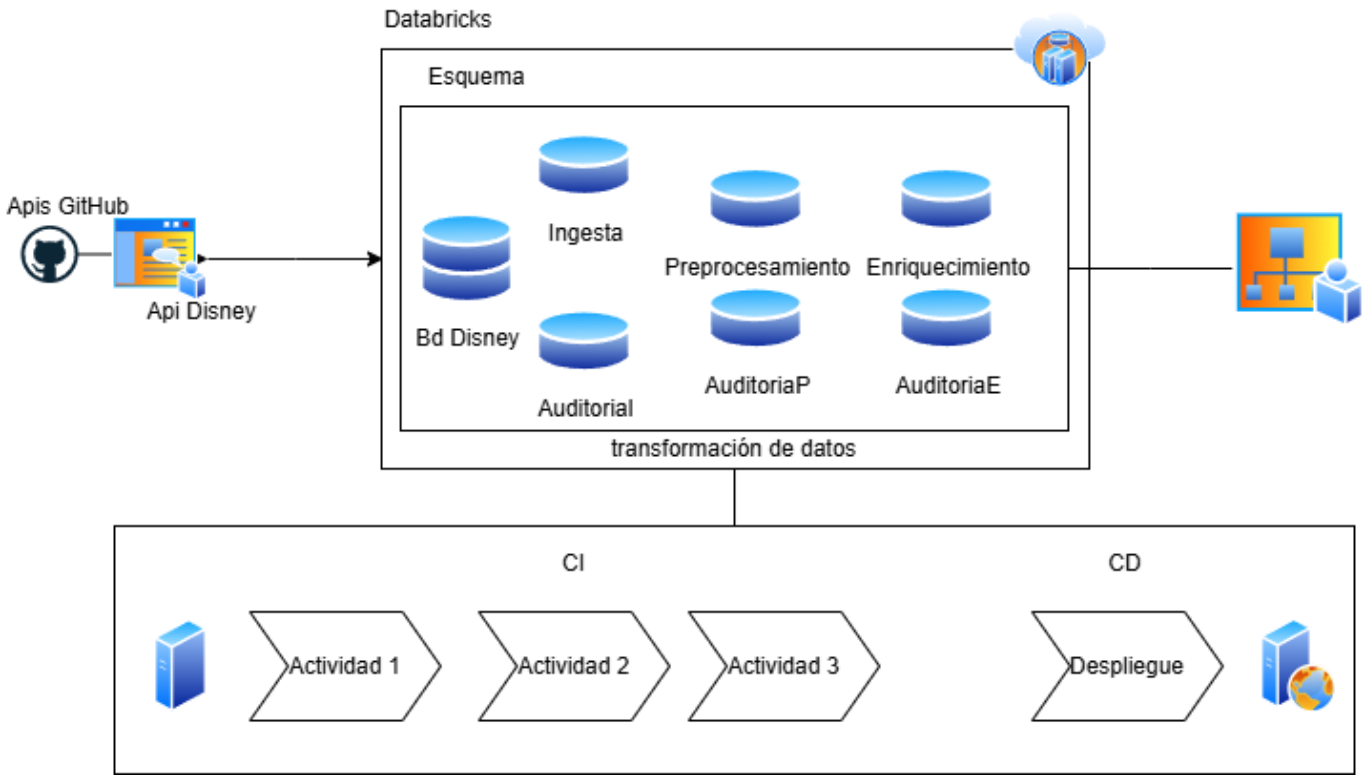
Componentes principales

Esta actividad esta apoyada por varios componentes claves los cuales permiten la simulación del procesamiento de datos en un entorno Big Data.

- **Base de datos analítica (SQLite):**
  Se utiliza SQLite como sistema de almacenamiento ya que es ligero, portátil y efectivo para este escenario de prueba. Permite realizar consultas sobre los datos para luego facilitar su análisis.
- **Scripts de procesamiento (Python):**
  Ingesta: Realiza la conexión al api creando la BD, archivo CSV inicial y txt para su respectiva auditoria.
  Preprocesamiento: Se analiza el dataset para encontrar inconsistencias sobre los datos obtenidos, para generar una versión limpia de estos con su respectivo informe limpieza.
  Enriquecimiento: Se agregan nuevos datos de la fuente original para que su análisis sea mejor con respecto a su calidad.

- **Automatización (GitHub Actions):**
  Se utiliza como orquestador de datos para simular un despliegue en un entorno BigData, enriqueciéndolo en cada etapa para lograr su correcta ejecución.
  Su configuración es para la ejecución automática sobre los Scripts definidos para que ante eventos como un Commit o actualización se realice sin la necesidad de intervenir manualmente.

Diagrama de la arquitectura



Modelo de datos

El modelo de datos para esta integración se compone de tres tablas:

Personajes, Información_personajes y Personajes_Disney

Tabla 1: Personajes

Esta tabla contiene los datos limpios recolectado mediante la primera y segunda fase de la integración

Campo	Tipo de Dato	Descripción
Id	Int	Identificador Personaje.
Name	Text	Nombre del personaje.
Films	Text	Películas en las que aparece el personaje.
TvShows	Text	Programas de televisión donde aparece el personaje.

Tabla 2: Información_personajes

Esta tabla tiene como Id principal el de la Tabla 1 “Personajes” que hace referencia al personaje para luego realizar el enriquecimiento a la información inicial de este personaje.

Campo	Tipo de Dato	Descripción
Id	Int	Identificador Personaje.
createAt	Text	Fecha de creación del registro.
updateAt	Text	Fecha de la última actualización.
url	Text	URL con más información del personaje.

Tabla 3: Personajes_Disney

Esta tabla es el resultado de la integración de la tabla 1 y 2 , consolidando todos los datos importantes para posteriores análisis.

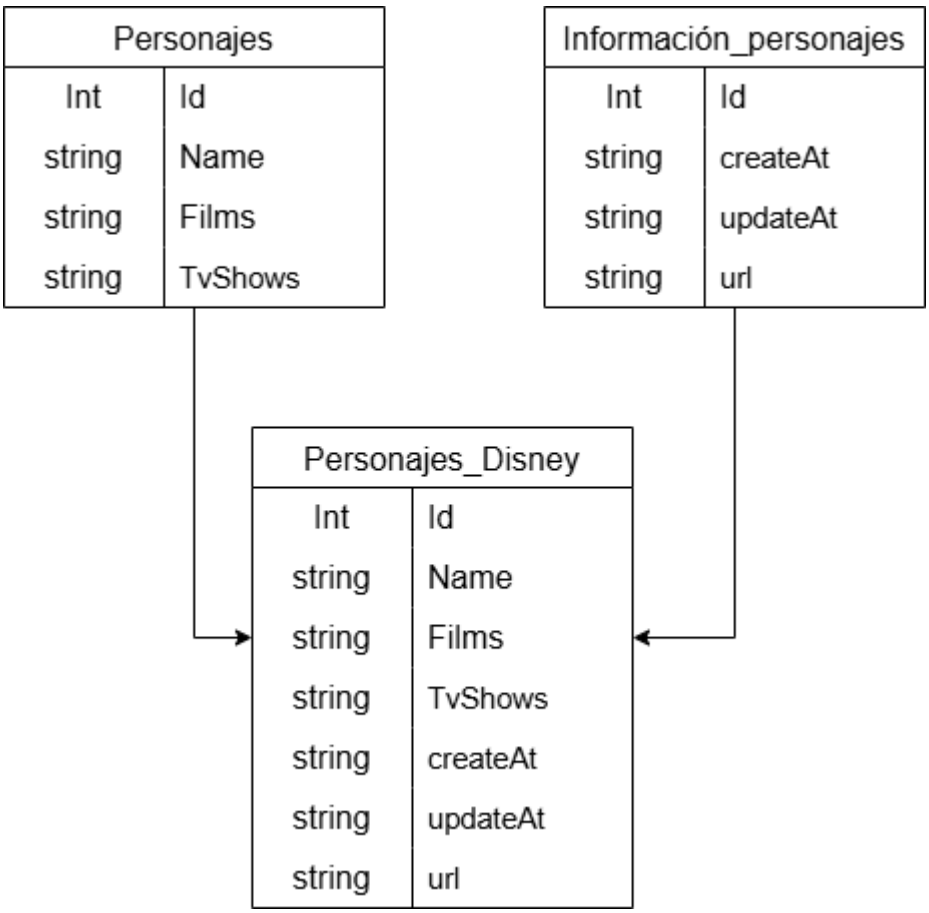
Campo	Tipo de Dato	Descripción
Id	Int	Identificador Personaje.
Name	Text	Nombre del personaje.
Films	Text	Películas en las que aparece el personaje.
TvShows	Text	Programas de televisión donde aparece el personaje.
createAt	Text	Fecha de creación del registro.
updateAt	Text	Fecha de la última actualización.
url	Text	URL con más información del personaje.

Justificación

Este modelo de datos se pensó para facilitar la limpieza de datos y el enriquecimiento de los datos de una manera más flexible y teniendo una calidad más precisa de cada parte.

En caso de requerir más datos a futuro solo se extendería la tabla 2 para luego realizar la ingesta en la tabla 3, sin necesidad de afectar los datos iniciales.

La tabla 3 quedaría como una tabla consolidad la cual se puede realizar uso para análisis, visualizaciones y exploraciones. Teniendo una buena estructura para entornos como pandas o herramientas como Power Bi.



Justificación de Herramientas y Tecnologías

SQLite:

Esta herramienta se eligio debido a que es una base de datos ligera, portable y fácil de integrar con desarrollos locales o para simulaciones de entornos de Big data.

No se requiere una instalación adicional de servidores, facilita consultas SQL sobre archivos acelerando las pruebas y desarrollo.

Pandas:

Pandas es una librería esencial para manipulación y análisis de datos en Python, la cual facilita la transformación de datos.

Es ideal para realizar análisis exploratorio, es intuitivo y se puede tener una escalabilidad para pasar a otras soluciones como lo es PySpark.

GitHub Actions:

Al GitHub ser el repositorio donde guardamos nuestro desarrollo, se aprovecha su herramienta Actions para la integración continua, realizando las ejecuciones al código bajo cada comit asegurándonos de un correcto funcionamiento y promoviendo buenas prácticas.

Simulación del Entorno Cloud

Base de Datos Integrada (Actividad 1 y 2)

Se genera una base de datos SQLite (disney.db) con tres tablas: personajes, información_personajes y personajes_disney.

Representa una base de datos almacenada en un entorno cloud accesible por múltiples servicios.

Scripts de procesamiento (Actividad 3)

Python y pandas procesan los datos para limpieza, integración y enriquecimiento.

Estos scripts simulan pipelines de datos que correrían en servicios tipo AWS Glue, Databricks o Google Dataflow.

GitHub Actions como entorno automatizado

Se configura un flujo (.yml) que simula un entorno serverless, ejecutando automáticamente tareas como:

Cargar datos desde JSON o APIs.

Procesar con Pandas/PySpark.

Guardar resultados en SQLite.

Emula la ejecución en la nube con cada push o pull request.

Conclusiones

Esta actividad permitió consolidar información de personajes del universo de Disney de manera eficiente y estructurada, esto gracias al uso de herramientas como SQLite, Pandas y GitHub Actions fue posible simular un entorno de procesamiento en la nube para automatizar flujos y garantizar los resultados.