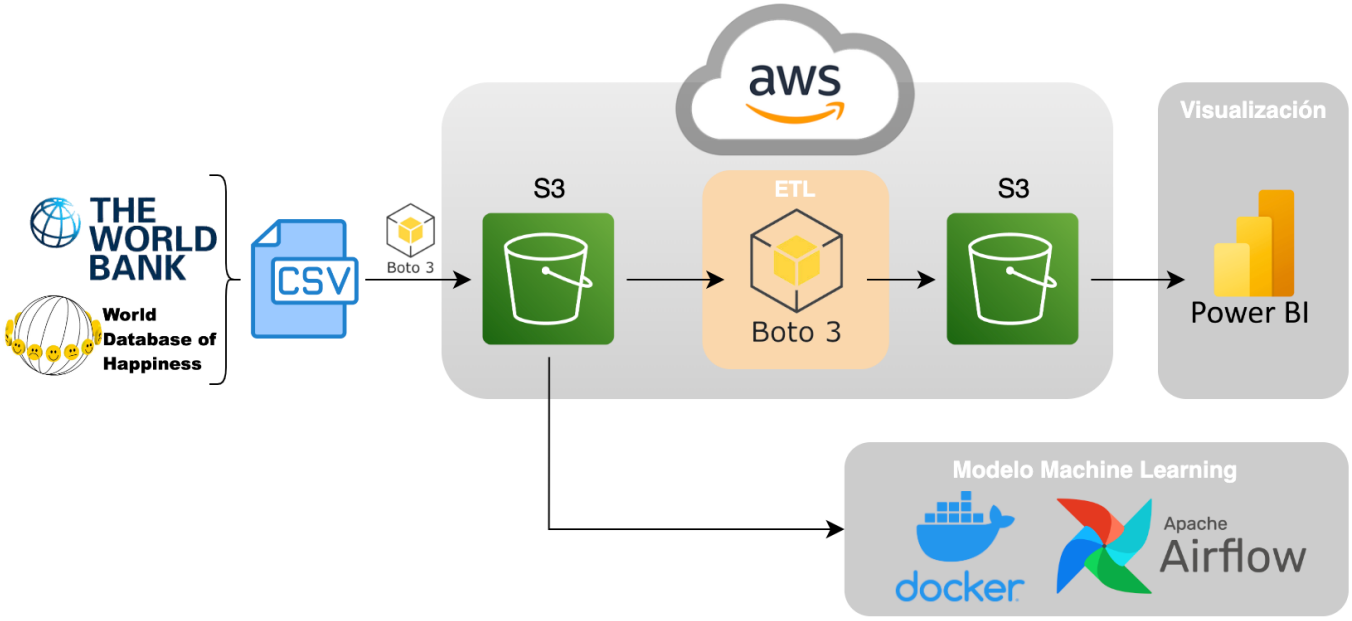
Semana 2

Ingeniería de Datos

Flujo de trabajo (workflow)



El workflow de nuestro proyecto fue el siguiente:

1. **Obtención de los datos crudos en formato CSV:** En primer lugar, se obtienen los datos en bruto de su fuente original, en formato CSV. Estos datos fueron obtenidos de diferentes fuentes (The World Bank, World Database of Happiness), y los proporcionados por Henry como bases de datos, archivos locales o servicios web.
2. **Transferencia de los datos a S3 utilizando Boto3:** Utilizando la biblioteca Boto3 de Python, se carga el archivo CSV crudo a un bucket de Amazon S3. Se aseguró que la autenticación sea adecuada y que la política de acceso esté configurada correctamente para que solo los usuarios autorizados puedan acceder a los datos almacenados en S3.
3. **Transformación de los datos mediante ETL automatizado en Boto3:** Una vez que los datos están almacenados en S3, se transformaron para ser utilizados en el análisis de datos y en la creación de modelos de machine learning. Esto se puede lograr mediante ETL automatizado utilizando Boto3 y otras bibliotecas de Python y Pandas. La transformación incluirá la limpieza y normalización de los datos para que puedan ser utilizados en diferentes plataformas de análisis y modelo ML.
4. **Almacenamiento de los datos transformados en S3:** Los datos transformados se almacenaron en S3 en un bucket separado y seguro que esté protegido por la autenticación y la política de acceso adecuadas. También es importante tener en cuenta la gestión de versiones y la retención de datos para asegurar que los datos estén disponibles para el análisis y la modelización en el futuro.

5. **Análisis de los datos mediante Power BI:** Los datos transformados se van a analizar utilizando herramientas de visualización de datos, como Power BI. Power BI se conecta directamente a los datos almacenados en S3 y permite la creación de gráficos, tablas y otras visualizaciones para identificar patrones y tendencias en los datos.
6. **Creación de modelos de machine learning utilizando Docker y Apache Airflow:** Finalmente, se pueden utilizar las bibliotecas de Python para crear modelos de machine learning para predecir el comportamiento futuro y tomar decisiones informadas. Para el modelo a construir vamos a utilizar Docker y Apache Airflow, que permiten la orquestación de la infraestructura y la automatización de los flujos de trabajo para la creación y entrenamiento de modelos de machine learning.

Este workflow nos proporciona una base sólida para la migración de datos y la creación de modelos de machine learning utilizando tecnologías modernas y seguras, lo que garantiza que los datos estén disponibles y sean útiles para la toma de decisiones informadas en el futuro.

Stack tecnológico

Las tecnologías elegidas para el desarrollo de este proyecto fueron:

- ***Python:*** lenguaje de programación, utilizado para desarrollar aplicaciones de todo tipo. Es un lenguaje interpretado, dinámico y multiplataforma. Es de código abierto y clasificado constantemente como uno de los lenguajes de programación más populares.

- ***Librerías de Python***
  - **Pandas:** librería escrita para el lenguaje Python que permite manipular y analizar datos. Ofrece estructuras de datos y operaciones para la manipulación de tablas numéricas y series temporales.
  - **Boto3:** librería de Python que proporciona una interfaz de programación de aplicaciones para interactuar con el servicio AWS de manera programática. Además nos permite automatizar tareas, administrar recursos y aprovechar la funcionalidad completa de AWS desde su aplicación.

- ***AWS S3(Simple Storage Service):*** servicio de almacenamiento de grandes volúmenes de datos, flexible, escalable, seguro y confiable. S3 proporciona almacenamiento de objetos organizados en buckets, donde cada objeto se identifica mediante una clave única asignada por el usuario. Este servicio lo utilizamos para almacenar archivos sin procesar (raw), archivos procesados para consulta y, que a su vez los archivos sin procesar sirvan de backup.

- ***Power BI:*** software desarrollado por Microsoft para la visualización de datos de manera interactiva, con enfoque principal en la inteligencia empresarial (*Business Intelligence)*

- ***Docker:*** plataforma de contenedores que permite empaquetar aplicaciones y sus dependencias en entornos aislados y portátiles. Su uso en el proyecto está orientado a

la disponibilización del modelo de machine Learning ya que proporciona portabilidad, reproducibilidad, escalabilidad y facilidad de mantenimiento y control de versiones. Además es una opción popular para empaquetar y distribuir modelos de maching Learning, permitiendo su ejecución de manera consistente en diferentes entornos.

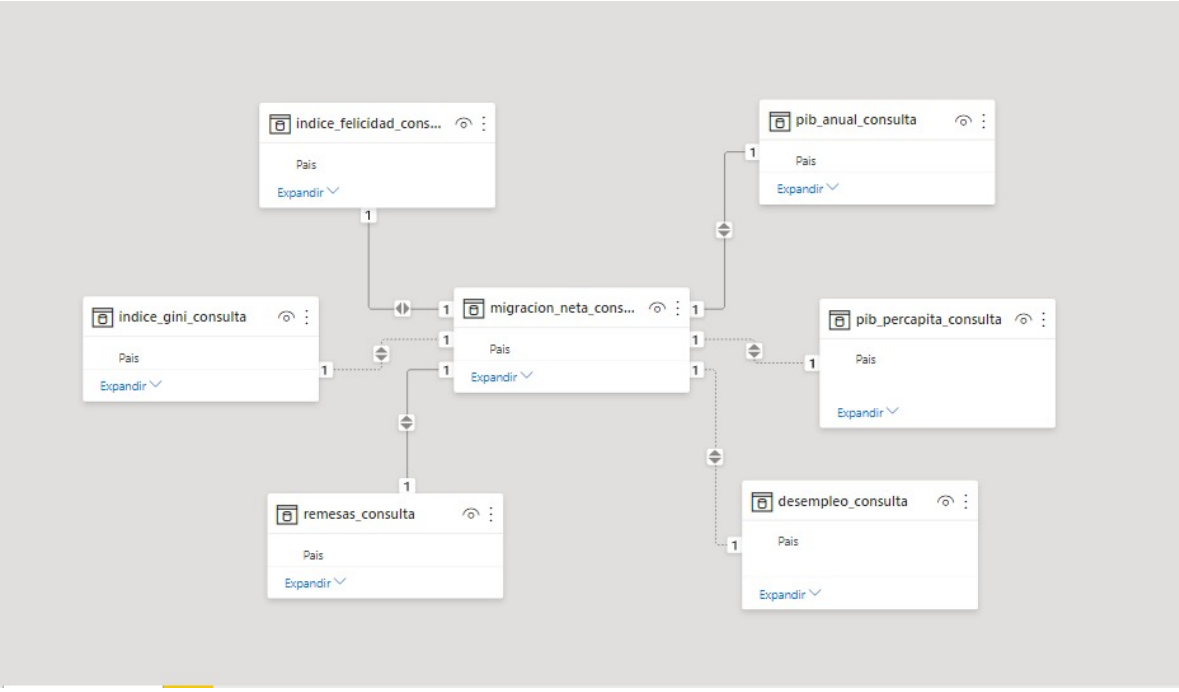
- ***Airflow:*** herramienta de flujo de trabajo y programación dirigida por datos, y si bien su enfoque principal es la orquestación de tareas, posee características útiles para la implementación y programación de modelos de machine Learning. También brinda una programación flexible lo que permite programar tareas según una programación determinada o desencadenarlas en función de eventos o condiciones específicas, lo que ayudará a gestionar y automatizar el flujo de trabajo necesario para disponibilizar el modelo en un entrono de producción.

Arquitectura de datos

- Carga de datos: el código utilizado para la tarea de cargar datos presenta el siguiente orden de ejecución:
  - Importar librerías necesarias para las tareas de carga y transformación de datos
  - Extracción de datos desde archivos CSV mediante la librería pandas. Utilizando funciones como 'read_csv()' para leer los archivos y convertirlos en dataframes.
- Extracción de datos:
  El origen de los datos proviene de 8 archivos CSV, los que han sido extraídos previamente de sitios web de datos mundiales (Banco Mundial, World Happiness Record).
- Ingesta de datos: utilizando la librería Boto3 de Python, los datos extraídos se carga en el servicio de almacenamiento AWS S3. Esta interfaz nos permite interactuar con el servicio de AWS, permitiéndonos cargar datos en S3 de forma programática.
- Proceso de Transformación: durante la etapa de transformación se aplicaron diversas acciones para preparar los datos para su uso, las que incluyeron:
  - Revisar y manejar los valores nulos y duplicados
  - Filtrar columnas por nombre de país y años de interés (2000-2021).
  - Eliminación de columnas que no contengan valores relevantes para el análisis
  - Filtrar filas por nombre de países de interés (Estados Unidos, México, Guatemala, Honduras, El Salvador, Colombia, Chile, Argentina, Brasil, Uruguay)
- Carga de datos transformados: una vez los datos han sido transformados según el paso anterior, se exportan a un archivo CSV para su posterior carga en AWS S3 para su almacenamiento en el bucket de consulta.

Esta arquitectura de datos implica la carga de datos desde archivos CSV, su transformación mediante operaciones de limpieza y filtrado, para finalizar en una carga de los datos transformados a AWS S3.

Modelo ER



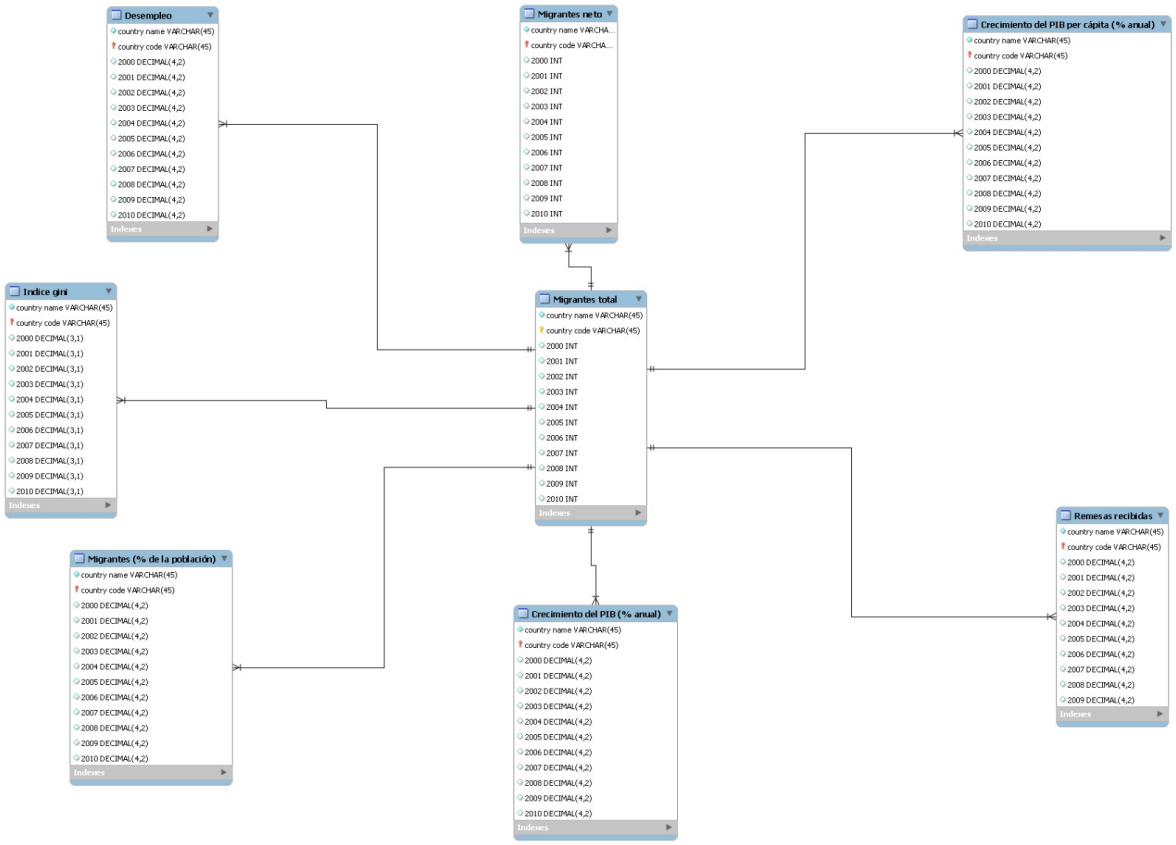
Nuestro modelo entidad relación es en forma de estrella, el cual se organiza en torno a una tabla central de hechos con los campos de Country Name (nombre del país) y Años comprendidos entre 2000 y 2021. Esta tabla de hechos se relaciona con las tablas dimensionales mediante la clave foránea. En este caso nuestra Primary Key (PK) seleccionada es el Nombre del país.

Cada tabla dimensional del modelo contiene medidas numéricas de los siguientes indicadores:

- Migración neta
- Volumen internacional de migrantes (total y % de la población)
- Crecimiento del Producto Interno Bruto (% anual y % anual percápita)
- Desempleo total
- Remesas
- Índice de Gini
- Índice de Felicidad

Esta estructura de diseño de base de datos nos proporciona un enfoque simple, eficiente y comprensible para el análisis de datos.

Diagrama entidad relación



Diccionario de datos

La base de datos consta de 8 archivos csv. Cada archivo contiene una tabla con el registro del nombre de los países, su código de país y un indicador respectivo de cada país en años comprendidos desde 2000 a 2021.

	Hecho	Tipo de dato	Descripción
PK	Country Name	Carácter	Nombre del País, sin valores nulos
	Años (2000-2021)	Int	Años comprendidos desde el año 2000 a 2021

Indicador	Tipo de dato	Descripción
Migración neta	float	La migración neta es el total neto de personas que migraron durante el período: la cantidad total de inmigrantes menos la cantidad anual de emigrantes, incluidos los ciudadanos y los no ciudadanos.

Volumen internacional de migrantes (% de la población)	float	Cantidad de personas nacidas en un país en el que no viven. Incluye refugiados. Representado como porcentaje de la población.
Volumen internacional de migrantes, total	float	Cantidad de personas nacidas en un país en el que no viven. Incluye refugiados.
Crecimiento del PIB (% anual)	float	Tasa de crecimiento anual porcentual del Producto Interno Bruto (PIB) a precios de mercado en moneda local, a precios constantes. Los agregados están expresados en USD a precios constantes del año 2010. El PIB es la suma del valor agregado bruto de todos los productores residentes en la economía más todo impuesto a los productos, menos todo subsidio no incluido en el valor de los productos.
Crecimiento del PIB per cápita (% anual)	float	Tasa de crecimiento anual porcentual del Producto Interno Bruto (PIB)per cápita en moneda local, a precios constantes. El PIB per cápita es el producto interno bruto dividido por la población a mitad de año.
Desempleo total (% de la población activa total)	float	Porcentaje de la población activa que no tiene trabajo pero que busca o está disponible para realizar un trabajo.
Remesas de trabajadores (% del PIB)	float	Indicador económico de los fondos que trabajadores migrantes que residen en el extranjero, envían a una persona o familia en su país de origen. Puede incluir envíos de efectivo, transferencias bancarias u otros medios de envío de dinero. Valor expresado como proporción del PIB.
Índice de Gini	float	El índice de Gini mide hasta qué punto la distribución del ingreso entre individuos u hogares dentro de una economía se aleja de una distribución perfectamente equitativa. Un índice de Gini de 0 representa una equidad perfecta, mientras que un índice de 100 representa una inequidad perfecta.
Índice de felicidad	float	El Índice de Felicidad evalúa el nivel de felicidad y bienestar subjetivo de los países y sus habitantes. Es elaborado anualmente por World Happiness Report en base a variables como el PIB per cápita, el apoyo social, la esperanza de vida, la libertad de elección, la generosidad y corrupción percibida. Puntaje por país basado en encuesta a personas donde evalúan la calidad de vida actual en una escala de 0 a 10.

Presentación Demo 2