ETL para la carga de datasets de Turismo Internacional en Argentina

Docker PostgreSQL Apache Superset pgAdmin

**Descarga de Datasets**

Los datasets utilizados en este proyecto pueden descargarse desde el portal oficial de datos abiertos del gobierno de Argentina:

<https://datos.gob.ar/dataset/turismo-turismo-internacional---total-pais>

Este portal proporciona información pública en formatos reutilizables, incluyendo datos relacionados con el turismo internacional - Total país en Argentina.

**Resumen del Tutorial**

Este tutorial guía al usuario a través de los pasos necesarios para desplegar una infraestructura ETL utilizando Docker, PostgreSQL, Apache Superset y pgAdmin. Se incluyen instrucciones detalladas para:

* Levantar los servicios con Docker.
* Configurar la conexión a la base de datos en Apache Superset.
* Ejecutar consultas SQL para analizar los datos de turismo internacional - Total país en Argentina.
* Crear gráficos y tableros interactivos para la visualización de datos.
* Palabras Clave
* Docker
* PostgreSQL
* Apache Superset
* pgAdmin
* ETL
* Visualización de Datos
* Mantenido Por Grupo 8

**Descargo de Responsabilidad**

El código proporcionado se ofrece "tal cual", sin garantía de ningún tipo, expresa o implícita. En ningún caso los autores o titulares de derechos de autor serán responsables de cualquier reclamo, daño u otra responsabilidad.

**Descripción del Proyecto**

Este proyecto implementa un proceso ETL (Extract, Transform, Load) para la carga y análisis de datos relacionados con datos del turismo internacional - Total país en Argentina. Utiliza herramientas modernas como Docker, PostgreSQL, Apache Superset y pgAdmin para facilitar la gestión, análisis y visualización de datos.

Desarrollar una solución escalable y reproducible para analizar el turismo internacional - Total país en Argentina, permitiendo la exploración dinámica de datos por país de origen y país destino donde tenemos tipos de transportes, balanzas, períodos temporales y diferentes tipos de visitantes, con el fin de generar tableros interactivos y gráficos personalizados que apoyen la toma de decisiones en el sector turístico.

**Características Principales**

* *Infraestructura Contenerizada:* Uso de Docker para simplificar la configuración y despliegue.
* *Base de Datos Relacional:* PostgreSQL para almacenar y gestionar los datos.
* *Visualización de Datos:* Apache Superset para crear gráficos y tableros interactivos.
* *Gestión de Base de Datos:* pgAdmin para administrar y consultar la base de datos.

**Requisitos Previos**

* ***Antes de comenzar, asegúrate de tener instalados los siguientes componentes:***

Docker

Docker Compose

Navegador web para acceder a Apache Superset y pgAdmin.

* ***Servicios Definidos en Docker Compose:***

networks:

net:

external: false

volumes:

postgres-db:

external: false

services:

db:

image: postgres:alpine

env\_file:

- .env.db

restart: unless-stopped

environment:

- POSTGRES\_INITDB\_ARGS=--auth-host=md5 --auth-local=trust

healthcheck:

test: [ "CMD-SHELL", "pg\_isready" ]

interval: 10s

timeout: 2s

retries: 5

ports:

- 5432:5432

volumes:

- postgres-db:/var/lib/postgresql/data

- ./scripts:/docker-entrypoint-initdb.d

- ./DATOS:/DATOS

networks:

- net

superset:

image: apache/superset:4.0.0

restart: unless-stopped

env\_file:

- .env.db

ports:

- 8088:8088

depends\_on:

db:

condition: service\_healthy

networks:

- net

pgadmin:

image: dpage/pgadmin4

restart: unless-stopped

env\_file:

- .env.db

ports:

- 5050:80

depends\_on:

db:

condition: service\_healthy

networks:

- net

**Instrucciones de Configuración**

* ***Clonar el repositorio:***

git clone <URL\_DEL\_REPOSITORIO>

cd ETL-Turismo

* ***Configurar el archivo .env.db: Crea un archivo .env.db en la raíz del proyecto con las siguientes variables de entorno:***

#Definimos cada variable

DATABASE\_HOST=db

DATABASE\_PORT=5432

DATABASE\_NAME=postgres

DATABASE\_USER=postgres

DATABASE\_PASSWORD=postgres

POSTGRES\_INITDB\_ARGS="--auth-host=scram-sha-256 --auth-local=trust"

# Configuracion para inicializar postgres

POSTGRES\_PASSWORD=${DATABASE\_PASSWORD}

PGUSER=${DATABASE\_USER}

# Configuracion para inicializar pgadmin

PGADMIN\_DEFAULT\_EMAIL=postgres@postgresql.com

PGADMIN\_DEFAULT\_PASSWORD=${DATABASE\_PASSWORD}

# Configuracion para inicializar superset

SUPERSET\_SECRET\_KEY=your\_secret\_key\_here

* ***Levantar los servicios:***

Ejecuta los siguientes comandos para iniciar los contenedores:

docker compose up -d

. init.sh

* ***Acceso a las herramientas:***

Apache Superset: <http://localhost:8088/>

* ***Credenciales predeterminadas:*** admin/admin

pgAdmin: <http://localhost:5050/>

(Configura la conexión a PostgreSQL utilizando las credenciales definidas en el archivo .env.db.)

**Uso del Proyecto**

***1. Configuración de la Base de Datos***

Accede a Apache Superset y crea una conexión a la base de datos PostgreSQL en la sección Settings. Asegúrate de que la conexión sea exitosa antes de proceder.

***2. Consultas SQL***

**Consulta 1: “Top 5 países que más turistas enviaron a Argentina”**

SELECT p.nombre, SUM(tnr.viajes\_de\_turistas\_no\_residentes) AS total\_viajes

FROM turismo\_no\_residentes tnr

JOIN pais p ON tnr.id\_pais\_origen = p.id

GROUP BY p.nombre

ORDER BY total\_viajes DESC

LIMIT 5;

**Consulta 2: “Ranking de países con mayor diferencia entre residentes que los visitan y turistas que envían”**

SELECT

p.nombre,

COALESCE(SUM(tr.viajes\_de\_turistas\_residentes), 0) AS emisivo,

COALESCE(SUM(tnr.viajes\_de\_turistas\_no\_residentes), 0) AS receptivo,

COALESCE(SUM(tnr.viajes\_de\_turistas\_no\_residentes), 0) - COALESCE(SUM(tr.viajes\_de\_turistas\_residentes), 0) AS diferencia

FROM pais p

LEFT JOIN turismo\_residentes tr ON tr.id\_pais\_destino = p.id

LEFT JOIN turismo\_no\_residentes tnr ON tnr.id\_pais\_origen = p.id

GROUP BY p.nombre

ORDER BY diferencia DESC;

**Consulta 3: “Países con mayor déficit o superávit turístico”**

SELECT p.nombre, SUM(bt.balanza) AS saldo

FROM balanza\_turistica bt

JOIN pais p ON bt.id\_pais = p.id

GROUP BY p.nombre

ORDER BY saldo DESC;

**Consulta 4: “Distribución por medio de transporte utilizado para viajes al exterior”**

SELECT mt.nombre AS medio, SUM(viajes\_de\_turistas\_residentes) AS total

FROM turismo\_residentes tr

JOIN medio\_transporte mt ON tr.id\_medio\_de\_transporte = mt.id

GROUP BY mt.nombre

ORDER BY total DESC;

***3. Creación de Gráficos y Tableros***

Ejecuta las consultas en SQL Lab de Apache Superset.

Haz clic en el botón CREATE CHART para crear gráficos interactivos.

Configura el tipo de gráfico y las dimensiones necesarias.

Guarda el gráfico en un tablero con el botón SAVE.

**Estructura del Proyecto**

ETL-Turismo/

├── docker-compose.yml # Configuración de Docker Compose

├──scripts #Carpeta contenedora del script de la base de datos

├── init.sh # Script de inicialización de base de datos

├──env.db #Variables de entorno

├── DATOS/ # Carpeta para almacenar datasets

└── README.md # Documentación del proyecto