

Documentación Técnica proyecto Tablero Covid 2021

Descripción

Utilizando los datos provistos, los cuales contienen información sobre los contagios, las recuperaciones y las muertes, dados por país y región. Se solicita que usted construya un data pipeline que procese los 3 archivos csv, los inserte a una base de datos, y luego basado en los datos procesados, debe construir un dashboard que permite analizar las estadísticas de cada uno de los archivos.

Ingesta de Datos

Con los 3 archivos csv debe construir uno o varios DAGs utilizando Airflow que permite leer, transformar e insertar los archivos en una base de datos, la cual puede ser un MySQL o cualquier otra de DB de su preferencia. Para la construcción de los DAGs pueden utilizar cualquier tipo de task y operadores que considere apropiados, sólo debe tomar en cuenta que al momento de la calificación debe poder ejecutar el archivo y que este actualice la base de datos.

Dashboard

Una vez la data ha sido procesada, debe construir un Dashboard sobre la data, el cual debe contener por lo menos los siguientes criterios:

- Un mapa que muestre por código de colores, burbujas o como mejor considere la totalidad de los casos, recuperaciones y muertes para una fecha dada y un país determinado.
- Mostrar estadísticas de los incrementos de los casos, recuperaciones y muertes por varios países.
- Cualquier otro tipo de estadística o gráfica que considere necesario.
- Para el dashboard puede utilizar cualquier herramienta vista en clase, Shiny Apps, Stream List, etc.

Docker

Todo el proyecto debe correr utilizando docker, por lo que se espera que usted pueda crear en docker una instancia de airflow (no necesariamente distribuida) que ejecute los DAGs creados, la base de datos que decida utilizar, también debe estar corriendo en Docker y finalmente la herramienta que utilice para ejecutar el dashboard debiera también correr en Docker y debe utilizar la base de datos con la data ingresada desde airflow. Si la herramienta de dashboard utiliza los csv directamente, será penalizado con 0 en esta rúbrica.

Arquitectura

Módulos

Como se menciona dentro de la descripción del proyecto, este se construyó dentro de contenedores según muestra el siguiente diagrama.

Proyecto Tablero Covid - Arquitectura

AIRFLOW			db	streamlit
posgre DB	container Files	Apache web Server	MySql DB	Streamlit web server
docker				

Básicamente el proyecto consiste en 3 módulos: AirFlow, Una base de datos y frontend con Streamlit.

Dentro del módulo de AirFlow se consideran dos contenedores de docker, uno que contiene una base de datos, y otro que contiene al servidor web. Así mismo se utiliza el sistema de archivos del contenedor para alojar temporalmente los archivos que serán procesados e insertados en la base de datos.

En lo que respecta a las versiones, se utilizó postgres 9.6, y la última versión disponible del servidor web.

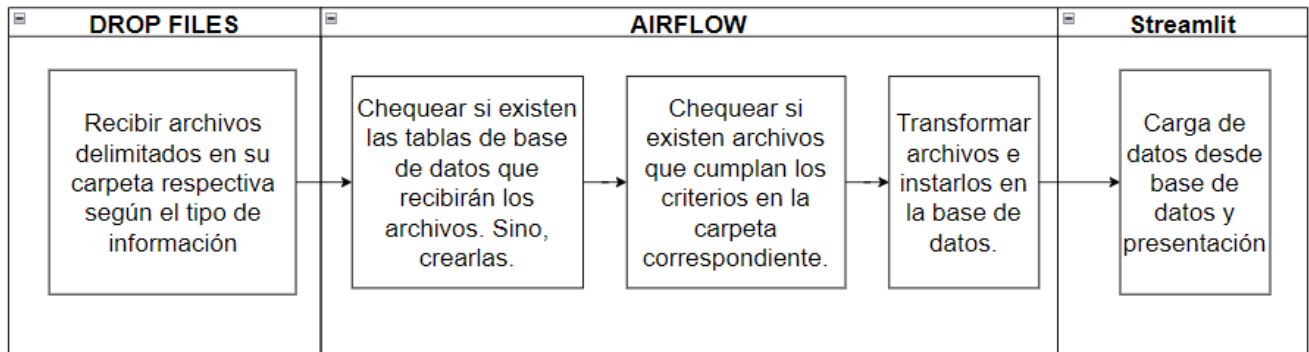
La base de datos funciona como un repositorio permanente de datos destinados a ser consumidos por la aplicación de Dashboard. En este módulo se utilizó MySql versión 5.7. Es importante mencionar que fue necesario eliminar los volúmenes anteriores para poder utilizar una imagen limpia de MySql.

Streamlit es la aplicación de Dashboard que se utilizó, misma que consume la información desde la base de datos. Esta se utilizó en su versión más reciente.

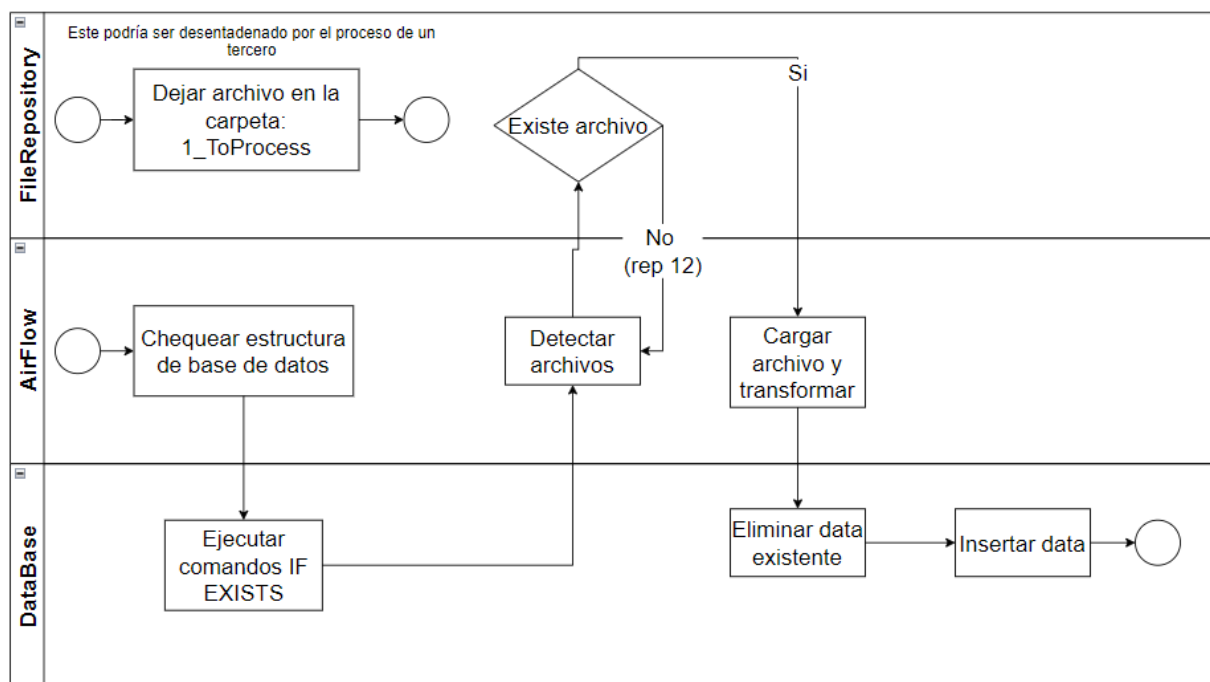
Procesos

A continuación se presenta el diagrama general del proceso.

Proyecto Tablero Covid - Procesos



Proyecto Tablero Covid - Ingesta de archivos



Desarrollo

Docker

Para el desarrollo del proyecto se utilizó como base el repositorio que fue compartido durante la clase, pero se hicieron los siguientes cambios:

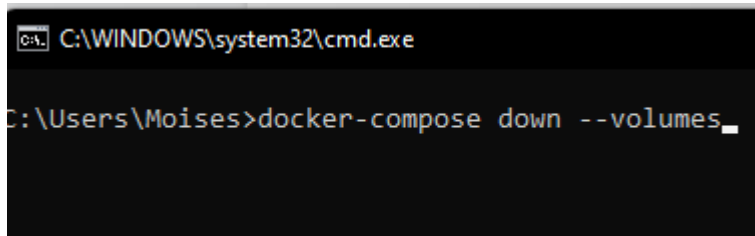
1. **Volúmenes contenedor AirFlow webserver**

Se agregó un volumen para almacenar los comandos de verificación de tablas en la base de datos.

2. **Archivo “schema.sql” para el contenedor db (MySQL 5.7)**

Se actualizó el esquema de base de datos para utilizar una base de datos con distinto nombre e incluir las tablas necesarias.

Es importante notar que fue necesario ejecutar el comando



```
C:\WINDOWS\system32\cmd.exe

C:\Users\Moises>docker-compose down --volumes_
```

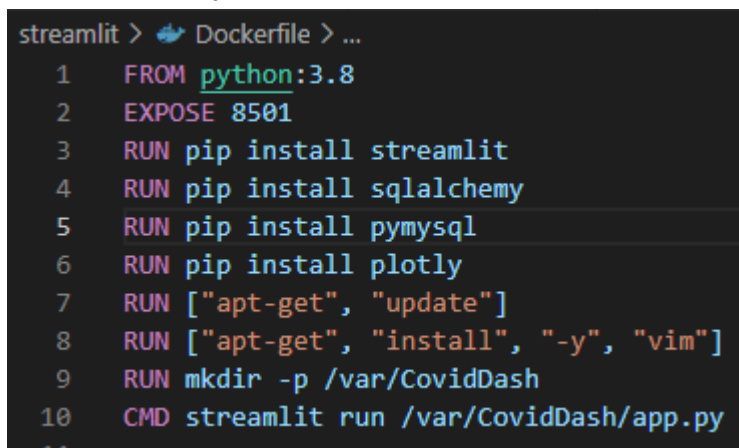
para poder utilizar una imagen “limpia” para este contenedor.

3. **Datos de conexión a base de datos**

Se actualizaron las credenciales para acceder a la base de datos.

4. **Servicio de streamlit**

Se agregó un Dockerfile, que tiene la instalación de streamlit junto con algunas dependencias necesarias para el dashboard. Finalizando con la ejecución del comando para ejecutar el servicio.



```
streamlit > Dockerfile > ...
1 FROM python:3.8
2 EXPOSE 8501
3 RUN pip install streamlit
4 RUN pip install sqlalchemy
5 RUN pip install pymysql
6 RUN pip install plotly
7 RUN ["apt-get", "update"]
8 RUN ["apt-get", "install", "-y", "vim"]
9 RUN mkdir -p /var/CovidDash
10 CMD streamlit run /var/CovidDash/app.py
11
```











Durante la instalación se utilizó vim para poder asegurarse que los archivos utilizados por el contenedor tenían el contenido correspondiente.

AirFlow DAGs

Conexiones

Las configuraciones necesarias para los operadores fueron básicamente la creación de las siguientes conexiones:

1. Conexiones hacia las carpetas que corresponden a la ingesta de archivos
2. Conexión hacia la carpeta que contiene los comandos de base de datos.

<input type="checkbox"/>	 	fs_ConfirmedFiles	fs
<input type="checkbox"/>	 	fs_DBSchema	fs
<input type="checkbox"/>	 	fs_DeathsFiles	fs
<input type="checkbox"/>	 	fs_default	fs
<input type="checkbox"/>	 	fs_RecoveredFiles	fs

3. Ajuste de conexión por defecto a MySQL

Funciones

Debido a lo similar de los archivos a procesar se definió una método genérico que parametriza la conexión AirFlow, el nombre del archivo, y el nombre de la tabla en la base de datos.

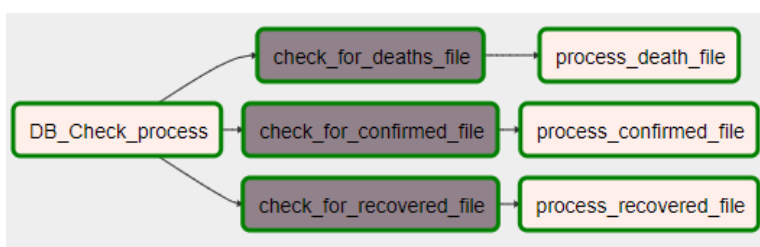
```
#-----TRANSFORM AND INSERT DATA INTO DATABASE-----  
> def ingest_file(connectionName, fileName, tableName):...
```

Posteriormente se crearon los operadores como llamadas a dicha función genérica.

Modelado Dag

Una vez desarrolladas las funciones se acomodaron los operadores de manera que el procesamiento de los tres tipos de archivo se ejecuten en paralelo.

```
189 dbCheckOperator >> sensorDeath >> operateDeathFile  
190 dbCheckOperator >> sensorConfirmed >> operateConfirmedFile  
191 dbCheckOperator >> sensorRecovered >> operateRecoveredFile
```



Dashboard

El dashboard generado consta de un panel de filtros y 4 secciones de presentacion de informacion

1. Casos de covid19 a nivel mundial
2. Casos por país
3. Mapa de casos
4. Detalla general de casos

Selecciones pais(es)

Choose an option

Seleccione de inicio

2020/01/22

Seleccione fecha de fin

2021/08/04

Covid19 - Analisis de datos

Casos de covid19 a nivel muncial

	Ubicacion	Casos Confirmados	Casos Recuperados	Casos Fallecidos
0	Global	260,069,295	137,546,993	5,180,005

Casos por pais

Detalle

Mapa de casos

Metricas por pais

Detalle general de casos

Detalles

Casos de covid19 a nivel mundial

Esta sección mostrará la información general de los casos reporta, recuperados y fallecidos a nivel mundial, aun cuando los filtros sean modificados esto no afectará la sección esta sección.

Casos de covid19 a nivel muncial

	Ubicacion	Casos Confirmados	Casos Recuperados	Casos Fallecidos
0	Global	260,069,295	137,546,993	5,180,005

Casos por país

Esta sección mostrará el resumen de los casos por país, al no tener ningún filtro aplicado mostrar la información en general.

Casos por país

Detalle				-	
	Casos_confirmados	Casos_recuperados	Casos_fallecidos		
Afghanistan	148935	82614	6836		
Albania	133310	130314	2457		
Algeria	176724	118523	4404		
Andorra	14797	14382	128		
Angola	43158	39889	1029		
Antigua and Barbuda	1313	1239	43		
Argentina	4975616	4615834	106747		
Armenia	230993	220438	4625		

Si nos vamos a la parte izquierda al panel de configuración podemos elegir qué países queremos mostrar y el rango de fecha a presentar, como se muestra en las siguientes imagen,

Selecciones país(es)

Guatemala X

Mexico X

El Salvador X

⊙ ▼

Seleccione de inicio

2020/06/01

Seleccione fecha de fin

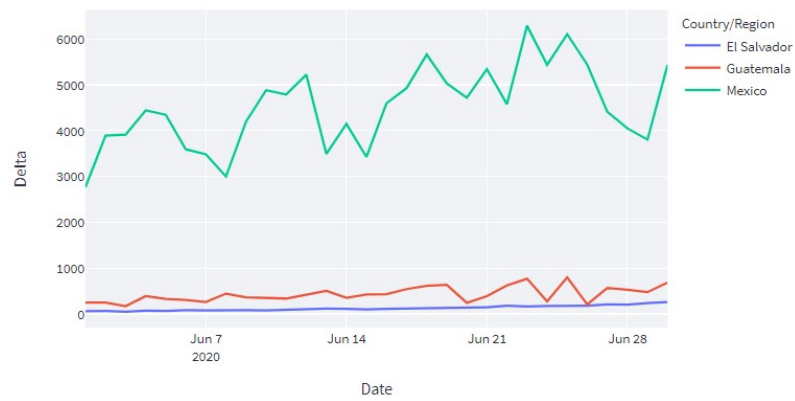
2020/06/30

Casos por país

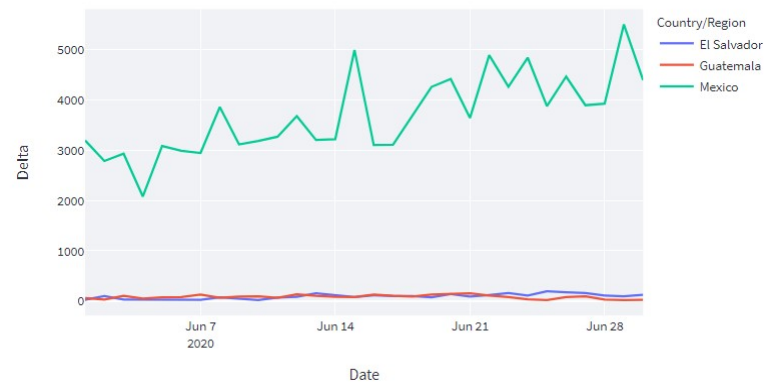
Detalle				-	
	Casos_confirmados	Casos_recuperados	Casos_fallecidos		
El Salvador	3921	2730	128		
Guatemala	13009	2459	665		
Mexico	135425	110766	17839		

Si continuamos con la misma sección podremos ver las gráficas de cómo se han comportado los casos durante el periodo elegido.

Grafica - Casos confirmados



Grafica - Casos recuperados

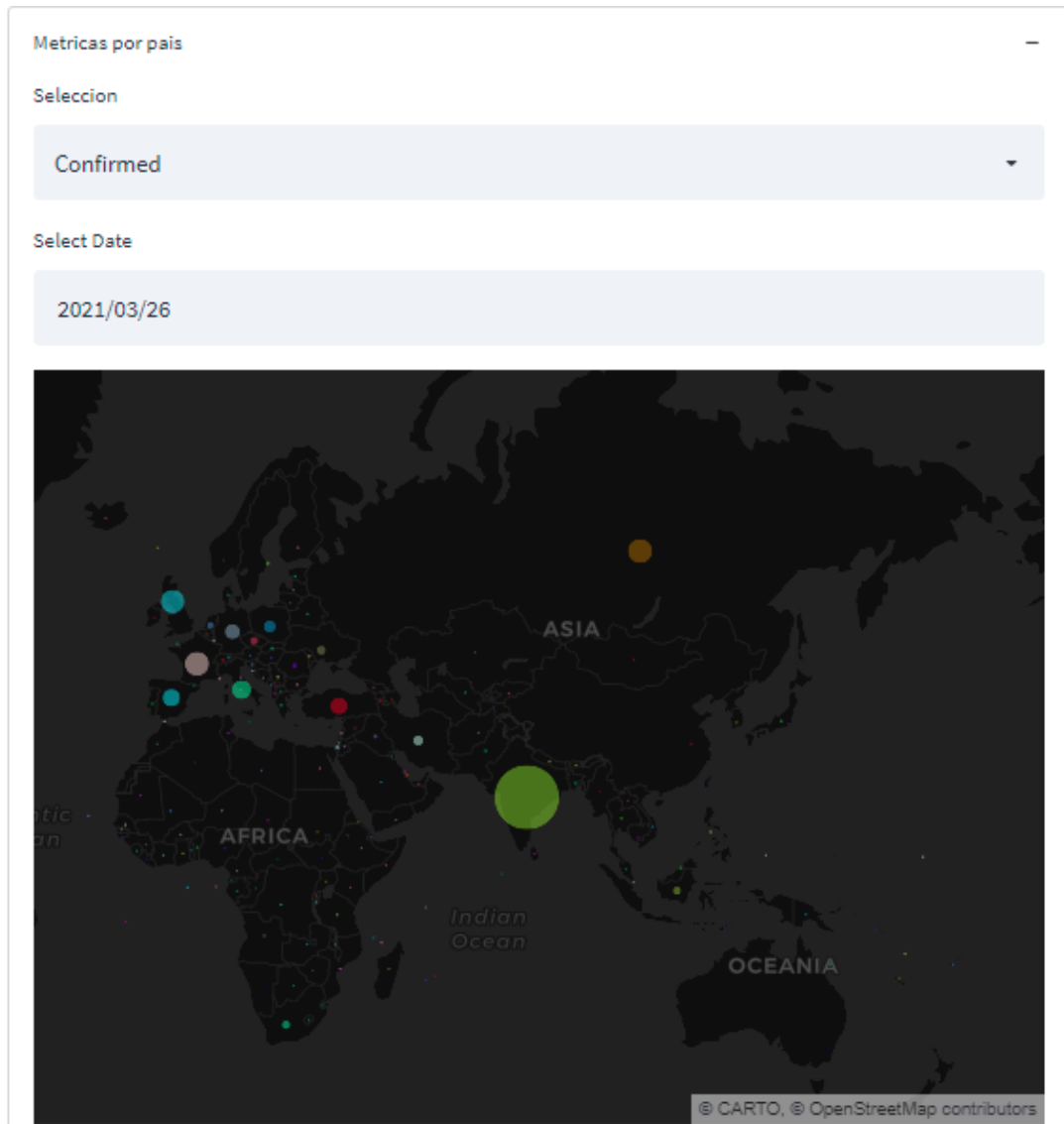


Grafica - Casos Fallecidos



Mapa de casos

Para cada una de las métricas relacionadas al COVID como lo son los casos confirmados, casos recuperados y fallecidos, se puede seleccionar la métrica de interés, así como también una fecha en específico para poder visualizar por medio de un mapa como es la relación de densidad entre cada país. A continuación, se muestra el mapa generado.



Para la generación de este mapa se utilizó la librería Pydeck, específicamente se creó una capa de diagrama de dispersión y por medio de las coordenadas de latitud y longitud incluidas en set se montó sobre un mapa.

Detalle general de casos

Esta sección mostrar la información general de los casos reportados, y funcionará con base a los parámetros indicados en la barra de configuraciones y de la opción seleccionada (Casos confirmado, Casos recuperado y Muertes), a continuación se muestra la consulta de 5 días de casos para el país de Guatemala

Selecciones país(es)

Guatemala X

Seleccione de inicio

2020/06/01

Seleccione fecha de fin

2020/06/05

Detalle: --

Selección:

Casos confirmados

Casos confirmados

Casos recuperados

Muertes

93008 | Guatemala | 15.7835 | -90.2308 | 6154 | 2020-06-04 | 394

Selección:

Casos confirmados

		Country/Region	Lat	Long	Acumulado	Date	Delta
93005		Guatemala	15.7835	-90.2308	5336	2020-06-01	249
93006		Guatemala	15.7835	-90.2308	5586	2020-06-02	250
93007		Guatemala	15.7835	-90.2308	5760	2020-06-03	174
93008		Guatemala	15.7835	-90.2308	6154	2020-06-04	394
93009		Guatemala	15.7835	-90.2308	6485	2020-06-05	331

Selección:

Casos recuperados

		Country/Region	Lat	Long	Acumulado	Date	Delta
82910		Guatemala	15.7835	-90.2308	795	2020-06-01	60
82911		Guatemala	15.7835	-90.2308	824	2020-06-02	29
82912		Guatemala	15.7835	-90.2308	929	2020-06-03	105
82913		Guatemala	15.7835	-90.2308	979	2020-06-04	50
82914		Guatemala	15.7835	-90.2308	1053	2020-06-05	74

Selección:

Muertes

		Country/Region	Lat	Long	Acumulado	Date	Delta
93005		Guatemala	15.7835	-90.2308	116	2020-06-01	8
93006		Guatemala	15.7835	-90.2308	123	2020-06-02	7
93007		Guatemala	15.7835	-90.2308	143	2020-06-03	20
93008		Guatemala	15.7835	-90.2308	158	2020-06-04	15
93009		Guatemala	15.7835	-90.2308	216	2020-06-05	58