laboratorio 2

Daniel Alejandro Soto Mogollón

19 de junio de 2024

universidad ecci

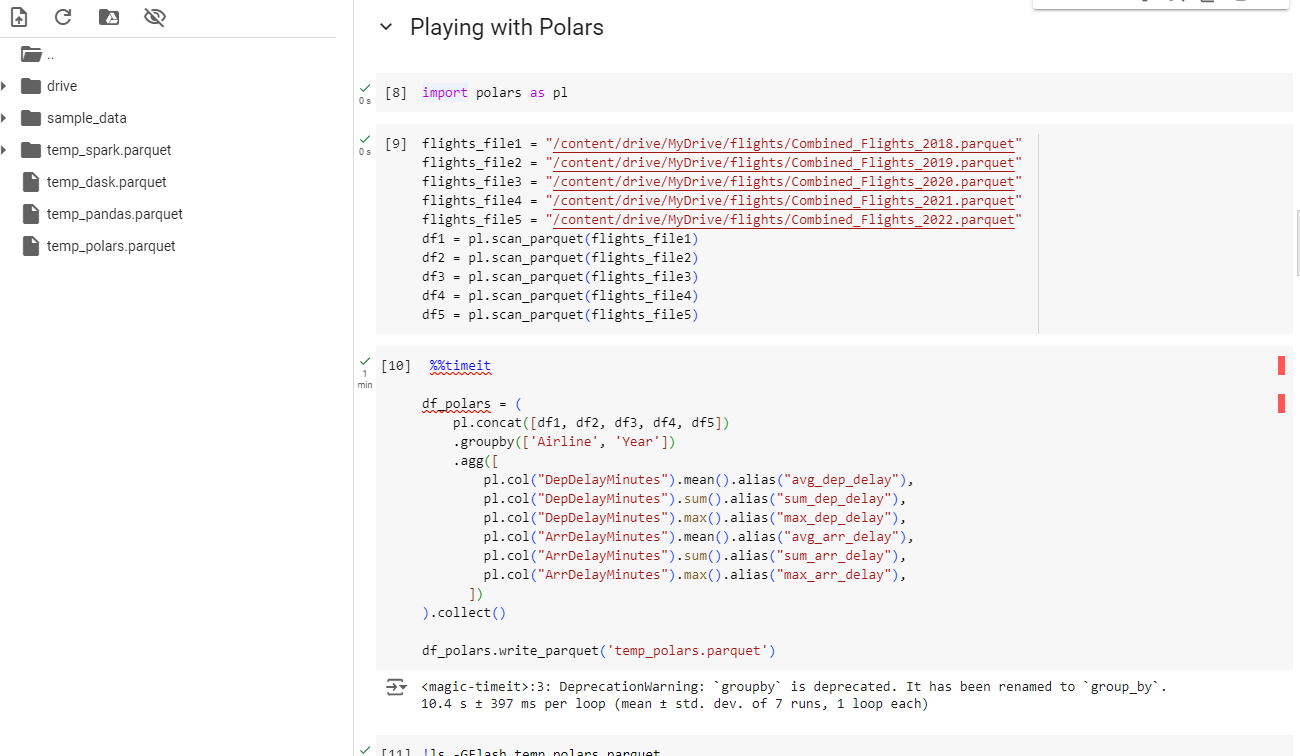
# Pandas



Se realizaron varias pruebas y se logra evidenciar que se pueden leer máximo 3 archivos con esta librería, a la hora de consolidar la capacidad solo permite consolidar 2 Dataframes

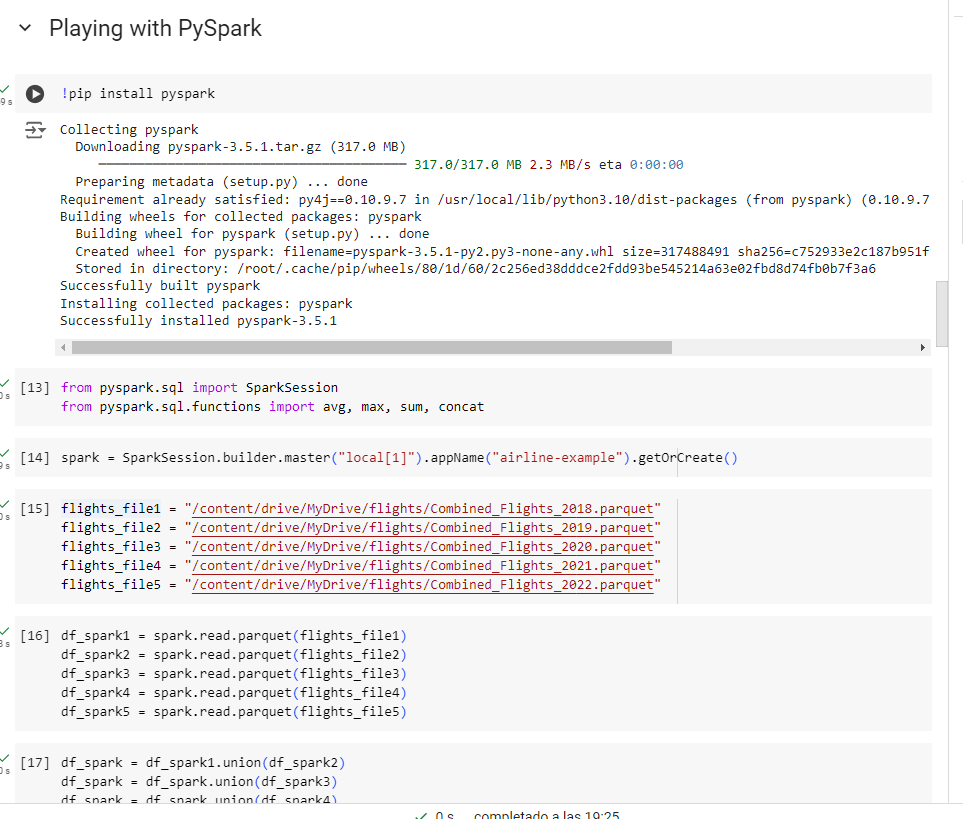
En conclusión vemos que esta libreria esta limitada cuando los dataframes tienen una gran dimension

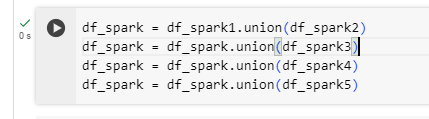
# Polars



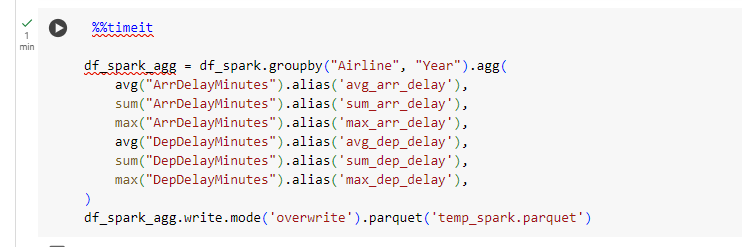
Con esta librería en este proceso se logran leer los 5 archivos , se consumen máximo 2GB de memoria Ram y poco disco ,vemos que a la hora de concatenar estos 5 dataframes vemos que el consumo de Ram sube a 3.2GB este proceso tarda cerca de 90 segundos

# Spark

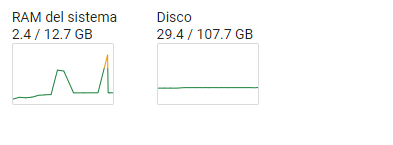


Para este caso vemos que el proceso de lectura de todos los Dataframes dura tan solo 8 segundos como se ve en la celda #16   
  


Vemos que el proceso de unión de los dataframes tardó muy poco .

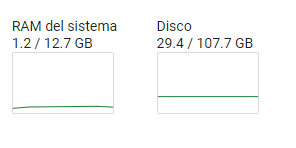


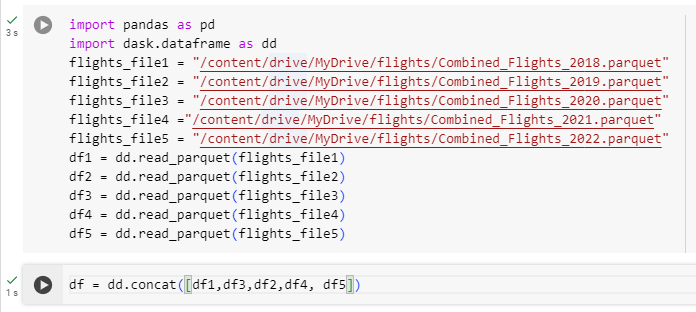
El proceso de agrupamiento tardo un minuto en el cual empleo la mayoría del tiempo exportando la data llamada “temp\_spark.parquet”

En total este proceso consumio los siguientes recursos:  


# Dask



Vemos que lee y consolida en muy poco tiempo consumiendo los siguientes recursos:  


Vemos que es muy bajo el consumo, al igual que realizarlo con todos los Dataframes:  


El proceso de lectura de la totalidad de los archivos y la unión de todos ellos solo tarda 4 segundos

Lectura de los archivos resultado de la transformación con todas las librerias.



Vemos que la lectura de estos archivos fue demasiado veloz con la librería de pandas .