Informe Librerías

Juan Sebastian Cortes Chia

Jhonatan Rios Tapiero

Seminario Big Data

Elías Buitrago Bolívar

Universidad ECCI

Facultad Ingeniería

Contenido

[**Tablas Comparativas** 3](#_Toc171251607)

[Pandas 3](#_Toc171251608)

[Polars 3](#_Toc171251609)

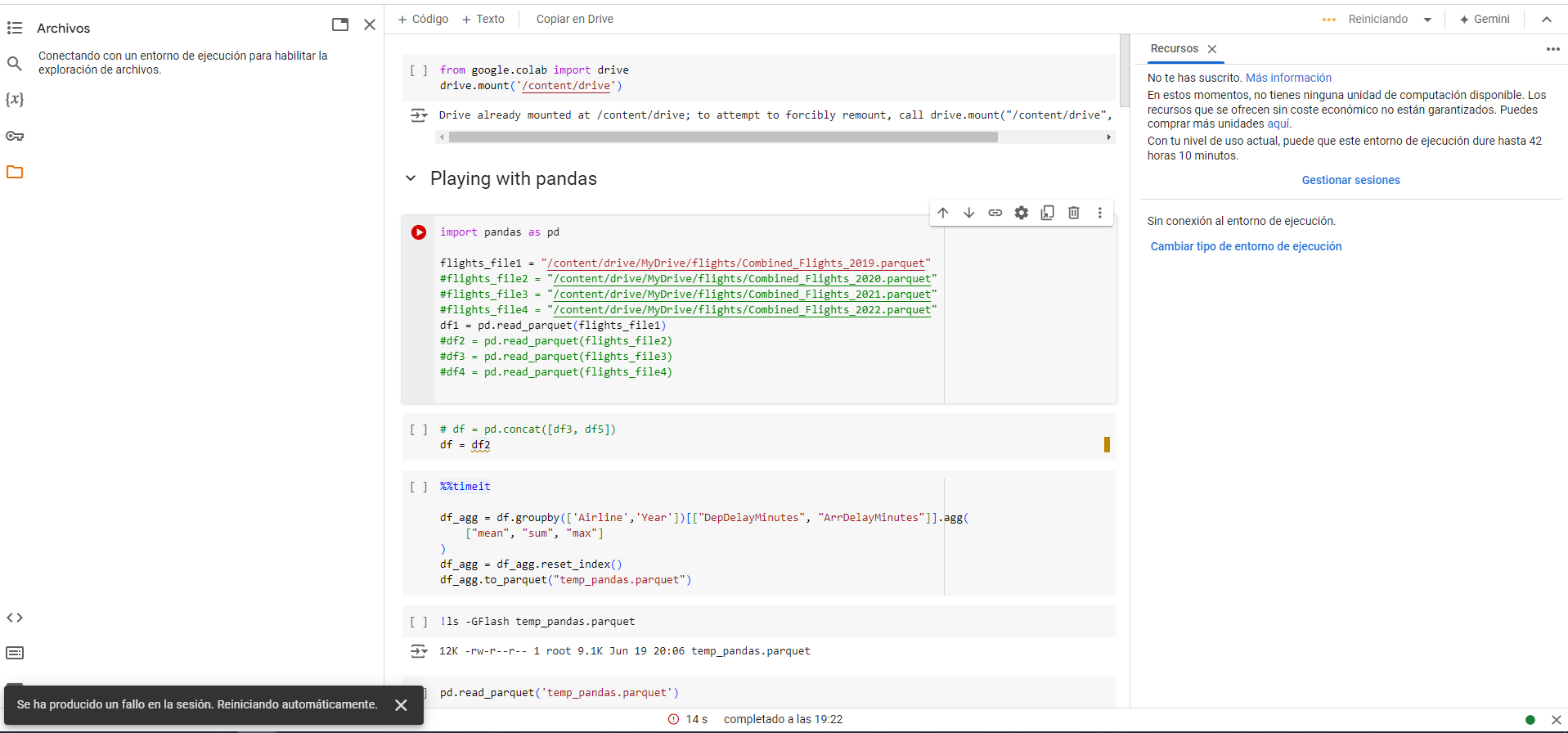
[PySpark 4](#_Toc171251610)

[Dask 5](#_Toc171251611)

[Análisis de resultados 7](#_Toc171251612)

# **Tablas Comparativas**

## Pandas

Con un único archivo, se evidencia que Pandas no realiza el proceso de importación correctamente, ya que por los recursos se limita:

## Polars

Se realiza el cargue con los 4 archivos, y se evidencia que con la librería polars se realiza el cargue exitosamente, como se visualiza en la imagen:

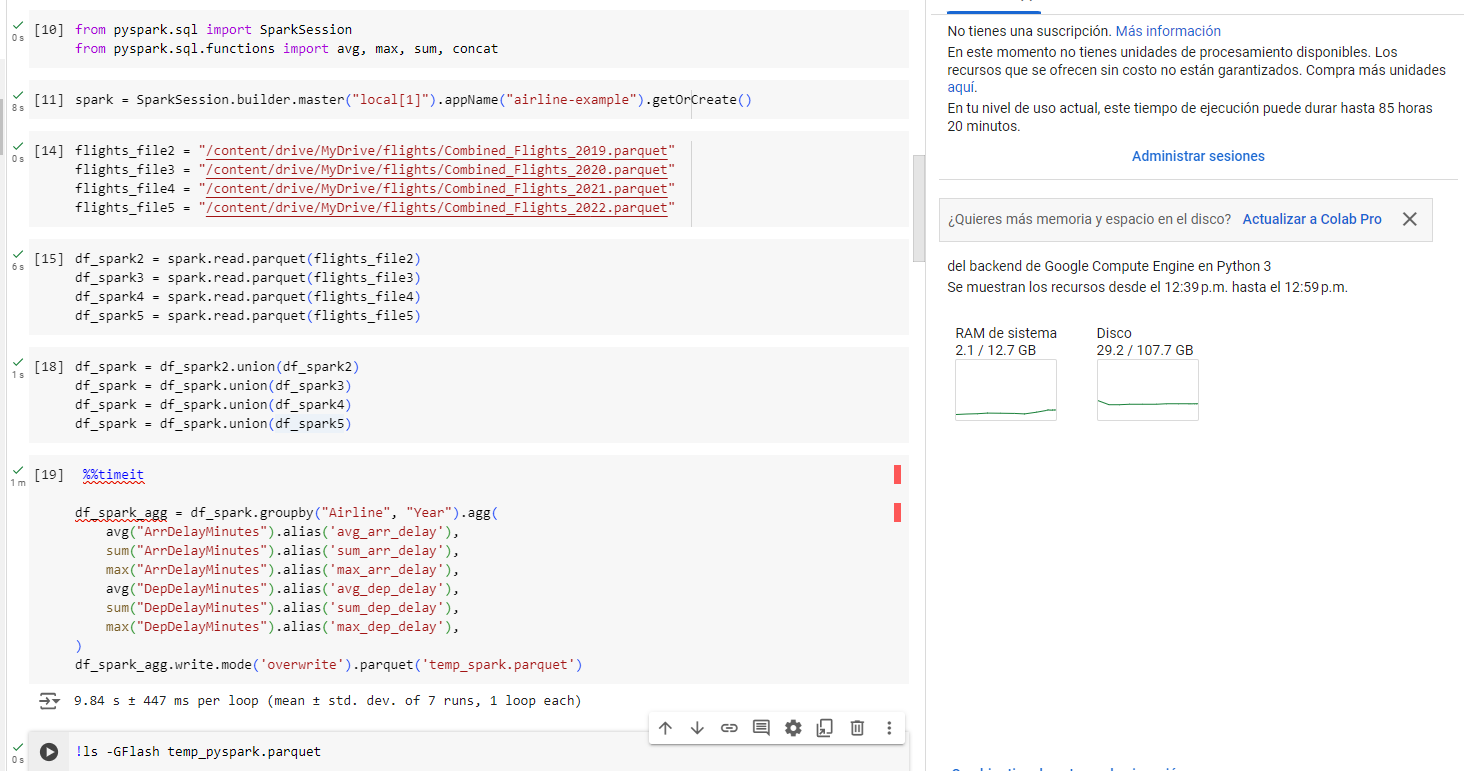


Al realizar la concatenación y agrupación de los dataframes, se evidencia que su tiempo de procesamiento es de 1min:



## PySpark

Se realiza el cargue con los 4 archivos, y se evidencia que con la librería PySpark se realiza el cargue exitosamente, se evidencia que el tiempo en el cuál tuvo mas demora en su procesamiento fue en el la agrupación de los data set, y en las columnas, ya que se demoro alrededor de minuto y medio:

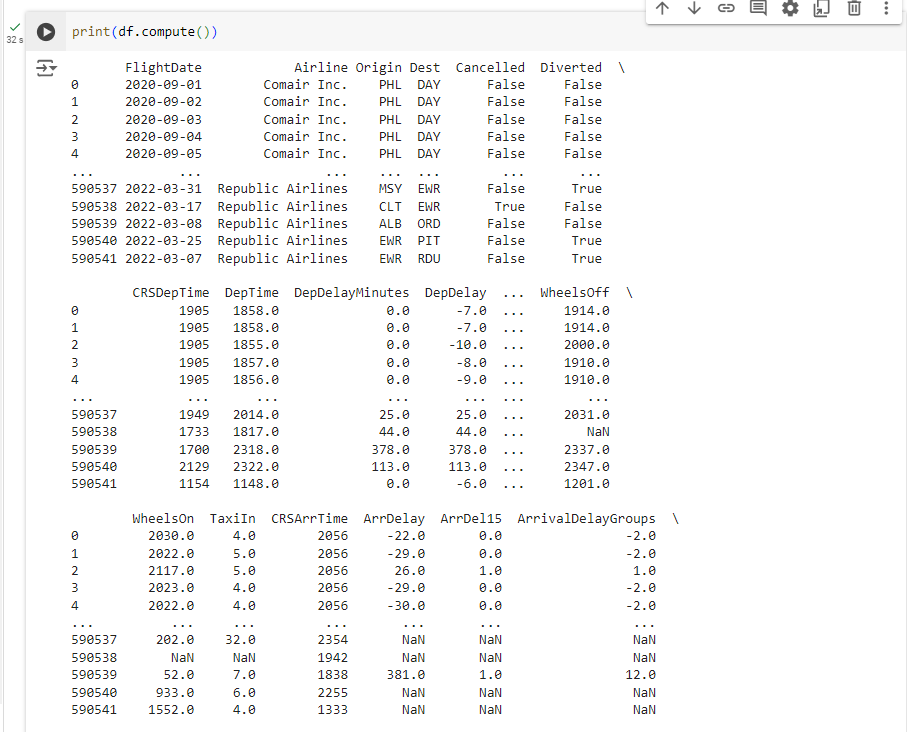


## Dask

Al realizar el cargue de archivos con la librería dask, se evidencia que el cargue es ágil, como se evidencia en la siguiente imagen:



En el procesamiento para visualizar los datos, se evidencia que el tiempo transcurrido es de alrededor de 32 segundos en el procesamiento de estos datos para visualizarlos posteriormente:

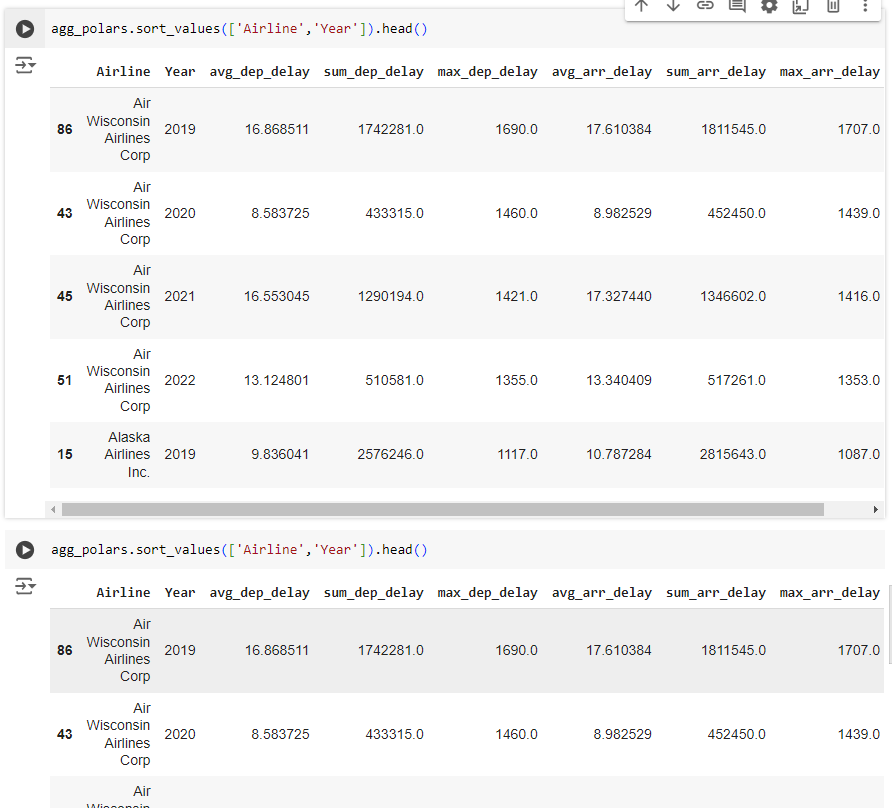


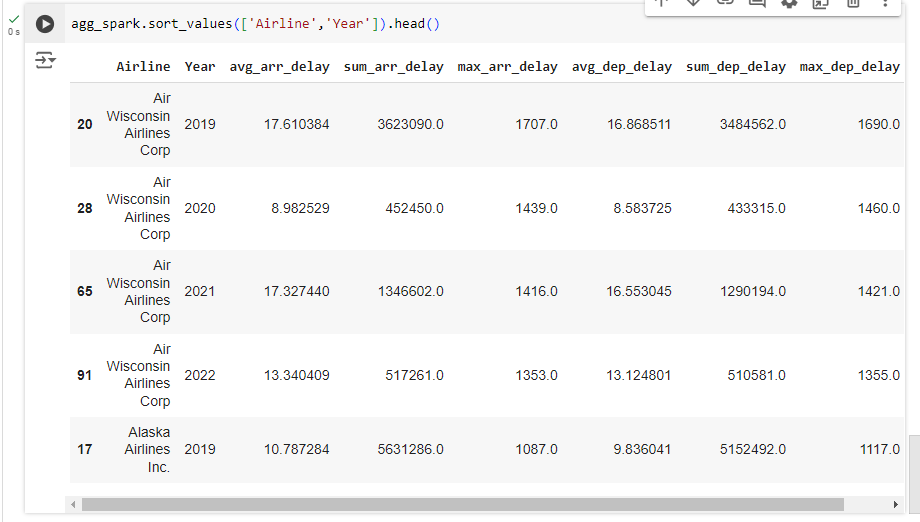
En el proceso de la agrupación de los dataframes se evidencia un mejor rendimiento en el procesamiento, ya que su tiempo es de 1s:

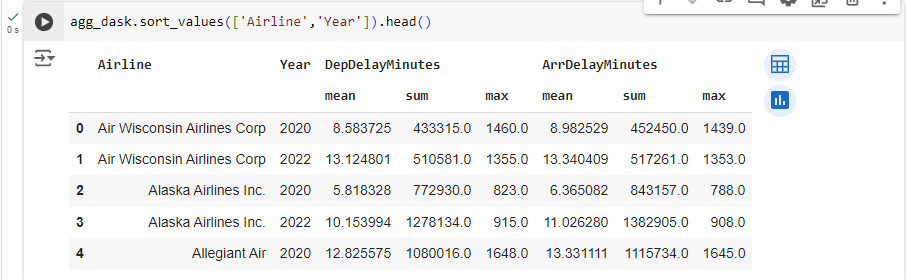


# Análisis de resultados

Al realizar el análisis de resultados con Pandas, no es exitoso debido a que como no fue posible realizar el cargue de los archivos, por los recursos del servidor de Google Collaboratory, únicamente dejo interactuar con las demás librerias Polars, , como se evidencia a continuación:







# Conclusiones

Después de realizar pruebas en Google Collaboratory con las bibliotecas Pandas, Polars, PySpark y Dask, se encontró que todas funcionaron bien dentro de los recursos limitados proporcionados. Sin embargo, Pandas experimentó dificultades al intentar exportar grandes conjuntos de datos en un solo archivo.

En términos de velocidad y capacidad para manejar tareas complejas de análisis de datos, Polars, PySpark y Dask demostraron ser sólidas y consistentes bajo condiciones de recursos restringidos como los que nos brindaba Google Collaboratory. Estas bibliotecas mostraron un rendimiento eficiente y fueron capaces de manejar tamaños de datos crecientes sin comprometer significativamente los tiempos de ejecución.

Por otro lado, Pandas, a pesar de ser ampliamente utilizada y potente, reveló limitaciones específicas al enfrentarse a la exportación de grandes archivos de datos en el entorno de Google Collaboratory. Este hallazgo destaca la importancia de considerar las capacidades de procesamiento y los recursos disponibles al seleccionar la herramienta adecuada para proyectos de análisis de datos en entornos con recursos limitados.

En resumen, mientras que Polars, PySpark y Dask se destacaron por su eficiencia y capacidad para manejar grandes volúmenes de datos en Google Collaboratory, Pandas mostró desafíos en escenarios que requerían la manipulación de archivos de datos extensos. Estos resultados enfatizan la necesidad de evaluar cuidadosamente las herramientas según las necesidades específicas y las limitaciones del entorno de trabajo.