LABORATORIO EN CLASE
COMPRENSIÓN DE LOS DATOS

Profesor: Elias Buitrago B.

Última actualización: 30-Noviembre-2023

OBJETIVOS DE LA CLASE

- Comparar el desempeño de librerías de Python para carga y manipulación de datos tabulares.

INTRODUCCIÓN

Este laboratorio busca comprar la eficiencia en el procesamiento básico de datos tabulares en Python de las siguientes librerías: *Pandas*, *Polars*, *Spark*, *Dask*. En cada caso se buscar que pueden leer la mayor cantidad de datos en Google Colaboratory evitando que se sature la memoria RAM disponible y se reinicie el entorno. Por lo tanto, deberán jugar con cada librería y con el conjunto de datos propuesto busca la más eficiente para cargar datos (que sature menos la memoria RAM) maximizando la cantidad de archivos que puedan procesar con cada uno. Al final tendrán que hacer un cuadro comparativo y gráficos de desempeño de los resultados, mediante un informe que entregarán vía Classroom. En el inicio de la clase se brindarán mayores detalles al respecto. Es importante que tengan en cuenta que la actividad es para desarrollar en clase.

DESCRIPCIÓN GENERAL

Se conformarán grupos de máximo 2 estudiantes. El taller en clase se apoya en las siguientes actividades:

1. Ejecutar cuaderno de Jupyter *"1PPvsSpark_01.ipynb"* y comprobar su correcto funcionamiento. Este cuaderno tiene como propósito realizar una comparación de librerías para carga y procesamiento básico de datos tabulares en Python, con el fin de identificar cuál es más eficiente en un caso de uso específico.
2. Experimentar con cada una de las librerías intentando cargar en Google Colaboratory la mayor cantidad de datos, evitando que se sature la memoria RAM disponible y se reinicie el entorno. Por lo tanto, se recomienda "jugar" con cada librería y con el conjunto de datos propuesto, buscando identificar la librería más eficiente para cargar datos (que sature menos la memoria RAM) y maximizar la cantidad de archivos que puedan procesar con cada uno. Al final tendrán que hacer un cuadro comparativo o graficar el desempeño de los resultados. Tengan en cuenta la tabla comparativa que se propuso en clase. Finalmente, hagan una discusión de los resultados.
3. Realizar una tabla comparativa que muestre el tamaño en disco de cada uno de los archivos de datos y su tamaño correspondiente, una vez ha sido cargado en Google Colaboratory.
4. Comparar las distintas librerías intentando cargar la mayor cantidad de archivos con cada una para concatenarlos, sin que se reinicie el entorno de ejecución por saturación de RAM. se deben evidenciar los resultados del experimento con tablas, gráficos y capturas de pantalla. Además, se espera que analicen y comenten los resultados obtenidos.

El desarrollo de los puntos se debe plasmar en un informe, bien organizado (al nivel de ingeniería) y con buena ortografía. No olviden escribir los nombres completos de los integrantes de cada grupo.

INFORME DE LABORATORIO, COMPRENSIÓN DE LOS DATOS

Jhoan Sebastian Riaño Herrera, 38525.

Brayan Camilo Salazar González, 70058.

1. DESARROLLO

1.1. Lo primero que se hizo para llevar a cabo el laboratorio fue cargar los archivos provenientes de la carpeta Flights en el Drive, para así, generar una ruta que permitiera la debida ejecución por parte del cuaderno Jupyter.

Posterior a eso se iniciaron las pruebas, en primera instancia se inició con la librería Pandas. Después de ejecutar la librería, el siguiente paso era cargar los archivos de manera individual y en diferentes combinaciones con el fin de identificar en qué momento se presentaba una *saturación* de la RAM del sistema.

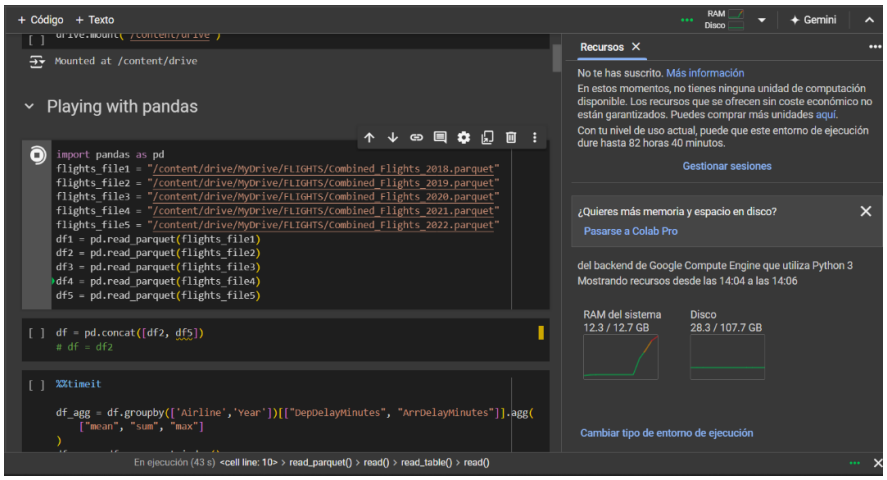


Fig. 1. Muestra de saturación de la RAM del sistema con librería Pandas

Al ejecutar la librería Pandas y cargar los archivos FLIGHTS de manera individual, se identifica que el archivo que ocupa más espacio en RAM del sistema es el #2, llegando aproximadamente a 8Gb en 14s.

En el momento que se empiezan a leer datos combinados, se logra hacer la lectura de la combinación 1-2, y, 1 a 3, llegando a ocupar un máximo de 11Gb de RAM en 27s, pero las combinaciones 1 a 4 y 1 a 5 ya saturan la RAM y generan reinicio del sistema como se puede ver en la figura 1.

Al ejecutar la función de concatenar, la combinación de variables 1:2 satura el sistema (fig. 2), 2:3 satura el sistema (fig. 3), la 3:4 lo ejecuta llegando a 11.8Gb de RAM en 12s, la 4:5 también la ejecuta con 10.9 de RAM en 9s.

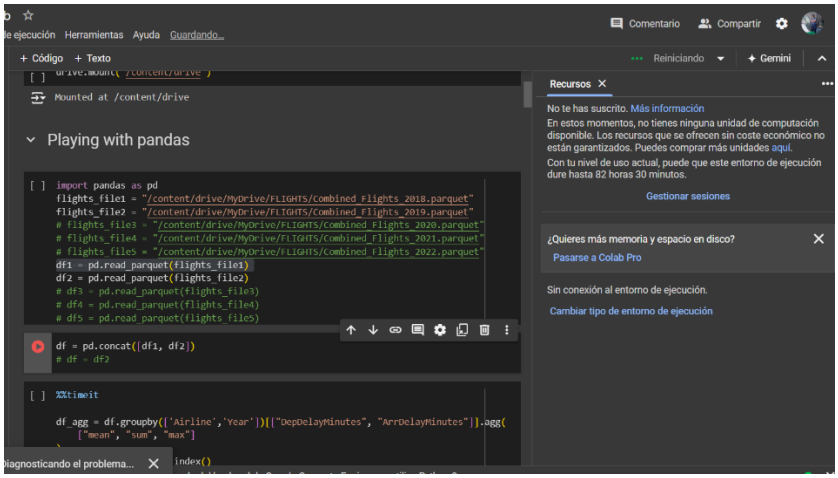


Fig. 2. Ejecución de función *concat* en Pandas con archivo 1 y 2.

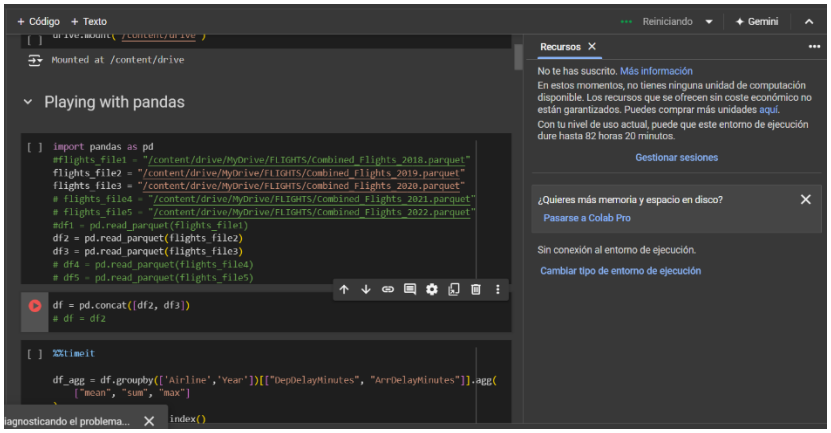


Fig. 3. Ejecución de función concat en Pandas con archivo 2 y 3.

1.2. La lectura de datos en la librería Polars es rápida, toma menos de 1s y no se evidencian cambios en la RAM del sistema, ahora bien, al ejecutar la función concat con todas las variables, toma cerca de 80s y alcanza un máximo de 5Gb de RAM aproximadamente.

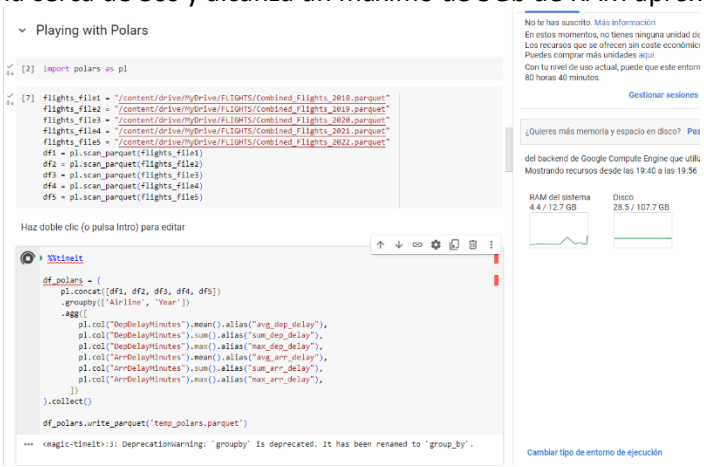


Fig. 4. Prueba de carga y lectura de los archivos en librería Polars.

Fig. 5. Ejecución de función concat en Polars.

1.3. Con la librería Sparks, la carga de los archivos en conjunto del 1 al 5 toma menos de 1s y no se evidencia cambios en la RAM del sistema. Así mismo, la ejecución de las diferentes uniones se realiza de una manera eficaz en cuanto a tiempo y RAM.

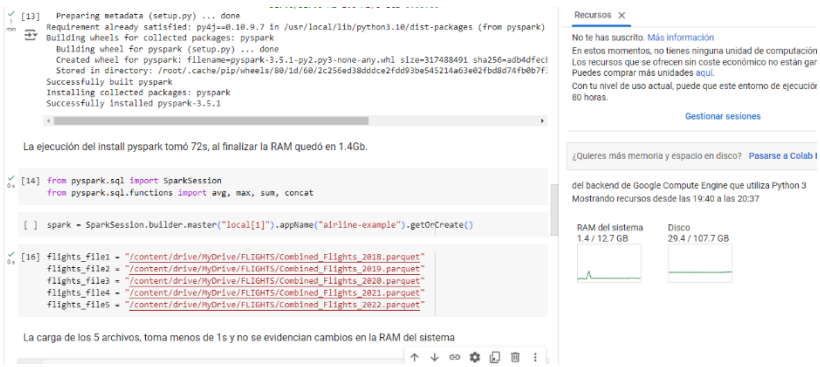


Fig. 6. Lectura, carga y unión de los archivos con la librería Sparks.

1.4. Finalmente, al hacer la prueba con la librería Dask, se encuentra que la carga y lectura de los archivos tanto de manera individual como en conjunto, se realiza en un tiempo menor a 1s con una variación máxima de 0.5Gb.

Con la ejecución de la función concat, se evidencia que así se utilicen las 5 variables, se lleva a cabo en un tiempo máximo de 2s y un aumento hasta 2Gb en la RAM del sistema.

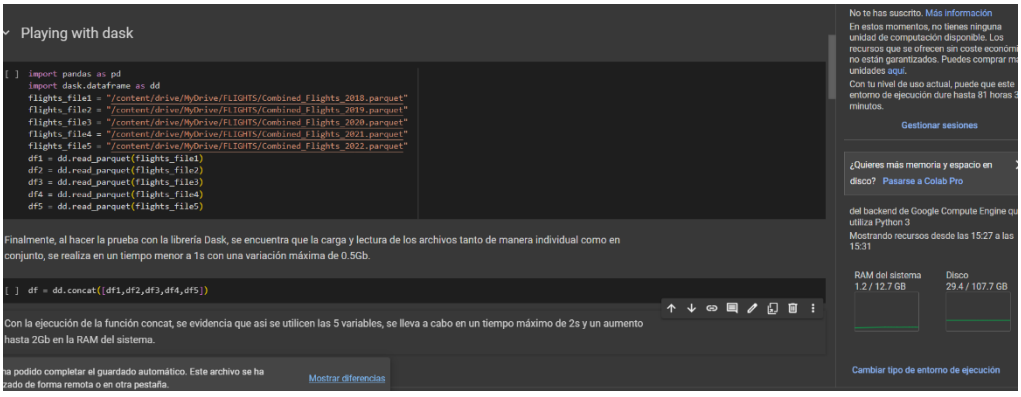
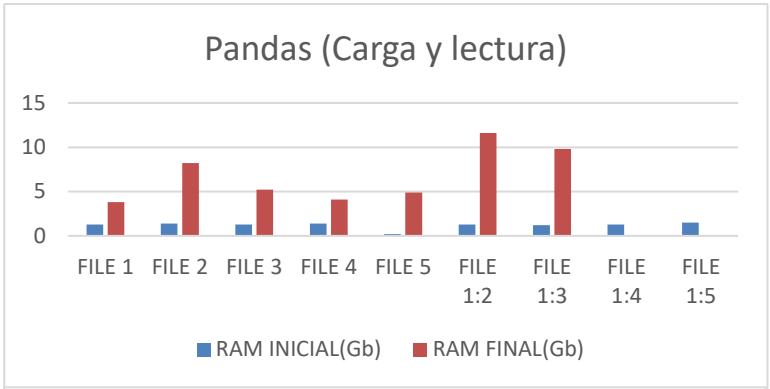
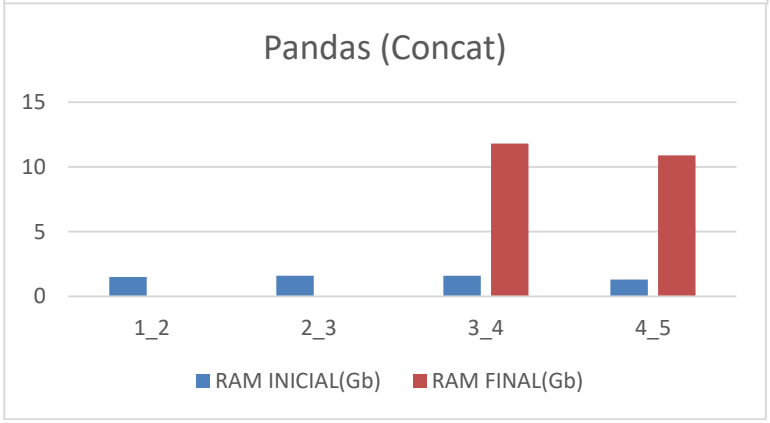


Fig. 7. Lectura, carga y concat de los archivos con la librería Dask.

2. **RESULTADOS**.

2.1. **PANDAS**: En las siguientes tablas, se puede evidenciar el rendimiento de la RAM del sistema durante la carga y lectura y en la ejecución de la función *concat* de cada uno de los archivos y en cada una de las combinaciones que se intentó. En aquellas variables donde se ve la barra correspondiente a la RAM final no aparece, representa la saturación del sistema.





3. **CONCLUSIONES**.

Después de interactuar con las diferentes librerías y realizar las pruebas indicadas, se puede concluir que todas tienen un buen desempeño, pero la librería Pandas presenta limitación cuando se cuenta con muchos datos, por lo menos en esta versión. Las demás librerías, cargan y leen los archivos sin presentar ninguna dificultad.

Esto permite deducir que es necesario conocer el desempeño de cada librería al momento de utilizarlas ya que finalmente lo que se busca es reducir tiempo y optimizar los recursos disponibles.