**LABORATORIO 2**

**SEMINARIO DE BIG DATA Y CIENCIAS DE DATOS**

**CRISTHIAN CAMILO BALAGUERA SOTO - 88913**

**DANIEL MAURICIO GORDILLO ALVAREZ - 89400**

**PROFESOR**

[ELIAS BUITRAGO BOLIVAR](mailto:ebuitragob@ecci.edu.co)

**UNIVERSIDAD ECCI**

**BOGOTÁ D.C., COLOMBIA**

**2024**

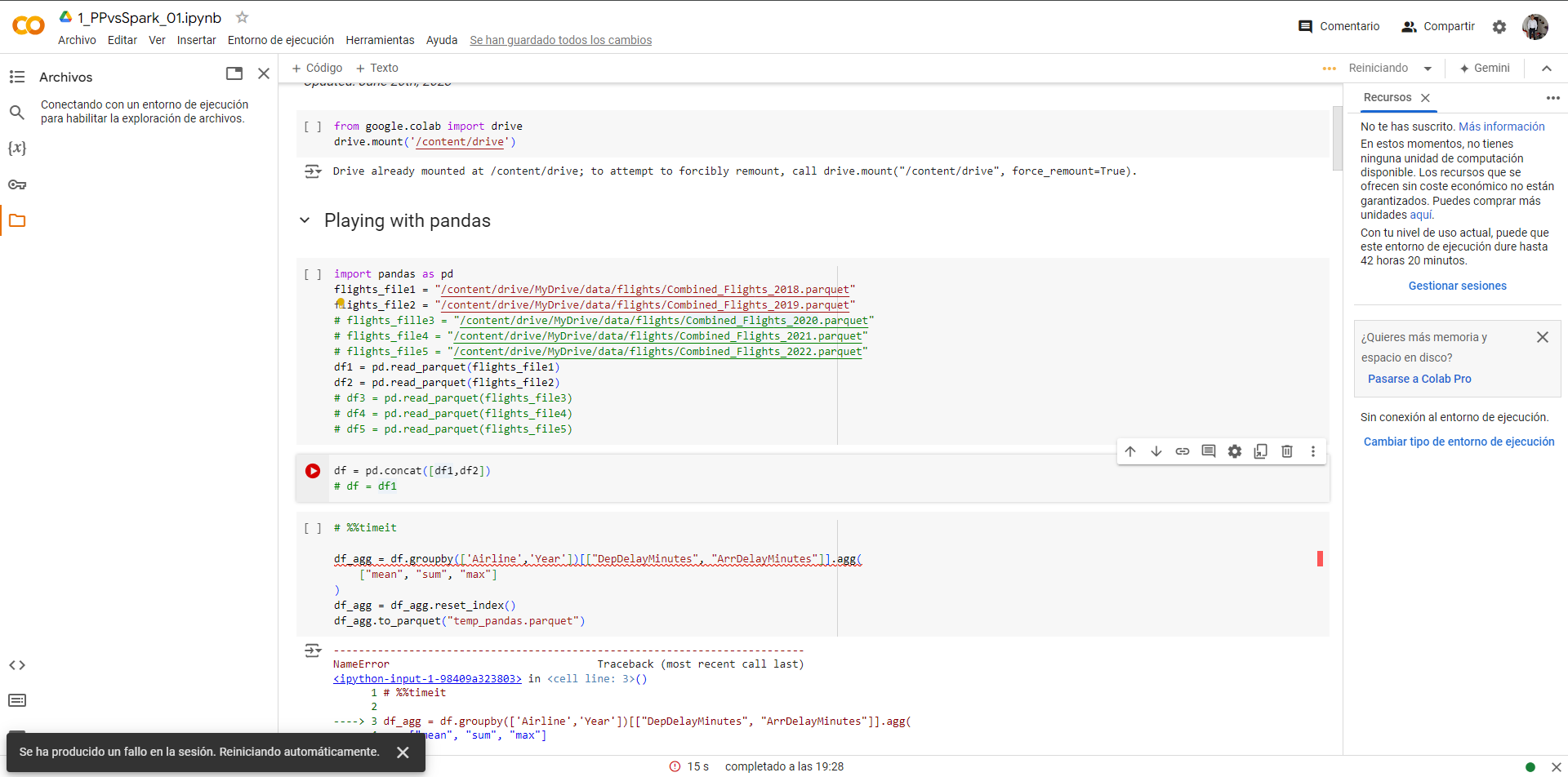
**Laboratorio 2**

**Introducción**

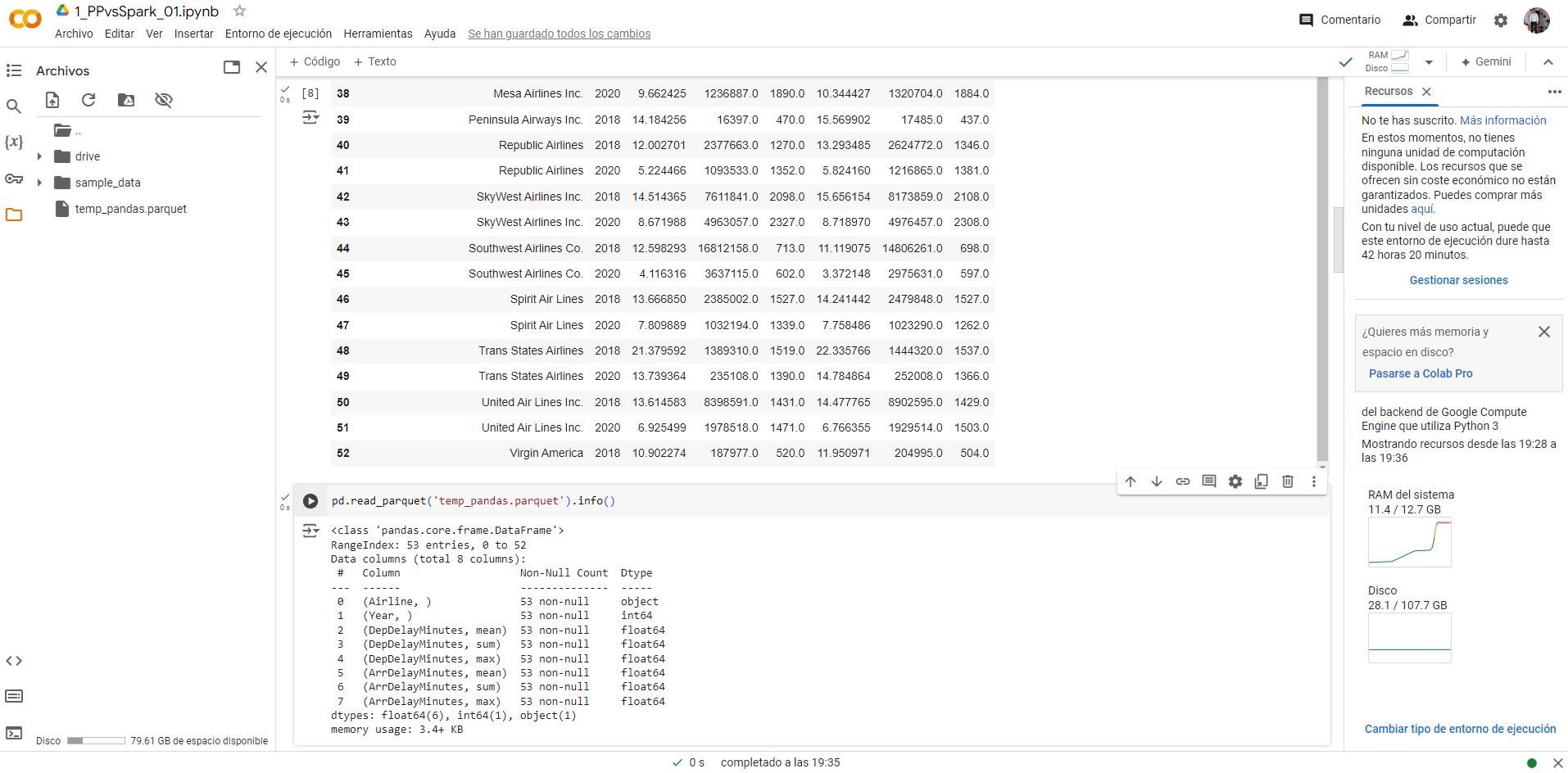
Este laboratorio tiene como objetivo evaluar la eficiencia de diversas librerías de Python para el procesamiento de las librerías: Pandas, Polars, Spark y Dask. El propósito es determinar cuál de estas librerías puede manejar la mayor cantidad de datos en Google Colaboratory sin que la memoria RAM se sature y provoque el reinicio del entorno. Se experimentaran con cada librería y el conjunto de datos proporcionado, identificando cuál es la más eficiente en términos de uso de memoria RAM y capacidad de procesamiento de archivos.

| **Método**  **Pandas** | **# de prueba** | **Datos utilizados** | **Resultados** |
| --- | --- | --- | --- |
| Prueba 1 | **flights\_file1**  **flights\_file2**  **df1**  **df2** | 42 segundos  Limite de RAM alcanzado en el paso 3 |
| Prueba 2 | **flights\_file1**  **flights\_file3**  **df1**  **df3** | 43 segundos  Proceso ejecutado con éxito  RAM : 11.4 Gb  Disco : 28.1 Gb |
| Prueba 3 | **flights\_file2**  **flights\_file4**  **df2**  **df4** | 47 segundos  Limite alcanzado en el paso 2 |
| Prueba 4 | **flights\_file2**  **flights\_file5**  **df2**  **df5** | 43 segundos  Limite alcanzado en el paso 3 |

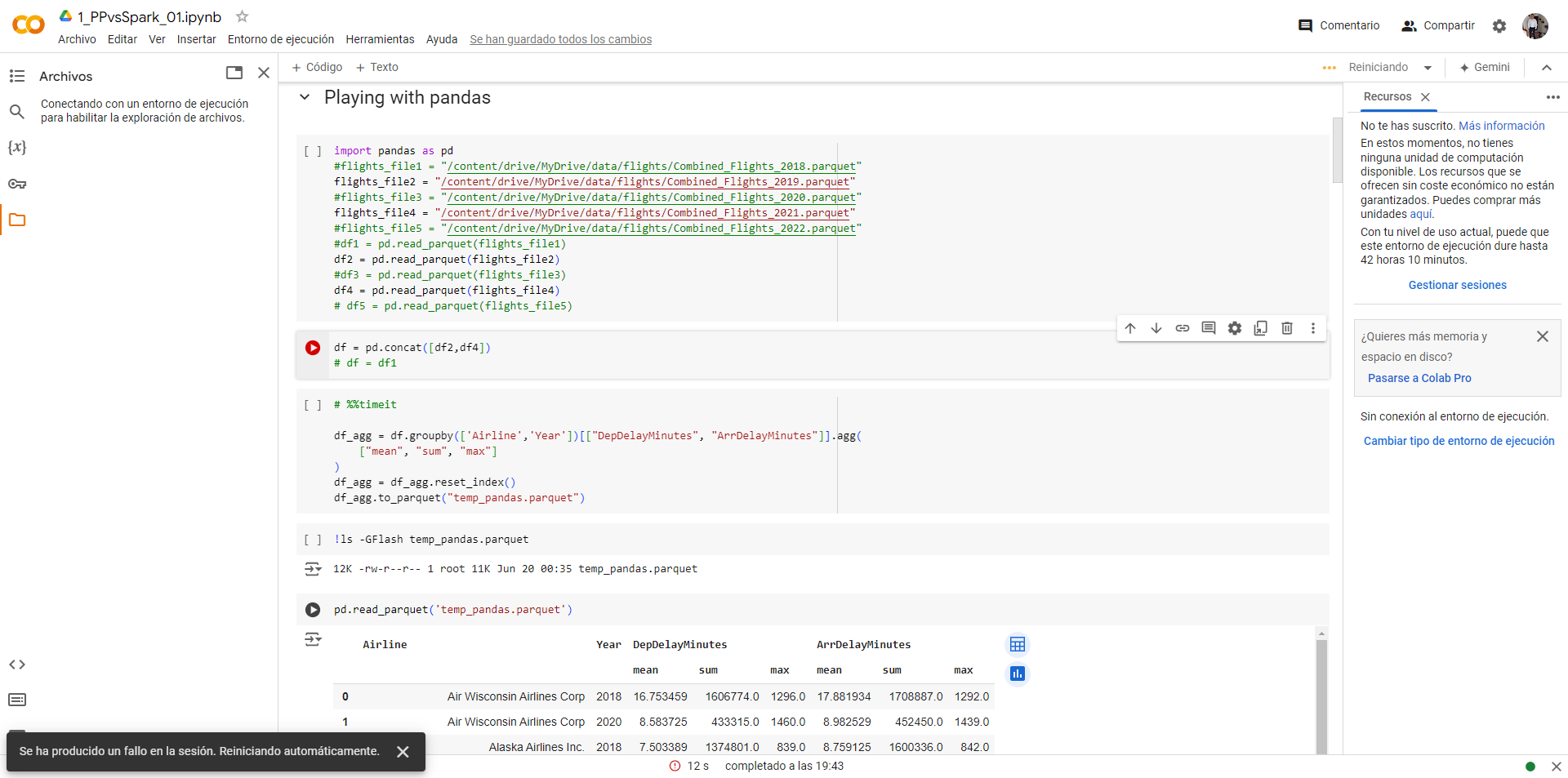
**Prueba 1**

****

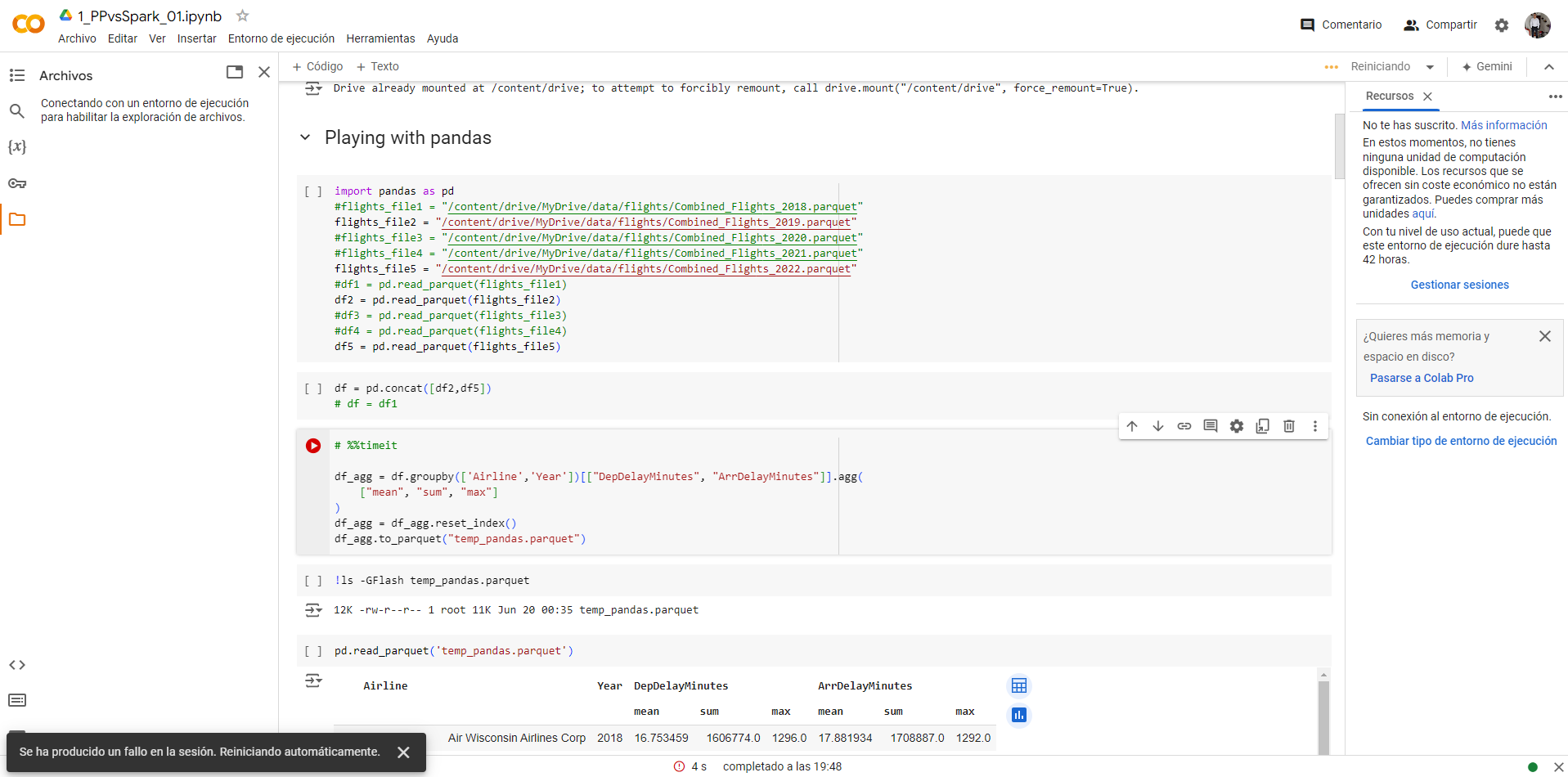
**Prueba 2**

****

**Prueba 3**

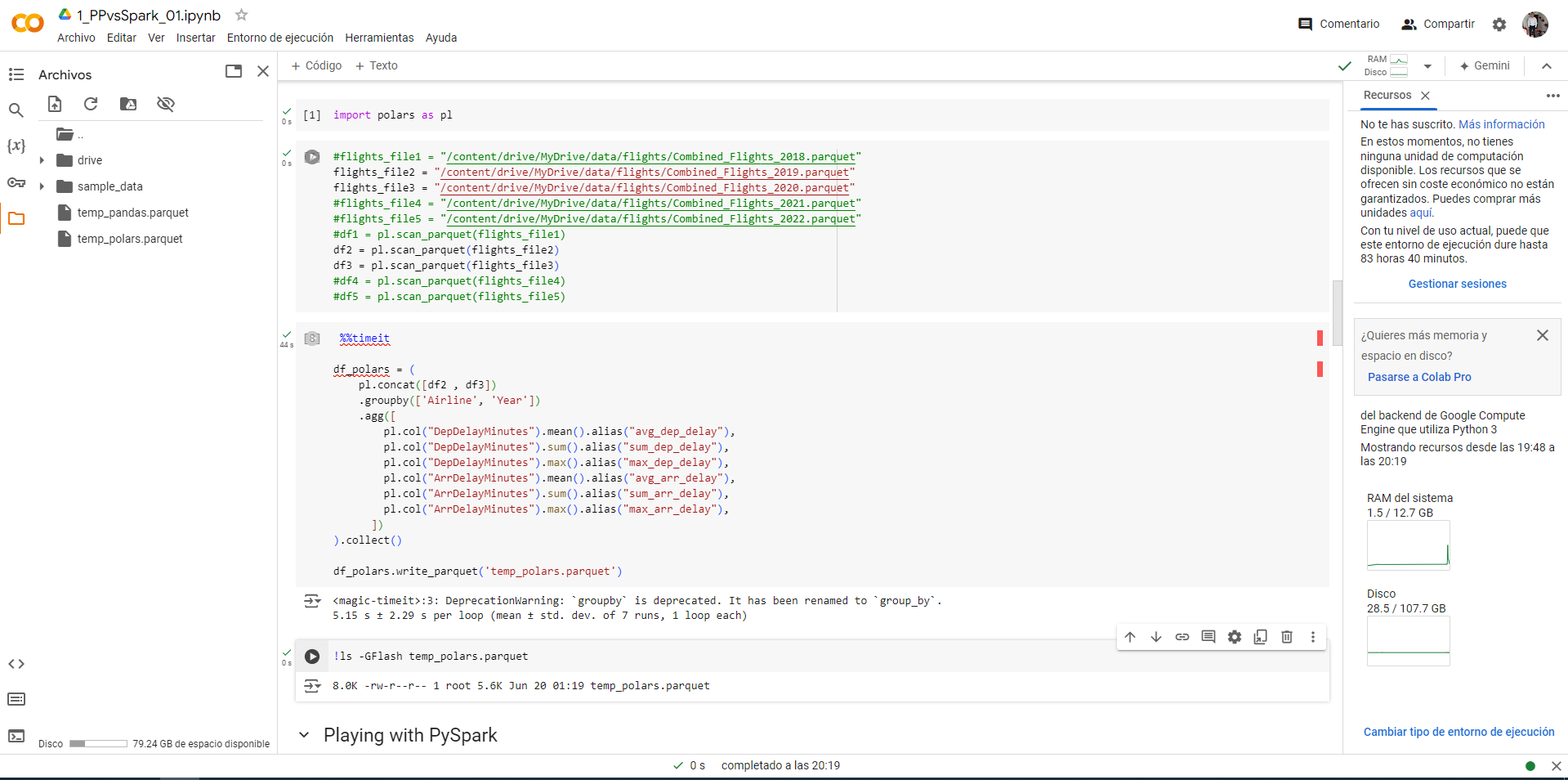
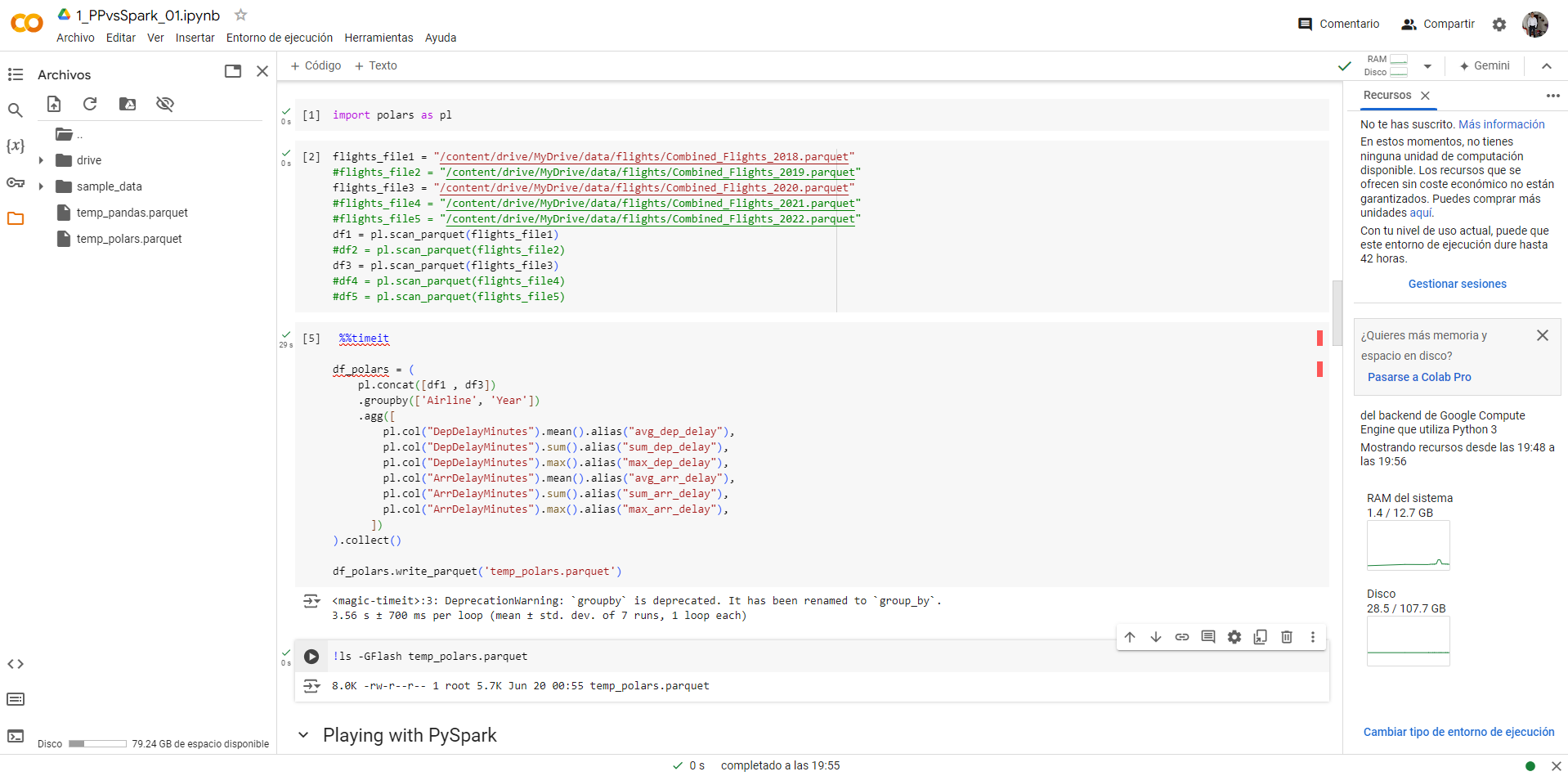
****

**Prueba 4**

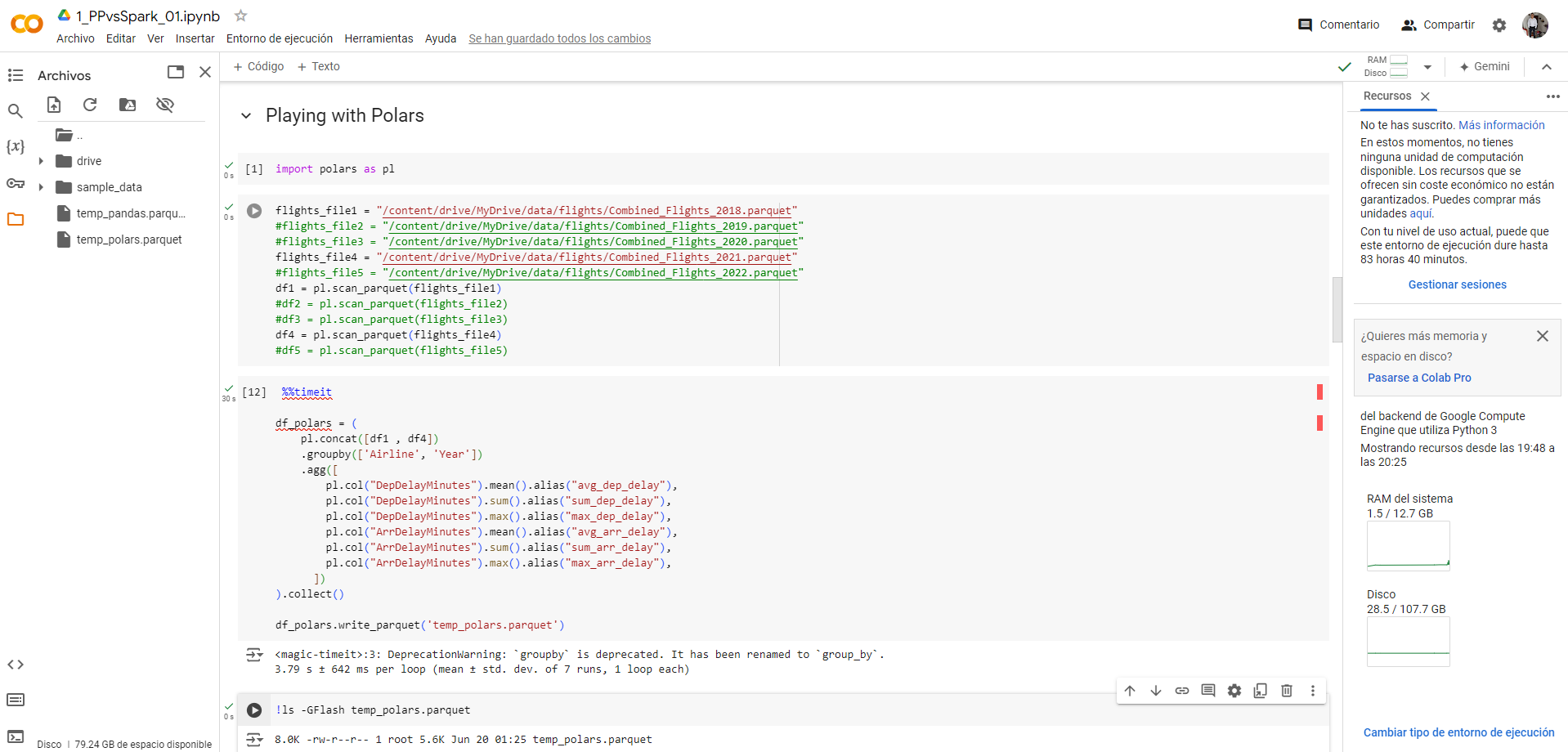
****

| **Método**  **Polars** | **# de prueba** | **Datos utilizados** | **Resultados** |
| --- | --- | --- | --- |
| Prueba 1 | **flights\_file1**  **flights\_file3**  **df1**  **df3** | 42 segundos  Proceso ejecutado con éxito  RAM: 1.4 Gb  Disco: 28.5 Gb |
| Prueba 2 | **flights\_file2**  **flights\_file3**  **df2**  **df3** | 44 segundos  Proceso ejecutado con éxito  RAM : 1.5 Gb  Disco : 28.5 Gb |
| Prueba 3 | **flights\_file1**  **flights\_file4**  **df1**  **df4** | 30 segundos  Proceso ejecutado con éxito.  RAM:1.5 Gb  Disco: 28.5 Gb |
| Prueba 4 | **flights\_file3**  **flights\_file5**  **df3**  **df5** | 24 segundos  Proceso ejecutado con éxito.  RAM:1.5 Gb  Disco: 28.5 Gb |

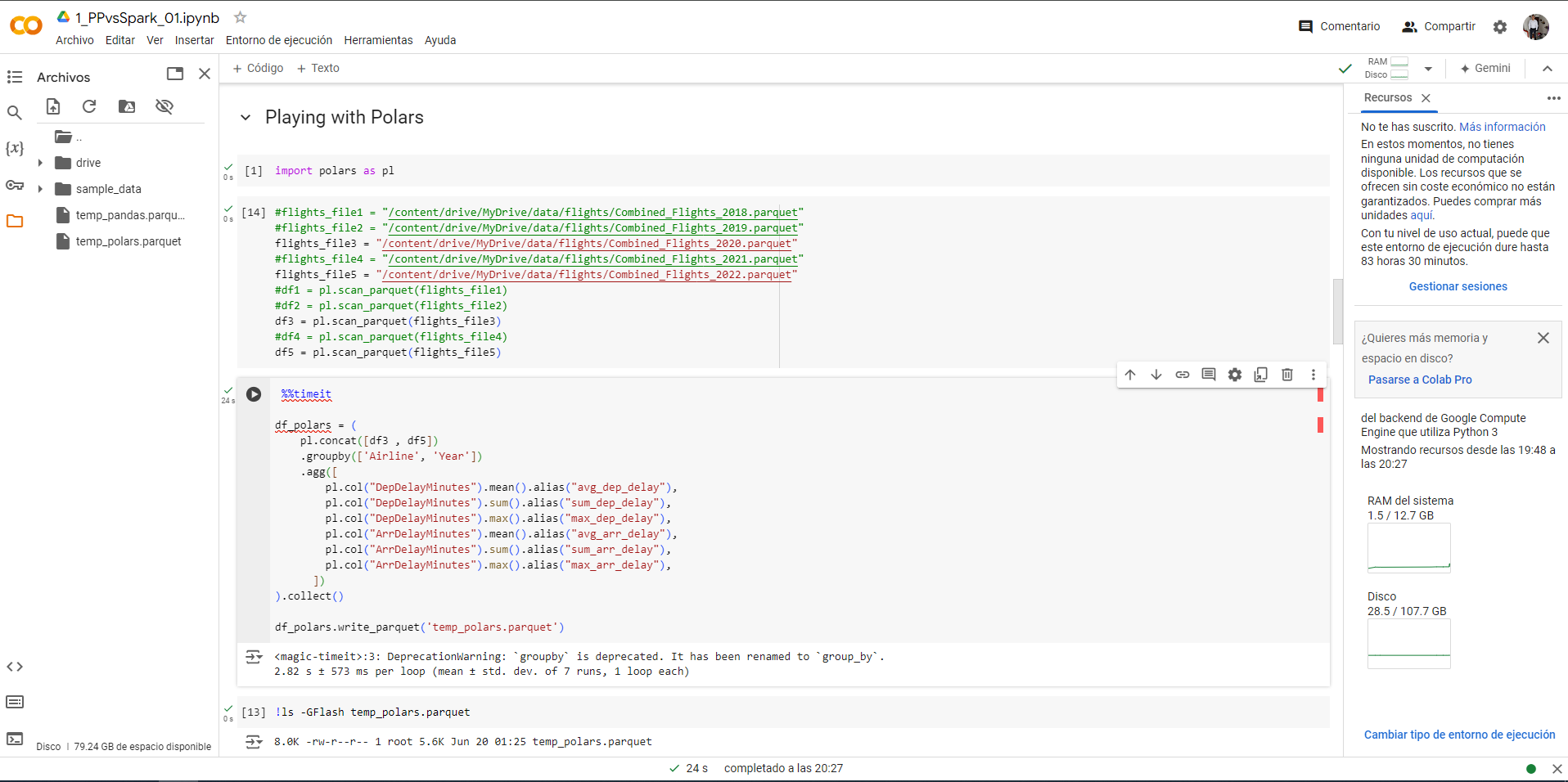
**Prueba 1**

**Prueba 2 **

**Prueba 3**

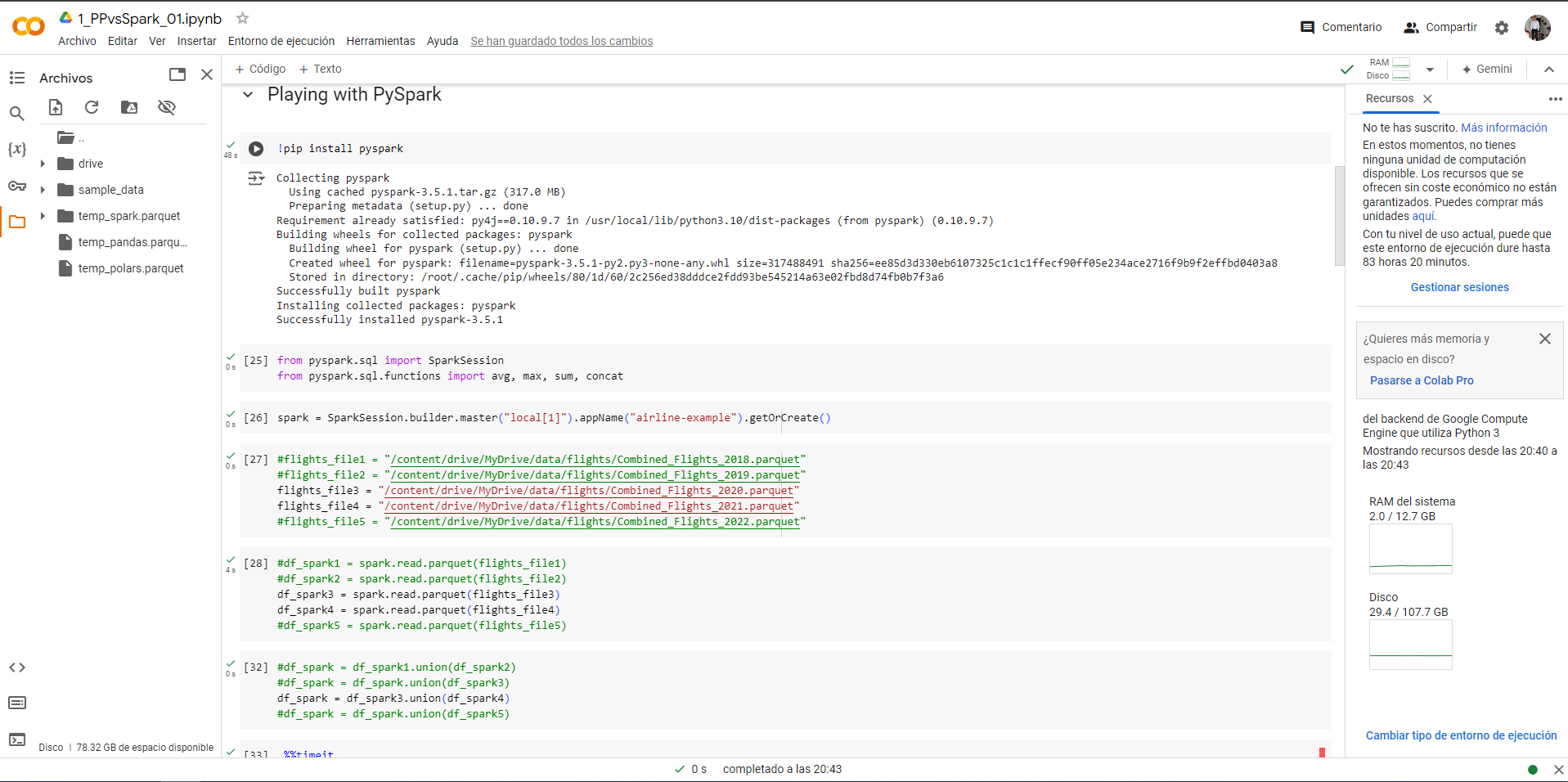
****

**Prueba 4**

****

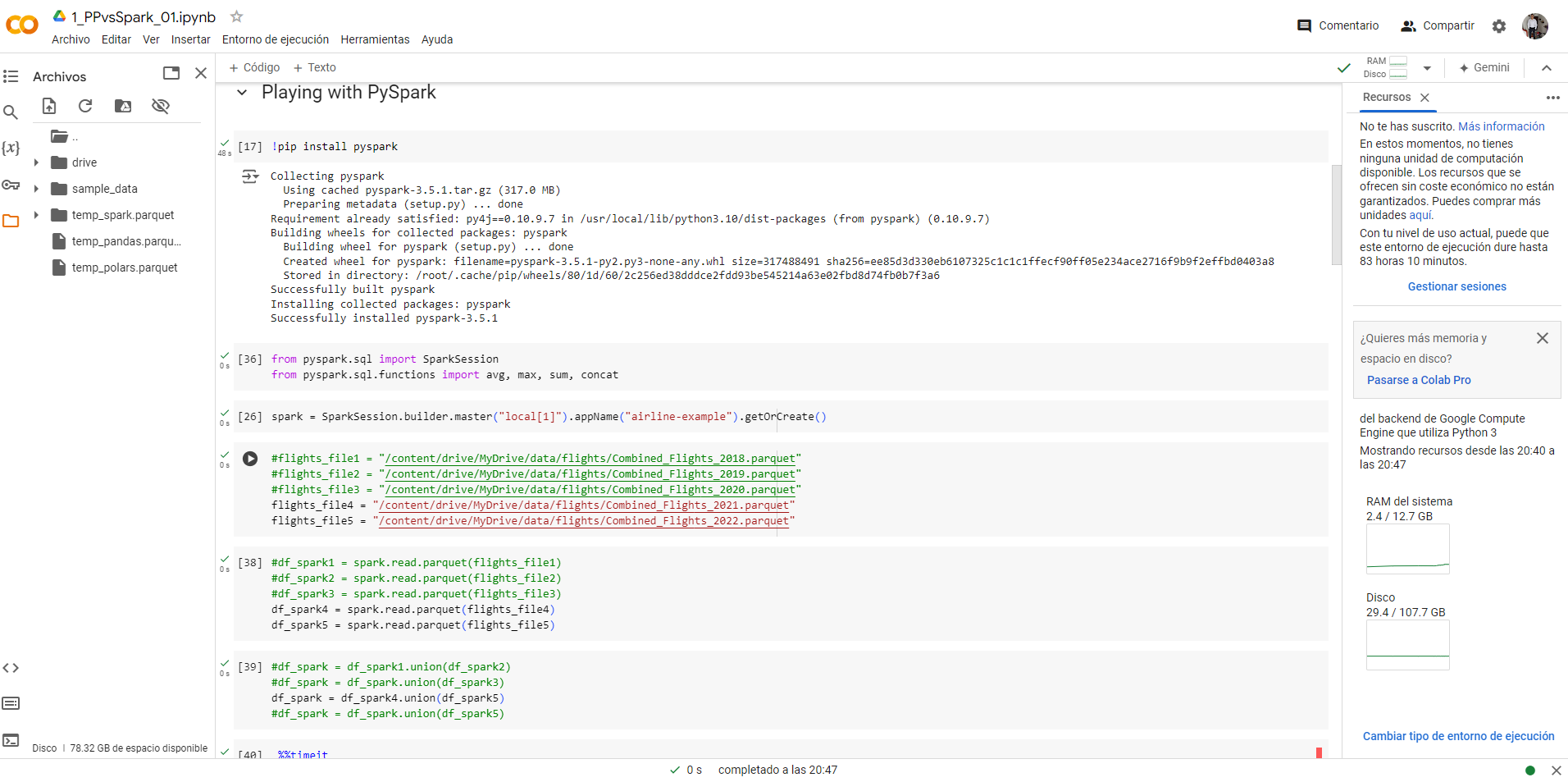
| **Método****PySpark** | **# de prueba** | **Datos utilizados** | **Resultados** |
| --- | --- | --- | --- |
| Prueba 1 | **flights\_file3**  **flights\_file4**  **df3**  **df4** | 46 segundos  Proceso ejecutado con éxito  RAM: 2.0Gb  Disco: 29.4 Gb |
| Prueba 2 | **flights\_file4**  **flights\_file5**  **df4**  **df5** | 29 segundos  Proceso ejecutado con éxito  RAM : 2.3 Gb  Disco : 29.4 Gb |
| Prueba 3 | **flights\_file1**  **flights\_file5**  **df1**  **df5** | 26 segundos  Proceso ejecutado con éxito  RAM : 2.3 Gb  Disco : 29.4 Gb |
| Prueba 4 | **flights\_file2**  **flights\_file3**  **df2**  **df3** | 33 segundos  Proceso ejecutado con éxito  RAM : 2.3 Gb  Disco : 29.4 Gb |

**Prueba 1**

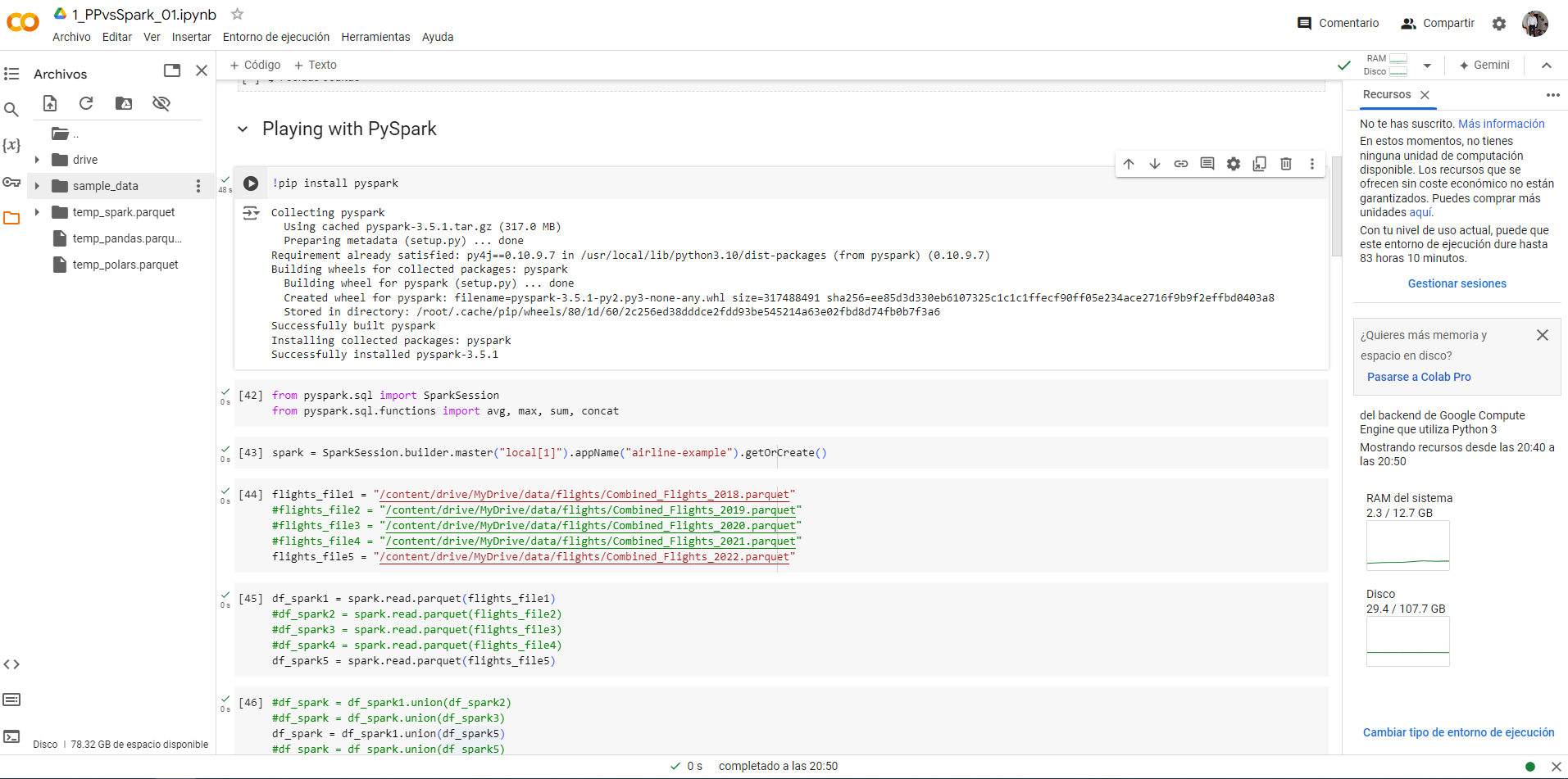




**Prueba 2**

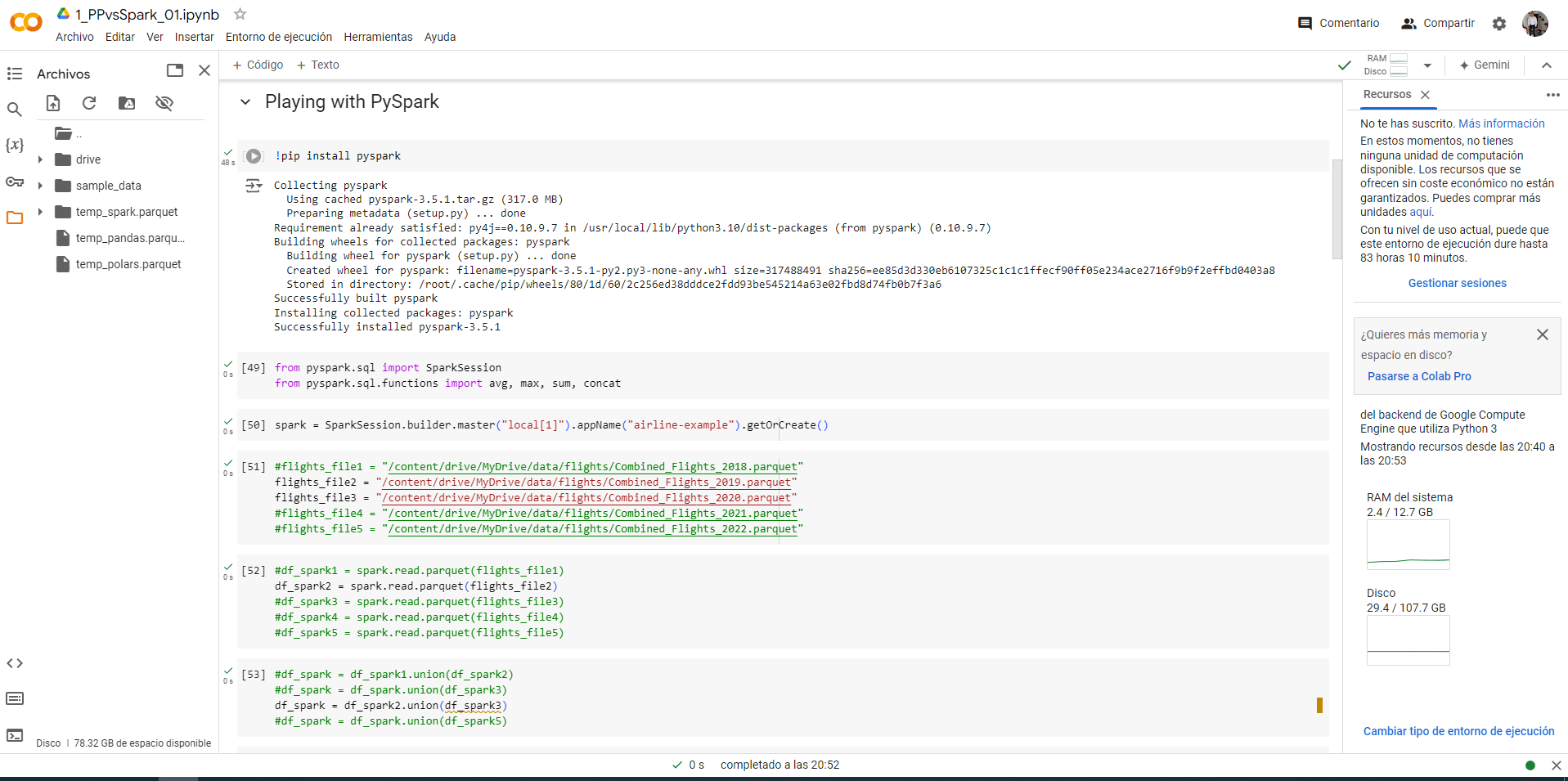
****

**Prueba 3**



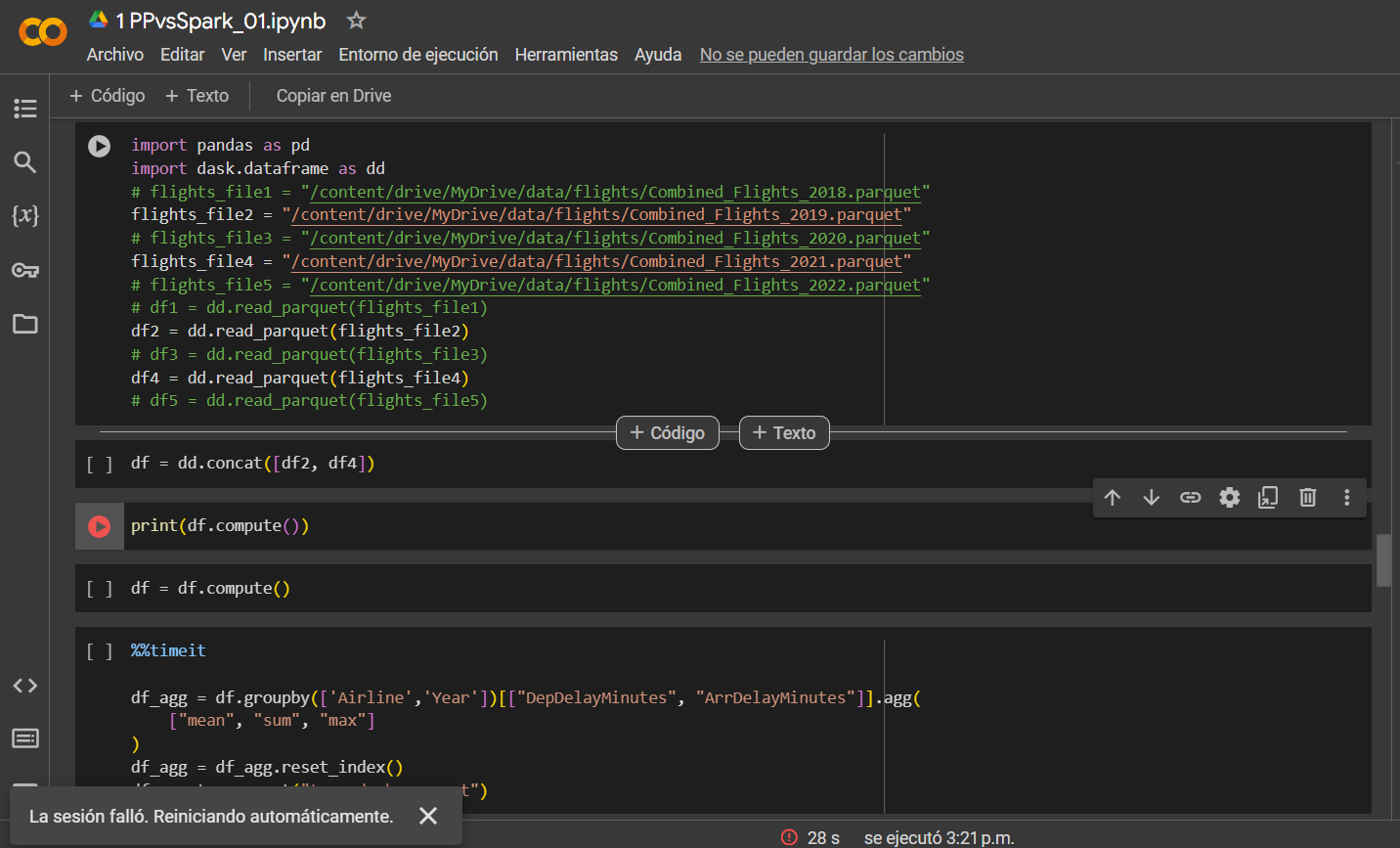


**Prueba 4**

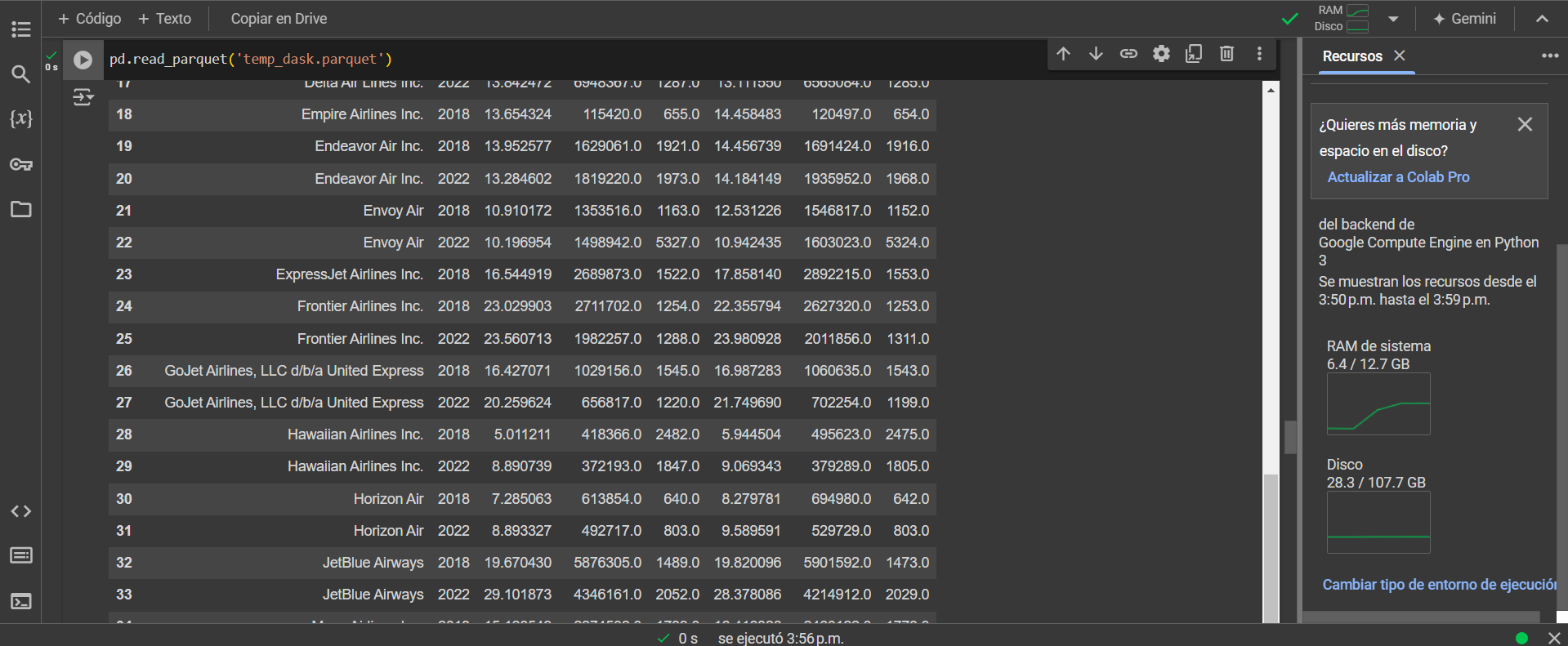
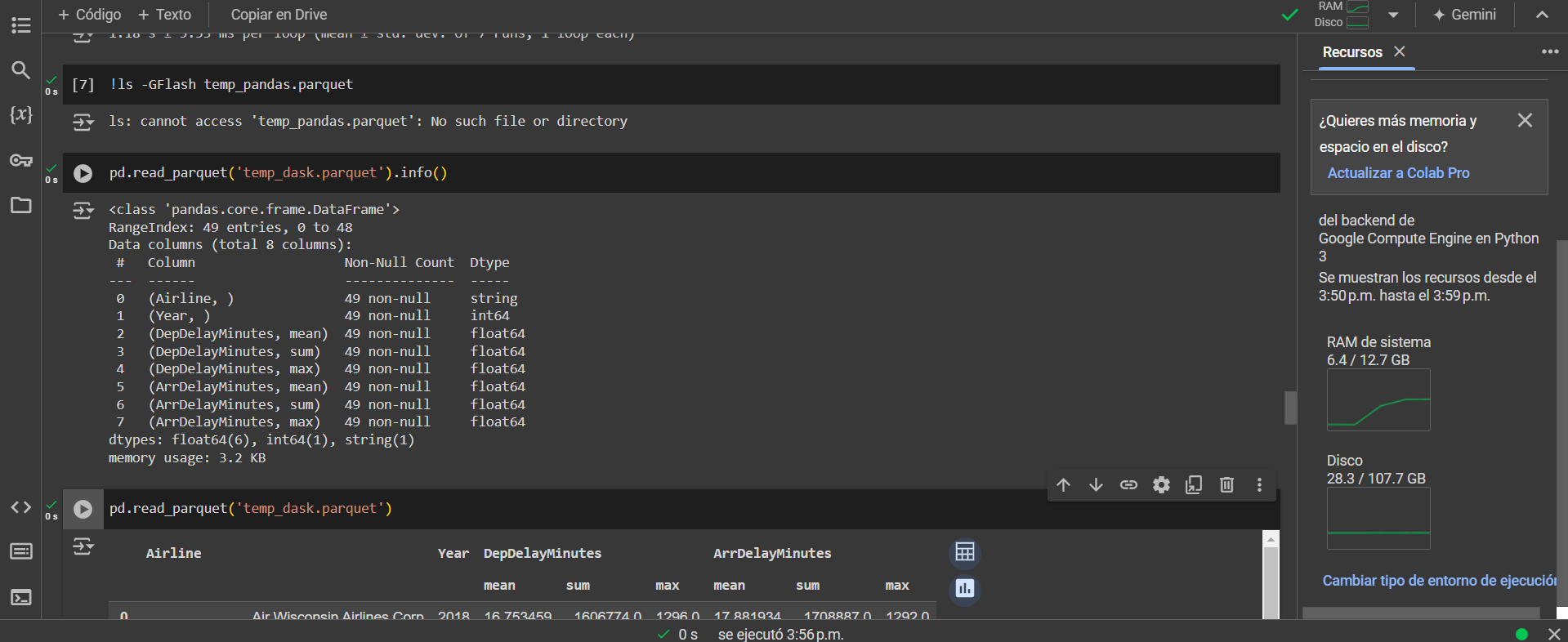
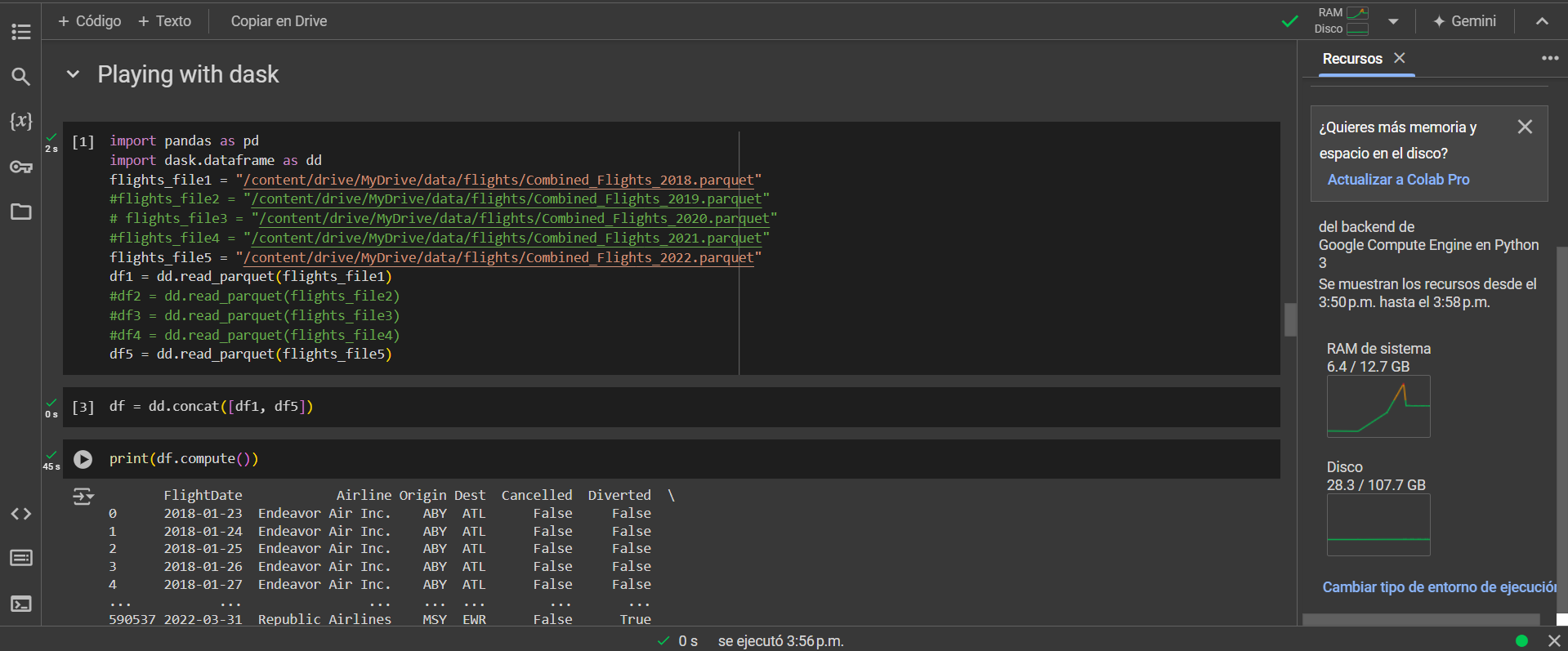


| **Método****Dask** | **# de prueba** | **Datos utilizados** | **Resultados** |
| --- | --- | --- | --- |
| Prueba 1 | **flights\_file2**  **flights\_file4**  **df3**  **df4** | 29 segundos  Limite de RAM alcanzado en el paso 3 |
| Prueba 2 | **flights\_file1**  **flights\_file5**  **df4**  **df5** | 77 segundos  Proceso ejecutado con éxito  RAM : 6.4 Gb  Disco : 28.3 Gb |
| Prueba 3 | **flights\_file2**  **flights\_file5**  **df2**  **df5** | 51 segundos  Proceso ejecutado con éxito  RAM : 6.5 Gb  Disco : 28.3 Gb |
| Prueba 4 | **flights\_file1**  **flights\_file3**  **df1**  **df3** | 65 segundos  Proceso ejecutado con éxito  RAM : 7.1 Gb  Disco : 28.5 Gb |

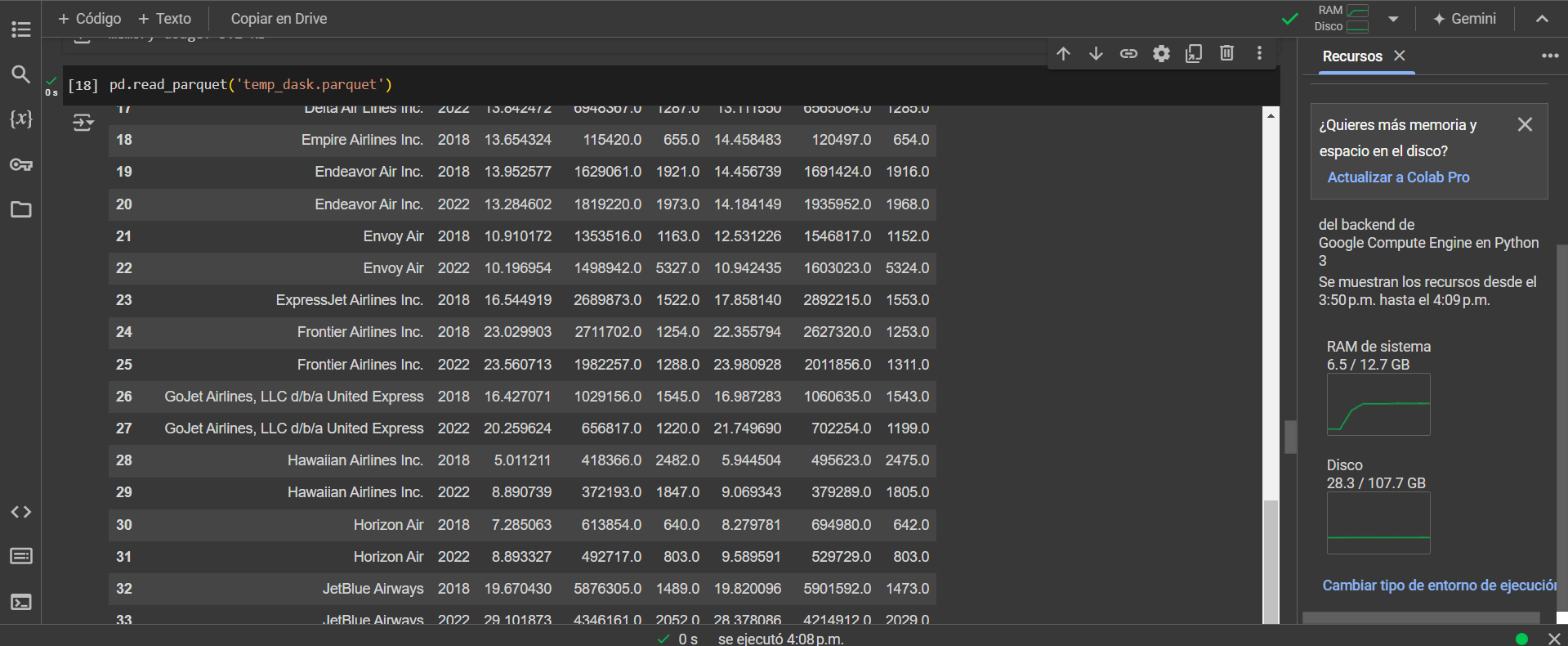
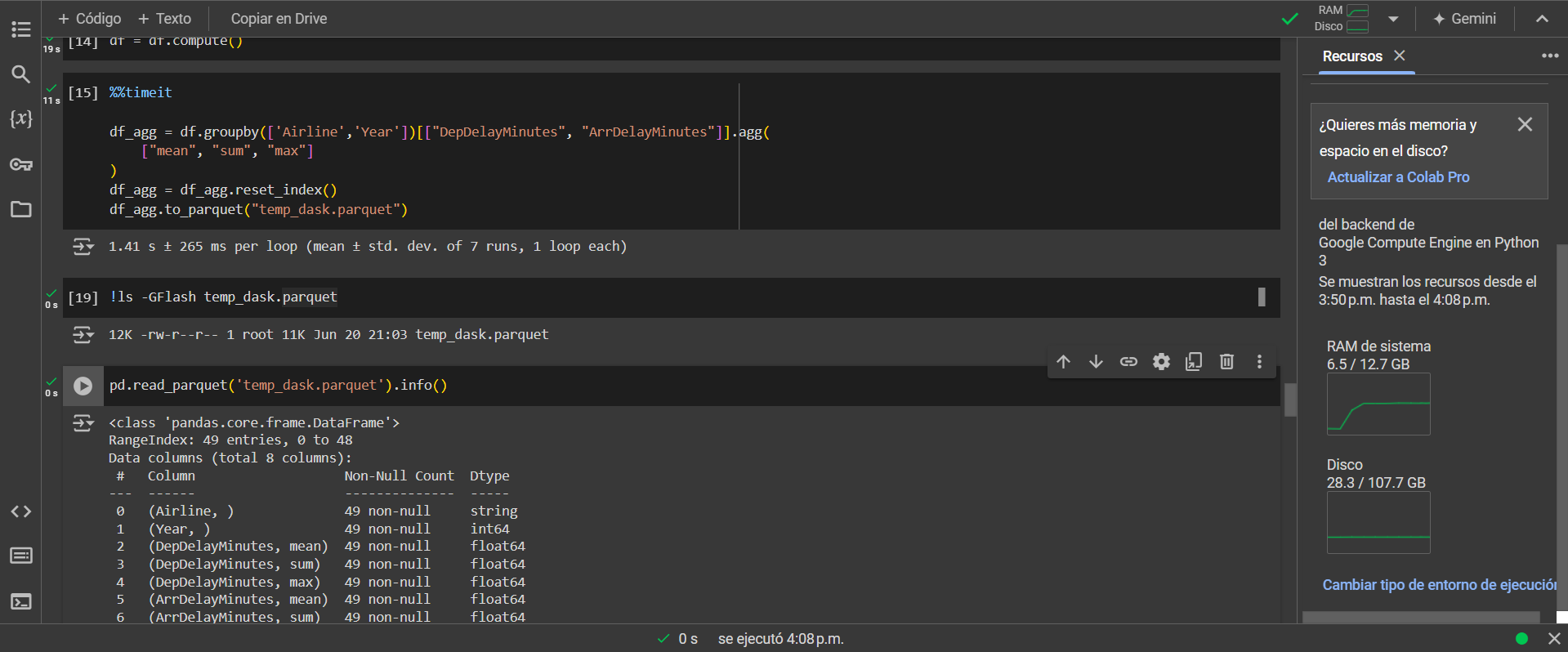
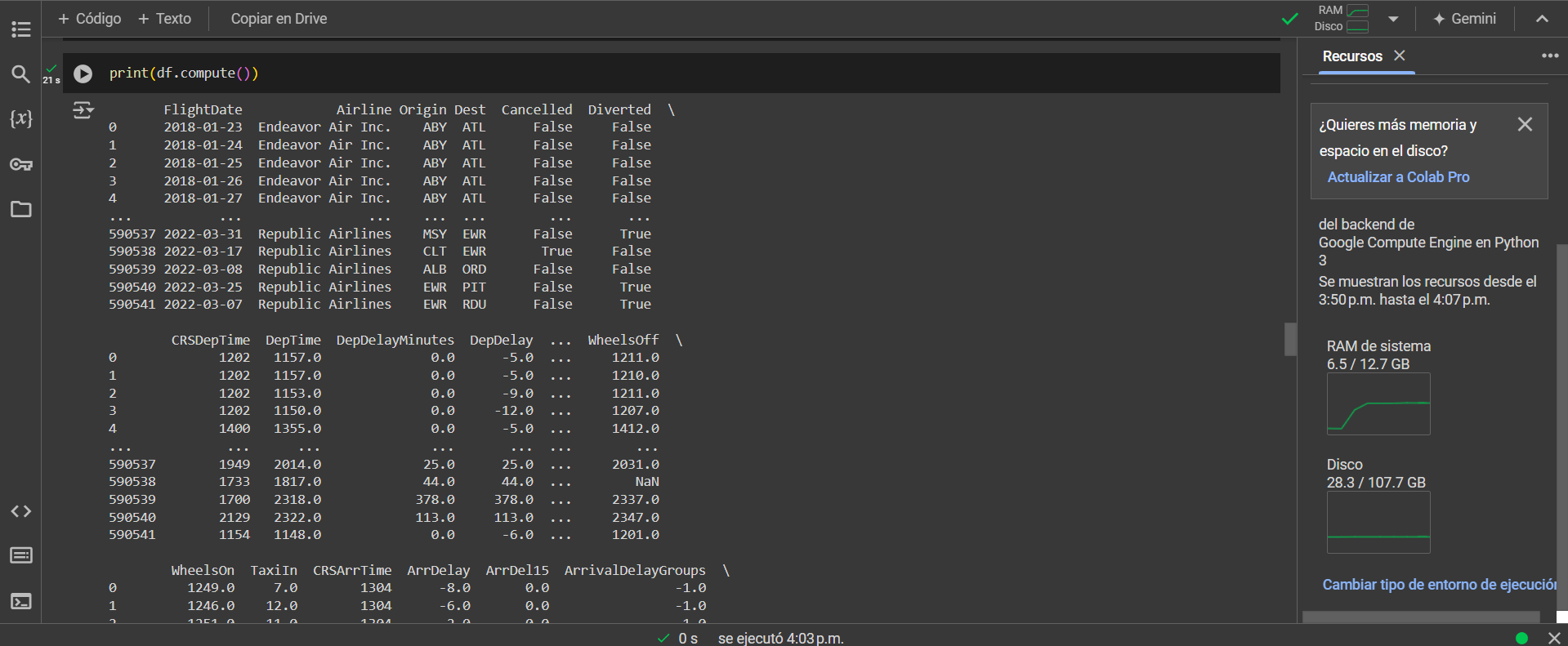
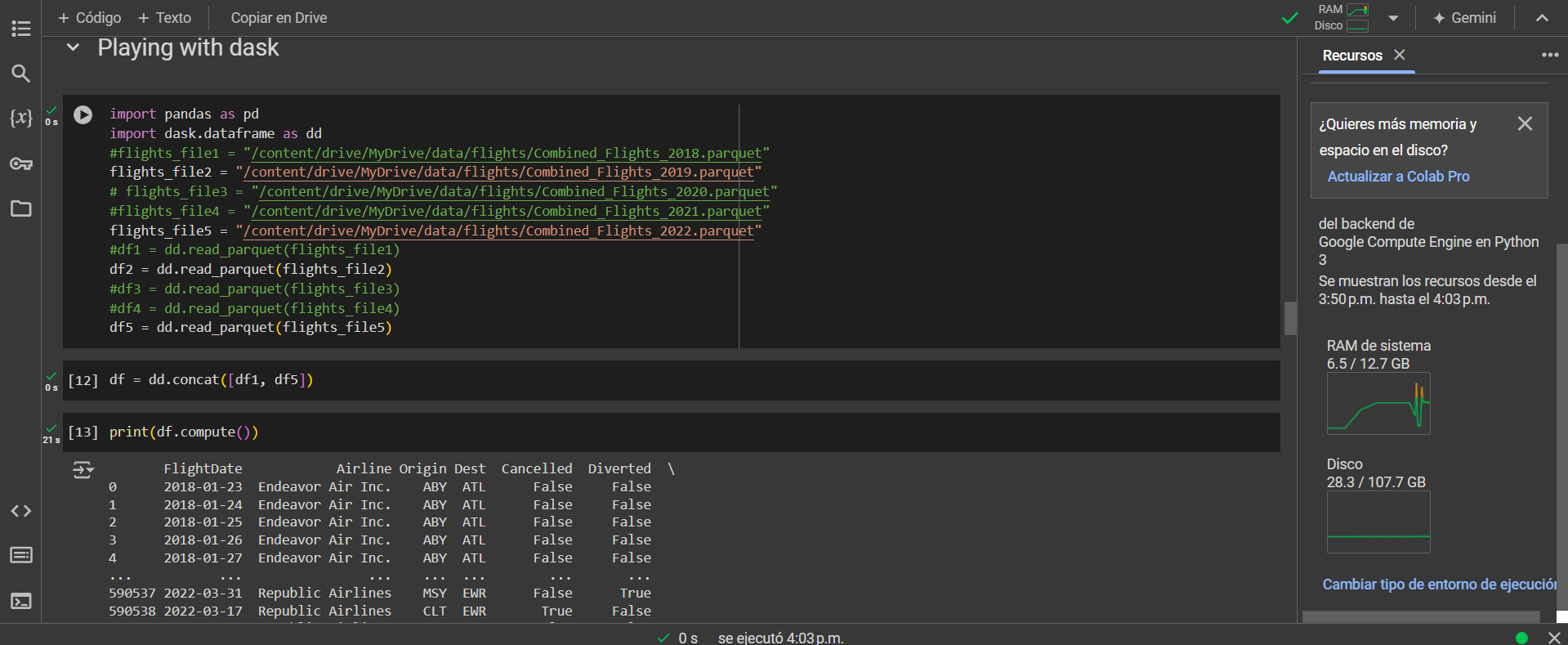
Prueba 1



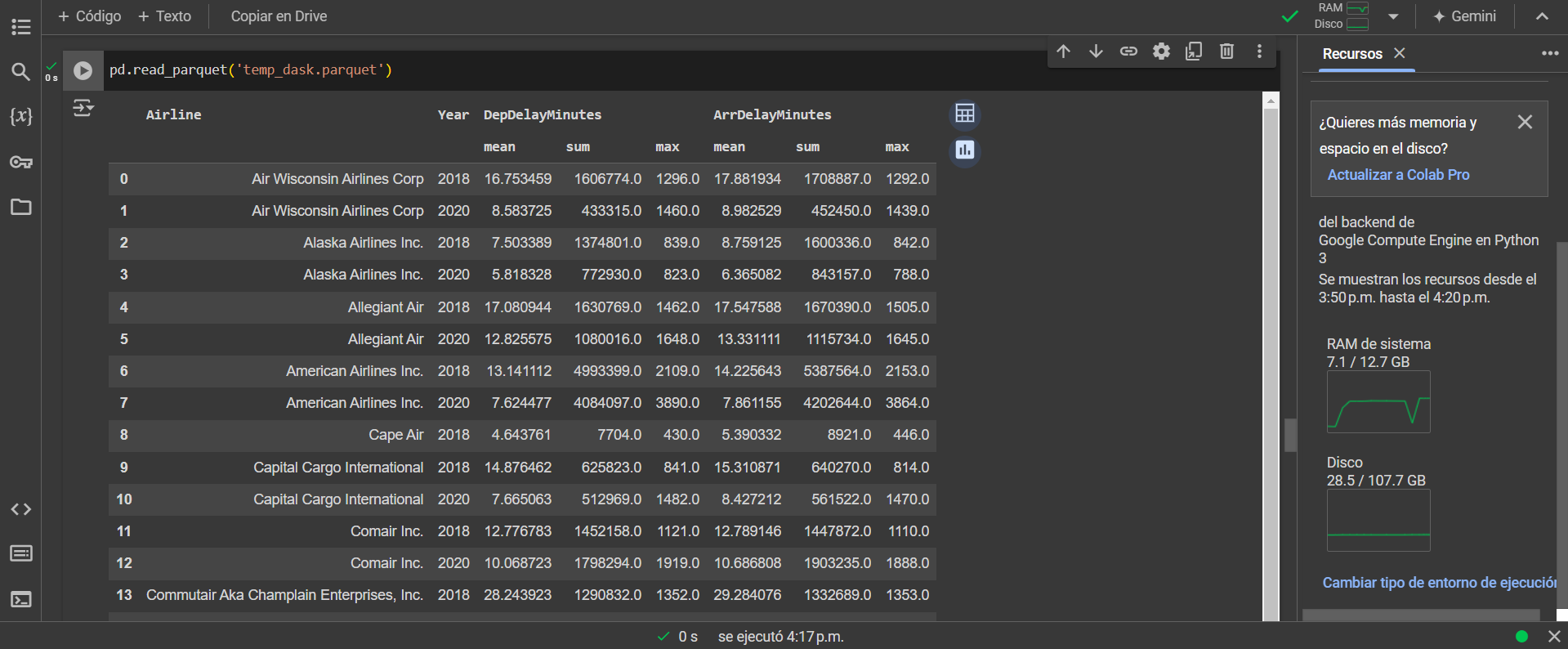
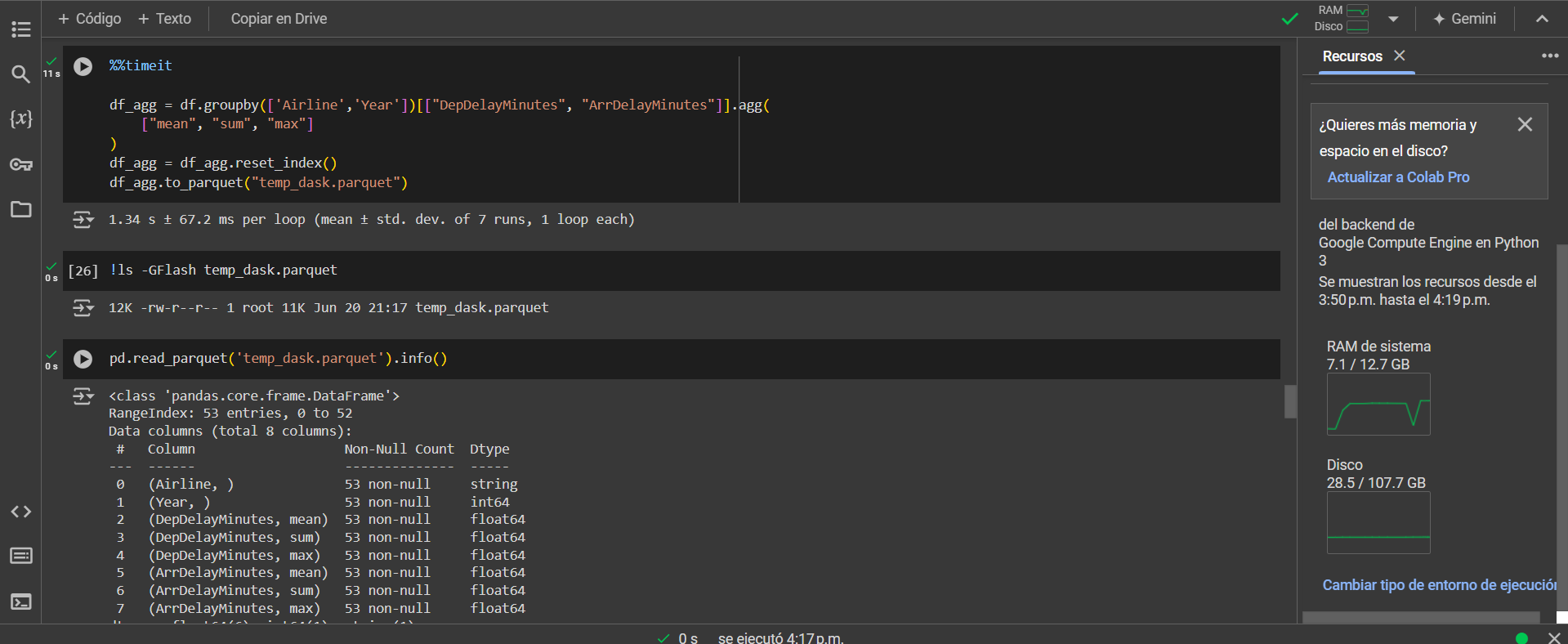
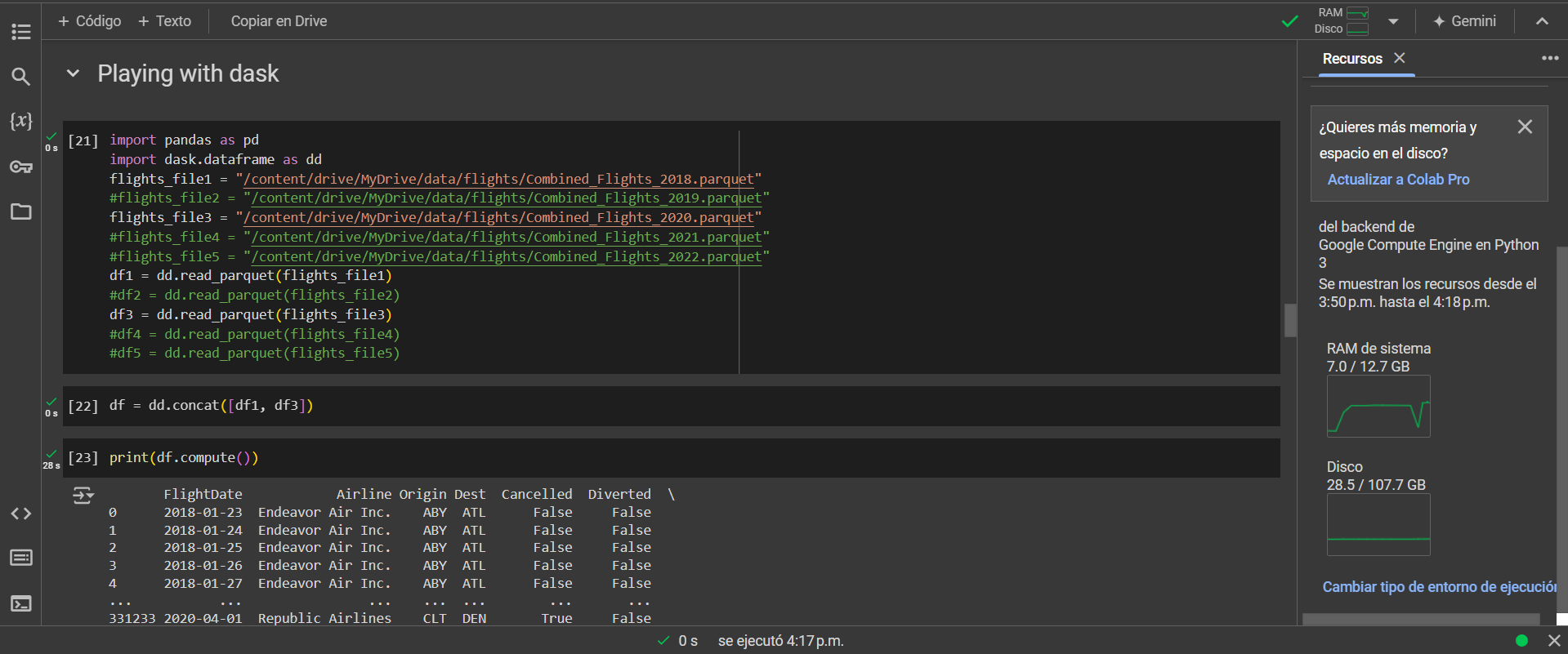
**Prueba 2**

****

**Prueba 3**

****

**Prueba 4**

****

**Conclusiones**

Luego de realizar los análisis con cada una de las librerías y con varias combinaciones de datos, podemos realizar las siguientes conclusiones de este laboratorio:

* El método polars es el más flexible y sencillo con los datos ya que este trabaja con pocos recursos tanto en programación como como en hardware
* El método pandas terminó siendo el que tuvo más complicaciones para realizar el análisis de los datos dado a que de las 4 pruebas, solo una pudo realizarse con éxito
* Los grupos de datos más flexibles de trabajar fueron los de 2018 (flights\_file1) y 2020 (flights\_file1)
* El grupo de datos con más inconvenientes para trabajar fueron los de 2019 (flights\_file2) debido a que estos mismos eran los que hacían sobrepasar el límite de RAM con más frecuencia a comparación de los demás grupos de datos
* El método PySpark fue el método más equilibrado en cuanto a uso de RAM con un promedio entre las 2 Gb y eficiencia de manejo de datos. Esto está demostrado con el hecho de que los métodos Pandas y Dusk superan las 7 Gb de RAM y el método Polars maneja usos entre la 1 y 2 Gb pero su manejo de datos es muy sencillo.