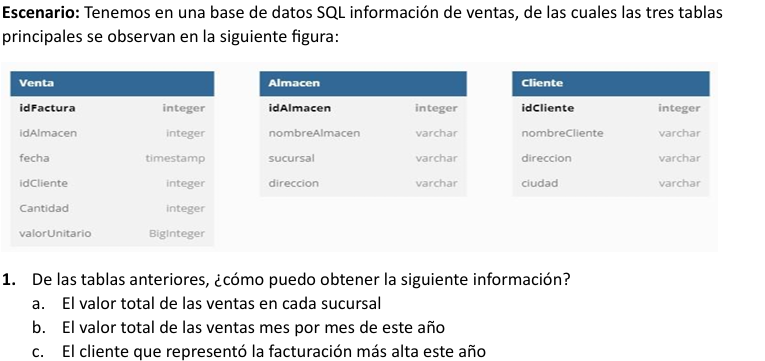
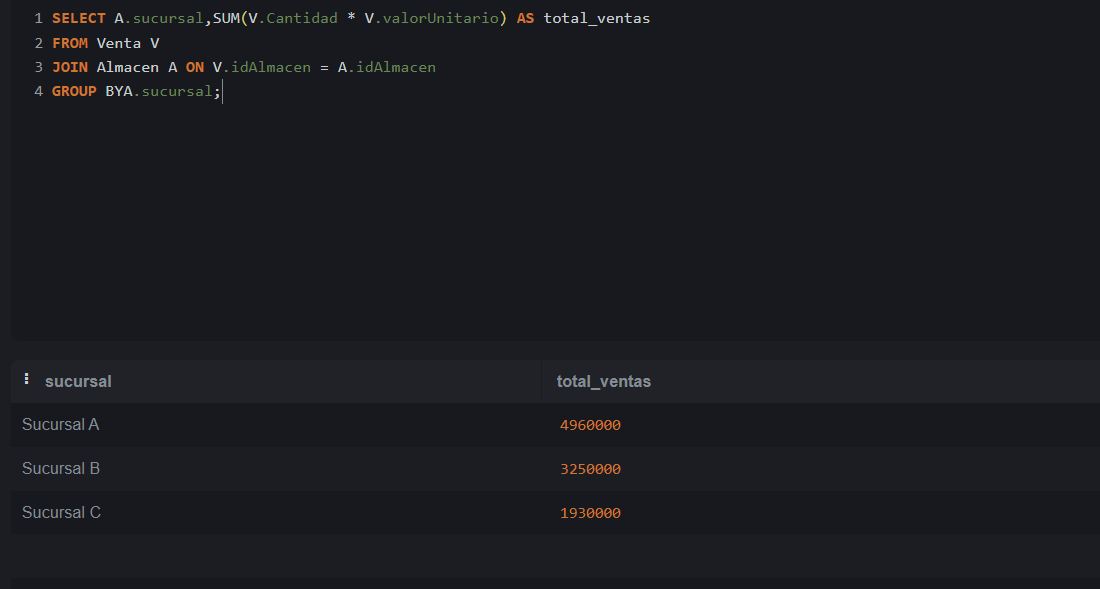
Prueba de Ingreso – Ingeniero de Datos CredibanCo

Solucionado por Diego chinchilla

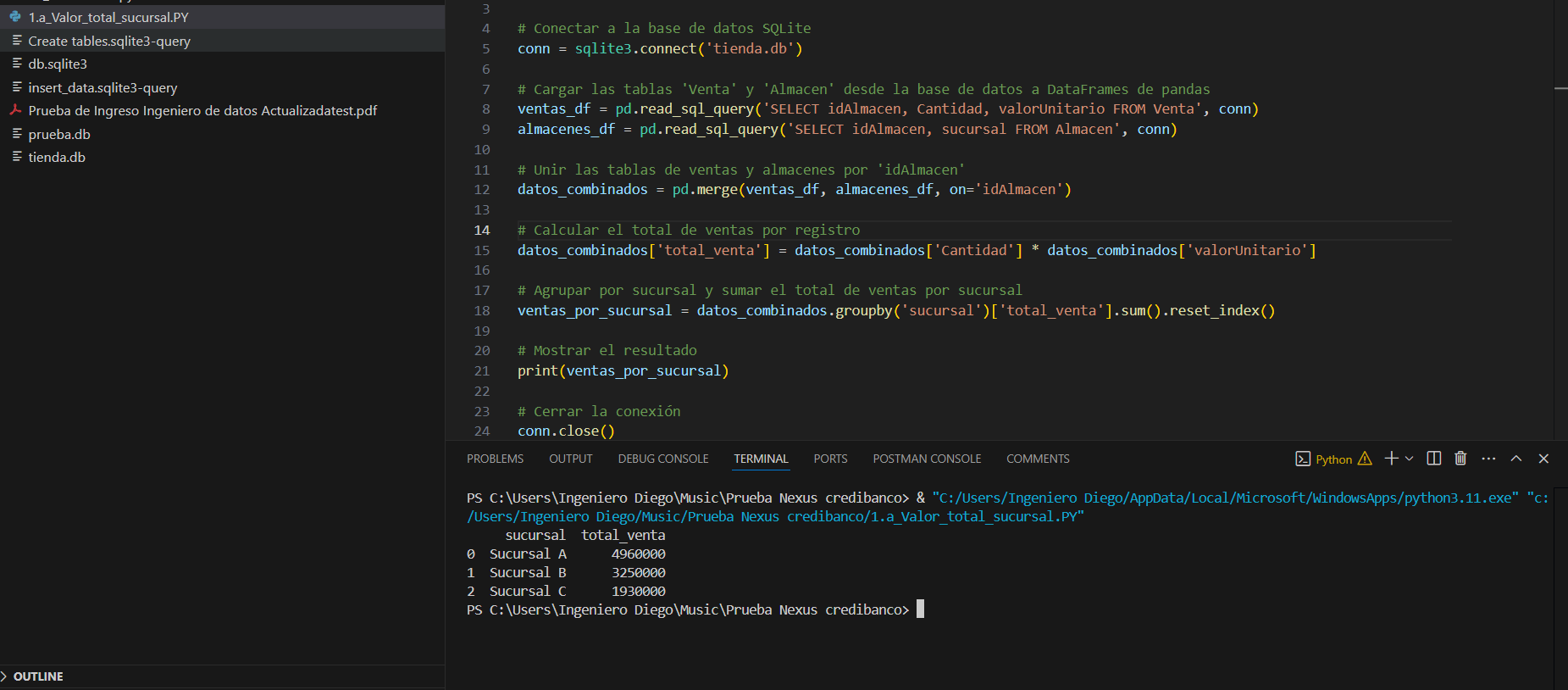
Para el desarrollo de la prueba se utiliza la herramienta online <https://sqliteonline.com/> en la cual se crean las tablas para el desarrollo de la prueba, esta herramienta modela y ejecuta código sql , también se genera el uso de sql lite en ambiente local recreadndo la creación de las tablas para su posterior consumo con Python como s con el fin de emular el escenario planteado a continuación:



Punto 1ª Vlaor toral de ventas por sucursal



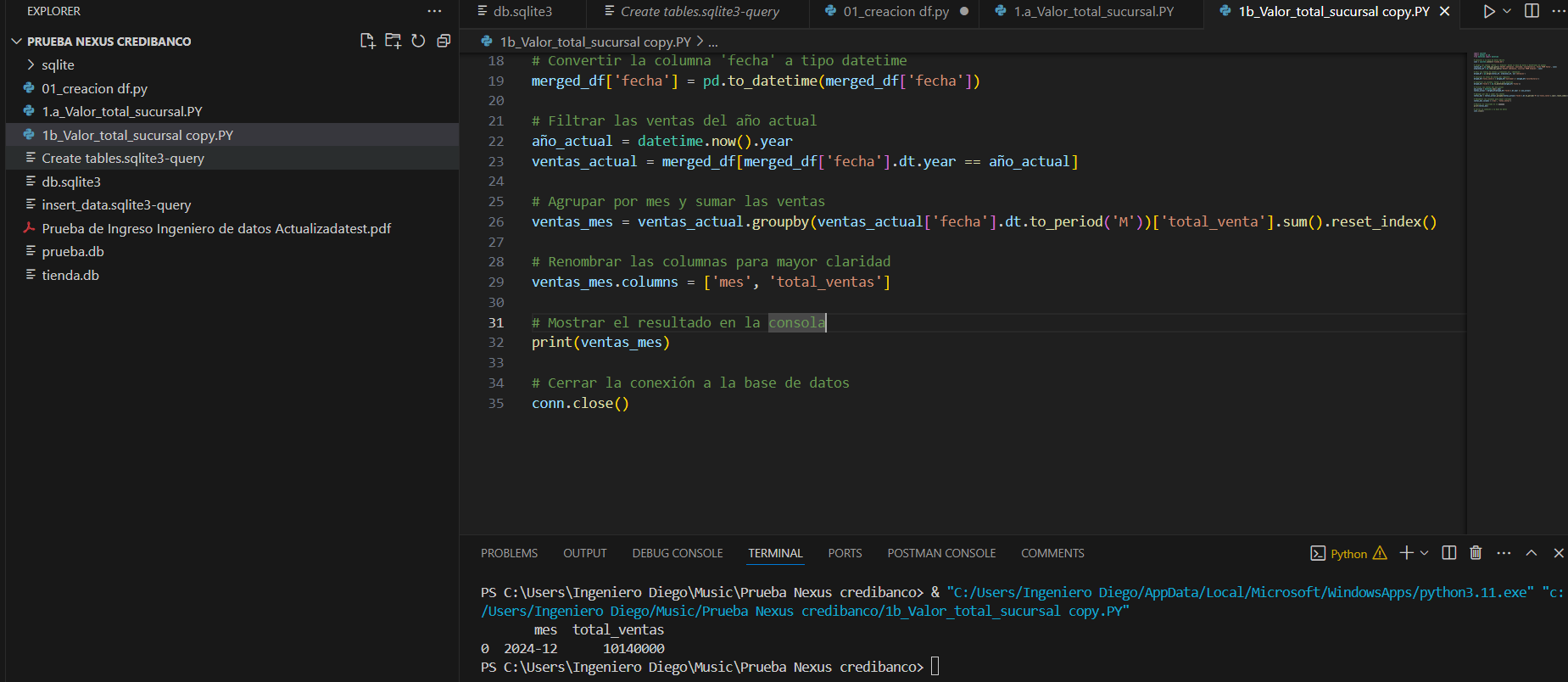
Solución en Python punto 1a



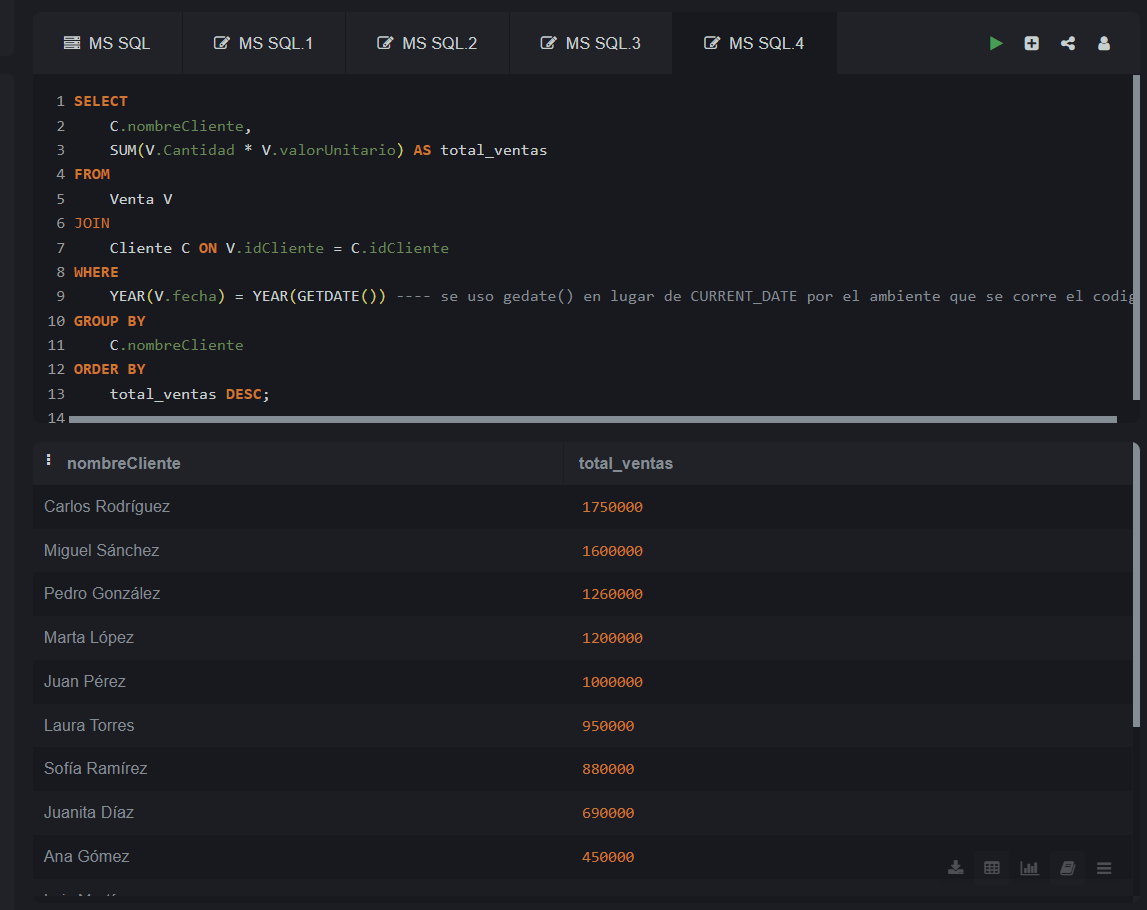
Punto 1.b valor total de ventas este año

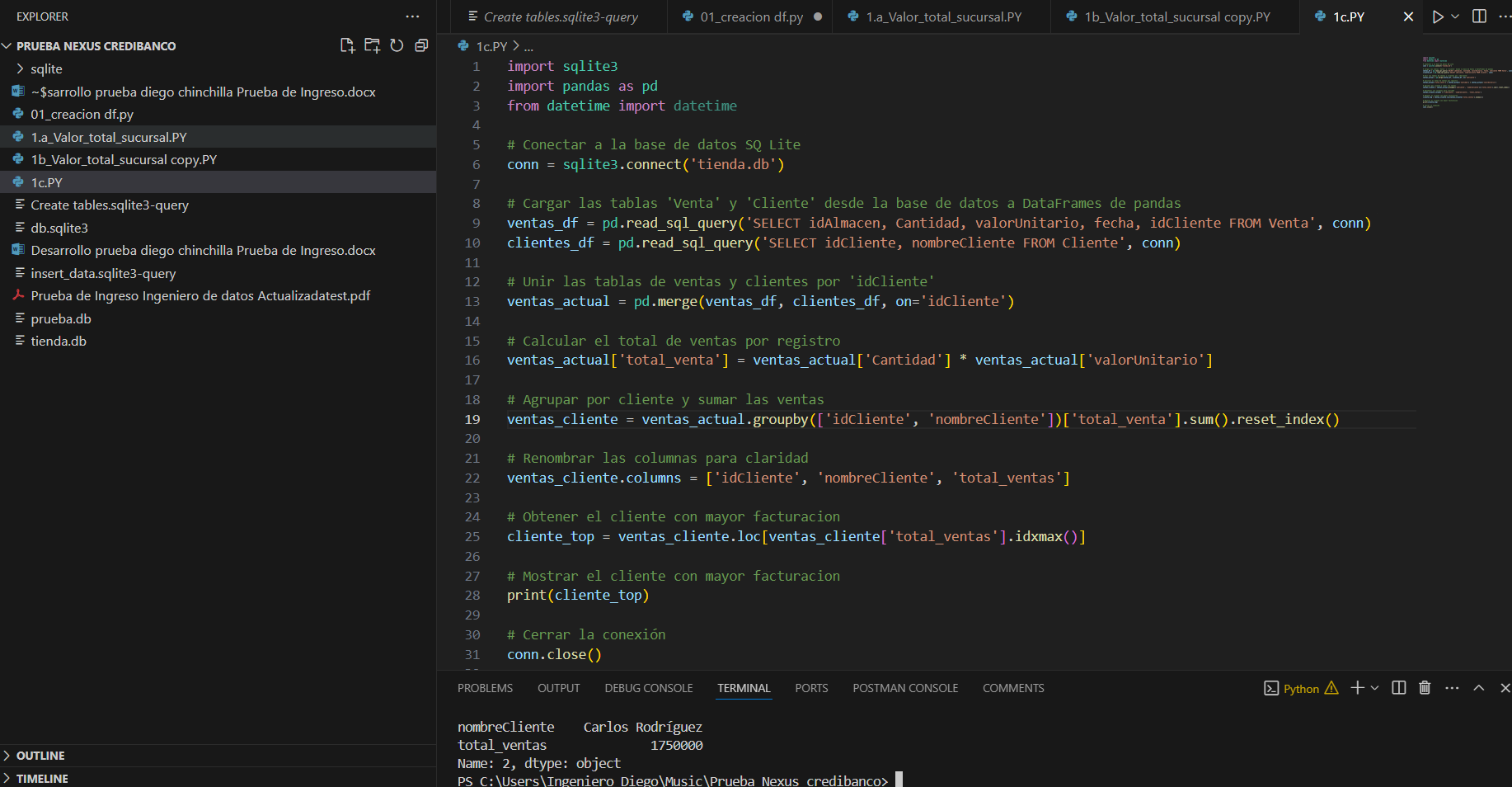


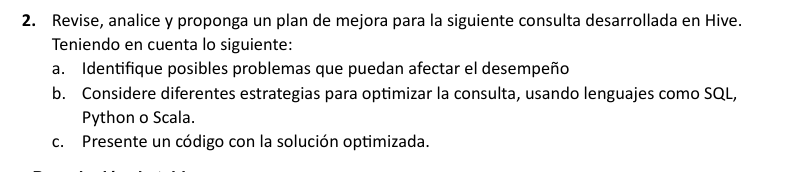
Solución en Python 1b



Punto 1.C Cliente que más Facturo



Prueba en Python 

Respuesta punto 2 

**Punto a:**

Se genera la identificación problema de desempeño revisando que de entrada el uso del distinct puede generar problemas de rendimiento al momento de su ejecución y un gran consumo de memoria y almacenamiento temporal en disco .

Se valida y se están usando subconsultas anidadas donde a lo largo de tiempo pueden ser difíciles de optimizar e incluso dañar el flujo de datos

Funciones de ventana donde el uso del rank () y el row\_number() generan sobrecosto de utlizacion de recursos que al momento de generar procesamiento incremental puede generar caídas en la arquitectura montada.

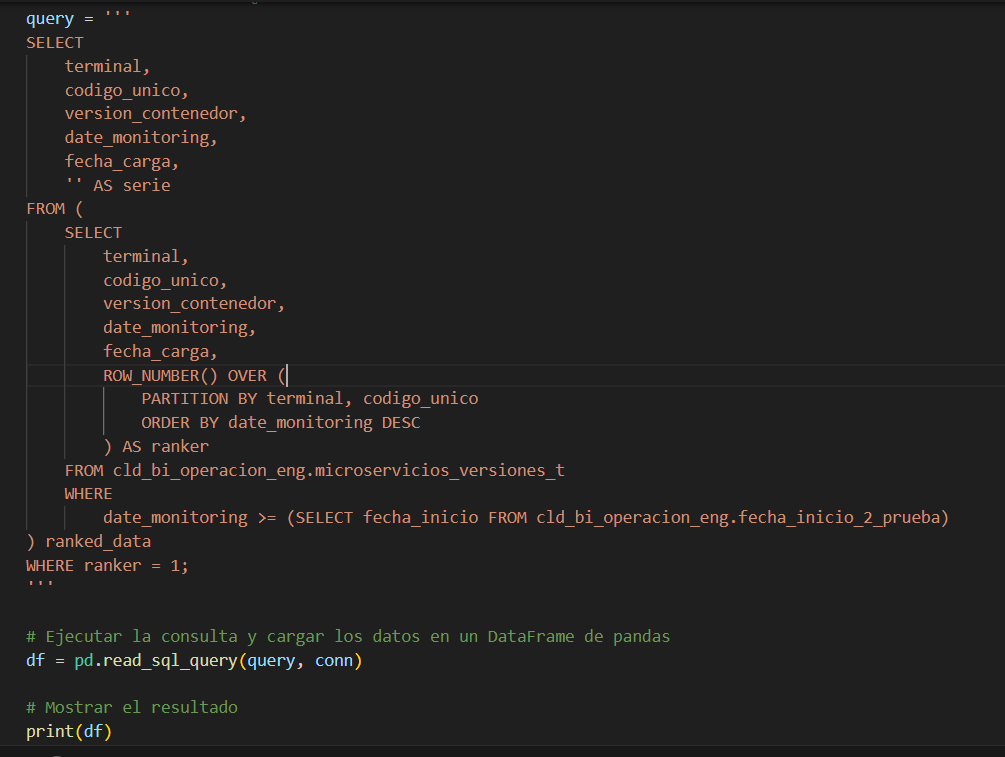
Formato inadecuado del uso de from\_unixtime y unix\_timestamp dadoa que se aplican repetidamente

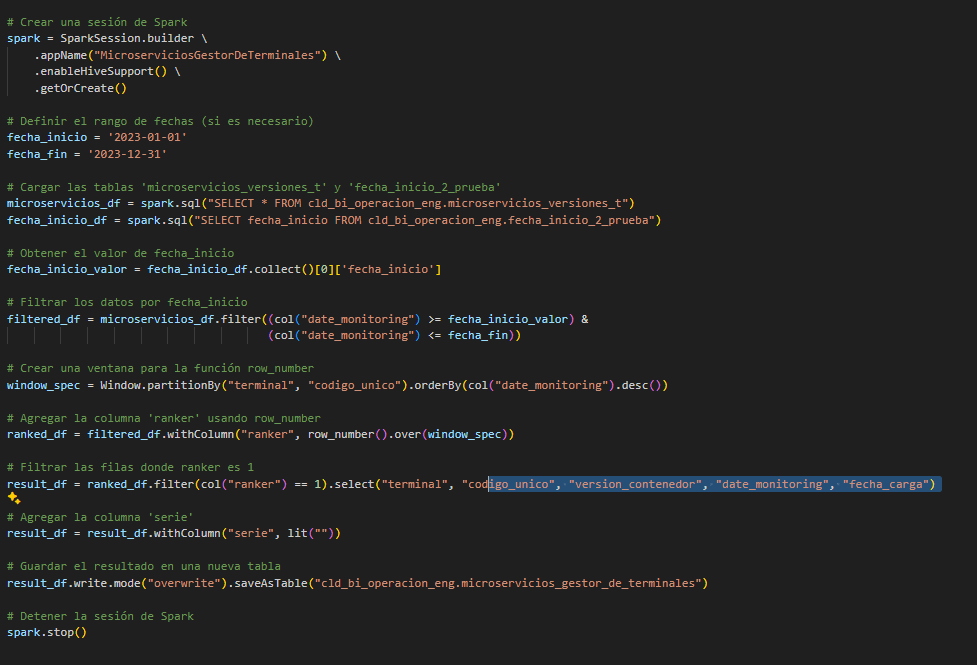
**Punto b**

Estrategias de mejora que se plantean

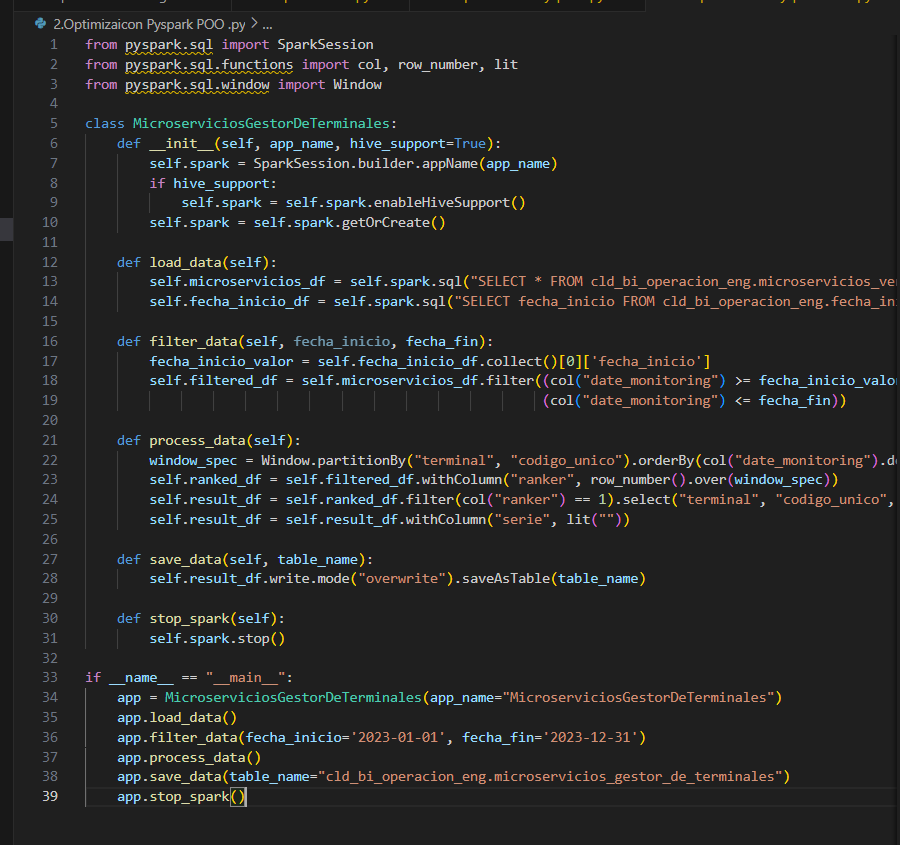
* Eliminar el y generar una consulta con combinación adecuada de llaves para poder aplicar un groupby .
* Eliminación de formatos de fecha que no sean necesarios en cada fila
* Generar un filtros necesarios para limitación de datos al momento de procesar y no generar sobrecargar al cluster de datos.
* La Función ROW\_NUMBER() para asignar un número único a cada fila en un grupo de registros, en función de combinaciones de terminal y código\_único. Los registros se ordenan por date\_monitoring en orden descendente, asegurando que la línea más reciente tenga el número de línea 1. Luego, los registros se filtran para mantener solo las líneas más recientes usando ranking = 1. Además, se incluye un filtro de fecha para garantizar que solo los registros con una fecha mayor o igual a un valor específico obtenido de una subconsulta, reduciendo la cantidad de datos procesados. Finalmente, se seleccionan las columnas requeridas, incluida una columna serial adicional inicializada como vacía. Esta optimización mejora el rendimiento al reducir la cantidad de datos procesados, hace que la consulta sea más clara y fácil de entender y proporciona mayor flexibilidad con los rangos de fechas. En resumen, la mejora garantiza consultas más eficientes y claras, con un rendimiento mejorado y menos carga en el sistema.

Solucion en Python consuminoto query

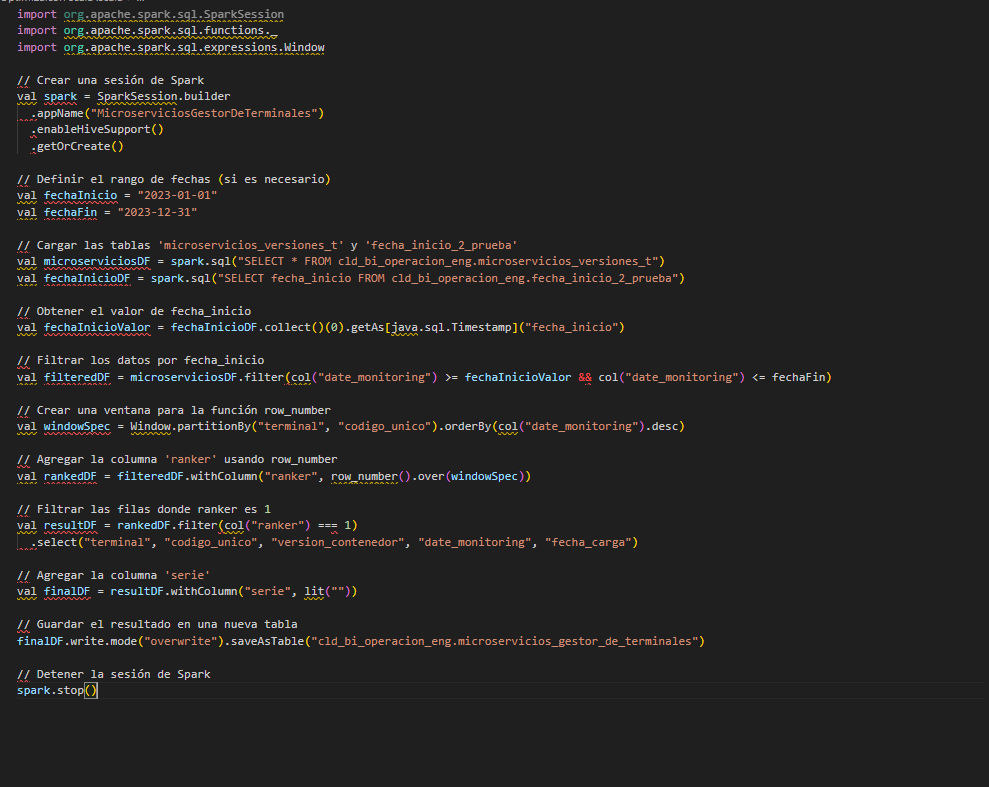


Solucion en pyspark 

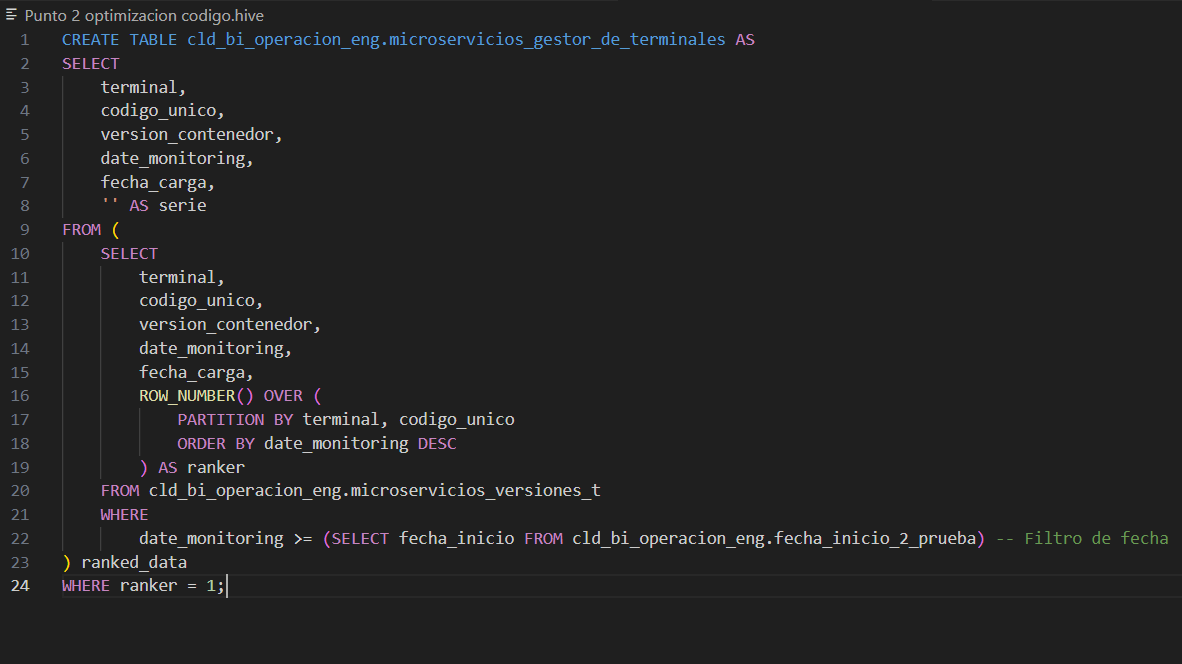
Solucion en pyspark con POO



Solucion scala



**Punto C**



En resumen

La Función ROW\_NUMBER() para asignar un número único a cada fila en un grupo de registros, en función de combinaciones de terminal y código\_único. Los registros se ordenan por date\_monitoring en orden descendente, asegurando que la línea más reciente tenga el número de línea 1. Luego, los registros se filtran para mantener solo las líneas más recientes usando ranking = 1. Además, se incluye un filtro de fecha para garantizar que solo los registros con una fecha mayor o igual a un valor específico obtenido de una subconsulta, reduciendo la cantidad de datos procesados. Finalmente, se seleccionan las columnas requeridas, incluida una columna serial adicional inicializada como vacía. Esta optimización mejora el rendimiento al reducir la cantidad de datos procesados, hace que la consulta sea más clara y fácil de entender y proporciona mayor flexibilidad con los rangos de fechas. En resumen, la mejora garantiza consultas más eficientes y claras, con un rendimiento mejorado y menos carga en el sistema.

Herramientas usadas

* <https://sqliteonline.com/>
* sqlite3
* vscode
* 