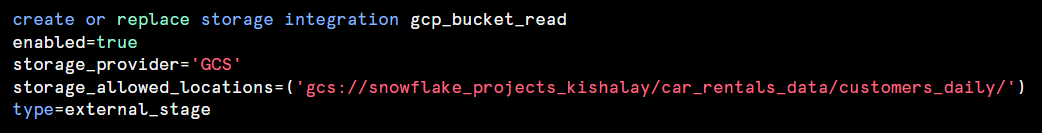
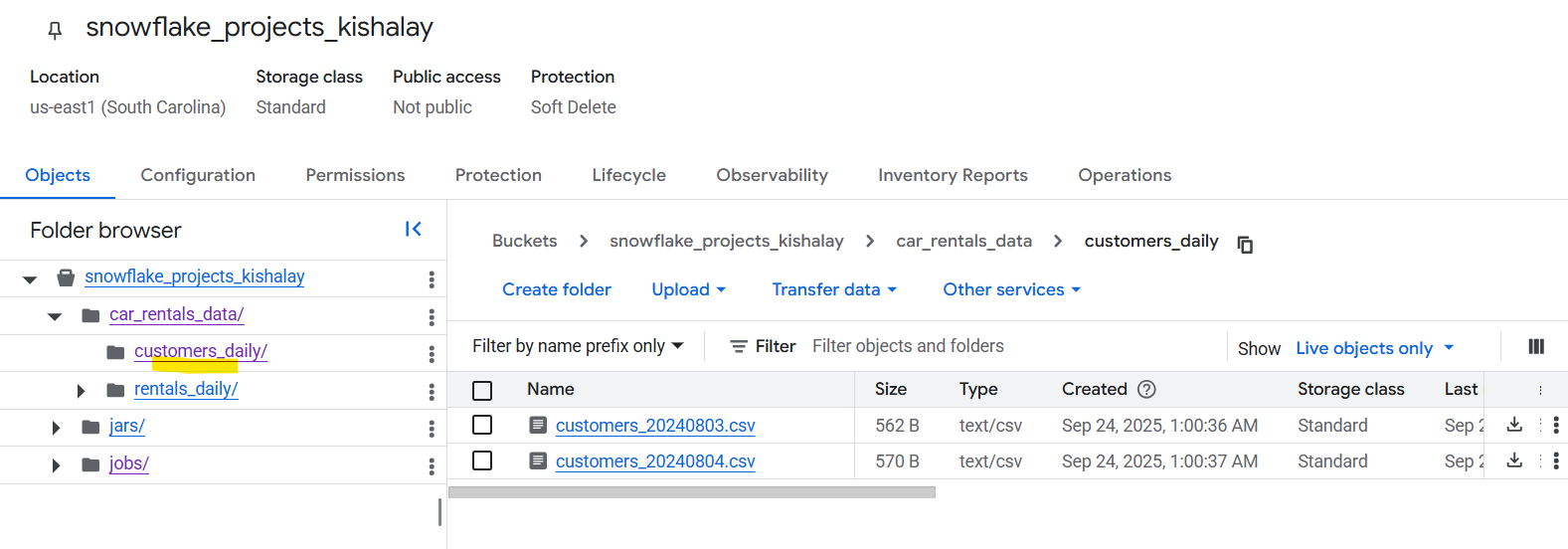
**Car\_rental\_data\_ingestion\_from\_GCS\_Storage\_to\_snowflake**

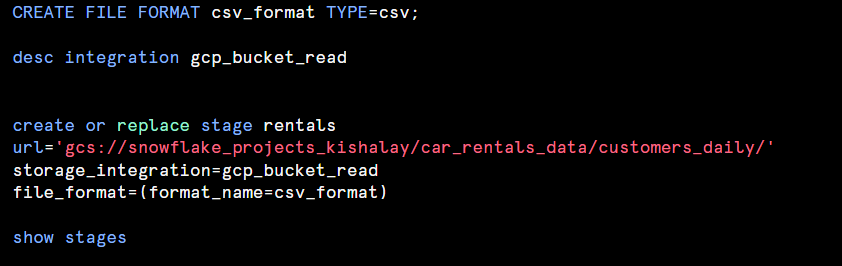
1. I created table definition of dim\_date,dim\_car,dim\_location,dim\_customer and fact\_rentals and established primary key and foreign key constraint. I also had inserted data for dim\_date,dim\_location,dim\_car from sql worksheet itself. It has been added in this repository.
2. I also created a storage integration in my snowflake database – car\_rental with a GCS bucket with the below prefix and created an external stage on top of it. The customers daily data which will be used to perform upsert operation will be placed in the given bucket location.





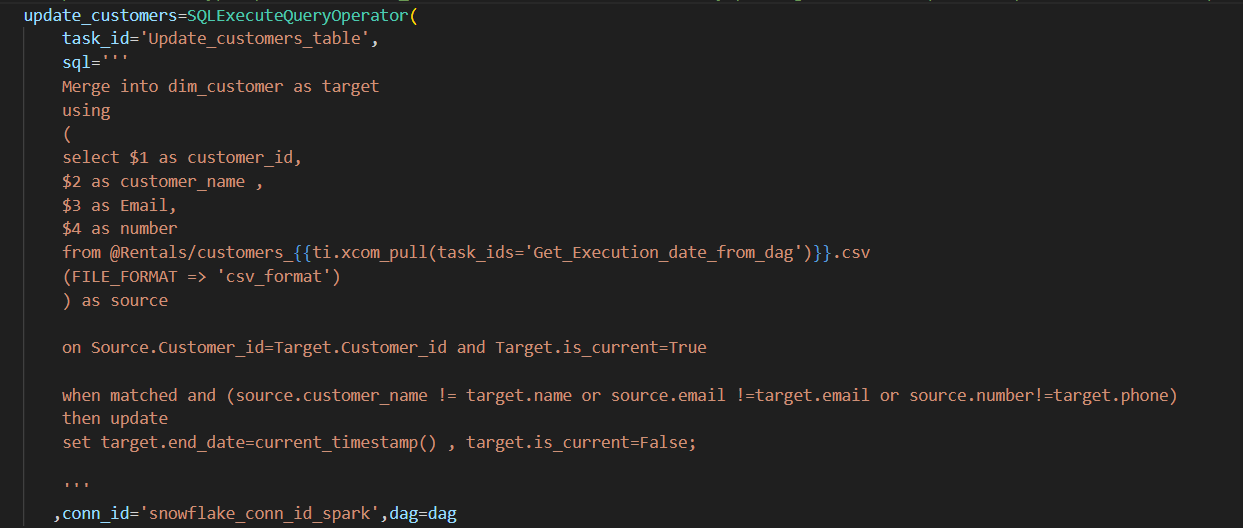
I provided necessary permissions to the service account returned by **desc integration gcp\_bucket\_read** so that snowflake service account can have read access to the above-mentioned gcs bucket.

1. I have created an external stage on top of this storage integration.

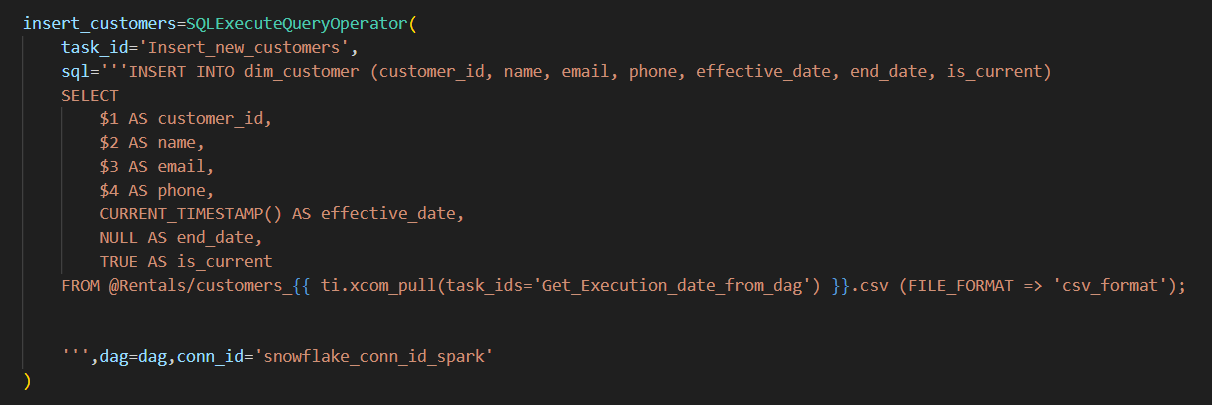


1. In airflow dag python file, I have configured the DAG with parameter which will take the date of the file user want to process in YYYYMMDD format, if user does not provide any date the PythonOperator is configured to get the date of execution date for the dag context and push it to xcom.

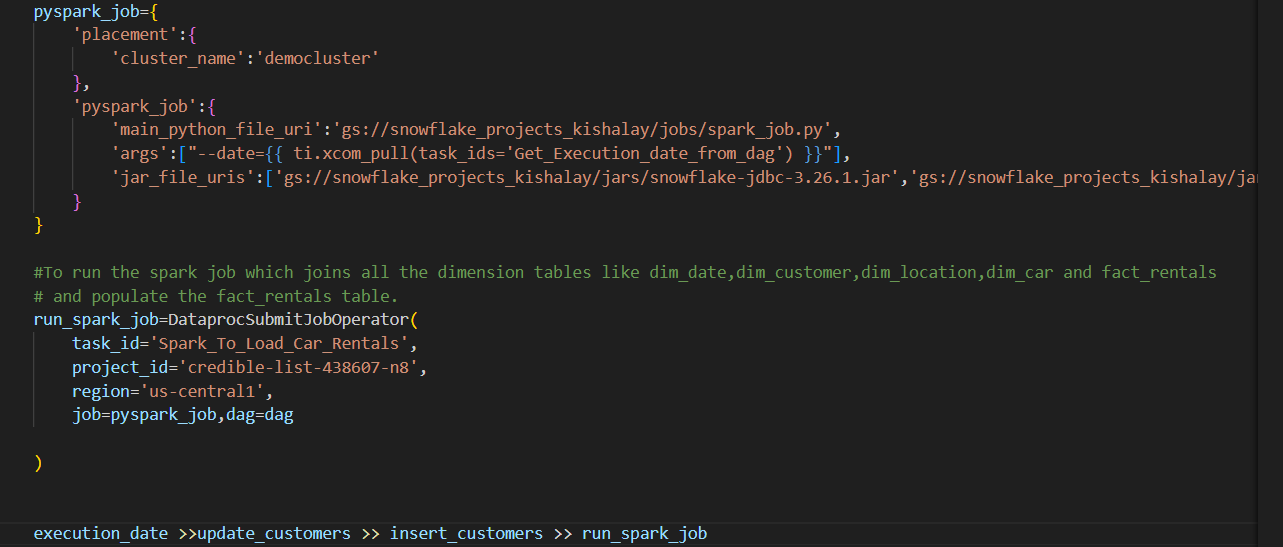
1. I am updating current records present in **dim\_customer** table pulling the data of customers table from the external stage(rentals) dynamically deciding the name of the file as per the parameter passed while running the DAG by user. I am assigning the end\_date for a given record and also marking the is\_current flag as False.



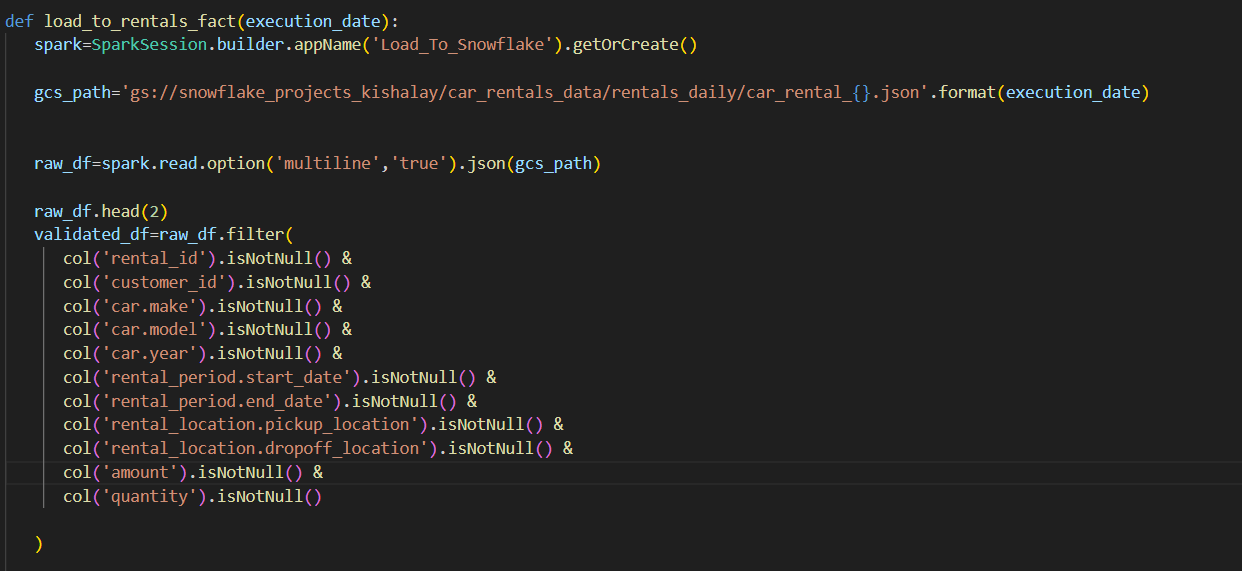
1. I am also inserting the new record for dim\_customer table with end\_date as Null and current timestamp as the start\_date for a given record.



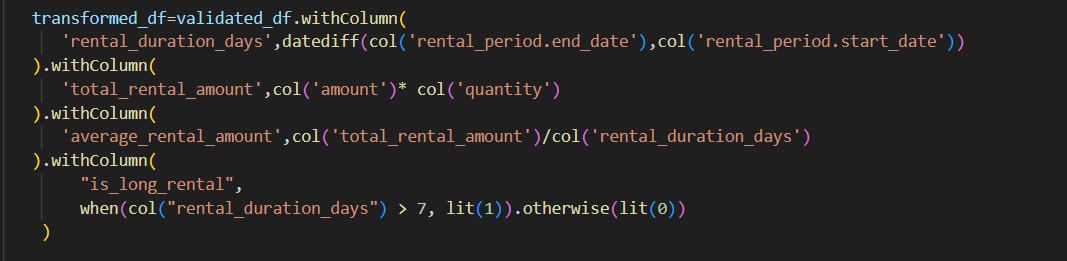
1. I am calling the spark job kept in GCS Storage bucket and also passing the storage path of the required jar files and submitting that to dataproc cluster.



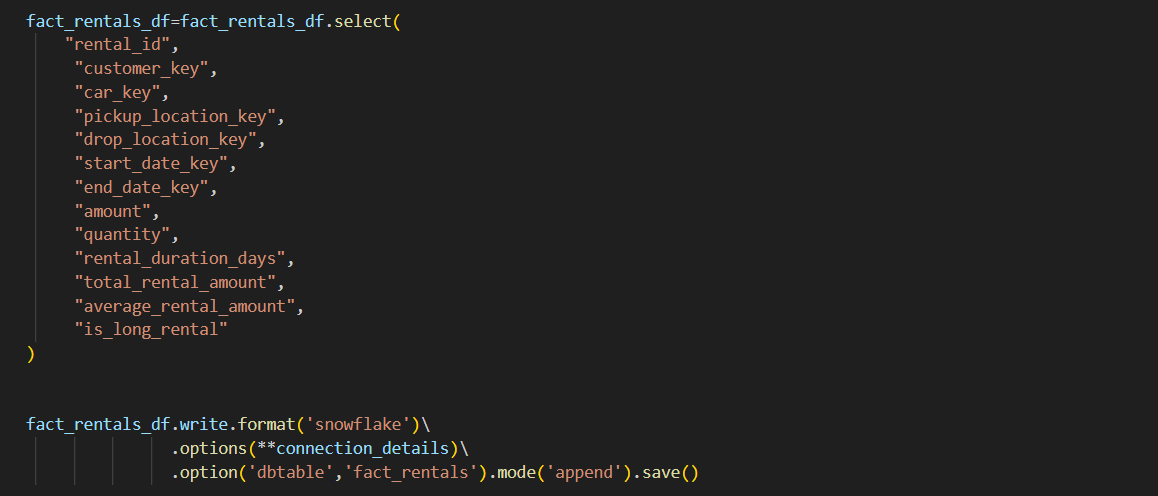
1. While calling the spark job from airflow, I had passed the date argument that will decide which file to pick from GCS Bucket path. In spark job I am reading the file as per the filepath from GCS bucket that is being dynamically set up as per the date passed from airflow job.Post reading , I am doing some validation checks to filter out records from the dataframe having null values for a given row.



1. Adding some more custom columns on the rentals dataframe.

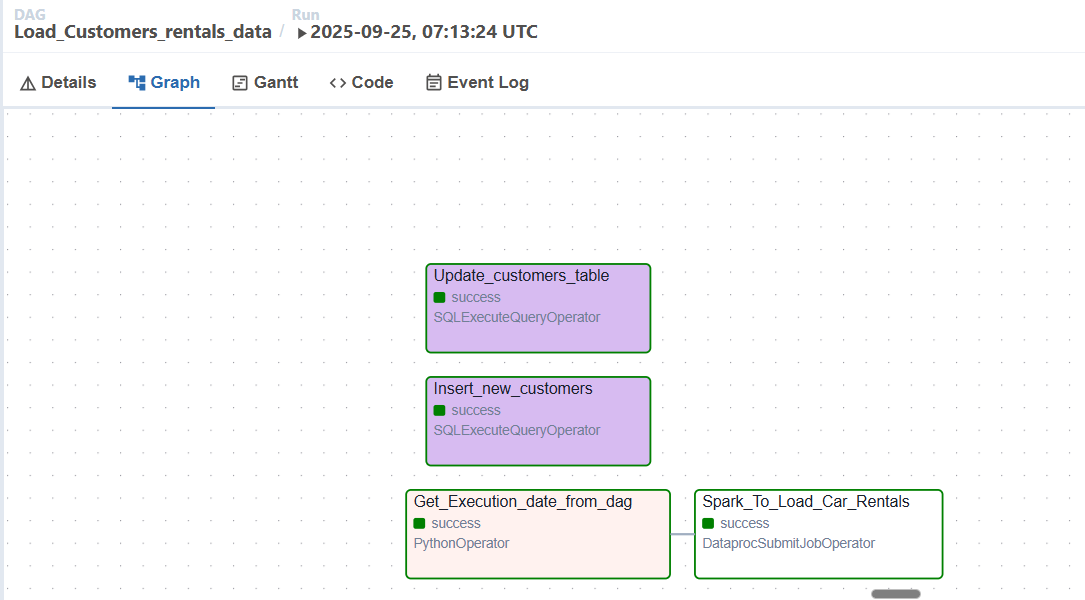


1. I read the dimension tables -dim\_date,dim\_customer,dim\_location,dim\_car and joined them with my rentals dataframe and only selected the foreign keys and the above-mentioned custom columns from the rentals dataframe and wrote them to snowflake table – fact\_rentals.



1. Results:

From **airflow ui**:



We also have data present in fact\_rentals table in our snowflake database:

