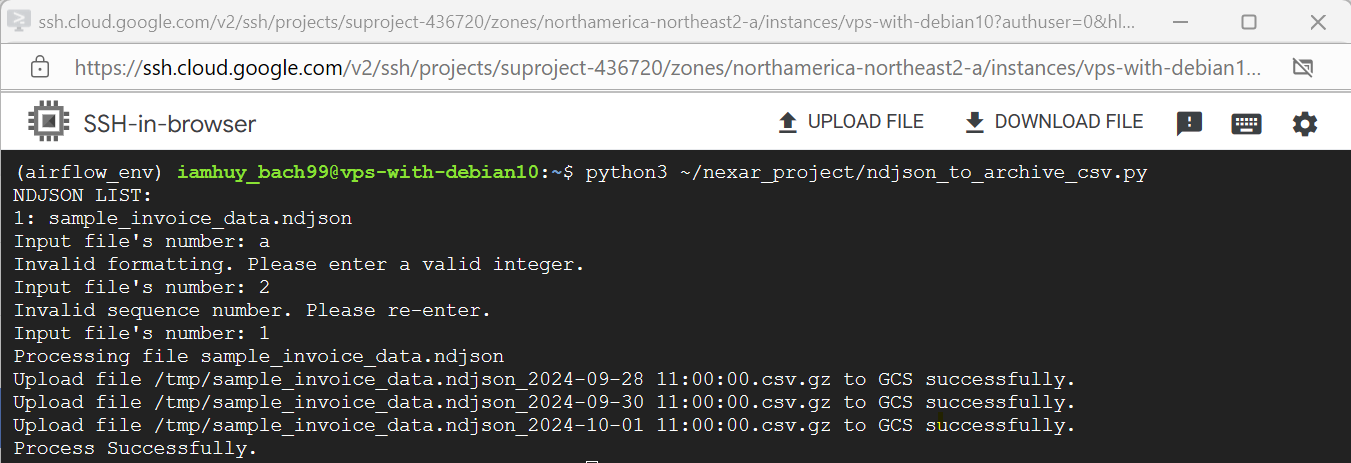
| NGUYỄN HUY BÁCH  **IMPORT DATA FROM NDJSON FILES INTO TABLE**  TEST CASE |  |
| --- | --- |

horizontal line

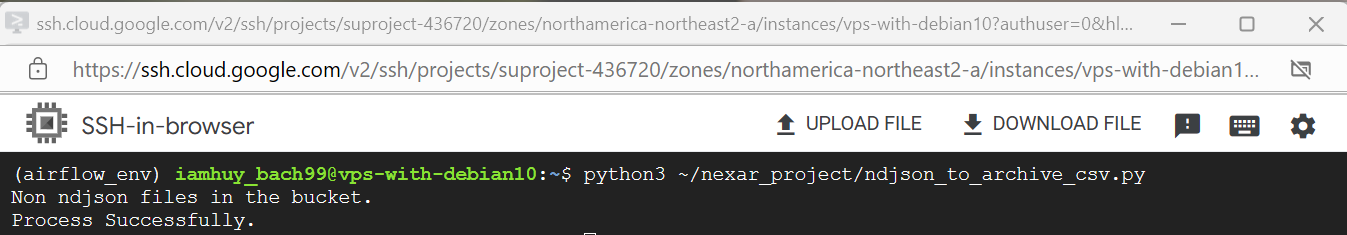
# Program 1: Convert data from NDJSON files to CSV files and compress

| **Test case num** | **Case** |
| --- | --- |
| TC1 | Manual load: Check input incorrect format and correct format |
| TC2 | Check non NDJSON files in the bucket |
| TC3 | Check if the loaded CSV files exist |
| TC4 | Automatic load by Airflow |
| TC5 | Convert data from NDJSON files to CSV files and compress successfully |

## TC1: Manual load: Check input incorrect format and correct format. - Done

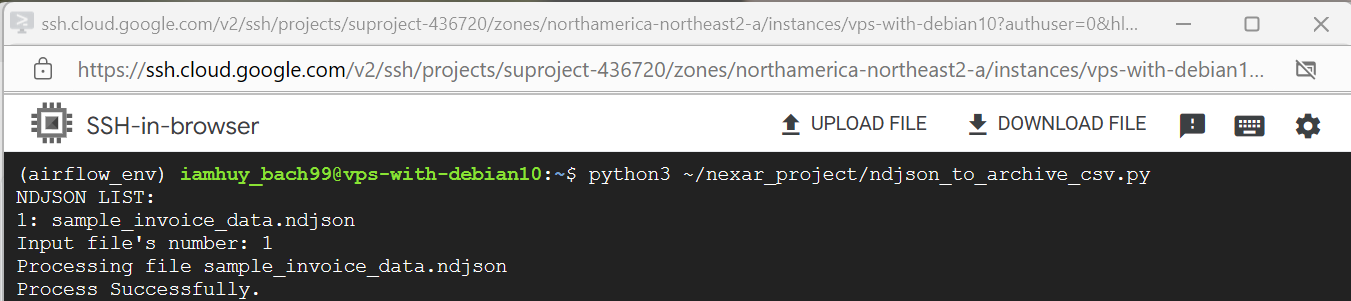


## TC2: Check non NDJSON files in the system. - Done



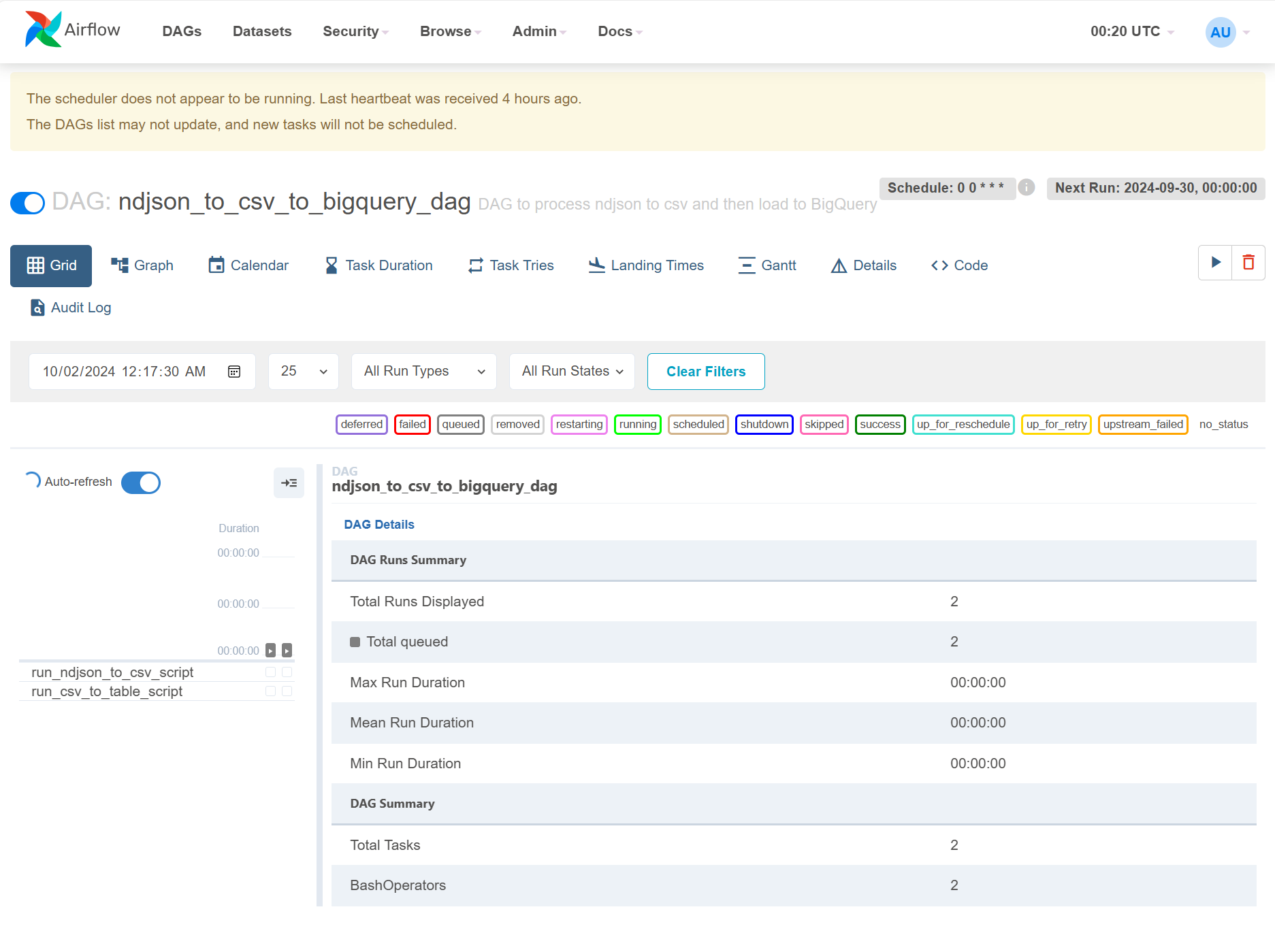
## TC3: Check if the loaded CSV files exist. - Done

In my NDJSON file, have a “timestamp” field. This field marks the date the data was loaded. I will compare this field with the timestamp in the file names of the CSV files. If they match, I will skip those lines of data and only load the new data.

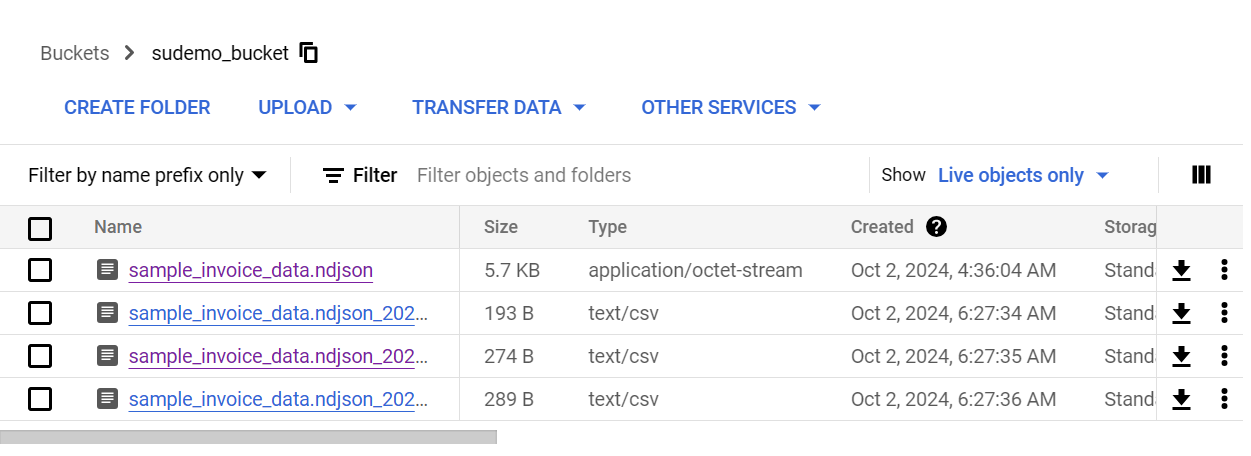


In the image, no new files will be created after running because those files already exist in the bucket.

## TC4: Automatic load by Airflow. - Done



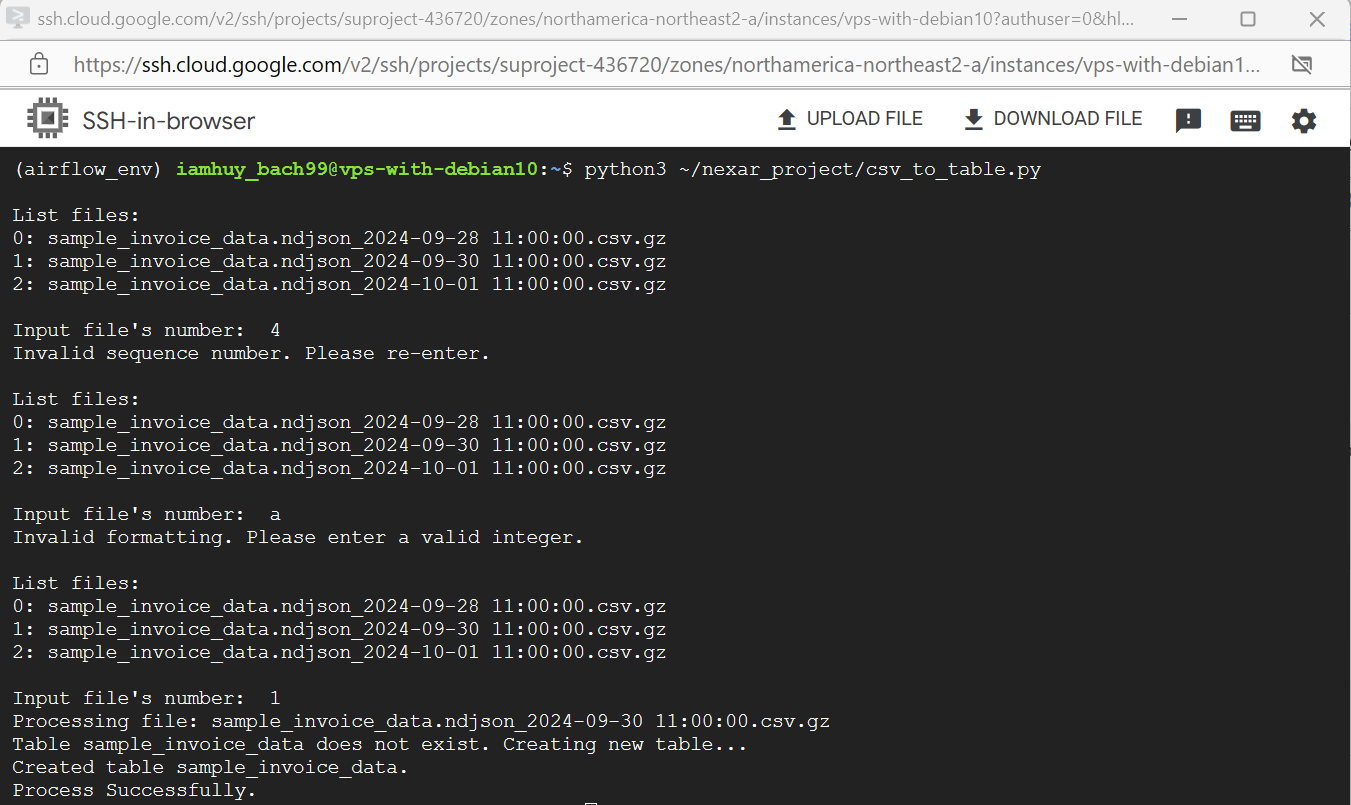
## TC5: Convert data from NDJSON files to CSV files and compress successfully. - Done



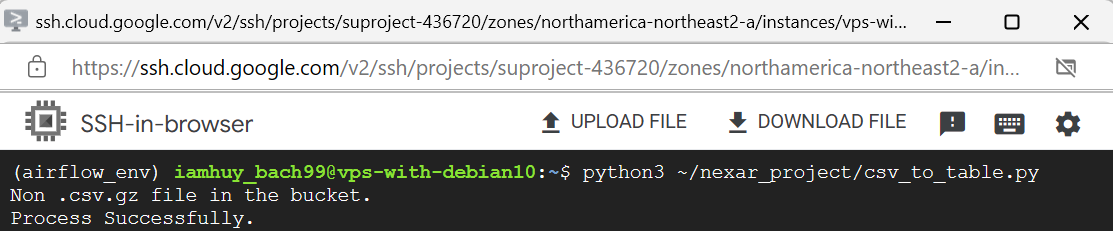
# Program 2: Import data from CSV files into Table in Bigquery

| **Test case num** | **Case** |
| --- | --- |
| TC1 | Manual load: Check input incorrect format and correct format. If the table does not exist, a new table will be created. |
| TC2 | Check non CSV files was compressed in the bucket |
| TC3 | Check if data exist |
| TC4 | Automatic load by Airflow |
| TC5 | Insert data to table successfully |

## TC1: Manual load: Check input incorrect format and correct format. If the table does not exist, a new table will be created. - Done

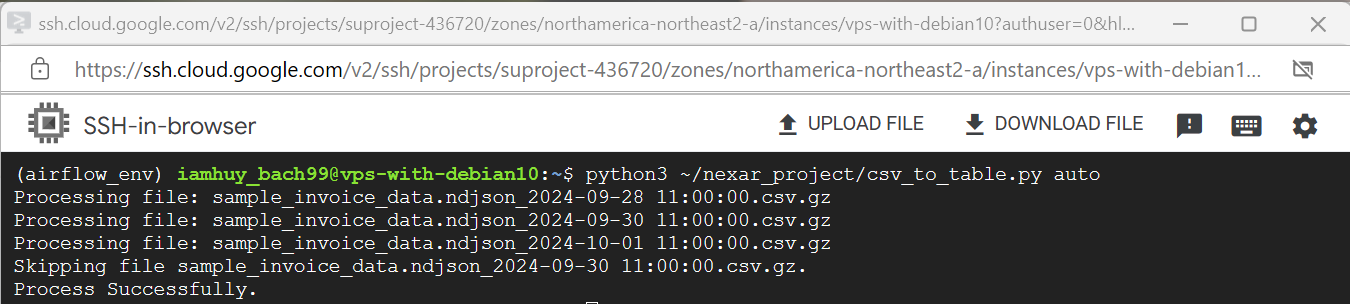


## TC2: Check non CSV files was compressed in the bucket. - Done

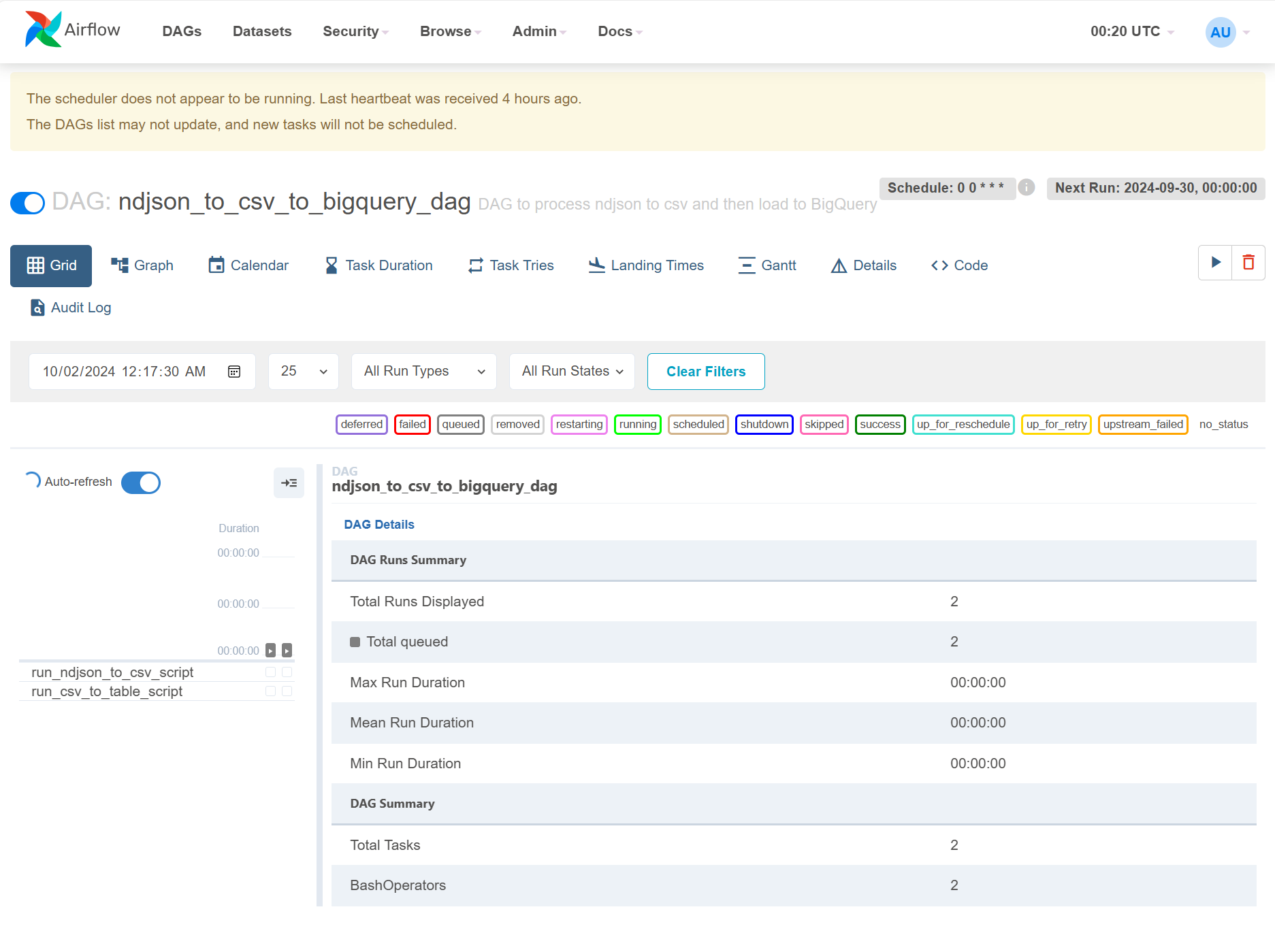


## TC3: Check if data exist. - Done

In the BigQuery table, have timestamp field. This field marks the date the data was loaded. I will compare this field with the timestamp in the file names of the CSV files. If they match, I will skip those lines of data and only load the new data. This helps the program run less, saving resources.

In the image, the file sample\_invoice\_data.ndjson\_2024-09-30 11:00:00.csv.gz has already been imported into the table. Therefore, the program will skip this file.

## TC4: Automatic load by Airflow. - Done



## TC5: Insert data to table successfully. - Done

