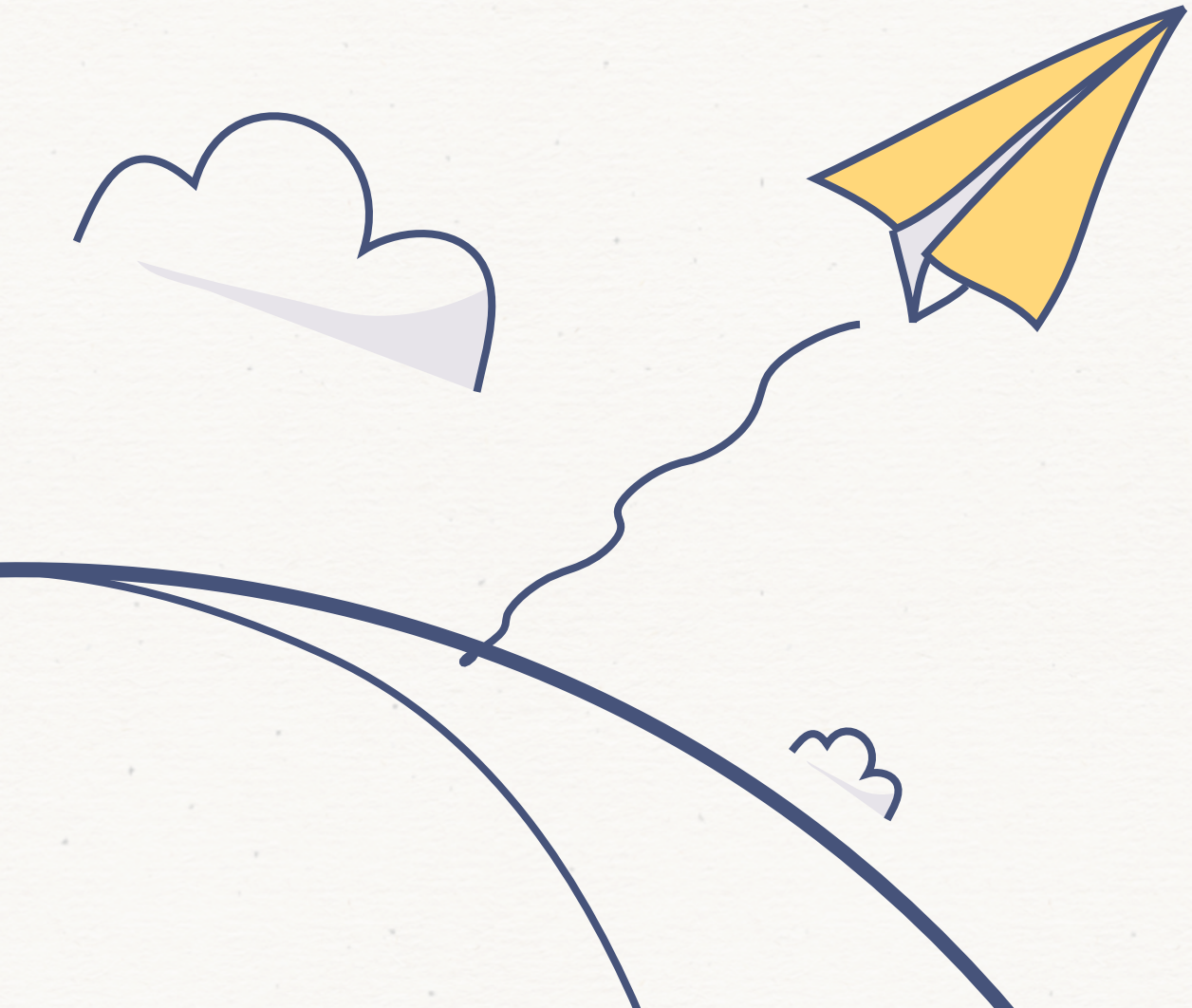


NYC TAXI TRIP
DATA PIPELINE

Link github: https://github.com/mufidnuha/nyc-taxi

OBJECTIVE

Build a data pipeline to retrieve and collect some taxi trip data such as yellow taxi trip, green taxi trip, and location zone data. Once collected, the data is transformed, cleaned, and combined according to business needs and then stored in the data warehouse so it can be easily used for analysis purposes.

DATA SOURCE

This project used taxi trip records in 2020-2021 such as:

1. Yellow Taxi Trip Records (>1 million records per month)
2. Green Taxi Trip Records (> 1 million records per month)
3. Taxi Zone Lookup Table

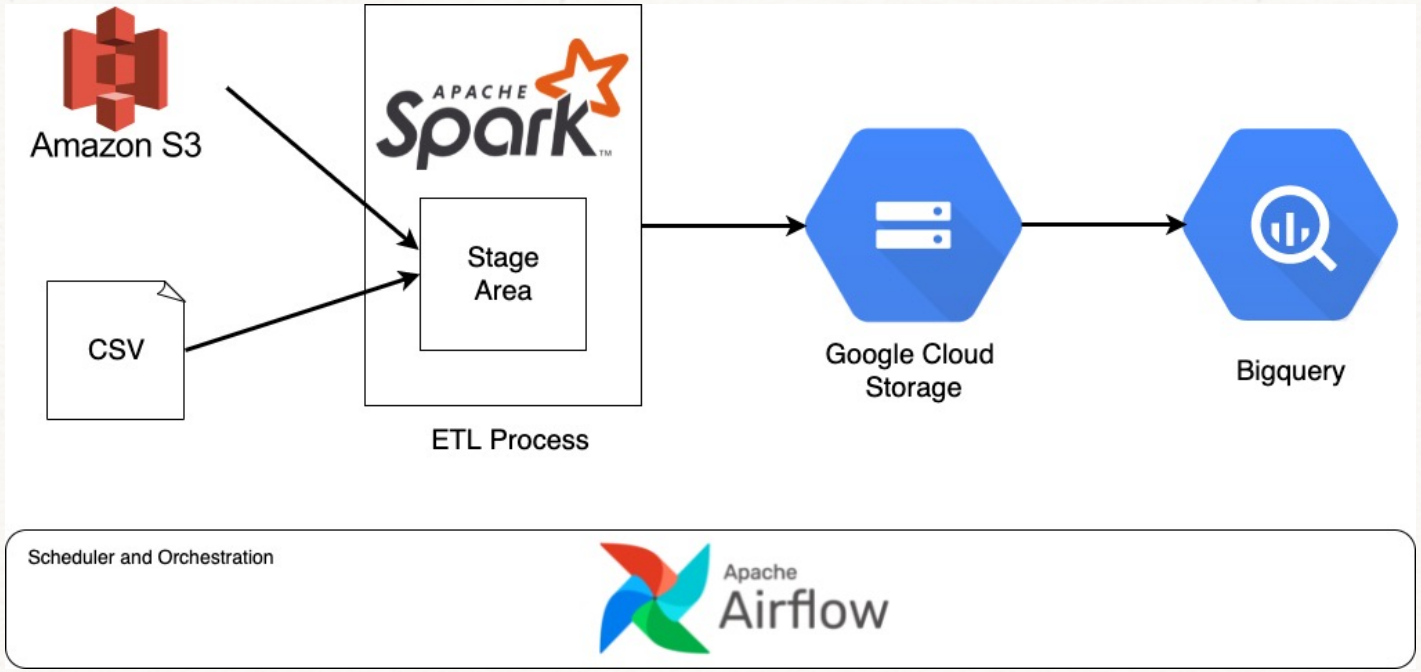
NYC Taxi data can access on https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

green_tripdata_2021_01

VendorID	lpep_pickup_datetime	lpep_dropoff_datetime	store_and_fwd_flag	RatecodeID	PULocationID	DOLocationID	passenger_count	trip_distance	fare_amount	extra	mta_tax	tip_amount	tolls_amount	ehail_fee	improvement_surcharge	total_amount	payment_type	trip_type	congestion_surcharge
2	2021-01-01 00:15:56	2021-01-01 00:19:52	N	1	43	151	1	1.01	5.5	0.5	0.5	0	0		0.3	6.8	2	1	0
2	2021-01-01 00:25:59	2021-01-01 00:34:44	N	1	166	239	1	2.53	10	0.5	0.5	2.81	0		0.3	16.86	1	1	2.75
2	2021-01-01 00:45:57	2021-01-01 00:51:55	N	1	41	42	1	1.12	6	0.5	0.5	1	0		0.3	8.3	1	1	0
2	2020-12-31 23:57:51	2021-01-01 00:04:56	N	1	168	75	1	1.99	8	0.5	0.5	0	0		0.3	9.3	2	1	0
2	2021-01-01 00:16:36	2021-01-01 00:16:40	N	2	265	265	3	.00	-52	0	-0.5	0	0		-0.3	-52.8	3	1	0
2	2021-01-01 00:16:36	2021-01-01 00:16:40	N	2	265	265	3	.00	52	0	0.5	0	0		0.3	52.8	2	1	0
2	2021-01-01 00:19:14	2021-01-01 00:19:21	N	5	265	265	1	.00	180	0	0	36.06	0		0.3	216.36	1	2	0
2	2021-01-01 00:26:31	2021-01-01 00:28:50	N	1	75	75	6	.45	3.5	0.5	0.5	0.96	0		0.3	5.76	1	1	0
2	2021-01-01 00:57:46	2021-01-01 00:57:57	N	1	225	225	1	.00	2.5	0.5	0.5	0	0		0.3	3.8	2	1	0
2	2021-01-01 00:58:32	2021-01-01 01:32:34	N	1	225	265	1	12.19	38	0.5	0.5	2.75	0		0.3	42.05	1	1	0
2	2021-01-01 00:31:14	2021-01-01 00:55:07	N	1	244	244	2	3.39	18	0.5	0.5	0	0		0.3	19.3	2	1	0
2	2021-01-01 00:08:50	2021-01-01 00:21:56	N	1	75	213	1	6.69	19.5	0.5	0.5	0	0		0.3	20.8	2	1	0
2	2021-01-01 00:35:13	2021-01-01 00:44:44	N	1	74	238	1	2.34	10	0.5	0.5	0	0		0.3	14.05	1	1	2.75
2	2021-01-01 00:39:57	2021-01-01 00:55:25	N	1	74	60	1	5.48	18	0.5	0.5	0	0		0.3	19.3	2	1	0
1	2021-01-01 00:51:27	2021-01-01 00:57:20	N	1	42	41	2	.90	6	0.5	0.5	0	0		0.3	7.3	1	1	0
2	2021-01-01 00:29:05	2021-01-01 00:29:07	N	5	42	264	1	.00	10	0	0	2.06	0		0.3	12.36	1	2	0
2	2021-01-01 00:32:07	2021-01-01 00:42:54	N	1	74	116	1	2.08	9.5	0.5	0.5	2.16	0		0.3	12.96	1	1	0
2	2021-01-01 00:49:59	2021-01-01 01:05:01	N	1	116	143	1	4.64	16.5	0.5	0.5	5.14	0		0.3	25.69	1	1	2.75
2	2021-01-01 00:07:20	2021-01-01 00:12:01	N	1	75	42	1	1.68	6.5	0.5	0.5	0	0		0.3	7.8	2	1	0
2	2021-01-01 00:25:54	2021-01-01 00:28:20	N	1	74	75	1	.68	4	0.5	0.5	0	0		0.3	5.3	2	1	0
2	2021-01-01 00:15:51	2021-01-01 00:30:34	N	1	7	82	5	2.70	12	0.5	0.5	0	0		0.3	13.3	1	1	0
2	2021-01-01 00:21:09	2021-01-01 01:06:01	N	1	152	117	1	29.07	77.5	0.5	0.5	0	6.12		0.3	84.92	2	1	0
2	2021-01-01 00:42:25	2021-01-01 00:43:06	N	1	82	82	1	.00	2.5	0.5	0.5	8	0		0.3	11.8	1	1	0
2	2021-01-01 00:51:52	2021-01-01 01:06:13	N	1	116	74	1	2.78	12	0.5	0.5	0	0		0.3	13.3	2	1	0
2	2021-01-01 00:23:19	2021-01-01 00:34:03	N	1	116	69	3	2.25	10	0.5	0.5	0	0		0.3	11.3	2	1	0
1	2021-01-01 00:56:41	2021-01-01 01:13:31	N	1	259	116	1	.00	28.2	0	0.5	0	0		0.3	29	1	1	0
2	2021-01-01 00:50:23	2021-01-01 00:55:11	N	1	247	167	1	1.03	5.5	0.5	0.5	0	0		0.3	6.8	1	1	0
2	2021-01-01 00:15:41	2021-01-01 00:18:57	N	1	166	41	1	.65	4.5	0.5	0.5	0	0		0.3	5.8	1	1	0

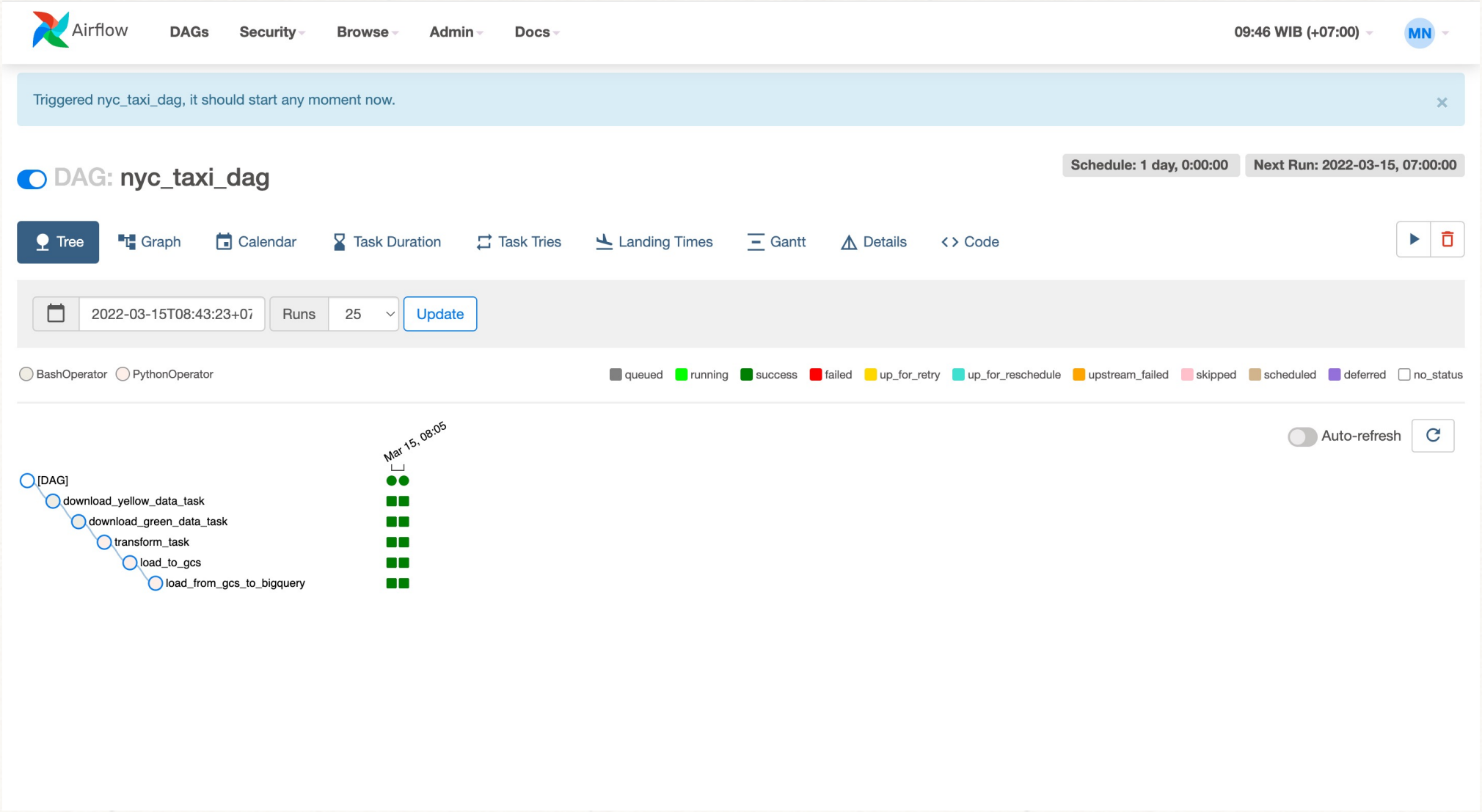
Sample Data

DATA PIPELINE DIAGRAM

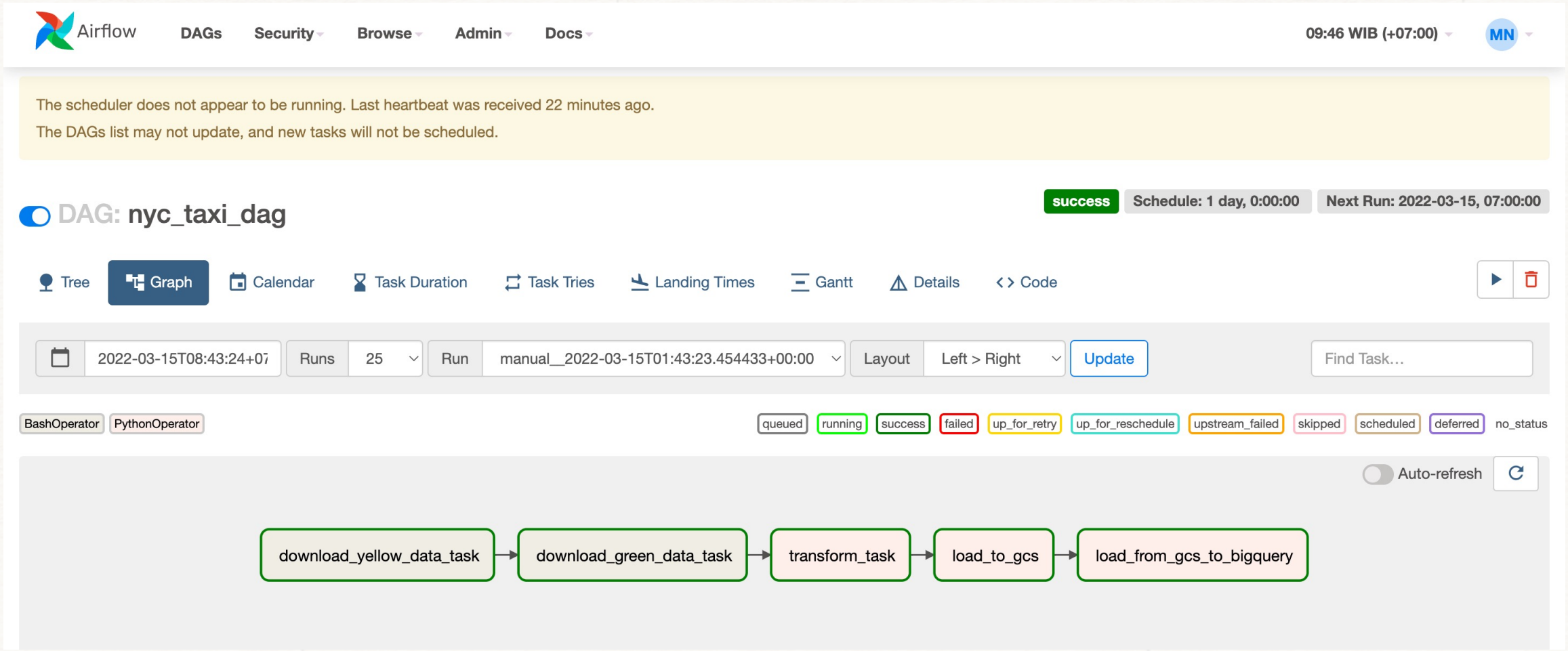


1. Data from AWS S3 is extracted and stored in the staging area
2. Data is processed and transformed using Apache Spark
3. Data is loaded Google Cloud Storage as a data lake in parquet format
4. From GCS, data is retrieved and stored to Bigquery as data warehouse
5. All workflows in this pipeline are managed and scheduled by Apache Airflow

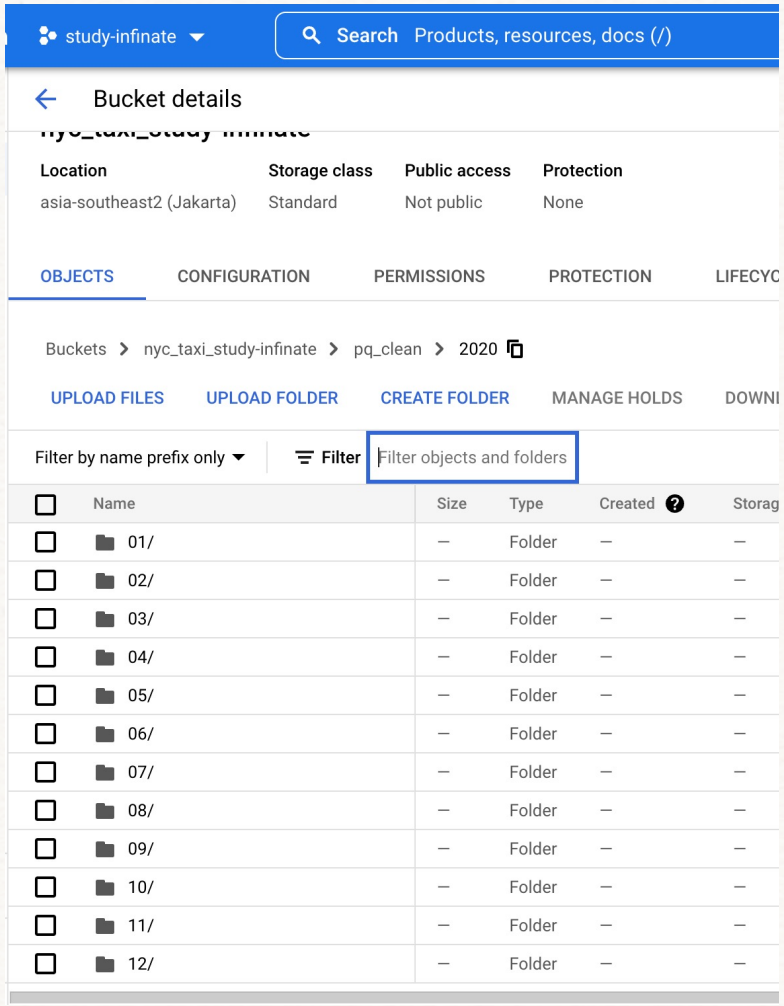
RESULTS:
AIRFLOW DAG IN TREE VIEW



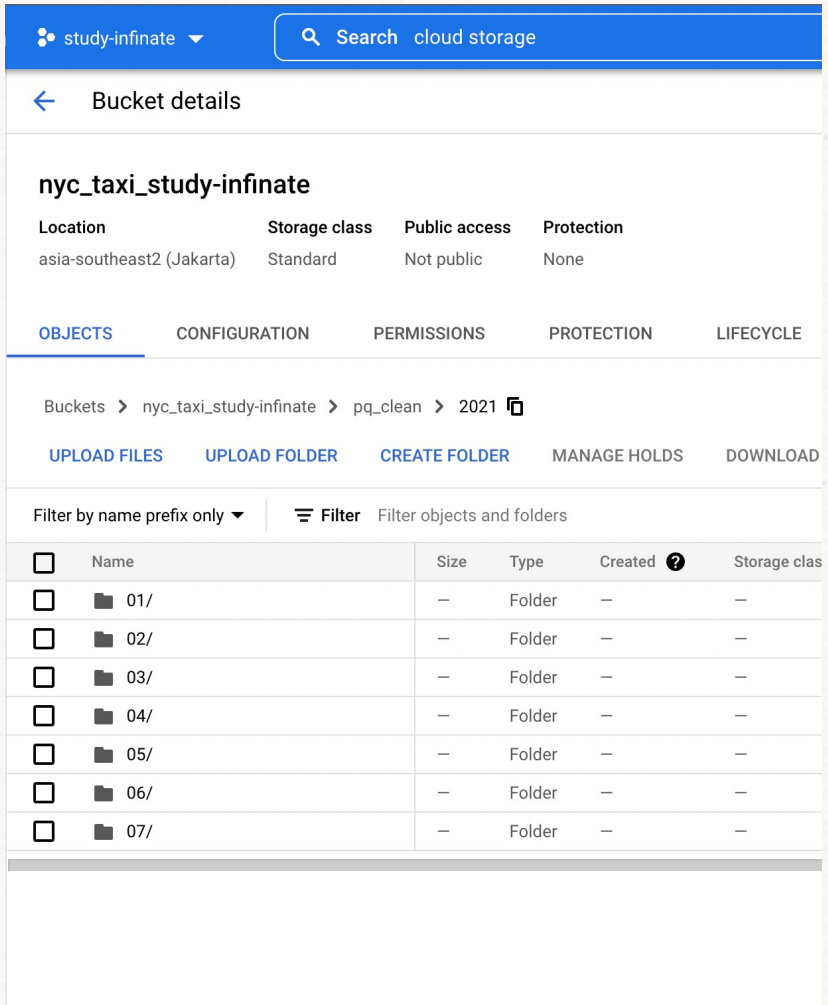
RESULTS:
AIRFLOW DAG IN GRAPH VIEW



RESULTS: GOOGLE CLOUD STORAGE



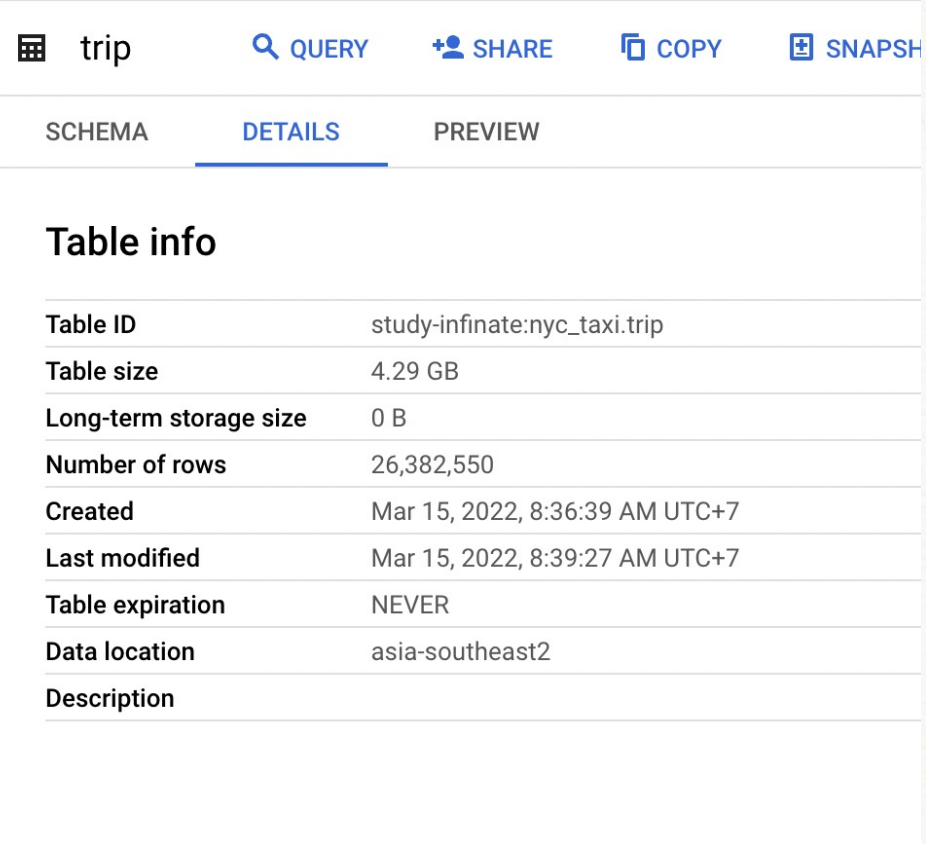
2020



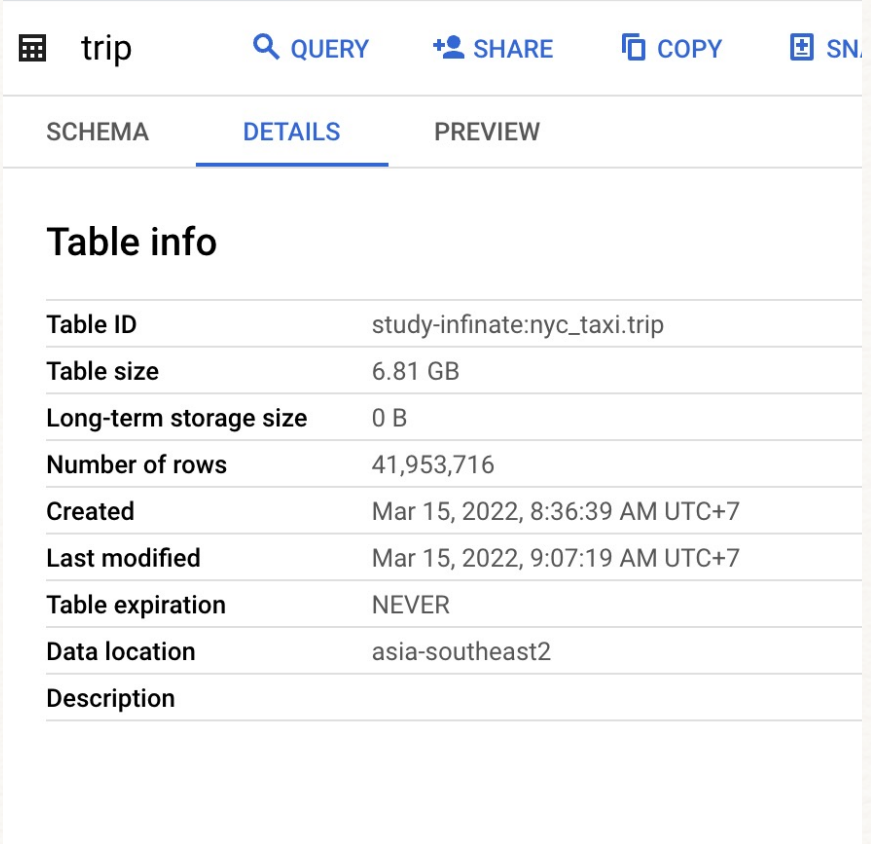
2021

- The results of the data transformation are stored in Google Cloud Storage.
- Data is stored by month and year.
- Data is stored in parquet format to support fast data processing for complex data.

RESULTS: BIGQUERY

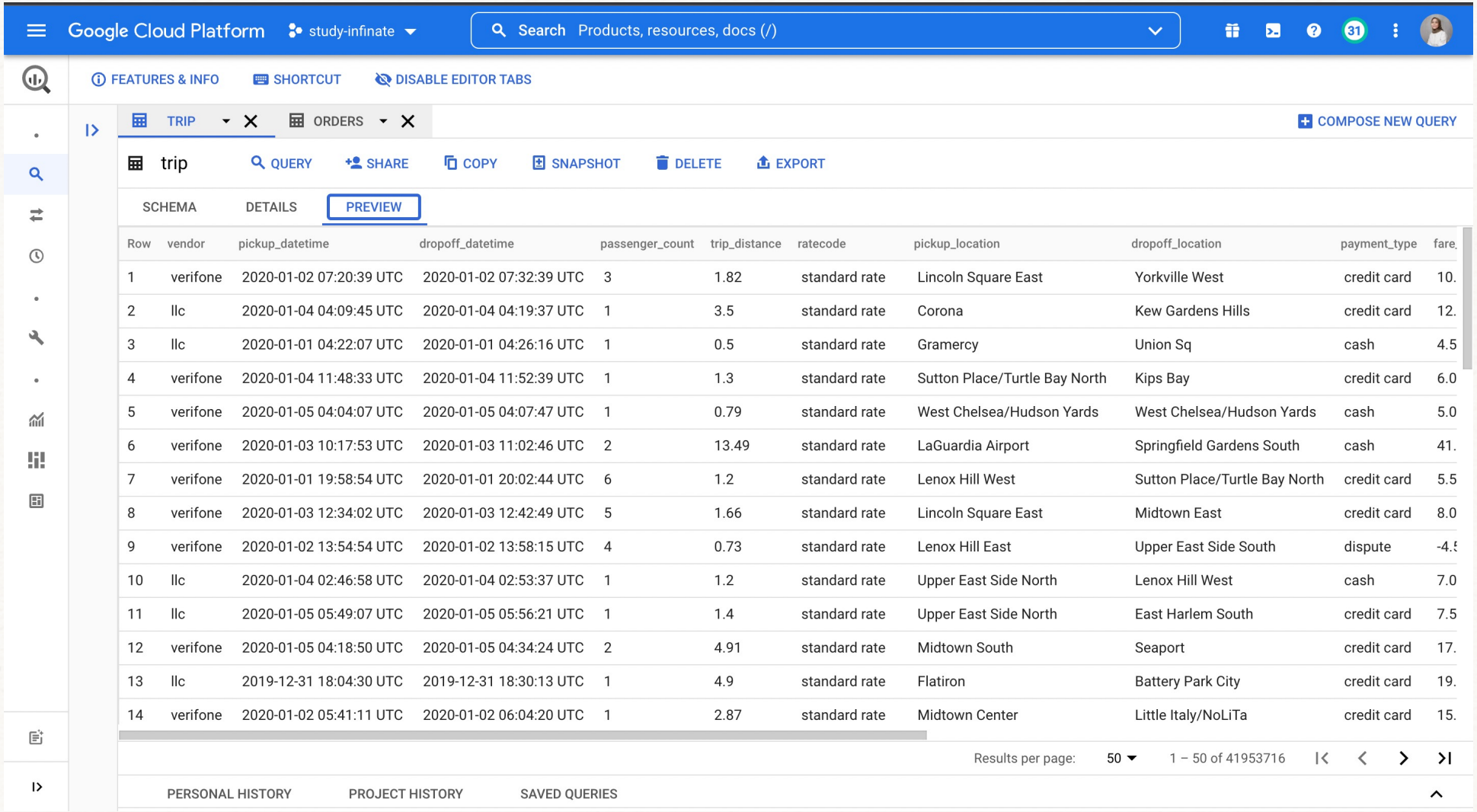


Details of the table after added taxi trip data in 2020
Number of records: > 26 million



Details of the table after added taxi trip data in 2021
Number of records: > 41 million

RESULTS: PREVIEW TABLE ON BIGQUERY



Thank You