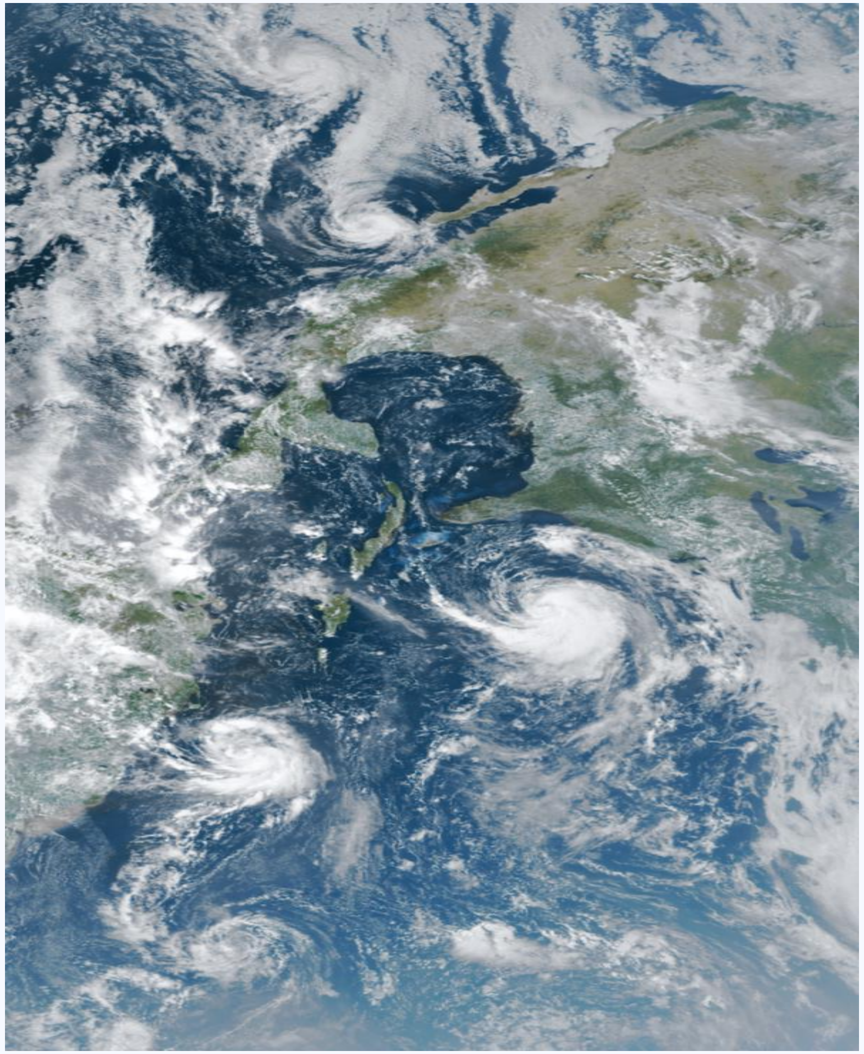
GLOBAL WEATHER REPOSITORY

GROUP 8:

- AISWARYA
- NEHA THAKUR
- SAUMYA VARSHNEY
- SHIKHA SINGH
- SREELEKHA



INTRODUCTION

- Global Historical Climatology Network (GHCN) data from the National Oceanic and Atmospheric Administration (NOAA) website has been utilized for implementing the data pipeline.
- In this project we are working on Cloud Analytics and Data Warehouse Implementation by building a data pipeline to analyze archival and real time weather data and gather useful insights from this data.
- Temperature, Precipitation, Snow and Snow Depth patterns over the years will be analyzed and visualized for drawing useful insights which can be further utilized by industries.

Dataset link:

https://www.ncdc.noaa.gov/cdo-web/webservices/v2#gettingStarted

https://noaa-ghcn-pds.s3.amazonaws.com/index.html

Data Sources

NOAA GHCN-D AWS Bucket - This public bucket contains all the data collected by NOAA for the GHCN-D dataset from year 1750 to 2023. We are only considering the data from 2010 to 2023 for this project.

Each year is stored as a CSV, and the headers comprise of

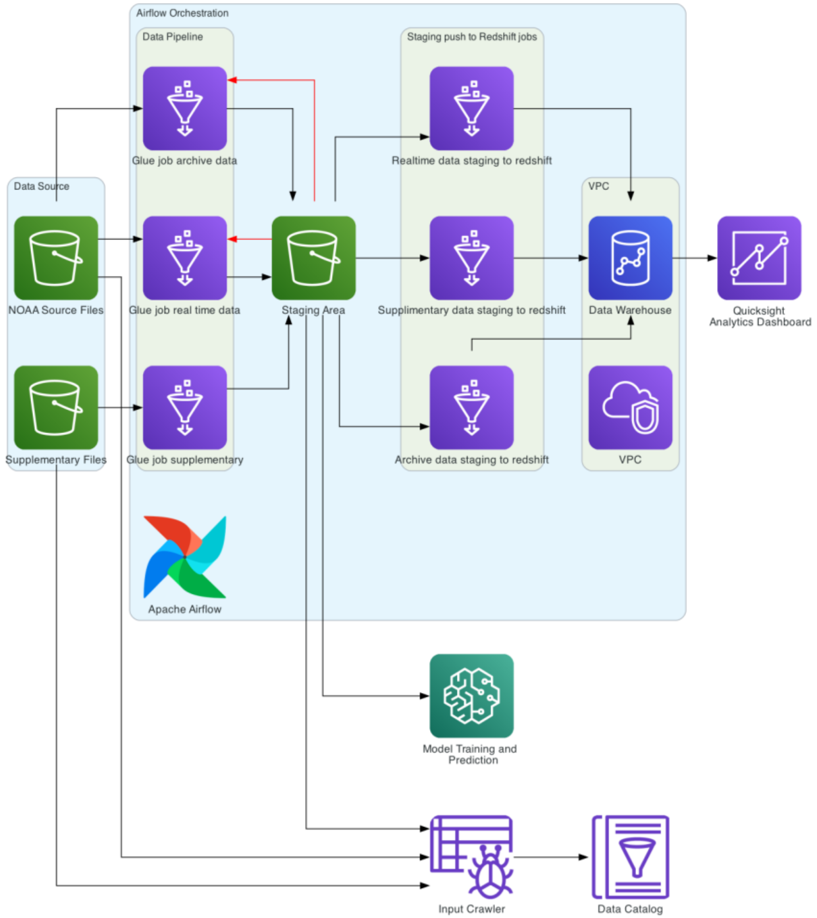
- station id - id for the weather station where the data was collected from.
- date id - date at which the data was collected in the format yyyymmdd (20231130).
- element id / datatype id - what category of data was recorded (maximum temperature, minimum temperature, precipitation, snow fall and snow depth).
- Data / value - the data recorded corresponding to element.

Additionally, we scraped a lot of supplemental data from https://www.ncdc.noaa.gov/cdo-web/webservices/v2#gettingStarted using REST APIs.

CLOUD ARCHITECTURE

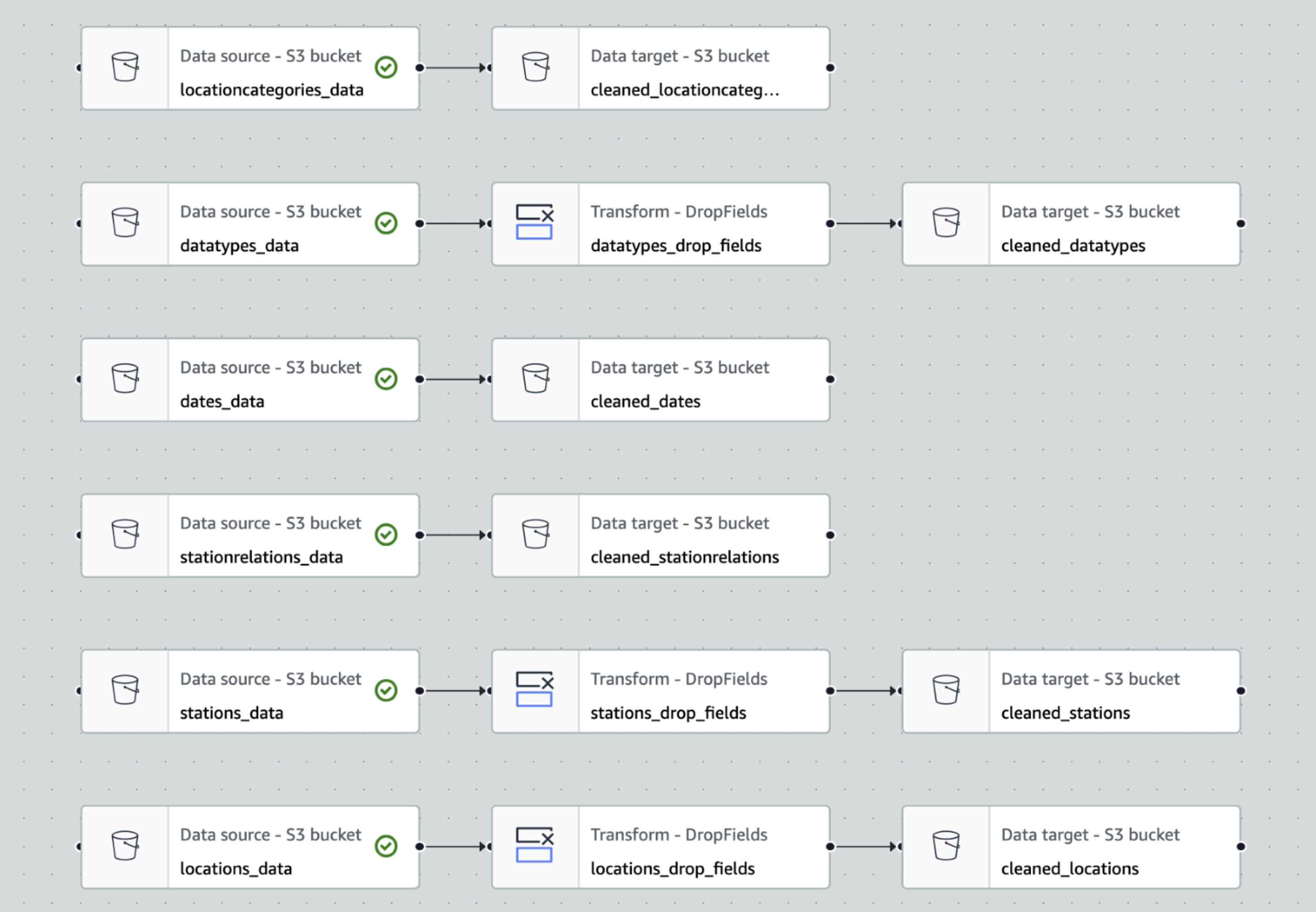
AWS Services used:

- VPC
- S3
- Glue
- Glue Crawler
- Glue Data Catalog
- Redshift
- Apache Airflow
- Quick Sight
- Sage Maker



AWS Data Pipeline for Global weather Repository

ETL (EXTRACT-TRANSFORM-LOAD)



Data source - S3 bucket
locationcategories_data
→ Data target - S3 bucket
cleaned_locationcateg...

Data source - S3 bucket
datatypes_data
→ Transform - DropFields
datatypes_drop_fields
→ Data target - S3 bucket
cleaned_datatypes

Data source - S3 bucket
dates_data
→ Data target - S3 bucket
cleaned_dates

Data source - S3 bucket
stationrelations_data
→ Data target - S3 bucket
cleaned_stationrelations

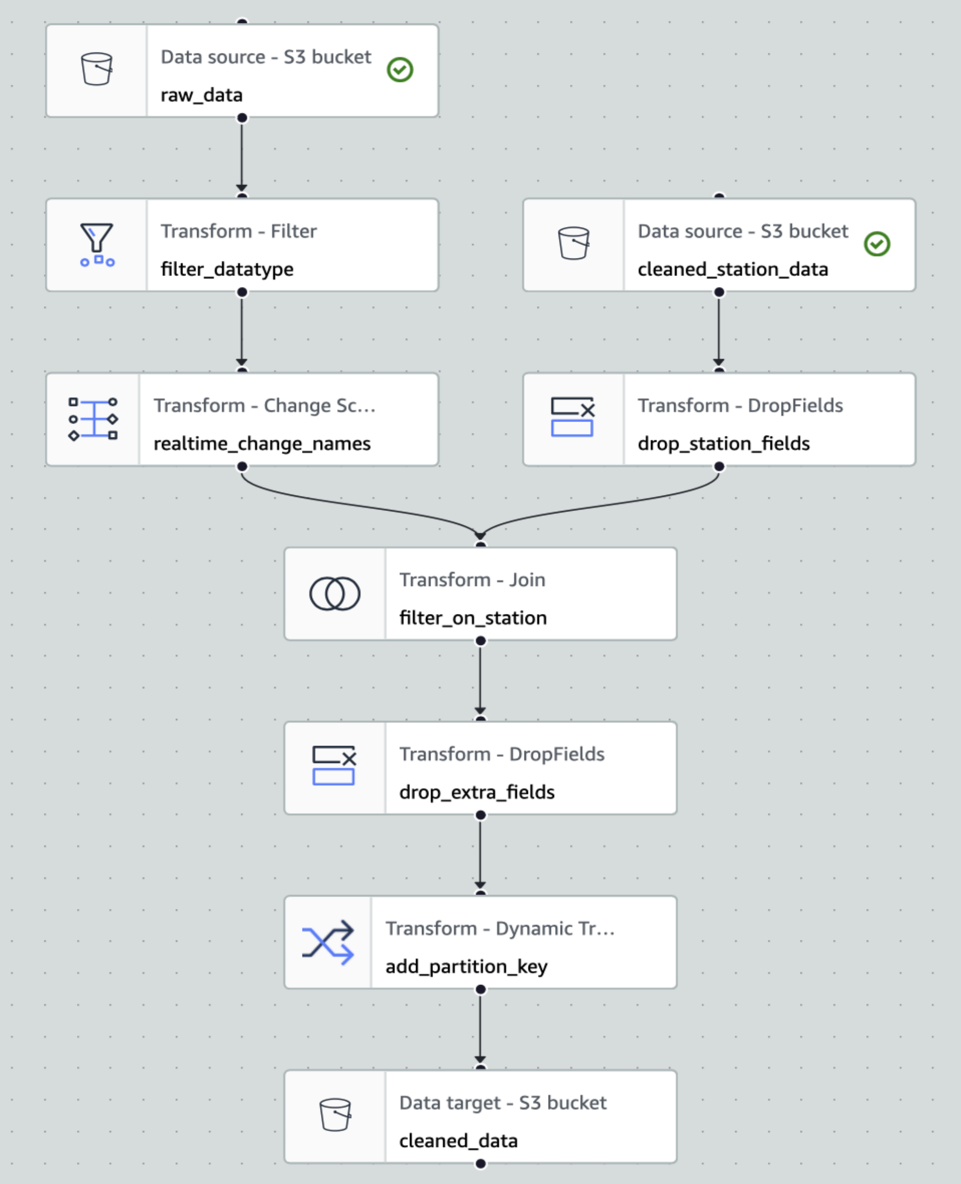
Data source - S3 bucket
stations_data
→ Transform - DropFields
stations_drop_fields
→ Data target - S3 bucket
cleaned_stations

Data source - S3 bucket
locations_data
→ Transform - DropFields
locations_drop_fields
→ Data target - S3 bucket
cleaned_locations

Supplementary Data Pipeline

Reads the raw CSV files from source S3 bucket, transforms and load the cleaned data to S3 staging area in parquet format.

- locationcategories.csv
- datatypes.csv
- dates.csv
- stationrelations.csv
- stations.csv
- locations.csv



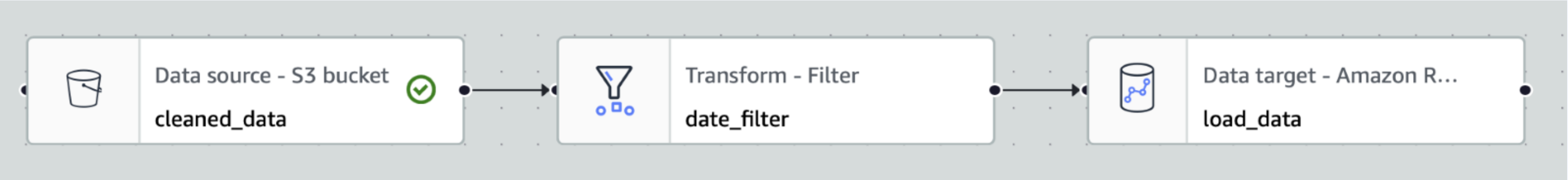
ETL (EXTRACT-TRANSFORM-LOAD)

Archive (2010- 2022) and Real time (2023) Data Pipeline

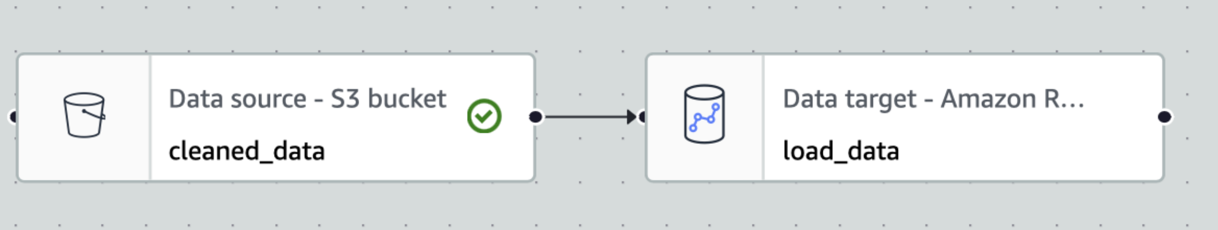
Reads the raw CSV files from source S3 bucket, transforms and load the cleaned data to S3 staging area in parquet format.

LOAD TO DATA WAREHOUSE

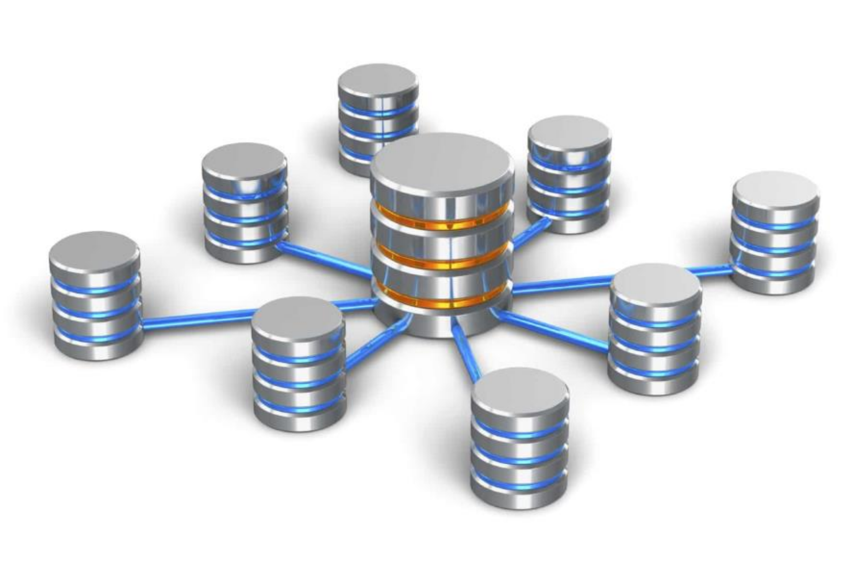
Load Archive data to Datawarehouse

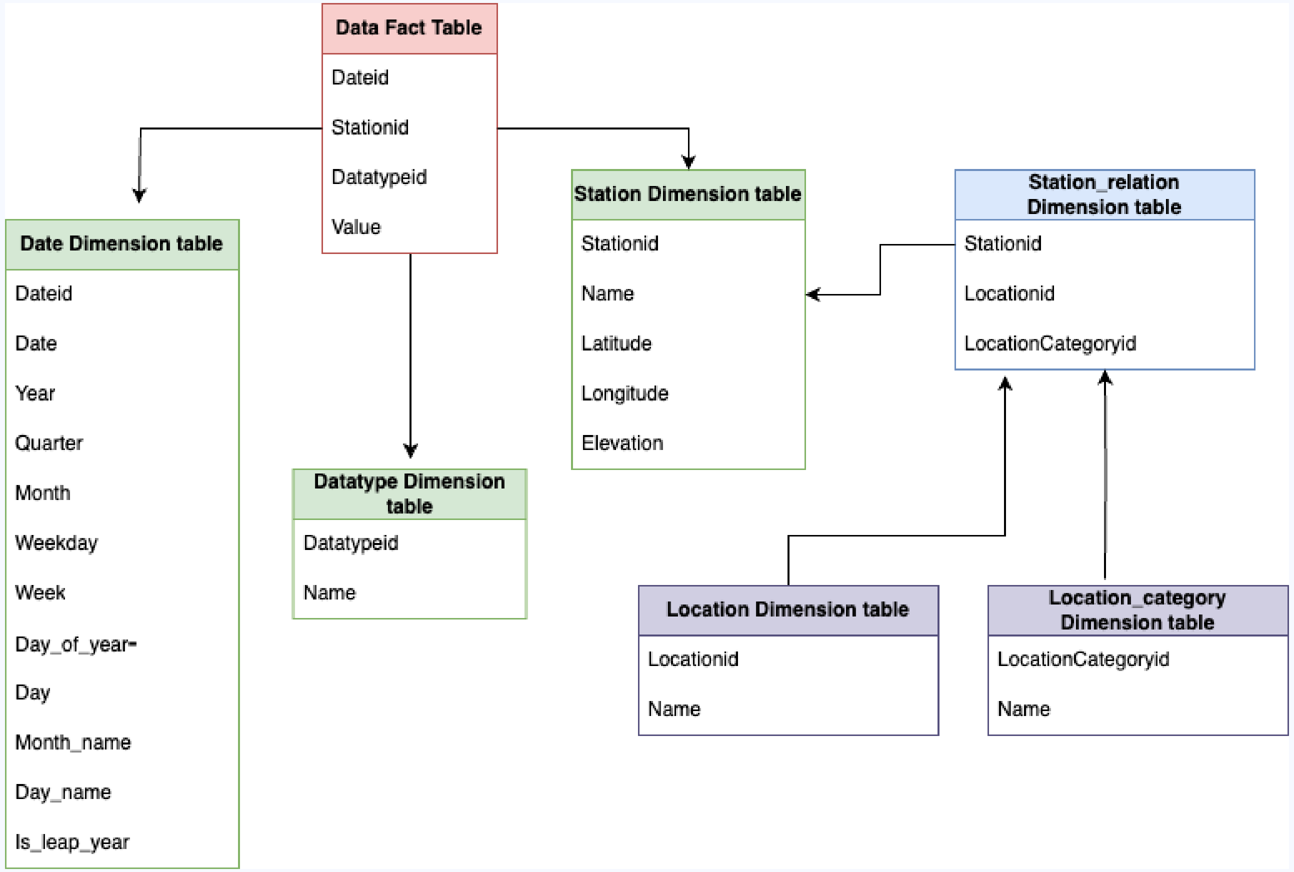


Load Real time data to Data Warehouse



Similarly, we have loaded supplementary data to data Warehouse





Data Fact Table
Dateid
Stationid
Datatypeid
Value

Date Dimension table
Dateid
Date
Year
Quarter
Month
Weekday
Week
Day_of_year-
Day
Month_name
Day_name
Is_leap_year

Datatype Dimension table
Datatypeid
Name

Station Dimension table
Stationid
Name
Latitude
Longitude
Elevation

Station_relation Dimension table
Stationid
Locationid
LocationCategoryid

Location Dimension table
Locationid
Name

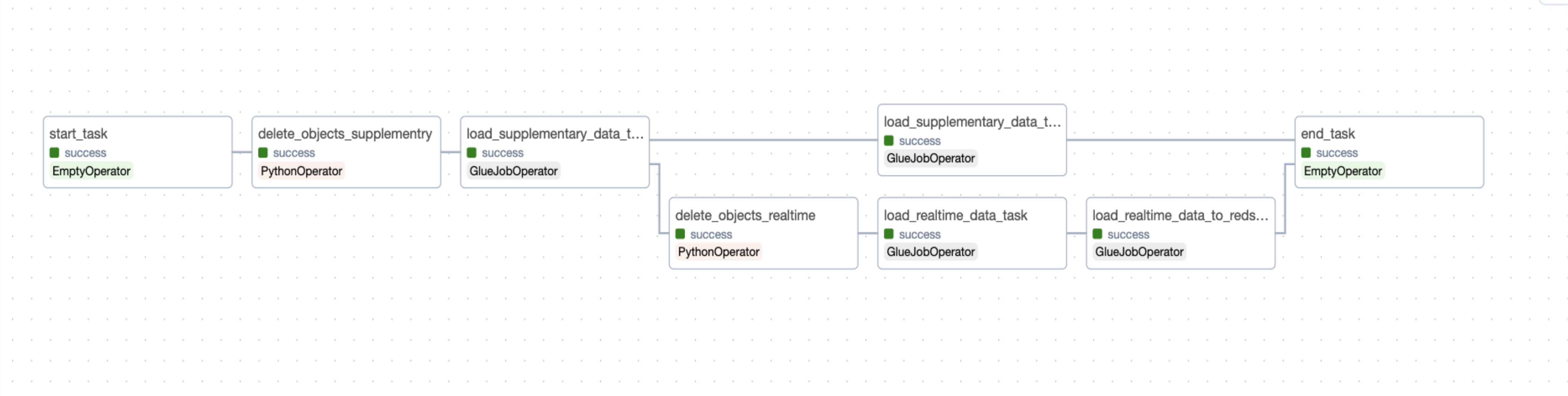
Location_category Dimension table
LocationCategoryid
Name

Data Warehouse Schema

- Snowflake schema
- Fact table – Data
- Dimension table – Date, Datatype and Station

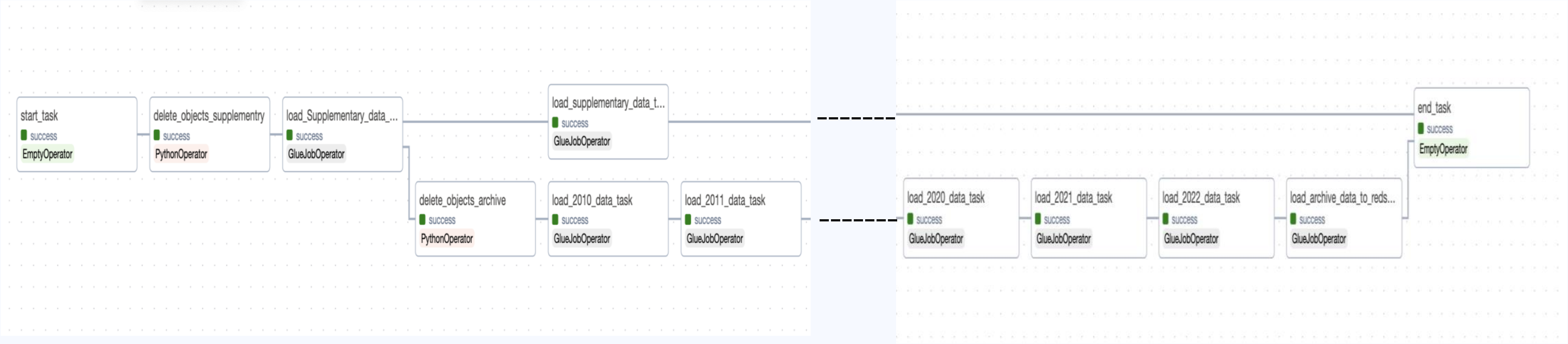
APACHE AIRFLOW

Real-time data pipeline



APACHE AIRFLOW

Archive data pipeline



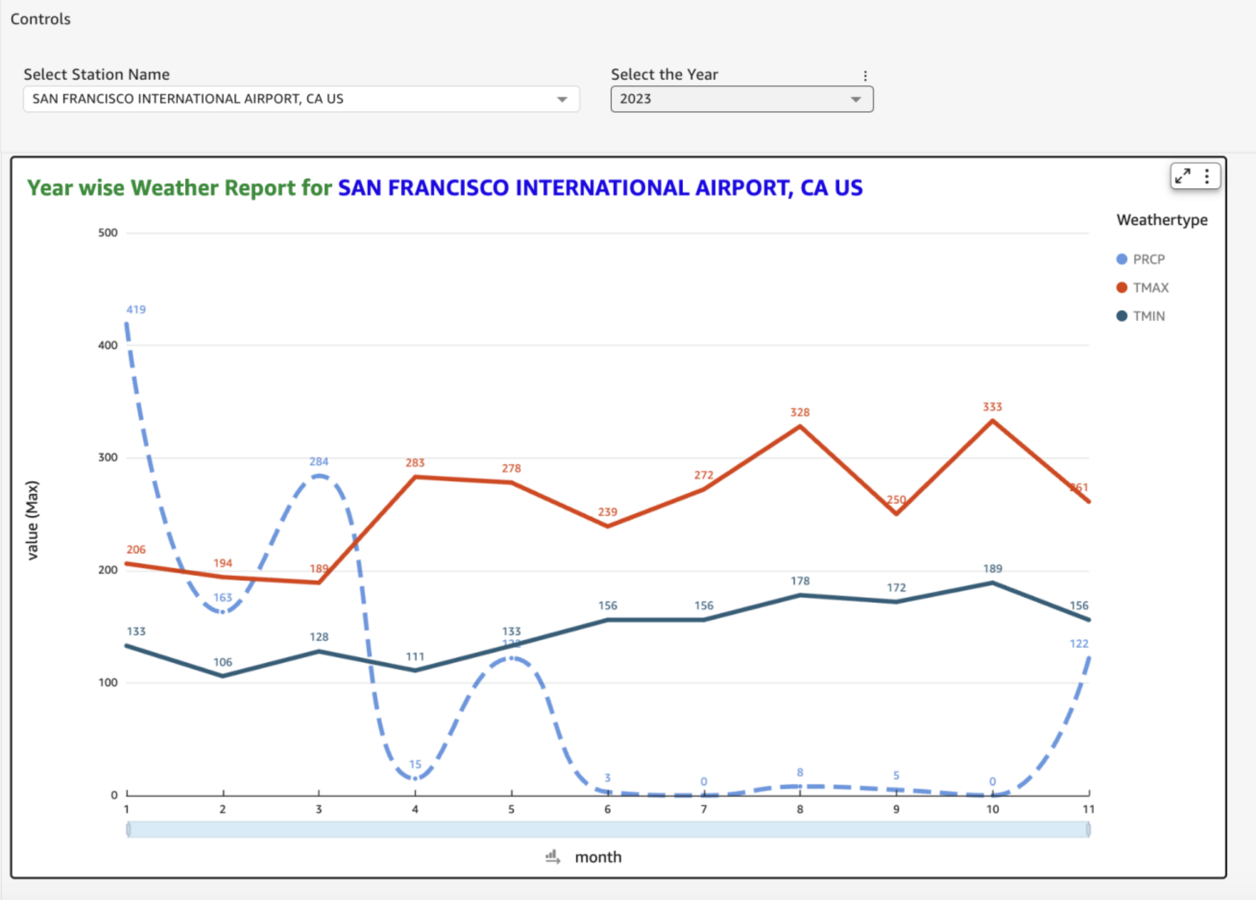
VISUALIZATION

amazon QuickSight

Analysis and Visualizations on Weather Parameters

Year wise Weather Analysis

VISUAL ANALYSIS



Controls

Select Station Name: SAN FRANCISCO INTERNATIONAL AIRPORT, CA US

Select the Year: 2023

Year wise Weather Report for SAN FRANCISCO INTERNATIONAL AIRPORT, CA US

PREDICTIVE ANALYTICS

Prediction Model

We have used the XGBoost Algorithm to train regression model to predict maximum and minimum temperature, precipitation, snow fall and snow depth. We used this model as this is known to work best for tabular data.

Final Schema for Training Data Features

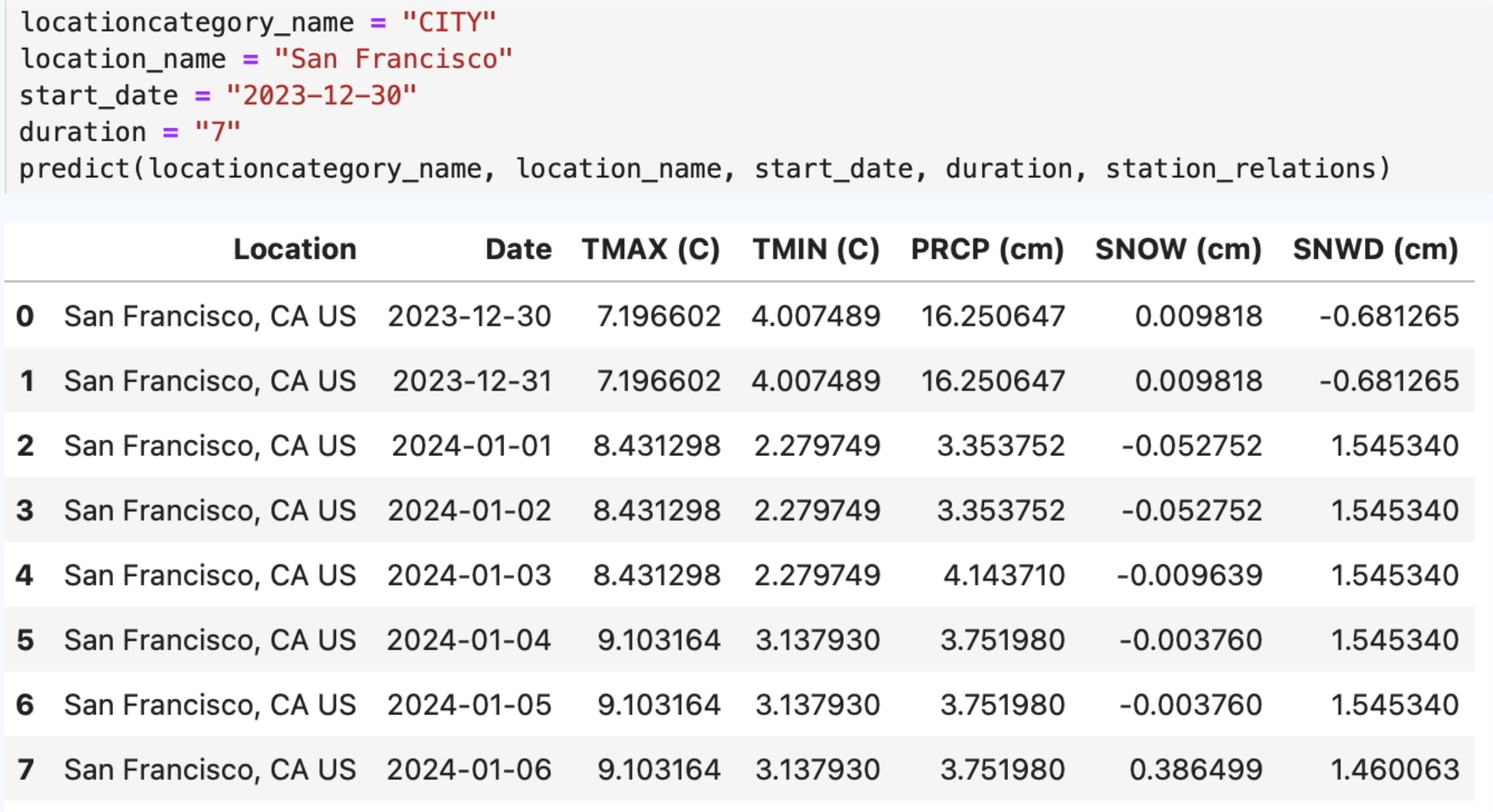
year	quarter	month	week	day_of_year	leap_year	latitude	longitude	elevation

Final Schema for Training Data Targets

TMAX	TMIN	PRCP	SNOW	SNWD

R2 score for your model is 0.398

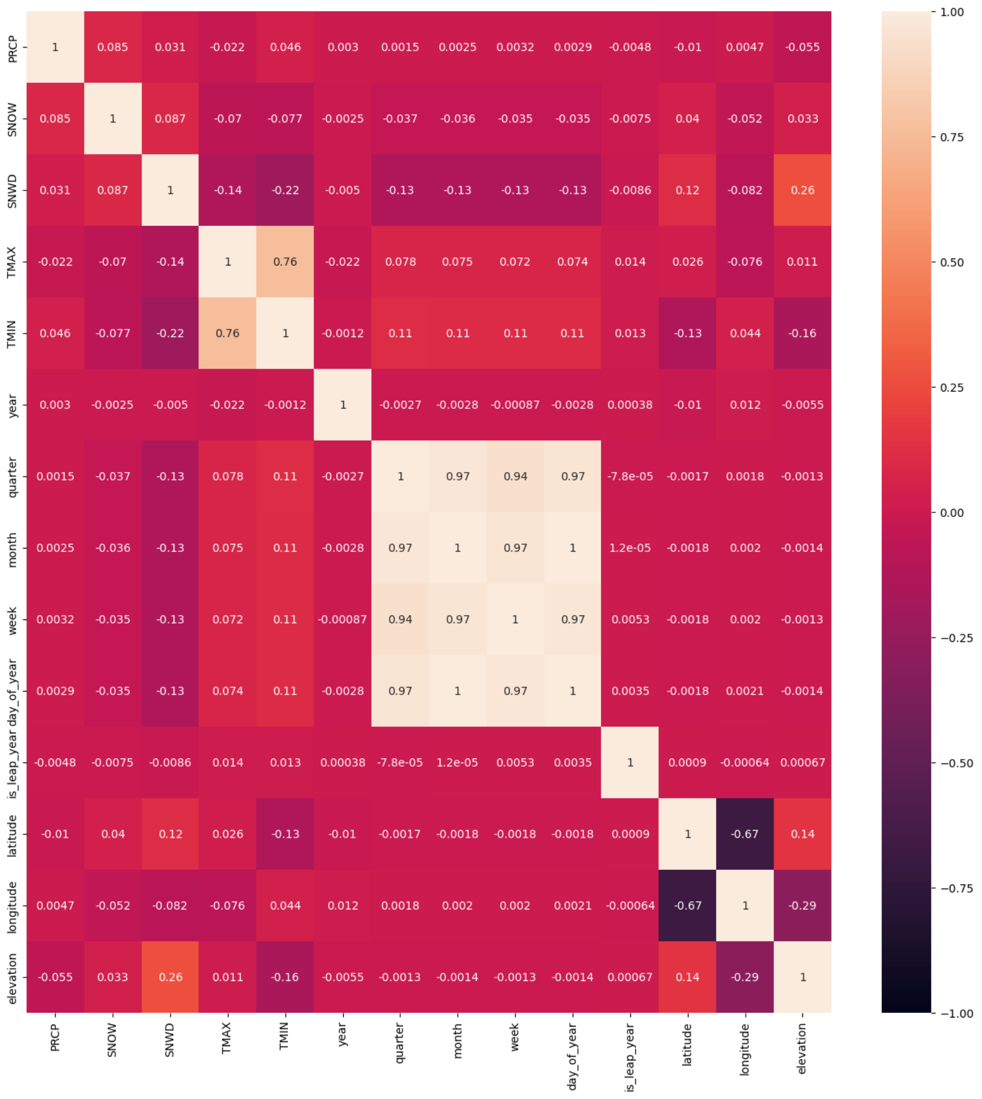
PREDICTIVE ANALYTICS



```
locationcategory_name = "CITY"
location_name = "San Francisco"
start_date = "2023-12-30"
duration = "7"
predict(locationcategory_name, location_name, start_date, duration, station_relations)
```

	Location	Date	TMAX (C)	TMIN (C)	PRCP (cm)	SNOW (cm)	SNWD (cm)
0	San Francisco, CA US	2023-12-30	7.196602	4.007489	16.250647	0.009818	-0.681265
1	San Francisco, CA US	2023-12-31	7.196602	4.007489	16.250647	0.009818	-0.681265
2	San Francisco, CA US	2024-01-01	8.431298	2.279749	3.353752	-0.052752	1.545340
3	San Francisco, CA US	2024-01-02	8.431298	2.279749	3.353752	-0.052752	1.545340
4	San Francisco, CA US	2024-01-03	8.431298	2.279749	4.143710	-0.009639	1.545340
5	San Francisco, CA US	2024-01-04	9.103164	3.137930	3.751980	-0.003760	1.545340
6	San Francisco, CA US	2024-01-05	9.103164	3.137930	3.751980	-0.003760	1.545340
7	San Francisco, CA US	2024-01-06	9.103164	3.137930	3.751980	0.386499	1.460063

Correlation Analysis



	PRCP	SNOW	SNWD	TMAX	TMIN	year	quarter	month	week	day_of_year	is_leap_year	latitude	longitude	elevation
PRCP	1	0.085	0.031	-0.022	0.046	0.003	0.0015	0.0025	0.0032	0.0029	-0.0048	-0.01	0.0047	-0.055
SNOW	0.085	1	0.087	-0.07	-0.077	-0.0025	-0.037	-0.036	-0.035	-0.035	-0.0075	0.04	-0.052	0.033
SNWD	0.031	0.087	1	-0.14	-0.22	-0.005	-0.13	-0.13	-0.13	-0.13	-0.0086	0.12	-0.082	0.26
TMAX	-0.022	-0.07	-0.14	1	0.76	-0.022	0.078	0.075	0.072	0.074	0.014	0.026	-0.076	0.011
TMIN	0.046	-0.077	-0.22	0.76	1	-0.0012	0.11	0.11	0.11	0.11	0.013	-0.13	0.044	-0.16
year	0.003	-0.0025	-0.005	-0.022	-0.0012	1	-0.0027	-0.0028	-0.00087	-0.0028	0.00038	-0.01	0.012	-0.0055
quarter	0.0015	-0.037	-0.13	0.078	0.11	-0.0027	1	0.97	0.94	0.97	-7.8e-05	-0.0017	0.0018	-0.0013
month	0.0025	-0.036	-0.13	0.075	0.11	-0.0028	0.97	1	0.97	1	1.2e-05	-0.0018	0.002	-0.0014
week	0.0032	-0.035	-0.13	0.072	0.11	-0.00087	0.94	0.97	1	0.97	0.0053	-0.0018	0.002	-0.0013
day_of_year	0.0029	-0.035	-0.13	0.074	0.11	-0.0028	0.97	1	0.97	1	0.0035	-0.0018	0.0021	-0.0014
is_leap_year	-0.0048	-0.0075	-0.0086	0.014	0.013	0.00038	-7.8e-05	1.2e-05	0.0053	0.0035	1	0.0009	-0.00064	0.00067
latitude	-0.01	0.04	0.12	0.026	-0.13	-0.01	-0.0017	-0.0018	-0.0018	-0.0018	0.0009	1	-0.67	0.14
longitude	0.0047	-0.052	-0.082	-0.076	0.044	0.012	0.0018	0.002	0.002	0.0021	-0.00064	-0.67	1	-0.29
elevation	-0.055	0.033	0.26	0.011	-0.16	-0.0055	-0.0013	-0.0014	-0.0013	-0.0014	0.00067	0.14	-0.29	1

- We can see that there is a white patch in the center.
- All those fields are highly positively correlated.
- Latitude and Longitude are negatively correlated.

CONCLUSION

AWS powered system efficiently manages data from various sources, making it easy for businesses . We used tools like AWS Glue for pipelining, Redshift as data warehouse , Airflow to Orchestrate the pipeline and many more. Our data system on AWS helps companies organize information easily, make smarter decisions using advanced analysis, future predictions and stay flexible as their needs evolve. It's like having a reliable assistant that keeps everything in order and helps the business grow smarter.

THANK YOU