ANALYSIS ON CLIMATE CHANGE AND IMPACT OF GREENHOUSE GAS EMISSION

Vijay Rama Raju Penmetsa
Aryama Ray
Vaishnavi Mocherla
Sai Sahithi Bethapudi
Ananya Varma Mudunuri

Abstract—

The primary aim of this project is to create a model that will be useful in predicting the patterns in environmental change in different areas of the world in the coming years and to utilize this data to study weather conditions, explicitly as to temperature, precipitation, and irregularities. We will also analyze the connection between the current shift in weather patterns and the historical production of carbon emissions by various countries. By expanding our understanding of climate change, extreme weather, and regional variations, the project aims to advance climate science. This will assist with asset portion, debacle planning, and environment variation independent direction, eventually encouraging strength and maintainability.

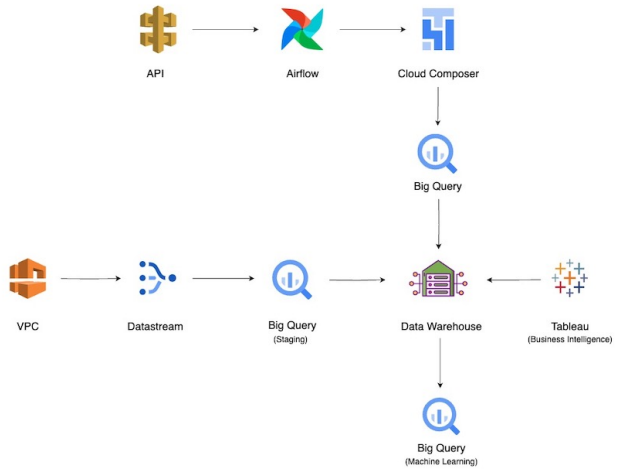
I. PROBLEM STATEMENT

The influence of human civilization on the climate is quite evident and at this point it is undeniable. Serious actions should be taken before it goes out of our hands and effect our future generations. The initial step is to analyze how is the climate being changing rapidly with respect to harmful gas emissions by humans. There are few long-term shift in temperature and weathers, these are called climate change. Whereas drastic change in climate due to greenhouse gas emissions like carbon dioxide and methane are the cause of human civilization.

II. SOLUTION REQUIREMENT

The approach of this project is that it provides this platform where all kinds of analyzing are readily available to understand and predict the climate changes. Here, we are going develop a predictive model using a pipeline with various google cloud services that helps in transforming the data , including real-time data into the warehouse and finally create a predictive model.

Pipeline using the google cloud services:



The data is first present in the VPC which is then sent to the big query using the data stream. Now this data is stored in the staging database in the big query. The warehouse is created using this staging database. On the other hand, the real-time data is sent through the API which is processed and scheduled to various transformations using airflow. This ETL is runned in cloud composer. This data is scheduled on daily basis and send to the warehouse. Now that the data is stored in the warehouse, we can perform various analysis and create a model to predict the future climate conditions. This data can be even used to perform visualizations using tableau.

III. DATA

There are two different types of data sources for this project. One is the archived data which is extracted from various websites. The other data source is real-time data which is updated on daily bases.

Cities

This table has all the data related to cities and their respective weather stations.

The station id provides the unique code to represent the station of the city. City, country and state has the names, iso2 and iso3 are the two letter and three letter

country code. Longitude and latitude hold the values corresponding to the city.

station_id	city_name	country	state	iso2	iso3	latitude	longitude
41515	Asadabad	Afghanistan	Kunar	AF	AFG	34.86600004	71.15000459
38954	Fayzabad	Afghanistan	Badakhshan	AF	AFG	37.12976076	70.57924719
41560	Jalalabad	Afghanistan	Nangarhar	AF	AFG	34.44152692	70.43610347
38947	Kunduz	Afghanistan	Kunduz	AF	AFG	36.72795066	68.87252966
38987	Qala i Naw	Afghanistan	Badghis	AF	AFG	34.98300013	63.13329964
38915	Sheberghan	Afghanistan	Jawzjan	AF	AFG	36.65798077	65.73830237
13577	Peshkopi	Albania	Dibër	AL	ALB	41.6833021	20.43330349
13461	Shkodër	Albania	Shkodër	AL	ALB	42.06845156	19.51884965
13615	Tirana	Albania	Durrës	AL	ALB	41.32754071	19.81888301
60620	Adrar	Algeria	Adrar	DZ	DZA	27.86999005	-0.2899670831
60369	Algiers	Algeria	Alger	DZ	DZA	36.7630648	3.05055253
60360	Annaba	Algeria	Annaba	DZ	DZA	36.92000612	7.759980834
60468	Batna	Algeria	Batna	DZ	DZA	35.56995933	6.170000365
60525	Biskra	Algeria	Biskra	DZ	DZA	34.85997683	5.73002722
60444	Bordj Bou Arréric	Algeria	Bordj Bou Arreri	DZ	DZA	36.05900401	4.629996466
60571	Béchar	Algeria	Béchar	DZ	DZA	31.61110537	-2.230003704
60402	Béjaïa	Algeria	Béjaïa	DZ	DZA	36.76037762	5.070015827

Countries

Like the city table, the country table also has information such as country name, language, iso2 and iso3 like the cities table iso and iso3. The population of the country, the area in square kilometers of the country, capital name, along with its latitude and longitude and finally the region of the country along with its continent.

country	native_name	iso2	iso3	population	area	capital	capital_lat	capital_lng	region
Afghanistan	افغانستان	AF	AFG	26023100	652230	Kabul	34.526011	69.177684	Southern and C
Albania	Shqipëria	AL	ALB	2895947	28748	Tirana	41.326873	19.818791	Southern Europ
Algeria	الجزائر	DZ	DZA	38700000	2381741	Algiers	36.775361	3.060188	Northern Africa
American Samo	American Samoa	AS	ASM	55519	199	Pago Pago	-14.275479	-170.70483	Polynesia
Angola	Angola	AO	AGO	24383301	1246700	Luanda	-8.82727	13.243951	Central Africa
Anguilla	Anguilla	AI	AIA	13452	91	The Valley	41.559572	-98.980548	Caribbean
Antigua and Bar	Antigua and Bar	AG	ATG	86295	442	Saint John's	47.561701	-52.715149	Caribbean
Argentina	Argentina	AR	ARG	42669500	2780400	Buenos Aires	-34.607568	-58.437089	South America
Armenia	Հայաստան	AM	ARM	3009800	29743	Yerevan	40.177612	44.512585	Middle East
Aruba	Aruba	AW	ABW	101484	180	Oranjestad	12.526874	-70.035684	Caribbean
Australia	Australia	AU	AUS		7692024	Canberra	-35.297591	149.101268	Australia and Ne
Austria	Österreich	AT	AUT	8527230	83871	Vienna	48.208354	16.372504	Western Europe
Azerbaijan	Azərbaycan	AZ	AZE	9552500	86600	Baku	40.375443	49.832675	Middle East
Bahrain	البحرين	BH	BHR	1316500	765	Manama	26.223504	50.582244	Middle East
Bangladesh	বাংলাদেশ	BD	BGD	157486000	147570	Dhaka	23.759357	90.378814	Southern and C
Belarus	Беларусь	BY	BLR	9475100	207600	Minsk	53.902334	27.561879	Eastern Europe
Belgium	België	BE	BEL	11225469	30528	Brussels	50.846557	4.351697	Western Europe
Belize	Belize	BZ	BLZ	349728	22966	Belmopan	17.250199	-88.770018	Central America

Daily Weather

This table provides the weather data of the countries daily, an archived dataset. These tables have all kinds of data like the station id, city name, date, season, average temperature that day along with the minimum and maximum temperature. It also documents the precipitation, the snow death, and other kinds of weather reports like these.

station_id	city_name	date	season	avg_temp_c	min_temp_c	max_temp_c	precipitation_mm	snow_depth_mm	avg_wind_dir_de	avg_wind_speed	peak_wind_gust	avg_sea_level_p	sunshine_
41515	Asadabad	6/30/1957, 5:00:	Summer	27	21.1	35.6	0	-	-	-	-	-	-
41515	Asadabad	7/1/1957, 5:00:0	Summer	22.8	18.9	32.2	0	-	-	-	-	-	-
41515	Asadabad	7/2/1957, 5:00:0	Summer	24.3	16.7	35.6	1	-	-	-	-	-	-
41515	Asadabad	7/3/1957, 5:00:0	Summer	26.6	16.1	37.8	4.1	-	-	-	-	-	-
41515	Asadabad	7/4/1957, 5:00:0	Summer	30.8	20	41.7	0	-	-	-	-	-	-
41515	Asadabad	7/5/1957, 5:00:0	Summer	30.2	22.8	41.1	0	-	-	-	-	-	-
41515	Asadabad	7/6/1957, 5:00:0	Summer	31	24.4	39.4	0	-	-	-	-	-	-
41515	Asadabad	7/7/1957, 5:00:0	Summer	30.9	24.4	38.9	0	-	-	-	-	-	-
41515	Asadabad	7/8/1957, 5:00:0	Summer	28.1	21.1	34.4	2	-	-	-	-	-	-
41515	Asadabad	7/9/1957, 5:00:0	Summer	26	-	35.6	0.3	-	-	-	-	-	-
41515	Asadabad	7/10/1957, 5:00:	Summer	26.3	17.2	36.1	2	-	-	-	-	-	-
41515	Asadabad	7/11/1957, 5:00:	Summer	28.8	21.7	36.7	0	-	-	-	-	-	-
41515	Asadabad	7/12/1957, 5:00:	Summer	27.2	21.1	36.1	0	-	-	-	-	-	-
41515	Asadabad	7/13/1957, 5:00:	Summer	28	20.6	36.1	0.3	-	-	-	-	-	-
41515	Asadabad	7/14/1957, 5:00:	Summer	28.6	21.1	37.2	0	-	-	-	-	-	-
41515	Asadabad	7/15/1957, 5:00:	Summer	31.7	22.8	41.7	0	-	-	-	-	-	-
41515	Asadabad	7/16/1957, 5:00:	Summer	33.1	23.3	46.1	0	-	-	-	-	-	-
41515	Asadabad	7/17/1957, 5:00:	Summer	33.3	26.1	41.1	0	-	-	-	-	-	-
41515	Asadabad	7/18/1957, 5:00:	Summer	30.1	25	35.6	1	-	-	-	-	-	-
41515	Asadabad	7/19/1957, 5:00:	Summer	27.6	21.1	34.4	3	-	-	-	-	-	-
41515	Asadabad	7/20/1957, 5:00:	Summer	28.8	22.2	35	0	-	-	-	-	-	-
41515	Asadabad	7/21/1957, 5:00:	Summer	27.4	21.7	33.9	0.3	-	-	-	-	-	-

Cumulative greenhouse gas emission

This data is extracted from the website that has data related to greenhouse gas emission.

CNTR_NAME	ISO3	Gas	Component	Year	Data	Unit
Afghanistan	AFG	3-GHG	Fossil	1851	0.000454704256	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1852	0.000913130775	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1853	0.001375296506	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1854	0.00184121966	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1855	0.002310915871	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1856	0.002784399916	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1857	0.003261685266	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1858	0.003742784294	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1859	0.004227706791	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1860	0.004716465226	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1861	0.005209240717	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1862	0.005706055555	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1863	0.006206932535	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1864	0.006711894976	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1865	0.007220965284	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1866	0.007734169016	Pg~CO[2]*-e[100]
Afghanistan	AFG	3-GHG	Fossil	1867	0.008251530954	Pg~CO[2]*-e[100]

Real time data

There is also real time data that is in json format that is going to be used for this project.

This data has information like city name, date, season, average, minimum and maximum temperature of the date. As well as related data about precipitation, wind sea level and finally sunshine time. All these data are real-time, meaning they get updated regularly and in this case daily.

city_name	date	season	avg_temp_c	min_temp_c	max_temp_c	precipitation_mm	avg_wind_dir_deg	avg_wind_speed_km	peak_wind_gust_km	avg_sea_level_pres	sunshine_total_min
Helena	2023-12-03	Winter	2.43	-2.49	4.54	0	249	26.712	66.744	1013	528.1333333
Montpelier	2023-12-03	Winter	3.53	1.34	3.53	0.99	169	8.244	29.808	1014	544.5333333
Bismarck	2023-12-03	Winter	3.51	-2.36	4.28	0	295	18.036	25.74	1006	526.6
Saint Paul	2023-12-03	Winter	1.79	-1.3	2.18	0	141	9.108	24.588	1010	539.85
Cheyenne	2023-12-03	Winter	-0.34	-3.81	3.3	0	267	63.036	85.4	1014	563.9833333
Madison	2023-12-03	Winter	15.53	9.59	16.23	0.55	272	17.676	27.828	1014	598.35
Denver	2023-12-03	Winter	3.75	2.67	5.49	0	284	23.616	57.708	1013	572.0666667
Des Moines	2023-12-03	Winter	5.36	-1.36	5.42	0	212	11.952	15.84	1009	561.2666667
Indianapolis	2023-12-03	Winter	7.1	2.48	7.42	0	242	23.004	42.372	1008	571.9666667
Lincoln	2023-12-03	Winter	4.94	-2.95	5.97	0	345	17.748	29.592	1007	566
Boise	2023-12-03	Winter	3.49	1.78	5.44	13.19	112	13.392	21.564	1020	548.6166667
Albany	2023-12-03	Winter	4.79	3.87	5.31	18.82	149	14.364	39.132	1011	555.2166667
Topeka	2023-12-03	Winter	7.87	-1.03	8.46	0	323	22.536	42.156	1008	575.9666667
Columbus	2023-12-03	Winter	10.98	5.88	11.79	0.56	249	26.532	53.244	1007	570.8833333
Springfield	2023-12-03	Winter	9.68	-0.82	10.9	0	262	26.784	51.48	1012	585.8166667
Jefferson City	2023-12-03	Winter	8.04	3.41	9.42	0	268	19.548	37.008	1013	578.55
Frankfort	2023-12-04	Winter	32.31	16.67	34.36	0	316	18.972	33.984	1011	625.3166667

IV: ELT PROCESS

The VPC is a virtual private cloud. Our archived data is present in this VPC. This provides a cloud portion which acts like traditional offline data center network. The VPC is first created. In which the SQL cloud instance is put. Now we can connect this to the big Query using the Data Stream.

DATA STREAM

This data stream is an end-to-end connection which basically means it has end to end encryption so there are not going to be any attacks or loss of data through the process of data transfer from the VPC to the Big Query.

This is important to remember to set up security authorizations to the allow DataStream to push data into the big query. Before we initialize the data stream, it is advisable to create a replication data stream process to completely understand if the stream is facing with any issues and resolve them so that these errors can be avoided in the actual data stream and let the data be pushed to the Big Query in a smoothy without any errors since the initialization.

After the initialization, we will be able to load the data, which is considered as raw data into the Big Query. In this big query, we will create a staging database in which the raw data is temporarily stored.

V.AIRFLOW

ETL Processing

Since the real-time data is huge and constantly keep changing, it is important to handle this data. This kind of data needs some scheduling and transformations whenever the API is updated with new data.

The Apache airflow is a type of service that can orchestrate the workflows that are complex. It is a powerful tool that helps in scheduling various tasks, manage data pipeline basically using DAGs.

For this project, the Apache airflow is used to create DAGs to define, schedule and execute a few tasks which basically form a pipeline.

This DAG perform various tasks:

Pulling data: It extracts the data from openweather API for a particular list of cities. Now this data is then stored in weather_data_all which is Xcom to be used in subsequent tasks.

Validate Data: This task basically validates the retrieved data, checks if there are any missing or errors in the data. It then processes the data after either correcting or filling of the data, subsequently updates the table in Big Query with the number of records or missing records.

Transform Data: This raw data is then structured into a data frame including changing few column datatypes to appropriate types. Which is then pushed into weather_df to Xcom.

Load Data:

Now this task is used to load the data into Big Query. This defines the schema data-225-group-project. climate_.real_time_data while converting to appropriate datatypes. This is loaded into a staging database.

Staging:

Small transformations are performed on the data where duplicates are dropped and the data is loaded inot the warehouse

This pipeline is schedules to run daily since the API is updated daily. This whole ETL is run using cloud composer.

VI. DATAWAREHOUSE

Datawarehouse:

Datawarehouse is one of the most important storing techniques used for any kind of efficient project that specifically have lot of data to work with. It helps in storing the data that is centralized and can intergrade different kinds of data. It is also mostly used to store historical data on which analytics are performed. One can decide in which structure to store the data in, for example in snowflake or star schema.

There are various Datawarehouse tools in GCP in which BigQuery is used for this project because of its infrastructure to store data and retrieve it to form various analytics.

Big Query is a GCP which provides a cloud-based warehouse that can be used for storing the data and using this data for further analysis. It can help in providing analysis for real time data in huge volumes.

For Big Query, the schema structured is used star scheme, which is one of the most famous techniques for structuring the data in warehouse. In a star scheme, there is a central table called fact table which basically connects with other tables called dimensions table with the help of foreign keys. This structure is called a start schema because of its representation which looks like a star with tables connecting to one fact table.

The raw data which is sent through the data stream is fetched and saved under one dataset called climate-data-staging. From this dataset, the data is transformed into star schema, performing various operations like changing the datatypes of the columns, expanding the date column and so on.

For this project, we have converted the data into start schema structure this way:

<u>climate_fact:</u>

This table is a fact table which consist of attributes like

Station id, record date, minimum, maximum , average temperature , snow information , precipitation , wind data , sunshine level and the greenhouse gas data

Field name	Type	Mode
stationid	STRING	NULLABLE
record_date	DATE	NULLABLE
avg_temp_c	FLOAT	NULLABLE
min_temp_c	FLOAT	NULLABLE
max_temp_c	FLOAT	NULLABLE
precipitation_mm	FLOAT	NULLABLE
snow_depth_mm	FLOAT	NULLABLE
avg_wind_dir_deg	FLOAT	NULLABLE
avg_wind_speed_kmh	FLOAT	NULLABLE
peak_wind_gust_kmh	FLOAT	NULLABLE
avg_sea_level_pres_hpa	FLOAT	NULLABLE
sunshine_total_min	FLOAT	NULLABLE
ghgas_data	FLOAT	NULLABLE

There are 4 primary dimension tables created.

Location dimension:

This table is extracted from the city and country tables and has all the information related to the location, hence its name.

Field name	Type	Mode
station_id	STRING	NULLABLE
city_name	STRING	NULLABLE
state	STRING	NULLABLE
country	STRING	NULLABLE
iso3	STRING	NULLABLE
capital	STRING	NULLABLE
region	STRING	NULLABLE
continent	STRING	NULLABLE

Date Dimension:

This tables is extracted from daily-weather dataset and after apply various transactions on the data like extracting month , year , date from record date column.

Field name	Type	Mode
date_id	INTEGER	NULLABLE
record_date	DATE	NULLABLE
record_week	INTEGER	NULLABLE
record_month	INTEGER	NULLABLE
record_quarter	INTEGER	NULLABLE
record_year	INTEGER	NULLABLE
season	STRING	NULLABLE

Greenhouse gas emission Dimension:

This tables are completely extracted from the greenhouse gas dataset.

Field name	Type	Mode
ghgas_id	INTEGER	NULLABLE
ghgas_country	STRING	NULLABLE
Gas	STRING	NULLABLE
Component	STRING	NULLABLE
Year	INTEGER	NULLABLE

Real Time Data:

This consist of data from the API, which provides the real-time data, that gets updated on daily basis.

Field name	Type	Mode
city_name	STRING	NULLABLE
date	DATE	NULLABLE
season	STRING	NULLABLE
avg_temp_c	FLOAT	NULLABLE
min_temp_c	FLOAT	NULLABLE
max_temp_c	FLOAT	NULLABLE
precipitation_mm	FLOAT	NULLABLE
avg_wind_dir_deg	FLOAT	NULLABLE
avg_wind_speed_kmh	FLOAT	NULLABLE
peak_wind_gust_kmh	FLOAT	NULLABLE
avg_sea_level_pres_hpa	FLOAT	NULLABLE
sunshine_total_min	FLOAT	NULLABLE

Data Tracking

This table basically records the real time on how many records are being transferring or missing. This is for the monitoring of real time data only.

Field name	Type	Mode
Date	DATETIME	NULLABLE
Recorded	INTEGER	NULLABLE
Missing_records	INTEGER	NULLABLE

VII. DATAWAREHOUSE

Various data analysis can be performed using this data like:

1-Top Cities by Average Temperature

```
SELECT
    city_name,
    AVG(avg_temp_c) AS avg_temperature
FROM
    `data-225-group-project.climate_dwh.climate_fact`
fact
JOIN
    `data-225-group-project.climate_dwh.location_dim`
location
ON
    fact.stationid = location.station_id
GROUP BY
    city_name
ORDER BY
    avg_temperature DESC
LIMIT 20;
```

Row	city_name	avg_temperature
1	Honolulu	24.89408247489...
2	Phoenix	22.91001673478...
3	Austin	20.45615246291...
4	Tallahassee	19.85505529485...
5	Montgomery	18.47826236387...
6	Columbia	17.68848571724...
7	Little Rock	17.07175005845...
8	Atlanta	16.77345754785...
9	Sacramento	15.70770115343...
10	Oklahoma City	15.67218438621...
11	Raleigh	15.49904015468...

2-Top Cities with the Most Extreme Weather Events

```
SELECT
   l.city_name,
   COUNT(*) AS extreme_events_count
FROM
   `data-225-group-project.climate_dwh.climate_fact`
c
JOIN
   `data-225-group-project.climate_dwh.location_dim`
l
ON
   c.stationid = l.station_id
WHERE
   c.max_temp_c > 35 OR c.min_temp_c < 0 OR
c.precipitation_mm > 100
GROUP BY
   l.city_name
ORDER BY
   extreme_events_count DESC
LIMIT 10;
```

1	Montpelier	20760
2	Cheyenne	17977
3	Helena	14539
4	Bismarck	14457
5	Carson City	14426
6	Concord	13664
7	Springfield	13483
8	Saint Paul	12669
9	Denver	12603
10	Madison	12417

3-Extreme Climate Events by City with Corresponding Dates

```
WITH ExtremeDates AS (
  SELECT
    stationid,
    MAX(max_temp_c) AS max_temperature,
    MIN(min_temp_c) AS min_temperature,
    MAX(peak_wind_gust_kmh) AS max_wind_speed,
    MAX(precipitation_mm) AS max_precipitation
  FROM
    `data-225-group-project.climate_dwh.climate_fact`
  GROUP BY
    stationid
)

SELECT
    location.city_name,
    dates.record_date AS date_max_temperature,
    dates_min.record_date AS date_min_temperature,
    dates_wind.record_date AS date_max_wind_speed,
    dates_precipitation.record_date AS
date_max_precipitation,
    extreme_dates.max_temperature,
    extreme_dates.min_temperature,
    extreme_dates.max_wind_speed,
    extreme_dates.max_precipitation
FROM
    ExtremeDates extreme_dates
JOIN
    `data-225-group-project.climate_dwh.climate_fact`
dates
    ON extreme_dates.stationid = dates.stationid
    AND extreme_dates.max_temperature =
dates.max_temp_c
JOIN
    `data-225-group-project.climate_dwh.climate_fact`
dates_min
    ON extreme_dates.stationid = dates_min.stationid
    AND extreme_dates.min_temperature =
dates_min.min_temp_c
JOIN
    `data-225-group-project.climate_dwh.climate_fact`
dates_wind
    ON extreme_dates.stationid = dates_wind.stationid
    AND extreme_dates.max_wind_speed =
dates_wind.peak_wind_gust_kmh
JOIN
    `data-225-group-project.climate_dwh.climate_fact`
dates_precipitation
    ON extreme_dates.stationid =
dates_precipitation.stationid
    AND extreme_dates.max_precipitation =
dates_precipitation.precipitation_mm
JOIN
```

`data-225-group-project.climate_dwh.location_dim` location
ON extreme_dates.stationid = location.station_id;

city_name	date_max_temperatu	date_min_temperatu	date_max_wind_spe	date_max_precipitati	max_temperature	min_temperature	max_wind_speed	max_precipitation
Des Moines	1983-08-16	1996-02-03	1993-07-24	1975-08-27	42.2	-32.2	133.2	157.0
Springfield	1954-07-14	1905-02-13	1975-11-29	2016-08-12	44.4	-31.1	122.0	142.0
Boston	2011-07-22	1943-02-15	1954-08-31	1955-08-19	39.4	-25.6	160.9	179.3
Concord	1966-07-03	1943-02-16	1995-04-05	2006-05-13	38.9	-38.3	97.9	130.0
Jackson	2012-06-29	1985-01-21	1985-12-01	2015-04-03	40.0	-27.8	96.1	109.0
Jackson	2012-06-29	1985-01-21	1988-02-15	2015-04-03	40.0	-27.8	96.1	109.0
Jackson	2012-06-29	1985-01-21	1993-06-04	2015-04-03	40.0	-27.8	96.1	109.0
Jackson	2012-06-29	1985-01-20	1985-12-01	2015-04-03	40.0	-27.8	96.1	109.0
Jackson	2012-06-29	1985-01-20	1988-02-15	2015-04-03	40.0	-27.8	96.1	109.0
Jackson	2012-06-29	1985-01-20	1993-06-04	2015-04-03	40.0	-27.8	96.1	109.0
Jackson	2012-06-29	1994-01-19	1985-12-01	2015-04-03	40.0	-27.8	96.1	109.0

4-Total Greenhouse Gas Emissions by Country

```
SELECT
   ghgas_country,
   SUM(ghgas_id) AS total_emissions
FROM
   `data-225-group-project.climate_dwh.ghgas_dim`
GROUP BY
   ghgas_country
ORDER BY
   total_emissions DESC;
```

Row	ghgas_country	total_emissions
1	Zimbabwe	419699886
2	Zambia	418727198
3	Yemen	417754510
4	Viet Nam	416568318
5	Venezuela	415595630
6	Vanuatu	414622942
7	Uzbekistan	413650254
8	USA	412677566
9	Uruguay	411704878
10	United Kingdom	410732190
11	United Arab Emirates	409759502

5-Gas Component Distribution

```
SELECT
   Gas,
   COUNT (*) AS component_count
FROM
   `data-225-group-project.climate_dwh.ghgas_dim`
GROUP BY
   Gas;
```

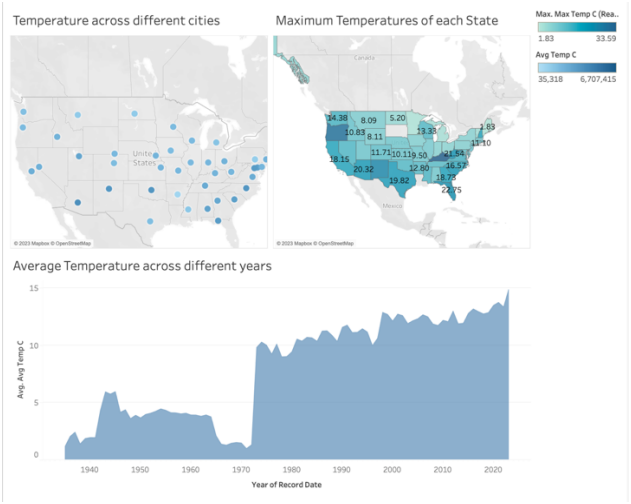
Gas	component_count
CH[4]	121152
CO[2]	126744
N[2]*O	121152

6-Years with the most gas emission in USA

```
SELECT
   Year,
   SUM(ghgas_id) AS total_emissions
FROM
   `data-225-group-project.climate_dwh.ghgas_dim`
WHERE
   ghgas_country = 'USA'  -- Adjust this condition based on your actual data
GROUP BY
   Year
ORDER BY
   total_emissions DESC
LIMIT 10;
```

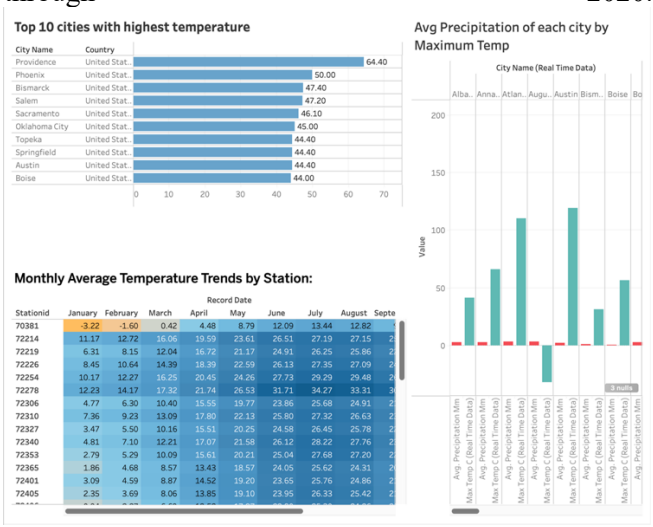
Year	total_emissions
2021	2175560
2020	2175551
2019	2175542
2018	2175533
2017	2175524
2016	2175515
2015	2175506
2014	2175497
2013	2175488
2012	2175479

VIII. VISUALIZATION



The graphic consists of three graphs that display temperature data for several American cities, states, and years. The first graph shows a map of the US with

various cities represented by blue dots. The dots' sizes reflect the temperature in the city. The second graph shows a map of the US with the highest temperatures in each state indicated by shades of blue and green. The temperature in such state is reflected in the shade's intensity. The average temperature is displayed on a line graph in the third graph, which spans the years 1940 through 2020.



This is a data visualization image that shows the top 10 cities with the highest temperature and the monthly average temperature trends by station. The image is divided into three sections: the top left section shows a table of the top 10 cities with the highest temperature, the top right section shows a bar graph of the average precipitation of each city by maximum temperature, and the bottom section shows a table of the monthly average temperature trends by station.

IX. MACHINE LEARNING IN BIG QUERY

From the data we can perform various kinds of techniques such as time series, linear regression modeling, etc.

For this project, we were able to perform time series to predict the future weather reports based on the archived data information.

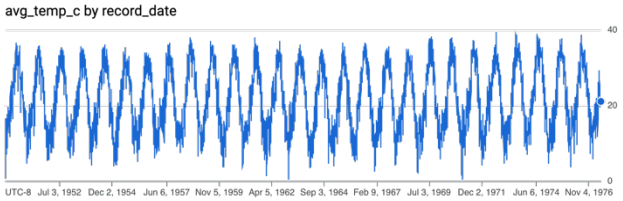
ARIMA model

In Big Query, with the Machine learning functions we can perform time series analysis. Using max_temp_c as the target variable and the record date as the timestamp, AutoRegressive Integrated moving average time series can be performed.

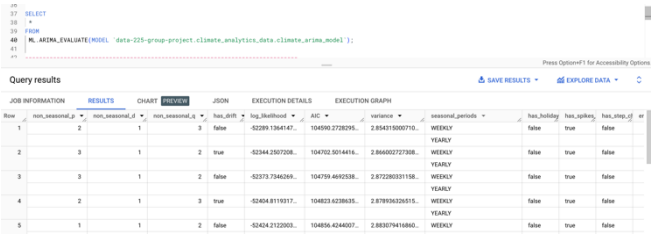
```
-- Time Series Model: ARIMA model builing
-------------------------------------------------------------------------------

CREATE OR REPLACE MODEL `data-225-group-project.climate_analytics_data.climate_arima_model`
OPTIONS
  (model_type = 'ARIMA_PLUS',
   time_series_timestamp_col = 'record_date',
   time_series_data_col = 'max_temp_c',
   auto_arima = TRUE,
   data_frequency = 'AUTO_FREQUENCY',
   decompose_time_series = TRUE
  ) AS
select record_date,max_temp_c
 from `data-225-group-project.climate_dwh.climate_fact` cl_fact inner join
`data-225-group-project.climate_dwh.location_dim` l_dim on cl_fact.stationid=l_dim.station_id
where state='Arizona' and  extract(year from record_date)>=1950;
```

Based on the maximum temperature in Arizona since 1950



This model can be evaluated by calculation different types of metrices like mean absolute error , mean squared error , etc



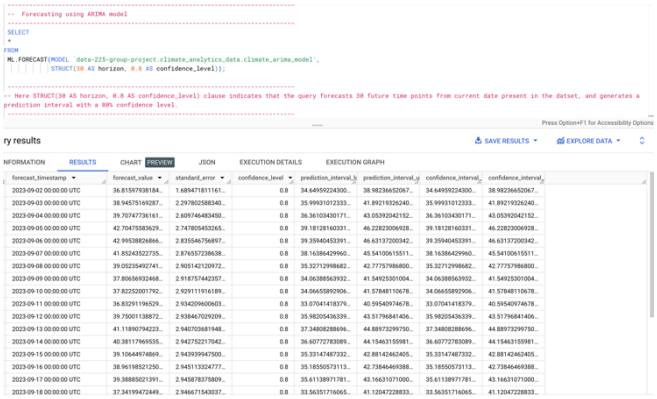
We can also retrieve the correlation of the model along with the autoregressive and moving average (MA) coefficient

```
SELECT
 *
FROM
 ML.ARIMA_COEFFICIENTS(MODEL `data-225-group-project.climate_analytics_data.climate_arima_model`);
```

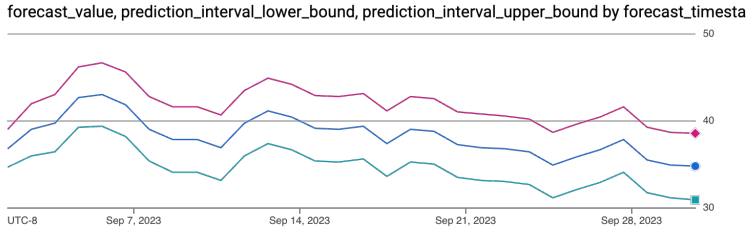
Results:

Model results can be seen using trained ARIMA model.

To display these results, we've used Arizona data from the year 1950 to August 2023 and tried predicting the weather of the next 30 days using the model.



Here, the results indicate the weather predictions of the next 30 days using the ARIMA model.



X. CONCLUSION

This is how we can use google cloud services to predict the weather change in the future. However, this project is only centered to the cities in USA. This is because of the limitation with the cloud storage for the version being used. Furthermore, the data in the archive data and real-time data are not consistent which led us to extract only 101 cities in USA.

The future scope of this project could be that, instead of predicting just for few cities in United Sates, it can be used for predicting globally.