DATA 228

BIG DATA BAYWATCH: CALIFORNIA FLOODS EDITION.

Maharsh Soni

Contents

Introduction

01

Flooding events have been increasing in severity and frequency in California due to climate change.

California has witnessed a 34% YoY increase in flooding events over the past decade. Despite the increase in flooding events, the number of flood insurance policies sought has dropped by 5.8%.

02

03

The 2023 floods in California resulted in financial losses worth $31 billion, six times higher than the national average of flood-related financial losses since 2009. Heavy rains are the leading cause of flooding events in California.

This project utilizes climate and flooding event data from NOAA and insurance policy data from FEMA to develop a data warehouse for efficient analytics and data-driven insights for better-informed flood-related policy-making.

Existing Solutions

FIRMs(Flood Insurance Rate Maps) are outdated (last updated in 2009) and lack interactivity

FEMA flood map visualizations are not user-friendly and informative to the general audience

FEMA National Risk Index map has limited interactivity and does not provide actual information such as numbers of floods or flood insurance policies sought by the counties

FEMA historical flood risk Tableau dashboards show floods only for which insurance claims have been sought and do not offer granular information

There is no one solution that integrates flood data and insurance policy data for comprehensive analysis

Problem & Solution Proposed

The lack of a real-time big data analytics monitoring system makes it difficult for residents in California to take early action in response to floods and subscribe to flood insurance programs.

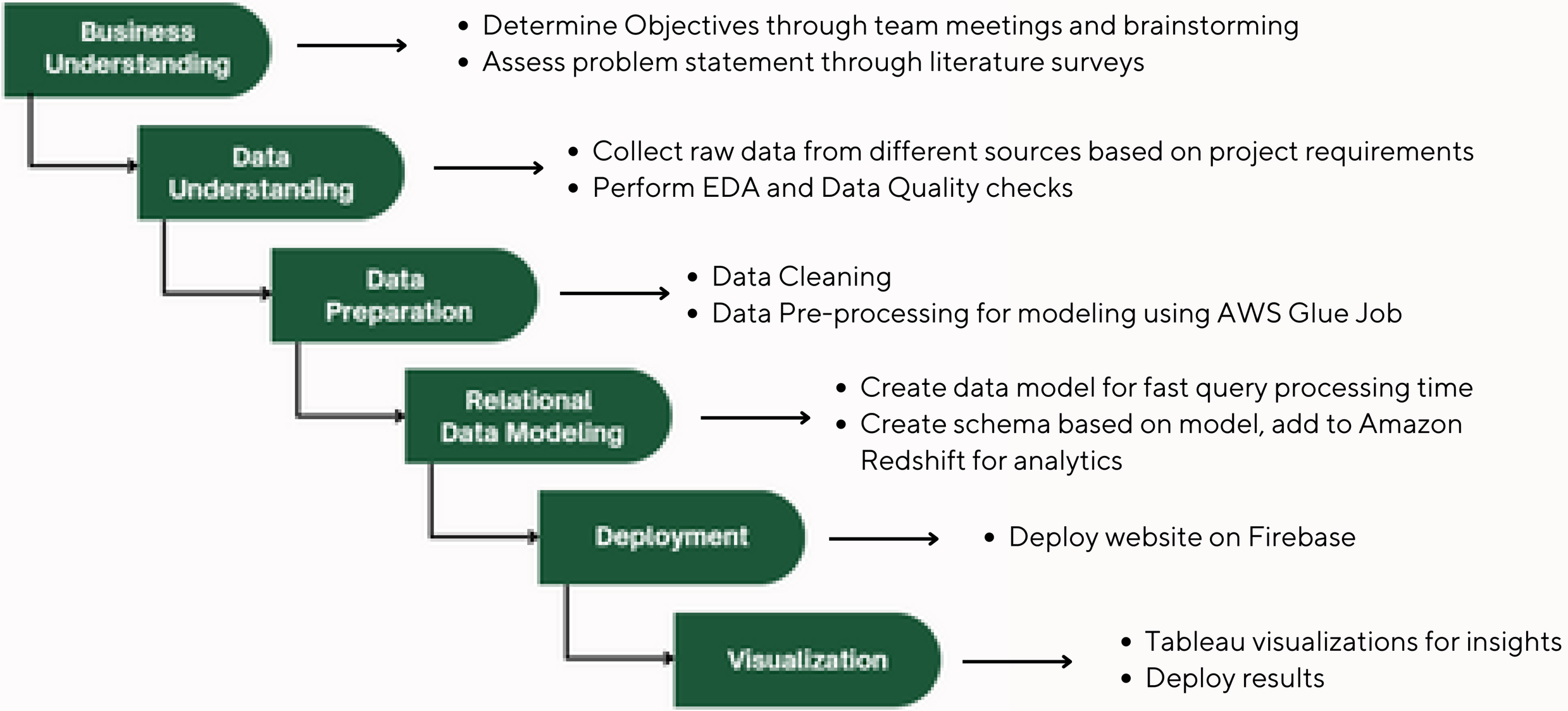
The proposed project aims to address this issue by delivering three academic/technical contributions.

Firstly, it proposes a big data analytics system design that can process and analyze large volumes of climate and flood insurance transaction data in real-time using advanced data processing techniques.

Secondly, the project aims to develop a data warehouse solution consisting of tables that can be used to create new flood risk assessment models based on climate movements, and provide monthly updates to reflect changes in the environment and climate.

Finally, the project aims to provide insights through Tableau dashboards to inform the development of new flood insurance policies and programs that can be customized to the needs of California residents and encourage more people to subscribe to flood insurance based on timely updated insights and predictions provided by the real-time big data analytics system.

CRISP-DM Methodology



- Determine Objectives through team meetings and brainstorming
- Assess problem statement through literature surveys

- Collect raw data from different sources based on project requirements
- Perform EDA and Data Quality checks

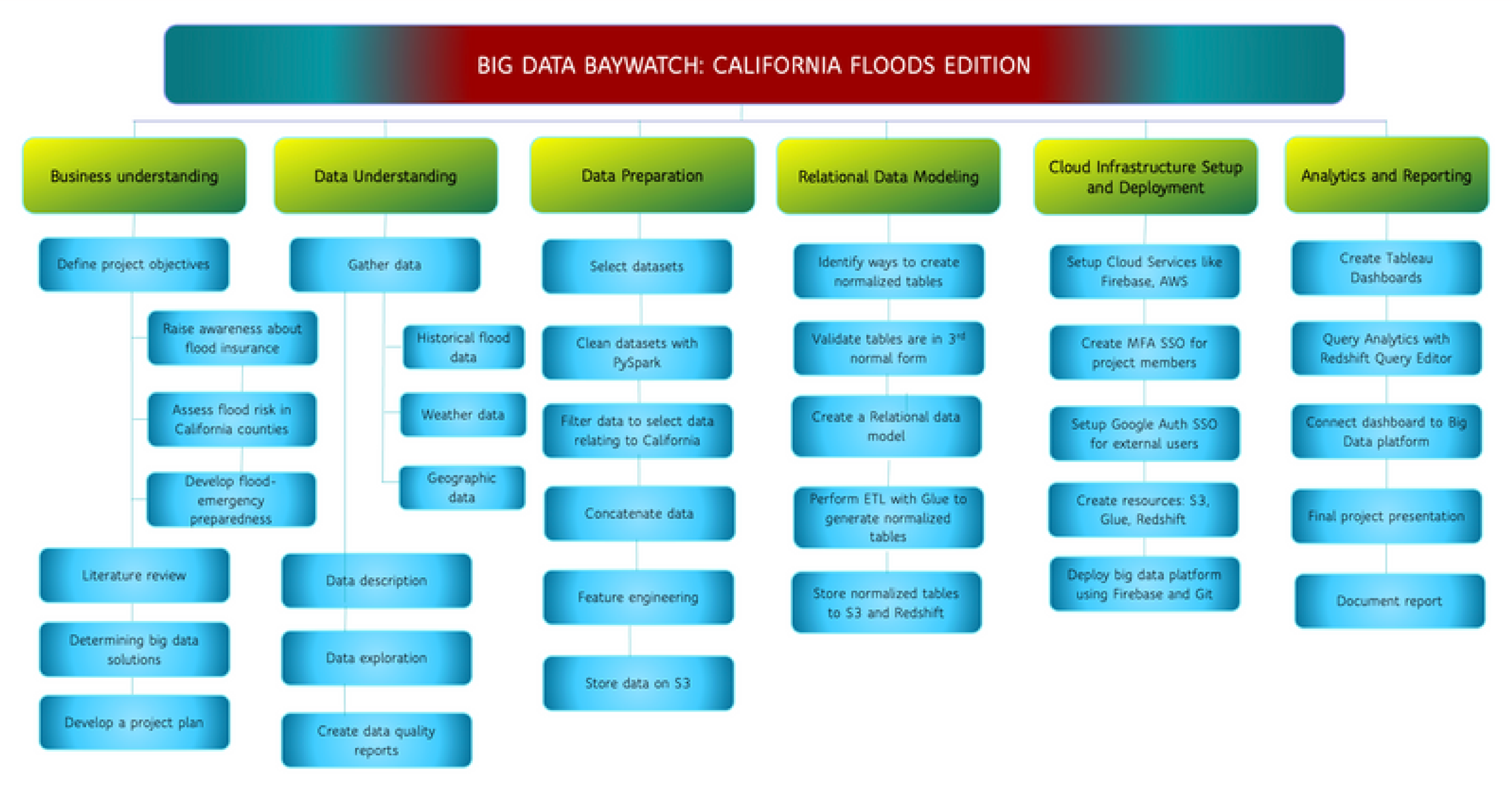
- Data Cleaning
- Data Pre-processing for modeling using AWS Glue Job

- Create data model for fast query processing time
- Create schema based on model, add to Amazon Redshift for analytics

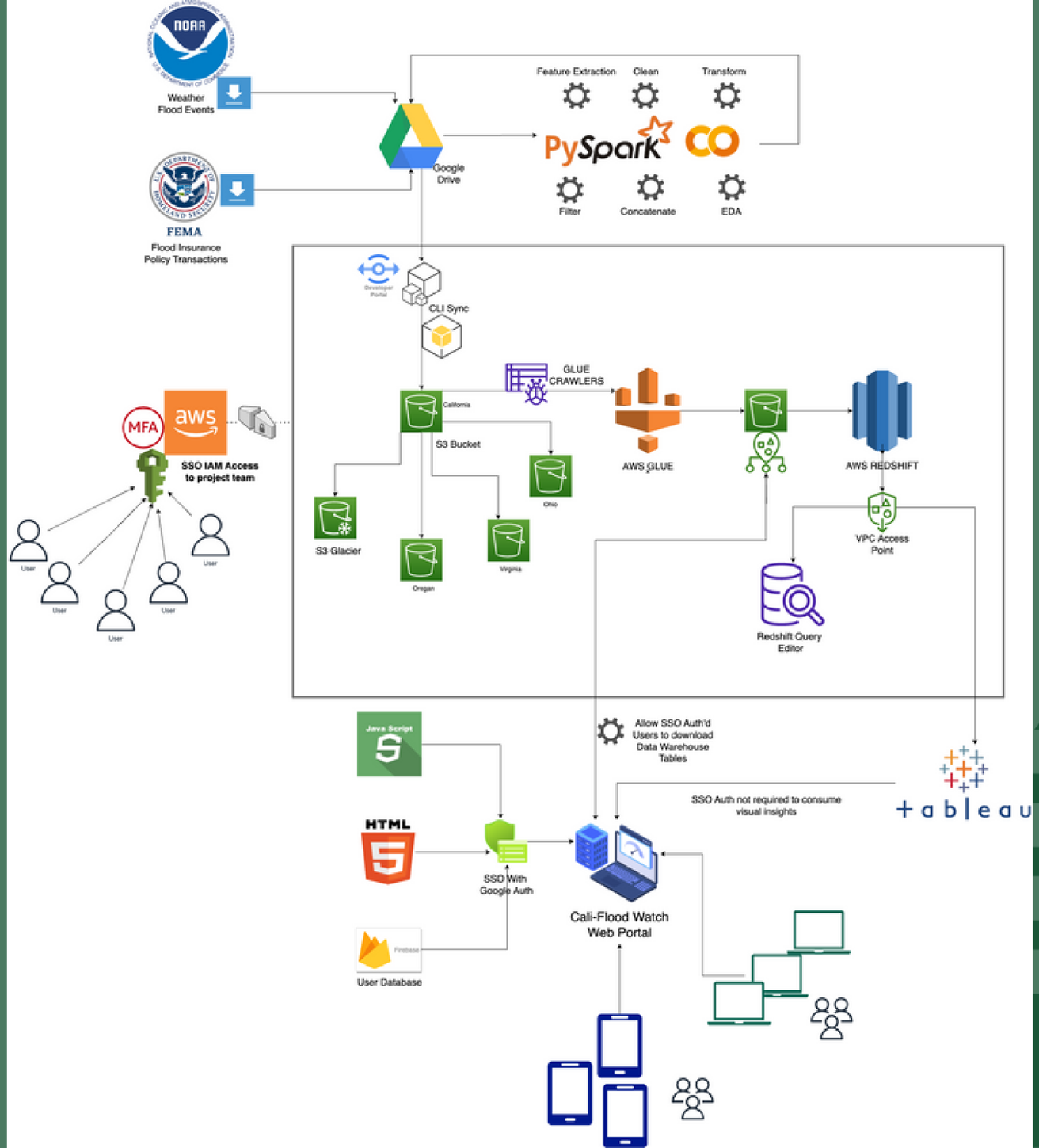
- Deploy website on Firebase

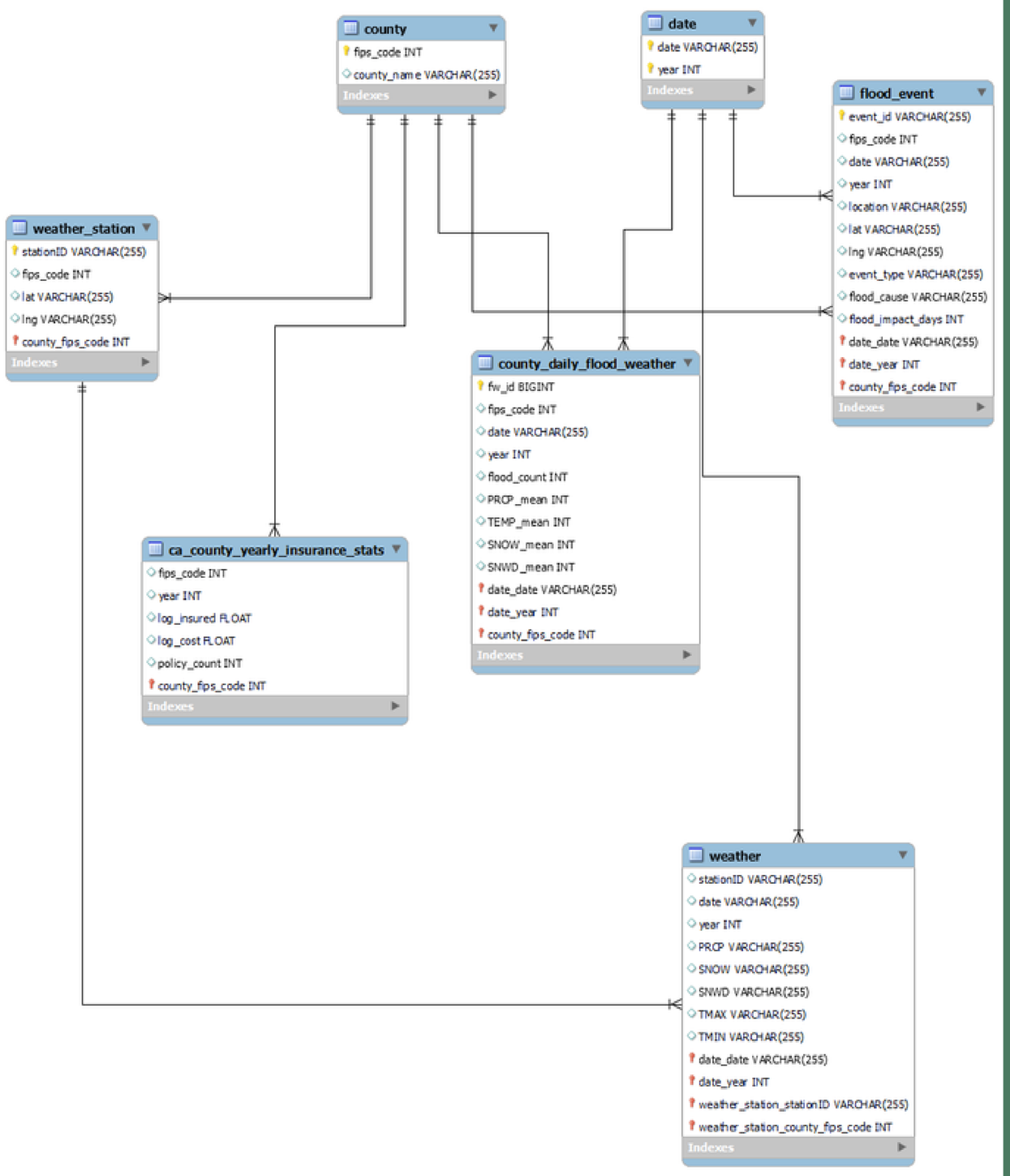
- Tableau visualizations for insights
- Deploy results

Work Breakdown Structure Chart



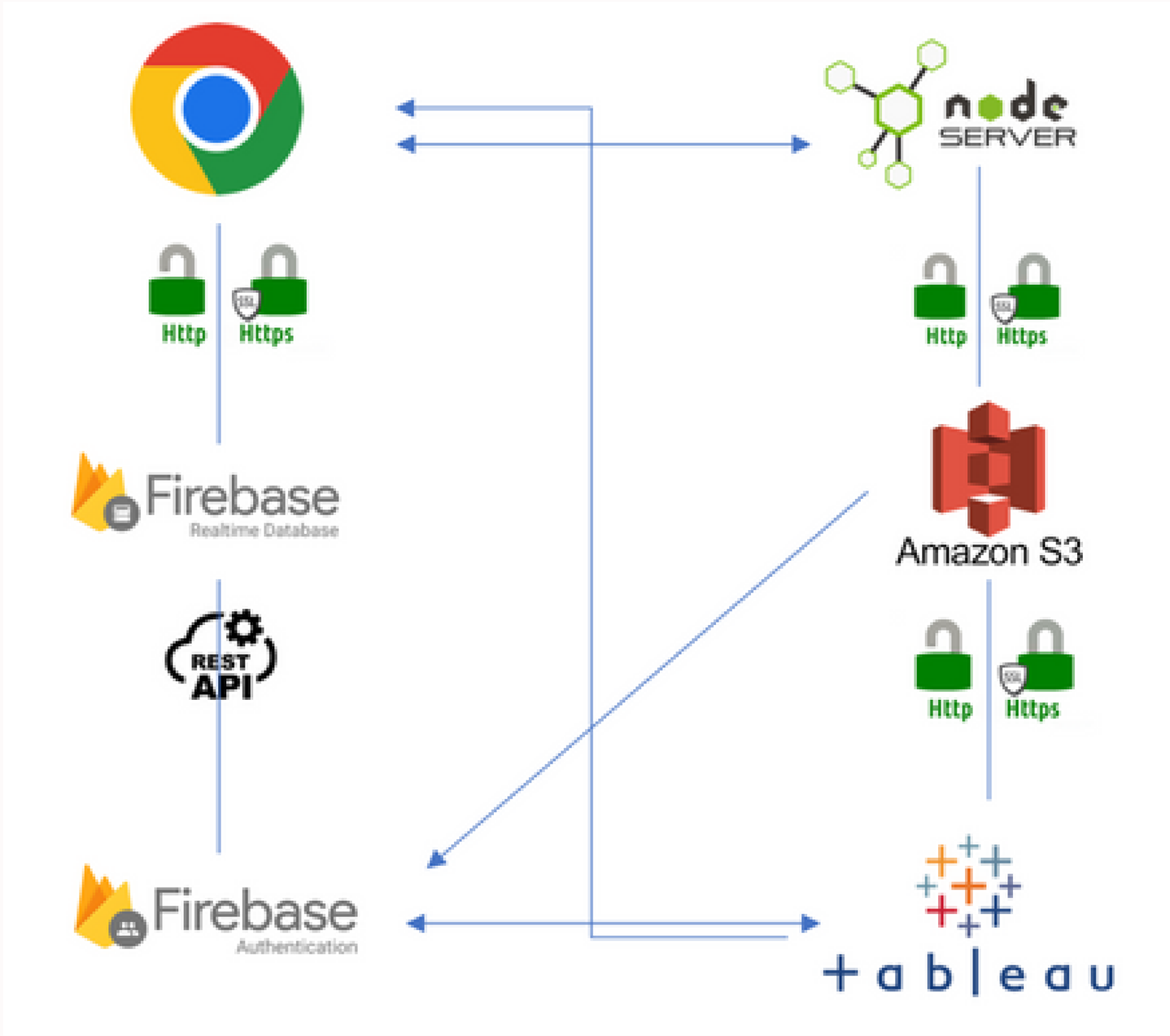
SYSTEM DESIGN



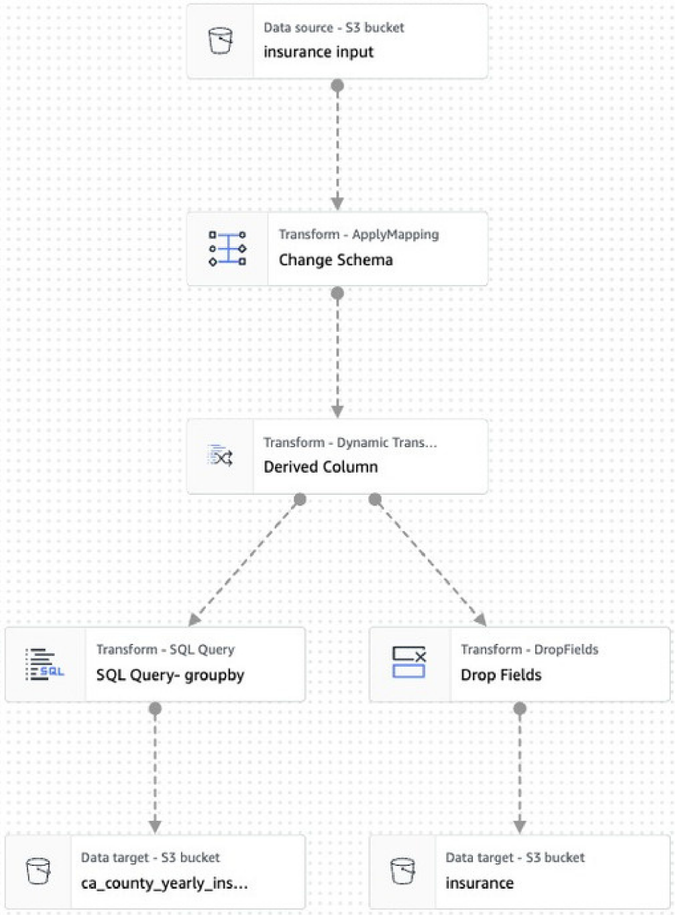

county
fips_code INT
county_name VARCHAR(255)
Indexes
date
date VARCHAR(255)
year INT
Indexes
flood_event
event_id VARCHAR(255)
fips_code INT
date VARCHAR(255)
year INT
location VARCHAR(255)
lat VARCHAR(255)
lng VARCHAR(255)
event_type VARCHAR(255)
flood_cause VARCHAR(255)
flood_impact_days INT
date_date VARCHAR(255)
date_year INT
county_fips_code INT
Indexes
weather_station
stationID VARCHAR(255)
fips_code INT
lat VARCHAR(255)
lng VARCHAR(255)
county_fips_code INT
Indexes
county_daily_flood_weather
fw_id BIGINT
fips_code INT
date VARCHAR(255)
year INT
flood_count INT
PRCP_mean INT
TEMP_mean INT
SNOW_mean INT
SNWD_mean INT
date_date VARCHAR(255)
date_year INT
county_fips_code INT
Indexes
ca_county_yearly_insurance_stats
fips_code INT
year INT
log_insured FLOAT
log_cost FLOAT
policy_count INT
county_fips_code INT
Indexes
weather
stationID VARCHAR(255)
date VARCHAR(255)
year INT
PRCP VARCHAR(255)
SNOW VARCHAR(255)
SNWD VARCHAR(255)
TMAX VARCHAR(255)
TMIN VARCHAR(255)
date_date VARCHAR(255)
date_year INT
weather_station_stationID VARCHAR(255)
weather_station_county_fips_code INT
Indexes

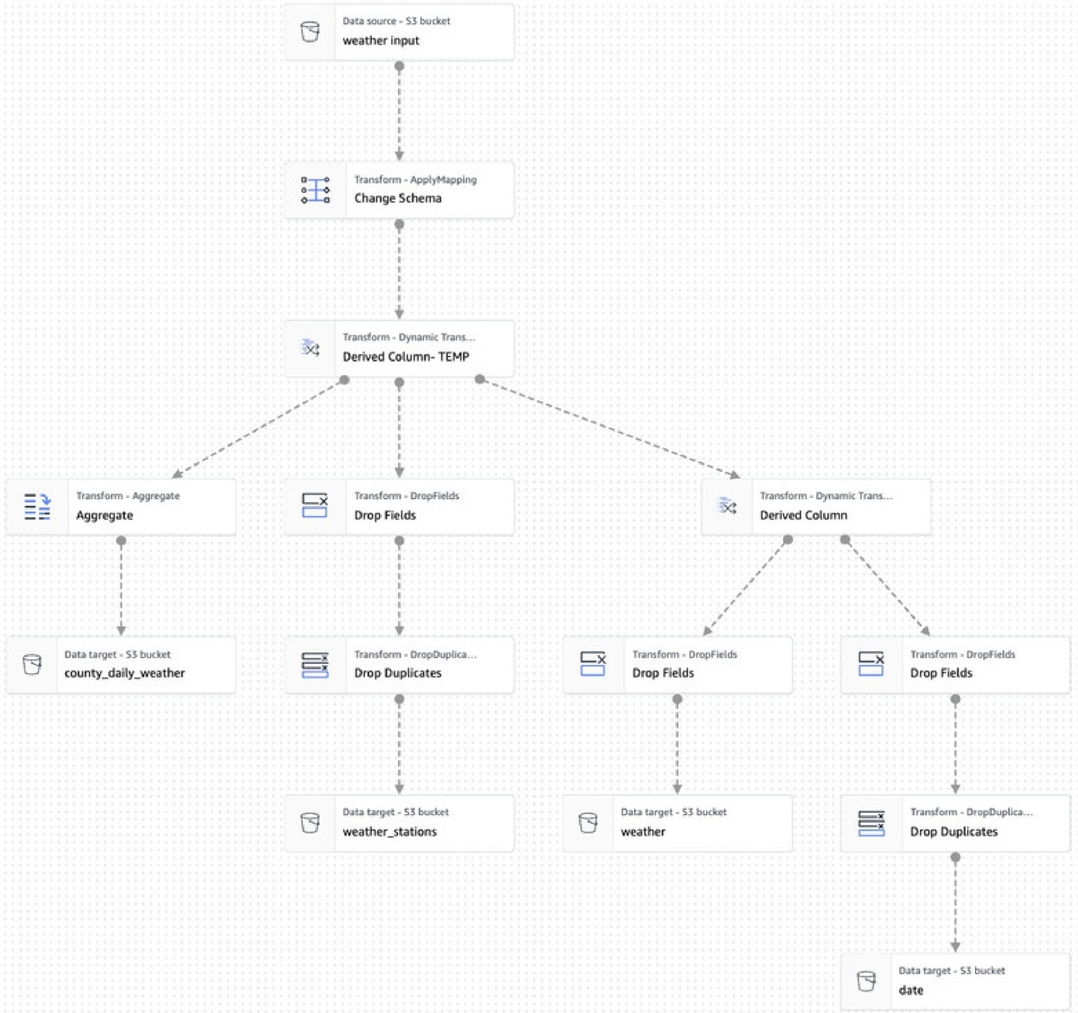
DATA MODEL

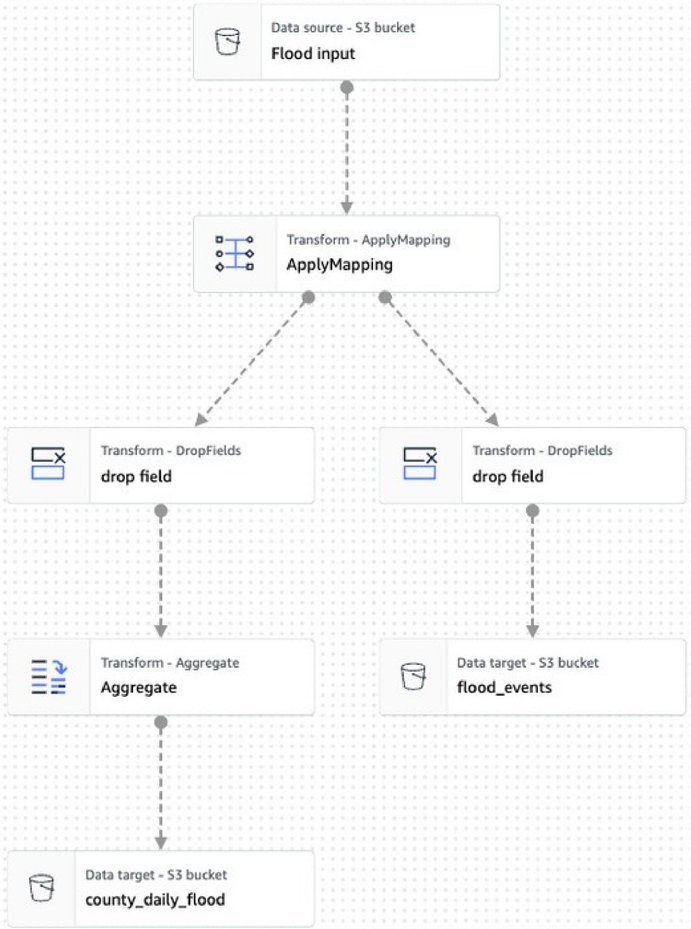
System Interface and Connectivity Design



Glue Job Architecture







Redshift Queries

```sql
SELECT c.county_name, d.year, SUM(f.flood_impact_days) AS total_flood_events, AVG(w.PRCP) AS avg_precipitation
FROM COUNTY c
JOIN flood_events f ON c.fips_code = f.fips_code
JOIN DATE d ON f.year = d.year AND f.date = d.date
JOIN WEATHER_STATIONS ws ON c.fips_code = ws.fips_code
JOIN WEATHER w ON ws.stationID = w.stationID AND d.year = w.year AND f.date = w.date
GROUP BY c.county_name, d.year
ORDER BY total_flood_events DESC;
```

Result 1 (100)

county_name	year	total_flood_events	avg_precipitation
Mariposa County	2017	663	279.5833333333333
San Bernardino County	2010	573	162.56658878504672
Fresno County	2019	436	97.04741379310344
Fresno County	2017	394	159.90828924162258
Tulare County	2019	385	171.46280991735537
Plumas County	2017	378	518.02
Inyo County	2017	348	10.24390243902439
Sacramento County	2023	338	251.70503597122303
Tulare County	2017	315	183.43975903614458
Tehama County	2023	288	230.9696969696997
San Diego County	2017	286	175.63177339901478
San Diego County	2021	281	102.12521008403361

```sql
SELECT c.county_name,
       i.year,
       COUNT(DISTINCT i.zip_code) AS num_insured_properties,
       SUM(i.policy_cost) AS total_policy_cost,
       AVG(w.flood_count) AS avg_flood_count,
       AVG(w.prcp_mean) AS avg_prcp_mean,
       AVG(w.temp_mean) AS avg_temp_mean,
       AVG(w.snow_mean) AS avg_snow_mean,
       AVG(w.snwd_mean) AS avg_snwd_mean
FROM insurance i
JOIN county c ON i.fips_code = c.fips_code
JOIN county_daily_flood_weather w ON i.fips_code = w.fips_code AND i.year = w.year
```

Result 1 (100) Export Chart

county_name	year	num_insured_properties	total_policy_cost	avg_flood_count	avg_prcp_mean	avg_tem
Alameda County	2009	64	2541770940	0	12	149
Alpine County	2009	11	337630840	0	25	60
Amador County	2009	20	113220810	0	21	152
Butte County	2009	51	3342467425	0	30	153
Calaveras County	2009	23	51995710	0	21	144
Colusa County	2009	41	2381777935	0	5	160
Contra Costa County	2009	72	4058579175	0	15	149
Del Norte County	2009	18	869239470	0	39	109

Elapsed time: 22542 ms Total rows: 100

Redshift Queries

```sql
SELECT c.county_name, i.year, SUM(i.log_cost) AS total_policy_cost_per_year, SUM(f.flood_impact_days) AS total_flood_events_overall
FROM COUNTY c
JOIN flood_events f ON c.fips_code = f.fips_code
JOIN CA_COUNTY_YEARLY_INSURANCE_STATS i ON c.fips_code = i.fips_code AND f.year = i.year
GROUP BY c.county_name, i.year
ORDER BY total_flood_events_overall DESC;
```

Result 1 (100)

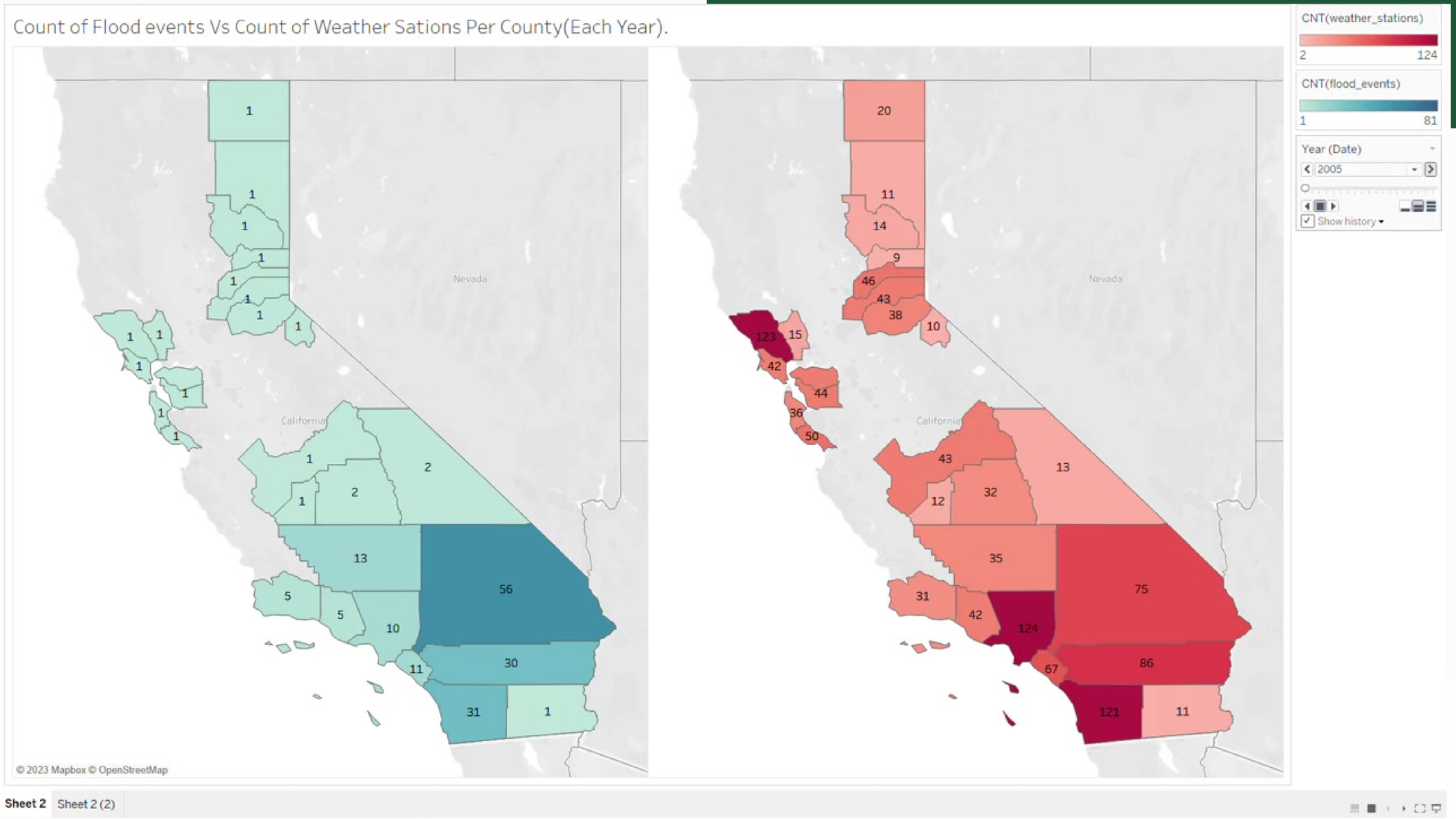
county_name	year	total_policy_cost_per_year	total_flood_events_overall
Inyo County	2017	6379.21200000006	2842
Plumas County	2017	3073.2555000000018	1519
Fresno County	2019	11196.617599999909	980
Mono County	2011	2286.4184000000037	882
Fresno County	2017	21757.396499999828	784
San Bernardino County	2010	24493.257599999557	735
Butte County	2019	17307.50560000021	588
Sacramento County	2017	14427.143499999775	539
Kern County	2020	8300.007100000052	490
Merced County	2019	3748.744999999989	490
Lassen County	2017	2376.1668000000013	490
Kings County	2017	2020.3826999999922	490

```sql
--county with the highest average annual flood impact days
SELECT c.county_name, AVG(fe.flood_impact_days) AS avg_flood_impact_days
FROM COUNTY c
JOIN FLOOD_EVENTS fe ON c.fips_code = fe.fips_code
GROUP BY c.county_name
ORDER BY avg_flood_impact_days DESC
LIMIT 1;
```

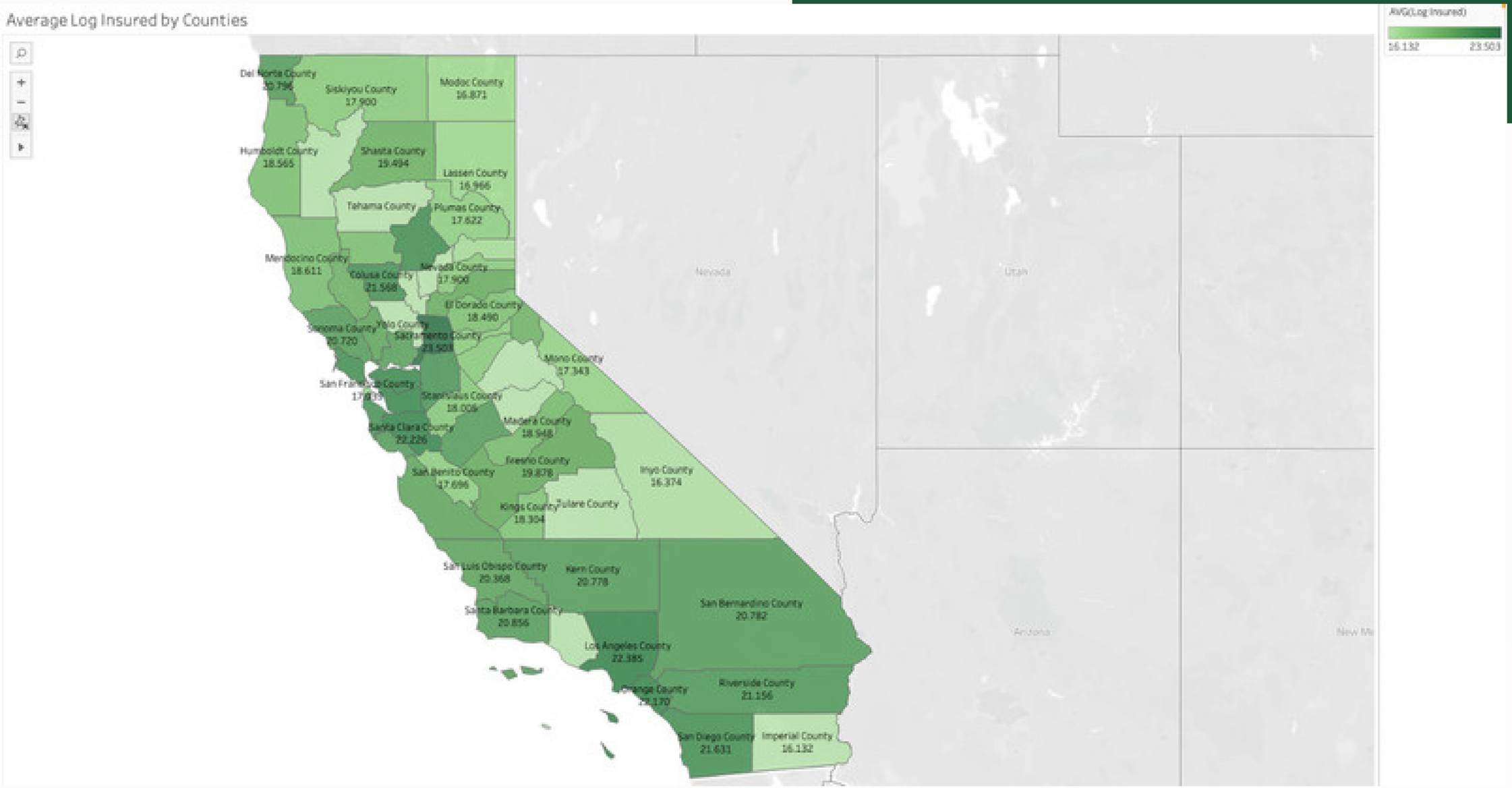
Result 1 (1)

county_name	avg_flood_impact_days
Plumas County	3

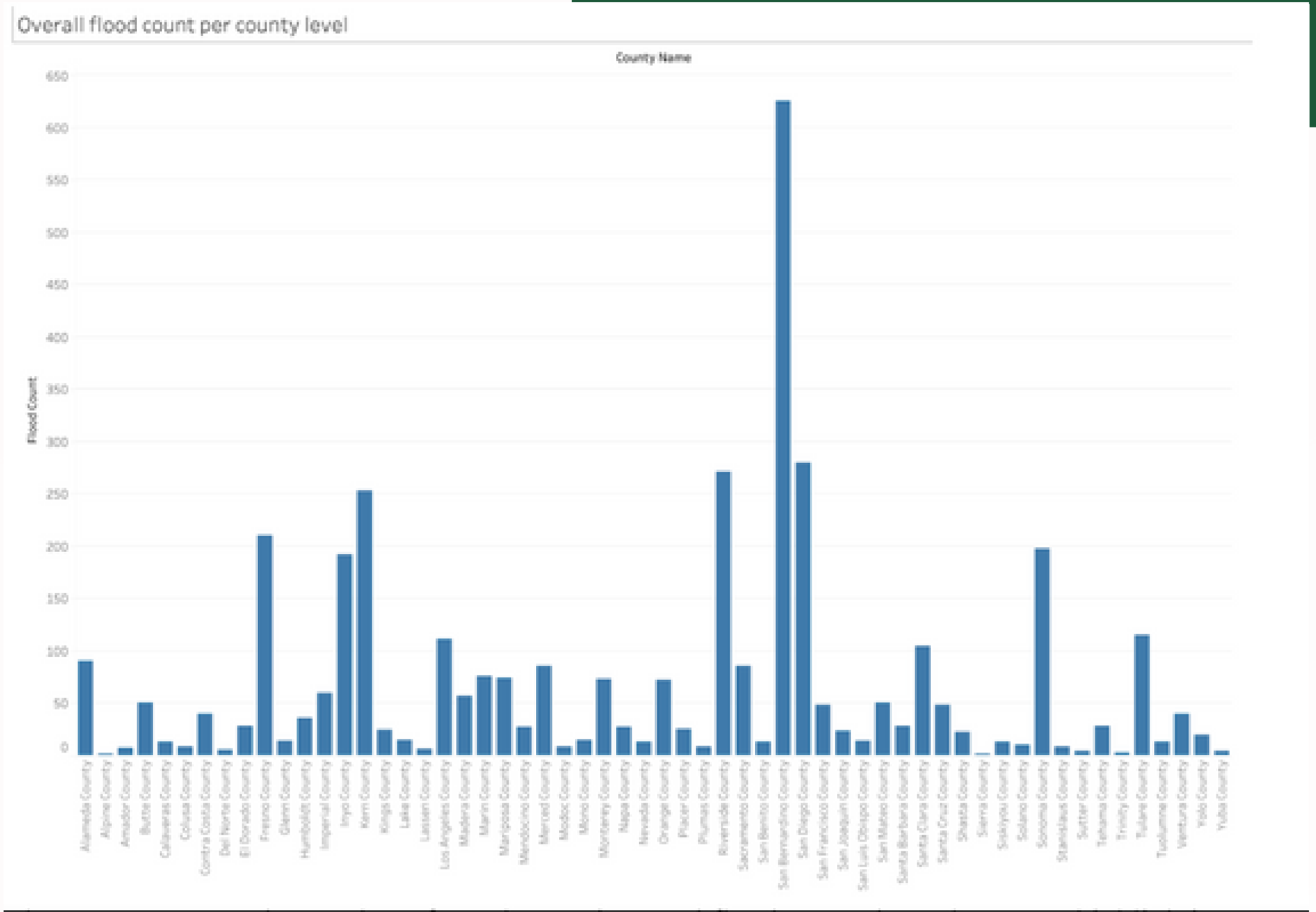
Visualizations



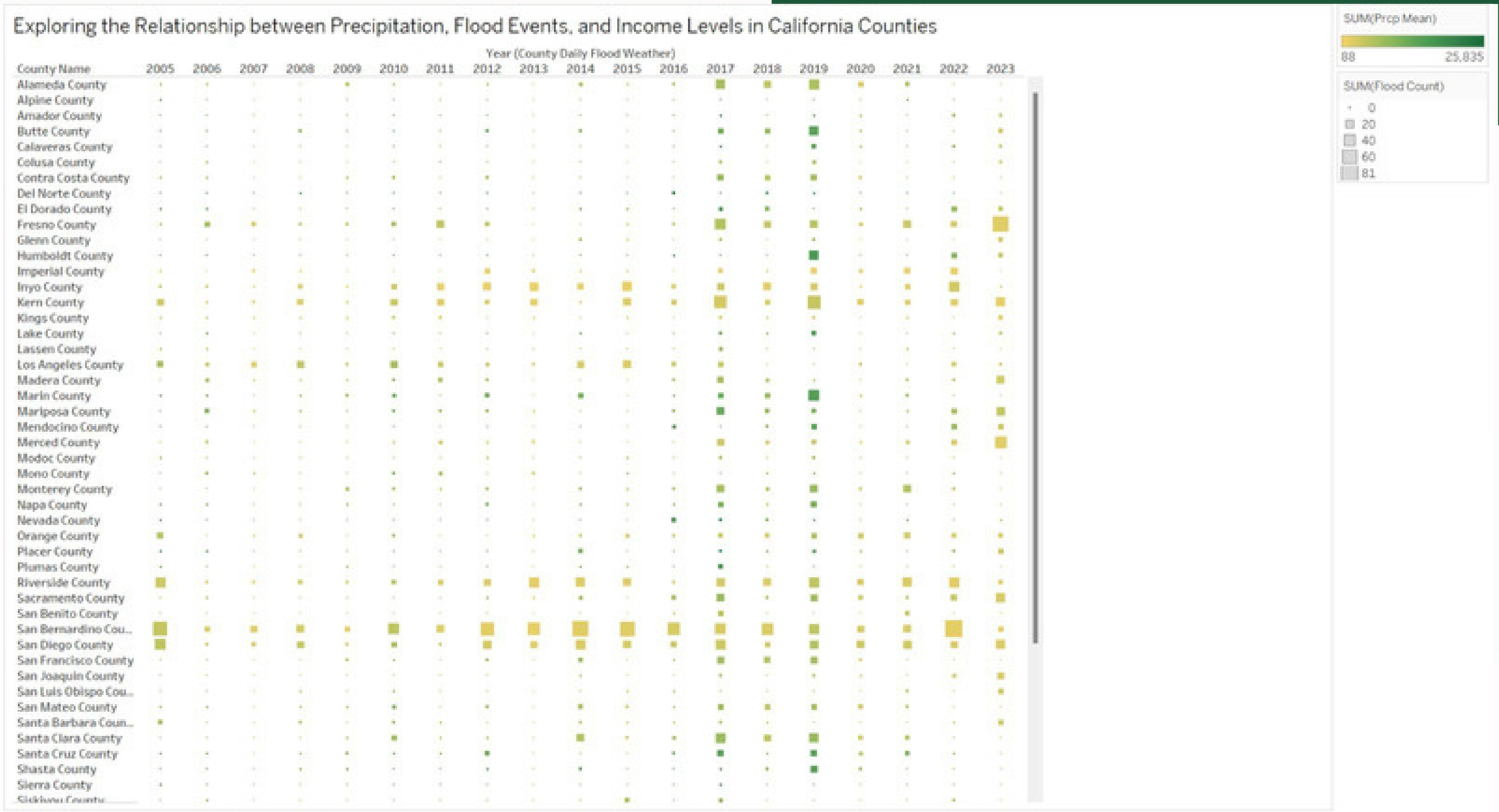
Visualizations



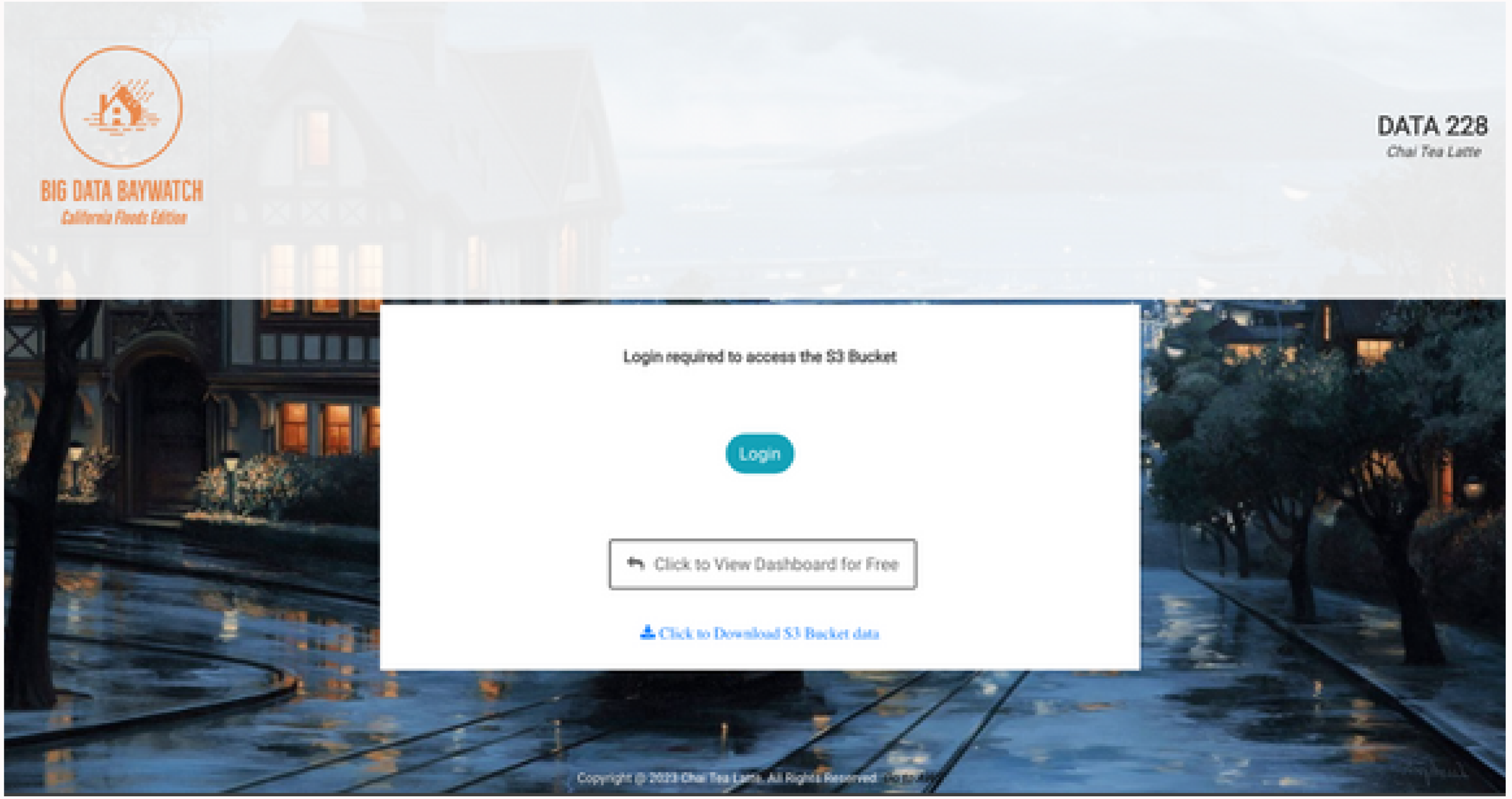
Visualizations



Visualizations



Website Demo



Conclusion & Future Work

This project developed a big data analytics platform that addresses flooding events and declining flood insurance policies in California. By utilizing climate data, flood events data, and insurance policies data, I created a data-driven solution that informs flood-emergency preparedness, promotes flood insurance awareness, and guides policy-making decisions. The project has both technical and societal impacts, offering a scalable solution for natural disaster risk management.

Expanding the project's scope to cover the entire USA and incorporating river and stormwater information can broaden the project's domain. This would be beneficial to more stakeholders and provide more comprehensive insights.

THANK YOU