DAV6100 FINAL PROJECT

NYC311 Complaints Analysis

Group 2

Agenda

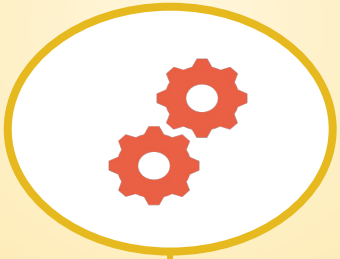
- Data Profile
- Conceptual Architecture
- Demo
- Project Milestones & Timeline
- Team Responsibilities
- Challenges
- Lessons Learned
- Next Steps

Project Overview

311 Requests
NYC Covid19
NYC Median Income

Many data sources

Combines data sources into DW that is easy to maintain



Data Warehouse

Tableau dashboard
Python

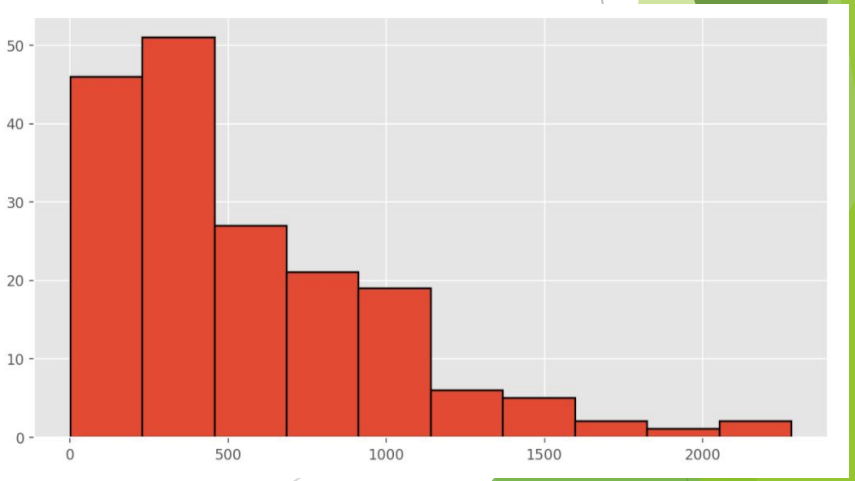
Data Visualization

Data Profile (1): NYC311 Dataset

Dataset Summary

Source of Information	https://portal.311.nyc.gov
Number of Records	1.3M (Original 25M)
Frequency of updates	Daily
Data type and structure	Text, Integers, Float, Date, Time API, CSV
Number of columns	16
Granularity	Each 311 complaint

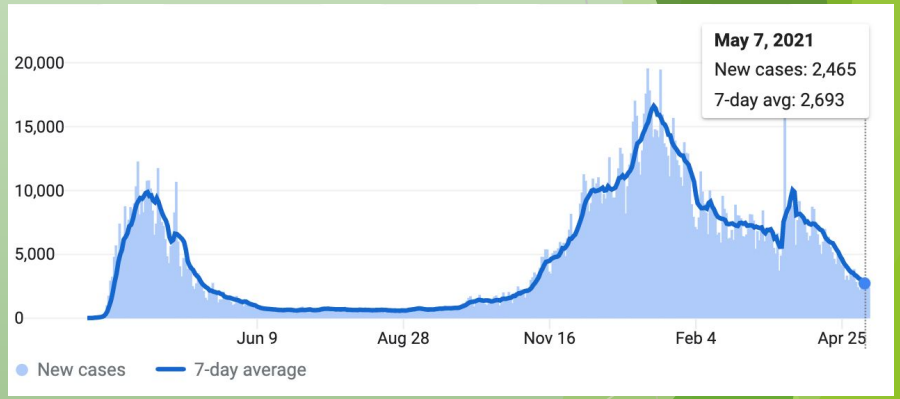
Frequency of Complaints



Data Profile (2): NYC COVID-19 2019

Dataset Summary

Source of Information	https://github.com/nychealth/coronavirus-data
Number of Records	6764
Frequency of updates	Daily
Data type and structure	Text, Float API, CSV
Number of columns	3
Granularity	Caserate grouping by ZIP Code

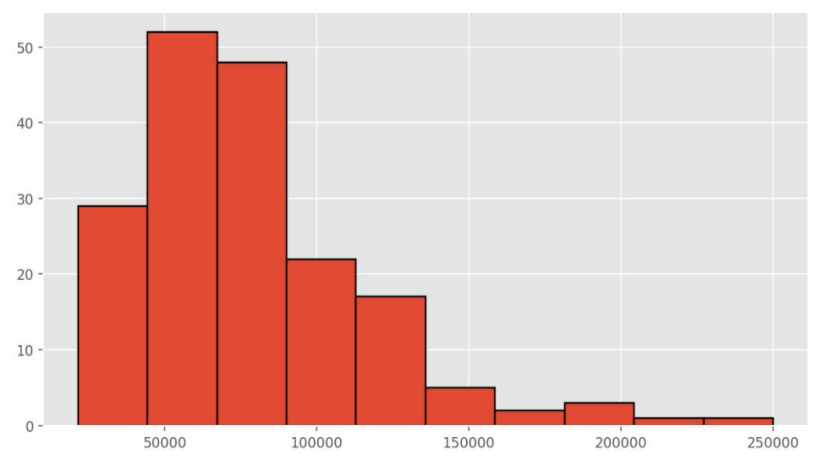


Data Profile (3): NYC Median Income

Dataset Summary

Source of Information	https://data.cccnewyork.org
Number of Records	181
Frequency of updates	Yearly
Data type and structure	Integers CSV
Number of columns	3
Granularity	Income by ZIP code

Average Income Per Household



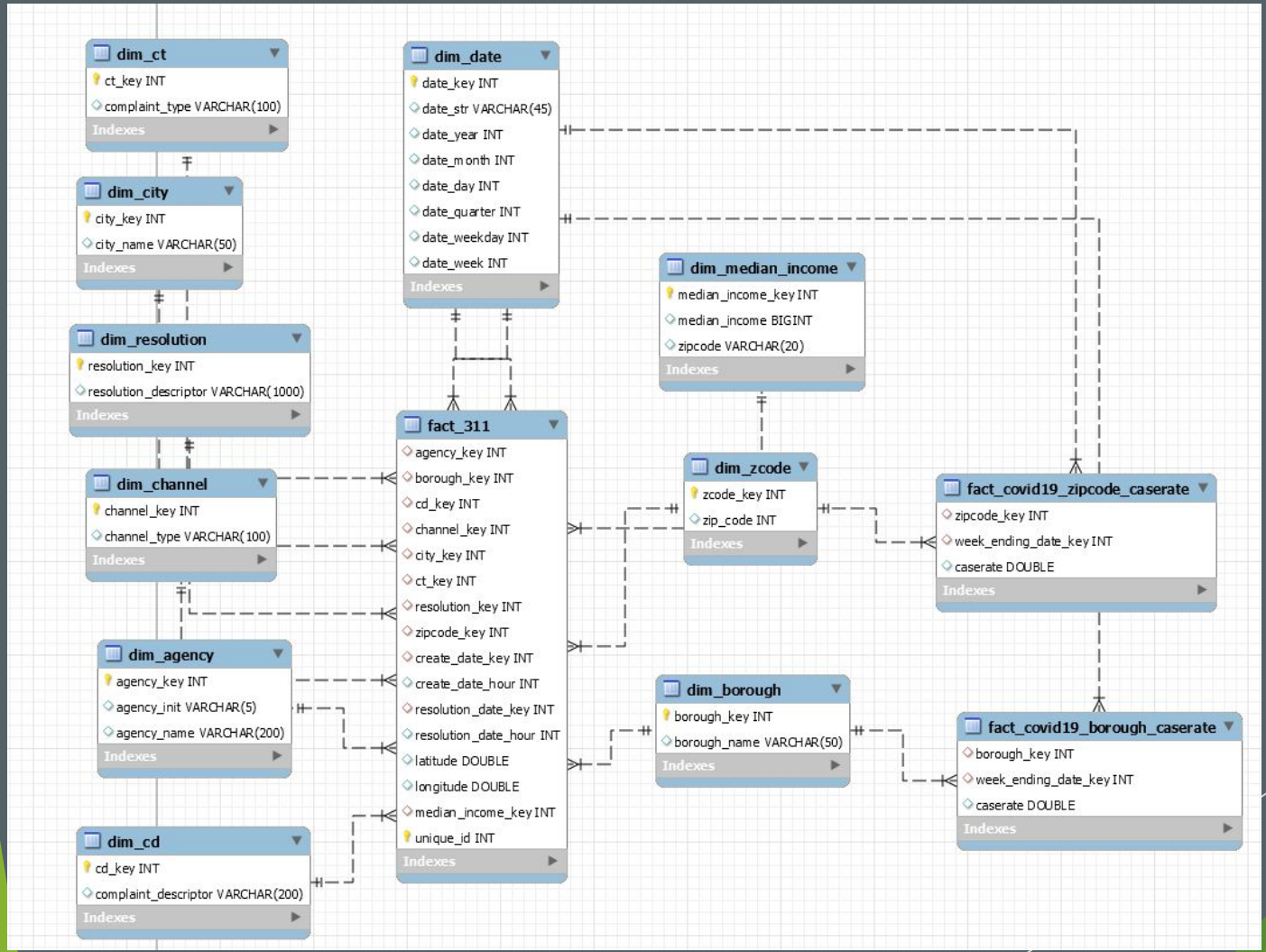
Schema Selection

The first thing to notice about the dimensional schema is its simplicity and symmetry. The simplicity of a dimensional model also has performance benefits. Dimensional models are gracefully extensible to accommodate

--- Kimball The Data Warehouse Toolkit by Ralph Kimball, Margy Ross

1. **Millions of Records**
2. **Aggregation**
3. **Not Only for Us**

ER Diagram


dim_ct
ct_key INT
complaint_type VARCHAR(100)
Indexes
dim_city
city_key INT
city_name VARCHAR(50)
Indexes
dim_resolution
resolution_key INT
resolution_descriptor VARCHAR(1000)
Indexes
dim_channel
channel_key INT
channel_type VARCHAR(100)
Indexes
dim_agency
agency_key INT
agency_init VARCHAR(5)
agency_name VARCHAR(200)
Indexes
dim_cd
cd_key INT
complaint_descriptor VARCHAR(200)
Indexes
dim_date
date_key INT
date_str VARCHAR(45)
date_year INT
date_month INT
date_day INT
date_quarter INT
date_weekday INT
date_week INT
Indexes
fact_311
agency_key INT
borough_key INT
cd_key INT
channel_key INT
city_key INT
ct_key INT
resolution_key INT
zipcode_key INT
create_date_key INT
create_date_hour INT
resolution_date_key INT
resolution_date_hour INT
latitude DOUBLE
longitude DOUBLE
median_income_key INT
unique_id INT
Indexes
dim_median_income
median_income_key INT
median_income BIGINT
zipcode VARCHAR(20)
Indexes
dim_zcode
zcode_key INT
zip_code INT
Indexes
dim_borough
borough_key INT
borough_name VARCHAR(50)
Indexes
fact_covid19_zipcode_caserate
zipcode_key INT
week_ending_date_key INT
caserate DOUBLE
Indexes
fact_covid19_borough_caserate
borough_key INT
week_ending_date_key INT
caserate DOUBLE
Indexes

Assumptions

1. The variables we are integrating to our warehouse affected the 311 complaint rate

2. Stakeholders would want daily updates

3. Two Choices:
   1. Combine API to csv & create a new file.
   2. Apply star schema to 311 data

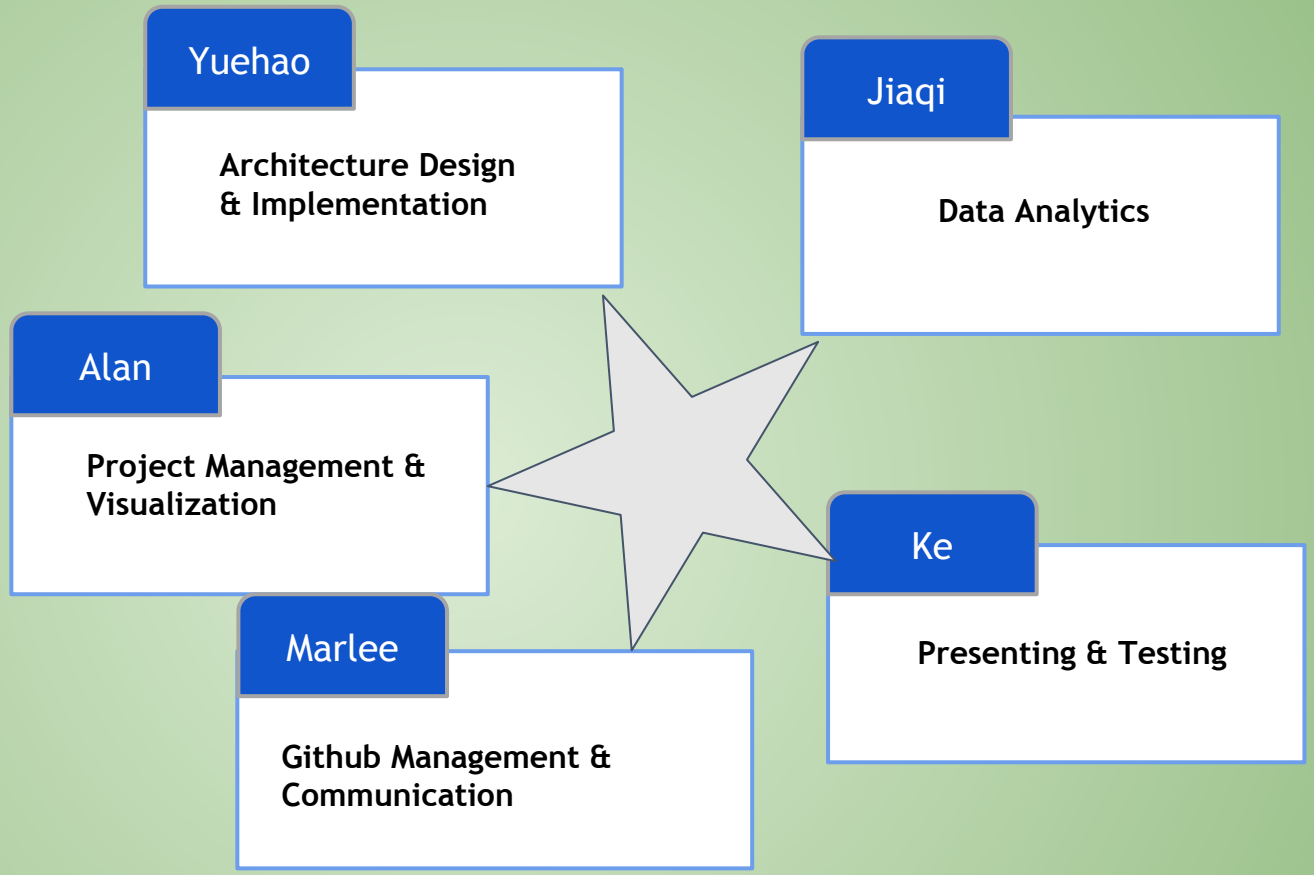
   For practicing and data updating purpose, the second choice was decided.

4. Easy to work asynchronously

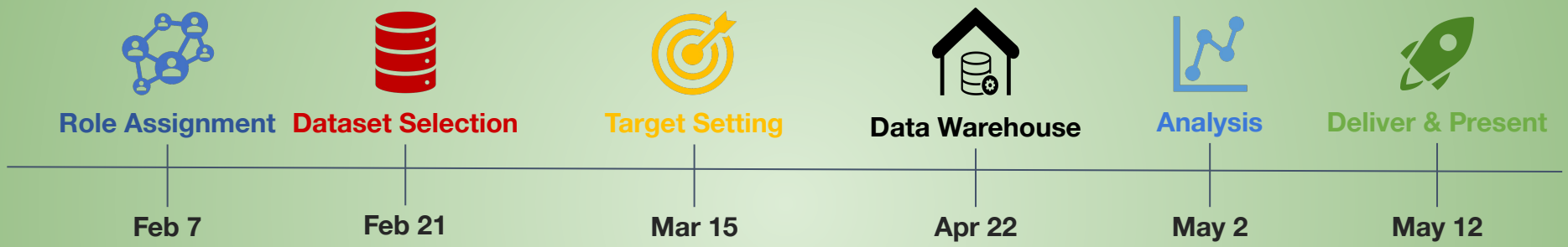
5. No problem with doing class while also learning how to do the project 'on the run'

6. Data amount not overwhelming

Team Leads

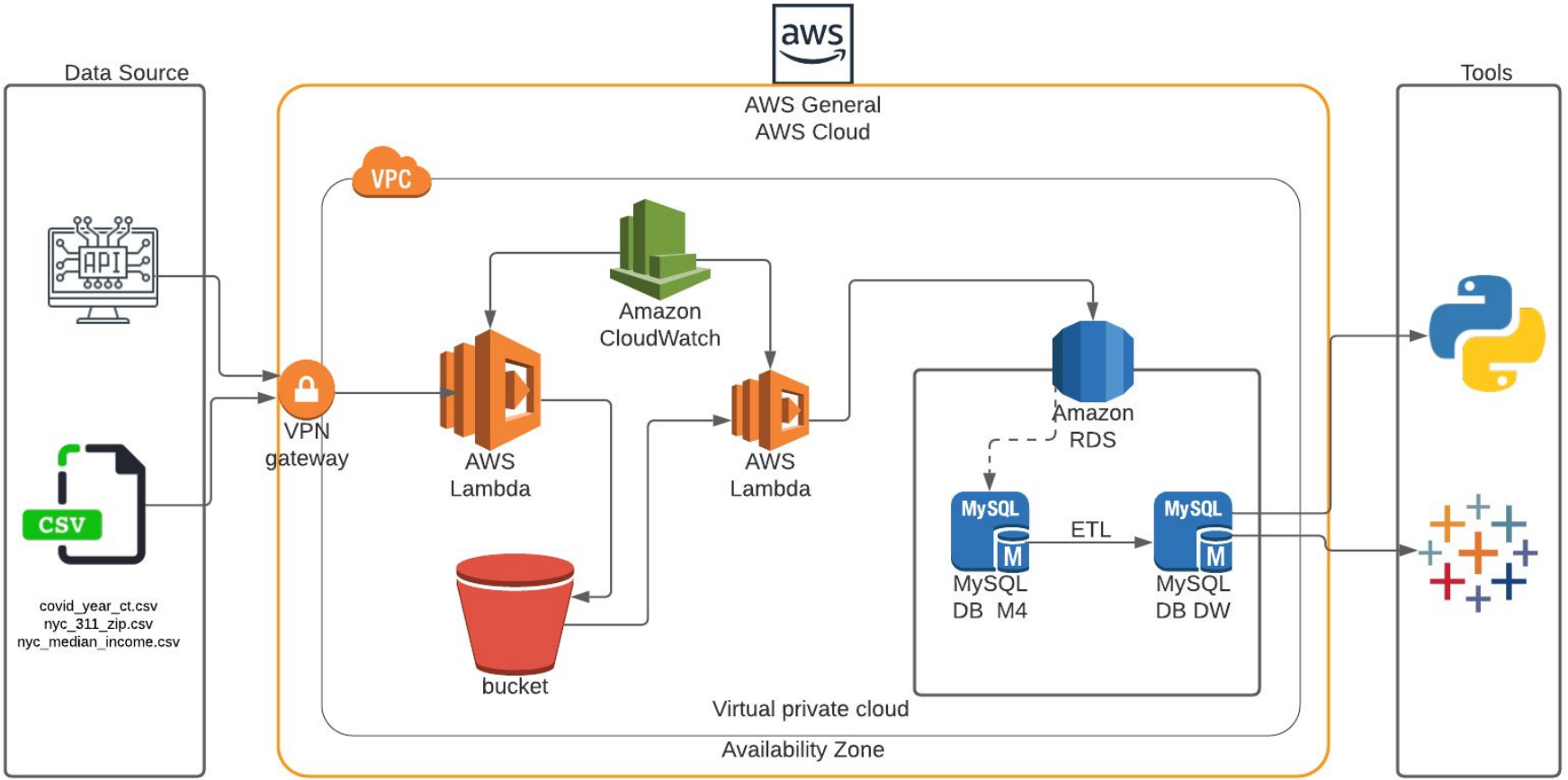

Yuehao
Architecture Design & Implementation
Jiaqi
Data Analytics
Alan
Project Management & Visualization
Ke
Presenting & Testing
Marlee
Github Management & Communication

Milestones



Timeline

Conceptual Architecture



DEMO

Yuehao: Architecture

Alan: Tableau

Marlee: Statistics & hypothesis testing

Challenges

Blank Slate
Asynchronous Team
Big Data

Lessons Learned

ENSURE CLARITY OF PROJECT PURPOSE, ROLE EXPECTATIONS, AND DELIVERABLES AS EARLY AS POSSIBLE

DIAGRAMs BEFORE IMPLEMENTATION IS GOOD, BUT NOT AS GOOD AS STARTING SMALL THEN TESTING

THERE IS A REAL TRADEOFF TO BUILDING AN ARCHITECTURE FOR READING VS WRITING

RESEARCH COSTS AND SET A BUDGET

Next Steps

Incorporate new datasets to measure new variables [race, twitter sentiments etc]

Build a more robust architecture

Leverage Glue and other AWS products to create a more robust architecture in instead of custom scripts. ETL and transferring is done with scripts now, instead of aws tools. All hail Amazon

Expand the dataset to NY state

Thank you!

Any Questions?