

DAV6100: NYC Service Request & Median Income

Group: Xiaolan Li, Bernard Cooper

Professor: Brandon Chiazza



Yeshiva University®

Agenda

- Overview
- Project Requirements
- Data Profile
- Conceptual Architecture
- Demo
- Project Milestones & Timeline
- Team Responsibilities
- Challenges
- Lessons Learned

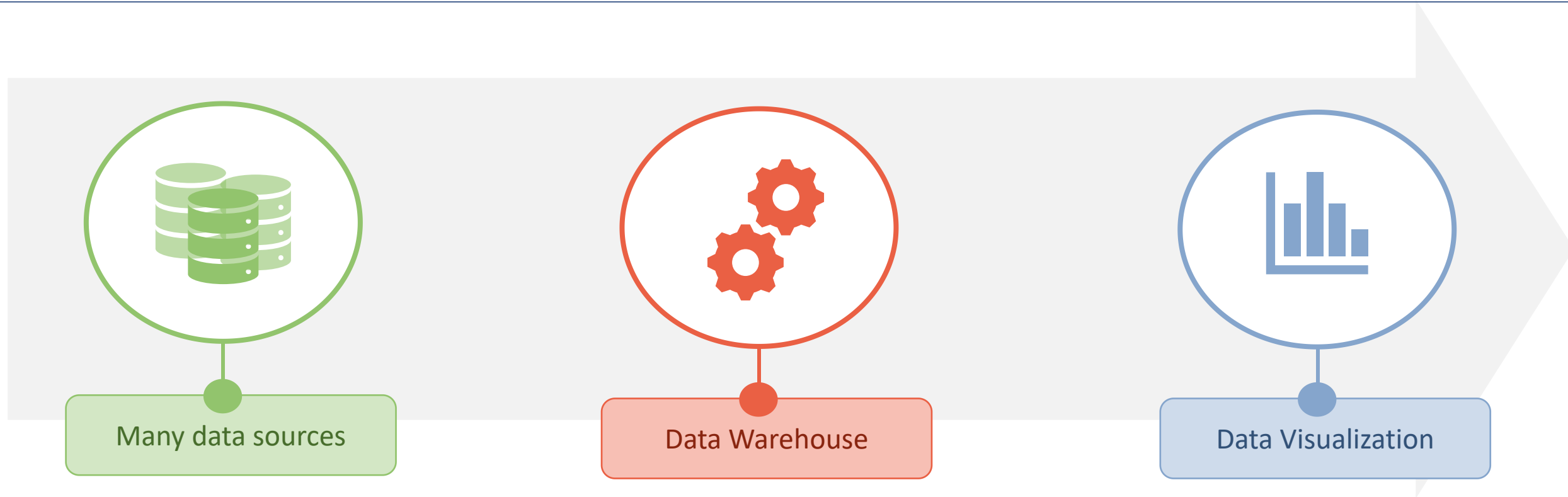


Overview

Using AWS services to store the Data Resources

Using My SQL to store the Data Warehouse

Using Tableau to do the Business Analysis



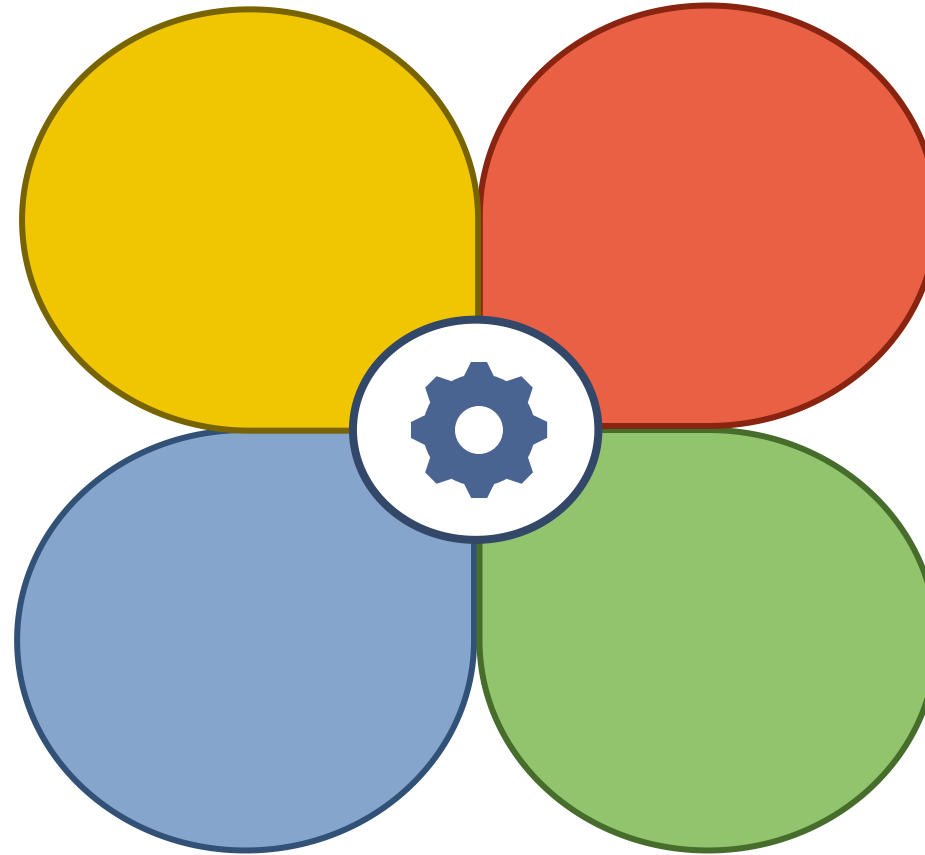
Project Requirements

1. Design, Document, & Plan

- ✓ Develop a conceptual design architectures
- ✓ Develop data flow diagrams and data models
- ✓ Define analytics concepts with bus matrix
- ✓ Define ETL Instructions
- ✓ Define data attributes

2. Develop and Build

- ✓ Develop the warehouse solution using Amazon Web Services as the platform
- ✓ Include two data structures:
 - Structured dataset
 - Semi/Unstructured dataset
- ✓ Integrations:
 - Batch/Migration
 - Real-time
- ✓ Data Visualization
- ✓ Code Repository (GitHub)



3. Test the Solution

- ✓ A prototype is to be test
- ✓ Break-testing and optimization of the database may be necessary (use of indexes)
- ✓ Ensure that error-handling scenarios are considered

4. Present and Deliver

- ✓ Deliver an executive presentation
- ✓ Demo the architectural components
- ✓ Demo the visualizations in a data visualization platform like Tableau

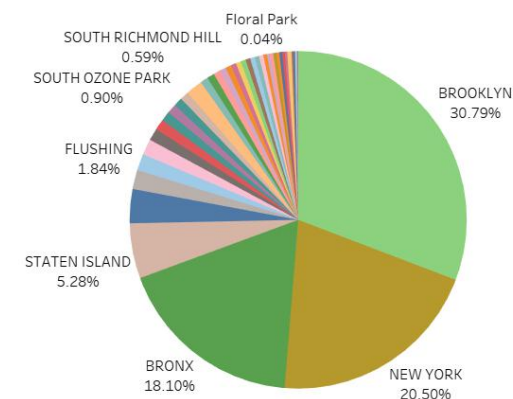
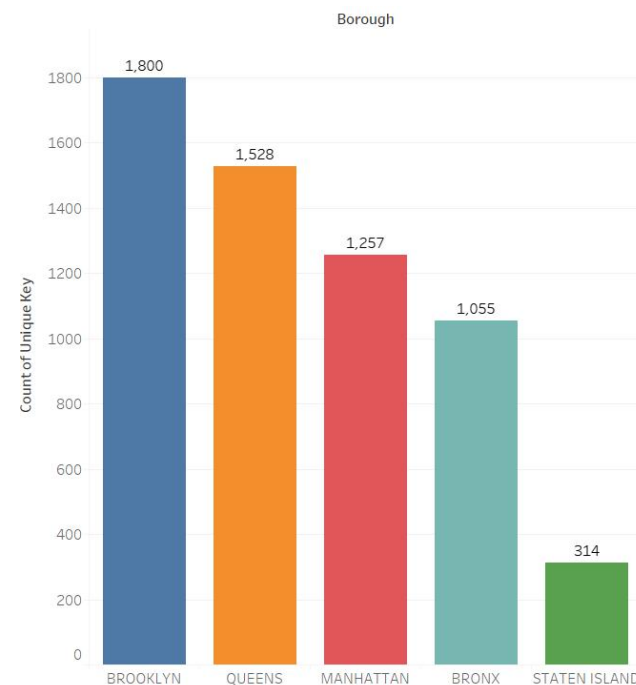
Data Profile 1: <<311 Service Request>>



Dataset Summary

Source of Information	https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
Number of Records	Around 2021 records
Frequency of updates	per day
Data type and structure	Structured Data
Number of columns	41
Granularity	Service request event with details

Borough Incident Count & District Incident Count



Descriptive Statistics

Data Profile 2: <<NYC Median Income>>

Dataset Summary

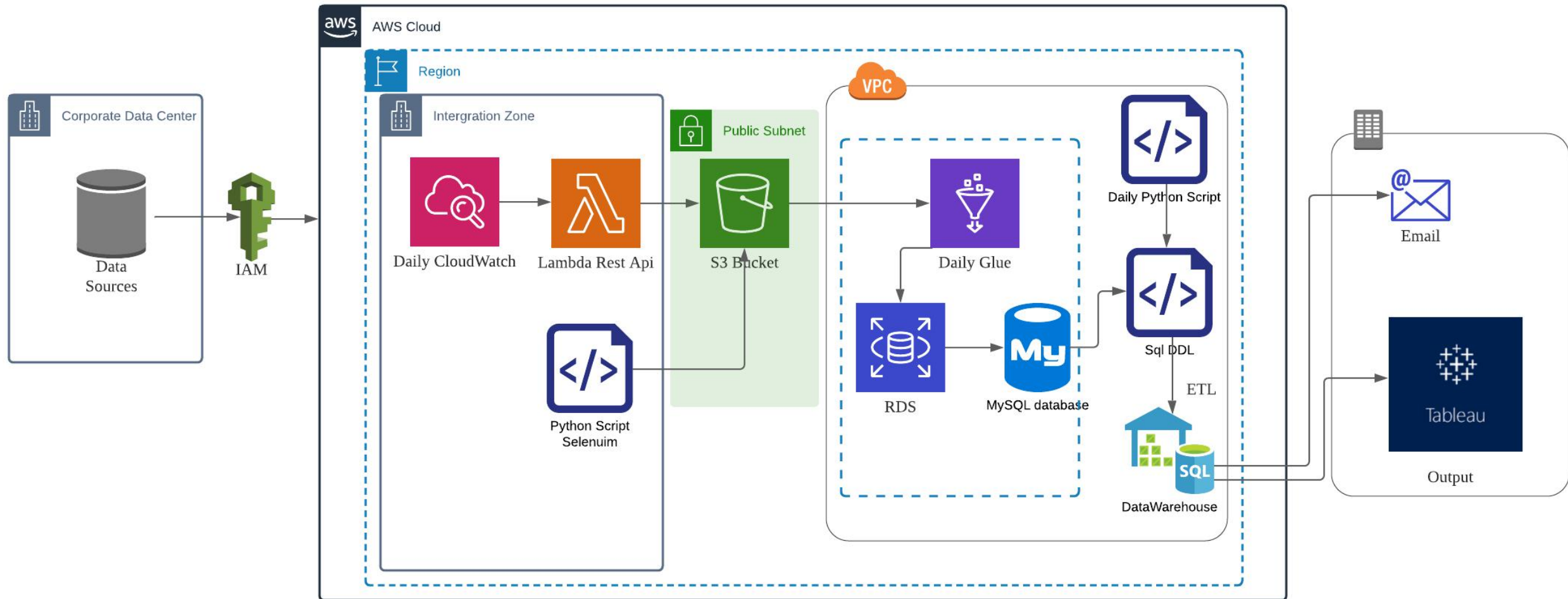
Source of Information	https://data.cccnewyork.org/data/table/66/median-incomes#66/107/62/a/a
Number of Records	62 districts, 5 boroughs, 181 zipcodes
Frequency of updates	per day
Data type and structure	Unstructured Data
Number of columns	5
Granularity	Median income in each location in NYC area

Show tables in different regions

Location (N..	All Households	Location (Nyc Distri..	All Households ..	Location ..	All Households..
Bronx	\$41,432	ASTORIA	\$79,180	10001	\$92,840
Brooklyn	\$66,937	BATTERY PARK/TRIB..	\$162,092	10002	\$36,982
Manhattan	\$93,651	BAY RIDGE	\$76,569	10003	\$118,161
Queens	\$73,696	BAYSIDE	\$92,682	10004	\$190,223
Staten Island	\$89,821	BEDFORD PARK	\$41,336	10005	\$189,702
		BEDFORD STUYVES..	\$61,186	10006	\$179,044
		BENSONHURST	\$57,139	10007	\$224,063
		BOROUGH PARK	\$55,071		
		BROWNSVILLE	\$31,345		
		BUSHWICK	\$66,275		

Descriptive Statistics

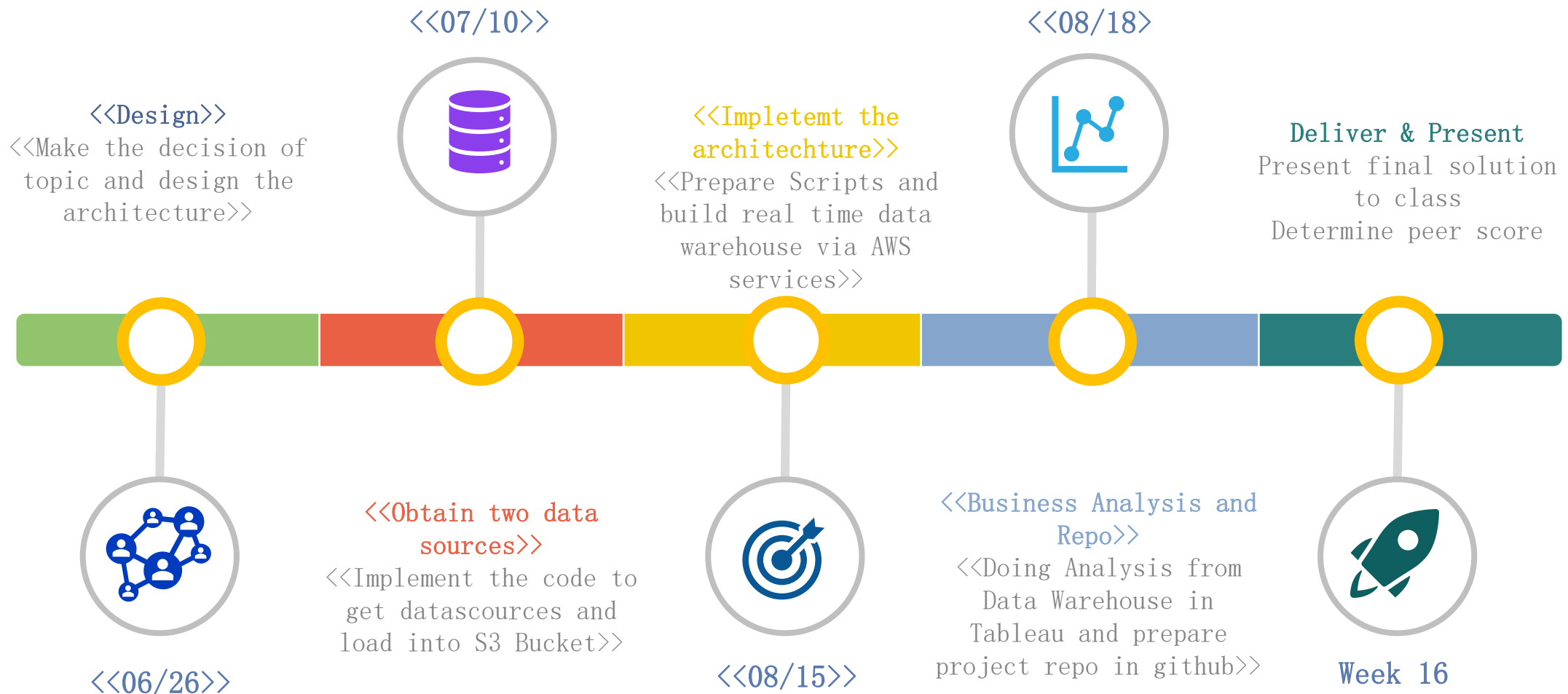
Conceptual Architecture



DEMO



Project Milestones & Timeline



Team Responsibilities

GROUP



Bernard Cooper

Created Research Questions, Obtained Data Sources, Created Data Warehouse, Tableau Data Analysis, Presentation



Xiaolan Li

Obtained Data Sources, Implement AWS Services, ETL Data Sources to Data Warehouse, Built Github Repo, Presentation



None

Assumptions

1

Illegal Parking is the highest frequent incident in service request.

2

Brooklyn, Newyork and Bronx are the highest frequent service request districts

3

The distribution of time to close in all incidents are right skew with a long tail

4

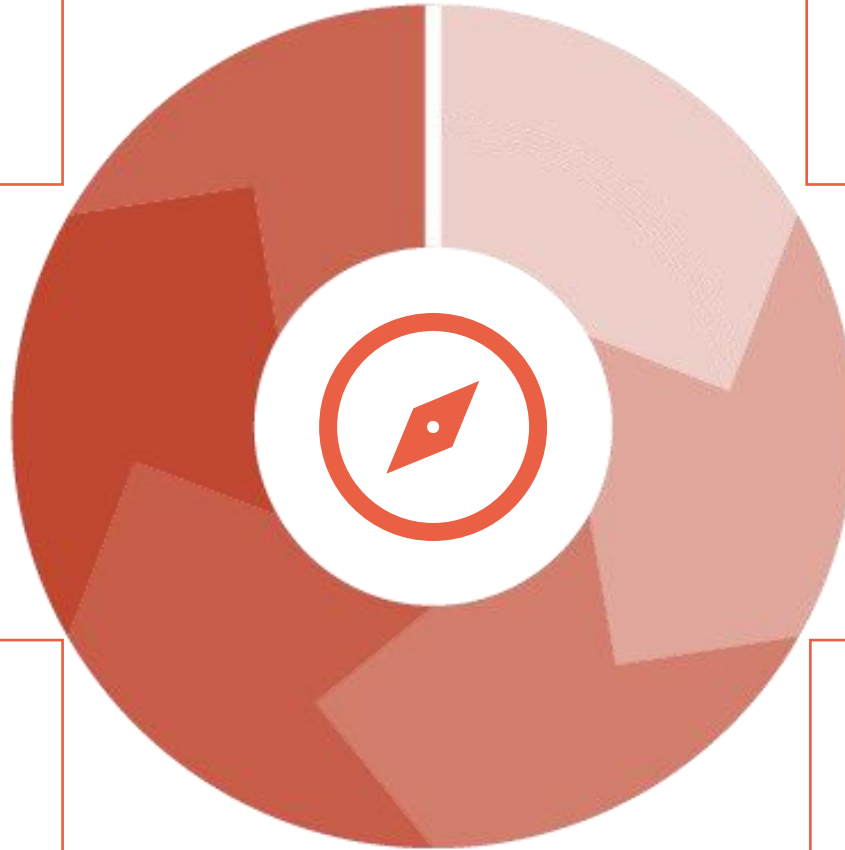
Zipcode 11226 has the highest frequent service request

5

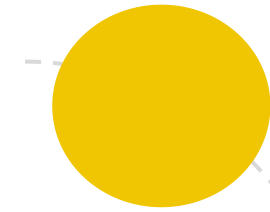
There's no strong correlation between median income and Borough as well as in zipcode but has a negative correlation with districts

6

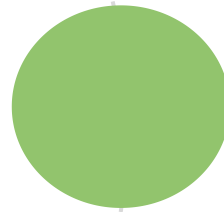
Borough Queens, District Ridgewood, zipcode 11411 have highest average time to close incidents



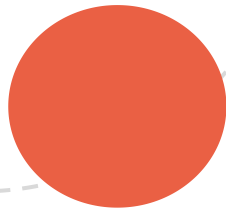
Challenges



Lambda Functions with required packages (add whl files to match linux env)
Security group rule when connect the GLUE with RDS (add `All TCP` rule to Sg)
The JOB in GLUE can not detect the columns from data sources
(drop index, rename columns and drop first row in DDL)



Data Sources can't match a lot district names between services request
and median income info (replace the names of districts in median
income data source)



Lessons Learned



The following are the key lessons learned from the project.

AWS Services:

- S3, RDS, GLUE, VPC, LAMBDA, IAM, CLOUDWATCH

ETL:

- Create Data Warehouse
- Update dimensional tables
- Update fact tables

Notification:

- Run SQL DDL
- Send email to notice

Tableau:

- Analyze the Data Warehouse
- Doing the regression and solve the research questions

