Tides of Taste: Personalized Seafood Restaurant Insights

Boyuan Chen
Gaurangi Agrawal
Mengxin Zhao
Varun Kaza

1

Introduction

Why We Are Interested in This Problem

1. A shared passion for dining out
2. An interest in the seafood culture of the Greater Boston Area
3. An ideal application of experience and expertise
4. Opportunity to work with various types of data and tools
5. Relevant to our immediate environment

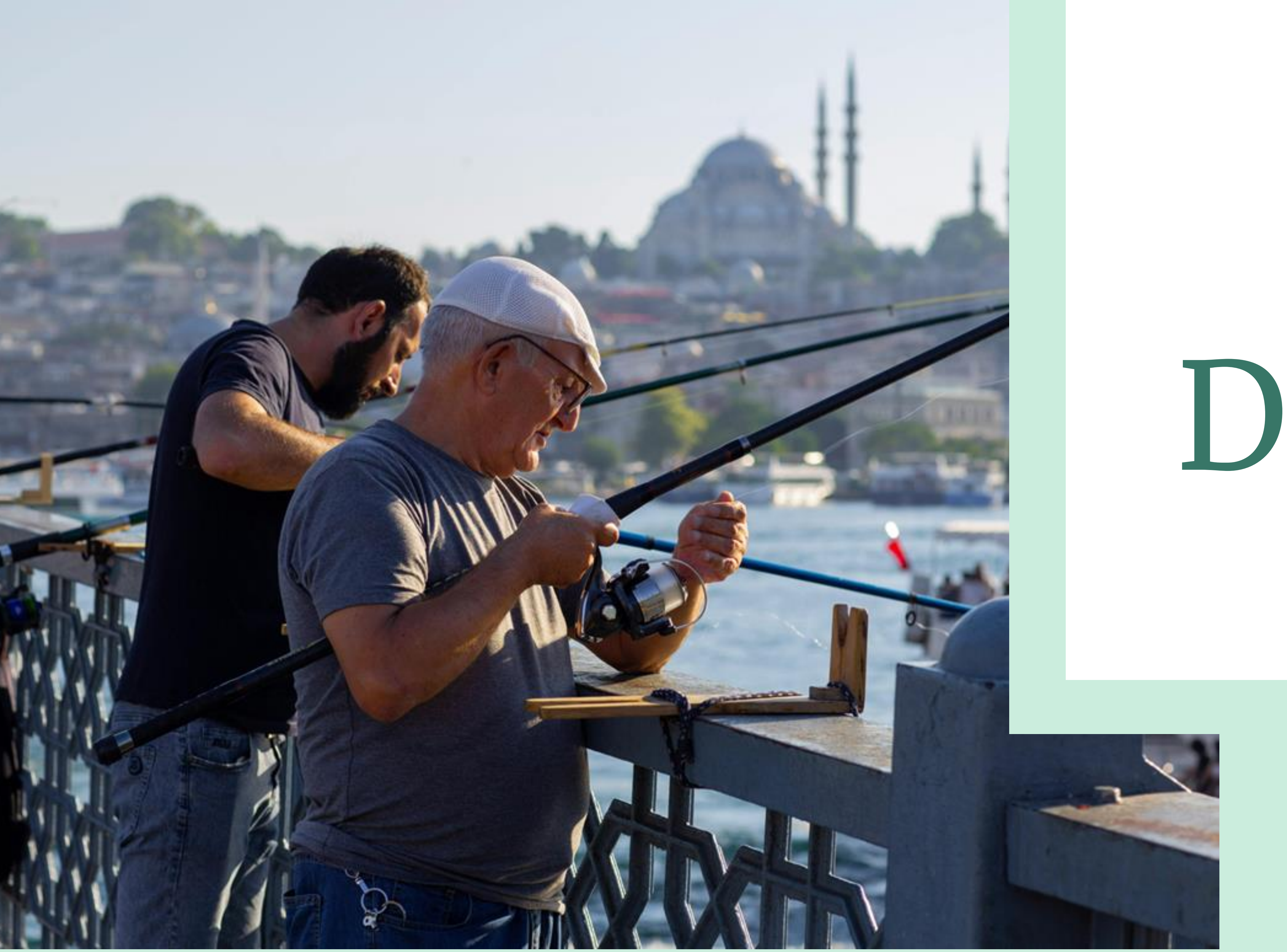
Who It Is Useful To

Usefulness to Customers

1. To find frequently occurring words used in good reviews
2. To create a consolidated guide for tourists to the Greater Boston area
3. To allow customers to plan their trips in consideration of prices and reviews
4. To find the best time of day to visit a restaurant

Usefulness to Restaurants

1. To find a correlation between words used in reviews and the ratings given
2. To know the best time of day to focus their offerings
3. To determine price offerings



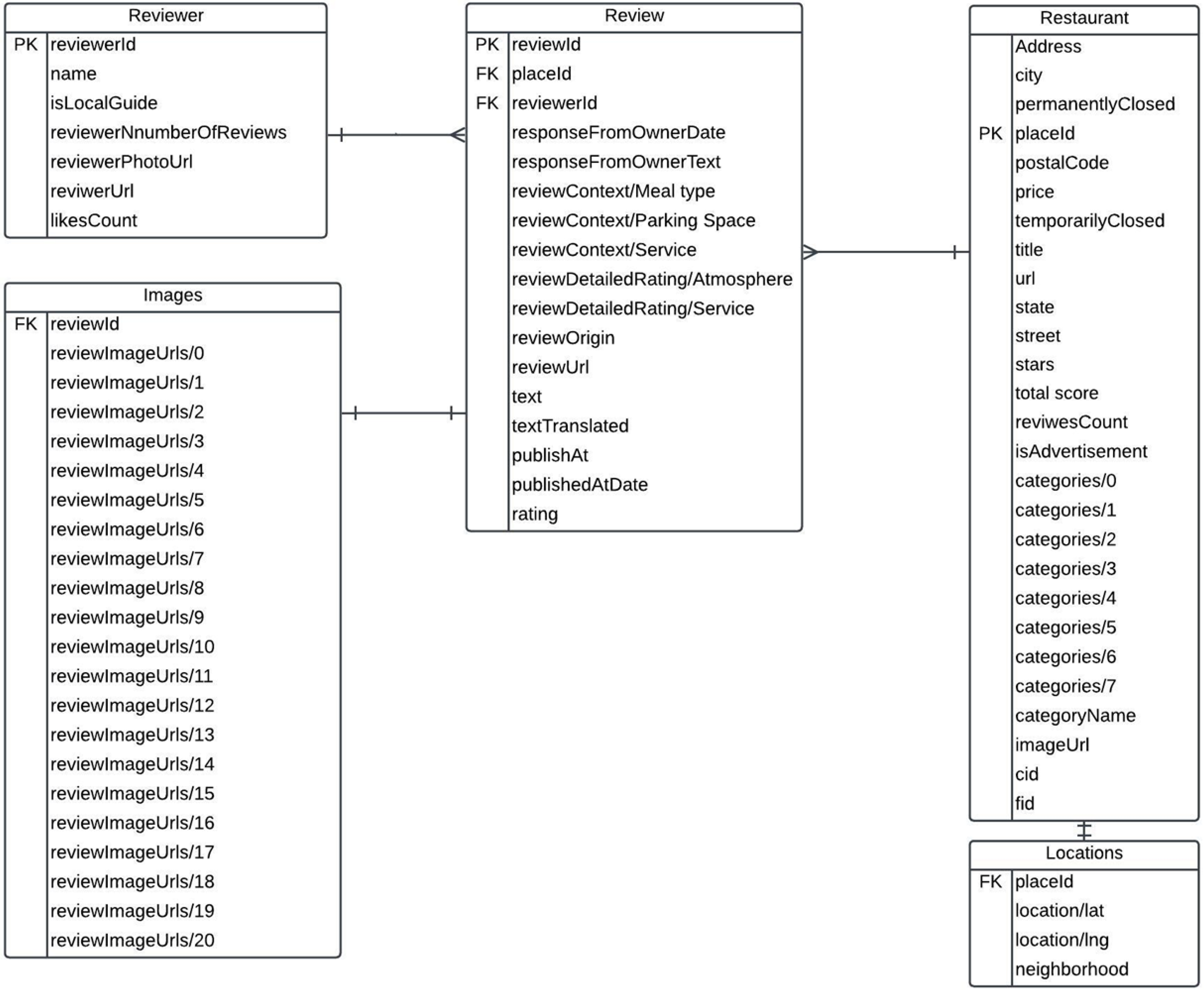
2

Data Structure

The Dataset

Dimensions	77 columns
Source	Google Reviews
Type of Data	Numeric, Boolean,Text, Vector, Events
Method of Extraction	Apify.com

Entity Relationship Diagram



	Reviewer
PK	reviewerId
	name
	isLocalGuide
	reviewerNnumberOfReviews
	reviewerPhotoUrl
	reviwerUrl
	likesCount

	Review
PK	reviewId
FK	placeId
FK	reviewerId
	responseFromOwnerDate
	responseFromOwnerText
	reviewContext/Meal type
	reviewContext/Parking Space
	reviewContext/Service
	reviewDetailedRating/Atmosphere
	reviewDetailedRating/Service
	reviewOrigin
	reviewUrl
	text
	textTranslated
	publishAt
	publishedAtDate
	rating

	Restaurant
	Address
	city
	permanentlyClosed
PK	placeId
	postalCode
	price
	temporarilyClosed
	title
	url
	state
	street
	stars
	total score
	reviwesCount
	isAdvertisement
	categories/0
	categories/1
	categories/2
	categories/3
	categories/4
	categories/5
	categories/6
	categories/7
	categoryName
	imageUrl
	cid
	fid

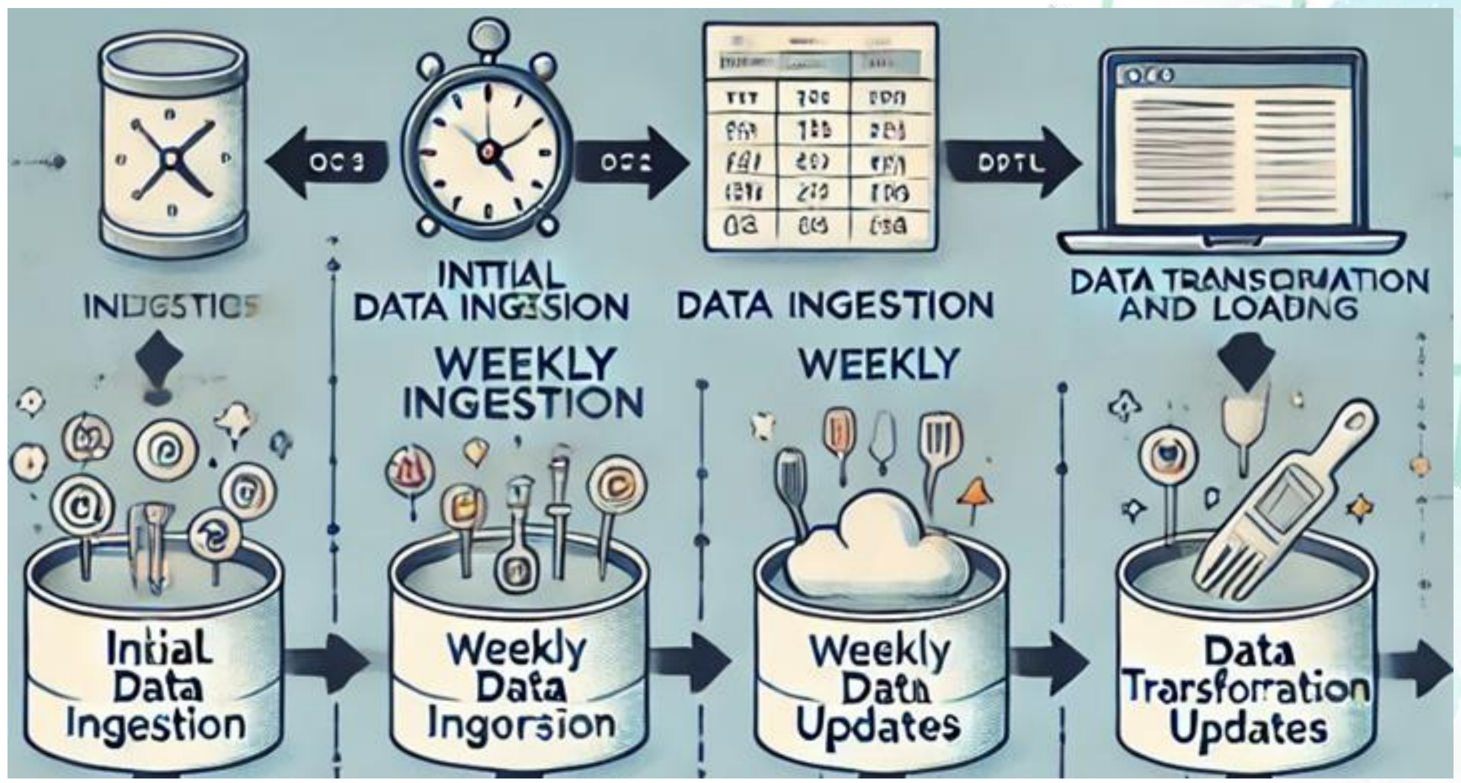
	Images
FK	reviewId
	reviewImageUrls/0
	reviewImageUrls/1
	reviewImageUrls/2
	reviewImageUrls/3
	reviewImageUrls/4
	reviewImageUrls/5
	reviewImageUrls/6
	reviewImageUrls/7
	reviewImageUrls/8
	reviewImageUrls/9
	reviewImageUrls/10
	reviewImageUrls/11
	reviewImageUrls/12
	reviewImageUrls/13
	reviewImageUrls/14
	reviewImageUrls/15
	reviewImageUrls/16
	reviewImageUrls/17
	reviewImageUrls/18
	reviewImageUrls/19
	reviewImageUrls/20

	Locations
FK	placeId
	location/lat
	location/lng
	neighborhood

3

Pipeline Structure

An EtTL pipeline automates the data flow from Google Map to the MotherDuck DB



Tools we used before SuperSet

Apify

Make API calls to get
the data ingested.

Prefect

Pipeline
Orchestration.

Google Cloud Storage Bucket

As our data lake to
store unprocessed
json files

MotherDuck DB

As our data
warehouse to store
processed data.

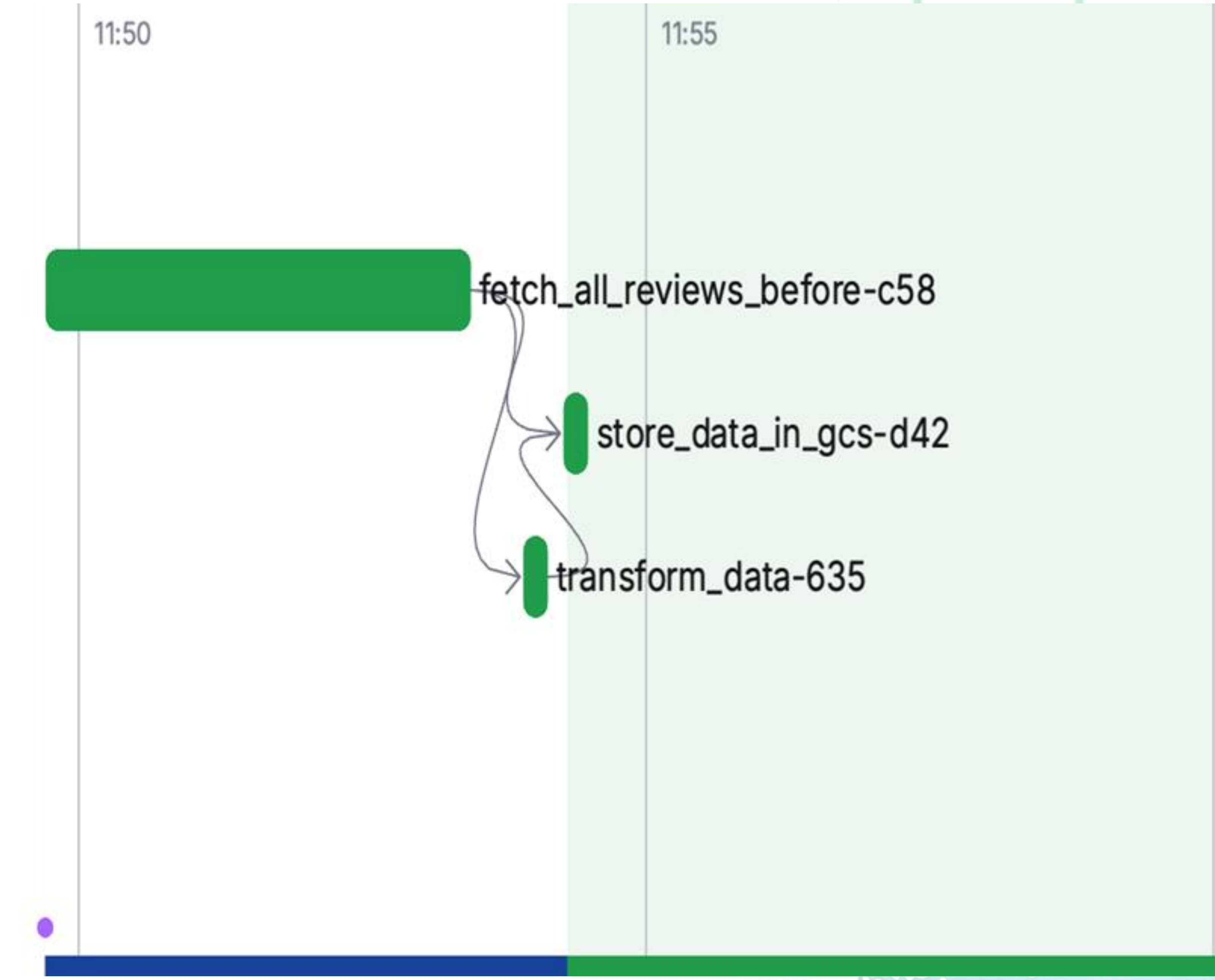
Phase 1: Initial Ingestion “E & t” of the EtTL

We used initial_data_ingestion_flow.py to

1. make the first scrape of review from 2022-01-01 to 2024-10-07
2. Remove columns like review ID, URL info, etc.
3. Convert some categorical data into binary form, like whether the restaurant is temporarily/permanently closed
4. Combine the translated text and original text into one column
5. Count the number of review images for one review

Example:
"reviewImageUrls": [
"https://lh5.googleusercontent.com/p/AF1QipONR0VH5dOSceyOfy
wFtqK9M1Y3lGW4E68xpfpp=w1920-h1080-k-no-p",
"......."
   ],

-> Number of photo in one review = n

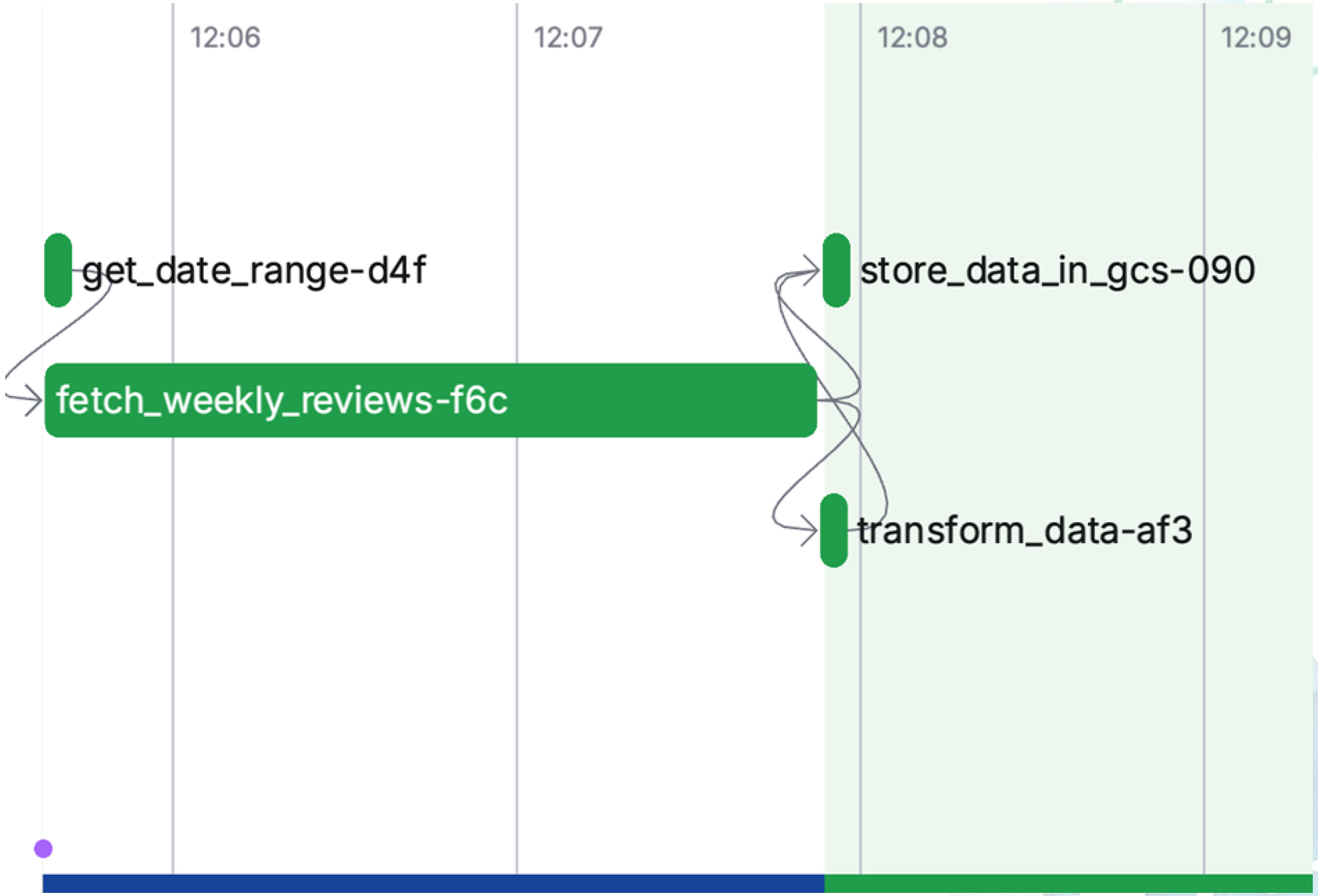


Initial Data Ingestion Flow

Phase 2: Weekly Review Ingestion "E & t" of the EtTL

We used deployment_cycle_scrape.py to

1. make the scrape of new reviews beginning from 2024-10-15 on a weekly basis
2. perform similar transformations towards the data and store it in the GCS bucket



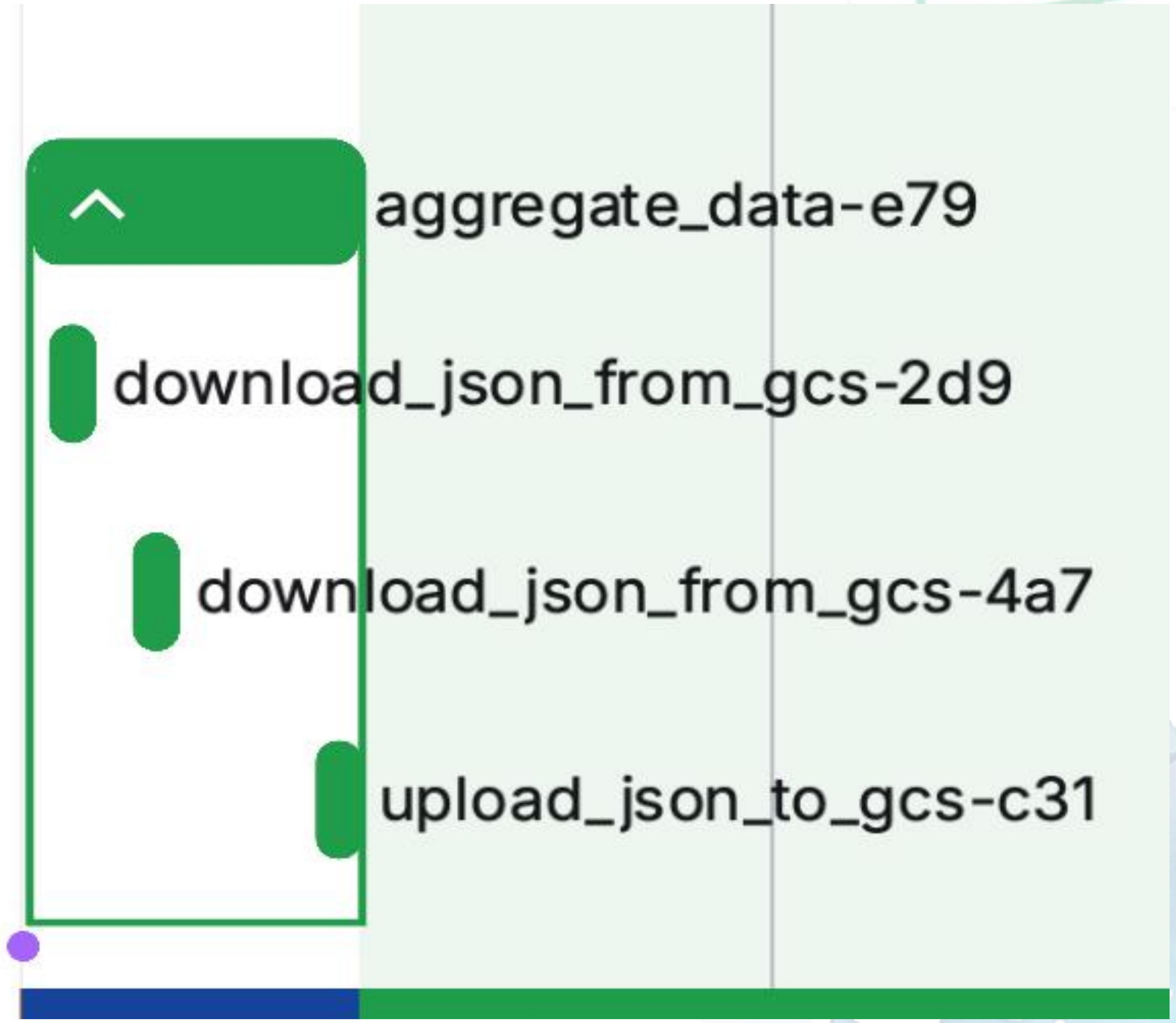
Weekly Data Ingestion Flow

Phase 3: Data Aggregation “E & t” of the EtTL

We used deployment_aggregation_flow.py to

- For the first two week:
  1. Identify the initial scraped data
  2. Identify the first weekly scraped data
  3. Append them into an aggregation file and remove duplicates using 2 fields: reviewerId & publishedAtDate
  4. Upload the file into GCS

- For the 3rd week and ahead:
  1. Identify the latest weekly scraped data
  2. Identify the most recent aggregation file
  3. Append them into an aggregation file and remove duplicates using 2 fields: reviewerId & publishedAtDate
  4. Upload the file into GCS

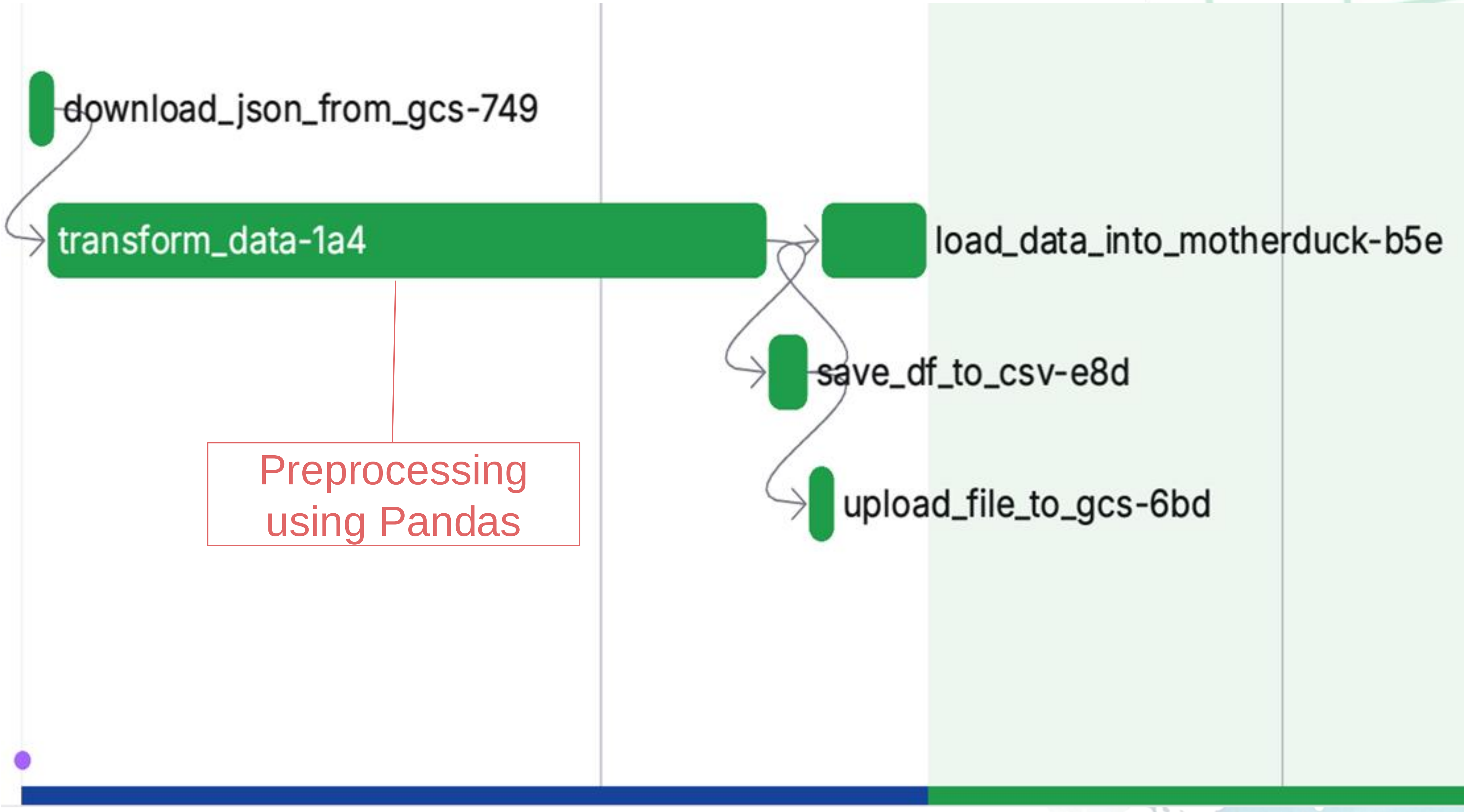


Aggregation Flow

Phase 4: Transformation & Load “T & L” of the EtTL

We used deployment_transform_and_load.py to

1. perform the most basic text preprocessing (remove special characteristics, lowercase the text, etc.)
2. Feature Splitting
   (for example, Review Rating-> food, service, atmosphere)
1. Save the dataframe into CSV format
2. Upload it into MotherDuck DB



Preprocessing using Pandas

Initial Data Ingestion Flow

The overall pipeline schedule

Initial Scrape

Oct 07

Weekly Scrape

Once a week, beginning
from Oct 15.

Aggregation

Execute right after
each weekly scrape

Transform & Load

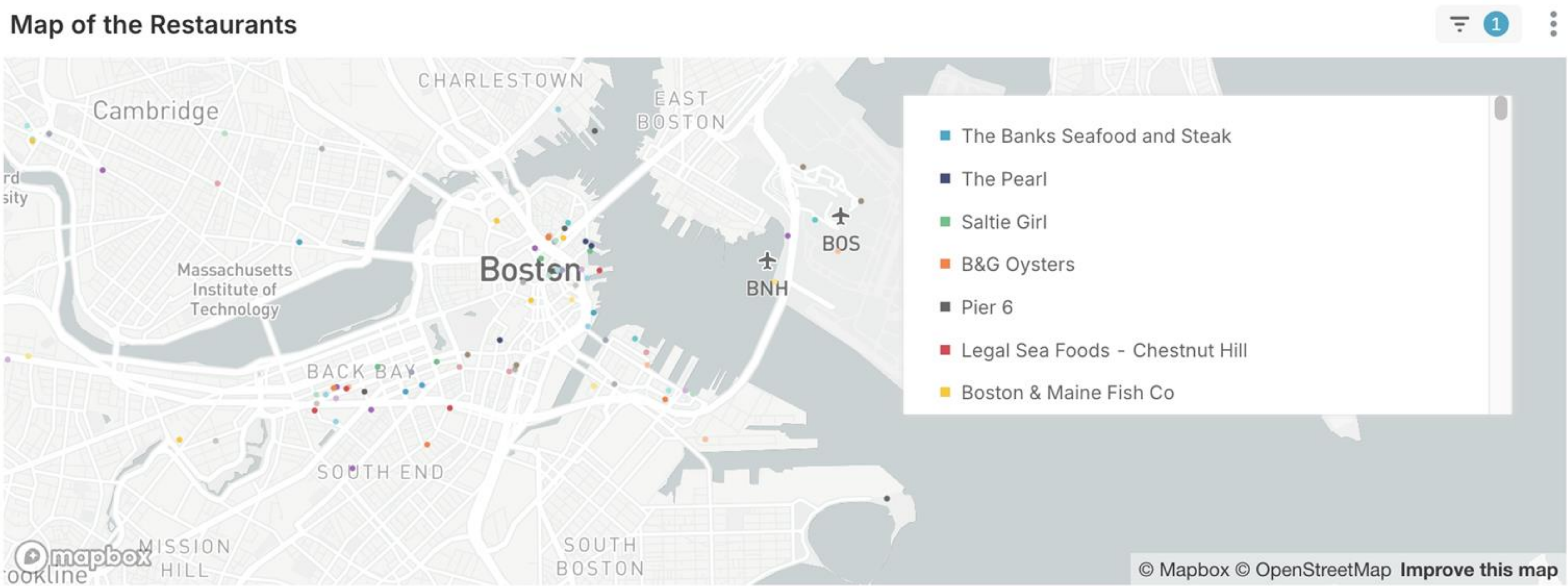
Execute right after
each weekly scrape.

4

Dashboard Insights & Plans

Dashboard on Superset & Potential improvements on the pipeline

Superset Dashboard



Map of the Restaurants

- 🟦 The Banks Seafood and Steak
- ■ The Pearl
- 🟩 Saltie Girl
- 🟧 B&G Oysters
- ■ Pier 6
- 🟥 Legal Sea Foods - Chestnut Hill
- 🟨 Boston & Maine Fish Co

Link:http://34.68.1.107:8088/superset/dashboard/p/dmR3QzZ5067/

Next Step:

- Text data: Implement NLP models to extract more insights from customer reviews. Analyze the text part of reviews to identify praises, complaints, and suggestions.

- Numerical data: Combined with text data, develop ML models that can represent and assess restaurant performance.

- Build the recommendation engine for customers and give restaurant owners insights to improve their food, service, decorations, etc.

- Enhance the pipeline by implementing monitoring and alert system for timely issue detection, utilize Generative AI tools in the next step.

Thanks For Listening!