Safe stay with Airbnb

A Recommendation Platform

APAN 5400 Final Presentation

Group 5: Yulu Sun, Zixuan Wang, Ziyi Weng, Ning Yang, Chao Zhang

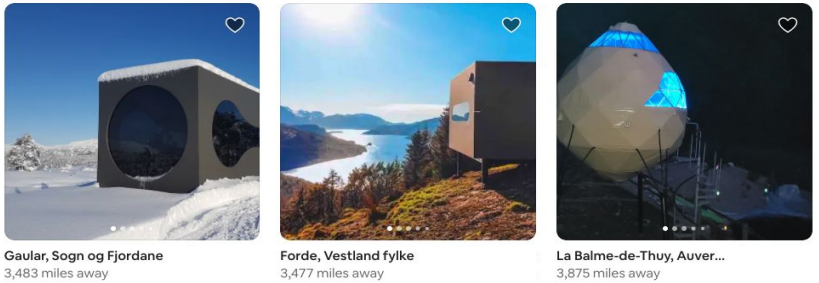
Background

A recommendation system that help users find the perfect airbnb to stay in with COVID-19 consideration.

-Covid
- Built rating system based on the number of COVID-19 case rate in last 14 days

-Airbnb
- Build a rating system based on the real customer feedback
- More freedom for customer to choose

- Business Use
- Encourage travel with COVID-19 awareness
- Recommendation given based on location
- Target the travelers who are concerned about COVID-19 when they make airbnb travel decision
- Potential Collaboration with Airbnb, making a world wide covid travel recommendation system.



Data Preparation

CSV for the project (demo)

- LA NY Coronavirus (COVID-19) Dashboard
  - Covid 19 statistics on each county
  - Cases - Incident_Rate(last 14 days)
  - LA csv file contain 280 community(19MB)
  - NY csv file contain 74 community(2.88MB)
- LA NY Airbnb Data
  - (151200 lists(48.5MB), 8348196 reviews(3GB))
  - Mostly last two years of housing information(eg.name,type)
  - Housing price on average
  - Housing reviews and rating (key words)
  - Housing region (county)

Data sources:http://dashboard.publichealth.lacounty.gov/covid19_surveillance_dashboard
https://health.data.ny.gov/d/xdss-u53e/visualization
https://www.kaggle.com/datasets/samyukthamurali/airbnb-ratings-dataset?resource=download&select=airbnb-reviews.csv

Data Source Specification

Important data explanation

Colum Name	Data type	Description
Positive 14 Days	Numeric	The case rate during last 14 days
Condition	Factor	Four different covid concerning levels
Listing_id	Numeric	Every listing had unique ID
Property_type	Factor	Airbnb type, house or apartment
Airbnb_scores	Numeric	Airbnb original rating score
Count(New score system)	Numeric	How many times key word in the customer review (great, friendly)

Data Quality Dimensions

1 Completeness

With fully supported by Airbnb, The world wide listing data and reviews will be very complete

2 Accuracy

Covid data and Airbnb data all come from official sources, the data is verified by state and airbnb data department.

3 Consistency

Data is very consistency with the regular trend, It is also will be updated on daily basis

4 Validity

Data type, range , and formate in a good shape

5 Uniqueness

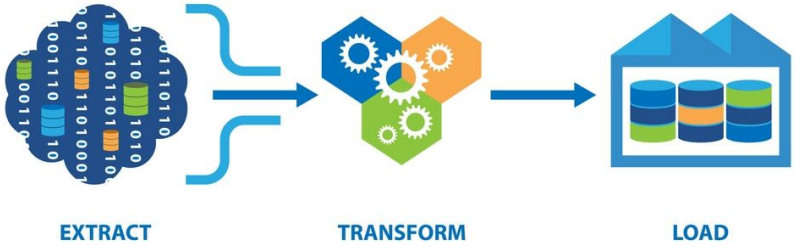
Each county and listing are unique

6 Integrity

The Covid data is open sourced for public, Airbnb data will require permission

Design Choices (ETL Design)

- **Extract:** Get raw data from official public health website, and convert csv files to sql files to store in database
- **Transform:** Interact with data using pgAdmin; Create datastores, filter out the unused data
- **Load:** Use datastores; Complete web development and return query values



The Selected Technologies

PostgreSQL

- PostgreSQL calculate COVID-19 case rate within each region, assign 4 different pandemic risk levels for each region(low, medium-low, medium-high and high).
- PostgreSQL returns the three regions with lowest case rate within each risk level to promote safe travel

Reason:

- PostgreSQL is very stable to manage data in a relational database
- Retention features are important for Web applications, PostgreSQL is better

MongoDB

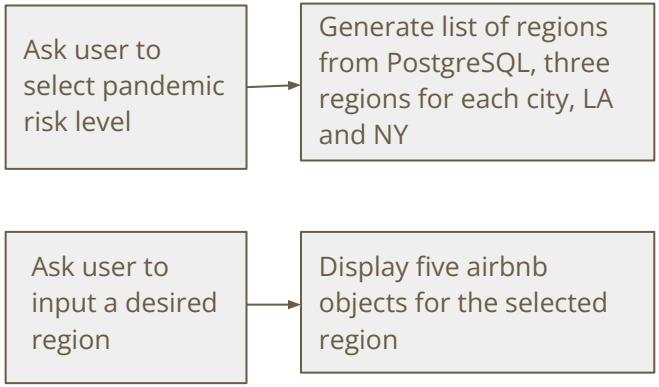
- Rank the houses in the selected area–based on the appearance frequency of keywords in past user reviews
- Find the top five house information in a given region

Reason:

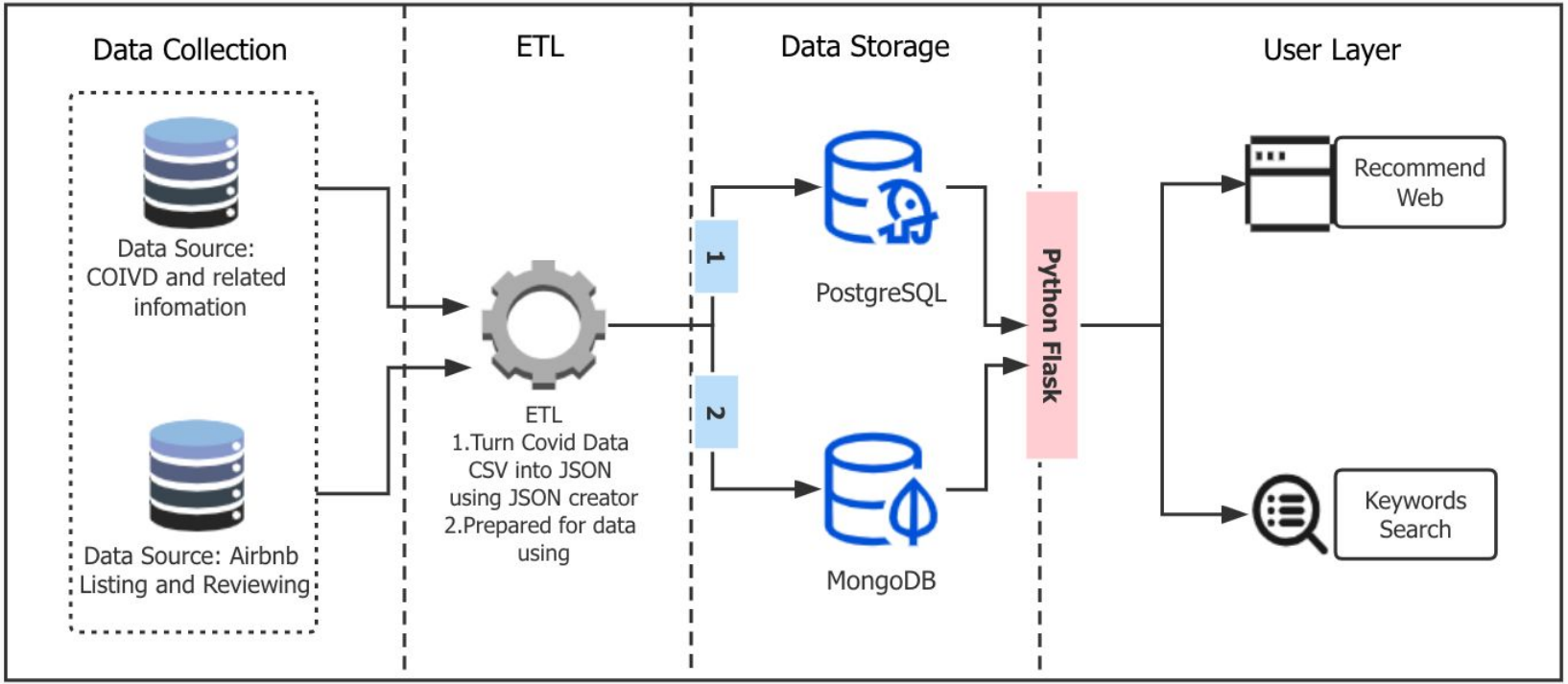
- It can better ensure the user's access speed.
- It is easier to obtain data.
- Support large-capacity storage.

Flask

- Interactive Web development with flask
- Designed HTML put in template file, use render_template() to render HTML file into Python, allowing requests to main/route using functions built



Data Pipeline



See Prototype Design : https://www.processon.com/diagraming/626183030e3e74074acfb01a

Data Governance and Cost Implications

Data Governance

- Timely updates the data of Covid cases , the Airbnbs' rent data and the reviews
- Allocate database permissions and roles, restrict accessibility to the database
- Avoid information leaks and crawlers
- Detect and screen out robert reviews/fake housing

Cost Implications

Regulation	• Data subscription fee from Airbnb,Covid • Develop and maintenance fee that can updates data in time and keep operation fluent • Server Cost with big data (API,Database,etc.)
Feature	• Develop automatic monitoring function to ensure the accuracy and authenticity of data, eg. Block the house-id which comments >10 consecutive reviews
Security	• Database security system contain complex passwords, two-factor authentication to protect our users and database

Scalability and Cost Details

API–Airbnb & Covid

- Fully supported by Airbnb and use their API, data access cost may be eliminated in the future
- Public Health Center which provide Covid data is an open data source

Flask

- Flask use a web server called *Werkzeug* which good for development. But we can use **Gunicorn** for production in the future, a broadly compatible with various web frameworks and fairly speedy.
- It has good support for multi-processing, which is good for our use case since we would be I/O constrained when reading data from the database, processing multiple requests at the same time would limit the number of Flask applications.
- Plus, we can start off with 10 workers per server and have a thread pool size of 10 for each worker which process those requests as fast as we can. Or choose a larger I/O in Amazon redshift.

mongoDB

Category : NoSQL
Spec: M200; vCPUs:64; Storage:1500GB; RAM:256GB
Cost: $**10,505** monthly (1 instance * $14.59/h*720 hrs)

Amazon EC2

Category : Sever
Spec: vCPUs:2; Memory:16 GiB; Up to 4,750 Mbps
Cost : $**96.1**/month(1 instance * $14.59/h*720 hrs)

amazon REDSHIFT

Category : Data Warehouse
Spec: vCPUs:32; Memory:244 GiB; I/O:7.5GB/s
Cost : $27,640 yearly OR
$**3,456** monthly (1 instance * $4.8/h*720 hrs)

Sources: https://aws.amazon.com/ec2/instance-types/; https://aws.amazon.com/redshift/pricing/; https://www.mongodb.com/pricing
https://flask.palletsprojects.com/en/2.1.x/becomingbig/#scale-like-a-pro https://gunicorn.org/#quickstart

Future Improvements and Demo Display

Improvements

- After being recognized by Airbnb, expanded horizontally into travel-related industries, such as hotels and entertainment centers
- NLP will be used for sensitive analysis to obtain more accurate evaluation words in reviewing and build better reviewing score, customer will be able to personalized their keyword search to find related reviews.
- Add broader global data related to COVID, such as access to the United Nations, WHO, to achieve Covid related data

Demo (Screen Share)

Thank you!

Welcome for any suggestions