# Group 6
# Yelp Business Intelligence Cloud Subscription Service

Group member::
Sheng Luo | Rocky D. Wang | Zhiyuan Liu | Yao Zhang | Harvey Wang

Based on yelp Customer's Review

# Background and Use case

## Background

- Importance of online reviews:
  - Online reputation is crucial in today's digital world
  - Yelp is a popular platform for rating and reviewing businesses
- Impact on restaurants:
  - Positive reputation drives more customers and revenue
  - Negative reputation deters potential customers

## Use case

- Automatic monthly report generation
- Analyzes Yelp review data to help restaurant owners
- Provides insights into customer preferences, opinions, and concerns
- Allows benchmarking against competitors and market trends

# Team Member's Responsibilities

01 **Sheng Luo - S3:** Store Raw and application data in storage

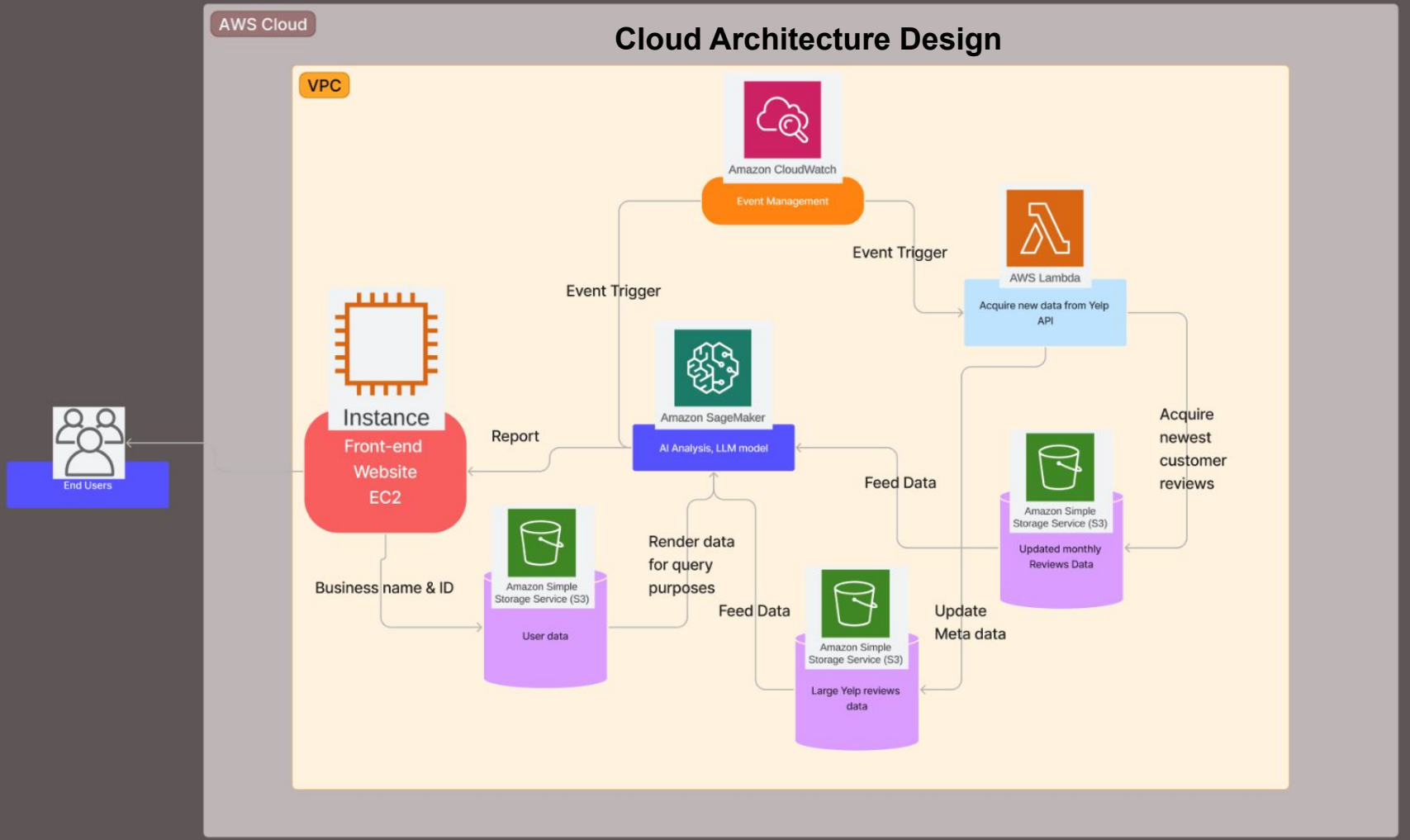
02 **Yao Zhang - Lambda & Cloudwatch:** To implement event-based cloud computation and services.

03 **Zhiyuan Liu - Sagemaker:** To preprocess data and train models for application.

04 **Harvey Wang - EC2:** To collect user information

05 **Rocky Wang - VPC:** To Manage the accessibility and security.

# Cloud Architecture Design



# Data source and details

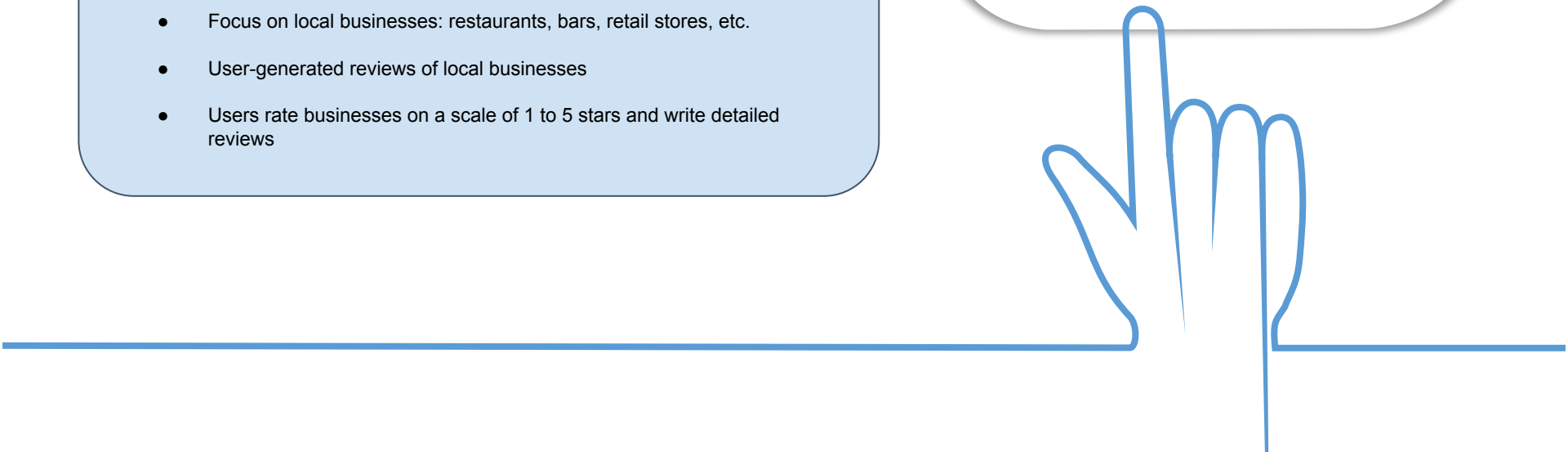
## Yelp datasets - 5 GB

This dataset is a subset of Yelp's businesses, reviews, and user data provided by Kaggle.

**Dataset Composition**

- Information about businesses across 8 metropolitan areas in the USA and Canada
- Focus on local businesses: restaurants, bars, retail stores, etc.
- User-generated reviews of local businesses
- Users rate businesses on a scale of 1 to 5 stars and write detailed reviews

- **yelp_business_attributes.csv**
- **yelp_business_hours.csv**
- **yelp_business.csv**
- **yelp_review.csv**
- **yelp_tip.csv**
- **yelp_user.csv**



# Virtual Server

## by EC2

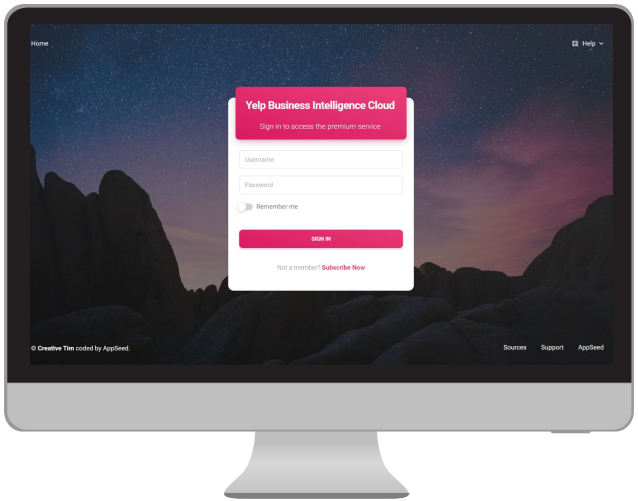
When new subscriber is added

- **Website Hosting**:

EC2 hosts the front-end website, allowing users to register and log in to their accounts.

- **Integration with AWS Services**:

EC2 integrates seamlessly with other AWS services.



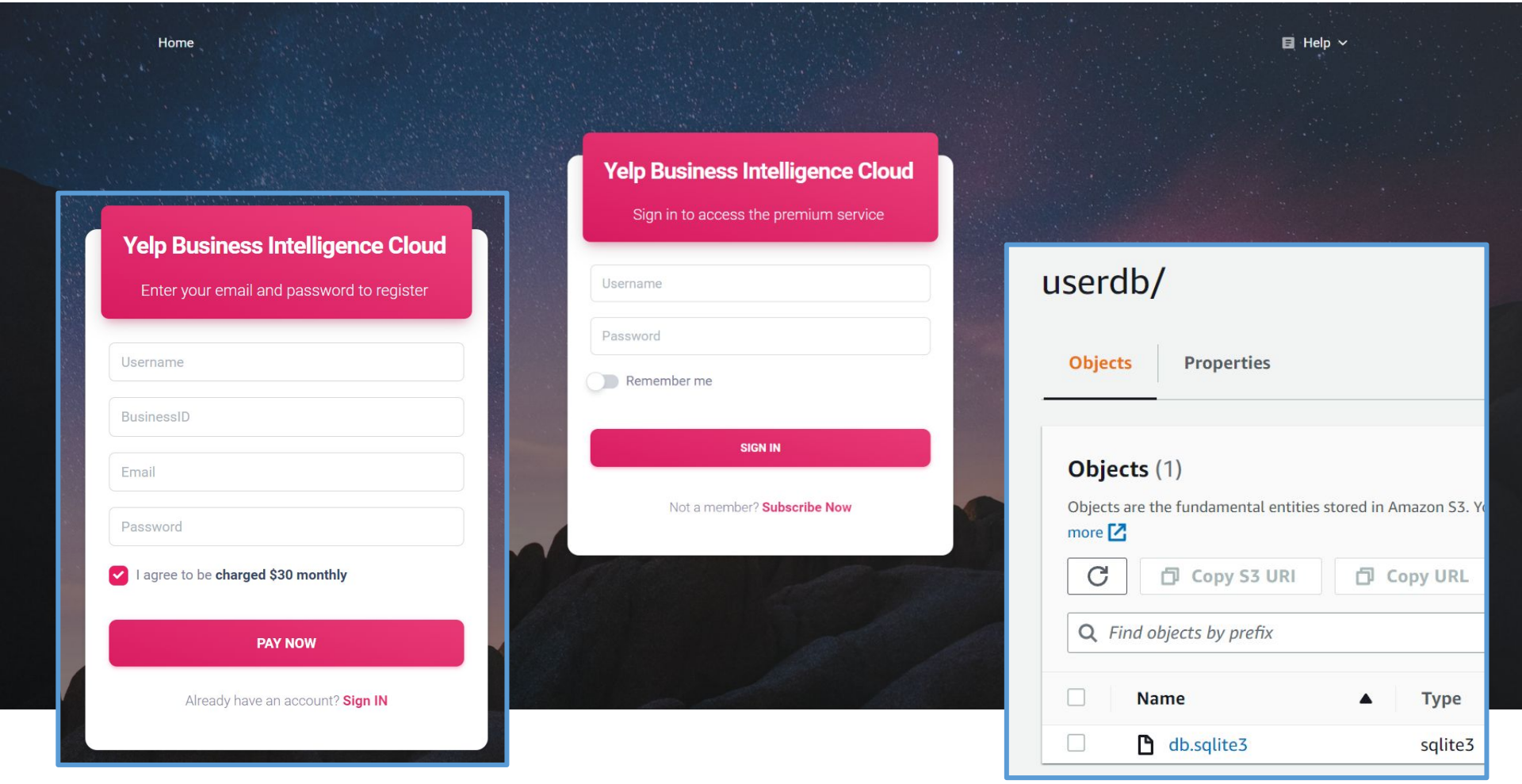

- **Management and Monitoring**:

EC2 provides various tools and APIs for managing and monitoring your instances.

- **Pay-as-you-go Pricing**:

EC2 follows a pay-as-you-go pricing model, where you pay only for the compute resources you consume.

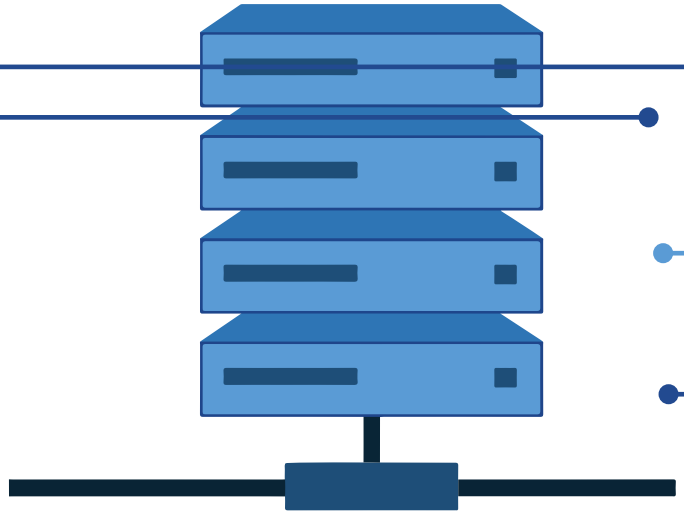
# EC2 Live Demo: http://44.204.87.217:5000/login



# Event Based Pipeline

## by Lamda and Event Bridge(Cloudwatch Event)

Monthly API request



**Data Ingestion**

We created Lambda functions that read from subscriber's data and constantly update database

**Severless APIs**

We created serverless APIs using Lambda. This is useful for exposing data and analysis results to other applications and services triggered by event.

**Report/survey tools**

By using Lambda, we can automate the process of sending report or surveys to restaurant owners.

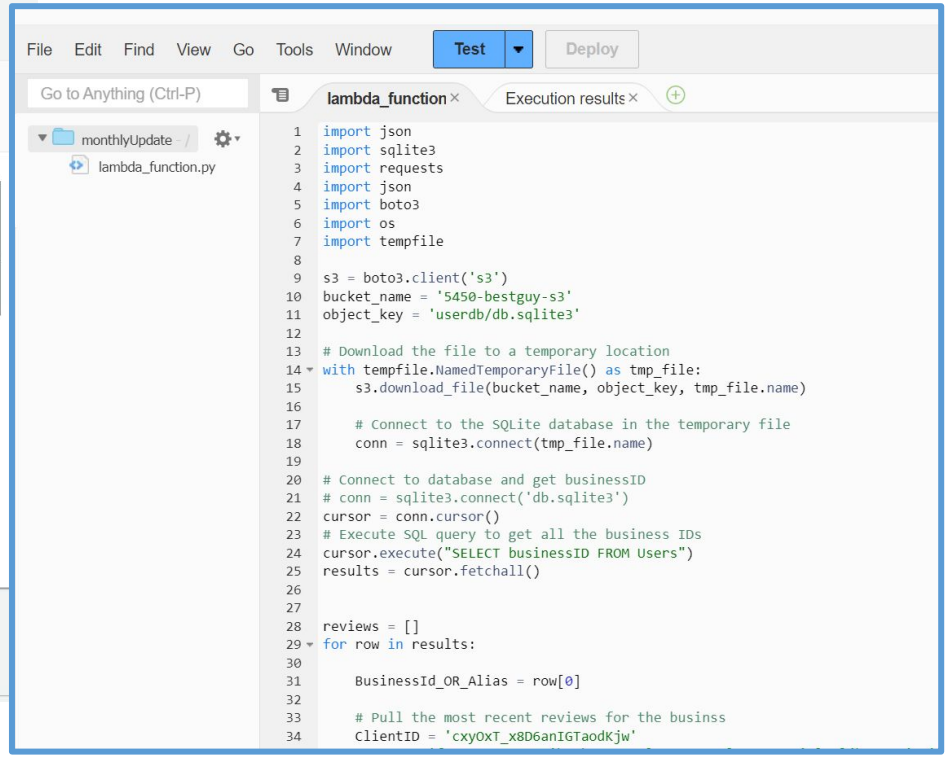
# monthlyUpdate

## Function overview Info



+ Add trigger

Code




## voc-codebuild-cw-rule

### Rule details Info

Rule name
voc-codebuild-cw-rule

Status
Enabled

Description
monthly update

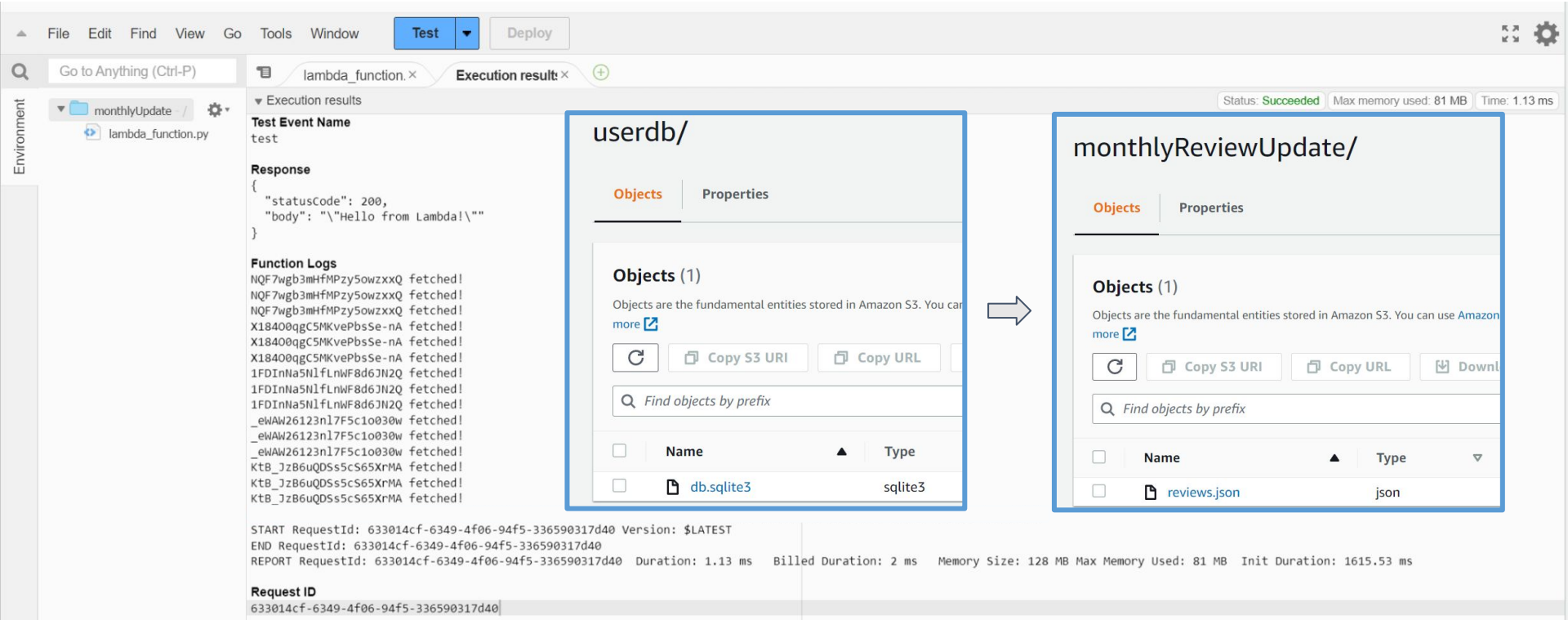
Rule ARN
arn:aws:events:us-east-1:925398600691:rule/voc-codebuild-cw-rule

Event schedule | Targets | Monitoring | Tags

### Event schedule Info

Fixed rate of
30 day

# Test/One time Running the Function



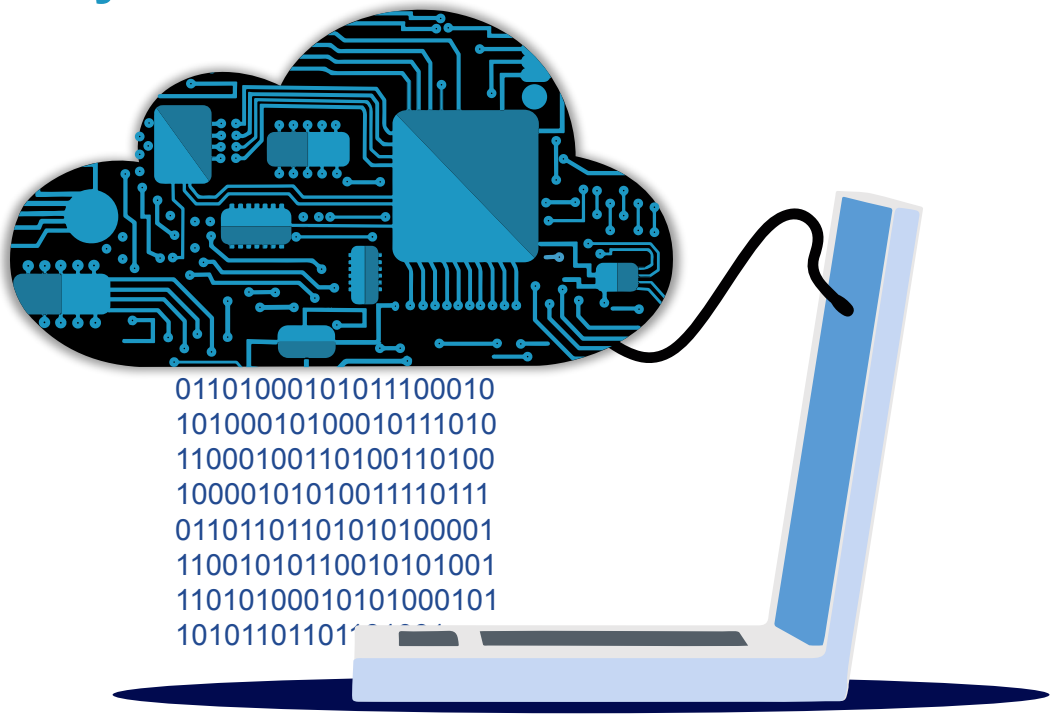
# Data Storage

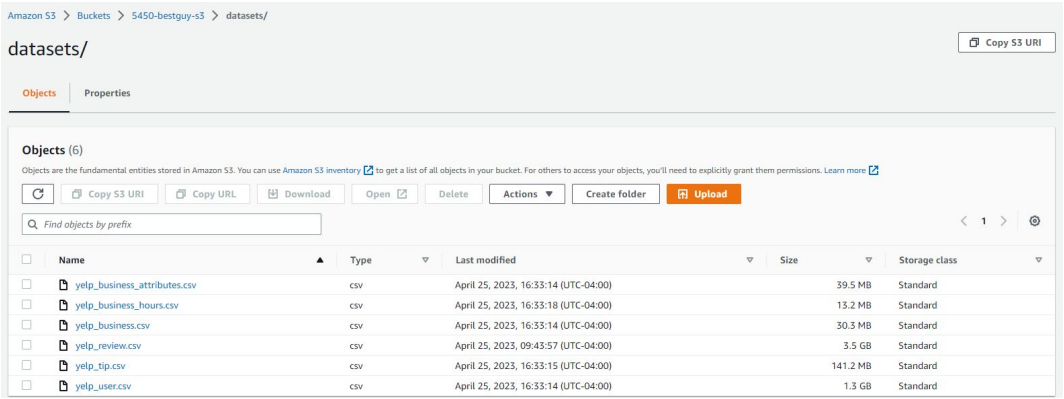
## by S3

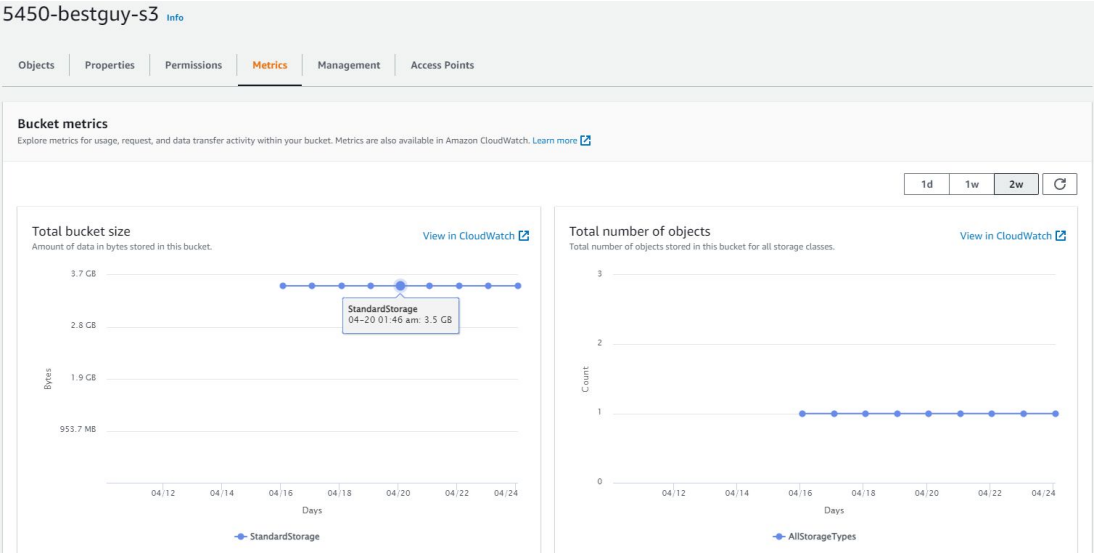
### Store Raw data in Amazon S3

- Upload the raw data
- Create a new bucket.
- Configure the bucket permissions
- Allow Lambda or Sagemaker to access it

- Sentiment analysis
- Store the processed data in S3 Amazon
- SageMaker for further analysis or visualization






# S3

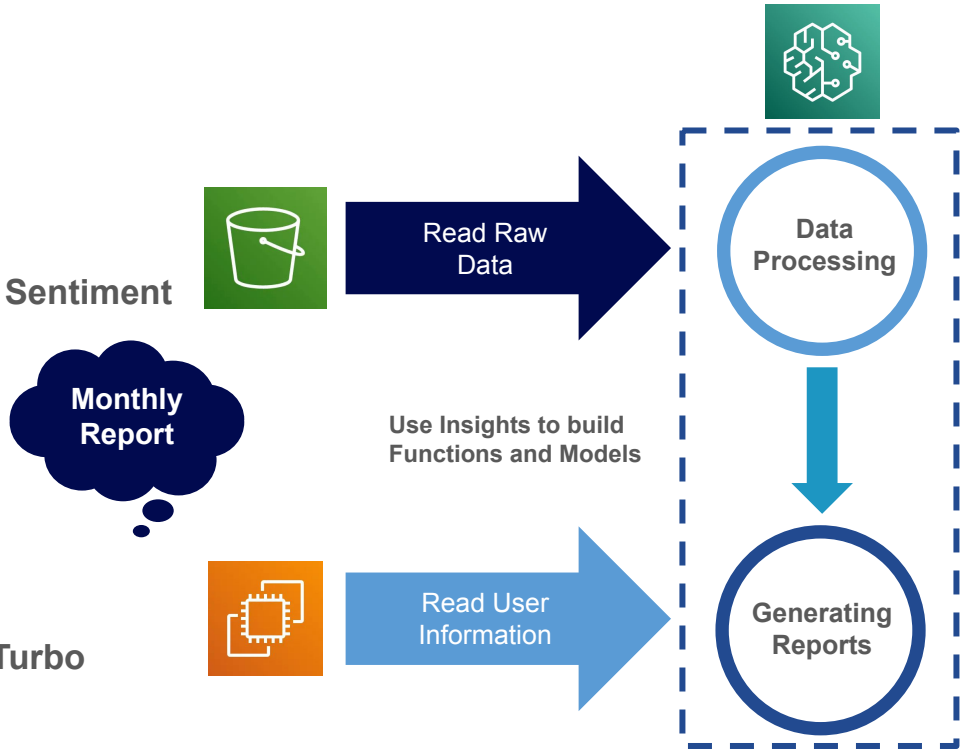
# Datasets And Metrics

With AWS S3 as the data storage solution, data is automatically replicated across multiple availability zones, which provides resiliency and ensures data availability. Additionally, EC2 and SageMaker both offer automatic backup and recovery options, which further enhances resiliency.

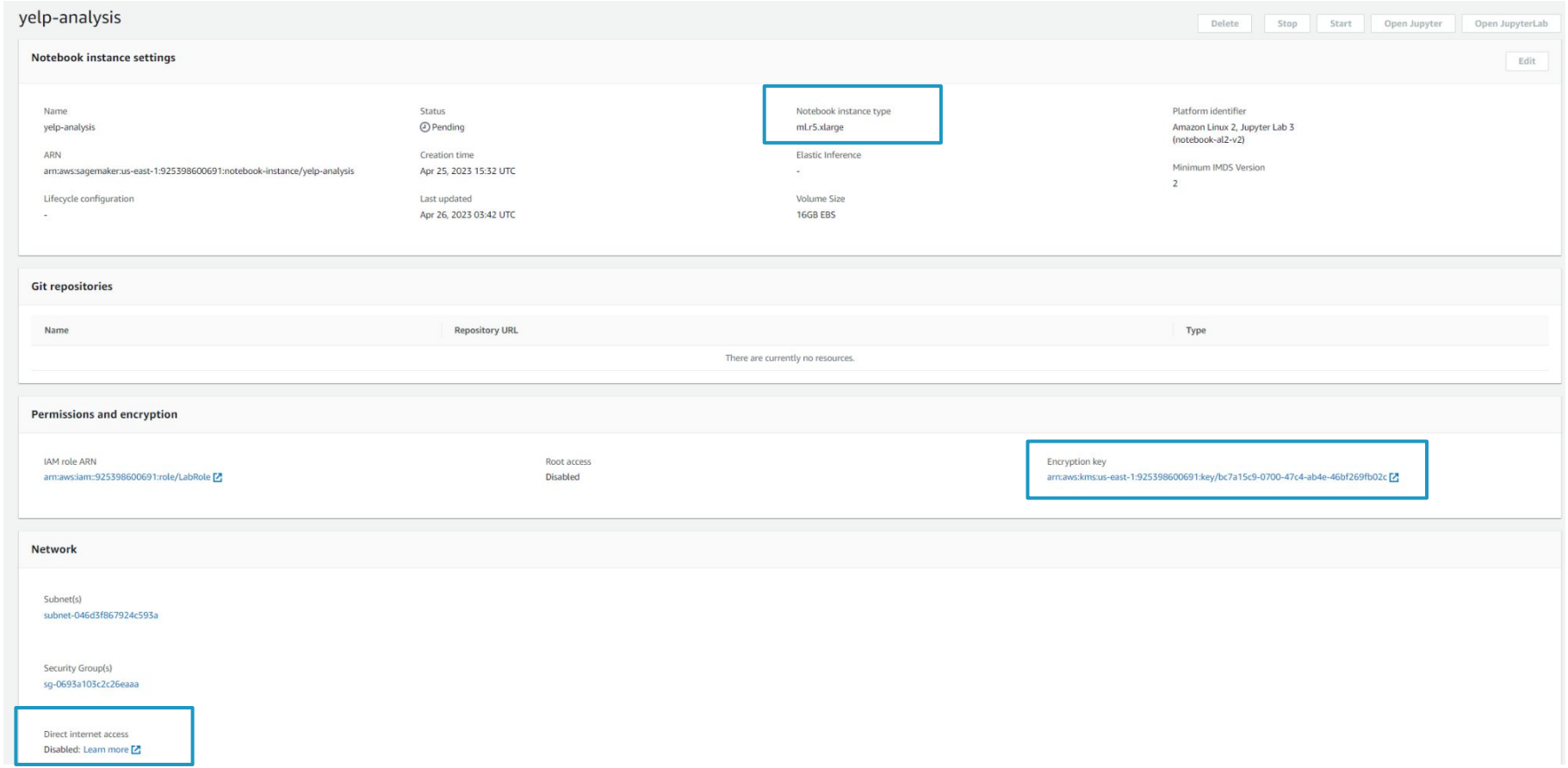
# Accelerating Data Analysis

## by SageMaker Notebook

- Data Processing
  - Raw Data: **Over 5GB** Yelp data
  - Analysis: **Competitor, Geographic and Sentiment**
- Generating Report:
  - Method: Python **`reportlab`** library
  - Visualization: **10+ graphs or charts**
  - Suggestions: Generating from **GPT 3.5 Turbo**



# Sagemaker Screenshots



# Sagemaker Screenshots

Click

```
# Section 2: Review Analysis
elements.append(Paragraph("2. Review Analysis", styles["Heading3"]))

elements.append(Paragraph("The following charts provide insights into the business's review trends. \
                          star rating distribution, and the 'useful', 'funny', and 'cool' properties of the reviews. \
                          By analyzing these charts, users can better understand the business's performance and areas for improvement.", custom_style))

image_files = [
    (review_path[0], "Figure 1: Number of Reviews per Month",
     "This chart shows the total number of reviews received by the business each month."),
------
    (review_path[1], "Figure 2: Average Stars per Month",
     "This chart shows the average star rating given by customers each month."),
------
    (review_path[2], "Figure 3: Star Distribution",
     "This chart displays the distribution of star ratings received by the business."),
------
    (review_path[3], "Figure 4: Useful, Funny, and Cool Distribution",
     "This chart shows the distribution of the 'useful', 'funny', and 'cool' properties in the reviews."),
-------
    (os.path.join(save_folder, "Afinn_Analysis.png"), "Figure 5: Sentiment Analysis using Afinn",
     "This bar chart shows the sentiment distribution using Afinn sentiment scores, categorized into Extreme Negative, Negative, Neutral, Positive, and Extreme Positi

    (os.path.join(save_folder, "NRC_Analysis.png"), "Figure 6: Sentiment Analysis using NRC",
     "This bar chart displays the frequency of NRC emotions detected in the reviews, providing insights into the emotional content of the reviews."),
-------
    (os.path.join(save_folder, "Combined_Word_Clouds_with_Titles.png"),
     "Figure 7: Combined Word Clouds for Different Sentiment Categories",
     "This combined word cloud image shows the most frequently used words in different sentiment categories. It can help users identify the most common themes and key

    (os.path.join(save_folder, "Combined_Bar_Charts.png"),
     "Figure 8: Top 20 Words Frequency for Different Sentiment Categories",
     "This combined horizontal bar plot image shows the frequency of the top 20 words for different sentiment categories. It can help users compare the most common wo
]



for image_file, title, description in image_files[0:2]:
    put_image(image_file, title, description, pagebreak=False)
------
for image_file, title, description in image_files[2:4]:
    put_image(image_file, title, description)
------
------
for image_file, title, description in image_files[4:6]:
    put_image(image_file, title, description)

for image_file, title, description in image_files[6:8]:
    put_image(image_file, title, description, height = 600)

elements.append(Paragraph("GPT Analysis of Negative Reviews", styles["Heading4"]))

gpt_answer = response['choices'][0]['message']['content'].replace('\n', '<br/>')

elements.append(Paragraph(gpt_answer, styles["BodyText"]))

# Add the copyright statement
copyright_statement = "© 2023 Bestguys Tech Ltd."
elements.append(Paragraph("<para align=center spaceAfter=10><font size=12><b>{}</b></font></para>".format(copyright_statement), styles["Normal"]))
```

## Yelp In-Depth Report

"Gangnam Asian BBQ Dining" (hihud--QRriCYZw1zZvW4g)

### *1. Market Analysis*

***Competitive Density Analysis***

Number of businesses within 5 km of business hihud--QRriCYZw1zZvW4g: 7613.
Number of similar businesses with similarity score > 0.3: 6

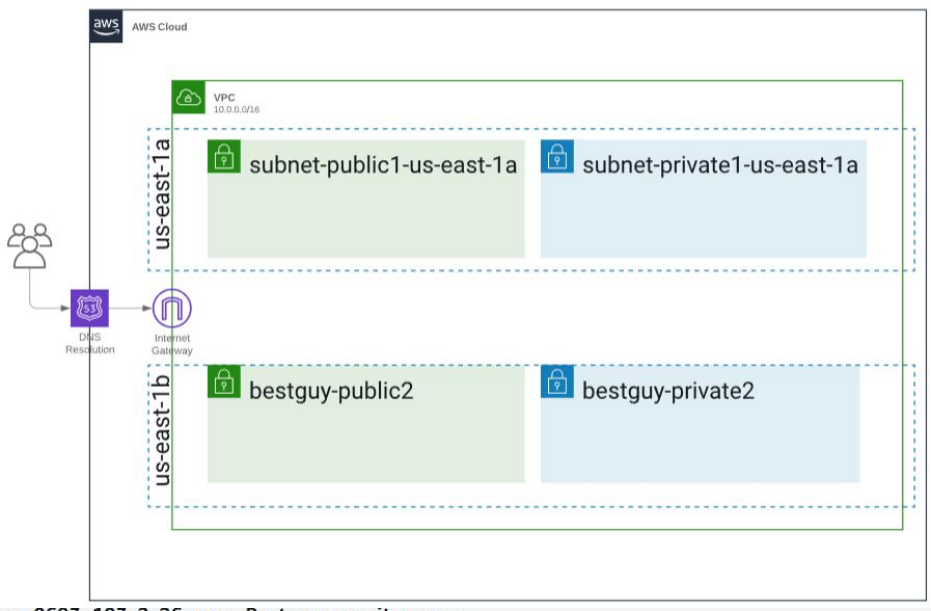
***Key Competitors***

1. "Azuza Hookah Lounge & Cafe" (Similarity: 0.53)
2. "Dog Haus" (Similarity: 0.53)
3. "Joe's New York Pizza" (Similarity: 0.44)
4. "Capriotti's Sandwich Shop" (Similarity: 0.44)
5. "Stussy" (Similarity: 0.37)
6. "High Roller Cigar & Smoke Shop" (Similarity: 0.37)
7. "Lok Acupuncture Clinic" (Similarity: 0.30)
8. "Djanel Spa" (Similarity: 0.30)
9. "Pharaoh Beads" (Similarity: 0.30)
10. "Oyster Bar" (Similarity: 0.30)

***Table 1: Key Competitors Information***

| name | neighborhood | address | stars | review_count | is_open |
|---|---|---|---|---|---|
| "Azuza Hookah Lounge & Cafe" | Eastside | "4480 Paradise Rd" | 4.5 | 239 | 1 |
| "Dog Haus" | Eastside | "4480 Paradise Rd" | 4.5 | 282 | 1 |
| "Joe's New York Pizza" | Eastside | "4480 Paradise Rd" | 3.5 | 384 | 1 |
| "Capriotti's Sandwich Shop" | Eastside | "4480 Paradise Rd" | 4.0 | 368 | 1 |
| "Stussy" | Eastside | "4480 Paradise Rd" | 3.5 | 25 | 0 |
| "High Roller Cigar & Smoke Shop" | Eastside | "4480 Paradise Rd, Ste 250" | 2.0 | 9 | 1 |
| "Lok Acupuncture Clinic" | Eastside | "1818 E Desert Inn Rd" | 4.5 | 34 | 1 |
| "Djanel Spa" | Eastside | "372 E Tropicana Ave" | 4.5 | 21 | 1 |
| "Pharaoh Beads" | Eastside | "3528 S Maryland Pkwy" | 4.5 | 10 | 1 |
| "Oyster Bar" | Eastside | "4455 Paradise Rd" | 4.5 | 61 | 1 |

# VPC Resource Map and Security Group




| | | | | |
|---|---|---|---|---|
| Inbound rules : | IPv4 | HTTP | TCP | 80 |
| Outbound rules : | IPv4 | All-traffic | All | All |

**sg-0693a103c2c26eaaa - Bestguy security group**

Details | Inbound rules | Outbound rules | Tags

## Details

| Security group name | Security group ID | Description | VPC ID |
|---|---|---|---|
| Bestguy security group | sg-0693a103c2c26eaaa | Enable HTTP access | vpc-0f1b48a104d6bef89 |

| Owner | Inbound rules count | Outbound rules count |
|---|---|---|
| 925398600691 | 1 Permission entry | 1 Permission entry |

# Cost Analysis

AWS COST - Here the upfront cost is **$42.05** and the monthly cost is **$40.12.** The long term cost for one year is **$523.49.** Our selection of services includes EC2, S3, SageMaker, VPC.


**Google Cloud Cost- The monthly cost is $52.17. The long term cost for one year is $626.04. Our selection of services contains Compute Engine, Cloud Storage, BigQuery ML,Cloud VPN**

# Success Criteria

## 01 Security

- ✔ Virtual Private Cloud (VPC) - a private network helps protect in transit data
- ✔ Security Groups - control inbound and outbound rules for our instance, separate security groups for different tiers in the future

## 02 Reliability

- ✔ Multiple availability zones - application is deployed across multiple data centers, ensuring high availability with a higher failure tolerance

## 03 Performance Efficiency

- ✔ Auto scaling - automatic scale computing resources to handle spikes or low demand

## 04 Cost Optimizer

- ✔ AWS Cost Explorer - identity opportunities for spending optimization
- ✔ AWS Budgets - control costs by setting a spending limits as well as receiving alerts when close to the limit