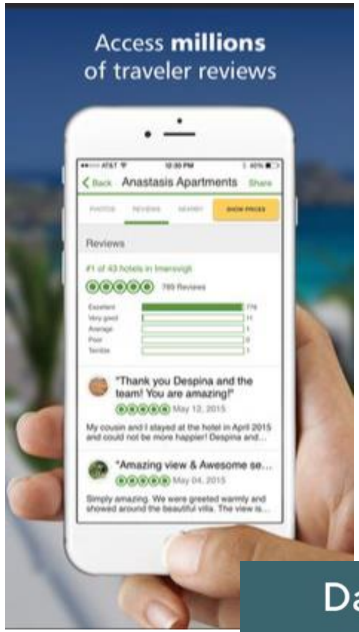
ACME INC.
Restaurant Advisor

(If Tripadvisor and Doordash had a baby…or Tinder)

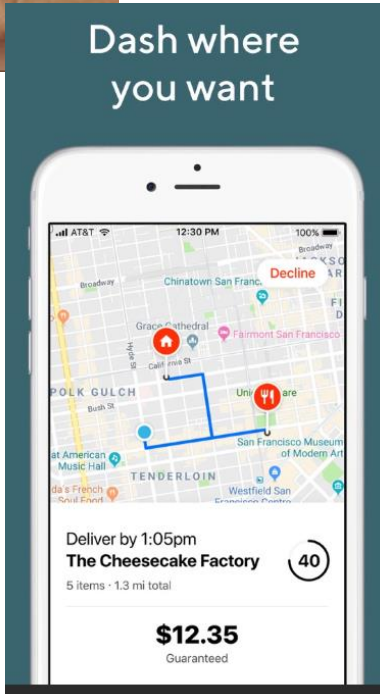
Project Team 1 Members: P. Marie Barrier, Alex Brooks, Indhu Kethireddy, Deevanshu Khatri, and Yolanda Zhou

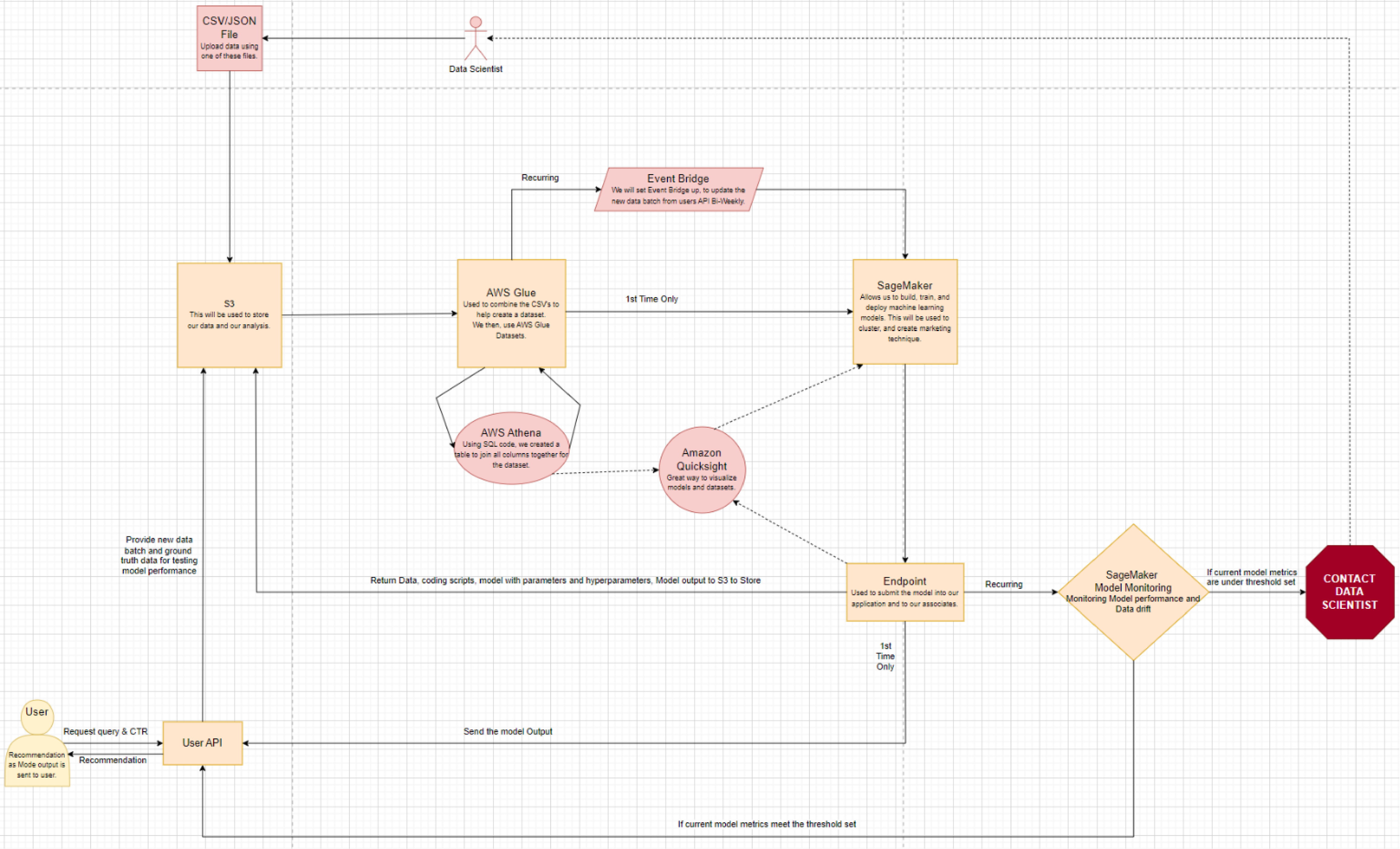


Project Background and Introduction

Team Role: Data Scientist & Market Strategist to compete with Doordash/Tripadvisor

- Goals:
  - Leverage customer database to create a targeted marketing campaign - specifically: "cross sell" to existing customers
  - Move the customer data to the cloud
  - Cluster the customers
  - Create a targeted campaign strategy
  - Continuous improvement over time
  - Create a measurement plan to prove the effectiveness of your model(s).





CSV/JSON File
Upload data using one of these files.
Data Scientist
Recurring
Event Bridge
We will set Event Bridge up, to update the new data batch from users API Bi-Weekly.
S3
This will be used to store our data and our analysis.
AWS Glue
Used to combine the CSV's to help create a dataset. We then, use AWS Glue Datasets.
1st Time Only
SageMaker
Allows us to build, train, and deploy machine learning models. This will be used to cluster, and create marketing technique.
AWS Athena
Using SQL code, we created a table to join all columns together for the dataset.
Amazon Quicksight
Great way to visualize models and datasets.
Provide new data batch and ground truth data for testing model performance
Return Data, coding scripts, model with parameters and hyperparameters, Model output to S3 to Store
Endpoint
Used to submit the model into our application and to our associates.
Recurring
SageMaker Model Monitoring
Monitoring Model performance and Data drift
If current model metrics are under threshold set
CONTACT DATA SCIENTIST
1st Time Only
User
Request query & CTR
User API
Recommendation
Recommendation as Mode output is sent to user.
Send the model Output
If current model metrics meet the threshold set

Meeting Project Goals - using our Infrastructure for Business Solution:

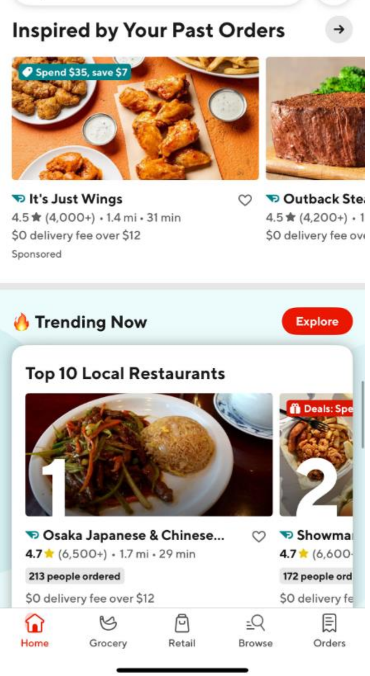
- ➔ Leverage customer database to create a **targeted marketing campaign - specifically: "cross sell" to existing customers**
- ➔ Move the customer data to the cloud
- ➔ **Cluster the customers**
- ➔ **Create a targeted campaign strategy**
- ➔ Continuous improvement over time
- ➔ Create a measurement plan to prove the effectiveness of your model(s).

Data Preprocessing ➔ **Data Exploration & Model Development** ➔ **PROD Scripts for Recurring Model Build and Use** ➔

Athena
SQL Editor

Sagemaker
Notebook
Instance

Lambda Function
for Integration
into Operations



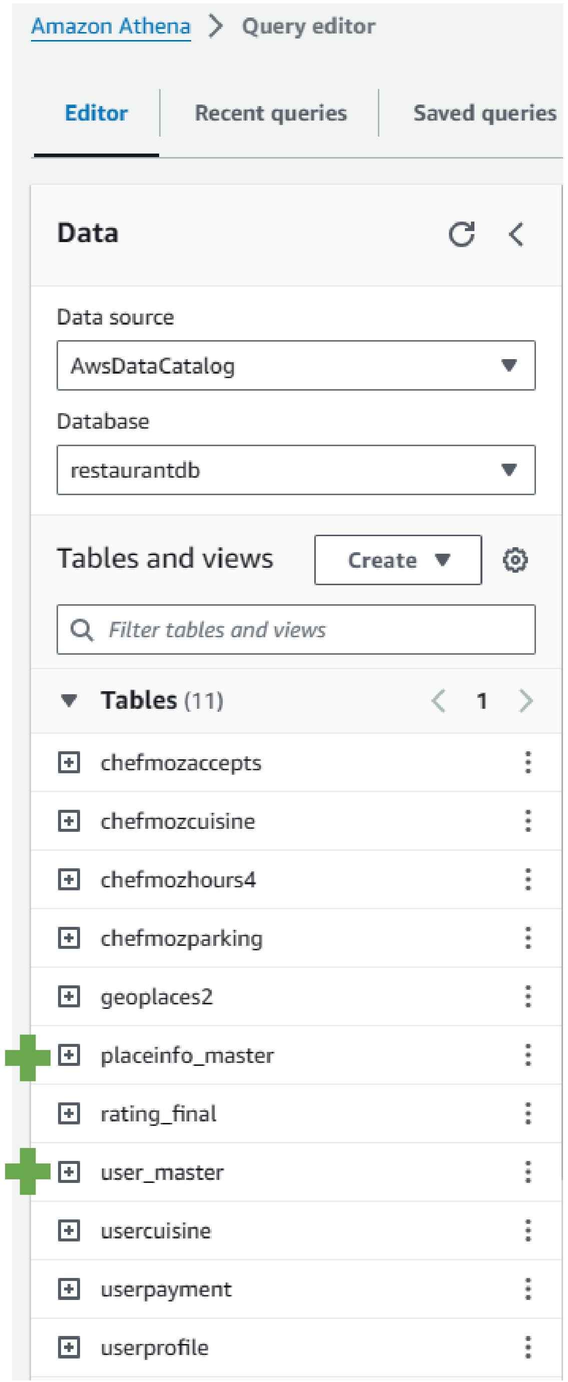
S3 -> Glue Crawler -> AWS Glue Data Catalog -> **Athena SQL Editor**:

1. Combine multiple data sources into 2 master tables

   placeinfo_master
   location information, cuisine type, dress code, accessibility, price, rating, food rating, service rating, alcohol, days, hours, etc

   user_master
   rating, food rating, service rating, smoker, drinking level, marital status, birth year, religion, weight, budget, etc

1. (For PROD) repeatable/easy cleaning here
   -append new data with saved scripts via Glue Job
   -impute empty fields using mode for categorical, mean for quantitative



Amazon SageMaker Notebook Instance (ml.t3.xlarge):

DEVELOPMENT and PRODUCTION scripts are both initially created via SageMaker. (This is also where model tuning and validation takes place for continual improvement.)

DEV:

1. Data Exploration, Feature Selection (PCA, feature importance, and correlation matrix were used) for Customer Data
2. **Clustering of Customers - we used 4 clusters** (Elbow Method)
   Features used for clustering (PCA and K-Means): rating, birth year, weight, marital status, religion, budget, user cuisine, drink level, and ambience

1. Data Exploration and Feature Selection were used for Restaurant Data
2. Created a **User-Item Matrix** where Cluster # are Row Headers, where columns are PlaceID Headers - rating is field value, and no rating defaults to 0
3. Created **2 Recommendation Models to use in conjunction** for Restaurant Recommendations
   a. Content-Based Recommender - **Features used:** cuisine, days open, alcohol, smoking area, dress code, accessibility, price, ambience, franchise, rating
   b. Collaborative Filtering Recommender - Based on user interactions (ratings) with restaurants, no explicit features, **user-item matrix**

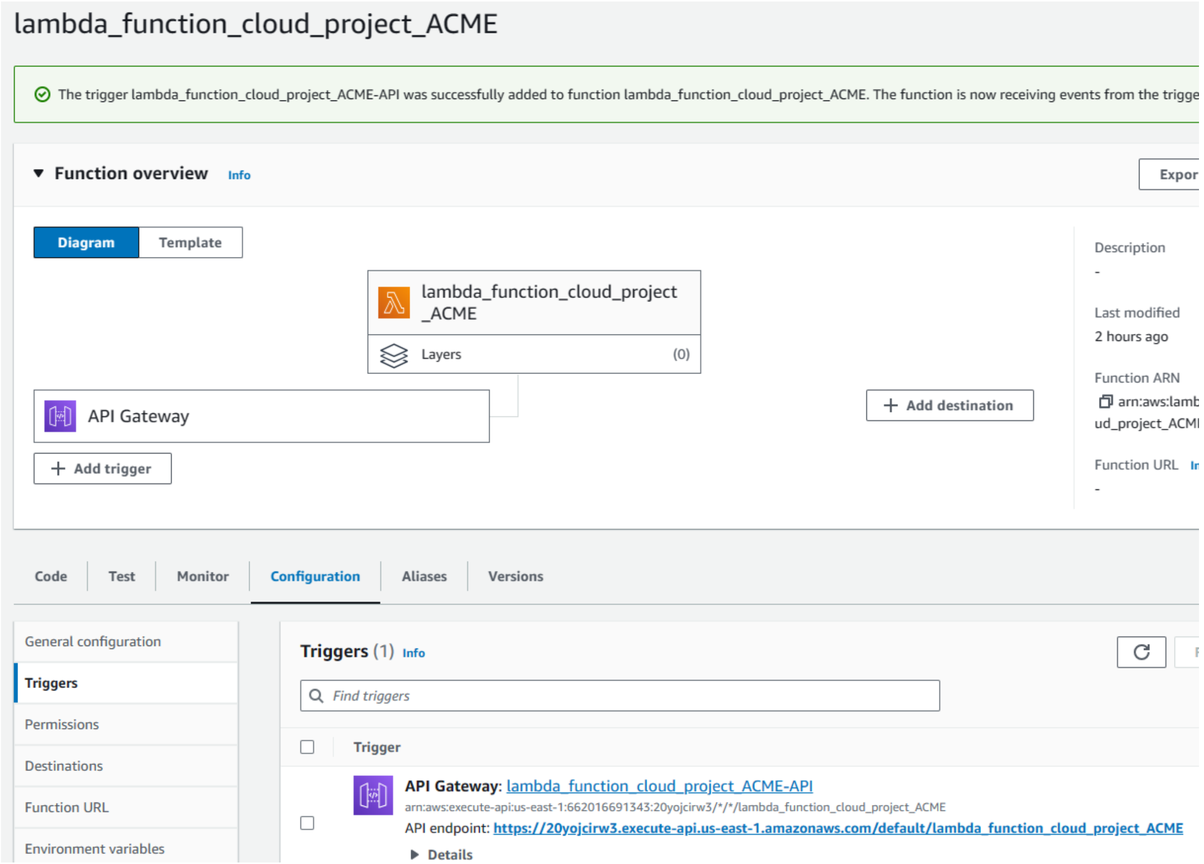
```
# Athena configuration
region_name = 'us-east-1'
database_name = 'restaurantdb'

# Create a connection to Athena using the IAM role associated with your SageMaker notebook instance
conn = connect(s3_staging_dir=f's3://groupproject-cloudcomputing/',
               region_name=region_name)

# SQL query to select data from your Athena table
query = 'SELECT * FROM "restaurantdb"."user_master"'

# Execute the query and fetch data into a DataFrame
df = pd.read_sql(query, conn)
```

Lambda Function for Integration into Operations via Mobile App:

PROD:

Production Environment:

- Development scripts are replicated in production with nonessential elements removed for efficiency
- Python scripts and dependencies are packaged for deployment
- Machine learning models, optimized and compressed, are integrated within **AWS Lambda Function** for seamless execution
- Our **AWS API Gateway** is triggered by user opening their mobile app, utilizing real-time location data to provide filtered restaurant recommendations



Planned Future Enhancements (we couldn't set these up with labrole):

- Implement AWS Cognito User Pool for robust user authentication, managing access and security using JWT tokens
- Configure API Gateway to employ JWT for enhanced endpoint security.
- Establish specific API routes for streamlined data processing
- Set up comprehensive deployment strategies and permission configurations through IAM for better access control and security

Measurement Plan

Predictive metrics

Confusion matrix of recommendation results

	Relevant	Not relevant
Recommended top k	True Positive (TP)	False Positive (FP)
Not recommended	False Negative (FN)	True Negative (TN)

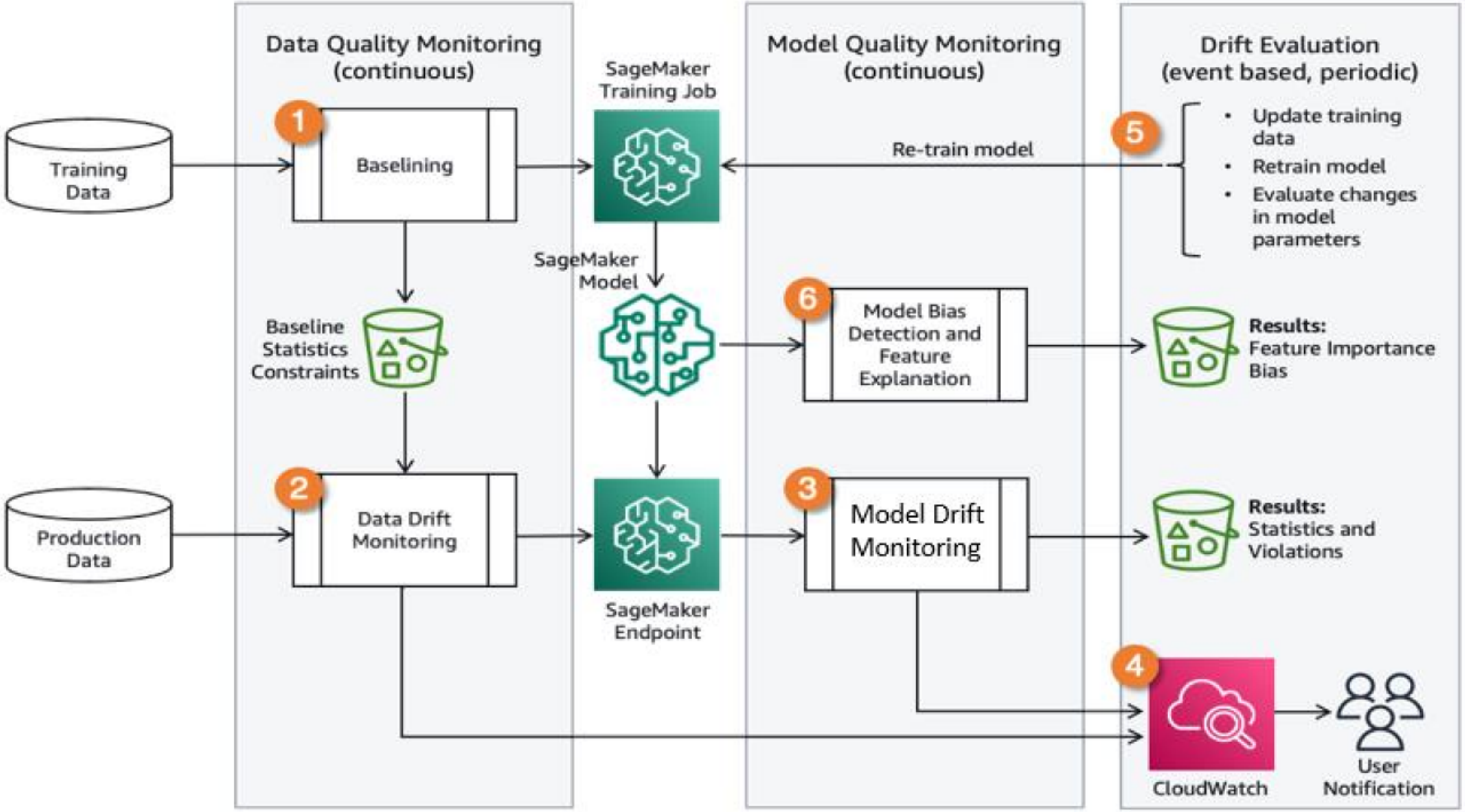
Precision@k
P = (# of top k recommendations that are relevant)/(# of items that are recommended)

Recall@k or HitRatio@k
R = (# of top k recommendations that are relevant)/(# of all relevant items)

Business Metrics

- Click-through rates
- Adoption and conversion

The Concept of Sagemaker Model Monitor Processing



- Using Sagemaker Python SDK by “sagemaker.model” package

Cost Estimation

Service	Cost per Month	Details
S3 Standard and S3 Standard Infrequent Access	$0.52	100000 PUT, COPY, POST, LIST requests to S3 Standard per month
Glue	$1.82	with data catalog and crawler features. 10 DPUs, apache spark ETL job runs for 10 minutes, 10 minutes of interactive session, 100000 objects per month, crawler running for 10 minutes
EventBridge	Free	Standard Event Bus: First 2 million events per month are free. Custom Event Bus: First 500,000 events per month are free
Athena	$0.24	Assuming 50 queries/month and each query scans 1Gb of Data

Cost Estimation

Service	Cost per Month	Details
Quicksight	$19.21	With 2 authors, with twice a month access frequency, and 10 monitoring alerts per month and annual commitment
Sagemaker	$10.19	With 2 data scientist, 1 notebook each, 10 days of usage and 1 hour of runtime per use. 2 bath transforms and training per month, 1Gb of Dedicated SSD and 2 Autopilot jobs per month
Sagemaker (Model Monitoring)	$2.59	With 1 data scientist access, 2 days per month and 1 hour of runtime, remaining same configurations as above
Total	$34.57	Entire Architecture