INSTACART RECOMMENDATION SYSTEM PROJECT

Presented by

Team Data Alchemists

TABLE OF CONTENT

INTRODUCTION

Company Background & Business Problem

- Instacart, an e-commerce business struggles to deliver personalised shopping experience
- Customer retention challenges
- Declining sales and revenue

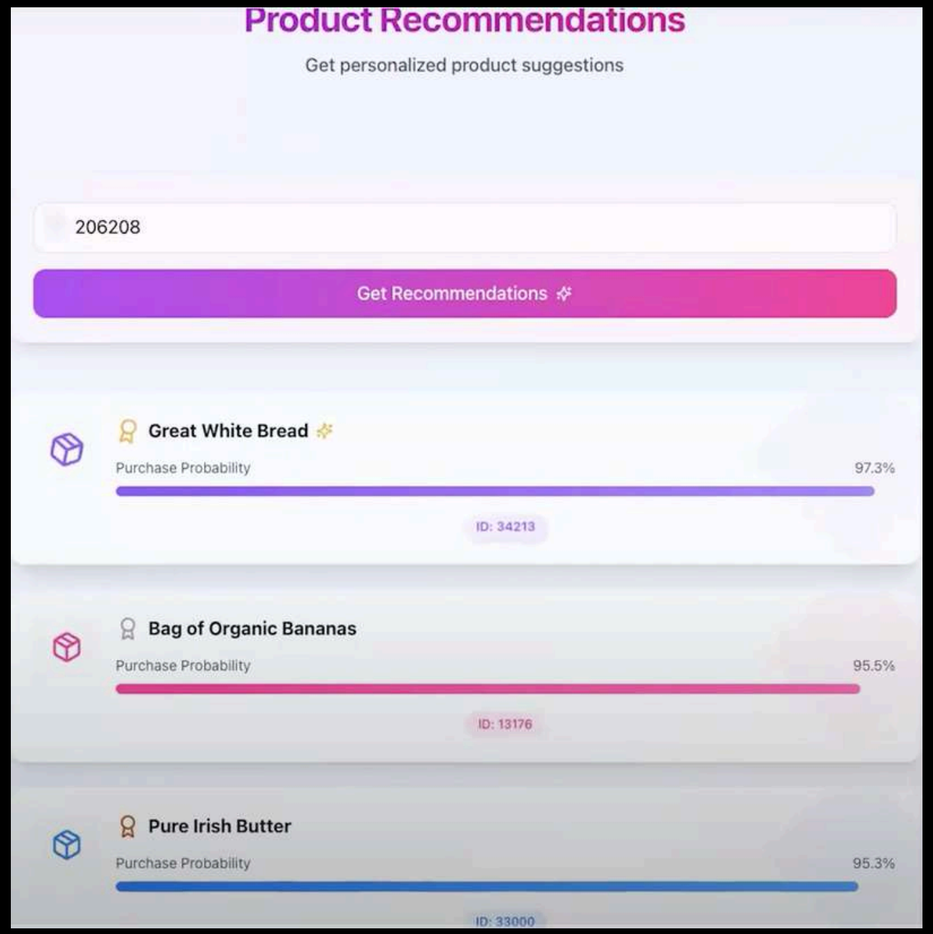
Our Recommendation System was

BORN

PRODUCT HIGHLIGHT

- Leveraged 3 million historical order data to develop a powerful e-commerce recommendation engine
- Built on a scalable AWS cloud infrastructure, accommodating growing user demands
- The system delivers high performance and cost efficiency, benefiting both the company and its customers

DEMO TIME!


Product Recommendations
Get personalized product suggestions
206208
Get Recommendations
Great White Bread
Purchase Probability
97.3%
ID: 34213
Bag of Organic Bananas
Purchase Probability
95.5%
ID: 13176
Pure Irish Butter
Purchase Probability
95.3%
ID: 33000

DEMO TIME!

https://ecomm-recomm-demo.com/

CLOUD ARCHITECTURE DESIGN

- **CRITICAL FACTORS**

HIGH PERFORMANCE

HIGH SCALABILITY

HIGH AVAILABLITY

HIGH MAINTAINABILITY

OBSERVABILITY

DATA BACKUP & RECOVERY

HIGH SECURITY

HIGH RELIABILITY

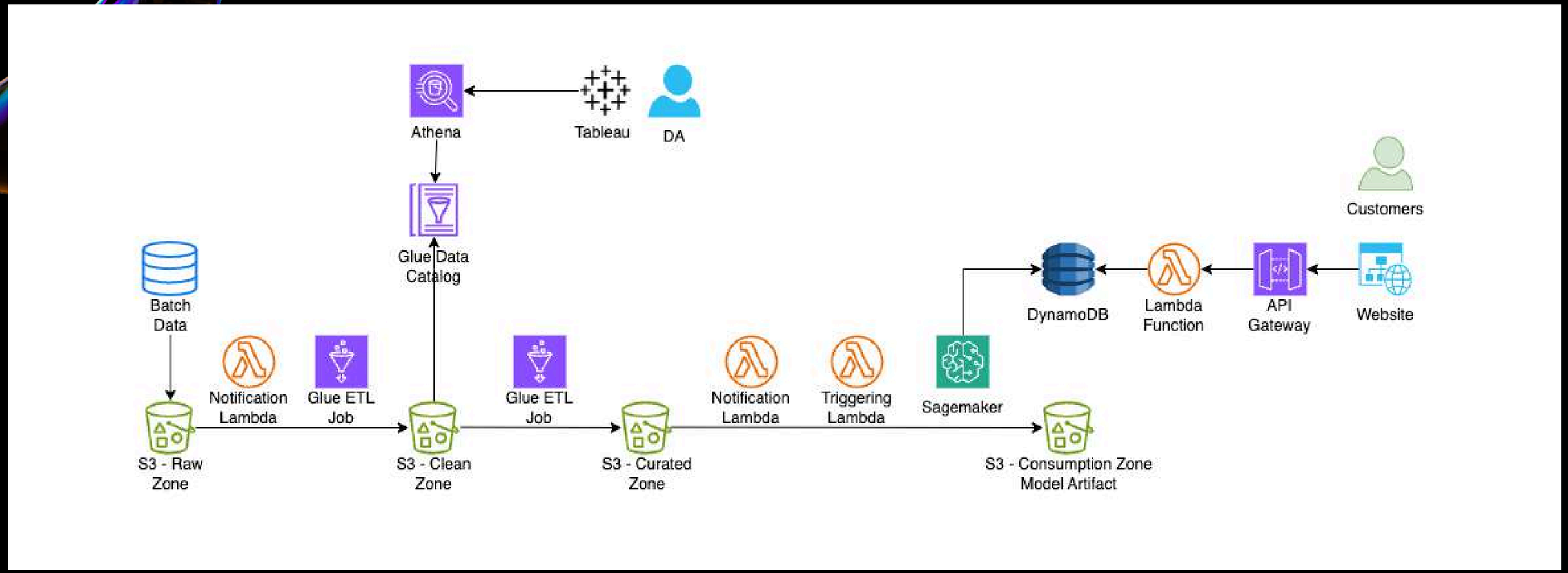
COST EFFICIENCY

AUTOMATION CAPABILITY

COMPLIANCE

GLOBAL REACH

CLOUD ARCHITECTURE DESIGN



CICD: GITHUB ACTIONS

IAC: AWS CLOUDFORMATION

CLOUD ARCHITECTURE DESIGN – DATA LAKE

jrde15-datalake-clean-zone	Asia Pacific (Sydney) ap-southeast-2
jrde15-datalake-consumption-zone	Asia Pacific (Sydney) ap-southeast-2
jrde15-datalake-raw-zone	Asia Pacific (Sydney) ap-southeast-2
jrde15-datalake-curated-zone-bucket	Asia Pacific (Sydney) ap-southeast-2

Bucket Versioning

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use both unintended user actions and application failures. Learn more

Bucket Versioning can't be suspended because Object Lock is enabled for this bucket.

Bucket Versioning
Enabled

Multi-factor authentication (MFA) delete
An additional layer of security that requires multi-factor authentication for changing Bucket Versioning settin
Disabled

Object Lock

Store objects using a write-once-read-many (WORM) model to help you prevent objects from being

Object Lock
Enabled

Default retention
Automatically protect new objects put into this bucket from being deleted or overwritten.
Enabled

Default retention mode
Governance

Default retention period
180 days

Default encryption Info

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption type Info
Server-side encryption with AWS Key Management Service keys (SSE-KMS)

Encryption key ARN
2633d782-97b9-4af6-95d

Bucket Key
When KMS encryption is used to encrypt new objects in this bucket, the bucket key reduces encryption costs by
Enabled

Server access logging

Log requests for access to your bucket. Use CloudWatch to check the health of your server access logging. Learn more

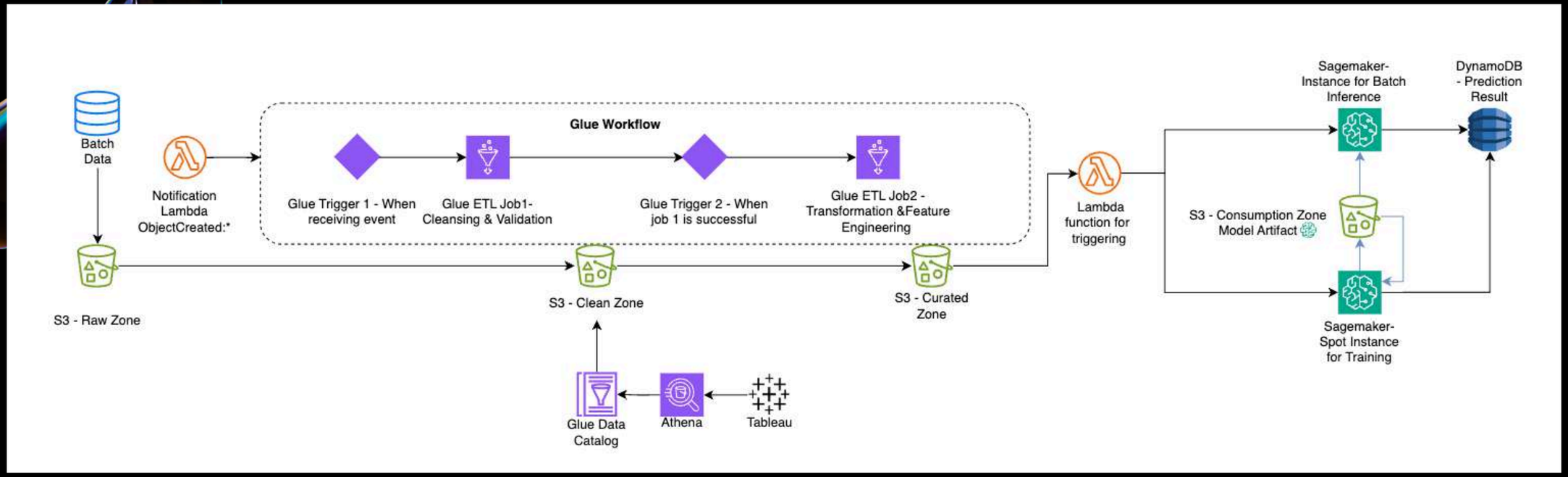
Server access logging
Enabled

Destination bucket
s3://jrde15-shared-logs

Log object key format
raw-zone-logs/[YYYY]-[MM]-[DD]-[hh]-[mm]-[ss]-[UniqueString]

- Security
  - KMS
  - Block ALL Public Access
  - Version Control
  - Object Lock
- Monitoring
  - Server Access Log
- Cost Management
  - Lifecycle Rule

CLOUD ARCHITECTURE DESIGN – DATA PIPELINE

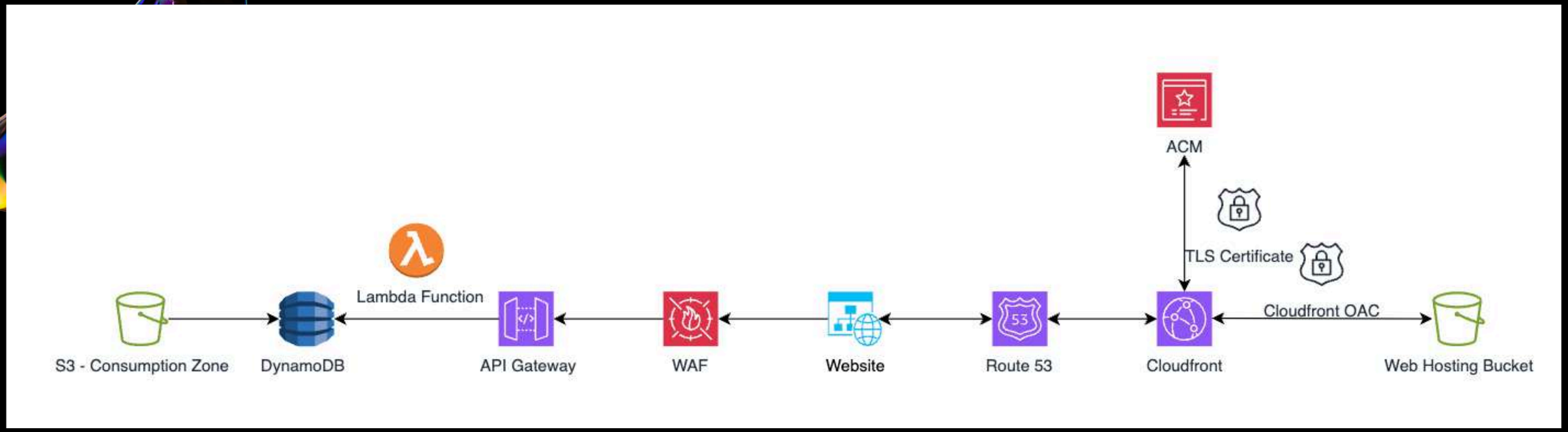


- Security
  - KMS For Glue
  - Private Network Setup for SageMaker
    - VPC
    - Private Subnet
    - Security Groups
    - VPC Endpoint
    - NAT Gateway

- Monitoring
  - CloudWatch
  - CloudTrail

- Cost Management
  - Spot Instance
  - Serverless

CLOUD ARCHITECTURE DESIGN – DATA CONSUMPTION



Security

- WAF
  - IP Rate Limit Block(100/IP/5min)
  - SQL Ingestion Block
- S3
  - Block All Public Access
  - Allows Cloudfront Only
  - HTTPS Only
- Cloudfront
  - ACM Certified
- API Gateway
  - API Key
  - Throttle Limit (Burst/Rate)

Monitoring

- CloudWatch
- CloudTrail

Cost Management

- Pay As You Go
- Lifecycle Rule & Retention for the Logs

CLOUD ARCHITECTURE DESIGN – CICD

JRDE15Project / jrde15-infrastructure

Code | Issues | Pull requests | Actions | Projects | Security | Insights | Settings

Deploy CloudFormation Stack

IaC Update - Adding WAF for FE - v34 #262

Summary

Jobs

deploy

Run details

Usage

Workflow file

Annotations
1 warning

deploy
succeeded 2 days ago in 1m 30s

- Set up job
- Checkout code
- Configure AWS credentials for ap-southeast-2
- Create ap-southeast-2 buckets
- Package and upload Lambda functions
- Upload Glue ETL scripts
- Upload and validate ap-southeast-2 templates
- Deploy main stack
- Monitor CloudFormation Events
- Check deployment status
- Post Configure AWS credentials for ap-southeast-2
- Post Checkout code
- Complete job

JRDE15Project / jrde15-Frontend

Code | Issues | Pull requests | Actions | Projects | Security | Insights | Settings

Deploy Frontend

added product name #7

Summary

Jobs

deploy

Run details

Usage

Workflow file

Annotations
1 warning

deploy
succeeded 2 days ago in 50s

- Set up job
- Checkout code
- Configure AWS credentials
- Get SSM Parameters
- Setup Node.js
- Install dependencies
- Build
- Deploy to S3
- Invalidate CloudFront cache
- Post Setup Node.js
- Post Configure AWS credentials
- Post Checkout code
- Complete job

CLOUD ARCHITECTURE DESIGN – SECURITY

Role name	Trusted entities	Last activity
AdminRole	Account: 841162709701	-
DARole	AWS Service: s3, and 2 more.	-
DERole	AWS Service: s3, and 8 more.	21 minutes ago

Permissions policies (7) Info

You can attach up to 10 managed policies.

Simulate | Remove | Add permissions

Filter by Type: All types

Search

Policy name	Type	Attached entities
AWSGlueServiceRole	AWS managed	3
EventBridgeAccess	Customer inline	0
GlueCloudWatchAccess	Customer inline	0
GlueKMSAccess	Customer inline	0
GlueS3Access	Customer inline	0
GlueSelfAccess	Customer inline	0
GlueTriggerAccess	Customer inline	0

GlueS3Access — Copy JSON | Edit

```
"Version": "2012-10-17",
"Statement": [
    {
        "Action": [
            "s3:GetObject",
            "s3:PutObject",
            "s3:DeleteObject",
            "s3:ListBucket"
        ],
        "Resource": [
            "arn:aws:s3:::jrde15-datalake-raw-zone/*",
            "arn:aws:s3:::jrde15-datalake-raw-zone",
            "arn:aws:s3:::jrde15-datalake-clean-zone/*",
            "arn:aws:s3:::jrde15-datalake-clean-zone",
            "arn:aws:s3:::jrde15-datalake-curated-zone-bucket/*",
            "arn:aws:s3:::jrde15-datalake-curated-zone-bucket",
            "arn:aws:s3:::jrde15-shared-script-bucket/*",
            "arn:aws:s3:::jrde15-shared-script-bucket"
        ],
        "Effect": "Allow"
```

Security

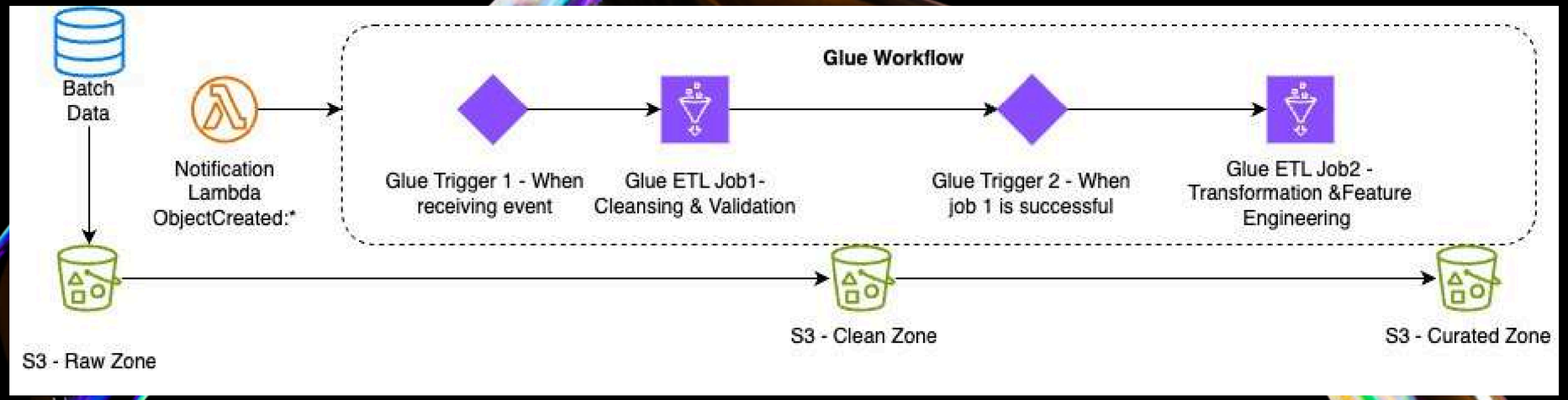
- **KMS**
  - IP Rate Limit Block(100/IP/5min)
  - SQL Ingestion Block
- **S3**
  - Block All Public Access
  - Allows Cloudfront Only
  - HTTPS Only
- **API Gateway**
  - API Key
  - Throttle Limit (Burst/Rate)
- **Cloudfront**
  - ACM Certified
- **WAF**
  - IP Rate Limit Block(100/IP/5min)
  - SQL Ingestion Block
- **VPC**
  - Private Subnet
  - Security Groups
  - VPC Endpoint
  - NAT Gateway
- **IAM**
  - RBAC
  - Least Privileged Rule
  - Avoid Using Root Account

DATA TRANSFORMATION (ETL)

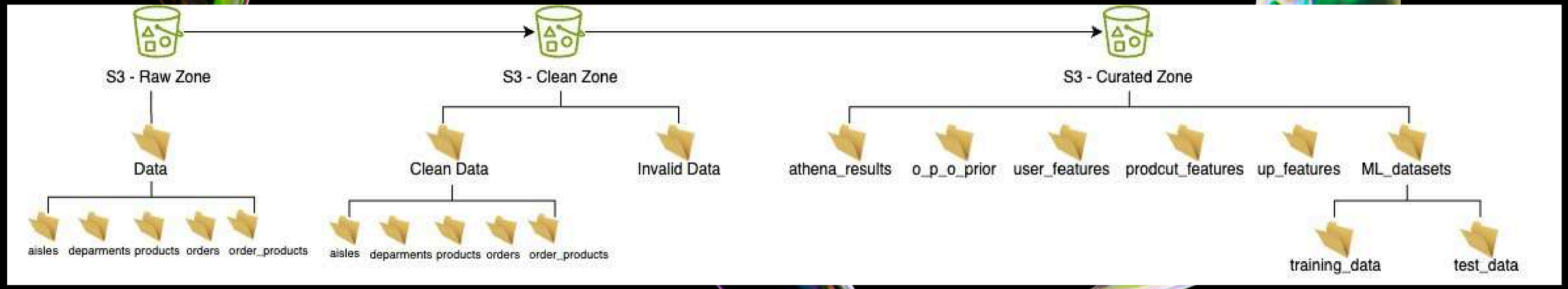
	AWS GLUE	AMAZON EMR
Architecture	Serverless, no infrastructure to provision or manage.	Cluster-based, requires manually setting up and managing cluster.
Frameworks	Primarily uses Apache Spark	Apache Spark, Hadoop, HBase, Presto, Flink, and other big data workloads
Scability	Automatically scales	Manual or automatic scaling of clusters
Cost	Pay-As-You-Go	Cluster-Based Pricing (EC2 instance)
Use Case	Serverless ETL pipelines (On demand), batch workloads	Big data analytics, Process real-time data streams

AWS GLUE ARCHITECTURE

Glue Architecture



S3 Bucket Sturcture



DATA TRANSFORMATION (ETL)

DATA CLEANING

- Missing values (“days_since_prior_order” in the orders table)
- Duplicates
- Normalisation
- Invalid values (Cross table validation)

FEATURE ENGINEERING

- New features (User/product/user-production interaction)
- Correlation analysis
- Converting categorical variables using One Hot Encoding **
- Scaling variables

Data for machine learning model training

- Training data: 8, 474, 661 rows × 25 columns
- Test data: 4, 833, 292 rows × 25 columns

Detailed Doc:
https://drive.google.com/file/d/1GUfyYRNzTNIb4ErvTiJ2TBGpKAhzFQCt/view?usp=sharing

DATA TRANSFORMATION (ETL)

EXPLORATION

- Slowly Changing dimension (SCD Type 2)
- Partition

S3 > Buckets > difan-imba-cleaned > aisles/ > 2025/ > 01/ > 18/

```
cleaned_df = cleaned_df.withColumn("year", lit(year)).withColumn("month", lit(month)).withColumn("day", lit(day))

# Write cleaned data to S3
glueContext.write_dynamic_frame.from_options(
    frame=DynamicFrame.fromDF(cleaned_df, glueContext, f"cleaned_{table}").coalesce(1),
    connection_type="s3",
    connection_options={
        "path": f"s3://{cleaned_bucket}/{table}/",
        "partitionKeys": ["year", "month", "day"]
    },
    format="parquet",
)
```

Hard

MACHINE LEARNING

Localization debugging

- ZERO cost
- Easier setup
- Quick adjustment
- Easier debugging
- Ideal for prototyping and testing models

AWS deployment

- Seamless integration with upstream ETL and downstream frontend
- Supports automation and scalability
- Consistent and reproducible in controlled AWS environments

MACHINE LEARNING

	XGBOOST	GRAPH-BASED
Model type	Decision tree	Graph
Data structure	Tabular data	Graph data
Cold start handling	Week	Strong
Deployment complexity	Easy	Hard
Hardware requirements	Moderate	High

XGBOOST

- **Efficient for Tabular Data**
- **Fast Development and Deployment**
- **Strong Performance for Recommendations**
- **Robustness and Scalability**
- **Feature Importance and Explainability**

MACHINE LEARNING

XGBOOST MODEL PERFORMANCE

Evaluation Metrics

- Precision
- Recall
- F1-score
- Accuracy

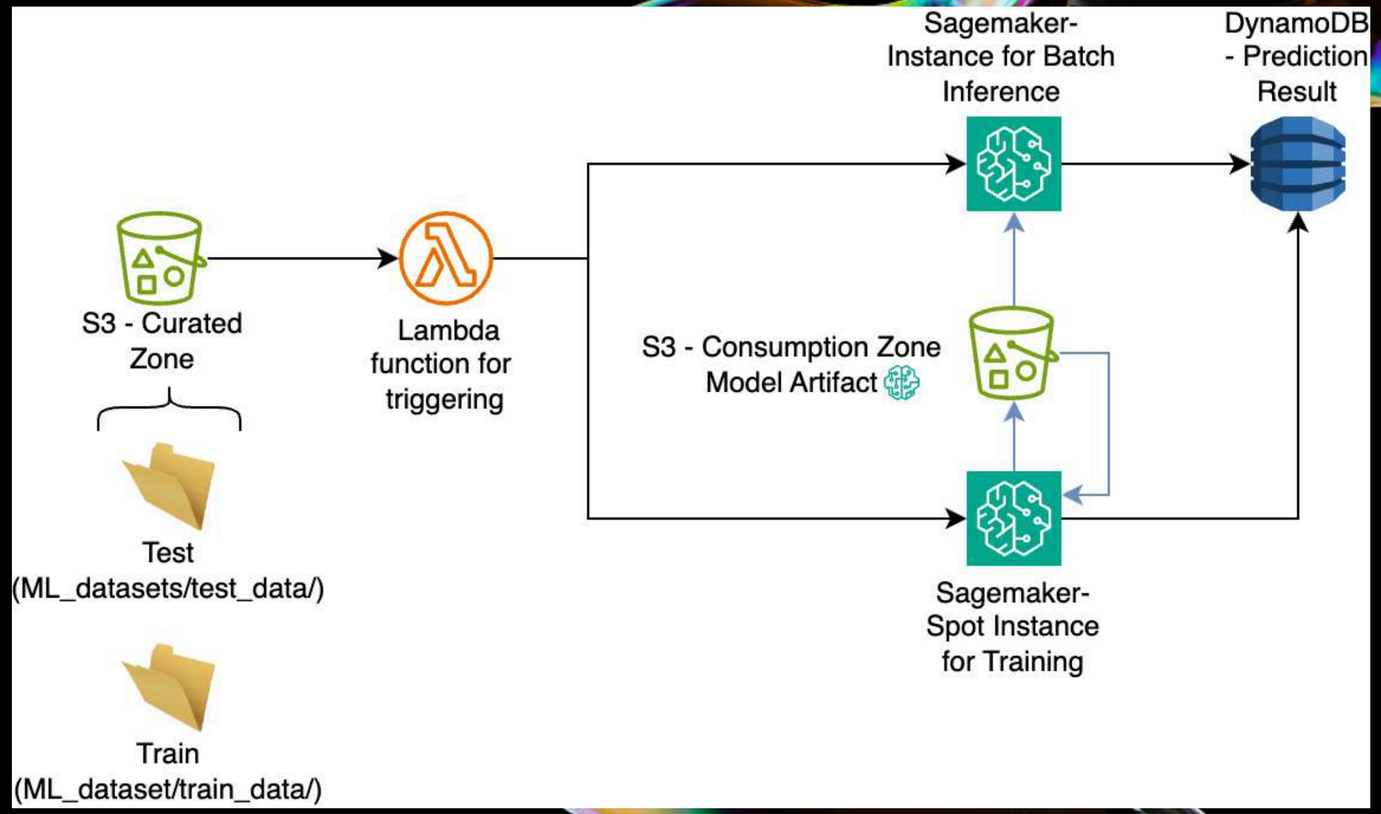
GRAPHSAGE MODEL PERFORMANCE

Evaluation Metrics

- Loss
- Cosine similarity
- R square

MACHINE LEARNING

MODEL TRAINING & INFERENCE ARCHITECTURE

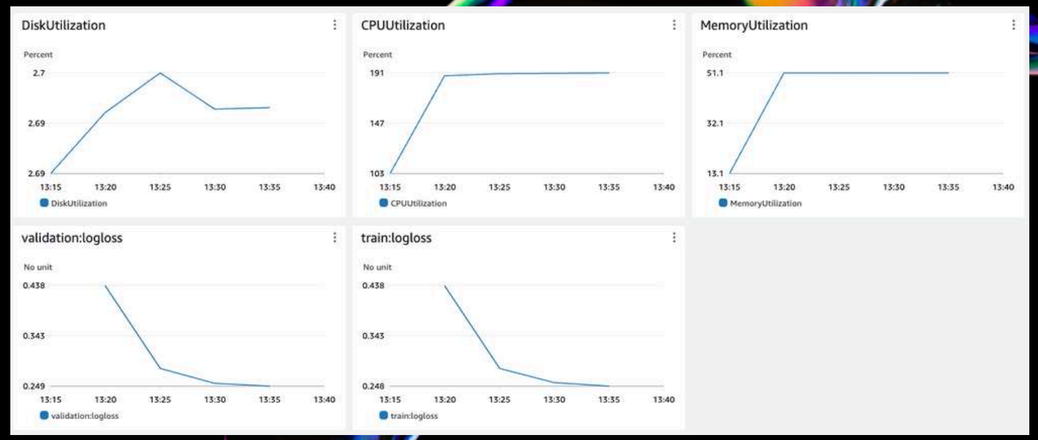


- Use batch transform strategy
- 2 sagemaker instances
- Use DynamoDB for Batch Transform Results

	DynamoDB	S3	RDS
Scalability	Automatic highly scalable	Highly scalable	manual scaling
Latency	Single-digit milliseconds	Higher latency (object store)	Low latency
Data Model	Key-value document	Object storage	Relational
Integration	Seamless with AWS services	Seamless with AWS services	Requires more setup

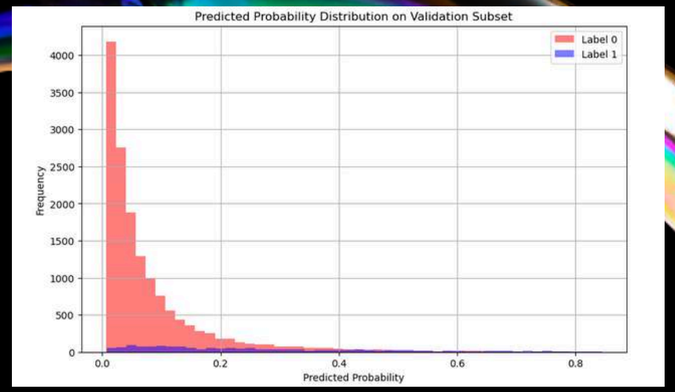
MACHINE LEARNING

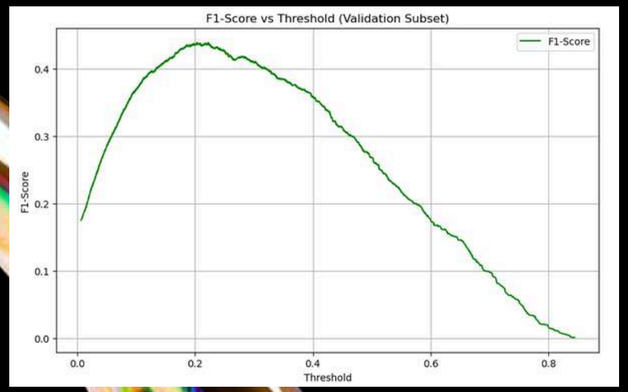
TRAINING MONITOR

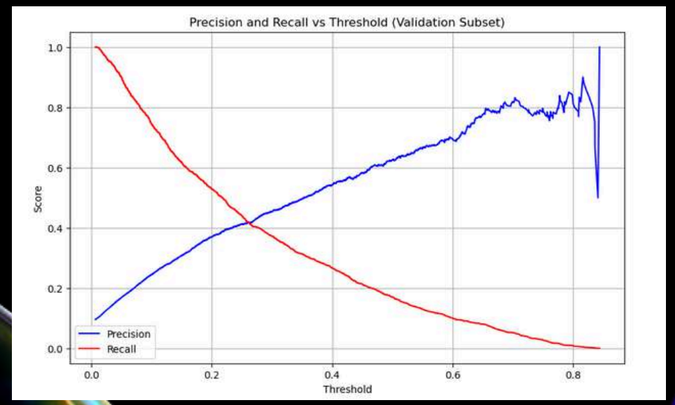



```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.95      0.94     15321
           1       0.45      0.37      0.41      1628

    accuracy                           0.90     16949
   macro avg       0.69      0.66      0.68     16949
weighted avg       0.89      0.90      0.89     16949
```

THRESHOLD ANALYSIS







MACHINE LEARNING

MODELS & BATCH TRANSFORM JOBS

Batch transform jobs

Actions | Create batch transform job

Search batch transform jobs

Name	Status	Duration	Creation time
sagemaker-xgboost-2025-01-22-08-17-27-485	Completed	9 minutes	1/22/2025, 4:17:27 PM
sagemaker-xgboost-2025-01-22-08-04-49-051	Completed	8 minutes	1/22/2025, 4:04:49 PM
sagemaker-xgboost-2025-01-22-05-19-37-608	Completed	15 minutes	1/22/2025, 1:19:37 PM

Models Info

Filter models or endpoints by property or value

Model Name	Endpoints	Last batch transform job	Last batch transform job runtime	Model creation time
xgboost-2025-01-19-08-18-57-033	-		-	1/19/2025, 4:18:57 PM
sagemaker-xgboost-2025-01-21-11-47-22-925	-		-	1/21/2025, 7:47:23 PM
xgboost-model-test	-	batch-inference-job-1737442508	1/21/2025, 2:58:02 PM	1/20/2025, 7:53:27 PM
sagemaker-xgboost-2025-01-21-08-18-00-237	-		-	1/21/2025, 4:18:00 PM
xgboost-2025-01-19-07-21-14-066	-		-	1/19/2025, 3:21:14 PM
sagemaker-xgboost-2025-01-21-15-29-21-766	-	sagemaker-xgboost-2025-01-21-15-46-54-736	1/21/2025, 11:50:11 PM	1/21/2025, 11:29:22 PM
sagemaker-xgboost-2025-01-22-08-04-44-603	-	sagemaker-xgboost-2025-01-22-08-04-49-051	1/22/2025, 4:07:52 PM	1/22/2025, 4:04:45 PM
sagemaker-xgboost-2025-01-22-03-39-37-898	-		-	1/22/2025, 11:39:38 AM
sagemaker-xgboost-2025-01-23-02-06-30-154	sagemaker-xgboost-2025-01-23-02-06-30-154		-	1/23/2025, 10:06:30 AM
sagemaker-xgboost-2025-01-23-02-23-19-236	-	sagemaker-xgboost-2025-01-23-02-23-35-385	1/23/2025, 10:26:52 AM	1/23/2025, 10:23:19 AM

Items returned (50)

Actions | Create item

user_id *(String)*	product_name *(String)*	predicted_probability
129333	The Second Generation Pr...	0.0570520274341106
172329	Spinach Peas & Pear Stage...	0.1082564890384674
91653	Banana	0.1298064887523651
74540	Baby Spinach	0.0699079111218452
167723	French Baguette Bread	0.0129656437784433
37657	Mild Red Pepper Sauce	0.111812949180603
133469	English Seedless Cucumber	0.0269189309328794
141520	Jet-Puffed Marshmallows	0.073224276304245
80559	Organic Cello Lettuce	0.2859939336776733
164894	Organic Whole Strawberries	0.0129374144598841

- Delete endpoint after training
- Inference results including: user id & product name & probability
- Inference results uploaded to DynamoDB to support Frontend usage

FRONT-END

Why DynamoDB

- Automatic scaling
- Cost-effective
- Optimized for quick queries
- Serverless architecture

FRONT-END

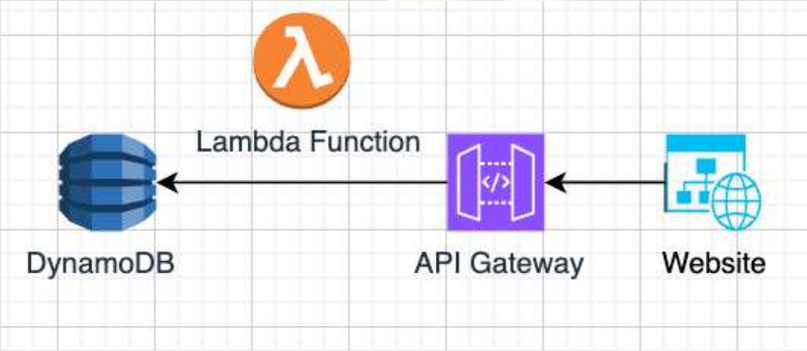
API Gateway Implementation

Two main methods:

- GET: For retrieving personalized recommendations
- OPTIONS: For handling CORS and preflight requests

Lambda function:

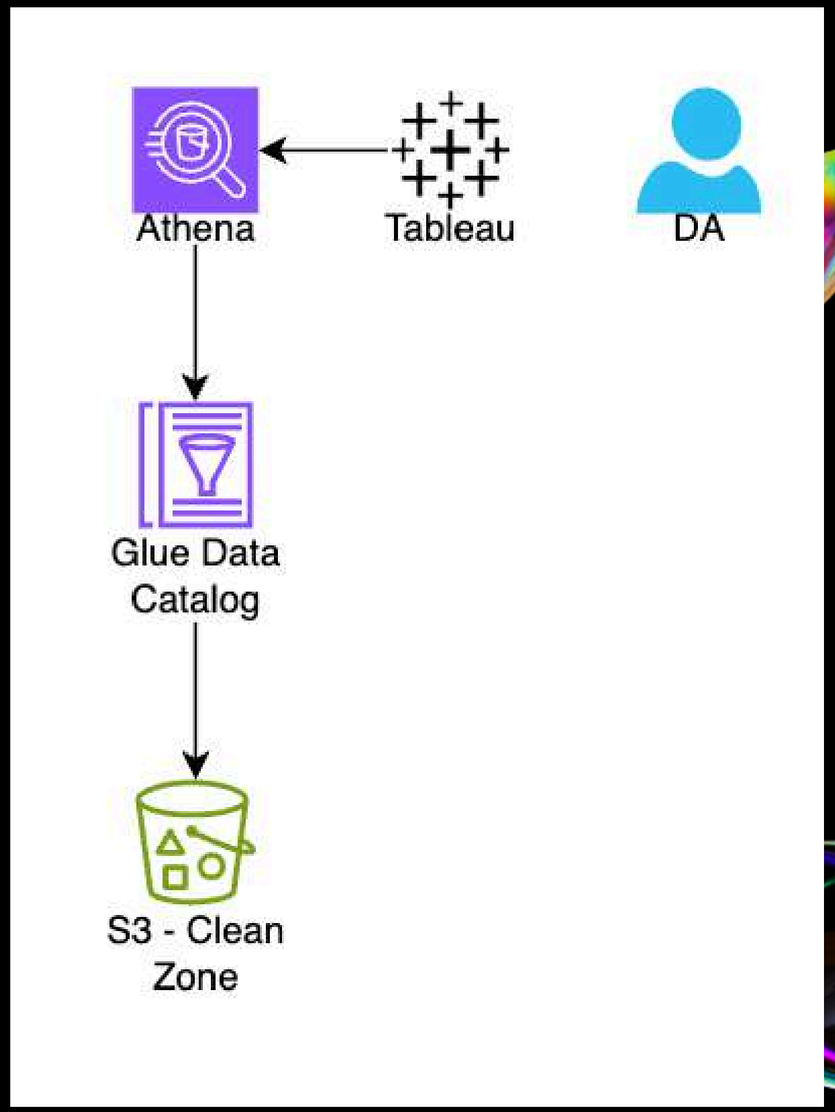
- Processes requests from API Gateway
- Queries DynamoDB
- Calculates top 3 recommendations
- Returns personalized results



Front-end implementation

- Built with React.js for dynamic user interface
- Responsive design with CSS

SECOND USER – DATA ANALYSTS



SANDBOX FOR DATA ANALYSTS

S3 Clean Zone - Data Storage

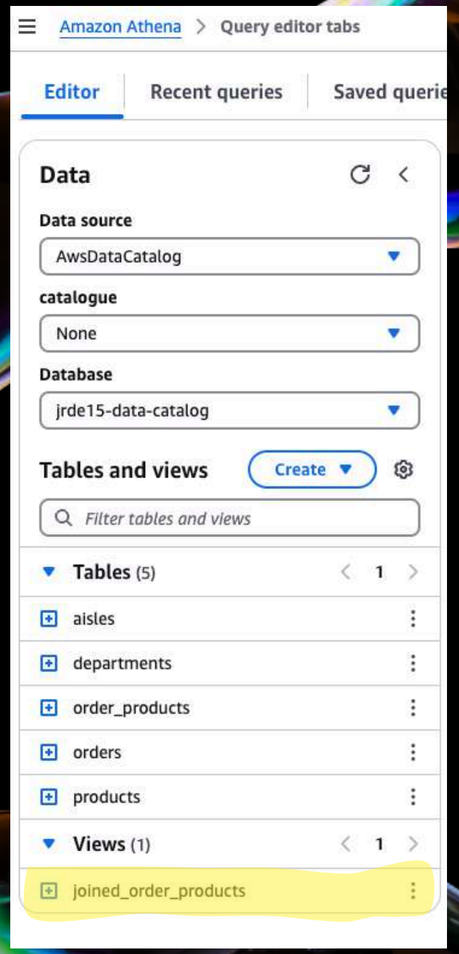
Glue Data Catalog - Metadata Management

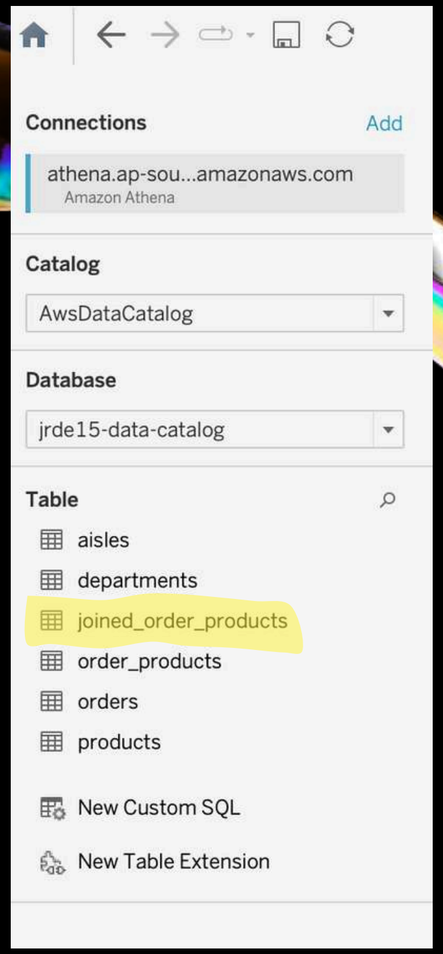
Athena - Query and Create Views

Tableau - Visualization

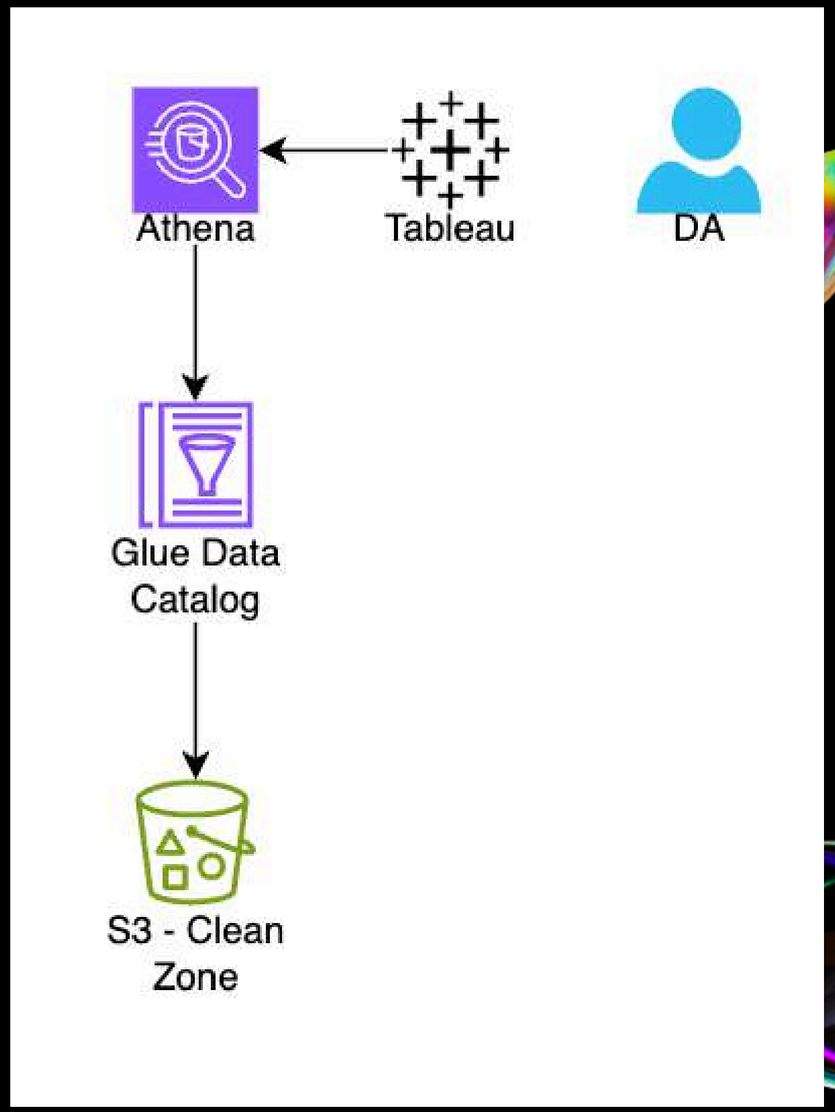
SECOND USER – DATA ANALYSTS

DEMO





SECOND USER – DATA ANALYSTS



Features of Sandbox for Data Analysts

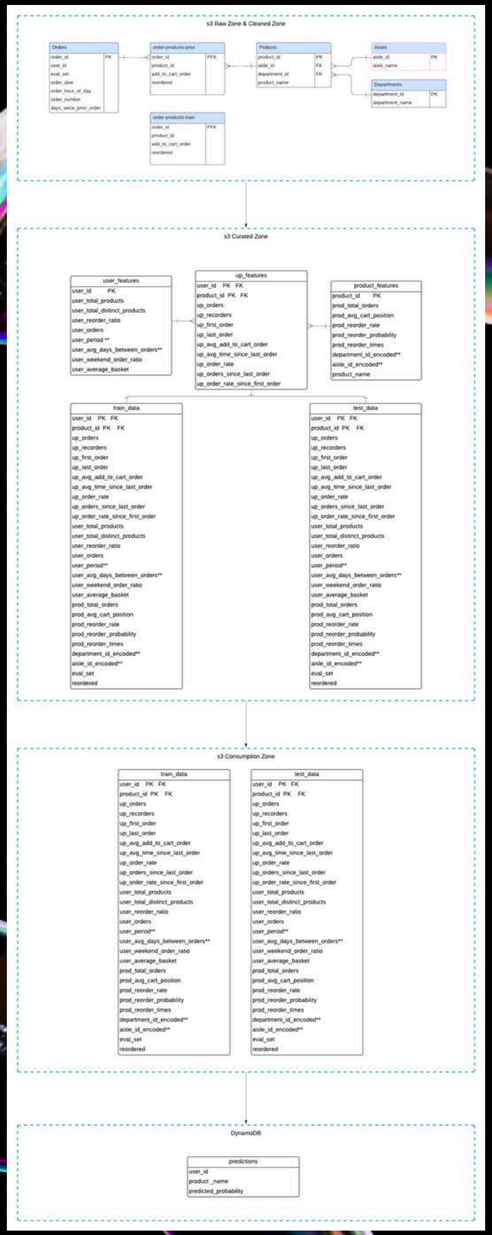
Isolation: Isolate from production system

Flexibility: Freedom to query, transform and analyse / Create view

Accessibility: Easy access to tools and data

Security & Governance: Fine-grained access control and personal identifiable information masked

DATA SCHEMA FLOW-CHART



Storage: Ar

FUTURE IMPROVEMENTS

Infrastructure

- Scalability: Increasing workload and incremental data?
- Fault-tolerance: Error-handling and retries.

IaC

- CI/CD: Separating deployment from script commit in Infrastructure as Code (IaC).
- IaC Modularization: Especially for Glue ETL and Sagemaker.

Security

- Least Privilege Rule: Reinforce the least privilege rule.

Disaster Recovery

- Disaster recovery plan: Back up infrastructure and data in another availability zone.

Simulation of Other Cases

- Data Warehouse as data source in a private subnet.
- Streaming

MEET THE TEAM

KEVIN

Team Leader
DevOps

JIE

Data Engineer
Machine Learning Engineer

JULY

Data Engineer
Project Manager

LEO

Project Supervisor

JEVY

Project Supervisor

TORAIN

Data Engineer
Data Scientist

IVAN

Data Engineer
Data Scientist

ZIBO

Data Engineer
Machine Learning Engineer

SHAWN

Data Engineer
Machine Learning Engineer

THANK YOU
Q&A

Team Data Alchemists