LSDE Coursework Part 1 – AWS Architecture for ArtAI

Rohan Anthony (2704500)

Overview

The task is to design an AWS architecture for "ArtAI", a public web service that allows users to upload images and receive AI-generated variations. The architecture must deliver low-latency global access, secure and private model access, automatic backup and versioning of images, and an automated alert if daily requests exceed 10,000. We focus on achieving high availability, reliability, and scalability by integrating AWS services.

Architecture and AWS Services

This architecture necessitates integrating several AWS services that collectively deliver performance, security, scalability, and cost efficiency. The key services and their roles are listed below, followed by an explanation of how they interact within the architecture.

1. Amazon Route 53 – Latency-based DNS routing for worldwide access.
2. Amazon CloudFront – Global CDN for caching and HTTPS delivery.
3. Amazon S3 – Object storage with versioning, replication and lifecycle management.
4. AWS Lamda – Serverless compute for backend processing.
5. Amazon WAF – Web application firewall in front of CloudFront.
6. Amazon Certificate Manger – Managed TLS certificates.
7. Amazon API Gateway – Secure API entry point for backend logic.
8. Amazon SageMaker Endpoint – Pre-trained model hosting in private VPC.
9. Amazon SQS – Queue to decouple uploads from inference jobs.
10. Amazon DynamoDB – Metadata store with Point in Time Recovery.
11. Amazon CloudWatch – Metrics and alarms.
12. Amazon SNS – Sends email when alarm is triggered.
13. AWS IAM – Access control
14. AWS Key Management Service – Encryption of data at rest and in transit.

Data Backup and Recovery

Monitoring and Alerts

Conclusion