R&D Document: Deploying Gen AI Application on AWS ECS

1. Introduction

Amazon Elastic Container Service (ECS) is a fully managed container orchestration service that makes it easy to run, stop, and manage Docker containers on AWS. This document provides a research-oriented guide for deploying a Generative AI (Gen AI) application on ECS using Dockerization, IAM Roles, and ECS deployment best practices.

2. Dockerization of Gen AI Application

1. Write a Dockerfile for the Gen AI application (Python/Node/Java etc.).
2. Include dependencies such as Hugging Face Transformers, PyTorch, TensorFlow, or LangChain.
3. Build Docker image locally: docker build -t genai-app .
4. Tag and push the Docker image to Amazon Elastic Container Registry (ECR):
  - aws ecr create-repository --repository-name genai-app
  - docker tag genai-app:latest <account_id>.dkr.ecr.<region>.amazonaws.com/genai-app:latest
  - docker push <account_id>.dkr.ecr.<region>.amazonaws.com/genai-app:latest

3. IAM Roles and Permissions

ECS requires IAM roles for execution and task operations:
- ecsTaskExecutionRole: Provides ECS tasks the permissions to pull container images and publish logs to Amazon CloudWatch.
  Policies:
    * AmazonECSTaskExecutionRolePolicy
    * AmazonEC2ContainerRegistryReadOnly
- Application Specific Role: If the Gen AI app uses S3, DynamoDB, or Secrets Manager, additional policies are required.
    * AmazonS3ReadOnlyAccess (or fine-grained bucket access)
    * SecretsManagerReadWrite (for API keys / credentials)
    * CloudWatchLogsFullAccess (for monitoring)

4. ECS Deployment Steps

1. Create an ECS Cluster.
2. Define a Task Definition referencing the Docker image in ECR.

3. Assign the ecsTaskExecutionRole and any custom application IAM Role.
4. Create a Service in ECS with the Task Definition.
5. Configure Auto Scaling (if required).
6. Expose the application through an Application Load Balancer (ALB).

5. Conclusion

By Dockerizing the Gen AI application and deploying it via ECS, AWS provides a scalable, secure, and manageable environment for running AI workloads. IAM roles ensure fine-grained security, while ECS handles orchestration and scaling.