High Availability & Auto Scaling – AWS Architecture Setup

Project Description

This project deploys a scalable and highly available application architecture on AWS.
The goal is to ensure that backend workloads remain secure in private subnets, while still handling public web traffic through an internet-facing Application Load Balancer (ALB).
Traffic is routed through the ALB and distributed to EC2 instances launched by an Auto Scaling Group (ASG), which maintains availability across multiple Availability Zones.
This setup prevents direct access to backend instances, supports automatic scaling, and ensures continued service during traffic spikes or instance failures.

Architecture Overview

- **Application Load Balancer (ALB) – Public entry point for incoming traffic**
- **Auto Scaling Group (ASG) – Automatically launches EC2 instances**
- **EC2 Instances in Private Subnets – Backend workload; not publicly reachable**
- **2 Public Subnets – Host ALB, NAT access**
- **2 Private Subnets – Host compute layer for application instances**
- **NAT Gateway – Allows private EC2 instances outbound access without exposing them**
- **Target Group – Connects ALB to backend EC2 instances**
- **Separate Route Tables – Ensure correct routing and subnet isolation**