AWS Architecture Documentation: Scalable Web Application with ALB and Auto Scaling

Created by: Marwan Tarik Madbouly

Table of Contents

1. Overview

This document describes the AWS architecture for a scalable web application utilizing an Application Load Balancer (ALB) and Auto Scaling. The architecture ensures high availability, fault tolerance, and security while efficiently managing traffic distribution.

2. Architecture Components

2.1 Virtual Private Cloud (VPC)

A VPC (10.16.0.0/16) is created to logically isolate resources and control networking configurations.

2.2 Subnet Configuration

The architecture is deployed across multiple Availability Zones (AZs) for high availability. Three main subnet types are used:

- **Reserved Subnet**: Allocated for future expansion or specialized workloads.
- **Database (DB) Subnet**: Hosts relational databases, ensuring restricted access from
- **application services**: Services Subnet: Contains core application services and business logic.
- **Web Subnet**: Contains frontend-facing instances handling user traffic.

2.3 Load Balancing and Auto Scaling

- **Application Load Balancer (ALB)**
  - Distributes incoming traffic across multiple instances in different AZs.
  - Ensures fault tolerance and redundancy.
- **Auto Scaling Groups**
  - Ensures dynamic scaling of web and service instances based on traffic demand.
  - Configured with health checks to automatically replace unhealthy instances.

2.4 Networking & Security

- **NAT Gateway**
  - Placed in public subnets to allow private instances (DB and services) to access external resources while blocking incoming traffic.
- **Internet Gateway**
  - Provides internet access to public-facing web instances.
- **Security Groups and Network ACLs**
  - Security groups restrict access to resources based on protocols and IP ranges.
  - Network ACLs add an extra layer of control at the subnet level.

2.5 AWS Account Integration

Monitoring and Alerts

- Integrated with AWS CloudWatch for performance monitoring.
- Configured with AWS SNS to send alerts to teams via Slack.
- Logging enabled for auditing and debugging.

3. Workflow

1. Users access the web application through the ALB, which routes traffic to the available web instances.
2. The ALB forwards requests to backend services hosted in the services subnet.
3. If backend services require data, they communicate with the database located in the DB subnet.
4. NAT Gateway allows backend services to interact with external APIs securely.
5. Auto Scaling dynamically adjusts the number of instances to handle traffic efficiently.

4. Benefits of the Architecture

- High Availability: Multi-AZ deployment ensures continuous operation.
- Scalability: Auto Scaling adapts to workload demands.
- Security: Isolated subnets, restricted access, and AWS security controls.

- Performance Optimization: ALB efficiently distributes traffic.
- Cost Efficiency: Scales resources based on demand, reducing costs.

5. Diagram