Module 9️⃣: Cloud Architecture and Module 🔟 : Auto Scaling and Monitoring

Module 9: Cloud Architecture

1. AWS Well-Architected Framework Design Principles

The AWS Well-Architected Framework provides a consistent approach to evaluating and improving the architecture of your cloud-based applications. It is based on five key pillars: Operational Excellence, Security, Reliability, Performance Efficiency, and Cost Optimization. Each pillar addresses different aspects of cloud architecture, guiding you in making informed decisions to build secure, high-performing, resilient, and efficient infrastructures for your applications. The framework is designed to help identify potential risks and implement best practices to mitigate them, ensuring your architecture aligns with your business objectives and customer needs.

Key Points:

- Five pillars: Operational Excellence, Security, Reliability, Performance Efficiency, Cost Optimization.
- Consistent approach to evaluating cloud architecture.
- Helps identify risks and best practices.
- Aligns architecture with business objectives.

2. Operational Excellence

Operational Excellence focuses on the ability to run and monitor systems effectively to deliver business value and continuously improve processes. It involves automating changes, responding to events, and defining standards to manage daily operations. By treating operations as code, making frequent, small, and reversible changes, and refining operational procedures regularly,

organizations can anticipate and learn from failures, ensuring continuous improvement and resilience in their operations.

Key Points:

- **Automate**: Implement operations as code to reduce human error.
- **Frequent Changes**: Make small, reversible changes to minimize risk.
- **Refinement**: Continuously improve operational procedures.
- **Anticipate Failures**: Test and prepare for potential failures.
- **Learning**: Learn from all operational events and failures.

3. Security

The Security pillar emphasizes protecting data, systems, and assets through risk assessments and mitigation strategies. It includes principles like implementing strong identity and access management, enabling traceability, automating security best practices, and protecting data at rest and in transit. By using a defense-in-depth approach, organizations can secure their applications and infrastructure from internal and external threats while meeting compliance requirements.

Key Points:

- **Identity and Access Management**: Implement strong controls.
- **Traceability**: Enable logging and monitoring for security events.
- **Automation**: Automate security best practices to ensure consistency.
- **Data Protection**: Encrypt data both at rest and in transit.
- **Defense in Depth**: Use multiple layers of security to protect assets.

4. Reliability

Reliability in the AWS Well-Architected Framework ensures that workloads perform their intended functions correctly and consistently. This pillar involves building architectures that automatically recover from failure, scaling based on demand, and monitoring systems to detect and address issues before they impact operations. By designing for failure and using automated recovery processes, organizations can maintain system availability and resilience.

Key Points:

- **Automatic Recovery**: Design architectures that recover from failures.
- **Scaling**: Implement automatic scaling to meet demand.
- **Monitoring**: Use monitoring tools to detect and address issues early.
- **Design for Failure**: Anticipate and plan for failures to maintain availability.

5. Performance Efficiency

Performance Efficiency is about using IT and computing resources efficiently to meet system requirements and maintain responsiveness under varying load conditions. This pillar focuses on selecting the right resource types and sizes, monitoring performance, and making informed decisions based on that data. By regularly reviewing and optimizing architecture, organizations can ensure they are using resources efficiently and effectively.

Key Points:

- **Resource Selection**: Choose the right resource types and sizes for workloads.
- **Performance Monitoring**: Continuously monitor and optimize performance.
- **Review**: Regularly review and optimize architecture.
- **Efficiency**: Ensure resources are used effectively to meet system demands.

6. Cost Optimization

Cost Optimization involves strategies to avoid unnecessary costs and make the most out of your investment in cloud infrastructure. This pillar includes understanding and controlling where money is spent, selecting the most appropriate services, and scaling to meet business needs without overspending. By implementing practices like using reserved instances, monitoring expenditure, and optimizing resources, organizations can maximize their cloud investment.

Key Points:

- **Cost Control**: Understand and control cloud spending.
- **Service Selection**: Choose cost-effective services and resources.
- **Scaling**: Scale efficiently to avoid overspending.

- **Reserved Instances**: Use reserved instances for predictable workloads.

7. Reliability

Reliability ensures that your application can recover from failures, scale with demand, and provide a consistent experience to users. This pillar involves designing systems with fault tolerance, automating recovery processes, and regularly testing for failures. By doing so, organizations can maintain high availability and minimize downtime, ensuring that services remain reliable and trustworthy.

Key Points:

- **Fault Tolerance**: Design systems that can withstand failures.
- **Automated Recovery**: Implement automated processes to recover from failures.
- **High Availability**: Ensure services are available when needed.
- **Regular Testing**: Test systems regularly to ensure reliability.

8. AWS Trusted Advisor

AWS Trusted Advisor is an online resource that provides real-time guidance to help you provision your resources according to AWS best practices. It offers recommendations across various areas, including cost optimization, security, fault tolerance, performance, and service limits. By using Trusted Advisor, organizations can ensure that their AWS environment is running efficiently and securely, making necessary adjustments to improve their architecture.

Key Points:

- **Real-Time Guidance**: Provides real-time advice on AWS best practices.
- **Recommendations**: Offers suggestions for cost optimization, security, and performance.
- **Service Limits**: Helps manage and monitor service limits.
- **Efficiency**: Ensures AWS resources are used effectively and securely.

Module 10 : Auto Scaling and Monitoring

This module covers three key AWS services: **Elastic Load Balancing**, **Amazon CloudWatch**, and **Amazon EC2 Auto Scaling**. It includes hands-on activities and a lab to demonstrate how these services work together to create a scalable and resilient architecture.

Module Objectives

After completing this module, you should be able to:

- **Distribute traffic** across Amazon EC2 instances using Elastic Load Balancing.
- **Monitor AWS resources** and applications in real-time with Amazon CloudWatch.
- **Launch and release servers** in response to workload changes with Amazon EC2 Auto Scaling.
- Perform **scaling and load balancing tasks** to improve an architecture.

Section 1: Elastic Load Balancing

Introduction to Elastic Load Balancing

- **Elastic Load Balancing (ELB)** is a service that distributes incoming application or network traffic across multiple targets, such as EC2 instances, containers, IP addresses, and Lambda functions, within a single Availability Zone or across multiple Availability Zones.
- **Scalability**: ELB scales the load balancer automatically as traffic increases or decreases, ensuring that your application can handle varying levels of traffic without manual intervention.

Types of Load Balancers

Elastic Load Balancing offers three types of load balancers, each designed for specific use cases:

1. **Application Load Balancer (ALB)**
    - Operates at the **application level** (OSI model layer 7).
    - Routes traffic based on the **content of the request**.
    - Ideal for **HTTP and HTTPS** traffic.

- Supports advanced request routing, beneficial for **microservices** and **container-based applications**.
- Enhances security by ensuring that the latest SSL/TLS ciphers and protocols are used.

2. **Network Load Balancer (NLB)**
   - Operates at the **network transport level** (OSI model layer 4).
   - Routes connections based on **IP protocol data**.
   - Designed for handling **TCP and UDP** traffic.
   - Capable of handling **millions of requests per second** with ultra-low latency.
   - Optimized for **volatile network traffic patterns**.
3. **Classic Load Balancer (CLB)**
   - Provides basic load balancing across multiple EC2 instances.
   - Operates at both the **application level and network transport level**.
   - Supports **HTTP, HTTPS, TCP, and SSL** protocols.
   - **Older implementation**; AWS recommends using ALB or NLB when possible.

Use Cases for Elastic Load Balancing

- **High Traffic Websites**: ELB is essential for modern high-traffic websites that must serve hundreds of thousands to millions of concurrent requests, ensuring that resources are used efficiently and reliably.
- **Scalable Applications**: ELB automatically adjusts to traffic changes, making it suitable for applications with unpredictable or varying levels of traffic.

Section 2: Amazon CloudWatch

Introduction to Amazon CloudWatch

- **Amazon CloudWatch** is a monitoring service for AWS resources and applications.

- It provides **real-time monitoring** of resource utilization, application performance, and operational health.
- **CloudWatch Logs**: Enables the collection, monitoring, and storage of log files from AWS resources.

Key Features

- **Metrics**: CloudWatch collects and tracks metrics, which are variables that are measured over time.
- **Alarms**: You can set alarms based on metrics to take automated actions or send notifications when thresholds are exceeded.
- **Logs**: CloudWatch Logs allow you to monitor, store, and access log files from various AWS resources.

Section 3: Amazon EC2 Auto Scaling

Introduction to Amazon EC2 Auto Scaling

- **Amazon EC2 Auto Scaling** ensures that you have the right number of EC2 instances available to handle your application's load.
- It automatically adjusts the number of EC2 instances in response to changes in demand, ensuring consistent performance at the lowest possible cost.

Key Concepts

- **Scaling Policies**: Define the conditions under which Auto Scaling will add or remove instances.
- **Auto Scaling Group**: A collection of EC2 instances treated as a logical grouping for the purposes of scaling and management.
- **Dynamic Scaling**: Automatically adjusts capacity based on demand patterns.
- **Predictive Scaling**: Uses machine learning to forecast future traffic and scale capacity accordingly.

Use Cases

- **Handling Traffic Spikes**: Automatically scales out during peak demand periods and scales in when demand drops.
- **Cost Optimization**: Ensures that you are only using the resources you need, reducing costs associated with over-provisioning.

Key Points Not Covered in Detail

- **Elastic Load Balancing**:
  - **Health Checks**: ELB performs health checks on registered targets and routes traffic only to healthy instances.
  - **Cross-Zone Load Balancing**: ELB can distribute traffic evenly across all registered instances in all enabled Availability Zones.
- **Amazon CloudWatch**:
  - **Dashboards**: CloudWatch provides customizable dashboards for visualizing metrics and logs in a single view.
  - **Events**: CloudWatch Events deliver a near real-time stream of system events that describe changes in AWS resources.
- **Amazon EC2 Auto Scaling**:
  - **Cooldown Period**: A configurable setting that helps ensure Auto Scaling doesn't launch or terminate additional instances before the previous scaling activity has had time to take effect.
  - **Termination Policies**: Determine which instances to terminate first during scale-in events.

These notes and points provide a comprehensive overview of the content from the AWS Academy Cloud Foundations Module 10, focusing on Auto Scaling and Monitoring.

Thanks

2081/05/14

Course Completed.

The END