Module 6 - Compute

Slides

Objectives / Topics

- Provide an overview of different AWS compute services in the cloud
- Demonstrate why to use Amazon Elastic Compute Cloud (Amazon EC2)
- Identify the functionality in the EC2 consoleâ€¢Perform basic functions in Amazon EC2 to build a virtual computing environment
- Identify Amazon EC2 cost optimization elements
- Demonstrate when to use AWS Elastic Beanstalk
- Demonstrate when to use AWS Lambda
- Identify how to run containerized applications in a cluster of managed servers

Labs / Activities

- Knowledge Check
- Lab: Introduction to Amazon EC2 --- Lab Instructions
- Lab: AWS Lambda --- Lab Instructions
- Lab: AWS Elastic Beanstalk --- Lab Instructions

Section 1: Compute Services Overview

AWS offers many compute services

- EC2
- EC2 Auto Scaling
- Elastic Container Registry (ECR)
- Elastic Container Service
- VMware Cloud on AWS
- Elastic Beanstalk
- Lambda
- Elastic Kubernetes Service (EKS)
- Lightsail
- Batch
- Fargate
- Outposts
- Serverless Application Repository

The optimal compute service or services that you use will depend on your use case, some aspects to consider:

- What is your application design?
- What are your usage patterns?
- Which configuration settings will you want to manage?

Section 2: Amazon Elastic Compute Cloud (EC2)

Overview

- Provides virtual machines â€” referred to as EC2 instances â€” in the cloud
- Gives you full control over the guest operating system (Windows or Linux) on each instance
- You can launch instances of any size into an Availability Zone anywhere in the world with just a few clicks or a line of code, and they are ready in minutes
- Resizable compute capacity

- You can launch instances from Amazon Machine Images (AMIs)
- You can control traffic to and from instances
- Provides tools to build failure resilient applications and isolate them from common failure scenarios

Launching an EC2 Instance

These are the nine key decisions to make when you create an EC2 instance by using the AWS Management Console Launch Instance Wizard.

1. Select Amazon Machine Image (AMI)
2. Is a template that is used to create an EC2 instance (a virtual machine)
3. Contains a Windows or Linux operating system and often has some software pre-installed
4. AMI choices:
    - Quick Start â€“ Linux and Windows AMIs that are provided by AWS
    - My AMIs â€“ Any AMIs that you created
    - AWS Marketplace â€“ Pre-configured templates from third parties
    - Community AMIs â€“ AMIs shared by others; use at your own risk
5. Select an Instance Type
6. Optimized to fit different use cases
7. The instance type that you choose determines
    - Memory (RAM)
    - Processing power (CPU)
    - Disk space and disk type (Storage)
    - Network performance
8. Instance type categories
    - General purpose
    - Compute optimized
    - Memory optimized
    - Storage optimized
    - Accelerated computing
9. Instance types offer family, generation, and size (Example - `t3.large`: Family - `T`, Generation - `3`, Size - `large`)
10. Networking Features
    - The network bandwidth (Gbps) varies by instance type.
    - To maximize networking and bandwidth performance of your instance type enable enhanced networking and if you have interdependent instances, launch them into a cluster placement group.
    - Enhanced networking types are supported on most instance types. Enhanced networking types:
    - Elastic Network Adapter (ENA): Supports network speeds of up to 100 Gbps.
    - Intel 82599 Virtual Function interface: Supports network speeds of up to 10 Gbps.
11. Specify Network Settings
12. Where should the instance be deployed? Identify the VPC and optionally the subnet
13. Should a public IP addressbe automatically assigned?
14. You can have multiple networks, such as different ones for development, testing and production
15. Attach IAM Role (optional)
16. Will software on the EC2 instance need to interact with other AWS services? If yes, attach an appropriate IAM Role. IAM Roles can be attached at any time, not just launch.
17. An AWS Identity and Access Management (IAM) role that is attached to an EC2 instance is kept in an instance profile.
18. User Data Script (optional)
19. Specify a user data script at instance launch. Use user data scripts to customize the runtime environment of your instance
20. Script executes the first time the instance starts
21. Specify Storage
22. Configure the root volume
23. Attach additional storage volumes(optional). For each volume, specify:
    - Disk Size (in GB)
    - Volume type (SSDs or HDDs)
    - If the volume will be deleted when the instance is terminated
    - If encryption should be used
24. Storage Options:
    - Amazon Elastic Block Store (Amazon EBS) â€“
    - Durable, block-level storage volumes.
    - You can stop the instance and start it again, and the data will still be there.
    - Amazon EC2 Instance Store
    - Storage is provided on disks that are attached to the host computer where the EC2 instance is running.

    - If the instance stops, data stored here is deleted.
    - Other options for storage (not for the root volume)
    - Mount an Amazon Elastic File System (EFS) file system.
    - Connect to Amazon Simple Storage Service (S3).
25. Add Tags
26. Consists of a key and an optional value.
27. Tagging is how you can attach metadata to an EC2 instance
28. Potential benefits: Filtering, automation, cost allocation, and access control.
29. Security Group Settings
30. A security group is a set of firewall rules that control traffic to the instance
31. When you launch an instance, you associate one or more security groups with it
32. Create rules that specify the source, which ports that network communications can use and the protocol (TPC, UDP, ICMP)
33. Modify the rules for a security group at any time; the new rules are automatically applied to all instances that are associated with the security group
34. Identify or Create the Key Pair
35. At instance launch, you specify an existing key pair or create a new key pair
36. A key pair consists of a public key that AWS stores and a private key file that you store
37. It enables secure connections to the instance
38. For Windows AMIs â€“ Use the private key to obtain the administrator password that you need to log in to your instance
39. For Linux AMIs â€“ Use the private key to use SSH to securely connect to your instance

Miscellaneous

Consider using an Elastic IP Address if you require a persistent public IP address.

Instance metadata can be viewed in browser or a terminal window, and can be used to configure or manage a running instance

Amazon CloudWatch can be used to monitor an EC2 instance to provide near-real-time metrics, charts, and 15 months of historical data. CloudWatch has basic monitoring (no additional cost) or detailed monitoring.

Lab: Introduction to Amazon EC2 --- Lab Instructions

Section 3: Amazon EC2 Cost Optimization

Amazon Pricing Models

On-Demand Instances

- Pay by the hour
- No long-term commitments.
- Eligible for the AWS Free Tier.

Dedicated Hosts

- A physical server with EC2 instance capacity fully dedicated to your use.

Dedicated Instances

- Instances that run in a VPC on hardware that is dedicated to a single customer.

Spot Instances

- Instances run as long as they are available and your bid is above the Spot Instance price.
- They can be interrupted by AWS with a 2-minute notification
- Interruption options include terminated, stopped or hibernated
- Prices can be significantly less expensive compared to On-Demand Instances
- Good choice when you have flexibility in when your applications can run.

Reserved Instances

- Full, partial, or no upfront payment for instance you reserve
- Discount on hourly charge for that instance
- 1-year or 3-year term

Scheduled Reserved Instances

- Purchase a capacity reservation that is always available on a recurring schedule you specify
- 1-year term

Per second billing available for On-Demand Instances, Reserved Instances, and Spot Instances that run Amazon Linux or Ubuntu

Four Pillars of Cost Optimization

Right Size

- Provision instances to match the need - CPU, memory, storage, and network throughput
- Use Amazon CloudWatch metrics to dowsize as needed - How idle are instances? When?
- Best practice: Right size, then reserve

Increase Elasticity

- Stop or hibernate Amazon EBS-backed instances that are not actively in use
- Use automatic scaling to match needs based on usage

Optimal Pricing Model

- Leverage the right pricing model for your use case
- Optimize and combine purchase types
- Examples: Use On-Demand Instance and Spot Instances for variable workloads, use Reserved Instances for predictable workloads
- Consider serverless solutions (AWS Lambda)

Optimize Storage Choices

- Reduce costs while maintaining storage performance and availability
- Resize EBS volumes and change EBS volume types
- Delete EBS snapshots that are no longer needed
- Identify the most appropriate destination for specific types of data

Wrap-up

Cost optimization is an ongoing process.

- Define and enforce cost allocation tagging
- Define metrics, set targets, and review regularly.
- Encourage teams to architect for cost
- Assign the responsibility of optimization to an individual or to a team

Section 4: Container Services

Container Basics

Containers are a method of operating system virtualization.

- Repeatable
- Self-contained execution environments
- Software runs the same in different environments
- Faster to launch and stop or terminate than virtual machines

Docker is a software platform that enables you to build, test, and deploy applications quickly. Containers are created from a template called an image.

Amazon Elastic Container Service (ECS)

- A highly scalable, fast, container management service
- Key benefits
- Orchestrates the execution of Docker containers
- Maintains and scales the fleet of nodes that run your containers

- Removes the complexity of standing up the infrastructure
- Integrated with features that are familiar to Amazon EC2 service users
- Elastic Load Balancing
- Amazon EC2 security groups
- Amazon EBS volumes
- IAM roles

Do you want to manage the Amazon ECS cluster that runs the containers?

- If yes, create an Amazon ECS cluster backed by Amazon EC2 (provides more granular control over infrastructure)
- If no, create an Amazon ECS cluster backed by AWS Fargate (easier to maintain, focus on your applications)

Amazon Elastic Container Registry (ECR) is a fully managed Docker container registry that makes it easy for developers to store, manage, and deploy Docker container images.

Amazon Elastic Kubernetes Service (EKS)

What is Kubernetes?

- Kubernetes is open source software for container orchestration
- Deploy and manage containerized applicationsat scale
- The same toolset can be used on premises and in the cloud
- Complements Docker
- Docker enables you to run multiple containers on a single OS host
- Kubernetes orchestrates multiple Docker hosts (nodes)
- Automates container provisioning, networking, load distribution and scaling

Amazon Elastic Kubernetes Service (Amazon EKS)

- Enables you to run Kubernetes on AWS
- Certified Kubernetes conformant (supports easy migration)
- Supports Linux and Windows containers
- Compatible with Kubernetes community tools and supports popular Kubernetes add-ons
- Use Amazon EKS to manage clusters of Amazon EC2 compute instances and run containers that are orchestrated by Kubernetes on those instances

Section 5: Introduction to AWS Lambda

- Serverless computing enables you to build and run applications and services without provisioning or managing servers.
- Supports multiple programming languages.
- Provides built-in fault tolerance and automatic scaling.
- An event source is an AWS service or developer-created application that triggers a Lambda function to run.
- Pay-per-use pricing
- The maximum memory allocation for a single Lambda function is 3,008 MB.
- The maximum execution time for a Lambda function is 15 minutes
- Deployment package size = 250 MB unzipped, including layers

Lab: AWS Lambda --- Lab Instructions

Section 6: Introduction to Elastic Beanstalk

- An easy way to get web applications up and running
- A managed service that automatically handles
- Infrastructure provisioning and configuration
- Deployment
- Load balancing
- Automatic scaling
- Health monitoring
- Analysis and debugging
- Logging

- No additional charge for Elastic Beanstalk, pay only for the underlying resources that are used
- It supports web applications written for common platforms
- You upload your code and Elastic Beanstalk automatically handles the deployment

Lab: AWS Elastic Beanstalk --- Lab Instructions

Knowledge Check

AWS Lambda Functions and Autoscaling Video --- Walkthrough Instructions

Build a Password-Protected Website with Lambda and CloudFront --- Accompanying Blog

Build, Train, and Deploy a ML Model to SageMaker --- Supporting Notebook

SageMaker Technical Deep Dive Playlist

Deploy a Python App with Plotly Dash and Elastic Beanstalk --- Accompanying Blog