SECTION 1

Let's Get Started!

© Digital Cloud Training | https://digitalcloud.training

The SAA-C02 Exam

© Digital Cloud Training | https://digitalcloud.training

The SAA-C02 Exam

Level: Associate

Length: 130 minutes

Format: 65 questions

Cost: $150 USD

Delivery Method: Testing center or online

Scoring:

- Scaled score between 100 – 1000
- Minimum passing score of 720

© Digital Cloud Training | https://digitalcloud.training

The SAA-C02 Exam

Recommended knowledge:

- One year of hands-on experience with AWS technology
- Experience deploying, managing, and operating workloads on AWS
- Familiarity with using both the AWS Management Console and the AWS Command Line Interface (CLI)
- Understanding of the AWS Well-Architected Framework, AWS networking, security services, and the AWS global infrastructure
- Ability to identify which AWS services meet a given technical requirement and to define technical requirements for an AWS-based application

© Digital Cloud Training | https://digitalcloud.training

The SAA-C02 Exam

Question format:

- Multiple-choice: Has one correct response and three incorrect responses
- Multiple-response: Has two or more correct responses out of five or more options

© Digital Cloud Training | https://digitalcloud.training

The SAA-C02 Exam

Domain 1: Design Resilient Architectures

- 1.1 Design a multi-tier architecture solution
- 1.2 Design highly available and/or fault-tolerant architectures
- 1.3 Design decoupling mechanisms using AWS services
- 1.4 Choose appropriate resilient storage

Domain 2: Design High-Performing Architectures

- 2.1 Identify elastic and scalable compute solutions for a workload
- 2.2 Select high-performing and scalable storage solutions for a workload
- 2.3 Select high-performing networking solutions for a workload
- 2.4 Choose high-performing database solutions for a workload

© Digital Cloud Training | https://digitalcloud.training

The SAA-C02 Exam

Domain 3: Design Secure Applications and Architectures

- 3.1 Design secure access to AWS resources
- 3.2 Design secure application tiers
- 3.3 Select appropriate data security options

Domain 4: Design Cost-Optimized Architectures

- 4.1 Identify cost-effective storage solutions
- 4.2 Identify cost-effective compute and database services
- 4.3 Design cost-optimized network architectures

© Digital Cloud Training | https://digitalcloud.training

The SAA-C02 Exam

Domain	% of Exam
Domain 1: Design Resilient Architectures	30%
Domain 2: Design High-Performing Architectures	28%
Domain 3: Design Secure Applications and Architectures	24%
Domain 4: Design Cost-Optimized Architectures	18%
TOTAL	100%

© Digital Cloud Training | https://digitalcloud.training

AWS Account Overview

© Digital Cloud Training | https://digitalcloud.training

AWS Account Overview

© Digital Cloud Training | https://digitalcloud.training

AWS Account Overview

© Digital Cloud Training | https://digitalcloud.training

Create Your AWS Free Tier Account

© Digital Cloud Training | https://digitalcloud.training

What you need…

Credit card for setting up the account and paying any bills

Unique email address for this account

Check if you can use a **dynamic alias** with an existing email address

john@gmail.com

john+ACCOUNT-ALIAS-1@gmail.com

john+ACCOUNT-ALIAS-2@gmail.com

AWS account name / alias

Phone to receive an **SMS** verification code

© Digital Cloud Training | https://digitalcloud.training

Configure Account and Setup Billing Alarm

© Digital Cloud Training | https://digitalcloud.training

Account Configuration

- Configure **Account Alias**
- Enable access to billing for **IAM users**
- Update **billing preferences**
- Create a **billing alarm**
- Confirm **SNS subscription**

© Digital Cloud Training | https://digitalcloud.training

Install Tools
(AWS CLI, VS Code, CloudShell)

© Digital Cloud Training | https://digitalcloud.training

Install Tools

- Install the **AWS Command Line Interface (CLI)**
- Install **Visual Studio Code**
- Launch **AWS CloudShell**

© Digital Cloud Training | https://digitalcloud.training

SECTION 2

AWS Identity and Access Management (IAM)

© Digital Cloud Training | https://digitalcloud.training

AWS IAM Overview

© Digital Cloud Training | https://digitalcloud.training

AWS Identity and Access Management (IAM)

© Digital Cloud Training | https://digitalcloud.training

IAM Users, Groups, Roles, and Policies

© Digital Cloud Training | https://digitalcloud.training

Users, Groups, Roles and Policies

IAM Users

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

IAM Groups

Admin Group

Eric

Sunil

Development Group

Ethan

Lee

Operations Group

Andrea

The user gains the **permissions** applied to the **group** through the **policy**

Groups are collections of users. Users can be members of up to 10 groups

The main reason to use groups is to apply **permissions** to users using **policies**

© Digital Cloud Training | https://digitalcloud.training

IAM Roles

An **IAM role** is an IAM **identity** that that has specific **permissions**

© Digital Cloud Training | https://digitalcloud.training

IAM Policies

Policies are **documents** that define **permissions** and are written in JSON

AdministratorAccess

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "*",
            "Resource": "*"
        }
    ]
}
```

All permissions are **implicitly denied** by default

User

Group

Role

Identity-based policies can be applied to users, groups, and roles

Bucket Policy

```
{
    "Version": "2012-10-17",
    "Id": "Policy1561964929358",
    "Statement":[
    {
        "Sid": "Stmt1561964454052",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::515148227241:user/Paul"
        },
        "Action": "s3:*",
        "Resource": "arn:aws:s3:::dctcompany",
        "Condition": {
            "StringLike": {
                "s3:prefix": "Confidential/*"
            }
        }
    }
    ]
}
```

Resource-based policies apply to **resources** such as S3 buckets or DynamoDB tables

S3 Bucket

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Create IAM User Account

© Digital Cloud Training | https://digitalcloud.training

Root User vs IAM User

User	Login Details	Permissions
Root User	Email address	Full - Unrestricted
IAM User	Friendly name: **John** + AWS account ID or Alias	IAM Permissions Policy

© Digital Cloud Training | https://digitalcloud.training

IAM Authentication and MFA

© Digital Cloud Training | https://digitalcloud.training

IAM Authentication Methods

© Digital Cloud Training | https://digitalcloud.training

Multi-Factor Authentication

Something you **know**:

EJPx!*21p9%

Password

Something you **have**:

Something you **are**:

© Digital Cloud Training | https://digitalcloud.training

Multi-Factor Authentication

Something you **know**:

IAM User

EJPx!*21p9%

Password

Something you **have**:

MFA Virtual MFA

e.g. Google Authenticator on your smart phone

MFA Physical MFA

Physical tokens can be purchased from **third parties**

© Digital Cloud Training | https://digitalcloud.training

Enable Multi-Factor Authentication (MFA)

© Digital Cloud Training | https://digitalcloud.training

AWS Security Token Service (STS)

© Digital Cloud Training | https://digitalcloud.training

AWS Security Token Service (STS)

Credentials include:
- AccessKeyId
- Expiration
- SecretAccessKey
- SessionToken

EC2 Instance

Application

S3 Bucket

Temporary security credentials are returned

AWS STS

Instance Profile

IAM Role

EC2 attempts to assume role (sts:AssumeRole API call)

Trust policies control who can assume the role

```
{
  "Effect": "Allow",
  "Principal": {
    "Service": "ec2.amazonaws.com"
  },
  "Action": "sts:AssumeRole"
}
```

Trust Policy

Permissions Policy

Temporary credentials are used with identity federation, delegation, cross-account access, and IAM roles

© Digital Cloud Training | https://digitalcloud.training

IAM Password Policy

© Digital Cloud Training | https://digitalcloud.training

Identity-Based Policies and Resource-Based Policies

© Digital Cloud Training | https://digitalcloud.training

Identity-Based IAM Policies

Identity-based policies are JSON permissions policy documents that control what actions an identity can perform, on which resources, and under what conditions

Managed policy. Either AWS managed or customer managed

AWS managed are created and managed by AWS; customer managed are created and managed by you

Managed policies are standalone policies that can be attached to multiple users, groups, or roles

User

Inline policy

Group

Role

Inline policies have a 1-1 relationship with the user, group, or role

© Digital Cloud Training | https://digitalcloud.training

Resource-Based Policies

Resource-based policies are JSON policy documents that you attach to a resource such as an Amazon S3 bucket

Resource-based policies grant the specified **principal** (Paul) **permission** to perform specific **actions** on the **resource**

```
{
    "Version": "2012-10-17",
    "Id": "Policy1561964929358",
    "Statement":[
    {
        "Sid": "Stmt1561964454052",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::515148227241:user/Paul"
        },
        "Action": "s3:*",
        "Resource": "arn:aws:s3:::dctcompany"
    }
    ]
}
```

© Digital Cloud Training | https://digitalcloud.training

Resource-Based Policies

© Digital Cloud Training | https://digitalcloud.training

Permissions Boundaries

© Digital Cloud Training | https://digitalcloud.training

Permissions Boundaries

© Digital Cloud Training | https://digitalcloud.training

Privilege Escalation

© Digital Cloud Training | https://digitalcloud.training

Preventing Privilege Escalation

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Evaluation Logic

© Digital Cloud Training | https://digitalcloud.training

Evaluation Logic

© Digital Cloud Training | https://digitalcloud.training

Steps for Authorizing Requests to AWS

1. Authentication – AWS authenticates the principal that makes the request

AWS IAM

User

Identity-based policy

3. Evaluating all policies within the account

Resource-based policy

Console

CLI

API

Request context:

- **Actions** – the actions or operations the principal wants to perform
- **Resources** – The AWS resource object upon which actions are performed
- **Principal** – The user, role, federated user, or application that sent the request
- **Environment data** – Information about the IP address, user agent, SSL status, or time of day
- **Resource data** – Data related to the resource that is being requested

s3:GetObject

S3 Bucket

4. Determining whether a request is allowed or denied

2. Processing the request context

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Types of Policy

- **Identity-based policies** – attached to users, groups, or roles
- **Resource-based policies** – attached to a resource; define permissions for a principal accessing the resource
- **IAM permissions boundaries** – set the maximum permissions an identity-based policy can grant an IAM entity
- **AWS Organizations service control policies (SCP)** – specify the maximum permissions for an organization or OU
- **Session policies** – used with AssumeRole* API actions

© Digital Cloud Training | https://digitalcloud.training

Evaluating Policies within an AWS Account

© Digital Cloud Training | https://digitalcloud.training

Determination Rules

1. By default, all requests are implicitly denied (though the root user has full access)
2. An explicit allow in an identity-based or resource-based policy overrides this default
3. If a permissions boundary, Organizations SCP, or session policy is present, it might override the allow with an implicit deny
4. An explicit deny in any policy overrides any allows

© Digital Cloud Training | https://digitalcloud.training

Evaluation Logic

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Structure

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Structure

An IAM policy is a JSON document that consists of one or more statements

```
{
    "Statement":[{
      "Effect":"effect",
      "Action":"action",
      "Resource":"arn",
      "Condition":{
        "condition":{
          "key":"value"
          }
        }
      }
    ]
  }
```

The **Effect** element can be Allow or Deny

The **Action** element is the specific API action for which you are granting or denying permission

The **Resource** element specifies the resource that's affected by the action

The **Condition** element is optional and can be used to control when your policy is in effect

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Example 1

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "*",
            "Resource": "*"
        }
    ]
}
```

The AdministratorAccess policy uses wildcards (*) to allow all actions on all resources

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Example 2

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": ["ec2:TerminateInstances"],
            "Resource": ["*"]
        },
        {
            "Effect": "Deny",
            "Action": ["ec2:TerminateInstances"],
            "Condition": {
                "NotIpAddress": {
                    "aws:SourceIp": [
                        "192.0.2.0/24",
                        "203.0.113.0/24"
                    ]
                }
            },
            "Resource": ["*"]
        }
    ]
}
```

The specific API action is defined

The effect is to deny the API action if the IP address is not in the specified range

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Example 3

```
{
    "Version": "2012-10-17",
    "Id": "ExamplePolicy01",
    "Statement": [
        {
            "Sid": "ExampleSatement01",
            "Effect": "Allow",
            "Principal": {
                "AWS": "*"
            },
            "Action": [
                "elasticfilesystem:ClientRootAccess",
                "elasticfilesystem:ClientMount",
                "elasticfilesystem:ClientWrite"
            ],
            "Condition": {
                "Bool": {
                    "aws:SecureTransport": "true"
                }
            }
        }
    ]
}
```

You can tell this is a resource-based policy as it has a principal element defined

The policy grants read and write access to an EFS file systems to all IAM principals ("AWS ": "*")

Additionally, the policy condition element requires that SSL/TLS encryption is used

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Example 4

```
{
    "Version": "2012-10-17",
    "Statement": [
      {
        "Action": ["s3:ListBucket"],
        "Effect": "Allow",
        "Resource": ["arn:aws:s3:::mybucket"],
        "Condition": {"StringLike": {"s3:prefix": ["${aws:username}/*"]}}
      },
      {
        "Action": [
          "s3:GetObject",
          "s3:PutObject"
        ],
        "Effect": "Allow",
        "Resource": ["arn:aws:s3:::mybucket/${aws:username}/*"]
      }
    ]
  }
```

A variable is used for the s3:prefix that is replaced with the user's friendly name

The actions are allowed only within the user's folder within the bucket

© Digital Cloud Training | https://digitalcloud.training

IAM Policy Simulator

© Digital Cloud Training | https://digitalcloud.training

IAM Best Practices

© Digital Cloud Training | https://digitalcloud.training

AWS IAM Best Practices

- Lock away your AWS account root user access keys
- Create individual IAM users
- Use groups to assign permissions to IAM users
- Grant least privilege
- Get started using permissions with AWS managed policies
- Use customer managed policies instead of inline policies
- Use access levels to review IAM permissions
- Configure a strong password policy for your users
- Enable MFA

© Digital Cloud Training | https://digitalcloud.training

AWS IAM Best Practices

- Use roles for applications that run on Amazon EC2 instances
- Use roles to delegate permissions
- Do not share access keys
- Rotate credentials regularly
- Remove unnecessary credentials
- Use policy conditions for extra security
- Monitor activity in your AWS account

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – AWS IAM

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – AWS IAM

Requirement	Solution
A select group of users only should be allowed to change their IAM passwords	Create a group for the users and apply a permissions policy that grants the iam:ChangePassword API permission
An Amazon EC2 instance must be delegated with permissions to an Amazon DynamoDB table	Create a role and assign a permissions policy to the role that grants access to the database service
A company has created their first AWS account. They need to assign permissions to users based on job function	Use AWS managed policies that are aligned with common job functions

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – AWS IAM

Requirement	Solution
A solutions architect needs to restrict access to an AWS service based on the source IP address of the requester	Create an IAM permissions policy and use the Condition element to control access based on source IP address
A developer needs to make programmatic API calls from the AWS CLI	Instruct the developer to create a set of access keys and use those for programmatic access
A group of users require full access to all Amazon EC2 API actions	Create a permissions policy that uses a wildcard for the Action element relating to EC2 (ec2:*)

© Digital Cloud Training | https://digitalcloud.training

SECTION 3

Amazon Elastic Compute Cloud (EC2)

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic Compute Cloud (EC2)

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2

EC2 instances run Windows, Linux, or MacOS

An **EC2 instance** is a virtual server

EC2 hosts are **managed by AWS**

EC2 Host Server

A selection of **instance types** come with varying combinations of CPU, memory, storage and networking

© Digital Cloud Training | https://digitalcloud.training

Launching an EC2 Instance

EC2 Instance

Select an **instance type**

A **snapshot** is a point-in-time backup of an instance

An **AMI** defines the configuration of the instance

Amazon Machine Image (AMI)

EBS Snapshot

Linux

Microsoft Windows

You can customize your instance and create a **custom AMI**

Customized AMI

Family	Type	vCPUs	Memory (GiB)
General purpose	t2.micro	1	1
Compute optimized	c5n.large	2	5.25
Memory optimized	r5ad.large	2	16
Storage optimized	d2.xlarge	4	30.5
GPU instances	g2.2xlarge	8	15

The **instance type** defines the hardware profile (and cost)

© Digital Cloud Training | https://digitalcloud.training

Benefits of Amazon EC2

- **Elastic computing** – easily launch hundreds to thousands of EC2 instances within minutes
- **Complete control** – you control the EC2 instances with full root/administrative access
- **Flexible** – Choice of instance types, operating systems, and software packages
- **Reliable** – EC2 offers very high levels of availability and instances can be rapidly commissioned and replaced
- **Secure** – Fully integrated with Amazon VPC and security features
- **Inexpensive** – Low cost, pay for what you use

© Digital Cloud Training | https://digitalcloud.training

Launch EC2 Instances (Windows + Linux)

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Instance in a Public Subnet

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Connect and SSH

© Digital Cloud Training | https://digitalcloud.training

RDP to Windows Instance

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 User Data and Metadata

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 User Data

AWS Management Console

The code is run when the instance starts for the **first time**

Batch and **PowerShell** scripts can be run on Windows

Limited to **16 KB**

EC2 Instance

EC2 Instance with a **web service** is launched

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Metadata

- Instance metadata is data about your EC2 instance
- Instance metadata is available at http://169.254.169.254/latest/meta-data
- Examples:

```
[ec2-user@ip-172-31-42-248 ~]$ curl http://169.254.169.254/latest/meta-data
ami-id
ami-launch-index
ami-manifest-path
block-device-mapping/
events/
hibernation/
hostname
identity-credentials/
instance-action
instance-id
instance-life-cycle
instance-type
local-hostname
local-ipv4
```

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Metadata

- Examples ctd.:

```
[ec2-user@ip-172-31-42-248 ~]$ curl http://169.254.169.254/latest/meta-data/local-ipv4
172.31.42.248[ec2-user@ip-172-31-42-248 ~]$
```

```
[ec2-user@ip-172-31-42-248 ~]$ curl http://169.254.169.254/latest/meta-data/public-ipv4
3.26.54.18[ec2-user@ip-172-31-42-248 ~]$
```

© Digital Cloud Training | https://digitalcloud.training

[HOL] Launch Instance with User Data and Metadata

© Digital Cloud Training | https://digitalcloud.training

Accessing Services – Access Keys and IAM Roles

© Digital Cloud Training | https://digitalcloud.training

Access Keys

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Instance Profiles (IAM Roles for EC2)

© Digital Cloud Training | https://digitalcloud.training

Access Keys and IAM Roles

© Digital Cloud Training | https://digitalcloud.training

Status Checks and Monitoring

© Digital Cloud Training | https://digitalcloud.training

EC2 Placement Groups

© Digital Cloud Training | https://digitalcloud.training

EC2 Placement Groups

- **Cluster** – packs instances close together inside an Availability Zone. This strategy enables workloads to achieve the **low-latency** network performance necessary for **tightly-coupled** node-to-node communication that is typical of **HPC applications**

- **Partition** – spreads your instances across logical partitions such that groups of instances in one partition **do not share the underlying hardware** with groups of instances in different partitions. This strategy is typically used by large **distributed and replicated workloads**, such as Hadoop, Cassandra, and Kafka

- **Spread** – strictly places a small group of instances across **distinct underlying hardware** to reduce correlated failures

© Digital Cloud Training | https://digitalcloud.training

Cluster Placement Group

Partition Placement Group

© Digital Cloud Training | https://digitalcloud.training

Spread Placement Group

© Digital Cloud Training | https://digitalcloud.training

EC2 Placement Group Use Cases

Cluster	Tightly-coupled application that requires low-latency, high throughput network traffic between instances
Partition	Distributed and replicated NoSQL database; requires separate hardware for node groups
Spread	Small number of critical instances that should be kept separate from each other

© Digital Cloud Training | https://digitalcloud.training

Network Interfaces (ENI, ENA, EFA)

© Digital Cloud Training | https://digitalcloud.training

Network Interfaces (ENI, ENA, EFA)

Network Interfaces (ENI, ENA, EFA)

Elastic network interface

- Basic adapter type for when you don't have any high-performance requirements
- Can use with all instance types

Elastic network adapter

- Enhanced networking performance
- Higher bandwidth and lower inter-instance latency
- Must choose supported instance type

Elastic Fabric Adapter

- Use with High Performance Computing and MPI and ML use cases
- Tightly coupled applications
- Can use with all instance types

© Digital Cloud Training | https://digitalcloud.training

Working with ENIs

© Digital Cloud Training | https://digitalcloud.training

Public, Private and Elastic IP Addresses

© Digital Cloud Training | https://digitalcloud.training

Public, Private and Elastic IP Addresses

Public, Private and Elastic IP Addresses

© Digital Cloud Training | https://digitalcloud.training

Public, Private and Elastic IP Addresses

© Digital Cloud Training | https://digitalcloud.training

Public, Private and Elastic IP addresses

Name	Description
Public IP address	Lost when the instance is stopped Used in Public Subnets No charge Associated with a private IP address on the instance Cannot be moved between instances
Private IP address	Retained when the instance is stopped Used in Public and Private Subnets
Elastic IP address	Static Public IP address You are charged if not used Associated with a private IP address on the instance Can be moved between instances and Elastic Network Adapters

© Digital Cloud Training | https://digitalcloud.training

NAT for Public Addresses

© Digital Cloud Training | https://digitalcloud.training

NAT for Public Addresses

```
[ec2-user@ip-172-31-32-63 ~]$ ip addr show eth0
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9001 qdisc pfifo_fast sta
te UP group default qlen 1000
    link/ether 06:06:db:ad:56:28 brd ff:ff:ff:ff:ff:ff
    inet 172.31.32.63/20 brd 172.31.47.255 scope global dynamic eth0
       valid_lft 3330sec preferred_lft 3330sec
    inet6 fe80::406:dbff:fead:5628/64 scope link
       valid_lft forever preferred_lft forever
```

© Digital Cloud Training | https://digitalcloud.training

Working with EC2 IP addresses

© Digital Cloud Training | https://digitalcloud.training

Private Subnets and Bastion Hosts

© Digital Cloud Training | https://digitalcloud.training

Private Subnets and Bastion Hosts

Public Subnet Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	igw-id

Private Subnet Route Table

Destination	Target
172.31.0.0/16	Local

© Digital Cloud Training | https://digitalcloud.training

Private Subnets and Bastion Hosts

© Digital Cloud Training | https://digitalcloud.training

Private Subnets and Bastion Hosts

Public Subnet Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	igw-id

Private Subnet Route Table

Destination	Target
172.31.0.0/16	Local

© Digital Cloud Training | https://digitalcloud.training

NAT Gateways and NAT Instances Overview

© Digital Cloud Training | https://digitalcloud.training

NAT Gateways

Main Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	igw-id

Private Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	nat-gateway-id

© Digital Cloud Training | https://digitalcloud.training

NAT Instances

Must **disable** source/destination checks

Uses a special AMI with the string "**amzn-ami-vpc-nat**" in the name

The **NAT instance ID** must be specified in the **private subnet** RT

Main Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	igw-id

Private Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	nat-instance-id

© Digital Cloud Training | https://digitalcloud.training

NAT Instance vs NAT Gateway

NAT Instance	NAT Gateway
Managed by you (e.g. software updates)	Managed by AWS
Scale up (instance type) manually and use enhanced networking	Elastic scalability up to 45 Gbps
No high availability – scripted/auto-scaled HA possible using multiple NATs in multiple subnets	Provides automatic high availability within an AZ and can be placed in multiple AZs
Need to assign Security Group	No Security Groups
Can use as a bastion host	Cannot access through SSH
Use an Elastic IP address or a public IP address with a NAT instance	Choose the Elastic IP address to associate with a NAT gateway at creation
Can implement port forwarding through manual customisation	Does not support port forwarding

© Digital Cloud Training | https://digitalcloud.training

Private Subnet with NAT Gateway

© Digital Cloud Training | https://digitalcloud.training

NAT Gateways

Main Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	igw-id

Private Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	nat-gateway-id

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Lifecycle

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Lifecycle

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Lifecycle

Stopping EC2 instances

- EBS backed instances only
- No charge for stopped instances
- EBS volumes remain attached (chargeable)
- Data in RAM is lost
- Instance is migrated to a different host
- Private IPv4 addresses and IPv6 addresses retained; public IPv4 addresses released
- Associated Elastic IPs retained

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Lifecycle

Hibernating EC2 instances

- Applies to on-demand or reserved Linux instances
- Contents of RAM saved to EBS volume
- Must be enabled for hibernation when launched
- Specific prerequisites apply
- When started (after hibernation):
  - The EBS root volume is restored to its previous state
  - The RAM contents are reloaded
  - The processes that were previously running on the instance are resumed
  - Previously attached data volumes are reattached and the instance retains its instance ID

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Lifecycle

Rebooting EC2 instances

- Equivalent to an OS reboot
- DNS name and all IPv4 and IPv6 addresses retained
- Does not affect billing

Retiring EC2 instances

- Instances may be retired if AWS detects irreparable failure of the underlying hardware that hosts the instance
- When an instance reaches its scheduled retirement date, it is stopped or terminated by AWS

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Lifecycle

Terminating EC2 instances

- Means deleting the EC2 instance
- Cannot recover a terminated instance
- By default root EBS volumes are deleted

Recovering EC2 instances

- CloudWatch can be used to monitor system status checks and recover instance if needed
- Applies if the instance becomes impaired due to underlying hardware / platform issues
- Recovered instance is identical to original instance

© Digital Cloud Training | https://digitalcloud.training

Nitro Instances and Nitro Enclaves

© Digital Cloud Training | https://digitalcloud.training

AWS Nitro System

- Nitro is the underlying platform for the next generation of EC2 instances
- Support for many virtualized and bare metal instance types
- Breaks functions into specialized hardware with a Nitro Hypervisor
- Specialized hardware includes:
  - Nitro cards for VPC
  - Nitro cards for EBS
  - Nitro for Instance Storage
  - Nitro card controller
  - Nitro security chip
  - Nitro hypervisor
  - Nitro Enclaves

© Digital Cloud Training | https://digitalcloud.training

AWS Nitro System

- Improves performance, security and innovation:
  - Performance close to bare metal for virtualized instances
  - Elastic Network Adapter and Elastic Fabric Adapter
  - More bare metal instance types
  - Higher network performance (e.g. 100 Gbps)
  - High Performance Computing (HPC) optimizations
  - Dense storage instances (e.g. 60 TB)

© Digital Cloud Training | https://digitalcloud.training

AWS Nitro Enclaves

- Isolated compute environments
- Runs on isolated and hardened virtual machines
- No persistent storage, interactive access, or external networking
- Uses cryptographic attestation to ensure only authorized code is running
- Integrates with AWS Key Management Service (KMS)
- Protect and securely process highly sensitive data:
  - Personally identifiable information (PII)
  - Healthcare data
  - Financial data
  - Intellectual Property data

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Pricing Options

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Pricing Options

On-Demand

Standard rate - no discount; no commitments; dev/test, short-term, or unpredictable workloads

Reserved

1 or 3-year commitment; up to 75% discount; steady-state, predictable workloads and reserved capacity

Spot Instances

Bid for unused capacity; up to 90% discount; can be terminated at any time; workloads with flexible start and end times

Dedicated Instances

Physical isolation at the host hardware level from instances belonging to other customers; pay per instance

Dedicated Hosts

Physical server dedicated for your use; Socket/core visibility, host affinity; pay per host; workloads with server-bound software licenses

Savings Plans

Commitment to a consistent amount of usage (EC2 + Fargate + Lambda); Pay by $/hour; 1 or 3-year commitment

DigitalCloud TRAINING

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Billing

Commercial Linux distros such as **Red Hat EL** and **SUSE ES** use **hourly** pricing

Billed per hour;
Minimum of 1 hour

Billed per second;
Minimum of 1 minute

Per-second billing is for **Amazon Linux** and **Ubuntu** in **On-Demand**, **Reserved**, and **Spot** forms

Volumes billed per second;
Minimum of 1 minute

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Reserved Instances (RIs)

Standard RI

Change **AZ**,
instance size (Linux),
networking type

Use **ModifyReservedInstances** API

Term is 1 or 3 years

Can pay **All Upfront**, **Partial Upfront**, **No Upfront**

Convertible RI

Change **AZ**,
instance size (Linux),
networking type

+

Change **family**, **OS**,
tenancy, **payment option**

Use **ExchangeReservedInstances** API

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Reserved Instances (RIs)

Tenancy: **Default** or **Dedicated**

When the **attributes** of a used instance match the **attributes** of an RI the discount is applied

Reserves capacity in specified AZ

Does **not** reserve capacity; discount applies to all AZs

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Reserved Instances (RIs)

Scheduled RI

- Match capacity reservation to **recurring schedule**
- Minimum **1200 hours** per year
- Example: Reporting app that runs 6 hours a day 4 days a week = 1248 hours per year

Important

We do not have any capacity for purchasing Scheduled Reserved Instances or any plans to make it available in the future. To reserve capacity, use On-Demand Capacity Reservations instead. For discounted rates, use Savings Plans.

This message started showing recently but exam may **not** reflect this yet

© Digital Cloud Training | https://digitalcloud.training

AWS Savings Plans

Compute Savings Plan

1 or 3-year; hourly commitment to usage of **Fargate**, **Lambda**, and **EC2**; Any Region, family, size, tenancy, and OS

EC2 Savings Plan

1 or 3-year; hourly commitment to usage of **EC2** within a **selected Region** and **Instance Family**; Any size, tenancy and OS

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Spot Instances

Bid for unused capacity at up to 90% discount

2-minute warning if AWS need to reclaim capacity – available via **instance metadata** and **CloudWatch Events**

Spot Instance: One or more EC2 instances

Spot Fleet: launches and maintains the number of Spot / On-Demand instances to meet specified target capacity

EC2 Fleet: launches and maintains specified number of Spot / On-Demand / Reserved instances in a **single API call**

Can define separate OD/Spot **capacity targets**, **bids**, **instance types**, and **AZs**

© Digital Cloud Training | https://digitalcloud.training

Spot Block

Requirement:
Uninterrupted for
1-6 hours

Solution: **Spot Block**

Pricing is **30% - 45%** less than On-Demand

```
$ aws ec2 request-spot-instances \
  --block-duration-minutes 360 \
  --instance-count 5 \
  --spot-price "0.25" ...
```

© Digital Cloud Training | https://digitalcloud.training

Dedicated Instances and Dedicated Hosts

Characteristic	Dedicated Instances	Dedicated Hosts
Enables the use of dedicated physical servers	X	X
Per instance billing (subject to a $2 per region fee)	X	
Per host billing		X
Visibility of sockets, cores, host ID		X
Affinity between a host and instance		X
Targeted instance placement		X
Automatic instance placement	X	X
Add capacity using an allocation request		X

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Pricing Use Cases

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Pricing Use Cases

- On-Demand
- Reserved
- Scheduled Reserved
- Spot Instances
- Dedicated Instances
- Dedicated Hosts

Developer working on a small project for several hours; cannot be interrupted

Compute-intensive, cost-sensitive distributed computing; can withstand interruption

Steady-state, business critical, line-of-business application; continuous demand

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Pricing Use Cases

- On-Demand
- Reserved
- Scheduled Reserved
- Spot Instances
- Dedicated Instances
- Dedicated Hosts

Reporting application, runs for 6 hours a day, 4 days per week

Database with per-socket licensing

Security-sensitive application, requires dedicated hardware; per-instance billing

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon EC2

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon EC2

Requirement	Solution
Company needs to run a short batch script to configure Amazon EC2 Linux instances after they are launched	Add the bash script to the user data of the EC2 instances
A tightly coupled High Performance Computing (HPC) workload requires low-latency between nodes and optimum network performance	Launch EC2 instances in a single AZ in a cluster placement group and use an Elastic Fabric Adapter (EFA)
LoB application receives weekly bursts of traffic and must scale for short periods – need the most cost-effective solution	Use reserved instances for minimum required workload and then use Spot instances for the bursts in traffic

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Amazon EC2

Requirement	Solution
A single instance application uses a static public IP address. In the event of failure, the address must be remapped to a failover instance	Attach an Elastic IP address to the EC2 instance. Remap the EIP in the event of failure
A fleet of Amazon EC2 instances run in private subnets across multiple AZs. Company needs a redundant path to the internet	Deploy NAT Gateways into multiple AZs and update route tables
A team of engineers must administer EC2 instances in private subnets from remote locations using SSH	Deploy a bastion host in a public subnet and instruct the engineers to use the bastion host to “jump” to the instances in private subnets

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Amazon EC2

Requirement	Solution
An application uses several EC2 instances. Architect must eliminate the risk of correlated hardware failures	Launch the instances in a spread placement group across distinct underlying hardware
Application requires enhanced networking capabilities	Choose an instance type that supports enhanced networking and ensure the ENA module is installed and ENA support is enabled
Instance needs close to bare metal performance, EFA, and high performance networking	Use an AWS Nitro instance type

© Digital Cloud Training | https://digitalcloud.training

SECTION 4

Elastic Load Balancing, and Auto Scaling

© Digital Cloud Training | https://digitalcloud.training

Elasticity: Scaling Up vs Out

© Digital Cloud Training | https://digitalcloud.training

Scaling Up (vertical scaling)

Scaling Up (vertical scaling)

Scaling up means **adding** resources to the instance

Limitation is that you have a **single point of failure** (SPOF)

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Scaling Out (horizontal scaling)

Scaling out provides greater **resiliency**

Scaling out can be used to add almost unlimited capacity

Scaling Up (vertical scaling)

© Digital Cloud Training | https://digitalcloud.training

Scaling Out (horizontal scaling)

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Auto Scaling

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Auto Scaling

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Auto Scaling

- EC2 Auto Scaling **launches** and **terminates** instances dynamically
- Scaling is horizontal (scales out)
- Provides **elasticity** and **scalability**
- Responds to EC2 status checks and CloudWatch metrics
- Can scale based on demand (performance) or on a schedule
- Scaling policies define how to respond to changes in demand
- Auto Scaling groups define collections of EC2 instances that are scaled and managed together

© Digital Cloud Training | https://digitalcloud.training

Configuration of an Auto Scaling Group

A **Launch Template** specifies the EC2 instance configuration

Launch Template

- AMI and instance type
- EBS volumes
- Security groups
- Key pair
- IAM instance profile
- User data
- Shutdown behavior
- Termination protection
- Placement group name
- Capacity reservation
- Tenancy
- Purchasing option (e.g. Spot)

Launch Config

- AMI and instance type
- EBS volumes
- Security groups
- Key pair
- Purchasing option (e.g. Spot)
- IAM instance profile
- User data

Launch Configurations are replaced by launch templates and have fewer features

Configure purchase options – On-demand vs Spot

Configure VPC and Subnets

Attach Load Balancer

Configure health checks EC2 & ELB

Group size and scaling policies

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Auto Scaling

- Health checks
  - EC2 = EC2 status checks
  - ELB = Uses the ELB health checks **in addition** to EC2 status checks
- Health check grace period
  - How long to wait before checking the health status of the instance
  - Auto Scaling does not act on health checks until grace period expires

© Digital Cloud Training | https://digitalcloud.training

Auto Scaling - Monitoring

- Group metrics (ASG)
  - Data points about the Auto Scaling group
  - 1-minute granularity
  - No charge
  - Must be enabled
- Basic monitoring (Instances)
  - 5-minute granularity
  - No Charge
- Detailed monitoring (Instances)
  - 1-minute granularity
  - Charges apply

© Digital Cloud Training | https://digitalcloud.training

Additional Scaling Settings

- **Cooldowns** – Used with simple scaling policy to prevent Auto Scaling from launching or terminating before effects of previous activities are visible. Default value is 300 seconds (5 minutes)

- **Termination Policy** – Controls which instances to terminate first when a scale-in event occurs

- **Termination Protection** – Prevents Auto Scaling from terminating protected instances

- **Standby State** – Used to put an instance in the `InService` state into the `Standby` state, update or troubleshoot the instance

© Digital Cloud Training | https://digitalcloud.training

Additional Scaling Settings

- **Lifecycle Hooks** – Used to perform custom actions by pausing instances as the ASG launches or terminates them
- Use case:
  - Run a script to download and install software after launching
  - Pause an instance to process data before a scale-in (termination

Create an Auto Scaling Group

© Digital Cloud Training | https://digitalcloud.training

Load Balancing and High Availability

© Digital Cloud Training | https://digitalcloud.training

Load Balancing and High Availability

© Digital Cloud Training | https://digitalcloud.training

Fault Tolerance

Redundant components allow the system to continue to operate

The system may fail if there is no built-in **redundancy**

© Digital Cloud Training | https://digitalcloud.training

High Availability and Fault Tolerance

© Digital Cloud Training | https://digitalcloud.training

High Availability and Fault Tolerance

© Digital Cloud Training | https://digitalcloud.training

High Availability and Fault Tolerance

© Digital Cloud Training | https://digitalcloud.training

Types of Elastic Load Balancer (ELB)

© Digital Cloud Training | https://digitalcloud.training

Types of Elastic Load Balancer (ELB)

Application Load Balancer

- Operates at the request level
- Routes based on the content of the request (layer 7)
- Supports path-based routing, host-based routing, query string parameter-based routing, and source IP address-based routing
- Supports instances, IP addresses, Lambda functions and containers as targets

Network Load Balancer

- Operates at the connection level
- Routes connections based on IP protocol data (layer 4)
- Offers ultra high performance, low latency and TLS offloading at scale
- Can have a static IP / Elastic IP
- Supports UDP and static IP addresses as targets

© Digital Cloud Training | https://digitalcloud.training

Types of Elastic Load Balancer (ELB)

Old and **shouldn't** be in the exam anymore

Classic Load Balancer

- Old generation; not recommended for new applications
- Performs routing at Layer 4 and Layer 7
- Use for existing applications running in EC2-Classic

Gateway Load Balancer

- Used in front of virtual appliances such as firewalls, IDS/IPS, and deep packet inspection systems.
- Operates at Layer 3 – listens for all packets on all ports
- Forwards traffic to the TG specified in the listener rules
- Exchanges traffic with appliances using the GENEVE protocol on port 6081

New and **starting** to appear in the exam

© Digital Cloud Training | https://digitalcloud.training

ELB Features

Feature	ALB	NLB
OSI Layer	7	4
Target Type	IP, Instance, Lambda, ECS	IP, Instance
Protocol Listeners	HTTP, HTTPS, gRPC	TCP, UDP, TLS
PrivateLink support	No	(TCP, TLS)
Static IP address	No	Yes
HTTP header based routing	Yes	No
Source IP preservation	x-forwarded-for	Native
SSL termination	Load Balancer	Load Balancer or target

© Digital Cloud Training | https://digitalcloud.training

ELB Use Cases

Application Load Balancer

- Web applications with L7 routing (HTTP/HTTPS)
- Microservices architectures (e.g. Docker containers)
- Lambda targets

Network Load Balancer

- TCP and UDP based applications
- Ultra-low latency
- Static IP addresses
- VPC endpoint services

© Digital Cloud Training | https://digitalcloud.training

ELB Use Cases

Gateway Load Balancer

- Load balance virtual appliances such as:
  - Intrusion detection systems (IDS)
  - Intrusion prevention systems (IPS)
  - Next generation firewalls (NGFW)
  - Web application firewalls (WAF)
  - Distributed denial of service protection systems (DDoS)
- Integrate with Auto Scaling groups for elasticity
- Apply network monitoring and logging for analytics

© Digital Cloud Training | https://digitalcloud.training

Routing with ALB and NLB

© Digital Cloud Training | https://digitalcloud.training

Application Load Balancer (ALB)

© Digital Cloud Training | https://digitalcloud.training

Network Load Balancer (NLB)

© Digital Cloud Training | https://digitalcloud.training

Create Target Groups

© Digital Cloud Training | https://digitalcloud.training

Network Load Balancer (NLB)

© Digital Cloud Training | https://digitalcloud.training

Application Load Balancer (ALB)

© Digital Cloud Training | https://digitalcloud.training

ALB - Query String Routing

© Digital Cloud Training | https://digitalcloud.training

Amazon EC2 Scaling Policies

© Digital Cloud Training | https://digitalcloud.training

Dynamic Scaling – Target Tracking

AWS recommend scaling on metrics with a **1-minute** frequency

Average CPU = 71.5%

© Digital Cloud Training | https://digitalcloud.training

Dynamic Scaling – Simple Scaling

© Digital Cloud Training | https://digitalcloud.training

Dynamic Scaling – Step Scaling

© Digital Cloud Training | https://digitalcloud.training

Scheduled Scaling

Attempts to maintain **desired** count ->

The **minimum** instances running at any time ->

The **maximum** instances that can run ->

© Digital Cloud Training | https://digitalcloud.training

Elastically Scale the Application

© Digital Cloud Training | https://digitalcloud.training

Cross-Zone Load Balancing

© Digital Cloud Training | https://digitalcloud.training

Cross-Zone Load Balancing

When cross-zone load balancing is enabled:

- Each load balancer node distributes traffic across the registered targets in all enabled Availability Zones

When cross-zone load balancing is disabled:

- Each load balancer node distributes traffic only across the registered targets in its Availability Zone

- With Application Load Balancers, cross-zone load balancing is **always enabled**
- With Network Load Balancers and Gateway Load Balancers, cross-zone load balancing is **disabled by default**

© Digital Cloud Training | https://digitalcloud.training

Cross-Zone Load Balancing - Disabled

If cross-zone load balancing is disabled:

- Each of the three targets in Availability Zone A receives 16.6% of the traffic
- Each of the two targets in Availability Zone B receives 25% of the traffic

Cross-Zone Load Balancing - Enabled

If cross-zone load balancing is enabled:

- Each of the five targets receives 20% of the traffic

Session State and Session Stickiness

© Digital Cloud Training | https://digitalcloud.training

Storing Session State

© Digital Cloud Training | https://digitalcloud.training

Sticky Sessions

Session data such as authentication details stored **locally**

Cookie is generated and client bound to instance for **cookie lifetime**

If an instance fails, session state is **lost** – use with **session state store** for more resiliency

Client is directed to another **instance**

© Digital Cloud Training | https://digitalcloud.training

Enable Sticky Sessions

© Digital Cloud Training | https://digitalcloud.training

Secure Listeners for ELB

© Digital Cloud Training | https://digitalcloud.training

SSL/TLS Termination

© Digital Cloud Training | https://digitalcloud.training

SSL/TLS Termination

© Digital Cloud Training | https://digitalcloud.training

Create a Secure Listener

© Digital Cloud Training | https://digitalcloud.training

Clean-up Resources

Architecture Patterns – Auto Scaling and ELB

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Auto Scaling and ELB

Requirement	Solution
High availability and elastic scalability for web servers	Use Amazon EC2 Auto Scaling and an Application Load Balancer across multiple AZs
Low-latency connections over UDP to a pool of instances running a gaming application	Use a Network Load Balancer with a UDP listener
Clients need to whitelist static IP addresses for a highly available load balanced application in an AWS Region	Use an NLB and create static IP addresses in each AZ

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Auto Scaling and ELB

Requirement	Solution
Application on EC2 in an Auto Scaling group requires disaster recovery across Regions	Create an ASG in a second Region with the capacity set to 0. Take snapshots and copy them across Regions (Lambda or DLM)
Application on EC2 must scale in larger increments if a big increase in traffic occurs, compared to small increases in traffic	Use Auto Scaling with a Step Scaling policy and configure a larger capacity increase
Need to scale EC2 instances behind an ALB based on the number of requests completed by each instance	Configure a target tracking policy using the ALBRequestCountPerTarget metric

DigitalCloud TRAINING

Architecture Patterns - Auto Scaling and ELB

Requirement	Solution
Application runs on EC2 behind an ALB. Once authenticated users should not need to reauthenticate if an instance fails	Use session state store such as DynamoDB or ElastiCache
Company is deploying an IDS/IPS system using virtual appliances and needs to scale horizontally	Deploy a Gateway Load Balancer in front of the virtual appliances

© Digital Cloud Training | https://digitalcloud.training

SECTION 5

AWS Organizations

© Digital Cloud Training | https://digitalcloud.training

AWS Organizations

© Digital Cloud Training | https://digitalcloud.training

AWS Organizations

- AWS organizations allows you to consolidate multiple AWS accounts into an organization that you create and centrally manage
- Available in two feature sets:
  - **Consolidated Billing**
  - **All features**
- Includes root accounts and organizational units
- Policies are applied to root accounts or OUs
- Consolidated billing includes:
  - **Paying Account** – independent and cannot access resources of other accounts
  - **Linked Accounts** – all linked accounts are independent

© Digital Cloud Training | https://digitalcloud.training

AWS Organizations

© Digital Cloud Training | https://digitalcloud.training

Create AWS Organization and Add Account

© Digital Cloud Training | https://digitalcloud.training

Account Configuration

Service Control Policies (SCPs)

© Digital Cloud Training | https://digitalcloud.training

Service Control Policies

SCPs control the maximum available permissions

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "*",
            "Resource": "*"
        }
    ]
}
```

AWS Organization

Root

Management Account

Users in the **management** account are not restricted

Tag policy applied to enforce tag standardization

Test

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "RequireMicroInstanceType",
            "Effect": "Deny",
            "Action": "ec2:RunInstances",
            "Resource": "arn:aws:ec2:*:*:instance/*",
            "Condition": {
                "StringNotEquals": {
                    "ec2:InstanceType": "t2.micro"
                }
            }
        }
    ]
}
```

Dev

Dev users can only launch **T2.micro** instances

Prod

Prod users **cannot** launch any instance other than t2.micro (denied above)

NOTE: SCPs do not grant **ANY** permissions, they control the **AVAILABLE** permissions

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Create Service Control Policy (SCP)

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – AWS Organizations

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – AWS Organizations

Requirement	Solution
A company needs a method of quickly creating AWS accounts programmatically	Use the Organizations API to create the accounts programmatically
Users in a member account in AWS Organizations should be restricted from making changes in IAM	User a Service Control Policy (SCP) to deny access to IAM actions
An AWS account must be moved between Organizations	Migrate the account using the AWS Organizations console

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – AWS Organizations

Requirement	Solution
A solutions architect created a new account through the Organizations console and needs to login to launch resources	The architect should switch roles to access the new account
Multiple member accounts in AWS Organizations require the same permissions to be restricted using SCPs	Create an OU and add the member accounts then attach the SCP to the OU
The developers in a company each have their own AWS accounts for testing. The security team wish to enable central governance	Create an AWS Organization and send an invite to each developer's AWS account to join the Organization

© Digital Cloud Training | https://digitalcloud.training

SECTION 6

Amazon Virtual Private Cloud (VPC)

© Digital Cloud Training | https://digitalcloud.training

IPv4 Addressing Primer

© Digital Cloud Training | https://digitalcloud.training

What is an IPv4 address?

Name	Type	Value
example.com	A	52.134.26.12
test.com	A	137.10.47.51

© Digital Cloud Training | https://digitalcloud.training

Structure of an IPv4 Address

Most significant bit →

1	1	1	1	1	1	1	1
128	64	32	16	8	4	2	1

← Least significant bit

Binary Values

© Digital Cloud Training | https://digitalcloud.training

Networks and Hosts

An IPv4 address has a **network** and **host ID**

Host ID

192 . 168 . 0 . 1

Network ID

255 . 255 . 255 . 0

Subnet Mask

The **subnet mask** is used to define the **network** and **host ID**

© Digital Cloud Training | https://digitalcloud.training

Networks and Hosts

Network	192	168	0	0
Subnet Mask	255	255	255	0

24 bits

192.168.0.0/24

A **network** and **subnet mask** can also be written in this format

All computers share the same **network ID** and have a unique **host ID**

© Digital Cloud Training | https://digitalcloud.training

Classes of IPv4 Address

Class A	10	0	0	0
	255	0	0	0
Class B	172	16	0	0
	255	255	0	0
Class C	192	168	0	0
	255	255	255	0

First assignable address = 10.0.0.1
Last assignable address = 10.255.255.254
Total networks = 126
Usable addresses per network = 16,777,214

First assignable address = 172.16.0.1
Last assignable address = 172.31.255.254
Total networks = 16,382
Usable addresses per network = 65,534

First assignable address = 192.168.0.1
Last assignable address = 192.168.0.254
Total networks = 2,097,150
Usable addresses per network = 254

DigitalCloud TRAINING

Private IP Address Ranges

10	0	0	0	→	10	255	255	255
172	16	0	0	→	172	31	255	255
192	168	0	0	→	192	168	0	255

These addresses are reserved for **private use** according to IETF RFC-1918

© Digital Cloud Training | https://digitalcloud.training

Classless Interdomain Routing (CIDR)

Network	192	168	0	0
/24 Subnet Mask	255	255	255	0
/16 Subnet Mask	255	255	0	0
/20 Subnet Mask	255	255	0	0

First Address	Last Address
192.168.0.1	192.168.0.254

8 host bits = 254 addresses

16 host bits = 65,534 addresses

First Address	Last Address
192.168.0.1	192.168.255.254

12 host bits = 4094 addresses

First Address	Last Address
192.168.0.1	192.168.15.254

Classless Interdomain Routing (CIDR) uses **variable length subnets masks** (VLSM)

© Digital Cloud Training | https://digitalcloud.training

Amazon VPC Overview

© Digital Cloud Training | https://digitalcloud.training

Amazon Virtual Private Cloud (VPC)

A **VPC** is a logically isolated portion of the AWS cloud within a region

Region

VPC

Availability Zone

Public subnet

EC2 Instance

Subnets are created within **AZs**

The **VPC router** takes care of routing within the VPC and outside of the VPC

Router

Internet gateway

Availability Zone

Private subnet

EC2 Instance

You can launch **EC2 instances** into your VPC subnets

An **Internet Gateway** is attached to a VPC and used to connect to the Internet

Main Route Table

Destination	Target
10.0.0.0/16	Local
0.0.0.0/0	igw-id

The **route table** is used to configure the VPC router

© Digital Cloud Training | https://digitalcloud.training

Amazon VPC

Each **VPC** has a different block of IP addresses

CIDR stands for Classless Interdomain Routing

Region

VPC

CIDR 10.0.0.0/16

Availability Zone

Public subnet

10.0.0.0/24

Private subnet

10.0.2.0/24

Availability Zone

Public subnet

10.0.1.0/24

Private subnet

10.0.3.0/24

VPC

CIDR 10.1.0.0/16

Availability Zone

Public subnet

10.1.0.0/24

Private subnet

10.1.3.0/24

Availability Zone

Public subnet

10.1.2.0/24

Private subnet

10.1.4.0/24

Each subnet has a block of **IP addresses** from the CIDR block

You can create **multiple VPCs** within each region

© Digital Cloud Training | https://digitalcloud.training

Amazon VPC Components

VPC Component	What it is
Virtual Private Cloud (VPC)	A logically isolated virtual network in the AWS cloud
Subnet	A segment of a VPC's IP address range where you can place groups of isolated resources
Internet Gateway/Egress-only Internet Gateway	The Amazon VPC side of a connection to the public Internet for IPv4/IPv6
Router	Routers interconnect subnets and direct traffic between Internet gateways, virtual private gateways, NAT gateways, and subnets
Peering Connection	Direct connection between two VPCs
VPC Endpoints	Private connection to public AWS services
NAT Instance	Enables Internet access for EC2 instances in private subnets managed by you)
NAT Gateway	Enables Internet access for EC2 instances in private subnets (managed by AWS)
Virtual Private Gateway	The Amazon VPC side of a Virtual Private Network (VPN) connection
Customer Gateway	Customer side of a VPN connection
AWS Direct Connect	High speed, high bandwidth, private network connection from customer to aws
Security Group	Instance-level firewall
Network ACL	Subnet-level firewall

DigitalCloud TRAINING

Amazon VPC Core Knowledge

- A virtual private cloud (VPC) is a virtual network dedicated to your AWS account
- Analogous to having your own data center inside AWS
- It is logically isolated from other virtual networks in the AWS Cloud
- Provides complete control over the virtual networking environment including selection of IP ranges, creation of subnets, and configuration of route tables and gateways
- You can launch your AWS resources, such as Amazon EC2 instances, into your VPC

© Digital Cloud Training | https://digitalcloud.training

Amazon VPC Core Knowledge

- When you create a VPC, you must specify a range of IPv4 addresses for the VPC in the form of a Classless Inter-Domain Routing (CIDR) block; for example, 10.0.0.0/16
- A VPC spans all the Availability Zones in the region
- You have full control over who has access to the AWS resources inside your VPC
- By default you can create up to 5 VPCs per region
- A default VPC is created in each region with a subnet in each AZ

© Digital Cloud Training | https://digitalcloud.training

Defining VPC CIDR Blocks

© Digital Cloud Training | https://digitalcloud.training

Rules and Guidelines

- CIDR block size can be between /16 and /28
- The CIDR block must not overlap with any existing CIDR block that's associated with the VPC
- You cannot increase or decrease the size of an existing CIDR block
- The first four and last IP address are not available for use
- AWS recommend you use CIDR blocks from the RFC 1918 ranges:

RFC 1918 Range	Example CIDR Block
10.0.0.0 - 10.255.255.255 (10/8 prefix)	Your VPC must be /16 or smaller, for example, 10.0.0.0/16
172.16.0.0 - 172.31.255.255 (172.16/12 prefix)	Your VPC must be /16 or smaller, for example, 172.31.0.0/16
192.168.0.0 - 192.168.255.255 (192.168/16 prefix)	Your VPC can be smaller, for example 192.168.0.0/20

© Digital Cloud Training | https://digitalcloud.training

VPC CIDR Blocks and Subnets

VPC CIDR Block	10	0	0	0
/16 Subnet Mask	255	255	0	0

→ 10.0.0.0/16

VPC subnets have a longer **subnet mask** than the **CIDR block** by using additional bits from the host portion

Subnets:

- 10.0.1.0/24
- 10.0.2.0/24
- 10.0.3.0/24

255	255	255	0

© Digital Cloud Training | https://digitalcloud.training

Additional Considerations

- Ensure you have enough networks and hosts
- **Bigger CIDR blocks** are typically better (more flexibility)
- **Smaller subnets** are OK for most use cases
- Consider deploying application tiers per subnet
- Split your HA resources across subnets in different AZs
- VPC Peering requires non-overlapping CIDR blocks
  - This is across all VPCs in all Regions / accounts you want to connect
- **Avoid overlapping CIDR blocks** as much as possible!

© Digital Cloud Training | https://digitalcloud.training

Example VPC CIDR Block and Subnets

VPC CIDR Block	10	0	0	0
/16 Subnet Mask	255	255	0	0

Subnet Name	IPv4 CIDR block	Availability Zone	Route Table	Auto-assign Public IP v4
private-1a	10.0.0.0/24	us-east-1a	Private-RT	No
private-1b	10.0.1.0/24	us-east-1b	Private-RT	No
private-1c	10.0.2.0/24	us-east-1c	Private-RT	No
public-1a	10.0.3.0/24	us-east-1a	MAIN	Yes
public-1b	10.0.4.0/24	us-east-1b	MAIN	Yes
public-1c	10.0.5.0/24	us-east-1c	MAIN	Yes

© Digital Cloud Training | https://digitalcloud.training

VPC Wizard

© Digital Cloud Training | https://digitalcloud.training

Create a Custom VPC with Subnets

© Digital Cloud Training | https://digitalcloud.training

VPC CIDR Block and Subnets

VPC CIDR Block	10	0	0	0
/16 Subnet Mask	255	255	0	0

Subnet Name	IPv4 CIDR block	Availability Zone	Route Table	Auto-assign Public IPv4
private-1a	10.0.3.0/24	us-east-1a	Private-RT	No
private-1b	10.0.4.0/24	us-east-1b	Private-RT	No
public-1a	10.0.1.0/24	us-east-1a	MAIN	Yes
public-1b	10.0.2.0/24	us-east-1b	MAIN	Yes

Has a route to an **Internet Gateway**

Automatically assign **IPv4 Public addresses**

© Digital Cloud Training | https://digitalcloud.training

Launch Instances and Test VPC

© Digital Cloud Training | https://digitalcloud.training

Security Groups and Network ACLs

© Digital Cloud Training | https://digitalcloud.training

Security Groups and Network ACLs

© Digital Cloud Training | https://digitalcloud.training

Stateful vs Stateless Firewalls

PROTOCOL	SOURCE IP	DESTINATION IP	SOURCE PORT	DESTINATION PORT
HTTP	10.1.1.1	10.2.1.10	65188	80
HTTP	10.2.1.10	10.1.1.1	80	65188

A stateful firewall allows the return traffic automatically

A stateless firewall checks for an allow rule for both connections

© Digital Cloud Training | https://digitalcloud.training

Security Group Rules

Security groups support **allow** rules only

Inbound rules

Type	Protocol	Port range	Source
SSH	TCP	22	0.0.0.0/0
RDP	TCP	3389	0.0.0.0/0
RDP	TCP	3389	::/0
HTTPS	TCP	443	0.0.0.0/0
HTTPS	TCP	443	::/0
All ICMP - IPv4	ICMP	All	0.0.0.0/0

Separate rules are defined for outbound traffic

A source can be an **IP address** or **security group ID**

© Digital Cloud Training | https://digitalcloud.training

Security Groups Best Practice

Public subnet(s)

Security group – PublicALB

Inbound: Protocol/Port HTTP/80 Source: 0.0.0.0/0
Outbound: Protocol/Port HTTP:80 Destination: PublicEC2

Internet-facing ALB

Security group – PublicEC2

Inbound: Protocol/Port HTTP/80 Source: PublicALB
Outbound: Protocol/Port HTTP/8080 Destination: PrivateALB

Web Front-End

Private subnet(s)

Security group – PrivateALB

Inbound: Protocol/Port HTTP/8080 Source: PublicEC2
Outbound: Protocol/Port HTTP/8080 Destination: PrivateEC2

Internal ALB

Security group – PrivateEC2

Inbound: Protocol/Port HTTP/8080 Source: PrivateALB

Application Layer

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Network ACLs

Inbound Rules

Rule #	Type	Protocol	Port Range	Source	Allow / Deny
100	ALL Traffic	ALL	ALL	0.0.0.0/0	ALLOW
101	ALL Traffic	ALL	ALL	::/0	ALLOW
*	ALL Traffic	ALL	ALL	0.0.0.0/0	DENY
*	ALL Traffic	ALL	ALL	::/0	DENY

Outbound Rules

Rule #	Type	Protocol	Port Range	Destination	
100	ALL Traffic	ALL	ALL	0.0.0.0/0	ALLOW
101	ALL Traffic	ALL	ALL	::/0	ALLOW
*	ALL Traffic	ALL	ALL	0.0.0.0/0	DENY
*	ALL Traffic	ALL	ALL	::/0	DENY

NACLs have an explicit deny

Rules are processed in order

© Digital Cloud Training | https://digitalcloud.training

Configure Security Groups and NACLs

© Digital Cloud Training | https://digitalcloud.training

VPC Peering

© Digital Cloud Training | https://digitalcloud.training

VPC Peering

© Digital Cloud Training | https://digitalcloud.training

VPC Peering

Security group (VPCPEER-MGMT)

Protocol	Port	Source
ICMP	All	10.1.0.0/16
TCP	22	0.0.0.0/0

Security group (VPCPEER-PROD)

Protocol	Port	Source
ICMP	All	10.0.0.0/16
TCP	22	0.0.0.0/0

Route Table

Destination	Target
10.1.0.0/16	peering-id

Route Table

Destination	Target
10.0.0.0/16	peering-id

© Digital Cloud Training | https://digitalcloud.training

Configure VPC Peering

© Digital Cloud Training | https://digitalcloud.training

VPC Peering

Security group (VPCPEER-MGMT)

Protocol	Port	Source
ICMP	All	10.1.0.0/16
TCP	22	0.0.0.0/0

Security group (VPCPEER-PROD)

Protocol	Port	Source
ICMP	All	10.0.0.0/16
TCP	22	0.0.0.0/0

Route Table

Destination	Target
10.1.0.0/16	peering-id

Route Table

Destination	Target
10.0.0.0/16	peering-id

© Digital Cloud Training | https://digitalcloud.training

VPC Endpoints

© Digital Cloud Training | https://digitalcloud.training

VPC Interface Endpoints

© Digital Cloud Training | https://digitalcloud.training

VPC Gateway Endpoints

Route Table

Destination	Target
pl-6ca54005 (com.amazonaws.ap-southeast-2.s3, 54.231.248.0/22, 54.231.252.0/24, 52.95.128.0/21)	*vpce-ID*

A **route table** entry is required with the prefix list for S3 and the **gateway ID**

© Digital Cloud Training | https://digitalcloud.training

VPC Endpoints

	Interface Endpoint	Gateway Endpoint
What	Elastic Network Interface with a Private IP	A gateway that is a target for a specific route
How	Uses DNS entries to redirect traffic	Uses prefix lists in the route table to redirect traffic
Which services	API Gateway, CloudFormation, CloudWatch etc.	Amazon S3, DynamoDB
Security	Security Groups	VPC Endpoint Policies

© Digital Cloud Training | https://digitalcloud.training

Service Provider Model

© Digital Cloud Training | https://digitalcloud.training

Create VPC Endpoint

© Digital Cloud Training | https://digitalcloud.training

VPC Gateway Endpoints

Route Table

Destination	Target
pl-6ca54005 (com.amazonaws.ap-southeast-2.s3, 54.231.248.0/22, 54.231.252.0/24, 52.95.128.0/21)	*vpce-ID*

© Digital Cloud Training | https://digitalcloud.training

AWS Client VPN

© Digital Cloud Training | https://digitalcloud.training

AWS Client VPN

© Digital Cloud Training | https://digitalcloud.training

AWS Site-to-Site VPN

© Digital Cloud Training | https://digitalcloud.training

AWS Site-to-Site VPN

AWS VPN is a managed IPSec VPN

VPC

CIDR: 10.0.0.0/16

Public subnet

A **VGW** is deployed on the AWS site

Supports **static routes** or **BGP** peering/routing

Corporate data center

CIDR: 192.168.0.0/16

Virtual Private Gateway (VGW)

VPN connection

Customer gateway

Private subnet

Route Table

Destination	Target
192.168.0.0/16	vgw-id

A **customer gateway** is deployed on the **customer** side

Route table points to the **VGW**

© Digital Cloud Training | https://digitalcloud.training

AWS VPN CloudHub

© Digital Cloud Training | https://digitalcloud.training

AWS VPN CloudHub

© Digital Cloud Training | https://digitalcloud.training

AWS Direct Connect (DX)

© Digital Cloud Training | https://digitalcloud.training

AWS Direct Connect (DX)

© Digital Cloud Training | https://digitalcloud.training

AWS Direct Connect Benefits

- **Private** connectivity between AWS and your data center / office
- Consistent network experience – increased **speed/latency** & **bandwidth/throughput**
- Lower costs for organizations that transfer **large** volumes of data

© Digital Cloud Training | https://digitalcloud.training

AWS Direct Connect (DX)

A **VIF** is a virtual interface (**802.1Q VLAN**) and a **BGP** session

A **Private VIF** connects to a **single** VPC in the **same** AWS Region using a VGW

A **Public VIF** can be used to connect to AWS **Public services** in any Region (but **not** the Internet)

© Digital Cloud Training | https://digitalcloud.training

AWS Direct Connect (DX)

© Digital Cloud Training | https://digitalcloud.training

AWS Direct Connect (DX)

- Speeds from 50Mbps to 500Mbps can also be accessed via an APN partner
- 100 Gbps is now featuring in select locations
- DX Connections are **NOT** encrypted!
- Use an **IPSec S2S VPN** connection over a VIF to add encryption in transit

© Digital Cloud Training | https://digitalcloud.training

AWS Direct Connect Gateway

© Digital Cloud Training | https://digitalcloud.training

Direct Connect - Multiple Regions

Example architecture **without** AWS Direct Connect Gateway

A **Private VIF** connects to a **single** VPC in the **same** AWS Region using a VGW

DX is a **regional** service so **multiple** DX locations must be used

Requires **regional** offices or **long distance** (expensive) links

© Digital Cloud Training | https://digitalcloud.training

Direct Connect - Multiple Regions

Example architecture **with** AWS Direct Connect Gateway

© Digital Cloud Training | https://digitalcloud.training

Direct Connect - Multiple Regions

Example architecture **with** AWS Direct Connect Gateway

DX Gateway **does not allow** VGWs to send traffic to each other

Network traffic can be routed from **on-premises** to **any VPC**

© Digital Cloud Training | https://digitalcloud.training

AWS Transit Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS Transit Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS Transit Gateway

Specify **one subnet** from **each AZ** to enable routing within the AZ

Example full mesh architecture **with** AWS Transit Gateway

Transit Gateway is a network transit hub that interconnects **VPCs** and **on-premises** networks

TGWs can be attached to **VPNs**, **Direct Connect Gateways**, **3rd party appliances** and **TGWs** in other Regions/accounts

VPCs are **attached** to Transit Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS TGW + DX Gateway

This architecture supports **full transitive** routing between **on-premises**, **TGW** and **VPCs**

© Digital Cloud Training | https://digitalcloud.training

Using IPv6 in a VPC

© Digital Cloud Training | https://digitalcloud.training

Using IPv6 in a VPC

An IPv4 address is **32 bits** long

11000000	10101000	00000000	00000001
192 .	168 .	0 .	1

Public **IPv4** addresses are close to being exhausted and **NAT** must be used extensively

IPv4 provides approximately **4.3 billon** addresses

© Digital Cloud Training | https://digitalcloud.training

Using IPv6 in a VPC

An IPv6 address is **128 bits** long

2020 : 0001 : 9d32 : 5bc2 : 1c48 : 32c1 : a93b : b12c

Network Part | Node Part

An **IPv6** addresses use **hexadecimal** whereas **IPv4** addresses use **dotted decimal**

That's enough to assign more than **100 IPv6 addresses** to **every atom** on earth!!!

IPv6 provides **340,282,366,920,938,463,463,374,607,431,768,211,456** addresses

© Digital Cloud Training | https://digitalcloud.training

Using IPv6 in a VPC

AWS assign a **/56** IPv6 address range to your VPC

Subnets receive a **/64** address range allowing **18 million trillion** addresses

Route Table

Destination	Target
10.0.0.0/16	Local
2406:da1c:f7b:ae00::/56	Local
0.0.0.0/0	igw-id
::/0	igw-id

Route table has an entry to send all **external IPv6 traffic** to the **IGW**

All **IPv6** addresses are **publicly routable** (no NAT)

A **hexadecimal pair** is assigned for each subnet – values from **00** – **FF** = **256 /64** subnets

© Digital Cloud Training | https://digitalcloud.training

Using IPv6 in a VPC

Route Table

Destination	Target
172.31.0.0/16	Local
0.0.0.0/0	igw-id
::/0	eo-igw-id

An **Egress-only** Internet Gateway allows IPv6 traffic **outbound** but **not inbound**

All **IPv6** addresses are **publicly routable** (no NAT)

© Digital Cloud Training | https://digitalcloud.training

VPC Flow Logs

© Digital Cloud Training | https://digitalcloud.training

VPC Flow Logs

- Flow Logs capture information about the IP traffic going to and from network interfaces in a VPC
- Flow log data is stored using Amazon CloudWatch Logs or S3
- Flow logs can be created at the following levels:
  - VPC
  - Subnet
  - Network interface

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon VPC

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon VPC

Requirement	Solution
An Amazon S3 bucket must only allow access from EC2 instances in a private subnet using private IPs	Create a VPC endpoint and configure a bucket policy that restricts access to the VPC endpoint ID using the Condition element
Malicious traffic is reaching some EC2 instances in a public subnet from a few identified public IP addresses	Use a Network ACL to deny access based on the source IP addresses
A company wants to connect their on-premises data center to AWS and requires consistent performance and encryption	Create an AWS Direct Connect connection and run an AWS S2S VPN over the DX connection to enable encryption

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon VPC

Requirement	Solution
A company requires private connectivity between VPCs in different Regions will full redundancy	Create VPC peering connections between the VPCs
Several remote office locations should be connected to an Amazon VPC and to each other over the internet with full encryption	Create VPG and attach multiple remote locations in a hub and spoke topology using AWS CloudHub
Microservices app requires instance-level firewall with different rules per application component	Create separate security groups for each application component and configure the appropriate rules

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon VPC

Requirement	Solution
A company is using IPv6 addresses with Amazon EC2 and needs to enable outbound internet connectivity	Configure an egress-only Internet Gateway
Subnet must be configured that allows internet connectivity using IPv4 with auto-address assignment	Attach an internet gateway to the VPC and update the route table for the new subnet. Enable the auto-assign public IPv4 setting for the subnet
An on-premises data center needs to establish S2S VPN connections to several VPCs in a full mesh architecture	Deploy an AWS Transit Gateway and attach the VPN connection from on-premises and each VPC

© Digital Cloud Training | https://digitalcloud.training

SECTION 7

Amazon Simple Storage Service (S3)

© Digital Cloud Training | https://digitalcloud.training

Amazon S3 Overview

© Digital Cloud Training | https://digitalcloud.training

Amazon Simple Storage Service (S3)

Accessing objects in a bucket:

```
https://bucket.s3.aws-region.amazonaws.com/key
https://s3.aws-region.amazonaws.com/bucket/key
```

The **HTTP protocol** is used with a **REST API** (e.g. GET, PUT, POST, SELECT, DELETE)

© Digital Cloud Training | https://digitalcloud.training

Amazon Simple Storage Service (S3)

- You can store any type of file in S3
- Files can be anywhere from 0 bytes to 5 TB
- There is unlimited storage available
- S3 is a universal namespace so **bucket names** must be **unique globally**
- However, you create your buckets within a **REGION**
- It is a best practice to create buckets in regions that are physically closest to your users to reduce latency
- There is no hierarchy for objects within a bucket
- Delivers strong read-after-write consistency

© Digital Cloud Training | https://digitalcloud.training

Buckets, Folders, and Objects

© Digital Cloud Training | https://digitalcloud.training

Buckets, Folders, and Objects

- Folders **can** be created within folders
- Buckets **cannot** be created within other buckets
- An objects consists of:
  - Key (the name of the object)
  - Version ID
  - Value (actual data)
  - Metadata
  - Subresources
  - Access control information

© Digital Cloud Training | https://digitalcloud.training

Accessing Amazon S3

© Digital Cloud Training | https://digitalcloud.training

Amazon S3 Storage Classes

© Digital Cloud Training | https://digitalcloud.training

S3 Storage Classes

	S3 Standard	S3 Intelligent Tiering	S3 Standard-IA	S3 One Zone-IA	S3 Glacier	S3 Glacier Deep Archive
Designed for durability	99.999999999%	99.999999999%	99.999999999%	99.999999999%	99.999999999%	99.999999999%
Designed for availability	99.99%	99.9%	99.9%	99.5%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99.9%	99.9%
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	Per GB retrieved	Per GB retrieved	Per GB retrieved	Per GB retrieved
First byte latency	milliseconds	milliseconds	milliseconds	milliseconds	select minutes or hours	select hours
Storage type	Object	Object	Object	Object	Object	Object
Lifecycle transitions	Yes	Yes	Yes	Yes	Yes	Yes

© Digital Cloud Training | https://digitalcloud.training

S3 Storage Class Use Cases

- S3 Standard
- S3 Intelligent Tiering
- S3 Standard-IA
- S3 One Zone-IA
- S3 Glacier
- S3 Glacier Deep Archive

Require long-term storage for production data with low cost and immediate access for occasional requests

Copy of backup data required in a separate Region with minimal access latency

Frequently accessed data; ad-hoc needs; short-term requirements (<30 days)

Lowest cost required for long-term archival of data for compliance purposes

© Digital Cloud Training | https://digitalcloud.training

Create Amazon S3 Bucket

© Digital Cloud Training | https://digitalcloud.training

IAM Policies, Bucket Policies and ACLs

© Digital Cloud Training | https://digitalcloud.training

IAM Policies

- IAM Policies are identity-based policies
- Specify what actions are allowed on what AWS resources

© Digital Cloud Training | https://digitalcloud.training

Bucket Policies

- Bucket Policies are resource-based policies
- Can only be attached to Amazon S3 buckets
- Also use the AWS access policy language

```
{
    "Version": "2012-10-17",
    "Id": "Policy1561964929358",
    "Statement":[
    {
        "Sid": "Stmt1561964454052",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::515148227241:user/Paul"
        },
        "Action": "s3:*",
        "Resource": "arn:aws:s3:::dctcompany"
    }
    ]
}
```

© Digital Cloud Training | https://digitalcloud.training

S3 Access Control Lists (ACLs)

- Legacy access control mechanism that predates IAM
- AWS generally recommends using **S3 bucket policies** or **IAM policies** rather than ACLs
- Can be attached to a **bucket** or directly to an **object**
- Limited options for grantees and permissions

© Digital Cloud Training | https://digitalcloud.training

When to use each access control mechanism

- **Use IAM policies if:**
  - You need to control access to AWS services other than S3
  - You have numerous S3 buckets each with different permissions requirements (IAM policies will be easier to manage)
  - You prefer to keep access control policies in the IAM environment
- **Use S3 bucket policies if:**
  - You want a simple way to grant cross-account access to your S3 environment, without using IAM roles
  - Your IAM policies are reaching the size limits
  - You prefer to keep access control policies in the S3 environment

© Digital Cloud Training | https://digitalcloud.training

Authorization Process

© Digital Cloud Training | https://digitalcloud.training

Access Control Lists (ACLs)

© Digital Cloud Training | https://digitalcloud.training

Bucket and User Policy Practice

© Digital Cloud Training | https://digitalcloud.training

S3 Versioning, Replication and Lifecycle Rules

© Digital Cloud Training | https://digitalcloud.training

S3 Versioning

- Versioning is a means of keeping multiple variants of an object in the same bucket
- Use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket
- Versioning-enabled buckets enable you to recover objects from **accidental deletion** or **overwrite**

© Digital Cloud Training | https://digitalcloud.training

S3 Replication

Cross-Region Replication (CRR)

Buckets must have **versioning** enabled

© Digital Cloud Training | https://digitalcloud.training

S3 Lifecycle Management

There are two types of actions:

- **Transition actions** - Define when objects transition to another storage class
- **Expiration actions** - Define when objects expire (deleted by S3)

© Digital Cloud Training | https://digitalcloud.training

S3 LM: Supported Transitions

You can transition from the following:

- The **S3 Standard** storage class to any other storage class
- Any storage class to the **S3 Glacier** or **S3 Glacier Deep** Archive storage classes
- The **S3 Standard-IA** storage class to the **S3 Intelligent-Tiering** or **S3 One Zone-IA** storage classes
- The **S3 Intelligent-Tiering** storage class to the **S3 One Zone-IA** storage class
- The **S3 Glacier** storage class to the **S3 Glacier Deep Archive** storage class

© Digital Cloud Training | https://digitalcloud.training

S3 LM: Unsupported Transitions

You can't transition from the following:

- Any storage class to the **S3 Standard** storage class
- Any storage class to the **Reduced Redundancy** storage class
- The **S3 Intelligent-Tiering** storage class to the **S3 Standard-IA** storage class
- The **S3 One Zone-IA** storage class to the **S3 Standard-IA** or **S3 Intelligent-Tiering** storage classes

© Digital Cloud Training | https://digitalcloud.training

Versioning and Replication

© Digital Cloud Training | https://digitalcloud.training

Lifecycle Rules

© Digital Cloud Training | https://digitalcloud.training

MFA with Amazon S3

© Digital Cloud Training | https://digitalcloud.training

S3 Multi-Factor Authentication Delete (MFA Delete)

- Adds MFA requirement for bucket owners to the following operations:
  - Changing the versioning state of a bucket
  - Permanently deleting an object version
- The `x-amz-mfa` request header must be included in the above requests
- The second factor is a token generated by a hardware device or software program
- Requires **versioning** to be enabled on the bucket

© Digital Cloud Training | https://digitalcloud.training

S3 Multi-Factor Authentication Delete (MFA Delete)

- **Versioning** can be enabled by:
  - Bucket owners (root account)
  - AWS account that created the bucket
  - Authorized IAM users
- **MFA delete** can be enabled by:
  - Bucket owner (root account)

© Digital Cloud Training | https://digitalcloud.training

MFA-Protected API Access

- Used to enforce another authentication factor (MFA code) when accessing AWS resources (not just S3)
- Enforced using the `aws:MultiFactorAuthAge` key in a bucket policy:

```
{
    "Version": "2012-10-17",
    "Id": "123",
    "Statement": [
      {
        "Sid": "",
        "Effect": "Deny",
        "Principal": "*",
        "Action": "s3:*",
        "Resource": "arn:aws:s3:::examplebucket/securedocuments/*",
        "Condition": { "Null": { "aws:MultiFactorAuthAge": true }}
      }
    ]
}
```

Denies any API operation that is not authenticated using MFA

© Digital Cloud Training | https://digitalcloud.training

S3 Encryption

© Digital Cloud Training | https://digitalcloud.training

S3 Encryption

Server-side encryption with S3 managed keys (SSE-S3)

- S3 managed keys
- Unique object keys
- Master key
- AES 256

Server-side encryption with AWS KMS managed keys (SSE-KMS)

- KMS managed keys
- Customer master keys
- CMK can be customer generated

© Digital Cloud Training | https://digitalcloud.training

S3 Encryption

Server-side encryption with client provided keys (SSE-C)

- Client managed keys
- Not stored on AWS

Client-side encryption

- Client managed keys
- Not stored on AWS
- OR you can use a KMS CMK

© Digital Cloud Training | https://digitalcloud.training

S3 Default Encryption

- Amazon S3 default encryption provides a way to set the default encryption behavior for an S3 bucket
- You can set default encryption on a bucket so that all new objects are encrypted when they are stored in the bucket
- The objects are encrypted using server-side encryption
- Amazon S3 encrypts objects before saving them to disk and decrypts them when the objects are downloaded
- There is no change to the encryption of objects that existed in the bucket before default encryption was enabled

© Digital Cloud Training | https://digitalcloud.training

Prevent uploads of unencrypted objects

```
{
    "Version": "2012-10-17",
    "Id": "PutObjPolicy",
    "Statement": [
        {
            "Sid": "DenyIncorrectEncryptionHeader",
            "Effect": "Deny",
            "Principal": "*",
            "Action": "s3:PutObject",
            "Resource": "arn:aws:s3:::<bucket_name>/*",
            "Condition": {
                "StringNotEquals": {
                    "s3:x-amz-server-side-encryption": "AES256"
                }
            }
        },
        {
            "Sid": "DenyUnEncryptedObjectUploads",
            "Effect": "Deny",
            "Principal": "*",
            "Action": "s3:PutObject",
            "Resource": "arn:aws:s3:::<bucket_name>/*",
            "Condition": {
                "Null": {
                    "s3:x-amz-server-side-encryption": tr
                }
            }
        }
    ]
}
```

Enforces encryption using **SSE-S3**

For **SSE-KMS** use "aws:kms"

Example PUT request

```
PUT /example-object HTTP/1.1
Host: myBucket.s3.amazonaws.com
Date: Wed, 8 Jun 2016 17:50:00 GMT
Authorization: authorization string
Content-Type: text/plain
Content-Length: 11434
x-amz-meta-author: Janet
Expect: 100-continue
x-amz-server-side-encryption: AES256
[11434 bytes of object data]
```

© Digital Cloud Training | https://digitalcloud.training

Enforce Encryption

© Digital Cloud Training | https://digitalcloud.training

S3 Event Notifications

© Digital Cloud Training | https://digitalcloud.training

S3 Event Notifications

- Sends notifications when events happen in buckets
- Destinations include:
  - Amazon Simple Notification Service (SNS) topics
  - Amazon Simple Queue Service (SQS) queues
  - AWS Lambda

S3 Presigned URLs

© Digital Cloud Training | https://digitalcloud.training

S3 Presigned URLs

AWS S3 CLI command to generate a presigned URL

```
aws s3 presign s3://dct-data-bucket/cool_image.jpeg
```

```
https://dct-data-bucket.s3.ap-southeast-2.amazonaws.com/cool_image.jpeg?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIA3KSVPHP6MAHNW5YH%2F20200909%2Fap-southeast-2%2Fs3%2Faws4_request&X-Amz-Date=20200909T053538Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=8b74653beee371da07a73dfdb4ff6883742383afa528aecd5c95c326c97764db
```

This is the response; the URL expires after 1 hour

© Digital Cloud Training | https://digitalcloud.training

Multipart Upload & Transfer Acceleration

© Digital Cloud Training | https://digitalcloud.training

S3 Multipart Upload

- Multipart upload uploads objects in parts independently, in parallel and in any order
- Performed using the S3 Multipart upload API
- It is recommended for objects of 100 MB or larger
- Can be used for objects from 5 MB up to 5 TB
- Must be used for objects larger than 5 GB

© Digital Cloud Training | https://digitalcloud.training

S3 Transfer Acceleration

- Uses CloudFront edge locations to improve performance of transfers from client to S3 bucket

```
http://bucketname.s3-accelerate.amazonaws.com

http://bucketname.s3-accelerate.dualstack.amazonaws.com
```

Transfer Acceleration is enabled at the bucket level

AWS only charges if there's a performance improvement

© Digital Cloud Training | https://digitalcloud.training

S3 Select and Glacier Select

© Digital Cloud Training | https://digitalcloud.training

S3 Select and Glacier Select

© Digital Cloud Training | https://digitalcloud.training

Server Access Logging

© Digital Cloud Training | https://digitalcloud.training

Server Access Logging

- Provides detailed records for the requests that are made to a bucket
- Details include the requester, bucket name, request time, request action, response status, and error code (if applicable)
- Disabled by default
- Only pay for the storage space used
- Must configure a separate bucket as the destination (can specify a prefix)
- Must grant write permissions to the Amazon S3 Log Delivery group on destination bucket

© Digital Cloud Training | https://digitalcloud.training

S3 Static Website

© Digital Cloud Training | https://digitalcloud.training

Cross-Origin Resource Sharing (CORS)

© Digital Cloud Training | https://digitalcloud.training

CORS with Amazon S3

- Allows requests from an origin to another origin
- Origin is defined by DNS name, protocol, and port

© Digital Cloud Training | https://digitalcloud.training

CORS with Amazon S3

- Enabled through setting:
  - Access-Control-Allow-Origin
  - Access-Control-Allow-Methods
  - Access-Control-Allow-Headers
- These settings are defined using rules
- Rules are added using JSON files in S3

© Digital Cloud Training | https://digitalcloud.training

Example CORS Rule

```
[
    {
        "AllowedHeaders": [
            "*"
        ],
        "AllowedMethods": [
            "PUT",
            "POST",
            "DELETE"
        ],
        "AllowedOrigins": [
            "http://www.mycompany.com"
        ],
        "ExposeHeaders": []
    }
]
```

© Digital Cloud Training | https://digitalcloud.training

Cross Account Access

© Digital Cloud Training | https://digitalcloud.training

Cross Account Access Methods

- **Resource-based policies** and **IAM policies** for **programmatic-only** access to S3 bucket objects
- **Resource-based ACL** and **IAM policies** for **programmatic-only** access to S3 bucket objects
- Cross-account **IAM roles** for **programmatic and console access** to S3 bucket objects

© Digital Cloud Training | https://digitalcloud.training

Cross Account Access

© Digital Cloud Training | https://digitalcloud.training

Configure Cross Account Access

© Digital Cloud Training | https://digitalcloud.training

Cross Account Access

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon S3

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon S3

Requirement	Solution
Company is concerned about accidental deletion of Amazon S3 objects	Enable S3 versioning
Data stored in S3 is frequently accessed for 30 days then is rarely accessed but must be immediately retrievable	Use a lifecycle policy to transition objects from S3 standard to S3 Standard-IA after 30 days
A backup of S3 objects within a specific folder in a bucket must be replicated to another Region	Configure cross-region replication and specify the folder name as a prefix

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon S3

Requirement	Solution
Previous versions of objects in a versioning-enabled S3 bucket must be stored long term at the lowest cost	Create a lifecycle rule that transitions previous versions to S3 Glacier Deep Archive
A company wishes to manage all encryption of S3 objects through their application with their own encryption keys	Use client-side encryption with client managed keys
Unencrypted objects in an Amazon S3 bucket must be encrypted	Re-upload the objects and specify the encryption an encryption key

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon S3

Requirement	Solution
An administrator requires a notification when objects are deleted from an Amazon S3 bucket	Configure an event notification that uses the SNS service
A group of customers without AWS credentials must be granted time-limited access to a software update that is stored in an Amazon S3 bucket	Generate a presigned URL
Solutions architects require both programmatic and console access across AWS accounts	Configure cross-account access using IAM roles

© Digital Cloud Training | https://digitalcloud.training

SECTION 8

DNS, Caching, and Performance Optimization

© Digital Cloud Training | https://digitalcloud.training

DNS and Amazon Route 53

© Digital Cloud Training | https://digitalcloud.training

The Domain Name System (DNS)

User enters **website address** in browser

Name	Type	Value
mycompany.local	A	192.168.0.1
emailserver.local	A	192.168.0.2

DNS Server

Domain name is resolved to the **IP address** of the webserver

Web Server

Computer connects to **192.168.0.1**

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 Routing Policies

Routing Policy	What it does
Simple	Simple DNS response providing the IP address associated with a name
Failover	If primary is down (based on health checks), routes to secondary destination
Geolocation	Uses geographic location you're in (e.g. Europe) to route you to the closest region
Geoproximity	Routes you to the closest region within a geographic area
Latency	Directs you based on the lowest latency route to resources
Multivalue answer	Returns several IP addresses and functions as a basic load balancer
Weighted	Uses the relative weights assigned to resources to determine which to route to

© Digital Cloud Training | https://digitalcloud.training

Amazon Route Features

Amazon Route 53

Domain Registration

.net
.com
.org

Hosted zone

example.com
dctlabs.com

Health Checks

EC2 Instances

Traffic Flow

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Register Domain with Route 53 (Optional)

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 Hosted Zones

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 Hosted Zones

Name	Type	Value	TTL
example.com	A	8.1.2.1	60
dev.example.com	A	8.1.2.2	60

Amazon Route 53

This is an example of a **public hosted zone**

A **hosted zone** represents a set of records belonging to a domain

example.com

What's the address for example.com?

Address is 8.1.2.1

HTTP GET to 8.1.2.1

Region

VPC

Availability Zone

Public subnet

Web Server: 8.1.2.1

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 Hosted Zones

Name	Type	Value	TTL
db.mycompany.local	A	10.0.0.10	60
app.mycompany.local	A	10.0.0.11	60

This is an example of a **private hosted zone**

© Digital Cloud Training | https://digitalcloud.training

Migration to/from Route 53

- You can migrate from **another DNS provider** and can import records
- You can migrate a hosted zone to **another AWS account**
- You can migrate from Route 53 to **another registrar**
- You can also associate a Route 53 hosted zone with a **VPC in another account**
  - Authorize association with VPC in the second account.
  - Create an association in the second account

© Digital Cloud Training | https://digitalcloud.training

Route 53 Routing Policies

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 Routing Policies

Routing Policy	What it does
Simple	Simple DNS response providing the IP address associated with a name
Failover	If primary is down (based on health checks), routes to secondary destination
Geolocation	Uses geographic location you're in (e.g. Europe) to route you to the closest region
Geoproximity	Routes you to the closest region within a geographic area
Latency	Directs you based on the lowest latency route to resources
Multivalue answer	Returns several IP addresses and functions as a basic load balancer
Weighted	Uses the relative weights assigned to resources to determine which to route to

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 – Simple Routing Policy

Name	Type	Value	TTL
simple.dctlabs.com	A	1.1.1.1 2.2.2.2	60
simple2.dctlabs.com	A	3.3.3.3	60

Amazon Route 53 – Weighted Routing Policy

Name	Type	Value	Health	Weight
weighted.dctlabs.com	A	1.1.1.1	ID	60
weighted.dctlabs.com	A	2.2.2.2	ID	20
weighted.dctlabs.com	A	3.3.3.3	ID	20

Simplified values - actually uses an integer between 0 and 255

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 – Latency Routing Policy

Name	Type	Value	Health	Region
latency.dctlabs.com	A	1.1.1.1	ID	ap-southeast-1
latency.dctlabs.com	A	2.2.2.2	ID	us-east-1
latency.dctlabs.com	A	alb-id	ID	ap-southeast-2

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 – Failover Routing Policy

Name	Type	Value	Health	Record Type
failover.dctlabs.com	A	1.1.1.1	ID	Primary
failover.dctlabs.com	A	alb-id		Secondary

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 – Geolocation Routing Policy

Name	Type	Value	Health	Geolocation
geolocation.dctlabs.com	A	1.1.1.1	ID	Singapore
geolocation.dctlabs.com	A	2.2.2.2	ID	Default
geolocation.dctlabs.com	A	alb-id	ID	Oceania

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 – Multivalue Routing Policy

Name	Type	Value	Health	Multi Value
multivalue.dctlabs.com	A	1.1.1.1	ID	Yes
multivalue.dctlabs.com	A	2.2.2.2	ID	Yes
multivalue.dctlabs.com	A	3.3.3.3	ID	Yes

© Digital Cloud Training | https://digitalcloud.training

Using Route 53 Routing Policies

© Digital Cloud Training | https://digitalcloud.training

Amazon Route 53 Resolver

© Digital Cloud Training | https://digitalcloud.training

Route 53 Resolver – Outbound Endpoints

© Digital Cloud Training | https://digitalcloud.training

Route 53 Resolver – Inbound Endpoints

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudFront Origins and Distributions

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudFront

© Digital Cloud Training | https://digitalcloud.training

CloudFront Origins and Distributions

CloudFront Distribution

Name: **d1schtd9zdwrm1.cloudfront.net**

S3 Origin

S3 static websites can also be origins

Behaviors

Path Pattern
Viewer Protocol Policy
Cache Policy
Origin Request Policy

Custom Origin

RTMP distributions were discontinued so only **web distributions** are currently available

CloudFront Web Distribution:

- Speed up distribution of static and dynamic content, for example, .html, .css, .php, and graphics files
- Distribute media files using HTTP or HTTPS
- Add, update, or delete objects, and submit data from web forms
- Use live streaming to stream an event in real time

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudFront Caching and Behaviors

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudFront Caching

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudFront Caching

Decreasing the TTL is best for **Dynamic content**, increasing TTL is better for performance (and reduces load on origin)

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudFront Caching

- You can define a maximum **Time To Live** (TTL) and a default TTL
- TTL is defined at the **behavior** level
- This can be used to define different TTLs for different file types (e.g. png vs jpg)
- After expiration, CloudFront checks the origin for any new requests (check the file is the latest version)
- Headers can be used to control the cache:
  - `Cache-Control max-age=(seconds)` - specify how long before CloudFront gets the object again from the origin server
  - `Expires` – specify an expiration date and time

© Digital Cloud Training | https://digitalcloud.training

CloudFront Path Patterns

Precedence	Path Pattern	Origin or Origin Group	Viewer Protocol Policy	Cache Policy Name
0	*.jpg	Origin 1	HTTP and HTTPS	Managed-CachingOptimized
1	*.mp4	Origin 2	HTTP and HTTPS	Managed-CachingOptimized
2	Default (*)	Origin 1	HTTP and HTTPS	Managed-CachingOptimized

© Digital Cloud Training | https://digitalcloud.training

Caching Based on Request Headers

- You can configure CloudFront to forward **headers** in the **viewer request** to the origin
- CloudFront can then cache multiple versions of an object based on the values in one or more request headers
- Controlled in a behavior to do one of the following:
  - Forward all headers to your origin (objects are not cached)
  - Forward a whitelist of headers that you specify
  - Forward only the default headers (doesn't cache objects based on values in request headers)

© Digital Cloud Training | https://digitalcloud.training

CloudFront Signed URLs and OAI

© Digital Cloud Training | https://digitalcloud.training

CloudFront Signed URLs

- Signed URLs provide more control over access to content.
- Can specify beginning and expiration date and time, IP addresses/ranges of users.

Signed URLs should be used for **individual files** and clients that don't support **cookies**.

© Digital Cloud Training | https://digitalcloud.training

CloudFront Signed Cookies

- Similar to Signed URLs
- Use signed cookies when you don’t want to change URLs
- Can also be used when you want to provide access to **multiple restricted files** (Signed URLs are for individual files)

© Digital Cloud Training | https://digitalcloud.training

CloudFront Origin Access Identity (OAI)

```
{
    "Version": "2008-10-17",
    "Id": "PolicyForCloudFrontPrivateContent",
    "Statement": [
        {
            "Sid": "1",
            "Effect": "Allow",
            "Principal": {
                "AWS": "arn:aws:iam::cloudfront:user/CloudFront Origin Access Identity E11A2JL2H306JJ"
            },
            "Action": "s3:GetObject",
            "Resource": "arn:aws:s3:::mybucket/*"
        }
    ]
}
```

Policy restricts access to the **OAI**

HTTP GET **https://d1schtd9zdwrm1.cloudfront.net**

Custom Origin

S3 Bucket configured as static website

Bucket Policy

Origin Access Identity (OAI)

Amazon CloudFront

Users

Blocked by **bucket policy**

GET **https://mybucket.s3.amazonaws.com/beach.jpg**

© Digital Cloud Training | https://digitalcloud.training

Cache and Behavior Settings

© Digital Cloud Training | https://digitalcloud.training

CloudFront SSL/TLS and SNI

© Digital Cloud Training | https://digitalcloud.training

CloudFront SSL/TLS

Viewer Protocol

For CloudFront certificate must be issued in **us-east-1**

Certificate can be **ACM** or a trusted **third-party CA**

SSL

AWS Certificate Manager

Default CF **domain name** can be changed using **CNAMES**

S3 has its **own** certificate (can't be changed)

Certificate can be **ACM** (ALB) or **third-party** (EC2)

SSL

SSL

Origin Protocol

S3 Origin

Origin certificates must be **public certificates**

Custom Origin

© Digital Cloud Training | https://digitalcloud.training

CloudFront Server Name Indication (SNI)

HTTP GET: https://**mypublicdomain.com**

HTTP GET: https://**myotherdomain.com**

Request URL includes domain name which **matches** certificate

Name: **myotherdomain.com**

Name: **mypublicdomain.com**

Multiple certificates share the **same IP** with SNI

Note: SNI works with browsers/clients released **after 2010** – otherwise need **dedicated IP**

© Digital Cloud Training | https://digitalcloud.training

Lambda@Edge

© Digital Cloud Training | https://digitalcloud.training

Lambda@Edge

- Run Node.js and Python Lambda functions to customize the content CloudFront delivers
- Executes functions closer to the viewer
- Can be run at the following points
  - After CloudFront receives a request from a viewer (viewer request)
  - Before CloudFront forwards the request to the origin (origin request)
  - After CloudFront receives the response from the origin (origin response)
  - Before CloudFront forwards the response to the viewer (viewer response)

© Digital Cloud Training | https://digitalcloud.training

AWS Global Accelerator

© Digital Cloud Training | https://digitalcloud.training

AWS Global Accelerator

© Digital Cloud Training | https://digitalcloud.training

Create a Global Accelerator

© Digital Cloud Training | https://digitalcloud.training

AWS Global Accelerator

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – DNS, Caching and Performance Optimization

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – DNS, Caching and Performance Optimization

Requirement	Solution
An Elastic Load Balancer must be resolvable using a company's public domain name. A Route 53 hosted zone exists	Create an Alias record that maps the domain name to the ELB
A website runs across two AWS Regions. All traffic goes to one Region and should be redirected only if the website is unavailable	Create a failover routing policy in AWS Route 53 and configure health checks on the primary
Websites run in several countries and distribution rights require restricting access to content based on the geographic source of the connection	Use AWS Route 53 geolocation routing and restrict distribution based on geographic location

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – DNS, Caching and Performance Optimization

Requirement	Solution
A CloudFront distribution has multiple S3 origins. Requests should be served from different origins based on file type being requested	Modify the CloudFront behavior and configure a path pattern
Content is accessed using an application and CloudFront distribution. Need to control access to multiple files on the distribution	Configure signed cookies and update the application
Application runs behind an Application Load Balancer in multiple Regions. Need to intelligently route traffic based on latency and availability	Create an AWS Global Accelerator and add the ALBs

© Digital Cloud Training | https://digitalcloud.training

SECTION 9

Block and File Storage

© Digital Cloud Training | https://digitalcloud.training

Block vs File vs Object Storage

© Digital Cloud Training | https://digitalcloud.training

Hard Drives

Hard drives are **block-based** storage systems

The Operating System (OS) can be used to create **volumes**. A volume can be partitioned and formatted

Hard Disk Drive (HDD)

Hard drives are block-based storage systems

© Digital Cloud Training | https://digitalcloud.training

Network Attached Storage

The Operating System (OS) sees a **filesystem** that is mapped to a local drive letter

The NAS "shares" **filesystems** over the network

NAS devices are file-based storage systems

© Digital Cloud Training | https://digitalcloud.training

Object Storage Systems

User uploads **objects** using a **web browser**

There is no **hierarchy** of objects in the container

The **HTTP protocol** is used with a REST API (e.g. GET, PUT, POST, SELECT, DELETE)

Object Storage Container

Objects can be files, videos, images etc.

© Digital Cloud Training | https://digitalcloud.training

Block, File, and Object Storage

The OS sees **volumes** that can be partitioned and formatted

A **filesystem** can be shared by many users/computers

Massively scalable, low cost

Block Storage

The OS reads/writes at the **block level**. Disks can be internal, or network attached

File Storage

A **filesystem** is "mounted" to the OS using a **network share**

Object Storage

Object Storage Container

There is **no hierarchy** of objects in the container

Uses a **REST API**

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

AWS Storage Services

Block Storage

Amazon Elastic Block Store

File Storage

Amazon Elastic File System

Object Storage

Amazon Simple Storage Service (S3)

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS Deployment and Volume Types

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS Deployment

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS Multi-Attach

May not be on the exam yet

Must be within a **single AZ**

Must be a **Provisioned IOPS** io1 volume

Up to **16 instances** can be attached to a single volume

Available for **Nitro system-based** EC2 instances

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS SSD-Backed Volumes

New and **may not** be on the exam yet

New and **may not** be on the exam yet

	General Purpose SSD		Provisioned IOPS SSD		
Volume type	gp3	gp2	io2 Block Express ‡	io2	io1
Durability	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.999% durability (0.001% annual failure rate)		99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)
Use cases	• Low-latency interactive apps • Development and test environments		Workloads that require sub-millisecond latency, and sustained IOPS performance or more than 64,000 IOPS or 1,000 MiB/s of throughput		• Workloads that require sustained IOPS performance or more than 16,000 IOPS • I/O-intensive database workloads
Volume size	1 GiB - 16 TiB		4 GiB - 64 TiB		4 GiB - 16 TiB
Max IOPS per volume (16 KiB I/O)	16,000		256,000		64,000 †
Max throughput per volume	1,000 MiB/s	250 MiB/s *	4,000 MiB/s		1,000 MiB/s †
Amazon EBS Multi-attach	Not supported		Supported		
Boot volume	Supported				

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS HDD-Backed Volumes

	Throughput Optimized HDD	Cold HDD
Volume type	st1	sc1
Durability	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)
Use cases	• Big data • Data warehouses • Log processing	• Throughput-oriented storage for data that is infrequently accessed • Scenarios where the lowest storage cost is important
Volume size	125 GiB - 16 TiB	125 GiB - 16 TiB
Max IOPS per volume (1 MiB I/O)	500	250
Max throughput per volume	500 MiB/s	250 MiB/s
Amazon EBS Multi-attach	Not supported	Not supported
Boot volume	Not supported	Not supported

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS

- EBS volume data persists **independently** of the life of the instance
- EBS volumes do not need to be attached to an instance
- You can attach multiple EBS volumes to an instance
- You can use multi-attach to attach a volume to multiple instances but with some constraints
- EBS volumes must be in the **same AZ** as the instances they are attached to
- Root EBS volumes **are deleted** on termination by default
- Extra non-boot volumes **are not deleted** on termination by default

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS Copying, Sharing and Encryption

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS Copying, Sharing and Encryption

© Digital Cloud Training | https://digitalcloud.training

Take Snapshot, Create AMI, Launch New Instance

© Digital Cloud Training | https://digitalcloud.training

Copying and Sharing AMIs and Snapshots

Volume → Snapshot

- Encryption state retained
- Same region

Snapshot —Copy→ Encrypted Snapshot

- Can be encrypted
- Can change regions

Unencrypted Snapshot → Encrypted Volume

- Can be encrypted
- Can change AZ

Unencrypted Snapshot → AMI

- Cannot be encrypted
- Can be shared with other accounts
- Can be shared publicly

Snapshot —Copy→ Snapshot

- Can change encryption key
- Can change regions

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Copying and Sharing AMIs and Snapshots

Encrypted Snapshot → Encrypted AMI

- Can be shared with other accounts (custom key only)
- Cannot be shared publicly

Encrypted AMI —Copy→ Encrypted AMI

- Can change encryption key
- Can change region

Encrypted AMI → EC2 Instance

- Can change encryption key
- Can change AZ

Unencrypted AMI → EC2 Instance

- Can change encryption state
- Can change AZ

Encrypted Snapshot → Encrypted Volume

- Can be encrypted
- Can change AZ

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Amazon EBS Snapshots and DLM

© Digital Cloud Training | https://digitalcloud.training

Amazon EBS Snapshots

© Digital Cloud Training | https://digitalcloud.training

Amazon Data Lifecycle Manager (DLM)

- DLM automates the creation, retention, and deletion of EBS snapshots and EBS-backed AMIs

- DLM helps with the following:
  - Protects valuable data by enforcing a regular backup schedule
  - Create standardized AMIs that can be refreshed at regular intervals
  - Retain backups as required by auditors or internal compliance
  - Reduce storage costs by deleting outdated backups
  - Create disaster recovery backup policies that back up data to isolated accounts

© Digital Cloud Training | https://digitalcloud.training

EC2 Instance Store Volumes

© Digital Cloud Training | https://digitalcloud.training

EBS vs instance store

© Digital Cloud Training | https://digitalcloud.training

EBS vs instance store

- Instance store volumes are high performance local disks that are physically attached to the host computer on which an EC2 instance runs
- Instance stores are ephemeral which means the data is lost when powered off (non-persistent)
- Instance stores are ideal for temporary storage of information that changes frequently, such as buffers, caches, or scratch data
- Instance store volume root devices are created from AMI templates stored on S3
- Instance store volumes cannot be detached/reattached

© Digital Cloud Training | https://digitalcloud.training

Creating and Attaching EBS Volumes

© Digital Cloud Training | https://digitalcloud.training

Using Amazon Machine Images (AMIs)

© Digital Cloud Training | https://digitalcloud.training

Amazon Machine Images (AMIs)

- An **Amazon Machine Image** (AMI) provides the information required to launch an instance
- An AMI includes the following:
  - One or more EBS snapshots, or, for instance-store-backed AMIs, a template for the root volume of the instance (for example, an operating system, an application server, and applications)
  - Launch permissions that control which AWS accounts can use the AMI to launch instances
  - A block device mapping that specifies the volumes to attach to the instance when it's launched
- AMIs come in three main categories:
  - **Community AMIs** - free to use, generally you just select the operating system you want
  - **AWS Marketplace AMIs** - pay to use, generally come packaged with additional, licensed software
  - **My AMIs** - AMIs that you create yourself

© Digital Cloud Training | https://digitalcloud.training

EC2 Image Builder

© Digital Cloud Training | https://digitalcloud.training

Migrate EBS Volume between AZs

© Digital Cloud Training | https://digitalcloud.training

Working with EBS Snapshots and DLM

© Digital Cloud Training | https://digitalcloud.training

Using RAID with EBS

© Digital Cloud Training | https://digitalcloud.training

Using RAID with EBS

- RAID stands for Redundant Array of Independent disks
- Not provided by AWS, you must configure through your operating system
- RAID 0 and RAID 1 are potential options on EBS
- RAID 5 and RAID 6 are not recommended by AWS

© Digital Cloud Training | https://digitalcloud.training

Using RAID with EBS

- RAID 0 is used for striping data across disks (performance):
  - Use 2 or more disks
  - If one disk fails, the entire RAID set fails

© Digital Cloud Training | https://digitalcloud.training

Using RAID with EBS

- RAID 1 is used for mirroring data across disks (redundancy / fault tolerance):
  - If one disk fails, the other disk is still working
  - Data gets sent to 2 EBS volumes at the same time

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic File System (EFS)

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic File System (EFS) Overview

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic File System (EFS) Overview

© Digital Cloud Training | https://digitalcloud.training

Create EFS Filesystem

© Digital Cloud Training | https://digitalcloud.training

Amazon FSx

© Digital Cloud Training | https://digitalcloud.training

FSx

Amazon FSx

- Amazon FSx provides fully managed third-party file systems
- Amazon FSx provides you with two file systems to choose from:
  - **Amazon FSx for Windows File Server** for Windows-based applications
  - **Amazon FSx for Lustre** for compute-intensive workloads

© Digital Cloud Training | https://digitalcloud.training

FSx

Amazon FSx for Windows File Server

- Provides a fully managed native Microsoft Windows file system
- Full support for the SMB protocol, Windows NTFS, and Microsoft Active Directory (AD) integration
- Supports Windows-native file system features:
  - Access Control Lists (ACLs), shadow copies, and user quotas.
  - NTFS file systems that can be accessed from up to thousands of compute instances using the SMB protocol
- **High availability:** replicates data within an Availability Zone (AZ)
- **Multi-AZ:** file systems include an active and standby file server in separate AZs

DigitalCloud TRAINING

Amazon FSx for Windows File Server

© Digital Cloud Training | https://digitalcloud.training

Amazon FSx for Lustre

- High-performance file system optimized for fast processing of workloads such as:
  - Machine learning
  - High performance computing (HPC)
  - Video processing
  - Financial modeling
  - Electronic design automation (EDA)
- Works natively with S3, letting you transparently access your S3 objects as files
- Your S3 objects are presented as files in your file system, and you can write your results back to S3
- Provides a POSIX-compliant file system interface

© Digital Cloud Training | https://digitalcloud.training

Amazon FSx for Lustre

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway – File Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway – File Gateway

- File gateway provides a virtual **on-premises file server**
- Store and retrieve files as objects in Amazon S3
- Use with on-premises applications, and EC2-based applications that need file storage in S3 for object-based workloads
- File gateway offers **SMB** or **NFS**-based access to data in Amazon S3 with local caching

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway - Volume Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway - Volume Gateway

- The volume gateway supports block-based volumes
- Block storage – iSCSI protocol
- **Cached Volume mode** – the entire dataset is stored on S3 and a cache of the most frequently accessed data is cached on-site
- **Stored Volume mode** – the entire dataset is stored on-site and is asynchronously backed up to S3 (EBS point-in-time snapshots). Snapshots are incremental and compressed

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway - Tape Gateway

© Digital Cloud Training | https://digitalcloud.training

AWS Storage Gateway - Tape Gateway

- Used for backup with popular backup software
- Each gateway is preconfigured with a media changer and tape drives. Supported by NetBackup, Backup Exec, Veeam etc.
- When creating virtual tapes, you select one of the following sizes: 100 GB, 200 GB, 400 GB, 800 GB, 1.5 TB, and 2.5 TB
- A tape gateway can have up to 1,500 virtual tapes with a maximum aggregate capacity of 1 PB
- All data transferred between the gateway and AWS storage is encrypted using SSL
- All data stored by tape gateway in S3 is encrypted server-side with Amazon S3-Managed Encryption Keys (SSE-S3)

© Digital Cloud Training | https://digitalcloud.training

AWS Backup

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Block and File Storage

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Block and File Storage

Requirement	Solution
Simple method of backing up Amazon EBS volumes is required that is fully automated	Use Data Lifecycle Manager to create a backup schedule
A distributed application has many nodes that each hold a copy of data that is synchronized between them. Need the best performance	Use instance stores for storing the data with the best performance
Application must startup quickly when launched by ASG but requires app dependencies and code to be installed	Create an AMI that includes the application dependencies and code

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Block and File Storage

Requirement	Solution
Many Linux instances must be attached to a shared filesystem that scales elastically	Create an Amazon EFS file system and mount from each instance
Company requires a managed file system that uses the NTFS file system	Use Amazon FSx for Windows File Server
On-premises servers must be able to attach a block storage system locally. Data should be backed up to S3 as snapshots	Deploy an AWS Storage Gateway volume gateway in stored volume mode

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Block and File Storage

Requirement	Solution
An Amazon EBS volume must be moved between Regions	Take a snapshot and copy the snapshot between Regions
Root EBS volumes for a critical application must not be deleted on termination	Modify the delete on termination attribute when launching the EC2 instances
On-premises servers use NFS to attach a file system. The file system should be replaced with an AWS service that uses Amazon S3 with a local cache	Deploy an AWS Storage Gateway file gateway

© Digital Cloud Training | https://digitalcloud.training

SECTION 10

Docker Containers and ECS

© Digital Cloud Training | https://digitalcloud.training

Docker Containers and Microservices

© Digital Cloud Training | https://digitalcloud.training

Server Virtualization vs Containers

Every VM/instance needs an **operating system** which uses significant resources

© Digital Cloud Training | https://digitalcloud.training

Docker Containers

Containers **start up** very **quickly**

A **container** includes all the code, settings, and dependencies for running the application

Containers are very resource **efficient**

Each container is **isolated** from other containers

Monolithic Application

© Digital Cloud Training | https://digitalcloud.training

Monolithic Application

Updates to, or failures of, any single component can take down the whole application

The user interface, business logic, and data access layer are combined on a single platform

© Digital Cloud Training | https://digitalcloud.training

Microservices Application

© Digital Cloud Training | https://digitalcloud.training

Microservices Application

© Digital Cloud Training | https://digitalcloud.training

Microservices Application

Microservices can also be **spread** across hosts

Many instances of each microservice can run on each host

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic Container Service (ECS)

© Digital Cloud Training | https://digitalcloud.training

Amazon ECS

```
{
"containerDefinitions": [
                {
                  "name": "wordpress",
                  "links": [
                    "mysql"
                ],
                "image": "wordpress",
                "essential": true,
                "portMappings": [
                  {
                    "containerPort": 80,
                    "hostPort": 80
                     }
                  ],
                  "memory": 500,
                  "cpu": 10
                }
```

© Digital Cloud Training | https://digitalcloud.training

Amazon ECS Key Features

- **Serverless with AWS Fargate** – managed for you and fully scalable
- **Fully managed container orchestration** – control plane is managed for you
- **Docker support** – run and manage Docker containers with integration into the Docker Compose CLI
- **Windows container support** – ECS supports management of Windows containers
- **Elastic Load Balancing integration** – distribute traffic across containers using ALB or NLB
- **Amazon ECS Anywhere (NEW)** – enables the use of Amazon ECS control plane to manage on-premises implementations

© Digital Cloud Training | https://digitalcloud.training

Amazon ECS Components

Elastic Container Service (ECS)	Description
Cluster	Logical grouping of EC2 instances
Container instance	EC2 instance running the the ECS agent
Task Definition	Blueprint that describes how a docker container should launch
Task	A running container using settings in a Task Definition
Service	Defines long running tasks – can control task count with Auto Scaling and attach an ELB

© Digital Cloud Training | https://digitalcloud.training

Amazon ECS Launch Types

© Digital Cloud Training | https://digitalcloud.training

Launch Types – EC2 and Fargate

EC2 Launch Type

- You explicitly provision EC2 instances
- You're responsible for managing EC2 instances
- Charged per running EC2 instance
- Docker volumes, EFS, and FSx for Windows File Server
- You handle cluster optimization
- More granular control over infrastructure

Fargate Launch Type

- Fargate automatically provisions resources
- Fargate provisions and manages compute
- Charged for running tasks
- EFS integration
- Fargate handles cluster optimization
- Limited control, infrastructure is automated

© Digital Cloud Training | https://digitalcloud.training

Launch Task on AWS Fargate

© Digital Cloud Training | https://digitalcloud.training

Amazon ECS and IAM Roles

© Digital Cloud Training | https://digitalcloud.training

ECS and IAM Roles

NOTE: container instances have access to **all of the permissions** that are supplied to the **container instance role** through instance metadata

AmazonEC2ContainerServiceforEC2Role

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ec2:DescribeTags",
                "ecs:CreateCluster",
                "ecs:DeregisterContainerInstance",
                "ecs:DiscoverPollEndpoint",
                "ecs:Poll",
                "ecs:RegisterContainerInstance",
                "ecs:StartTelemetrySession",
                "ecs:UpdateContainerInstancesState",
                "ecs:Submit*",
                "ecr:GetAuthorizationToken",
                "ecr:BatchCheckLayerAvailability",
                "ecr:GetDownloadUrlForLayer",
                "ecr:BatchGetImage",
                "logs:CreateLogStream",
                "logs:PutLogEvents"
            ],
            "Resource": "*"
        }
    ]
}
```

© Digital Cloud Training | https://digitalcloud.training

ECS and IAM Roles

Scaling Amazon ECS

© Digital Cloud Training | https://digitalcloud.training

Auto Scaling for ECS

Two types of scaling:

1. Service auto scaling
2. Cluster auto scaling

- **Service auto scaling** automatically adjusts the desired task count up or down using the Application Auto Scaling service
- **Service auto scaling** supports target tracking, step, and scheduled scaling policies
- **Cluster auto scaling** uses a Capacity Provider to scale the number of EC2 cluster instances using EC2 Auto Scaling

© Digital Cloud Training | https://digitalcloud.training

Service Auto Scaling

- 1. Metric reports CPU > 80%
- 2. CloudWatch notifies Application Auto Scaling
- 3. ECS launches additional task

© Digital Cloud Training | https://digitalcloud.training

Service Auto Scaling

- Amazon ECS Service Auto Scaling supports the following types of scaling policies:
  - **Target Tracking Scaling Policies**—Increase or decrease the number of tasks that your service runs based on a target value for a specific CloudWatch metric
  - **Step Scaling Policies**—Increase or decrease the number of tasks that your service runs in response to CloudWatch alarms. Step scaling is based on a set of scaling adjustments, known as step adjustments, which vary based on the size of the alarm breach
  - **Scheduled Scaling**—Increase or decrease the number of tasks that your service runs based on the date and time

DigitalCloud TRAINING

Cluster Auto Scaling

- 1. Metric reports target capacity > 80%
- 2. CloudWatch notifies ASG
- 3. AWS launches additional container instance

A **Capacity provider reservation** metric measures the total percentage of cluster resources needed by all ECS workloads in the cluster

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Cluster Auto Scaling

- Uses an ECS resource type called a **Capacity Provider**
- A Capacity Provider can be associated with an EC2 **Auto Scaling Group** (ASG)
- ASG can automatically scale using:
  - **Managed scaling** - with an automatically-created scaling policy on your ASG
  - **Managed instance termination protection** - which enables container-aware termination of instances in the ASG when scale-in happens

© Digital Cloud Training | https://digitalcloud.training

Amazon ECS with ALB

© Digital Cloud Training | https://digitalcloud.training

Amazon ECS with ALB

© Digital Cloud Training | https://digitalcloud.training

Deploy ECS Cluster (EC2 Launch Type)

© Digital Cloud Training | https://digitalcloud.training

Use ALB with Fargate Cluster

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic Kubernetes Service (EKS)

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic Kubernetes Service (EKS)

© Digital Cloud Training | https://digitalcloud.training

Amazon EKS Use Cases

- Use when you need to **standardize** container orchestration across multiple environments using a **managed Kubernetes** implementation
- **Hybrid Deployment** - manage Kubernetes clusters and applications across hybrid environments (AWS + On-premises)
- **Batch Processing** - run sequential or parallel batch workloads on your EKS cluster using the Kubernetes Jobs API. Plan, schedule and execute batch workloads
- **Machine Learning** - use Kubeflow with EKS to model your machine learning workflows and efficiently run distributed training jobs using the latest EC2 GPU-powered instances, including Inferentia
- **Web Applications** - build web applications that automatically scale up and down and run in a highly available configuration across multiple Availability Zones

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon ECS

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon ECS

Requirement	Solution
Application will be deployed on Amazon ECS and must scale based on memory	Use service auto-scaling and use the memory utilization
Application will run on Amazon ECS tasks across multiple hosts and needs access to an Amazon S3 bucket	Use a task execution IAM role to provide permissions to S3 bucket.
Company requires standard Docker container automation and management service to be used across multiple environments	Deploy Amazon EKS

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon ECS

Requirement	Solution
Company plans to deploy Docker containers on AWS at the lowest cost	Use Amazon ECS with a cluster of Spot instances and enable Spot instance draining
Company plans to migrate Docker containers to AWS and does not want to manage operating systems	Migrate to Amazon ECS using the Fargate launch type
Multiple microservices applications running on Amazon ECS. Need to route based on information in the HTTP header	Deploy an Application Load Balancer in front of ECS and use query string parameter based routing

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Amazon ECS

Requirement

Containerized app runs on Amazon EKS. Need to collect and centrally view metrics and logs including EKS namespaces and EKS services

Solution

Configure CloudWatch Container Insights and view data in CloudWatch console

© Digital Cloud Training | https://digitalcloud.training

SECTION 11

Serverless Applications

© Digital Cloud Training | https://digitalcloud.training

Serverless Services and Event-Driven Architecture

© Digital Cloud Training | https://digitalcloud.training

Serverless Services and Event-Driven Architecture

© Digital Cloud Training | https://digitalcloud.training

Serverless Services

- With serverless there are **no instances** to manage
- You don’t need to provision hardware
- There is no management of operating systems or software
- Capacity provisioning and patching is handled automatically
- Provides automatic scaling and high availability
- Can be very cheap!

© Digital Cloud Training | https://digitalcloud.training

AWS Lambda

© Digital Cloud Training | https://digitalcloud.training

AWS Lambda

© Digital Cloud Training | https://digitalcloud.training

AWS Lambda

- AWS Lambda executes code only when needed and scales automatically
- You pay only for the compute time you consume (you pay nothing when your code is not running)
- Benefits of AWS Lambda:
  - No servers to manage
  - Continuous scaling
  - Millisecond billing
  - Integrates with almost all other AWS services

© Digital Cloud Training | https://digitalcloud.training

AWS Lambda

- Primary use cases for AWS Lambda:
  - Data processing
  - Real-time file processing
  - Real-time stream processing
  - Build serverless backends for web, mobile, IOT, and 3rd party API requests

© Digital Cloud Training | https://digitalcloud.training

Lambda Function Invocations

Synchronous:

- CLI, SDK, API Gateway
- Wait for the function to process the event and return a response
- Error handling happens client side (retries, exponential backoff etc.)

Asynchronous:

- S3, SNS, CloudWatch Events etc.
- Event is queued for processing and a response is returned immediately
- Lambda retries up to 3 times
- Processing must be idempotent (due to retries)

Event source mapping:

- SQS, Kinesis Data Streams, DynamoDB Streams
- Lambda does the polling (polls the source)
- Records are processed in order (except for SQS standard)

SQS can also **trigger** Lambda

© Digital Cloud Training | https://digitalcloud.training

Lambda Function Concurrency

Additional functions are initialized up to the **burst** or **account limit**

If the concurrency **limit** is **exceeded** throttling occurs with error **"Rate exceeded"** and a 429 **"TooManyRequestsException"**

Burst **concurrency** quotas:

- 3000 – US West (Oregon), US East (N. Virginia), Europe (Ireland)
- 1000 – Asia Pacific (Tokyo), Europe (Frankfurt), US East (Ohio)
- 500 – Other Regions

© Digital Cloud Training | https://digitalcloud.training

Create Simple Lambda Function

© Digital Cloud Training | https://digitalcloud.training

Create a Simple Lambda Function

© Digital Cloud Training | https://digitalcloud.training

Application Integration Services Overview

© Digital Cloud Training | https://digitalcloud.training

Application Integration Services Overview

Service	What it does	Example use cases
Simple Queue Service	Messaging queue; store and forward patterns	Building distributed / decoupled applications
Simple Notification Service	Set up, operate, and send notifications from the cloud	Send email notification when CloudWatch alarm is triggered
Step Functions	Out-of-the-box coordination of AWS service components with visual workflow	Order processing workflow
Simple Workflow Service	Need to support external processes or specialized execution logic	Human-enabled workflows like an order fulfilment system or for procedural requests Note: AWS recommends that for new applications customers consider Step Functions instead of SWF
Amazon MQ	Message broker service for Apache Active MQ and RabbitMQ	Need a message queue that supports industry standard APIs and protocols; migrate queues to AWS
Amazon Kinesis	Collect, process, and analyze streaming data.	Collect data from IoT devices for later processing

Kinesis vs SQS vs SNS

Amazon Kinesis	Amazon SQS	Amazon SNS
Consumers pull data	Consumers pull data	Push data to many subscribers
As many consumers as you need	Data is deleted after being consumed	Publisher / subscriber model
Routes related records to same record processor	Can have as many workers (consumers) as you need	Integrates with SQS for fan-out architecture pattern
Multiple applications can access stream concurrently	No ordering guarantee (except with FIFO queues)	Up to 10,000,000 subscribers
Ordering at the shard level	Provides messaging semantics	Up to 100,000 topics
Can consume records in correct order at later time	Individual message delay	Data is not persisted
Must provision throughput	Dynamically scales	No need to provision throughput

© Digital Cloud Training | https://digitalcloud.training

Amazon SQS

© Digital Cloud Training | https://digitalcloud.training

Decoupling with SQS Queues

© Digital Cloud Training | https://digitalcloud.training

SQS Queue Types

© Digital Cloud Training | https://digitalcloud.training

SQS Queue Types

Standard Queue	FIFO Queue
Unlimited Throughput: Standard queues support a nearly unlimited number of transactions per second (TPS) per API action.	**High Throughput:** FIFO queues support up to 300 messages per second (300 send, receive, or delete operations per second). When you batch 10 messages per operation (maximum), FIFO queues can support up to 3,000 messages per second
Best-Effort Ordering: Occasionally, messages might be delivered in an order different from which they were sent	**First-In-First-out Delivery:** The order in which messages are sent and received is strictly preserved
At-Least-Once Delivery: A message is delivered at least once, but occasionally more than one copy of a message is delivered	**Exactly-Once Processing:** A message is delivered once and remains available until a consumer processes and deletes it. Duplicates are not introduced into the queue

© Digital Cloud Training | https://digitalcloud.training

SQS Queue Types

- FIFO queues require the **Message Group ID** and **Message Deduplication ID** parameters to be added to messages
- **Message Group ID**:
  - The tag that specifies that a message belongs to a specific message group Messages that belong to the same message group are guaranteed to be processed in a FIFO manner
- **Message Deduplication ID**:
  - The token used for deduplication of messages within the deduplication interval

© Digital Cloud Training | https://digitalcloud.training

SQS – Dead Letter Queue

Message not processed successfully (**ReceiveCount** exceeds **maxReceiveCount** for queue)

Dead-letter queue is a **standard** or **FIFO** queue that has been specified as a **dead-letter queue**

© Digital Cloud Training | https://digitalcloud.training

SQS – Dead Letter Queue

- The main task of a dead-letter queue is handling message failure
- A dead-letter queue lets you set aside and isolate messages that can't be processed correctly to determine why their processing didn't succeed
- It is not a queue type, it is a **standard** or **FIFO** queue that has been specified as a dead-letter queue in the configuration of **another** standard or FIFO queue

Enable Redrive Policy

Specify the queue to use as a dead-letter queue

Specify the maximum receives before a message is sent to the dead-letter queue

© Digital Cloud Training | https://digitalcloud.training

SQS – Delay Queue

© Digital Cloud Training | https://digitalcloud.training

SQS Long Polling vs Short Polling

Long polling waits for the `WaitTimeSeconds` and eliminates empty responses

Short polling checks a subset of servers and may not return all messages

© Digital Cloud Training | https://digitalcloud.training

SQS Long Polling vs Short Polling

- SQS **Long polling** is a way to retrieve messages from SQS queues – waits for messages to arrive
- SQS **Short polling** returns immediately (even if the message queue is empty)
- SQS Long polling can lower costs
- SQS Long polling can be enabled at the queue level or at the API level using `WaitTimeSeconds`
- SQS Long polling is in effect when the Receive Message Wait Time is a value greater than 0 seconds and up to 20 seconds

The maximum amount of time that a long polling receive call will wait for a message to become available before returning an empty response.

© Digital Cloud Training | https://digitalcloud.training

Amazon SNS

© Digital Cloud Training | https://digitalcloud.training

Amazon SNS Notifications

© Digital Cloud Training | https://digitalcloud.training

Amazon SNS

- Amazon SNS is a highly available, durable, secure, fully managed pub/sub messaging service
- Amazon SNS provides topics for high-throughput, push-based, many-to-many messaging
- Publisher systems can fan out messages to a large number of subscriber endpoints:
- Endpoints include:
  - Amazon SQS queues
  - AWS Lambda functions
  - HTTP/S webhooks
  - Mobile push
  - SMS
  - Email

© Digital Cloud Training | https://digitalcloud.training

Amazon SNS

- Multiple recipients can be grouped using Topics
- A topic is an “access point” for allowing recipients to dynamically subscribe for identical copies of the same notification
- One topic can support deliveries to multiple endpoint types
- Simple APIs and easy integration with applications
- Flexible message delivery over multiple transport protocols

© Digital Cloud Training | https://digitalcloud.training

Amazon SNS + Amazon SQS Fan-Out

- You can subscribe one or more Amazon SQS queues to an Amazon SNS topic
- Amazon SQS manages the subscription and any necessary permissions
- When you publish a message to a topic, Amazon SNS sends the message to every subscribed queue

© Digital Cloud Training | https://digitalcloud.training

Simple Event-Driven App

© Digital Cloud Training | https://digitalcloud.training

Simple Event-Driven App

```
aws sqs send-message --queue-url https://sqs.us-east-1.amazonaws.com/778642014689/
ProductVisitsDataQueue --message-body file://message-body.json
```

Manually add message to queue with **AWS CLI**

CW Logs

SQS Queue

productVisitsDataHandler

DynamoDB Table

Item is written to **DynamoDB table**

SQS triggers Lambda

```
JSON
{
  "ProductId": "c96b49bb-c378-4a15-b2e3-842a9850b23d",
  "ProductName": "Gloves",
  "Category": "Accessories",
  "PricePerUnit": "10",
  "CustomerId": "be44af0a-74f9-438e-a3ac-e3e21d84259f",
  "CustomerName": "John Doe",
  "TimeOfVisit": "2021-02-21T16:23:42.389Z"
}
```

This is the **message body**

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

AWS Step Functions

© Digital Cloud Training | https://digitalcloud.training

AWS Step Functions

© Digital Cloud Training | https://digitalcloud.training

AWS Step Functions

- AWS Step Functions is used to build distributed applications as a series of steps in a visual workflow.
- You can quickly build and run state machines to execute the steps of your application

How it works:

1. Define the steps of your workflow in the **JSON-based Amazon States Language**. The visual console automatically graphs each step in the order of execution
2. Start an execution to visualize and verify the steps of your application are operating as intended. The console highlights the real-time status of each step and provides a detailed history of every execution
3. AWS Step Functions **operates and scales** the steps of your **application** and **underlying compute** for you to help ensure your application executes reliably under increasing demand

© Digital Cloud Training | https://digitalcloud.training

Create a State Machine

© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge

© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge

EventBridge used to be known as **CloudWatch Events**

© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge Example 1

© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge Example 1

Event matching pattern

You can use pre-defined pattern provided by a service or create a custom pattern

- (●) Pre-defined pattern by service
- ( ) Custom pattern

Service provider

AWS services or custom/partner services

AWS

Service name

The name of partner service selected as the event source

EC2

Event type

The type of events as the source of the matching pattern

EC2 Instance State-change Notification

- ( ) Any state
- (●) Specific state(s)

terminated ×

- ( ) Any instance
- (●) Specific instance Id(s)

i-1234567890abcdef0 Remove

Add

Event pattern Copy Edit

```
1 {
2   "source": ["aws.ec2"],
3   "detail-type": ["EC2 Instance State-change N
4   "detail": {
5     "state": ["terminated"],
6     "instance-id": ["i-1234567890abcdef0"]
7   }
8 }
```


```
{
  "version": "0",
  "id": "6a7e8feb-b491-4cf7-a9f1-bf3703467718",
  "detail-type": "EC2 Instance State-change Notification",
  "source": "aws.ec2",
  "account": "111122223333",
  "time": "2017-12-22T18:43:48Z",
  "region": "us-west-1",
  "resources": [
    "arn:aws:ec2:us-west-1:123456789012:instance/i-1234567890abcdef0"
  ],
  "detail": {
    "instance-id": "i-1234567890abcdef0",
    "state": "terminated"
    }
  }
```


© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge Example 2

```
{
   "source": ["aws.cloudtrail"],
   "detail-type": ["AWS API Call via CloudTrail"],
   "detail": {
     "eventSource": ["cloudtrail.amazonaws.com"],
     "eventName": ["s3:PutBucketPolicy"]
   }
 }
```

© Digital Cloud Training | https://digitalcloud.training

Create Event Bus and Rule

© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge Example 1

Event matching pattern

You can use pre-defined pattern provided by a service or create a custom pattern

- (●) Pre-defined pattern by service
- ( ) Custom pattern

Service provider

AWS services or custom/partner services

AWS

Service name

The name of partner service selected as the event source

EC2

Event type

The type of events as the source of the matching pattern

EC2 Instance State-change Notification

- ( ) Any state
- (●) Specific state(s)

terminated ×

- ( ) Any instance
- (●) Specific instance Id(s)

i-1234567890abcdef0 Remove

Add

Event pattern Copy Edit

```
1 {
2   "source": ["aws.ec2"],
3   "detail-type": ["EC2 Instance State-change N
4   "detail": {
5     "state": ["terminated"],
6     "instance-id": ["i-1234567890abcdef0"]
7   }
8 }
```


```
{
  "version": "0",
  "id": "6a7e8feb-b491-4cf7-a9f1-bf3703467718",
  "detail-type": "EC2 Instance State-change Notification",
  "source": "aws.ec2",
  "account": "111122223333",
  "time": "2017-12-22T18:43:48Z",
  "region": "us-west-1",
  "resources": [
    "arn:aws:ec2:us-west-1:123456789012:instance/i-1234567890abcdef0"
  ],
  "detail": {
    "instance-id": "i-1234567890abcdef0",
    "state": "terminated"
    }
  }
```


© Digital Cloud Training | https://digitalcloud.training

Amazon API Gateway

© Digital Cloud Training | https://digitalcloud.training

Amazon API Gateway Overview

© Digital Cloud Training | https://digitalcloud.training

Amazon API Gateway Deployment Types

Edge-optimized endpoint

Key benefits:
- Reduced latency for requests from around the world

Regional endpoint

Key benefits:
- Reduced latency for requests that originate in the same region
- Can also configure your own CDN and protect with WAF

Private endpoint

Key benefits:
- Securely expose your REST APIs only to other services within your VPC or connect via Direct Connect

© Digital Cloud Training | https://digitalcloud.training

Structure of a REST API

© Digital Cloud Training | https://digitalcloud.training

API Gateway Integrations

For a **Lambda function** you can have:

- Lambda proxy integration
- Lambda custom integration

For an **HTTP endpoint** you can have:

- HTTP proxy integration
- HTTP custom integration

- For an **AWS service action** you have the AWS integration of the non-proxy type only

© Digital Cloud Training | https://digitalcloud.training

API Gateway - Caching

- You can add caching to API calls by provisioning an Amazon API Gateway cache and specifying its size in gigabytes
- Caching allows you to cache the endpoint's response
- Caching can reduce number of calls to the backend and improve latency of requests to the API

© Digital Cloud Training | https://digitalcloud.training

API Gateway - Throttling

- API Gateway sets a limit on a steady-state rate and a burst of request submissions against all APIs in your account
- Limits:
  - By default API Gateway limits the steady-state request rate to 10,000 requests per second
  - The maximum concurrent requests is 5,000 requests across all APIs within an AWS account
  - If you go over 10,000 requests per second or 5,000 concurrent requests you will receive a 429 Too Many Requests error response
- Upon catching such exceptions, the client can resubmit the failed requests in a way that is rate limiting, while complying with the API Gateway throttling limits

© Digital Cloud Training | https://digitalcloud.training

API Gateway – Usage Plans and API Keys

Users connect to **specific public endpoint** with **API key** that is configured in a usage plan

Can also configure **per-method** throttling limits

© Digital Cloud Training | https://digitalcloud.training

Simple REST API

© Digital Cloud Training | https://digitalcloud.training

Web App with HTTP API

© Digital Cloud Training | https://digitalcloud.training

Web App with HTTP API

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Serverless

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Serverless

Requirement	Solution
Application includes EC2 and RDS. Spikes in traffic causing writes to be dropped by RDS	Decouple EC2 and RDS database with an SQS queue; use Lambda to process records in the queue
Web app includes a web tier and processing tier. Must be decoupled and processing tier should dynamically scale based on number of jobs	Decouple the web tier and processing tier with SQS and scale with Auto Scaling based on the queue length
Lambda function execution time has increased significantly as the payload size increased.	Optimize execution time by increasing memory available to the function which will proportionally increase CPU

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Serverless

Requirement	Solution
Statistical data is stored in RDS and will be accessed using a REST API. Demand will range from no traffic to sudden bursts of traffic and is unpredictable	Create a REST API using Amazon API Gateway and integrate with an AWS Lambda function for connecting to RDS
New application processes customer orders and consists of multiple decoupled tiers. Orders must be processed in the order they are received	Implement an Amazon SQS FIFO queue to preserve the record order
App uses API Gateway and Lambda. During busy periods, many requests fail multiple times before succeeding. No errors reported in Lambda	Throttle limit could be configured with a value that is too low. Increase the throttle limit

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Serverless

Requirement	Solution
EC2 instance processes images using JavaScript code and stores results in S3. Load is highly variable. Need a more cost-effective solution	Replace EC2 with AWS Lambda function
App uses API Gateway regional REST API. Just gone global and performance has suffered	Convert API to an edge-optimized API to optimize for the global user base
Legacy application uses many batch scripts that process data and pass on to next script. Complex and difficult to maintain	Migrate scripts to AWS Lambda functions and use AWS Step Functions to coordinate components

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Serverless

Requirement

Objects uploaded to an S3 bucket must be processed by AWS Lambda

Solution

Create an event source notification to notify Lambda function to process new objects

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

SECTION 12

Databases and Analytics

© Digital Cloud Training | https://digitalcloud.training

Types of Database

© Digital Cloud Training | https://digitalcloud.training

Relational vs Non-Relational

Key differences are how data are ***managed*** and how data are ***stored***

Relational	Non-Relational
Organized by tables, rows and columns	Varied data storage models
Rigid schema (SQL)	Flexible schema (NoSQL) – data stored in key-value pairs, columns, documents or graphs
Rules enforced within database	Rules can be defined in application code (outside database)
Typically scaled vertically	Scales horizontally
Supports complex queries and joins	Unstructured, simple language that supports any kind of schema
Amazon RDS, Oracle, MySQL, IBM DB2, PostgreSQL	Amazon DynamoDB, MongoDB, Redis, Neo4j

© Digital Cloud Training | https://digitalcloud.training

Relational Database

EmployeeID	FirstName	LastName	JobRole	Location
00001	Paul	Peterson	Senior Developer	Sydney
00002	Kaleigh	Annette	Assistant Manager	Brisbane
00003	Carl	Wood	Sales Support	Sydney
00004	Vinni	Jones	Customer Services	Melbourne
00005	Stefanie	Howard	IT Architect	Brisbane

Relational Database

Relational Database – Multiple Tables

Structured Query Language (SQL) query:

```
SELECT FirstName
FROM employees
WHERE Location = Sydney
```

© Digital Cloud Training | https://digitalcloud.training

Types of Non-Relational DB (NoSQL)

Key-value – e.g. Amazon DynamoDB

Document – e.g. MongoDB

```JSON
[
    {
        "year" : 2013,
        "title" : "Turn It Down, Or Else!",
        "info" : {
            "directors" : [ "Alice Smith", "Bob Jones"],
            "release_date" : "2013-01-18T00:00:00Z",
            "rating" : 6.2,
            "genres" : ["Comedy", "Drama"],
            "image_url" : "http://ia.media-imdb.com/images/N/O9ERWAU7FS797AJ7LU8HN09AMUP908RLlo5JF90EWR7LJKQ7@@._V1_SX400_.jpg",
            "plot" : "A rock band plays their music at high volumes, annoying the neighbors.",
            "actors" : ["David Matthewman", "Jonathan G. Neff"]
        }
    },
    {
        "year": 2015,
        "title": "The Big New Movie",
        "info": {
            "plot": "Nothing happens at all.",
            "rating": 0
        }
    }
]
```

Graph – e.g. Amazon Neptune

© Digital Cloud Training | https://digitalcloud.training

Operational vs Analytical

Key differences are ***use cases*** and how the database is ***optimized***

Operational / transactional	Analytical
Online Transaction Processing (OLTP)	Online Analytics Processing (OLAP) – the source data comes from OLTP DBs
Production DBs that process transactions. E.g. adding customer records, checking stock availability (INSERT, UPDATE, DELETE)	Data warehouse. Typically, separated from the customer facing DBs. Data is extracted for decision making
Short transactions and simple queries	Long transactions and complex queries
Relational examples: Amazon RDS, Oracle, IBM DB2, MySQL	Relational examples: Amazon RedShift, Teradata, HP Vertica
Non-relational examples: MongoDB, Cassandra, Neo4j, HBase	Non-relational examples: Amazon EMR, MapReduce

© Digital Cloud Training | https://digitalcloud.training

Operational vs Analytical

© Digital Cloud Training | https://digitalcloud.training

AWS Databases

Data Store	Use Case
Database on EC2	• Need full control over instance and database • Third-party database engine (not available in RDS)
Amazon RDS	• Need traditional relational database • e.g. Oracle, PostgreSQL, Microsoft SQL, MariaDB, MySQL • Data is well-formed and structured
Amazon DynamoDB	• NoSQL database • In-memory performance • High I/O needs • Dynamic scaling
Amazon RedShift	• Data warehouse for large volumes of aggregated data
Amazon ElastiCache	• Fast temporary storage for small amounts of data • In-memory database
Amazon EMR	• Analytics workloads using the Hadoop framework

© Digital Cloud Training | https://digitalcloud.training

Amazon Relational Database Service (RDS)

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS

RDS runs on **EC2 instances**, so you must choose an **instance type**

RDS is a **managed**, relational database

RDS supports the following database engines:

- Amazon Aurora
- MySQL
- MariaDB
- Oracle
- Microsoft SQL Server
- PostgreSQL

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Scaling Up (vertically)

© Digital Cloud Training | https://digitalcloud.training

Disaster Recovery (DR) and Scaling Out (Horizontally)

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS

- RDS uses EC2 instances, so you must choose an instance family/type
- Relational databases are known as Structured Query Language (SQL) databases
- RDS is an Online Transaction Processing (OLTP) type of database
- Easy to setup, highly available, fault tolerant, and scalable
- Common use cases include online stores and banking systems
- You can encrypt your Amazon RDS instances and snapshots at rest by enabling the encryption option for your Amazon RDS DB instance (during creation)
- Encryption uses AWS Key Management Service (KMS)

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS

- Amazon RDS supports the following database engines:
  - SQL Server
  - Oracle
  - MySQL Server
  - PostgreSQL
  - Aurora
  - MariaDB
- Scales up by increasing instance size (compute and storage)
- Read replicas option for read heavy workloads (scales out for reads/queries only)
- Disaster recovery with Multi-AZ option

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Backup and Recovery

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Automated Backups

Restore can be to **any** point in time during the **retention period**

New DB Instance

Restore

Backup window Info

Select the period you want automated backups of the database to be created by Amazon RDS.

- (●) Select window
- ( ) No preference

Start time				Duration	
00	:	00	UTC	0.5	hours

DB Instance

Backup

Snapshot

Retention is **0–35** days

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Amazon RDS Manual Backups (Snapshot)

- Backs up the entire DB instance, not just individual databases
- For single-AZ DB instances there is a brief suspension of I/O
- For Multi-AZ SQL Server, I/O activity is briefly suspended on primary
- For Multi-AZ MariaDB, MySQL, Oracle and PostgreSQL the snapshot is taken from the standby
- Snapshots do not expire (no retention period)

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Maintenance Windows

- Operating system and DB patching can require taking the database offline
- These tasks take place during a maintenance window
- By default a weekly maintenance window is configured
- You can choose your own maintenance window

© Digital Cloud Training | https://digitalcloud.training

Create Amazon RDS Database

© Digital Cloud Training | https://digitalcloud.training

Read Replicas and Multi-AZ

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Security

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Security

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

Amazon RDS Security

- Encryption **at rest** can be enabled – includes DB storage, backups, read replicas and snapshots
- You can only enable encryption for an Amazon RDS DB instance when you create it, not after the DB instance is created
- DB instances that are encrypted can't be modified to disable encryption
- Uses AES 256 encryption and encryption is transparent with minimal performance impact
- RDS for Oracle and SQL Server is also supported using Transparent Data Encryption (TDE) (may have performance impact)
- AWS KMS is used for managing encryption keys

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Security

- You can't have:
  - An **encrypted** read replica of an **unencrypted** DB instance
  - An **unencrypted** read replica of an **encrypted** DB instance
- Read replicas of encrypted primary instances are encrypted
- The same KMS key is used if in the same Region as the primary
- If the read replica is in a different Region, a different KMS key is used
- You can't restore an unencrypted backup or snapshot to an encrypted DB instance

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Security

© Digital Cloud Training | https://digitalcloud.training

Create Encrypted Copy of RDS Database

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Security

© Digital Cloud Training | https://digitalcloud.training

Amazon Aurora

© Digital Cloud Training | https://digitalcloud.training

Amazon Aurora

- Amazon Aurora is an AWS database offering in the RDS family
- Amazon Aurora is a MySQL and PostgreSQL-compatible relational database built for the cloud
- Amazon Aurora is up to five times faster than standard MySQL databases and three times faster than standard PostgreSQL databases
- Amazon Aurora features a distributed, fault-tolerant, self-healing storage system that auto-scales up to 128TB per database instance

© Digital Cloud Training | https://digitalcloud.training

Amazon Aurora

Aurora Fault Tolerance

- Fault tolerance across 3 AZs
- Single logical volume
- Aurora Replicas scale-out read requests
- Can **promote** Aurora Replica to be a new primary or create new primary
- Can use **Auto Scaling** to add replicas

© Digital Cloud Training | https://digitalcloud.training

Amazon Aurora Key Features

Aurora Feature	Benefit
High performance and scalability	Offers high performance, self-healing storage that scales up to 128TB, point-in-time recovery and continuous backup to S3
DB compatibility	Compatible with existing MySQL and PostgreSQL open source databases
Aurora Replicas	In-region read scaling and failover target – up to 15 (can use Auto Scaling)
MySQL Read Replicas	Cross-region cluster with read scaling and failover target – up to 5 (each can have up to 15 Aurora Replicas)
Global Database	Cross-region cluster with read scaling (fast replication / low latency reads). Can remove secondary and promote
Multi-Master	Scales out writes within a region. In preview currently and will not appear on the exam
Serverless	On-demand, autoscaling configuration for Amazon Aurora - does not support read replicas or public IPs (can only access through VPC or Direct Connect - not VPN)

© Digital Cloud Training | https://digitalcloud.training

Amazon Aurora Replicas

Feature	Aurora Replica	MySQL Replica
Number of replicas	Up to 15	Up to 5
Replication type	Asynchronous (milliseconds)	Asynchronous (seconds)
Performance impact on primary	Low	High
Replica location	In-region	Cross-region
Act as failover target	Yes (no data loss)	Yes (potentially minutes of data loss)
Automated failover	Yes	No
Support for user-defined replication delay	No	Yes
Support for different data or schema vs. primary	No	Yes

© Digital Cloud Training | https://digitalcloud.training

Amazon Aurora Deployment Options

© Digital Cloud Training | https://digitalcloud.training

Aurora Fault Tolerance and Aurora Replicas

Aurora Fault Tolerance

- Fault tolerance across 3 AZs
- Single logical volume
- Aurora Replicas scale-out read requests
- Up to 15 Aurora Replicas with **sub-10ms** replica lag
- Aurora Replicas are independent endpoints
- Can **promote** Aurora Replica to be a new primary or create new primary
- Set priority (tiers) on Aurora Replicas to control order of promotion
- Can use **Auto Scaling** to add replicas

© Digital Cloud Training | https://digitalcloud.training

Cross-Region Replica with Aurora MySQL

© Digital Cloud Training | https://digitalcloud.training

Aurora Global Database

© Digital Cloud Training | https://digitalcloud.training

Aurora Multi-Master

Aurora Multi-Master

- All nodes allow reads/writes
- Available for MySQL only
- Up to four read/write nodes
- Single Region only
- Cannot have cross-Region replicas
- Can work with active-active and active-passive workloads
- Can restart read/write DB instance without impacting other instances

© Digital Cloud Training | https://digitalcloud.training

Aurora Serverless

© Digital Cloud Training | https://digitalcloud.training

Aurora Serverless Use Cases

- Infrequently used applications
- New applications
- Variable workloads
- Unpredictable workloads
- Development and test databases
- Multi-tenant applications

© Digital Cloud Training | https://digitalcloud.training

Amazon RDS Anti-Patterns and Alternatives

© Digital Cloud Training | https://digitalcloud.training

When NOT to use Amazon RDS (anti-patterns)

- Anytime you need a **DB type** other than:
  - MySQL
  - MariaDB
  - SQL Server
  - Oracle
  - PostgreSQL
- You need **root access** to the OS (e.g. install software such as management tools)

© Digital Cloud Training | https://digitalcloud.training

When NOT to use Amazon RDS (anti-patterns)

Requirement	More Suitable Service
Lots of large binary objects (BLOBs)	S3
Automated Scalability	DynamoDB
Name/Value Data Structure	DynamoDB
Data is not well structured or unpredictable	DynamoDB
Other database platforms like IBM DB2 or SAP HANA	EC2
Complete control over the database	EC2

© Digital Cloud Training | https://digitalcloud.training

Database on Amazon EC2

- You can run any database you like with **full control** and **ultimate flexibility**
- You must **manage everything** like backups, redundancy, patching and scaling

Amazon ElastiCache

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache

- Fully managed implementations **Redis** and **Memcached**
- ElastiCache is a **key/value** store
- In-memory database offering high performance and low latency
- Can be put in front of databases such as RDS and DynamoDB
- ElastiCache nodes run on Amazon EC2 instances, so you must choose an instance family/type

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache

Feature	Memcached	Redis (cluster mode disabled)	Redis (cluster mode enabled)
Data persistence	No	Yes	Yes
Data types	Simple	Complex	Complex
Data partitioning	Yes	No	Yes
Encryption	No	Yes	Yes
High availability (replication)	No	Yes	Yes
Multi-AZ	Yes, place nodes in multiple AZs. No failover or replication	Yes, with auto-failover. Uses read replicas (0-5 per shard)	Yes, with auto-failover. Uses read replicas (0-5 per shard)
Scaling	Up (node type); out (add nodes)	Up (node type); out (add replica)	Up (node type); out (add shards)
Multithreaded	Yes	No	No
Backup and restore	No (and no snapshots)	Yes, automatic and manual snapshots	Yes, automatic and manual snapshots

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache Use Cases

- Data that is relatively **static** and **frequently accessed**
- Applications that are tolerant of stale data
- Data is slow and expensive to get compared to cache retrieval
- Require push-button scalability for memory, writes and reads
- Often used for storing session state

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache Examples

Use Case	Benefit
Web session store	In cases with load-balanced web servers, store web session information in Redis so if a server is lost, the session info is not lost, and another web server can pick it up
Database caching	Use Memcached in front of AWS RDS to cache popular queries to offload work from RDS and return results faster to users
Leaderboards	Use Redis to provide a live leaderboard for millions of users of your mobile app
Streaming data dashboards	Provide a landing spot for streaming sensor data on the factory floor, providing live real-time dashboard displays

© Digital Cloud Training | https://digitalcloud.training

Scaling ElastiCache

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache - Scalability

Memcached

- Add nodes to a cluster
- Scale vertically (node type) – must create a **new cluster** manually

Redis

- Cluster mode **disabled**:
  - Add replica or change node type – creates a new cluster and migrates data
- Cluster mode **enabled**:
  - Online resharding to add or remove shards; vertical scaling to change node type
  - Offline resharding to add or remove shards change node type or upgrade engine (more flexible than online)

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache Memcached

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache Redis (Cluster mode enabled)

© Digital Cloud Training | https://digitalcloud.training

Amazon ElastiCache Redis (Cluster mode disabled)

© Digital Cloud Training | https://digitalcloud.training

Create ElastiCache Cluster

© Digital Cloud Training | https://digitalcloud.training

Amazon DynamoDB

© Digital Cloud Training | https://digitalcloud.training

Amazon DynamoDB

- Fully managed NoSQL database service
- Key/value store and document store
- It is a non-relational, key-value type of database
- Fully serverless service
- Push button scaling

DynamoDB Table

© Digital Cloud Training | https://digitalcloud.training

Amazon DynamoDB

- DynamoDB is made up of:
  - Tables
  - Items
  - Attributes

userid	orderid	book	price	date
user001	1000092	ISBN100..	9.99	2020.04..
user002	1000102	ISBN100..	24.99	2020.03..
user003	1000168	ISBN2X0..	12.50	2020.04..

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Time to Live (TTL)

- TTL lets you define when items in a table expire so that they can be automatically deleted from the database
- With TTL enabled on a table, you can set a timestamp for deletion on a per-item basis
- No extra cost and does not use WCU / RCU
- Helps reduce storage and manage the table size over time

© Digital Cloud Training | https://digitalcloud.training

Amazon DynamoDB

This is known as a **composite key** as it has a **partition key + sort key**

Primary Key		Attributes				
Partition Key	Sort Key					
clientid	created	sku	category	size	colour	weight
john@example.com	1583975308	SKU-S523	T-Shirt	Small	Red	Light
chris@example.com	1583975613	SKU-J091	Pen		Blue	
chris@example.com	1583975449	SKU-A234	Mug			
sarah@example.com	1583976311	SKU-R873	Chair			4011
jenny@example.com	1583976323	SKU-I019	Plate	30		

Useful when the data structure is **unpredictable**

© Digital Cloud Training | https://digitalcloud.training

Amazon DynamoDB

DynamoDB Feature	Benefit
Serverless	Fully managed, fault tolerant, service
Highly available	99.99% availability SLA – 99.999% for Global Tables!
NoSQL type of database with Name / Value structure	Flexible schema, good for when data is not well structured or unpredictable
Horizontal scaling	Seamless scalability to any scale with push button scaling or Auto Scaling
DynamoDB Streams	Captures a time-ordered sequence of item-level modifications in a DynamoDB table and durably stores the information for up to 24 hours. Often used with Lambda and the Kinesis Client Library (KCL)
DynamoDB Accelerator (DAX)	Fully managed in-memory cache for DynamoDB that increases performance (microsecond latency)
Transaction options	Strongly consistent or eventually consistent reads, support for ACID transactions
Backup	Point-in-time recovery down to the second in last 35 days; On-demand backup and restore
Global Tables	Fully managed multi-region, multi-master solution

DigitalCloud TRAINING

Create DynamoDB Table

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Streams

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Streams

1. Application inserts / updates /deletes item
2. A record is written to the DynamoDB stream
3. A Lambda function is triggered
4. The Lambda function writes to CloudWatch Logs

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Streams

- Captures a **time-ordered** sequence of **item-level** modifications in any DynamoDB table and stores this information in a log for up to **24 hours**
- Can configure the information that is written to the stream:
  - **KEYS_ONLY** — Only the key attributes of the modified item
  - **NEW_IMAGE** — The entire item, as it appears after it was modified
  - **OLD_IMAGE** — The entire item, as it appeared before it was modified
  - **NEW_AND_OLD_IMAGES** — Both the new and the old images of the item

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Accelerator (DAX)

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Accelerator (DAX)

- DAX is a fully managed, highly available, **in-memory** cache for DynamoDB
- Improves performance from milliseconds to microseconds
- Can be a read-through cache and a write-through cache
- Used to improve **READ** and **WRITE** performance
- You do not need to modify application logic, since DAX is compatible with existing DynamoDB API calls

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Accelerator (DAX)

© Digital Cloud Training | https://digitalcloud.training

DAX vs ElastiCache

- DAX is **optimized** for DynamoDB
- With ElastiCache you have more **management overhead** (e.g. invalidation)
- With ElastiCache you need to **modify application code** to point to cache
- ElastiCache supports more datastores

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Global Tables

© Digital Cloud Training | https://digitalcloud.training

DynamoDB Global Tables

© Digital Cloud Training | https://digitalcloud.training

Create DynamoDB Global Table

© Digital Cloud Training | https://digitalcloud.training

Amazon RedShift

© Digital Cloud Training | https://digitalcloud.training

Amazon RedShift

- Amazon Redshift is a fast, fully managed data warehouse
- Analyze data using standard SQL and existing Business Intelligence (BI) tools
- RedShift is a SQL based data warehouse used for analytics applications
- RedShift is a relational database that is used for Online Analytics Processing (OLAP) use cases
- RedShift uses Amazon EC2 instances, so you must choose an instance family/type
- RedShift always keeps three copies of your data
- RedShift provides continuous/incremental backups

© Digital Cloud Training | https://digitalcloud.training

OLTP vs OLAP (refresher)

Key differences are ***use cases*** and how the database is ***optimized***

Operational / transactional	Analytical
Online Transaction Processing (OLTP)	Online Analytics Processing (OLAP) – the source data comes from OLTP DBs
Production DBs that process transactions. E.g. adding customer records, checking stock availability (INSERT, UPDATE, DELETE)	Data warehouse. Typically, separated from the customer facing DBs. Data is extracted for decision making
Short transactions and simple queries	Long transactions and complex queries
Examples: Amazon RDS, DynamoDB	Examples: Amazon RedShift, Amazon EMR

© Digital Cloud Training | https://digitalcloud.training

Reporting and Analytics Use Cases

© Digital Cloud Training | https://digitalcloud.training

Redshift Data Sources

RedShift Spectrum can run SQL queries on data directly in **S3**

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

RedShift Use Cases

- Perform **complex queries** on massive collections of **structured** and **semi-structured** data and get fast performance
- Frequently accessed data that needs a consistent, highly structured format
- Use **Spectrum** for direct access of **S3 objects** in a data lake
- Managed data warehouse solution with:
  - Automated provisioning, configuration and patching
  - Data durability with continuous backup to S3
  - Scales with simple API calls
  - Exabyte scale query capability

© Digital Cloud Training | https://digitalcloud.training

Amazon Elastic Map Reduce (EMR)

© Digital Cloud Training | https://digitalcloud.training

Amazon EMR

- Managed cluster platform that simplifies running big data frameworks including **Apache Hadoop** and **Apache Spark**
- Used for processing data for analytics and business intelligence
- Can also be used for transforming and moving large amounts of data
- Performs extract, transform, and load (ETL) functions

© Digital Cloud Training | https://digitalcloud.training

Amazon EMR

© Digital Cloud Training | https://digitalcloud.training

Amazon Kinesis Core Knowledge

© Digital Cloud Training | https://digitalcloud.training

Amazon Kinesis Services

© Digital Cloud Training | https://digitalcloud.training

Amazon Kinesis Data Streams

- Producers send data to Kinesis, data is stored in Shards for 24 hours (by default, up to 7 days)
- Consumers then take the data and process it - data can then be saved into another AWS service
- Real time (~200ms)

© Digital Cloud Training | https://digitalcloud.training

Kinesis Client Library (KCL)

- The Kinesis Client Library (KCL) helps you consume and process data from a Kinesis data stream

© Digital Cloud Training | https://digitalcloud.training

Kinesis Client Library (KCL)

- Each shard is processed by exactly one KCL worker and has exactly one corresponding record processor
- One worker can process any number of shards, so it's fine if the number of shards exceeds the number of instances

Amazon Kinesis Data Streams

© Digital Cloud Training | https://digitalcloud.training

Kinesis Data Firehose

- Producers send data to Firehose
- There are no Shards, completely automated (scalability is elastic)
- Firehose data is sent to another AWS service for storing, data can be optionally processed/transformed using AWS Lambda
- Near real-time delivery (~60 seconds latency)

© Digital Cloud Training | https://digitalcloud.training

Kinesis Data Firehose

Kinesis Data Firehose destinations:

- RedShift (via an intermediate S3 bucket)
- Elasticsearch
- Amazon S3
- Splunk
- Datadog
- MongoDB
- New Relic
- HTTP Endpoint

© Digital Cloud Training | https://digitalcloud.training

Kinesis Data Analytics

- Provides real-time SQL processing for streaming data
- Provides analytics for data coming in from Kinesis Data Streams and Kinesis Data Firehose
- Destinations can be Kinesis Data Streams, Kinesis Data Firehose, or AWS Lambda

Amazon Athena and AWS Glue

© Digital Cloud Training | https://digitalcloud.training

Amazon Athena and AWS Glue

© Digital Cloud Training | https://digitalcloud.training

Amazon Athena and AWS Glue

Amazon Athena

SQL queries can also be run against these data sources

Data from source is **mapped** in **tables** in Athena and is queryable

A **Lambda function** connects Athena to **data source**

Lambda function

© Digital Cloud Training | https://digitalcloud.training

Amazon Athena

- Athena queries data in S3 using SQL
- Can be connected to other data sources with Lambda
- Data can be in CSV, TSV, JSON, Parquet and ORC formats
- Uses a managed Data Catalog (AWS Glue) to store information and schemas about the databases and tables

© Digital Cloud Training | https://digitalcloud.training

Optimizing Athena for Performance

- **Partition your data**
- **Bucket your data** – bucket the data within a single partition
- **Use Compression** – AWS recommend using either Apache Parquet or Apache ORC
- **Optimize file sizes**
- **Optimize columnar data store generation** – Apache Parquet and Apache ORC are popular columnar data stores
- **Optimize ORDER BY and Optimize GROUP BY**
- **Use approximate functions**
- **Only include the columns that you need**

© Digital Cloud Training | https://digitalcloud.training

AWS Glue

- Fully managed extract, transform and load (ETL) service
- Used for preparing data for analytics
- AWS Glue runs the ETL jobs on a fully managed, scale-out Apache Spark environment
- AWS Glue discovers data and stores the associated metadata (e.g. table definition and schema) in the AWS Glue Data Catalog
- Works with data lakes (e.g. data on S3), data warehouses (including RedShift), and data stores (including RDS or EC2 databases)

© Digital Cloud Training | https://digitalcloud.training

AWS Glue

- You can use a **crawler** to populate the AWS Glue Data Catalog with tables
- A crawler can crawl multiple data stores in a single run
- Upon completion, the crawler creates or updates one or more tables in your Data Catalog.
- ETL jobs that you define in AWS Glue use the Data Catalog tables as sources and targets

© Digital Cloud Training | https://digitalcloud.training

Query S3 ALB Access Logs with Athena

© Digital Cloud Training | https://digitalcloud.training

Amazon OpenSearch Service (Elasticsearch)

© Digital Cloud Training | https://digitalcloud.training

Amazon OpenSearch Service

Successor to **Amazon Elasticsearch Service**

Search, visualize, and analyze **text and unstructured data**

Amazon OpenSearch Service

Deploy **nodes** and **replicas** across AZs

Fully Managed

Petabyte Scale

Secure

Highly Available

Scalable

Deploy to **Amazon VPC** and integrates with **IAM**

© Digital Cloud Training | https://digitalcloud.training

Amazon OpenSearch Service

- Distributed search and analytics suite
- Based on the popular open source Elasticsearch
- Supports queries using SQL syntax
- Integrates with open-source tools
- Scale by adding or removing instances
- Availability in up to three Availability Zones
- Backup using snapshots
- Encryption at-rest and in-transit

© Digital Cloud Training | https://digitalcloud.training

OpenSearch Service Deployment

- Clusters are created (Management Console, API, or CLI)
- Clusters are also known as OpenSearch Service domains
- You specify the number of instances and instance types
- Storage options include UltraWarm or Cold storage

© Digital Cloud Training | https://digitalcloud.training

Ingesting Data into OpenSearch Service Domains

OpenSearch in an Amazon VPC

- Clusters can be deployed in a VPC for secure intra-VPC communications
- VPN or proxy required to connect from the internet (public domains are directly accessible)
- Cannot use IP-based access policies

OpenSearch in an Amazon VPC

- Limitations of VPC deployments:
  - You can’t switch from VPC to a public endpoint. The reverse is also true
  - You can’t launch your domain within a VPC that uses dedicated tenancy
  - After you place a domain within a VPC, you can’t move it to a different VPC, but you can change the subnets and security group settings

© Digital Cloud Training | https://digitalcloud.training

The ELK Stack

- ELK stands for Elasticsearch, Logstash, and Kibana

- This is a popular combination of projects
- Aggregate logs from systems and applications, analyze these logs, and create visualizations
- Use cases include:
  - Visualizing application and infrastructure monitoring data
  - Troubleshooting
  - Security analytics

© Digital Cloud Training | https://digitalcloud.training

OpenSearch Access Control

- **Resource-based policies** – often called a domain access policy
- **Identity-based policies** – attached to users or roles (principals)
- **IP-based policies** – Restrict access to one or more IP addresses or CIDR blocks
- **Fine-grained access control** – Provides:
  - Role-based access control
  - Security at the index, document, and field level
  - OpenSearch Dashboards multi-tenancy
  - HTTP basic authentication for OpenSearch and OpenSearch Dashboards

© Digital Cloud Training | https://digitalcloud.training

OpenSearch Access Control

- Authentication options include:
  - Federation using SAML to on-premises directories
  - Amazon Cognito and social identity providers

© Digital Cloud Training | https://digitalcloud.training

OpenSearch Best Practices

- Deploy OpenSearch data instances across three Availability Zones (AZs) for the best availability
- Provision instances in multiples of three for equal distribution across AZs
- If three AZs are not available use two AZs with equal numbers of instances

© Digital Cloud Training | https://digitalcloud.training

OpenSearch Best Practices

- Use three dedicated master nodes
- Configure at least one replica for each index
- Apply restrictive resource-based access policies to the domain (or use fine-grained access control)
- Create the domain within an Amazon VPC
- For sensitive data enable node-to-node encryption and encryption at rest

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases and Analytics

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement	Solution
Relational database running on MySQL must be migrated to AWS and must be highly available	Use Amazon RDS MySQL and configure a Multi-AZ standby node for HA
Amazon RDS DB has high query traffic that is causing performance degradation	Create a Read Replica and configure the application to use the reader endpoint for database queries
Amazon RDS DB is approaching its storage capacity limits and/or is suffering from high write latency	Scale up the DB instance to an instance type that has more storage / CPU

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement	Solution
Amazon RDS database is unencrypted and a cross-Region read replica must be created with encryption	Encrypt a snapshot of the main DB and create a new encrypted DB instance from the encrypted snapshot. Create a encrypted cross-Region read replica
Amazon Aurora DB deployed and requires a cross-Region replica	Deploy an Aurora MySQL Replica in the second Region
Amazon Aurora DB deployed and requires a read replica in the same Region with minimal synchronization latency	Deploy an Aurora Replica in the Region in a different Availability Zone

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement	Solution
Aurora deployed and app in another Region requires read-only access with low latency – synchronization latency must also be minimized	Use Aurora Global Database and configure the app in the second Region to use the reader endpoint
Application and DB migrated to Aurora and requires the ability to write to the DB across multiple nodes	Use Aurora Multi-Master for an in-Region multi-master database
Application requires a session-state data store that provides low-latency	Use either Amazon ElastiCache or DynamoDB

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement	Solution
Multi-threaded in-memory datastore required for unstructured data	Use Amazon ElastiCache Memcached
In-memory datastore required that offers microsecond performance for unstructured data	Use Amazon DynamoDB DAX (DAX)
In-memory datastore required that supports data persistence and high availability	Use Amazon ElastiCache Redis

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement	Solution
Serverless database required that supports No-SQL key-value store workload	Use Amazon DynamoDB
Serverless database required that supports MySQL or PostgreSQL	Use Amazon Aurora Serverless
Relational database required for a workload with an unknown usage pattern (usage expected to be low and variable)	Use Amazon Aurora Serverless

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement	Solution
Application requires a key-value database that can be written to from multiple AWS Regions	Use DynamoDB Global Tables
Athena is being used to analyze a large volume of data based on date ranges. Performance must be optimized	Store data using Apache Hive partitioning with a key based on the data. Use the Apache Parquet and ORC storage formats
Lambda is processing streaming data from API Gateway and is generating a TooManyRequestsException as volume increases	Stream the data into a Kinesis Data Stream from API Gateway and process in batches

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement	Solution
Lambda function is processsing streaming data that must be analyzed with SQL	Load data into a Kinesis Data Stream and then analyze with Kinesis Data Analytics
Security logs generated by AWS WAF must be sent to a third-party auditing application	Send logs to Kinesis Data Firehose and configure the auditing application using an HTTP endpoint
Real-time streaming data must be stored for future analysis	Ingest data into a Kinesis Data Stream and then use Firehose to load to a data store for later analysis

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Databases & Analytics

Requirement

Company runs several production databases and must run complex queries across consolidated data set for business forecasting

Solution

Load the data from the OLTP databases into a RedShift data warehouse for OLAP

© Digital Cloud Training | https://digitalcloud.training

SECTION 13

Deployment and Management

© Digital Cloud Training | https://digitalcloud.training

AWS CloudFormation

© Digital Cloud Training | https://digitalcloud.training

AWS CloudFormation

Infrastructure patterns are defined in a **template** file using **code**

CloudFormation **builds** your infrastructure according to the **template**

AWS CloudFormation

```
"AWSTemplateFormatVersion" : "2010-09-09",

  "Description" : "AWS CloudFormation Sample Template WordPress_Multi_AZ: WordPress is web

  "Parameters" : {
    "VpcId" : {
      "Type" : "AWS::EC2::VPC::Id",
      "Description" : "VpcId of your existing Virtual Private Cloud (VPC)",
      "ConstraintDescription" : "must be the VPC Id of an existing Virtual Private Cloud."
    },

    "Subnets" : {
      "Type" : "List<AWS::EC2::Subnet::Id>",
      "Description" : "The list of SubnetIds in your Virtual Private Cloud (VPC)",
      "ConstraintDescription" : "must be a list of at least two existing subnets associated
    },
```

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

AWS CloudFormation - Benefits

- Infrastructure is provisioned consistently, with fewer mistakes (human error)
- Less time and effort than configuring resources manually
- You can use version control and peer review for your CloudFormation templates
- Free to use (you're only charged for the resources provisioned)
- Can be used to manage updates and dependencies
- Can be used to rollback and delete the entire stack as well

© Digital Cloud Training | https://digitalcloud.training

AWS CloudFormation

Component	Description
Templates	The JSON or YAML text file that contains the instructions for building out the AWS environment
Stacks	The entire environment described by the template and created, updated, and deleted as a single unit
StackSets	AWS CloudFormation StackSets extends the functionality of stacks by enabling you to create, update, or delete stacks across multiple accounts and regions with a single operation
Change Sets	A summary of proposed changes to your stack that will allow you to see how those changes might impact your existing resources before implementing them

© Digital Cloud Training | https://digitalcloud.training

Deploy Stack and Change Set

© Digital Cloud Training | https://digitalcloud.training

Complex CloudFormation Template

© Digital Cloud Training | https://digitalcloud.training

AWS Elastic Beanstalk

© Digital Cloud Training | https://digitalcloud.training

AWS Elastic Beanstalk

© Digital Cloud Training | https://digitalcloud.training

AWS Elastic Beanstalk

- Supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker web applications
- Supports the following languages and development stacks:
  - Apache Tomcat for Java applications
  - Apache HTTP Server for PHP applications
  - Apache HTTP Server for Python applications
  - Nginx or Apache HTTP Server for Node.js applications
  - Passenger or Puma for Ruby applications
  - Microsoft IIS 7.5, 8.0, and 8.5 for .NET applications
  - Java SE
  - Docker
  - Go

© Digital Cloud Training | https://digitalcloud.training

AWS Elastic Beanstalk

There are several **layers**

Applications:

- Contain environments, environment configurations, and application versions
- You can have multiple application versions held within an application

APPLICATION

© Digital Cloud Training | https://digitalcloud.training

AWS Elastic Beanstalk

Application version

- A specific reference to a section of deployable code
- The application version will point typically to an Amazon S3 bucket containing the code

© Digital Cloud Training | https://digitalcloud.training

AWS Elastic Beanstalk

Environments:

- An application version that has been deployed on AWS resources
- The resources are configured and provisioned by AWS Elastic Beanstalk
- The environment is comprised of all the resources created by Elastic Beanstalk and not just an EC2 instance with your uploaded code

Web Servers and Workers

- **Web servers** are standard applications that listen for and then process HTTP requests, typically over port 80
- **Workers** are specialized applications that have a background processing task that listens for messages on an Amazon SQS queue
- **Workers** should be used for long-running tasks

© Digital Cloud Training | https://digitalcloud.training

AWS Elastic Beanstalk

© Digital Cloud Training | https://digitalcloud.training

Deploy Elastic Beanstalk Web Server and Worker

© Digital Cloud Training | https://digitalcloud.training

AWS Systems Manager Parameter Store

© Digital Cloud Training | https://digitalcloud.training

AWS SSM Parameter Store

- Parameter Store provides secure, hierarchical storage for configuration data and secrets
- Highly scalable, available, and durable
- Store data such as passwords, database strings, and license codes as parameter values
- Store values as plaintext (unencrypted data) or ciphertext (encrypted data)
- Reference values by using the unique name that you specified when you created the parameter
- No native rotation of keys (difference with AWS Secrets Manager which does it automatically)

© Digital Cloud Training | https://digitalcloud.training

AWS Config

© Digital Cloud Training | https://digitalcloud.training

AWS Config

- Evaluate your AWS resource configurations for desired settings
- Get a snapshot of the current configurations of resources that are associated with your AWS account
- Retrieve configurations of resources that exist in your account
- Retrieve historical configurations of one or more resources
- Receive a notification whenever a resource is created, modified, or deleted
- View relationships between resources

© Digital Cloud Training | https://digitalcloud.training

AWS Config

AWS Config evaluates the **configuration** against desired configurations

Example Services:

Amazon EC2

Elastic Load Balancing

Amazon RDS

Configuration changes occur and information is sent to **AWS Config**

AWS Config

Amazon S3

Amazon Simple Notification Service

Send notifications with **SNS**

CloudWatch Events

Alert via **CW Events** when changes occur

Systems Manager Automation

Automatic **remediation**

© Digital Cloud Training | https://digitalcloud.training

DigitalCloud TRAINING

AWS Config

Example Rule	Description
s3-bucket-server-side-encryption-enabled	Checks that your Amazon S3 bucket either has S3 default encryption enabled or that the S3 bucket policy explicitly denies put-object requests without server side encryption
restricted-ssh	Checks whether security groups that are in use disallow unrestricted incoming SSH traffic
rds-instance-public-access-check	Checks whether the Amazon Relational Database Service (RDS) instances are not publicly accessible
cloudtrail-enabled	Checks whether AWS CloudTrail is enabled in your AWS account

© Digital Cloud Training | https://digitalcloud.training

AWS Config Rule with Remediation

© Digital Cloud Training | https://digitalcloud.training

AWS Secrets Manager

© Digital Cloud Training | https://digitalcloud.training

AWS Secrets Manager

- Stores and rotate secrets safely without the need for code deployments
- Secrets Manager offers automatic rotation of credentials (built-in) for:
  - Amazon RDS (MySQL, PostgreSQL, and Amazon Aurora)
  - Amazon Redshift
  - Amazon DocumentDB
- For other services you can write your own AWS Lambda function for automatic rotation

© Digital Cloud Training | https://digitalcloud.training

AWS Secrets Manager vs SSM Parameter Store

	Secrets Manager	SSM Parameter Store
Automatic Key Rotation	Yes, built-in for some services, use Lambda for others	No native key rotation; can use custom Lambda
Key/Value Type	String or Binary (encrypted)	String, StringList, SecureString (encrypted)
Hierarchical Keys	No	Yes
Price	Charges apply per secret	Free for standard, charges for advanced

© Digital Cloud Training | https://digitalcloud.training

AWS OpsWorks

© Digital Cloud Training | https://digitalcloud.training

AWS OpsWorks

- AWS OpsWorks is a configuration management service that provides managed instances of Chef and Puppet
- Updates include patching, updating, backup, configuration and compliance management

AWS Resource Access Manager (RAM)

© Digital Cloud Training | https://digitalcloud.training

AWS RAM

- Shares resources:
  - Across AWS accounts
  - Within AWS Organizations or OUs
  - IAM roles and IAM users
- Resource shares are created with:
  - The AWS RAM Console
  - AWS RAM APIs
  - AWS CLI
  - AWS SDKs

© Digital Cloud Training | https://digitalcloud.training

AWS RAM

RAM can be used to share:

- AWS App Mesh
- Amazon Aurora
- AWS Certificate Manager Private Certificate Authority
- AWS CodeBuild
- Amazon EC2
- EC2 Image Builder
- AWS Glue
- AWS License Manager
- AWS Network Firewall
- AWS Outposts
- Amazon S3 on Outposts
- AWS Resource Groups
- Amazon Route 53
- AWS Systems Manager Incident Manager
- Amazon VPC

© Digital Cloud Training | https://digitalcloud.training

Share a Subnet Across Accounts

© Digital Cloud Training | https://digitalcloud.training

AWS RAM

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Deployment and Management

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Deployment and Management

Requirement	Solution
Application must authenticate to Amazon Aurora and need to securely store credentials. Automatic credential rotation is required on a monthly basis	Use AWS Secrets Manager to store the credentials. Update the app to retrieve credentials from Secrets Manager; enable automatic monthly rotation
Company currently uses Chef cookbooks to manage infrastructure and is moving to the cloud. Need to minimize migration complexity	Use AWS OpsWorks for Chef Automate
Need a managed environment for running a simple web application. App processes incoming data which can take several minutes per task	Use an Elastic Beanstalk environment with a web server for the app front-end and a decoupled worker tier for the long running process

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Deployment and Management

Requirement	Solution
Systems integrator deploys standardized Amazon VPC configuration for many customers. Need to increase efficiency and reduce errors	Create an AWS CloudFormation template based on the standard config and use CloudFormation to deploy to new customers
Company has experienced issues rolling out updates to a CloudFormation Stack and needs to preview changes before the next update	Preview the change by creating a change set with the updated template before updating the stack
Manager wishes to monitor and enforce configuration compliance for AWS resources including S3 buckets and security groups	Use AWS Config to create rules to monitor compliance and use auto remediation to enforce compliance

© Digital Cloud Training | https://digitalcloud.training

SECTION 14

Monitoring, Logging, and Auditing

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch Overview

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch

CloudWatch is used for performance monitoring, alarms, log collection and automated actions

Use cases / benefits include:

- Collect performance metrics from AWS and on-premises systems
- Automate responses to operational changes
- Improve operational performance and resource optimization
- Derive actionable insights from logs
- Get operational visibility and insight

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch

CloudWatch Core Features:

- **CloudWatch Metrics** – services send time-ordered data points to CloudWatch
- **CloudWatch Alarms** – monitor metrics and initiate actions
- **CloudWatch Logs** – centralized collection of system and application logs
- **CloudWatch Events** – stream of system events describing changes to AWS resources and can trigger actions

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch Metrics

- Metrics are sent to CloudWatch for many AWS services
- EC2 metrics are sent every **5 minutes** by default (free)
- Detailed EC2 monitoring sends every **1 minute** (chargeable)
- Unified CloudWatch Agent sends system-level metrics for EC2 and on-premises servers
- System-level metrics include memory and disk usage

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch Metrics

- Can publish custom metrics using CLI or API
- Custom metrics are one of the following resolutions:
  - **Standard resolution** – data having a one-minute granularity
  - **High resolution** – data at a granularity of one second
- AWS metrics are standard resolution by default

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch Alarms

Two types of alarms

- **Metric alarm** – performs one or more actions based on a single metric
- **Composite alarm** – uses a rule expression and takes into account multiple alarms
- Metric alarm states:
  - OK – Metric is within a threshold
  - ALARM – Metric is outside a threshold
  - INSUFFICIENT_DATA – not enough data

Create a Custom Metric

© Digital Cloud Training | https://digitalcloud.training

Create a CloudWatch Alarm

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch Logs

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch Logs

- Gather application and system logs in CloudWatch
- Defined expiration policies and KMS encryption
- Send to:
  - Amazon S3 (export)
  - Kinesis Data Streams
  - Kinesis Data Firehose

Unified CloudWatch Agent installed on EC2 and on-premises servers

© Digital Cloud Training | https://digitalcloud.training

Review CloudWatch Logs

© Digital Cloud Training | https://digitalcloud.training

The Unified CloudWatch Agent

© Digital Cloud Training | https://digitalcloud.training

The Unified CloudWatch Agent

The unified CloudWatch agent enables you to do the following:

- Collect internal system-level metrics from Amazon **EC2 instances** across operating systems
- Collect system-level metrics from **on-premises servers**
- Retrieve custom metrics from your applications or services using the StatsD and collectd protocols
- Collect logs from Amazon EC2 instances and on-premises servers (Windows / Linux)

© Digital Cloud Training | https://digitalcloud.training

The Unified CloudWatch Agent

- Agent must be installed on the server
- Can be installed on:
  - Amazon EC2 instances
  - On-premises servers
  - Linux, Windows Server, or macOS

© Digital Cloud Training | https://digitalcloud.training

AWS CloudTrail

© Digital Cloud Training | https://digitalcloud.training

AWS CloudTrail

- CloudTrail logs **API activity** for auditing
- By default, management events are logged and retained for 90 days
- A **CloudTrail Trail** logs any events to S3 for indefinite retention
- Trail can be within Region or all Regions
- CloudWatch Events can triggered based on API calls in CloudTrail
- Events can be streamed to CloudWatch Logs

© Digital Cloud Training | https://digitalcloud.training

AWS CloudTrail

© Digital Cloud Training | https://digitalcloud.training

CloudTrail – Types of Events

- **Management events** provide information about management operations that are performed on resources in your AWS account
- **Data events** provide information about the resource operations performed on or in a resource
- **Insights events** identify and respond to unusual activity associated with write API calls by continuously analyzing CloudTrail management events

© Digital Cloud Training | https://digitalcloud.training

Create a CloudTrail Trail

© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge (Refresher)

© Digital Cloud Training | https://digitalcloud.training

Amazon CloudWatch Events / EventBridge

EventBridge used to be known as **CloudWatch Events**

© Digital Cloud Training | https://digitalcloud.training

Create EventBridge rule for CloudTrail API calls

© Digital Cloud Training | https://digitalcloud.training

Amazon EventBridge Example 2

```
{
   "source": ["aws.cloudtrail"],
   "detail-type": ["AWS API Call via CloudTrail"],
   "detail": {
     "eventSource": ["cloudtrail.amazonaws.com"],
     "eventName": ["s3:PutBucketPolicy"]
   }
 }
```

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Monitoring, Logging and Auditing

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Monitoring, Logging and Auditing

Requirement	Solution
Need to stream logs from Amazon EC2 instances in an Auto Scaling Group	Install the unified CloudWatch Agent and collect log files in Amazon CloudWatch
Need to collect metrics from EC2 instances with a 1 second granularity	Create a custom metric with high resolution
The application logs from on-premises servers must be processed by AWS Lambda in real time	Install the unified CloudWatch Agent on the servers and use a subscription filter in CloudWatch to connect to a Lambda function

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Monitoring, Logging and Auditing

Requirement	Solution
CloudWatch Logs entries must be transformed with Lambda and then loaded into Amazon S3	Configure a Kinesis Firehose destination, transform with Lambda and then load into an S3 bucket
Access auditing must be enabled, and records must be stored for a minimum of 5 years. Any attempts to modify the log files must be identified	Create a trail in CloudTrail that stores the data in an S3 bucket and enable log file integrity validation
API activity must be captured from all accounts in an Organization. Admins in member accounts must not be able to modify or delete the trail	Create an Organization trail in AWS CloudTrail that applies to the management account and all member accounts

DigitalCloud TRAINING

Architecture Patterns – Monitoring, Logging and Auditing

Requirement	Solution
Company requires API events that involve the root user account to generate a notification	Create a CloudTrail trail and an EventBridge rule that looks for API events that involve root and configure an SNS notification
For compliance reasons all S3 buckets must have encryption enabled and any non-compliant buckets must be auto remediated	Use AWS Config to check the encryption status of the buckets and use auto remediation to enable encryption as required

© Digital Cloud Training | https://digitalcloud.training

SECTION 15

Security in the Cloud

© Digital Cloud Training | https://digitalcloud.training

AWS Directory Service

© Digital Cloud Training | https://digitalcloud.training

AWS Managed Microsoft AD

Managed implementation of **Microsoft Active Directory** running on Windows Server

HA pair of Windows Server **Domain Controllers** (DCs)

One or two-way **trust relationship**

aws

Console

Workspaces

RDS

Workdocs

QuickSight

Workmail

VPC

MS

MS

ADSync ADFS

VPN

Corporate Office / DC

Microsoft AD

Securely connect to Amazon EC2 **Linux** and **Windows** instances

Also Allows you to:

- Apply group policy
- Use single sign-on to apps and services
- Enable MFA with RADIUS

Example apps and services that support **authentication** and **authorization** using AWS Directory Service

Azure

Azure AD

Office 365

Synchronize users and federate identities with **Azure/O365**

DigitalCloud TRAINING

© Digital Cloud Training | https://digitalcloud.training

AWS Managed Microsoft AD

- Fully managed AWS service
- Best choice if you have more than 5000 users and/or need a trust relationship set up
- Can perform schema extensions
- Can setup trust relationships to with on-premises Active Directories:
  - On-premise users and groups can access resources in either domain using SSO
  - Requires a VPN or Direct Connect connection
- Can be used as a standalone AD in the AWS cloud

© Digital Cloud Training | https://digitalcloud.training

AD Connector

© Digital Cloud Training | https://digitalcloud.training

AD Connector

- Redirects directory requests to your on-premise Active Directory
- Best choice when you want to use an existing Active Directory with AWS services
- AD Connector comes in two sizes:
  - Small – designed for organizations up to 500 users
  - Large – designed for organizations up to 5000 users
- Requires a VPN or Direct Connect connection
- Join EC2 instances to your on-premise AD through AD Connector
- Login to the AWS Management Console using your on-premise AD DCs for authentication

© Digital Cloud Training | https://digitalcloud.training

Simple AD

- Inexpensive Active Directory-compatible service with common directory features
- Standalone, fully managed, directory on the AWS cloud
- Simple AD is generally the least expensive option
- Best choice for less than 5000 users and don’t need advanced AD features
- Features include:
  - Manage user accounts / groups
  - Apply group policies
  - Kerberos-based SSO
  - Supports joining Linux or Windows based EC2 instances

© Digital Cloud Training | https://digitalcloud.training

Identity Federation

© Digital Cloud Training | https://digitalcloud.training

Identity Federation

1. Client application attempts to authenticate using IdP
2. IdP authenticates the user
3. IdP sends client SAML assertion
4. App calls sts:AssumeRoleWithSAML
5. AWS return temporary security credentials
6. App uses credentials to access S3 bucket

© Digital Cloud Training | https://digitalcloud.training

IAM Federation

© Digital Cloud Training | https://digitalcloud.training

AWS Single Sign-on (SSO)

© Digital Cloud Training | https://digitalcloud.training

AWS Single Sign-on (SSO)

Corporate Office / DC

Active Directory (self-managed)

Identity sources can be AWS SSO, Active Directory and standard providers using **SAML 2.0**

Enables centralized **permissions** management

AWS Directory Service

AWS Single Sign-On

AWS Organizations

AWS Account A

AWS Account B

Azure

Azure AD

AWS Account C

AWS Account D

Connect **AWS accounts** and **Organizations**

Built-in **SSO** integrations to business applications

DigitalCloud TRAINING

© Digital Cloud Training | https://digitalcloud.training

Amazon Cognito

© Digital Cloud Training | https://digitalcloud.training

Cognito User Pools

© Digital Cloud Training | https://digitalcloud.training

Cognito Identity Pool

© Digital Cloud Training | https://digitalcloud.training

User Pools + Identity Pools

© Digital Cloud Training | https://digitalcloud.training

Encryption Primer

© Digital Cloud Training | https://digitalcloud.training

Encryption In Transit vs At Rest

Encryption In Transit

Data is protected by **SSL/TLS** in transit

Encryption At Rest

Amazon S3 **encrypts** the object as it is **written** to the bucket it

© Digital Cloud Training | https://digitalcloud.training

Asymmetric Encryption

- Asymmetric encryption is also known as public key cryptography
- Messages encrypted with the public key can only be decrypted with the private key
- Messages encrypted with the private key can be decrypted with the public key
- Examples include SSL/TLS and SSH

© Digital Cloud Training | https://digitalcloud.training

Symmetric Encryption

Encryption

Decryption

© Digital Cloud Training | https://digitalcloud.training

AWS Key Management Service (KMS)

© Digital Cloud Training | https://digitalcloud.training

Customer Master Keys (CMKs)

- Customer master keys are the primary resources in AWS KMS
- The CMK also contains the key material used to encrypt and decrypt data
- CMKs are created in AWS KMS
- Symmetric CMKs and the private keys of asymmetric CMKs never leave AWS KMS unencrypted
- By default, AWS KMS creates the key material for a CMK
- Can also import your own key material
- A CMK can encrypt data up to 4KB in size
- A CMK can generate, encrypt and decrypt Data Encryption Keys
- Data Encryption Keys can be used for encrypting large amounts of data

© Digital Cloud Training | https://digitalcloud.training

AWS Managed CMKs

- Created, managed, and used on your behalf by an AWS service that is integrated with AWS KMS
- You cannot manage these CMKs, rotate them, or change their key policies
- You also cannot use AWS managed CMKs in cryptographic operations directly; the service that creates them uses them on your behalf

© Digital Cloud Training | https://digitalcloud.training

Data Encryption Keys

- Data keys are encryption keys that you can use to encrypt data, including large amounts of data and other data encryption keys
- You can use AWS KMS customer master keys (CMKs) to generate, encrypt, and decrypt data keys
- AWS KMS does not store, manage, or track your data keys, or perform cryptographic operations with data keys
- You must use and manage data keys outside of AWS KMS

© Digital Cloud Training | https://digitalcloud.training

Customer Master Keys (CMKs)

Type of CMK	Can view	Can manage	Used only for my AWS account	Automatic rotation
Customer managed CMK	Yes	Yes	Yes	Optional. Every 365 days
AWS managed CMK	Yes	No	Yes	Required. Every 1095 days
AWS owned CMK	No	No	No	Varies

© Digital Cloud Training | https://digitalcloud.training

Create Encryption Key

© Digital Cloud Training | https://digitalcloud.training

AWS CloudHSM

© Digital Cloud Training | https://digitalcloud.training

AWS CloudHSM

- AWS CloudHSM is a cloud-based hardware security module (HSM)
- Generate and use your own encryption keys on the AWS Cloud
- CloudHSM runs in your Amazon VPC
- Uses FIPS 140-2 level 3 validated HSMs
- Managed service and automatically scales
- Retain control of your encryption keys - you control access (and AWS has no visibility of your encryption keys)

© Digital Cloud Training | https://digitalcloud.training

AWS CloudHSM Use Cases

- Offload SSL/TLS processing from web servers
- Protect private keys for an issuing certificate authority (CA)
- Store the master key for Oracle DB Transparent Data Encryption
- Custom key store for AWS KMS – retain control of the HSM that protects the master keys

© Digital Cloud Training | https://digitalcloud.training

AWS CloudHSM vs KMS

	CloudHSM	AWS KMS
Tenancy	Single-tenant HSM	Multi-tenant AWS service
Availability	Customer-managed durability and available	Highly available and durable key storage and management
Root of Trust	Customer managed root of trust	AWS managed root of trust
FIPS 140-2	Level 3	Level 2 / Level 3
3rd Party Support	Broad 3rd Party Support	Broad AWS service support

© Digital Cloud Training | https://digitalcloud.training

AWS Certificate Manager (ACM)

© Digital Cloud Training | https://digitalcloud.training

AWS Certificate Manager (ACM)

- Create, store and renew SSL/TLS X.509 certificates
- Single domains, multiple domain names and wildcards
- Integrates with several AWS services including:
  - **Elastic Load Balancing**
  - **Amazon CloudFront**
  - **AWS Elastic Beanstalk**
  - **AWS Nitro Enclaves**
  - **AWS CloudFormation**

© Digital Cloud Training | https://digitalcloud.training

AWS Certificate Manager (ACM)

- **Public certificates** are signed by the AWS public Certificate Authority
- You can also create a Private CA with ACM
- Can then issue private certificates
- You can also import certificates from third-party issuers

© Digital Cloud Training | https://digitalcloud.training

SSL/TLS Certificate in ACM

© Digital Cloud Training | https://digitalcloud.training

AWS Web Application Firewall (WAF)

© Digital Cloud Training | https://digitalcloud.training

AWS WAF

- AWS WAF is a web application firewall
- WAF lets you create rules to filter web traffic based on conditions that include IP addresses, HTTP headers and body, or custom URIs
- WAF makes it easy to create rules that block common web exploits like SQL injection and cross site scripting

© Digital Cloud Training | https://digitalcloud.training

AWS WAF

© Digital Cloud Training | https://digitalcloud.training

AWS WAF

- **Web ACLs** – You use a web access control list (ACL) to protect a set of AWS resources
- **Rules** – Each rule contains a statement that defines the inspection criteria, and an action to take if a web request meets the criteria
- **Rule groups** – You can use rules individually or in reusable rule groups

© Digital Cloud Training | https://digitalcloud.training

AWS WAF

- **IP Sets** - An IP set provides a collection of IP addresses and IP address ranges that you want to use together in a rule statement
- **Regex pattern set** - A regex pattern set provides a collection of regular expressions that you want to use together in a rule statement

© Digital Cloud Training | https://digitalcloud.training

AWS WAF

A **rule action** tells AWS WAF what to do with a web request when it ***matches*** the criteria defined in the rule:

- **Count** – AWS WAF counts the request but doesn't determine whether to allow it or block it. With this action, AWS WAF continues processing the remaining rules in the web ACL
- **Allow** – AWS WAF allows the request to be forwarded to the AWS resource for processing and response
- **Block** – AWS WAF blocks the request and the AWS resource responds with an HTTP 403 (Forbidden) status code

© Digital Cloud Training | https://digitalcloud.training

AWS WAF

Match statements compare the web request or its origin against conditions that you provide

Match Statement	Description
Geographic match	Inspects the request's country of origin
IP set match	Inspects the request against a set of IP addresses and address ranges
Regex pattern set	Compares regex patterns against a specified request component
Size constraint	Checks size constraints against a specified request component
SQLi attack	Inspects for malicious SQL code in a specified request component
String match	Compares a string to a specified request component
XSS scripting attack	Inspects for cross-site scripting attacks in a specified request component

DigitalCloud TRAINING

AWS Shield

© Digital Cloud Training | https://digitalcloud.training

AWS Shield

- AWS Shield is a managed Distributed Denial of Service (DDoS) protection service
- Safeguards web application running on AWS with always-on detection and automatic inline mitigations
- Helps to minimize application downtime and latency
- Two tiers –
  - **Standard** – no cost
  - **Advanced** - $3k USD per month and 1 year commitment
- Integrated with Amazon CloudFront (standard included by default)

© Digital Cloud Training | https://digitalcloud.training

Amazon Macie

© Digital Cloud Training | https://digitalcloud.training

Amazon Macie

- Macie is a fully managed data security and data privacy service
- Uses machine learning and pattern matching to discover, monitor, and help you protect your sensitive data on Amazon S3
- Macie enables security compliance and preventive security as follows:
  - Identify a variety of data types, including PII, Protected Health Information (PHI), regulatory documents, API keys, and secret keys
  - Identify changes to policy and access control lists
  - Continuously monitor the security posture of Amazon S3
  - Generate security findings that you can view using the Macie console, AWS Security Hub, or Amazon EventBridge
  - Manage multiple AWS accounts using AWS Organizations

© Digital Cloud Training | https://digitalcloud.training

Amazon Macie

© Digital Cloud Training | https://digitalcloud.training

Amazon Inspector

© Digital Cloud Training | https://digitalcloud.training

Amazon Inspector

- Runs assessments that check for security exposures and vulnerabilities in EC2 instances
- Can be configured to run on a schedule
- Agent must be installed on EC2 for host assessments
- Network assessments do not require an agent

© Digital Cloud Training | https://digitalcloud.training

Amazon Inspector

Network Assessments

- Assessments: Network configuration analysis to check for ports reachable from outside the VPC
- If the Inspector Agent is installed on your EC2 instances, the assessment also finds processes reachable on port
- Price based on the number of instance assessments

© Digital Cloud Training | https://digitalcloud.training

Amazon Inspector

Host Assessments

- Assessments: Vulnerable software (CVE), host hardening (CIS benchmarks), and security best practices
- Requires an agent (auto-install with SSM Run Command)
- Price based on the number of instance assessments

© Digital Cloud Training | https://digitalcloud.training

AWS GuardDuty

© Digital Cloud Training | https://digitalcloud.training

AWS GuardDuty

- Intelligent threat detection service
- Detects account compromise, instance compromise, malicious reconnaissance, and bucket compromise
- Continuous monitoring for events across:
  - **AWS CloudTrail Management Events**
  - **AWS CloudTrail S3 Data Events**
  - **Amazon VPC Flow Logs**
  - **DNS Logs**

© Digital Cloud Training | https://digitalcloud.training

Defense In-Depth

© Digital Cloud Training | https://digitalcloud.training

Secure Multi-Tier Architecture

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns - Security

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Security

Requirement	Solution
Need to enable custom domain name and encryption in transit for an application running behind an Application Load Balancer	Use AWS Route 53 to create an Alias record to the ALB's DNS name and attach an SSL/TLS certificate issued by Amazon Certificate Manager (ACM)
Website running on EC2 instances behind and ALB must be protected against well known web exploits	Create a Web ACL in AWS WAF to protect against web exploits and attach to the ALB
Need to block access to an application running on an ALB from connections originating in a specific list of countries	Create a Web ACL in AWS WAF with a geographic match and block traffic that matches the list of countries

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Security

Requirement	Solution
Company needs to encrypt large volumes of data using a CMK in AWS KMS	Create a data encryption key using the CMK to encrypt large volumes of data
Mobile app requires authorized access to AWS services. Users authenticate using social IdPs and a preconfigured Web UI is required for logging in	Create an Amazon Cognito User Pool that leverages the social IdPs and an Identity Pool for gaining temporary credentials for AWS
Company has an on-premises Microsoft AD and AWS DX connection. Requires joining EC2 instances to on-premises domain	Configure an AD Connector that uses the on-premises AD

DigitalCloud TRAINING

SECTION 16

Migration and Transfer Services

© Digital Cloud Training | https://digitalcloud.training

AWS Migration Tools Overview

© Digital Cloud Training | https://digitalcloud.training

AWS Migration Tools

© Digital Cloud Training | https://digitalcloud.training

AWS Application Discovery Service

© Digital Cloud Training | https://digitalcloud.training

AWS Application Discovery Service

© Digital Cloud Training | https://digitalcloud.training

AWS Application Discovery Service

	Discovery Connector	Discovery Agent
Supported Servers	VMware	VMware, Physical
Deployment	Per vCenter	Per Server
Collected data	Static configuration data VM utilization metrics	Static configuration data Time Series Performance info (export) Network Inbound/outbound (export) Running processes (export)
Supported OS	Any OS running in VMware vCenter (v5.5, v6, & v6.5	Amazon Linux, Ubuntu, RHEL, CentOS, SUSE, Windows Server

© Digital Cloud Training | https://digitalcloud.training

AWS Database Migration Service (DMS)

© Digital Cloud Training | https://digitalcloud.training

AWS Database Migration Service (DMS)

© Digital Cloud Training | https://digitalcloud.training

AWS DMS Use Cases

- **Cloud to Cloud** – EC2 to RDS, RDS to RDS, RDS to Aurora
- **On-Premises to Cloud**
- **Homogeneous migrations** – Oracle to Oracle, MySQL to RDS MySQL, Microsoft SQL to RDS for SQL Server
- **Heterogeneous migrations** – Oracle to Aurora, Oracle to PostgreSQL, Microsoft SQL to RDS MySQL (must convert schema first wit the **Shema Conversion Tool** (SCT))

© Digital Cloud Training | https://digitalcloud.training

AWS DMS Use Cases

- **Development and Test** – use the cloud for dev/test workloads
- **Database consolidation** – consolidate multiple source DBs to a single target DB
- **Continuous Data Replication** – use for DR, dev/test, single source multi-target or multi-source single target

© Digital Cloud Training | https://digitalcloud.training

AWS Server Migration Service (SMS)

© Digital Cloud Training | https://digitalcloud.training

AWS Server Migration Service (SMS)

- AWS SMS migrates VMware vSphere, Microsoft Hyper-V/SCVMM, and Azure virtual machines to Amazon EC2

© Digital Cloud Training | https://digitalcloud.training

AWS SMS – Application Migration

© Digital Cloud Training | https://digitalcloud.training

AWS DataSync

© Digital Cloud Training | https://digitalcloud.training

AWS DataSync

© Digital Cloud Training | https://digitalcloud.training

AWS Snow Family

© Digital Cloud Training | https://digitalcloud.training

AWS Snowball Family

- **AWS Snowball and Snowmobile** are used for migrating large volumes of data to AWS
- **Snowball Edge Compute Optimized**
  - Provides block and object storage and optional GPU
  - Use for data collection, machine learning and processing, and storage in environments with intermittent connectivity (edge use cases)
- **Snowball Edge Storage Optimized**
  - Provides block storage and Amazon S3-compatible object storage
  - Use for local storage and large-scale data transfer
- **Snowcone**
  - Small device used for edge computing, storage and data transfer
  - Can transfer data offline or online with AWS DataSync agent

© Digital Cloud Training | https://digitalcloud.training

AWS Snowball Family

- Uses a secure storage device for physical transportation
- Snowball Client is software that is installed on a local computer and is used to identify, compress, encrypt, and transfer data
- Uses 256-bit encryption (managed with the AWS KMS) and tamper-resistant enclosures with TPM
- Snowball (80TB) (50TB) “petabyte scale”
- Snowball Edge (100TB) “petabyte scale”
- Snowmobile – “exabyte scale” with up to 100PB per Snowmobile

© Digital Cloud Training | https://digitalcloud.training

AWS Snowball Family

Ways to optimize the performance of Snowball transfers:

1. Use the latest Mac or Linux Snowball client
2. Batch small files together
3. Perform multiple copy operations at one time
4. Copy from multiple workstations
5. Transfer directories, not files

© Digital Cloud Training | https://digitalcloud.training

AWS Snowball Use Cases

- **Cloud data migration** – migrate data to the cloud
- **Content distribution** – send data to clients or customers
- **Tactical Edge Computing** – collect data and compute
- **Machine learning** – run ML directly on the device
- **Manufacturing** – data collection and analysis in the factory
- **Remote locations with simple data** – pre-processing, tagging, compression etc.

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Migration and Transfer

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Migration and Transfer

Requirement	Solution
Company is migrating Linux and Windows VMs in VMware to the cloud. Need to determine performance requirements for right-sizing	Install the Application Discovery Service discovery connector in VMware vCenter to gather data
Company has a mixture of VMware VMs and physical servers to migrate to AWS. Need to identify dependencies for grouping applications for migration	Install the Application Discovery Service discovery connector in VMware vCenter and the discovery agent on physical servers
Need to migrate an Oracle data warehouse to AWS	Migrate using AWS DMS and SCT to a RedShift data warehouse

© Digital Cloud Training | https://digitalcloud.training

Architecture Patterns – Migration and Transfer

Requirement	Solution
Snowball Edge used to transfer millions of small files using a shell script. Transfer times are very slow	Perform multiple copy operations at one time by running each command from a separate terminal, in separate instances of the Snowball client
Need to minimize downtime for servers that must be migrated to AWS	Use AWS SMS and perform a final synchronization before cutting over in a short outage window
Need to migrate 50TB of data and company only has a 1Gbps internet link. Requirement is urgent	Use AWS Snowball to transfer the data

© Digital Cloud Training | https://digitalcloud.training