Module 1. The AWS Cloud Defined

This Module covers the following subjects:

- **Introduction to the Cloud**: Cloud computing is one of the hottest topics in Information Technology—and not surprisingly, one of the most significant areas of demand in tech employment. This module section introduces you to cloud technologies and details why they are so important and exciting.
- **Introduction to the AWS Cloud**: This module overviews crucial service categories and services. While these services are given in greater detail later in this course, this early look is critical for building your AWS understanding and vocabulary.

In this critical Module, we discuss the various characteristics of technology that would qualify a solution as a "cloud." As you might guess, this module also examines the specifics of Amazon Web Services (AWS) that help make it the most popular (by far) public cloud offering.

FOUNDATION TOPICS

INTRODUCTION TO THE CLOUD

To help us define the "cloud," we turn to the US National Institute for Standards and Technology (NIST). You can find this beneficial site at https://www.nist.gov/. According to the NIST, cloud computing is "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

This statement says a lot, and we need to break it down. Fortunately, the NIST helps us with this as well. Here are the essential **cloud characteristics** you should be aware of:

- **On-demand self-service:** This characteristic means that a customer of cloud technologies (even if you are a customer of your own company's private cloud) can provision and manage resources without the intervention of cloud-hosting administrative personnel. For example, you might need a new webserver to advertise a product or service. You can completely provision, configure, and deploy this web server without contacting anyone responsible for hosting the cloud solution.
- **Broad network access**: This aspect of the cloud states that your cloud resources should be available over the network and accessed through standard mechanisms. These standard access approaches (such as HTTPS) promote the

use of cloud by thin or thick client platforms (for example, mobile phones, tablets, laptops, and workstations).

- **Resource pooling**: The provider's computing resources are pooled to serve multiple clients using a multitenant model. This model allows numerous customers to use the provider's physical hardware securely. At any time, the cloud provider can use different physical and virtual resources dynamically assigned and reassigned according to consumer demand. You should note that this approach provides a sense of location independence because the customer generally has no control or knowledge over the exact location of the provided resources. If required, the customer can typically specify location at a higher level of abstraction (such as country, state, or geographical zone). Examples of resources that are usually pooled include storage, processing, memory, and network bandwidth.
- **Rapid elasticity**: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward in accordance with customer demand. To the consumer, the capabilities available for provisioning often appear unlimited and can be appropriated in any quantity at any time.

- **Measured service**: Cloud systems automatically control and optimise resource use by leveraging a metering capability. This is done by the provider at some level of abstraction appropriate to the type of service. For example, the metering might be based on storage, processing, bandwidth, or active user accounts. Resource usage can be monitored, controlled, and reported, providing transparency for both the service provider and the consumer. This is where the cloud services your IT department pays for are often compared to a utility bill. Like with the electric bill, you can be billed monthly for just those services you used.

Another excellent way to make sense of the many cloud technologies today is to break them down by the "**as a Service"** category they fall under. "as a Service" means that customers "subscribe" to IT resources as needed. The "as a Service" technologies we see today include the following:

- **Software as a Service (SaaS)**: This is currently the most popular cloud model. In this model, customers access a provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface, such as a web browser or a program interface. Note that the customer does not manage or control the underlying cloud infrastructure in this model except for limited user-specific application configuration settings.
- **Platform as a Service (PaaS)**: This model provides a cloud infrastructure to the customer. This permits the customer to deploy onto the cloud infrastructure consumer-created or acquired applications developed for the cloud. The provider ensures the required programming languages, libraries, services, and tools are available for the customer. Typically, this is done on a pay-per-use or charge-per-use basis. Note that in this case, a cloud infrastructure is the

collection of hardware and software that enables the five essential characteristics of cloud computing. The cloud infrastructure can be viewed as containing both a physical layer and an abstraction layer. The physical layer consists of the hardware resources necessary to support the cloud services being provided, and typically includes server, storage, and network components. The abstraction layer consists of the software deployed across the physical layer, which manifests the essential cloud characteristics. Conceptually the abstraction layer sits above the physical layer. Notice also that the customer does not manage or control the underlying cloud infrastructure, including the network, servers, operating systems, and storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

- **Infrastructure as a Service (IaaS)**: Allows the customer to provision processing, storage, networks, and other fundamental computing resources. The customer is then able to deploy and run arbitrary software, which can include operating systems and applications. The customer does not manage or control the underlying cloud infrastructure, but has control over operating systems, storage, and deployed applications. The customer might also have limited control of select networking components such as host firewalls.

How are cloud technologies commonly deployed? These **deployment models** are all in practice today:

- **Private cloud:** The cloud infrastructure is provisioned for exclusive use by a single organisation comprising multiple consumers, which might be business units. It might be owned, managed, and operated by the organisation, a third party, or some combination of both, and it might exist on or off premises.
- **Community cloud:** The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organisations with shared concerns (for example, an overall mission or shared security requirements). It might be owned, managed, and operated by one or more organisations in the community, a third party, or some combination of both. It might exist on or off-premises.
- **Public cloud:** The cloud infrastructure is provisioned for open use by the general public. It might be owned, managed, and operated by a business, academic institution, government organisation, or some combination of the three. It exists on the premises of the cloud provider.
- **Hybrid cloud:** A hybrid cloud is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities but are bound together by standardised or proprietary technology that enables data and application portability. This is a widespread deployment model today.

INTRODUCTION TO THE AWS CLOUD

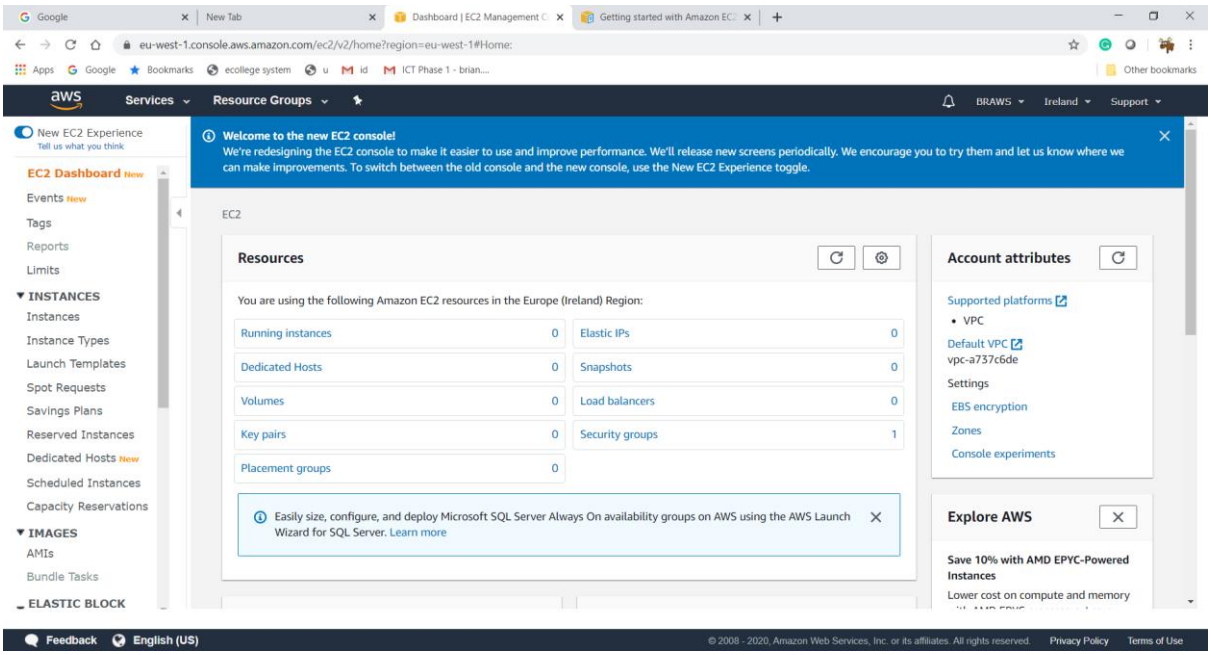
It is time to examine (at a high level) just some of the service categories **in the AWS Cloud** and the services and tools in each. This section provides a vital introduction.

Compute Service

AWS offers many different options for your acquisition and execution of computing resources. This section provides an overview of these many services:

- **Elastic Compute Cloud (EC2):** EC2 is a web service that provides secure and resizable compute resources in the AWS Cloud. The EC2 service allows you to provision and configure capacity with minimal effort. It provides you with easy control of your computing resources. EC2 reduces the time required to obtain and boot new servers (EC2 instances) to minutes. This efficiency allows you to scale capacity vertically (up and down, making your server resources bigger or smaller, respectively) and horizontally (out and in, adding more capacity in the form of more instances) as your computing requirements change. We call this remarkable quality "elasticity," and we cover that in greater detail in Module 2, "Advantages of the AWS Cloud." Figure 1-1 shows two virtual machines running in AWS EC2.

New View EC2 in AWS



- **Lambda**: AWS Lambda lets you run code without the burden of provisioning or managing servers. This code you run against Lambda can be for various aspects of an application or service. When you use Lambda, you upload your code, and Lambda does everything required to run and scale your code with high availability and fault tolerance. Again, you are not required to provision or configure any server infrastructure. Figure 1-2 shows AWS's Lambda graphical user interface (GUI).

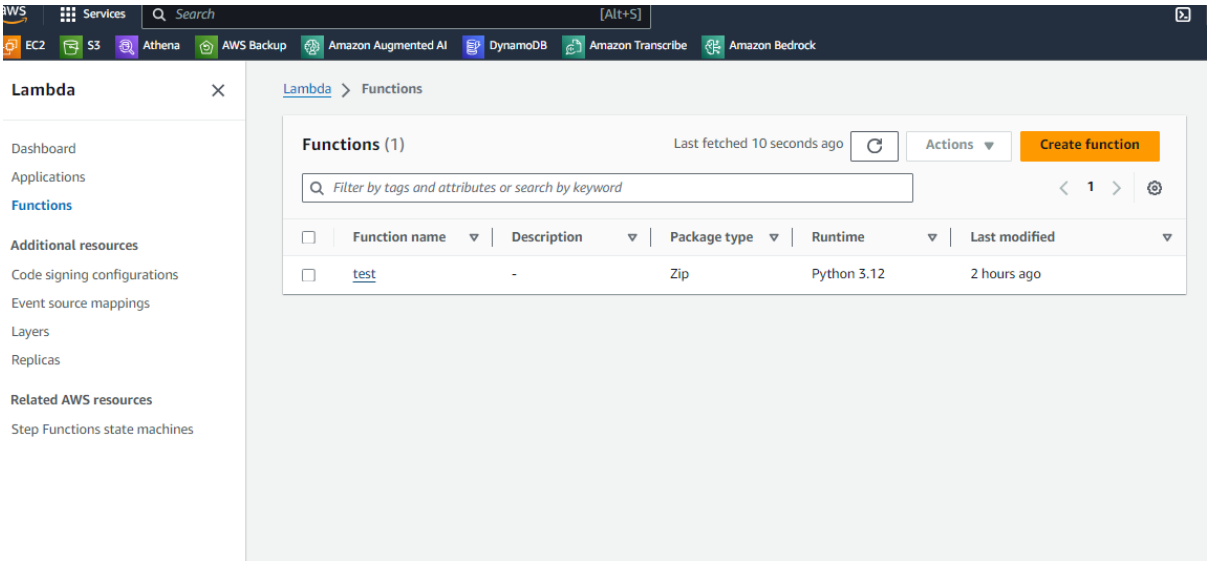


Figure 1-2 AWS Lambda

- **Elastic Beanstalk**: AWS Elastic Beanstalk is an easy-to-use service for deploying and scaling web applications and services developed with popular languages such as Java, PHP, and Python, to name a few. These web applications are run on familiar servers such as Apache, Nginx, Passenger, and Internet Information Services (IIS). Amazingly, with this service, you upload your code, and Elastic Beanstalk automatically handles the deployment, from capacity provisioning to load balancing, auto-scaling, and application health monitoring. Figure 1-3 shows the GUI interface of Elastic Beanstalk.

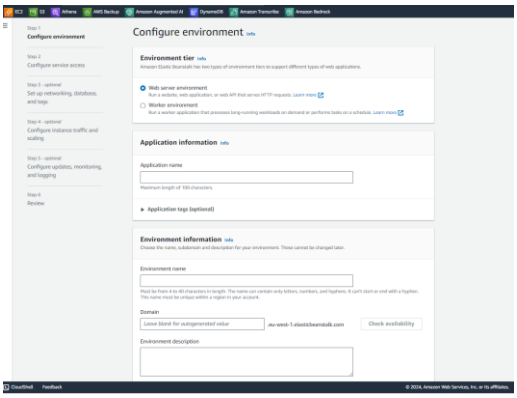


Figure 1-3 Elastic Beanstalk in AWS

- **Elastic Container Service (ECS)**: Amazon Elastic Container Service is a highly scalable, high-performance container management service that supports Docker containers. ECS allows you to run applications efficiently on a managed cluster of EC2 instances, eliminating the need to install, operate, and scale your cluster management infrastructure.

Storage Services

The demands placed on storage for digital information today are higher than ever—and getting bigger all the time. It is no wonder that AWS offers many services in this regard. The list that follows highlights the important ones we will discuss further in this text:

- **Simple Storage Service (S3)**: The AWS Simple Storage Service is object storage with a simple web service interface to store and retrieve any amount of data from anywhere on the web. It is designed to deliver 99.999999999% durability. You can use Amazon S3 for many purposes, such as primary storage for cloud-native applications or a bulk repository (or "data lake") for analytics. It is so flexible and so easy to work with, there are far too many potential uses to list!
- **Elastic Block Store (EBS)**: Elastic Block Store provides persistent block storage volumes for use with EC2 instances in the AWS Cloud. Think of it as being the disk drive in a computer. Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability. EBS volumes provide the consistent and low-latency performance needed to run your workloads. With Amazon EBS, you can scale your usage up or down within minutes while paying a low price for only what you provision.
- **S3 Glacier**: Glacier is a secure, durable, and extremely low-cost storage service for data archiving and long-term backup. With Glacier, you can reliably store large or small amounts of data for as little as $0.004 per gigabyte per month*. Glacier provides three options for access to archives, from a few minutes to several hours. (always check current pricing for your region)
- **Elastic File System (EFS)**: Amazon Elastic File System provides simple, scalable file storage for use with Amazon EC2 instances in the AWS Cloud or by on-premises servers in your organisation. EFS is easy to use and offers a simple interface that allows you to create and configure file systems quickly and easily.

Network Services

Where would we be without the network? Well, back to the Sneakernet, I suppose. Here are some of the critical networking services we discuss in this text:

- **Virtual Private Cloud (VPC)**: Amazon Virtual Private Cloud lets you provision a logically isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking environment, including a selection of your IP address range, the creation of subnets, and configuration of route tables and network gateways. You can use both IPv4 and IPv6 in your VPC for secure and easy access to resources and applications. Figure 1-4 shows elements inside the VPC.

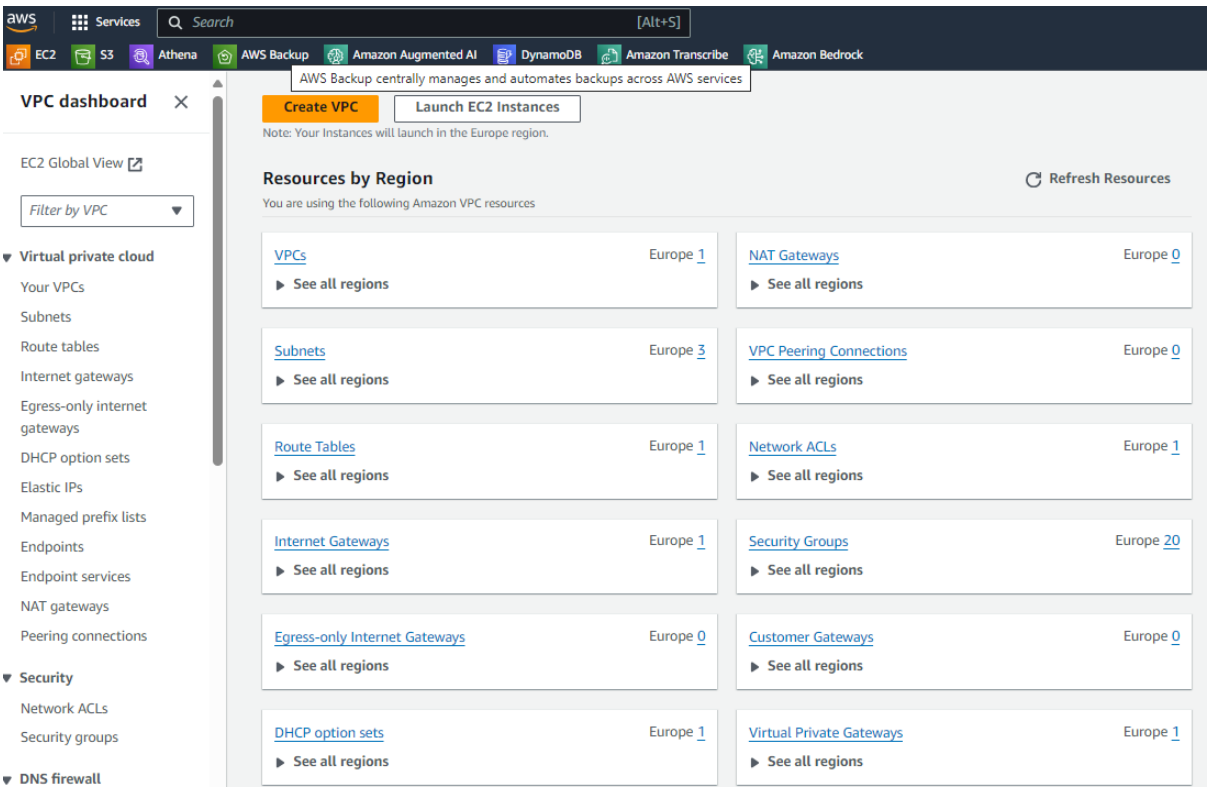


Figure 1-4 The Components of the AWS VPC

- **Route 53**: Amazon Route 53 is a highly available and scalable cloud Domain Name System (DNS) web service. Route 53 effectively directs user requests to infrastructure running in AWS—such as EC2 instances, Elastic Load Balancing load balancers, or S3 buckets—and can also route users to infrastructure outside of AWS. You can use Route 53 to configure DNS health checks to route traffic to healthy endpoints or to independently monitor your application's health and its endpoints.
- **CloudFront**: Amazon CloudFront is a global content delivery network (CDN) service. This service accelerates delivery of your websites, APIs, video content, or other web assets. The service automatically routes requests for your content to the nearest edge location, so it delivers content with the best possible performance.
- **API Gateway**: Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. With a few clicks in the AWS Management Console, you can create an API that acts as a "front door" for applications to access data, business logic, or functionality from your back-end services, such as workloads running on EC2, code running on AWS Lambda, or any web application.
- **Direct Connect**: AWS Direct Connect is a solution that makes it easy to establish a dedicated network connection from your premises to AWS. Using AWS Direct Connect, you can establish private connectivity between AWS and your private network. In many cases, AWS Direct Connect can reduce your network costs,

increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

Database Services

There are many different approaches to databases these days as our data needs have grown more varied and complex. Fortunately, AWS does a great job of keeping up with the advancements in a variety of services:

- **Relational Database Service (RDS)**: RDS makes it easy to set up, operate, and scale a relational database in the cloud. It provides six database engines to choose from: Aurora, PostgreSQL, MySQL, MariaDB, Oracle, and Microsoft SQL Server.
- **DynamoDB**: Amazon DynamoDB is a fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale. It is an excellent fit for mobile, web, gaming, ad tech, Internet of Things (IoT), and many other applications.
- **ElastiCache**: ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud. The service improves the performance of web applications by allowing you to retrieve information from fast, managed, in-memory caches instead of relying entirely on slower disk-based databases. Interestingly, ElastiCache is not an AWS proprietary solution and runs the standardised Redis or Memcached solutions.
- **Redshift**: Redshift is a fast, fully managed, petabyte-scale data warehouse that makes it simple and cost-effective to analyse all your data using your existing business intelligence tools.
-

Security Services

If you follow best practices, you can be more secure with the cloud than with any approach you could take in your own data centre. Here are the major technologies in the security area you should be aware of:

- **Identity and Access Management (IAM)**: AWS Identity and Access Management (IAM) enables you to securely control access to AWS services and resources for your users. Using IAM, you can create and manage AWS users and groups and use permissions to allow and deny their access to AWS resources. Figures 1-5 shows IAM in AWS's GUI.

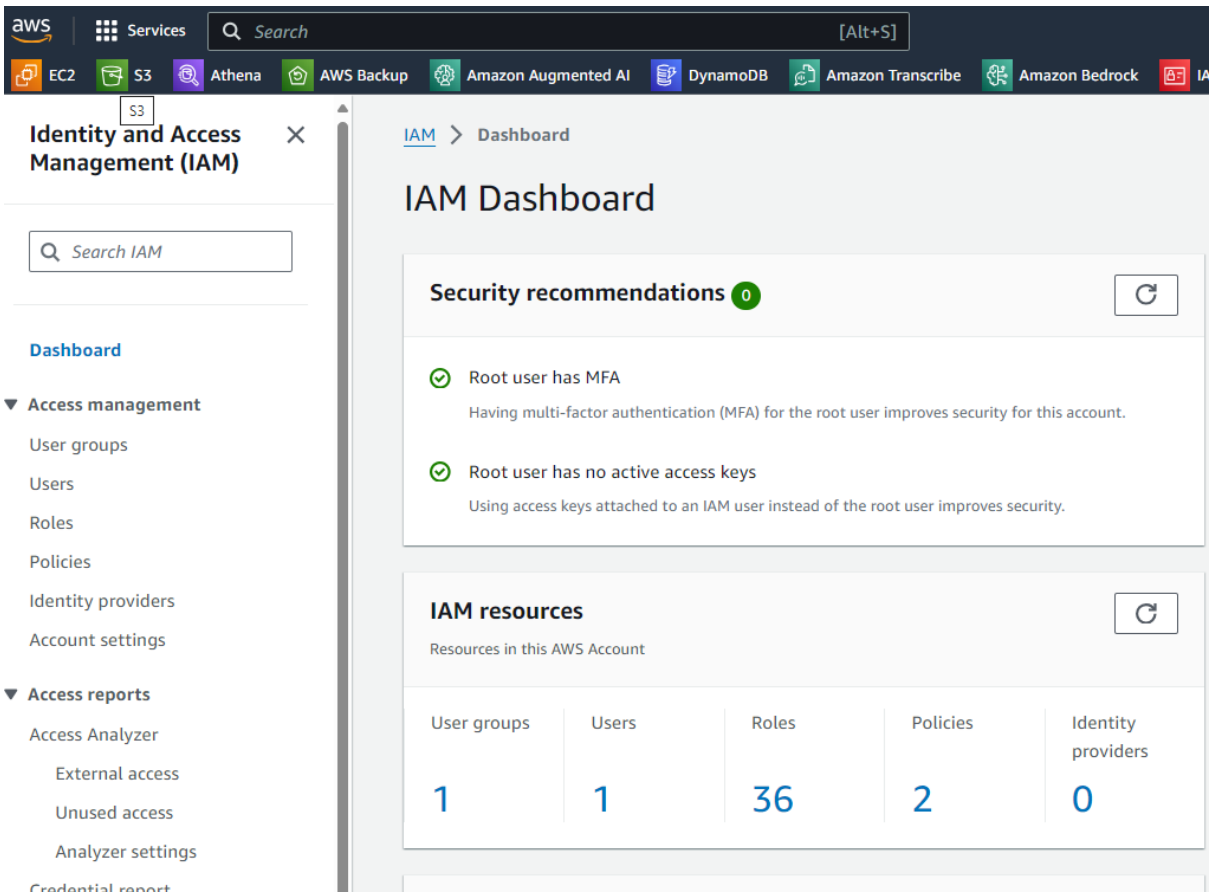


Figure 1-5 IAM in AWS

- **Security groups**: AWS security groups are associated with EC2 instances and provide security at the protocol and port access level. Each security group contains a set of rules that filter traffic coming into and out of an EC2 instance. If no rule explicitly permits a particular data packet, it will be dropped. Security groups can also be applied to many other services within your VPC, including ELB, RDS, Redshift, ElastiCache, and others.
- **Network ACLs**: Network access control lists control traffic moving between your AWS VPC subnets. They function like traditional access control lists and consist of permit and deny entries for various addresses and ports.

Automation and Application Support

Many tools foster the deployment of applications and automation in AWS. Here are some of the major ones:

- **CodeDeploy**: AWS CodeDeploy is a fully managed deployment service that automates software deployments to various compute services such as EC2, Lambda, and your on-premises servers. It makes it easier for you to rapidly release new features, helps you avoid downtime during application deployment, and handles the complexity of updating your applications.

- **CloudFormation**: AWS CloudFormation makes it easy to provision and configure related AWS resources based on a template. The tool even offers a designer that permits you to build architectures in templated code from your "sketches" using the design tool.
-
- ***OpsWorks****: (deprecated; see Systems Manager/CloudFormation) AWS OpsWorks is a configuration management service that uses Chef or Puppet. These automation platforms treat server configurations as code. OpsWorks uses Chef or Puppet to automate configuring, deploying, and managing servers across your EC2 instances or on-premises compute environments. Chef and Puppet can still be used on AWS without using OpsWorks:*

https://www.chef.io/

https://www.puppet.com/

-

Management Tools

Here are some tools to help you manage all that important stuff in the AWS cloud.

- **Service Catalog**: AWS Service Catalog allows organisations to create and manage catalogues of IT services approved for AWS use. These IT services can include everything from virtual machine images, servers, software, and databases to complete multi-tier application architectures. AWS Service Catalog allows you to centrally manage commonly deployed IT services. It helps you achieve consistent governance and meet your compliance requirements while enabling users to deploy only the approved IT services they need quickly.
- **Systems Manager**: AWS Systems Manager gives you visibility and control of your infrastructure on AWS. It provides a unified user interface so you can view operational data from multiple AWS services and automate operational tasks across your AWS resources. With Systems Manager, you can group resources (EC2 instances, S3 buckets, or RDS instances) by application, view operational data for monitoring and troubleshooting, and take action on your groups of resources.
- **Trusted Advisor**: AWS Trusted Advisor is an online resource to help you reduce cost, increase performance, and improve security by optimising your AWS environment. Trusted Advisor provides real-time guidance to help you provision your resources following AWS best practices.

Monitoring

Do you need to accurately track the performance and status of your resources and services? AWS offers tools for that, too.

- **CloudWatch**: Amazon CloudWatch is a monitoring service for AWS Cloud resources and the applications you run on AWS. CloudWatch can collect and track metrics, collect and monitor log files, set alarms, and automatically react to changes in your AWS resources.
- **CloudTrail**: AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. Features include detailed reports of recorded information, which can include the API caller's identity, the time of the API call, the source IP address of the API caller, the request parameters, and the response elements returned by the AWS service.

EXAM PREPARATION TASKS

REVIEW ALL KEY TOPICS

Review the all topics in this Module. Table 1-2 lists these key topics and the page numbers on which each is found.

Define all of the key terms – and check your answers in the glossary

Then do the quiz.

DEFINE KEY TERMS

Define the following key terms from this Module and check your answers in the Glossary:

Elasticity	S3	Redshift
SaaS	EBS	IAM
PaaS	S3 Glacier	security groups
IaaS	EFS	network ACLs
private cloud	VPC	CodeDeploy
community cloud	Route 53	CloudFormation
public cloud	CloudFront	OpsWorks
hybrid cloud	API Gateway	Service Catalog
EC2	Direct Connect	Systems Manager
Lambda	RDS	Trusted Advisor
Elastic Beanstalk	DynamoDB	CloudWatch
ECS	ElastiCache	CloudTrail