**Chapter 2**

**What is Amazon Elastic Compute Cloud (Amazon EC2)?**

* Provides secure, resizable compute capacity in the cloud as Amazon EC2 instances.
* They serve client request
* Provides a wide selection of instance types to optimized to fit different use cases.

**With an Amazon EC2 instance you can use a virtual server to run applications in the AWS Cloud**

* You can provision and launch an Amazon EC2 instance within minutes.
* You can stop using it when you have finished running a workload.
* You pay only for the compute time you use when an instance is running, not when it is stopped or terminated.
* You can save costs by paying only for server capacity that you need or want.

**How Amazon EC2 Works**

Graphical user interface, application, Teams

Description automatically generated

1. Launch the instance, begin by selecting the template with basics configuration for your instance. The configurations include the operating system, application server, or application, instance type which is the specific hardware configuration.
2. Connect to the instance
3. After you connected to the instance you can begin using it.

**5 types of Amazons EC2 instance**

Are optimized for different tasks.

1. **General Purpose Instances**

Provide a balance of compute memory, memory and networking resources, and can be used for variety of diverse workloads, such as application server, gaming servers, backup server, small and medium databases.

These instances are ideal for applications that use these resources in equal proportions such as web servers and code repositories.

Example -> Amazon EC2 Mac Instances, Amazon EC2 Tg4 instances

1. **Compute optimised instances**

Are ideal for compute-bound applications that benefits from high-performance processors. Like general purpose instances you can use compute optimized instances for workloads such as web, application, and gaming servers

**Difference with compute-bound applications**

Are ideal for high-performance web servers, compute-intensive applications servers, and dedicated gaming servers. You can also use compute optimised instance for batch processing working loads that require processing many transactions in a single group.

1. **Memory Optimized Instances**

Memory optimised instances are designed to deliver fast performance for workloads that process large datasets in memory. Enable you to run workloads with high memory needs and receive great performance.

1. **Accelerated computing instances**

Use hardware accelerators, or coprocessors, to perform, some functions more efficiently than is possible in software running on CPU’s.

Examples -> include floating-point number calculations, graphics processing, and data pattern matching.

1. **Storage optimized instances**

Are designed for workloads that require high, sequential read and write access to large datasets on local storage.

Example -> of workloads suitable for storage optimized instances include distributed file systems, data warehousing application, and high-frequency online transaction processing (OLTP) systems.

**5 types of Amazon EC2 Pricing**

1. On-demand -> are ideal short term, instances run continuously until you stop them, and you pay for only the compute time you use
2. Amazon EC2 Savings Plans -> enable you to reduce your computes costs by committing to a consistent amount of usage for a 1-year or 3-year term. You can save up to 72%
3. Reserved Instances -> are billing discount to the use of On-Demand instance in your account. 1 -3 years
4. Spot Instances -> are ideally for workloads with flexible start and end time, or that can withstand interruptions.
5. Dedicated Hots -> are physical servers with Amazon EC2 instances capacity that is fully dedicated to your use

**Scaling Amazon EC2**

Scalability -> involves beginning with only resources you need and designing your architecture to automatically respond to changing demand by scaling out or in.

**Amazon EC2 Auto Scaling**

Enables you to automatically add or remove Amazon EC2 instances in response to changing application demand.

Within Amazon EC2 Auto Scaling you can use two approaches:

* Dynamic scaling -> responds to changing demand.
* Predictive scaling -> automatically schedules the right number of Amazon EC2 instances based on predicted demand.

**Maximum Amazon EC2 instances.**

When you create auto scaling group, you can set the minimum number of Amazon EC2 instances.

**What is minimum capacity?**

Is the number of Amazon EC2 instances that launch immediately after you have created the Auto Scaling group, Has minimum of one Amazon EC2 instance

**Directing traffic with Elastic Load Balancing.**

**What is Elastic Load Balancing?**

* Is the AWS service that automatically distributes incoming application across multiple resources, such as Amazon EC2 instances.

Elastic Load Balancing and Amazon EC2 Auto scaling are separate services, they work together to help ensure that applications running in Amazon EC2 can provide high performance availability.

**Messaging and queuing**

To help maintain application availability when a single component fails, you can design your application through a microservices approach.

**Microservice approach**

* Applications are loosely coupled. In this case, if a single component fails the other components continue to work because they are communicating with each other.
* The loose coupling prevents the entire application from failing.

**Two services facilitate application integration**

1. Amazon Simple Notification Service (Amazon SNS)
2. Amazon Simple Queue Service (Amazon SQS)
3. **Amazon Simple Notification Service (Amazon SNS)**

* Is a publish/subscribe service. Using Amazon SNS topic, a publisher publishes messages to subscribers.
* In Amazon SNS, subscribers can be web servers, email address, AWS Lambda functions, or several other options.

1. **Amazon Simple Queue Service (Amazon SQS)**

* Is a message queuing service. You can send, store, and retrieve messages between software components, without losing messages or requiring other services to be available.
* In Amazon SQS, an application sends messages into a queue. A user or service retrieve a message from the queue, process it, then deletes is from the queue

**Additional Compute Services**

Amazon EC2 -> a service that lets you run virtual service in the cloud.

If you have applications that you want to run in Amazon EC2, you must do the following:

1. Provision instances.
2. Upload your code.
3. Continue to manage the instances while your application is running.

Serverless -> means that your code runs on servers, but you do not need to provision or manage these servers.

**Benefits of serverless computing**

1. You can focus more on innovating new products and features instead of maintaining the servers
2. Flexibility to scale serverless applications automatically.
3. Serverless computing can adjust the applications capacity by modifying the units of consumptions, such as throughput and memory.

**AWS Lambda**

* Is AWS service for serverless
* Is a service that lets you run code without needing to provision or manage servers

**How AWS Lambda works?**

1. You upload your code to Lamda
2. You set your code to trigger from an event source, such as AWS services, mobile applications or HTTP endpoints
3. Code runs only when triggered
4. Pay only for the compute time use.

**Containers**

Containers provides you with a standard way to package your applications code and dependencies into a single object.

**Amazon Elastic Container Service (Amazon ECS)**

Is a highly scalable, high performance container management system that enables you to run and scale containerised application on AWS

**Amazon ECS supports Docker containers.**

Docker -> is a software platform that enables you to build, test and deploy applications quickly.

AWS supports the use of open-source Docker Community Edition and subscription-based Docker Enterprise Edition.

**Amazon Elastic Kubernetes Services (Amazon EKS)**

* Is a fully managed service that you can use to run Kubernetes on AWS

**Kubernetes**

* Is an open-source software that enables you to deploy and manage containerised application at scale.

**What is AWS Fargate**

* Is a serverless compute engine for containers. It works both Amazon ECS and Amazon EKS
* Manages server infrastructure for you.