Assignment-7

1. Explain Cloud Watch? Why we need Monitoring?

Any application you run on your AWS cloud can be monitored by using the Amazon CloudWatch, it will monitor your resource files on the cloud and collects the log metrics files. It is basically a metrics repository like Amazon EC2 and you retrieve the statistics based on those metrics.

Amazon CloudWatch provides you a most reliable, scalable and flexible way to monitor your resource or application on cloud. You can manage resource utilization of your servers.

The main reasons of monitoring are:

* To make sure that your website/cloud and app are always online.
* To make your application secure for your customers.
* It can help you monitor the resultant performance and cost of your application.
* Troubleshooting and recommendation for future and how to avoid the existing errors.

1. What is VPC? What are the key components of VPC?

A virtual private cloud (VPC) is defined as an isolated private cloud environment typically hosted and secured within another cloud, which is usually a public cloud.

* Subnet: A segment of a VPC’s where you can place groups to isolated resources.
* Internet Gateway: VPC side of a connection to utilize public Internet.
* NAT Gateway: A highly available, managed Network Address Translation (NAT) service for your resources in a private subnet to access the Internet.
* Virtual private gateway: The Amazon VPC side of a VPN connection for secure transactions.
* Peering Connection: To route traffic via private IP addresses between two peered VPCs.
* VPC Endpoints: Enables private connectivity for your service in AWS without using an Internet Gateway, VPN, Network Address Translation (NAT) devices, or firewall proxies.
* Egress-only Internet Gateway: A stateful gateway that provides egress only access for IPv6 traffic from the VPC to the Internet.

1. Explain Route table, Internet gateway, NAT gateway?

Internet Gateway

* Internet Gateway (IGW) is a horizontally scaled, redundant, and highly available VPC component that allows communication between your VPC and the internet.
* Internet Gateway enables resources (like EC2 instances) in public subnets to connect to the internet. Similarly, resources on the internet can initiate a connection to resources in your subnet using the public.
* If a VPC does not have an Internet Gateway, then the resources in the VPC cannot be accessed from the Internet (unless the traffic flows via a Corporate Network and VPN/Direct Connect).

NAT Gateway

* NAT Gateway (NGW) is a managed Network Address Translation (NAT) service.
* NAT Gateway does something similar to Internet Gateway (IGW), but it only works one way: Instances in a private subnet can connect to services outside your VPC but external services cannot initiate a connection with those instances.
* If you have resources in multiple Availability Zones and they share one NAT gateway, and if the NAT gateway’s Availability Zone is down, resources in the other Availability Zones lose internet access.
* You can associate exactly one Elastic IP address with a public NAT gateway.

Route Table

* Each subnet in your VPC must be associated with a route table, which controls the routing for the subnet (subnet route table). A subnet can be explicitly associated with custom route table.
* Main route table automatically comes with VPC. It controls the routing for all subnets that are not explicitly associated with any other route table.
* Custom route table is empty, by default and you add routes as needed.
* A subnet can only be associated with one route table at a time, but you can associate multiple subnets with the same subnet route table.
* Every route table contains a local route for communication within the VPC.
* Gateway route table — it’s associated with an internet gateway or a virtual private gateway (gateway route table).
* A gateway route table supports routes where the target is local or an elastic network interface in VPC.
* Each route in a table specifies a destination and a target.

1. Different types of IP and What is VPC peering?

Public IP Address

A public IP address is an address where one primary address is associated with your whole network. In this type of IP address, each of the connected devices has the same IP address.

This type of public IP address is provided to your router by your ISP.

Private IP Address

A private IP address is a unique IP number assigned to every device that connects to your home internet network, which includes devices like computers, tablets, smartphones, which is used in your household.

It also likely includes all types of Bluetooth devices you use, like printers or printers, smart devices like TV, etc.

VPC Peering

A VPC peering connection is a networking connection between two VPCs that enables you to route traffic between them using private IPv4 addresses or IPv6 addresses. Instances in either VPC can communicate with each other as if they are within the same network. You can create a VPC peering connection between your own VPCs, or with a VPC in another AWS account.

1. What is Autoscaling? Different types of scaling?

An EC2 Auto Scaling functionality mainly centres around maintaining a specified number of instances in a group, or automatically increasing and decreasing the size of a group according to different aspects, such as application loads or predetermined schedules.

The different types of Scaling are: -

* 1. Vertical Scaling: - **Vertical scaling means that you scale by adding more power (CPU, RAM) to an existing machine**.
  2. Horizontal Scaling: - **Horizontal scaling means that you scale by adding more machines** into your pool of resources

1. Why do we need Autoscaling?

When creating an Auto Scaling group, you have to specify a number of parameters, including the minimum and maximum number of instances to have in the clusters, and a criterion for triggering adding or removing an instance from the cluster. The choice of parameter values will determine the cost of running the cluster.

The minimum number of instances should be enough to meet the base application load, but not have too much unused capacity. A single instance configured to meet these low-end needs may seem like the optimal choice for the instance type, but that’s not necessarily the case. In addition to thinking about the minimum resources required, you should consider the optimal increment for adding instances.

1. What is load balancer? Explain different types of load balancer.

Load balancing is the process of distributing network traffic across multiple servers. This ensures no single server bears too much demand. By spreading the work evenly, load balancing improves application responsiveness. It also increases availability of applications and websites for users. Modern applications cannot run without load balancers.

The different types of load balancers

* Application load balancer

Application load balancer is an elastic load balancer that runs on the application layer (layer 7 of the OSI layer model) and receives only HTTP and HTTPS connection requests. It acts as a single point of entry for the incoming requests and distributes the traffic among targets in a target group.

* Gateway load balancer

Gateway load balancer can be used to exchange traffic securely across VPCs by using the gateway load balance endpoints. Gateway load balancer endpoint is a VPC endpoint that is used for private connectivity between the virtual appliance and application server in different VPCs.

* Network load balance

Network load balancer can send requests to multiple targets in different availability zones. While creating the load balancer, will ask for the availability zones in which you want to deploy the load balancer. The elastic load balancer then creates nodes in all the selected availability zones. The network load balancer distributes the traffic to its own availability zone only. Still, if you have enabled the **cross-zone load balancing**feature, it will distribute the load across different availability zones.

1. What is Amazon RDS?

Amazon Relational Database Service (RDS) is a managed SQL database service provided by Amazon Web Services (AWS). Amazon RDS supports an array of database engines to store and organize data. It also helps with [relational database](https://www.techtarget.com/searchdatamanagement/definition/relational-database) management tasks, such as data migration, backup, recovery and patching.

1. What is launch template, auto scaling group, Target group?

Launch Template: -

Launch template that contains the configuration information to launch an instance. You can use launch templates to store launch parameters so that you do not have to specify them every time you launch an instance.

Auto Scaling Group: -

An EC2 auto scaling group is a logical collection of several Amazon EC2 instances used for management and scaling purposes. Auto scaling groups let you use core features of the EC2 Auto Scaling service, including health checks, minimum/maximum instances and scaling policies.

Target Group: -

Target group is used to route requests to one or more registered targets. When you create each listener rule, you specify a target group and conditions. When a rule condition is met, traffic is forwarded to the corresponding target group.