AWS-Compute

AWS Scalability:
Vertical Scaling: Here we upgrade our machine by adding extra resources which can meet up our requirements. Risk to downtime.
Horizontal Scaling: Here we add multiple instances which could share the load across them. High availability.

Auto scaling: is a service that allows you to automatically adjust the number of EC2 instances in your application's fleet based on defined policies.
To Create Autoscaling group:
- Create AMI with Applications pre installed
- Decide which metrics to use when your ec2 has to be scaled out or in
- Choose the AZ's (subnets) where scaling has to happned
- Choose the scaling methodologies

Fixed Scaling: fixed count is maintained by AWS. For eg: you have give fixed scaling count to 5, it creates 5 ec2 instances in chosen subnets (AZ's). By any chance if the ec2 instance is down, aws will create one more to maintain count of 5.

Auto scaling: In this case we specify min ec2 instances, max ec2 instances, desired count. The count can be scale in or scale out, desired count will be dynamic depending.

What is elastic bean stack?
Using this aws service we can easily deploy our applications without worrying about underlying the infrastructure details. This service will provide all the facilities like autoscaling, loadbalancing, monitoring& logging, health check etc for our application.

What are different types of instances?
Burstable Performance Instances (T3):

General Purpose Instances (T2, M5, M6):These instances provide a balance of compute, memory, and network resources. They are suitable for a variety of workloads, including web servers, small databases, and development environments.

Compute Optimized Instances (C5, C6): These instances are designed for compute-intensive workloads that require high-performance processors. They are ideal for applications such as scientific modeling, batch processing, and gaming servers.

Memory Optimized Instances (R5, R6, X1): These instances offer high memory capacity, making them suitable for memory-intensive workloads such as in-memory databases, real-time big data processing, and high-performance computing.

Storage Optimized Instances (I3, D2): These instances are optimized for high-speed, low-latency storage. They are well-suited for data warehousing, large-scale databases, and distributed file systems.

Accelerated Computing: Accelerated computing instances use hardware accelerators, or co-processors, to perform functions, such as floating point number calculations, graphics processing, or data pattern matching, more efficiently than is possible in software running on CPUs.

What is cloud watch?
CloudWatch collects monitoring and operational data in the form of logs, metrics, and events, and visualizes it using automated dashboards so you can get a unified view of your AWS resources, applications, and services that run in AWS and on-premises.

Simple Routing: Sends traffic to a single resource (e.g., an IP address or a load balancer).

Weighted Routing: Distributes traffic based on specified weights assigned to different resources, allowing you to split traffic between multiple endpoints.

Latency-based Routing: Routes traffic to the resource with the lowest latency based on the user's geographic location.

Geolocation Routing: Routes traffic based on the user's geographic location.

Failover Routing: Routes traffic to a secondary resource in case the primary resource becomes unavailable.

Multi-Value Answer Routing: Returns multiple healthy records for a single DNS query, providing high availability and load balancing at the DNS level.

Routing Policies:
Geo-proximity routing policy:
IP based routing policy:

Network Load Balancer: Best suited for extreme performance. TCP/UPD/TLS. Operate at layer 4 maintains ultra low latency new generation V2.

What is load balancers?
It helps to distribute traffic across multiple targets and varying load of our application traffic

What is Route 53?
Amazon Route 53 is a scalable and highly available domain name system (DNS) web service provided by AWS. It manages the routing of internet traffic to various AWS resources, such as EC2 instances, load balancers, S3 buckets, and more.

Domain Registration: Route 53 allows you to register and manage domain names. You can either register a new domain directly through Route 53 or transfer an existing domain from another registrar.

DNS Management: Once you have a registered domain, Route 53 acts as your DNS service provider. You can create hosted zones in Route 53, which represent the domain names you want to manage.

DNS Record Management: Within a hosted zone, you can create DNS records that define the routing of traffic for specific domain names. Some commonly used DNS record types include:

A Record: Associates a domain name with an IPv4 address.
AAAA Record: Associates a domain name with an IPv6 address.
CNAME Record: Maps a domain name to another domain name (canonical name).
Alias Record: Maps a domain name to an AWS resource (e.g., an ELB, S3 bucket, or CloudFront distribution).

Features of route 53:
Health Checks: Route 53 can perform health checks on your resources to ensure they are available and responding correctly. You can configure health checks to monitor the health of your resources and Route 53 can automatically route traffic away from unhealthy resources.

Instance Type Pricing:
On-Demand Instances: It is highly available and costly and pay as you use.
Reserved Instance: Need to reserve for 1 year or 3 years our instance we will get 72% discount compared to on demand instance. We can pay like full upfront, no upfront, partial upfront.
Savings plan: It is better version of reserved instance, here we can upgrade instance family size everything.
Dedicated Host: Here sever will create on single host. It shares same hardware between servers, less security, more pricing
Spot Instance: It will be allocated depends upon the biding and it is not permanent if some other instances requires aws will be remove the spot instance to allocate them.