ARYATECHNO

AWS EC2 Auto Scaling

Topics : AWS
Written on December 01, 2023

AWS Auto Scaling is a service provided by Amazon Web Services (AWS) that enables you to automatically adjust the number of Amazon EC2 instances in a group to accommodate changes in demand for your applications. This service helps you maintain application availability, optimize performance, and minimize costs by automatically adjusting the number of instances based on policies and conditions you define.

Here are key features and concepts related to AWS Auto Scaling:

1. **Auto Scaling Groups (ASGs):**
   - An Auto Scaling Group is a logical grouping of EC2 instances that share similar characteristics and are created from a common Amazon Machine Image (AMI). ASGs define the scaling policies and other parameters for the instances.
   - ASGs are associated with launch configurations or launch templates, which specify the configuration details for the instances (e.g., instance type, AMI, security groups).
2. **Scaling Policies:**
   - Scaling policies define the conditions under which Auto Scaling should scale the group. These policies can be based on various metrics, such as CPU utilization, network traffic, or custom CloudWatch alarms.
   - There are two types of scaling policies: **Simple Scaling Policies** and **Step Scaling Policies.**
     - *Simple Scaling Policies:* Increase or decrease the desired capacity of the group based on a single scaling adjustment.
     - *Step Scaling Policies:* Adjust the capacity based on a set of predefined scaling adjustments that vary with the size of the alarm breach.
3. **Dynamic Scaling and Predictive Scaling:**
   - **Dynamic Scaling:** Auto Scaling dynamically adjusts the number of running instances based on demand. It adds or removes instances as needed, following the scaling policies.
   - **Predictive Scaling:** This feature uses machine learning algorithms to predict future demand for your application. It automatically adjusts the number of instances in advance of expected changes in demand.
4. **Cooldown Period:**
   - The cooldown period is a configurable time interval during which Auto Scaling waits

before allowing further scaling actions. This helps prevent the group from launching or terminating additional instances before the effects of previous actions are visible.

5. **Integration with Elastic Load Balancing (ELB):**
   - Auto Scaling works seamlessly with Elastic Load Balancing to distribute incoming traffic across multiple EC2 instances. This helps ensure that the load is evenly distributed and provides fault tolerance.
6. **Integration with Launch Templates:**
   - Auto Scaling Groups can use launch templates to specify the settings for new instances, providing more flexibility and the ability to use the latest features of EC2 instances.
7. **Lifecycle Hooks:**
   - Lifecycle hooks enable you to perform custom actions before instances are launched or terminated. This is useful for tasks such as validating application deployments or performing cleanup operations.
8. **Termination Policies:**
   - Termination policies define the criteria for selecting instances to terminate when scaling down. You can choose from options like OldestInstance, NewestInstance, OldestLaunchConfiguration, and Default.
9. **Auto Scaling Plans:**
   - Auto Scaling supports the creation of scaling plans, which are predefined scaling strategies based on predicted demand patterns. These plans help simplify the process of configuring Auto Scaling for applications with predictable load changes.
10. **Integration with AWS CloudWatch:**
    - Auto Scaling leverages Amazon CloudWatch for monitoring and triggering scaling actions based on CloudWatch alarms.

Auto Scaling helps you achieve better fault tolerance, availability, and cost optimization by automatically adjusting your EC2 capacity based on real-time demand. It is particularly beneficial for applications with varying workloads or those experiencing periodic traffic fluctuations.