Amazon EC2 Auto Scaling

Inhalt

Was ist EC2 Auto Scaling?

- Zustandsprüfung (Health Checks)
- Benutzerdefinierte Richtlinien (von CloudWatch gesteuert)
- Zeitpläne
- Andere Kriterien (progammgesteuert)
- Manuell mit eingestellter gewünschter Kapazität

-> **Skalieren um die Nachfrage zu befriedigen und dabei Kosten zu senken**

Welche Schritte sind notwendig?

- Um zu Regeln welche Insances gestartet werden sollen ist eine Vorlage nötig
- Es gibt den legacy (alt) Weg über Launch Configuration -> Nicht verwenden, da nicht von Amazon empfohlen
- Benutze Launch Template, da neuer und empfohlen

Konfigurieren und Verwalten von Launch Templates

Um eine Startvorlage (Launch Template) zu erstellen, sind folgende Konfigurationen wichtig:

- Amazon Machine Image (AMI)
- Instance Typ
- VPC
- Sicherheitsgruppe
- Speicher
- Instance Schlüsselpaar
- IAM Rolle (Instance Role)
- Benutzerdaten (User Data)
- Markierung

Auto Scaling in Action

Logische Gruppen von EC2 Instances

Automatisch Skalieren zwischen:

- Minimum

- Erwünscht (optional)
- Maximum

Führt Health Checks durch um die Gruppengröße beizubehalten

Instances über AZs verteilen und ausgleichen

Auto Scaling Richtlinie (Policy)

Parameter zum durchführen der Auto Scaling Aktion

Wie werden Richtlinien ausgelöst?

- CloudWatch Alarm
- Zielverfolgung (Target Tracking)
- Geplant
- Manuell

Auf- oder runterskalieren und um wie viel:

- ChangeInCapacity (+/- #)
- ExactCapacity
- ChangeInPercent (+/- %)

Health Checks

- Behält den Gesundheitszustand bei
- Beendet ungesunde (unhealthy) Instances
- Verwendet standardmäßige EC2 Statusprüfungen
- Erweiterbar durch Load Balancer
  - Load Balancer Health Checks
  - EC2 Instance Checks

Kann manuell gesetzt werden:

```
aws autoscaling set-instance-health
```

Kündigungsrichtlinie

- Bestimmt welche Instance bei *scale in* beendet wird
- Basierend auf 2 Faktoren:
  - AZ mit den meisten Instances
  - Mehrere Richtlinien (In der aufgeführten Reihenfolge)

Kündigungsrichtlinie	Beschreibung
OldestInstance	Wählt die am längsten laufende Instance
NewestInstance	Wählt die am kürzesten laufende Instance
OldestLaunchTemplate	Beendet die Instance mit der ältesten Startvorlage (Standard)
ClosestToNextInstanceHour	Beendet die Instance, die der nächsten abrechenbaren Stunde am nächsten kommt (Standard)

Steady-State Gruppe

- Auto Scaling Group mit gleichem:
  - Minimal
  - Maximal
  - Gewünschter Wert
- Die Instance wird automatisch neu erstellt, wenn sie ungesund wird oder wenn AZ ausfällt
- Trotzdem noch mögliche Ausfallzeiten beim Recycling von Instances

Skalierungsarten

Geplante Skalierung

- Skalierung anhand eines Zeitplan
- ZB. Freitag große Sale Aktion

Dynamische Skalierung

- Target Tracking:
  - Basierend auf Zielwert für bestimmte Metrik (z.B. CPU)
- Step Scaling:
  - Basierend auf CloudWatch Alarm (z.B. CPU > 70%)
- Simple Scaling:
  - Auf Grundlage von Skalierungsanpassung

Prädiktive Skalierung

- Last prognostizieren
- Mindestkapazität planen
- Basierend auf ML (Nutzungsmuster)

Zusammenfassung

- Hilft bei der Aufrechterhaltung der Anwendungsverfügbarkeit
- EC2 Instancen werden gemäß Launch Template hinzugefügt oder entfernt
- Besteht aus 3 Teilen:
  - Launch Template
  - Auto Scaling Group
  - Skalierungsrichtlinie (Policy)
- Skalierung kann basieren auf:
  - Instance Health
  - CloudWatch Alarms
  - Zeitplan oder Vorhersage