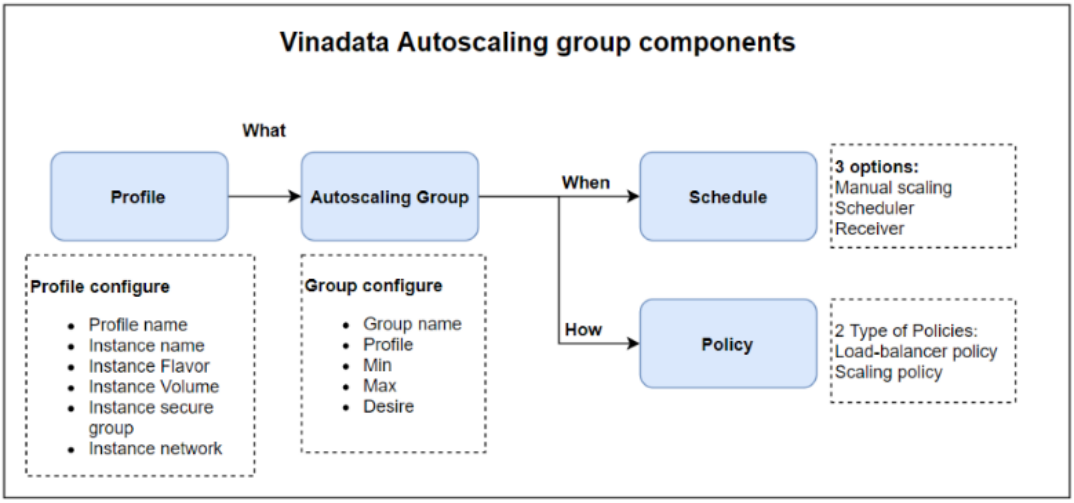
Dịch Vụ Auto Scaling HCM 02

Auto-scaling là dịch vụ tự động tăng thêm hoặc giảm bớt (scale) máy chủ (VM) theo điều kiện do user đặt ra. Auto-scaling đảm bảo tính sẵn sàng cao của hệ thống mà vẫn tối ưu chi phí

Các thành phần của Autoscaling



Profile: Là thiết lập cấu hình mẫu của instance bao gồm Flavor, Volume, Secure group, network. Auto scaling group sẽ dựa trên thiết lập của profile để scal e Intance có cấu hình tương ứng.

Mỗi profile có thể gắn nhiều scaling group, tuy nhiên mỗi scaling group chỉ có thể được gắn với 1 profile duy nhất.

Cấu hình mẫu của profile sau khi tạo thành công không thể edit. Nếu có nhu cầu thay đổi thì bạn sẽ cần tạo profile mới.

Auto Scaling group: Một auto scaling group sẽ gồm nhiều instances có cùng một loại cấu hình được định nghĩa tại profile. Vì vậy mỗi auto scaling group cần phải được gắn với 1 profile và không được có nhiều hơn 1 profile cho mỗi group.

Auto scaling group sẽ tạo số lượng instance tương ứng với thiết lập “desired capacity” và duy trì số lượng này cho dù có 1 instance bị lỗi (un-healthy). Các tính năng “healthy check” & “Recover node” sẽ kiểm tra, terminate các instance bị lỗi đồng thời tạo instance mới, giúp cho số lượng instance luôn đảm bảo .

Schedule: Component này sẽ xác định thời điểm nào auto-scale sẽ thực hiện, Có 3 cách thiết lập là:

- Manual: Bằng cách thay đổi số “desired capacity” tại giao diện cấu hình Auto scaling group, số lượng instance sẽ tăng thêm / giảm xuống tùy theo nhu cầu.
- Scheduler: thiết lập lịch chạy tự động tùy theo nhu cầu. Ví dụ thiết lập lịch tăng lượng instance khi ở giờ cao điểm và thiết lập lịch giảm số instanc e xuống ở giờ thấp điểm để tối ưu chi phí.
- Receiver: Cho phép bạn scale (In/out) một auto-scaling group thông qua http url. Webhook receiver thường được sử dụng khi bạn có sẵn một mo nitor tool.

Policy: Khi gắn vào 1 auto-scaling group, policy sẽ định nghĩa cách thức scale của group. Policy gồm 2 loại là:

- Load ballancer policy: Policy này giúp các instance tạo ra bởi auto-scaling tự động được thêm vào pool của Load ballancer.
- Scaling policy: Policy này sẽ định nghĩa các tính chất gồm: Scaling out (Tăng thêm) , scaling in (giảm xuống), mỗi lần tăng / giảm bao nhiêu instan ce, cooldown giữa mỗi lần scale

Một policy có thể gắn vào nhiều auto-scaling group khác nhau (trừ load balancer policy), giúp bạn tiết kiệm thời gian khi có nhiều auto scaling group. Mỗi a uto-scaling group có thể gắn nhiều hơn 1 policy, tuy nhiên không thể gắn cùng loại policy (ví dụ: mỗi group chỉ được có 1 scaling out hoặc 1 scaling in policy).