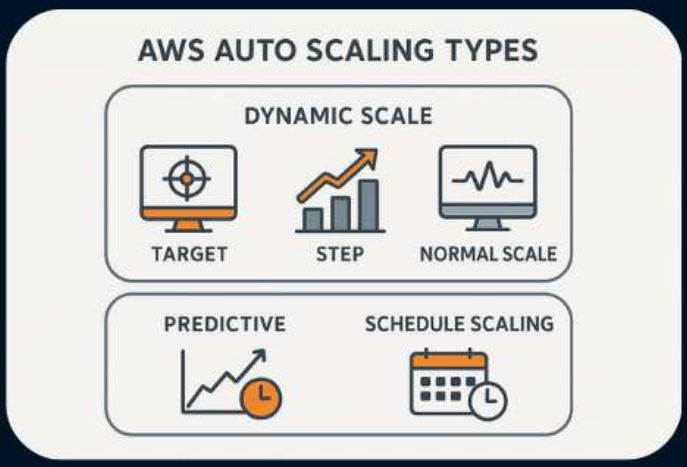
AWS Auto Scaling Types

1. Dynamic Scaling - AWS adjusts capacity automatically based on metrics (like CPU, requests).



1a) Target Tracking Scaling

Scenario: You want your EC2 Auto Scaling Group to always keep average CPU at 50%.
If CPU goes above 50% → scale out.
If CPU goes below 50% → scale in.
Use case: Best for steady applications like web servers.

1b) Step Scaling

Scenario: If CPU > 70% → add 2 instances, and if CPU > 90% → add 5 instances.
Scaling happens in steps based on thresholds.
Use case: For workloads where you need finer-grained control over scaling.

1c) Simple Scaling

Scenario: If CPU > 70%, add 1 instance. If CPU < 40%, remove 1 instance.
A straightforward if condition → then action.
Use case: When you need basic scaling rules without complexity.

2. Predictive Scaling - Uses machine learning to forecast future traffic & scale ahead of time.
Scenario: An e-commerce app sees traffic rise every day at 9 AM.
Predictive scaling analyzes past patterns and launches extra instances at 8:50 AM before traffic spikes.
Use case: Perfect for repetitive traffic patterns (Umbrella sales in rainy days, office hours).

3. Scheduled Scaling - Scales based on a defined schedule (like cron jobs).
Scenario: Your company runs payroll only on the 1st of every month, so traffic spikes that day.
You schedule scaling to add instances on 1st at 9 AM and scale back on 2nd at midnight.
Use case: Ideal for known fixed schedules (end of month, daily batch jobs).