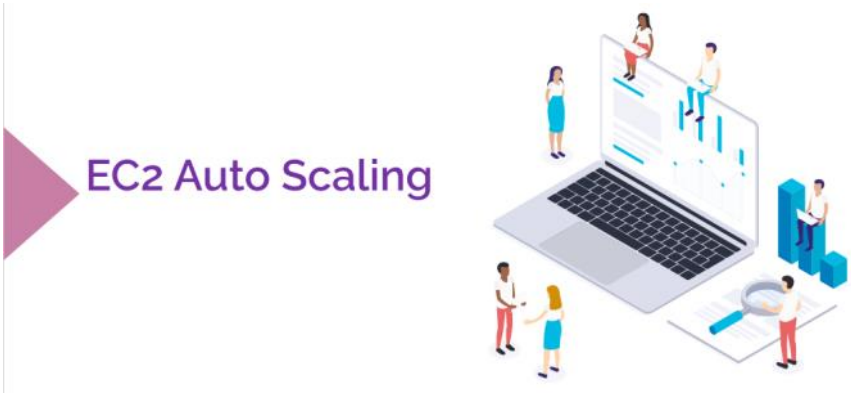
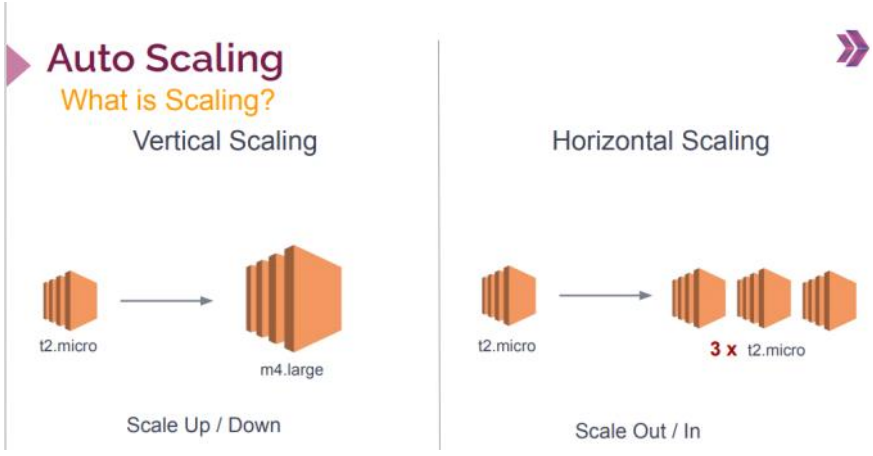
AUTO SCALING

Monday, August 5, 2024 11:29 AM

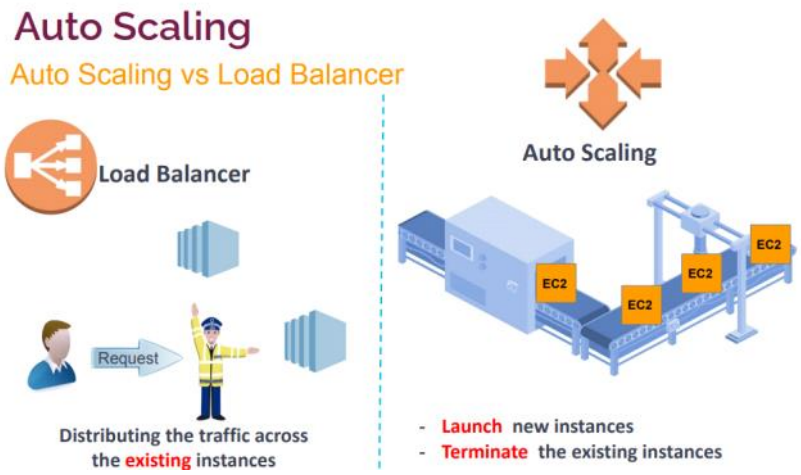




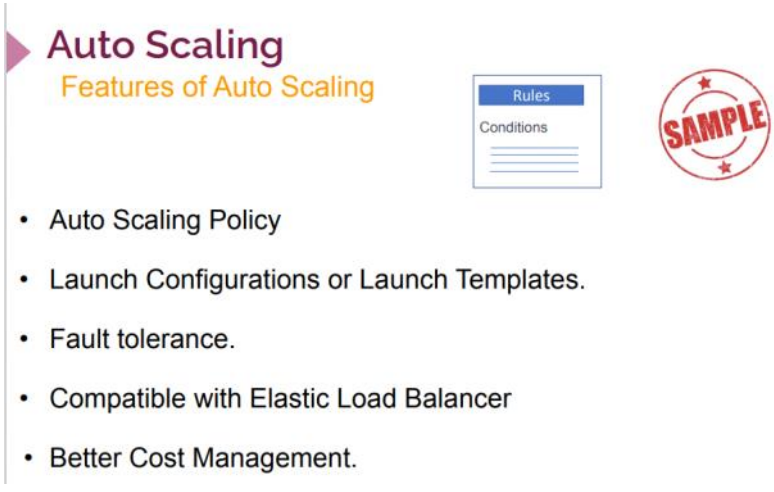
Scaling konusu bizim heryerde karsimiza cikaabilecek bir konu. Sadece AWS e has birsey degil. O yuzden bu konuyu iyi anlamak lazim.
2 cesit olceklendirme vardir: Vertical ve Horizontal scaling
Vertical scalingte bir bina var mesela uzerine kat cikmak gibi. Dikey bir buyume var. T2.micro az geldi mesela diyelimki o zaman m4.large ile buyutuyoruz. Kapladigi alan ayni ama ozelligi farkli mesela memoryi arttiriyoruz ya da cpusunu arttiriyoruz gibi. Literaturde buna scale up veya scale down diyoruz. Burada quality/nitelik arttiriyoruz.
Horizontalda ise bir instance yetmediyese mesela onun kapasitesini arttirmak yerine sayisini arttiriyoruz. 1 instance degilde mesela 3 instance run ediyoruz gibi. Burada nicelik arttiriyoruz. Bunu da literaturde scale out, scale in diye kullaniyoruz.
Aws cogunlukla horizontal scaling yapiyor ama launch template ettigimde hem horizontal hem vertical scaling yapabiliyorum.

- Amazon EC2 Auto Scaling is a component that helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application.
- Auto Scaling adds or removes instances to keep your system steady state.
- You can automate the increasing or decreasing of virtual machines depending on your policy.

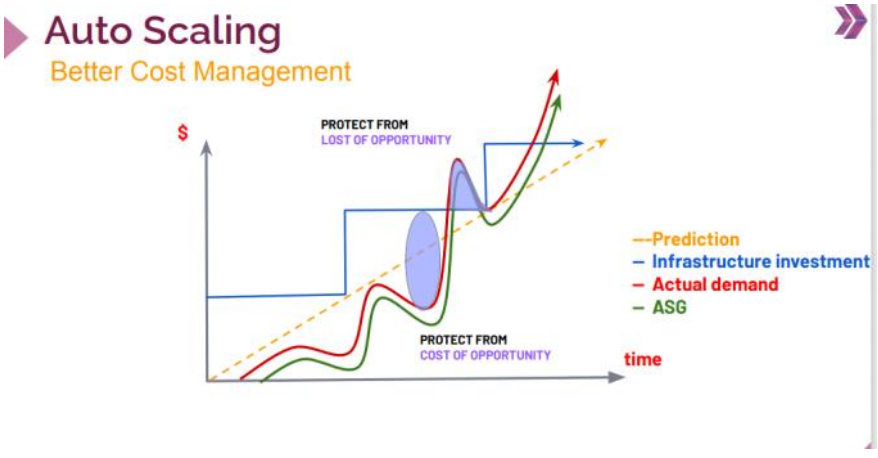
Auto scaling ise bizim bir yukumuz var diyelim. Bu yuke yetecek kadar isin icine instance katmamiz demek scaling yapmak demek.



Load balancer yoktan var etme gucu yok. Elinde kac instance varsa onu yonetir. Mesela 3 instanceta birisi bozuldu diyelim diger ikisine gonderir. Ikisi bozulduysa yuku kalan instancea gonderir ama o da bozulursa shuttle down yapar ve tum yuku hepsine gonderir hangisi neyi alirsa artik mantigiyla.
Ama auto scalingte ihtiyac hasil oldu diyelim hemen verilen algoritmalar dahilinde instance ayaga kaldirir ya da instance azaltir.
Auto scaling target grouplara ihtiyac olan instancelari sagliyor. Load balancerda bu instancelara dengeli yuk dagilimlarini yapiyor.
Auo scaling uretme ve terminate etmek yetkisi var.
Load balancer ise bu auto scalingin urettigi veya terminate ettigi instancelar arasinda yuk dagilimini sagliyor.



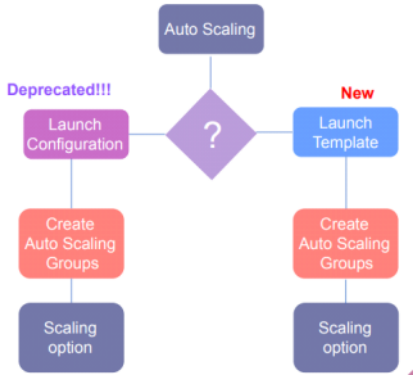
Bu yonetme sekli policylerle oluyor. Bizim belirledigimiz policylere gore yani bizim belirledigimiz kriterlere gore auto scaling uretime geciyor belirli zamanlarda ya da terminate ediyor instancelari kullanim azligina gore. Ama bunlarin hepsini policylerde biz belirliyoruz cunku gorunmez bi makine var ortada onu yonlendirmek icin bir kurallar butununе ihtiyacimiz var.
Auto scalingin bir instance ayaga kaldirabilmesi icin launch configuration ve launch tepmlate ihtiyaci var ki hangi kistaslara gore bir instance kaldiracagini bilsin-
Her zaman instance dengesini tutturamayabiliyoruz yani bazen cok bazen eksik instance ayaga kaldirtabiliyoruz auto scalingde. Bunun dengesiz oldugu durumlarda zararlari olabilir.
ELB ile isbirlikci calisir.



Bu bizim tahmini yatirim ve zaman cizelgemiz. Sari olan cizgi bizim tahminimiz.
Mavi olan kisimda dikeye gectigi zamanlar instance ihtiyacinin oldugu zamanlar. Yukari dogru gittigi vakit yeni instancelar eklemis demek. Ongorulebilir belirli araliklarla infrastructurea yatirim yapmis oluyoruz.
Ama bazen bizim yaotigimiz yatirimla actual demand yani gercek ihtiyaclar uyusmaya biliyo kirmizi cizgideki gibi.
Sari bizim tahmini egri. Mavi bizim yatiririmimiz. Kirmizi ise gerceklesen egri.
Yesil olan bizim auto scaling groupumuz. Bizim mavi cizgimizden ziyade kendi iyaptigi ongoruye gore gercege en yakin olan instance araligini belirleyip instance uretiyor ya da terminate ediyor. Yani yeterli yatirim yapamama ya da fazla gereksiz yatirim yapma olayini cozmus oluyor talep etdilme veya edilmeme durumu dahilinde ve policy ve kriterlere gore.

Auto Scaling
Auto Scaling Creating Process

- First, you need to select either the Launch Template or the Launch Configuration option and create it.
- Then, create an Auto Scaling Group.
- Finally, Finish Creating Auto Scaling



CLARUSWAY

Auto scaling yaratmanin iki yolu var: Launch configuration (deprecated olmus bu) ve Launch template

Bir launch template yaratiyoruz instance yaratacagi zaman neye gore yaratacagini bilsin diye. Daha spnra bir auto scaling grubu olusturuyoruz ve bittikten sonra scaling option ekliyoruz.

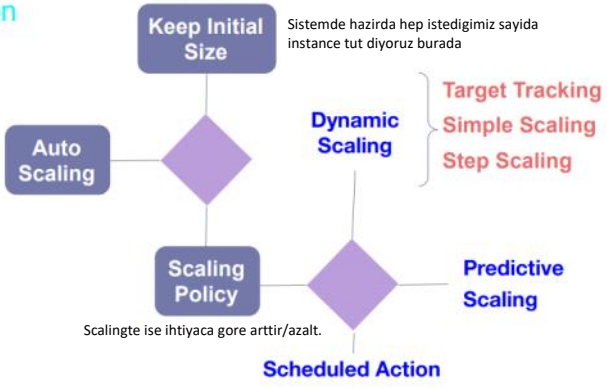
Auto Scaling
Launch Configuration Deprecation Schedule

The announcement from AWS

- March 31, 2023 - New accounts created after this date will not be able to create new launch configurations via the console. API and CLI access will remain available to support customers with automation use cases.
- December 31, 2023 - New accounts created after this date will not be able to create new launch configurations.

Launch configuration artik tedavulden kalkmis.

Auto Scaling
Auto Scaling Options
Next session

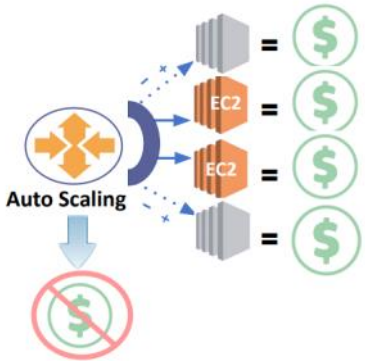


CLARUSWAY
WAY TO REINVENT YOURSELF

Auto scalingi iki sekilde kullanacagiz yani iki turunu de kullanacagiz: keep initial size ve scaling policy.
Mesela biz auto scalingten her zaman 3 saglikli instance olusturmasini isteriz. Cunku birisinin basina bir is gelmesi dahilinde load balancer sadece iki instance arasinda yuk dagitir yuk ayni olmasina ragmen. O yuzden auto scalingin her zaman hazirda uc instance tutmasini istiyoruz diyelim ki. O hemen bir saglikli instance daha uretir 3 istedigimiz icin biz hep en az.
Scaling policyde ise sunu diyoruz:trafik arttiginda arttir azaldiginda instancelari azalt.
Burada da 3 sekilde scaling yapiliyor: scheduled, predictive, dynamic.
Scheduledta zaman belli. Mesela her gun saat 03.00 da benim ekstra instance ihtiyacim va sen arttir. Ve aksam 5te de kaldir instancelari ihtiyac yok diyorum mesela. Benim ayarladigim saate gore ben ongoruyorum ve bildiriyorum.
Predictivede ise kendi el atiyo mevzuya. Diyoki mben sistemi en az bi 48 saat inceleyeyim, bakayim bir hangi saatlerde ihtiyac fazla ya da az. Ona gore kendi kendine o araliklarda ec2 arttirip azaltmaya basliyor. Bir analiz var burada.
Dynamicte ise anlik olarak responce var. Anlik olarak ihtiyac artti mesela aninda ec2 kaldiriyor ya da ihtiyac bitti diyelim hemen terminate ediyor. Burada bir netlik yok. Urunun ne kadar satilacagi bilinmiyordur mesela bu kullanilir. Burada surekli bir gozlem var. Hep dinamik. Burada da 3 mesele var:
Target trackingte mesela diyoruzki bir pcpu belirliyoruz mesela. Bunu hep %40in altinda tut diyoruz. O da ona gore anlik reaksiyon gosteriyor.yani %30-35 oldugunda aws harekete geciyo hemen aktif bir sekilde gozlem yapiyor ve reaksiyon gosteriyor.
Simpe scalingte, eger cpu %40i gecer ise harekete gec diyoruz yani %40 olana kadar hareket etme bi dur diyoruz. Targetta gecmeden once reaksiyon basliyodu. Burada ise gectikten sonra basliyor.
Stepte ise simplein bir ust hali. Yanimesela diyoruz ki %60-80 arasinda 1 instance ayaga kaldir ama %80 gecerse 3 instance ya da %90 gecerse ne varsa kullan diyoruz mesela bu sekilde steplendiriyoruz.
SINAVLARDA SIMPLE VE TARGETI AYNI ANDA VERMEZLER AMA AWS TARGET TAVSIYE EDER. ORAN VS VARSA TARGETTIR
SPESIFIK BIR SAAT ISE SCHEDULED SECIYORUZ.
ORAN VERIRSE TARGET, SAAT VERIRSE SCHEDULED

Auto Scaling

Pricing for Amazon EC2 Auto Scaling



Load Balancer

CLARUSWAY
WAY TO REINVENT YOURSELF

Auto scaling ucretsiz. Arkada bunu yapan makineyi biz gormuyoruz.
Ama urettigi nihai ciktilardan ucret aliyor.
Yani auto scaling gece ucte arttirip azaltmaktan degil, kullanidigi kadar instance parasi aliyor.
Load balancer ve cloudwatch icinde para aliyor belli bir miktardan sonra.
Cloudwatch auto scalingin onemli bir parcasi. Buradan alarm kuruyor arttirmasi gerektigi zamanlari anlamasi icin.