PROJEKTOWANIE SYSTEMÓW OBIEKTOWYCH I ROZPROSZONYCH

LABORATORIUM 3

Auto-scaling, Elastic Load Balancer (ELB)

wersja1.0

przygotował:
Radosław Adamus
Instytut Informatyki Stosowanej

.

HISTORIA WERSJI

DATA	WERSJA	AUTOR	OPIS
17.02.2014	0.1	Radosław Adamus	Pierwsza robocza wersja dokumentu
24.03.2014	1.0	Radosław Adamus	Pierwsza oficjalna wersja instrukcji

Cel:

Celem laboratorium jest:

1. Zapoznanie się z mechanizmami skalowania oraz równoważenia obciążenia dla aplikacji webowych uruchamianych w usłudze AWS.

Efekty:

Po ukończeniu laboratorium student potrafi:

1. Definiować i wykonywać podstawową konfigurację usługi równoważenia obciążenia
2. Definiować i konfigurować najistotniejsze ustawienia usługi auto-skalowania.
3. Modyfikować ustawienia za pośrednictwem *AWS Management Console* oraz programowo z wykorzystaniem AWS-API.

Wymagania wstępne:

1. Posiadanie konta na platformie Github.
2. Skonfigurowane konto AWS

Narzędzia:

1. AWS Management Console
2. Platforma nodeJS oraz edytor.

Reguły wykonywania ćwiczeń laboratoryjnych:

1. Po ukończeniu laboratorium należy wyłączyć wszystkie działające instancje EC2, oraz wyzerować ustawienia dotyczące docelowej, minimalnej i maksymalnej liczby instancji w usłudze ASG.
2. Otrzymane dane autoryzacyjne (hasła oraz klucze dostępu) są danymi wrażliwymi i muszą być chronione. W szczególności nie można dodawać do repozytorium kontroli wersji oraz pozwolić na wysłanie na usługi hostujące repozytoria kontroli wersji kodu źródłowego (GitHub) plików zawierających konfigurację autoryzacji dostępu do API AWS.

Wstęp:

0 Stosowane nazewnictwo - skróty

W ramach instrukcji wykorzystywane są następujące skróty:
ASG - Auto Scaling Group - usługa pozwalająca na tworzenie grup automatycznie skalujących się poziomo instancji EC2.
ELB - Elastic Load Balancer - usługa pozwalająca na równoważenie obciążenia instancji EC2.

1. Automatyczne skalowanie aplikacji webowej - usługa Auto-scaling

Usługa Auto Scaling pozwala na automatyczne uruchamianie i zatrzymywanie instancji EC2 na podstawie predefiniowanych ustawień dotyczących ich bieżącego statusu obciążenia oraz w oparciu o harmonogramowanie. Skalowanie ma istotne znaczenie w zakresie kontroli kosztów oraz zarządzania zasobami aplikacji uruchomionych w chmurze.
Dzięki usłudze Auto Scaling'u uruchomiona aplikacja może automatycznie uzyskać dodatkowe zasoby w momencie zwiększonego zapotrzebowania oraz zwolnić zasoby, gdy nie są one już potrzebne.

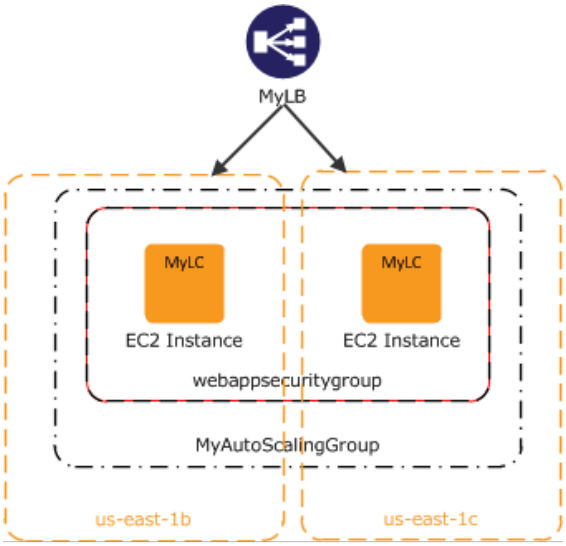
Konfiguracja:

Usługa Auto Scaling wymaga utworzenia tzw. Auto Scaling Group (ASG). Grupa reprezentuje zbiór instancji EC2, w ramach których usługa będzie dokonywała automatycznego zwiększania lub zwalniania zasobów w zależności od potrzeb i przy założeniu minimalizacji kosztów. Grupa taka stanowi jednostkę z punktu widzenia zarządzania i reguł uruchamiania. Z grupą musi być powiązana konfiguracja uruchomieniowa (ang. launch configuration), która poza standardowymi ustawieniami związanymi z instancjami EC2 zawiera również ustawienia specyficzne dla grupy. W celu uzyskania wysokiej dostępności grupy możliwe jest również aby instancje w jej ramach były uruchamiane w różnych strefach dostepności (Avaliabilty Zones).

2. Równoważenie obciążenia - usługa Elastic Load Balancing (ELB)

Równoważenie obciążenia to technika pozwalająca na optymalizację wykorzystania infrastruktury poprzez rozpraszanie obciążenia pomiędzy dostępne zasoby (węzły sieci, procesory, dyski). Najczęściej wykorzystuje się ją w przypadku pojedynczej usługi uruchomionej na wielu serwerach (np. usługi WWW, DNS).

Chmura Amazon posiada usługę Elastic Load Balancing (ELB), która automatycznie rozpraszać ruch pomiędzy wiele instancji EC2. ELB pozwala również na zwiększenie tolerancji na awarię przez wykrywanie problemów z instancjami i automatyczne przekierowywanie ruchu do działających instancji.



Rys 1. Wysokodostępna konfiguracja Auto Scaling Group wykorzystująca Elastic Load Balancer.

Opis laboratorium:

1. Uruchomienie ASG

Zadanie polega na utworzeniu i skonfigurowaniu ASG dostępnej za pośrednictwem usługi ELB. Po utworzeniu należy poeksperymentować z ustawieniami polityki skalowania (ang. Scaling policy) w taki sposób aby wymusić, poprzez wzrost obciążenia, zwiększenie, uruchomionych w ramach ASG, liczby instancji EC2.

Konfiguracja usług ELB i ASG:

0. W celu uniknięcia pomyłek definiowane usługi należy nazywać stosując konwencję: [Nazwisko][TypUsługi], np. *AdamusASG*.

1. Aby automatycznie skalowane instancje mogły być dostępne za pośrednictwem usługi ELB musi zostać ona wstępnie utworzona i skonfigurowana (większość z tych ustawień jest również edytowalna po zdefiniowaniu usługi). Usługa równoważenia obciążenia powinna, na wstępie nie

mieć przypisanych żadnych instancji oraz pozwalać na przyjmowanie ruchu na porcie 80 i przekierowywanie go na docelowy port 8080.

2. Utworzenie ASG wymaga wstępnego zdefiniowania konfiguracji uruchomieniowej (Launch Configuration), który stanowi szablon wykorzystywany przy uruchamianiu instancji EC2 w ramach grupy. Konfiguracja taka powinna być dostępna na koncie AWS, dostępnego dla studentów. Została on zdefiniowana w oparciu o obraz instancji EC2 z preinstalowaną aplikacją wyliczającą hashe, poznaną w ramach poprzedniego laboratorium. Instancja skonfigurowana jest w taki sposób, aby aplikacja uruchamiała się automatycznie w trakcie startu systemu.
W trakcie konfiguracji ASG należy:

a. wybrać opcję “Receive traffic from Elastic Load Balancer(s)” (w ramach Advanced details) i wybrać zdefiniowaną uprzednio usługę ELB.

b. Ustalić, że wymagana i minimalna liczba instancji jest równa 1 a maksymalna 2.

c. Wybrać domyślną podsieć dla Avaliability Zone: us-west-2c

d. pozostawić ustawienia dotyczące polityki skalowania oraz notyfikacji bez zmian.

Eksperymentowanie z ASG:

Po utworzeniu ASG, automatycznym uruchomieniu pierwszej instancji w jej obrębie należy sprawdzić czy dostęp do niej jest możliwy za pośrednictwem adresu DNS przypisanego usłudze ELB. Następnie należy:

1. Poprzez zdefiniowanie odpowiedniej polityki skalowania (oraz wymuszenie wzrostu obciążenia) należy wymusić zwiększenie liczby instancji w ramach grupy.
2. Poprzez zdefiniowanie odpowiedniej polityki skalowania należy wymusić zmniejszenie się liczby instancji do poziomu pierwotnego. Należy również zaobserwować jaka akcja wykonywana jest na instancji usuwanej z grupy oraz znaleźć miejsce definiowania polityki usuwania instancji z ASG.

Każde zdarzenie skalowania powinno wysyłać informację na adres email studenta oraz prowadzącego zajęcia. Ustawienia te definiuje się w postaci tzw. notyfikacji przypisanych następnie do polityki skalowania.

2. Programowe zwiększanie liczby instancji EC2 w ramach ASG

1. Zaimplementuj, rozwijając przykładową aplikację poznaną na poprzednim laboratorium, możliwość programowego zwiększania liczby instancji w ramach ASG z wykorzystaniem parametru sterującego wymaganą/oczekiwaną liczbą aktywnych instancji (ang. desired). W celu

realizacji zadania należy wykorzystać funkcję AWS - API o nazwie *setDesiredCapacity*.
http://docs.aws.amazon.com/AWSJavaScriptSDK/latest/AWS/AutoScaling.html#setDesiredCapacity-property

3. Wysoka dostępność (opcjonalnie)

1. Zlokalizuj opcje pozwalające na tworzenie instancji w ramach ASG w różnych stref dostępności (Avaliability Zone). Zmień ustawienia ASG, tak aby instancje mogły pochodzić z różnych stref. Wymuś skalowanie grupy i sprawdź efekt.

Materiały:

1. Co to jest Auto Scaling:
http://docs.aws.amazon.com/AutoScaling/latest/DeveloperGuide/WhatIsAutoScaling.html
2. Konfiguracja ELB i ASG:
http://docs.aws.amazon.com/AutoScaling/latest/DeveloperGuide/US_SetUpASLBApp.html