Odporne architektury

Domena 2: strefy, równoważniki, kolejki, bazy, kopie i DR.

- Odporne architektury, część 1: strefy, ELB, Auto Scaling, Route 53 i luźne powiązania
- Odporne architektury, część 2: RDS Multi-AZ, Aurora, AWS Backup i strategie DR

Odporne architektury, część 1: strefy, ELB, Auto Scaling, Route 53 i luźne powiązania

Poziom: średnio

Dotyczy: SAA-C03 domena 2, zadania 2.1 (skalowalne i luźno powiązane architektury) i 2.2 (wysoka dostępność i odporność na awarie), Elastic Load Balancing, EC2 Auto Scaling, Amazon Route 53, Amazon SQS, Amazon SNS, Amazon EventBridge, AWS Step Functions, Amazon API Gateway, dni 10 do 13 planu

Czas: 4 dni po 1,5 do 2 godzin

Domena odpornych architektur to 26 procent egzaminu. Ta strona opisuje warstwę aplikacji: jak rozłożyć ją na strefy, jak rozdzielać ruch równoważnikiem, jak dokładać i zdejmować instancje, jak kierować użytkowników przez DNS i jak rozdzielić części systemu kolejką, żeby awaria jednej nie zatrzymała reszty. Dane i kopie opisuje część druga.

ZAŁOŻENIA SCENARIUSZA

ćwiczenia robimy w VPC laboratorium 10.60.0.0/16 z podsieciami `pub` i `app` w trzech strefach, z bramą NAT włączoną na czas ćwiczeń (`WlaczNat=true`, patrz Plan 28 dni i laboratorium AWS w Zielonej). Strefa prywatna Route 53 `lab.aws.zielona.example` istnieje tylko w VPC laboratorium. Strona aplikacji to prosty serwer WWW na Amazon Linux 2023, który wypisuje nazwę strefy.

Skróty użyte na tej stronie (29)

Skrót	Co znaczy
ALB	Application Load Balancer: równoważnik AWS warstwy 7, kieruje żądania HTTP i HTTPS według ścieżki i nazwy hosta
API	Application Programming Interface: sposób, w jaki programy rozmawiają ze sobą
AWS	Amazon Web Services: chmura Amazonu
CLI	Command Line Interface: praca przez wpisywanie komend w terminalu

Skrót	Co znaczy
CNAME	Canonical Name: rekord DNS będący aliasem, drugą nazwą wskazującą na już istniejący rekord A
DNS	Domain Name System: książka telefoniczna internetu, zamienia nazwy na adresy IP
DynamoDB	baza danych typu klucz-wartość w chmurze AWS; Lore trzyma w niej fragmenty, stan, magazyn zmienny i blokady, gdy działa na AWS
EC2	Elastic Compute Cloud: usługa AWS wynajmująca gotowe maszyny wirtualne na godziny
EFS	Elastic File System: współdzielony system plików NFS w AWS, dostępny z wielu instancji i stref
FIFO	First In, First Out: kolejność obsługi, w której komunikat, który przyszedł pierwszy, pierwszy też wychodzi; tak działa kolejka SQS FIFO
GWLB	Gateway Load Balancer: równoważnik AWS, który przepuszcza ruch przez urządzenia zabezpieczające innych firm
HTTP	Hypertext Transfer Protocol: protokół stron WWW
IP	Internet Protocol: adres komputera w sieci; w elektryce: stopień ochrony obudowy (Ingress Protection)
JSON	JavaScript Object Notation: tekstowy format danych z klamrami i cudzysłowami
NAT	Network Address Translation: podmiana adresów IP na granicy sieci
NLB	Network Load Balancer: równoważnik AWS warstwy 4 dla TCP i UDP, ze stałym adresem IP w strefie
RPO	Recovery Point Objective: ile danych wolno stracić przy awarii
RTO	Recovery Time Objective: jak długo wolno nie działać przy awarii
S3	Simple Storage Service: magazyn plików w chmurze AWS
SAA-C03	kod wersji egzaminu AWS Solutions Architect Associate zdawanej od 30 sierpnia 2022, aktualnej w październiku 2026
SNS	Simple Notification Service: usługa AWS, która rozsyła jeden komunikat do wielu odbiorców naraz (kolejki, e-mail, SMS)
SQS	Simple Queue Service: kolejka komunikatów AWS, która rozdziela nadawcę i odbiorcę, żeby jeden nie czekał na drugiego
SSM	Systems Manager: usługa AWS do zarządzania instancjami (Session Manager, Parameter Store); skrót od dawnej nazwy Simple Systems Manager
TLS	Transport Layer Security: szyfrowanie połączeń, np. w HTTPS
VPC	Virtual Private Cloud: prywatna sieć wirtualna klienta w chmurze AWS, w której stawiamy maszyny i zakładamy tablice tras
VPN	Virtual Private Network: szyfrowany tunel do sieci firmowej
WAF	Web Application Firewall: zapora aplikacji WWW, która filtruje żądania HTTP, na przykład wstrzykiwanie SQL i zalew żądań

Skrót	Co znaczy
WMS	Warehouse Management System: system zarządzania magazynem
WWW	World Wide Web: strony i aplikacje otwierane w przeglądarce po HTTP albo HTTPS

Fiszka

Komendy wykonujemy w CloudShell z profilem konta laboratorium. `IdVpc`, `IdPodsieci`, `IdGrupy`, `ArnRownowaznika`, `ArnGrupyDocelowej`, `ArnCertyfikatu`, `IdStrefyHostowanej` i podobne zastępujemy wartościami z poprzednich poleceń.

◫◫ ZMIENIA SYSTEM

Utworzenie równoważnika aplikacji w trzech strefach: aws elbv2 create-load-balancer

```
aws elbv2 create-load-balancer --name alb-lab --type application --scheme internet-facing --subnets IdPodsieciPubA IdPodsieciPubB IdPodsieciPubC --security-groups IdGrupyAlb
```

ALB potrzebuje podsieci w co najmniej dwóch strefach. Każda podana podsieć dostaje węzeł równoważnika, a nazwa DNS równoważnika wskazuje na wszystkie.

◫◫ **ALB** = Application Load Balancer: równoważnik AWS warstwy 7, kieruje żądania HTTP i HTTPS według ścieżki i nazwy hosta · **DNS** = Domain Name System: książka telefoniczna internetu, zamienia nazwy na adresy IP · **aws** = AWS CLI: narzędzie chmury Amazonu

◫◫ ZMIENIA SYSTEM

Utworzenie grupy docelowej z kontrolą zdrowia: aws elbv2 create-target-group

```
aws elbv2 create-target-group --name tg-lab-www --protocol HTTP --port 80 --vpc-id IdVpc --target-type instance --health-check-path /zdrowie --healthy-threshold-count 2 --unhealthy-threshold-count 2
```

Równoważnik wysyła ruch tylko do celów, które odpowiadają na ścieżkę kontroli zdrowia. Ścieżka `/zdrowie` powinna sprawdzać aplikację, a nie tylko serwer WWW.

WWW = World Wide Web: strony i aplikacje otwierane w przeglądarce po HTTP albo HTTPS · **aws** = AWS CLI: narzędzie chmury Amazonu · **HTTP** = Hypertext Transfer Protocol: protokół stron WWW

ZMIENIA SYSTEM

Utworzenie nasłuchu HTTPS z certyfikatem ACM: aws elbv2 create-listener

```
aws elbv2 create-listener --load-balancer-arn ArnRownowaznika --protocol HTTPS --port 443 --certificates CertificateArn=ArnCertyfikatu --default-actions Type=forward,TargetGroupArn=ArnGrupyDocelowej
```

W laboratorium bez domeny zakładamy nasłuch HTTP na porcie 80 (`--protocol HTTP --port 80` i bez `--certificates`). TLS kończy się na równoważniku, a do instancji idzie HTTP w VPC, chyba że wymagania mówią o szyfrowaniu do końca.

HTTPS = HTTP zabezpieczony szyfrowaniem TLS · **ACM** = AWS Certificate Manager: usługa wydająca i odnawiająca certyfikaty TLS dla równoważników, CloudFront i API Gateway · **HTTP** = Hypertext Transfer Protocol: protokół stron WWW · **TLS** = Transport Layer Security: szyfrowanie połączeń, np. w HTTPS · **VPC** = Virtual Private Cloud: prywatna sieć wirtualna klienta w chmurze AWS, w której stawiamy maszyny i zakładamy tablice tras · **aws** = AWS CLI: narzędzie chmury Amazonu

TYLKO ODCZYT

Sprawdzenie zdrowia celów: aws elbv2 describe-target-health

```
aws elbv2 describe-target-health --target-group-arn ArnGrupyDocelowej --query "TargetHealthDescriptions[].[Target.Id,TargetHealth.State]" --output table
```

Stan `healthy` jest dobry, `unhealthy` z powodem `Target.ResponseCodeMismatch` albo `Target.Timeout` mówi, czy aplikacja odpowiada źle, czy wcale.

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Utworzenie szablonu uruchamiania: aws ec2 create-launch-template

```
aws ec2 create-launch-template --launch-template-name lt-lab-www --launch-template-data file://szablon-www.json
```

Szablon opisuje obraz, typ instancji, grupę zabezpieczeń, profil z rolą i skrypt startowy. Grupa Auto Scaling bierze z niego przepis na każdą nową instancję.

aws = AWS CLI: narzędzie chmury Amazonu

PLIK KONFIGURACYJNY

Dane szablonu uruchamiania: szablon-www.json

```
{
  "ImageId": "resolve:ssm:/aws/service/ami-amazon-linux-latest/al2023-ami-kernel-default-x86_64",
  "InstanceType": "t3.micro",
  "SecurityGroupIds": ["IdGrupyAplikacji"],
  "IamInstanceProfile": { "Name": "NazwaProfiluInstancji" },
  "MetadataOptions": { "HttpTokens": "required" },
  "UserData": "TrescSkryptuWBase64"
}
```

`resolve:ssm:` pobiera najnowszy obraz Amazon Linux 2023 z publicznego parametru AWS przy każdym uruchomieniu. `HttpTokens: required` wymusza IMDSv2, czyli metadane instancji tylko z tokenem sesji. `TrescSkryptuWBase64` to wynik `base64 -w0 www.sh` dla skryptu z ćwiczenia 1.

AWS = Amazon Web Services: chmura Amazonu

ZMIENIA SYSTEM

Utworzenie grupy Auto Scaling w trzech strefach: aws autoscaling create-auto-scaling-group

```
aws autoscaling create-auto-scaling-group --auto-scaling-group-name asg-lab-www --launch-template "LaunchTemplateName=lt-lab-www,Version=\$Latest" --min-size 2 --max-size 6 --desired-capacity 3 --vpc-zone-identifier "IdPodsieciAppA,IdPodsieciAppB,IdPodsieciAppC" --target-group-arns ArnGrupyDocelowej --health-check-type ELB --health-check-grace-period 120
```

Grupa rozkłada instancje równo między strefy i po awarii strefy uruchamia brakujące w pozostałych. `--health-check-type ELB` sprawia, że instancję wymienia także wtedy, gdy działa, ale nie przechodzi kontroli zdrowia równoważnika.

ELB = Elastic Load Balancing: rodzina równoważników obciążenia AWS (ALB, NLB, GWLB) · **aws** = AWS CLI: narzędzie chmury Amazonu

TYLKO ODCZYT

Polityka skalowania według średniego obciążenia procesora: aws autoscaling put-scaling-policy

```
aws autoscaling put-scaling-policy --auto-scaling-group-name asg-lab-www --policy-name cpu-50 --policy-type TargetTrackingScaling --target-tracking-configuration file://cel-cpu.json
```

aws = AWS CLI: narzędzie chmury Amazonu

PLIK KONFIGURACYJNY

Plik celu skalowania: cel-cpu.json

```
{
  "PredefinedMetricSpecification": { "PredefinedMetricType": "ASGAverageCPUUtilization"
},
  "TargetValue": 50.0
}
```

Śledzenie celu (target tracking) samo zakłada alarmy CloudWatch i dokłada albo zdejmuje instancje, żeby średnie obciążenie wracało do 50 procent. Inne gotowe metryki to między innymi liczba żądań na cel równoważnika (`ALBRequestCountPerTarget`).

TYLKO ODCZYT

Historia działań grupy: aws autoscaling describe-scaling-activities

```
aws autoscaling describe-scaling-activities --auto-scaling-group-name asg-lab-www --max-items 5 --query "Activities[].[StartTime,StatusCode,Description]" --output table
```

Pierwsze miejsce, do którego patrzymy, gdy grupa nie uruchamia instancji albo wymienia je w kółko.

aws = AWS CLI: narzędzie chmury Amazonu

◫◫ ZMIENIA SYSTEM

Utworzenie prywatnej strefy DNS dla VPC: aws route53 create-hosted-zone

```
aws route53 create-hosted-zone --name lab.aws.zielona.example --vpc VPCRegion=eu-central-1,VPCId=IdVpc --hosted-zone-config Comment=laboratorium,PrivateZone=true --caller-reference "lab-$(date +%s)"
```

Strefa prywatna odpowiada tylko zapytaniom z przypiętych VPC. `--caller-reference` musi być unikatowy przy każdym wywołaniu, stąd znacznik czasu.

◫◫ **DNS** = Domain Name System: książka telefoniczna internetu, zamienia nazwy na adresy IP · **VPC** = Virtual Private Cloud: prywatna sieć wirtualna klienta w chmurze AWS, w której stawiamy maszyny i zakładamy tablice tras · **aws** = AWS CLI: narzędzie chmury Amazonu

◫◫ ZMIENIA SYSTEM

Utworzenie kontroli zdrowia Route 53: aws route53 create-health-check

```
aws route53 create-health-check --caller-reference "hc-$(date +%s)" --health-check-config Type=HTTPS,FullyQualifiedDomainName=sklep.zielona.example,ResourcePath=/zdrowie,Port=443,RequestInterval=30,FailureThreshold=3
```

Kontrolerzy Route 53 sprawdzają adres z wielu miejsc w internecie, więc kontrola zdrowia działa tylko dla punktów publicznych. Dla zasobów prywatnych używamy kontroli opartej na alarmie CloudWatch.

◫◫ **aws** = AWS CLI: narzędzie chmury Amazonu · **HTTPS** = HTTP zabezpieczony szyfrowaniem TLS

◫◫ TYLKO ODCZYT

Zmiana rekordów DNS z pliku: aws route53 change-resource-record-sets

```
aws route53 change-resource-record-sets --hosted-zone-id IdStrefyHostowanej --change-batch file://failover.json
```

DNS = Domain Name System: książka telefoniczna internetu, zamienia nazwy na adresy IP ·**aws** = AWS CLI: narzędzie chmury Amazonu

PLIK KONFIGURACYJNY

Rekordy przełączenia awaryjnego: failover.json

```
{
  "Comment": "Sklep: Frankfurt glowny, strona zastepcza w S3 zapasowa",
  "Changes": [
    {
      "Action": "UPSERT",
      "ResourceRecordSet": {
        "Name": "sklep.zielona.example",
        "Type": "A",
        "SetIdentifier": "frankfurt",
        "Failover": "PRIMARY",
        "HealthCheckId": "IdKontroliZdrowia",
        "AliasTarget": { "HostedZoneId": "IdStrefyRownowaznika", "DNSName":
"NazwaDnsRownowaznika", "EvaluateTargetHealth": true }
      }
    },
    {
      "Action": "UPSERT",
      "ResourceRecordSet": {
        "Name": "sklep.zielona.example",
        "Type": "A",
        "SetIdentifier": "zastepcza",
        "Failover": "SECONDARY",
        "AliasTarget": { "HostedZoneId": "IdStrefyStronyS3", "DNSName":
"NazwaDnsStronyS3", "EvaluateTargetHealth": false }
      }
    }
  ]
}
```

Rekord alias wskazuje zasób AWS (ALB, CloudFront, stronę S3) bez adresu IP i działa także na wierzchołku domeny, gdzie zwykły CNAME jest niedozwolony. `IdStrefyRownowaznika` to identyfikator strefy samego równoważnika z pola `CanonicalHostedZoneId` w `describe-load-balancers`, a nie naszej strefy.

AWS = Amazon Web Services: chmura Amazonu · **ALB** = Application Load Balancer: równoważnik AWS warstwy 7, kieruje żądania HTTP i HTTPS według ścieżki i nazwy hosta · **S3** = Simple Storage Service: magazyn plików w chmurze AWS · **IP** = Internet Protocol: adres komputera w sieci; w elektryce: stopień ochrony obudowy (Ingress Protection) · **CNAME** = Canonical Name: rekord DNS będący aliasem, drugą nazwą wskazującą na już istniejący rekord A

ZMIENIA SYSTEM

Utworzenie kolejki z kolejką martwych komunikatów: aws sqs create-queue

```
aws sqs create-queue --queue-name lab-zamowienia-dlq
aws sqs create-queue --queue-name lab-zamowienia --attributes file://kolejka-zamowienia.json
```

aws = AWS CLI: narzędzie chmury Amazonu

PLIK KONFIGURACYJNY

Atrybuty kolejki zamówień: kolejka-zamowienia.json

```
{
  "VisibilityTimeout": "60",
  "MessageRetentionPeriod": "345600",
  "ReceiveMessageWaitTimeSeconds": "20",
  "RedrivePolicy": "{\"deadLetterTargetArn\":\"arn:aws:sqs:eu-central-1:IdKonta:lab-zamowienia-dlq\",\"maxReceiveCount\":\"5\"}"
}
```

Komunikat, którego odbiorca nie przetworzy pięć razy, trafia do kolejki `lab-zamowienia-dlq`, zamiast blokować resztę. `ReceiveMessageWaitTimeSeconds` włącza długie odpytywanie, które zmniejsza liczbę pustych odpowiedzi i koszt. Wartość `RedrivePolicy` jest napisem z JSON w środku, stąd ukośniki.

JSON = JavaScript Object Notation: tekstowy format danych z klamrami i cudzysłowami

ZMIENIA SYSTEM

Utworzenie tematu powiadomień z subskrypcją kolejki: aws sns create-topic

```
aws sns create-topic --name lab-nowe-zamowienie
aws sns subscribe --topic-arn ArnTematu --protocol sqs --notification-endpoint
ArnKolejki
```

SNS rozsyła jeden komunikat do wielu odbiorców naraz (fan-out): kolejki dla magazynu, dla faktur i e-mail dla Kasi. Kolejka musi mieć politykę, która pozwala tematowi wysyłać do niej komunikaty.

SNS = Simple Notification Service: usługa AWS, która rozsyła jeden komunikat do wielu odbiorców naraz (kolejki, e-mail, SMS) · **aws** = AWS CLI: narzędzie chmury Amazonu

Z życia wzięte

Z ŻYCIA WZIĘTE

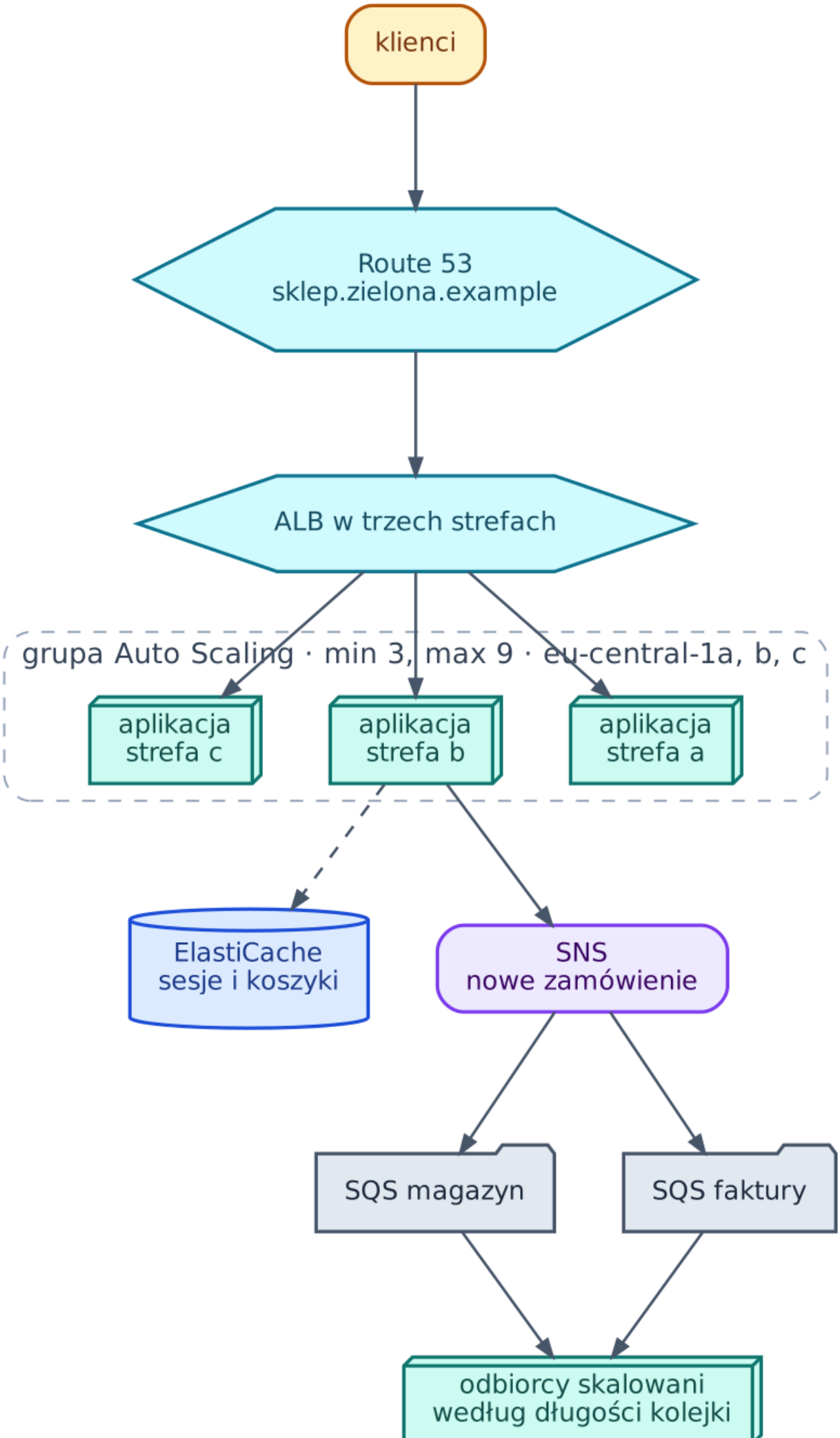
Piątek, 16:05. W laboratorium Tomek wyłącza na próbę wszystkie instancje w strefie `eu-central-1a`, a Ola odświeża stronę testową co sekundę. Przez kilkanaście sekund co trzecie odświeżenie kończy się błędem 502, potem równoważnik przestaje wysyłać ruch do martwych celów, a Auto Scaling uruchamia dwie nowe instancje w strefach b i c. Ola pyta, czy klienci sklepu w listopadzie zauważyliby te kilkanaście sekund, i Tomek zapisuje w projekcie niższy próg niezdrowych kontroli i krótszy interwał.

Jak to działa

Rozwiń: jak to działa pod spodem

Wiele stref zamiast jednego mocnego serwera

Architektura odporna na awarię strefy ma tę samą aplikację w co najmniej dwóch strefach, równoważnik przed nią i dane, które przeżywają utratę strefy. Instancje są bezstanowe: sesje trzymamy w ElastiCache albo DynamoDB, pliki w S3 albo EFS, a nie na dysku instancji, więc każdą można w każdej chwili wymienić.


klienci
Route 53
sklep.zielona.example
ALB w trzech strefach
grupa Auto Scaling · min 3, max 9 · eu-central-1a, b, c
aplikacja
strefa c
aplikacja
strefa b
aplikacja
strefa a
ElastiCache
sesje i koszyki
SNS
nowe zamówienie
SQS magazyn
SQS faktury
odbiorcy skalowani
według długości kolejki

Rysunek 1. Odporna warstwa aplikacji: Route 53, ALB w trzech strefach, grupa Auto Scaling, sesje w ElastiCache, zamówienia przez kolejkę SQS do odbiorców, którzy skalują się osobno (kliknięcie otwiera pełny rozmiar)

> **Pojęcie: skalowanie poziome i pionowe.** Poziome to dokładanie kolejnych instancji, pionowe to zamiana instancji na większą. Egzamin prawie zawsze nagradza poziome, bo nie wymaga przestoju i daje odporność.

> **Pojęcie: luźne powiązanie (loose coupling).** Części systemu nie wołają się bezpośrednio, tylko przez kolejkę, temat albo szynę zdarzeń. Gdy odbiorca jest wolny albo leży, nadawca dalej przyjmuje zamówienia, a komunikaty czekają w kolejce.

Elastic Load Balancing

Rodzaj	Warstwa	Kiedy jest odpowiedzią
Application Load Balancer (ALB)	7, HTTP i HTTPS	reguły po ścieżce i nazwie hosta, mikrousługi, kontenery, uwierzytelnianie przez Cognito, WAF
Network Load Balancer (NLB)	4, TCP, UDP, TLS	bardzo duży ruch, małe opóźnienie, stały adres IP w strefie (Elastic IP), protokoły inne niż HTTP, PrivateLink
Gateway Load Balancer (GWLB)	3, protokół GENEVE	przepuszczenie ruchu przez urządzenia zabezpieczające innych firm (zapory, IDS)

Classic Load Balancer to poprzednia generacja, której nie wybieramy w nowych projektach. Lepkie sesje (sticky sessions) na ALB to ratunek dla starej aplikacji trzymającej sesję w pamięci, a nie wzorzec.

EC2 Auto Scaling

Sposób skalowania	Kiedy
śledzenie celu (target tracking)	domyślny wybór: utrzymujemy metrykę na poziomie
krokowe (step)	różne reakcje na różne progi alarmu
zaplanowane (scheduled)	znany wzorzec, na przykład start zmiany w Swarzędzu o 6:00

Sposób skalowania	Kiedy
predykcyjne (predictive)	powtarzalny dzienny albo tygodniowy wzorzec, uczony na historii

Ciepła pula (warm pool) trzyma przygotowane, zatrzymane instancje, gdy start aplikacji trwa długo. Okres karencji kontroli zdrowia (`health-check-grace-period`) daje instancji czas na start, zanim zostanie oceniona.

Route 53 i polityki routingu

Polityka	Co robi	Przykład pytania
prosta (simple)	jeden rekord, bez kontroli zdrowia	jeden serwer WWW
ważona (weighted)	ruch w proporcjach	10 procent ruchu na nową wersję
opóźnienia (latency)	region z najmniejszym opóźnieniem do użytkownika	aplikacja w dwóch regionach dla klientów z Europy i Azji
przełączenie awaryjne (failover)	główny i zapasowy, według kontroli zdrowia	strona zastępcza w S3 przy awarii regionu
geolokalizacja (geolocation)	według kraju albo kontynentu użytkownika	treść zgodna z prawem danego kraju
bliskość geograficzna (geoproximity)	według odległości, z przesunięciem granic	przeniesienie części ruchu do drugiego regionu
wielowartościowa (multivalue answer)	do ośmiu zdrowych adresów w odpowiedzi	prosty rozkład bez równoważnika
oparta na IP (IP-based)	według zakresu adresów klienta	ruch od konkretnego operatora do wybranego punktu

Kolejki, tematy, zdarzenia i przepływy

Usługa	Model	Kiedy jest odpowiedzią
Amazon SQS standard	kolejka, co najmniej jedno dostarczenie, kolejność niegwarantowana	rozdzielenie nadawcy i odbiorcy, bufor przed bazą, praca w tle
Amazon SQS FIFO	kolejka z kolejnością i deduplikacją	zamówienia, które muszą być przetworzone po kolei
Amazon SNS	publikacja i subskrypcja, wysyłka do wielu	fan-out do kilku kolejek, e-mail, SMS

Usługa	Model	Kiedy jest odpowiedzią
Amazon EventBridge	szyna zdarzeń z regułami i harmonogramem	reakcja na zdarzenia usług AWS i aplikacji SaaS, zadania o stałej porze
AWS Step Functions	przepływ pracy z krokami, ponowieniami i stanem	proces z wieloma krokami i obsługą błędów, na przykład zamówienie, płatność, wysyłka
Amazon API Gateway	brama API z limitami i kluczami	API REST albo WebSocket przed Lambdą, limitowanie żądań klientów
Amazon MQ	broker ActiveMQ albo RabbitMQ	migracja aplikacji, która już używa tych protokołów, bez przepisywania

Liczby SQS według dokumentacji w październiku 2026 roku: komunikat do 1 MiB (od sierpnia 2025 roku, wcześniej 256 KiB), przechowywanie domyślnie 4 dni, najwyżej 14 dni, czas niewidoczności domyślnie 30 sekund, najwyżej 12 godzin, opóźnienie do 15 minut. Starsze materiały do nauki podają 256 KiB i tę liczbę może jeszcze mieć część pytań.

Pułapki egzaminacyjne

- „Najmniejszy wysiłek operacyjny” przy odporności warstwy WWW: ALB i Auto Scaling w kilku strefach, a nie skrypt, który sprawdza serwer i uruchamia nowy.
- Aplikacja traci zamówienia, gdy odbiorca nie nadąża: kolejka SQS między nimi, a odbiorcy skalowani według liczby komunikatów w kolejce.
- Komunikat jest przetwarzany dwa razy: czas niewidoczności krótszy niż czas przetwarzania. Zwiększamy go, a gdy duplikaty są niedopuszczalne, kolejka FIFO.
- Stały adres IP dla klienta, który wpisuje adresy na swojej zaporze: NLB z Elastic IP albo Global Accelerator, a nie ALB.
- Rekord dla wierzchołka domeny `zielona.example` wskazujący ALB: alias, a nie CNAME.
- Skalowanie przed znanym szczytem o 6:00: zaplanowane działanie, a nie reaktywne śledzenie celu, które zareaguje za późno.
- Instancje są wymieniane w kółko zaraz po starcie: za krótki okres karencji kontroli zdrowia.
- Jedna instancja w jednej strefie z Elastic IP to nie jest wysoka dostępność, nawet z automatycznym odzyskiwaniem instancji.

Częste błędy

Objaw	Przyczyna	Co robimy
ALB zwraca 503	brak zdrowych celów w grupie	`describe-target-health`, ścieżka kontroli i grupa zabezpieczeń instancji

Objaw	Przyczyna	Co robimy
ALB zwraca 504	aplikacja odpowiada dłużej niż limit bezczynności równoważnika	optymalizacja albo dłuższy limit w atrybutach równoważnika
Cele `unhealthy` z `Target.Timeout`	grupa instancji nie wpuszcza ruchu z grupy ALB	reguła z `--source-group` grupy równoważnika
Grupa Auto Scaling nie uruchamia instancji	brak obrazu, błędny profil instancji albo limit vCPU	`describe-scaling-activities`
Rekord failover nie przełącza	kontrola zdrowia sprawdza adres niedostępny z internetu	kontrola na alarm CloudWatch albo publiczna ścieżka
Komunikaty wracają do kolejki bez końca	odbiorca nie usuwa komunikatu po przetworzeniu albo za krótki czas niewidoczności	`DeleteMessage` po sukcesie, dłuższy czas, kolejka martwych komunikatów

Ćwiczenia

1. Dzień 10 i 11. Stawiamy ALB, grupę docelową i grupę Auto Scaling z trzema instancjami, a skrypt startowy wypisuje nazwę strefy. Jak wygląda skrypt i jak sprawdzić, że ruch trafia do wszystkich stref?

Odpowiedź

Skrypt `www.sh` dla Amazon Linux 2023:

```
#!/bin/bash
dnf -y install nginx
token=$(curl -s -X PUT "http://169.254.169.254/latest/api/token" -H "X-aws-ec2-metadata-token-ttl-seconds: 300")
strefa=$(curl -s -H "X-aws-ec2-metadata-token: $token" http://169.254.169.254/latest/meta-data/placement/availability-zone)
echo "Zielona lab: $strefa" > /usr/share/nginx/html/index.html
echo "ok" > /usr/share/nginx/html/zdrowie
systemctl enable --now nginx
```

Po starcie odpytujemy równoważnik kilka razy:

```
for i in $(seq 1 9); do curl -s http://NazwaDnsRownowaznika/; done
```

W wynikach powinny się pojawić trzy strefy. Brama NAT jest potrzebna, bo `dnf` pobiera pakiety z internetu.

1. Dzień 11. Symulujemy awarię strefy: kończymy instancje w `eu-central-1a`. Co robi równoważnik, a co Auto Scaling, i ile to trwa?

Odpowiedź

```
aws ec2 describe-instances --filters "Name=tag:aws:autoscaling:groupName,Values=asg-lab-www" "Name=availability-zone,Values=eu-central-1a" --query "Reservations[].Instances[].InstanceId" --output text
aws ec2 terminate-instances --instance-ids IdInstancji
aws autoscaling describe-scaling-activities --auto-scaling-group-name asg-lab-www --max-items 3
```

Równoważnik przestaje kierować ruch do celu po kolejnych nieudanych kontrolach zdrowia (u nas dwie co 30 sekund, czyli około minuty). Auto Scaling zauważa brak instancji i uruchamia nową, starając się wyrównać strefy. Czas od zakończenia do nowej zdrowej instancji mierzymy sami i zapisujemy w dzienniku.

1. Dzień 13. Wysyłamy dziesięć komunikatów do kolejki `lab-zamowienia`, odbieramy je bez usuwania sześć razy i sprawdzamy, gdzie są. Jakie polecenia?

Odpowiedź

```
url=$(aws sqs get-queue-url --queue-name lab-zamowienia --query QueueUrl --output text)
for i in $(seq 1 10); do aws sqs send-message --queue-url "$url" --message-body "zamowienie-$i"; done
aws sqs receive-message --queue-url "$url" --max-number-of-messages 10 --visibility-timeout 1
aws sqs get-queue-attributes --queue-url "$url" --attribute-names ApproximateNumberOfMessages
```

Polecenie `receive-message` powtarzamy co kilka sekund. Po przekroczeniu `maxReceiveCount` równego 5 komunikaty przechodzą do `lab-zamowienia-dlq`, co sprawdzamy tym samym `get-queue-attributes` na drugiej kolejce. Jedno wywołanie może zwrócić mniej niż dziesięć komunikatów, bo standardowa kolejka odpytuje część serwerów.

1. Pytanie w stylu egzaminu: sklep przyjmuje zamówienia, które zapisuje do bazy, a w szczycie baza nie nadąża i klienci dostają błędy. Zamówienia mogą być zapisane z kilkuminutowym opóźnieniem. Co z najmniejszym wysiłkiem operacyjnym?

Odpowiedź

Kolejka SQS między aplikacją a bazą: aplikacja wysyła zamówienie do kolejki i od razu odpowiada klientowi, a odbiorcy (instancje w grupie Auto Scaling skalowanej według liczby komunikatów albo funkcja Lambda) zapisują je do bazy w tempie, które baza wytrzyma. Większa instancja bazy jest odpowiedzią gorszą, bo droższą i bez gwarancji przy następnym szczycie.

1. Pytanie w stylu egzaminu: firma ma aplikację w `eu-central-1` i chce, żeby przy awarii regionu klienci widzieli stronę z informacją o przerwie, bez ręcznej zmiany DNS. Co wybiera?

Odpowiedź

Route 53 z polityką failover: rekord główny alias na ALB z kontrolą zdrowia, rekord zapasowy alias na statyczną stronę w S3 (albo CloudFront) w innym regionie. Polityka ważona albo opóźnień nie da zachowania „tylko przy awarii”.

Krok projektu: sklep B2B

Aplikacja sklepu z `app01` trafia do grupy Auto Scaling w trzech strefach za ALB z WAF, minimum trzy instancje, maksimum dziewięć, śledzenie celu na liczbę żądań na cel i zaplanowane zwiększenie w listopadzie. Sesje klientów przechodzą do ElastiCache, zdjęcia do S3. Zamówienia idą przez SNS do trzech kolejek SQS: magazyn (synchronizacja z WMS na `ora01` przez VPN), faktury i powiadomienia, każda z kolejką martwych komunikatów. Route 53 trzyma `sklep.zielona.example` z polityką failover na stronę zastępczą w drugim regionie. Wymaganie Kasi to RTO 1 godzina i RPO 5 minut, a dane opisuje Odporne architektury, część 2: RDS Multi-AZ, Aurora, AWS Backup i strategie DR.

Powiązane: Bezpieczne architektury, część 2: VPC, szyfrowanie KMS, sekrety i ochrona aplikacji · Wysoka dostępność: pojęcia, liczenie dostępności i poziomy rezerwy · Macierz wysokiej dostępności usług Zielonej: poziom rezerwy każdej usługi, zależności i kolejność prac

Źródła: przewodnik SAA-C03, domena 2, zadania 2.1 i 2.2; dokumentacja Elastic Load Balancing (ALB, NLB, GWLB, kontrole zdrowia); dokumentacja Amazon EC2 Auto Scaling (szablony uruchamiania, polityki skalowania, ciepłe pule); dokumentacja Amazon Route 53 (polityki routingu, rekordy alias, kontrole zdrowia, strefy prywatne); dokumentacja Amazon SQS (limity komunikatów, kolejki martwych komunikatów), komunikat AWS z sierpnia 2025 o komunikatach do 1 MiB; dokumentacja Amazon SNS, EventBridge, Step Functions, API Gateway; dokumentacja Amazon EC2

(metadane instancji IMDSv2, publiczne parametry obrazów w SSM). Parametry poleceń sprawdzone z modelem usług AWS CLI (botocore 1.43) 19 października 2026.

Odporne architektury, część 2: RDS Multi-AZ, Aurora, AWS Backup i strategie DR

Poziom: średnio

Dotyczy: SAA-C03 domena 2, zadanie 2.2 (wysoka dostępność i odporność na awarie) oraz część zadań 1.3 i 3.3, Amazon RDS, Amazon Aurora, RDS Proxy, AWS Backup, AWS Elastic Disaster Recovery, replikacja S3, dni 15 do 17 planu

Czas: 3 dni po 1,5 do 2 godzin

Aplikacja bezstanowa przeżyje awarię strefy, jeśli przeżyją ją dane. Ta strona opisuje bazy RDS w trybie Multi-AZ i z replikami odczytu, Aurorę, kopie zapasowe w AWS Backup i cztery strategie odtwarzania po katastrofie z ich czasami RTO i RPO. Na egzaminie te słowa pojawiają się w co drugim pytaniu z tej domeny.

ZAŁOŻENIA SCENARIUSZA

ćwiczenia robimy w VPC laboratorium 10.60.0.0/16, a bazy stoją w podsieciach `db-a`, `db-b` i `db-c`. Region zapasowy w projekcie sklepu to `eu-west-1` (Irlandia); w laboratorium SCP `tylko-frankfurt` go blokuje, więc kopie między regionami opisujemy, a w laboratorium robimy tylko kopie w obrębie Frankfurtu. Pojęcia RPO i RTO w Zielonej opisuje strona Strategia kopii zapasowych Zielonej: RPO, RTO, reguła 3-2-1-1-0 i dwie lokalizacje.

Skróty użyte na tej stronie (23)

Skrót	Co znaczy
ACU	Aurora Capacity Unit: jednostka mocy Aurora Serverless, w której baza rośnie i maleje według obciążenia
API	Application Programming Interface: sposób, w jaki programy rozmawiają ze sobą
AWS	Amazon Web Services: chmura Amazonu
CLI	Command Line Interface: praca przez wpisywanie komend w terminalu
CRR	Cross-Region Replication: replikacja obiektów S3 do kubełka w innym regionie

Skrót	Co znaczy
DNS	Domain Name System: książka telefoniczna internetu, zamienia nazwy na adresy IP
DR	Disaster Recovery: plan przywrócenia działania po poważnej awarii
DynamoDB	baza danych typu klucz-wartość w chmurze AWS; Lore trzyma w niej fragmenty, stan, magazyn zmienny i blokady, gdy działa na AWS
EBS	Elastic Block Store: dyski blokowe maszyn wirtualnych EC2 w chmurze AWS, rozliczane osobno od maszyny
EC2	Elastic Compute Cloud: usługa AWS wynajmująca gotowe maszyny wirtualne na godziny
EFS	Elastic File System: współdzielony system plików NFS w AWS, dostępny z wielu instancji i stref
IP	Internet Protocol: adres komputera w sieci; w elektryce: stopień ochrony obudowy (Ingress Protection)
KMS	Key Management Service: usługa chmurowa przechowująca i zarządzająca kluczami szyfrującymi
PITR	Point-In-Time Recovery: odtworzenie bazy do dowolnej chwili, np. 14:31
RDS	Remote Desktop Services: zdalne pulpity i aplikacje Windows Server
RPO	Recovery Point Objective: ile danych wolno stracić przy awarii
RTO	Recovery Time Objective: jak długo wolno nie działać przy awarii
S3	Simple Storage Service: magazyn plików w chmurze AWS
SAA-C03	kod wersji egzaminu AWS Solutions Architect Associate zdawanej od 30 sierpnia 2022, aktualnej w październiku 2026
SCP	Service Connection Point: obiekt w Active Directory, po którym komputery domenowe znajdują swój tenant Entra ID przy dołączaniu hybrydowym
TiB	tebibajt: 2 do potęgi 40 bajtów, jednostka, w której liczą Windows i PowerShell (1TB w PowerShellu); dysk z etykietą 2 TB ma około 1,82 TiB
UTC	Coordinated Universal Time: uniwersalny czas bez strefy, punkt odniesienia dla wszystkich stref czasowych na świecie
VPC	Virtual Private Cloud: prywatna sieć wirtualna klienta w chmurze AWS, w której stawiamy maszyny i zakładamy tablice tras

Fiszka

Komendy wykonujemy w CloudShell z profilem konta laboratorium. `NazwaInstancji`, `NazwaKlastra`, `IdGrupyDb`, `NazwaSejfu`, `IdPlanu` i podobne zastępujemy własnymi. Instancje baz są płatne za godzinę, także zatrzymane po siedmiu dniach same się uruchamiają, więc usuwamy je w dniu ćwiczenia.

ZMIENIA SYSTEM

Utworzenie grupy podsieci dla bazy: aws rds create-db-subnet-group

```
aws rds create-db-subnet-group --db-subnet-group-name lab-db --db-subnet-group-description "Podsieci baz laboratorium" --subnet-ids IdPodsieciDbA IdPodsieciDbB IdPodsieciDbC
```

RDS wybiera z tej grupy podsieć dla instancji głównej i zapasowej. Grupa musi obejmować co najmniej dwie strefy.

RDS = Remote Desktop Services: zdalne pulpity i aplikacje Windows Server · **aws** = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Utworzenie bazy MariaDB w trybie Multi-AZ: aws rds create-db-instance --multi-az

```
aws rds create-db-instance --db-instance-identifier lab-sklep --engine mariadb --db-instance-class db.t4g.micro --allocated-storage 20 --storage-type gp3 --master-username admin_sklep --manage-master-user-password --multi-az --db-subnet-group-name lab-db --vpc-security-group-ids IdGrupyDb --no-publicly-accessible --storage-encrypted --kms-key-id alias/lab-dane --backup-retention-period 7
```

`--manage-master-user-password` zapisuje hasło administratora w Secrets Manager zamiast w poleceniu. `--multi-az` dokłada synchroniczną instancję zapasową w innej strefie, do której nie da się podłączyć, dopóki nie nastąpi przełączenie. Dostępność klasy `db.t4g.micro` w trybie Multi-AZ dla danej wersji silnika sprawdzamy w konsoli przy tworzeniu.

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Wymuszenie przełączenia na instancję zapasową: aws rds reboot-db-instance --force-failover

```
aws rds reboot-db-instance --db-instance-identifier lab-sklep --force-failover
aws rds describe-events --source-identifier lab-sklep --source-type db-instance --duration 30
```

Restart z przełączeniem to próba awarii strefy. Adres końcowy bazy się nie zmienia, zmienia się tylko adres IP za nim, więc aplikacja z krótkim czasem pamięci DNS łączy się ponownie sama.

IP = Internet Protocol: adres komputera w sieci; w elektryce: stopień ochrony obudowy (Ingress Protection) · **DNS** = Domain Name System: książka telefoniczna internetu, zamienia nazwy na adresy IP · **aws** = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Utworzenie repliki odczytu: aws rds create-db-instance-read-replica

```
aws rds create-db-instance-read-replica --db-instance-identifier lab-sklep-odczyt --source-db-instance-identifier lab-sklep
```

Replika dostaje dane asynchronicznie, ma własny adres i przyjmuje tylko odczyty. Może stać w innym regionie, a w razie potrzeby można ją awansować na samodzielną bazę poleceniem `aws rds promote-read-replica --db-instance-identifier lab-sklep-odczyt`.

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Odtworzenie bazy na wybraną chwilę: aws rds restore-db-instance-to-point-in-time

```
aws rds restore-db-instance-to-point-in-time --source-db-instance-identifier lab-sklep --target-db-instance-identifier lab-sklep-odtw --restore-time 2026-10-19T09:15:00Z --db-subnet-group-name lab-db --vpc-security-group-ids IdGrupyDb
```

Odtworzenie zawsze tworzy nową instancję z nowym adresem. Zamiast `--restore-time` można podać `--use-latest-restorable-time`. Najpóźniejszy możliwy czas pokazuje pole `LatestRestorableTime` w `describe-db-instances`.

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Ręczna migawka bazy: aws rds create-db-snapshot

```
aws rds create-db-snapshot --db-instance-identifier lab-sklep --db-snapshot-identifier
lab-sklep-przed-zmiana
```

Kopie automatyczne znikają razem z okresem przechowywania (do 35 dni) i domyślnie przy usunięciu instancji, a migawka ręczna zostaje, dopóki jej nie usuniemy.

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Utworzenie klastra Aurora MySQL z instancją zapisującą: aws rds create-db-cluster

```
aws rds create-db-cluster --db-cluster-identifier lab-aurora --engine aurora-mysql --
master-username admin_sklep --manage-master-user-password --db-subnet-group-name lab-db
--vpc-security-group-ids IdGrupyDb --storage-encrypted
aws rds create-db-instance --db-instance-identifier lab-aurora-1 --db-cluster-identifier
lab-aurora --engine aurora-mysql --db-instance-class db.t4g.medium
```

Klaster Aurora to wspólny wolumen danych i instancje, które z niego korzystają. Druga instancja w innej strefie (to samo polecenie z inną nazwą) staje się repliką, na którą klaster przełącza się przy awarii.

aws = AWS CLI: narzędzie chmury Amazonu

TYLKO ODCZYT

Przełączenie klastra Aurora: aws rds failover-db-cluster

```
aws rds failover-db-cluster --db-cluster-identifier lab-aurora
```

Adres zapisu klastra zaczyna wskazywać awansowaną replikę. Bez repliki Aurora tworzy nową instancję, co trwa dłużej.

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Utworzenie sejfu kopii: aws backup create-backup-vault

```
aws backup create-backup-vault --backup-vault-name lab-sejf --encryption-key-arn ArnKluczaKms
```

Sejf to kontener na punkty odtworzenia, szyfrowany kluczem KMS i z własną polityką dostępu.

KMS = Key Management Service: usługa chmurowa przechowująca i zarządzająca kluczami szyfrującymi · **aws** = AWS CLI: narzędzie chmury Amazonu

TYLKO ODCZYT

Zablokowanie sejfu przed skróceniem retencji: aws backup put-backup-vault-lock-configuration

```
aws backup put-backup-vault-lock-configuration --backup-vault-name lab-sejf --min-retention-days 7 --max-retention-days 35 --changeable-for-days 3
```

Przez trzy dni blokadę można jeszcze usunąć (tryb zarządzania), potem staje się nieodwracalna (tryb zgodności) i nikt, także użytkownik główny, nie usunie kopii przed czasem. W laboratorium pomijamy `--changeable-for-days`, żeby blokada została w trybie zarządzania i dała się usunąć.

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Utworzenie planu kopii: aws backup create-backup-plan

```
aws backup create-backup-plan --backup-plan file://plan-kopii.json
```

aws = AWS CLI: narzędzie chmury Amazonu

PLIK KONFIGURACYJNY

Plan codziennej kopii z kopią do drugiego regionu: plan-kopii.json

```
{
  "BackupPlanName": "sklep-codziennie",
  "Rules": [
    {
```

```
      "RuleName": "codziennie-0100",
      "TargetBackupVaultName": "lab-sejf",
      "ScheduleExpression": "cron(0 1 * * ? *)",
      "StartWindowMinutes": 60,
      "Lifecycle": { "DeleteAfterDays": 35 },
      "CopyActions": [
        {
          "DestinationBackupVaultArn": "arn:aws:backup:eu-west-1:IdKonta:backup-
vault:sejf-irlandia",
          "Lifecycle": { "DeleteAfterDays": 35 }
        }
      ]
    }
  ]
}
```

Wyrażenie `cron` w AWS ma sześć pól (z rokiem) i liczy czas w UTC. W laboratorium usuwamy blok `CopyActions`, bo SCP blokuje `eu-west-1`.

AWS = Amazon Web Services: chmura Amazonu · **UTC** = Coordinated Universal Time: uniwersalny czas bez strefy, punkt odniesienia dla wszystkich stref czasowych na świecie · **SCP** = Service Connection Point: obiekt w Active Directory, po którym komputery domenowe znajdują swój tenant Entra ID przy dołączaniu hybrydowym

ZMIENIA SYSTEM

Wskazanie zasobów do kopii po tagu: aws backup create-backup-selection

```
aws backup create-backup-selection --backup-plan-id IdPlanu --backup-selection
file://wybor-kopii.json
```

aws = AWS CLI: narzędzie chmury Amazonu

PLIK KONFIGURACYJNY

Wybór zasobów po tagu: wybor-kopii.json

```
{
  "SelectionName": "tag-kopia-codziennie",
```

```
  "IamRoleArn": "arn:aws:iam::IdKonta:role/service-role/AWSBackupDefaultServiceRole",
  "ListOfTags": [
    { "ConditionType": "STRINGEQUALS", "ConditionKey": "kopia", "ConditionValue":
"codziennie" }
  ]
}
```

Każdy zasób z tagiem `kopia=codziennie` (baza, dysk EBS, system plików EFS, tabela DynamoDB) trafia do planu bez zmiany planu. Rola domyślna powstaje przy pierwszym użyciu AWS Backup w konsoli, a jej istnienie sprawdzamy przed wywołaniem.

EBS = Elastic Block Store: dyski blokowe maszyn wirtualnych EC2 w chmurze AWS, rozliczane osobno od maszyny · **EFS** = Elastic File System: współdzielony system plików NFS w AWS, dostępny z wielu instancji i stref · **DynamoDB** = baza danych typu klucz-wartość w chmurze AWS; Lore trzyma w niej fragmenty, stan, magazyn zmienny i blokady, gdy działa na AWS · **AWS** = Amazon Web Services: chmura Amazonu

TYLKO ODCZYT

Lista punktów odtworzenia w sejfie: aws backup list-recovery-points-by-backup-vault

```
aws backup list-recovery-points-by-backup-vault --backup-vault-name lab-sejf --query
"RecoveryPoints[].[ResourceType,CreationDate,Status]" --output table
```

aws = AWS CLI: narzędzie chmury Amazonu

ZMIENIA SYSTEM

Włączenie odtwarzania na chwilę dla tabeli DynamoDB: aws dynamodb update-continuous-backups

```
aws dynamodb update-continuous-backups --table-name NazwaTabeli --point-in-time-
recovery-specification PointInTimeRecoveryEnabled=true
```

Tabelę można potem odtworzyć do dowolnej sekundy z okresu przechowywania, domyślnie 35 dni, do nowej tabeli.

DynamoDB = baza danych typu klucz-wartość w chmurze AWS; Lore trzyma w niej fragmenty, stan, magazyn zmienny i blokady, gdy działa na AWS · **aws** = AWS CLI: narzędzie chmury Amazonu

Z życia wzięte

Z ŻYCIA WZIĘTE

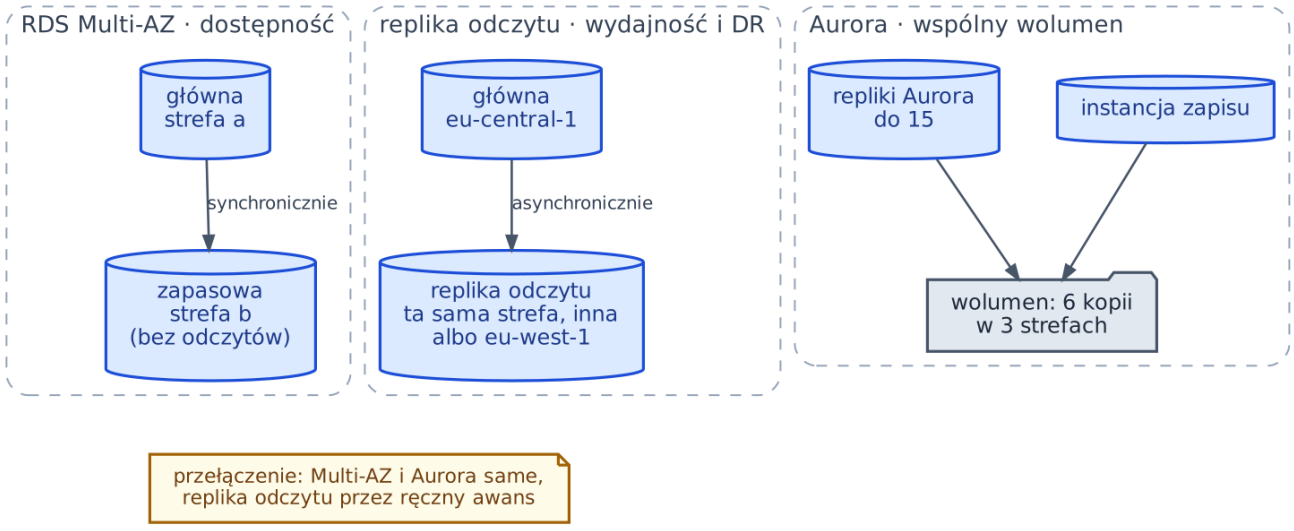
Czwartek, 11:45. Tomek pokazuje Kasi bazę sklepu w laboratorium z repliką odczytu w innej strefie i mówi, że to jest wysoka dostępność. Ola pyta, co się stanie, gdy w nocy padnie strefa z bazą główną, a nikt nie będzie przy komputerze. Tomek sprawdza w dokumentacji i odkrywa, że replika odczytu sama się nie awansuje, a Multi-AZ przełącza się bez nas. W projekcie zostają oba: Multi-AZ na awarię strefy, replika odczytu dla raportów Kasi.

Jak to działa

Rozwiń: jak to działa pod spodem

RDS: Multi-AZ i repliki odczytu

Multi-AZ służy dostępności, a repliki odczytu wydajności. To najczęstsze rozróżnienie w tej domenie i warto je mieć jako jedno zdanie w dzienniku.



Rysunek 1. Trzy sposoby odporności bazy: RDS Multi-AZ z synchroniczną instancją zapasową, repliki odczytu z asynchroniczną kopią, także w innym regionie, i Aurora ze wspólnym wolumenem w trzech strefach (kliknięcie otwiera pełny rozmiar)

Cecha	RDS Multi-AZ (instancja)	RDS Multi-AZ (klaster, dwie zapasowe)	Replika odczytu	Aurora
Cel	dostępność	dostępność i odczyty	odczyty, kopia w innym regionie	dostępność, odczyty, wydajność
Replikacja	synchroniczna	półsynchroniczna	asynchroniczna	wspólny wolumen, sześć kopii w trzech strefach
Odczyt z kopii	nie	tak	tak	tak, do 15 replik
Przełączenie	automatyczne, ten sam adres	automatyczne	ręczne awansowanie	automatyczne, adres zapisu klastra
Inny region	nie	nie	tak	Aurora Global Database
Silniki	wszystkie RDS	MySQL i PostgreSQL	większość silników	Aurora MySQL i Aurora PostgreSQL

Czasy przełączenia podajemy z dokumentacji jako rzędy wielkości: Multi-AZ instancji to zwykle jedna do dwóch minut, Aurora z repliką zwykle poniżej minuty. Na egzaminie liczy się porównanie, a nie dokładna liczba.

> **Pojęcie: RDS Proxy.** Zarządzana pula połączeń przed bazą RDS albo Aurora. Przydaje się, gdy wiele krótkich połączeń (funkcje Lambda) zabija bazę liczbą sesji, i skraca przełączenie, bo aplikacja rozmawia z proxy, a nie z adresem bazy.

> **Pojęcie: Aurora Serverless v2.** Tryb Aurory, w którym moc instancji rośnie i maleje sama w jednostkach ACU, według obciążenia. Dobry dla zmiennego albo nieprzewidywalnego ruchu.

Aurora rośnie sama do 256 TiB (od lipca 2025 roku dla nowszych wersji silników, wcześniej 128 TiB, starsze materiały podają tę drugą liczbę). Aurora Global Database replikuje do innych regionów z opóźnieniem typowo poniżej sekundy, a przełączenie regionu trwa według AWS około minuty.

Kopie zapasowe w AWS

Co	Mechanizm	Uwagi
RDS i Aurora	kopie automatyczne z odtwarzaniem na chwilę (1 do 35 dni) i migawki ręczne	odtwarzanie tworzy nową instancję
EBS	migawki przyrostowe w S3	kopia do innego regionu, Data Lifecycle Manager do harmonogramu
S3	wersjonowanie, replikacja w regionie (SRR) i między regionami (CRR), Object Lock	replikacja wymaga wersjonowania po obu stronach i nie kopiuje wstecz obiektów sprzed jej włączenia bez Batch Replication
DynamoDB	odtwarzanie na chwilę (PITR) i kopie na żądanie	odtworzenie do nowej tabeli
EFS, FSx	AWS Backup	
Wszystko naraz	AWS Backup: plany, sejfy, blokada sejfu, kopie między kontami i regionami	jedno miejsce na politykę kopii i raport dla audytora
Całe serwery z serwerowni	AWS Elastic Disaster Recovery	ciągła replikacja dysków, uruchomienie w AWS w minutach

Strategie odtwarzania po katastrofie

Pojęcie: RPO (Recovery Point Objective). Ile danych, liczonych w czasie, możemy stracić. RPO 5 minut oznacza, że po katastrofie brakuje najwyżej ostatnich pięciu minut.

Pojęcie: RTO (Recovery Time Objective). Jak długo system może nie działać, od awarii do przywrócenia.



Rysunek 2. Cztery strategie DR od najtańszej do najdroższej: kopia i odtworzenie, pilot light, ciepła rezerwa i aktywne w wielu miejscach, z typowymi RTO i RPO (kliknięcie otwiera pełny rozmiar)

Strategia	Co działa w regionie zapasowym	Typowe RTO i RPO	Koszt
kopia i odtworzenie (backup and restore)	tylko kopie	godziny	najniższy
pilot light	dane replikowane na bieżąco, serwery wyłączone albo w szablonie	dziesiątki minut, RPO minuty	niski
ciepła rezerwa (warm standby)	pomniejszona, działająca kopia całego systemu	minuty	średni
aktywne w wielu miejscach (multi-site active/active)	pełny system przyjmujący ruch	bliskie zera	najwyższy

Typowe wartości to rzędy wielkości z dokumentacji AWS o DR, a nie gwarancje. W pytaniu szukamy najtańszej strategii, która spełnia podane RTO i RPO.

Pułapki egzaminacyjne

- „Wysoka dostępność bazy” albo „awaria strefy”: Multi-AZ. Replika odczytu jest odpowiedzią na „raporty spowalniają bazę”.
- Replika odczytu w innym regionie to dobry element strategii DR, ale bez awansowania nie przyjmie zapisów.
- RPO w sekundach i RTO w minutach dla bazy w wielu regionach: Aurora Global Database (albo DynamoDB global tables dla NoSQL).

- Wiele funkcji Lambda wyczerpuje połączenia do bazy: RDS Proxy.
- Szyfrowanie istniejącej bazy: migawka, kopia migawki z szyfrowaniem, odtworzenie. Replika odczytu nie zaszyfruje się sama, jeśli źródło nie jest szyfrowane.
- Kopie mają być nieusuwalne nawet przez administratora: AWS Backup Vault Lock w trybie zgodności albo S3 Object Lock w trybie zgodności.
- Stary system z serwerowni ma mieć DR w AWS bez przepisywania: AWS Elastic Disaster Recovery.
- Przy usunięciu instancji RDS kopie automatyczne domyślnie znikają; migawkę końcową trzeba zrobić albo zachować kopie automatyczne.

Częste błędy

Objaw	Przyczyna	Co robimy
Aplikacja po przełączeniu Multi-AZ łączy się dalej ze starym adresem IP	pamięć DNS w aplikacji albo w JVM	krótki czas pamięci DNS, ponowne łączenie w kodzie, RDS Proxy
Nie da się wybrać `--multi-az` dla klasy instancji	ta klasa albo wersja silnika nie obsługuje Multi-AZ w regionie	inna klasa, sprawdzenie w konsoli
Odtworzona baza nie ma tych samych grup i parametrów	odtworzenie tworzy nową instancję z domyślnymi ustawieniami	podajemy grupę podsieci, grupy zabezpieczeń i grupę parametrów w poleceniu
Plan AWS Backup nie robi kopii	zasób nie ma tagu albo rola nie ma uprawnień	tag `kopia=codziennie`, `AWSBackupDefaultServiceRole`
Kopia do drugiego regionu kończy się błędem uprawnień	SCP blokuje region albo klucz KMS w regionie docelowym	wyjątek w SCP dla regionu zapasowego, sejf z kluczem w tamtym regionie
Rachunek za bazę po zakończeniu ćwiczeń	instancja zatrzymana uruchamia się sama po siedmiu dniach	usuwamy instancję z `--skip-final-snapshot` po ćwiczeniu

□□ Ćwiczenia

1. Dzień 15. Tworzymy bazę `lab-sklep` w trybie Multi-AZ, wymuszamy przełączenie i mierzymy, jak długo baza nie przyjmowała połączeń. Jak to zmierzyć z instancji w podsieci aplikacji?

Odpowiedź

Na instancji z klientem MariaDB uruchamiamy pętlę, która co sekundę wypisuje czas i wynik prostego zapytania, a w drugim oknie wymuszamy przełączenie:

```
while true; do printf '%s ' "$(date +%T)"; mysql -h AdresBazy -u admin_sklep -p"$HASLO" -e "select 1" -sN 2>&1 | head -1; sleep 1; done
```

```
aws rds reboot-db-instance --db-instance-identifier lab-sklep --force-failover
```

Liczymy sekundy z błędami połączenia. Hasło pobieramy wcześniej z Secrets Manager do zmiennej `HASLO`. Po ćwiczeniu: `aws rds delete-db-instance --db-instance-identifier lab-sklep --skip-final-snapshot --delete-automated-backups`.

1. Dzień 17. Kasia wymaga dla sklepu RTO 1 godzina i RPO 5 minut przy utracie całego regionu Frankfurt. Która najtańsza strategia DR to spełni i z jakich usług ją zbudujemy?

Odpowiedź

Pilot light: replika odczytu bazy MariaDB w `eu-west-1` (albo Aurora Global Database, jeśli przejdziemy na Aurorę), kubełek zdjęć z replikacją CRR, szablon CloudFormation całej warstwy aplikacji gotowy do uruchomienia, Route 53 z polityką failover. Przy awarii awansujemy replikę, uruchamiamy stos i przełączamy DNS. Replika asynchroniczna daje RPO poniżej pięciu minut przy normalnym opóźnieniu, a uruchomienie stosu i awans mieszczą się w godzinie, co sprawdzamy próbą. Sama kopia i odtworzenie nie da RPO 5 minut, a ciepła rezerwa jest droższa niż potrzeba.

1. Dzień 17. Zakładamy sejf `lab-sejf`, plan `sklep-codziennie` bez kopii do drugiego regionu i wybór po tagu, oznaczamy tagiem dysk EBS i uruchamiamy kopię na żądanie. Jakie polecenie uruchamia kopię od razu?

Odpowiedź

```
aws backup start-backup-job --backup-vault-name lab-sejf --resource-arn arn:aws:ec2:eu-central-1:IdKonta:volume/IdDysku --iam-role-arn arn:aws:iam::IdKonta:role/service-role/AWSBackupDefaultServiceRole
aws backup list-backup-jobs --by-backup-vault-name lab-sejf --query "BackupJobs[].[ResourceType,State]" --output table
```

Punkt odtworzenia pojawia się w `list-recovery-points-by-backup-vault`. Przed usunięciem sejfu trzeba usunąć wszystkie punkty, a sejfu z blokadą w trybie zgodności nie da się wyczyścić przed końcem retencji.

1. Pytanie w stylu egzaminu: aplikacja raportowa wykonuje ciężkie zapytania na produkcyjnej bazie RDS PostgreSQL i spowalnia sklep. Co robimy z najmniejszym wysiłkiem?

Odpowiedź

Replika odczytu RDS i skierowanie aplikacji raportowej na jej adres. Multi-AZ z jedną instancją zapasową nie pomoże, bo instancja zapasowa nie przyjmuje odczytów. ElastiCache też nie jest odpowiedzią przy ciężkich, różnorodnych zapytaniach raportowych.

1. Pytanie w stylu egzaminu: firma potrzebuje relacyjnej bazy z odczytami w trzech regionach o małym opóźnieniu i przełączeniem regionu w około minutę przy RPO rzędu sekundy. Co wybiera?

Odpowiedź

Aurora Global Database z klastrem głównym w jednym regionie i klastrami wtórnymi w dwóch pozostałych. Repliki odczytu RDS między regionami dałyby odczyty, ale nie takie RPO i RTO przy przełączeniu.

Krok projektu: sklep B2B

Baza `sklep_b2b` z `my01` przechodzi do RDS for MariaDB w trybie Multi-AZ w podsieciach `db`, szyfrowana kluczem zarządzanym przez klienta, z kopiami automatycznymi przez 14 dni i repliką odczytu w `eu-west-1` dla strategii pilot light. Katalog z `mongo01` ma dwie drogi do decyzji: Amazon DocumentDB (zgodny z API MongoDB, zarządzany) albo MongoDB na EC2 z własną replikacją, co opisujemy przy wyborze baz na stronie Wydajne architektury: obliczenia, magazyn, bazy, CloudFront i strumienie danych. AWS Backup z planem po tagu kopiuje bazę i dyski do sejfu z blokadą, z kopią do drugiego regionu. Próbę przełączenia regionu robimy raz na kwartał, poza okresem zakazu zmian od 15 listopada do 6 stycznia.

Powiązane: Odporne architektury, część 1: strefy, ELB, Auto Scaling, Route 53 i luźne powiązania · Strategia kopii zapasowych Zielonej: RPO, RTO, reguła 3-2-1-1-0 i dwie lokalizacje · Replikacja · Zimne archiwum w chmurze: Azure Blob Storage i Amazon S3 Glacier · Azure Site Recovery dla Hyper-V bez VMM: replikacja, test, przełączenie i powrót

Źródła: przewodnik SAA-C03, domena 2, zadanie 2.2; dokumentacja Amazon RDS (Multi-AZ instancji i klastra, repliki odczytu, odtwarzanie na chwilę, RDS Proxy, automatyczne uruchamianie zatrzymanych instancji po siedmiu dniach); dokumentacja Amazon Aurora (wolumen w trzech strefach, Global Database, Serverless v2), komunikaty AWS z lipca 2025 o wolumenach do 256 TiB;

dokumentacja AWS Backup (plany, wybory, Vault Lock); whitepaper AWS „Disaster Recovery of Workloads on AWS” (cztery strategie); dokumentacja DynamoDB (PITR); dokumentacja AWS Elastic Disaster Recovery. Parametry poleceń sprawdzone z modelem usług AWS CLI (botocore 1.43) 19 października 2026.