VIGIL GUARD

Profile Skalowania 1.9

Jak skaluje się Vigil Guard Enterprise 1.9 w środowiskach produkcyjnych

PRZEWIDYWALNE SKALOWANIE WDROŻEŃ

Vigil Guard Enterprise 1.9 skaluje środowiska produkcyjne za pomocą gotowych profili skalowania. Profil jest dobierany i walidowany względem zasobów hosta podczas instalacji lub zatwierdzonej zmiany konfiguracji, a następnie utrwalany.

Profile są dobrane do ruchu mieszanego zbliżonego do produkcyjnego. Wyższy profil uruchamia więcej instancji modeli detekcji i analizy semantycznej, a budżety współbieżności wynikają z liczby rdzeni CPU hosta. Wdrożenie działa powtarzalnie po restartach, aktualizacjach i serwisie.

KLUCZOWE KORZYŚCI

- Cztery gotowe profile dopasowane do klasy hosta
- Analiza semantyczna skaluje się z profilem (1–4 instancje)
- Budżety żądań dopasowane do hosta
- Pełna jakość detekcji PII pod obciążeniem
- Walidacja profilu względem zasobów i pojemności CPU hosta
- Powtarzalne wdrożenia po restartach, aktualizacjach i serwisie
- Aktualizacje zachowują wybrany profil
- Sygnał, gdy potrzebna jest większa klasa hosta

4
profile skalowania

20-52
rdzenie CPU (zakres klas)

30-256 GB
RAM (zakres klas)

1 host
wdrożenie jednowęzłowe Linux

Wdrożenie lokalne (on-premise) • Sieć odizolowana (air-gapped) • Jeden serwer Linux

PROFILE SKALOWANIA

Każde wdrożenie korzysta z jednego z czterech profili skalowania, dobranego do klasy hosta. Profil określa minimalne wymagania serwera oraz docelowe przeznaczenie produkcyjne.

PROFIL	KLASA WDROŻENIA	CPU (MIN.)	RAM (MIN.)	DYSK (ZALECANY)	PRZEZNACZENIE
prod-32-balanced	Wejściowa	20 rdzeni	30 GB	150 GB+ NVMe SSD	Niski ruch, środowiska wewnętrzne lub kontrolowane pilotaże.
prod-64-balanced	Bazowa (zalecana)	28 rdzeni	64 GB*	150-200 GB NVMe SSD	Domyślny wybór dla większości wdrożeń produkcyjnych.
prod-128-balanced	Wzrostowa	36 rdzeni	128 GB*	300 GB+ NVMe SSD	Wyższa, utrzymująca się współbieżność i większe obciążenie analityczne.
prod-256-balanced	Duża jednowęzłowa	52 rdzenie	256 GB*	500 GB+ NVMe SSD	Największe wdrożenia jednowęzłowe z dużym zapasem zasobów.

* Minima RAM od 64 GB akceptują 95% wartości nominalnej (60, 121 i 243 GiB widoczne w systemie, który zawsze raportuje mniej niż zainstalowano), więc standardowy serwer 64, 128 lub 256 GB kwalifikuje się do swojego profilu.

ZALECANY PUNKT STARTOWY

Dla większości nowych wdrożeń produkcyjnych zalecanym profilem jest **prod-64-balanced** na dedykowanym serwerze Linux. Zapewnia bezpieczny zapas dla danych, analityki i inferencji modeli oraz miejsce na wzrost. Profil prod-32-balanced to najmniejszy zgodny kształt, nie domyślny wybór enterprise.

JAK DZIAŁAJĄ PROFILE

Każdy profil skaluje oba wąskie gardła platformy, a budżety są dobrane do ruchu mieszanego zbliżonego do produkcyjnego (krótkie, średnie i długie prompty jednocześnie).

- **Analiza semantyczna skaluje się z profilem.** Profile uruchamiają 1, 2, 4 i 4 instancje analizy semantycznej, każdą z ograniczoną pulą wątków, dzięki czemu CPU wykonuje realną pracę zamiast jałowego oczekiwania.
- **Modele detekcji stale obciążone.** Każdy profil utrzymuje do 16 żądań w toku na instancjach modeli detekcji, więc procesy modeli są stale zasilane pracą.
- **Przyjmowanie ruchu dopasowane do hosta.** API przyjmuje jedno równoległe żądanie na rdzeń CPU (20–52 przy minimach profili, maks. 64).
- **Detekcja PII gotowa na skoki ruchu.** Kolejka silnika PII mieści pełną serię żądań, a potok językowy działa tylko, gdy wymaga go typ encji, więc detekcja PII zachowuje pełną jakość pod obciążeniem.
- **Bramki pojemności.** Planowany kształt musi zmieścić się w CPU hosta z zapasem dla reszty platformy, w tym funkcji opcjonalnych, takich jak moderacja treści i scope drift.

CO URUCHAMIA KAŻDY PROFIL

PROFIL	INSTANCJE MODELI DETEKCJI	INSTANCJE ANALIZY SEMANTYCZNEJ	ŻĄDANIA RÓWNOLEGŁE*	ŻĄDANIA DETEKCJI W TOKU
prod-32-balanced	2	1	20	16
prod-64-balanced	2	2	28	16
prod-128-balanced	3	4	36	16
prod-256-balanced	4	4	52	16

* Przy minimalnej liczbie rdzeni profilu. Budżety podążają za faktyczną liczbą rdzeni hosta, więc mocniejszy serwer dostaje większe budżety (do 64 żądań).

JAK DOBRAĆ HOSTA

- **Zacznij od prod-64-balanced** dla większości wdrożeń produkcyjnych. Profil prod-32-balanced stosuj tylko do pilotaży i środowisk o niskim ruchu.
- **Wybierz prod-128-balanced** przy wyższej, utrzymującej się współbieżności: dodaje trzecią instancję modeli detekcji i uruchamia cztery instancje analizy semantycznej.
- **Wybierz prod-256-balanced** dla największych wdrożeń jednowęzłowych, z czterema instancjami modeli detekcji.
- **CPU ponad minimum jest wykorzystywane**: budżety przyjmowania ruchu i procesów roboczych rosną z faktyczną liczbą rdzeni.
- **RAM**: dobieraj według zainstalowanej pamięci. Standardowy serwer 64, 128 lub 256 GB kwalifikuje się do swojego profilu, bo walidacja uwzględnia pamięć raportowaną przez system operacyjny.
- **Dysk**: stosuj NVMe SSD w zalecanym rozmiarze; zostawia miejsce na retencję zdarzeń i wzrost analityki.

CO ZMIENIA WYŻSZY PROFIL

Skalowanie w wersji 1.9 to nie tylko większy serwer. Wraz z profilem rośnie realna zdolność przetwarzania, a system pozostaje stabilny pod obciążeniem.

Większa równoległość analizy

Wyższy profil uruchamia więcej instancji modeli detekcji i analizy semantycznej, zwiększając przepustowość najcięższych etapów oraz odporność na ruch szczytowy i zapytania unikalne.

Współbieżność dopasowana do hosta

Budżety przyjmowania ruchu, żądań w toku i wątków wynikają z faktycznej liczby rdzeni CPU hosta, dzięki czemu procesy modeli są stale zajęte bez przeciążania procesora.

Ochrona przed przeciążeniem

Mechanizmy dopuszczania ruchu i ograniczone kolejki pracy ograniczają degradację pod obciążeniem i chronią środowisko. Silnik PII obsługuje pełne serie żądań, więc detekcja zachowuje pełną jakość.

Zapas dla danych i analityki

Większa klasa hosta zwiększa zapas pamięci, dysku i pamięci podręcznej dla magazynu zdarzeń i obciążeń analitycznych.

WARSTWY PLATFORMY A SKALOWANIE

Platforma działa jako jeden spójny stos na serwerze Linux. Poszczególne warstwy w różnym stopniu wpływają na dobór klasy hosta.

WARSTWA	ROLA	WPŁYW NA KLASĘ HOSTA
Reverse proxy / brama TLS	Wejście ruchu i terminacja TLS	Lekka, stała ścieżka wejścia.
Rdzeń danych i analityka	Magistrala komunikatów, pamięć podręczna, baza danych analityczna, magazyn wektorowy, kolektor metryk	Wyznacza zapotrzebowanie na pamięć, dysk i wydajność I/O.
Aplikacja i API	Obsługa żądań i panel zarządzania	Umiarkowana pamięć, wrażliwość na opóźnienia użytkownika. Budżet przyjmowania ruchu rośnie z liczbą rdzeni.
Przetwarzanie i analiza	Równoległe warstwy detekcji, analiza semantyczna	Główne zużycie CPU w trakcie analizy. Analiza semantyczna skaluje się z profilem.
Usługi modeli ML	Inferencja modeli detekcji	Skalowane wraz z profilem; główne wąskie gardło wydajności.

WALIDACJA I POWTARZALNOŚĆ

Każdy profil jest walidowany względem dostępnych zasobów hosta: liczby rdzeni CPU, pamięci RAM i wolnego miejsca na dysku. Dodatkowo bramki pojemności sprawdzają, czy planowane instancje modeli detekcji wraz z ich budżetami wątków mieszczą się w CPU hosta z zapasem dla reszty platformy. Jeśli serwer nie spełnia wymagań danego profilu, profil nie zostanie zastosowany. Zatwierdzona konfiguracja jest utrwalana, dzięki czemu kolejne uruchomienia środowiska są powtarzalne, a stan działający nie rozjeżdża się z tym, co zatwierdzono.

AKTUALIZACJE I ZMIANA PROFILU

Aktualizacja zachowuje wybrany profil skalowania i ponownie dopasowuje konfigurację uruchomieniową do aktualnych zasobów hosta oraz nowej wersji platformy. Dzięki temu aktualizacja nie zmienia charakterystyki działającego środowiska w niekontrolowany sposób.

Każda aktualizacja automatycznie generuje konfigurację skalowania dla docelowej wersji; konfiguracja nigdy nie jest przenoszona ręcznie. Narzędzia administracyjne rekomendują też właściwy profil dla bieżącego hosta.

Zmiana profilu odbywa się przez dostarczone narzędzia administracyjne i jest potwierdzana procedurą weryfikacji wdrożenia. Utrzymuje to konfigurację w zgodzie z procesem instalacji i zachowuje powtarzalność wdrożenia.

> **WYDAJNOŚĆ I JEJ UWARUNKOWANIA**
>
> Wydajność poszczególnych wdrożeń zależy od charakteru ruchu, udziału zapytań unikalnych, konfiguracji retencji danych, obciążenia analitycznego, włączonych funkcji detekcji, parametrów hosta oraz warunków środowiska klienta. Z tego powodu wartości przepustowości i opóźnień nie powinny być traktowane jako uniwersalna gwarancja. Dla większych klas wdrożeniowych końcową ocenę wydajności należy potwierdzić w docelowym środowisku klienta.

KIEDY PRZEJŚĆ NA WIĘKSZY PROFIL

Następujące sygnały wskazują, że warto zaplanować większą klasę hosta i wyższy profil skalowania:

- Zużycie pamięci często zbliża się do skonfigurowanych limitów usług analitycznych i modeli.
- Rośnie obciążenie analityczne, wolumen zdarzeń lub wymagania dotyczące retencji danych.
- Opóźnienie analizy lub czas odpowiedzi rosną przy normalnym ruchu biznesowym. Przy przeciążeniu system najpierw wydłuża czas odpowiedzi, zamiast odrzucać żądania.
- Monitoring raportuje częściową utratę pojemności lub ostrzeżenia o zasobach hosta.
- Zużycie dysku zbliża się do skonfigurowanych progów.

W razie potrzeby przejdź na kolejną, większą klasę hosta i odpowiadający jej profil skalowania zamiast doraźnych zmian konfiguracji.

OPCJE WDROŻENIA

- **On-premise**: jeden dedykowany serwer Linux, pełna kontrola nad danymi.
- **Air-gapped**: pełna funkcjonalność bez dostępu do internetu.
- **Bez wymogu GPU**: inferencja modeli działa na CPU.

Skalowanie dopasowane do Twojego wdrożenia

Cztery gotowe profile skalowania pozwalają dopasować Vigil Guard Enterprise 1.9 do klasy hosta i obciążenia, zachowując powtarzalne i przewidywalne wdrożenia.

ZALECANY STANDARD: PROD-64-BALANCED

Dedykowany serwer Linux, 28 rdzeni CPU, 64 GB RAM, dysk NVMe SSD min. 150-200 GB. Najbardziej zrównoważony punkt startowy dla wdrożeń produkcyjnych: dwie instancje modeli detekcji i dwie analizy semantycznej, z bezpiecznym zapasem dla analityki, inferencji modeli oraz wzrostu.

VIGIL GUARD ENTERPRISE 1.9

Cztery gotowe profile skalowania. Walidacja względem zasobów hosta. Powtarzalne wdrożenia.

Wdrożenie lokalne, sieć odizolowana, jeden serwer Linux, bez wymogu GPU.

Zespoły Wdrożeniowe

Powtarzalne, przewidywalne uruchomienia środowiska

Operacje / SRE

Dobór profilu i kontrolowane aktualizacje

Architektura / Infrastruktura

Jasne wymagania hosta i planowanie pojemności

vigilguard.ai | contact@vigilguard.ai