Rekomendacja infrastruktury serwerowej dla AsWiseAI

Producent:
Firma Informatyczna EnterSoft
AsWiseAI.pl, EnterSoft.pl

Wersja v 1.0 <-> 2026

Spis treści

Firma Informatyczna EnterSoft
e-mail: biuro@entersoft.pl, entersoft.pl, kom.: **696 481 842**

Przykładowe zestawienia serwerów według skali wdrożenia, liczby dokumentów i oczekiwanej równoległości pracy

Cel dokumentu

Dokument przedstawia orientacyjne warianty infrastruktury serwerowej dla wdrożeń AsWiseAI.

Pozycja	Opis
Opracowanie	Rekomendacja orientacyjna dla wdrożeń on-premise AsWiseAI
Wersja	1.0
Data	2026-05-11
Charakter dokumentu	Dokument prezentacyjno-rekomendacyjny, nie zastępuje finalnego audytu infrastruktury

Firma Informatyczna EnterSoft

e-mail: biuro@entersoft.pl, entersoft.pl, kom.: **696 481 842**

EnterSoft

1. Wstęp

AsWiseAI jest platformą klasy on-premise RAG, czyli systemem AI działającym na infrastrukturze klienta i analizującym firmowe dokumenty bez konieczności przesyłania ich do zewnętrznych usług chmurowych. Z tego powodu dobór serwera powinien uwzględniać nie tylko liczbę użytkowników, ale przede wszystkim wolumen dokumentów, liczbę stron, udział skanów OCR, wybrany model językowy oraz oczekiwaną liczbę równoległych zapytań do AI.

Najważniejsza zasada sizingu jest prosta: dokumenty obciążają bazę wiedzy i dyski, natomiast modele AI obciążają głównie GPU oraz pamięć VRAM. Dlatego dwa wdrożenia z taką samą liczbą dokumentów mogą wymagać różnych serwerów, jeśli jedno obsługuje proste pliki tekstowe, a drugie duże skany PDF z OCR i wieloma równoległymi użytkownikami.

Rekomendacja bazowa

Dla większości małych i średnich wdrożeń produkcyjnych rekomendowanym punktem startowym jest serwer z 16-32 rdzeniami CPU, 128-256 GB RAM, kartą GPU 24-48 GB VRAM oraz szybkim dyskiem NVMe 2-8 TB. Dla większych organizacji zalecane jest rozdzielenie usług na osobne serwery lub maszyny wirtualne: aplikacja/API, baza wektorowa Qdrant, PostgreSQL, storage oraz inference GPU.

2. Co najbardziej wpływa na wymagania serwera

Czynnik	Wpływ na infrastrukturę	Komentarz
Liczba stron	Dysk, baza Qdrant, czas indeksacji	Lepsza miara niż sama liczba dokumentów. Jeden plik może mieć 2 strony albo 400 stron.
Udział skanów OCR	CPU/GPU, czas przetwarzania, kolejki zadań	Skanowane PDF-y i zdjęcia wymagają rozpoznawania tekstu, co jest znacznie cięższe niż zwykły PDF tekstowy.
Wybrany model LLM	GPU/VRAM, czas odpowiedzi	Większy model zwykle daje lepszą jakość, ale wymaga więcej VRAM i jest wolniejszy.
Równoległość użytkowników	GPU/VRAM, CPU, kolejki, limity LLM	To nie liczba kont, lecz liczba osób pytających AI w tym samym czasie wpływa na obciążenie.
Workflow i automatyzacje	CPU, RAM, I/O, storage	Import z e-maila, katalogów i automatyczna analiza dokumentów wymagają zapasu zasobów.
Retencja i backup	Dysk podstawowy i przestrzeń backupowa	Oryginalne pliki, baza SQL, Qdrant i logi powinny mieć regularne kopie zapasowe.

3. Założenia przeliczeniowe

Poniższe wartości są założeniami orientacyjnymi. Finalny sizing należy potwierdzić podczas POC na rzeczywistych dokumentach klienta.

Parametr	Wartość orientacyjna
Średnia liczba stron w jednym dokumencie	5-10 stron
Liczba fragmentów wiedzy na jedną stronę	3-5 chunków
Uproszczony wzór	liczba chunków = liczba stron x 4
Mały model LLM	7B-9B, szybki i ekonomiczny
Średni model LLM	14B-32B, lepsza jakość odpowiedzi
Duży model LLM	70B, wymagający dużej pamięci VRAM i osobnego podejścia

4. Zestawienie wariantów serwerowych

Wariant	Skala użycia	Dokumenty / strony	Równoległość AI	Rekomendowany profil
POC / Starter	Testy, demo, pierwszy dział	1 000-5 000 dokumentów / 5 000-25 000 stron	1 zapytanie AI	8-12 CPU, 64 GB RAM, GPU 16-24 GB, 1 TB NVMe
Business	Mała produkcja	5 000-20 000 dokumentów / 25 000-150 000 stron	1-2 zapytania AI	16 CPU, 128 GB RAM, GPU 24 GB, 2-4 TB NVMe
Professional	Kilka działów, workflow, OCR	20 000-100 000 dokumentów / 150 000-750 000 stron	2-5 zapytań AI	24-32 CPU, 256 GB RAM, GPU 48 GB lub 2 x 24 GB, 6-10 TB NVMe
Enterprise	Duża organizacja, wiele integracji	100 000+ dokumentów / 750 000+ stron	5-15 zapytań AI	48-64 CPU, 512 GB RAM, 2 x 48 GB lub 80 GB VRAM, 15-30 TB NVMe, architektura rozdzielona

Firma Informatyczna EnterSoft
e-mail: biuro@entersoft.pl, entersoft.pl, kom.: **696 481 842**

5. Wariant 1: POC / Starter

Wariant przeznaczony do prezentacji możliwości systemu, krótkiego pilotażu lub wdrożenia w jednym małym dziale. Pozwala zweryfikować jakość odpowiedzi, działanie OCR, obsługę cytowań oraz podstawowe scenariusze AI Fakt i AI Agent.

Obszar	Rekomendacja
Zakres dokumentów	1 000-5 000 dokumentów lub około 5 000-25 000 stron
Użytkownicy	3-10 użytkowników
Równoległe zapytania do AI	1 aktywne zapytanie generujące odpowiedź
CPU	8-12 rdzeni
RAM	64 GB
GPU	16-24 GB VRAM
Dysk	1 TB NVMe
Model LLM	7B-9B w wariancie zoptymalizowanym pod szybkość
Backup	Minimum 1-2 TB przestrzeni backupowej

- Dobry wariant do POC, demonstracji i zebrania pierwszych opinii użytkowników.
- Nie jest rekomendowany jako długoterminowa platforma dla wielu działów.
- Przy większym udziale OCR należy liczyć się z dłuższym czasem indeksacji dokumentów.

6. Wariant 2: Business

Wariant Business jest rekomendowany jako poziom produkcyjny dla małej organizacji, kancelarii, działu prawnego, księgowości, HR lub zespołu operacyjnego. Zapewnia większy komfort pracy i rezerwę zasobów na typowe przetwarzanie dokumentów.

Obszar	Rekomendacja
Zakres dokumentów	5 000-20 000 dokumentów lub około 25 000-150 000 stron
Użytkownicy	10-30 użytkowników
Równoległe zapytania do AI	1-2 aktywne zapytania generujące odpowiedź
CPU	16 rdzeni
RAM	128 GB
GPU	24 GB VRAM, np. NVIDIA L4 / RTX 4090 / karta klasy 24 GB
Dysk	2-4 TB NVMe
Model LLM	8B-14B, zależnie od oczekiwanej jakości i szybkości
Backup	4-8 TB przestrzeni backupowej

- Najlepszy wybór dla pierwszej produkcji w ograniczonym zakresie.
- Dobrze sprawdza się, gdy liczba równoległych użytkowników jest niewielka.
- Pozwala uruchomić typowe workflow, ale bez bardzo dużego bufora na masowe OCR.

Firma Informatyczna EnterSoft
e-mail: biuro@entersoft.pl, entersoft.pl, kom.: **696 481 842**

7. Wariant 3: Professional

Wariant Professional jest rekomendowany dla organizacji, które chcą używać AsWiseAI szerzej: w kilku działach, z większą bazą wiedzy, automatyzacjami, importem z e-maili lub katalogów oraz regularnym OCR dokumentów skanowanych.

Obszar	Rekomendacja
Zakres dokumentów	20 000-100 000 dokumentów lub około 150 000-750 000 stron
Użytkownicy	30-100 użytkowników
Równoległe zapytania do AI	2-5 aktywnych zapytań generujących odpowiedź
CPU	24-32 rdzenie
RAM	256 GB
GPU	48 GB VRAM albo 2 x 24 GB VRAM
Dysk	6-10 TB NVMe / enterprise SSD
Model LLM	14B-32B, przy zachowaniu kontroli równoległości
Backup	10-20 TB przestrzeni backupowej

- Rekomendowany wariant docelowy dla większości wdrożeń MŚP i średnich organizacji.
- Zapewnia większy komfort przy OCR, automatyzacjach i pracy wielu użytkowników.
- Dobry punkt startowy, jeśli klient od początku zakłada kilka działów lub większą bazę dokumentów.

8. Wariant 4: Enterprise

Wariant Enterprise jest przeznaczony dla dużych organizacji, wielu jednostek organizacyjnych, dużych zbiorów dokumentów, wielu widgetów WWW i scenariuszy integracyjnych. Przy tej skali rekomenduje się rozdzielenie komponentów systemu na osobne serwery lub maszyny wirtualne.

Obszar	Rekomendacja
Zakres dokumentów	100 000-500 000+ dokumentów lub 750 000-3 000 000+ stron
Użytkownicy	100-500 użytkowników
Równoległe zapytania do AI	5-15 aktywnych zapytań generujących odpowiedź
CPU	48-64 rdzenie lub więcej
RAM	512 GB lub więcej
GPU	2 x 48 GB VRAM lub akcelerator klasy 80 GB VRAM
Dysk	15-30 TB NVMe / enterprise SSD, zależnie od retencji
Model LLM	32B lub 70B w wariancie zoptymalizowanym, po testach POC
Backup	30-60 TB lub polityka backupowa zgodna z retencją klienta

Rekomendacja architektoniczna dla Enterprise

Przy dużej skali nie rekomenduje się uruchamiania wszystkiego na jednym serwerze bez planu rozwoju. Lepszym podejściem jest osobny serwer lub VM dla GPU inference, osobna baza Qdrant, osobny PostgreSQL, osobny storage oraz wydzielone workery do OCR i zadań automatycznych.

Firma Informatyczna EnterSoft

e-mail: biuro@entersoft.pl, entersoft.pl, kom.: **696 481 842**

EnterSoft

9. Macierz wyboru wariantu

Sytuacja klienta	Rekomendowany wariant	Uzasadnienie
Chcemy tylko sprawdzić, czy AI dobrze odpowiada na nasze dokumenty	POC / Starter	Najniższy koszt wejścia i wystarczające zasoby do walidacji jakości.
Jeden dział, kilkanaście osób, dokumenty głównie PDF tekstowe	Business	Dobry balans kosztu, jakości i szybkości odpowiedzi.
Kilka działów, OCR, integracja e-mail, workflow	Professional	Większy zapas CPU/RAM/GPU i dysku na zadania w tle.
Duża baza dokumentów i wymagania enterprise	Enterprise	Wymagana rozdzielona architektura, większa równoległość i lepsza odporność operacyjna.
Klient chce najlepszą jakość odpowiedzi i większe modele LLM	Professional / Enterprise	Większe modele wymagają większej pamięci VRAM i starannego zarządzania równoległością.

10. Rekomendacje dotyczące dysków i backupu

Dysk w systemie AsWiseAI pełni kilka ról: przechowuje oryginalne pliki, bazę relacyjną PostgreSQL, bazę wektorową Qdrant, logi, tymczasowe pliki OCR oraz kopie eksportów. Z tego powodu należy unikać konfiguracji z pojedynczym małym dyskiem systemowym.

Obszar	Rekomendacja
Dysk systemowy	Osobny wolumen 500 GB - 1 TB NVMe, najlepiej RAID1
Dane aplikacji	Osobny wolumen NVMe/SSD 2-30 TB zależnie od wariantu
Baza Qdrant	Preferowany szybki dysk NVMe; przy dużej skali osobny wolumen
PostgreSQL	Osobny wolumen lub przynajmniej wydzielony katalog danych
Backup	Minimum 2x przestrzeń aktywnych danych przy retencji kilku kopii
Snapshoty	Zalecane snapshoty przed aktualizacją i migracją danych

11. Rekomendacje dotyczące GPU i modeli AI

GPU dobieramy przede wszystkim pod wybrany model językowy i liczbę równoległych odpowiedzi. Liczba dokumentów nie oznacza automatycznie potrzeby większego GPU; większa baza dokumentów bardziej wpływa na Qdrant, storage i czas indeksacji. GPU odpowiada głównie za komfort generowania odpowiedzi przez LLM.

Klasa GPU	Typowe zastosowanie	Komentarz
16 GB VRAM	POC i małe modele	Dobre do testów, ale z ograniczonym zapasem na przyszłość.
24 GB VRAM	Mała produkcja	Rozsądny punkt startowy dla modeli 8B-14B i małej równoległości.
48 GB VRAM	Średnia produkcja	Lepszy wybór dla większych modeli, dłuższego kontekstu i większego komfortu.
80 GB VRAM lub więcej	Enterprise / duże modele	Wariant dla dużych modeli, większej równoległości i bardziej wymagających klientów.

Praktyczna zasada

Jeżeli klient oczekuje szybkich odpowiedzi, większej liczby równoległych użytkowników lub modeli powyżej 14B, warto przejść z GPU 24 GB na 48 GB VRAM. Jeżeli oczekuje modeli klasy 70B, sizing należy wykonać indywidualnie.

12. Rekomendacje sieciowe i bezpieczeństwa

- Dla środowisk produkcyjnych zalecane jest połączenie 1 GbE jako minimum, a 10 GbE przy dużych wolumenach dokumentów, backupach i osobnym storage.
- Serwer powinien działać w wydzielonej sieci lub VLAN z ograniczonym dostępem administracyjnym.
- Dostęp do panelu administracyjnego powinien być ograniczony do zaufanych adresów lub przez VPN.
- Dla integracji z Active Directory rekomendowane jest LDAPS oraz dedykowane konto serwisowe o minimalnych uprawnieniach.
- Dla widgetów WWW i API należy stosować limity zapytań, audyt użycia tokenów oraz cykliczną rotację tokenów serwisowych.
- Warto przewidzieć monitoring CPU, RAM, GPU, VRAM, temperatur, dysków, kolejek Celery oraz czasu odpowiedzi LLM.

13. Minimalny zestaw pytań przed finalnym sizingiem

Przed przygotowaniem finalnej oferty infrastrukturalnej należy zebrać poniższe informacje. Odpowiedzi pozwalają uniknąć niedoszacowania serwera.

1. Ile dokumentów ma zostać zaimportowanych na start?
2. Jaka jest orientacyjna liczba stron w tych dokumentach?
3. Jaki procent dokumentów to skany wymagające OCR?

4. Ilu użytkowników będzie miało konto w systemie?
5. Ilu użytkowników może jednocześnie zadawać pytania do AI?
6. Czy system ma obsługiwać automatyczny import z e-maila lub katalogów?
7. Czy mają działać widgety czatowe na stronach WWW?
8. Jakie są wymagania dotyczące retencji dokumentów i historii zapytań?
9. Czy klient wymaga integracji z Active Directory / LDAP?
10. Czy wymagane jest środowisko testowe obok produkcyjnego?

14. Zastrzeżenia i zakres odpowiedzialności

Na podstawie zakładanej skali wdrożenia rekomendujemy uruchomienie AsWiseAI na serwerze klasy produkcyjnej wyposażonym w szybkie dyski NVMe, odpowiedni zapas pamięci RAM oraz kartę GPU dobraną do lokalnej inferencji modeli językowych. Dla wdrożeń pilotażowych wystarczający będzie wariant POC/Starter, natomiast dla regularnej pracy kilku działów rekomendowany jest wariant Professional z 24-32 rdzeniami CPU, 256 GB RAM, kartą GPU 48 GB VRAM oraz przestrzenią dyskową 6-10 TB NVMe. Finalny dobór zasobów powinien zostać potwierdzony podczas etapu POC na rzeczywistych dokumentach klienta, szczególnie przy dużym udziale skanów OCR, automatyzacjach oraz większej liczbie równoległych użytkowników.

- Przedstawione konfiguracje mają charakter orientacyjny i służą do planowania budżetu oraz rozmów technicznych z klientem.
- Finalne wymagania mogą się zmienić po analizie rzeczywistego zbioru dokumentów, jakości skanów, oczekiwanego modelu LLM oraz liczby równoległych użytkowników.
- W przypadku środowisk krytycznych, wymagań HA lub wielu organizacji należy przygotować osobną architekturę produkcyjną.
- Wdrożenia z dużą liczbą dokumentów skanowanych powinny uwzględniać dodatkowy zapas CPU/GPU oraz harmonogram indeksacji poza godzinami pracy.

15. Rekomendacja końcowa

Najbardziej uniwersalnym wariantem startowym dla komercyjnej instalacji AsWiseAI jest konfiguracja Professional: 24-32 rdzenie CPU, 256 GB RAM, GPU 48 GB VRAM oraz 6-10 TB szybkiej przestrzeni NVMe. Taki zestaw daje rozsądny zapas na rozwój bazy wiedzy, obsługę OCR, automatyzacje i kilku równoległych użytkowników AI bez konieczności natychmiastowej rozbudowy infrastruktury.