cloudware

AI w praktyce: IBM Fusion na przykładzie wdrożenia

Maciej Osowski, Cloudware Polska
Common 2026, Wisła, 20 października 2026

Plan na 45 minut

1	Dwa zadania z wdrożenia w firmie ICT	5 min
2	Siedem problemów, które ma każdy RAG na dokumentach firmy	15 min
3	Dane, które zostają u klienta: IBM Fusion	10 min
4	Jak prowadzimy wdrożenie i mierzymy jakość	8 min
5	Co dalej: monitoring aplikacji i agent ankietowy	3 min
6	Pytania	4 min

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

Dwa zadania AI: sprzedaż i prowadzenie projektów

Uzupełnianie ankiet dla klientów

- Klienci przysyłają ankiety i kwestionariusze przy zapytaniach i ofertach.
- Odpowiedzi leżą w dokumentacji firmy i w poprzednich ankietach.
- Zespół wypełniał je ręcznie: dużo pytań, szczegóły techniczne, dane poufne.

Specjalistyczny RAG

- Zespoły projektowe szukają odpowiedzi w obszernej dokumentacji procesów: PDF, prezentacje, arkusze, diagramy.
- Dokumentacja się zmienia, a starych zapisów nie wolno usuwać.
- Odpowiedź ma wskazać źródło, żeby dało się ją sprawdzić.

Ankiety wspierają sprzedaż: szybsza odpowiedź klientowi to szybsza oferta. RAG wspiera prowadzenie projektów: odpowiedź ze źródłem bez przeszukiwania dokumentacji.

Klient i szczegóły zanonimizowane. Przykłady na kolejnych slajdach są wymyślone na potrzeby prezentacji.

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

WDROŻENIE

Ankieta: model przygotowuje odpowiedź, ekspert ją zatwierdza

Ankieta od klienta (arkusz z pytaniami) → **Pytania wyodrębnione z ankiety** → **Propozycja odpowiedzi + źródła + poziom pewności** → **Ekspert akceptuje albo poprawia**

Dwa źródła odpowiedzi:

Poprzednie ankiety
wzór stylu i precedens: jak firma odpowiadała wcześniej

Dokumenty firmy
PDF, DOCX, PPTX, skany JPG: stan faktyczny, z wersją i datą

Pytania z niską pewnością trafiają do przeglądu z oznaczeniem. Bez dostatecznie dobrego źródła model nie pisze odpowiedzi i oddaje pytanie człowiekowi.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

Siedem problemów, które ma każdy RAG na dokumentach firmy

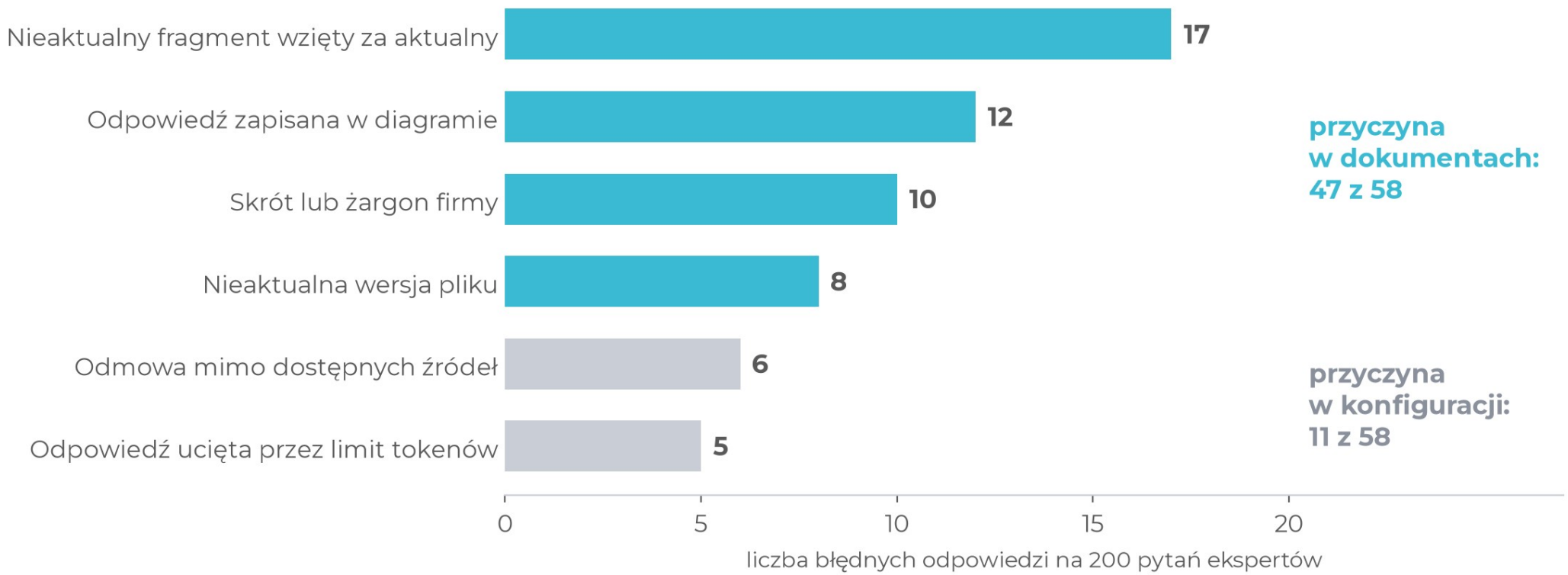
cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PIERWSZA WERSJA ASYSTENTA

Na 200 pytań ekspertów 58 odpowiedzi jest błędnych, 47 z nich przez dokumenty

Przyczyny błędnych odpowiedzi, liczba pytań



Dane ilustracyjne: liczby wymyślone na potrzeby prezentacji, proporcje zbliżone do obserwowanych we wdrożeniach RAG.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROBLEM 1 Z 7

Wycofane fragmenty zostają w dokumentach i wyglądają jak aktualne

W organizacji

Dokumenty mają przekreślone punkty, dopiski „nieaktualne” i odwołania do nowych zapisów. Obieg dokumentów nie pozwala ich usunąć ani wydać czystej wersji. Wyszukiwarka traktuje stary i nowy zapis jako równorzędne.

Przykład

Punkt 4.2: „Wniosek akceptuje kierownik działu”. Dopisek: „od 01.2025 nieaktualne, zob. 4.5”. Na pytanie, kto akceptuje wniosek, model cytuje 4.2.

Co robimy

- Asystent znajduje wycofane fragmenty i zapisy, które je zastąpiły
- Odpowiada według aktualnego zapisu, wycofany podaje z datą
- Dokumentacja klienta zostaje bez zmian

Przykład wymyślony na potrzeby prezentacji.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROBLEM 2 Z 7

Proces zapisany w diagramie ginie przy zamianie na tekst

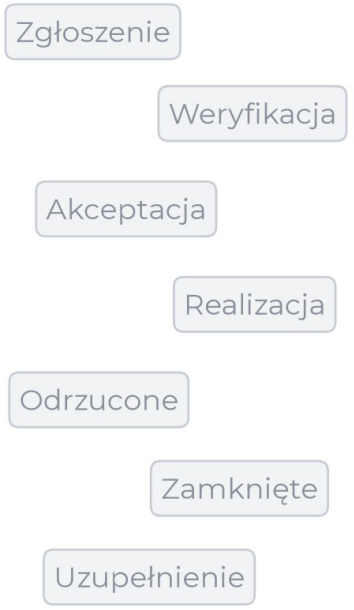
W organizacji

Diagramy w plikach Visio i jako obrazy (JPG, PNG) trafiają do indeksu przez OCR. Zostają nazwy etapów, giną strzałki i warunki przejść.

Co robimy

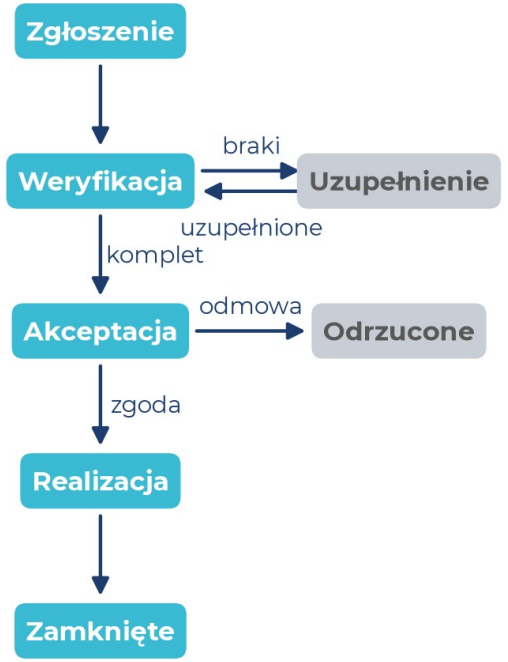
- Diagram czytamy jako graf: etapy, przejścia, legenda kolorów
- Graf dołączamy do kontekstu pytań o proces
- Pliki źródłowe zostają u klienta

Co widzi OCR



Nazwy etapów bez kolejności i bez warunków przejść

Co widzi asystent: graf



„Co po weryfikacji z brakami?"
→ Uzupełnienie, potem znów Weryfikacja

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROBLEM 3 Z 7

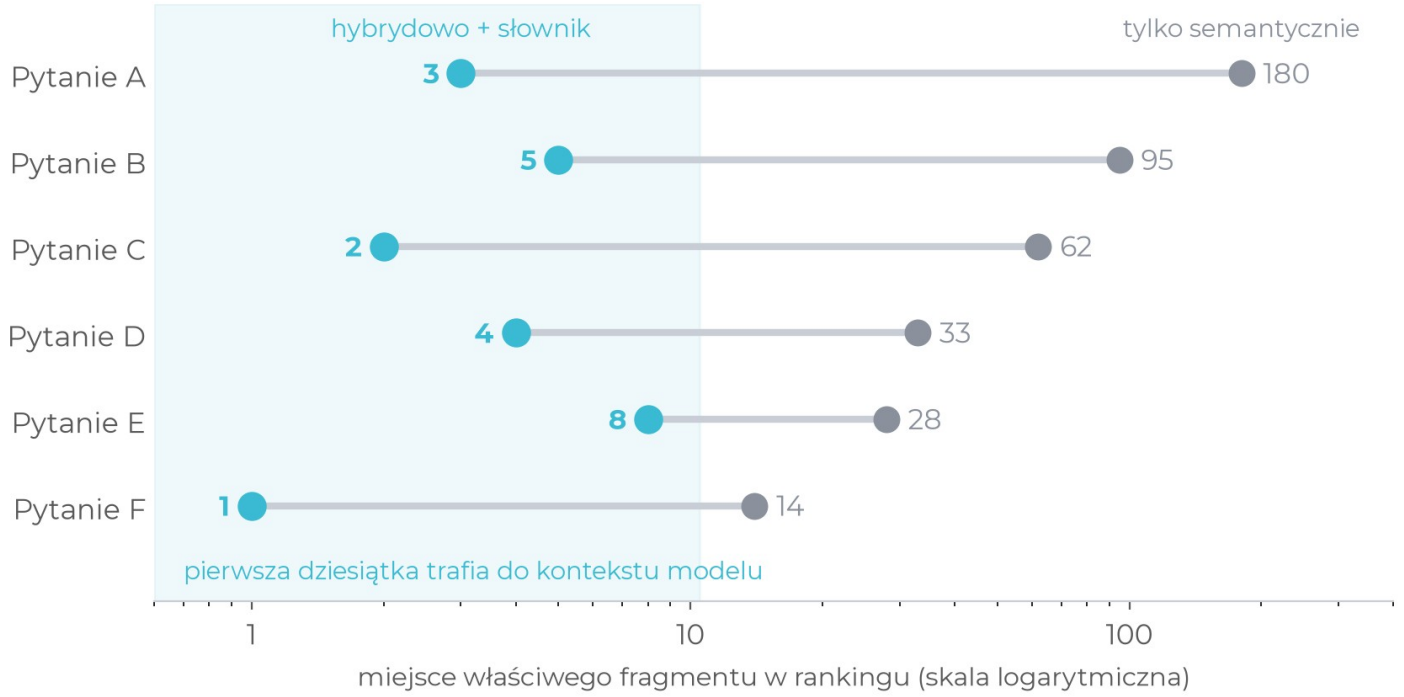
Firmowe skróty wypychają właściwy fragment poza kontekst modelu

W organizacji

Wyszukiwanie semantyczne słabo radzi sobie z wewnętrznymi skrótami i kodami. Właściwy fragment jest w indeksie, ale daleko w rankingu, więc model go nie dostaje.

Co robimy

- Wyszukiwanie hybrydowe: semantyczne i leksykalne (BM25)
- Słownik skrótów organizacji
- Pomiar pozycji właściwego fragmentu przed zmianą i po niej



Dane ilustracyjne: liczby wymyślone na potrzeby prezentacji, proporcje zbliżone do obserwowanych we wdrożeniach RAG.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROBLEM 4 Z 7

Ten sam temat opisuje kilka plików z różnych lat

W organizacji

Ten sam temat jest w kilku plikach: PDF, prezentacja, arkusz, skan JPG, każdy w kilku wersjach. Do tego dochodzą odpowiedzi udzielone kiedyś klientom, sprzed ostatniej zmiany.

Przykład

PDF z 2025 r. podaje czas obsługi zgłoszenia 4 godziny. Ankieta z 2023 r. podawała 8 godzin. Model widzi oba teksty.

Co robimy

- Kolejność ważności źródeł ustalona z klientem
- Przy sprzeczności obowiązuje nowszy dokument
- Aktualny dokument ma pierwszeństwo przed starą odpowiedzią z ankiety, która służy jako wzór stylu
- Każde źródło podane z nazwą pliku, wersją i datą

Przykład wymyślony na potrzeby prezentacji.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROBLEM 5 Z 7

To samo słowo znaczy w ankiecie co innego niż w dokumentach firmy

W organizacji

W ankiecie od klienta „dostawcą” jest firma, która odpowiada. W jej dokumentach wewnętrznych „dostawca” to firma zewnętrzna, którą nadzoruje.

Przykład

„Czy dostawca prowadzi rejestr zmian w usłudze?” Wyszukiwarka znajduje rozdział o wymaganiach wobec dostawców, a model podaje wymaganie stawiane innym jako opis własnej praktyki.

Co robimy

- Przed wyszukiwaniem model ustala, kogo dotyczy pytanie
- Pytanie przepisujemy na słownictwo dokumentów firmy
- Przy pytaniu o samą firmę pomijamy rozdziały o dostawcach zewnętrznych

Przykład wymyślony na potrzeby prezentacji.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

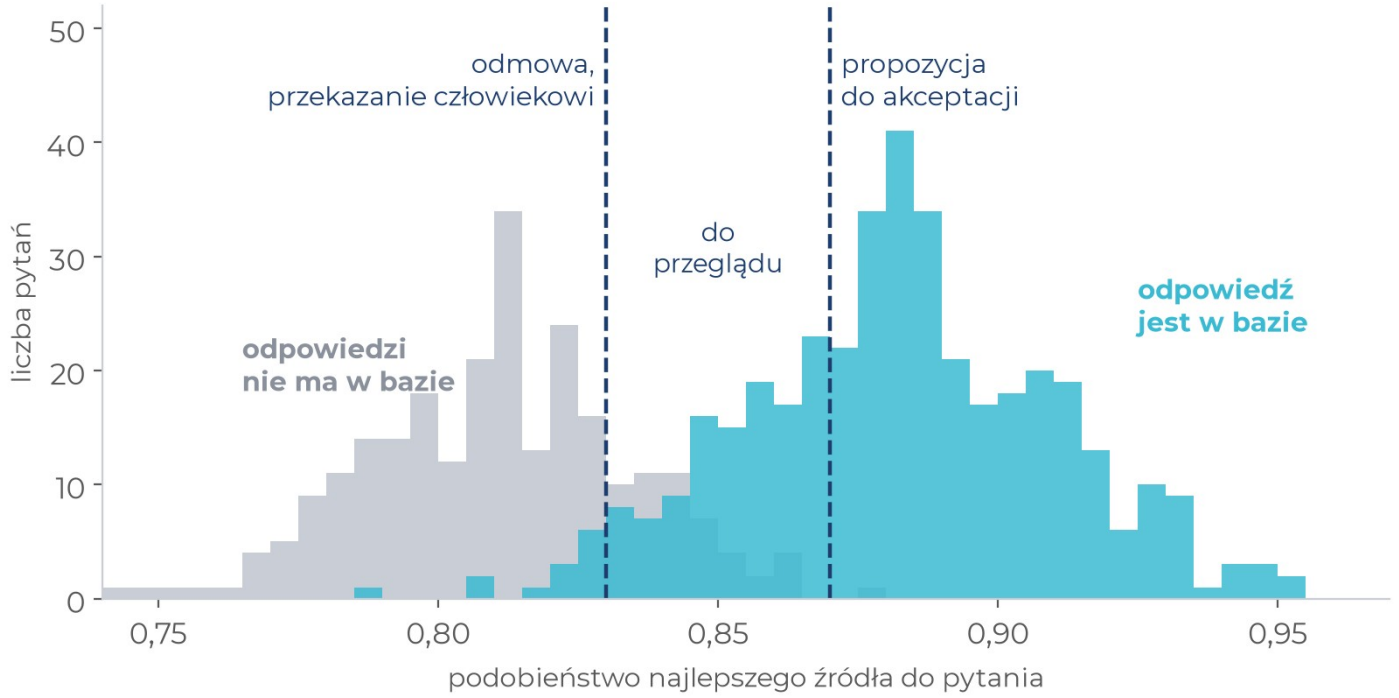
Asystent musi umieć nie odpowiedzieć

W organizacji

Model zawsze coś odpowie. W ankiecie wysłanej do klienta zmyślona odpowiedź kosztuje więcej niż pusta.

Co robimy

- Progi pewności skalibrowane na pytaniach klienta
- Poniżej progu: pytanie do człowieka z najbliższym materiałem
- Materiał z innego tematu: „brak materiału”
- Tajemnica przedsiębiorstwa: parafraza bez nazw systemów



Dane ilustracyjne: liczby wymyślone na potrzeby prezentacji, proporcje zbliżone do obserwowanych we wdrożeniach RAG.

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROBLEM 7 Z 7

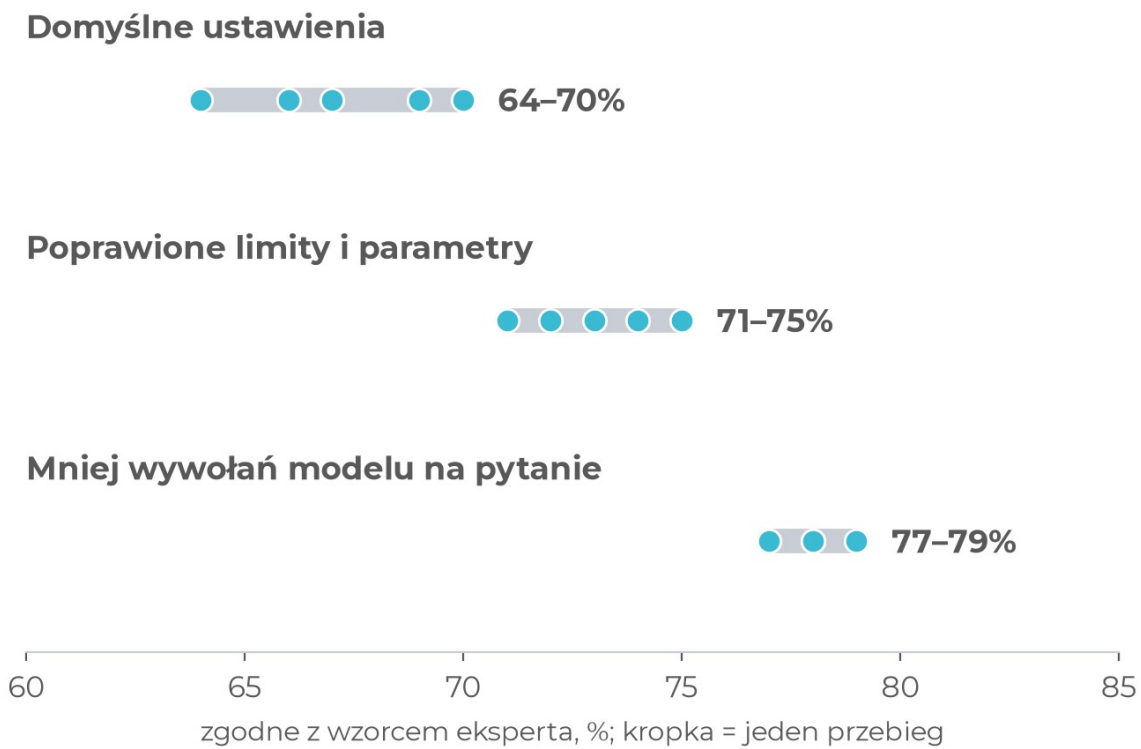
Jakości nie widać bez pomiaru, a ten sam model daje różne wyniki

W organizacji

Odpowiedzi brzmią poprawnie, a ekspert ocenia je jako błędne. Ten sam model przy temperaturze 0 odpowiada w kolejnych przebiegach inaczej. Domyślne ustawienia wywołań potrafią uciąć odpowiedź.

Co robimy

- Pytania ekspertów z odpowiedziami wzorcowymi
- Automatyczny oceniający (grader) i kilka przebiegów na każdą zmianę
- Ślad każdego kroku agenta: gdzie zginęła odpowiedź



Dane ilustracyjne: liczby wymyślone na potrzeby prezentacji, proporcje zbliżone do obserwowanych we wdrożeniach RAG.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

WNIOSEK

Każdy z tych problemów rozwiązuje się na danych konkretnej organizacji

Problem	Co trzeba ustawić u klienta
Wycofane fragmenty	sposób oznaczania zmian w jej dokumentach
Diagramy	legenda kolorów i kształtów
Skróty	słownik organizacji
Wersje plików	kolejność ważności źródeł i daty obowiązywania
Perspektywa	słownik ról w ankietach i dokumentach
Odmowa odpowiedzi	progi pewności na jej zbiorze testowym
Pomiar	pytania i odpowiedzi wzorcowe od jej ekspertów

Do tego asystent potrzebuje pełnego dostępu do dokumentów organizacji, także poufnych.

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

Dane, których organizacja nie wyśle do chmury

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PLATFORMA

Do asystenta trafia to, co firma chroni najbardziej

Dokumenty wewnętrzne (PDF, DOCX, skany)

Odpowiedzi udzielone klientom

Dokumenty z klauzulą tajemnicy przedsiębiorstwa

Dokumentacja projektów i procesów

Wymaganie

Model, indeks wektorowy i dokumenty pracują w infrastrukturze organizacji. Pytania i odpowiedzi nie opuszczają jej sieci.

Dla kogo

Podmioty objęte NIS2 i DORA, firmy z tajemnicą przedsiębiorstwa w dokumentacji technicznej, administracja publiczna.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PLATFORMA

IBM Fusion: usługi danych dla aplikacji na OpenShift

Wydajny storage dla aplikacji

Bloki, pliki i obiekty, skalowane razem z klastrem OpenShift.

Odporność

Wysoka dostępność, odtwarzanie po awarii, kopie zapasowe spójne ze stanem aplikacji.

Bezpieczeństwo

Szyfrowanie danych i ich przechowywania, integracja z cyberodpornością.

Mobilność aplikacji

Te same aplikacje od brzegu sieci, przez serwerownię, po chmurę.

Katalogowanie danych

Wgląd w dane w dowolnym miejscu organizacji, przed wpuszczeniem ich do AI.

Dla RAG

Dokumenty, indeks i model na tej samej platformie, w serwerowni klienta.

Źródło: IBM, prezentacja „IBM Fusion – wprowadzenie” (2026), udostępniona przez IBM Polska.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PLATFORMA

watsonx na Fusion HCI: AI na infrastrukturze klienta

watsonx.ai

Modele językowe i embeddingi, serwowanie usług AI. Tego używaliśmy we wdrożeniu.

watsonx.data

Baza wektorowa Milvus dla wyszukiwania w RAG; lakehouse dla danych z hurtowni i jezior danych. Tego używaliśmy.

watsonx.governance

Zarządzanie cyklem życia modeli i zgodnością. Dostępne na Fusion HCI, poza zakresem wdrożenia.

IBM Fusion HCI: OpenShift, storage i GPU pod wszystkimi trzema usługami, w jednym systemie

Źródło: IBM, prezentacja „IBM Fusion – wprowadzenie" (2026), udostępniona przez IBM Polska.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PLATFORMA

Fusion HCI: GPU, storage i sieć w jednym systemie

Węzły GPU

Do 7 w szafie; NVIDIA H200, RTX 6000 Pro albo AMD MI210.

Węzły obliczeniowe i storage

32 albo 64 rdzenie, 256–2048 GB RAM, trwały storage dla aplikacji i maszyn wirtualnych.

Sieć

200 Gb Ethernet dla storage i aplikacji, osobna sieć 1 Gb do zarządzania.

Skalowanie

Od 3 węzłów na start do 16 w szafie i kilku szaf w klastrze.

Jeden punkt wsparcia

IBM odpowiada za sprzęt, oprogramowanie Fusion i platformę OpenShift.

Opcjonalnie

Węzły IBM Storage AFM dla obciążeń z dużą liczbą operacji wejścia-wyjścia.

Źródło: IBM, prezentacja „IBM Fusion – wprowadzenie” (2026), udostępniona przez IBM Polska. Konfiguracje zależą od wersji i producenta serwerów.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PLATFORMA

Fusion CAS: przygotowanie danych do RAG po stronie storage

Co robi

Przetwarza nieustrukturyzowane dane do wyszukiwania w aplikacjach RAG, na GPU, z NVIDIA NeMo Retriever.

Skąd bierze dane

Z istniejących zasobów plikowych i obiektowych (SAN, NFS, S3) przez cache IBM Storage Scale.

Konfiguracja

Dwa węzły GPU (do 8× RTX 6000 Pro albo H200) i węzeł obliczeniowy.

Kierunek rozwoju platformy według IBM. Wycofania, diagramy, wersje i perspektywa i tak wymagają pracy na danych klienta.

Źródło: IBM, prezentacja „IBM Fusion – wprowadzenie” (2026), udostępniona przez IBM Polska.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PLATFORMA

Układ we wdrożeniu: wszystkie warstwy w sieci klienta

- Modele językowe i embeddingi na GPU platformy klienta
- Baza wektorowa Milvus w watsonx.data, na tym samym klastrze
- Usługa AI wdrożona z notatnika jako serwis watsonx.ai, wołana przez API aplikacji
- Nowa wersja asystenta to nowe wdrożenie; aplikacja przełącza się bez zmian w infrastrukturze
- Ten sam klaster obsłuży kolejne zadania: monitoring aplikacji i agentów



Wszystkie warstwy w sieci klienta

Schemat uproszczony.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

Platforma zapewnia bezpieczeństwo i moc obliczeniową, wdrożenie odpowiada za jakość

Platforma (IBM Fusion + watsonx.ai)

- Dane i modele w infrastrukturze klienta
- GPU i serwowanie modeli
- Skalowanie, kopie zapasowe, odtwarzanie
- Zarządzanie cyklem życia modeli

Wdrożenie (Cloudware)

- Przygotowanie dokumentów: wycofania, diagramy, wersje
- Wyszukiwanie dopasowane do języka firmy
- Reguły odpowiedzi i progi pewności
- Pomiar jakości na pytaniach ekspertów

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

Jak prowadzimy wdrożenie

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROCES

Od warsztatu do wdrożenia: sześć kroków



gdzie AI odciąży ludzi

wolumen, poufność, dane, mierzalność

z odpowiedziami wzorcowymi

na danych klienta, w jego infrastrukturze

kilka przebiegów na każdą zmianę

testy użytkowników, ślad odpowiedzi

Kryteria wyboru pierwszego zadania

Dużo powtarzalnej pracy na dokumentach; dane, których nie można wysłać na zewnątrz; dokumenty dostępne w formie elektronicznej; ekspert, który oceni odpowiedzi i poda wzorce.

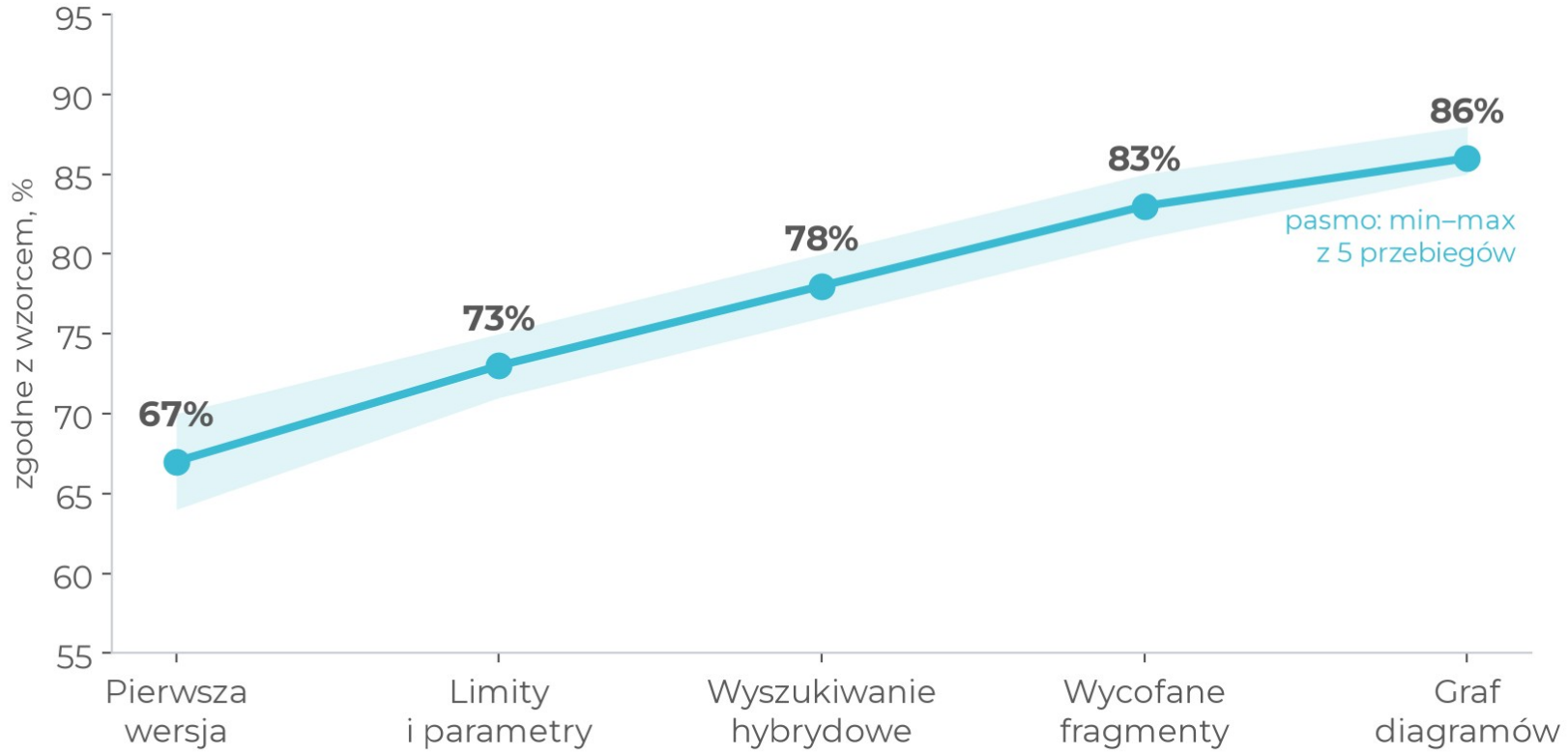
cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

PROCES

Każda poprawka ma zmierzony wynik

Odpowiedzi zgodne z wzorcem po kolejnych zmianach, %; pasmo z pięciu przebiegów



Dane ilustracyjne: liczby wymyślone na potrzeby prezentacji, proporcje zbliżone do obserwowanych we wdrożeniach RAG.

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

CO DALEJ

W planach: AI w monitoringu aplikacji i agent ankietowy

Monitoring aplikacji (APM)

- IBM Instana zbiera metryki, logi i ślady aplikacji
- Model na tej samej platformie opisuje anomalie i wskazuje prawdopodobną przyczynę
- Dane z monitoringu zostają w sieci klienta

Agent ankietowy

- Wypełnia całą ankietę i oznacza pytania bez źródła
- Pytania bez odpowiedzi wysyła do właścicieli procesów
- Zatwierdzone odpowiedzi wracają do bazy

cloudware

Copyright © 2026 Cloudware. Wszelkie prawa zastrzeżone.

cloudware

Jakie dokumenty u was blokują AI?

Maciej Osowski, Cloudware Polska
maciej.osowski@cloudware.pl

Materiały z sesji i kontakt