O'REILLY®

Dane gotowe na AI

Praktyczne wzorce projektowania nowoczesnej architektury danych



Navnit Shukla
Kien Pham
Srikanth Sopirala
Harsha Tadiparthi

Helion

Tytuł oryginału: AI-Ready Data Blueprints: From Raw Data to AI-Driven Innovation

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-289-4259-2

© 2027 Helion S.A.

Authorized Polish translation of the English edition of *AI-Ready Data Blueprints*
ISBN 9798341631793 © 2026 Navnit Kumar Shukla, AZ25 Lab, Harsha Tadiparthi,
and Srikanth Sopirala.

This translation is published and sold by permission of O'Reilly Media, Inc.,
which owns or controls all rights to publish and sell the same.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiejkolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz wydawca dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz wydawca nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
helion.pl/user/opinie/dagoai
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

- Kup książkę
- Poleć książkę
- Oceń książkę

- Księgarnia internetowa
- Lubię to! » Nasza społeczność

Spis treści

Przedmowa **9**

Wprowadzenie **13**

1. Wprowadzenie do fundamentów danych przystosowanych do AI **19**

Wprowadzenie i kontekst rynkowy 20
Czym wyróżnia się generatywna AI? 22
Architektura transformera: techniczne podstawy generatywnej AI 23
Dane przedsiębiorstwa jako główny czynnik różnicujący 24
Inteligencja kontekstowa: prosty przykład 25
Reprezentacja znaczenia w przestrzeni wektorowej 26
Przykład praktyczny z organizacji: zrozumienie wielu dokumentów w dziale obsługi klienta 27
Ewolucja zastosowań generatywnej AI 28
Od asystentów do agentów: cztery etapy ewolucji generatywnej AI 30
Rosnąca złożoność wymagań dotyczących danych 32
Jakie rozwiązania są obecnie rozwijane w wiodących organizacjach? 33
Weryfikacja w środowisku produkcyjnym 33
Nowe wzorce architektoniczne dla generatywnej AI 34
Od wzorców do praktyki 40
Cykle rozwoju rozwiązań: uczenie maszynowe a generatywna AI 40
Tradycyjny cykl rozwoju rozwiązań z obszaru uczenia maszynowego 40
Cykl rozwoju rozwiązań opartych na generatywnej AI 43
Implikacje dla infrastruktury danych 46
Ewolucja architektury w praktyce: od tradycyjnych procesów ETL do potoków przystosowanych do generatywnej AI 47
Najważniejsze różnice między podstawową warstwą danych w tradycyjnych rozwiązaniach i w generatywnej AI 48
Przykład praktyczny: ewolucja od modelu Kimballa do architektury medalionowej i generatywnej AI 49

Kup książkę Poleć książkę

Przygotowanie się na przyszłość z wiodącą rolą agentów 55
„Jeeves robi zakupy" — doświadczenia użytkowników korzystających z agentów 55
SEO: od słów kluczowych do optymalizacji pod kątem agentów 56
Agenty nie są lojalne wobec marek — przygotowania do ery radykalnej przejrzystości 56
Autopilot w działalności biznesowej — autonomiczne operacje i podejmowanie decyzji 57
Lista kontrolna do oceny dostosowania danych do przyszłości opartej na agentach 58
Plan wdrożenia: wskazówki dotyczące implementacji 59
Od czego zacząć? 59
Drzewo decyzyjne: który wzorzec wybrać jako pierwszy? 60
Plan działania 61
Podsumowanie 62

2. Architektura danych dla aplikacji generatywnej i agentowej AI 63
Wstęp: budowa fundamentów dla danych przystosowanych do AI 63
Ewolucja architektur danych 64
Okres początkowy (2015 – 2019) 65
Okres przejściowy (2020 – 2022) 66
Era generatywnej AI (od 2022 do dziś) 66
Era agentowej AI (od 2024 roku) 67
Potrzeba nowego podejścia: podstawowe wymagania dotyczące danych przystosowanych do AI 68
Uchwytywanie logiki biznesowej i kontekstu 69
Zapewnianie jakości i spójności danych 70
Zarządzanie złożonością i różnorodnością 70
Utrzymywanie bezpieczeństwa, zgodności z regulacjami i prywatności 70
Umożliwianie udostępniania informacji i współpracy 71
Ułatwianie skalowalności i zapewnianie wydajności 71
Zarządzanie danymi jako produktem strategicznym 71
Zapewnianie użytkownikom dokumentacji i wskazówek 72
Podstawowa architektura danych przystosowanych do AI 72
Uchwytywanie logiki biznesowej i kontekstu 72
Zapewnianie jakości i spójności danych 77
Zarządzanie złożonością i różnorodnością 81
Utrzymywanie bezpieczeństwa, zgodności z regulacjami i prywatności 85
Umożliwianie udostępniania informacji i współpracy 89
Ułatwianie skalowalności i zapewnianie wydajności 92
Zarządzanie danymi jako produktem strategicznym 95
Zapewnianie użytkownikom dokumentacji i wskazówek 99

Kup książkę Poleć książkę

Wzorce architektury danych przystosowanych do AI: praktyczny przewodnik wdrażania 102
Wzorzec nr 1. Silnik analizy kontekstu biznesowego 102
Wzorzec nr 2. Adaptacyjne zarządzanie jakością danych 104
Wzorzec nr 3. Zarządzanie różnorodnością i złożonością danych 105
Wzorzec nr 4. Platforma danych dla AI zorientowana na bezpieczeństwo 106
Podsumowanie 109

3. Obróbka i przygotowywanie danych na potrzeby zastosowań w generatywnej i agentowej AI 110
Problem w przedsiębiorstwie 110
Cel i odbiorcy 110
Imperatyw rynkowy 111
Przewaga strategiczna 112
Przegląd procesu transformacji 112
Na czym polega zmiana paradygmatu? 113
Tradycyjny potok danych w uczeniu maszynowym 113
Paradygmat danych tabelarycznych i jego ograniczenia 114
Rewolucja w sposobie myślenia w przetwarzaniu danych 115
Nowe paradygmaty przetwarzania i wymagania techniczne 116
Wpływ zmiany paradygmatu na działalność biznesową 117
Formularz samooceny: obecny poziom organizacji 118
Elementy składowe danych przystosowanych do generatywnej AI 119
Podstawy rozumienia semantycznego 120
Grafy wiedzy i modelowanie relacji 122
Wymagania dotyczące przetwarzania danych w czasie rzeczywistym 123
Model dojrzałości przygotowywania danych na potrzeby generatywnej AI 124
Studium przypadku: transformacja firmy handlowej 125
Architektura warstwy semantycznej 126
Przegląd architektury i główne zasady 126
Podstawowa warstwa danych: wykorzystanie architektury typu lakehouse 128
Zarządzanie metadanymi i ontologią: umożliwianie zrozumienia semantycznego 129
Potok transformacji i wzbogacania: dodawanie inteligencji do danych 130
Infrastruktura wektoryzacji i indeksowania: umożliwianie wyszukiwania semantycznego 131
API i warstwa wnioskowania: umożliwianie systemom AI korzystania z wiedzy 132
Interakcje między komponentami i przepływy danych 133
Kwestie dotyczące implementacji i popularne wzorce 134
Architektura referencyjna i implementacja w AWS 136
Przegląd ekosystemu AWS w obszarze przygotowywania danych na potrzeby generatywnej AI 136
Najważniejsze usługi i funkcje AWS 138
Wzorce integracji i wybór usług 140

Kup książkę Poleć książkę

Ład, jakość i obserwowalność 141
Rozszerzony zestaw wymiarów jakości dla danych do obsługi generatywnej AI 141
Podstawy modelu ram ładu 143
Ramy ładu i monitorowania jakości 145
Podejścia do zapewniania monitorowania i obserwowalności 148
Zapewnianie i weryfikowanie zgodności z regulacjami 148
Studium przypadku: wdrożenie ładu w firmie z branży usług finansowych 149
Zaawansowane zagadnienia: wymagania dotyczące danych dla agentowej AI 150
Podstawy agentowej AI i jej wpływ na dane 150
Przykłady implementacji w przedsiębiorstwach 152
Przyszłość: autonomiczna obróbka danych wspomagana przez AI 153
Odkrywanie danych i inteligentna klasyfikacja z wykorzystaniem AI 153
Przygotowywanie danych konwersacyjnych i interfejsy oparte na języku naturalnym 154
Autonomiczne zapewnianie jakości i systemy samokorygujące się 154
Przyszłe kierunki rozwoju i najnowsze obszary badań 154
Plan działania 155
Najważniejsze zasady 155
Bezpośrednie kolejne kroki na różnych poziomach dojrzałości 156
Długoterminowe kwestie strategiczne 157
Podsumowanie 157

4. Ład danych, bezpieczeństwo, zgodność z regulacjami i orkiestracja w generatywnej AI 158
Ład i bezpieczeństwo danych w aplikacjach AI 159
Model operacyjny dotyczący ładu danych w obszarze AI 160
Różnice w ładzie dla danych ustrukturyzowanych i nieustrukturyzowanych 162
Opieka nad danymi i zarządzanie metadanymi 163
Udostępnianie danych dla ludzi i agentowej AI 163
Testowanie narzędzi 166
Jakość danych 167
Odpowiedzialna sztuczna inteligencja 169
Sprawiedliwość 170
Przejrzystość 171
Rozliczalność 175
Prywatność i bezpieczeństwo 176
Niezawodność 178
Nadzór ze strony człowieka 178
Prywatność i bezpieczeństwo danych w aplikacjach generatywnej AI 179
Ochrona informacji wrażliwych 180
Ograniczanie tematów i filtrowanie słów 182
Narzędzia do filtrowania treści 184
Ochrona danych w całym potoku ich przetwarzania 187

Kup książkę Poleć książkę

LLMOps: orkiestracja przepływów pracy AI 188
Wzorce orkiestracji dla aplikacji AI 188
Wzorce dla agentowej AI 192
Podsumowanie 196

5. Bazy wiedzy i wektorowe bazy danych 197
Wyzwania związane z danymi dla generatywnej AI 197
Bazy wiedzy: organizacja i przechowywanie danych 198
Bazy wektorowe: reprezentacja danych i wyszukiwanie na podstawie podobieństwa 198
Metoda RAG: wyszukiwanie danych i łączenie kontekstu 199
Dlaczego wymienione technologie są niezbędne? 199
Podstawy bazy wiedzy: od danych do wiedzy 200
Architektura danych dla baz wiedzy 200
Potok przygotowywania danych dla baz wiedzy 204
Typy danych i strategie zarządzania nimi 205
Jakość danych: czynnik decydujący o sukcesie lub porażce 209
Podstawy wektorowych baz danych: reprezentacja danych 211
Osadzenia wektorowe i jakość danych 211
Infrastruktura wektorowych baz danych 213
Narzędzia otwartoźródłowe 214
Porównanie algorytmów indeksowania danych 215
Techniki kompresji danych 216
Porównanie funkcji z obszaru ładu danych 217
Wybór na podstawie wymagań dotyczących danych 217
RAG: warstwa wyszukiwania danych 219
Po co stosować metodę RAG? Problem dostępu do danych 220
Architektura RAG: przepływ danych 220
Podział danych na fragmenty jest krytyczny dla skuteczności systemów RAG 223
Efektywność wykorzystania danych i optymalizacja kosztów 231
Przepływ danych od początku do końca: baza wiedzy → wektorowa baza danych → RAG 237
Podsumowanie 241

6. Optymalizacja aplikacji AI pod kątem gotowości produkcyjnej 243
Droga od prototypu do środowiska produkcyjnego 243
Przygotowanie aplikacji AI do udanego działania w środowisku produkcyjnym 245
Dwa filary optymalizacji aplikacji AI 245
Inżynieria wydajnych przepływów danych dla produkcyjnych systemów AI 246
Ocena jakości danych i podstawy wstępnego przetwarzania 246
Zrozumienie okien kontekstowych i implikacji ich stosowania 249

Kup książkę Poleć książkę

Obsługa dużych kontekstów wejściowych 250
Optymalizacja jakości wnioskowania z wykorzystaniem automatycznego rozumowania 251
Gotowość metadanych dla systemów agentowej AI 257
Wyzwania związane ze skalą i złożonością 257
Wzorzec: bezpośrednie wykorzystanie plików w formatach natywnych 259
Budowanie agentów świadomych danych dzięki inteligentnym metadanym semantycznym 260
Przykładowa implementacja: bezpośrednie wykorzystanie surowych plików 263
Od surowych danych do inteligencji semantycznej: budowa warstwy metadanych 267
Najważniejsze możliwości platform agentowej AI gotowych do wdrożenia w środowisku produkcyjnym 271
Wiodące platformy agentowej AI 273
Podstawowe środowisko uruchomieniowe i strategia wdrażania 273
Frameworki, integracja narzędzi i obsługa protokołów 273
Możliwości w zakresie zarządzania pamięcią i wiedzą 274
Możliwości w zakresie bezpieczeństwa i zgodności z regulacjami 275
Wsparcie w zakresie obserwowalności i zapewniania jakości 276
Podsumowanie 276

Kup książkę Poleć książkę

ROZDZIAŁ 1.

Wprowadzenie do fundamentów danych przystosowanych do AI

W tym rozdziale omawiamy szybki rozwój technologii z obszaru generatywnej AI i krytyczną infrastrukturę danych niezbędną do ich pomyślnego wdrażania. Bezprecedensowe wprowadzanie modeli bazowych odbywa się zdecydowanie szybciej, niż organizacje uczą się odpowiednio przygotowywać dane. W efekcie powstaje poważna luka między eksperymentami a wdrażaniem rozwiązań w środowisku produkcyjnym.

Zbadamy, jak ta luka objawia się w praktyce. Według ostatnich badań przeprowadzonych przez firmę McKinsey (*https://oreil.ly/EQyIR*) 79 procent organizacji regularnie korzysta z generatywnej AI w co najmniej jednej funkcji biznesowej, ale tylko 7 procent w pełni przeskalowało wykorzystanie AI w środowisku produkcyjnym. Powszechnie wiadomo, że główną przyczyną tej rozbieżności nie jest wybór modelu, lecz nieodpowiednie przygotowanie danych. Tradycyjne architektury danych zoptymalizowane pod kątem analityki i uczenia maszynowego nie są w stanie zapewnić zrozumienia semantycznego, kontekstu w czasie rzeczywistym i wnioskowania międzydomenowego, których wymagają aplikacje generatywnej AI.

W miarę jak generatywna AI ewoluuje od prostych asystentów do autonomicznych agentów, wymagania dotyczące infrastruktury danych stają się coraz bardziej złożone. Każdy etap ewolucji, od podstawowych asystentów AI, przez asystenty typu copilot i agenty RAG, aż po agentową sztuczną inteligencję, wiąże się z fundamentalnie odmiennymi wymaganiami z zakresu strukturyzowania danych, dostępu do nich i nadzoru nad nimi.

W tym rozdziale zidentyfikowaliśmy pięć wzorców architektonicznych, które konsekwentnie pojawiają się w organizacjach potrafiących skutecznie pokonać przestrzeń między eksperymentami a środowiskiem produkcyjnym: grafy wiedzy dla inteligencji kontekstowej, architektura sterowana zdarzeniami do obsługi AI w czasie rzeczywistym, platformy typu lakehouse dla ujednoliconych danych, wyszukiwanie semantyczne dla wyszukiwania opartego na znaczeniu oraz właściwości przystosowane do agentów na potrzeby autonomicznego dostępu. Za pomocą praktycznych przykładów, w tym zmian wprowadzonych przez globalnego sprzedawcę detalicznego, u którego początkowo używane były tradycyjne modele wymiarowe, pokazujemy, jak te wzorce współdziałają ze sobą i umożliwiają osiąganie sukcesów w środowisku produkcyjnym.

Kup książkę Poleć książkę

Wprowadzenie i kontekst rynkowy

W dniu 30 listopada 2022 roku firma OpenAI udostępniła ChatGPT 3.5 jako „wersję zapoznawczą do celów badawczych" (*https://oreil.ly/F4lBD*) w celu zebrania opinii oraz zrozumienia jego możliwości i ograniczeń.

Jednak premiera spotkała się z ogromnym zainteresowaniem, a nowy model szybko stał się fenomenem. Pojawiła się technologiczna siła, która w krótkim czasie miała zmienić świat w sposób, jaki dopiero zaczynamy rozumieć. W ciągu dwóch miesięcy liczba użytkowników ChatGPT urosła do ponad 100 milionów (*https://oreil.ly/ZYwwZ*). Żadna inna aplikacja konsumencka w historii nie zyskała takiej popularności w tak krótkim czasie. W zaledwie 60 dni ChatGPT osiągnął bazę użytkowników, na której zbudowanie Instagram potrzebował dwóch i pół roku, a TikTok dziewięciu miesięcy.

Wcześniejsze przełomowe wynalazki technologiczne, takie jak silnik spalinowy w latach 60. XIX wieku, internet w latach 90. XX wieku, smartfony w 2007 roku i chmura obliczeniowa na początku drugiej dekady XXI wieku, były wdrażane stopniowo. Jednak możliwości generatywnej AI niemal natychmiast stały się dostępne prawie dla każdego.

Generatywna AI oferuje każdemu niezwykle rozbudowane możliwości w zakresie tworzenia treści. Często wystarczy wpisać prosty prompt w języku naturalnym lub nawet wydać polecenie głosowe. Obecnie każdy może generować treści w swoim obszarze zainteresowań lub wiedzy specjalistycznej niezależnie od tego, czy chodzi o tworzenie obrazów, pisanie skryptów, produkcję filmów, czy rozwój kodu. Ten nowy poziom dostępności doprowadził do rewolucji w produktywności. Ludzie niezależnie od posiadanej wiedzy merytorycznej mogą bezproblemowe tworzyć wszelkiego rodzaju treści.

Pracownicy umysłowi w różnych branżach, w tym programiści, twórcy treści, analitycy, badacze, konsultanci i inni specjaliści, którzy pracują głównie z informacjami, doświadczyli radykalnego przyspieszenia procesów pracy. Zadania, które kiedyś zajmowały całe dni, są obecnie wykonywane w zaledwie kilka godzin lub nawet minut. Ta płynna integracja generatywnej AI jest możliwa dzięki intuicyjnym interfejsom opartym na języku naturalnym, które (w przeciwieństwie do wcześniejszego oprogramowania dla przedsiębiorstw, z koniecznym czasochłonnym wdrażaniem nowych pracowników) nie wymagają specjalistycznego szkolenia. Prowadzi to do bezprecedensowej fali wprowadzania narzędzi AI zarówno w środowiskach konsumenckich, jak i korporacyjnych.

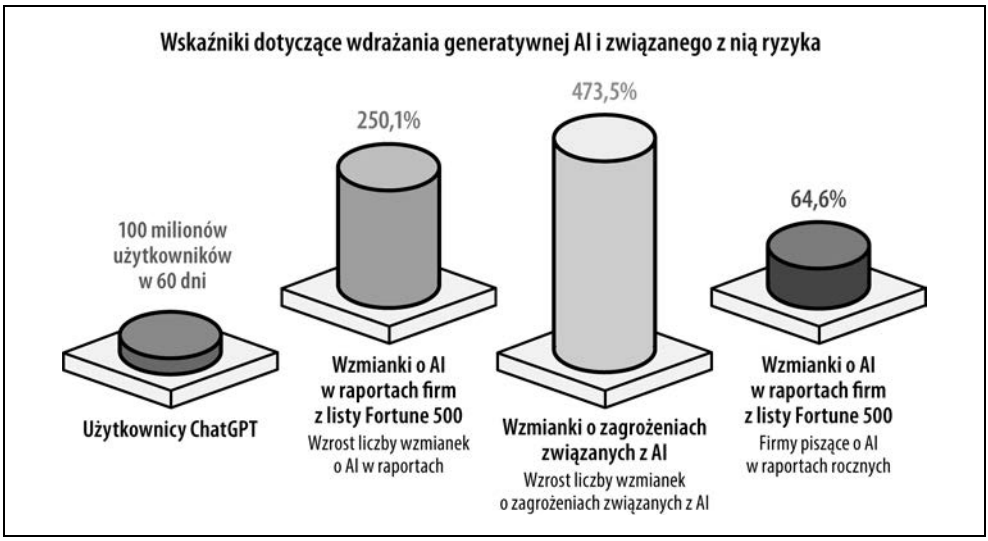
Na podstawie obserwacji transformacyjnego potencjału generatywnej AI organizacje szybko weszły w tryb, który analitycy branżowi nazwali *panicznym wdrażaniem AI*. Skutkuje to próbami pośpiesznego wdrażania generatywnej AI bardziej z powodu presji konkurencji niż na podstawie strategicznych planów. Kadra kierownicza wyższego szczebla w różnych sektorach domagała się natychmiastowej realizacji inicjatyw w zakresie generatywnej AI, co zmuszało zespoły do realizacji projektów PoC (ang. *proof of concept*) wkrótce po debiucie ChatGPT w listopadzie 2022 roku.

Analiza z 2024 roku (*https://oreil.ly/ERa_8*) przeprowadzona przez Arize AI na podstawie danych z dokumentów składanych w amerykańskiej Komisji Papierów Wartościowych i Giełd (SEC)

Kup książkę Poleć książkę

wykazała, że 64,6 procent firm z listy Fortune 500 wspomniało o sztucznej inteligencji w swoich najnowszych raportach rocznych. Stanowiło to wzrost o 250,1 procent względem roku 2022.

Ponad jedna piąta firm bezpośrednio pisze o generatywnej AI, a ponad połowa (281) wymienia ją jako czynnik ryzyka (w porównaniu z zaledwie 49 w 2022 roku). Rysunek 1.1 obrazuje ten szybki wzrost popularności generatywnej AI i jej świadomości.



Rysunek 1.1. Statystyki wdrażania generatywnej AI, w tym bezprecedensowe wskaźniki wzrostu liczby użytkowników ChatGPT i obserwacje Arize AI dotyczące trendów we wzmiankach o sztucznej inteligencji w raportach rocznych firm z listy Fortune 500 w latach 2022 – 2024

W trakcie rozmów w salach konferencyjnych nagle zaczęto używać terminologii technicznej, która wcześniej pojawiała się wyłącznie w artykułach naukowych na temat AI. Terminy takie jak „model bazowy", „baza wiedzy", „inżynieria promptów", „architektura RAG", „wektorowa baza danych", „agenty", „konfabulacje" (inaczej „halucynacje") i „agentowa sztuczna inteligencja" niemal z dnia na dzień stały się elementami słownictwa stosowanego przez kadrę kierowniczą.

Ten korporacyjny entuzjazm (lub FOMO — ang. *fear of missing out*, czyli strach przed tym, co nas omija) jeszcze się nasilił, gdy na rynek w szybkim tempie wchodziły kolejne modele bazowe. W ciągu sześciu miesięcy po premierze ChatGPT kilka konkurencyjnych firm udostępniło własne rozwiązania. Oto kilka z nich:

- Claude firmy Anthropic,
- Bard firmy Google,
- Llama firmy Meta,
- Titan i Bedrock firmy Amazon,
- Tongyi Qianwen firmy Alibaba,
- modele Diffusion firmy Stability.ai.

Kup książkę Poleć książkę

Pojawianie się kolejnych modeli dodatkowo podsycało w organizacjach poczucie pilności, by wprowadzać technologie generatywnej AI i nie zostać w tyle za konkurencją. Firmy aktywnie poszukiwały wskazówek dotyczących dostępnych opcji, aby przyspieszyć rozwój projektów PoC, a szybki wzrost popularności rozwiązań AI wśród konsumentów przekładał się na bezprecedensowe wdrożenia w przedsiębiorstwach.

Grupa finansowa JPMorgan Chase (*https://oreil.ly/2Xejh*) wdrożyła generatywną AI dla 200 000 pracowników na potrzeby globalnej działalności operacyjnej. Stanowi to jedno z największych wdrożeń AI w przedsiębiorstwach w historii usług finansowych. Podobnie firma Siemens (*https://oreil.ly/jVi1N*) zintegrowała usługę Amazon Bedrock ze swoją platformą Mendix, obsługującą obecnie przeszło 50 milionów użytkowników korzystających z ponad 200 000 aplikacji. Szybkość osiąganych postępów jest równie niezwykła. Na przykład BT Group (*https://oreil.ly/WAdfn*) wdrożyła asystenta Amazon CodeWhisperer (obecnie zintegrowanego z asystentem Q Developer) i udostępniła go 1200 inżynierom. Pozwoliło to zautomatyzować 12 procent ich pracy i wygenerować ponad 100 000 wierszy kodu w zaledwie cztery miesiące.

Jednak z tym szalonym tempem wprowadzania rozwiązań AI wiąże się poważne wyzwanie. *Otóż większość organizacji odkrywa, że ich obecna infrastruktura danych nie jest w stanie obsłużyć produkcyjnych aplikacji generatywnej AI*. Chociaż firmy prowadzą wiele eksperymentów, próby skalowania rozwiązań w celu zastosowania ich w środowisku produkcyjnym ujawniają fundamentalne braki w sposobie strukturyzowania danych, dostępu do nich i zarządzania nimi w przedsiębiorstwie.

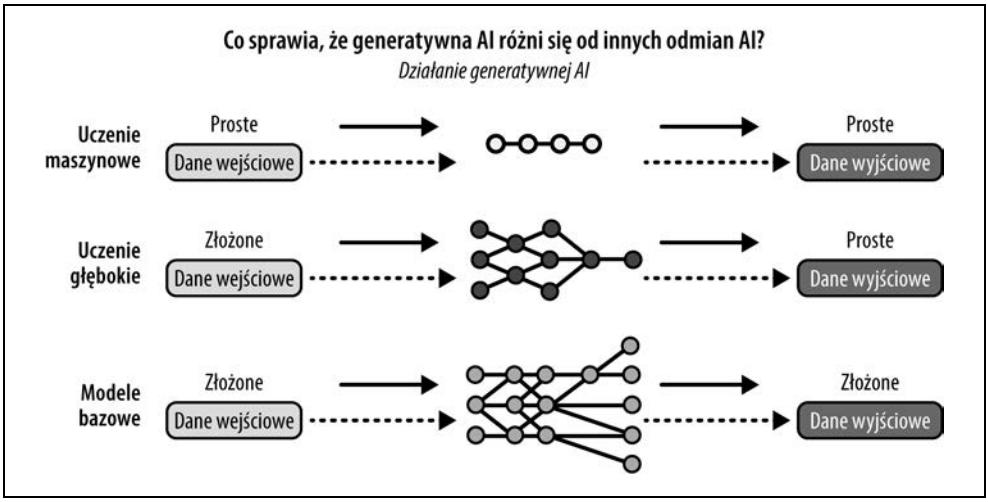
Aby zrozumieć, dlaczego tradycyjne architektury danych zawodzą w kontekście generatywnej AI, należy najpierw przeanalizować, co sprawia, że technologia ta zasadniczo różni się od starszych podejść do AI. Istniejące różnice techniczne są bezpośrednią przyczyną problemów infrastrukturalnych, z którymi borykają się organizacje podczas przechodzenia od fazy eksperymentów do zastosowań produkcyjnych.

Czym wyróżnia się generatywna AI?

Generatywna AI wiąże się z fundamentalnymi zmianami w sposobach budowania systemów AI i ich możliwościach. Tradycyjne systemy AI zazwyczaj służą przede wszystkim do klasyfikowania, przewidywania lub rekomendowania na podstawie ustrukturyzowanych danych wejściowych. Bazują na wąskich modelach specyficznych dla danego zadania, których usprawnianie wymaga oznaczonych zbiorów danych lub sygnałów reprezentujących wzmocnienia. Z kolei modele generatywne, a przede wszystkim duże modele bazowe, są w stanie tworzyć całkowicie nowe treści, w tym tekst, obrazy, muzykę, wideo, a nawet kod.

Modele bazowe są przełomowe nie tylko ze względu na skalę lub różnorodność wyników, ale też z powodu metod uczenia (rysunek 1.2). Zamiast polegać głównie na danych oznaczonych przez człowieka (jak ma to miejsce w uczeniu nadzorowanym) lub pętlach z nagrodami, gdzie stosowana jest metoda prób i błędów (jak w uczeniu przez wzmacnianie), nowe modele wykorzystują uczenie *samonadzorowane*. Jest to metoda, w której model sam uczy się rozumieć i generować język, obrazy lub kod przez przewidywanie zamaskowanych lub brakujących elementów w surowych, nieoznaczonych danych.

Kup książkę Poleć książkę



Rysunek 1.2. Porównanie tradycyjnego uczenia nadzorowanego (z oznakowanymi danymi i uczeniem jawnym), stosowanego w uczeniu maszynowym i uczeniu głębokim, z uczeniem samonadzorowanym

Proces ten odzwierciedla sposób uczenia się typowy dla ludzi. Uczymy się przez obserwację, wyciąganie wniosków i tworzenie wewnętrznych reprezentacji świata bez konieczności ciągłego otrzymywania bezpośrednich instrukcji. Na przykład gdy opowiadasz małemu dziecku o psie, kocie lub innym zwierzęciu, nie zdobywa ono wiedzy wyłącznie o tym konkretnym gatunku. Wkrótce zaczyna rozróżniać poszczególne gatunki zwierząt, a ostatecznie także rasy.

Architektura transformera: techniczne podstawy generatywnej AI

Podstawą współczesnej generatywnej AI jest architektura transformera. Stanowi ona przełom w uczeniu głębokim. Zrewolucjonizowała sposób, w jaki systemy AI przetwarzają i generują treści. Transformery, opisane po raz pierwszy w przełomowej publikacji z 2017 roku *Attention Is All You Need* (*https://oreil.ly/Lf6li*) przez Ashisha Vaswaniego i zespół z Google'a, umożliwiają modelom ujmowanie złożonych relacji w danych za pomocą mechanizmu zwanego *samouwagą*.

Mechanizm samouwagi pozwala modelowi określić znaczenie różnych elementów danych wejściowych dla generowania każdego aspektu danych wyjściowych. Umożliwia to modelowi utrzymanie spójności i kontekstu przy przetwarzaniu długich sekwencji tekstu lub innych danych, co stanowiło problem dla starszych architektur. Oto najważniejsze aspekty techniczne transformerów, które wpływają na wymagania dotyczące danych:

Przetwarzanie równoległe

W przeciwieństwie do wcześniejszych modeli sekwencyjnych (na przykład rekurencyjnych sieci neuronowych i sieci LSTM) transformery przetwarzają wszystkie tokeny wejściowe jednocześnie, co umożliwia wydajne uczenie modelu z wykorzystaniem ogromnych zbiorów danych.

Kup książkę Poleć książkę

Okna kontekstowe

Transformery działają w oknach kontekstowych o stałej długości, co wymaga infrastruktury danych, która pozwala skutecznie zarządzać kontekstem i pobierać jego odpowiednie porcje.

Kodowanie pozycyjne

Transformery wykorzystują kodowanie pozycyjne do zrozumienia kolejności elementów w sekwencji, co wymaga przygotowania danych, które zachowują sensowne relacje w sekwencjach.

Mechanizmy uwagi

Mechanizm samouwagi tworzy graf obliczeniowy łączący każdy token z wszystkimi innymi. Pozwala to na zaawansowane zrozumienie kontekstu, ale wymaga znacznych zasobów obliczeniowych.

Te cechy architektury bezpośrednio wpływają na to, jak dane muszą być ustrukturyzowane, przechowywane i pobierane w celu uzyskania skutecznych wdrożeń generatywnej AI. Tradycyjne architektury danych zoptymalizowane na potrzeby wzorców dostępu opartych na wierszach lub kolumnach są często źle dopasowane do kontekstowych, opartych na tokenach wzorców dostępu niezbędnych w modelach bazujących na transformerach.

Dane przedsiębiorstwa jako główny czynnik różnicujący

Chociaż modele bazowe mają duże ogólne możliwości, ich prawdziwa wartość biznesowa ujawnia się dopiero wtedy, gdy połączy się je z unikalnymi danymi organizacji. Pozwala to przekształcić systemy sztucznej inteligencji ogólnego przeznaczenia w wyspecjalizowane narzędzia biznesowe, które rozumieją kontekst z określonej branży, terminologię firmy i wiedzę specyficzną dla danej dziedziny.

Organizacje, które potrafią skutecznie połączyć własne zastrzeżone dane z modelami bazowymi, zyskują znaczącą przewagę konkurencyjną dzięki dokładniejszym, trafniejszym i bardziej godnym zaufania rozwiązaniom AI. W tym celu powszechnie stosuje się cztery strategie:

Inżynieria kontekstu

Wykorzystanie techniki RAG w celu dostarczenia zastrzeżonych informacji jako kontekstu dla modeli bazowych. Wymaga to wyszukiwania semantycznego w osadzeniach wektorowych z niskim opóźnieniem i aktualizacjami w czasie rzeczywistym.

Dostrajanie

Dostosowywanie wstępnie wyuczonych modeli bazowych za pomocą zestawów danych specyficznych dla określonej dziedziny. Wymaga wysokiej jakości wersjonowanych i oznaczonych zestawów danych z precyzyjnym śledzeniem pochodzenia.

Trening modeli niestandardowych

Budowanie modeli specjalnego przeznaczenia zoptymalizowanych pod kątem konkretnych zastosowań i danych. Wymaga masowego równoległego dostępu do różnorodnych zestawów danych, zoptymalizowanego pod kątem przepustowości, a nie opóźnień.

Kup książkę Poleć książkę

Optymalizacja modeli

Używanie technik takich jak destylacja wiedzy i przycinanie modeli do tworzenie mniejszych, bardziej wydajnych modeli, które mają możliwości większych rozwiązań, a jednocześnie wymagają mniejszych zasobów.

W rozdziale 2. wyjaśniamy, że te wzorce implementacji wyznaczają kontinuum dotyczące poziomu strukturalnego powiązania między danymi przedsiębiorstwa a działaniem modelu. W miarę jak organizacje wykorzystują te techniki, wprowadzanie generatywnej AI zazwyczaj ewoluuje i obejmuje etapy, z których każdy stawia inne wymagania infrastrukturze danych. Te etapy omawiamy w podrozdziale „Ewolucja zastosowań generatywnej AI".

Inteligencja kontekstowa: prosty przykład

Zrozumienie technicznych podstaw generatywnej AI, w tym inteligencji kontekstowej, uczenia samonadzorowanego i rozumienia semantycznego, jest niezbędne, ponieważ mechanizmy te umożliwiają rozwój coraz bardziej wyrafinowanych aplikacji. Aby zilustrować znaczenie mechanizmu samouwagi, omówimy prosty, ale wymowny eksperyment. Poproś model generatywny o utworzenie obrazu przedstawiającego *a man sitting by a bank*[1], a możesz otrzymać zdjęcie człowieka siedzącego przed instytucją finansową. Teraz zmień prompt na *a man fishing by a bank*, a wynikowy obraz prawdopodobnie będzie przedstawiał mężczyznę nad brzegiem rzeki (rysunek 1.3). Model nie zgaduje losowo. Wykorzystuje kontekst wyznaczany przez sąsiednie słowa, aby doprecyzować znaczenie wyrazu *bank*.



Rysunek 1.3. Przykładowe obrazy wygenerowane przez narzędzie Amazon Nova Canvas na podstawie promptów „a man sitting by a bank" i „a man fishing by a bank"

Teraz pokażemy, że w procesie rozumienia języka pojedyncze słowo może całkowicie zmienić to, czego dotyczy zaimek. Ilustruje to rysunek 1.4. W zdaniu *The trophy doesn't fit in the suitcase because it is too big* (puchar nie mieści się w walizce, ponieważ jest zbyt duży) zaimek *it* (to) zdecydowanie dotyczy pucharu. Ale wystarczy zmienić jedno słowo, *The trophy doesn't fit in the suitcase because it is too small* (puchar nie mieści się w walizce, ponieważ jest zbyt mała),

[1] Zdanie może oznaczać „mężczyznę siedzącego przy banku" lub „mężczyznę łowiącego ryby przy brzegu", ponieważ po angielsku *bank* to zarówno bank, jak i brzeg — *przyp. tłum.*

Kup książkę Poleć książkę

a *it* (to) będzie dotyczyć walizki. Zdanie ma tę samą strukturę gramatyczną i obejmuje ten sam zaimek, ale kontekst zapewniany przez słowo *big* lub *small* zasadniczo zmienia to, czego ten zaimek dotyczy. To pokazuje, w jaki sposób modele muszą wykorzystywać wskazówki kontekstowe i wiedzę o świecie rzeczywistym na temat fizycznych ograniczeń, aby poprawnie interpretować znaczenie.



Rysunek 1.4. Przykład określania na podstawie kontekstu, czego dotyczy zaimek

Zdolność wnioskowania o znaczeniu z kontekstu dowodzi, że model wewnętrznie rozumie relacje semantyczne. Stanowi to kluczową różnicę w stosunku do tradycyjnych systemów, które bazują na dopasowywaniu słów kluczowych lub sztywnych etykietach. W miarę jak modele generatywne są trenowane na podstawie coraz bardziej zróżnicowanych i obszernych danych, ich poziom zrozumienia się pogłębia, co pozwala na wykonywanie zadań, których nigdy nie były bezpośrednio uczone.

Takie wnioskowanie kontekstowe to nie magia. Model uczy się go podczas treningu. Aby zrozumieć, w jaki sposób modele generatywnej AI tak precyzyjnie potrafią interpretować znaczenie, trzeba przyjrzeć się temu, jak wewnętrznie reprezentują język.

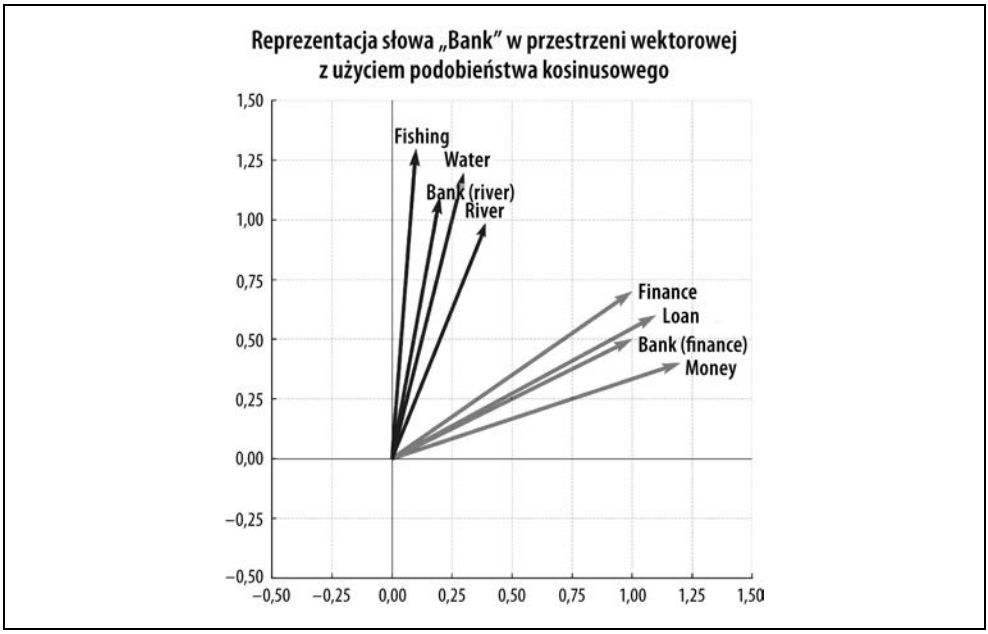
Reprezentacja znaczenia w przestrzeni wektorowej

Skąd model generatywnej AI wie, czy słowa są sobie bliskie semantycznie? Zrozumienie kontekstu matematycznie bazuje na sposobie, w jaki modele reprezentują słowa, frazy, a nawet pojęcia jako wektory w przestrzeni wielowymiarowej (rysunek 1.5). Słowa o podobnym znaczeniu lub zastosowaniu kontekstowym znajdują się bliżej siebie, podczas gdy terminy niepowiązane lub używane w inny sposób są od siebie oddalone w większym stopniu.

W tym przykładzie:

- Słowo *bank* (bank) znajduje się blisko takich terminów jak *loan* (pożyczka), *money* (pieniądze) i *finance* (finanse).
- Słowo *bank* (brzeg rzeki) pojawia się w pobliżu słów *water* (woda), *fishing* (wędkowanie) i *river* (rzeka).
- Kąt między dwoma wektorami dla słowa *bank* jest duży, co wskazuje na niskie podobieństwo kosinusowe (czyli niewielkie pokrywanie się kontekstów).

Kup książkę Poleć książkę



Rysunek 1.5. Wizualizacja dwuwymiarowej przestrzeni wektorowej obrazująca, jak różne znaczenia słowa „bank" są umiejscowione w stosunku do innych wyrazów

Taka struktura umożliwia generatywnym systemom AI wnioskowanie semantyczne (zamiast syntaktycznego), dzięki czemu potrafią wykonywać zadania związane z wyszukiwaniem kontekstowym, podsumowywaniem, odpowiadaniem na pytania i kreatywnym generowaniem treści w sposób niedostępny dla tradycyjnych odmian sztucznej inteligencji.

Zdolność wyszukiwania semantycznego, którą w dalszej części tego rozdziału analizujemy jako jeden z pięciu pojawiających się wzorców architektonicznych, staje się fundamentalnym wymogiem dla korporacyjnych systemów AI.

Przykład praktyczny z organizacji: zrozumienie wielu dokumentów w dziale obsługi klienta

Aby zilustrować, jak rozumienie oparte na wektorach sprawdza się w środowisku biznesowym, omówimy bazę wiedzy obsługi klienta z następującymi wpisami:

- Produkt X wymaga oprogramowania układowego w wersji 2.1 lub nowszej, gdy jest używany z kontrolerem Y3000.
- Kontroler Y3000 został wycofany w listopadzie 2023 roku i zastąpiony modelem Y3500.
- Kontrolery Y3500 są wstecznie kompatybilne ze wszystkimi produktami wymagającymi kontrolerów Y3000.

Kup książkę Poleć książkę

Gdy agent obsługi klienta zapyta: „Jakiego oprogramowania układowego wymaga Produkt X?”, tradycyjne systemy oparte na słowach kluczowych prawdopodobnie zwrócą tylko pierwszy dokument. Jednak system z generatywną AI wykorzystujący rozumienie semantyczne potrafi:

- rozpoznać, że wszystkie trzy dokumenty są powiązane kontekstowo z powodu wspólnych informacji o kontrolerach;
- zrozumieć relację czasową między modelami Y3000 a nowszymi modelami Y3500;
- uwzględnić informację o kompatybilności, aby wygenerować kompletną odpowiedź, która obejmuje zarówno pierwotny wymóg, jak i zaktualizowane informacje o kontrolerze.

Zdolność korzystania z kontekstu z wielu dokumentów, co ilustruje rysunek 1.6, sprawia, że generatywna AI zasadniczo różni się od tradycyjnych systemów danych, w których architektury danych często nie pozwalają na skuteczny rozwój takich funkcji.

Zrozumienie sposobu, w jaki modele generatywnej AI przetwarzają i reprezentują znaczenie, jest kluczem do projektowania inteligentnych systemów. Stanowi to jednak tylko część całej historii. W miarę jak technologie ewoluują od prostych asystentów do autonomicznych agentów, organizacje reagują rozwojem nowych architektur danych, które mają zniwelować luki między tradycyjnymi systemami a infrastrukturą dostosowaną do AI (tabela 1.1).

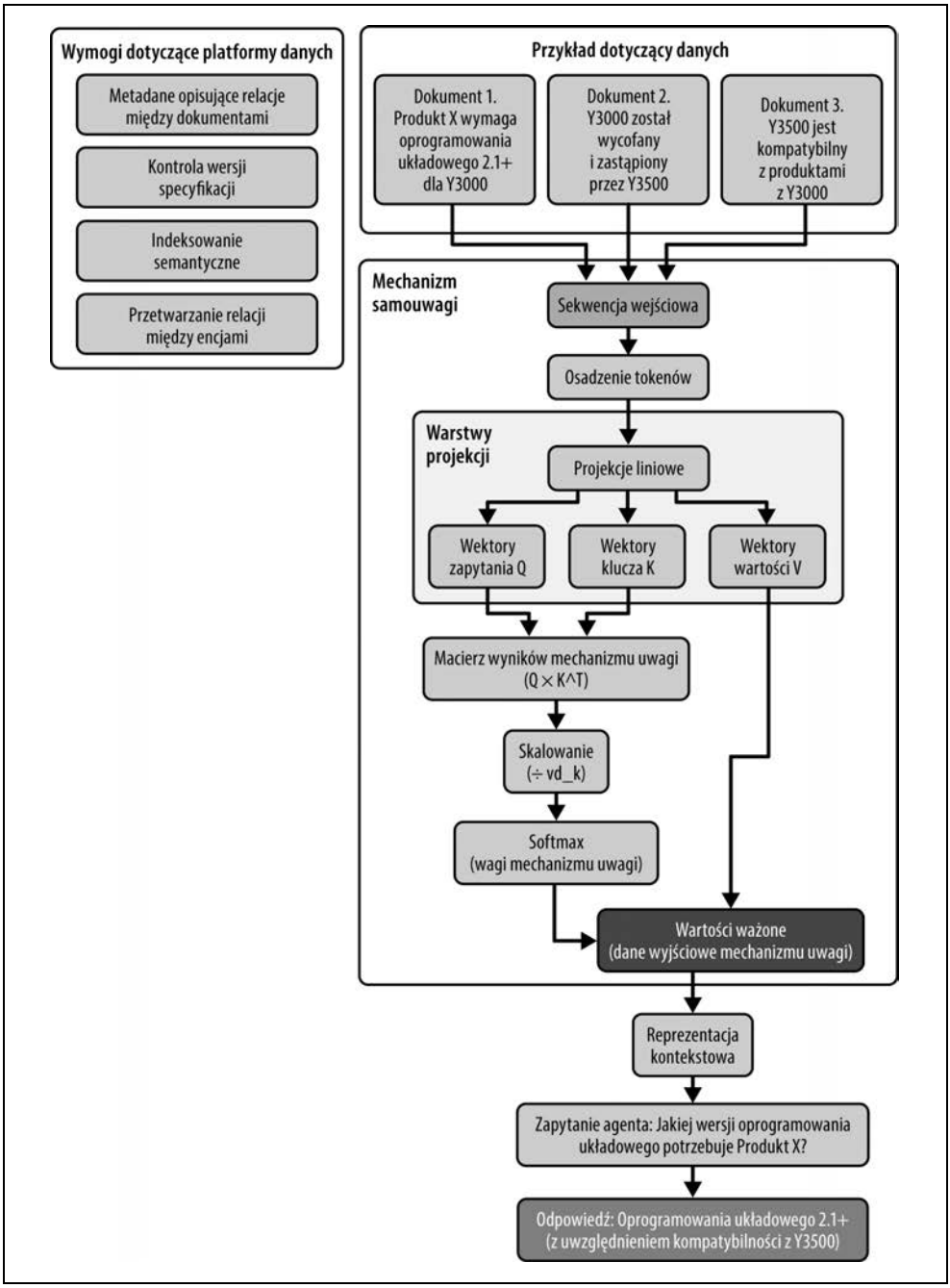
Tabela 1.1. Różnice między tradycyjną i generatywną AI

Aspekt	Tradycyjna AI	Generatywna AI
Metoda uczenia	Uczenie nadzorowane z użyciem oznaczonych danych	Uczenie samonadzorowane na podstawie surowych danych
Rozumienie	Syntaktyczne dopasowywanie do wzorca	Rozumienie semantyczne na podstawie kontekstu
Reprezentacja	Wektory cech	Osadzenia kontekstowe w przestrzeni wektorowej
Możliwości	Klasyfikowanie, prognozowanie, rekomendowanie	Generowanie, wnioskowanie, rozumienie kontekstowe
Wymogi dotyczące danych	Ustrukturyzowane i oznaczone zbiory danych	Zróżnicowane dane kontekstowe z relacjami
Wymogi dotyczące infrastruktury	Przetwarzanie wsadowe, hurtownie danych	Pobieranie w czasie rzeczywistym, wektorowe bazy danych, zarządzanie kontekstem

Ewolucja zastosowań generatywnej AI

W miarę dojrzewania technologii generatywnej AI obserwujemy szybką ewolucję jej zastosowań od prostych asystentów po coraz bardziej autonomiczne agenty. Ten postęp jest związany nie tylko z osiągnięciami technicznymi, ale również z fundamentalną zmianą w sposobie interakcji systemów AI z danymi, użytkownikami i światem. Organizacje muszą zrozumieć tę zmianę, aby przygotować infrastrukturę danych zarówno na obecne, jak i przyszłe możliwości sztucznej inteligencji.

Kup książkę Poleć książkę

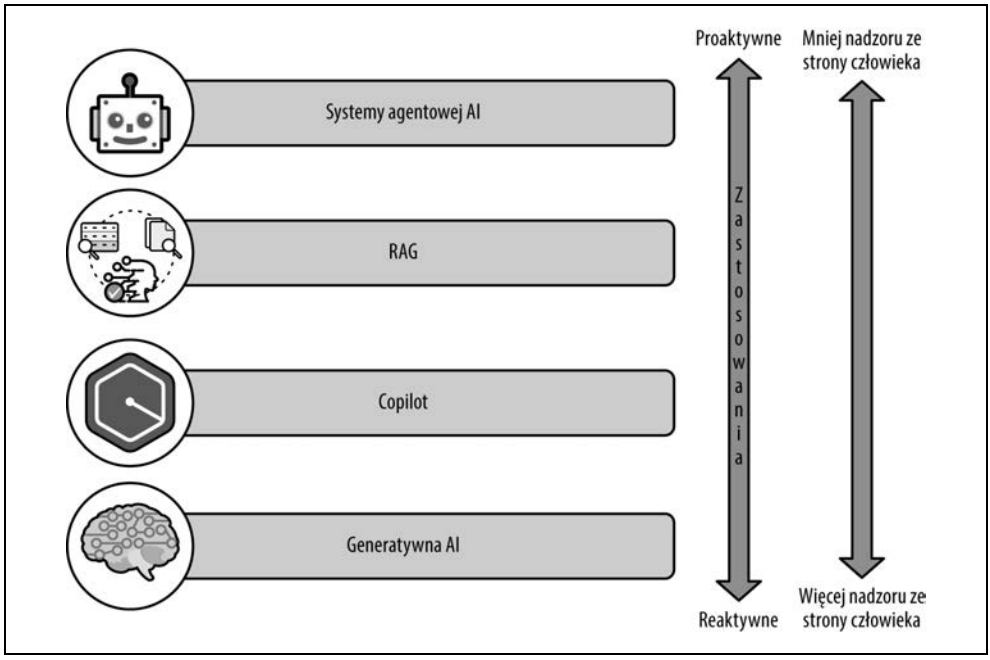


Rysunek 1.6. Proces, w jakim samouwaga umożliwia zaawansowane rozumienie wielu dokumentów i łączenie informacji z trzech tekstów, aby zwrócić kompletną odpowiedź na temat wymagań dotyczących oprogramowania układowego

Kup książkę Poleć książkę

Od asystentów do agentów: cztery etapy ewolucji generatywnej AI

Ewolucję zastosowań generatywnej AI można podzielić na cztery etapy przedstawione na rysunku 1.7. Każdy z nich wykorzystuje możliwości z poprzedniego, a dodatkowo skutkuje nowymi wymogami dotyczącymi danych i wyzwaniami na poziomie organizacji.



Rysunek 1.7. Cztery etapy ewolucji generatywnej AI

Etap 1. Asystenty AI — wyszukiwanie informacji i realizowanie podstawowych zadań

Pierwsza generacja zastosowań generatywnej AI skupiała się na dostarczaniu informacji i wykonywaniu prostych zadań w reakcji na bezpośrednie zapytania użytkowników. Te asystenty, na przykład wczesne wersje narzędzi ChatGPT i Gemini, świetnie radziły sobie z:

- odpowiadaniem na pytania oparte na faktach na podstawie danych treningowych;
- generowaniem treści takich jak e-maile, podsumowania lub teksty kreatywne;
- dostarczaniem wyjaśnień pojęć lub procedur;
- tłumaczeniem między językami.

Systemy te korzystały głównie z wiedzy uzyskanej w procesie uczenia. Posiadały ograniczoną zdolność dostępu do nowych informacji lub wnioskowania na ich temat. Miały stosunkowo proste wymagania dotyczące danych. Potrzebowały wysokiej jakości danych treningowych, ale tylko minimalnego poziomu integracji danych w czasie rzeczywistym lub złożonego zarządzania wiedzą.

Kup książkę Poleć książkę

Etap 2. Asystenty typu copilot — współpraca z ludźmi

W drugim etapie systemy generatywnej AI ewoluowały do postaci asystentów typu copilot, które współpracują z ludźmi i rozszerzają ich możliwości w określonych dziedzinach. Przykłady to GitHub Copilot w obszarze programowania, Microsoft 365 Copilot w obszarze pracy biurowej i Adobe Firefly w obszarze pracy twórczej. Asystenty typu copilot:

- rozumieją konteksty i terminologię specyficzną dla danej dziedziny;
- generują treści i sugestie odpowiednie dla danej dziedziny;
- uczą się na podstawie opinii i preferencji użytkowników;
- integrują się z istniejącymi narzędziami i przepływami pracy.

Asystenty typu copilot mają bardziej złożone wymogi dotyczące danych. Niezbędne są między innymi dostęp do baz wiedzy specyficznych dla danej dziedziny, rozumienie zastrzeżonych formatów danych i utrzymywanie kontekstu w trakcie sesji użytkownika. Na tym etapie zaciera się granica między ogólnymi a wyspecjalizowanymi systemami AI i konieczne są bardziej wyrafinowane strategie integracji danych.

Etap 3. Agenty oparte na metodzie RAG — rozszerzona wiedza i rozumienie kontekstu

Na trzecim etapie pojawiła się metoda RAG, czyli generowanie wspomagane wyszukiwaniem (ang. *retrieval-augmented generation*), która pomaga przezwyciężyć ograniczenia związane z wiedzą zapisywaną w modelu na podstawie wstępnego uczenia. Agenty oparte na RAG potrafią:

- uzyskiwać dostęp do firmowych baz wiedzy w czasie rzeczywistym i wyciągać wnioski na ich podstawie;
- dostarczać aktualne informacje wykraczające poza moment zakończenia nauki danego modelu;
- generować odpowiedzi na podstawie konkretnych dokumentów lub źródeł danych;
- łączyć informacje z wcześniej odizolowanych od siebie źródeł.

Metoda RAG znacznie zwiększa złożoność wymagań dotyczących danych. Konieczne jest stosowanie:

- baz wektorowych do wyszukiwania semantycznego;
- potoków generowania osadzeń w celu przekształcania dokumentów w reprezentacje wektorowe;
- zarządzania metadanymi w celu określania źródeł danych;
- mechanizmów odświeżania treści w celu utrzymania jej aktualności.

Organizacje wdrażające agenty oparte na metodzie RAG odkryły, że dotychczasowa architektura danych często nie była w stanie sprostać nowym wymogom, co prowadziło do problemów w fazie wdrażania. Trudności te omawiamy w dalszej części rozdziału.

Kup książkę Poleć książkę

Etap 4. Agentowa AI — autonomiczne podejmowanie decyzji i działanie

Czwartą i obecną granicą ewolucji generatywnej AI jest agentowa AI. Do tej kategorii należą systemy, które potrafią autonomicznie planować i wykonywać złożone zadania przy minimalnym nadzorze ze strony człowieka. Agenty tego typu:

- rozkładają złożone cele na wykonalne kroki;
- uzyskują dostęp do wielu narzędzi i API w celu realizowania zadań;
- podejmują decyzje na podstawie informacji pozyskiwanych w czasie rzeczywistym;
- uczą się na sukcesach i błędach, aby poprawić skuteczność;
- działają z coraz większym stopniem autonomii.

Agentowa AI oznacza znaczny skok w zakresie wymogów dotyczących danych i gotowości organizacyjnej. Systemy tego typu potrzebują nie tylko dostępu do informacji, ale także:

- modelu uprawnień do kontrolowania działań;
- systemów monitorowania na potrzeby nadzoru;
- mechanizmów sprzężenia zwrotnego w celu uczenia się;
- zabezpieczeń zapobiegających nadużyciom;
- weryfikacji wiarygodności źródeł danych.

Etap czwarty już się pojawia w wąskich dziedzinach. Ponad milion robotów Amazona (*https://oreil.ly/NLDUe*) potrafi autonomicznie podejmować decyzje w obszarze logistyki, gdzie podejmują niezależne decyzje dotyczące obsługi paczek, optymalizacji tras i zarządzania zapasami.

Rosnąca złożoność wymagań dotyczących danych

Każdy etap ewolucji wiąże się z nowymi wymogami i wyzwaniami dotyczącymi danych, co obrazuje tabela 1.2.

Tabela 1.2. Ewolucja generatywnej AI

Etap rozwoju generatywnej AI	Główny sposób wykorzystywania danych	Najważniejsze wymogi dotyczące danych	Wyzwania organizacyjne
Asystenty AI	Wiedza uzyskana w wyniku wstępnej nauki	Wysokiej jakości dane treningowe	Zarządzanie oczekiwaniami dotyczącymi ograniczeń wiedzy
Asystenty typu copilot	Wiedza specyficzna dla dziedziny	Integrowanie z zastrzeżonymi systemami i formatami	Uzyskanie równowagi między pomocą ze strony asystentów i ludzką wiedzą ekspercką
Agenty oparte na metodzie RAG	Bazy wiedzy w przedsiębiorstwie	Wektorowe bazy danych, potoki tworzenia osadzeń, zarządzanie metadanymi	Eliminowanie silosów danych, gwarantowanie jakości danych
Agentowa AI	Wiedza z wielu źródeł i uprawnienia do podejmowania działań	Modele uprawnień, systemy monitorowania, weryfikacja wiarygodności	Bezpieczeństwo, ład, odpowiedzialność, obserwowalność i kontrola

Kup książkę Poleć książkę

Ta związana z ewolucją złożoność wyjaśnia, dlaczego tak wiele organizacji ma trudności z wyjściem poza fazę eksperymentów i dlaczego firmy, które z powodzeniem przeprowadzały wdrożenia na wcześniejszych etapach generatywnej AI, mogą napotkać trudności przy bardziej zaawansowanych rozwiązaniach. Każdy etap wymaga uwzględnienia innych kwestii dotyczących architektury danych. Na podstawie analizy wzorców powtarzających się w udanych wdrożeniach można zidentyfikować konkretne podejścia architektoniczne, które pomagają skutecznie radzić sobie z rosnącymi wymogami.

Jakie rozwiązania są obecnie rozwijane w wiodących organizacjach?

Chociaż luka między eksperymentami a rozwiązania produkcyjnymi nadal jest znaczna i obecnie niecałe 10 procent organizacji z sukcesem skaluje generatywną AI, wśród firm, które odnoszą na tym polu sukcesy, widoczny jest wyraźny trend. Zamiast czekać na idealne rozwiązania, wiodące organizacje budują systemy zgodne z określonymi wzorcami architektonicznymi. Wzorce te pozwalają przekształcić tradycyjną infrastrukturę danych w fundamenty dostosowane do AI, co pomaga firmom przejść od projektów PoC do rozwiązań produkcyjnych.

Weryfikacja w środowisku produkcyjnym

Badania McKinsey Global Survey on AI z 2025 roku (*https://oreil.ly/ryNP6*) ujawniają zaskakujący problem na etapie wdrożeń. Choć 88 procent organizacji korzysta obecnie z AI w co najmniej jednej funkcji (w porównaniu z 72 procentami w 2024 roku), przejście od fazy projektów PoC do środowiska produkcyjnego pozostaje trudne. Sukces w generowaniu istotnej wartości biznesowej różni się w znacznym stopniu zależnie od złożoności zastosowań. Otóż 15 – 20 procent organizacji z powodzeniem przeskalowało AI do przetwarzania dokumentów i obsługi klienta, 10 – 15 procent wyszło poza etap eksperymentów w generowaniu treści i analizie danych, a mniej niż 10 procent (obecnie szacowane na 6 procent „liderów") z powodzeniem zastosowało AI do podejmowania złożonych decyzji lub wdrożyło autonomiczne agenty.

W swojej własnej pracy mieliśmy okazję zaobserwować te wyzwania na własne oczy. Jeden z klientów musiał wycofać wdrożenie generatywnej AI nie przez problemy z modelem lub promptami, ale z powodu niskiej jakości danych. Choć w środowisku testowym wszystko działało poprawnie, w środowisku produkcyjnym ujawniły się niespójności w danych, co doprowadziło do całkowitego wstrzymania prac. Obecnie organizacja skupia się na przebudowie podstawowych aspektów architektury danych.

Prowadziliśmy również udane wdrożenia. Jedno z nich dotyczyło rozwiązania do analizy odchyleń dla biura dyrektora finansowego. Narzędzie to przetwarza dane z systemów ERP (ang. *enterprise resource planning*), oblicza odchylenia i wykorzystuje AI do tworzenia wersji wstępnej komentarzy finansowych. W innym scenariuszu pomogliśmy klientowi zbudować marketingowego agenta konwersacyjnego, który generuje lokalne treści dla zespołów działających w terenie.

Kup książkę Poleć książkę

W tego typu praktycznych wdrożeniach pojawiło się pięć odrębnych wzorców architektonicznych. Nie są to tylko teoretyczne modele, lecz sprawdzone podejścia zaprojektowane w celu przezwyciężenia ograniczeń tradycyjnych architektur danych i spełnienia wymogów specyficznych dla generatywnej AI.

Nowe wzorce architektoniczne dla generatywnej AI

W tej części szczegółowo omawiamy pięć nowych wzorców architektonicznych.

Wzorzec 1. Grafy wiedzy na potrzeby inteligencji kontekstowej

Czym są? Grafy wiedzy (*https://oreil.ly/0ppyn*) łączą rozproszone źródła danych przez bezpośrednie odwzorowywanie relacji. Tworzą sieć semantyczną, która umożliwia wnioskowanie na poziomie wielu domen. W przeciwieństwie do tradycyjnych baz danych, które przechowują izolowane rekordy, grafy wiedzy reprezentują encje (klientów, produkty, transakcje) i relacje między nimi (kupiono, zwrócono, polecono) jako połączone ze sobą węzły.

Dlaczego mają znaczenie dla AI? Agenty generatywnej AI muszą rozumieć kontekst wykraczający poza granice organizacji. Kiedy agent do obsługi klienta zadaje pytanie: „Dlaczego dany klient zwrócił trzy zamówienia w zeszłym miesiącu?", odpowiedź wymaga połączenia danych na temat klienta, historii zamówień, specyfikacji produktów, logistyki wysyłki i zgłoszeń do pomocy technicznej. Są to dane, które zazwyczaj znajdują się w oddzielnych systemach. Grafy wiedzy sprawiają, że takie dane są bezpośrednio powiązane ze sobą i można je przeszukiwać.

Praktyczny przykład. Firma świadcząca usługi finansowe wdrożyła graf wiedzy łączący profile klientów, historie transakcji, dane rynkowe i wymogi regulacyjne. Jej agent ds. zgodności z regulacjami oparty na architekturze RAG może dzięki temu odpowiadać na pytania typu „Których klientów dotyczy nowe rozporządzenie UE?" na podstawie badania relacji między danymi, które wcześniej znajdowały się w odizolowanych zbiorach. Dzięki temu czas analizy skrócił się z dni do minut.

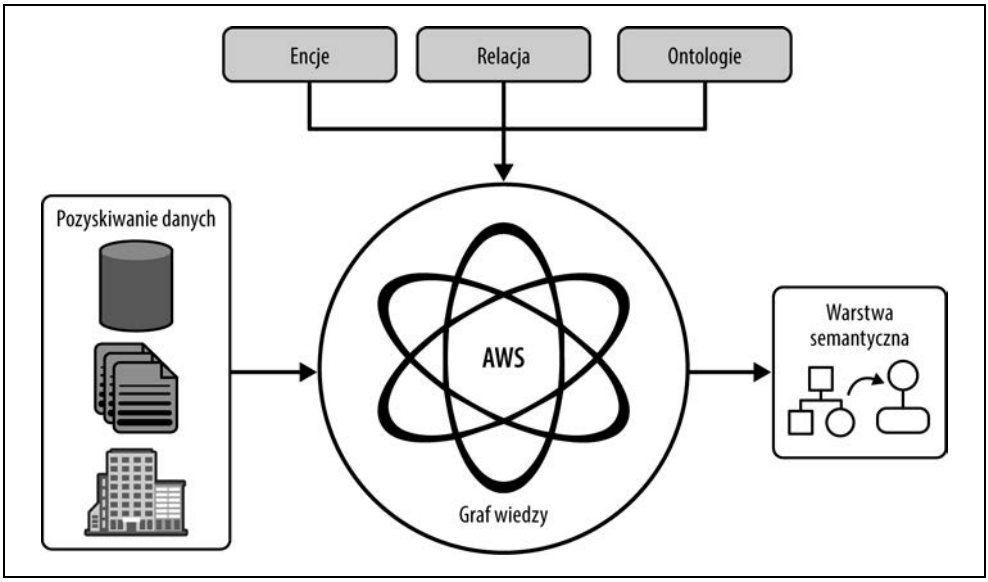
Wymogi dotyczące danych:

- bazy grafowe (Neo4j, Amazon Neptune) do przechowywania relacji;
- potoki rozpoznawania encji do identyfikowania i łączenia zduplikowanych encji;
- zarządzanie ontologią w celu spójnego definiowania relacji;
- ciągła synchronizacja danych z systemów źródłowych.

Ten wzorzec zilustrowaliśmy na rysunku 1.8.

Ewolucja w stronę siatki danych. Wiodące organizacje rozszerzają ten wzorzec — traktują grafy wiedzy jako federacyjne produkty danych. Zamiast centralizować wszystkie dane w jednym grafie, tworzą grafy wiedzy specyficzne dla domeny (graf klientów, graf produktów, graf łańcucha dostaw), które agenty mogą samodzielnie odkrywać i przeszukiwać. Takie podejście oparte na siatce danych umożliwia skalowanie rozwiązań, ponieważ nie występują w nim wąskie gardła spowodowane centralizacją.

Kup książkę Poleć książkę



Rysunek 1.8. Grafy wiedzy tworzone na potrzeby inteligencji kontekstowej

Wzorzec 2. Architektura oparta na zdarzeniach jako fundament AI działającej w czasie rzeczywistym

Czym jest? Architektura oparta na zdarzeniach (*https://oreil.ly/PS62X*) zastępuje przetwarzanie wsadowe potokami ze strumieniowaniem danych. Takie potoki wykrywają i uwzględniają zmiany w miarę ich występowania. Każde zdarzenie biznesowe, na przykład zakup dokonany przez klienta, aktualizacja zapasów, zmiana ceny lub zgłoszenie do pomocy technicznej, przepływa przez strumienie zdarzeń, które mogą być wykorzystywane przez wiele systemów w czasie rzeczywistym.

Dlaczego ma znaczenie dla AI? Agenty potrzebują bieżącego kontekstu, a nie nieaktualnych danych z wczorajszego zadania wsadowego. Kiedy klient pyta: „Czy ten produkt znajduje się w magazynie?", odpowiedź musi odzwierciedlać zmiany w stanie zapasów sprzed kilku sekund, a nie wczorajszy snapshot. Architektury oparte na zdarzeniach gwarantują, że agenty zawsze korzystają z aktualnych informacji.

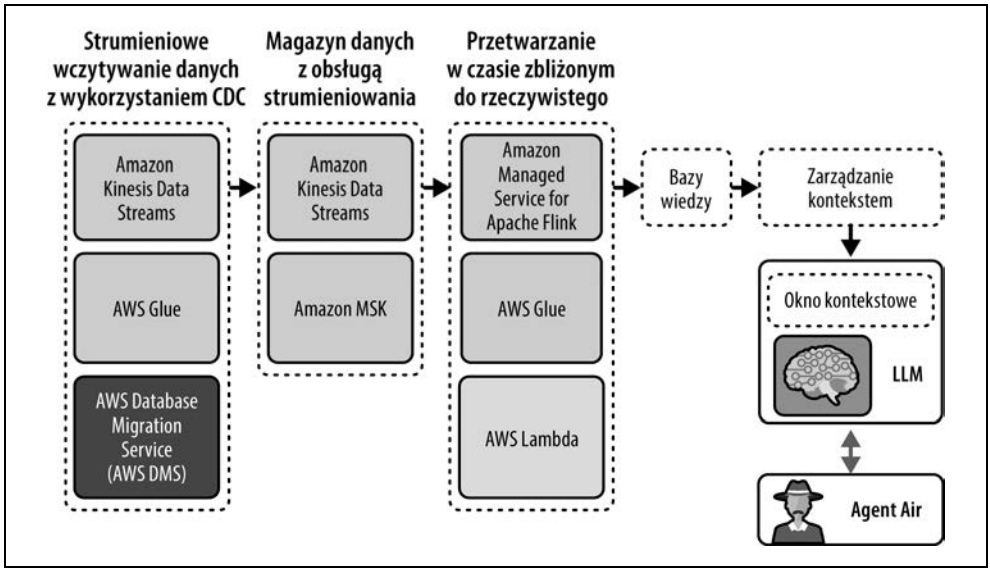
Praktyczny przykład. Firma z sektora handlu detalicznego zastąpiła przeprowadzany co nocy proces wyodrębniania, przekształcania i ładowania danych (ETL) strumieniem zdarzeń opartym na platformie Kafka. Zdarzenia dotyczące działań klientów (odsłony stron, dodanie produktu do koszyka, zakup) stale napływają do wektorowej bazy danych, dzięki czemu system może aktualizować reprezentujące produkty osadzenia w czasie zbliżonym do rzeczywistego. Agent generujący rekomendacje jest w stanie uwzględniać popularne produkty w ciągu kilku minut, a nie dni, co skutkuje wzrostem współczynnika konwersji.

Kup książkę Poleć książkę

Wymogi dotyczące danych:

- platformy przetwarzania strumieniowego (Kafka, Amazon Kinesis);
- przechwytywanie zmian w danych (ang. *change data capture* — CDC) z systemów źródłowych;
- platformy przetwarzania strumieniowego (Flink, Spark Streaming) na potrzeby transformacji;
- aktualizacje wektorowej bazy danych z niskim opóźnieniem.

Rysunek 1.9 ilustruje ten wzorzec.



Rysunek 1.9. Architektura oparta na zdarzeniach na potrzeby AI działającej w czasie rzeczywistym

Odświeżanie kontekstu w czasie rzeczywistym. W najbardziej zaawansowanych implementacjach ten wzorzec jest rozszerzany o ciągłe aktualizowanie osadzeń wektorowych w reakcji na zmiany w dokumentach źródłowych. Gdy specyfikacja produktu zostaje zaktualizowana, potok generowania osadzeń automatycznie odtwarza wektory i aktualizuje indeks wyszukiwania semantycznego, dzięki czemu agenty nigdy nie pobierają nieaktualnych informacji.

Wzorzec 3. Architektura typu lakehouse jako ujednolicona platforma sztucznej inteligencji

Czym jest? Architektura typu lakehouse łączy jeziora danych i hurtownie danych w zunifikowaną platformę, która obsługuje zarówno procesy analityczne, jak i operacje związane z AI. Zbudowane z wykorzystaniem otwartych formatów tabel, takich jak Apache Iceberg, Delta Lake i Apache Hudi, repozytoria typu lakehouse obsługują transakcje ACID, ewolucję schematów i dostęp do historycznych wersji danych (możliwość przesyłania zapytań dotyczących historycznych snapshotów danych) za pomocą jeziora danych. To podejście łączy elastyczność jezior z niezawodnością hurtowni.

Kup książkę Poleć książkę

Dlaczego ma znaczenie dla AI? Tradycyjne architektury zmuszają organizacje do wyboru między jeziorami danych (elastyczne i nieustrukturyzowane) a hurtowniami danych (ustrukturyzowane i wydajne). Generatywna AI wymaga obu tych podejść: nieustrukturyzowanych dokumentów dla metody RAG, ustrukturyzowanych danych do analiz oraz możliwości ich płynnego łączenia. Repozytoria typu lakehouse na to pozwalają.

Praktyczny przykład. Organizacja z branży opieki zdrowotnej skonsolidowała dane z 15 lat badań klinicznych (ustrukturyzowane wyniki badań, nieustrukturyzowane publikacje naukowe, dokumentację pacjentów i dokumenty regulacyjne) za pomocą architektury typu lakehouse opartej na formacie Iceberg. Agenty wspomagające badania mogą teraz wysyłać zapytania o dowolne rodzaje danych za pomocą jednego interfejsu. Pozwala to łączyć ustrukturyzowane wyniki badań z nieustrukturyzowanymi notatkami badaczy. Czas przygotowywania danych skrócił się z tygodni do godzin.

Wymogi dotyczące danych:

- otwarte formaty tabel (Iceberg, Delta Lake, Hudi) do ujednoliconego przechowywania danych;
- dostęp za pomocą wielu systemów (Spark, Trino, Athena) na potrzeby różnorodnych procesów;
- ujednolicone zarządzanie (Lake Formation, Unity Catalog) wszystkimi danymi;
- obsługa różnorodnych typów danych (ustrukturyzowane, częściowo ustrukturyzowane, nieustrukturyzowane).

Rysunek 1.10 ilustruje ten wzorzec.

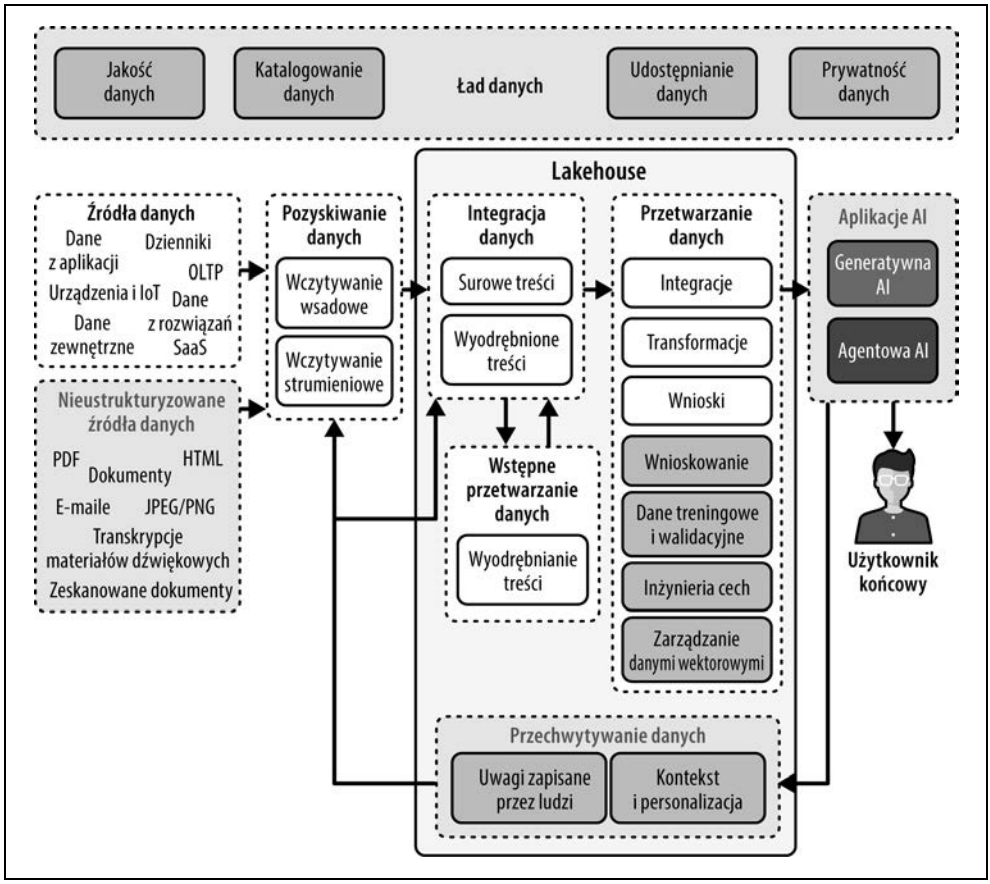
Ewolucja w kierunku rozwiązań multimodalnych. Wiodące organizacje rozszerzają architektury typu lakehouse, aby obsługiwały dane multimodalne (tekst, obraz, dźwięk, wideo i dane ustrukturyzowane) w ramach jednej platformy. Umożliwia to agentom wnioskowanie z wykorzystaniem różnych modalności, analizowanie obrazów produktów wraz z opiniami klientów i danymi o sprzedaży lub łączenie transkrypcji audio z ustrukturyzowanymi danymi z centrum obsługi telefonicznej.

Wzorzec 4. Wyszukiwanie semantyczne jako nowa warstwa zapytań

Czym jest? W wyszukiwaniu semantycznym tradycyjne wyszukiwanie oparte na słowach kluczowych jest zastępowane analizą podobieństwa z wykorzystaniem wektorów. Dokumenty są przekształcane w wielowymiarowe osadzenia, które reprezentują znaczenie danych. Umożliwia to ich wyszukiwanie na podstawie podobieństwa pojęciowego, a nie dokładnego dopasowywania słów.

Dlaczego ma znaczenie dla AI? Skuteczność agentów stosujących metodę RAG zależy od używanych w nich systemów wyszukiwania. Gdy użytkownik pyta: „Jak radzimy sobie z europejskimi wymogami dotyczącymi prywatności danych?”, wyszukiwanie oparte na słowach kluczowych może pominąć dokumenty omawiające „zgodność z RODO” lub „unijne przepisy o ochronie danych”. Jednak system z wyszukiwaniem semantycznym potrafi wykryć, że takie dokumenty dotyczą powiązanych koncepcji, dlatego pobierze wszystkie odpowiednie informacje.

Kup książkę Poleć książkę



Rysunek 1.10. Architektura typu lakehouse jako ujednolicona platforma dla AI

Przykład praktyczny. Firma produkcyjna zbudowała warstwę wyszukiwania semantycznego dla specyfikacji inżynieryjnych powstałych na przestrzeni 30 lat. Agent generujący projekty potrafi teraz znajdować podobne komponenty w dokumentacji z kilkudziesięciu lat, nawet jeśli terminologia przez ten czas uległa zmianom. Inżynierowie mogą zadawać pytania w języku naturalnym („Znajdź lekkie materiały odpowiednie do zastosowań w wysokich temperaturach") i otrzymują odpowiednie specyfikacje z dokumentów, w których nie użyto dokładnie tych samych słów.

Wymogi dotyczące danych:

- wektorowe bazy danych (Pinecone, OpenSearch, pgvector itp.) do przechowywania osadzeń;
- modele osadzeń (na przykład text-embedding-ada-002, Titan Embeddings) do wektoryzacji dokumentów;
- mechanizmy wyszukiwania hybrydowego łączące wyszukiwanie semantyczne z wyszukiwaniem słów kluczowych;
- filtrowanie metadanych w celu kontroli dostępu i poprawy trafności.

Kup książkę Poleć książkę

Wskazówka implementacyjna. Najskuteczniejsze implementacje wykorzystują wyszukiwanie hybrydowe, które łączy podobieństwo semantyczne z dopasowywaniem słów kluczowych i filtrami metadanych. Dzięki temu agenty pobierają dokumenty, które są zarówno koncepcyjnie istotne, jak i precyzyjnie pasują do określonych kryteriów (zakresy dat, typy dokumentów, kategorie bezpieczeństwa).

Wzorzec 5. Zasoby cyfrowe dostosowane do agentów

Czym są? Organizacje przeprojektowują swoje zasoby cyfrowe, w tym witryny internetowe, produkty, API i dokumentację, aby mogły z nich korzystać agenty AI, a nie tylko ludzie. Wprowadzają między innymi znaczniki danych ustrukturyzowanych (ze słownika Schema.org), API czytelne dla maszyn oraz dokumentację zrozumiałą dla agentów, która umożliwia autonomiczne odkrywanie danych i interakcję z nimi.

Dlaczego mają znaczenie dla AI? W powstającej obecnie gospodarce opartej na agentach klienci nie będą przeglądać Twojej witryny. Będą to robić ich agenty. Jeśli informacje o produkcie nie będą czytelne dla maszyn, agent konkurenta poleci jego produkt zamiast Twojego. Zasoby dostosowane do agentów sprawiają, że Twoja firma pozostanie widoczna i dostępna w świecie, w którym w kontaktach z klientami pośredniczy sztuczna inteligencja.

Przykład praktyczny. Firma z branży handlu internetowego zrekonstruowała katalog produktów przez zastosowanie znaczników Schema.org, utworzyła specyfikacje OpenAPI dla wszystkich usług i zbudowała warstwę API dostępną dla agentów. Gdy agenty zakupowe wysyłają zapytanie „Znajdź ergonomiczne krzesła biurowe w cenie poniżej 500 dolarów z dostawą następnego dnia", system w odpowiedzi przesyła ustrukturyzowane dane produktów, które agenty mogą bezpośrednio porównać i na tej podstawie podjąć działania. Sprzedaż generowana przez agenty odpowiada obecnie za 12 procent przychodów, a poziom ten wciąż rośnie.

Wymogi dotyczące danych:

- znaczniki danych ustrukturyzowanych (Schema.org, JSON-LD) dla treści;
- projektowanie zorientowane na API z wykorzystaniem specyfikacji OpenAPI lub GraphQL;
- dokumentacja dostosowana do agentów (czytelna dla maszyn, a nie tylko dla ludzi);
- uwierzytelnianie dostępu agentów i ograniczanie liczby przesyłanych przez nie żądań;
- analityka użytkowania w celu zrozumienia wzorców zachowań agentów.

Strategiczna konieczność. Ten wzorzec reprezentuje fundamentalną zmianę w sposobie, w jaki organizacje myślą o swojej działalności cyfrowej. Podobnie jak projektowanie zorientowane na urządzenia mobilne (ang. *mobile-first*) stało się niezbędne w erze smartfonów, tak w epoce AI niezwykle istotne staje się projektowanie zorientowane na agenty. Organizacje, które odraczają transformację w tym obszarze, ryzykują, że staną się niewidoczne dla agentów pośredniczących w przyszłych interakcjach z klientami.

Kup książkę Poleć książkę

Od wzorców do praktyki

Pięć opisanych wzorców nie wyklucza się wzajemnie. Najbardziej udane wdrożenia łączą w sobie kilka z nich. Na przykład organizacja z sektora handlu detalicznego może wykorzystywać grafy wiedzy do łączenia danych o klientach z danymi o produktach (wzorzec 1.), strumieniowe przetwarzanie zdarzeń do utrzymywania aktualności tych danych (wzorzec 2.), architekturę typu lakehouse do ujednolicania ustrukturyzowanych i nieustrukturyzowanych źródeł danych (wzorzec 3.), wyszukiwanie semantyczne do pozyskiwania danych przez agenty (wzorzec 4.) oraz API dostosowane do agentów do dostępu zewnętrznego (wzorzec 5.).

Najważniejszym wnioskiem jest to, że opisane wzorce rozwiązują problem fundamentalnego niedopasowania tradycyjnych architektur danych do wymagań generatywnej AI. Tradycyjne systemy były optymalizowane na potrzeby analityków uruchamiających zapytania SQL-owe dotyczące danych ustrukturyzowanych. Jednak systemy generatywnej AI potrzebują rozumienia semantycznego, kontekstu w czasie rzeczywistym, wnioskowania w różnych domenach i dostępu autonomicznego. Funkcje te wymagają fundamentalnie odmiennego podejścia do architektury.

Organizacje, które z powodzeniem wdrażają przedstawione wzorce, mają zazwyczaj kilka cech wspólnych: zaczynają od pojedynczego wzorca dostosowanego do scenariusza o największej wartości dla firmy, szybko udowadniają jego przydatność (przeważnie w ciągu 3 – 6 miesięcy), a następnie wprowadzają kolejne wzorce w miarę rozwoju kompetencji w zakresie AI. Takie przedsiębiorstwa nie czekają na idealne rozwiązania, tylko budują, uczą się i iteracyjnie powtarzają ten proces.

To iteracyjne podejście do wdrażania architektury odzwierciedla głębszą różnicę między projektami generatywnej AI a tradycyjnymi projektami uczenia maszynowego. Aby z powodzeniem wprowadzić opisane tu wzorce architektoniczne, organizacje muszą rozumieć, w jaki sposób cykle rozwoju generatywnej AI różnią się od tradycyjnych podejść stosowanych w uczeniu maszynowym. Różnice w procesach mają bezpośredni wpływ na to, jak i kiedy należy wdrażać poszczególne wzorce architektoniczne.

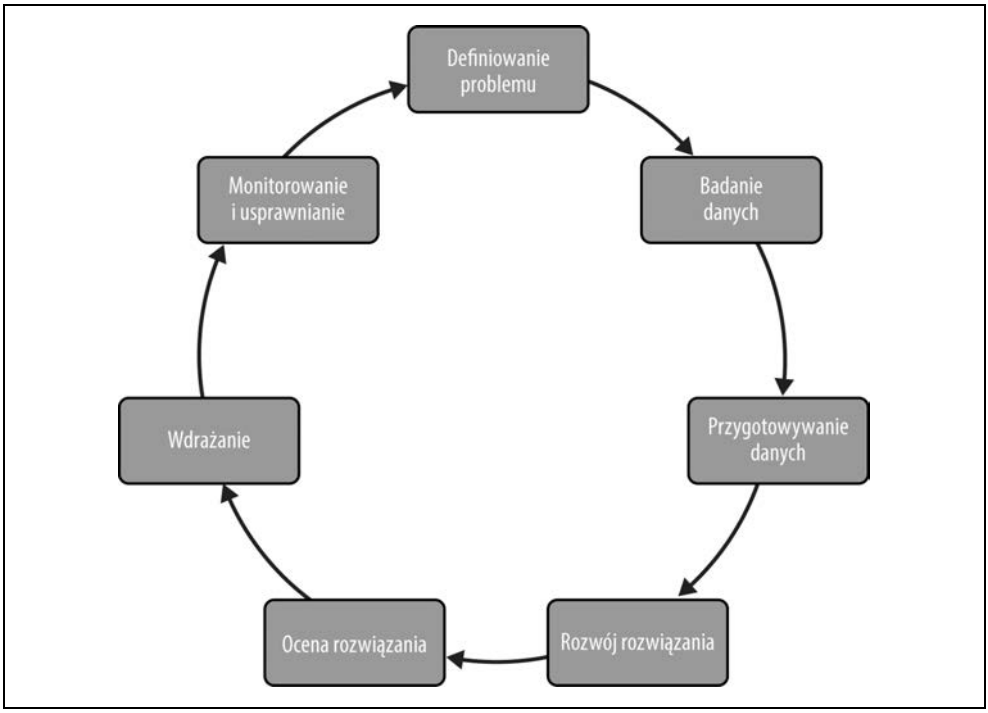
Cykle rozwoju rozwiązań: uczenie maszynowe a generatywna AI

Zrozumienie podstawowych różnic między cyklami rozwoju rozwiązań w tradycyjnym uczeniu maszynowym a w generatywnej AI jest niezbędne dla organizacji, które chcą przejść od fazy eksperymentów do wdrożeń produkcyjnych. Różnice te tłumaczą, dlaczego praktyki sprawdzone w uczeniu maszynowym często zawodzą przy wdrażaniu generatywnej AI.

Tradycyjny cykl rozwoju rozwiązań z obszaru uczenia maszynowego

Tradycyjny cykl tworzenia modeli uczenia maszynowego, pokazany na rysunku 1.11, przebiega według dobrze znanego wzorca, który ewoluował przez dziesięciolecia tworzenia praktycznych rozwiązań.

Kup książkę Poleć książkę



Rysunek 1.11. Tradycyjny cykl tworzenia systemów uczenia maszynowego

Oto kroki tego procesu:

1. *Definiowanie problemu*. Na tym początkowym etapie zespoły jasno określają problem biznesowy i definiują konkretne, mierzalne skutki. Ustanawia to zakres i kryteria sukcesu dla całego projektu. Kluczowe wymagania na tym etapie obejmują:
   - jasno zdefiniowany problem biznesowy z konkretnymi wskaźnikami;
   - wąski zakres ograniczony do pojedynczego zadania (na przykład predykcji lub klasyfikacji);
   - jednoznacznie zdefiniowane wejścia i wyjścia;
   - ściśle określone kryteria sukcesu oparte na dokładności, precyzji, kompletności itp.
2. *Badanie danych.* Podczas tej fazy, związanej z odkrywaniem, danolodzy eksplorują dostępne źródła danych, poznają ich strukturę oraz oceniają jakość i adekwatność danych w kontekście zdefiniowanego problemu. Oto najważniejsze działania na tym etapie:
   - analizowanie ustrukturyzowanych danych z systemów korporacyjnych;
   - identyfikowanie i wybieranie odpowiednich cech;
   - przeprowadzanie analizy statystycznej w celu zrozumienia rozkładów danych;
   - definiowanie zbiorów danych treningowych, walidacyjnych i testowych.
3. *Przygotowywanie danych.* Ten krytyczny etap wstępnego przetwarzania ma na celu przekształcenie surowych danych na format odpowiedni dla algorytmów uczenia maszynowego.

Kup książkę Poleć książkę

To w tej fazie rozwiązywane są problemy z jakością danych i powstają zoptymalizowane dane wejściowe. Oto najważniejsze działania z tego etapu:

- inżynieria cech w celu utworzenia danych wejściowych modelu;
- oczyszczanie i normalizacja danych;
- radzenie sobie z brakującymi wartościami i wartościami odstającymi;
- tworzenie zrównoważonych zbiorów danych na potrzeby procesu uczenia.

4. *Rozwój rozwiązania*. W fazie rozwoju modelu danolodzy wybierają i wdrażają odpowiednie algorytmy, dostrajają parametry i iteracyjnie zwiększają skuteczność rozwiązania na podstawie wyników procesu walidacji. Ten etap obejmuje następujące działania:
   - wybór odpowiednich algorytmów (lasy losowe, wzmacnianie gradientowe itp.);
   - dostrajanie hiperparametrów i optymalizacja;
   - uczenie modelu na podstawie zbiorów danych z etykietami;
   - iteracyjne usprawnianie modelu na podstawie wyników walidacji.
5. *Ocenianie rozwiązania.* Na etapie oceny należy rygorystycznie przetestować skuteczność modelu względem obiektywnych wskaźników przy użyciu wyodrębnionych danych. Celem jest uzyskanie niezawodnego i skutecznego modelu. Oto najważniejsze zadania z tego etapu:
   - rygorystyczne testowanie rozwiązania z użyciem wyodrębnionych danych testowych;
   - pomiar skuteczności przy użyciu ustalonych wskaźników;
   - porównanie z modelami bazowymi;
   - testowanie istotności statystycznej.
6. *Wdrażanie.* Podczas wdrażania sprawdzony model jest przygotowywany do użytku w środowisku produkcyjnym i integrowany z istniejącymi systemami. Powstaje również dokumentacja dla zespołów operacyjnych. Oto najważniejsze działania na tym etapie:
   - serializacja i pakowanie modelu;
   - integracja z systemami produkcyjnymi;
   - konfiguracja wnioskowania wsadowego lub w czasie rzeczywistym;
   - tworzenie dokumentacji i przekazanie jej do działu operacyjnego.
7. *Monitorowanie i usprawnianie.* Ostatni, stale realizowany etap wymaga opracowania procesów ciągłego monitorowania w celu zapewnienia trwałej skuteczności i umożliwienia iteracyjnego wprowadzania usprawnień. Obejmuje on następujące działania:
   - monitorowanie skuteczności z uwzględnieniem ustalonych wskaźników;
   - wykrywanie dryfu i ponowne uczenie modelu;
   - testy A/B usprawnień modelu;
   - wprowadzenie pętli sprzężenia zwrotnego w celu ciągłego doskonalenia rozwiązania.

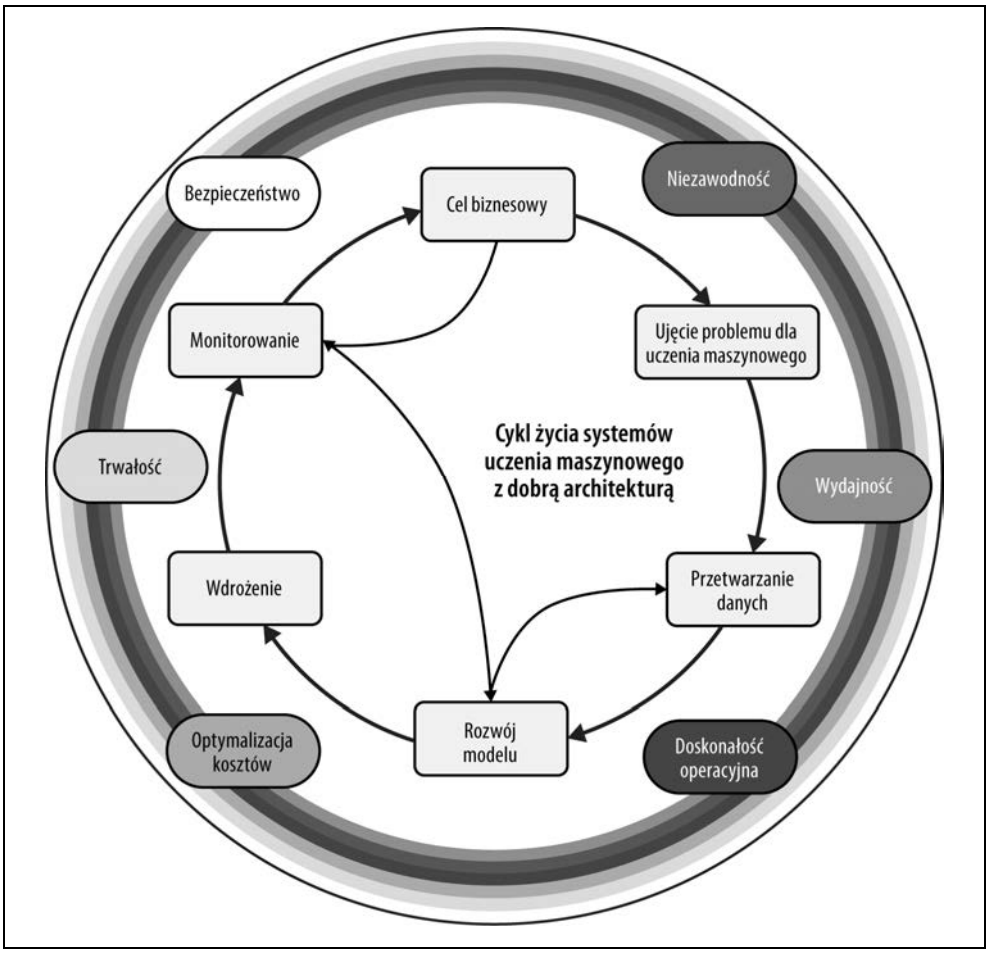
Ten tradycyjny cykl jest dobrze dostosowany do zadań predykcyjnych i klasyfikacyjnych z jasno zdefiniowanymi danymi wejściowymi i wyjściowymi. Opiera się w dużej mierze na ustrukturyzowanych danych i bezpośredniej inżynierii cech, a sukces jest mierzony za pomocą obiektywnych

Kup książkę Poleć książkę

wskaźników działania. Proces ten ma z natury charakter iteracyjny, a wnioski z etapu monitorowania i usprawniania są wykorzystywane do ponownego definiowania problemu w celu ciągłego udoskonalania rozwiązania.

Cykl rozwoju rozwiązań opartych na generatywnej AI

Cykl rozwoju systemów opartych na generatywnej AI, przedstawiony na rysunku 1.12, obejmuje zasadniczo inne wzorce i wymagania.



Rysunek 1.12. Cykl rozwoju w generatywnej AI

Także ten cykl można podzielić na siedem etapów:

1. *Definiowanie scenariusza użycia i wybór modelu.* Pierwszy etap obejmuje przede wszystkim definiowanie ogólnych celów (na przykład „poprawa obsługi klienta") i wybór odpowiednich modeli bazowych (GPT, Claude, Llama itp.), które są zgodne z wymaganiami scenariusza użycia i ograniczeniami organizacyjnymi. Dodatkowe zadania obejmują:

Kup książkę Poleć książkę

— rozważenie możliwości, ograniczeń i uprzedzeń modelu;
— ocenę sposobów hostingu (usługi API lub hosting samodzielny);
— zdefiniowanie kryteriów sukcesu wykraczających poza tradycyjne wskaźniki.

2. *Integracja wiedzy i inżynieria kontekstu.* Ten unikalny dla generatywnej AI etap obejmuje łączenie korporacyjnych źródeł wiedzy z modelem bazowym. Dzięki temu można sprawić, że istotne informacje będą odpowiednio reprezentowane i dostępne. To na tym etapie ważnych jest pięć wzorców architektonicznych opisanych w podrozdziale „Jakie rozwiązania są obecnie rozwijane w wiodących organizacjach?". Oto najważniejsze zadania z tego etapu:
   — zidentyfikowanie korporacyjnych źródeł wiedzy;
   — opracowanie strategii przetwarzania dokumentów i ich podziału;
   — generowanie osadzeń i wybór wektorowej bazy danych;
   — wzbogacanie metadanych i projektowanie strategii wyszukiwania;
   — integracja ustrukturyzowanych i nieustrukturyzowanych źródeł danych.
3. *Inżynieria promptów i projektowanie systemu.* Ta krytyczna faza projektowania obejmuje przede wszystkim tworzenie skutecznych instrukcji dla modelu i wprowadzanie struktur gwarantujących odpowiednie odpowiedzi i zabezpieczenia. Obejmuje ona:
   — projektowanie skutecznych promptów i instrukcji systemowych;
   — tworzenie przykładów typu few-shot i szablonów;
   — opracowywanie mechanizmów oceny i filtrowania;
   — integrację z zewnętrznymi narzędziami i API;
   — implementowanie mechanizmów zabezpieczających i środków bezpieczeństwa.
4. *Dostrajanie i dostosowywanie (opcjonalne).* W tym opcjonalnym, ale ważnym etapie modele bazowe są dostosowywane do określonych dziedzin lub zadań za pomocą dodatkowej nauki wyspecjalizowanej wiedzy. Ten etap obejmuje:
   — tworzenie zbiorów danych na potrzeby dostrajania modelu;
   — wybór odpowiednich technik dostrajania;
   — wykorzystanie metod dostrajania efektywnych ze względu na parametry (LoRA, QLoRA);
   — ocenę skuteczności dostrojonego modelu;
   — analizę kompromisów między podejściem opartym na dostrajaniu a metodą RAG.
5. *Ocena i zgodność z celami.* Ta faza obejmuje zarówno obiektywne wskaźniki, jak i ocenę dokonywaną przez ludzi, aby zagwarantować, że wyniki spełniają standardy jakości, bezpieczeństwa i zgodności z celami w różnych scenariuszach. Obejmuje to:
   — ocenę wyników przez ludzi pod kątem jakości i adekwatności;
   — symulowane ataki w celu wykrywania luk w bezpieczeństwie i ochronie;
   — zapewnianie zgodności z wartościami i wytycznymi organizacji;
   — testowanie w różnych scenariuszach i przypadkach brzegowych;
   — ocena wskaźników halucynacji i precyzji merytorycznej.

Kup książkę Poleć książkę

6. *Wdrażanie i integracja.* Na etapie wdrażania najważniejsza jest integracja systemu generatywnej AI z istniejącymi przepływami pracy przy jednoczesnym wprowadzeniu odpowiedniego monitorowania, mechanizmów informacji zwrotnej i szkoleń użytkowników. Najistotniejsze zadania z tej fazy obejmują:
   - integrację z istniejącymi przepływami pracy i systemami;
   - wdrożenie monitorowania i rejestrowania zdarzeń;
   - wprowadzenie mechanizmów gromadzenia informacji zwrotnych;
   - szkolenie użytkowników i zarządzanie zmianą;
   - wdrożenie ładu i kontroli zgodności z regulacjami.
7. *Ciągłe uczenie się i doskonalenie.* Ten ciągły etap podkreśla dynamiczny charakter rozwiązań w obszarze generatywnej AI, gdzie niezbędne są ciągłe gromadzenie informacji zwrotnych i regularne aktualizowanie zarówno systemu, jak i baz wiedzy. Ta faza obejmuje:
   - gromadzenie i analizę opinii użytkowników;
   - monitorowanie działania modelu i wzorców jego użytkowania;
   - identyfikowanie scenariuszy awarii i przypadków brzegowych;
   - regularne aktualizowanie baz wiedzy i kontekstu;
   - dostosowywanie rozwiązań do zmieniających się potrzeb i oczekiwań użytkowników.

Tabela 1.3 podsumowuje różnice strukturalne między cyklami rozwoju systemów tradycyjnego uczenia maszynowego i generatywnej AI.

Tabela 1.3. Najważniejsze różnice między cyklami życia systemów uczenia maszynowego i generatywnej AI

Aspekt	Cykl rozwoju tradycyjnych systemów uczenia maszynowego	Cykl rozwoju systemów generatywnej AI
Zakres problemu	Wąskie, dobrze zdefiniowane zadania predykcyjne	Szerokie, otwarte zadania z obszaru generowania i wnioskowania
Wymogi dotyczące danych	Ustrukturyzowane zbiory danych z etykietami	Różnorodne źródła wiedzy, zarówno ustrukturyzowane, jak i nieustrukturyzowane
Tworzenie modelu	Niestandardowe modele budowane od podstaw	Dostosowanie wstępnie wytrenowanych modeli bazowych
Nacisk w obszarze inżynierii	Inżynieria cech	Inżynieria kontekstu i projektowanie promptów
Metody oceny	Obiektywne wskaźniki (dokładność, F1 itd.)	Subiektywna ocena i weryfikacja z udziałem człowieka
Wzorzec wdrażania	Wdrażanie modelu statycznego z okresowym ponownym uczeniem	Systemy dynamiczne z ciągłym aktualizowaniem wiedzy
Kryteria sukcesu	Ilościowe miary skuteczności	Miary zadowolenia użytkowników i wartości biznesowej
Scenariusze awarii	Stopniowy spadek skuteczności	Katastrofalne awarie (halucynacje, szkodliwe wyniki)
Potrzeby w zakresie ładu	Dokumentowanie i monitorowanie modelu	Kompleksowe mechanizmy kontroli bezpieczeństwa, etyki i zgodności z celami biznesowymi

Kup książkę Poleć książkę

Implikacje dla infrastruktury danych

Różnice w cyklach rozwoju opisane w tabeli 1.3 mają bezpośredni wpływ na wymogi dotyczące infrastruktury danych, co prowadzi do poważnych wyzwań w organizacjach próbujących wykorzystać dawne inwestycje w uczenie maszynowe do zastosowań w obszarze generatywnej AI. Pięć wzorców architektonicznych, które wymieniliśmy wcześniej, pomaga uwzględnić te wyzwania:

Zarządzanie wiedzą a magazyny cech

W tradycyjnym uczeniu maszynowym wykorzystuje się magazyny cech zawierające ustrukturyzowane, wyodrębnione cechy, natomiast generatywna AI wymaga kompleksowych systemów zarządzania wiedzą, które zachowują kontekst i relacje. Pewna firma telekomunikacyjna odkryła, że jej starannie wyodrębnione cechy (na przykład „średni czas trwania połączenia") były nieodpowiednie dla asystenta obsługi klienta opartego na generatywnej AI, ponieważ pozbawiały dane bogactwa kontekstowego niezbędnego do udzielania pomocnych odpowiedzi.

Rozwiązanie: Grafy wiedzy (wzorzec 1.) zachowują relacje i kontekst, które są eliminowane w magazynach cech.

Statyczny a dynamiczny dostęp do danych

Modele uczenia maszynowego korzystają ze statycznych zbiorów danych, które zmieniają się tylko podczas okresowego ponownego uczenia. Natomiast generatywna AI wymaga aktualizowania danych w czasie rzeczywistym. System generatywnej AI globalnego banku nadal dostarczał nieaktualne informacje o zasadach przez tygodnie po aktualizacji przepisów, co groziło naruszeniami w zakresie zgodności z regulacjami.

Rozwiązanie: Architektury bazujące na zdarzeniach (wzorzec 2.) sprawiają, że agenty zawsze korzystają z aktualnych informacji.

Przetwarzanie wsadowe a interaktywne

Przepływy pracy w uczeniu maszynowym optymalizują przepustowość dzięki przetwarzaniu wsadowemu, ale generatywna AI wymaga interaktywnego przetwarzania danych z niskim opóźnieniem. Nocny potok przetwarzania danych w firmie produkcyjnej nie był dostosowany do asystenta utrzymania ruchu opartego na generatywnej AI, który wymagał czasu reakcji poniżej sekundy.

Rozwiązanie: Architektury typu lakehouse (wzorzec 3.) obsługują zarówno analitykę wsadową, jak i interaktywne zadania wykonywane przez AI.

Ustrukturyzowane a kontekstowe magazyny danych

Dane w tradycyjnych systemach uczenia maszynowego są zoptymalizowane pod kątem przechowywania w formie tabelarycznej, ale generatywna AI wymaga nieustrukturyzowanych, bogatych w kontekst informacji. Systemy dostawcy usług opieki zdrowotnej przechowywały kody diagnoz, ale nie potrafiły rejestrować kontekstu z opisowej dokumentacji medycznej pacjentów.

Rozwiązanie: Wyszukiwanie semantyczne (wzorzec 4.) umożliwia pobieranie informacji na podstawie znaczenia, a nie tylko struktury.

Kup książkę Poleć książkę

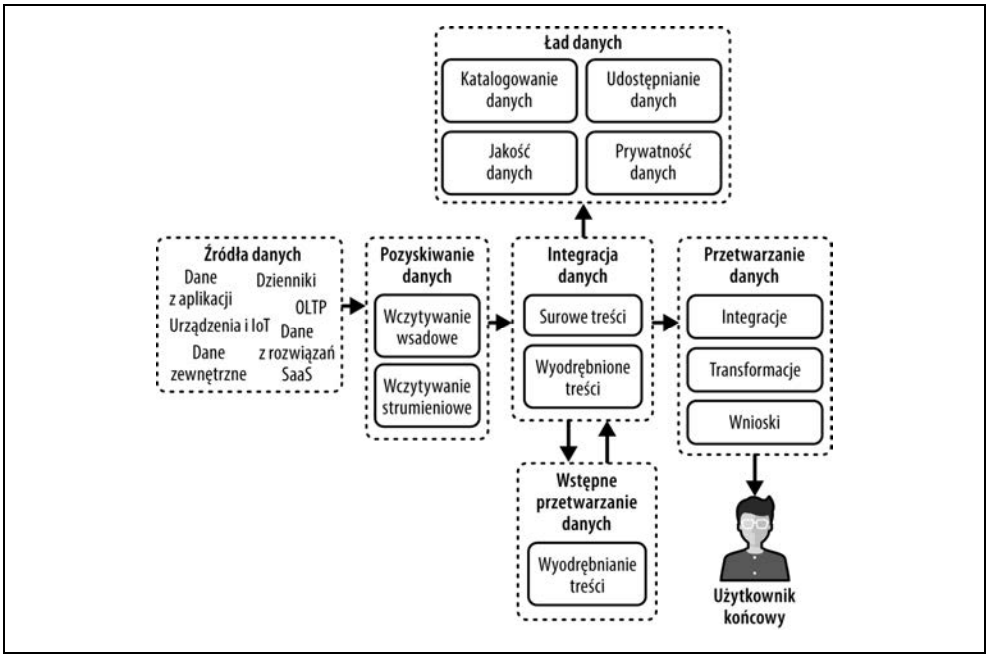
Usprawnienia oparte na wskaźnikach a udoskonalanie oparte na informacjach zwrotnych

Systemy uczenia maszynowego są doskonalone na podstawie optymalizacji wskaźników, ale systemy generatywnej AI ewoluują dzięki stale napływającym informacjom zwrotnym od użytkowników i ciągłemu uczeniu. System rekomendacji w firmie handlowej poprawił się według współczynników konwersji, ale asystent zakupowy bazujący na generatywnej AI musiał uwzględniać subiektywne opinie i stale aktualizować wiedzę o produktach.

Rozwiązanie: Architektury dostosowane do agentów (wzorzec 5.) obejmują pętle informacji zwrotnej i mechanizmy ciągłego uczenia.

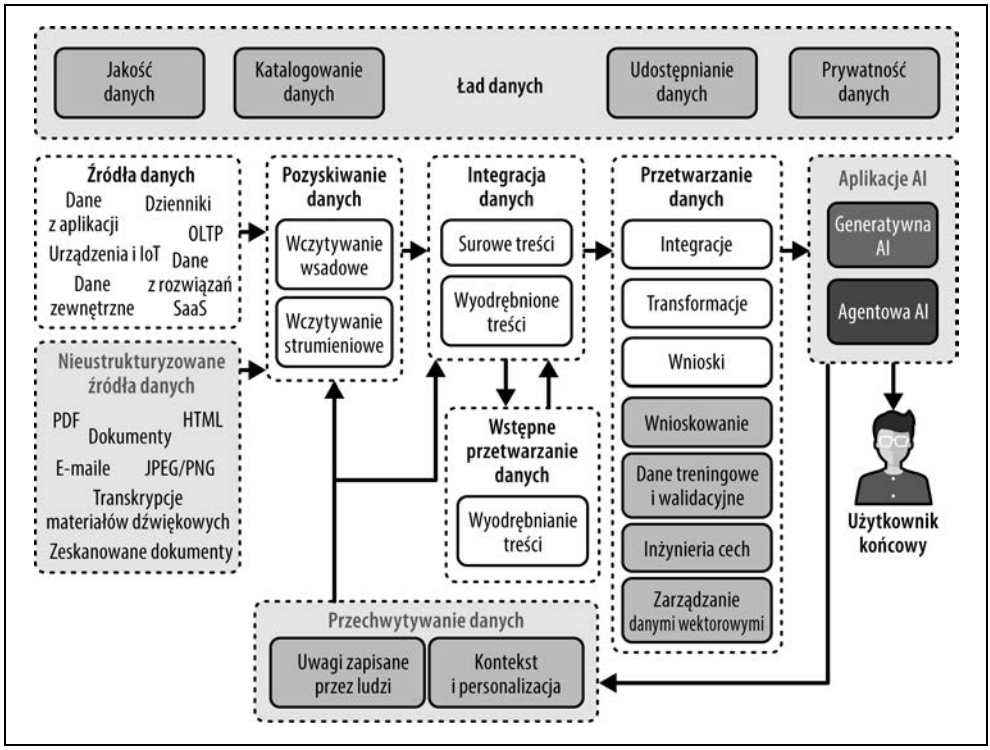
Ewolucja architektury w praktyce: od tradycyjnych procesów ETL do potoków przystosowanych do generatywnej AI

Aby w pełni uwolnić transformacyjny potencjał generatywnej AI, przedsiębiorstwa muszą wyjść poza tradycyjne paradygmaty sztucznej inteligencji, uczenia maszynowego i big data przedstawione na rysunkach 1.13 i 1.14. Generatywna AI wymaga zupełnie nowego podejścia do przechowywania danych, ich przetwarzania, wyszukiwania i zarządzania nimi. Wykorzystanie potencjału tej technologii wymaga zmiany myślenia. Trzeba przejść od ustrukturyzowanych potoków dostosowanych do przetwarzania wsadowego do elastycznych, działających w czasie rzeczywistym architektur kontekstowych zaprojektowanych z myślą o nieustrukturyzowanych i dynamicznych treściach.



Rysunek 1.13. Tradycyjna podstawowa struktura danych

Kup książkę Poleć książkę



Rysunek 1.14. Podstawowa struktura danych dla generatywnej AI

Najważniejsze różnice między podstawową warstwą danych w tradycyjnych rozwiązaniach i w generatywnej AI

Zanim przejdziemy do praktycznego przykładu, warto pokrótce przypomnieć najważniejsze różnice. Tabela 1.4 zawiera ich zestawienie.

Tabela 1.4. Najważniejsze różnice między podstawową warstwą danych w tradycyjnych rozwiązaniach i w generatywnej AI

Aspekt	Tradycyjne systemy	Generatywna AI
Typy danych	Głównie ustrukturyzowane, czasem nieustrukturyzowane	W dużej mierze nieustrukturyzowane ze złożonymi relacjami
Wymagania dotyczące kontekstu	Kontekst ograniczony, często przetwarzanie pojedynczych rekordów	Bogate relacje kontekstowe między źródłami danych
Wzorce zapytań	Przewidywalne, bezpośrednio tworzone zapytania	Wyszukiwanie semantyczne, oparte na podobieństwie i uwzględniające kontekst
Tryb przetwarzania	Głównie wsadowe z elementami strumieniowego	W czasie rzeczywistym, interaktywne, z dynamicznymi oknami kontekstowymi
Paradygmat przechowywania	Oparty na tabelach (wiersze i kolumny)	Oparty na wektorach z relacjami semantycznymi

Kup książkę Poleć książkę

Tabela 1.4. Najważniejsze różnice między podstawową warstwą danych w tradycyjnych rozwiązaniach i w generatywnej AI (ciąg dalszy)

Aspekt	Tradycyjne systemy	Generatywna AI
Częstotliwość aktualizacji	Okresowe odświeżanie	Ciągła integracja nowej wiedzy
Wymagania dotyczące schematu	Wymagane sztywne schematy	Elastyczne schematy z naciskiem na osadzenia
Zakres ładu	Kontrola dostępu i zgodność z regulacjami	Etyka, wykrywanie uprzedzeń i odpowiedzialne użytkowanie
Metoda wyszukiwania	Dokładne dopasowanie oparte na słowach kluczowych	Oparte na podobieństwie, zrozumienie semantyczne
Wymiar skalowania	Ilość (petabajty ustrukturyzowanych danych)	Kontekst (utrzymywanie relacji w dużej skali)

Przykład praktyczny: ewolucja od modelu Kimballa do architektury medalionowej i generatywnej AI

W jaki sposób organizacje w praktyce wprowadzają zmiany architektoniczne niezbędne do odniesienia sukcesu w obszarze generatywnej AI? Przeanalizujemy teraz studium przypadku, które obrazuje transformację tego typu.

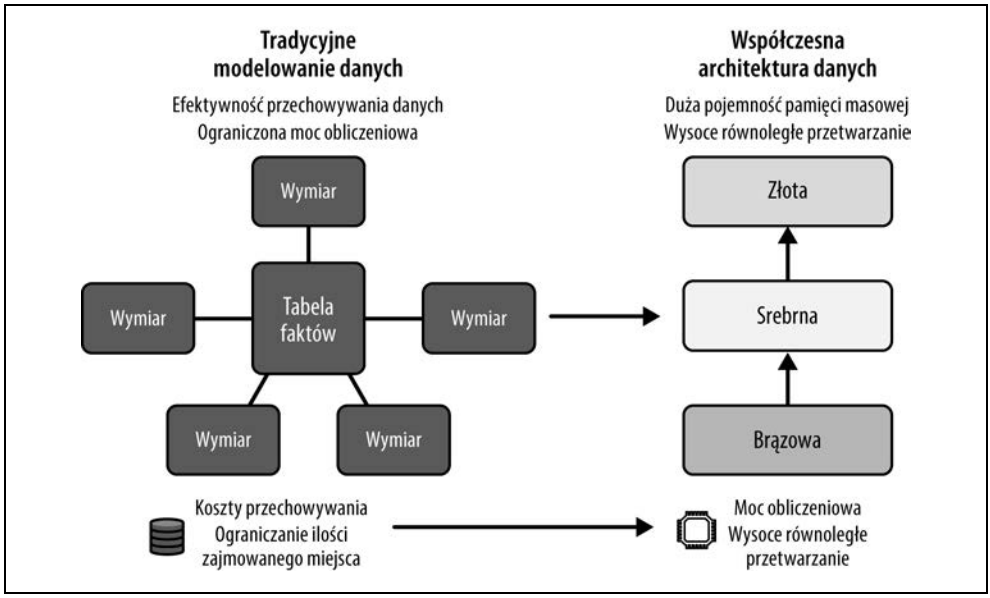
Wyzwanie: tabele faktów z miliardami wierszy w erze generatywnej AI

Globalny sprzedawca detaliczny, którego historia transakcji przekracza 3 miliardy wierszy, przeprowadził migrację z tradycyjnego schematu gwiazdy Kimballa do architektury medalionowej z warstwami danych: brązową (dane surowe), srebrną (dane oczyszczone i ustandaryzowane) i złotą (dane wyselekcjonowane), co obrazuje rysunek 1.15. Jednocześnie rozwijał swoje pierwsze aplikacje generatywnej AI.

Istniejący model wyglądał następująco:

```
-- Tradycyjny model wymiarowy z odrębnymi tabelami faktów i wymiarów
CREATE TABLE gold.fact_sales (
  transaction_id BIGINT,
  date_key INT,
  product_key INT,
  store_key INT,
  customer_key INT,
  quantity INT,
  amount DECIMAL(12,2)
);
CREATE TABLE gold.dim_product (
  product_key INT,
  product_id VARCHAR(50),
  product_name VARCHAR(100),
  category VARCHAR(50),
  subcategory VARCHAR(50),
  brand VARCHAR(50)
);
```


Kup książkę Poleć książkę



Rysunek 1.15. Przejście z tradycyjnego schematu gwiazdy do architektury medalionowej

```
CREATE TABLE gold.dim_store (
  store_key INT,
  store_id VARCHAR(20),
  store_name VARCHAR(100),
  region_id VARCHAR(10),
  city VARCHAR(50),
  state VARCHAR(2)
);
```

To tradycyjne podejście, choć skuteczne w przypadku raportowania dotyczącego danych historycznych, wiązało się z kilkoma wyzwaniami przy tworzeniu aplikacji generatywnej AI:

- Opisy produktów i szczegółowe informacje znajdowały się wyłącznie w ustrukturyzowanych polach.
- Brakowało semantycznego zrozumienia relacji między produktami.
- Nie obsługiwano zapytań w języku naturalnym na temat produktów.
- System nie umożliwiał obsługi danych w postaci obrazów ani bogatych treści.
- Aktualizacje były wprowadzane za pomocą przetwarzania wsadowego, co mogło prowadzić do używania nieaktualnych danych.

Rozwiązanie: podejście hybrydowe dostosowane do generatywnej AI

Organizacja wdrożyła podejście hybrydowe, dzięki czemu zachowało podstawowe zasoby wymiarowe, co pozwoliło utrzymać spójne definicje biznesowe w całym przedsiębiorstwie. Jednocześnie dodano struktury danych zoptymalizowane pod kątem generatywnej AI:

Kup książkę Poleć książkę

- specjalnie zbudowane, zdenormalizowane tabele faktów dostosowane do określonych domen biznesowych;
- projekty zoptymalizowane zarówno pod kątem analityki z udziałem człowieka, jak i do wykorzystania przez AI.

Na potrzeby wdrożenia utworzono dwa typy zasobów w złotej warstwie:

```
-- Dla danych w limicie silnika SPICE (< 1 miliard wierszy)
CREATE TABLE gold.daily_aggregated AS
SELECT
  date_key,
  region_id,
  product_category,
  SUM(amount) as daily_total,
  COUNT(DISTINCT customer_id) as customer_count
FROM silver.transactions
GROUP BY 1,2,3;
-- Dla szczegółowych danych wykraczających poza limit silnika SPICE
-- Korzystaj z trybu zapytań bezpośrednich
CREATE VIEW gold.transaction_details AS
SELECT * FROM silver.transactions;
```

Aby uzyskać płaskie dane na potrzeby sztucznej inteligencji, utworzono zdenormalizowane widoki specjalnie dla aplikacji generatywnej AI:

```
CREATE TABLE gold.store_performance AS
SELECT
  t.transaction_id,
  t.transaction_date,
  t.amount,
  p.product_name,
  p.category,
  p.subcategory,
  s.store_name,
  s.region_id,
  s.city,
  s.state
FROM silver.transactions t
JOIN silver.products p ON t.product_id = p.product_id
JOIN silver.stores s ON t.store_id = s.store_id;
```

Architektura potoku danych rozszerzona o obsługę generatywnej AI

Aby dodać obsługę aplikacji generatywnej AI, w organizacji wdrożono kompleksowy potok danych, który wykracza poza tradycyjne procesy ETL. W tym potoku zaimplementowano wszystkich pięć wzorców przedstawionych wcześniej:

```
# Kod w Pythonie dla potoku danych dostosowanego do generatywnej AI w branży handlowej
# 1. Wyodrębnianie kompleksowych danych na temat produktu i transakcji
def extract_retail_data():
    # Wyodrębnianie ustrukturyzowanych danych z tradycyjnych źródeł
    structured_data = extract_from_database("SELECT * FROM source.products")
    # Wyodrębnianie nieustrukturyzowanych danych z wielu źródeł
    product_descriptions = extract_from_cms("product_descriptions")
    product_images = extract_from_asset_manager("product_images")
    customer_reviews = extract_from_reviews_system("customer_reviews")
```


Kup książkę Poleć książkę

```
    social_media_mentions = extract_from_social_platforms("brand_mentions")
    return {
        "structured": structured_data,
        "descriptions": product_descriptions,
        "images": product_images,
        "reviews": customer_reviews,
        "social": social_media_mentions
    }
# 2. Przetwarzanie i wzbogacanie danych na potrzeby generatywnej AI
def process_for_genai(raw_data):
    # Przetwarzanie danych ustrukturyzowanych
    products_df = process_structured_data(raw_data["structured"])
    # Generowanie osadzeń tekstowych reprezentujących opisy produktów (wzorzec 4., wyszukiwanie
    # semantyczne)
    description_embeddings = generate_embeddings( raw_data["descriptions"],
        model="text-embedding-ada-002"
    )
    # Wyodrębnianie cech ze zdjęć produktów (wzorzec 3.,
    # wielomodalna architektura typu lakehouse)
    image_features = process_images_with_vision_model(raw_data["images"])
    # Analiza sentymentu i wyodrębnianie motywów na podstawie recenzji
    review_insights = analyze_reviews_with_nlp(raw_data["reviews"])
    # Przetwarzanie informacji z mediów społecznościowych w celu oceny sentymentu wobec marki
    social_sentiment = analyze_social_sentiment(raw_data["social"])
    # Tworzenie ujednoliconych osadzeń dla produktów
    unified_embeddings = create_multimodal_embeddings(
        text=description_embeddings,
        images=image_features,
        reviews=review_insights,
        social=social_sentiment
    )
    return {
        "structured_data": products_df,
        "embeddings": unified_embeddings,
        "metadata": extract_metadata(raw_data),
        "relationships": identify_product_relationships(products_df, \
        unified_embeddings)
    }
# 3. Zapisywanie danych z użyciem hybrydowej architektury
# zoptymalizowanej na potrzeby generatywnej AI
def store_genai_data(processed_data):
    # Zapisywanie ustrukturyzowanych danych w tradycyjnej hurtowni (na potrzeby generowania raportów)
    store_in_warehouse(processed_data["structured_data"])
    # Zapisywanie osadzeń w wektorowej bazie danych (wzorzec 4., wyszukiwanie semantyczne)
    store_in_vector_db(
        embeddings=processed_data["embeddings"],
        metadata=processed_data["metadata"]
    )
    # Zapisywanie relacji w grafowej bazie danych (wzorzec 1., grafy wiedzy)
    store_in_graph_db(processed_data["relationships"])
    # Aktualizowanie indeksu wyszukiwania w czasie rzeczywistym na potrzeby aplikacji AI
    update_search_index(processed_data)
    # Tworzenie grafu wiedzy na potrzeby złożonego wnioskowania (wzorzec 1.)
    update_knowledge_graph(processed_data["relationships"])
# 4. Implementacja aktualizacji w czasie rzeczywistym na potrzeby
# aplikacji generatywnej AI
def setup_realtime_genai_pipeline():
```


Kup książkę Poleć książkę

```
# Konfigurowanie odbiorników zdarzeń do aktualizowania danych o produktach
# (wzorzec 2., architektura oparta na zdarzeniach)
setup_event_listeners([
    "product_updates",
    "inventory_changes",
    "price_updates",
    "new_reviews",
    "social_mentions"
])
# Implementowanie metody CDC dla danych ustrukturyzowanych (wzorzec 2.)
setup_cdc_pipeline()
# Monitorowanie aktualizacji opisów w systemie zarządzania treścią
monitor_cms_changes()
```

Skutki i korzyści

Gdy później wdrożono asystenta AI przeznaczonego do odpowiadania na pytania biznesowe, podejście hybrydowe okazało się nieocenione:

- Sztuczna inteligencja potrafiła skutecznie uzyskać dostęp do danych specyficznych dla dziedziny, a przy tym zachowano spójne definicje w całym przedsiębiorstwie.
- Dzięki wyszukiwaniu semantycznemu (wzorzec 4.) możliwe stało się odpowiadanie na zapytania w języku naturalnym, na przykład: „Pokaż mi produkty podobne do naszych najlepiej sprzedających się kurtek zimowych".
- Kontekst w postaci określanych w czasie rzeczywistym stanów magazynowych umożliwił sztucznej inteligencji dostarczanie dokładnych informacji o dostępności (wzorzec 2.).
- Uzyskano wnioski wynikające z uwzględnienia różnych domen. Stało się tak dzięki połączeniu danych o produktach z opiniami klientów i nastrojami w mediach społecznościowych (wzorzec 1.).

Na przykład gdy kierownik sklepu zapytał asystenta AI: „Dzisiejsza sprzedaż jest o 15% niższa od prognozy — co powinienem zrobić?", system mógł odpowiedzieć: „Dzisiejszy wzorzec sprzedaży jest podobny do tego, co zaobserwowano w zeszłym kwartale podczas zakłóceń w łańcuchu dostaw. Z danych historycznych wynika, że promowanie produktów komplementarnych w kategoriach X i Y zazwyczaj pozwala odzyskać 8 – 12 procent spadku. Obecne poziomy zapasów pozwalają zrealizować tę strategię".

Strategie skalowania w erze AI

W miarę jak wolumeny danych sklepu detalicznego zwiększały się z miliardów do bilionów wierszy, wdrożono kilka strategii w celu utrzymania wydajności:

Modelowanie oparte na dziedzinach

Tworzenie w warstwie złotej oddzielnych zbiorów danych dla zespołów merchandisingu, operacji sklepowych i finansów (wzorzec 1., ewolucja siatki danych).

Przemyślany podział

Podział danych transakcyjnych według dat i regionów (wzorzec 3.).

Kup książkę Poleć książkę

Warstwy agregacji

Budowanie dla typowych zapytań zagregowanych wyników dziennych, tygodniowych i miesięcznych.

Wzbogacanie metadanych

Dodawanie jednoznacznych opisów biznesowych do wszystkich pól (wzorzec 5.).

Optymalizacja wektorów

Wdrażanie hierarchicznych indeksów wektorowych w celu szybszego wyszukiwania na podstawie podobieństw (wzorzec 4.).

Strategie buforowania

Zastosowanie inteligentnego buforowania często używanych osadzeń (wzorzec 2.).

Dlaczego ma to znaczenie? Podstawowa architektura danych umożliwiająca sukces w obszarze AI

Opisane podejście sprawiło, że logika biznesowa organizacji była dostępna zarówno dla tradycyjnych narzędzi z obszaru analityki biznesowej, jak i dla aplikacji generatywnej AI. Przejrzystość wynikająca z umieszczenia obliczeń w warstwie transformacji danych, a nie w procedurach bazodanowych, pomogła ulepszyć dokumentację, a ostatecznie także poprawić dokładność pulpitów nawigacyjnych i wniosków generowanych przez sztuczną inteligencję.

Przeniesienie złożonej logiki biznesowej z nieprzejrzystych procedur do transparentnych transformacji w warstwie złotej pozwoliło stworzyć jedno spójne źródło prawdy, z którego mogły korzystać zarówno narzędzia do wizualizacji, jak i modele AI.

W miarę rozszerzania wdrożenia generatywnej AI u sprzedawcy odkryto, że decyzje dotyczące architektury danych miały dalekosiężne konsekwencje:

- *Jakość danych stała się najważniejsza.* Systemy AI potęgowały wszelkie niespójności lub błędy w danych.
- *Zachowanie kontekstu było konieczne.* Utrzymywanie relacji między elementami danych umożliwiło bardziej wyrafinowane wnioskowanie przez AI (wzorzec 1.).
- *Niezbędne stały się aktualizacje w czasie rzeczywistym.* Przestarzałe dane prowadziły do nieaktualnych odpowiedzi AI, co podważało zaufanie użytkowników (wzorzec 2.).
- *Bogate metadane umożliwiły poprawę działania AI.* Szczegółowo udokumentowane dane pomogły systemom AI udzielać dokładniejszych i bardziej trafnych odpowiedzi (wzorzec 5.).

Opisany tu proces ewolucji od tradycyjnego modelu wymiarowego Kimballa do architektury dostosowanej do generatywnej AI pokazuje praktyczne kroki, jakie organizacje muszą wykonać, aby zniwelować lukę między eksperymentalnymi aplikacjami AI a systemami gotowymi do wdrożenia produkcyjnego, które dostarczają rzeczywistą wartość biznesową. Co ważniejsze, proces ten obrazuje, jak pięć wzorców architektonicznych wprowadzonych w tym rozdziale współdziała ze sobą w praktyce, by powstała kompleksowa bazowa architektura danych dla generatywnej AI.

Kup książkę Poleć książkę

Przygotowanie się na przyszłość z wiodącą rolą agentów

Chociaż omówione dotychczas techniki są dostosowane do obecnych wymogów generatywnej AI, to gdy spojrzymy poza dzisiejsze jej wdrożenia, w kierunku przyszłości w coraz większym stopniu kształtowanej przez autonomiczne agenty AI, organizacje muszą przygotować infrastrukturę danych na bardziej fundamentalne zmiany w sposobie dostępu do informacji, ich przetwarzania i podejmowania działań na ich podstawie. W tym podrozdziale omawiamy główne trendy, które będą definiować dostosowanie danych do świata sterowanego przez agenty. Ponadto przedstawiamy wskazówki dla organizacji dążących do zdobycia korzystnej pozycji w tym nowym otoczeniu. Omówione wcześniej wzorce architektoniczne mogą pomóc także w przeprowadzeniu tej transformacji.

„Jeeves robi zakupy" — doświadczenia użytkowników korzystających z agentów

Jednym z najbardziej przełomowych scenariuszy wynikających z wdrażania agentowej AI jest coś, co można nazwać paradygmatem „Jeeves robi zakupy". W tym modelu agenty AI pełnią rolę pośredników między konsumentami a firmami, co skutkuje całkowitą zmianą sposobu odkrywania, oceny i zakupu produktów i usług.

W tym scenariuszu:

- Konsumenci delegują podejmowanie decyzji zakupowych agentom AI, które rozumieją ich preferencje, ograniczenia budżetowe i potrzeby.
- Agenty negocjują bezpośrednio z firmami w imieniu konsumentów i porównują oferty wielu dostawców.
- Tradycyjny marketing staje się mniej skuteczny, ponieważ agenty filtrują treści promocyjne na podstawie obiektywnych kryteriów.
- Cechy produktów i usług stają się wymogami czytelnymi dla maszyn, uwzględnianymi zamiast argumentów emocjonalnych.
- Przejrzystość cen staje się znacznie większa, ponieważ agenty potrafią błyskawicznie porównywać oferty różnych dostawców.

Dla organizacji, szczególnie w branżach handlu detalicznego i usług konsumenckich, ta zmiana wymaga kompleksowego przemyślenia gotowości danych w różnych wymiarach:

- *Informacje o produkcie.* Wszystkie atrybuty produktu muszą być ustrukturyzowane w formatach czytelnych dla maszyn, tak by agenty mogły analizować i porównywać dane (wzorzec 5.).
- *Modele cenowe.* Systemy dynamicznego ustalania cen muszą być zaprojektowane w taki sposób, by można było modyfikować ceny na podstawie interakcji z agentami AI za pośrednictwem standardowych interfejsów (wzorzec 5.).
- *Systemy zarządzania magazynem.* Agenty muszą mieć dostęp do danych o stanach magazynowych w czasie rzeczywistym, aby zapobiegać błędom w realizacji zamówień (wzorzec 2.).

Kup książkę Poleć książkę

- *Umowy o gwarantowanym poziomie świadczenia usług (SLA).* Gwarancje poziomu działania muszą być sformalizowane w taki sposób, by agenty mogły je zweryfikować i wyegzekwować (wzorzec 5.).

SEO: od słów kluczowych do optymalizacji pod kątem agentów

W miarę jak agenty AI będą w coraz większym stopniu pośredniczyć w odkrywaniu informacji i podejmowaniu decyzji, tradycyjna optymalizacja pod kątem wyszukiwarek (ang. *search engine optimization* — SEO) przekształci się w nową dyscyplinę, którą można nazwać *optymalizacją pod kątem agentów* (ang. *agent optimization* — AO).

Ta transformacja całkowicie zmieni sposób, w jaki organizacje muszą strukturyzować dane, aby były one łatwe do znalezienia i adekwatne, co pokazaliśmy w tabeli 1.5.

Tabela 1.5. Tradycyjna SEO a optymalizacja pod kątem agentów

Tradycyjna SEO	Optymalizacja pod kątem agentów
Optymalizacja słów kluczowych	Optymalizacja danych ustrukturyzowanych
Czytelność treści dla ludzi	Standaryzacja atrybutów czytelnych dla maszyn
Wartość linków przychodzących	Weryfikowalne potwierdzenie zaufania
Wskaźniki zaangażowania użytkowników	Obiektywne wskaźniki działania
Odwołania do emocji	Wymierne korzyści biznesowe

Ta zmiana wymaga od organizacji:

- wdrożenia kompleksowych znaczników dla danych ustrukturyzowanych przy użyciu standardów takich jak Schema.org i ontologii specyficznych dla branży (wzorzec 5.);
- opracowania czytelnych dla maszyn wskaźników działania, które mogą być niezależnie weryfikowane przez agenty;
- utworzenia na potrzeby interakcji z agentami ustandaryzowanych interfejsów, które umożliwiają zautomatyzowane porównywanie danych i prowadzenie negocjacji (wzorzec 5.);
- wprowadzenia weryfikowalnych mechanizmów oceny zaufania, których agenty mogą używać do oceny niezawodności i jakości (wzorzec 1.);
- utrzymania doskonałej higieny danych, ponieważ niespójności, które ludzie mogliby przeoczyć, zdyskwalifikują ofertę z perspektywy agenta.

Agenty nie są lojalne wobec marek — przygotowania do ery radykalnej przejrzystości

Być może najpoważniejszą implikacją rozwoju świata sterowanego przez agenty jest to, że agenty AI nie mają wbudowanej lojalności wobec określonych marek, dostawców czy platform. W odróżnieniu od ludzi, u których w wyniku powstawania więzi emocjonalnych, wcześniejszych doświadczeń lub zwykłej inercji rozwija się lojalność, agenty podejmują decyzje wyłącznie na podstawie kryteriów optymalizacji i dostępnych danych. Tworzy to rynek cechujący się radykalną przejrzystością i obiektywnymi porównaniami. W takim otoczeniu:

Kup książkę Poleć książkę

- Koszty zmiany dostawcy spadają do zera, ponieważ agenty potrafią błyskawicznie ocenić dostępne oferty bez emocjonalnego przywiązania.
- Historyczne relacje zapewniają minimalną przewagę, chyba że przekładają się na obiektywnie lepszą ofertę.
- Wartość marki wynika z weryfikowalnego poziomu działania, a nie z emocjonalnych skojarzeń.
- Ukryte opłaty lub problemy z jakością nie mogą zostać zatuszowane przez marketing lub manipulacje psychologiczne.
- Wartość musi być jednoznacznie mierzalna w kategoriach, które agenty potrafią przetwarzać i porównywać.

Dla organizacji przyzwyczajonych do konkurowania z wykorzystaniem lojalności wobec marki, więzi emocjonalnych lub asymetrii informacji stanowi to wyzwanie egzystencjalne. Sukces w świecie sterowanym przez agenty wymaga zmian w strukturze danych i działalności pod kątem obiektywnego dostarczania wartości i algorytmicznej analizy ofert. To właśnie w tym obszarze niezbędnych jest pięć wymienionych wcześniej wzorców architektonicznych. Stanowią one punkt wyjścia do konkurowania na radykalnie przejrzystym rynku.

Autopilot w działalności biznesowej — autonomiczne operacje i podejmowanie decyzji

Poza obsługą interakcji z konsumentami agentowa AI umożliwia coś, co można nazwać „biznesowym autopilotem", czyli autonomiczne wykonywanie podstawowych procesów biznesowych przy minimalnej ingerencji człowieka. Funkcje tego typu można wykorzystać w różnych działach:

- *łańcuch dostaw:* autonomiczne zarządzanie zapasami, wybór dostawców i optymalizacja logistyki (wzorce 1., 2., 3.);
- *marketing:* zautomatyzowane tworzenie kampanii, targetowanie i optymalizacja skuteczności (wzorce 2., 4., 5.);
- *obsługa klienta:* kompleksowe rozwiązywanie problemów bez przekazywania do poziomu człowieka (wzorce 1., 4., 5.);
- *operacje finansowe:* zautomatyzowane zarządzanie przepływami pieniężnymi, podejmowanie decyzji inwestycyjnych i zabezpieczanie przed ryzykiem (wzorce 1., 2., 3.);
- *rozwój produktu:* określanie priorytetowych funkcji i iteracyjny rozwój projektu przez sztuczną inteligencję (wzorce 1., 3., 4.).

Aby te autonomiczne operacje były wykonywane efektywnie, organizacje muszą opracować infrastrukturę danych, która pozwala utworzyć:

- jednoznaczne granice decyzyjne i modele autoryzacji, które określają, w jakich obszarach i w jaki sposób agenty mogą podejmować autonomiczne działania (wzorzec 5.);
- systemy monitorowania w czasie rzeczywistym, które śledzą działania i wyniki agentów (wzorzec 2.);

Kup książkę Poleć książkę

- mechanizmy informacji zwrotnych, które umożliwiają ciągłe uczenie się i doskonalenie (wzorzec 5.);
- narzędzia do objaśniania działania AI, które pomagają ludziom zrozumieć decyzje agentów (wzorzec 1.);
- mechanizmy zmiany decyzji, które pozwalają na ingerencję człowieka, gdy jest to konieczne (wzorzec 5).

Lista kontrolna do oceny dostosowania danych do przyszłości opartej na agentach

Aby przygotować się na opisaną przyszłość opartą na agentach, organizacje powinny móc ocenić dostosowanie danych do niej w następujących kluczowych wymiarach:

Dostępność danych ustrukturyzowanych

- Wszystkie atrybuty produktów i usług są dostępne w formatach czytelnych dla maszyn.
- API zapewniają ustandaryzowany dostęp do wszystkich istotnych danych biznesowych.
- Aktualizacje danych w czasie rzeczywistym są dostępne za pośrednictwem strumieni zdarzeń.
- Kompletne metadane opisują znaczenie i kontekst wszystkich elementów danych.

Przejrzysta weryfikacja działania

- Kluczowe wskaźniki działania są obiektywnie zdefiniowane i mierzalne.
- Dane historyczne dotyczące działania są dostępne i weryfikowalne.
- Gwarancje poziomu świadczenia usług są sformalizowane i czytelne dla maszyn.
- Monitorowanie działania odbywa się w ciągły i przejrzysty sposób.

Zaufanie i weryfikacja

- Pochodzenie danych jest rejestrowane i weryfikowalne.
- Procesy zapewnienia jakości są udokumentowane i powszechnie dostępne.
- Certyfikaty stron trzecich są dostępne w formatach odczytywalnych maszynowo.
- Poświadczenia zaufania są ustandaryzowane i weryfikowalne.

Możliwości interakcji z agentami

- Dostępne są ustandaryzowane interfejsy do obsługi zapytań i transakcji z udziałem agentów.
- Zdefiniowane są protokoły negocjacji cen i warunków.
- Istnieją mechanizmy przekazywania informacji zwrotnych dotyczących jakości usług.
- Ustanowiono procesy obsługi wyjątków dla nietypowych sytuacji.

Ład danych i kontrola

- Zdefiniowano jednoznaczne modele uprawnień dla działań agentów.
- Rejestrowane są ścieżki audytu dla wszystkich interakcji z udziałem agentów.

Kup książkę Poleć książkę

- Dostępne są mechanizmy zmiany decyzji agentów umożliwiające ingerencję człowieka.
- Zapewniona jest zgodność ze zmieniającymi się wymaganiami dotyczącymi ładu i regulacji.

Organizacje, które proaktywnie zadbają o te aspekty dostosowania danych, będą dobrze przygotowane na to, by odnosić sukcesy w rodzącej się gospodarce opartej na agentach. Natomiast firmy, które będą zwlekać ze zmianami, mogą tracić udziały w rynku, w miarę jak transakcje realizowane za pośrednictwem agentów staną się normą. Pięć wzorców architektonicznych przedstawionych w tym rozdziale określa techniczne podstawy do osiągnięcia niezbędnej gotowości.

Plan wdrożenia: wskazówki dotyczące implementacji

Błyskawiczny rozwój generatywnej sztucznej inteligencji z jednej strony otwiera przed organizacjami z różnych branż bezprecedensowe możliwości, a z drugiej stawia przed nimi wyzwania. W tym rozdziale wyjaśniliśmy, że przepaść między eksperymentalnymi projektami PoC a udanymi wdrożeniami produkcyjnymi wynika w dużej mierze z nieodpowiedniej podstawowej architektury danych. Można jednak wyeliminować tę lukę.

Od czego zacząć?

Punkt wyjścia zależy od obecnej dojrzałości organizacji w zakresie zarządzania danymi i ambicji związanych z generatywną AI.

Jeśli dopiero zaczynasz (etap eksperymentów)

Zacznij od: wzorzec 4. (wyszukiwanie semantyczne) + wzorzec 3. (architektura typu lakehouse).

Dlaczego? Wyszukiwanie semantyczne natychmiast generuje wartość w zastosowaniach związanych z wyszukiwaniem dokumentów i zarządzaniem wiedzą. Z kolei repozytorium typu lakehouse zapewnia ujednolicony podstawowy magazyn danych, który będzie potrzebny w pozostałych wzorcach.

Harmonogram prac: 3 – 4 miesiące do pierwszego produkcyjnego scenariusza użycia.

Przykładowe zastosowania: wyszukiwanie w wewnętrznej bazie wiedzy, pytania i odpowiedzi dotyczące dokumentów, asystenty ds. zasad i procedur.

Wskaźnik sukcesu: skrócenie czasu wyszukiwania informacji o ponad 50 procent.

Jeśli masz już uruchomione proste narzędzia generatywnej AI (etap projektów pilotażowych)

Dodaj: wzorzec 2. (architektura oparta na zdarzeniach) + wzorzec 1. (grafy wiedzy).

Dlaczego? Architektura oparta na zdarzeniach sprawia, że agenty korzystają z aktualnych informacji, a grafy wiedzy umożliwiają wnioskowanie z uwzględnieniem wielu dziedzin, co otwiera drogę do bardziej zaawansowanych zastosowań.

Kup książkę Poleć książkę

Harmonogram prac: 4 – 6 miesięcy na ulepszenie istniejących aplikacji.

Przykładowe zastosowania: agenty obsługi klienta z kontekstem określanym w czasie rzeczywistym, optymalizacja łańcucha dostaw, analityka obejmująca wiele obszarów biznesowych.

Wskaźnik sukcesu: przejście od zastosowań w jednej dziedzinie do zastosowań na poziomie wielu obszarów.

Jeśli skalujesz rozwiązania produkcyjne (etap wzrostu)

Dodaj: wzorzec 5. (dane przystosowane do agentów) + zaawansowane możliwości we wszystkich wzorcach.

Dlaczego? Projekty przygotowane do obsługi agentów pomogą Ci poradzić sobie w nowej gospodarce opartej na tej technologii. Zaawansowane możliwości (siatka danych, multimodalność, aktualizowanie kontekstu w czasie rzeczywistym) umożliwiają operacje autonomiczne.

Harmonogram prac: 6 – 12 miesięcy do pełnej transformacji w kierunku przystosowania do agentów.

Przykładowe zastosowania: autonomiczne agenty zakupowe, funkcje biznesowego autopilota, doświadczenia klientów z udziałem agentów.

Wskaźnik sukcesu: umożliwienie agentom autonomicznym działania przy minimalnej ingerencji człowieka.

Drzewo decyzyjne: który wzorzec wybrać jako pierwszy?

Zacznij od określenia najbardziej wartościowego zastosowania. To naprowadzi Cię na wzorzec, od którego warto zacząć:

Wyszukiwanie dokumentów i wiedzy? → Zacznij od wzorca 4. (wyszukiwanie semantyczne)

- szybkie korzyści dzięki istniejącym repozytoriom dokumentów;
- podstawa do tworzenia agentów korzystających z metody RAG;
- czas do wdrożenia w środowisku produkcyjnym: 3 – 4 miesiące.

Wnioski generowane z uwzględnieniem różnych dziedzin? → Zacznij od wzorca 1. (grafy wiedzy)

- łączenie odizolowanych źródeł danych;
- umożliwianie zaawansowanego wnioskowania;
- czas do wdrożenia w środowisku produkcyjnym: 4 – 6 miesięcy.

Umożliwianie podejmowania decyzji w czasie rzeczywistym? → Zacznij od wzorca 2. (systemy oparte na zdarzeniach)

- gwarantuje, że agenty korzystają z aktualnych danych;
- zapobiega problemom związanym z nieaktualnymi informacjami;
- czas do wdrożenia w środowisku produkcyjnym: 3 – 5 miesięcy.

Kup książkę Poleć książkę

Ujednolicona analityka + AI? → Zacznij od wzorca 3. (architektura typu lakehouse)

- konsoliduje infrastrukturę danych;
- obsługuje zarówno procesy analityki biznesowej, jak i związane z AI;
- czas do wdrożenia w środowisku produkcyjnym: 4 – 6 miesięcy.

Dostęp dla agentów zewnętrznych? → Zacznij od wzorca 5. (przygotowanie do obsługi agentów)

- przygotowuje na gospodarkę opartą na agentach;
- umożliwia autonomiczne interakcje;
- czas do wdrożenia w środowisku produkcyjnym: 6 – 9 miesięcy.

Pamiętaj, że najbardziej udane wdrożenia łączą wiele wzorców. Zacznij od jednego z nich, szybko udowodnij jego wartość (w okresie 3 – 6 miesięcy), a następnie, w miarę jak kompetencje organizacji w zakresie sztucznej inteligencji będą dojrzewać, wprowadzaj kolejne wzorce.

Plan działania

W ciągu najbliższych 30 dni:

1. Oceń obecny poziom dojrzałości organizacji w zakresie danych. Uwzględnij pięć opisanych wzorców.
2. Zidentyfikuj zastosowanie generatywnej AI o najwyższej wartości.
3. Wybierz wzorzec początkowy na podstawie drzewa decyzyjnego.
4. Zbuduj zespół wielofunkcyjny (inżynieria danych, AI, uczenie maszynowe, interesariusze biznesowi).
5. Zdefiniuj wskaźniki sukcesu, które wykraczają poza jakość techniczną i obejmują wartość biznesową.

W ciągu najbliższych 90 dni:

1. Zaimplementuj pierwszy wzorzec dla wybranego zastosowania.
2. Wprowadź pętle sprzężenia zwrotnego obejmujące użytkowników końcowych.
3. Udokumentuj wnioski i decyzje architektoniczne.
4. Rozpocznij planowanie wprowadzenia drugiego wzorca.
5. Rozwijaj kompetencje organizacji dzięki szkoleniom i zarządzaniu zmianami.

W ciągu najbliższych 6 – 12 miesięcy:

1. Wprowadź wiele wzorców w miarę dojrzewania rozwiązań.
2. Skaluj udane wdrożenia w całej organizacji.
3. Wprowadź ramy ładu dotyczące działania agentów.
4. Przygotuj się na przyszłość opartą na agentach przez udostępnienie danych dostosowanych do odczytu przez maszyny.
5. Mierz i komunikuj wartość biznesową, aby zapewnić ciągłość inwestycji.

Kup książkę Poleć książkę

Podsumowanie

W rozdziale 2. szczegółowo omawiamy techniczną implementację frameworku danych przystosowanych do AI. Przedstawiamy wzorce architektoniczne, wybór technologii oraz modele ładu danych, które umożliwiają pomyślne wdrożenia w środowisku produkcyjnym. Analizujemy konkretne usługi Amazon Web Services (AWS) i sposoby wdrażania, a także najlepsze praktyki dla organizacji na różnych etapach rozwoju rozwiązań AI.

Czas na uzyskanie przewagi konkurencyjnej jest krótki. Organizacje, które już teraz przygotowują architekturę danych przystosowaną do AI, będą liderami w gospodarce opartej na agentach. Natomiast firmy, które ograniczają się do tradycyjnego podejścia, ryzykują, że staną się niewidoczne dla agentów pośredniczących w przyszłych interakcjach z klientami.

Pytanie nie brzmi, czy warto przygotowywać się na przyszłość opartą na agentach. Istotne jest to, czy Twoja organizacja będzie gotowa, gdy ta przyszłość już nadejdzie.

Kup książkę Poleć książkę


PROGRAM
PARTNERSKI
GRUPY HELION
1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ
Zmień swoją stronę WWW w działający bankomat!
Dowiedz się więcej i dołącz już dzisiaj!
http://program-partnerski.helion.pl
GRUPA
Helion

Od surowych danych do zaawansowanej AI. Zbuduj architekturę przyszłości!

Rozwój generatywnej sztucznej inteligencji to jeden z najważniejszych trendów we współczesnym IT. Mimo ogromnego entuzjazmu większość firm napotyka poważne przeszkody przy przechodzeniu od fazy eksperymentów do wdrożeń produkcyjnych. Główną przyczyną niepowodzeń nie są ograniczenia samych modeli językowych (LLM), lecz nieodpowiednio przygotowane dane. Ta książka pokazuje, że w erze AI przewagę konkurencyjną zyskają nie te organizacje, które dysponują najpotężniejszymi algorytmami, ale te, które potrafią zaprojektować i wdrożyć nowoczesną, skalowalną architekturę danych.

Oto kompleksowy przewodnik, który przeprowadza czytelnika przez proces transformacji surowych, nieustrukturyzowanych informacji w wiedzę gotową do wykorzystania przez AI. Autorzy szczegółowo omawiają pięć kluczowych wzorców architektonicznych, w tym grafy wiedzy, platformy lakehouse i systemy sterowane zdarzeniami. Krok po kroku wyjaśniają mechanizmy działania wektorowych baz danych, wyszukiwania semantycznego i techniki RAG (*retrieval-augmented generation*). Książka wykracza poza samą technologię — dostarcza również gotowe ramy dla ładu danych *data governance*, bezpieczeństwa, zgodności z regulacjami i orkiestracji w podejściu LLMOps. To praktyczny plan działania, który pozwala zbudować niezawodne, bezpieczne i autonomiczne systemy agentowe klasy korporacyjnej.

Ta książka przekłada najtrudniejszy problem związany z generatywną AI w organizacjach (...) w jasny i możliwy do wdrożenia plan inżynieryjny. Jeśli tworzysz rozwiązania, a nie tylko o nich mówisz, będzie to Twój praktyczny przewodnik.

Magesh Varadharajan, menedżer zespołu inżynieryjnego, Gen Digital Inc.

Jest to rzadkie połączenie omówienia strategii i jej realizacji. Książka ta stanowi praktyczny przewodnik, który wyjaśnia, jak przełożyć szum wokół AI na realia organizacji.

Iskander Sanchez-Rola, menedżer do spraw AI i innowacji

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-4259-2
9 788328 942592
Cena: 89,00 zł

Najważniejsze zagadnienia

- Projektowanie architektury danych dla generatywnej i agentowej AI
- Wdrażanie wektorowych baz danych i wyszukiwania semantycznego
- Optymalizacja systemów RAG i zaawansowana inżynieria promptów
- Ramy ładu danych, bezpieczeństwa i zgodności z regulacjami
- Orkiestracja przepływów pracy i praktyki LLMOps

Navnit Shukla jest głównym architektem rozwiązań w firmie Snowflake, w której wspiera klientów w pozyskiwaniu informacji z danych przy użyciu AI. Jest również autorem książki *Data Wrangling on AWS*.

Kien Pham jest głównym architektem rozwiązań AWS wspierającym firmy z branży cyfrowej. Ma przeszło dziesięcioletnie doświadczenie w inżynierii oprogramowania.

Srikanth Sopirala jest głównym specjalistą do spraw AI w AWS, gdzie pomaga globalnym przedsiębiorstwom radzić sobie z istotnymi wyzwaniami związanymi z AI i danymi.

Harsha Tadiparthi jest głównym specjalistą do spraw AI w AWS. Wspomaga firmy z listy Fortune 500 w rozwiązywaniu złożonych problemów w zakresie danych i AI.