

Databricks: GenAI – RAG z AI Search

Szkolenie Databricks: GenAI – RAG z AI Search to praktyczne wprowadzenie do budowy rozwiązań Retrieval-Augmented Generation na danych firmowych. Program koncentruje się na pełnym przepływie RAG: od przygotowania dokumentów, przez wyszukiwanie w AI Search (wcześniej Databricks Vector Search), po wygenerowanie odpowiedzi przez model językowy i ocenę jej jakości.

Uczestnicy przechodzą przez cały proces pracy z dokumentami, metadanymi, indeksem, modelem językowym oraz ewaluacją odpowiedzi w MLflow. Dzięki temu szkolenie pokazuje nie tylko, jak zbudować działający RAG, ale też jak zmierzyć jego jakość, znaleźć przyczyny złych odpowiedzi i ocenić, kiedy RAG jest właściwym rozwiązaniem, a kiedy nie. Ćwiczenia opierają się na dokumentach RetailOps (fikcyjna firma handlowa, na której danych realizujemy wszystkie ćwiczenia).



Odbiorcy szkolenia

Szkolenie jest przeznaczone dla osób technicznych, które chcą budować rozwiązania generatywnej AI oparte na dokumentach i wiedzy firmowej w Databricks.

W szczególności kurs sprawdzi się dla:

- inżynierów ML i AI budujących rozwiązania oparte na modelach językowych,
- programistów tworzących asystentów dokumentacyjnych, chatboty Q&A i wyszukiwarki semantyczne,
- inżynierów danych przygotowujących dokumenty i dane dla rozwiązań AI,
- zespołów, które mają za sobą pierwsze eksperymenty z LLM i chcą przejść do rozwiązania z mierzalną jakością.

Szkolenie będzie szczególnie wartościowe dla osób, które chcą świadomie ocenić, czy ich przypadek użycia nadaje się do RAG, zanim zainwestują w pełne wdrożenie.



Korzyści

- Architektura RAG – poznasz elementy rozwiązania RAG, typowe ryzyka i sposób definiowania kryteriów sukcesu.
- Przygotowanie dokumentów – nauczysz się dobierać strategię dzielenia dokumentów (chunking) i wzbogacać fragmenty o metadane.
- AI Search – dowiesz się, jak zbudować indeks, wybrać model embeddingów i wyszukiwać z filtrowaniem po metadanych.
- Generowanie odpowiedzi – zbudujesz łańcuch RAG z promptem i modelem udostępnionym przez Model Serving.
- Ewaluacja i analiza błędów – nauczysz się śledzić działanie RAG w MLflow Tracing i oceniać odpowiedzi na zestawie pytań testowych.
- Kompletne rozwiązanie – przejdziesz przez proces od korpusu dokumentów do działającego mini-RAG z oceną jakości odpowiedzi.



Program szkolenia

Moduł – Architektura RAG i kryteria sukcesu

Cel modułu: uczestnik rozumie, jak działa RAG i potrafi ocenić, czy pasuje do jego przypadku użycia.

- Czym jest RAG – model językowy odpowiada na pytania na podstawie dokumentów firmy, które najpierw wyszukujemy
- Elementy rozwiązania w Databricks: dokumenty w Volumes, tabele Delta, AI Search, Model Serving, MLflow
- AI Playground – szybkie porównanie modeli i promptów przed budową rozwiązania
- Ryzyka – zmyślane odpowiedzi (halucynacje), nieaktualne dokumenty, ujawnienie danych poufnych, koszt
- Kiedy RAG ma sens, a kiedy wystarczy zwykłe wyszukiwanie lub zapytanie SQL
- Gotowy agent bez programowania (Genie Agent, który odpowiada też na pytania o dokumenty w Volumes – Beta) a własny RAG – kiedy który; dotychczasowy Knowledge Assistant ma od października 2026 status legacy
- Karta przypadku użycia – kto pyta, o co, na podstawie jakich źródeł i po czym poznamy, że rozwiązanie działa
- Lab: uczestnik przygotowuje kartę przypadku użycia dla asystenta dokumentacji RetailOps

2. Moduł – Przygotowanie dokumentów

Cel modułu: uczestnik przygotowuje dokumenty tak, żeby wyszukiwanie zwracało trafne fragmenty.

- Odczyt dokumentów PDF i DOCX z Volumes, w tym wyciąganie tekstu i tabel (`ai_parse_document()`) –

pokaz; w laboratorium pracujemy na dokumentach już przetworzonych)

- Chunking, czyli podział dokumentów na fragmenty – po stałej liczbie znaków, po nagłówkach lub według znaczenia; automatyczny podział z dopisaniem tytułu i sekcji do każdego fragmentu (`ai_prep_search()`, Beta – pokaz)
- Jak rozmiar fragmentu wpływa na trafność odpowiedzi i koszt
- Metadane fragmentów – źródło, sekcja, data, kategoria, poziom poufności; później służą do filtrowania
- Aktualizacja bazy wiedzy – jak obsłużyć nowe i zmienione dokumenty bez przeliczania wszystkiego
- Lab: uczestnik dzieli dokumenty RetailOps na fragmenty z metadanymi; efektem jest tabela gotowa do indeksowania

3. Moduł – AI Search: indeks i wyszukiwanie

Cel modułu: uczestnik buduje indeks dokumentów i sprawdza, czy wyszukiwanie zwraca właściwe fragmenty.

- AI Search (wcześniej Databricks Vector Search) – wyszukiwanie po znaczeniu, a nie tylko po słowach kluczowych
- Embeddingi – zamiana tekstu na wektory liczb; wybór modelu i jego wpływ na jakość i koszt
- Rodzaje indeksów – indeks synchronizowany automatycznie z tabelą Delta lub indeks zasilany bezpośrednio
- Wyszukiwanie hybrydowe – połączenie wyszukiwania po znaczeniu i po słowach kluczowych
- Filtrowanie wyników po metadanych (np. tylko aktualne dokumenty z danej kategorii) i ponowne sortowanie wyników (reranking)
- Lab: uczestnik testuje wyszukiwanie na przygotowanym indeksie z filtrami po kategorii i dacie

4. Moduł – Generowanie odpowiedzi

Cel modułu: uczestnik łączy wyszukiwanie z modelem językowym w działający łańcuch RAG.

- Przebieg łańcucha – pytanie → wyszukanie fragmentów → zbudowanie kontekstu → odpowiedź modelu ze wskazaniem źródeł
- Modele dostępne w Databricks (Foundation Model APIs) i modele zewnętrzne udostępniane i kontrolowane przez Unity Gateway
- Prompt – instrukcja dla modelu, sposób podania kontekstu, odmowa odpowiedzi, gdy w dokumentach brak informacji
- Wywołanie modelu z SQL (`ai_query()`) – np. do przetwarzania wielu dokumentów naraz; wyszukanie fragmentów i odpowiedź ze wskazaniem źródeł w jednym zapytaniu SQL (`ai_search()`, Beta) (pokaz)
- Zapis łańcucha w Unity Catalog – wersjonowanie i kontrola, która wersja jest używana (pokaz)
- Lab: uczestnik buduje łańcuch RAG odpowiadający na pytania o dokumenty RetailOps

5. Moduł – Śledzenie, ocena jakości i analiza błędów

Cel modułu: uczestnik potrafi zmierzyć jakość odpowiedzi i znaleźć przyczynę złych odpowiedzi.

- MLflow Tracing – zapis każdego kroku: jakie fragmenty znaleziono, co trafiło do modelu, ile to trwało i kosztowało
- Zestaw testowy – pytania z oczekiwanymi odpowiedziami i źródłami
- Automatyczna ocena (`mlflow.genai.evaluate()`) – czy odpowiedź jest trafna, oparta na dokumentach, poprawna i bezpieczna

- Własne kryteria oceny zapisane zwykłym językiem (np. „odpowiedź musi podawać numer procedury”)
- Analiza błędów – czy problem leży w podziale dokumentów, w wyszukiwaniu czy w prompcie
- Ocena odpowiedzi przez ekspertów merytorycznych bez znajomości kodu – Review App i kolejki przeglądu (review queues, Beta) (pokaz)
- Lab: uczestnik ocenia łańcuch na przygotowanym zestawie 15–20 pytań, zmienia jeden parametr i porównuje wynik

6. Moduł – Warsztat zintegrowany i droga do produkcji

Cel modułu: uczestnik składa kompletne rozwiązanie i wie, co jest potrzebne do jego wdrożenia.

- Warsztat: uczestnicy dopracowują łańcuch zbudowany w modułach 4 i 5 (tabela fragmentów → indeks AI Search → łańcuch RAG → śledzenie → ocena jakości): na podstawie wyników oceny poprawiają wyszukiwanie lub prompt i ponownie mierzą jakość. Efektem jest działający mini-RAG z raportem jakości i źródłami odpowiedzi.
- Opcje wdrożenia – Databricks Apps (zalecane dla nowych rozwiązań); wdrożenie jako endpoint Model Serving ma status legacy i dotyczy już istniejących rozwiązań; zapis zapytań, odpowiedzi i śladów MLflow w Unity Catalog (omówienie i pokaz)
- Monitorowanie jakości i kosztu po wdrożeniu



Oczekiwane przygotowanie uczestnika

- Ukończone szkolenie DBX-FND (Fundamenty Databricks) lub praktyka z Unity Catalog i tabelami Delta
- Podstawowa znajomość Python (poziom skryptowy)
- Ogólna znajomość pojęć model językowy (LLM) i embedding – nie jest wymagane doświadczenie w ich wdrażaniu

Dla zwiększenia komfortu pracy oraz efektywności szkolenia zalecamy skorzystanie z dodatkowego ekranu. Brak dodatkowego ekranu nie jest przeciwwskazaniem do udziału w szkoleniu, ale w znaczący sposób wpływa na komfort pracy podczas zajęć

Informacje oraz wymagania dotyczące uczestniczenia w szkoleniach w formule zdalnej dostępne na: <https://www.altkomakademia.pl/jak-dzialamy/o-szkoleniach/?distance-learning>



Szkolenie obejmuje

* dostęp do portalu słuchacza Altkom Akademii

Metoda szkolenia:

- przygotowane środowisko Databricks z korpusem dokumentów, indeksem AI Search i endpointem modelu na czas szkolenia

- notebooki i materiały do ćwiczeń w formie elektronicznej

Uwaga: W trakcie wybranych szkoleń Akademia zapewnia Uczestnikom dostęp do płatnych narzędzi i usług niezbędnych do realizacji programu szkolenia (w szczególności usług Azure, oraz innych zasobów chmurowych). Uczestnik zobowiązany jest do korzystania z udostępnionych zasobów wyłącznie w zakresie wynikającym z agendy szkolenia oraz zgodnie z poleceniami trenera. Wykorzystywanie tych zasobów w sposób wykraczający poza zakres szkoleniowy, w szczególności do realizacji projektów prywatnych, testowania dodatkowych usług, modyfikowania konfiguracji środowiska lub podejmowania działań generujących koszty niezwiązane z realizacją ćwiczeń – skutkować będzie obciążeniem Uczestnika kosztami przekraczającymi przewidziany limit, na podstawie wydatków poniesionych przez Akademię (<https://www.altkomakademia.pl/ogolne-warunki-uczestnictwa-w-szkoleniach/>)



Język

Wykład: polski

Materiały: angielski

Czas trwania

1 dni / 7 godzin

Opis egzaminu

Szkolenie nie zawiera formalnego egzaminu. Zaleca się jednak podejście do egzaminu certyfikacyjnego Databricks Certified Data Engineer Associate jako kolejny krok rozwojowy.