kod szkolenia: DBX-RAG / ENG DL 1d

Databricks: GenAI – RAG with AI Search

Databricks: GenAI – RAG with AI Search is a hands-on introduction to building Retrieval-Augmented Generation solutions on company data. The program focuses on the full RAG flow: from preparing documents, through retrieval with AI Search (formerly Databricks Vector Search), to generating an answer with a language model and evaluating its quality.

Participants go through the whole process of working with documents, metadata, the index, the language model and answer evaluation in MLflow. Along with building a working RAG solution, the course shows how to measure its quality, find the causes of bad answers and judge when RAG is the right solution and when it is not. The exercises use RetailOps documents (a fictional retail company whose data we use for all exercises).

Training recipients

The training is intended for technical staff who want to build generative AI solutions in Databricks based on company documents and knowledge.

The course is particularly suitable for:

- ML and AI engineers building solutions based on language models,
- developers creating documentation assistants, Q&A chatbots and semantic search,
- data engineers preparing documents and data for AI solutions,
- teams that have done their first LLM experiments and want to move to a solution with measurable quality.

The training will be especially valuable for people who want to make an informed assessment of whether their use case is a good fit for RAG before investing in a full deployment.

Benefits

- RAG architecture – you will learn the parts of a RAG solution, the typical risks and how to define success criteria.
- Document preparation – you will learn to choose a document splitting strategy (chunking) and enrich chunks with metadata.
- AI Search – you will learn how to build an index, choose an embedding model and search with metadata filters.
- Answer generation – you will build a RAG chain with a prompt and a model served through Model Serving.
- Evaluation and error analysis – you will learn to trace RAG behavior in MLflow Tracing and evaluate answers on a set of test questions.
- A complete solution – you will go through the process from a document corpus to a working mini-RAG with answer quality evaluation.

Training program

Module – RAG architecture and success criteria

Module goal: the participant understands how RAG works and can judge whether it fits their use case.

- What RAG is – a language model answers questions based on company documents that we first search through
- Building blocks in Databricks: documents in Volumes, Delta tables, AI Search, Model Serving, MLflow
- AI Playground – a quick comparison of models and prompts before building the solution
- Risks – made-up answers (hallucinations), outdated documents, exposure of confidential data, cost
- When RAG makes sense and when a regular search or a SQL query is enough
- A ready-made agent with no programming (a Genie Agent, which can also answer questions about documents in Volumes – Beta) vs. a custom RAG – when to use which; the earlier Knowledge Assistant has been legacy since October 2026
- Use case card – who asks, about what, based on which sources, and how we will know the solution works
- Lab: the participant prepares a use case card for the RetailOps documentation assistant

2. Module – Preparing documents

Module goal: the participant prepares documents so that search returns relevant passages.

- Reading PDF and DOCX documents from Volumes, including extracting text and tables (ai_parse_document() – demo; in the lab we work on documents that are already processed)
- Chunking, that is splitting documents into passages – by a fixed number of characters, by headings, or by meaning; automatic splitting that adds the title and section to each passage (ai_prep_search(), Beta – demo)

- How passage size affects answer relevance and cost
- Passage metadata – source, section, date, category, confidentiality level; used later for filtering
- Updating the knowledge base – how to handle new and changed documents without reprocessing everything
- Lab: the participant splits RetailOps documents into passages with metadata; the result is a table ready for indexing

3. Module – AI Search: the index and search

Module goal: the participant builds a document index and checks whether search returns the right passages.

- AI Search (formerly Databricks Vector Search) – searching by meaning, in addition to keywords
- Embeddings – turning text into vectors of numbers; choosing a model and how it affects quality and cost
- Index types – an index synced automatically with a Delta table, or an index you write to directly
- Hybrid search – combining search by meaning with keyword search
- Filtering results by metadata (for example, only current documents in a given category) and re-sorting results (reranking)
- Lab: the participant tests search on a prepared index with filters by category and date

4. Module – Generating answers

Module goal: the participant connects search with a language model into a working RAG chain.

- How the chain runs – question → passage search → building the context → model answer with sources cited
- Models available in Databricks (Foundation Model APIs) and external models served and governed through Unity Gateway
- The prompt – instructions for the model, how context is supplied, declining to answer when the documents have no information
- Calling a model from SQL (ai_query()) – for example, to process many documents at once; searching for passages and getting an answer with cited sources in a single SQL query (ai_search(), Beta) (demo)
- Registering the chain in Unity Catalog – versioning and control over which version is in use (demo)
- Lab: the participant builds a RAG chain that answers questions about RetailOps documents

5. Module – Tracing, quality evaluation and error analysis

Module goal: the participant can measure answer quality and find the cause of bad answers.

- MLflow Tracing – a record of every step: which passages were found, what was sent to the model, how long it took and what it cost
- Test set – questions with expected answers and sources
- Automatic evaluation (mlflow.genai.evaluate()) – whether an answer is relevant, grounded in the documents, correct and safe
- Custom evaluation criteria written in plain language (for example, “the answer must give the procedure number”)
- Error analysis – whether the problem lies in document splitting, in search or in the prompt
- Answer review by subject matter experts who do not need to know code – Review App and review

queues (Beta) (demo)

- Lab: the participant evaluates the chain on a prepared set of 15–20 questions, changes one parameter and compares the results

6. Module – End-to-end workshop and the path to production

Module goal: the participant assembles a complete solution and knows what is needed to deploy it.

- Workshop: participants refine the chain built in modules 4 and 5 (passage table → AI Search index → RAG chain → tracing → quality evaluation): based on the evaluation results they improve search or the prompt and measure quality again. The outcome is a working mini-RAG with a quality report and answer sources.
- Deployment options – Databricks Apps (recommended for new solutions); deployment as a Model Serving endpoint is legacy and applies to existing solutions; saving requests, responses and MLflow traces in Unity Catalog (overview and demo)
- Monitoring quality and cost after deployment

Expected preparation of the participant

- Completion of DBX-FND (Databricks Fundamentals) or hands-on experience with Unity Catalog and Delta tables
- Basic knowledge of Python (scripting level)
- A general understanding of the concepts of a language model (LLM) and an embedding – experience deploying them is not required

To increase the comfort of work and training's effectiveness we suggest using an additional monitor. The lack of additional monitor does not exclude participation in the training, however, it significantly influences the comfort of work during classes.

You can find information and requirements of participation in Distance Learning trainings at: https://www.altkomakademia.pl/en/how-it-works/about-courses/

Training Includes

Training includes

- access to the Altkom Akademia student portal
- a prepared Databricks environment with a document corpus, an AI Search index and a model endpoint for the duration of the training
- notebooks and exercise materials in electronic form

Note: During selected training sessions, the Academy provides Participants with access to paid tools and services necessary to complete the training program (in particular Azure, Fabric, and other cloud resources). The Participant is required to use the provided resources solely within the scope of the

training agenda and in accordance with the trainer’s instructions. Use of these resources beyond the training scope—particularly for private projects, testing additional services, modifying the environment configuration, or undertaking activities that generate costs unrelated to the exercises—will result in the Participant being charged for costs exceeding the allocated limit, based on expenses incurred by the Academy.

Czas trwania

1 dni / 7 godzin

Language

- Training: English
- Materials: English