kod szkolenia: DBX-PFE / PL DL 1d

Databricks Data Transformation

Szkolenie Databricks (Transformation) to ostatni krok w spójnej ścieżce szkoleniowej Fundamental → Explorer → Lakehouse → Transformation. Uczestnicy uczą się projektować modularne pipeline’y, łączyć batch i streaming, stosować zaawansowane transformacje, korzystać z Delta Live Tables oraz integrować przepływy z Git i CI/CD.

Odbiorcy szkolenia

Szkolenie przeznaczone jest dla inżynierów danych i zespołów DataOps, które odpowiadają za wdrażanie i utrzymanie produkcyjnych procesów przetwarzania danych w architekturze Lakehouse

Korzyści

Korzyści dla uczestników
– potrafią projektować modularne pipeline’y Silver → Gold
– rozumieją jak łączyć batch i stream w jednym przepływie
– umieją stosować funkcje okienkowe w PySpark do transformacji danych
– potrafią korzystać z Delta Live Tables do automatyzacji pipeline’ów
– znają dobre praktyki orkiestracji i CI/CD w Databricks
– umieją kontrolować jakość danych za pomocą expectations i monitorować lineage
– są przygotowani do utrzymania procesów produkcyjnych w Databricks

Program szkolenia

1.Architektura przetwarzania danych

- Przypomnienie koncepcji Bronze–Silver–Gold w kontekście pipeline’ów transformacyjnych
- Projektowanie przepływów danych w warstwach Silver i Gold
- Modularność i separacja logiki przetwarzania (load → transform → save)

2.Batch i Stream Load w praktyce

- Różnice między przetwarzaniem wsadowym a strumieniowym
- Batch ingest przy użyciu COPY INTO i zapis do tabel Delta
- Strumieniowy ingest z Auto Loader (cloudFiles)
- Structured Streaming: readStream, writeStream, checkpointing i fault tolerance
- Integracja batch i stream

3.Zaawansowane transformacje danych

- Tworzenie cech numerycznych, tekstowych i binarnych
- Transformacje logiczne (case when, when, otherwise)
- Funkcje okienkowe (lag, lead, row_number, rolling average)
- Tworzenie cech czasowych i sesyjnych

4.Delta Live Tables – automatyzacja pipeline’ów

- Deklaratywne podejście do przetwarzania: CREATE LIVE TABLE
- Tworzenie DAG-ów i ustalanie harmonogramów w DLT
- Integracja DLT z Auto Loader i Structured Streaming
- Expectations – kontrola jakości danych w czasie rzeczywistym
- Monitoring i lineage danych w interfejsie DLT

5.Orkiestracja i automatyzacja

- Workflows w Databricks – multi-task joby, zależności, retry
- Parametryzacja pipeline’ów (dbutils.widgets, dbutils.notebook.run)
- Dobre praktyki CI/CD i utrzymania kodu (Repos, wersjonowanie notebooków)

6.CI/CD – praktyczny demo z Git (Repos)

- Klonowanie repozytorium w Databricks Repos
- Commit i push notebooka do Gita
- Uruchomienie pipeline’u z Workflows w oparciu o repo
- Best practices DevOps dla Databricks

7.Projekt końcowy

- Zaprojektowanie i uruchomienie pipeline’u Silver → Gold z wykorzystaniem batch i stream load, Delta Live Tables, reguł kontroli jakości i integracji z Git

Oczekiwane przygotowanie uczestnika

– Ukończone szkolenie Databricks Lakehouse lub równoważna wiedza
– Znajomość SQL i PySpark
– Podstawowe doświadczenie w implementacji pipeline’ów danych

Szkolenie obejmuje

* dostęp do portalu słuchacza Altkom Akademii
Szkolenie prowadzone jest w środowisku Databricks w chmurze. Każdy uczestnik otrzymuje własny workspace z dostępem do Unity Catalog, SQL Editor, Notebooków oraz katalogu z danymi testowymi.
Uwaga:W trakcie wybranych szkoleń Akademia zapewnia Uczestnikom dostęp do płatnych narzędzi i usług niezbędnych do realizacji programu szkolenia (w szczególności usług Azure, Fabric 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

Metoda egzaminacyjna

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

Czas trwania

1 dni / 7 godzin

Opis egzaminu

Szkolenie nie zawiera formalnego egzaminu. Zaleca się jednak podejście do egzaminu Databricks Certified Data

Engineer Associate jako kolejny krok rozwojowy.