kod szkolenia: DBX-DEA / PL DL 3d

Databricks:Data Engineering Associate

To intensywne, trzydniowe szkolenie prowadzi uczestników przez pełny proces pracy inżyniera danych na platformie Databricks. Łącząc treści z modułów Fundamentals, Lakehouse i Data Transformation, uczestnicy nauczą się importować i eksplorować dane, budować multi-etapowe pipeline’y w architekturze Medallion oraz projektować zaawansowane przetwarzanie i automatyzację z użyciem Delta Live Tables. Program od początku przygotowuje do egzaminu Databricks Certified Data Engineer Associate, dlatego obejmuje zarówno podstawy pracy z Data Intelligence Platform, jak i zagadnienia związane z governance, jakością danych i orkiestracją zadań.

Odbiorcy szkolenia

- Inżynierowie danych przygotowujący się do egzaminu Data Engineer Associate.
- Analitycy danych i zespoły BI przenoszące przetwarzanie do Databricks.
- Data scientistci, którzy potrzebują zautomatyzowanych pipeline’ów i czystych, ustrukturyzowanych danych.
- Osoby odpowiedzialne za projektowanie i standaryzację przetwarzania danych (lakehouse, medallion).

Korzyści

- Pełne zrozumienie platformy Databricks Data Intelligence: workspace, klastry, DBFS, Repos i Workflows.
- Umiejętność wczytywania i oczyszczania danych z różnych formatów (CSV, JSON, Parquet) przy użyciu SQL i PySpark.

- Poznanie architektury Lakehouse i warstw Medallion (Bronze, Silver, Gold) oraz projektowanie linii przetwarzania.
- Opanowanie operacji na Delta Lake: ACID, time travel, merge, update, delete, vacuum, optimize, zorder.
- Budowa modularnych pipeline'ów transformacyjnych z użyciem funkcji okienkowych, DML/DDL oraz logiki warunkowej.
- Automatyzacja przetwarzania dzięki Delta Live Tables (DLT) i Databricks Workflows oraz parametryzacja zadań.
- Wprowadzenie do Unity Catalog, Data Governance i Delta Sharing dla zapewnienia bezpieczeństwa, kontroli dostępu i jakości danych.

Program szkolenia

1. Dzień 1 – Fundamentals & Exploration
   • Wprowadzenie do platformy: workspace, klastry, DBFS, Repos, Rejestr funkcji (Repos)
   • Import danych z CSV, JSON i Parquet: read.format, inferSchema, nagłówki, delimitery; typy danych i casty
   • Eksploracja i wizualizacja danych: display(), show(), describe(), summary(); tworzenie statystyk i wykresów w GUI
   • Czyszczenie danych: usuwanie duplikatów (dropDuplicates, distinct), obsługa wartości null (dropna, fillna), rzutowanie i transformacje logiczne (when, otherwise)
   • Podstawowe transformacje w SQL i PySpark: select, withColumn, filter, groupBy, orderBy, agg
   • Tworzenie widoków tymczasowych i praca z SQL w notebooku (createOrReplaceTempView)
   • Wstęp do Workflows: tworzenie prostych zadań, planowanie i uruchamianie cykliczne
2. Dzień 2 – Lakehouse & Delta Lake
   • Architektura Lakehouse i warstwy Medallion (Bronze → Silver → Gold); strategie projektowe i best practices
   • Delta Lake: ACID, enforcement i evolucja schematu, time travel, MERGE INTO, UPDATE, DELETE, CLONE
   • Batch vs streaming load: COPY INTO, Auto Loader, structured streaming (readStream, writeStream), trigger times
   • Budowa pipeline'ów Bronze → Silver → Gold: przetwarzanie surowych danych, czyszczenie, standaryzacja i agregacje
   • Operacje DDL i DML na tabelach Delta; różnice między tabelami managed i external
   • Optymalizacja zapytań: partycjonowanie, OPTIMIZE, ZORDER BY, VACUUM, analiza planu wykonania
   • Wprowadzenie do Unity Catalog i podstaw zarządzania danymi: katalogi, metastore, uprawnienia, lineage

Dzień 3 – Data Transformation & Delta Live Tables

• Zaawansowane transformacje: tworzenie cech binarnych, tekstowych i numerycznych; wykorzystanie

funkcji warunkowych (case when, when/otherwise)
• Funkcje okienkowe: lag, lead, row_number, rank, rolling average; tworzenie cech czasowych i sekwencyjnych
• Modularne pipeline'y w notebookach: struktura kodu (load → transform → save), parametryzacja i uruchamianie z dbutils.notebook.run()
• Przetwarzanie Silver → Gold: reguły transformacji, agregacji i czyszczenia; reużywalność i separacja logiki transformacyjnej
• Delta Live Tables (DLT): tworzenie i deklaracja pipeline'ów (CREATE LIVE TABLE), zarządzanie DAG-ami i harmonogramami, parametryzacja, expectations jako testy jakości danych
• Orkiestracja w Databricks Workflows: multi-task jobs, parametry wejściowe, alerty i monitorowanie
• Data Governance & Quality: Unity Catalog (rol-based access control, metastore), Delta Sharing, linega danych, cost considerations; implementacja audit logów i testów jakości
• Dobre praktyki: testowanie kodu transformacyjnego, modularność notebooków, dokumentacja i wersjonowanie repozytoriów

Oczekiwane przygotowanie uczestnika

- Podstawowa umiejętność pracy z plikami danych (CSV, JSON, Excel).
- Podstawowa znajomość pojęć relacyjnych (kolumna, rekord, typ danych).
- Znajomość SQL i/lub Pythona na poziomie podstawowym jest zalecana, ale nie wymagana.
- Doświadczenie w pracy z Databricks (co najmniej 6 miesięcy) ułatwi udział w warsztatach, ale nie jest wymagane.

Szkolenie obejmuje

* dostęp do portalu słuchacza Altkom Akademii
Produkt zawiera
Wykład (30 %) – wprowadzenie do platformy, architektury Lakehouse i zagadnień transformacji danych
Warsztaty (30 %) – projektowanie i implementacja pipeline'ów w notebookach oraz Workflows
Ćwiczenia (40 %) – praktyczne zadania w SQL, PySpark, Delta Lake, Auto Loader, DLT
Główne narzędzia: Databricks Data Intelligence Platform, SQL, PySpark, Delta Lake, Auto Loader, Unity Catalog, Delta Live Tables, Databricks Workflows, notebooks.
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. Jego celem jest przygotowanie uczestników do certyfikacji Databricks Certified Data Engineer Associate; udział w szkoleniu wraz z samodzielnym utrwaleniem materiału znacząco zwiększa szanse na pozytywny wynik egzaminu.

Czas trwania

3 dni / 21 godzin

Opis egzaminu

Szkolenie nie zawiera formalnego egzaminu. Jego celem jest przygotowanie uczestników do certyfikacji Databricks Certified Data Engineer Associate; udział w szkoleniu wraz z samodzielnym utrwaleniem materiału znacząco zwiększa szanse na pozytywny wynik egzaminu.