

kod szkolenia: DBX-EXP / PL DL 1d

Databricks (Explorer) Data Exploration

Szkolenie Databricks (Explorer) to drugi krok w spójnej ścieżce szkoleniowej Fundamental → Explorer → Lakehouse → Transformation. Uczestnicy wchodzi głębiej w analizę i eksplorację danych z wykorzystaniem SQL i PySpark. Poznają różnice między podejściem deklaratywnym i imperatywnym, uczą się łączyć dane z różnych źródeł, wykonywać agregacje, joiny oraz analizować jakość danych.



Odbiorcy szkolenia

Szkolenie przeznaczone jest dla inżynierów danych, analityków oraz specjalistów BI, którzy chcą nauczyć się eksplorować i analizować dane w Databricks z wykorzystaniem SQL i PySpark. To naturalny krok po ukończeniu szkolenia Databricks Fundamentals.



Korzyści

Korzyści dla uczestników

- rozumieją różnice między SQL a PySpark DataFrame API
- potrafią wczytywać i eksplorować dane z różnych źródeł
- umieją wykonywać agregacje, joiny i transformacje logiczne
- potrafią profilować jakość danych i wykrywać braki, wartości odstające i niezgodności
- znają zalety i wady różnych formatów danych (CSV, JSON, Parquet, Delta)
- potrafią tworzyć wizualizacje i proste dashboardy w Databricks
- są przygotowani do kolejnego etapu ścieżki - szkolenia Databricks Lakehouse



Program szkolenia

1. Analiza danych w środowisku Databricks

- Przypomnienie środowiska: Unity Catalog, notebooki, SQL Editor
- Wczytywanie danych z różnych źródeł: CSV, JSON, Parquet, Delta
- Tworzenie tabel i widoków w Unity Catalog
- Eksploracja danych: `display()`, `show()`, `summary()`, `describe()`

2. SQL i PySpark w analizie danych

- Różnice między podejściem SQL i DataFrame API
- Tworzenie zapytań SQL w notebookach Databricks
- Łączenie SQL i PySpark w jednej analizie (mixed cells)
- Najczęstsze błędy i sposoby optymalizacji zapytań

3. Operacje i transformacje danych

- Filtrowanie (WHERE, filter) i sortowanie (ORDER BY, sort)
- Grupowanie i agregacje (GROUP BY, agg, count, sum, avg)
- Tworzenie i modyfikacja kolumn (withColumn, alias)
- Joiny: INNER, LEFT, RIGHT, FULL, SEMI, ANTI
- Transformacje logiczne (case when, when, otherwise)

4. Analiza jakości i profilowanie danych

- Analiza braków, unikalności i typów danych (na, distinct, count)
- Walidacja i typowanie danych (cast, printSchema)
- AI Functions – wspomaganie analizy i czyszczenia danych
- Generowanie statystyk i profilowanie kolumn

5. File formats trade-offs (light)

- CSV, JSON, Parquet, Delta – różnice i kiedy stosować
- Delta vs Parquet: ACID, schema evolution, time travel
- Koszty i wydajność poszczególnych formatów

6. Wizualizacja i prezentacja wyników

- Tworzenie wykresów i dashboardów w Databricks GUI
- Porównanie rozkładów i wartości kluczowych
- Dokumentowanie wyników i wniosków w notebookach

7. Projekt końcowy

- Przygotowanie analizy jakości danych oraz transformacji łączącej SQL i PySpark, uzupełnione o prostą wizualizację wyników w notebooku Databricks



Oczekiwane przygotowanie uczestnika

- Ukończone szkolenie Databricks Fundamentals lub równoważna wiedza

- Podstawowa znajomość SQL
- Podstawowe doświadczenie w pracy z danymi



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.



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.