

Fundamenty Databricks – platforma i pierwsze produkty danych

Szkolenie Databricks: Fundamenty Databricks to praktyczne wejście w pracę z Databricks Data Intelligence Platform. Program koncentruje się na tym, jak w kontrolowany sposób przejść od surowego pliku do tabeli gotowej do analizy, z wykorzystaniem Unity Catalog, Volumes, SQL i tabel Delta.

Uczestnicy przechodzą przez cały podstawowy przepływ pracy: od organizacji danych w Unity Catalog, przez załadowanie pliku do Volume i pierwsze przekształcenia w SQL, po zapis wyniku jako tabeli Delta i uruchomienie go jako Lakeflow Job. Ćwiczenia opierają się na jednym scenariuszu biznesowym RetailOps (fikcyjna firma handlowa, na której danych realizujemy wszystkie ćwiczenia), dzięki czemu na koniec dnia uczestnik ma kompletny, powtarzalny przepływ danych zbudowany od początku do końca.

Szkolenie prowadzimy w formule hands-on. Trener pokazuje każdy krok na ekranie, a uczestnicy wykonują go razem z nim we własnych notebookach, więc nikt nie zostaje sam z nowym narzędziem. Na koniec dnia, w 45-60-minutowym warsztacie, uczestnicy samodzielnie scalają wszystkie elementy w jeden działający przepływ.



Odbiorcy szkolenia

Szkolenie jest przeznaczone dla osób, które zaczynają pracę z Databricks i chcą zrozumieć, jak platforma jest zbudowana i jak w niej pracować z danymi.

W szczególności kurs sprawdzi się dla:

- nowych osób w zespołach Data, BI i IT, które będą korzystać z Databricks,
- analityków danych, którzy przechodzą z hurtowni danych lub narzędzi BI na Lakehouse,

- programistów i inżynierów rozpoczynających projekty na Databricks,
- osób technicznych, które potrzebują wspólnego słownika pojęć przed kolejnymi szkoleniami ze ścieżki Databricks.

Szkolenie będzie szczególnie wartościowe dla zespołów, które chcą od pierwszego dnia pracować zgodnie z aktualnymi standardami platformy (Unity Catalog, Volumes, serverless compute).



Korzyści

- Zrozumienie platformy Databricks - poznasz rolę workspace, compute i Unity Catalog oraz dowiesz się, jak te elementy współpracują ze sobą.
- Organizacja danych w Unity Catalog - nauczysz się poruszać po hierarchii katalog → schemat → tabela/Volume i rozumieć, kto i do czego ma dostęp.
- Praca z plikami w Volumes - dowiesz się, jak załadować plik do Volume, odczytać go i sprawdzić jego schemat.
- Przekształcenia w SQL - nauczysz się filtrować, łączyć i agregować dane oraz tworzyć kolumny pochodne.
- Tabele Delta - poznasz, jak zapisać wynik jako tabelę Delta, opisać ją i korzystać z historii zmian.
- Powtarzalny przepływ danych - zbudujesz minimalny przepływ plik → tabela Bronze → zestaw danych do analizy i uruchomisz go jako Lakeflow Job.
- Wybór narzędzia do zasilania danych - dowiesz się, kiedy wystarczy plik, kiedy warto użyć Lakeflow Connect, a kiedy Lakeflow Jobs.



Program szkolenia

1. Moduł - Databricks Data Intelligence Platform

Cel modułu: uczestnik wie, z jakich elementów składa się platforma i gdzie w niej pracuje.

- Architektura platformy - co działa po stronie Databricks, a co w chmurze klienta i gdzie fizycznie leżą dane
- Workspace - miejsce pracy zespołu: notebooki do kodu, SQL editor do zapytań, Catalog Explorer do przeglądania danych, asystent AI Genie Code (wcześniej Databricks Assistant) do pomocy w pisaniu i poprawianiu kodu
- Compute, czyli moc obliczeniowa - czym różni się serverless (uruchamiany na żądanie, bez konfiguracji) od klastrów i SQL warehouse oraz jak to wpływa na koszt
- Notebook w praktyce - łączenie SQL i Python w jednym dokumencie, wyświetlanie wyników, parametry (display(), %sql, widgety)
- Git folders - praca z repozytorium kodu bezpośrednio z workspace (omówienie)
- Razem z trenerem: sprawdzenie środowiska - uruchomienie pierwszego notebooka i dostęp do

danych

2. Moduł - Unity Catalog: organizacja danych i uprawnienia

Cel modułu: uczestnik rozumie, jak w Databricks są uporządkowane dane i jak kontroluje się do nich dostęp.

- Hierarchia danych - katalog → schemat → tabela, widok lub Volume; analogia do bazy danych i folderów
- Pełne nazwy obiektów (katalog.schemat.tabela) - jak odwołać się do danych innego zespołu
- Tabele managed i external - kto zarządza plikami tabeli i co się dzieje przy jej usunięciu
- Uprawnienia - nadawanie i odbieranie dostępu poleceniami GRANT / REVOKE, właściciel obiektu
- Wyszukiwanie i pochodzenie danych - strona Discover, na której dane są pogrupowane według obszarów biznesowych (domeny), oraz lineage: z jakich tabel powstała dana tabela i kto z niej korzysta
- Razem z trenerem: uczestnik tworzy własny schemat; nadawanie uprawnień innym osobom pokazuje trener

3. Moduł - Volumes i pierwsze ładowanie danych

Cel modułu: uczestnik potrafi przechowywać pliki w Databricks i wczytać je do analizy.

- Volumes - kontrolowane miejsce na pliki (CSV, JSON, PDF) objęte uprawnieniami Unity Catalog
- DBFS i mounts - starszy sposób pracy z plikami; jak go rozpoznać w istniejącym kodzie i dlaczego się od niego odchodzi
- Wgrywanie pliku przez interfejs i przeglądanie zawartości Volume
- Odczyt plików CSV i JSON w SQL (read_files()) i w Python (spark.read)
- Schemat danych - automatyczne rozpoznanie typów kolumn a schemat zdefiniowany ręcznie; typowe problemy: separator, nagłówki, puste wartości
- Razem z trenerem: uczestnik wgrywa plik zamówień RetailOps i sprawdza, czy typy kolumn zostały rozpoznane poprawnie

4. Moduł - SQL i podstawowe przekształcenia

Cel modułu: uczestnik potrafi przygotować dane do analizy za pomocą SQL.

- Przeglądanie danych - wybór kolumn, filtrowanie, sortowanie
- Łączenie tabel (JOIN) - np. zamówienia z danymi klientów i produktów
- Agregacje - sumy, średnie i liczności w podziale na grupy (GROUP BY)
- Kolumny wyliczane - warunki (CASE WHEN), zmiana typów, operacje na datach i tekstach
- Brakujące wartości i duplikaty - jak je znaleźć i obsłużyć
- Widoki - zapisanie zapytania pod nazwą, żeby inni mogli z niego korzystać
- Razem z trenerem: uczestnik łączy zamówienia z klientami i produktami i liczy sprzedaż według kategorii i miesiąca

5. Moduł - Delta Lake: tabela, historia i jakość

Cel modułu: uczestnik zapisuje wynik pracy jako tabelę Delta i wie, jak z niej bezpiecznie korzystać.

- Czym jest tabela Delta - format tabel w Databricks, który gwarantuje spójny zapis i przechowuje historię zmian
- Zapis wyniku zapytania jako nowej tabeli i dopisywanie danych
- Opisy tabel i kolumn - dokumentacja dla innych użytkowników danych

- Historia zmian i powrót do poprzedniej wersji tabeli (time travel, RESTORE) – np. po błędnej aktualizacji
- Ochrona schematu – co się dzieje, gdy ktoś próbuje zapisać dane w złym formacie
- Proste reguły jakości – kolumny wymagane i warunki poprawności wartości
- Razem z trenerem: uczestnik zapisuje tabelę, celowo wprowadza błąd i przywraca poprzednią wersję

6. Moduł – Lakeflow Jobs i Lakeflow Connect

Cel modułu: uczestnik potrafi uruchamiać przygotowany proces automatycznie i wie, jakie są sposoby zasilania danych.

- Lakeflow Jobs – uruchamianie notebooka według harmonogramu, z powiadomieniem o błędzie
- Historia uruchomień – jak sprawdzić, czy proces się udał i co poszło nie tak
- Lakeflow Connect – gotowe konektory do systemów źródłowych (aplikacje SaaS, bazy danych, pliki w SharePoint i Google Drive) i ich wymagania (omówienie)
- Wybór sposobu zasilania – kiedy wystarczy plik, kiedy konektor, a kiedy zaplanowany job
- Razem z trenerem: utworzenie joba z notebooka, ręczne uruchomienie i podgląd historii uruchomień (harmonogram uczestnicy ustawiają samodzielnie w warsztacie)

7. Moduł – Warsztat zintegrowany i dobre praktyki

Cel modułu: uczestnik samodzielnie składa wszystkie elementy dnia w jeden działający przepływ.

- Warsztat samodzielny (45–60 min): uczestnicy bez prowadzenia krok po kroku łączą to, co robili z trenerem w modułach 1–6: plik w Volume → tabela surowych danych → oczyszczony zestaw do analizy → job z harmonogramem. Trener jest dostępny do konsultacji i na koniec omawia rozwiązania. Efektem jest kompletny, powtarzalny przepływ danych z opisanymi tabelami.
- Kontrola kosztów – automatyczne wyłączanie compute, wybór między serverless a klastrem, podgląd zużycia
- Bezpieczeństwo na co dzień – minimalne uprawnienia, brak haseł w kodzie
- Rekomendacja dalszej ścieżki zależnie od roli uczestnika



Oczekiwane przygotowanie uczestnika

- Brak wymagań formalnych
- Pomocna jest elementarna znajomość pojęcia tabeli i podstaw SQL (SELECT, WHERE)
- Nie jest wymagana znajomość Python ani PySpark

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 Unity Catalog i danymi scenariusza RetailOps na czas szkolenia
- notebooki z przebiegiem wszystkich ćwiczeń wykonywanych razem z trenerem oraz zadanie warsztatowe, 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/>)



Czas trwania

1 dni / 7 godzin

Język

- Szkolenie: polski
- Materiały: angielski