

Databricks: Lakeflow – pipeline'y Medallion, jakość i orkiestracja

Szkolenie Databricks: Lakeflow to praktyczny kurs budowy deklaracyjnych pipeline'ów danych i ich orkiestracji. Program koncentruje się na Lakeflow pipelines (Spark Declarative Pipelines, wcześniej Delta Live Tables) oraz Lakeflow Jobs, czyli na tym, jak zamienić działający proces przetwarzania danych w pipeline z zależnościami, kontrolą jakości i monitorowaniem.

Uczestnicy przechodzą przez cały proces: od wyboru typu obiektu (streaming table lub materialized view), przez obsługę zmian z Auto CDC i reguły jakości (expectations), po uruchomienie pipeline'u w Lakeflow Job z parametrami, ponowieniami i alertami. Dzięki temu szkolenie pokazuje, jak budować pipeline'y, które da się stabilnie utrzymywać w środowisku produkcyjnym. Pipeline budujemy dla RetailOps (fikcyjna firma handlowa, na której danych realizujemy wszystkie ćwiczenia)



Odbiorcy szkolenia

Szkolenie jest przeznaczone dla osób technicznych, które budują i utrzymują produkcyjne przepływy danych w Databricks.

W szczególności kurs sprawdzi się dla:

- inżynierów danych przenoszących procesy z notebooków do deklaracyjnych pipeline'ów,
- zespołów DataOps odpowiedzialnych za harmonogramy, monitorowanie i obsługę awarii,
- osób, które korzystały z Delta Live Tables i chcą poznać aktualne możliwości Lakeflow,
- architektów danych projektujących orkiestrację w Lakehouse.

Szkolenie będzie szczególnie wartościowe dla zespołów, które mają już działające procesy i chcą je uporządkować pod kątem jakości danych i obserwowalności.



Korzyści

- Model Lakeflow pipelines – poznasz flows, streaming tables i materialized views oraz dowiesz się, kiedy stosować każdy z nich.
- Przetwarzanie przyrostowe – nauczysz się wybierać między trybem wsadowym i strumieniowym dla poszczególnych etapów pipeline'u.
- Obsługa zmian z Auto CDC – zobaczysz, jak modelować zmiany i historię (SCD typu 1 i 2) w warstwie Silver.
- Jakość danych w pipeline – nauczysz się definiować expectations i kierować błędne rekordy do osobnej tabeli.
- Orkiestracja w Lakeflow Jobs – zbudujesz job z zależnościami, parametrami, ponowieniami i alertami.
- Obserwowalność – dowiesz się, jak analizować wykonanie pipeline'u przez event log i system tables.



Program szkolenia

1. Moduł – Lakeflow pipelines: jak to działa

Cel modułu: uczestnik rozumie, czym pipeline deklaracyjny różni się od ręcznie pisanego procesu.

- Podejście deklaratywne – opisujemy, jakie tabele mają powstać, a platforma sama ustala kolejność, ponowienia i przyrosty
- Lakeflow pipelines a wcześniejsze Delta Live Tables – co zmieniło się w nazwach i składni
- Rodzaje obiektów: streaming table (dane dopływające przyrostowo) i materialized view (wynik przeliczany z innych tabel)
- Flow – przepływ zasilający tabelę; kilka źródeł może zasilać jedną tabelę
- Definicja pipeline'u w SQL (CREATE OR REFRESH STREAMING TABLE) i w Python (from pyspark import pipelines as dp, @dp.table, @dp.materialized_view)
- Edytor pipeline'ów – pliki, graf zależności między tabelami i podgląd danych
- Pokaz: pierwszy pipeline Bronze → Silver (kontynuowany w warsztacie 1)

2. Moduł – Przetwarzanie wsadowe i strumieniowe

Cel modułu: uczestnik wie, jak często i w jakim trybie odświeżać pipeline, żeby pogodzić koszt z aktualnością danych.

- Tryb wyzwalany (triggered, np. co godzinę) a tryb ciągły – opóźnienie danych kontra koszt; pipeline ciągły można uruchamiać z joba ciągłego, co Databricks zaleca od sierpnia 2026
- Przetwarzanie przyrostowe a pełne przeliczenie – kiedy które jest właściwe
- Przyrostowe odświeżanie materialized views – przeliczanie tylko zmienionej części danych
- Ustawienia pipeline'u – docelowy katalog i schemat, compute, parametry (np. ten sam kod zapisujący do innego katalogu na dev i prod)
- Pokaz: porównanie opóźnienia i kosztu obu trybów

3. Moduł – Auto CDC: obsługa zmian z systemów źródłowych

Cel modułu: uczestnik potrafi przenieść do Lakehouse zmiany z systemu źródłowego, z historią lub bez.

- Strumień zmian (CDC) – źródło przysyła informację o dodaniu, zmianie lub usunięciu rekordu
- Auto CDC – deklaratywne zastosowanie zmian do tabeli docelowej, z właściwą kolejnością zdarzeń (wymaga pipeline'u serverless lub w wersji Pro/Advanced)
- SCD typu 1 i 2 – nadpisanie wartości albo zachowanie pełnej historii zmian (np. zmiana adresu klienta)
- Obsługa usunięć i rekordów, które przyszły w złej kolejności
- Lab (zadanie dodatkowe dla chętnych): uczestnik buduje wymiar klienta z historią zmian (SCD typu 2)

4. Moduł – Warsztat 1: pipeline Silver → Gold

Cel modułu: uczestnik samodzielnie buduje kompletny pipeline deklaratywny.

- Warsztat: pipeline od przygotowanych danych Bronze do tabel Gold ze wskaźnikami sprzedaży. Efektem jest działający pipeline z widocznym grafem zależności i lineage.
- Najczęstsze błędy przy pierwszym uruchomieniu i jak je diagnozować

5. Moduł – Jakość danych i obserwowalność

Cel modułu: uczestnik wbudowuje kontrolę jakości w pipeline i potrafi sprawdzić, co się w nim dzieje.

- Expectations – reguły jakości zapisane przy tabeli; do wyboru: tylko ostrzeżenie, odrzucić rekord albo zatrzymaj pipeline
- Kwarantanna – osobna tabela dla odrzuconych rekordów, żeby można je było przeanalizować – porównanie z ręcznym wzorcem z DBX-ETL
- Event log – dziennik pipeline'u: ile rekordów przeszło, ile odrzucono, ile trwał każdy krok; zdarzenia wszystkich pipeline'ów w regionie trafiają też do tabeli systemowej pipeline_events (Beta)
- System tables – czas działania i koszt pipeline'ów w jednym miejscu
- Lab: uczestnik dodaje reguły jakości i sprawdza w dzienniku, które rekordy zostały odrzucone i dlaczego

6. Moduł – Lakeflow Jobs: orkiestracja i monitorowanie

Cel modułu: uczestnik łączy pipeline'y i inne zadania w zaplanowany, monitorowany proces.

- Job jako zestaw zadań – notebooki, pipeline'y, zapytania SQL i skrypty Python w jednym procesie
- Zależności między zadaniami, warunki (wykonaj, jeśli...) i pętle po liście wartości
- Parametry – jeden job uruchamiany dla różnych dat, krajów lub środowisk
- Obsługa błędów – ponowienia, limity czasu, powiadomienia e-mail i na komunikatory
- Wyzwalacze – harmonogram, pojawienie się nowego pliku, aktualizacja tabeli źródłowej; kilka wyzwalaczy w jednym jobie (Beta)
- Ponowne uruchomienie tylko nieudanych zadań (repair run)
- Warsztat 2: uczestnik buduje job, który uruchamia pipeline z warsztatu 1 i zadanie kontrolne, z parametrami, ponowieniami i alertem. Efektem jest zaplanowany i monitorowany proces.

7. Moduł – Przygotowanie do wdrożenia

Cel modułu: uczestnik wie, co trzeba przygotować, zanim pipeline trafi na produkcję.

- Pipeline i job jako kod (Declarative Automation Bundles) – wprowadzenie; pełne wdrożenia omawiamy w DBX-OPS
- Środowisko deweloperskie a produkcyjne – różnice w ustawieniach pipeline'u
- Kryteria przekazania do wdrożenia i podstawowa instrukcja postępowania przy awarii



Oczekiwane przygotowanie uczestnika

- Ukończone szkolenie DBX-ETL (Data Engineering end-to-end) lub równoważne doświadczenie: Delta Lake, przyrostowe ładowanie danych, MERGE
- Praktyczna znajomość SQL i podstawowa znajomość Python
- Podstawowa znajomość Git

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 danymi Bronze i szkieletem pipeline'u na czas szkolenia
- notatki i materiały do ćwiczeń 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

Wykład: polski

Materiały: angielski