

Databricks: Data Engineering end-to-end – od źródła do warstwy Gold

Większość procesów ETL działa – do pierwszego ponownego uruchomienia. Na tym szkoleniu budujesz proces, który da się uruchomić drugi raz bez duplikatów i przekazać do utrzymania.

Szkolenie Databricks: Data Engineering end-to-end to warsztat budowy kompletnego procesu przetwarzania danych w Databricks. Program koncentruje się na jednym realnym problemie: zasileniu danych z pliku źródłowego i jego przyrostów, przetworzeniu ich do warstwy Silver, kontroli jakości oraz przygotowaniu stabilnej tabeli Gold dla odbiorcy biznesowego.

Uczestnicy przechodzą przez cały proces: od kontraktu danych i decyzji projektowych, przez Bronze, Silver z deduplikacją i MERGE, po tabelę Gold ze wskaźnikami, osobną ścieżkę dla błędnych rekordów i uruchomienie całości jako Lakeflow Job. Dzięki temu szkolenie pokazuje nie tylko, jak napisać transformacje w SQL i PySpark, ale też jak zbudować proces, który można bezpiecznie uruchamiać wielokrotnie i przekazać do utrzymania. Proces budujemy dla RetailOps (fikcyjna firma handlowa, na której danych realizujemy wszystkie ćwiczenia).



Odbiorcy szkolenia

Szkolenie jest przeznaczone dla osób technicznych, które budują lub utrzymują procesy ETL/ELT w Databricks.

W szczególności kurs sprawdzi się dla:

- inżynierów danych odpowiedzialnych za przepływy od źródła do warstwy Gold,
- programistów ETL migrujących procesy z narzędzi klasycznych do Databricks,

- analityków technicznych, którzy przygotowują dane dla raportów i modeli,
- zespołów, które chcą ujednolicić sposób pisania transformacji i obsługi błędów.

Szkolenie będzie szczególnie wartościowe dla osób, które znają już Delta Lake i chcą zobaczyć, jak poszczególne elementy składają się w jeden powtarzalny proces.



Korzyści

- Projekt procesu end-to-end – nauczysz się zaczynać od kontraktu danych, terminu dostarczenia danych (SLA) i klucza biznesowego, zanim powstanie pierwsza linia kodu.
- Ingestion z metadanymi – dowiesz się, jak ładować dane wsadowo i przyrostowo oraz jak dobrać wzorzec dla API, JDBC i CDC.
- Transformacje w warstwie Silver – poznasz standaryzację, deduplikację, łączenie danych i obsługę danych spóźnionych.
- Powtarzalność procesu (idempotentność) – ponowne uruchomienie nie dubluje ani nie psuje danych; nauczysz się tak projektować MERGE i odczyt zmian.
- Jakość danych – zbudujesz reguły walidacji i osobną ścieżkę dla rekordów odrzuconych – tabelę błędów (kwarantanna).
- Warstwa Gold dla biznesu – przygotujesz tabelę ze wskaźnikami i sprawdzisz ją na przypadkach brzegowych.
- Uruchomienie i utrzymanie – dowiesz się, jak parametryzować proces, ustawić ponowienia i przekazać go do utrzymania.



Program szkolenia

1. Moduł – Projekt procesu: kontrakt danych i wzorzec end-to-end

Cel modułu: uczestnik zaczyna od ustaleń ze źródłem danych i odbiorcą, zanim napisze kod.

- Kontrakt danych – co dostajemy ze źródła: struktura, typy, klucz, jak często i z jakim opóźnieniem
- Termin dostarczenia danych (SLA) – na kiedy odbiorca potrzebuje danych i co się dzieje, gdy źródło się spóźni
- Podział odpowiedzialności – które błędy poprawia system źródłowy, a które obsługuje platforma danych
- Ponowne uruchomienia i dane spóźnione – decyzje, które trzeba podjąć na etapie projektu
- Wzorzec procesu RetailOps: dane surowe → oczyszczone → raportowe oraz osobna tabela na błędne rekordy
- Lab: uczestnik opisuje źródło danych zamówień i zapisuje decyzje projektowe

2. Moduł – Ładowanie danych do warstwy Bronze

Cel modułu: uczestnik ładuje dane ze źródła w sposób powtarzalny i z pełną informacją o pochodzeniu.

- Ładowanie jednorazowe i przyrostowe – kiedy wystarczy wczytać plik, a kiedy potrzebne jest wykrywanie nowych plików (mechanizmy ładowania omówione szczegółowo w DBX-LKH)
- Metadane – nazwa pliku, czas załadowania i identyfikator partii przy każdym rekordzie
- Inne źródła: API, bazy danych, strumień zmian (CDC) – jaki wzorzec wybrać (pokaz gotowego rozwiązania)
- Zasada warstwy Bronze – zapisujemy dane tak, jak przyszły, bez poprawiania
- Lab: uczestnik łączy dane zamówień do Bronze razem z metadanymi

3. Moduł – Transformacje w warstwie Silver (SQL i PySpark)

Cel modułu: uczestnik zamienia surowe dane w spójne, oczyszczone tabele.

- Ujednolicenie danych – typy, formaty dat, wielkość liter, kody i nazwy kolumn
- Wartości puste i błędne – uzupełnianie, oznaczanie lub odrzucanie
- Usuwanie duplikatów – wybór najnowszej wersji rekordu po kluczu (funkcje okna)
- Łączenie z tabelami słownikowymi i obsługa rekordów bez dopasowania
- Dane zagnieżdżone z JSON i API – rozpakowanie list i struktur do zwykłych kolumn (omówienie na przykładzie)
- Dane spóźnione – aktualizacja okresów, które zostały już wcześniej przetworzone
- Lab: uczestnik buduje transformację, która poprawnie obsługuje powtórzone i spóźnione rekordy

4. Moduł – Aktualizacja tabel i powtarzalność procesu

Cel modułu: uczestnik buduje proces, który przy ponownym uruchomieniu daje ten sam wynik.

- Krótkie przypomnienie MERGE z DBX-LKH; nacisk na aktualizację dla każdej partii danych przyrostowych (foreachBatch)
- Change Data Feed – przekazanie do kolejnej warstwy tylko zmienionych rekordów; nowa, automatyczna wersja nie wymaga włączania tej funkcji na każdej tabeli (Databricks Runtime 19+, row tracking); w ćwiczeniach używamy wersji klasycznej, działającej na Runtime LTS
- Test powtarzalności – kilkukrotne uruchomienie procesu na tych samych danych i porównanie wyniku
- Lab: uczestnik uruchamia proces trzy razy i sprawdza, że liczba rekordów się nie zmienia

5. Moduł – Jakość danych i ścieżka błędów

Cel modułu: uczestnik wie, jak wykrywać błędne dane i co z nimi zrobić, żeby nie zatrzymywały całego procesu.

- Reguły jakości – kompletność, zakres wartości, format, istnienie powiązanego klucza
- Ograniczenia na tabelach Delta – co blokują przy zapisie, a czego nie wykryją
- Wzorzec kwarantanny – błędne rekordy trafiają do osobnej tabeli z opisem przyczyny, poprawne idą dalej
- Metryki jakości zapisywane po każdym uruchomieniu – podstawa do raportowania jakości danych
- Lab: uczestnik dodaje reguły walidacji i kieruje błędne rekordy do tabeli błędów

6. Moduł – Warstwa Gold i uruchomienie procesu

Cel modułu: uczestnik przygotowuje dane dla biznesu i uruchamia cały proces jako zaplanowane zadanie.

- Tabele Gold – wskaźniki sprzedaży, porównania okres do okresu, średnie kroczące
- Testy przypadków brzegowych – pusta dostawa, zwroty, korekty zamówień
- Lakeflow Jobs – uruchomienie procesu jako job z parametrem daty – pełna orkiestracja w DBX-LKF

- Warsztat: uczestnicy składają kroki zbudowane w modułach 2-5 w jeden proces Bronze → Silver → Gold z tabelą błędów i uruchamiają go jako Lakeflow Job. Efektem jest proces, który można uruchamiać wielokrotnie z tym samym wynikiem.
- Przegląd kodu i lista dobrych praktyk do repozytorium zespołu



Oczekiwane przygotowanie uczestnika

- Ukończone szkolenie DBX-LKH (Lakehouse i tabele Delta) lub równoważna wiedza: Delta Lake, MERGE, Auto Loader
- Praktyczna znajomość SQL
- Podstawowa znajomość Python; elementy PySpark (DataFrame, funkcje okna) omawiamy na zajęciach, kod startowy zapewniamy

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

- przygotowane środowisko Databricks z danymi scenariusza RetailOps na czas szkolenia
- noteboki 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, 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 certyfikacyjnego 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 certyfikacyjnego Databricks Certified Data Engineer Associate jako kolejny krok rozwojowy.