kod szkolenia: DBX-DE-ML2D / PL DL 2d

Databricks Data Engineering & Data Preparation for ML

Dwudniowe szkolenie prowadzi uczestników przez cały proces przygotowania i transformacji danych w środowisku Databricks na potrzeby uczenia maszynowego. Pierwszy dzień stanowi pigułkę z inżynierii danych – od wczytywania danych i podstawowych transformacji po wykorzystanie Delta Lake, Auto Loader i architektury Medallion z deklaratywnymi pipeline'ami Delta Live Tables. Drugi dzień skupia się na czystym przygotowaniu danych do modeli – eksploracji, czyszczeniu, inżynierii cech, podziale na zbiory uczące i testowe, a także logowaniu i udostępnianiu cech przy użyciu MLflow i Feature Store. Szkolenie bazuje na oficjalnych materiałach Databricks Academy oraz dokumentacji Databricks.

Odbiorcy szkolenia

Szkolenie przeznaczone jest dla:

- data scientistów i inżynierów uczenia maszynowego, którzy chcą tworzyć własne pipeline'y ML w Databricks,
- zespołów AI/ML i DataOps odpowiedzialnych za jakość i przygotowanie danych,
- inżynierów danych przygotowujących zestawy danych pod modele ML,
- uczestników znających podstawy Pythona/PySpark i modelowania ML.

Korzyści

- Zdobycie solidnych podstaw inżynierii danych w środowisku Databricks: wczytywanie i walidacja danych, użycie DataFrame API, operacje ACID w Delta Lake, wersjonowanie i time travel.
- Poznanie narzędzi do strumieniowego i incrementalnego ładowania danych (Auto Loader) oraz

korzyści takich jak skalowalność, kontrola schematu i dokładnie raz przetwarzane pliki.
- Zrozumienie architektury Medallion (Bronze→Silver→Gold) i budowy deklaratywnych pipeline'ów z Delta Live Tables, które automatycznie zarządzają infrastrukturą i gwarantują niezawodność.
- Nabycie umiejętności eksploracji, czyszczenia i wzbogacania danych na potrzeby uczenia maszynowego (EDA, imputacja, kodowanie, skalowanie, funkcje okienkowe).
- Nauczenie się budowania kompletnych pipeline'ów ML z wykorzystaniem Spark MLlib i MLflow, a także zarządzania cechami w Feature Store dla spójności trenowania i serwowania modeli.

Program szkolenia

Dzień 1 – Data Engineering for Machine Learning (pigułka)

1.Wprowadzenie do platformy Databricks: przegląd architektury Lakehouse, workspace, klastry, notebooki, DBFS.

2.Ingest i podstawowe transformacje:
- Tworzenie DataFrame'ów w SQL i PySpark; operacje select, filter, join, groupBy.
- Ładowanie danych z różnych formatów (CSV, JSON, Parquet) oraz zapis/odczyt tabel Delta.
- Zarządzanie schematem, ACID, time travel, MERGE, UPDATE i DELETE w Delta Lake.

3.Strumieniowe i incrementalne ładowanie danych:
- Wprowadzenie do Auto Loader i cloudFiles do automatycznego wykrywania nowych plików i przetwarzania ich dokładnie raz.
- Porównanie trybów „trigger once" i „continuous"; obsługa schema evolution i wielu formatów plików.
- Structured Streaming: readStream / writeStream, checkpointing i fault tolerance.

4.Architektura Medallion i Delta Live Tables:
- Omówienie koncepcji warstw Bronze, Silver, Gold i ich zastosowania w przygotowaniu danych.
- Praca z Delta Live Tables – deklaratywne definiowanie tabel i widoków, tworzenie pipeline'u w GUI, ustalanie harmonogramów, oczekiwań jakości danych, monitoring i lineage.
- Łączenie DLT z Auto Loader i Structured Streaming w jednym przepływie.

5.Orkiestracja zadań i governance:
- Użycie Databricks Workflows do tworzenia zadań i multi-task jobów.
- Podstawy zarządzania dostępem i katalogiem danych w Unity Catalog.
- Dobre praktyki wersjonowania kodu (Repos) i monitorowania wydajności (Spark UI).

Dzień 2 – Data Preparation & Feature Engineering

1.Eksploracja danych (EDA):
- Użycie polecenia display, summary i profile do eksploracji rozkładów i zależności.
- Tworzenie wizualizacji w notebookach Databricks; identyfikacja odstających obserwacji.

2.Podział danych:
- Podział na zbiory treningowe, walidacyjne i testowe; losowy i warstwowy sampling.
- Cross-validation i koncepcja time-based split.

3.Imputacja brakujących wartości:
- Metody usuwania braków (dropna) i uzupełniania (fillna).
- Użycie klasy Imputer z MLlib do zastępowania wartości średnią/medianą oraz tworzenia flag braków.

4.Kodowanie i transformacja cech:
- Kodowanie kategorii za pomocą StringIndexer i OneHotEncoder; wprowadzenie do kodowania docelowego.
- Skalowanie zmiennych numerycznych przy użyciu StandardScaler, MinMaxScaler i RobustScaler.
- Funkcje okienkowe (window functions): lag, lead, row_number, rolling average dla tworzenia cech sekwencyjnych.

5.Feature engineering i selekcja:
- Różnica między ekstrakcją cech a selekcją; tworzenie nowych zmiennych przez agregacje, transformacje logarytmiczne, interakcje.
- Łączenie zmiennych w wektor za pomocą VectorAssembler

6.Budowa pipeline'ów ML:
- Definiowanie etapów Pipeline (imputacja, kodowanie, skalowanie, model).
- Dopasowanie modeli i ewaluacja; logowanie metryk, hiperparametrów i modeli w MLflow.

7.Feature Store & MLflow:
- Tworzenie i rejestrowanie tabel cech w Databricks Feature Store; utrzymanie wersji i udostępnianie cech dla wielu modeli.
- Zastosowanie mlflow.log_model i mlflow.register_model do śledzenia eksperymentów oraz gwarancja spójności między treningiem a serwowaniem modeli.

8.Dobre praktyki i testowanie:
- Projektowanie modularnych notebooków oraz wykorzystanie testów jednostkowych dla funkcji transformacyjnych.
- Dokumentowanie pipeline'ów, monitorowanie jakości danych i zarządzanie lineage.

Oczekiwane przygotowanie uczestnika

Uczestnik powinien znać podstawowe koncepcje uczenia maszynowego (np. regresja i klasyfikacja), posługiwać się PySpark lub SQL oraz mieć doświadczenie w pracy z danymi lub pipeline'ami.

Szkolenie obejmuje

* dostęp do portalu słuchacza Altkom Akademii

Czas trwania

2 dni / 14 godzin

Język

Wykład: polski
Materiały: angielski