Data Engineering v Azure Databricks

Kód kurzu: MOC DP-750

Kurz je určený pre dátových inžinierov a dátových profesionálov, ktorí sa chcú naučiť navrhovať, implementovať a prevádzkovať kompletné riešenia dátového inžinierstva s využitím platformy Azure Databricks a služby Unity Catalog. Na školení pochopíte kľúčové koncepty platformy Azure Databricks, naučíte sa vyberať a konfigurovať vhodné výpočtové prostriedky a vyskúšate si organizovať dátové objekty v Unity Catalog s dôrazom na zabezpečenie, governance a sledovanie pôvodu dát (data lineage). Naučíte sa navrhovať dátové modely vrátane dimenzionálneho modelovania a Slowly Changing Dimensions, načítavať dáta rôznymi spôsobmi (Lakeflow Connect, Auto Loader, Spark Structured Streaming, Lakeflow Spark Declarative Pipelines), čistiť a transformovať dáta a vynucovať dátovú kvalitu pomocou pipeline expectations. Ďalej sa naučíte navrhovať a implementovať dátové pipeline v rámci medallion architektúry, automatizovať ich prostredníctvom Lakeflow Jobs, aplikovať osvedčené postupy vývojového cyklu (Git, testovanie, Declarative Automation Bundles, Databricks CLI) a monitorovať a optimalizovať záťaž vrátane diagnostiky problémov. Kurz je zároveň komplexnou prípravou na skúšku DP-750 Microsoft Certified: Azure Databricks Data Engineer Associate.

Čo Vás naučíme

- Zoznámite sa s platformou Azure Databricks a jej kľúčovými konceptmi
- Naučíte sa vyberať a konfigurovať vhodné výpočtové prostriedky pre rôzne scenáre
- Vyskúšate si vytvárať a organizovať objekty v Unity Catalog vrátane schém, tabuliek, pohľadov a volumes
- Pochopíte, ako zabezpečiť dáta pomocou fine-grained access control, row filtering, column masking a Azure Key Vault
- Dozviete sa, ako aplikovať data governance cez riadenie prístupu na základe atribútov, retenčné politiky, data lineage, audit logging a Delta Sharing
- Naučíte sa navrhovať dátové modely vrátane partitioningu, clusteringu a Slowly Changing Dimensions (SCD Type 2)
- Vyskúšate si extrahovať a načítavať dáta s využitím Lakeflow Connect, Auto Loader, Spark Structured Streaming a Lakeflow Spark Declarative Pipelines
- Naučíte sa čistiť a transformovať dáta pomocou PySpark a SQL operácií (joiny, agregácie, pivoty, merge)
- Pochopíte, ako vynucovať dátovú kvalitu pomocou pipeline expectations a riadiť schema drift
- Naučíte sa navrhovať a implementovať medallion architektúru (Bronze › Silver › Gold)
- Vyskúšate si automatizovať dátové pipeline pomocou Lakeflow Jobs s triggermi, plánovaním, alertmi a retry policies
- Zoznámite sa s vývojovým cyklom v Azure Databricks: Git, testovanie pomocou pytest, Declarative Automation Bundles a Databricks CLI
- Naučíte sa monitorovať a optimalizovať záťaž a diagnostikovať problémy s caching, data skew, memory spill a shuffle

Pre koho je kurz určený

- Dátovým inžinierom, ktorí chcú navrhovať a implementovať riešenia dátového inžinierstva na platforme Azure Databricks s využitím Unity Catalog.
- Dátovým a BI architektom, ktorí chcú pochopiť architektúru moderného lakehouse riešenia postaveného nad Azure Databricks a Delta Lake.
- Dátovým profesionálom, ktorí sa chcú pripraviť na certifikačnú skúšku Microsoft DP-750.

Požadované vstupné znalosti

- Základná znalosť jazyka SQL a relačných databáz
- Základná znalosť jazyka Python a frameworku Apache Spark (najmä PySpark)
- Základná znalosť princípov návrhu dátových skladov a implementácie ETL/ELT procesov
- Odporúčaná základná znalosť dátových služieb v Microsoft Azure na úrovni kurzu MOC DP-900
- Odporúčaná základná orientácia v platforme Azure Databricks a formáte Delta Lake

GOPAS Praha
Na Strži 2097/63
140 00 Praha 4 - Krč
Tel.: +420 226 201 390
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 530 513 590
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 902 903 132
info@gopas.sk

Copyright © 2026 GOPAS, a.s.,
All rights reserved

Data Engineering v Azure Databricks

Metódy výuky

- Odborný výklad s praktickými ukážkami, cvičenia na počítačoch

Študijné materiály

- Oficiálne elektronické študijné materiály spoločnosti Microsoft v anglickom jazyku

Osnova kurzu

1 Zoznámenie s Azure Databricks

- Zoznámite sa s platformou Azure Databricks a zorientujete sa vo workspace UI
- Spoznáte typické workloady, na ktoré je Azure Databricks určený
- Pochopíte kľúčové koncepty platformy
- Zoznámite sa s data governance cez Unity Catalog a Microsoft Purview
- Lab: Vyskúšate si nahranie datasetu do Unity Catalog volume, prácu v notebooku a využitie Databricks Assistant na scenári CityMoves Transit

2 Voľba a konfigurácia výpočtových prostriedkov

- Naučíte sa vybrať vhodný typ výpočtového prostriedku (compute) pre danú úlohu
- Dozviete sa, ako konfigurovať výpočtový výkon a behové prostredie na spúšťanie rôznych typov výpočtových úloh
- Zistíte, ako inštalovať knižnice na úrovni clustra aj notebooku
- Naučíte sa nastaviť prístup k výpočtovým zdrojom
- Lab: Vyskúšate si vytvorenie clustra, inštaláciu knižníc a generovanie syntetických dát pomocou PySpark a knižnice faker

3 Vytváranie a organizácia objektov v Unity Catalog

- Zoznámite sa s mennými konvenciami objektov v Unity Catalogu
- Vyskúšate si vytváranie katalógov, schém, tabuliek, pohľadov a volumes
- Pochopíte, ako vykonávať DDL operácie a implementovať foreign catalogs na pripojenie k externým dátovým zdrojom
- Dozviete sa, ako konfigurovať inštrukcie pre AI/BI Genie
- Lab: Zostavíte kompletný namespace pre univerzitnú dátovú platformu — medallion schémy, managed tabuľky s PK/FK, pohľady, volume a SQL funkcie

4 Zabezpečenie objektov v Unity Catalog

- Pochopíte query lifecycle a stratégie riadenia prístupu (access control)
- Naučíte sa implementovať jemne granulárne riadenie prístupu (fine-grained access control), row filtering a column masking
- Dozviete sa, ako pracovať s uloženými tajomstvami (secrets) cez Azure Key Vault
- Naučíte sa autentizovať prístup k dátam cez service principals a k zdrojom cez managed identities
- Lab: Vyskúšate si nastavenie oprávnení, row filtrov na obmedzenie prístupu k dátam podľa regiónu a maskovanie e-mailov a ochránite citlivé prístupové údaje pomocou Azure Key Vault

5 Správa a riadenie objektov v Unity Catalog

- Naučíte sa vytvárať a uchovávať definície tabuliek a konfigurovať Attribute-Based Access Control (riadenie prístupu na základe atribútov) pomocou tagov a politík
- Zistíte, ako aplikovať politiky retencie dát (vrátane VACUUM a predictive optimization)
- Naučíte sa nastaviť a spravovať data lineage a audit logging
- Dozviete sa, ako navrhnúť bezpečnú stratégiu zdieľania dát pomocou protokolu Delta Sharing
- Lab: Vyskúšate si governance pre connected vehicle platform — PII tagy, retenčné politiky, dotazovanie systémových tabuliek na lineage a analýzu audit logu

6 Návrh a implementácia dátového modelovania

GOPAS Praha
Na Strži 2097/63
140 00 Praha 4 - Krč
Tel.: +420 226 201 390
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 530 513 590
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 902 903 132
info@gopas.sk

Copyright © 2026 GOPAS, a.s.,
All rights reserved

- Naučíte sa navrhnúť logiku na načítanie dát, vybrať vhodné nástroje a zvoliť vhodný tabuľkový formát
- Pochopíte, ako navrhnúť a implementovať partitioning a clustering stratégie
- Dozviete sa, ako vybrať a implementovať typ Slowly Changing Dimension (najmä SCD Type 2) a temporálne (history) tabuľky
- Naučíte sa rozhodovať medzi managed a unmanaged tabuľkami a voliť správnu granularitu agregácie dát
- Lab: Navrhnete Delta Lake model pre retail banking — customer dimenziu s SCD Type 2, faktovú tabuľku s liquid clustering, Change Data Feed a vyskúšate si Delta time travel

7 Načítanie dát do Unity Catalog

- Naučíte sa extrahovať a načítavať dáta cez Lakeflow Connect, notebooky a SQL metódy
- Dozviete sa, ako pracovať s CDC feedom a Spark Structured Streaming
- Zistíte, ako využívať Auto Loader na automatické spracovanie súborov z cloudového úložiska
- Vyskúšate si Lakeflow Spark Declarative Pipelines na deklaratívny popis načítania dát
- Lab: Načítate CSV súbory z Unity Catalog volume do Delta tabuliek cez PySpark, COPY INTO a CTAS a nakonfigurujete Auto Loader na spracovanie nových súborov

8 Čistenie, transformácia a načítanie dát do Unity Catalog

- Naučíte sa profilovať dáta a vyberať správne dátové typy stĺpcov
- Zistíte, ako riešiť duplicity v dátach a NULL hodnoty
- Vyskúšate si transformáciu dát pomocou filtrov, agregácií, joinov, množinových operátorov, denormalizácie a pivotov
- Naučíte sa načítavať dáta cez operácie merge, insert a append
- Lab: Vyčistíte a reštrukturalizujete dáta o nehnuteľnostiach — zvolíte správne dátové typy, odstránite duplicitné dáta a skombinujete dáta z rôznych tabuliek pre potreby analýzy trendov

9 Implementácia a správa obmedzení kvality dát

- Naučíte sa implementovať validačné kontroly a kontroly dátových typov
- Dozviete sa, ako detegovať a riadiť schema drift
- Zistíte, ako spravovať kvalitu dát pomocou pipeline expectations
- Lab: Postavíte Lakeflow Spark Declarative Pipeline pre poisťovňu ClearCover — ktorá bude vynucovať potrebnú kvalitu vstupných dát a vyskúšate si monitoring metrík kvality dát

10 Návrh a implementácia dátových pipelines

- Naučíte sa navrhnúť poradie operácií v rámci pipeline a rozhodovať medzi notebookmi a Lakeflow Pipelines
- Pochopíte, ako navrhnúť logiku Lakeflow jobov a riešiť ošetrenie chýb
- Vyskúšate si vytváranie pipeline pomocou notebookov aj pomocou Lakeflow Spark Declarative Pipelines
- Lab: Postavíte medallion architektúru (Bronze › Silver › Gold) pre dáta z hotelov GlobStay — deduplikácia, validácia, agregácia dát, parametrizácia notebookov a konfigurácia Lakeflow Job so sekvenčnými závislosťami a retry policies

11 Implementácia Lakeflow Jobov

- Naučíte sa nakonfigurovať Lakeflow Joby
- Zistíte, ako konfigurovať triggery (časové aj event-based) a plánovanie úloh
- Dozviete sa, ako nastaviť alerty pre úspech/zlyhanie a automatické reštarty
- Lab: Zautomatizujete data pipeline pre TelConnect — parametrizovaný notebook spracovania dát o hovoroch cez bronze/silver/gold vrstvy, nakonfigurujete závislosti taskov, plánované aj event-based triggery, notifikácie a retry politiky

12 Implementácia procesov riadenia vývoja

- Naučíte sa aplikovať Git verzovanie a správu vetiev a pull requestov
- Dozviete sa, ako implementovať testovaciu stratégiu pre dátové pipeline
- Zistíte, ako konfigurovať a baliť Declarative Automation Bundles
- Vyskúšate si nasadzovanie bundles pomocou Databricks CLI

GOPAS Praha
Na Strži 2097/63
140 00 Praha 4 - Krč
Tel.: +420 226 201 390
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 530 513 590
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 902 903 132
info@gopas.sk

Copyright © 2026 GOPAS, a.s.,
All rights reserved

- Lab: Vyskúšate si implementovať testovaciu stratégiu pomocou knižnice pytest a zabalíte a nasadíte transformačnú pipeline ako Declarative Automation Bundle cez Databricks CLI

13 Monitorovanie, riešenie problémov a optimalizácia záťaže

- Naučíte sa monitorovať a riadiť spotrebu výpočtových prostriedkov clustrov
- Dozviete sa, ako diagnostikovať a opravovať Lakeflow Joby, Spark joby a notebooky
- Zistíte, ako diagnostikovať problémy s caching, data skew, memory spill a shuffle pomocou Spark UI
- Naučíte sa implementovať streamovanie logov do Azure Log Analytics
- Lab: Vygenerujete syntetické workloady s úmyselným data skew a nadmerným shuffle, diagnostikujete ich v Spark UI a aplikujete cielené opravy pomocou broadcast joinov, Adaptive Query Execution a technikami na redukciu shuffle

GOPAS Praha
Na Strži 2097/63
140 00 Praha 4 - Krč
Tel.: +420 226 201 390
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 530 513 590
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 902 903 132
info@gopas.sk

Copyright © 2026 GOPAS, a.s.,
All rights reserved