Data Engineering v Azure Databricks

Kód kurzu: MOC DP-750

Kurz je určen pro datové inženýry a datové profesionály, kteří se chtějí naučit navrhovat, implementovat a provozovat kompletní řešení datového inženýrství s využitím platformy Azure Databricks a služby Unity Catalog. Na školení pochopíte klíčové koncepty platformy Azure Databricks, naučíte se vybírat a konfigurovat vhodné výpočetní prostředky a vyzkoušíte si organizovat datové objekty v Unity Catalog s důrazem na zabezpečení, governance a sledování původu dat (data lineage). Naučíte se navrhovat datové modely včetně dimenzionálního modelování a Slowly Changing Dimensions, načítat data různými způsoby (Lakeflow Connect, Auto Loader, Spark Structured Streaming, Lakeflow Spark Declarative Pipelines), čistit a transformovat data a vynucovat datovou kvalitu pomocí pipeline expectations. Dále se naučíte navrhovat a implementovat datové pipeline v rámci medallion architektury, automatizovat je prostřednictvím Lakeflow Jobs, aplikovat osvědčené postupy vývojového cyklu (Git, testování, Declarative Automation Bundles, Databricks CLI) a monitorovat a optimalizovat zátěž včetně diagnostiky problémů. Kurz je zároveň komplexní přípravou na zkoušku DP-750 Microsoft Certified: Azure Databricks Data Engineer Associate.

Co Vás naučíme

- Seznámíte se s platformou Azure Databricks a jejími klíčovými koncepty
- Naučíte se vybírat a konfigurovat vhodné výpočetní prostředky pro různé scénáře
- Vyzkoušíte si vytvářet a organizovat objekty v Unity Catalog včetně schémat, tabulek, pohledů a volumes
- Pochopíte, jak zabezpečit data pomocí fine-grained access control, row filtering, column masking a Azure Key Vault
- Dozvíte se, jak aplikovat data governance přes řízení přístupu na základě atributů, retenční politiky, data lineage, audit logging a Delta Sharing
- Naučíte se navrhovat datové modely včetně partitioningu, clusteringu a Slowly Changing Dimensions (SCD Type 2)
- Vyzkoušíte si extrahovat a načítat data s využitím Lakeflow Connect, Auto Loader, Spark Structured Streaming a Lakeflow Spark Declarative Pipelines
- Naučíte se čistit a transformovat data pomocí PySpark a SQL operací (joiny, agregace, pivoty, merge)
- Pochopíte, jak vynucovat datovou kvalitu pomocí pipeline expectations a řídit schema drift
- Naučíte se navrhovat a implementovat medallion architekturu (Bronze › Silver › Gold)
- Vyzkoušíte si automatizovat datové pipeline pomocí Lakeflow Jobs s triggery, plánováním, alerty a retry policies
- Seznámíte se s vývojovým cyklem v Azure Databricks: Git, testování pomocí pytest, Declarative Automation Bundles a Databricks CLI
- Naučíte se monitorovat a optimalizovat zátěž a diagnostikovat problémy s caching, data skew, memory spill a shuffle

Pro koho je kurz určen

- Datovým inženýrům, kteří chtějí navrhovat a implementovat řešení datového inženýrství na platformě Azure Databricks s využitím Unity Catalog.
- Datovým a BI architektům, kteří chtějí pochopit architekturu moderního lakehouse řešení postaveného nad Azure Databricks a Delta Lake.
- Datovým profesionálům, kteří se chtějí připravit na certifikační zkoušku Microsoft DP-750.

Požadované vstupní znalosti

- Základní znalost jazyka SQL a relačních databází
- Základní znalost jazyka Python a frameworku Apache Spark (zejména PySpark)
- Základní znalost principů návrhu datových skladů a implementace ETL/ELT procesů
- Doporučena základní znalost datových služeb v Microsoft Azure na úrovni kurzu MOC DP-900
- Doporučena základní orientace v platformě Azure Databricks a formátu Delta Lake

Osnova kurzu

1 Seznámení s Azure Databricks

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

- Seznámíte se s platformou Azure Databricks a zorientujete se ve workspace UI
- Poznáte typické workloady, na které je Azure Databricks určen
- Pochopíte klíčové koncepty platformy
- Seznámíte se s data governance přes Unity Catalog a Microsoft Purview
- Lab: Vyzkoušíte si nahrání datasetu do Unity Catalog volume, práci v notebooku a využití Databricks Assistant na scénáři CityMoves Transit

2 Volba a konfigurace výpočetních prostředků

- Naučíte se vybrat vhodný typ výpočetního prostředku (compute) pro danou úlohu
- Dozvíte se, jak konfigurovat výpočetní výkon a běhové prostředí pro spouštění různých typů výpočetních úloh
- Zjistíte, jak instalovat knihovny na úrovni clusteru i notebooku
- Naučíte se nastavit přístup k výpočetním zdrojům
- Lab: Vyzkoušíte si vytvoření clusteru, instalaci knihoven a generování syntetických dat pomocí PySpark a knihovny faker

3 Vytváření a organizace objektů v Unity Catalog

- Seznámíte se s jmennými konvencemi objektů v Unity Catalogu
- Vyzkoušíte si vytváření katalogů, schémat, tabulek, pohledů a volumes
- Pochopíte, jak provádět DDL operace a implementovat foreign catalogs pro připojení k externím datovým zdrojům
- Dozvíte se, jak konfigurovat instrukce pro AI/BI Genie
- Lab: Sestavíte kompletní namespace pro univerzitní datovou platformu — medallion schémata, managed tabulky s PK/FK, pohledy, volume a SQL funkce

4 Zabezpečení objektů v Unity Catalog

- Pochopíte query lifecycle a strategie řízení přístupu (access control)
- Naučíte se implementovat jemně granulární řízení přístupu (fine-grained access control), row filtering a column masking
- Dozvíte se, jak pracovat s uloženými tajemstvími přes Azure Key Vault
- Naučíte se autentizovat přístup k datům přes service principals a ke zdrojům přes managed identities
- Lab: Vyzkoušíte si nastavení oprávnění, row filtrů pro omezení přístupu k datům podle regionu a maskování e-mailů a ochráníte citlivé přístupové údaje s pomocí Azure Key Vault

5 Správa a řízení objektů v Unity Catalog

- Naučíte se vytvářet a uchovávat definice tabulek a konfigurovat Attribute-Based Access Control (řízení přístupu na základě atributů) pomocí tagů a politik
- Zjistíte, jak aplikovat politiky retence dat (včetně VACUUM a predictive optimization)
- Naučíte se nastavit a spravovat data lineage a audit logging
- Dozvíte se, jak navrhnout bezpečnou strategii sdílení dat s pomocí protokolu Delta Sharing
- Lab: Vyzkoušíte si governance pro connected vehicle platform — PII tagy, retenční politiky, dotazování systémových tabulek na lineage a analýzu audit logu

6 Návrh a implementace datového modelování

- Naučíte se navrhnout logiku pro načtení dat, vybrat vhodné nástroje a zvolit vhodný tabulkový formát
- Pochopíte, jak navrhnout a implementovat partitioning a clustering strategie
- Dozvíte se, jak vybrat a implementovat typ Slowly Changing Dimension (zejména SCD Type 2) a temporální (history) tabulky
- Naučíte se rozhodovat mezi managed a unmanaged tabulkami a volit správnou granularitu agregace dat
- Lab: Navrhnete Delta Lake model pro retail banking — customer dimenzi s SCD Type 2, faktovou tabulku s liquid clustering, Change Data Feed a vyzkoušíte si Delta time travel

7 Načtení dat do Unity Catalog

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 se extrahovat a načítat data přes Lakeflow Connect, notebooky a SQL metody
- Dozvíte se, jak pracovat s CDC feedem a Spark Structured Streaming
- Zjistíte, jak využívat Auto Loader pro automatické zpracování souborů z cloudového úložiště
- Vyzkoušíte si Lakeflow Spark Declarative Pipelines pro deklarativní popis načtení dat
- Lab: Načtete CSV soubory z Unity Catalog volume do Delta tabulek přes PySpark, COPY INTO a CTAS a nakonfigurujete Auto Loader pro zpracování nových souborů

8 Čištění, transformace a načítání dat do Unity Catalog

- Naučíte se profilovat data a vybírat správné datové typy sloupců
- Zjistíte, jak řešit duplicity v datech a NULL hodnoty
- Vyzkoušíte si transformaci dat pomocí filtrů, agregací, joinů, množinových operátorů, denormalizace a pivotů
- Naučíte se načítat data přes operace merge, insert a append
- Lab: Vyčistíte a restrukturalizujete data o nemovitostech — zvolíte správné datové typy, odstraníte duplicitní data a zkombinujete data z různých tabulek pro potřeby analýzy trendů

9 Implementace a správa omezení kvality dat

- Naučíte se implementovat validační kontroly a kontroly datových typů
- Dozvíte se, jak detekovat a řídit schema drift
- Zjistíte, jak spravovat kvalitu dat pomocí pipeline expectations
- Lab: Postavíte Lakeflow Spark Declarative Pipeline pro pojišťovnu ClearCover — která bude vynucovat potřebnou kvalitu vstupních dat a zkusíte si monitoring metrik kvality dat

10 Návrh a implementace datových pipelines

- Naučíte se navrhnout pořadí operací v rámci pipeline a rozhodovat mezi notebooky a Lakeflow Pipelines
- Pochopíte, jak navrhnout logiku Lakeflow jobů a řešit ošetření chyb
- Vyzkoušíte si vytváření pipeline jak pomocí notebooků, tak pomocí Lakeflow Spark Declarative Pipelines
- Lab: Postavíte medallion architekturu (Bronze › Silver › Gold) pro data z hotelů GlobStay — deduplikace, validace, agregace dat, parametrizace notebooků a konfigurace Lakeflow Job se sekvenčními závislostmi a retry policies

11 Implementace Lakeflow Jobů

- Naučíte se nakonfigurovat Lakeflow Joby
- Zjistíte, jak konfigurovat triggery (časové i event-based) a plánování úloh
- Dozvíte se, jak nastavit alerty pro úspěch/selhání a automatické restarty
- Lab: Zautomatizujete data pipeline pro TelConnect — parametrizovaný notebook zpracování dat o hovorech přes bronze/silver/gold vrstvy, nakonfigurujete závislosti tasků, plánované i event-based triggery, notifikace a retry politiky

12 Implementace procesů řízení vývoje

- Naučíte se aplikovat Git verzování a správu větví a pull requestů
- Dozvíte se, jak implementovat testovací strategii pro datové pipeline
- Zjistíte, jak konfigurovat a balíčkovat Declarative Automation Bundles
- Vyzkoušíte si nasazování bundles pomocí Databricks CLI
- Lab: Vyzkoušíte si implementovat testovací strategii s pomocí knihovny pytest a zabalíte a nasadíte transformační pipeline jako Declarative Automation Bundle přes Databricks CLI

13 Monitorování, řešení problémů a optimalizace zátěže

- Naučíte se monitorovat a řídit spotřebu výpočetních prostředků clusterů
- Dozvíte se, jak diagnostikovat a opravovat Lakeflow Joby, Spark joby a notebooky
- Zjistíte, jak diagnostikovat problémy s caching, data skew, memory spill a shuffle pomocí Spark UI
- Naučíte se implementovat streamování logů do Azure Log Analytics
- Lab: Vygenerujete syntetické workloady s úmyslným data skew a nadměrným shuffle, diagnostikujete je ve Spark UI a aplikujete cílené opravy s pomocí broadcast joinů, Adaptive Query Execution a technikami na redukci 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