



DP-750T00

Implement data engineering solutions using Azure Databricks



Información general.

Este curso va desde la configuración básica hasta el despliegue en producción, cubriendo la configuración del entorno y la gobernanza a nivel empresarial. Aprende a construir pipelines de ingesta robustos, implementar seguridad con Unity Catalog y desplegar cargas de trabajo optimizadas. Al final, tendrás las habilidades prácticas para implementar, asegurar y mantener soluciones escalables de casas en el lago que cumplan con los rigurosos requisitos empresariales.

Perfil del público.

El público objetivo son ingenieros de datos que tienen conocimientos fundamentales sobre conceptos de análisis de datos, un conocimiento básico del almacenamiento en la nube y familiaridad con los principios de organización de datos. Deberían sentirse cómodos trabajando con SQL y tener experiencia usando Python, incluidos los cuadernos, para tareas de ingeniería de datos. Se espera que los estudiantes tengan un buen conocimiento de los espacios de trabajo de Azure Databricks y del Catálogo de Unity, junto con familiaridad con los patrones de acceso a datos y conceptos básicos de ingeniería de datos y almacenes de datos. Además, deben tener conocimientos básicos de seguridad en Azure, incluyendo Microsoft Entra ID, y estar familiarizados con los fundamentos del control de versiones de Git.

Duración.

4 Días.

Examen.

DP-750: Implementing Data Engineering Solutions Using Azure Databricks

Temario.

Ruta de aprendizaje: Configurar un entorno Azure Databricks.

Construye una base sólida en Azure Databricks comprendiendo su arquitectura, integraciones, opciones de cálculo y capacidades de organización de datos. Descubre cómo Azure Databricks ofrece una plataforma unificada para cargas de trabajo de ingeniería de datos, análisis e IA en la nube.

En esta ruta de aprendizaje, explorarás los fundamentos de Azure Databricks y cómo encaja en el ecosistema moderno de las plataformas de datos. Empezarás aprovisionando espacios de trabajo y entendiendo las cargas de trabajo principales, luego profundizarás en los conceptos arquitectónicos que separan los planos de control y de cálculo. Descubrirás cómo Azure Databricks se integra perfectamente con Microsoft Fabric, Power BI, Visual Studio Code y otros servicios de Microsoft para crear soluciones integrales. Aprenderás a seleccionar y configurar los recursos de cómputo adecuados para tus cargas de trabajo, optimizando tanto en rendimiento como en coste. Por último, dominarás la estructura organizativa de Unity Catalog para gestionar eficazmente tus activos de datos. Al final, tendrás los conocimientos básicos necesarios para construir soluciones de datos escalables en Azure Databricks.

Módulo 1: Explore Azure Databricks.

Azure Databricks es un servicio en la nube que proporciona una plataforma escalable para el análisis de datos utilizando Apache Spark.

- Introducción.
- Empieza con Azure Databricks.
- Identificar cargas de trabajo Azure Databricks.
- Entiende los conceptos clave.
- Gobernanza de datos usando Unity Catalog y Microsoft Champview.

Ejercicio: Explorar Azure Databricks.

Módulo 2: Comprender la arquitectura de Azure Databricks.

La arquitectura de Azure Databricks separa los planos de control y de cómputo mientras organiza los recursos mediante una estructura jerárquica. Este módulo explora cómo funciona la jerarquía de cuentas, las diferencias entre los planos de cómputo serverless y los clásicos, y las diversas opciones de almacenamiento disponibles, incluyendo almacenamiento por defecto, almacenamiento externo y almacenamiento gestionado por Unity Catalog para organizar y gobernar tus datos.

- Comprender la arquitectura de Azure Databricks.
- Entiende el almacenamiento gestionado del Unity Catalog.
- Entender el almacenamiento externo.
- Entender el almacenamiento por defecto.

Módulo 3: Comprender las integraciones de Azure Databricks.

Azure Databricks se integra con múltiples servicios de Microsoft para ofrecer capacidades integrales de ingeniería de datos, análisis e IA. Este módulo explora cómo Azure Databricks funciona con Microsoft Fabric, Power BI, Visual Studio Code, Power Platform, Copilot Studio, Microsoft Purview y Microsoft Foundry para habilitar soluciones integrales que combinan capacidades de data lake con inteligencia empresarial, desarrollo de aplicaciones e inteligencia artificial conversacional.

- Entender la integración con Microsoft Fabric.
- Entender la integración con Power BI.

- Entender la integración con VS Code.
- Comprender la integración con Power Platform.
- Entender la integración con Copilot Studio.
- Entender la integración con Microsoft Purview.
- Entender la integración con Microsoft Foundry.

Módulo 4: Seleccionar y configurar computación en Azure Databricks.

Azure Databricks ofrece múltiples opciones de cómputo optimizadas para diferentes cargas de trabajo. Este módulo explora cómo elegir el tipo de cómputo adecuado, configurar la configuración de rendimiento, gestionar permisos de acceso e instalar bibliotecas. Aprenderás cuándo usar computación sin servidor frente a computación clásica, cómo optimizar clústeres en coste y rendimiento, y mejores prácticas para asegurar recursos de cómputo.

- Elige un tipo de cómputo adecuado.
- Configurar el rendimiento de cómputo.
- Configurar características de cómputo.
- Instalar bibliotecas para computación.
- Configurar acceso de cómputo.

Ejercicio: Seleccionar y configurar el cálculo en Azure Databricks

Módulo 5: Crear y organizar objetos en el Catálogo de Unity.

El espacio de nombres de tres capas del Catálogo Unity — catálogos, esquemas y objetos— proporciona una base flexible para organizar los activos de datos manteniendo una gobernanza centralizada. Este módulo explora cómo crear catálogos para aislar el entorno, organizar esquemas dentro de esos catálogos y crear tablas, vistas y volúmenes para datos estructurados y no estructurados. Aprenderás a implementar catálogos extranjeros para acceder a bases de datos externas, aplicar convenciones de nombres efectivas y configurar instrucciones de AI/BI Genie para mejorar la detección de datos.

- Aplicar convenciones de nombres.
- Crear catálogo.
- Crear esquema.
- Crear tablas y vistas.





- Crear volúmenes.
- Implementar operaciones DDL.
- Implementar catálogo extranjero.
- Configurar instrucciones de AI/BI Genie.

Ejercicio: Crear y organizar objetos en el catálogo de Unity.

Ruta de aprendizaje: Asegurar y gobernar los objetos de Unity Catalog en Azure Databricks.

Aprende a implementar seguridad y gobernanza integrales para tus activos de datos en Azure Databricks usando Unity Catalog. Estrategias de control de acceso maestras, permisos detallados, gestión de credenciales y prácticas de gobernanza para construir una plataforma de datos segura y conforme.

En este camino de aprendizaje, aprenderás a proteger y gobernar tu patrimonio de datos con Unity Catalog. Comenzarás implementando controles de seguridad robustos, incluyendo permisos a nivel de tabla y esquema, filtrado de filas y enmascaramiento de columnas. Aprenderás a gestionar las credenciales de forma segura usando Azure Key Vault, los principales de servicio y las identidades gestionadas. Después, explorarás capacidades de gobernanza como el seguimiento de la línea de datos, el registro de auditorías y el control de acceso basado en atributos. Por último, aprenderás a compartir datos de forma segura con terceros usando Delta Sharing. Al final, estarás preparado para construir una plataforma de datos segura, conforme y bien gobernada que cumpla con los requisitos empresariales.

Módulo 6: Objetos del Catálogo Secure Unity.

Unity Catalog proporciona gobernanza centralizada y seguridad para los activos de datos en Azure Databricks. Este módulo explora cómo proteger los objetos del Catálogo Unity mediante estrategias de control de acceso, permisos detallados, gestión de credenciales y mecanismos de autenticación. Aprenderás a implementar seguridad a nivel de tabla y esquema, hacer cumplir el filtrado de filas y columnas, acceder de forma segura a secretos de Azure Key Vault y autenticar el acceso a datos usando principales de servicio e identidades gestionadas.

- Entender el ciclo de vida de las consultas.
- Implementar estrategias de control de acceso.

- Entender el control de acceso de gran detalle.
- Implementar filtrado de filas y enmascaramiento de columnas.
- Access Azure Key Vault secrets.
- Autenticar el acceso a los datos con los principales del servicio.
- Autenticar el acceso a recursos con identidades gestionadas.

Ejercicio: Objetos del catálogo Secure Unity.

Módulo 7: Gobernar objetos del Catálogo Unity.

Este módulo cubre prácticas esenciales de gobernanza en Unity Catalog, permitiéndote proteger, monitorizar y gestionar eficazmente tu patrimonio de datos. Aprenderás a implementar un control de acceso detallado, rastrear la línea de datos, configurar registros de auditoría y compartir datos de forma segura.

- Crear y preservar definiciones de tablas.
- Configurar ABAC con etiquetas y políticas.
- Aplicar políticas de retención de datos.
- Configurar y gestionar la línea de datos.
- Configurar el registro de auditoría.
- Diseña una estrategia segura de Delta Sharing.

Ejercicio: Gobernar objetos del catálogo de Unity.

Ruta de aprendizaje: Preparar y procesar datos con Azure Databricks.

Domina las habilidades esenciales para construir soluciones robustas y escalables de ingeniería de datos con Azure Databricks y Unity Catalog. Aprende a diseñar modelos de datos efectivos, a absorber datos de fuentes diversas, a transformar datos en bruto en formatos listos para análisis y a garantizar la calidad de los datos en toda la arquitectura de tu casa del lago.

En este camino de aprendizaje, aprenderás a construir un flujo de trabajo de ingeniería de datos usando Azure Databricks y Unity Catalog. Partiendo de conceptos fundamentales de modelado de datos, diseñarás esquemas y estrategias de particiones optimizadas para cargas de trabajo analíticas. Luego explorarás múltiples patrones de ingestión—desde conectores gestionados hasta pipelines de streaming—para llevar datos a tu casa del lago. A continuación, aplicarás técnicas de transformación para limpiar y remodelar datos para uso empresarial. Finalmente, implementarás

controles de calidad para mantener la integridad de los datos en todas tus canalizaciones. Al final, tendrás las habilidades prácticas para diseñar y construir soluciones de datos listas para producción en Unity Catalog.

Módulo 8: Diseñar e implementar modelado de datos con Azure Databricks.

Un modelado de datos eficaz constituye la base de una plataforma de datos eficiente y mantenible. Este módulo explora cómo diseñar lógica de ingestión, seleccionar herramientas y formatos de tabla apropiados, implementar esquemas de particionamiento, gestionar dimensiones que cambian lentamente, elegir la granularidad adecuada de los datos y optimizar el rendimiento de las tablas mediante estrategias de agrupamiento en Azure Databricks con Unity Catalog.

- Lógica de ingestión de diseño y configuración de la fuente de datos.
- Elige una herramienta de gestión de datos.
- Elige un formato de tabla de datos.
- Diseñar e implementar un esquema de partición de datos.
- Elige un tipo de dimensión que cambia lentamente (SCD).
- Implementar una dimensión que cambia lentamente (SCD) tipo 2.
- Diseñar e implementar una tabla temporal (historial) para registrar los cambios a lo largo del tiempo.
- Elige la granularidad en una columna o tabla según los requisitos.
- Elige tablas gestionadas o no gestionadas.
- Diseñar e implementar una estrategia de agrupamiento.

Ejercicio: Diseñar e implementar modelado de datos con Azure Databricks.

Módulo 9: Ingesta de datos en el Catálogo de Unity.

La ingestión de datos es una capacidad fundamental para cualquier plataforma de datos. Este módulo explora el conjunto completo de técnicas disponibles en Azure Databricks para cargar datos en tablas del Catálogo de Unity. Aprenderás a usar conectores gestionados con Lakeflow Connect, escribir código de ingestión personalizado en cuadernos, aplicar comandos SQL para cargar archivos por lotes, procesar fuentes de captura de datos

de cambios, configurar la ingestión en streaming desde buses de mensajes, configurar Auto Loader para detección automática de archivos y orquestar flujos de trabajo de ingestión con Lakeflow Spark Declarative Pipelines.

- Ingesta de datos con Lakeflow Connect.
- Ingesta de datos con cuadernos.
- Ingesta de datos con métodos SQL.
- Ingesta de datos con la fuente CDC.
- Ingesta de datos con Spark Structured Streaming.
- Ingesta de datos con Auto Loader.
- Ingesta de datos con Lakeflow Spark Declarative Pipelines.

Ejercicio: Ingesta de datos en el catálogo de Unity.

Módulo 10: Limpiar, transformar y cargar datos en el Catálogo de Unity.

La ingeniería de datos requiere transformar los datos en bruto en formatos limpios y bien estructurados, listos para su análisis. Este módulo explora técnicas para perfilar la calidad de los datos, seleccionar los tipos de columnas apropiados, resolver duplicados y valores nulos, aplicar transformaciones de filtrado y agregación, combinar conjuntos de datos con uniones y operadores de conjunto, remodelar datos mediante pivoteo y desnormalización, y cargar datos transformados mediante estrategias de añadido, sobreescritura y fusión.

- Datos de perfil.
- Elegir tipos de datos de columna.
- Resolver duplicados y nulos.
- Transformar datos con filtros y agregaciones.
- Transformar datos con uniones y operadores de conjunto.
- Datos de transformación con desnormalización y pivotes.
- Carga de datos con merge, insert y add.

Ejercicio: Limpiar, transformar y cargar datos en el catálogo de Unity.

Módulo 11: Implementar y gestionar restricciones de calidad de datos con Azure Databricks.

Este módulo explora estrategias para mantener una alta calidad de datos en Azure Databricks. Aprenderás a implementar comprobaciones de validación, hacer cumplir esquemas,





gestionar la deriva de esquemas y usar expectativas de pipeline para garantizar la integridad de los datos en toda la pipeline.

- Implementar comprobaciones de validación.
- Implementar comprobaciones de tipo de datos.
- Detectar y gestionar la deriva del esquema.
- Gestionar la calidad de los datos con las expectativas del pipeline.

Ejercicio: Implementar y gestionar restricciones de calidad de datos con Azure Databricks.

Ruta de aprendizaje: Despliega y mantiene canalizaciones y cargas de trabajo de datos con Azure Databricks.

Domina todo el ciclo de vida de construir, desplegar y mantener pipelines de datos listos para producción en Azure Databricks, desde el diseño y orquestación hasta la monitorización y optimización.

Al final de este camino de aprendizaje, podrás:

- Diseñar e implementar canalizaciones de datos robustas utilizando cuadernos y canalizaciones declarativas Lakeflow Spark
- Crear y orquestar Trabajos Lakeflow con disparadores, horarios y gestión de errores
- Aplicar control de versiones y desplegar pipelines entre entornos usando Git y Declarative Automation Bundles
- Monitorizar, solucionar problemas y optimizar las cargas de trabajo de datos para garantizar fiabilidad y rendimiento

Módulo 12: Diseñar e implementar pipelines de datos con Azure Databricks.

Aprende a diseñar e implementar pipelines de datos robustos en Azure Databricks usando notebooks y Lakeflow Spark Declarative Pipelines, cubriendo orquestación, manejo de errores y lógica de tareas.

- Orden de diseño de operaciones para una Pipelines.
- Elige cuaderno frente a Oleoductos de flujo de lago.
- Lógica de trabajo de flujo de lago de diseño.
- Gestión de errores de diseño en pipelines y trabajos.

- Crear pipeline con notebook.
- Crear tubería con Lakeflow Spark Declarative Pipelines.

Ejercicio: Diseñar e implementar canalizaciones de datos con Azure Databricks.

Módulo 13: Implementar trabajos de Lakeflow con Azure Databricks.

Este módulo te guía en el proceso de implementación de Lakeflow Jobs en Azure Databricks. Aprenderás a crear trabajos, configurar disparadores y horarios, configurar alertas y gestionar reinicios automáticos para garantizar una ejecución fiable de la canalización de datos.

- Crear configuración y configuración de trabajos.
- Configurar disparadores de trabajo.
- Programa un trabajo.
- Configurar alertas de empleo.
- Configurar reinicios automáticos.

Ejercicio: Implementar trabajos de Lakeflow con Azure Databricks.

Módulo 14: Implementar procesos del ciclo de vida del desarrollo en Azure Databricks.

Azure Databricks se integra con las prácticas de desarrollo establecidas a través de carpetas Git para control de versiones y Declarative Automation Bundles para despliegues de infraestructura como código. Este módulo explora las mejores prácticas de control de versiones en Git, flujos de trabajo de ramificación y pull requests, estrategias de pruebas integrales y el despliegue de paquetes basados en LI en entornos intervinientes.

- Aplicar las mejores prácticas de control de versiones de Git.
- Gestionar las ramificaciones y las pull requests.
- Implementar la estrategia de pruebas.
- Configurar y empaquetar paquetes de automatización declarativa.
- Bundle de despliegue con la CLI de Databricks.

Ejercicio: Implementar procesos del ciclo de vida del desarrollo en Azure Databricks.



Módulo 15: Monitorizar, solucionar problemas y optimizar cargas de trabajo en Azure Databricks.

La monitorización y la optimización son esenciales para ejecutar cargas de trabajo de datos fiables y rentables en Azure Databricks. Este módulo explora métricas de consumo de clústeres, resolución de problemas de Lakeflow Jobs, diagnósticos de trabajos de Spark, optimización del rendimiento para caché, problemas de sceo, spill y shuffle, y la transmisión de registros a Azure Log Analytics.

- Monitorizar y gestionar el consumo de clústeres.
- Resolución de problemas y reparación de trabajos en Lakeflow.
- Solucionar trabajos y cuadernos de Spark.
- Investiga el caché, el sexo, el desbordamiento, el barajador.
- Implementar la transmisión de registros con Azure Log Analytics.

Ejercicio: Monitorizar, solucionar problemas y optimizar cargas de trabajo en Azure Databricks.

