



DP-203T00

Data Engineering on Microsoft Azure



Información general.

En este curso, se aprenderá a implementar y administrar cargas de trabajo de ingeniería de datos en Microsoft Azure, mediante servicios como Azure Synapse Analytics, Azure Data Lake Storage Gen2, Azure Stream Analytics, Azure Databricks y otros.

El curso se centra en tareas comunes de ingeniería de datos, como la orquestación de canalizaciones de transferencia y transformación de datos, el trabajo con archivos de datos en un lago de datos, la creación y carga de almacenes de datos relacionales, la captura y la agregación de flujos de datos en tiempo real y el seguimiento de recursos y linaje de datos.

Duración.

4 Días.

Perfil del público.

La audiencia principal de este curso son profesionales de datos, arquitectos de datos y profesionales de business intelligence que quieren obtener información sobre la ingeniería de datos y la creación de soluciones analíticas mediante las tecnologías de plataforma de datos que hay en Microsoft Azure.

La audiencia secundaria de este curso incluye analistas y científicos de datos que trabajan con soluciones analíticas basadas en Microsoft Azure.

Examen.

DP-203: Data Engineering on Microsoft Azure.

Temario.

Módulo 1: Introducción a la ingeniería de datos en Azure.

Microsoft Azure proporciona una plataforma completa para la ingeniería de datos; pero ¿qué es la ingeniería de datos? Complete este módulo para averiguarlo.

- ¿En qué consiste la ingeniería de datos?
- Conceptos importantes de ingeniería de datos.
- Ingeniería de datos de Microsoft Azure.

En este módulo, aprenderá a realizar lo siguiente:

- Identificación de tareas comunes de ingeniería de datos.
- Descripción de los conceptos comunes de ingeniería de datos.
- Identificación de servicios de Azure para la ingeniería de datos.

Módulo 2: Introducción a Azure Data Lake Storage Gen2.

Los lagos de datos son un elemento clave de las arquitecturas de análisis de datos. Azure Data Lake Storage Gen2 proporciona una solución escalable, segura y basada en la nube para el almacenamiento en lagos de datos.

- Descripción de Azure Data Lake Storage Gen2.
- Habilitación de Azure Data Lake Storage Gen2 en Azure Storage.
- Comparación entre Azure Data Lake Storage y Azure Blob Storage.
- Fases del procesamiento de macrodatos.
- Uso de Azure Data Lake Storage Gen2 en cargas de trabajo de análisis de datos.



En este módulo, aprenderá a realizar lo siguiente:

- Describir las principales características y ventajas de Azure Data Lake Storage Gen2.
- Habilitar Azure Data Lake Storage Gen2 en una cuenta de Azure Storage.
- Comparar Azure Data Lake Storage Gen2 con Azure Blob Storage.
- Describir dónde encaja Azure Data Lake Storage Gen2 en las fases del procesamiento analítico.
- Describir cómo se usa Azure Data Lake Storage Gen2 en cargas de trabajo analíticas comunes.

Módulo 3: Introducción a Azure Synapse Analytics.

Obtenga información sobre las características y funcionalidades de Azure Synapse Analytics: una plataforma basada en la nube para el procesamiento y el análisis de macrodatos.

- Qué es Azure Synapse Analytics.
- Cómo funciona Azure Synapse Analytics.
- Cuándo usar Azure Synapse Analytics.

Ejercicio: exploración de Azure Synapse Analytics.**En este módulo aprenderá a:**

- Identificación de los problemas empresariales que Azure Synapse Analytics soluciona.
- Descripción de las funcionalidades principales de Azure Synapse Analytics.
- Determinación de cuándo usar Azure Synapse Analytics.

Módulo 4: Uso de un grupo de SQL sin servidor de Azure Synapse para consultar archivos en un lago de datos.

Con un grupo de SQL sin servidor de Azure Synapse, puede aprovechar sus aptitudes de SQL para explorar y analizar datos en archivos, sin la necesidad de cargar los datos en una base de datos relacional.

- Descripción de las capacidades y los casos de uso de los grupos de SQL sin servidor de Azure Synapse.
- Consulta de archivos mediante un grupo de SQL sin servidor.
- Creación de objetos de base de datos externos.

Ejercicio: Consulta de archivos mediante un grupo de SQL sin servidor.**Al finalizar este módulo podrá hacer lo siguiente:**

- Identificación de las capacidades y los y casos de uso de los grupos de SQL sin servidor en Azure Synapse Analytics.
- Consulta de archivos CSV, JSON y Parquet mediante un grupo de SQL sin servidor.
- Creación de objetos de base de datos externos en un grupo de SQL sin servidor.

Módulo 5: Uso de grupos de SQL sin servidor de Azure Synapse para transformar datos en un lago de datos.

Mediante el uso de un grupo de SQL sin servidor en Azure Synapse Analytics, puede usar el lenguaje SQL omnipresente para transformar los datos en archivos de un lago de datos.

- Transformación de archivos de datos con la instrucción CREATE EXTERNAL TABLE AS SELECT.
- Encapsulación de transformaciones de datos en un procedimiento almacenado.
- Inclusión de un procedimiento almacenado de transformación de datos en una canalización.

Ejercicio: Transformación de archivos mediante un grupo de SQL sin servidor.**Después de completar este módulo, podrá:**

- Uso de la instrucción CREATE EXTERNAL TABLE AS SELECT (CETAS) para transformar datos.
- Encapsulación de una instrucción CETAS en un procedimiento almacenado.
- Inclusión de un procedimiento almacenado de transformación de datos en una canalización.

Módulo 6: Creación de una base de datos de lago en Azure Synapse Analytics.

¿Por qué elegir entre trabajar con archivos en un lago de datos o un esquema de base de datos relacional? Con las bases de datos de lago de Azure Synapse Analytics, puede combinar las ventajas de ambos.





- Entender los conceptos de la base de datos de lago.
- Exploración de las plantillas de base de datos.
- Creación de una base de datos de lago.
- Usar una base de datos de lago.

Ejercicio: Análisis de los datos en una base de datos de lago.**Tras finalizar este módulo, podrá:**

- Entender los conceptos y componentes de una base de datos de lago.
- Describir las plantillas de base de datos en Azure Synapse Analytics.
- Creación de una base de datos de lago.

Módulo 7: Análisis de datos con Apache Spark en Azure Synapse Analytics.

Apache Spark es una de las tecnologías principales para el análisis de datos a gran escala. Aprenda a usar Spark en Azure Synapse Analytics para analizar y visualizar datos en un lago de datos.

- Introducción a Apache Spark.
- Uso de Spark en Azure Synapse Analytics.
- Análisis de datos con Spark.
- Visualización de datos con Spark.

Ejercicio: Análisis de datos con Spark.**Tras finalizar este módulo, podrá:**

- Identificar las características y las capacidades principales de Apache Spark.
- Configurar un grupo de Spark en Azure Synapse Analytics.
- Ejecutar código para cargar, analizar y visualizar datos en un cuaderno de Spark.

Módulo 8: Transformación de datos con Spark en Azure Synapse Analytics.

Normalmente, los ingenieros de datos necesitan transformar grandes volúmenes de datos. Los grupos de Apache Spark de Azure Synapse Analytics proporcionan una plataforma de procesamiento distribuido que pueden usar para lograr este objetivo.

- Modificación y guardado de marcos de datos.
- Creación de particiones de archivos de datos.
- Transformación de datos con SQL.
-
- Ejercicio: Transformación de datos con Spark en Azure Synapse Analytics.
-

- En este módulo, aprenderá a lo siguiente:
- Uso de Apache Spark para modificar y guardar marcos de datos.
- Cree particiones de archivos de datos para mejorar el rendimiento y la escalabilidad.
- Transformación de datos con SQL.

Módulo 9: Uso de Delta Lake en Azure Synapse Analytics.

Delta Lake es un área de almacenamiento relacional de código abierto para Spark que puede usar para implementar una arquitectura de lago de datos en Azure Synapse Analytics.

- Descripción de Delta Lake.
- Creación de tablas de Delta Lake y tablas de catálogo.
- Uso de Delta Lake con datos de streaming.
- Uso de Delta Lake en un grupo de SQL.

Ejercicio: Uso de Delta Lake en Azure Synapse Analytics.**En este módulo aprenderá a:**

- Describir las características y funcionalidades principales de Delta Lake.
- Crear y usar tablas de Delta Lake en un grupo de Spark de Synapse Analytics.
- Crear tablas de catálogo de Spark para datos de Delta Lake.
- Usar tablas de Delta Lake para transmitir datos.
- Consultar tablas de Delta Lake desde un grupo de SQL de Synapse Analytics.

Módulo 10: Análisis de datos en un almacenamiento de datos relacional.

Los almacenes de datos relacionales son un elemento principal de la mayoría de las soluciones empresariales de Business Intelligence (BI) y se usan como base para los modelos de datos, los informes y el análisis.

- Diseñar un esquema de almacenamiento de datos.
- Crear tablas de almacenamiento de datos.
- Carga de tablas de almacenamiento de datos.
- Consultar un almacenamiento de datos.

Ejercicio: Exploración de un almacenamiento de datos.

En este módulo aprenderá a:

- Diseñar un esquema para un almacenamiento de datos relacional.
- Crear tablas de hechos, dimensiones y almacenamiento provisional.
- Usar SQL para cargar datos en tablas de almacenamiento de datos.
- Usar SQL para consultar tablas de almacenamiento de datos relacionales.

Módulo 11: Carga de datos en un almacenamiento de datos relacional.

Una responsabilidad principal de un ingeniero de datos es implementar una solución de ingesta de datos que cargue nuevos datos en un almacenamiento de datos relacional.

- Cargar tablas de almacenamiento provisional.
- Cargar tablas de dimensiones.
- Cargar tablas de dimensiones de tiempo.
- Cargar dimensiones de variación lenta.
- Cargar tablas de hechos.
- Realizar la optimización posterior a la carga.

Ejercicio: Carga de datos en un almacenamiento de datos relacional.

En este módulo aprenderá a:

- Cargar tablas de almacenamiento provisional en un almacenamiento de datos.
- Cargar tablas de dimensiones en un almacenamiento de datos.
- Cargar dimensiones de tiempo en un almacenamiento de datos.
- Cargar dimensiones de variación lenta en un almacenamiento de datos.
- Cargar tablas de hechos en un almacenamiento de datos.
- Realizar optimizaciones posteriores a la carga en un almacenamiento de datos.

Módulo 12: Creación de una canalización de datos en Azure Synapse Analytics.

Las canalizaciones son la esencia de una solución de análisis de datos. Las canalizaciones son la esencia de una solución de análisis de datos.

- Descripción de las canalizaciones en Azure Synapse Analytics.
- Creación de una canalización en Azure Synapse Studio.
- Definición de flujos de datos.
- Ejecución de una canalización.

Ejercicio: Creación de una canalización de datos en Azure Synapse Analytics.

En este módulo, aprenderá a lo siguiente:

- Describir los conceptos básicos de las canalizaciones de Azure Synapse Analytics.
- Crear una canalización en Azure Synapse Studio.
- Implementar una actividad de flujo de datos en una canalización.
- Iniciar y supervisar ejecuciones de canalización.

Módulo 13: Uso de cuadernos de Spark en una canalización de Azure Synapse.

Apache Spark proporciona a los ingenieros de datos una plataforma de procesamiento de datos escalable y distribuida, que se puede integrar en una canalización de Azure Synapse Analytics.

- Descripción de los cuadernos y canalizaciones de Synapse.
- Uso de una actividad de cuaderno de Synapse en una canalización.
- Uso de parámetros en un cuaderno.

Ejercicio: Uso de un cuaderno de Apache Spark en una canalización.

En este módulo, aprenderá a lo siguiente:

- Describa la integración de los cuadernos y las canalizaciones.
- Use una actividad de cuaderno de Synapse en una canalización.
- Use parámetros con una actividad de cuaderno.





Módulo 14: Planeamiento del procesamiento analítico y transaccional híbrido mediante Azure Synapse Analytics.

Obtenga información sobre cómo el procesamiento analítico y transaccional híbrido (HTAP) puede ayudarle a realizar análisis operativos con Azure Synapse Analytics.

- Comprensión de los patrones de procesamiento analítico y transaccional híbrido.
- Descripción de Azure Synapse Link.

Después de completar este módulo, podrá:

- Describir los patrones del procesamiento analítico y transaccional híbrido.
- Identificar los servicios de Azure Synapse Link para HTAP.

Módulo 15: Implementación de Azure Synapse Link con Azure Cosmos DB.

Azure Synapse Link para Azure Cosmos DB permite la integración de HTAP entre los datos operativos de los entornos de ejecución de Azure Cosmos DB y Azure Synapse Analytics para Spark y SQL.

- Habilitación de una cuenta de Cosmos DB para usar Azure Synapse Link.
- Crear un contenedor habilitado para el almacén analítico.
- Creación de un servicio vinculado para Cosmos DB.
- Consulta de datos de Cosmos DB con Spark.
- Consulta de Cosmos DB con Synapse SQL.

Ejercicio: Implementación de Azure Synapse Link para Cosmos DB.

Después de completar este módulo, podrá:

- Configurar una cuenta de Azure Cosmos DB para usar Azure Synapse Link.
- Crear un contenedor habilitado para el almacén analítico.
- Crear un servicio vinculado para Azure Cosmos DB.
- Analizar datos vinculados mediante Spark y Synapse SQL.

Módulo 16: Implementación de Azure Synapse Link para SQL.

Azure Synapse Link para SQL permite la sincronización de baja latencia de datos operativos en una base de datos relacional con Azure Synapse Analytics.

- ¿Qué es Azure Synapse Link para SQL?
- Configuración de Azure Synapse Link para Azure SQL Database.
- Configuración de Azure Synapse Link para SQL Server 2022.

Ejercicio: Implementación de Azure Synapse Link para SQL.

En este módulo aprenderá a:

- Comprender los conceptos clave y las funcionalidades de Azure Synapse Link para SQL.
- Configurar Azure Synapse Link para Azure SQL Database.
- Configurar Azure Synapse Link para Microsoft SQL Server.

Módulo 17: Introducción a Azure Stream Analytics.

Azure Stream Analytics le permite procesar flujos de datos en tiempo real e integrar los datos que contienen en aplicaciones y soluciones analíticas.

- Definición de los flujos de datos.
- Descripción del procesamiento de eventos.
- Comprender las funciones de ventana.

Ejercicio: Introducción a Azure Stream Analytics.

En este módulo aprenderá a:

- Comprender los orígenes de datos.
- Comprender el procesamiento de eventos.
- Comprender las funciones de ventana.
- Introducción a Azure Stream Analytics.

Módulo 18: Ingesta de datos de streaming mediante Azure Stream Analytics y Azure Synapse Analytics.

Azure Stream Analytics proporciona un motor de procesamiento de datos en tiempo real que puede usar para ingerir datos de eventos de streaming en Azure Synapse Analytics para realizar análisis e informes adicionales.

- Escenarios de ingesta de flujos.
- Configuración de las entradas y salidas.
- Definición de una consulta para seleccionar, filtrar y agregar datos.
- Ejecución de un trabajo para ingerir datos.

Ejercicio: Ingesta de datos de streaming en Azure Synapse Analytics.

Después de completar este módulo, podrá:

- Describir escenarios comunes de ingesta de flujos para Azure Synapse Analytics.
- Configurar entradas y salidas para un trabajo de Azure Stream Analytics.
- Definir una consulta para ingerir datos en tiempo real en Azure Synapse Analytics.
- Ejecutar un trabajo para ingerir datos en tiempo real y consumirlos en Azure Synapse Analytics.

Módulo 19: Visualización de datos en tiempo real con Azure Stream Analytics y Power BI.

Al combinar las funcionalidades de procesamiento de flujos de Azure Stream Analytics y las funcionalidades de visualización de datos de Microsoft Power BI, puede crear paneles de datos en tiempo real.

- Uso de una salida de Power BI en Azure Stream Analytics.
- Creación de una consulta para la visualización en tiempo real.
- Creación de visualizaciones de datos en tiempo real con Power BI.

Ejercicio: Creación de una visualización de datos en tiempo real.

En este módulo aprenderá a:

- Configure una salida de Stream Analytics para Power BI.
- Use una consulta de Stream Analytics para escribir datos en Power BI.
- Cree una visualización de datos en tiempo real en Power BI.

Módulo 20: Introducción a Microsoft Purview.

En este módulo, evaluará si Azure Purview es la opción adecuada para sus necesidades de gobernanza y detección de datos.

- ¿Qué es Microsoft Purview?
- Funcionamiento de Microsoft Purview.
- Cuándo usar Microsoft Purview.

Al término de este módulo, podrá hacer lo siguiente:

- Evalúe si Microsoft Purview es adecuado para sus necesidades de gobernanza y detección de datos.
- Describir cómo funcionan las características de Azure Purview para proporcionar gobernanza y detección de datos.

Módulo 21: Integración de Microsoft Purview y Azure Synapse Analytics.

Aprenda a integrar Microsoft Purview con Azure Synapse Analytics para mejorar la detectabilidad de datos y el seguimiento de linaje.

- Catalogación de recursos de datos de Azure Synapse Analytics en Microsoft Purview.
- Conexión de Microsoft Purview con un área de trabajo de Azure Synapse Analytics.
- Búsquedas en un catálogo de Purview en Synapse Studio.
- Seguimiento del linaje de datos en las canalizaciones.

Ejercicio: Integrar Azure Synapse Analytics y Microsoft Purview.

Después de completar este módulo, podrá:

- Catalogar los recursos de base de datos de Azure Synapse Analytics en Microsoft Purview.
- Configurar la integración de Microsoft Purview en Azure Synapse Analytics.
- Hacer búsquedas en el catálogo de Microsoft Purview desde Synapse Studio.
- Hacer un seguimiento del linaje de datos en las actividades de canalizaciones de Azure Synapse Analytics.

Módulo 22: Explorar Azure Databricks.

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

- Introducción a Azure Databricks.
- Identificación de las cargas de trabajo de Azure Databricks.
- Descripción de los conceptos clave.

Ejercicio: Explorar Azure Databricks.



En este módulo aprenderá a:

- Aprovechone un área de trabajo de Azure Databricks.
- Identifique las cargas de trabajo principales y los roles de Azure Databricks.
- Describir los conceptos clave de una solución de Azure Databricks.

Módulo 23: Uso de Apache Spark en Azure Databricks.

Azure Databricks se basa en Apache Spark y permite a los ingenieros y analistas de datos ejecutar trabajos de Spark para transformar, analizar y visualizar datos a escala.

- Descubra Spark.
- Creación de un clúster de Spark.
- Uso de Spark en cuadernos.
- Uso de Spark para trabajar con archivos de datos.
- Visualización de datos.

Ejercicio: Uso de Spark en Azure Databricks.**En este módulo aprenderá a:**

- Describir los elementos clave de la arquitectura de Apache Spark.
- Cree y configure un clúster de Spark.
- Describir los casos de uso de Spark.
- Use Spark para procesar y analizar los datos almacenados en archivos.
- Use Spark para visualizar los datos.

Módulo 24: Ejecución de cuadernos de Azure Databricks con Azure Data Factory.

El uso de canalizaciones en Azure Data Factory para ejecutar cuadernos en Azure Databricks permite automatizar procesos de ingeniería de datos a escala de la nube.

- Descripción de los cuadernos y las canalizaciones de Azure Databricks.
- Creación de un servicio vinculado para Azure Databricks.
- Uso de una actividad de cuaderno en una canalización.
- Uso de parámetros en un cuaderno.

Ejercicio: Ejecución de un cuaderno de Azure Databricks con Azure Data Factory.**En este módulo aprenderá a:**

- Describa cómo se pueden ejecutar cuadernos de Azure Databricks en una canalización.
- Cree un servicio vinculado de Azure Data Factory para Azure Databricks.
- Use una actividad de cuaderno en una canalización.
- Pase parámetros a un cuaderno.

