DP-203: Data Engineering on Microsoft Azure

Duración: 4 días (32 hrs)

Descripción general

Los ingenieros de datos Azure diseñan e implementan la gestión, la monitorización, la seguridad y privacidad de los datos mediante la pila completa de los servicios Azure para satisfacer las necesidades comerciales.

En este curso, el estudiante aprenderá sobre los patrones y prácticas de la ingeniería de datos en lo que respecta al trabajo con soluciones analíticas por lotes y en tiempo real utilizando tecnologías de plataforma de datos de Azure.

Objetivos

- Explorar las opciones de procesamiento y almacenamiento para cargas de trabajo de ingeniería de datos en Azure
- Diseñar e implementar la capa de servicio
- Comprender las consideraciones de ingeniería de datos
- Ejecutar consultas interactivas utilizando grupos SQL sin servidor
- Explorar, transformar y cargar datos en el almacén de datos con Apache Spark
- Realizar exploración y transformación de datos en Azure Databricks
- Ingesta y carga datos en el almacén de datos
- Transformar datos con Azure Data Factory o Azure Synapse Pipelines
- Integrar datos de portátiles con Azure Data Factory o Azure Synapse Pipelines
- Optimizar el rendimiento de las consultas con grupos de SQL dedicados en Azure Synapse
- Analizar y optimizar el almacenamiento del almacén de datos
- Admitir el procesamiento analítico transaccional híbrido (HTAP) con Azure Synapse Link
- Realizar la seguridad de un extremo a otro con Azure Synapse Analytics
- Realizar el procesamiento de transmisiones en tiempo real con Stream Analytics
- Crear una solución de procesamiento de transmisión con Event Hubs y Azure Databricks
- Crear informes mediante la integración de Power BI con Azure Synpase Analytics
- Realizar procesos integrados de aprendizaje automático en Azure Synapse Analytics

Prerrequisitos del curso

- AZ-900 Microsoft Azure Fundamentals.
- DP-900: Microsoft Azure Data Fundamentals.

Dirigido a:

Este curso está dirigido a Data Professionals, Data Architects y profesionales de Business Intelligence que desean aprender sobre las tecnologías de plataforma de datos que existen en Microsoft Azure e implementar soluciones de datos de Azure. También está enfocado, aunque en segunda instancia, a personas que desarrollan aplicaciones que entregan contenido de las tecnologías de la plataforma de datos que existen en Microsoft Azure.

Esquema del curso

Módulo 1: Explorar las opciones de computación y almacenamiento para cargas de trabajo de ingeniería de datos

- Introducción a Azure Synapse Analytics
- Describir Azure Databricks
- Introducción al almacenamiento de Azure Data Lake
- Describir la arquitectura de Delta Lake
- Trabajar con flujos de datos mediante Azure Stream Analytics
- Laboratorio: Explorar las opciones de computación y almacenamiento para cargas de trabajo de ingeniería de datos
  - Combinar la transmisión y el procesamiento por lotes con una única pipeline
  - Organizar el data lake en niveles de transformación de archivos
  - Indexación de almacenamiento de data lake para la aceleración de consultas y cargas de trabajo

Módulo 2: Diseño e implementación de la capa de servicio

- Diseñar un esquema multidimensional para optimizar las cargas de trabajo analíticas
- Transformación sin código a escala con Azure Data Factory
- Rellenar dimensiones que cambian lentamente en las pipelines de Azure Synapse Analytics
- Laboratorio: Diseño e implementación de la capa de servicio
  - Diseñar un esquema en estrella para cargas de trabajo analíticas

- Rellenar dimensiones que cambian lentamente con Azure Data Factory y mapeo de flujos de datos

Módulo 3: Consideraciones de ingeniería de datos para archivos fuente

- Diseñar un almacén de datos moderno con Azure Synapse Analytics
- Proteger un almacén de datos en Azure Synapse Analytics
- Laboratorio: Consideraciones de ingeniería de datos
- Administrar archivos en un data lake de Azure
- Protección de archivos almacenados en un data lake de Azure

Módulo 4: Ejecutar consultas interactivas con grupos de SQL sin servidor de Azure Synapse Analytics

- Explorar las capacidades de los grupos SQL sin servidor de Azure Synapse
- Consultar datos en el lake mediante grupos de SQL sin servidor de Azure Synapse
- Crear objetos de metadatos en grupos SQL sin servidor de Azure Synapse
- Proteger los datos y administrar a los usuarios en los grupos SQL sin servidor de Azure Synapse
- Laboratorio: Ejecutar consultas interactivas utilizando grupos de SQL sin servidor
- Consultar datos de Parquet con grupos SQL sin servidor
- Crear tablas externas para archivos Parquet y CSV
- Crear vistas con grupos de SQL sin servidor
- Acceso seguro a los datos en un data lake cuando se utilizan grupos de SQL sin servidor
- Configurar la seguridad del data lake mediante el control de acceso basado en roles (RBAC) y la lista de control de acceso

Módulo 5: Explorar, transformar y cargar datos en el almacén de datos usando Apache Spark

- Comprender la ingeniería de big data con Apache Spark en Azure Synapse Analytics
- Ingestar datos con los cuadernos de Apache Spark en Azure Synapse Analytics
- Transformar datos con DataFrames en Apache Spark Pools en Azure Synapse Analytics
- Integrar grupos de SQL y Apache Spark en Azure Synapse Analytics
- Laboratorio: Explorar, transformar y cargar datos en el almacén de datos usando Apache Spark
- Realizar exploración de datos en Synapse Studio
- Ingestar datos con cuadernos Spark en Azure Synapse Analytics
- Transformar datos con DataFrames en grupos de Spark en Azure Synapse Analytics
- Integrar grupos de SQL y Spark en Azure Synapse Analytics

Módulo 6: Exploración y transformación de datos en Azure Databricks

- Describir Azure Databricks
- Leer y escribir datos en Azure Databricks
- Trabajar con DataFrames en Azure Databricks

- Trabajar con métodos avanzados de DataFrames en Azure Databricks
- Laboratorio: Exploración y transformación de datos en Azure Databricks
  - Usar DataFrames en Azure Databricks para explorar y filtrar datos
  - Almacenar en caché un DataFrame para consultas posteriores más rápidas
  - Eliminar datos duplicados
  - Manipular valores de fecha / hora
  - Eliminar y cambiar el nombre de las columnas DataFrame
  - Agregar datos almacenados en un DataFrame

Módulo 7: Ingesta y carga datos en el almacén de datos.

- Utilizar las mejores prácticas de carga de datos en Azure Synapse Analytics
- Ingestión a escala de petabytes con Azure Data Factory
- Laboratorio: Ingesta y carga datos en el almacén de datos
  - Realizar la ingestión a escala de petabytes con Azure Synapse Pipelines
  - Importar datos con PolyBase y COPY usando T-SQL
  - Utilizar las mejores prácticas de carga de datos en Azure Synapse Analytics

Módulo 8: Transformar datos con Azure Data Factory o Azure Synapse Pipelines

- Integración de datos con Azure Data Factory o Azure Synapse Pipelines
- Transformación sin código a escala con Azure Data Factory o Azure Synapse Pipelines
- Laboratorio: Transformar datos con Azure Data Factory o Azure Synapse Pipelines
  - Ejecutar transformaciones sin código a escala con Azure Synapse Pipelines
  - Crear un pipeline de datos para importar archivos CSV con formato deficiente
  - Crear flujos de datos de mapeo

Módulo 9: Orquestar el movimiento y la transformación de datos en Azure Synapse Pipelines

- Organizar el movimiento y la transformación de datos en Azure Data Factory
- Laboratorio: Orquestar el movimiento y la transformación de datos en Azure Synapse Pipelines
  - Integrar datos de portátiles con Azure Data Factory o Azure Synapse Pipelines

Módulo 10: Optimizar el rendimiento de las consultas con grupos de SQL dedicados en Azure Synapse

- Optimizar el rendimiento de las consultas del almacén de datos en Azure Synapse Analytics
- Comprender las características para desarrolladores de almacenamiento de datos de Azure Synapse Analytics
- Laboratorio: Optimizar el rendimiento de las consultas con grupos de SQL dedicados en Azure Synapse
  - Comprender las características para desarrolladores de Azure Synapse Analytics

- Optimizar el rendimiento de las consultas del almacén de datos en Azure Synapse Analytics
- Mejorar el rendimiento de las consultas

Módulo 11: Analizar y optimizar el almacenamiento del data wharehouse

- Analizar y optimizar el almacenamiento del data wharehouse de datos en Azure Synapse Analytics
- Laboratorio: Analizar y optimizar el almacenamiento del data wharehouse
- Comprobar si hay datos sesgados y uso de espacio
- Comprender los detalles de almacenamiento de la tienda de columnas
- Estudiar el impacto de las vistas materializadas
- Explorar las reglas para operaciones mínimamente registradas

Módulo 12: Soporte del procesamiento analítico transaccional híbrido (HTAP) con Azure Synapse Link

- Diseñar procesamiento transaccional y analítico híbrido con Azure Synapse Analytics
- Configurar Azure Synapse Link con Azure Cosmos DB
- Consultar Azure Cosmos DB con grupos de Apache Spark
- Consultar Azure Cosmos DB con grupos de SQL sin servidor
- Laboratorio: Soporte de procesamiento analítico transaccional híbrido (HTAP) con Azure Synapse Link
- Configurar Azure Synapse Link con Azure Cosmos DB
- Consultar Azure Cosmos DB con Apache Spark para Synapse Analytics
- Consultar Azure Cosmos DB con un grupo de SQL sin servidor para Azure Synapse Analytics

Módulo 13: Seguridad de un extremo a otro con Azure Synapse Analytics

- Proteger un almacén de datos en Azure Synapse Analytics
- Configurar y administrar secretos en Azure Key Vault
- Implementar controles de cumplimiento para datos confidenciales
- Laboratorio: Seguridad de un extremo a otro con Azure Synapse Analytics
- Asegurar la infraestructura de soporte de Azure Synapse Analytics
- Asegurar el área de trabajo de Azure Synapse Analytics y los servicios administrados
- Proteger los datos del área de trabajo de Azure Synapse Analytics

Módulo 14: Procesamiento de transmisión en tiempo real con Stream Analytics

- Habilitar la mensajería confiable para aplicaciones de Big Data con Azure Event Hubs
- Trabajar con flujos de datos mediante Azure Stream Analytics
- Ingesta flujos de datos con Azure Stream Analytics
- Laboratorio: Procesamiento de transmisión en tiempo real con análisis de transmisión
- Utilizar Stream Analytics para procesar datos en tiempo real de Event Hubs

- Utilizar las funciones de ventana de Stream Analytics para crear agregados y resultados en Synapse Analytics
- Escalar el trabajo de Azure Stream Analytics para aumentar el rendimiento mediante la partición
- Repartir la entrada de flujo para optimizar la paralelización

Módulo 15: Crear una solución de procesamiento de transmisión con Event Hubs y Azure Databricks

- Procesar datos de streaming con transmisión estructurada de Azure Databricks
- Laboratorio: Crear una solución de procesamiento de transmisión con Event Hubs y Azure Databricks
- Explorar las características y usos clave de la transmisión estructurada
- Transmitir datos desde un archivo y escríbirlos en un sistema de archivos distribuido
- Utilizar ventanas deslizantes para agregar fragmentos de datos en lugar de todos los datos
- Aplicar marcas de agua para eliminar datos obsoletos
- Conectarse a transmisiones de lectura y escritura de Event Hubs

Módulo 16: Generar informes mediante la integración de Power BI con Azure Synpase Analytics

- Crear informes con Power BI utilizando su integración con Azure Synapse Analytics
- Laboratorio: Generar informes utilizando la integración de Power BI con Azure Synpase Analytics
- Integrar un área de trabajo de Azure Synapse y Power BI
- Optimizar la integración con Power BI
- Mejorar el rendimiento de las consultas con vistas materializadas y almacenamiento en caché de conjuntos de resultados
- Visualizar datos con SQL sin servidor y crear un informe de Power BI

Módulo 17: Realizar procesos integrados de aprendizaje automático en Azure Synapse Analytics

- Utilizar el proceso de aprendizaje automático integrado en Azure Synapse Analytics
- Laboratorio: Realizar procesos integrados de aprendizaje automático en Azure Synapse Analytics
- Crear un servicio vinculado de Azure Machine Learning
- Activar un experimento de Auto ML con datos de una tabla Spark
- Enriquecer los datos utilizando modelos entrenados
- Ofrecer resultados de predicción con Power BI