Data Engineering on Microsoft Azure
Curso DP-203

Sobre este curso

En este curso, el estudiante aprenderá acerca de la ingeniería de datos en lo que se refiere a trabajar con soluciones analíticas por lotes y en tiempo real utilizando las tecnologías de la plataforma de datos de Azure. Los estudiantes comenzarán entendiendo las tecnologías básicas de computación y almacenamiento que se utilizan para construir una solución analítica. Los estudiantes aprenderán a explorar interactivamente los datos almacenados en archivos en un lago de datos. Aprenderán las distintas técnicas de ingesta que pueden utilizarse para cargar datos utilizando la capacidad de Apache Spark que se encuentra en Azure Synapse Analytics o Azure Databricks, o cómo ingerir utilizando las canalizaciones de Azure Data Factory o Azure Synapse. Los estudiantes también aprenderán las diversas formas en que pueden transformar los datos utilizando las mismas tecnologías que se utilizan para la ingesta de datos. Entenderán la importancia de implementar la seguridad para garantizar la protección de datos en reposo o en tránsito. A continuación, el estudiante mostrará cómo crear un sistema analítico en tiempo real para crear soluciones analíticas en tiempo real.

Duración

28 hs

Perfil de los alumnos

El público principal de este curso son profesionales de datos, arquitectos de datos y profesionales de la inteligencia empresarial que quieren aprender sobre ingeniería de datos y construcción de soluciones analíticas utilizando tecnologías de plataformas de datos que existen en Microsoft Azure. El asistente secundario de este curso son analistas y científicos de datos que trabajan con soluciones analíticas construidas en Microsoft Azure.

Requisitos Previos

Los alumnos que aprueben este curso tendrán conocimientos sobre computación en cloud y conceptos básicos de datos, así como experiencia profesional con soluciones de datos.
Específicamente completar:

- AZ-900 - Azure Fundamentals

Data Engineering on Microsoft Azure
Curso DP-203

- DP-900 - Microsoft Azure Data Fundamentals

Esquema del curso

Módulo 1: Explorar las opciones de computación y almacenamiento para las cargas de trabajo de ingeniería de datos
Este módulo proporciona un resumen de las opciones tecnológicas de computación y almacenamiento de Azure que están disponibles para ingenieros de datos que construyen cargas de trabajo analíticas. Este módulo enseña maneras de estructurar el lago de datos y de optimizar los archivos para las cargas de trabajo de exploración, streaming y batch. El alumno aprenderá a organizar el lago de datos en niveles de refinamiento de datos mientras transforma los archivos a través del procesamiento por lotes y de flujo. Luego aprenderá a crear índices en sus conjuntos de datos, tales como archivos CSV, JSON y Parquet, y a utilizarlos para una posible aceleración de las consultas y las cargas de trabajo.
Lecciones
- 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 utilizando Azure Stream Analytics

Laboratorio : Explora opciones de computación y almacenamiento para cargas de trabajo de ingeniería de datos
- Combinar el procesamiento en flujo y por lotes con una única canalización
- Organizar el lago de datos en niveles de transformación de archivos
- Indexar el almacenamiento del lago de datos para acelerar las consultas y las cargas de trabajo

Módulo 2: Ejecutar consultas interactivas utilizando los pools SQL sin servidor de Azure Synapse Analytics
En este módulo, los alumnos aprenderán a trabajar con archivos almacenados en el lago de datos y con fuentes de archivos externas, mediante declaraciones T-SQL ejecutadas por un pool SQL sin servidor en Azure Synapse Analytics. Los alumnos consultarán archivos Parquet almacenados en un lago de datos, así como archivos CSV almacenados en un almacén de datos externo. A continuación, crearán grupos de seguridad de Azure Active Directory e impondrán el acceso a los archivos del lago

de datos mediante el Control de Acceso Basado en Funciones (RBAC) y las Listas de Control de Acceso (ACL).

Lecciones

- Explorar capacidades de los pools SQL sin servidor de Azure Synapse
- Consultar de datos en el lago utilizando pools SQL sin servidor de Azure Synapse
- Crear objetos de metadatos en pools SQL sin servidor de Azure Synapse
- Proteger datos y gestionar usuarios en pools SQL sin servidor de Azure Synapse

Laboratorio : Realiza consultas interactivas utilizando pools SQL sin servidor

- Consultar datos de Parquet con pools SQL sin servidor
- Crear tablas externas para archivos Parquet y CSV
- Crear vistas con pools SQL sin servidor
- Asegurar el acceso a los datos de un lago de datos utilizando pools SQL sin servidor
- Configurar la seguridad del lago de datos utilizando el Control de Acceso Basado en Cargos (RBAC) y la Lista de Control de Acceso

Módulo 3: Exploración y transformación de datos en Azure Databricks

Este módulo enseña a utilizar varios métodos de Apache Spark DataFrame para explorar y transformar datos en Azure Databricks. El estudiante aprenderá a realizar métodos estándar de DataFrame para explorar y transformar datos. También aprenderá a realizar tareas más avanzadas, tales como eliminar datos duplicados, manipular valores de fecha/hora, renombrar columnas y agregar datos.

Lecciones

- 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

- Utiliza DataFrames en Azure Databricks para explorar y filtrar datos
- Almacena un DataFrame para agilizar las consultas posteriores
- Eliminar datos duplicados
- Manipular valores de fecha/hora
- Eliminar y renombrar columnas de DataFrame
- Agregar datos almacenados en un DataFrame

Data Engineering on Microsoft Azure
Curso DP-203

Módulo 4: Explorar, transformar y cargar datos en el Almacén de Datos utilizando Apache Spark

Este módulo enseña cómo explorar los datos almacenados en un lago de datos, transformar los datos y cargarlos en un almacén de datos relacional. El estudiante explorará archivos Parquet y JSON y utilizará técnicas para consultar y transformar archivos JSON con estructuras jerárquicas. A continuación, el estudiante utilizará Apache Spark para cargar datos en el almacén de datos y unir los datos Parquet del lago de datos con los datos del grupo SQL dedicado.

Lecciones

- Entender la ingeniería de big data con Apache Spark en Azure Synapse Analytics
- Ingesta de datos con cuadernos de Apache Spark en Azure Synapse Analytics
- Transformar datos con DataFrames en grupos de Apache Spark 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 utilizando Apache Spark

- Realiza la Exploración de Datos en Synapse Studio
- Ingesta de 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 5: Ingesta y carga de datos en el almacén de datos

Este módulo enseña a los estudiantes cómo ingerir datos en el almacén de datos mediante escrituras T-SQL y canalizaciones de integración de Synapse Analytics. El estudiante aprenderá a cargar datos en grupos SQL dedicados de Synapse con PolyBase y COPY utilizando T-SQL. El estudiante también aprenderá a utilizar la gestión de la carga de trabajo junto con una actividad de copia en una canalización de Azure Synapse para la ingestión de datos a escala de petabytes.

Lecciones

- Utilizar las mejores prácticas de carga de datos en Azure Synapse Analytics
- Ingesta a escala de petabytes con Azure Data Factory

Laboratorio : Ingesta y carga de datos en el Almacén de Datos

- Realiza la ingestión a escala de petabytes con Azure Synapse Pipelines
- Importar datos con PolyBase y COPY utilizando T-SQL

- Utilizar las mejores prácticas de carga de datos en Azure Synapse Analytics

Módulo 6: Transformar datos con Azure Data Factory o Azure Synapse Pipelines
Este módulo enseña a los estudiantes a construir canalizaciones de integración de datos para ingerir desde múltiples fuentes de datos, transformar los datos utilizando flujos de datos de mapeo y realizar el movimiento de datos a uno o más receptores de datos.
Lecciones

- 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 : Transformación de datos con Azure Data Factory o Azure Synapse Pipelines

- Ejecutar transformaciones sin código a escala con Azure Synapse Pipelines
- Crear una canalización de datos para importar archivos CSV mal formateados
- Crear Flujos de Datos de Mapeo

Módulo 7: Orquestar el movimiento y la transformación de datos en Azure Synapse Pipelines
En este módulo, aprenderás a crear servicios vinculados y a orquestar el movimiento y la transformación de datos utilizando cuadernos en Azure Synapse Pipelines.
Lecciones

- Orquestar 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 desde Cuadernos con Azure Data Factory o Azure Synapse Pipelines

Después de completar este módulo, los estudiantes serán capaces de:

- Orquestar el movimiento y la transformación de datos en Azure Synapse Pipelines

Módulo 8: Seguridad integral con Azure Synapse Analytics
En este módulo, los estudiantes aprenderán a asegurar un espacio de trabajo de Synapse Analytics y su infraestructura de apoyo. El estudiante observará la SQL Active Directory Admin, gestional las reglas IP firewalls, gestionar los secretos con

Azure Key Vault y acceder a esos secretos a través de un servicio vinculado a Key Vault y las actividades de canalización. El estudiante entenderá cómo implementar la seguridad a nivel de columna, la seguridad a nivel de fila y el enmascaramiento dinámico de datos al utilizar grupos SQL dedicados.

Lecciones

- Asegurar un almacén de datos en Azure Synapse Analytics
- Configurar y gestionar los secretos en Azure Key Vault
- Implementar controles de cumplimiento para datos sensibles

Laboratorio : Seguridad integral con Azure Synapse Analytics

- Asegura la infraestructura compatible con Azure Synapse Analytics
- Asegura el espacio de trabajo de Azure Synapse Analytics y los servicios gestionados
- Asegurar los datos del espacio de trabajo de Azure Synapse Analytics

Módulo 9: Soportar Hybrid Transactional Analytical Processing (HTAP) con Azure Synapse Link

En este módulo, los estudiantes aprenderán cómo Azure Synapse Link permite la conectividad perfecta de una cuenta de Azure Cosmos DB con un espacio de trabajo de Synapse. El estudiante entenderá cómo habilitar y configurar Synapse link, y luego cómo consultar el almacén analítico de Azure Cosmos DB utilizando Apache Spark y SQL serverless**.**

Lecciones

- Diseñar un procesamiento transaccional y analítico híbrido utilizando Azure Synapse Analytics
- Configurar Azure Synapse Link con Azure Cosmos DB
- Consulta de Azure Cosmos DB con grupos de Apache Spark
- Consulta de Azure Cosmos DB con grupos de SQL sin servidor

Laboratorio : Soportar Hybrid Transactional Analytical Processing (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

Data Engineering on Microsoft Azure
Curso DP-203

Módulo 10: Procesamiento de flujos en tiempo real con Stream Analytics
En este módulo, los estudiantes aprenderán a procesar datos de streaming con Azure Stream Analytics. El estudiante ingestará datos de telemetría de vehículos en Event Hubs, y luego procesará esos datos en tiempo real, utilizando varias funciones de ventana en Azure Stream Analytics. Enviarán los datos a Azure Synapse Analytics. Por último, el estudiante aprenderá a escalar el trabajo de Stream Analytics para aumentar el rendimiento.

Lecciones

- Habilitar la mensajería fiable para las aplicaciones de Big Data utilizando Azure Event Hubs
- Trabajar con flujos de datos utilizando Azure Stream Analytics
- Ingerir flujos de datos con Azure Stream Analytics

Laboratorio : Procesamiento de flujos en tiempo real con Stream Analytics

- Utiliza Stream Analytics para procesar datos en tiempo real desde Event Hubs
- Utiliza las funciones de ventana de Stream Analytics para crear agregados y enviarlos a Synapse Analytics
- Escala el trabajo de Azure Stream Analytics para aumentar el rendimiento mediante la partición
- Repartición de la aportación del flujo para optimizar la paralelización

Módulo 11: Crear una Solución de Procesamiento de Flujo con Event Hubs y Azure Databricks
En este módulo, los estudiantes aprenderán a ingerir y procesar datos de streaming a escala con Event Hubs y Spark Structured Streaming en Azure Databricks. El estudiante aprenderá las principales funciones y usos de Structured Streaming. El estudiante implementará ventanas deslizantes para agregar sobre trozos de datos y aplicar marcas de agua para eliminar los datos obsoletos. Por último, el estudiante se conectará a Event Hubs para leer y escribir flujos.

Lecciones

- Procesar datos en streaming con Azure Databricks structured streaming

Laboratorio : Crear una solución de procesamiento de streaming con Event Hubs y Azure Databricks

- Explora las principales funciones y usos del Structured Streaming
- Transmitir datos desde un archivo y escribirlos en un sistema de archivos distribuido

Data Engineering on Microsoft Azure
Curso DP-203

- Utilizar ventanas deslizantes para agregar trozos de datos en lugar de todos los datos
- Aplicar marcas de agua para eliminar los datos obsoletos
- Conectar con flujos de lectura y escritura de Event Hubs