Challenge Data engineer

Azure

Contexto

El siguiente challenge está pensado para que el desarrollador utilice las herramientas que crea necesarias para llevarlo a cabo. Para el desarrollo sobre las herramientas de azure, se debe utilizar la evaluación gratuita de 30 días que micrsoft ofrece con cualquier cuenta de email.

Tareas

Parte 1: Extracción y Transformación de Datos con PySpark

Extracción de Datos desde una API:

- Utilizar Spark para extraer datos de una API pública de tu elección
- Aplica transformaciones necesarias para preparar los datos para su posterior análisis (ej: procesar archivos json a tablas).
- Consolidar un archivo único con el resultado de la extracción.

Lectura de Datos desde una Base de Datos:

- Conéctate a una base de datos SQL en Azure utilizando PySpark.
- Realiza una consulta para extraer un conjunto de datos específico.
- Realiza transformaciones en los datos según sea necesario.

Parte 2: Integración de Servicios en Azure

Escribe los datos procesados en la Parte 1.1 en un almacenamiento de datos persistente en Azure (por ejemplo, Azure Data Lake Storage, Azure Blob Storage, etc.).

Escribe los datos de la parte 1.2 en un Azure Blob Storage en formato parquet y con las particiones que consideres adecuadas.

Documentación y Entrega

Documentación:

Documenta el proceso completo, incluyendo código, configuraciones y pasos utilizados.

Proporciona una breve explicación de las decisiones de diseño y enfoque técnico.

Entrega:

Por favor, envía el código y la documentación en un repositorio Git (por ejemplo, GitHub) y comparte el enlace con nosotros.

Notas Adicionales:

Siéntete libre de utilizar cualquier lenguaje o librería que consideres apropiada para las tareas.

No dudes en utilizar los servicios de Azure que consideres necesarios para completar las tareas.