Práctica de Ingesta Directa en un Data Lake

Enunciado de la práctica: Ingesta Directa de Actas de Notas La universidad está implementando un sistema de almacenamiento centralizado de datos mediante un Data Lake, con el objetivo de mejorar la gestión de la información académica. Como parte de este proyecto, es necesario diseñar un proceso de ingesta directa de archivos de actas de notas, que contienen los resultados de los estudiantes en distintas materias y programas académicos.

Implementar también un proceso de ingesta por lotes para la información de informes de proyectos de investigación

Objetivo:

Desarrollar un proceso de ingesta directa de archivos de actas de notas de al menos dos materias correspondientes a dos programas académicos distintos dentro de la universidad, asegurando la integración y almacenamiento eficiente en el Data Lake.

Instrucciones:

1. Tipos de archivos:

Recibirás archivos en formato CSV o Excel que contienen los datos de las actas de notas. Cada archivo debe contener, al menos, los siguientes campos:
- Nombre del estudiante
- Código del estudiante
- Materia
- Nota
- Periodo académico
- Programa académico

Para la información de los informes de investigación se recibirán archivos pdf o Word en el cual en el nombre se indicara el nombre del proyecto y nombre del grupo de investigación.

2. Materias y Programas académicos y informes de investigación:

Debes procesar datos de al menos dos materias que pertenezcan a dos programas académicos distintos. Los programas pueden ser, por ejemplo, Ingeniería de Sistemas y Administración de Empresas, y las materias podrían ser 'Programación I' y 'Contabilidad

Financiera'.
Debe procesar al menos 10 informes de investigación de diferentes proyectos.

3. Ingesta de datos:

Implementa un proceso de ingesta directa que cargue los archivos de las actas de notas en el Data Lake. Puedes utilizar herramientas como Azure Data Factory, o AWS Glue, según el entorno que se utilice.

Implementar un proceso por lotes que cargue los datos a sus respectivos repositorios, el tamaño del lote debe ser 5 archivos o 100 MB.

4. Estrategia de almacenamiento:

Los datos deben almacenarse en el Data Lake de manera que se mantenga la flexibilidad para consultas posteriores. Ten en cuenta la estructura lógica para los directorios y archivos en el Data Lake (por ejemplo, particionamiento por programa o periodo académicos) que diseño en el trabajo de almacenamiento, para esta práctica, en caso de que lo desee, puede simplificarlo a un tipo de particionamiento.

5. Validación y manejo de errores:

El proceso debe incluir una etapa de validación de datos. Si se encuentran errores en el formato o datos faltantes, estos deben ser registrados en un archivo de log para revisión.

6. Metadatos:

Debe en el proceso de ingesta cargar los metadatos de los archivos a los que de lugar, entre ellos al menos la fecha, facultad, periodo académico y demás que considero en la primera entrega

7. Entregables:

- Guía de la construcción del proceso de ingesta usando Código o pipeline para la ingesta directa.
- Logs de validación y cualquier error capturado.
- Evidencia de los datos correctamente almacenados en el Data Lake.