**Orquestación automatizada para consolidación masiva de registros clínicos**

**1. Necesidad principal del flujo**

La red de clínicas necesita consolidar diariamente grandes volúmenes de datos médicos provenientes de múltiples sedes y sistemas. Los tipos de datos incluyen:

* Información de pacientes (demográficos, historial clínico)
* Estudios realizados (laboratorio, imágenes)
* Recetas médicas
* Turnos y citas

**Objetivo de la automatización**:  
Reducir errores, duplicaciones y demoras mediante un flujo automatizado que permita la recolección, transformación y carga eficiente de los datos en una plataforma centralizada.

**2. Plataforma seleccionada: Apache Airflow**

**Justificación**:

* **Tipo de procesamiento**: Batch diario, ideal para Airflow por su capacidad de orquestar tareas programadas y dependientes.
* **Perfil técnico del equipo**: Airflow permite definir flujos en Python, lo que se adapta a equipos con conocimientos técnicos intermedios.
* **Complejidad del flujo**: Airflow facilita la gestión de dependencias, reintentos, alertas y monitoreo, ideal para flujos clínicos con múltiples etapas.

**3. Flujo paso a paso**

**Ingesta**

* Conectores FTP y APIs REST para extraer datos desde los sistemas de cada sede.
* Airflow ejecuta tareas programadas para recolectar los datos cada noche.
* Los datos se almacenan temporalmente en un bucket S3 (o similar) para procesamiento.

**Procesamiento**

* Validación de estructura y campos obligatorios (ID de paciente, fecha, tipo de estudio).
* Normalización de formatos (fechas, nombres de medicamentos).
* Detección y eliminación de duplicados.
* Enriquecimiento con datos del sistema maestro de pacientes.

**Entrega**

* Carga final en un **Data Lake** para almacenamiento histórico.
* Inserción en una **base de datos relacional** para consultas operativas.
* Exposición mediante **API REST** para consumo por sistemas clínicos y BI.

**4. Buenas prácticas aplicadas**

1. **Versionado de datos**: mantener trazabilidad de cada carga para auditoría.
2. **Validaciones estrictas**: aplicar reglas clínicas para evitar inconsistencias.
3. **Alertas y monitoreo**: configurar notificaciones en Airflow ante fallos o retrasos.
4. **Separación de ambientes**: desarrollo, pruebas y producción bien definidos.
5. **Documentación del flujo**: cada tarea del DAG documentada para mantenimiento.

**5. Esquema textual del flujo**

1. Airflow inicia el DAG de consolidación cada noche.
2. Se ejecutan tareas de ingesta desde APIs y FTP.
3. Los datos se validan, transforman y enriquecen.
4. Se eliminan duplicados y se aplican reglas clínicas.
5. Los datos se cargan en el Data Lake y en la base de datos.
6. Se expone una API para consumo externo.
7. Se registra el estado del flujo y se envían alertas si hay errores.