Orquestación automatizada para consolidación masiva de registros clínicos

1) Necesidad principal

Los datos a procesar incluyen: pacientes, episodios/visitas, diagnósticos (CIE-10), procedimientos, recetas, estudios e imágenes (metadatos DICOM), turnos, facturación y logs clínicos.
El objetivo es consolidar diariamente (y en near-real-time para eventos críticos) los registros desde múltiples sedes y sistemas a un Data Lake y DWH clínico único, estandarizado y trazable, habilitando analítica, tableros y reporting regulatorio.

2) Plataforma propuesta

- Apache NiFi: ingesta y normalización de datos desde múltiples fuentes heterogéneas.
- Apache Airflow: orquestación de tareas complejas, dependencias y control de calidad.
- Talend: opcional, para escenarios que requieran ETL low-code.

La elección de NiFi y Airflow permite separar ingesta (NiFi) y orquestación (Airflow), facilitando la escalabilidad y el mantenimiento del flujo de datos.

3) Flujo de orquestación

A. Ingesta (NiFi)

1. Captura de fuentes: FHIR/REST, HL7 TCP, SFTP/SMB, JDBC, DICOM.
2. Validación y normalización: uso de esquemas (Avro/JSON Schema), reglas de negocio, diccionarios ICD10/SNOMED.
3. Enriquecimiento: unión con catálogos maestros.
4. Estandarización: salida en formato Parquet particionado en un Data Lake.
5. Control de flujo: back-pressure y dead-letter queue para registros inválidos.

B. Orquestación (Airflow)

DAG clínico diario:

1. Verificación de datos Bronze.
2. Deduplicación de pacientes con estrategia SCD2.
3. Curación de episodios, órdenes, estudios y recetas.
4. Validaciones de calidad (unicidad, nulidad, cobertura de catálogos).
5. Carga incremental en DWH clínico.
6. Publicación de Data Marts.
7. Notificación en Slack/Teams.

C. Entrega

- Data Lake (Bronze/Silver/Gold en Parquet/Delta).
- DWH clínico para BI y reporting.
- APIs FHIR opcionales.
- Auditoría y reprocesos trazables.

4) Buenas prácticas

1. Gobernanza y linaje de datos (Atlas/Glue).
2. Data contracts y SLAs con sistemas fuente.
3. Seguridad y cifrado de PHI.
4. Validaciones automáticas en cada paso.
5. Idempotencia y reintentos.
6. Particionamiento eficiente.
7. Observabilidad con métricas y dashboards.
8. Pruebas unitarias e integración.
9. Manejo de colas y DLQ en NiFi.
10. Políticas de ciclo de vida y retención.

5) Pseudocódigo del DAG

```
with DAG('clinical_consolidation_dag', schedule='0 1 * * *') as dag:
  wait_bronze >> dedup_pacientes >> [curate_episodios, curate_ordenes]
  >> quality >> load_dwh >> publish_marts >> notify
```

6) Métricas de éxito

- Tasa de duplicados < 0.5%
- Completitud de datos > 98%
- Latencia diaria < 2h
- Errores de carga < 0.1%
- Disponibilidad > 99.5%
- Cobertura de catálogos > 97%