

Orquestación automatizada para consolidación masiva de registros clínicos

Elaborado por: Jonathan Vásquez

1. Necesidad principal del flujo

La red de clínicas privadas requiere consolidar diariamente grandes volúmenes de información médica proveniente de diversas sedes y sistemas. Los datos incluyen:

- Información de pacientes (nombre, RUT, edad, sexo, historial clínico).
- Estudios clínicos (laboratorio, imagenología, resultados).
- Recetas médicas.
- Turnos y agendas médicas.

Objetivos de la automatización:

- Reducir errores, duplicaciones y demoras en la consolidación de datos.
- Garantizar la **calidad, consistencia y disponibilidad** de la información.
- Facilitar análisis clínicos, reportes internos y toma de decisiones estratégicas.
- Escalar el flujo para manejar grandes volúmenes de datos en tiempo real y batch diario.

2. Selección de plataforma

Plataforma elegida: Apache NiFi

Justificación:

- Permite **procesamiento mixto**: datos en streaming y batch.

- Maneja **fuentes heterogéneas**: bases de datos SQL/NoSQL, APIs, archivos CSV/Excel.
- **Interfaz gráfica** simplifica el diseño de flujos complejos para el equipo de datos.
- Ofrece **control de errores, deduplicación y trazabilidad completa** del lineaje de los datos.
- Facilita la **integración con data lakes, warehouses y APIs** sin necesidad de programación avanzada.

Nota: Airflow es ideal para orquestación basada en tareas programadas, pero NiFi se ajusta mejor a flujos continuos y heterogéneos de datos clínicos.

3. Flujo automatizado de datos

3.1. Ingesta

- **Origen de datos:**
 - Sistemas internos de cada clínica.
 - APIs de laboratorios externos y farmacias.
 - Archivos CSV/Excel diarios exportados de sistemas antiguos.
- **Método:**
 - NiFi captura datos en tiempo real desde APIs y streams.
 - Procesamiento batch para archivos CSV/Excel mediante programaciones diarias automáticas.

3.2. Procesamiento

- **Transformaciones aplicadas:**
 - Normalización de nombres, formatos de fecha y códigos de paciente.
 - Validación de campos obligatorios (ej.: RUT, fecha de nacimiento).

- Eliminación de registros duplicados.
- Enriquecimiento de datos: integración de resultados de laboratorios y recetas con la ficha del paciente.
- **Gestión de errores:**
 - Registro de incidencias y alertas automáticas.
 - Posibilidad de reintentos automáticos y flujo de cuarentena para datos inválidos.

3.3. Entrega

- **Destino de datos procesados:**
 - **Data Warehouse centralizado** para análisis corporativo.
 - **Data Lake** para almacenamiento histórico de todos los registros médicos.
 - APIs internas para que otros sistemas consuman datos limpios en tiempo real.

4. Buenas prácticas aplicadas

1. **Lineaje completo de datos:** Documentar origen, transformaciones y destino para auditorías y cumplimiento normativo.
2. **Alertas y monitoreo en tiempo real:** Notificar al equipo de datos ante errores o inconsistencias.
3. **Diseño modular:** Cada flujo puede actualizarse o escalarse sin afectar el sistema completo.
4. **Seguridad y privacidad:** Cifrado de datos sensibles y cumplimiento con normativas de salud (ej. HIPAA, si aplica).
5. **Validación continua:** Revisar consistencia y calidad de los datos antes y después de cada transformación.

Conclusión:

La implementación de este flujo automatizado con **Apache NiFi** permitirá a la red de clínicas consolidar grandes volúmenes de información médica de forma eficiente, segura y escalable, reduciendo errores humanos y mejorando la disponibilidad de datos para análisis y toma de decisiones.