Implementación de una Base de Datos Analítica In-Process con Arquitectura Medallion para la Gestión Eficiente de Listas de Espera en Salud Pública

Autor: Paulo Villarroel, Data Scientist, Ministerio de Salud. Chile

Palabras claves: data warehouse, duckdb, analítica, lista de espera, ETL

Abstract

Las listas de espera en atención médica representan un desafío crítico para los sistemas de salud pública, especialmente en Chile, donde a marzo de 2024, más de 2.8 millones de pacientes esperan atención médica especializada o intervenciones quirúrgicas. Este trabajo presenta una solución innovadora para la gestión eficiente de grandes volúmenes de datos, utilizando R y DuckDB, una base de datos analítica in-process.

La implementación de un Data Warehouse con arquitectura Medallion (bronce, plata y oro) permite organizar y refinar más de 60 millones de registros anuales, mejorando significativamente la calidad, trazabilidad y accesibilidad de los datos. La capa "Oro", optimizada con DuckDB, ha permitido reducir los tiempos de procesamiento de consultas avanzadas de días a horas y automatizar reportes mensuales, lo que facilita la generación de estadísticas en tiempo real y el monitoreo continuo de pacientes en espera.

Este enfoque no solo optimiza el análisis de listas de espera, sino que también sienta una base sólida para aplicaciones futuras de inteligencia artificial y analítica avanzada en salud pública. La solución es escalable y adaptable, lo que la hace aplicable a otros sistemas de salud a nivel global, ofreciendo un modelo de gestión de datos robusto y eficiente para enfrentar desafíos similares.

Las listas de espera de atención médica son uno de los factores más relevantes en prácticamente cualquier sistema de salud del mundo, y representan uno de los mayores desafíos de los Estados y Gobiernos en salud pública. A marzo de 2024, en Chile hay 349.581 casos en espera de una intervención quirúrgica y 2.512.798 para atención de alguna especialidad médica u odontológica, con más de 610.000 casos resueltos durante el primer trimestre de 2024[1].

[1] Glosa 6 Listas de Espera ORD-1789-Glosa-06-letras-a-b-c-i-j-l-trimestre-2024.pdf (minsal.cl)

Los datos de lista de espera se gestionan localmente en cada institución y luego se cargan en una plataforma nacional denominada SIGTE[2], donde se consolida un repositorio mensual de casos tanto de ingresos como de egresos de lista de espera. La gestión eficiente de esta información es fundamental no solo para monitorear el desempeño en la atención de pacientes, sino también para la generación de reportes y estadísticas nacionales.

Dada la gran cantidad de registros y el constante movimiento de casos en listas de espera, así como la necesidad de contar con datos oportunos, se hace necesario implementar estrategias robustas para el manejo masivo de datos. Para ello, se ha implementado un Data Warehouse que permite la ingesta de datos desde todas las instituciones públicas de salud y otras fuentes complementarias.

El volumen de datos en un corte mensual típico de lista de espera es de aproximadamente 5 millones de registros, lo que asciende a unos 60 millones de registros al analizar un año completo. Con una alta dimensionalidad (cerca de 70 variables por caso), el análisis de varios años o estudios complejos puede involucrar centenas de millones de registros, lo cual requiere soluciones eficientes para el manejo de datos masivos.

En respuesta a estos desafíos, se ha adoptado una estrategia innovadora utilizando R y la librería DuckDB[3], una base de datos analítica in-process. DuckDB se destaca por su simplicidad y eficiencia en el procesamiento de consultas analíticas, integrándose profundamente con R para permitir consultas directas sobre datos sin necesidad de importación o copia. Sus principales ventajas incluyen la ejecución vectorizada de consultas, lo que mejora significativamente el rendimiento en análisis OLAP, y su extensibilidad para definir nuevos tipos de datos y funciones.

Esta solución está dentro de una estructura de Medallion usada por el Ministerio de Salud para manejar los datos a nivel nacional de listas de espera.

La arquitectura Medallion (también conocida como arquitectura de bronce, plata y oro) es un enfoque utilizado en la ingeniería de datos para estructurar y organizar datos en diferentes capas o etapas de refinamiento dentro de un lago de datos (data lake). Cada capa tiene un propósito específico en el proceso de transformación y almacenamiento de los datos, lo que facilita la gestión, la calidad, y el acceso a los datos para distintos casos de uso.

A continuación, se explican las tres capas principales del flujo de trabajo habitual (pipeline):

1. Bronce (Raw Data Layer):

Descripción: En esta capa se almacenan los datos crudos, tal como se reciben desde las fuentes. Estos datos no han sido procesados ni transformados de ninguna manera significativa.

---

[2] SIGTE https://sigte.minsal.cl/
[3] DuckDB https://duckdb.org/

Objetivo: La capa de bronce se utiliza para guardar una copia fiel de los datos originales, asegurando que se mantenga un registro completo y exacto de la información como fue recibida. En nuestro caso, se reciben los registros desde los distintos hospitales u servicios de salud.

2. Plata (Cleaned Data Layer):

Descripción: En esta capa, los datos son limpiados y transformados para que sean más consistentes y manejables. Esto incluye la eliminación de duplicados, la corrección de errores y el formateo de datos.

Objetivo: Preparar los datos para que sean más fáciles de analizar, manteniendo un balance entre la limpieza de datos y la flexibilidad para adaptarse a diferentes necesidades analíticas.

En el modelo de datos, lo que hacemos es organizar los datos, implementar estrategias de calidad, conexión con otras bases de datos y una serie de transformaciones para limpiar los datos y brindar calidad.

3. Oro (Curated Data Layer):

Descripción: En esta capa, los datos están altamente refinados y transformados para satisfacer necesidades analíticas específicas, modelos de machine learning, o para ser consumidos directamente por usuarios finales.

Objetivo: Proporcionar datos finales de alta calidad que puedan ser utilizados para análisis avanzados, reportes, cuadros de mando (dashboards), o modelos predictivos.

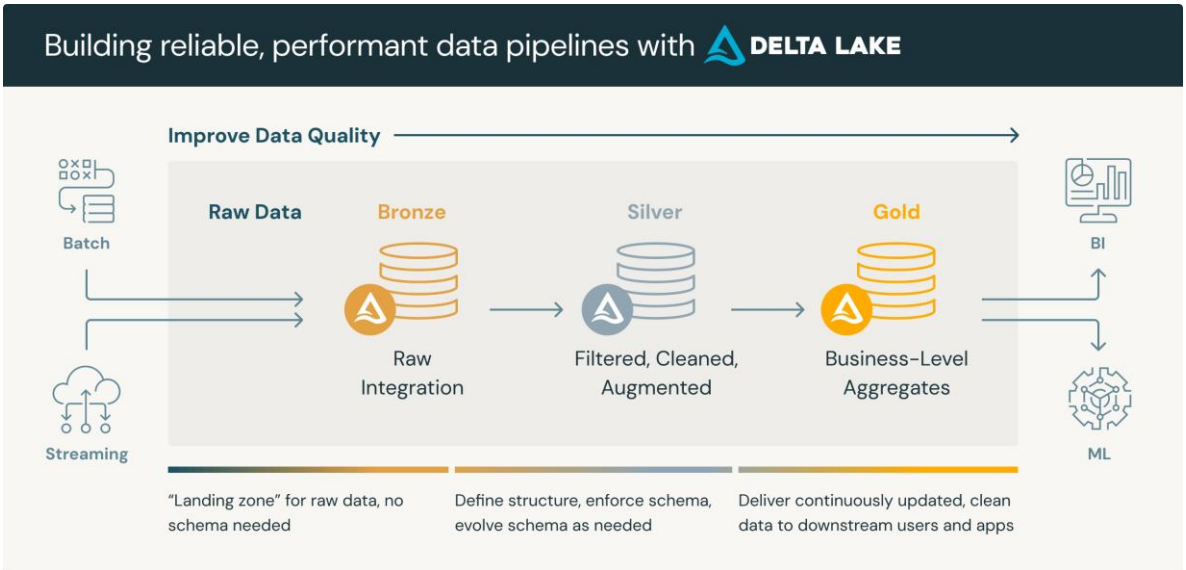
Es en esta capa donde integramos DuckDB para generar el manejo eficiente y de alto desempeño en la generación de reportes y analíticas avanzadas de listas de espera.

Además, desde esta capa se proveen datos validados para alimentar visores públicos[4], plataformas de autoconsultas[5], reportes oficiales y RDBMS para gestión de datos por distintos analistas.

La arquitectura Medallion ayuda a mantener un flujo claro de datos desde su ingestión hasta su análisis final, permitiendo a los equipos de datos trabajar de manera más estructurada y eficiente, algo fundamental en la gestión de registros de pacientes en listas de espera

[4] Visor ciudadano MNSAL https://public.tableau.com/app/profile/tableau.minsal/viz/PortadaLE/PortadaLE

[5] Portal Paciente https://portalpaciente.minsal.cl/



Medallion Architecture https://www.databricks.com/glossary/medallion-architecture

La integración de DuckDB con R para la analítica avanzada en tiempo real se presenta como una solución innovadora que supera las limitaciones de herramientas tradicionales en el procesamiento de grandes volúmenes de datos en sistemas de salud. La metodología adoptada incluyó la creación de pipelines automatizados para la ingesta de datos, optimización de consultas SQL mediante DuckDB, y la integración de estos procesos dentro de scripts en R, lo que resultó en un rendimiento mejorado en el análisis de datos de listas de espera.

Gracias a esta estrategia, además, se logró automatizar gran parte de los reportes mensuales de listas de espera, reduciendo notablemente el tiempo de procesamiento y permitiendo la generación de estadísticas en streaming. Esto facilita el seguimiento de pacientes en espera, el análisis estadístico y la evaluación de estrategias para reducir los tiempos de espera.

Este trabajo no solo aborda el análisis de grandes datos con R, sino que también contribuye al campo de la inteligencia artificial favoreciendo el acceso a datos de mejor calidad, y a la visualización de datos mediante dashboards interactivos, proporcionando un enfoque holístico y eficiente para la gestión de listas de espera en salud pública.

Para la generación de reportes se ha usado, principalmente, la librería *Tidyverse* para la limpieza y transformación de datos. Además, la exportación de tablas se ha realizado con la librería *writexl*.

Resultados

La implementación de la arquitectura Medallion con DuckDB y R para la gestión de listas de espera en el sistema de salud pública chileno ha producido mejoras significativas en varios aspectos clave del procesamiento y análisis de datos.

A continuación, se detallan los principales resultados obtenidos:

1. Reducción en el tiempo de generación de cortes mnsuales: La generación de cortes mensuales oficiales de listas de espera, un proceso crítico para el monitoreo y la toma de decisiones ha experimentado una mejora sustancial en eficiencia. El tiempo requerido para este proceso se ha reducido de aproximadamente 10 días a menos de 4 días. Esta reducción del 60% en el tiempo de procesamiento permite una respuesta más ágil a las necesidades del sistema de salud y facilita una toma de decisiones más oportuna.

2. Implementación de actualización diaria de dashboards: Gracias a la implementación del data warehouse y los pipelines automatizados (ETL), se ha logrado generar datos diarios de forma automática para la actualización de dashboards. Esta capacidad, que anteriormente no era factible, proporciona una visión en tiempo casi real del estado de las listas de espera. La disponibilidad de información actualizada diariamente mejora significativamente la capacidad de monitoreo y permite una gestión más proactiva de las listas de espera.

3. Optimización de la generación de reportes oficiales: La integración de DuckDB para el manejo masivo de datos ha permitido una notable mejora en la eficiencia de la generación de reportes oficiales. Este proceso, que anteriormente podía tomar varios días, ahora se completa en solo 1 día. Esta optimización no solo ahorra tiempo y recursos, sino que también permite una difusión más rápida de información crítica a los tomadores de decisiones y al público.

4. Automatización de sistemas de detección de anomalías y tests de calidad: Se han implementado sistemas automatizados de detección de anomalías y tests de calidad utilizando R. Esta nueva capacidad, que no existía previamente, permite una identificación temprana de problemas potenciales en los datos y asegura un alto estándar de calidad en la información procesada. La automatización de estos procesos mejora la confiabilidad de los datos y reduce el riesgo de errores no detectados en los análisis subsiguientes.

Estos resultados demuestran cómo la integración de tecnologías avanzadas como DuckDB y R, dentro del marco de la arquitectura Medallion, ha transformado significativamente la capacidad del sistema de salud chileno para manejar y analizar datos de listas de espera. La

mejora en la velocidad de procesamiento, la frecuencia de actualización de datos, la eficiencia en la generación de reportes y la implementación de controles de calidad automatizados contribuyen colectivamente a una gestión más efectiva y oportuna de las listas de espera en el sistema de salud pública.

Estas mejoras no solo optimizan los procesos internos, sino que también tienen un impacto directo en la capacidad del sistema para responder a las necesidades de los pacientes, permitiendo una planificación más precisa y una asignación más eficiente de los recursos de salud.

Además, permiten preparar el terreno para las futuras integraciones de los proyectos de interoperabilidad planificados en el mediano plazo.

Discusión

La implementación de una base de datos analítica in-process con DuckDB, integrada a la arquitectura Medallion, ha demostrado ser altamente eficiente para la gestión de grandes volúmenes de datos en el sistema de salud pública de Chile. La estructura de capas (bronce, plata y oro) permitió un manejo ordenado y progresivo de los datos, desde su forma cruda hasta un estado refinado y listo para el análisis.

En particular, la capa "Oro" ha sido fundamental para la generación de reportes y estadísticas oficiales del Ministerio de Salud de Chile. En esta etapa, R se utiliza junto con DuckDB para procesar los datos refinados y realizar análisis complejos, consultas avanzadas y la generación de reportes automáticos. La integración de DuckDB con R en la capa 'Oro' ha reducido los tiempos de procesamiento de consultas analíticas de varios días a solo unas horas, lo que ha mejorado considerablemente la eficiencia operativa y la capacidad de respuesta del sistema

La integración de DuckDB con R en esta fase final optimiza la interacción con grandes volúmenes de datos sin necesidad de importar o duplicar la información, mejorando la eficiencia en la creación de informes y análisis. Sin embargo, algunos desafíos persisten, como la capacitación del personal para usar estas herramientas de manera efectiva y la integración con sistemas heterogéneos de datos a nivel regional. A pesar de estas limitaciones, la solución es adaptable y escalable, lo que sugiere un potencial significativo para replicarla en otros contextos de sistemas de salud.

Desafíos

La gestión de datos de listas de espera es una tarea compleja, requiere de importantes aspectos de colaboración y mejoras de calidad de los registros locales en los hospitales. Pero

quizás el factor más desafiante es implementar un modelo de gobernanza de datos. Esto no es un problema solo técnico, sino que requiere de cambios culturales y de personas con competencias específicas al interior de las instituciones, algo que aún es escaso.

El diseño de la arquitectura de datos de MINSAL es una ventaja importante para soportar y adaptarse a cambios locales, pero es insuficiente sin afectar las áreas de atención clínica, que es donde realmente se capturan los datos. Y si bien se ha avanzado en ese camino, es necesario seguir madurando en temas de calidad, gobernanza y uso secundario de datos.

Conclusiones

- Eficiencia en la generación de reportes con R: Aunque el Data Lake y el Data Warehouse son gestionados en otras plataformas (Oracle, PostgreSQL), el uso de R junto a DuckDB en la capa "Oro" ha permitido la generación de reportes y estadísticas de manera eficiente, asegurando que los datos estén listos para el análisis y la toma de decisiones.

- Reducción de tiempos de procesamiento: La integración de DuckDB en scripts de R en la etapa de generación de reportes ha reducido significativamente los tiempos de procesamiento de consultas analíticas, mejorando la capacidad de respuesta del sistema de salud ante las necesidades de monitoreo y análisis de listas de espera.

- Automatización de reportes oficiales: La metodología adoptada ha permitido automatizar gran parte de los reportes oficiales del Ministerio de Salud de Chile en temas de listas de espera, lo que no solo ahorra tiempo y recursos, sino que también mejora la precisión y consistencia de la información presentada.

- Integración de tests de calidad de datos con R: Se han incorporado pruebas de calidad de datos usando R como parte de los flujos de trabajo habituales en la capa "Oro", lo que garantiza que la información utilizada en los análisis y reportes cumpla con los estándares de calidad necesarios. Esto ha facilitado la identificación temprana de inconsistencias y ha mejorado la confianza en los datos procesados.

- Escalabilidad y adaptabilidad: La solución basada en la arquitectura Medallion, con el uso de R en la fase final de análisis, es adaptable a otros sistemas de salud a nivel global. La capacidad de gestionar grandes volúmenes de datos de manera escalable y la flexibilidad de R para trabajar con DuckDB ofrecen un modelo replicable para enfrentar desafíos similares.

- Potencial para análisis avanzados con R: El uso de R para el análisis de datos refinados en la capa 'Oro' no solo facilita la implementación de modelos predictivos y

aplicaciones de inteligencia artificial, sino que también posiciona al sistema de salud para adaptarse rápidamente a nuevos desafíos, potenciando la toma de decisiones basada en datos

Finalmente, este trabajo resalta cómo la integración de R en la etapa final del procesamiento de datos dentro de la arquitectura Medallion, junto con DuckDB, puede transformar la gestión de datos masivos en salud pública. Esto ha permitido mejorar la generación de informes y estadísticas oficiales, incorporar controles de calidad de datos, y contribuir a una planificación más precisa y a una toma de decisiones más informada para el bienestar de los pacientes en listas de espera.

Bibliografía

1. Native Delta Lake Support in DuckDB https://duckdb.org/2024/06/10/delta.html
2. Data Mesh Principles and Logical Architecture https://martinfowler.com/articles/data-mesh-principles.html
3. Documentación DuckDB https://github.com/duckdb/duckdb