Diseño de un prototipo de integración de datos de salud mediante arquitecturas de Big Data, utilizando fuentes de información pública en el sistema de salud colombiano

Gabriel Jair Buendia Diaz

Asesor

Jorge Luis Quintero López

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Dedicatoria

Quiero dedicar este logro a mi familia, pilares fundamentales en cada etapa de mi vida, quienes con su paciencia y ejemplo me han enseñado el valor de la perseverancia y la disciplina.

Dedico este logro a quienes ya no están físicamente, pero dejaron huellas imborrables en mi camino. Su recuerdo y enseñanzas siguen acompañándome, guiando mis pasos y dándome fortaleza en los momentos de mayor dificultad. Este logro no debe ser el último, y espero que muchos más puedan ser dedicados a ustedes como testimonio de su influencia en mi vida.

A mis seres cercanos, que sin saberlo me brindaron gran parte de su apoyo incondicional, ayudándome a superar adversidades y obstáculos que parecían imposibles. Sus palabras, gestos y compañía fueron un impulso silencioso que me permitió mantenerme firme y avanzar con determinación.

Este logro está dedicado a todos aquellos que confiaron en mí y en mis capacidades, incluso cuando las circunstancias parecían adversas. Es el reflejo de la unión, la confianza y el compromiso que han marcado este recorrido, y demuestra que el compromiso, la lealtad y la convicción siempre se convierten en triunfo.

Agradecimientos

Quiero expresar mi sincero agradecimiento a las personas con las que trabajé, quienes me inspiraron y motivaron con sus ideas, y me brindaron su apoyo y las facilidades necesarias para continuar con este proyecto. Su acompañamiento fue decisivo para mantener la constancia y avanzar en cada etapa del proceso.

Un especial agradecimiento a mi tutor y director Jorge Luis Quintero López, por su asesoría, acompañamiento y paciencia. Su guía fue fundamental para superar las dificultades y hacer posible la culminación de este trabajo.

A la Universidad Nacional Abierta y a Distancia (UNAD), por brindarme el espacio académico y metodológico que permitió el desarrollo de este trabajo. Extiendo también mi gratitud a todos los docentes con quienes compartí este recorrido, por sus enseñanzas y aportes que enriquecieron mi formación.

Finalmente, agradezco a mis compañeros de trabajo y estudio, quienes con sus comentarios, discusiones y apoyo contribuyeron a fortalecer mi crecimiento profesional.

Resumen

La falta de eficiencia en el sistema de salud colombiano constituye un desafío social de gran magnitud, reflejado en retrasos en la atención, deficiencias en los procesos, errores administrativos y un creciente descontento ciudadano. La carencia de canales centralizados de información dificulta el seguimiento de los pacientes, reduce la capacidad de análisis de las instituciones y afecta la calidad de vida, en especial de las comunidades más vulnerables. Ante esta situación, este proyecto plantea el desarrollo de un prototipo de base central de datos estructurados mediante técnicas de Big Data, con el propósito de integrar y consolidar la información del sistema de salud nacional. Bajo un enfoque de Investigación Aplicada y Design Science Research (DSR), se definen los requerimientos, la arquitectura, los procesos de gestión de datos y los mecanismos de validación necesarios para una solución que optimice la disponibilidad, calidad y uso estratégico de la información, cimentando las bases para un ecosistema que favorezca la analítica avanzada, la predicción clínica y la toma de decisiones fundamentadas en datos, ofreciendo una alternativa tecnológica para reducir la ineficiencia estructural y fortalecer la gestión en salud en Colombia.

Palabras claves: Big data, analítica, repositorio, salud, eficiencia.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

En la actualidad, los datos se han consolidado como activos estratégicos para organizaciones y gobiernos. En el sector salud, su integración y análisis resulta esencial para fortalecer la toma de decisiones, optimizar recursos y responder a problemáticas epidemiológicas. En este contexto, las arquitecturas de Big Data permiten procesar información heterogénea de manera escalable, aunque la interoperabilidad continúa siendo un desafío para los sistemas modernos.

En Colombia, entidades como el Instituto Nacional de Salud (INS), el Ministerio de Salud y Protección Social y diversas plataformas de datos abiertos generan grandes volúmenes de información. Sin embargo, la fragmentación en formatos y niveles de calidad limita la construcción de una visión integral y reduce la capacidad analítica para identificar patrones epidemiológicos y necesidades de atención.

Ante esta problemática, el proyecto propone un prototipo de integración inteligente de datos de salud basado en arquitecturas Big Data y principios Lakehouse, implementando el modelo Medallion en capas Bronze, Silver y Gold. El enfoque se sustenta en la metodología Design Science Research (DSR), orientada a validar artefactos tecnológicos y demostrar la viabilidad de una infraestructura escalable y de calidad para el sistema de salud colombiano.

Finalmente, el documento expone progresivamente la descripción del problema, el marco de referencia, la metodología aplicada y los resultados obtenidos, concluyendo con recomendaciones y líneas de trabajo futuro en integración de datos y analítica avanzada.

Justificación

La transformación digital del sector salud ha incrementado la producción de información clínica, epidemiológica y administrativa, convirtiendo los datos en un recurso estratégico para la toma de decisiones y políticas públicas. Sin embargo, en Colombia persisten limitaciones en interoperabilidad, centralización y calidad de los sistemas de información. Desde la Ley 100 de 1993, se han evidenciado problemas de fragmentación, duplicidad y ausencia de estándares para el intercambio de datos (Myriam & Guerrero, 2023).

La dispersión de fuentes heterogéneas administradas por distintos actores institucionales dificulta la consolidación de una visión integral del sistema de salud, afectando procesos administrativos y reduciendo la capacidad de análisis avanzado para identificar patrones epidemiológicos y optimizar recursos. Esta situación limita la generación de conocimiento basado en evidencia y la respuesta oportuna ante escenarios críticos de salud pública.

Ante ello, las arquitecturas de Big Data y modelos tipo Lakehouse ofrecen una alternativa tecnológica para gestionar grandes volúmenes de datos estructurados y no estructurados, facilitando integración, limpieza y análisis avanzado. La adopción de arquitecturas Medallion asegura trazabilidad y calidad, mientras que la inteligencia artificial generativa aporta innovación en el descubrimiento y clasificación automatizada de fuentes de datos. Este proyecto busca sentar las bases que permitan evolucionar de un modelo fragmentado hacia uno proactivo, predictivo y eficiente, en línea con los avances de la medicina y la gestión administrativa.

Objetivos

Objetivo General

Desarrollar un prototipo de integración de datos de salud mediante arquitecturas de Big Data, utilizando fuentes de información pública.

Objetivos Específicos

Caracterizar fuentes de datos abiertos para establecer las estructuras de información reales sobre las cuales operará el prototipo.

Definir la estructura de ingesta de datos mediante la aplicación de técnicas de limpieza y normalización en el ecosistema de big data, asegurando la calidad y consistencia de la información.

Construir un prototipo de integración funcional, que permita consolidar las fuentes de información pública y sirva como base estructurada para visualización y consulta.

Planteamiento del Problema

Las organizaciones modernas generan grandes volúmenes de datos en sus procesos internos, lo que supera las capacidades humanas para analizarlos y transformarlos en conocimiento útil para la toma de decisiones (Alarcón García, 2021). En el sector salud, la eficiencia depende en gran medida de la calidad y el flujo de la información; sin embargo, esta gestión sigue siendo un reto global. Incluso en países de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), cerca del 20% del gasto en salud resulta ineficiente por demoras en la atención y errores en la medicación, generando cuellos de botella y poniendo en riesgo vidas humanas (Felipe et al., 2023).

En Colombia, el sistema de salud funciona como una red compleja de múltiples organizaciones. La ausencia de un diseño unificado ha originado fuentes de información desconectadas, lo que provoca duplicidad de datos, registros erróneos y dificultades para que las Entidades Promotoras de Salud (EPS) e Instituciones Prestadoras de Salud (IPS) interpreten información común. Esta fragmentación tecnológica incrementa el consumo innecesario de recursos administrativos y financieros (Rodríguez-Páez et al., 2024), afectando directamente el bienestar social: deteriora la calidad de vida, aumenta la mortalidad prevenible y debilita la confianza ciudadana en el sistema.

En comunidades rurales y vulnerables, la atención se percibe como un proceso desgastante y poco accesible (Urdinola et al., 2023). Además, la falta de oportunidad en la atención vulnera el derecho fundamental a la salud y profundiza las desigualdades sociales, al privilegiar la rentabilidad financiera sobre los derechos de las personas. Ante este panorama de dispersión de datos e ineficiencia administrativa, se evidencia la necesidad de una solución tecnológica basada en arquitecturas de datos modernas que integre la información entre los

distintos actores del sistema y permita una visión única de la realidad sanitaria. Para dar respuesta a la problemática planteada, se formula la siguiente pregunta de investigación:

¿Cómo implementar un prototipo de base de datos estructurada con técnicas de Big Data que consolide la información del sistema de salud colombiano y optimice la gestión de la atención médica?

Marco de Referencia

Marco Teórico

La presente propuesta de trabajo se estructura a partir de tres pilares fundamentales que se interconectan para dar sustento, rigor técnico y viabilidad al proyecto. En primer lugar, se analiza el sistema de salud en Colombia, lo que permite entender el contexto normativo, los desafíos actuales y las dinámicas del sector donde se aplicará la solución. Como segundo eje, se abordan los fundamentos de arquitectura, que proveen el diseño estructural y las bases tecnológicas necesarias para garantizar un sistema robusto, escalable y eficiente. Finalmente, se integra la analítica avanzada, aportando las herramientas metodológicas y estadísticas clave para transformar los datos en conocimiento estratégico y decisiones de alto impacto.

A continuación, se detalla el desarrollo de cada uno de estos componentes esenciales:

- El sistema de Salud en Colombia: El sistema de salud colombiano, derivado de la Ley 100 de 1993, opera bajo un modelo de competencia regulada (García Hernández & Esquer Bojorquez, 2024), donde la financiación pública coexiste con una operación predominantemente privada a través de las EPS e IPS(Castano et al., 2024). Posteriormente, la Ley 1438 de 2011 buscó fortalecer la coordinación intersectorial para la atención integral, la promoción de la salud y la prevención de la enfermedad(Sepúlveda Correa et al., 2021). Este sistema se describe como complejo, caracterizado por interacciones no lineales y fenómenos emergentes, lo que dificulta la supervisión y el control (Rodríguez-Páez et al., 2024). Las fallas estructurales, incluyendo la ausencia de un sistema de información unificado, han provocado fragmentación y segmentación, resultando en una atención poco integral y oportuna (Myriam & Guerrero, 2023) Esta fragmentación se refleja en la deficiente transferencia de información clínica entre niveles de atención, asociada a estrategias de control de costos de los administradores (Vargas et al., 2016).

- Fundamentos de la arquitectura Big Data: El proyecto se sustenta en la teoría de que la tecnología Big Data permite extraer información útil para la toma de decisiones estratégicas (García Espinosa & Figueroa Alvarado, 2021). Su aplicación es crucial para reconfigurar sistemas de salud deficientes (Felipe et al., 2023). La duplicidad de datos y la falta de un diseño unificado (Rodríguez-Páez et al., 2024) pueden mitigarse mediante arquitecturas de Big Data que consoliden información robusta y útil a partir de diversas fuentes históricas (Felipe et al., 2023).
- Analítica Avanzada: La centralización de datos habilita la analítica avanzada y el Aprendizaje Automático (ML). Modelos basados en ML han demostrado eficacia en el diagnóstico temprano de enfermedades (Mejia et al., 2023) y en la predicción de resultados clínicos adversos, como mortalidad u hospitalización en pacientes con Enfermedades No Transmisibles (ENT). Estrategias como la estratificación de riesgo son fundamentales para optimizar recursos y avanzar hacia modelos predictivos (Hernández-Arango et al., 2025).

Marco Conceptual

Para garantizar el correcto desarrollo y la viabilidad técnica del prototipo, es fundamental delimitar el terreno conceptual sobre el cual se construye. Esta sección reúne los conceptos operativos y técnicos esenciales que sirven como lenguaje común y base de diseño. A nivel operativo, se definen los procesos, flujos de trabajo y reglas de negocio que determinan el comportamiento del sistema; a nivel técnico, se establecen las herramientas, estándares informáticos y metodologías de datos necesarios para su implementación. De este modo, se asegura que cada componente del prototipo responda de manera precisa a las necesidades y arquitecturas planteadas en el proyecto.

- Big Data y Repositorios Centrales Estructurados: El Big Data comprende herramientas y metodologías para gestionar y procesar grandes volúmenes de datos que superan la capacidad de las bases de datos convencionales. Su objetivo es extraer valor de la información masiva, transformándola en insumos útiles y de alta calidad para la toma de decisiones (García Espinosa & Figueroa Alvarado, 2021). Para ello, se requiere el uso de tecnologías que permitan la recopilación, manipulación, almacenamiento y análisis de datos (Alarcón García, 2021). La implementación de Big Data se traduce en la creación de Repositorios Centrales de Datos o Data Warehouse, que funcionan como puntos estratégicos de integración de información proveniente de múltiples fuentes heterogéneas. Estos repositorios permiten almacenar datos históricos estructurados y no estructurados, utilizando sistemas distribuidos como Hadoop, modelos de programación como MapReduce y frameworks de computación en clúster como Apache Spark, consolidando insumos de alta calidad para el análisis avanzado en salud (Felipe et al., 2023)
- Procesamiento de Big Data: El procesamiento de Big Data es un procedimiento técnico orientado a la obtención de información de valor, el cual se divide en tres fases críticas para garantizar la exactitud de los análisis (Malla Valdiviezo et al., 2023). La primera fase corresponde a la limpieza de datos, enfocada en la eliminación o corrección de información inconsistente, duplicada o faltante, así como en la normalización para asegurar homogeneidad y consistencia. A esto le sigue la transformación de datos, etapa en la que se realiza la conversión de los registros a un formato óptimo para el análisis mediante la agregación de variables y la eliminación de información irrelevante. Finalmente, se ejecuta el análisis de datos, donde se aplican técnicas especializadas que optimizan los tiempos de respuesta y facilitan la toma de decisiones estratégicas. (Malla Valdiviezo et al., 2023)

- Gobernanza de Datos: La gobernanza de datos es un marco de políticas, procedimientos, roles y responsabilidades que asegura que los datos sean gestionados como un activo estratégico dentro de una organización. En sistemas complejos como el de salud, este marco resulta esencial (Rodríguez-Páez et al., 2024). En el sector salud, dada la sensibilidad y heterogeneidad de la información, la gobernanza implica la definición de normativas obligatorias para garantizar seguridad, privacidad y acceso a los datos masivos. Este enfoque está íntimamente ligado a la Gestión del Conocimiento (Rodríguez-Páez et al., 2024). Su objetivo no es solo la protección legal y ética de la información, sino también asegurar la calidad de los datos, garantizando que los registros sean consistentes, completos y útiles para la toma de decisiones clínicas y administrativas en todos los niveles del sistema (Felipe et al., 2023).
- Aprendizaje Automático (Machine Learning): El aprendizaje automático (ML) se refiere al uso de modelos predictivos basados en técnicas de inteligencia artificial aplicadas en medicina. Estos modelos apoyan el diagnóstico y permiten predecir resultados clínicos adversos, como mortalidad u hospitalización (Mejia et al., 2023). Una de sus aplicaciones clave es la estratificación de riesgo, que consiste en identificar y agrupar pacientes según su severidad o nivel de riesgo. El objetivo es asignar intervenciones adaptadas a sus futuras necesidades, optimizando recursos y mejorando la atención (Hernández-Arango et al., 2025).
- Inteligencia Artificial Generativa: La inteligencia artificial generativa representa una aplicación avanzada del aprendizaje automático, cuya base es el Big Data consolidado. Esta tecnología tiene un potencial significativo para revolucionar la medicina, ya que permite analizar grandes volúmenes de datos de salud, identificar patrones únicos y predecir respuestas individuales a los tratamientos. De esta manera, se busca mejorar los resultados clínicos y personalizar la atención al paciente (Bhuyan et al., 2025).

Metodología

Tipo de Metodología Aplicada

El presente proyecto se fundamenta en la metodología *Design Science Research (DSR),* un marco de trabajo especializado en la creación y evaluación de artefactos tecnológicos diseñados para resolver problemas organizacionales complejos (Peffers et al., 2007). Bajo este enfoque, el prototipo de integración de datos no se considera únicamente un producto de software, sino un aporte al conocimiento técnico en el área de la ingeniería de datos aplicada a la salud.

El diseño del artefacto se estructura bajo un modelo iterativo que garantiza la alineación entre las necesidades del sistema de salud colombiano y las capacidades de las arquitecturas de Big Data. Este proceso se desglosa en cuatro fases fundamentales:

Fase 1: Identificación y Definición de Requerimientos

Esta fase inicial constituye el diagnóstico técnico y operativo. Se centra en comprender la naturaleza de los datos públicos y las limitaciones de interoperabilidad actuales.

- Caracterización de fuentes de información: Se realiza un inventario técnico de los activos de datos presentes en RIPS, registros de EPS/IPS y repositorios de datos abiertos. Se analizan metadatos, esquemas de almacenamiento y protocolos de transferencia.
- Definición de requerimientos funcionales y analíticos: Establecimiento de las capacidades de procesamiento necesarias para soportar análisis predictivos y descriptivos, priorizando la estandarización semántica de los datos clínicos.
- Diagnóstico de calidad de datos (Data Profiling): Aplicación de técnicas de perfilamiento para identificar inconsistencias, valores nulos, duplicidad y problemas de integridad referencial que comprometen la veracidad de la información sanitaria.

Fase 2: Diseño de la Arquitectura y Modelado de Datos

Representa la etapa de ingeniería conceptual y lógica, donde se define la infraestructura que soportará el prototipo.

- Arquitectura Big Data (Medallion Architecture): Diseño de un ecosistema basado en niveles (Bronze, Silver, Gold) que permita el flujo de datos desde su estado crudo hasta su refinamiento analítico, utilizando paradigmas de procesamiento distribuido.
- Selección de componentes tecnológicos: Evaluación y selección de motores de procesamiento de alta escala (como Apache Spark) y entornos de almacenamiento escalable (Data Lakes), garantizando la capacidad de respuesta ante grandes volúmenes de registros.
- Modelado multidimensional: Implementación de estructuras de datos bajo esquemas de Copo de Nieve (*Snowflake Schema*). Este enfoque optimiza el rendimiento de consultas complejas y facilita la normalización de entidades médicas altamente relacionadas.

Fase 3: Especificación de los Procesos ETL/ELT y Mecanismos de Tratamiento

En esta fase se implementa la lógica de transformación que garantiza la "única fuente de verdad" del sistema.

- Limpieza y Curación: Desarrollo de algoritmos para la normalización de registros, corrección automática de errores y manejo de datos faltantes mediante técnicas estadísticas.
- Transformación y Enriquecimiento: Estandarización de vocabularios clínicos y códigos diagnósticos. Se preparan los datos mediante procesos de ingeniería de características (*Feature Engineering*) para facilitar su uso posterior en modelos de analítica avanzada.
- Carga y Orquestación: Definición de los mecanismos de persistencia en la base central, asegurando la trazabilidad del dato (*Data Lineage*) y su disponibilidad inmediata para el consumo informativo.

Fase 4: Definición del Protocolo de Pruebas y Validación

La fase final asegura que el artefacto desarrollado cumpla con los objetivos de diseño y sea técnicamente robusto.

- Pruebas de estrés y desempeño: Evaluación del comportamiento de la arquitectura bajo cargas de datos masivas y consultas concurrentes, midiendo tiempos de respuesta y latencia.
- Validación de integridad estructural: Verificación de la consistencia del modelo de datos y la eficacia de los mecanismos de indexación implementados.
- Validación funcional mediante visualización: Diseño de una interfaz analítica (Dashboard) que actúa como prueba de concepto, demostrando cómo la información integrada permite realizar análisis epidemiológicos y estratificación de riesgos en salud de manera eficiente.

Tipo de Estudio

La presente investigación se clasifica como un *estudio de carácter aplicado y tecnológico*, con un enfoque descriptivo-propositivo. Esta categorización se sustenta en los siguientes ejes:

- Investigación Aplicada: A diferencia de la investigación básica, este estudio no busca únicamente la generación de teorías abstractas, sino la aplicación de conocimientos científicos de la Ciencia de Datos para resolver una problemática social y administrativa específica: la fragmentación de la información en el sistema de salud colombiano (Alarcón García, 2021). El propósito final es la utilidad práctica y la mejora de procesos de toma de decisiones.
- Investigación Tecnológica: El estudio se centra en el desarrollo de un "artefacto" (el prototipo de arquitectura Big Data). Según los principios de la ingeniería, este tipo de estudio

sigue un proceso de síntesis donde se integran herramientas existentes (Spark, Databricks, IA Generativa) para crear una solución innovadora que actualmente no existe en el contexto analizado.

- Enfoque Descriptivo y Propositivo: El estudio es descriptivo en su fase inicial, ya que caracteriza el estado actual de los datos abiertos y las brechas de interoperabilidad. Es propositivo en su fase central, pues fórmula y construye una arquitectura técnica basada en el modelo Medallón para superar dichas limitaciones.

Este tipo de estudio permite que el investigador asuma un rol activo en la construcción de la solución, utilizando ciclos de prueba y error (iteraciones) propios de la metodología DSR, asegurando que la tecnología desarrollada sea escalable y responda a estándares de gobernanza de datos modernos.

Recolección de Datos

La estrategia de recolección de datos para este proyecto se fundamenta en la obtención de información de fuentes secundarias de carácter público. Dado que el prototipo busca la integración de datos abiertos del sistema de salud colombiano, se han definido protocolos técnicos para garantizar que la captura sea eficiente, reproducible y respete la integridad de los metadatos originales.

Las técnicas y herramientas empleadas para la recolección se dividen en tres categorías principales:

- Consumo de APIs y Microdatos: Se prioriza el acceso a través de interfaces de programación de aplicaciones (APIs) de portales como *Datos Abiertos Colombia* y el *Socrata Open Data API (SODA)*. Esto permite la extracción de datasets estructurados relacionados con la

vigilancia epidemiológica, indicadores de gestión de EPS y registros de capacidad instalada de IPS, garantizando una conexión directa y actualizada con la fuente.

- Web Scraping y Automatización: Para repositorios institucionales que no cuentan con APIs abiertas, se implementan técnicas de extracción automatizada mediante scripts de Python. Este proceso permite la captura de datos semiestructurados y metadatos alojados en sitios web oficiales, facilitando su posterior conversión a formatos compatibles con el Data Lake.
- Revisión y Extracción Documental: Se realiza un análisis detallado de diccionarios de datos, manuales técnicos de los RIPS (Registros Individuales de Prestación de Servicios de Salud) y resoluciones normativas. Esta técnica es fundamental para la fase de "mapeo", asegurando que los datos recolectados sean interpretados correctamente según el estándar clínico y administrativo nacional.

Los datos recolectados son almacenados inicialmente en su formato original en la Capa Bronze del prototipo, cumpliendo con los principios de gobernanza que exigen mantener la trazabilidad (*data lineage*) desde el origen hasta el procesamiento final.

Cronograma

El siguiente cronograma muestra las fases principales del proyecto aplicado y sus tiempos de ejecución. Su objetivo es ofrecer una visión clara y ordenada del desarrollo previsto, facilitando el seguimiento del avance y el cumplimiento de los plazos establecidos.

Tabla 1

Cronograma de Actividades

Actividad	Mes 1	Mes 2	Mes 3	Mes 4
Fase 1: Identificación y definición de requerimientos	X			
Fase 2: Diseño de la arquitectura y modelado de datos		X		
Fase 3: Especificación de los procesos ETL/ELT y mecanismos de tratamiento			X	
Fase 4: Definición del protocolo de pruebas y validación				X

Resultados

En este capítulo se expone la ejecución técnica y el desarrollo del prototipo de arquitectura de datos diseñado para el análisis del sector salud en Colombia. La implementación se organiza siguiendo el ciclo de vida de un proyecto de analítica de datos, desde la identificación de los requerimientos técnicos hasta la validación de la calidad de la información procesada.

A continuación, se presentan los resultados obtenidos en cada una de las fases de desarrollo, destacando la integración de fuentes de datos abiertos y la aplicación de un modelo de arquitectura de datos. Esta sección evidencia la viabilidad técnica de la solución propuesta al demostrar su capacidad para transformar datos crudos en activos de información confiables que respalden la toma de decisiones estratégicas en el ámbito de la salud pública. Finalmente, se establece una base de datos que no sólo consolida la información procesada, sino que también sirve como insumo para la construcción de modelos analíticos avanzados.

Primer Resultado: Diagnóstico y Requerimientos del Prototipo

Este primer bloque de resultados corresponde a la culminación de la Fase 1 de la metodología. Se centra en la comprensión del ecosistema de datos y la definición de las necesidades técnicas que debe cubrir el artefacto desarrollado.

Caracterización y Selección de Fuentes de Información

Este primer resultado representa la transición de la revisión documental hacia una caracterización técnica robusta, estableciendo el cimiento del prototipo. El proceso inició con una búsqueda exhaustiva en el portal de Datos Abiertos Colombia, identificando este repositorio como el eje central que concentra registros estratégicos del Instituto Nacional de Salud (INS) y del Ministerio de Salud y Protección Social.

Como producto de este análisis, se seleccionaron 16 fuentes de información (datasets) que conforman el insumo inicial del proyecto. Estas fuentes abarcan registros de vigilancia epidemiológica, prestación de servicios, talento humano y capacidad instalada. A continuación, se detallan las fuentes integradas en la fase inicial del prototipo:

Tabla 2

Fuentes de Información Seleccionadas del Portal de Datos Abiertos

Entidad	Nombre del Dataset / Tabla	Temática Principal
INS	DA-SIVIGILA 2021	Vigilancia en Salud Pública
INS	Vigilancia en Salud Pública de Colombia	Histórico de vigilancia
MinSalud	Indicadores de mortalidad y morbilidad	Salud pública por territorio
MinSalud	Número de afiliaciones (Protección Social)	Afiliaciones al sistema
MinSalud	Registros Individuales de Prestación (RIPS)	Prestación de servicios
MinSalud	Planilla Integrada de Liquidación (PILA)	Aportes y parafiscales
MinSalud	Reporte de prescripción (MIPRES)	Tecnologías no financiadas UPC
MinSalud	Talento Humano en Salud (RETHUS)	Registro de profesionales

MinSalud	Reporte de novedades de afiliados	Cambios en el SGSSS
MinSalud	Prestadores por departamento	Oferta de servicios
MinSalud	Ocupación de capacidad instalada	Disponibilidad hospitalaria
MinSalud	Indicadores Talento Humano (Saludatos)	Seguimiento THS
MinSalud	IPS públicas y privadas	Capacidad y niveles de atención
MinSalud	Afiliados por municipio y régimen	Cobertura territorial
MinSalud	Registro Especial de Prestadores (REPS)	Sedes y servicios habilitados
MinSalud	Indicadores de calidad IPS (ClicSalud)	Calidad percibida y técnica

Nota. Las URLs de acceso y los IDs únicos de cada dataset se encuentran documentados detalladamente en el Apéndices

Detección Automática y Caracterización Técnica

Un resultado diferenciador de este prototipo es el desarrollo de un módulo de caracterización técnica. A través de la extracción de metadatos mediante la API de Socrata, se logró describir de forma automática las estructuras, formatos y niveles de completitud. Este módulo permite:

- Identificación Unívoca: Extracción automática del ID del dataset desde la URL.

- Consulta Dinámica de Metadatos: Obtención de información sobre el número de columnas, tipos de datos, cantidad de registros y frecuencia de actualización.
- Análisis Comparativo: Generación de un DataFrame consolidado que permite evaluar la heterogeneidad de las fuentes antes de su procesamiento.

Finalmente, el prototipo fue escalado para habilitar la detección automática de nuevas fuentes mediante técnicas de inteligencia artificial. Este mecanismo facilita el descubrimiento dinámico de repositorios externos y el monitoreo continuo de las fuentes ya integradas, asegurando que la base de información se mantenga actualizada frente a cambios en los orígenes de datos gubernamentales. El código fuente de esta implementación se detalla en el apartado de evidencias técnicas.

Requerimientos Funcionales de Integración y Analítica

Tras la caracterización de las fuentes, se procedió a definir las especificaciones técnicas que guían la ejecución del prototipo. El diseño fue concebido no solo como un pipeline de datos, sino como un ecosistema inteligente que equilibra la automatización mediante inteligencia artificial con la supervisión humana necesaria en entornos críticos como el sector salud.
A continuación, se presentan los Requerimientos Funcionales (RF) que garantizan la capacidad operativa del prototipo:

- RF01 (Descubrimiento Automatizado): Capacidad del sistema para identificar de manera autónoma fuentes de datos en repositorios externos (Socrata/Datos Abiertos) y detectar actualizaciones en los sistemas internos.
- RF02 (Analítica y Revisión con IA): Integración de Modelos de Lenguaje de Gran Escala (LLM) para asistir en la auditoría de calidad y la generación de *insights* sobre la semántica de los registros.

- RF03 (Control de Calidad Humano): Flujo de control que requiere validación por parte de un analista antes de la promoción de datos hacia la capa *Silver*, garantizando la gobernanza.
- RF04 (Optimización de Limpieza): Aplicación de reglas de negocio automatizadas para la mitigación de duplicidades, normalización de formatos y corrección de errores de captura.
- RF05 (Escalabilidad): Infraestructura diseñada sobre *Databricks/Spark* para soportar el procesamiento de volúmenes masivos de datos sin degradación del rendimiento.
- RF06 (Monitoreo y Visualización): Integración de alarmas ante fallos en la ingesta y tableros de control para el seguimiento en tiempo real del estado de salud de la información.
- RF07 (Interoperabilidad): Diseño modular y flexible para conectar con diversas estructuras (APIs, archivos planos), asegurando una estandarización ágil de nuevos orígenes.

Este conjunto de requerimientos define un "Pipeline de Datos con Supervisión Inteligente". El diseño asegura que, aunque el sistema sea capaz de realizar descubrimientos automáticos mediante IA (RF01, RF02), la integridad de la información clínica se mantiene protegida bajo un flujo de aprobación humana (RF03), factor crítico para cumplir con los estándares de calidad del sistema de salud colombiano.

Diagnóstico de Calidad Inicial

Como cierre de la fase de diagnóstico, se ejecutó un proceso de *Data Profiling* sobre las 16 fuentes seleccionadas. Este análisis tuvo como objetivo cuantificar las brechas de información que impactan la analítica en salud pública. Los hallazgos se organizaron bajo cuatro ejes críticos que definen las reglas de negocio para la futura limpieza de datos:

Tabla 3

Ejes críticos de Calidad de Datos en Fuentes de Salud

Eje de Calidad	Descripción del Hallazgo	Impacto en la Toma de Decisiones
Completitud	Presencia recurrente de valores nulos en variables clave (Edad, Ubicación, Fechas).	Sesgo en indicadores epidemiológicos y subestimación de riesgos.
Duplicidad	Redundancia de registros por reportes paralelos (Entidades territoriales vs. Nacionales).	Inflación artificial de cifras y distorsión de la demanda de servicios.
Consistencia	Errores de captura (fechas lógicamente imposibles, códigos DIVIPOLA inexistentes).	Invalidez de las tendencias temporales y errores en la georreferenciación.
Trazabilidad	Ausencia de identificador único global (ID único) entre bases de datos del INS y MinSalud.	Imposibilidad de seguimiento longitudinal del paciente o correlación entre eventos de salud.

Estándares Definidos para el Prototipo

A partir de estos hallazgos, se establecieron los siguientes lineamientos técnicos que guían la ejecución del pipeline:

1. Cuantificación de Vacíos: Todo proceso de ingesta debe incluir un reporte de "calidad de carga" que determine el porcentaje de nulos. Si este supera el umbral definido [por ejemplo, >30%], el registro es marcado para revisión antes de su paso a la capa *Silver*.

2. Lógica de Desduplicación: Se implementó una función de llaves compuestas (*ID paciente + Código evento + Fecha*) para identificar y consolidar duplicados automáticamente durante la transformación.

3. Normalización DIVIPOLA: Se integró un catálogo maestro de códigos administrativos oficiales para validar y corregir automáticamente la georreferenciación.

4. Normalización de Entidades: Ante la falta de un ID único global, se diseñó un proceso de *Record Linkage* (vinculación de registros) que estandariza la identificación mediante un esquema de normalización que permite la interoperabilidad técnica entre las fuentes.

Este diagnóstico valida técnicamente la necesidad de implementar una arquitectura de datos que no solo almacene, sino que depure y sanee la información en su tránsito hacia la analítica. Los hallazgos descritos aquí actúan como los "filtros" de calidad que se aplicarán en los procesos ETL/ELT explicados en el siguiente apartado.

Segundo Resultado: Implementación de la Arquitectura

Teniendo en cuenta que se requiere contar con una arquitectura capaz de manejar grandes volúmenes de información la cual será procesada de manera heterogénea provenientes de múltiples fuentes públicas se plantea una arquitectura Big Data que permita:

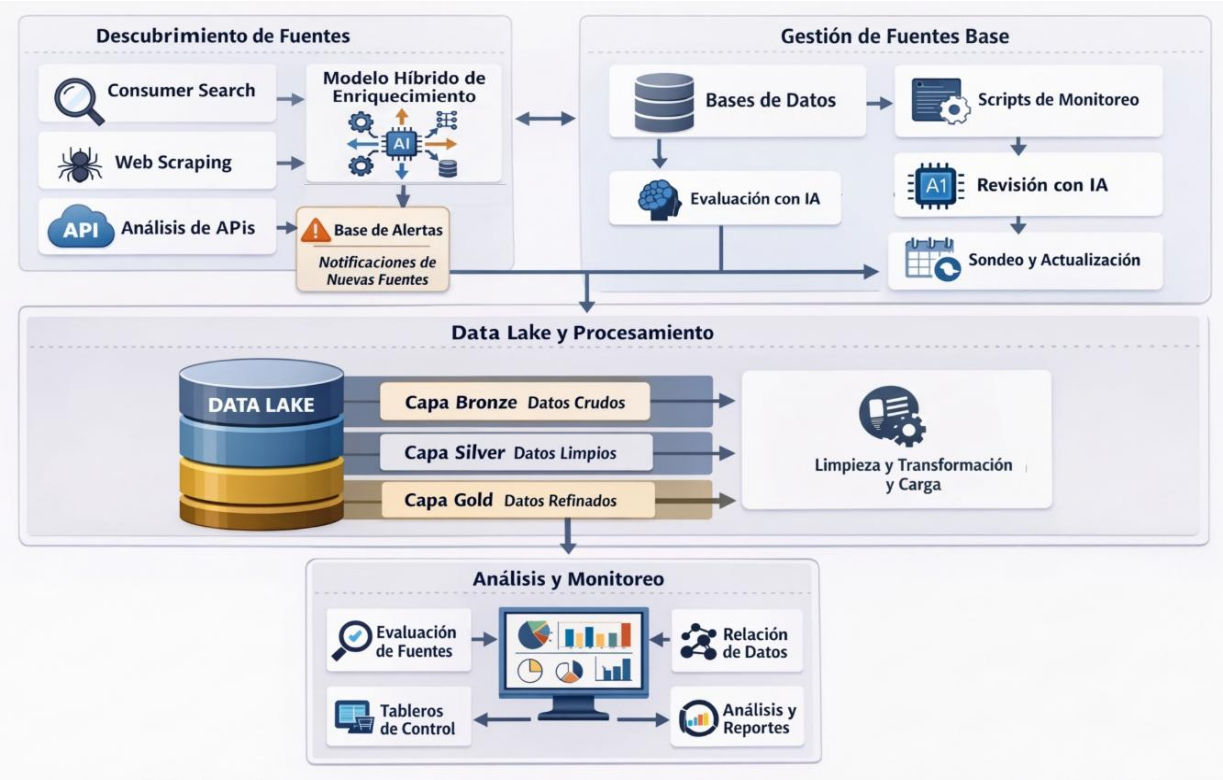
- Capturar datos desde diversas fuentes (APIs, repositorios abiertos, scraping).
- Almacenarlos en un entorno distribuido y flexible.
- Procesarlos bajo esquemas de transformación y enriquecimiento.
- Organizar la información en capas que soporten tanto la trazabilidad como el consumo analítico.

Arquitectura Propuesta

A continuación, se presenta la arquitectura diseñada para el prototipo:

Figura 1

Propuesta de Arquitectura Empleada



- Descubrimiento de fuentes de información: El proceso inicia con la identificación de fuentes relevantes mediante técnicas de búsqueda avanzada y web scraping, validando la existencia de APIs que faciliten la integración y aplicando modelos de inteligencia artificial para evaluar la pertinencia de los datos recolectados. Las fuentes clasificadas como útiles se registran en una base de alertas, cuyo propósito es notificar la disponibilidad de nuevas fuentes y mantener el sistema dinámico y actualizado.

- Gestión de fuentes base: Posteriormente, se consolida una base inicial a partir de repositorios de datos abiertos previamente caracterizados. En esta etapa se define un modelo preliminar de relacionamiento entre las distintas fuentes y se implementan scripts automatizados para monitorear cambios en estructura, contenido o disponibilidad. Los cambios detectados son evaluados mediante IA para determinar su impacto y relevancia, mientras que un sondeo periódico asegura la vigencia y calidad de las fuentes integradas.
- Construcción del Data Lake: Las fuentes validadas se integran en un Data Lake centralizado, organizado bajo una arquitectura de capas que permite diferenciar los niveles de procesamiento. La capa Bronze almacena los datos crudos tal como fueron capturados, la capa Silver contiene datos limpios y estructurados tras procesos de transformación, y la capa Gold concentra información lista para consumo analítico
- Monitoreo y evolución del modelo: Finalmente, la arquitectura contempla un sistema de monitoreo continuo que evalúa la relevancia de las fuentes integradas y la incorporación de nuevas fuentes que puedan mejorar la cobertura o calidad. Se desarrollan tableros de control que permiten analizar la calidad de los datos, las relaciones entre fuentes y el valor analítico generado, asegurando que la arquitectura evolucione de manera adaptativa frente a las necesidades cambiantes del sistema de salud

Propuestas de Herramientas

De manera inicial, se utiliza Google Colab junto con Python como entorno principal para la caracterización de fuentes y la ejecución de pruebas de código. A partir de esta base preliminar, se plantean diferentes alternativas de herramientas para cada etapa de la arquitectura. Para orientar la selección tecnológica en el marco del proyecto aplicado, se definieron criterios que aseguran la viabilidad, sostenibilidad y escalabilidad de la solución:

- Escalabilidad y rendimiento: Capacidad de la herramienta para crecer junto con el proyecto, soportando el incremento progresivo del volumen de datos sin comprometer eficiencia o desempeño. Este criterio es clave para evitar futuras migraciones o reemplazos.
- Facilidad de integración: Nivel de compatibilidad con otros sistemas y componentes tecnológicos. Una buena integración evita silos de información y garantiza un flujo continuo de datos dentro del ecosistema del proyecto.
- Costo y modelo de precios: Evaluación de la estructura de costos asociada al uso de la herramienta (almacenamiento, procesamiento y consultas). Permite valorar la sostenibilidad financiera y la relación inversión–beneficio.
- Facilidad de uso: Grado de accesibilidad e intuición para el equipo de trabajo. Una solución fácil de usar reduce la curva de aprendizaje y facilita la adopción, evitando depender de conocimientos altamente especializados.
- Seguridad y cumplimiento: Capacidad de la herramienta para proteger la información y cumplir con normativas de privacidad y manejo de datos. Es esencial cuando se trabaja con información sensible o regulada.
- Mantenimiento y operación: Esfuerzo requerido para administrar y sostener la herramienta en el tiempo. Se priorizan soluciones que reduzcan la carga operativa y permitan una gestión eficiente del sistema.
- Funcionalidades adicionales: Capacidades complementarias que aportan valor, como auditoría de datos, control de versiones o validación de calidad de la información, especialmente relevantes en etapas de crecimiento y madurez del sistema.

Con el fin de garantizar que el prototipo pueda identificar, evaluar y registrar nuevos repositorios de salud de manera autónoma, se realizó una comparativa técnica de las

herramientas disponibles. La selección se fundamentó en la capacidad de cada opción para operar en un entorno de Big Data y en su facilidad de integración con modelos de inteligencia artificial.

Tabla 4

Propuestas de Herramientas Identificadas

Categoría	Opciones Consideradas	Criterios de Selección	Capacidad de Integración
Descubrimiento de Fuentes	Google Search API, Bing Search API (SerpApi).	Automatización de búsqueda por palabras clave, límites de cuota y precisión de indexación.	Alta (vía Python)
Extracción y Rastreo	BeautifulSoup, Playwright, Scrapy, Apify.	Capacidad para navegar sitios dinámicos y escalabilidad de los rastreos en la nube.	Alta
Clasificación y Análisis (IA)	Google Gemini, OpenAI GPT-4o, Hugging Face.	Razonamiento contextual para determinar la relevancia de la	Excelente

		fuente y ventana de contexto.	
Gestión y Registro	MongoDB Atlas, Firebase, PostgreSQL.	Flexibilidad de esquemas NoSQL para metadatos variables y sistemas de alertas por triggers.	Alta
Procesamiento y Orquestación	Databricks, AWS Glue, Azure Data Factory.	Manejo de arquitectura Medallion, soporte Spark y capacidad de escalado elástico.	Líder (Core)

Selección de Componentes Base

Tras el análisis de los criterios de selección, se determinó que el flujo operativo de gestión de fuentes se basará principalmente en la combinación de *Databricks y Google Gemini:*

- Databricks: Se establece como el núcleo de la solución. Su capacidad para gestionar procesos de ingeniería de datos a gran escala permite que el descubrimiento de fuentes no sea un proceso aislado, sino una parte integrada del pipeline de datos. Esto facilita que la información pase rápidamente de la identificación a la limpieza (capa Silver) de manera eficiente.

- Google Gemini: Se integra como el motor de toma de decisiones. Gracias a su capacidad para procesar y clasificar grandes volúmenes de texto, actúa como un filtro inteligente que valida si las fuentes encontradas por las APIs de búsqueda son pertinentes para el sector salud, optimizando así el uso de recursos de almacenamiento y procesamiento.

Estructura de Componentes por Capas de Arquitectura

Teniendo en cuenta las capas de alto nivel de la arquitectura (descubrimiento de fuentes de información, gestión de fuentes base, construcción del Data Lake y monitoreo y evolución del modelo); se desarrolló la siguiente estructura a nivel de componentes, donde cada bloque cumple una función general dentro del sistema:

1. Descubrimiento de Fuentes de Información.

- 01_DESCUBRIMIENTO_EXTERNO: notebooks, scripts y resultados para búsqueda avanzada, validación de APIs y documentación del proceso.
- 08_ORQUESTACION_JOBS (Descubrimiento, Ingesta, IA): automatización de tareas de exploración y clasificación de fuentes.
- 13_MODELOS_IA (Clasificación_Fuentes, Evaluación_Gemini): modelos de inteligencia artificial para evaluar pertinencia y calidad.
- 14_REGISTROS_CENTRALES (Catálogo_Fuentes_Global): registro consolidado de fuentes descubiertas y su disponibilidad.

2. Gestión de Fuentes Base

- 02_GESTION_DE_FUENTES: catálogo, monitoreo, validación y alertas para consolidar la base inicial de fuentes.
- 07_GOBIERNO_Y_VALIDACION: reglas de negocio, contratos de datos y cumplimiento normativo.

- 09_COMPARTIDO: librerías, conectores y funciones comunes que soportan la gestión.
- 14_REGISTROS_CENTRALES (Versionado_Datasets, Histórico_Decisiones): trazabilidad de cambios y decisiones sobre las fuentes integradas.

3. Construcción del Data Lake

- 03_CAPA_BRONZE: almacenamiento de datos crudos, esquemas y particionamiento.
- 04_CAPA_SILVER: transformaciones, procesos de calidad y datos curados.
- 05_CAPA_GOLD: datamarts, analítica avanzada y modelos listos para consumo.
- 10_SQL: soporte transversal para consultas en todas las capas.
- 09_COMPARTIDO: utilidades y plantillas que facilitan la integración de procesos.

4. Monitoreo y Evolución del Modelo

- 06_OBSERVABILIDAD_Y_MONITOREO: métricas, alertas, linaje, auditoría y reportes para control continuo.
- 11_DASHBOARDS: tableros de control para alertas, monitoreo de fuentes, calidad de datos y analítica de salud.
- 12_DOCUMENTACION: arquitectura, metodología, manuales y evidencias para trazabilidad.
- 13_MODELOS_IA (Scoring_Modelos, Experimentos, Feature Engineering): evolución mediante experimentación y mejora continua.
- 14_REGISTROS_CENTRALES (Data Lineage, Versionado, Histórico): soporte al linaje y evolución adaptativa del sistema.

Todas las capas del modelo serán alimentadas de manera evolutiva, en función de la cantidad y diversidad de información que se suministre. Esto implica que cada componente se fortalecerá progresivamente a medida que nuevas fuentes sean integradas, validadas y transformadas dentro de la arquitectura.

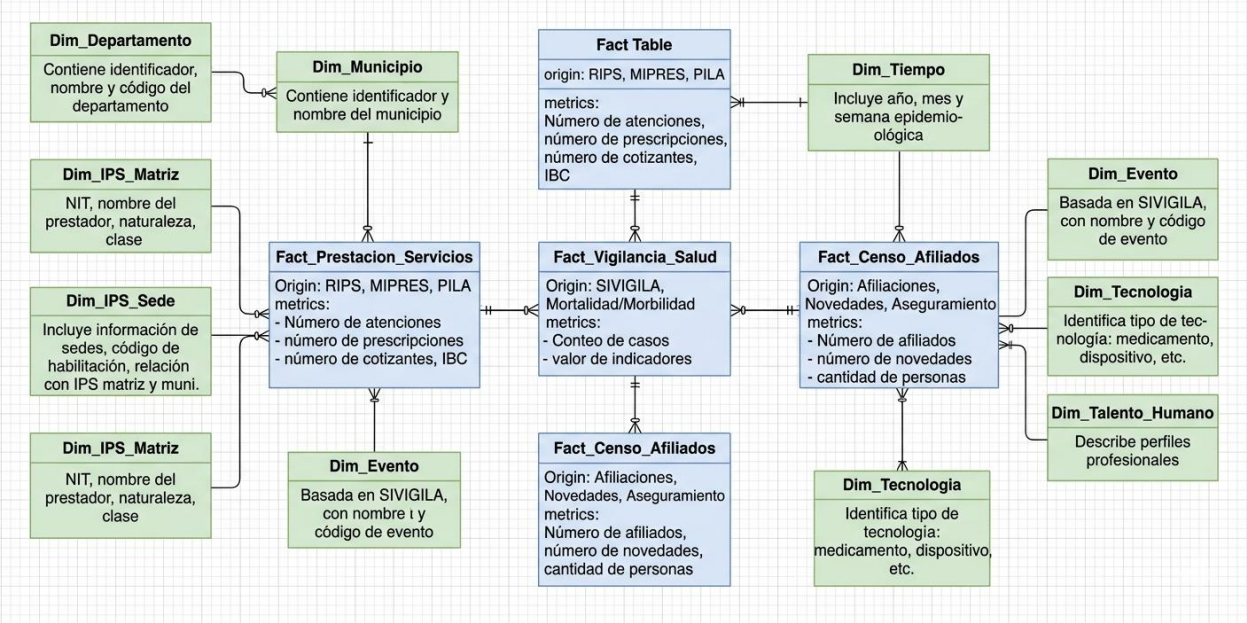
El Descubrimiento de fuentes de información se ampliará conforme se identifiquen y clasifiquen repositorios adicionales, enriqueciendo el catálogo global. La Gestión de fuentes base evolucionará con la incorporación de reglas de negocio, contratos de datos y procesos de validación que aseguren consistencia y trazabilidad. La Construcción del Data Lake crecerá en volumen y complejidad, pasando de datos crudos en la capa Bronze a información curada en Silver y analítica avanzada en Gold. Finalmente, el Monitoreo y evolución del modelo se adaptará continuamente mediante tableros de control, métricas de calidad y experimentación con modelos de inteligencia artificial, garantizando que la arquitectura responda de manera dinámica a las necesidades cambiantes del sistema.

Propuesta de Esquema del Modelo de Datos Inicial

Se plantea un modelo de datos que permita integrar de manera ordenada las 16 fuentes identificadas. La propuesta se basa en un Esquema de Copo de Nieve (Snowflake Schema), que busca normalizar las dimensiones y reducir redundancias. En este modelo, las tablas de hechos concentran las métricas principales (conteos, valores, número de afiliados), mientras que las dimensiones se descomponen en tablas relacionadas que permiten manejar jerarquías geográficas, administrativas y de prestación de servicios.

Figura 2

Esquema del Modelo de Datos Inicia



Esta propuesta se plantea como un primer diseño para organizar las 16 fuentes en un esquema de Copo de Nieve. La idea es normalizar las dimensiones y dividir las métricas en tablas de hechos, evitando redundancias y facilitando el análisis. Es importante aclarar que este modelo sigue siendo preliminar y está sujeto a revisión. Todavía debemos evaluar si trabajaremos con esta misma estructura o si será necesario ajustarla hacia otra más adecuada. También queda abierta la posibilidad de ampliar las relaciones entre tablas, aprovechando mejor la información disponible en las bases iniciales. Se aclara que este diseño es un bosquejo inicial que nos permite relacionar cómo integrar los datos, pero no constituye una decisión definitiva del prototipo.

Componentes del Modelo

1. Tablas de Hechos (Facts): Actúan como el núcleo del modelo y almacenan las métricas transaccionales de los tres pilares del proyecto: la prestación de servicios

(RIPS/MIPRES), la vigilancia epidemiológica (SIVIGILA) y el control de población (Censo de Afiliados). Estas tablas permiten realizar cálculos de volumen, frecuencia y costos de manera centralizada.

2. Dimensiones Normalizadas: A diferencia de un esquema en estrella, este modelo descompone las dimensiones complejas en subdimensiones. Por ejemplo, la dimensión de ubicación se divide en Dim_Municipio y Dim_Departamento, lo que elimina redundancias y asegura que los reportes territoriales sigan la codificación oficial DIVIPOLA sin inconsistencias de escritura.

3. Jerarquía de Prestadores: Se ha diseñado una relación entre la Dim_IPS_Sede y la Dim_IPS_Matriz. Esto permite realizar análisis granulares por punto de atención físico, o análisis agregados por entidad legal y naturaleza jurídica (Pública vs. Privada), facilitando la evaluación de la capacidad instalada del país.

Especificación de Procesos ETL/ELT

En esta fase se definen los procedimientos técnicos que garantizan la integridad, estandarización y trazabilidad de los datos dentro del Data Lake. Los procesos se organizan en tres etapas fundamentales, implementadas sobre Apache Spark y estructuradas en modelos analíticos:

1. Limpieza

- Implementación de algoritmos en Spark para la detección y corrección de duplicados, inconsistencias y registros faltantes.
- Uso de DataFrames y funciones distribuidas para asegurar eficiencia en grandes volúmenes de datos.

- Generación de datasets curados que alimentan la Capa Silver, garantizando calidad mínima antes de la transformación.

2. Transformación

- Normalización y estandarización de vocabularios clínicos mediante mapeos y diccionarios de referencia.
- Conversión de datos a esquemas analíticos (estrella, copo de nieve) que soportan consultas complejas y modelos de Machine Learning.
- Aplicación de funciones de Spark SQL y PySpark para homogenizar estructuras y preparar datasets para la Capa Gold.

3. Carga

- Definición de mecanismos de consolidación en la base central, con escritura en formatos optimizados (Parquet, Delta Lake).
- Inclusión de checkpoints y particionamiento para garantizar trazabilidad y eficiencia en consultas posteriores.
- Disponibilidad inmediata de los datos para procesos analíticos, dashboards y modelos de IA.

Arquitectura Medallón y Capas de Explotación Analítica

La arquitectura se organiza en capas que permiten un flujo progresivo de los datos, desde su estado crudo hasta su explotación analítica y uso en modelos de inteligencia artificial. Cada componente cumple una función específica dentro de este ciclo:

- 03_CAPA_BRONZE: Recibe los datos crudos tal como fueron capturados desde las fuentes originales. En esta capa se almacenan sin modificaciones, preservando su estructura inicial para garantizar trazabilidad y servir como punto de partida para procesos posteriores.

- 04_CAPA_SILVER: Almacena los datos curados tras procesos de limpieza y transformación. Aquí se eliminan duplicados, se corrigen inconsistencias y se normalizan vocabularios, generando datasets estructurados y listos para análisis intermedios.
- 05_CAPA_GOLD: Concentra los modelos analíticos y datasets finales preparados para consumo. Esta capa integra datamarts, información consolidada y estructuras optimizadas para dashboards, reportes y aplicaciones de inteligencia de negocio.
- 10_SQL: Habilita consultas transversales sobre todas las capas del Data Lake. Permite acceder a datos en Bronze, Silver y Gold mediante un lenguaje unificado, facilitando la explotación analítica y la integración con herramientas externas.
- 13_MODELOS_IA: Aprovecha los datos transformados y curados para entrenamiento, evaluación y experimentación con algoritmos de inteligencia artificial. Esta capa se nutre de la información disponible en Silver y Gold, potenciando la generación de modelos predictivos y de clasificación

Tercer Resultado: Validación y Visualización del Prototipo

El presente resultado valida la robustez técnica del artefacto desarrollado mediante un protocolo de pruebas exhaustivo aplicado a traves de la finalización de la Fase 4, diseñado para certificar que el prototipo no solo integra datos, sino que lo hace con niveles de rendimiento y confiabilidad aptos para el sector salud.

Pruebas Funcionales y de Desempeño

Las pruebas funcionales y de desempeño tienen como objetivo verificar que la arquitectura sea capaz de soportar consultas complejas y cargas masivas de datos clínicos, garantizando escalabilidad y tiempos de respuesta adecuados. Para ello, se articulan las siguientes capas:

- 03_CAPA_BRONZE: Se utilizan los datos crudos como insumo inicial para pruebas de ingestión y validación de volumen. Aquí se simulan escenarios de alta carga para comprobar la capacidad de almacenamiento y lectura distribuida.
- 04_CAPA_SILVER: Los datos curados permiten evaluar la eficiencia de los procesos de limpieza y transformación. Se realizan pruebas de consistencia y desempeño en consultas que requieren normalización y estandarización de vocabularios clínicos.
- 05_CAPA_GOLD: Los datamarts y datasets analíticos son sometidos a pruebas de rendimiento en consultas complejas, verificando que la capa soporte análisis epidemiológicos, estratificación de riesgo y modelos predictivos sin degradación significativa.
- 10_SQL: Se habilitan pruebas transversales de desempeño mediante consultas SQL sobre Bronze, Silver y Gold. Esto permite medir tiempos de respuesta y optimización de índices en escenarios de explotación analítica.
- 06_OBSERVABILIDAD_Y_MONITOREO: Registra métricas de desempeño, genera alertas y reportes sobre el comportamiento de la arquitectura bajo diferentes cargas. Esta capa asegura trazabilidad y evidencia de los resultados obtenidos.

Validación Estructural y Técnica

La validación estructural y técnica busca garantizar que el modelo de datos y la arquitectura mantengan coherencia, integridad y eficiencia en el manejo de la información clínica. Se centra en verificar la correcta organización de las tablas, los mecanismos de indexación y la consistencia de los esquemas definidos. Para ello, se articulan las siguientes capas:

- 03_CAPA_BRONZE: Se revisa la integridad de los datos crudos, asegurando que los esquemas iniciales capturen fielmente la estructura de las fuentes originales y que los checkpoints preserven trazabilidad:
- 04_CAPA_SILVER: Se valida la consistencia de los datos curados, comprobando que las transformaciones aplicadas mantengan integridad referencial y que los vocabularios clínicos normalizados se encuentren correctamente estandarizados.
- 05_CAPA_GOLD: Se evalúa la robustez de los datamarts y datasets analíticos, verificando que las estructuras soporten consultas complejas y que los modelos analíticos estén alineados con los requerimientos de explotación.
- 10_SQL: Se realizan pruebas de indexación y optimización de consultas transversales, garantizando que las tablas en Bronze, Silver y Gold respondan de manera eficiente y sin pérdida de integridad.
- 07_GOBIERNO_Y_VALIDACION: Se aplican reglas de negocio, contratos de datos y mecanismos de cumplimiento normativo, asegurando que la arquitectura cumpla con estándares técnicos y regulatorios.
- 06_OBSERVABILIDAD_Y_MONITOREO: Registra evidencias de validación, métricas de integridad y auditorías sobre los procesos, permitiendo trazabilidad completa de los resultados.

Diseño Conceptual de Visualización

El diseño conceptual de visualización busca demostrar cómo la base central puede aportar valor en escenarios clínicos y analíticos, mediante la construcción de dashboards demostrativos que integren datos curados y modelos de inteligencia artificial. Esta dimensión se articula con las siguientes capas:

- 05_CAPA_GOLD: Proporciona los datasets analíticos y datamarts que sirven como insumo principal para la construcción de tableros. Aquí se concentran los datos listos para consumo, estructurados en modelos que facilitan la estratificación de riesgo y el análisis epidemiológico.
- 11_DASHBOARDS: Es la capa dedicada a la visualización. Se diseñan tableros que muestran indicadores de calidad de datos, métricas de desempeño y resultados analíticos. Los dashboards permiten evidenciar cómo la información puede apoyar la toma de decisiones clínicas y de gestión.
- 13_MODELOS_IA: Aporta capacidades predictivas y de clasificación que se integran en los dashboards. Los modelos de IA enriquecen la visualización con recomendaciones, análisis de tendencias y escenarios de riesgo, potenciando el soporte a la decisión clínica.
- 06_OBSERVABILIDAD_Y_MONITOREO: Complementa la visualización con métricas de desempeño, auditoría y linaje de datos. Los reportes generados permiten contextualizar los resultados mostrados en los dashboards y asegurar trazabilidad.
- 12_DOCUMENTACION: Registra la propuesta conceptual, incluyendo diagramas, metodología y evidencias que sustentan el diseño de los tableros. Esta capa asegura que el proceso de visualización esté documentado y pueda ser replicado o validado.

Conclusiones

Tras el desarrollo y validación del prototipo de arquitectura de datos para el sector salud, se presentan las siguientes conclusiones:

Sostenibilidad y Escalabilidad del Artefacto: El prototipo desarrollado se consolida como una arquitectura de Big Data de alta adaptabilidad. Su capacidad para evolucionar progresivamente ante la creciente diversidad de fuentes públicas garantiza que el sistema no sea una solución estática, sino una infraestructura flexible capaz de responder a las exigencias cambiantes del sistema de salud colombiano.

Fundamentación Técnica mediante Caracterización: La fase de caracterización técnica de fuentes abiertas resultó ser el pilar del proyecto. Este proceso permitió transitar de una revisión documental superficial a una comprensión profunda de la heterogeneidad y complejidad de los datos públicos, asegurando que la arquitectura final se construyera sobre bases sólidas, veraces y representativas del ecosistema sanitario nacional.

Gobernanza y Calidad como Activo Estratégico: La implementación de protocolos de ingesta, limpieza y normalización no solo asegura la consistencia técnica, sino que eleva el dato de salud a la categoría de activo estratégico. Se logró establecer un marco de trabajo que garantiza la integridad y trazabilidad, transformando registros fragmentados en insumos analíticos confiables para la toma de decisiones clínicas y de gestión pública.

Viabilidad del Modelo Medallón para la Interoperabilidad: La consolidación de la información en un *Data Lake* organizado por capas (Bronze, Silver, Gold) demostró ser una solución altamente eficiente. Este diseño garantiza no solo la trazabilidad total del dato, sino también la optimización de tiempos en consultas complejas, permitiendo la disponibilidad

inmediata de información para explotación analítica, superando así la barrera histórica de la fragmentación tecnológica.

Recomendaciones

A partir de los hallazgos del estudio y la naturaleza de las fuentes de información pública, se presentan las siguientes recomendaciones para futuras líneas de investigación y desarrollo:

Dado que las fuentes de salud pública en Colombia reportan mayoritariamente cifras agregadas, se recomienda fortalecer los procesos de análisis de consistencia temporal y espacial. Esto implica diseñar mecanismos que validen la coherencia entre niveles de agregación (por ejemplo, asegurar que la suma de datos municipales coincida con el reporte departamental), garantizando que las agregaciones no pierdan integridad al ser integradas en el *Data Lake*.

Incorporar progresivamente nuevas fuentes de datos abiertos y repositorios especializados, asegurando la interoperabilidad entre datasets con diferentes granularidades. Esto permitirá que el prototipo no solo consolide cifras, sino que cree índices sintéticos que combinen indicadores de distintas fuentes.

Ajustar los mecanismos de limpieza para que sean dinámicos y capaces de manejar *schemas* variables. Al trabajar con agregaciones, es fundamental implementar reglas adaptativas que detecten cambios en los métodos de reporte de las entidades gubernamentales, reduciendo la intervención manual ante actualizaciones en los portales de origen.

Consolidar un entorno para la experimentación de modelos de *Machine Learning* diseñados para series temporales y datos de panel. Dado que las fuentes públicas permiten analizar tendencias históricas, se recomienda el uso de modelos que aprovechen la naturaleza agregada de los datos para predecir brotes epidemiológicos o proyecciones de demanda asistencial a nivel territorial.

Diseñar flujos automatizados de *Machine Learning Operations* (MLOps) que permitan el reentrenamiento de los modelos a medida que se liberan nuevas actualizaciones en los portales públicos.

Dada la complejidad de trabajar con agregaciones de diversas fuentes, es imperativo reforzar el registro del linaje de datos. Esto permitirá vincular cada indicador analítico final con la fuente, la fecha de corte y el nivel de agregación original, garantizando la transparencia en la toma de decisiones.

Desarrollar tableros avanzados que permitan la navegación jerárquica (*drill-down*), desde el nivel nacional hasta el nivel municipal. La visualización debe facilitar la interpretación del impacto de las políticas públicas basándose en la evolución temporal de los indicadores agregados.

Ajustar la arquitectura para soportar consultas analíticas sobre grandes volúmenes de indicadores históricos, optimizando los motores de procesamiento para que las consultas de agregación sean eficientes incluso cuando el histórico de datos crezca exponencialmente.

Bibliografía

Alarcón García, R. E. (2021). *Sistema analítico basado en un modelo predictivo de procesamiento de datos en la big data en la educación superior*. 1. https://dialnet.unirioja.es/servlet/tesis?codigo=369557&info=resumen&idioma=SPA

Bhuyan, S. S., Sateesh, V., Mukul, N., Galvankar, A., Mahmood, A., Nauman, M., Rai, A., Bordoloi, K., Basu, U., & Samuel, J. (2025). Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. *Journal of Medical Systems*, *49*(1). https://doi.org/10.1007/S10916-024-02136-1

Castano, R., Prada, S. I., Maldonado, N., & Soto, V. (2024). Managed competition in Colombia: Convergence of public and private insurance and delivery. *Health Economics, Policy and Law*. https://doi.org/10.1017/S1744133123000348

Felipe, A., Copete, L., Rodríguez Martínez, L., & Ramírez Gómez, D. A. (2023). Aplicación de big data en sistemas de salud pública. *Publicaciones e Investigación*, *17*(1). https://doi.org/10.22490/25394088.6446

García Espinosa, S., & Figueroa Alvarado, G. B. (2021). Big Data: Análisis del Transporte en la Ciudad Actual: Percepción de la población en el contexto de la pandemia. *Congreso Internacional de Investigación Academia Journals*, *13*(3), 293–398. https://openurl-ebsco-com.bibliotecavirtual.unad.edu.co/contentitem/fap:161269106?sid=ebsco:plink:crawler&id=ebsco:fap:161269106&crl=c

García Hernández, H., & Esquer Bojorquez, D. (2024). Análisis comparativo de los sistemas de salud de México y Colombia. *Población y Salud en Mesoamérica, ISSN-e 1659-0201, Vol. 21, Nº. 2 (enero-junio), 2024*, *21*(2), 11. https://doi.org/10.15517/psm.v21i2.54151

Hernández-Arango, A., Arias, M. I., Pérez, V., Chavarría, L. D., & Jaimes, F. (2025). Prediction of the Risk of Adverse Clinical Outcomes with Machine Learning Techniques in Patients with Noncommunicable Diseases. *Journal of Medical Systems*, *49*(1). https://doi.org/10.1007/S10916-025-02140-Z

Malla Valdiviezo, R. O., López Gorozabel, O., Arévalo Indio, J. A., & Tóala Briones, C. H. (2023). Mecanismos para el procesamiento de big data: Limpieza, transformación y análisis de Datos. *Polo del Conocimiento: Revista científico - profesional, ISSN-e 2550-682X, Vol. 8, Nº. 4 (ABRIL 2023), 2023, págs. 656-675*, *8*(4), 656–675. https://dialnet.unirioja.es/servlet/citart?info=link&codigo=9152273&orden=0

Mejia, J., Oviedo Benálcazar, M. A., Ordoñez, J. A., & Valencia, J. F. (2023). Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud. *Facultad Nacional de Salud Pública: El escenario para la salud pública desde la ciencia, ISSN-e 2256-3334, ISSN 0120-386X, Vol. 41, Nº. 2, 2023*, *41*(2), 3. https://doi.org/10.17533/udea.rfnsp.e351168

Myriam, B., & Guerrero, C. (2023). Aportes para la transformación del sistema de salud colombiano. *Facultad Nacional de Salud Pública: El escenario para la salud pública desde la ciencia, ISSN-e 2256-3334, ISSN 0120-386X, Vol. 41, Nº. 1, 2023*, *41*(1), 7. https://doi.org/10.17533/udea.rfnsp.e348269

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Rodríguez-Páez, F. G., Cabrera-Moya, D., & Herrera-Cuartas, J. A. (2024). Proposal of a Knowledge Management Model for Complex Systems: Case of the Supervision and Control

Subsystem of the Colombian Health System. *Journal of Market Access and Health Policy*, *12*(3), 224–251. https://doi.org/10.3390/JMAHP12030019/S1

Sepúlveda Correa, D., Montaña Vásquez, J. R., & Vargas Guette, M. L. (2021). Comparación de los modelos de Atención Primaria en Salud desde un enfoque sanitario en Colombia y sus países fronterizos. *Movimiento Científico, ISSN-e 2011-7191, Vol. 15, Nº. 1, 2021, págs. 42-54*, *15*(1), 42–54. https://doi.org/10.33881/2011-7197.mct.15107

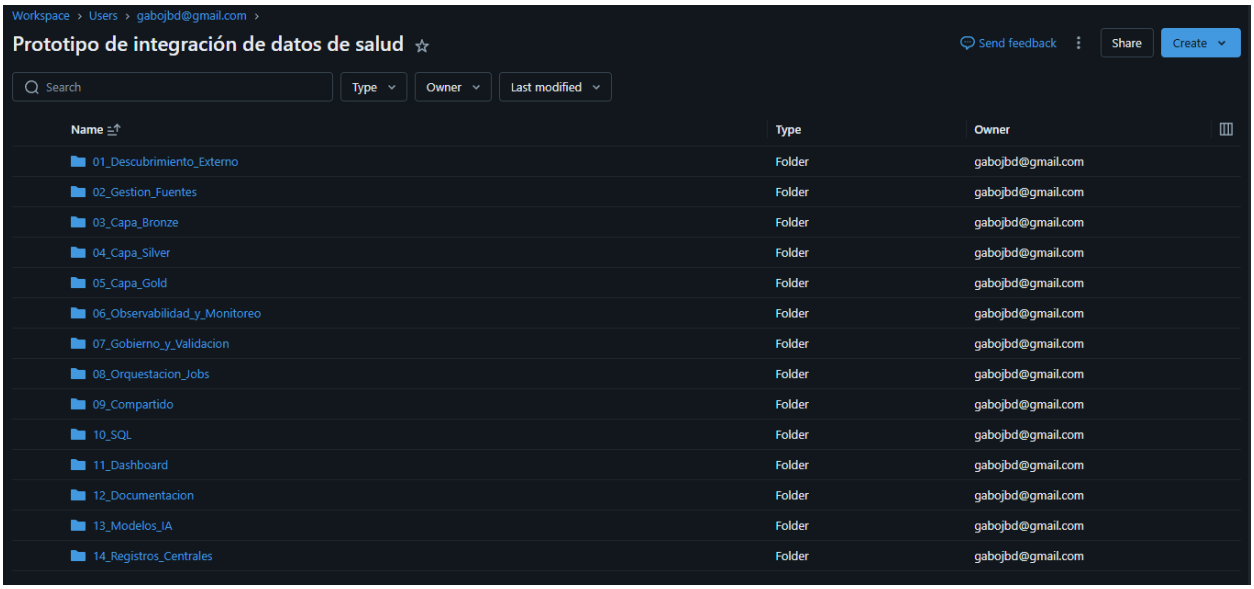
Urdinola, P., Bejarano, V., Espinosa, O., Do, P. L., Silva, N., Urdinola -Valeria, P., & Do, B.-O. E.-P. L. (2023). Estudio de caracterización socioeconómica y demográficade los afiliados al régimen subsidiado de salud, Colombia 2019-2020. *Ensayos de economía, ISSN 0121-117X, Vol. 33, Nº. 63 (julio-diciembre 2023), 2023, págs. 85-116*, *33*(63), 85–116. https://doi.org/10.15446/ede.v33n63.104918

Vargas, I., Mogollón-Pérez, A. S., De Paepe, P., Ferreira Da Silva, M. R., Unger, J. P., & Vázquez, M. L. (2016). Barriers to healthcare coordination in market-based and decentralized public health systems: a qualitative study in healthcare networks of Colombia and Brazil. *Health Policy and Planning*, *31*(6), 736–748. https://doi.org/10.1093/HEAPOL/CZV126

Apéndices

Apéndice A

Prototipo de Integración de Datos de Salud



Nota. Prototipo de integración de datos: enlace compartido en Databricks con permisos de acceso configurados (visualización y gestión). El archivo contiene notebooks y ficheros asociados al desarrollo del proyecto.