INFORME DE ACTIVIDADES

Plataforma de Analítica de Datos del Observatorio de Salud

Dirección Territorial de Salud de Caldas

Centro de Bioinformática y Biología Computacional de Colombia – BIOS

Octubre de 2026

1. Resumen Ejecutivo

La Dirección Territorial de Salud de Caldas (DTSC), en convenio con el Centro de Bioinformática y Biología Computacional de Colombia (BIOS), desarrolla un Observatorio de Salud Pública que reúne en un solo lugar las fuentes de información sanitaria, demográfica, socioeconómica y educativa del departamento. Cada fuente llega en su propio formato y con sus propias inconsistencias, por lo que el proyecto construye un Data Lakehouse que las lleva, paso a paso, a un formato común, limpio, trazable y seudonimizado, sin modificar nunca los datos originales. El fin último es habilitar tableros de control, vigilancia epidemiológica y, en una fase posterior, modelos de AutoML.

El objetivo de este informe es dar a conocer a la DTSC y a BIOS la arquitectura, el procedimiento de procesamiento, el estado de las fuentes integradas y los resultados de calidad del Data Lakehouse, así como los pendientes que condicionan su cierre. Su alcance comprende las once fuentes con pipeline construido: prestación de servicios de salud (RIPS), afiliación (BDUA), caracterización de familias (Equipos Básicos), registros vitales (RUAF-ND), matrícula escolar (SIMAT), clasificación socioeconómica (SISBÉN IV), vigilancia nutricional (SISVAN) y estadísticas y cartografía del DANE. Quedan fuera de este alcance las doce fuentes que cuentan con datos crudos pero aún sin pipeline, y la construcción de tableros y de modelos de AutoML.

2. Contexto y arquitectura

El Data Lakehouse de la Dirección Territorial de Salud de Caldas se construye sobre Delta Lake, un formato abierto de almacenamiento que registra cada cambio realizado a las tablas. Esto garantiza que las escrituras se apliquen de forma completa, permite consultar versiones anteriores de los datos («viaje en el tiempo»), admite la evolución controlada de los esquemas y es compatible con distintas herramientas de análisis, como Polars, DuckDB y Spark.

La organización de los datos sigue la arquitectura Medallion, que los distribuye en tres capas de refinamiento progresivo: Bronze, Silver y Gold. Cada capa cumple una función específica y es la base de la siguiente, de modo que ningún dato llega al modelo analítico sin haber pasado por la ingesta fiel y por la validación (Tabla 1).

La Figura 1 presenta la arquitectura general de la solución. En ella se observa el recorrido de la información desde las fuentes de datos hasta su uso analítico, a través de las tres capas descritas. También se identifican la cuarentena, donde se aíslan los registros que incumplen alguna regla de calidad, y los diccionarios, anexos y catálogos oficiales que definen dichas reglas. En la parte inferior se muestran los controles transversales que aplican a todas las fuentes.

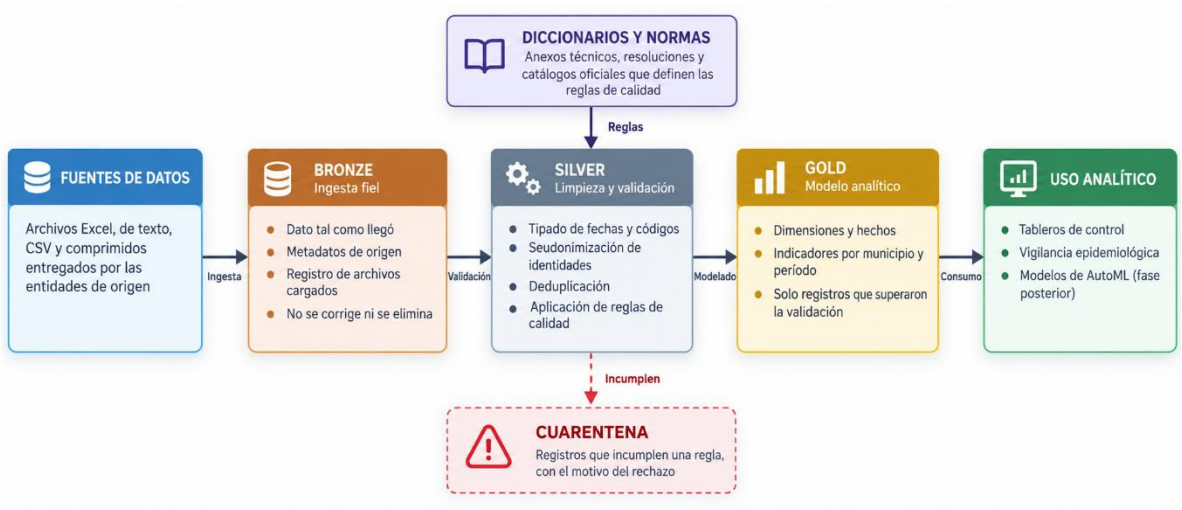


Figura 1. Arquitectura del Data Lakehouse DTSC Caldas

Capa	Contenido	Principio rector
Bronze	Dato tal como llegó, con metadatos de linaje (archivo, lote, fecha de carga)	Nunca se corrige ni se elimina: es copia fiel de la fuente
Silver	Dato tipado, validado contra catálogos oficiales y seudonimizado	Lo que incumple una regla no se borra: pasa a una tabla de cuarentena con el motivo
Gold	Modelo dimensional: dimensiones, hechos e indicadores agregados	Utiliza únicamente registros que superaron la validación de Silver

Tabla 1. Definición de capas en el Delta Lake

Decisiones técnicas relevantes

- **Motor:** deltalake (delta-rs) con Polars y DuckDB, sin clúster Spark, dado que el volumen por fuente (de cientos de miles a decenas de millones de filas) es manejable en un solo equipo.
- **Memoria acotada:** lectura por lotes, ejecución de cada etapa (y de cada archivo en Bronze) en un proceso independiente, y un vigilante basado en psutil que detiene el proceso al superar el tope definido.
- **Patrón uniforme:** cada fuente se implementa con cuatro módulos (schema, bronce, silver, gold), un orquestador y su documentación propia; la lógica transversal (E/S Delta, linaje, hash, calidad, DIVIPOLA).

- **Seudonimización consistente:** el hash HMAC-SHA256 de TIPO|NÚMERO, calculado con la llave DTSC_HASH_KEY, es idéntico entre fuentes, lo que permite cruzar a una misma persona sin exponer su documento.
- **Reglas con origen declarado:** cada regla de calidad indica si proviene de un diccionario oficial, de una norma (Res. 3374 o Res. 2275), de una fuente de referencia del proyecto o de un criterio propio pendiente de validación.

Gobernanza. No se modifica una fuente construida por otro integrante sin su aprobación; toda construcción se documenta con cifras reales; las pruebas automáticas emplean exclusivamente datos ficticios; y todo plan se presenta y aprueba antes de escribir código.

3. Procesamiento paso a paso

El procesamiento de cada fuente sigue la misma secuencia de diecinueve pasos, agrupados en seis fases. Cada etapa se ejecuta de forma independiente y bajo un límite de memoria definido, lo que permite reanudar el trabajo en caso de interrupción.

3.1 Preparación

- **Configuración del entorno.** Se define la llave secreta que se usa para proteger las identidades, obligatoria en producción, y se fija el límite de memoria de cada ejecución (1,5 GB por defecto, ajustable).
- **Recepción de los datos crudos.** Los archivos originales se depositan sin modificarlos en el repositorio de datos de cada fuente, junto con los diccionarios y anexos técnicos que sirven de referencia.

3.2 Capa Bronze: ingesta fiel

- **Clasificación de archivos.** Cada archivo se dirige a la tabla que le corresponde según su nombre. Se excluyen los reportes agregados, el período se obtiene de la carpeta o del nombre y, cuando el proveedor entregó el mismo mes dos veces, se conserva la versión comprimida y se descarta la repetida.
- **Lectura y detección del diseño.** Los libros de Excel se leen con un motor de alto rendimiento y todo el contenido se trata como texto, para evitar interpretaciones erróneas de tipo. Los archivos de texto se leen por lotes con su codificación original. El diseño de columnas se identifica contando los campos de la cabecera o de la primera línea, y los registros que el formulario partió en varias líneas se reconstruyen.
- **Normalización de columnas.** Cada diseño se lleva a un esquema común. Las columnas sobrantes se conservan en un campo auxiliar y las faltantes quedan vacías. No se descarta información con contenido, solo las filas totalmente vacías.
- **Registro con trazabilidad.** Los datos se agregan a la capa Bronze junto con sus metadatos de origen: archivo, lote, fecha de carga, firma del registro, período y diseño. Cada archivo cargado se anota en un manifiesto que impide ingerirlo dos veces.

- **Verificación de la ingesta.** El número de filas cargadas se contrasta con un conteo independiente de cada archivo fuente.

3.3 Definición de las reglas de calidad

Cada regla de validación se deriva de la documentación oficial disponible para la fuente y se registra con su origen declarado, de modo que sea posible auditarla y que nunca se introduzca un umbral sin indicarlo.

- **Identificación del diccionario de referencia.** Se determina qué documento oficial describe la estructura de la fuente (lineamiento técnico, anexo técnico, resolución o diccionario del proveedor) y se verifica que corresponda a la versión de los datos recibidos, no solo al tema.
- **Extracción y clasificación de las condiciones.** De ese documento se recogen, campo por campo, la obligatoriedad, el formato y la longitud, los catálogos de valores admitidos, los rangos y las dependencias entre campos. Cada condición se asigna a una dimensión del modelo de gobernanza: completitud, validez, consistencia o integridad.
- **Asignación del origen.** Cada regla declara su procedencia: el diccionario o anexo técnico de la fuente, una norma complementaria que amplía los catálogos, las reglas heredadas de las fuentes de referencia del proyecto, o un criterio propio (con su base legal cuando existe) pendiente de validación con el referente funcional.
- **Decisión del efecto y contraste con los datos.** Si el diccionario respalda la condición y los datos la sustentan, la regla es obligatoria y separa los registros incumplidos hacia cuarentena. Si el catálogo no está documentado, o la regla bloquearía una proporción excesiva de registros clínicamente válidos, se mantiene como informativa: se reporta, pero no separa registros. Antes de darla por definitiva se mide cuántos registros incumple.
- **Registro de la regla.** Cada regla queda documentada con su identificador, dimensión y origen, y el resultado de cada corrida (registros evaluados, fallos y porcentaje) se guarda en el reporte de calidad.

3.3 Capa Silver: limpieza, protección y validación

- **Limpieza y tipado.** Se eliminan caracteres ocultos, espacios y sufijos generados por Excel; se normalizan los códigos; las fechas se convierten a su tipo, con ajuste a la hora de Colombia cuando corresponde; los valores que indican "no aplica" pasan a vacío; y se calculan variables derivadas, como la edad a la fecha de referencia de cada registro.
- **Seudonimización.** A partir del tipo y número de documento se genera un identificador anónimo mediante un cifrado irreversible con llave secreta. Solo después se eliminan nombres, apellidos, documentos originales, direcciones, teléfonos y textos libres, de modo que ningún dato personal directo llega a esta capa.
- **Deduplicación.** Se aplica el criterio propio de cada fuente: por identificador único del registro, por la combinación de persona y fecha cuando debe conservarse la historia (SISVAN), o únicamente sobre filas idénticas entre archivos que se traslapan (RIPS). En BDUA no se deduplica entre períodos, porque cada mes es un censo completo.

- **Aplicación de las reglas y cuarentena.** Se evalúan las reglas definidas en la fase anterior: las obligatorias separan registros y las informativas solo se reportan. Los registros que incumplen alguna regla no se eliminan: pasan a una tabla de inconsistencias con el detalle de las reglas incumplidas.

3.4 Capa Gold: modelo analítico

- **Modelado dimensional.** Con los registros válidos se reduce el conjunto de columnas a las de interés analítico y se construyen las dimensiones, las tablas de hechos y los indicadores agregados por municipio, período, diagnóstico u otra categoría. Las tablas se organizan por año o por municipio, según la fuente, para optimizar las consultas.

3.5 Cierre y control

- **Verificación final.** Se ejecutan pruebas automáticas con datos ficticios, se contrastan los resultados con una reimplementación o un recálculo independiente, se alteran deliberadamente detalles del código para confirmar que las pruebas los detectan, se comprueba que ningún documento original aparezca en las capas Silver y Gold, y se confirma que dos corridas con los mismos datos producen resultados idénticos.
- **Documentación.** Se registra por fuente el diseño, las reglas con su origen, las cifras reales, los hallazgos y los pendientes.

Operación y reanudación. La ejecución puede realizarse completa o por etapas, y es posible modificar el límite de memoria o reingerir archivos ya cargados. Si el proceso se interrumpe, se retoma desde el punto en que quedó, porque la capa Bronze omite los archivos con manifiesto completo. Dado que el formato conserva las versiones anteriores de cada tabla, se recomienda depurarlas periódicamente en Silver y Gold para recuperar espacio de almacenamiento.

4. Estado de las Fuentes

Diez fuentes están concluidas (Bronze, Silver, Gold, pruebas y documentación); BDUA aguarda la ejecución de Silver y Gold sobre datos reales, y otras 12 cuentan con datos crudos pero sin pipeline.

Fuente	Contenido	Estado
rips	Prestación de servicios de salud (7 tipos de atención, 19 bases)	Concluida
bdua	Afiliación a EPS, régimen contributivo y subsidiado (54 archivos mensuales)	Bronze verificado; Silver y Gold pendientes
equipos_basicos	Caracterización de familia y entorno, 15 municipios (21 bases)	Concluida
ruaf_nd_nacimientos, ruaf_nd_defunciones	Registros vitales del RUAF	Concluida

Fuente	Contenido	Estado
simat_matricula	Matrícula escolar	Concluida
sisben	Clasificación socioeconómica SISBÉN IV (256 columnas de origen)	Concluida
sisvan	Vigilancia nutricional (gestantes, menores y mayores de 5 años)	Concluida
dane_nacimientos, dane_defunciones, dane_cartografia	Estadísticas vitales y cartografía del DANE	Concluida

Tabla 2. Descripción de las fuentes con su estado

Sin pipeline: ámbito laboral, discapacidad, geohelmintiasis, gestantes, habitante calle, ICBF, indígenas, migrantes, participación social, PIC, SIVIGILA y vacunación.

5. Calidad, Verificación y Hallazgos Transversales

El proyecto aplica a todas las fuentes un mismo esquema de verificación en seis niveles, independiente del código del pipeline, de modo que la confiabilidad de cada tabla pueda demostrarse y no solo afirmarse.

- **Conteos contra la fuente.** Las filas cargadas en Bronze se contrastan con un conteo independiente del contenido de cada archivo original (hojas, líneas o archivos comprimidos), archivo por archivo, para confirmar que no se perdió ni se duplicó ningún registro durante la ingesta.
- **Reimplementación independiente.** Las reglas y transformaciones de Silver se reescriben a partir de la documentación, sin reutilizar el código del pipeline, y se comparan con el resultado oficial: deben coincidir en los registros válidos, en los registros en cuarentena, en sus motivos y en el identificador anónimo de cada persona. De igual forma, las dimensiones e indicadores de Gold se recalculan con otra herramienta y deben ser idénticos.
- **Pruebas de mutación.** Se alteran deliberadamente detalles del código (umbrales, operadores, zonas horarias, órdenes de prioridad) para comprobar que las pruebas automáticas detectan cada alteración. Esto demuestra que las pruebas realmente protegen la lógica y no solo la ejecutan.
- **Pruebas automáticas.** Cada fuente cuenta con un conjunto de pruebas que cubre la lectura, la ingesta, cada regla de calidad con casos que deben pasar y casos que deben fallar, los casos límite y la construcción de las tablas finales. Siempre se emplean datos ficticios y nunca datos reales de personas.

- **Protección de la privacidad.** Se verifica de forma sistemática que ningún número de documento original presente en Bronze aparezca en las capas Silver o Gold, y que los datos personales directos hayan sido eliminados tras generar el identificador anónimo.
- **Determinismo.** Dos corridas consecutivas con los mismos datos deben producir resultados idénticos en cada tabla, condición necesaria para la reproducibilidad y la auditoría.

6. Glosario

Término	Descripción
ACID (transacciones)	Garantía de que cada escritura sobre una tabla se aplica completa o no se aplica, de modo que nunca queda una tabla a medias.
AutoML	Construcción automatizada de modelos predictivos, prevista para una fase posterior del proyecto.
BDUA	Base de Datos Única de Afiliados al sistema de salud.
Cuarentena	Tabla que conserva, junto con el motivo, los registros que incumplen una regla obligatoria, en lugar de eliminarlos.
Data Lakehouse	Almacén de datos que combina la flexibilidad de guardar archivos de cualquier formato con el orden y las garantías de una base de datos.
Deduplicación	Eliminación de registros repetidos según un criterio definido para cada fuente.
Delta Lake	Formato abierto de tablas, construido sobre archivos Parquet, que registra cada cambio realizado y puede ser leído por distintas herramientas de análisis.
Determinismo	Propiedad por la cual dos corridas con los mismos datos producen resultados idénticos.
Dimensiones, hechos e indicadores	Tablas del modelo analítico (Gold): las dimensiones describen entidades, como viviendas o usuarios; los hechos registran eventos, como consultas; y los indicadores son resultados agregados, por ejemplo por municipio y período.
DIVIPOLA	Catálogo oficial de códigos de departamentos y municipios del DANE.
Equipos Básicos de Salud	Equipos de atención primaria que caracterizan a las familias y su entorno en los municipios.
Linaje	Información que indica de qué archivo y en qué fecha se cargó cada dato.
Manifiesto	Registro de los archivos ya cargados, que impide ingerir dos veces el mismo.
Medallion (arquitectura)	Organización de los datos en capas de refinamiento progresivo, descritas en la Tabla 1.
Orquestador	Componente que ejecuta en orden las etapas de procesamiento de una fuente.

Término	Descripción
Pipeline	Secuencia automatizada de pasos que lleva los datos desde la fuente hasta el modelo analítico.
Polars, DuckDB y Spark	Herramientas de procesamiento de datos; las dos primeras operan en un solo equipo y Spark distribuye el trabajo entre varios.
Reglas obligatorias e informativas	Las obligatorias excluyen del análisis el registro que las incumple; las informativas solo se reportan.
Resoluciones 3374 de 2000 y 2275 de 2023	Normas de referencia de RIPS: la primera define el lineamiento técnico de registro y envío, y la segunda aporta las tablas oficiales de catálogos.
RIPS	Registro Individual de Prestación de Servicios de Salud, reportado por las instituciones prestadoras.
RUAF-ND	Registro Único de Afiliados, componente de nacimientos y defunciones.
Seudonimización	Reemplazo del documento de identidad por un código anónimo irreversible (HMAC-SHA256 con llave secreta), que permite cruzar datos de una misma persona sin exponerla.
SIMAT	Sistema Integrado de Matrícula Estudiantil.
SISBÉN IV	Sistema de Identificación de Potenciales Beneficiarios de programas sociales, cuarta versión.
SISVAN	Sistema de Vigilancia Alimentaria y Nutricional.
Viaje en el tiempo	Consulta de una tabla tal como estaba en una fecha o versión anterior, útil para auditoría.

Tabla 3. Glosario de términos