

Proyecto Curso

Herramientas de Computación en la Nube

Documento Base

Febrero 7 de 2026

por

Alvaro Mauricio Montenegro Díaz

CEO, Aprendizaje Profundo S.A.S

Documento académico ficticio para fines de formación en arquitectura de datos y analítica en la nube. Todos los datos, cifras y escenarios aquí descritos son simulados.

CASO DE ESTUDIO: Marco Lógico, Operativo y Analítico para la Gestión Multicanal de Solicitudes Ciudadanas (PQRS)

1. Introducción y contexto institucional

Una organización pública nacional de carácter administrativo, con cobertura en más de 900 municipios, atiende de manera permanente solicitudes ciudadanas clasificadas como Peticiones, Quejas, Reclamos y Sugerencias (PQRS).

Durante los últimos cinco años, la organización ha experimentado un crecimiento sostenido en el volumen de interacciones ciudadanas, impulsado por la digitalización de canales, la ampliación de cobertura territorial y la mayor exigencia normativa en materia de transparencia y tiempos de respuesta.

Actualmente, la gestión de PQRS constituye un proceso crítico, tanto desde el punto de vista operativo como desde la perspectiva de control, auditoría y toma de decisiones institucionales.

2. Diagnóstico de la situación actual

2.1. Volumen y crecimiento de solicitudes

Con base en estimaciones internas:

- Promedio histórico anual: 420.000 PQRS
- Promedio mensual: 35.000 PQRS
- Promedio diario: 1.150 PQRS
- Crecimiento anual estimado: 12 %

Durante eventos coyunturales (campañas institucionales, cambios normativos, crisis sectoriales), se han registrado picos de hasta:

- 3.000 – 4.000 PQRS/día, durante períodos de 5 a 10 días consecutivos.

2.2. Distribución por canal de ingreso (estimada)

Canal	Porcentaje
Correo electrónico	38 %
Formulario web	31 %
Mensajería instantánea	14 %
Atención telefónica (transcripción)	11 %
Otros canales digitalizados	6 %

Cada canal presenta estructuras de datos diferentes, niveles variables de calidad y distintos patrones temporales de llegada.

3. Problemas identificados

El análisis del proceso actual (simulado) evidencia los siguientes problemas estructurales:

3.1. Fragmentación de la información

- Los datos se almacenan en sistemas heterogéneos sin integración efectiva.
- No existe una vista consolidada y trazable del ciclo de vida de una PQRS.

3.2 Dificultad para el control de tiempos

- El cálculo del tiempo real de atención requiere consolidar múltiples fuentes.

- Se presentan discrepancias entre reportes operativos y reportes de control.

3.3 Limitaciones analíticas

- Los reportes se generan de forma manual o semiautomática.
- No es posible reconstruir escenarios históricos de manera confiable.
- La identificación de cuellos de botella es reactiva y tardía.

3.4 Riesgos de cumplimiento

- La ausencia de trazabilidad completa dificulta auditorías posteriores.
- No se cuenta con mecanismos robustos para justificar el cumplimiento de SLA ante entes de control (simulados).

4. Marco normativo y acuerdos de nivel de servicio (SLA)

De manera ficticia, la organización opera bajo los siguientes compromisos de atención:

Tipo de PQRS	SLA objetivo
Peticiones	15 días hábiles
Quejas	10 días hábiles
Reclamos	8 días hábiles
Sugerencias	20 días hábiles

Incumplimientos reiterados generan:

- alertas internas,
- planes de mejora,
- reportes de seguimiento trimestrales.

5. Objetivo general del caso de estudio

Diseñar y validar una arquitectura analítica moderna, capaz de:

- Integrar PQRS multicanal.

- Garantizar trazabilidad completa de cada caso.
- Calcular indicadores operativos y de cumplimiento.
- Permitir simulación de escenarios de carga.
- Producir información reproducible y auditável.

El alcance se limita explícitamente a una prueba de concepto (PoC).

6. **Modelo operativo del proceso PQRS (simplificado)**

Cada solicitud recorre las siguientes etapas:

1. Recepción
2. Radicación
3. Clasificación y priorización
4. Asignación
5. Gestión
6. Respuesta
7. Cierre
8. Archivo y evaluación

Cada transición genera un evento con timestamp, permitiendo reconstruir la historia completa del ticket.

7. **Modelo conceptual de datos (alto nivel)**

7.1.1. Entidades principales

Ticket

- Identificador único
- Canal de origen
- Tipo PQRS
- Prioridad

- Estado actual
- Fecha de radicación
- Ubicación geográfica (municipio / región)
- SLA asignado

Mensaje

- Ticket asociado
- Fecha/hora
- Rol (ciudadano / agente)
- Texto del mensaje
- Longitud del contenido

Evento de estado

- Ticket asociado
- Estado anterior
- Estado nuevo
- Timestamp
- Usuario/rol que genera el evento (simulado)

Adjunto

- Ticket asociado
- Tipo de archivo
- Tamaño
- Canal de origen

8. Indicadores clave de desempeño (KPIs)

8.1. Indicadores operativos

- PQRS recibidas por día y canal

- Backlog activo por tipo y región
- Tiempo promedio de primera respuesta
- Tiempo promedio de cierre
- Distribución de PQRS por prioridad

8.2 Indicadores de cumplimiento

- Porcentaje de PQRS atendidas dentro del SLA
- Número de PQRS vencidas por período
- Tiempo promedio de retraso en casos vencidos

8.3 Indicadores analíticos avanzados (opcionales)

- Tendencias semanales y mensuales
- Identificación de picos anómalos
- Distribución geográfica de carga operativa
- Análisis de flujos de estado (OPEN → IN_PROGRESS → CLOSED)

9. Enfoque arquitectónico propuesto (conceptual)

Se adopta un enfoque Lakehouse, con separación clara de responsabilidades:

9.1. Capa de datos crudos (Raw)

- Almacenamiento inmutable de eventos originales.
- Preservación del formato original.
- Versionado temporal.

9.2. Capa Bronze

- Normalización ligera.
- Conversión a formatos analíticos (Parquet).

- Particionamiento por fuente y fecha.

9.3. Capa Silver

- Datos curados, consistentes y deduplicados.
- Esquema estable.
- Preparados para análisis.

9.4. Capa Gold

- Productos analíticos:
 - KPIs
 - agregados
 - cubos de consulta

9.5. Capa de metadatos

- Registro de ejecuciones.
- Control de calidad.
- Trazabilidad de datos.

10. Simulación de datos para la PoC

Con fines académicos, se simulará:

- Llegada diaria de PQRS durante períodos de 30, 90 y 180 días.
- Variación controlada de volumen (+/- 30 %).
- Distribución realista por canal y tipo.
- Eventos de estado con duraciones variables.
- Incumplimientos de SLA en un 7–12 % de los casos.

La simulación será determinista, permitiendo reproducibilidad exacta.

11. Alcance y limitaciones de la prueba de concepto

11.1. Incluye

- Pipeline completo de datos.
- Cálculo de indicadores.
- Visualización analítica.
- Registro de trazabilidad.

11.2. Excluye

- Interfaces ciudadanas reales.
- Automatización de respuestas.
- Integración con sistemas productivos.
- Modelos avanzados de inteligencia artificial.

12. Consideraciones éticas y de uso

- No se utilizan datos reales.
- No se simulan personas reales.
- El caso no representa a ninguna entidad específica.
- Uso exclusivo académico y demostrativo.

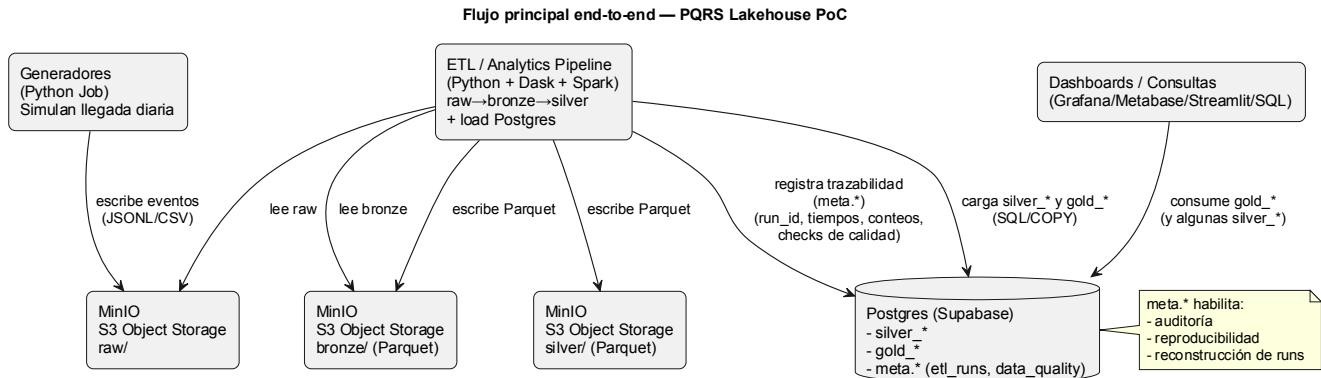
13. Valor pedagógico del caso

Este caso permite:

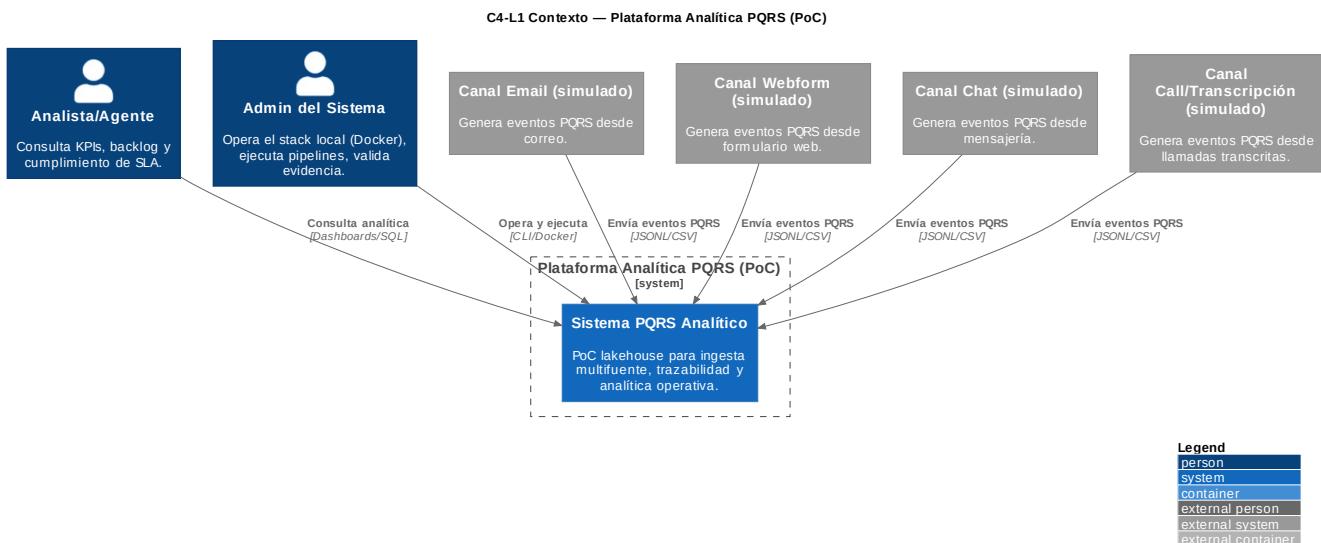
- Comparar estructuras de datos (lake, relacional, documental, grafo).
- Enseñar decisiones arquitectónicas justificadas.
- Evaluar reproducibilidad y operación.
- Simular escenarios de estrés y auditoría.

14. Arquitectura Mínima Inicial

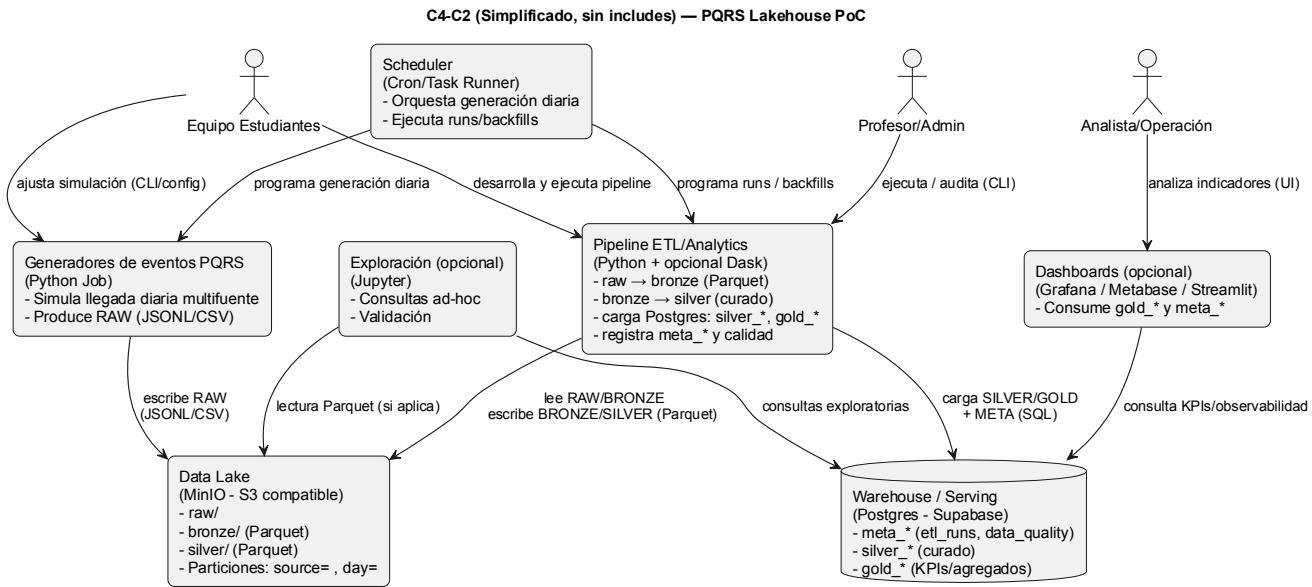
14.1. Flujo Principal



14.2. Diagrama C4-L1 - Contexto



14.3. C4-C2 Diagrama de Contenedores



15. Nota final

Este documento no busca describir un sistema real, sino modelar un problema realista para el aprendizaje de arquitectura de datos y analítica aplicada en la nube.

16. Anexo Diccionario Técnico

A. CONCEPTOS GENERALES DE ARQUITECTURA

16.1. Arquitectura Lakehouse

Modelo de arquitectura de datos que combina las ventajas del **data lake** (flexibilidad, bajo costo, datos crudos) con las del **data warehouse** (consistencia, rendimiento analítico, modelos estructurados). Se organiza en capas lógicas (Raw, Bronze, Silver, Gold) y permite trazabilidad y reproducibilidad.

16.2. Prueba de Concepto (PoC)

Implementación limitada y controlada cuyo objetivo es **demostrar viabilidad técnica**, decisiones arquitectónicas y capacidad analítica, sin pretender cubrir todos los requerimientos de un sistema en producción.

16.3. Pipeline de Datos

Conjunto de procesos automatizados que permiten **ingestar, transformar, validar y almacenar datos** desde su origen hasta su consumo analítico.

16.4. Trazabilidad

Capacidad de **reconstruir el recorrido completo de los datos**, desde su fuente original hasta los productos analíticos finales, incluyendo transformaciones, versiones y ejecuciones del pipeline.

16.5. Reproducibilidad

Propiedad de un sistema de datos que permite **repetir exactamente los mismos resultados** a partir de los mismos insumos, versiones de código y parámetros de ejecución.

B. CAPAS DEL MODELO DE DATOS (LAKEHOUSE)

16.6. Capa Raw (Datos Crudos)

Definición:

Zona del data lake que almacena los datos **exactamente como fueron recibidos**, sin transformación estructural.

Características:

- Inmutable
- Versionada por fecha
- Formato original (JSONL, CSV, TXT)
- Alta granularidad

Ejemplos en el caso PQRS:

- Correos electrónicos completos
- Formularios web sin normalizar
- Transcripciones simuladas
- Eventos originales de estado

Propósito:

- Auditoría
- Reprocesamiento
- Evidencia histórica

16.7. Capa Bronze (Normalización ligera)

Definición:

Zona del data lake donde los datos crudos son **limpiados mínimamente** y convertidos a formatos analíticos.

Características:

- Conversión a Parquet
- Tipos de datos normalizados
- Nombres de campos estandarizados
- Particionamiento Hive-style

Ejemplos:

- bronze/pqrs/source=email/day=2026-02-01/* .parquet

Propósito:

- Preparación para procesamiento analítico
- Reducción de ambigüedad estructural
- Conservación de granularidad

16.8. Capa Silver (Datos curados)

Definición:

Zona donde los datos alcanzan un **nivel de calidad y consistencia suficiente** para análisis confiable.

Características:

- Esquema estable
- Duplicados controlados
- Reglas de calidad aplicadas
- Claves y relaciones definidas

Ejemplos en PQRS:

- Tickets consolidados
- Mensajes normalizados
- Eventos de estado estructurados

Propósito:

- Base para métricas
- Integración entre fuentes
- Consistencia semántica

16.9. Capa Gold (Productos analíticos)

Definición:

Capa que contiene **indicadores, métricas y agregados**, optimizados para consumo por dashboards, reportes y toma de decisiones.

Características:

- Datos agregados
- Alto rendimiento de consulta
- Orientada a negocio/operación

Ejemplos:

- KPIs diarios de backlog
- Cumplimiento de SLA
- Volumen por canal y región

Propósito:

- Analítica operativa
- Reporting
- Visualización

C. ESTRUCTURAS DE DATOS Y MODELOS

16.10. Ticket

Unidad central del sistema PQRS que representa una solicitud ciudadana individual.

16.11. Mensaje

Registro textual asociado a un ticket, proveniente del ciudadano o del agente de atención.

16.12. Evento de Estado

Registro que captura una **transición de estado** de un ticket (por ejemplo, de “abierto” a “en gestión”), con timestamp.

16.13. SLA (Service Level Agreement)

Compromiso temporal máximo para atender una PQRS, definido según su tipo y prioridad.

16.14. Backlog

Conjunto de tickets que se encuentran **abiertos o en gestión** en un momento dado.

16.15. KPIs (Key Performance Indicators)

Indicadores cuantitativos que permiten evaluar el desempeño operativo y el cumplimiento de objetivos.

D. PARTICIONAMIENTO Y ALMACENAMIENTO

16.16. Hive-style Partitioning

Convención de organización de archivos en un data lake que utiliza rutas del tipo `clave=valor`, permitiendo poda eficiente de datos.

Ejemplo:

`bronze/pqrs/source=email/day=2026-02-01/`

16.17. Parquet

Formato de archivo columnar optimizado para análisis, ampliamente utilizado en data lakes por su eficiencia de lectura y compresión.

16.18. Object Storage

Modelo de almacenamiento basado en objetos (archivos), ideal para data lakes. Compatible con APIs S3.

E. METADATOS Y GOBERNANZA

16.19. Metadatos

Información que describe los datos: origen, fecha, versión, calidad, responsable y contexto de uso.

16.20. etl_runs

Registro estructurado de cada ejecución del pipeline de datos, incluyendo tiempos, volumen procesado y estado.

16.21. Calidad de Datos

Conjunto de reglas y métricas que evalúan la validez, completitud y consistencia de los datos.

16.22. Auditoría

Capacidad de demostrar, a posteriori, **cómo y cuándo** se procesaron los datos y qué resultados se produjeron.

F. ANALÍTICA Y OBSERVABILIDAD

16.23. Analítica Operativa

Uso de datos para monitorear procesos diarios y apoyar decisiones tácticas (ej. backlog, tiempos, carga).

16.24. Analítica Histórica

Análisis de tendencias, patrones y evolución temporal de los indicadores.

16.25. Observabilidad del Pipeline

Capacidad de medir y entender el comportamiento del pipeline de datos mediante métricas, logs y eventos.

G. SIMULACIÓN DE DATOS

16.26. Datos Sintéticos

Datos generados artificialmente que imitan características estadísticas y estructurales de datos reales, sin contener información personal.

16.27. Simulación Determinista

Simulación que utiliza semillas y reglas fijas para garantizar que los mismos insumos produzcan los mismos resultados.

16.28. Evento

Unidad mínima de información que representa una ocurrencia puntual en el sistema (recepción, cambio de estado, mensaje).

H. ALCANCE ACADÉMICO

16.29. Caso Apócrifo

Escenario ficticio diseñado para enseñanza, inspirado en múltiples realidades, sin representar a ninguna entidad real.

16.30. Evaluación Arquitectónica

Proceso de valoración centrado en la **calidad de las decisiones de diseño**, no en la cantidad de tecnologías utilizadas.

17. FRASES CANÓNICAS

- *Raw conserva la verdad; Gold entrega valor.*
- *Parquet vive en el lake, no en el warehouse.*
- *La trazabilidad es más importante que la herramienta.*
- *Si no es reproducible, no es ingeniería.*