UAX

UNIVERSIDAD ALFONSO X EL SABIO

El Caso: práctica sobre casos reales de BIG DATA

PLATAFORMAS TECNOLÓGICAS DE BIG DATA EN LA NUBE

FinTech: detección de fraude en tiempo real

Diseño de Arquitectura Híbrida

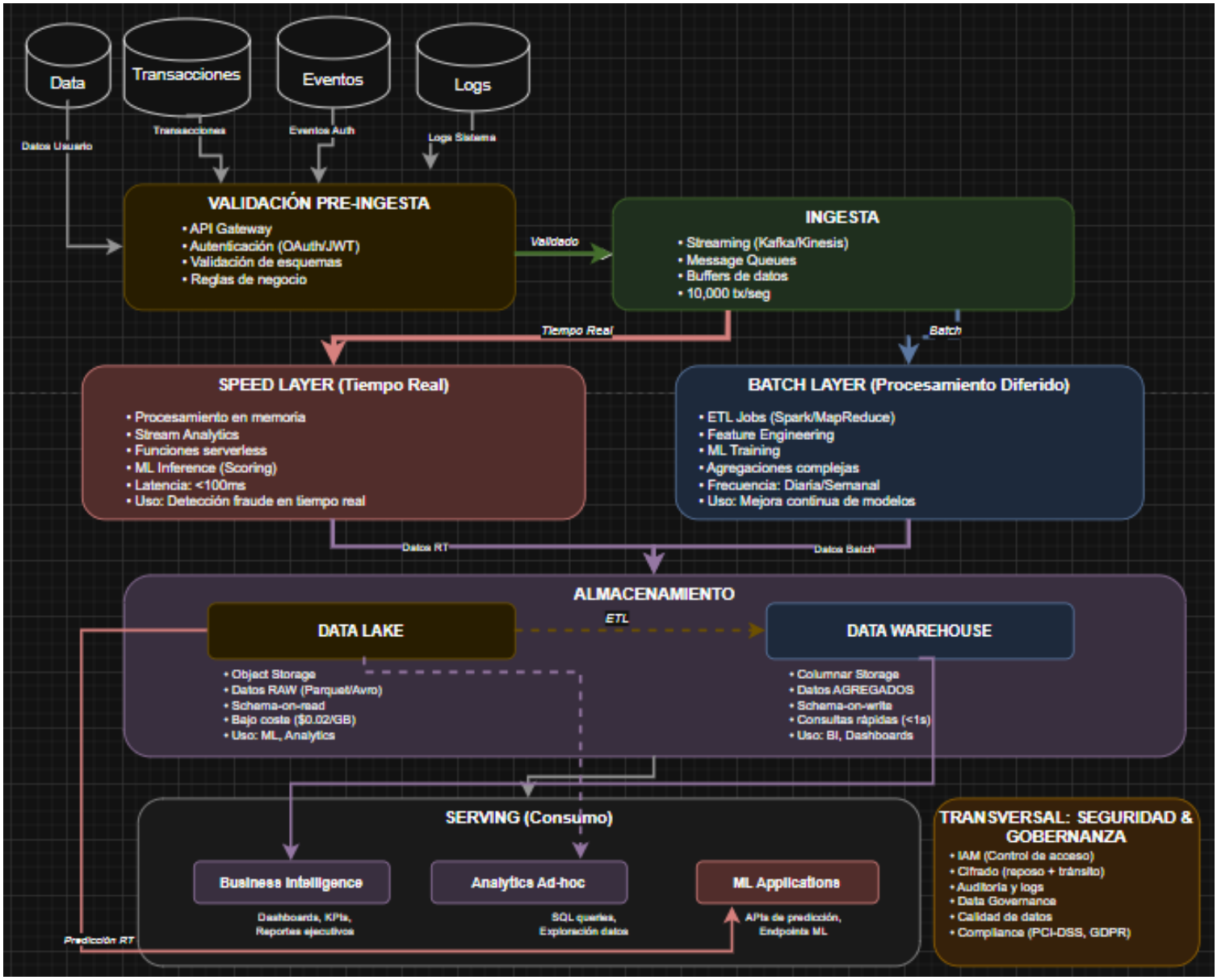
Big Data y Despliegue en AWS

Autor: José Luis Cendán Guzmán

Fecha: 17/02/2026

Contenido

Arquitectura.

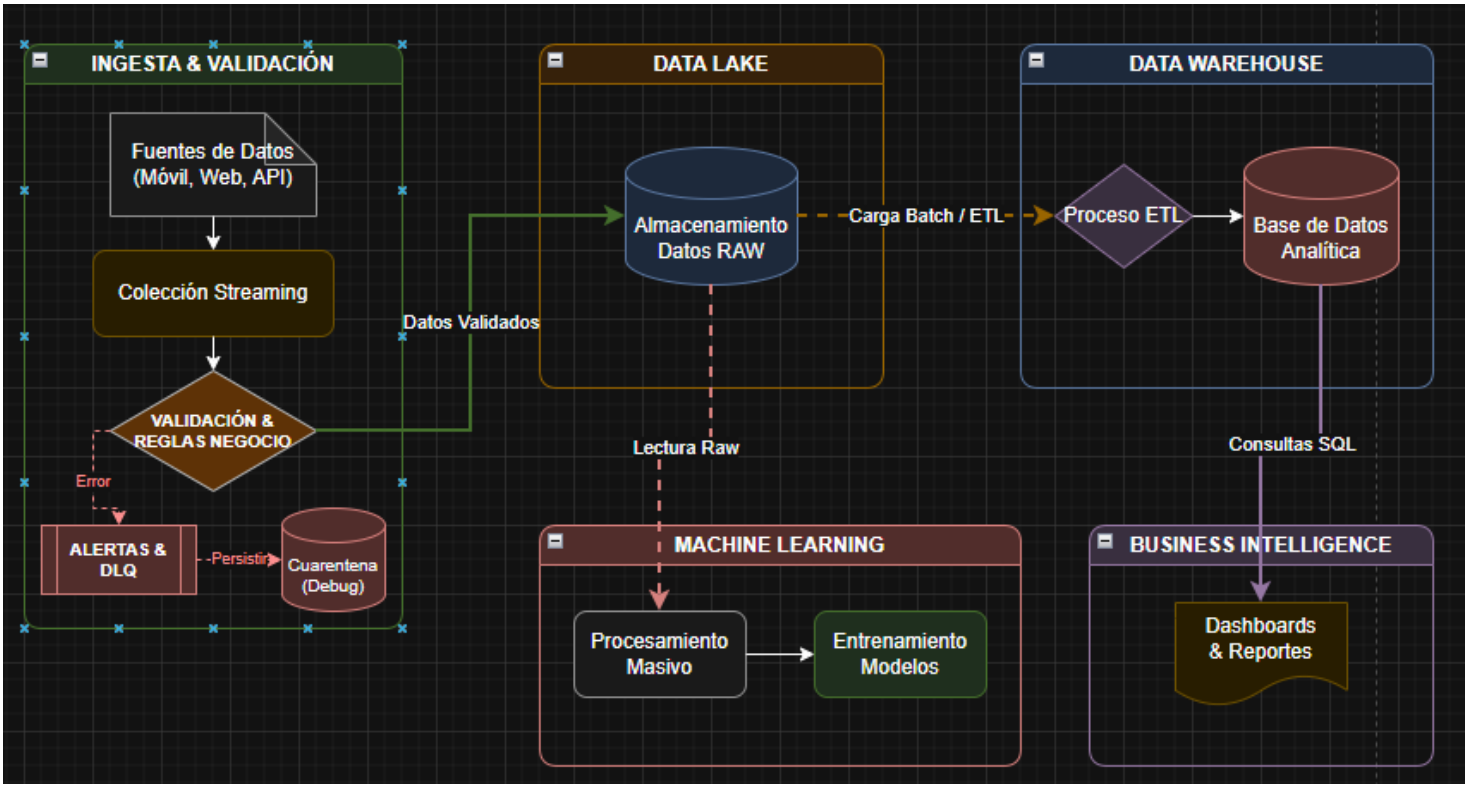


La arquitectura propuesta implementa el patrón **Lambda** para conciliar dos requisitos opuestos: **baja latencia** (<100ms) para bloquear fraudes en tiempo real y **alta precisión** basada en análisis histórico (7 años). A diferencia de enfoques puros como **Kappa** (cuya complejidad técnica se dispara al intentar reprocesar 50TB de histórico como un stream continuo) o **Solo Batch** (cuya latencia intrínseca impide bloquear transacciones en <100ms), este modelo híbrido asigna la detección inmediata a una **Speed Layer** y el reentrenamiento exhaustivo a una **Batch Layer**. Esto garantiza prevención instantánea, mejora continua de modelos y cumplimiento regulatorio sin comprometer el rendimiento.

El diseño prioriza la **Disponibilidad y Tolerancia a Fallos (AP)** del Teorema CAP en la ingesta para asegurar la captura de datos, manteniendo Consistencia eventual en el almacenamiento.

- **Escalabilidad (Horizontal y Vertical):** La arquitectura está diseñada para escalar sin límites teóricos mediante diversas estrategias. La ingesta se distribuye horizontalmente a través del particionamiento (sharding) del stream de datos, mientras que la computación serverless se ajusta automáticamente a la carga de transacciones concurrentes. El almacenamiento de objetos proporciona capacidad ilimitada y desacoplada del cómputo, complementado por un Data Warehouse híbrido con arquitectura MPP que escala ambos recursos de forma independiente. Finalmente, la elasticidad se aplica mediante **auto-scaling** en los recursos de procesamiento batch para optimizar costes operativos.
- **Disponibilidad (SLA 99.99%):** Para garantizar una alta disponibilidad, se implementa un despliegue multi-zona con todos los servicios críticos replicados en múltiples ubicaciones físicas, respaldado por mecanismos de replicación automática de datos para asegurar su durabilidad. El uso de componentes de procesamiento **stateless** permite un **failover** transparente en caso de incidencias, todo ello reforzado con políticas de backup automático y objetivos de recuperación (RTO/RPO) de minutos.
- **Resiliencia y Tolerancia a Fallos:** La resiliencia del sistema se logra mediante una estrategia de defensa en profundidad que incluye buffers de ingesta para retener datos temporalmente ante picos o fallos de consumidores, y mecanismos de degradación gradual para que el sistema siga operando con reglas deterministas simples si falla el modelo ML avanzado (sin downtime). Se garantiza un procesamiento **at-least-once** para evitar la pérdida de datos, gestionando duplicados en la capa de consumo, y se aplican patrones de **Circuit Breakers** para prevenir fallos en cascada entre los distintos componentes de la arquitectura.

Enfoque Data Lake vs Data Warehouse



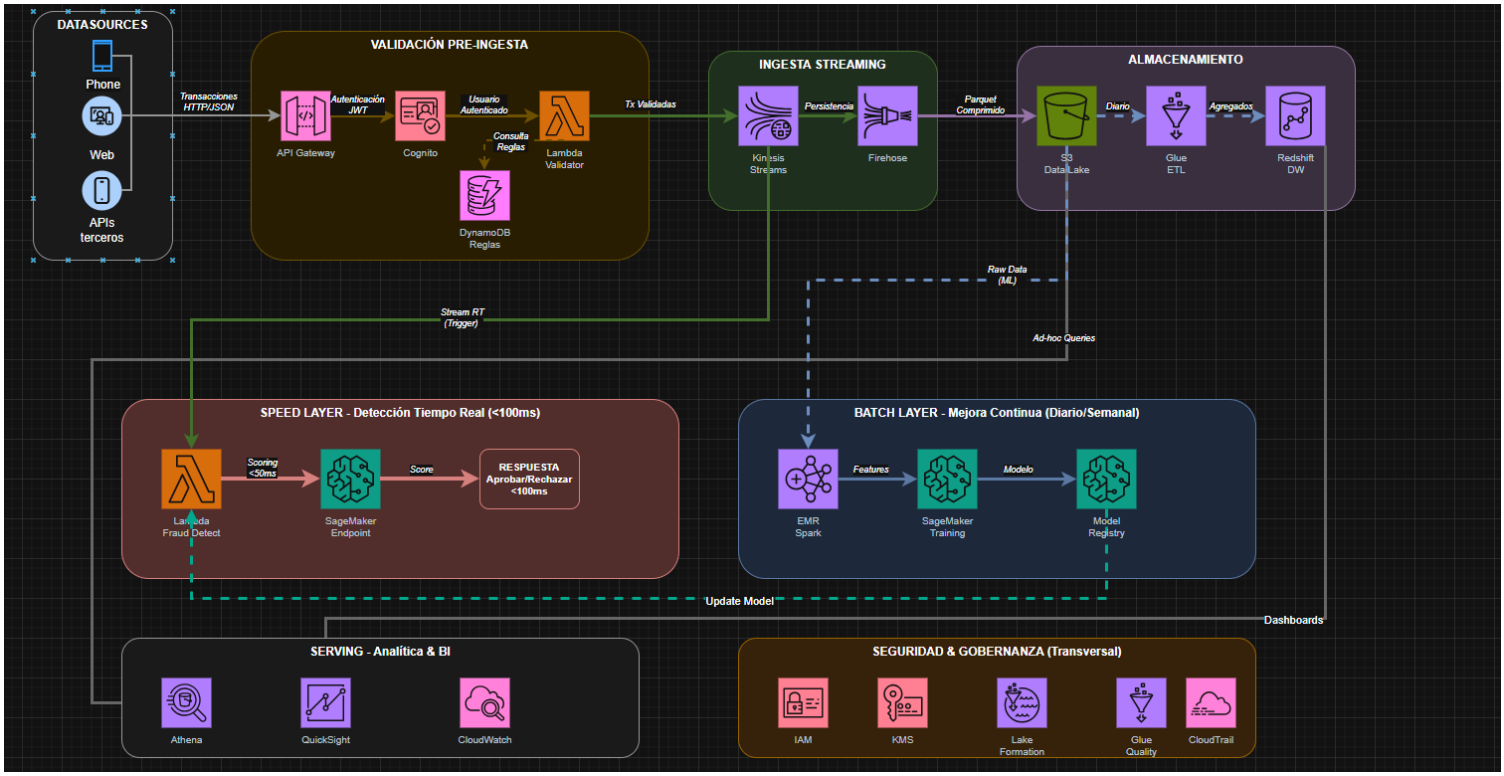
Para la estrategia de almacenamiento y análisis, se ha seleccionado un **enfoque híbrido** que combina las fortalezas de un Data Lake y un Data Warehouse, evitando las limitaciones de elegir uno solo. Esta decisión se fundamenta en la necesidad de gestionar datos crudos masivos para Machine Learning y, simultáneamente, ofrecer consultas estructuradas de alto rendimiento para Business Intelligence.

La elección híbrida se considera la más adecuada porque permite separar cargas de trabajo con requisitos opuestos. Por un lado, la **Ingesta y Validación** gestiona datos en streaming que pasan inmediatamente por un filtro de reglas de negocio: los datos válidos continúan hacia el almacenamiento, mientras que los inválidos se desvían a una cola de "cuarentena" (Dead Letter Queue) para su posterior análisis y corrección, evitando la pérdida de información valiosa.

El **Data Lake** (Almacenamiento RAW) actúa como el punto central donde aterriza toda la información validada sin procesar, sirviendo como la "fuente única de verdad" del sistema. Esto habilita que los procesos de **Machine Learning** lean directamente los datos en crudo (**raw**) para entrenar modelos con el máximo nivel de detalle, sin la pérdida de granularidad que implicaría una agregación previa.

Paralelamente, para el consumo de negocio, los datos fluyen hacia el **Data Warehouse** tras pasar por un proceso ETL que los estructura y optimiza. Así, la capa de **Business Intelligence** consume exclusivamente del Data Warehouse, garantizando rapidez y consistencia en los datos de negocio mediante consultas SQL, lo que permite generar dashboards y reportes ejecutivos con un rendimiento que el Data Lake por sí solo no podría ofrecer.

Implantación en AWS



La solución se mapea a los siguientes servicios nativos de AWS, seleccionados por su madurez, integración y capacidad de escala automática.

Ingesta de Datos (Streaming)

Amazon API Gateway + Amazon Cognito + Amazon Kinesis Data Streams. Se utiliza una entrada segura mediante **API Gateway** protegida por **Cognito** (JWT) y validada por **Lambda** antes de escribir en **Kinesis**. Esto garantiza que solo peticiones auténticas y con formato correcto lleguen al stream. Kinesis actúa como buffer elástico capaz de absorber picos de tráfico masivos sin perder datos.

El uso de componentes serverless (API Gateway, Lambda, Kinesis) reduce el coste de infraestructura ociosa a cero, pagando solo por petición/tráfico. Cognito elimina el desarrollo de un sistema de identidad propio (seguridad delegada). La validación temprana ahorra costes de almacenamiento al descartar datos inválidos antes de persistirlos.

Data Lake (Almacenamiento)

Amazon S3 (Simple Storage Service) + Intelligent-Tiering. S3 sirve como repositorio central inmutable para datos crudos y procesados. Su durabilidad extrema (11 nueves) y escalabilidad infinita lo hacen ideal. Se integra nativamente con todos los servicios de analítica y ML sin necesidad de mover datos.

En cuanto a costes, esta decisión reduce automáticamente los costes moviendo datos antiguos a capas de acceso infrecuente sin impacto operativo. El cifrado por defecto (SSE-S3/KMS) y las políticas de bucket aseguran el cumplimiento de normativas financieras (PCI-DSS) sobre la protección del dato en reposo.

Procesamiento (Batch & Speed Layers)

Amazon EMR (Spark) + AWS Glue + AWS Lambda. Se adopta una arquitectura **Lambda real.** Para la capa **Speed**, funciones **AWS Lambda** analizan eventos de **Kinesis** en milisegundos. Para la capa **Batch**, clústeres efímeros de **Amazon EMR** con Apache Spark procesan terabytes históricos para reentrenar modelos complejos, mientras AWS Glue maneja el catálogo de metadatos y transformaciones ETL ligeras.En cuanto a costes, los clústeres EMR efímeros (que se apagan al terminar el job) evitan pagar computación 24/7 innecesaria. AWS Glue al ser serverless elimina el mantenimiento de servidores ETL. Lambda ofrece la latencia mínima necesaria para detener fraudes en tiempo real, maximizando el ROI del sistema de prevención.

Analítica y Data Warehouse (Serving)

Amazon Redshift + Amazon Athena. Por un lado, **Amazon Redshift** ofrece rendimiento de Data Warehouse empresarial para reportes complejos y concurrentes. **Amazon Athena** permite a los científicos de datos lanzar consultas SQL ad-hoc directamente sobre S3 (serverless) sin necesidad de cargar los datos, ideal para exploración rápida.En cuanto a costes, Redshift permite escalar almacenamiento y cómputo de forma independiente (RA3 nodes), ajustando el coste a la carga real. Athena cobra solo por TB escaneado, incentivando el uso de formatos columnares (Parquet) que reducen drásticamente la factura y mejoran el tiempo de respuesta.

Machine Learning

Amazon SageMaker. Es una plataforma MLOps completa. Gestiona desde el etiquetado (Ground Truth) hasta el despliegue. Permite entrenar modelos distribuidos con infraestructura gestionada (spot instances) y desplegar endpoints de inferencia con auto-scaling para responder en <100ms a las transacciones. El uso de instancias Spot para entrenamiento puede reducir costes hasta un 90%. SageMaker asegura que los modelos se ejecutan en entornos aislados (VPC) y cifrados, protegiendo tanto la propiedad intelectual del algoritmo como los datos sensibles de los clientes.

Business Intelligence

Amazon QuickSight. Esta herramienta de cloud escala automáticamente de decenas a miles de usuarios. Su motor SPICE en memoria acelera los dashboards visuales sin impactar la base de datos subyacente. Se integra con la seguridad de IAM y Cognito para el control de acceso a los reportes. En cuanto a de precios se paga por sesión por lo que es muy económico para usuarios lectores ocasionales (ej. directivos), evitando licencias fijas caras. La seguridad integrada permite definir permisos a nivel de fila (Row-Level Security), asegurando que cada gestor vea solo los datos de su región o departamento.