Arquitectura Híbrida - Educación Virtual

1) Capas y componentes

- Ingesta: LMS, formularios, CRM, pagos, Google Analytics, Git repos.
- Data Lake (raw/bronze + curated/silver): datos crudos y estandarizados.
- Procesamiento / Orquestación: batch, validación DQ, enriquecimiento.
- Data Warehouse: modelo estrella (hechos + dimensiones).
- Data Marts: académico, admisión, finanzas, soporte.
- Consumo: dashboards BI, SQL ad-hoc, notebooks ML, export API.

2) Datos en cada entorno

- Data Lake (raw): todo lo crudo (CSV/JSON, logs, multimedia, eventos de app/web).
- Data Lake (curated): estandarizados, tipados, con PII protegida.
- Data Warehouse: tablas de hechos/dimensiones, métricas confiables, históricos.
- Data Marts: vistas/tablas filtradas y agregadas por dominio (académico, finanzas, marketing).

3) Conexión de entornos

1. Fuentes → Lake (raw) vía conectores/API/CDC.
2. Lake raw → Lake curated (limpieza, tipado, calidad, tokenización).
3. Curated → DW (ELT: staging → dimensiones/hechos).
4. DW → Data Marts (vistas/materialized views por área).
5. DM/DW → BI / Notebooks / APIs.

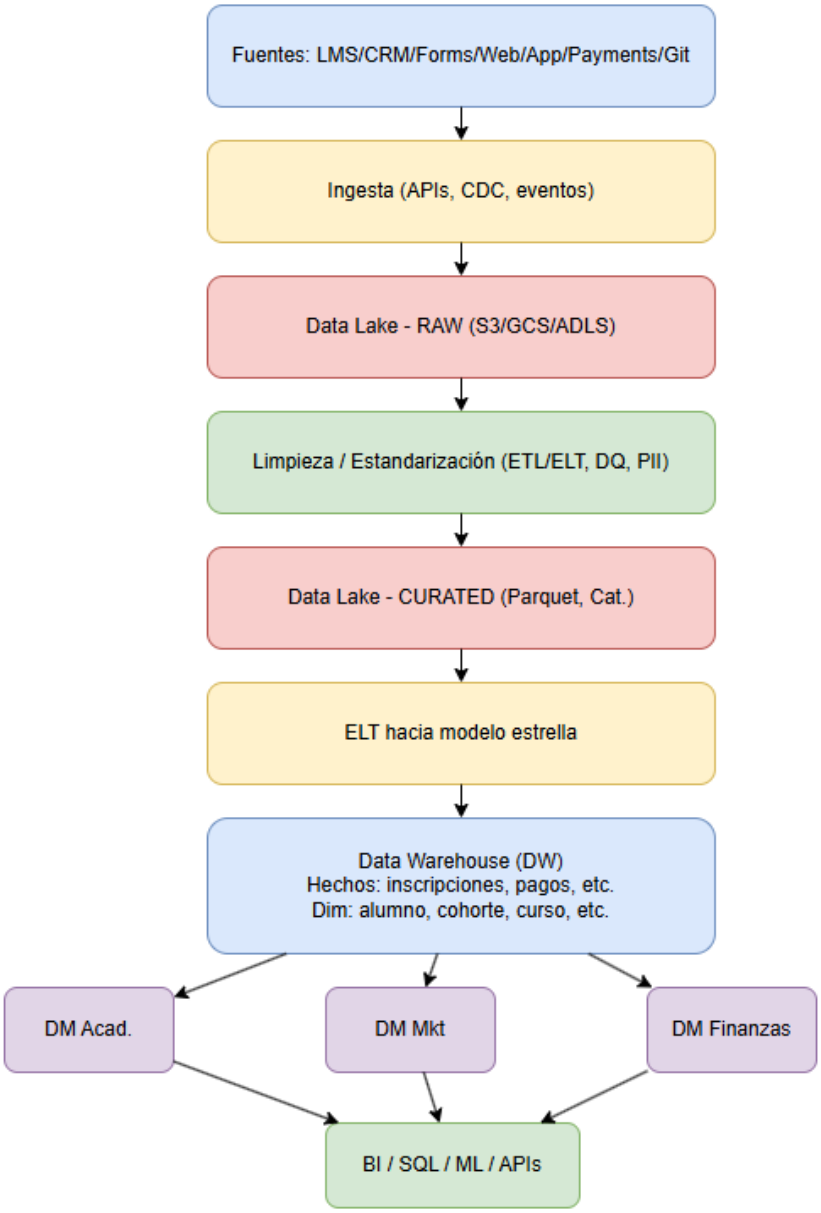
4) Herramientas por nube

- AWS: S3, Glue/Athena, EMR, Kinesis, Redshift, Lake Formation, QuickSight/Power BI.
- GCP: Cloud Storage, Dataflow, Pub/Sub, BigQuery, Dataplex, Looker/Power BI.
- Azure: ADLS Gen2, Data Factory, Event Hubs, Synapse, Purview, Power BI.

5) Usuarios y propósitos

- UTP/Académico: alertas de riesgo, tasa de aprobación, asistencia, progreso por módulo.
- Admisión/Marketing: CAC, ROI por canal, embudo de conversión, cohortes.
- Finanzas: ingresos por cohorte/carrera, morosidad, ARPU/LTV.
- Dirección: KPIs ejecutivos (retención, NPS, crecimiento).
- Data/ML: notebooks para modelos de deserción/éxito y segmentación.

6) Diagrama



7) Gobernanza y seguridad

- Catálogo & linaje: Glue Catalog / Data Catalog / Purview; etiquetas de sensibilidad.
- Accesos por rol: IAM/RBAC, columnas enmascaradas (PII), row-level security en BI/DW.
- Calidad: pruebas en pipelines (Great Expectations/Deequ), métricas DQ (completitud, unicidad, validez).
- Costos: particiones por fecha/cohorte, formatos columnares (Parquet), cache de consultas, auto-suspend en DW.