

# Gerardo Toboso

Ingeniero de Datos

Buenos Aires, Argentina | gerardotoboso1909@gmail.com | 011 15-4045-6207  
linkedin.com/in/gerardo-toboso-512a48290 | github.com/Gerardo1909

## Resumen

---

Ingeniero de Datos con experiencia en el diseño y automatización de pipelines robustos para entornos de datos complejos. He desarrollado proyectos propios end-to-end, desde la orquestación y monitoreo de flujos complejos de datos hasta la integración y análisis de datos transaccionales para responder preguntas de negocio. Me especializo en transformar grandes volúmenes de datos en modelos analíticos listos para tomar decisiones críticas, priorizando la calidad, trazabilidad y eficiencia operativa. Apasionado por la documentación, la mejora continua y la entrega de soluciones confiables para equipos de producto y analítica. Actualmente cursando la Licenciatura en Ciencia de Datos (UNSAM, Buenos Aires) y buscando aportar en equipos de ingeniería de datos innovadores y colaborativos.

## Habilidades

---

**Orquestación y Pipelines:** Apache Airflow (DAGs, monitoreo, alertas, reintentos), uso de operadores (DockerOperator, BashOperator, etc.), Scheduling (Cron) y versionado de datos

**Procesamiento y Modelado de Datos:** PySpark, Python (pandas, numpy, duckdb), Modelado dimensional (Star Schema - OLAP), Manejo de formato Parquet para particionamiento eficiente y reducción de tamaño

**Cloud & Data Lake:** AWS S3 (data lake), Docker, CI/CD (GitHub Actions), despliegue reproducible, seguridad y control de accesos

**Monitoreo y Calidad:** Logging estructurado, alertas automáticas, testing de pipelines (Pytest), control de calidad (Great Expectations) y trazabilidad de datos

**SQL & BI:** SQL (PostgreSQL), Generación de Dashboards (Looker Studio) y analítica avanzada orientada a producto

**Idiomas:** Español (Nativo) | Inglés (Fluido - C2)

## Experiencia

---

**Ingeniero de Datos, ETL de Reportes E-commerce (Proyecto Personal) – [GitHub](#)**

Nov 2025 – Dic 2025

- Diseñé e implementé un pipeline ETL end-to-end en Python para un dataset de e-commerce (10+ tablas relacionadas), reduciendo el tiempo de reportes manuales de 2 horas a menos de 3 minutos.
- Definí y calculé métricas de negocio clave (mejores clientes, productos más vendidos, tendencias mensuales) para soportar la toma de decisiones basada en datos.
- Identifiqué y resolví problemas de calidad de datos, manejando ~15% de valores faltantes y eliminando ~3% de registros duplicados.
- Optimicé el almacenamiento de datos migrando outputs de CSV a Parquet, reduciendo el tamaño de archivos en 8x.

**Ingeniero de Datos, ETL para datos IoT (Proyecto Personal) – [GitHub](#)**

Dic 2025 – presente

- Diseñé e implementé un pipeline ETL orquestado con Airflow para procesar millones de lecturas IoT industriales diariamente provenientes de una API externa, logrando 99.9% de uptime y detección de fallas en menos de 5 minutos.
- Implementé la etapa de transformación en PySpark, logrando procesar 50GB/minuto sin problemas de memoria.
- Implementé particionamiento y versionado de datos en Parquet, optimizando queries y reduciendo costos de almacenamiento.
- Diseñé esquema de almacenamiento en AWS S3 (data lake), asegurando potencial escalabilidad ante un mayor volumen de datos (> 1TB diario).

## Educación

---

**Universidad Nacional de San Martín (UNSAM), Licenciatura en Ciencia de Datos**

Jul 2022 – presente

- 75% completado — Promedio: 9.0 / 10