Gerardo Toboso

Data Engineer

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

Summary

Data Engineer with experience designing and automating robust pipelines for complex data environments. I have developed end-to-end personal projects, from orchestration and monitoring of complex data flows to the integration and analysis of transactional data to answer business questions. I specialize in transforming large volumes of data into analytics-ready models for critical decision-making, prioritizing quality, traceability, and operational efficiency. Passionate about documentation, continuous improvement, and delivering reliable solutions for product and analytics teams. Currently pursuing a BSc in Data Science (UNSAM, Buenos Aires) and seeking to contribute to innovative and collaborative data engineering teams.

Skills

Orchestration & Pipelines: Apache Airflow (DAGs, monitoring, alerts, retries), use of operators (DockerOperator, BashOperator, etc.), Scheduling (Cron), and data versioning

Data Processing & Modeling: PySpark, Python (pandas, numpy, duckdb), Dimensional modeling (Star Schema - OLAP), Parquet format management for efficient partitioning and size reduction

Cloud & Data Lake: AWS S3 (data lake), Docker, CI/CD (GitHub Actions), reproducible deployment, security and access control

Monitoring & Quality: Structured logging, automated alerts, pipeline testing (Pytest), data quality control (Great Expectations), and data traceability

SQL & BI: SQL (PostgreSQL), Dashboard generation (Looker Studio), and advanced analytics oriented to product

Languages: Spanish (Native) | English (Fluent - C2)

Experience

Data Engineer, E-commerce Reporting ETL (Personal Project) – GitHub Nov 2025 – Dec 2025

- Designed and implemented an end-to-end ETL pipeline in Python for an e-commerce dataset (10+ related tables), reducing manual reporting time from 2 hours to under 3 minutes.
- Defined and calculated key business metrics (top customers, best-selling products, monthly trends) to support data-driven decision-making.
- Identified and resolved data quality issues, handling ~15% missing values and removing ~3% duplicate records.
- Optimized data storage by migrating outputs from CSV to Parquet, reducing file size by 8x.

Data Engineer, IoT Data ETL (Personal Project) – GitHub Dec 2025 – present

- Designed and implemented an ETL pipeline orchestrated with Airflow to process millions of industrial IoT readings daily from an external API, achieving 99.9% uptime and failure detection in under 5 minutes.
- Implemented the transformation stage in PySpark, processing 50GB/minute without memory issues.
- Implemented data partitioning and versioning in Parquet, optimizing queries and reducing storage costs.
- Designed a storage schema in AWS S3 (data lake), ensuring potential scalability for higher data volumes (> 1TB daily).

Education

Universidad Nacional de San Martín (UNSAM), BSc in Data Science July 2022 – present

- 75% completed — GPA: 9.0 / 10