Hamza Anwar

Data Engineer · Python Developer · AI Automation & Platform Specialist

Szczecin, Poland | +48 505 687 830 | hamzaraja983@gmail.com | linkedin.com/in/hamzaraja983

github.com/mhamzaanwar | EU Blue Card

PROFESSIONAL SUMMARY

Data Engineer and Python Developer with 6+ years of production experience building reliable, high-throughput ETL pipelines and distributed data platforms. At X10 Logistics, architected Python-based data infrastructure processing 2M+ multilingual product listings across multiple e-commerce platforms, with full CI/CD, Docker containerisation, and Airflow orchestration. Hands-on with Snowflake, BigQuery, Databricks, Apache Spark, and medallion lakehouse architecture, complemented by cloud deployment on AWS and GCP. Experienced in AI automation workflows (n8n, Make.com, Zapier), RAG pipelines, and LLM integration.

PROFESSIONAL EXPERIENCE

Data Engineer & Python Developer | X10 Logistics sp. z o.o.

Dec 2023 – Mar 2026 · Szczecin, Poland

Data Platform & ETL Engineering

- Architected end-to-end ETL pipelines in Python for product, inventory, and sales data, reliably ingesting and synchronising 2M+ multilingual product listings across multiple e-commerce platforms.
- Implemented data validation, transformation, and normalisation logic (Pandas, NumPy) against PostgreSQL and MySQL schemas, reducing downstream reporting errors.
- Orchestrated scheduled pipeline workflows with Apache Airflow; diagnosed processing failures and maintained stable production delivery.
- Deployed containerised pipeline environments using Docker and Jenkins CI/CD, enabling repeatable, auditable deployments.
- Built distributed data processing workflows with Apache Spark and PySpark for large-scale transformation over millions of records.
- Integrated Snowflake & BigQuery with medallion architecture (Bronze / Silver / Gold) for reliable, incremental data processing.

Data Quality, Observability & Collaboration

- Implemented structured logging and monitoring across all pipeline stages; sustained SLA compliance at scale.
- Delivered KPI dashboards and reporting pipelines for sales, logistics, and operations leadership.
- Applied agile engineering practices: feature branching, pull request reviews, TDD, and iterative CI/CD via Jenkins and GitHub.

Python Developer | GulzarSoft

Jul 2019 – Mar 2023 · Gujrat, Pakistan

ETL, Automation & Team Leadership

- Designed and built modular, high-performance web scraping systems using Scrapy, Selenium, and BeautifulSoup to handle dynamic content, proxy rotation, and structured parsing at scale.
- Built and deployed automated ETL ingestion workflows across multiple industries with strong schema compliance standards.
- Led a team of developers; established code review workflows and version control practices (Git/GitHub/GitLab).
- Mentored junior developers on Python, Scrapy architecture, and ETL design, elevating team-wide technical quality.

TECHNICAL SKILLS

Languages & Core: Python, SQL (advanced), Bash, HTML/CSS/JavaScript, asyncio, Jupyter Lab

Data Engineering: Apache Spark, PySpark, Apache Kafka, Apache Airflow, Databricks, Delta Lake, ETL/ELT, Event-Driven Processing

Cloud & Warehouses: Snowflake, BigQuery, AWS S3, GCP Cloud Run, Azure Data Lake

Databases: PostgreSQL, MySQL, MongoDB, SQLite — query optimisation, schema design, large-scale processing

Analytics & BI: Pandas, NumPy, scikit-learn, Matplotlib, Power BI, Tableau, Looker Studio, KPI dashboards

Scraping & Ingestion: Scrapy, Selenium, BeautifulSoup, asyncio, headless browsers, proxy management

Backend & APIs: FastAPI, Flask, REST APIs, JSON — eBay API, Allegro API, TecDoc, Bootstrap

DevOps & CI/CD: Docker, Jenkins, Git (GitHub / GitLab), CI/CD pipelines, containerised deployments

Architecture: Medallion (Bronze/Silver/Gold), Data Lakehouse, Microservices, Event-Driven Processing

AI & Automation: n8n, Make.com, Zapier, RAG Pipelines, LLM Integration, AI Agents, LangChain, Vector Databases, Prompt Engineering

SELECTED PROJECTS

DSF Data Platform — 30+ production systems built at X10 Logistics (private / NDA)

DSF-Scrapers v1 + v2.0 · Automotive Parts Ingestion Engine *X10 Logistics · Private*

Web data ingestion engine capturing 10,000+ product records daily. v2.0 rebuilt with modular spider architecture and proxy orchestration, cutting processing time by 40%. Spiders parse JSON from marketplace scripts, rotate through 3 proxy providers via ScraperAPI, translate via DeepL, and feed a dual-database pipeline (MySQL + MongoDB) with Scrapyd deployment and AWS S3 photo storage.

Stack: Python, Scrapy, asyncio, MySQL, MongoDB, AWS S3, DeepL, ScraperAPI | **Impact:** 10K+ records/day · 40% faster v2.0

DSF-Cleaners v1 + v2.0 · Automotive Data Validation Pipeline *X10 Logistics · Private*

Multi-stage validation and cleaning pipeline between raw scrape output and marketplace upload. v2.0 added ML-based anomaly detection and automated error recovery, achieving 99%+ data accuracy. Validates and enriches 14 part categories: deduplicates SKUs, extracts production years, matches manufacturers, and resolves part numbers against a 22 MB compatibility database. Master-slave MySQL with transactional rollback.

Stack: Python, Pandas, NumPy, scikit-learn, MySQL, MongoDB | **Impact:** 99%+ data accuracy

DSF-Uploaders 2.0 · Marketplace Upload System *X10 Logistics · Private*

Next-gen batch upload system handling 1,000+ files per batch with validation and error reporting. Manages OAuth2 lifecycle, batches 20-item API calls, and runs category-specific crons for price, quantity, and description sync. Delists out-of-stock items with 1,024 concurrent requests across multi-shop environments (X10, X102, ATS).

Stack: Python, eBay Sell API, MySQL, MongoDB, asyncio | **Impact:** 1,024 concurrent requests

DSF-SalesReturns · Returns Analytics Platform *X10 Logistics · Private*

Returns analytics platform identifying root causes of product returns. Correlates return reasons with listing quality, product descriptions, and category attributes. Data-backed improvements drove a 15% reduction in return rate through targeted listing and product changes.

Stack: Python, Pandas, SQL, Matplotlib | **Impact:** 15% returns reduction

DSF-NP-StockPrices · Multi-Account Stock & Price Management *X10 Logistics · Private*

Multi-account dynamic pricing management system for new parts inventory. Rule-based pricing engine adjusts prices based on competition, margin targets, and stock levels. Dynamic pricing rules reduced overstock by 30%.

Stack: Python, MySQL, Pricing Engine, Automation | **Impact:** 30% overstock reduction

DSF-MarketingProject v1 + v2.0 · eBay Campaign Automation *X10 Logistics · Private*

eBay campaign automation with percentile-based customer segmentation. Automates Promoted Listings campaigns based on sales velocity, margin, and competitive positioning. v2.0 added cohort analysis and A/B test evaluation, improving campaign performance by 25%.

Stack: Python, eBay API, Pandas, Data Analysis | **Impact:** 25% campaign improvement

INVENTORY-MANAGEMENT-API · Inventory Orchestration Layer *X10 Logistics · Private*

Inventory orchestration layer synchronizing stock state across all downstream systems (eBay, Allegro, Express-Teile). Serves as the source of truth for available quantities, coordinates stock reservations between marketplace channels, and emits events to trigger replenishment workflows.

Stack: Python, FastAPI, MySQL, Event-Driven, Microservices

Polish Law RAG Assistant *Open Source · polishlawguide.streamlit.app*

Production-grade RAG system answering questions about Polish business law, tax setup, and B2B contracting, with cited source paragraphs on every answer. Built to avoid hallucination.

Stack: Python, RAG, LLM, Streamlit

Synthetic Data Pipeline · MSc Thesis · Högskolan Dalarna *Open Source · github.com/mhamzaanwar*

End-to-end generative synthesis pipeline for IoT sensor data in smart-home environments. Trains deep generative models (CTGAN, TVAE, TimeGAN) and enforces post-generation privacy filtering (DCR-based). Evaluated on statistical fidelity, downstream ML utility, and privacy risk.

Stack: Python, CTGAN, TVAE, TimeGAN, scikit-learn, Jupyter

EDUCATION

M.Sc. Business Intelligence *Jan 2023 – Jan 2024*
Högskolan Dalarna, Sweden
Thesis: Synthetic Data Generation for Privacy-Aware Homecare Support. Implemented GANs, CTGAN, VAEs, and GTransformers to create synthetic sensor data for downstream predictive ML models.

B.Sc. Information Technology *Sep 2015 – Jul 2019*
University of Gujrat, Pakistan

CERTIFICATIONS

Data Engineering Essentials, IBM | Nov 2025
ETL and Data Pipelines with Shell, Airflow & Kafka, IBM | Nov 2025
Introduction to PySpark, DataCamp | Nov 2025
Microsoft Office Specialist (Word / Excel / PowerPoint 2013), Microsoft | Dec 2018