AMR MUHAMED FATHY

Data Engineer

amrmuhamed86@gmail.com | +201111276198 | El-Sayeda Zeinab, Cairo.
Github | Linkedin | Portfolio

Data Engineer with extensive experience building scalable data pipelines, aggregating large-scale datasets, and optimizing data processing workflows. Develop and maintain robust ETL/ELT systems using Python, leveraging tools like Pandas, NumPy, and more for data manipulation and visualization. **Create advanced datahub** *(MedData)* and scraping frameworks, integrating APIs for machine learning and analytics applications. **Design and implement** static code analysis tools with rich-terminal dashboards to enhance data quality and pipeline observability. Driving data engineering innovation. **Demonstrate** strong team leadership, coordinating teams to deliver high-impact data projects under tight deadlines.

Work Experience

- Full Stack Developer — Kab — Jan. 2025 - Present
  - Remotely Cairo, Egypt.
  - Design and maintain data-driven full-stack applications, integrating Python-based ETL pipelines for efficient data processing.
  - Develop responsive dashboards using Next.js, enabling real-time data insights.

Education

- **Bachelor** of Science in Computer Science
  - *Benha University — Expected Graduation (2026)*
- **High School**: Gamal Abd El-Nasser Languages School

Training

- **DEPI Program** — Government Data Engineering Track — Jun. 2025 – Present
  - Advanced Python programming with focus on data engineering methodologies and pipeline construction
  - Database design, management, and optimization for large-scale data processing applications
  - ETL process development with industry best practices and performance optimization techniques

- **Benha University** — Internship program — Jun. 2024 – Jul. 2024
  - Build Python-based data applications with database integration, focusing on data aggregation and processing.

- Lead ChitChat project team, developing data-driven chat applications with socket programming and MVC architecture.
- Implement data pipelines for real-time communication, contributing ~85% of codebase while managing project timelines.

Technical Skills

- **Programming:** Python (build ETL pipelines, data processing, and ML dataset generation), JavaScript (integrate APIs for data aggregation), Ruby/Jekyll (generate data hub websites).
- **Data Processing:** Pandas, NumPy (data cleaning, transformation), Plotly, Chart.js (interactive visualizations).
- **Orchestration:** Javascript & Streamlit (build dashboards for data analytics).
- **Databases/Storage:** MongoDB (manage complex data relationships), Supabase (real-time data features), Parquet (efficient ML dataset storage).
- **Testing/Quality:** Implement error handling, rate limiting, and retry mechanisms for robust data pipelines.

Personal Skills

- **Leadership:** Coordinate teams, delegate tasks, and drive project delivery.
- **Adaptability:** Transition seamlessly between data engineering tools and frameworks based on project needs.
- **Time Management:** Balance concurrent data engineering projects with academic and open-source commitments.

Linguistics Skills

- **Arabic:** Mother tongue.
- **English:** Professional proficiency (daily use in technical documentation and collaboration).

Projects

- **CodeLyzer** — Static Code Analysis & Visualization Platform
  Python, Pandas, NumPy, Javascript, Rich-terminal, tree_sitter
  - Develop feature-rich static code analysis tools to identify data quality issues and complexity hotspots in large-scale repositories.
  - Implement rich-terminal dashboards with Plotly and Chart.js for actionable data metrics and pipeline observability.
  - Create aggregation algorithms to process repositories.
  - Deliver reporting system for data pipeline health, reducing technical debt through actionable insights.

- **MedData** — Multi-Source ML Dataset Hub — *Official Webpage*
  Python, Ruby/Jekyll, HuggingFace, Kaggle, SSG
  - Build data hub aggregating datasets from Hugging Face, Kaggle, Medium, and Dev.to for machine learning research.
  - Implement automated ETL pipelines with Pandas and NumPy for data cleaning and preprocessing across diverse formats.
  - Generate accessible dataset documentation using Ruby/Jekyll, improving data usability for researchers.

- **GHRepoLens** — GitHub Repositories Analysis & Intelligence
  Python, Pandas, Javascript, GitHub API
  - Develop tool for analyzing GitHub repositories, generating detailed reports and interactive visualizations.
  - Implement web scraping and GitHub API integration to process data from repositories, enabling trend analysis.
  - Create configurable Static dashboard for repository management and data-driven development insights.
  - Build CLI interface with exportable reports, enhancing data observability for development teams.

- **PixCrawler** — Configurable Image Dataset Builder
  Python, Pandas, JSON, AI
  - Build scalable image dataset builder using web crawling from Google and Bing and more; with JSON-based configuration.
  - Implement concurrent ETL pipelines with comprehensive error handling and rate limiting for reliable data collection.
  - Design data processing pipeline with Pandas for metadata extraction and organization, supporting large-scale datasets.
  - Create a flexible configuration system for custom search parameters, improving dataset quality for ML applications.

- **DevToHarvest** — Technical Content Dataset Generator
  Python, Pandas, Parquet, Streamlit
  - Develop Python scraper to create ML-ready datasets from Dev.to articles, focusing on technical content.
  - Implement parallelized ETL pipelines with intelligent rate limiting and retry mechanisms for robust data collection.
  - Create Parquet export functionality for efficient storage, optimizing datasets for NLP training pipelines.
  - Extract comprehensive metadata, enhancing dataset usability for machine learning applications.

- **windsurf-scraper** — Multi-Protocol Web Scraping Framework
  Python, Pandas, Faker, Selenium
  - Build a flexible scraping framework supporting multiple protocols and authentication methods for data extraction.
  - Implement robust error handling and data validation, ensuring reliable collection for data pipelines.

Accomplishments

- **Publish** 5+ PyPI packages, enhancing data engineering workflows with tools for dataset generation and processing.
  - ***C4f:*** *Github* | *PyPi*
  - ***Colab-print:*** *Github* | *PyPi*
  - ***Jsdfile:*** *Github* | *Pypi*
  - ***True-storage:*** *Github* | *PyPi*
  - ***True-core:*** *Github* | *PyPi*
- **Driving** data engineering and transformation innovation across organizations like JsonAlchemy and T2F-Labs.
- **Develop** scalable ETL pipelines for MedData, aggregating datasets from multiple sources for ML research.
- **Implement** rich-terminal dashboards in CodeLyzer, improving data quality observability for large-scale codebases.
- **Lead** data-driven projects like TeamUp, reducing communication overhead and enabling efficient team formation.

References upon your request...