Ossama Taha

Data Engineer

6th of October City, Egypt | ossama.y.taha@gmail.com | +20 10 26458878 | linkedin.com/in/ossamataha

github.com/OssamaTaha

Summary

Data Engineer with 2+ years building production ETL pipelines, orchestrated data workflows, and real-time streaming systems. Proficient in Python, SQL, Apache Airflow, Kafka, dbt, and AWS. Proven track record of reducing reporting time by 10x through automation and building scalable data platforms from scratch.

Experience

Nile for Cables and Electronics, BI & Sales Data Analyst — Giza, Egypt — Nov 2024 – present

- Engineered a semi-realtime automated reporting system replacing Power BI, improving reporting speed by 10x
- Designed end-to-end ETL pipelines using PostgreSQL and Dolt (version-controlled SQL database)
- Built executive dashboards supporting sales forecasting and strategic business decisions
- Led quarterly sales performance evaluations using consolidated KPIs
- Developed a local AI chatbot using open-source Hugging Face models for natural language data querying

Samsung Innovation Campus, Big Data Engineer — Giza, Egypt — July 2024 – Oct 2024

- Built scalable ETL pipelines using Apache Spark, Kafka, and Hadoop ecosystem
- Performed exploratory data analysis and feature engineering for analytical use cases
- Deployed data workflows on AWS EC2 and S3 for production environments

Projects

Airflow ETL Pipeline — Multi-Source Data Orchestration

- **Problem:** Manual data extraction from multiple APIs was error-prone and lacked scheduling, monitoring, or failure recovery
- **Process:** Designed 4 production DAGs with task dependencies, dynamic task mapping, XCom data passing, and automated retry logic
- **Solution:** End-to-end orchestrated pipeline extracting from 3 API sources, transforming in staging, and loading into PostgreSQL
- **Tech Stack:** Apache Airflow, Docker, PostgreSQL, Python, Streamlit, GitHub Actions
- **Business Impact:** Fully automated daily/weekly data ingestion with monitoring dashboard, zero manual intervention required

Kafka Streaming Pipeline — Real-Time Data Processing

- **Problem:** Batch processing couldn't meet real-time analytics requirements for stock monitoring and IoT telemetry
- **Process:** Built event-driven architecture with Kafka producers, consumer groups, windowed aggregation, and dead letter queues
- **Solution:** Real-time pipeline processing 1000+ events/sec with live dashboard, alerting, and PostgreSQL sink
- **Tech Stack:** Apache Kafka, Docker, Python, PostgreSQL, Streamlit, Pydantic, GitHub Actions
- **Business Impact:** Sub-second data latency, automated alerting on threshold breaches, real-time operational visibility

AWS Cloud Data Pipeline — Serverless ETL with IaC

- **Problem:** On-premise data infrastructure lacked scalability, reliability, and cost efficiency for growing data volumes
- **Process:** Designed S3 data lake architecture with Glue ETL, Athena serverless queries, and Step Functions orchestration
- **Solution:** Fully serverless pipeline with Terraform IaC — deploy entire infrastructure with single command
- **Tech Stack:** AWS (S3, Glue, Athena, Lambda, Step Functions, CloudWatch), Terraform, Python, PySpark, Docker
- **Business Impact:** 90% cost reduction vs traditional infra, auto-scaling, zero server management, sub-second query performance

dbt Data Warehouse — Star Schema Design & Transformation

- **Problem:** Raw operational data lacked dimensional structure for efficient analytics and BI tool consumption
- **Process:** Designed star schema with fact and dimension tables, built 15+ dbt models across staging/intermediate/marts layers
- **Solution:** Production-grade data warehouse with automated testing, documentation, SCD Type 2 tracking, and incremental processing
- **Tech Stack:** dbt-core, PostgreSQL, Docker, Jinja, Python, Faker, Streamlit, GitHub Actions
- **Business Impact:** Reduced query complexity by 70%, enabled self-service analytics, 99.9% data quality via dbt tests

Agentic AI Workflow — Multi-Agent DataOps Automation

- **Problem:** Data engineering operations required manual monitoring, debugging, and intervention across multiple systems
- **Process:** Built multi-agent system with orchestrator pattern, 5 specialized agents, tool integration, and persistent memory
- **Solution:** Natural language interface for DataOps — agents autonomously monitor pipelines, run quality checks, debug failures
- **Tech Stack:** LangChain, LangGraph, Python, PostgreSQL, Streamlit, Docker, OpenAI/Ollama, GitHub Actions
- **Business Impact:** Reduced mean-time-to-resolution for pipeline failures by 70%, automated 90% of routine monitoring tasks

Atlas Analytics Platform

- **Problem:** Enterprise needed unified analytics platform for multi-department data visibility
- **Process:** Built full-stack platform with Python backend, real-time ETL pipelines, and interactive dashboards
- **Solution:** Deployed production system with role-based access control, automated data sync, and real-time reporting
- **Tech Stack:** Python, PostgreSQL, Docker, AWS EC2/S3, Streamlit
- **Business Impact:** Enabled data-driven decisions across 3 departments, reduced manual reporting by 80%

Skills

Programming & Scripting: Python, SQL, Bash, JavaScript, TypeScript, HTML/CSS

Data Engineering: ETL/ELT Pipelines, Apache Airflow, Apache Spark, Apache Kafka, Hadoop, dbt, Data Modeling (Star Schema), PostgreSQL, MySQL, SQLite

Cloud & Infrastructure: AWS (S3, Glue, Athena, Lambda, Step Functions, EC2), Terraform, Docker, Docker Compose, Linux Server Administration, Nginx, GitHub Actions (CI/CD)

BI & Visualization: Power BI, DAX, Tableau, Excel, Dash, Streamlit, Plotly

AI & Machine Learning: LangChain, LangGraph, Hugging Face Transformers, Scikit-learn, TensorFlow, NLP, Local LLM Deployment

Data Quality & Monitoring: Great Expectations, Data Validation, Schema Testing, Anomaly Detection, Pipeline Monitoring

Automation & Tools: Git, n8n, OpenClaw, KiloCode, Hermes, Dolt, Jupyter, VS Code

Education

Al-Azhar University, Accounting & Business Management — Sept 2019 – June 2023

Bachelor of Commerce

- Self-taught in programming, data engineering, and cloud technologies through professional certifications

Certifications

AWS Certified Data Engineer – Associate (In Progress)

IBM AI Engineering Professional Certificate (In Progress)

IBM Data Engineer Professional Certificate

- Covered ETL pipelines, data warehousing, big data technologies (Spark, Hadoop, Kafka), and cloud deployment — Verify

Google Data Analytics Professional Certificate

- Trained in data cleaning, analysis, visualization, and storytelling using spreadsheets, SQL, Tableau, and R programming — Verify

IBM Data Science Professional Certificate

- Built data science skills including machine learning, data analysis, Python, SQL, and applied data science capstone projects — Verify

Samsung Innovation Campus Big Data Engineering Certificate

- "**Top Achiever (#1)** in the Big Data Engineering track"
- "Completed intensive training in big data infrastructure,distributed computing, and data engineering fundamentals"

Languages

Languages: Arabic: Native | English: Professional | Russian: Basic