

Sameh Sharaf

Berlin, Germany · (+49) 0152 5894 0798 · sameh.shar@gmail.com

WORK EXPERIENCE

Senior Data Engineer | Zalando | Berlin, Germany

Sep 2020 – Present

Senior Data Engineer building, developing, and maintaining large-scale ETL pipelines on Airflow, AWS, and Databricks. Led the successful migration of 100+ production pipelines from Oracle and Exasol to a modern data lakehouse, collaborating cross-functionally with engineering, marketing, analytics, and product teams. Designed and implemented robust pipeline infrastructure with comprehensive monitoring, fault handling, and data quality governance to ensure high reliability at scale. Drove cost optimisation initiatives by modernising legacy data infrastructure and streamlining cloud resource utilisation.

Data Engineer | Sertis | Bangkok, Thailand

Jun 2018 – Aug 2020

Built and delivered cloud and on-premise data solutions for 10+ enterprise and mid-market clients across retail, finance, healthcare, and telco industries on AWS, GCP, and Azure. Optimised client ETL pipelines on Databricks, delivering measurable reductions in processing latency and infrastructure costs. Developed internal self-serve tooling for data scientists and analysts covering infrastructure provisioning and data wrangling, containerised with Docker and Kubernetes.

Data Engineer | iflix | Kuala Lumpur, Malaysia

Dec 2016 – Jun 2018

Sole data engineer responsible for end-to-end data infrastructure handling petabytes of data across real-time streaming and batch pipelines. Designed and administered AWS Redshift data warehouse, delivering significant improvements in query performance and pipeline reliability. Built executive-facing dashboards and reporting layers covering the full data lifecycle — collection, transformation, storage, and visualisation. Managed Qubole platform including Apache Hive, Spark, and Airflow clusters, scaling infrastructure to support growing data volumes while maintaining data governance and quality standards.

Full Stack / Data Engineer | Jirnexu | Kuala Lumpur, Malaysia

Aug 2015 – Dec 2016

Sole data and software engineer driving the full development of SORA, an in-house CRM system serving marketing, tele-sales, and customer service teams. Engineered near real-time data synchronisation from AWS DynamoDB to Redshift using DynamoDB Streams, enabling more reliable data availability and improved lead conversion tracking. Built a dynamic query builder translating UI inputs to SQL, significantly accelerating access to customer data for non-technical teams. Administered AWS Redshift data warehouse and delivered new reporting capabilities across tens of thousands of customer records.

BI Engineer | MTN Syria | Damascus, Syria

Nov 2010 – Jun 2015

Contributed to revenue assurance initiatives, identifying and recovering millions in lost revenue caused by billing inaccuracies and system failures. Automated manual reconciliation processes and built monitoring and alerting systems to proactively detect and reduce revenue leakage. Improved billing accuracy through Oracle PL/SQL development and full administration of roaming services on BSCS iX. Ensured ongoing compliance with information security policies and regional regulatory requirements.

TECHNICAL EXPERIENCE

Big Data: Apache Spark, Databricks, AWS Redshift, PostgreSQL, Oracle PL/SQL, AWS DynamoDB, Google BigQuery, Elastic Stack
Orchestration: Apache Airflow, MLflow
Cloud Services: Amazon Web Services (AWS), Google Cloud (GCP), Microsoft Azure, Databricks

Machine Learning: LLMs, Agentic Engineering, Deep Learning (RNN, CNN), NLP, PyTorch, SciKit-Learn, AWS SageMaker, AWS Bedrock

Programming: Python, Go

DevOps: Docker, Gitlab CI, Kubernetes (GKE), Terraform

Monitoring: DataDog, Grafana, Bugsnag, Rollbar, Sentry

Software Eng.: ITIL v3.0, UML, ERD, Flowchart Design

Other: FastAPI, Falcon, HTML, CSS, Bootstrap, Semantic UI

Telecommunication: BSCS IX R2, CS5 Ericsson Warehouse Model, Interconnect v7.1

CERTIFICATES

Google Cloud Certified Data Engineer Professional

Mar 2019

AWS Certified Solution Architect – Associate

Dec 2017

EDUCATION

Georgia Institute of Technology – Machine Learning

Atlanta, USA | Present

Damascus University – Software Engineering

Syria | 2005 – 2010

WORK / ACADEMIC PROJECTS

Podcast Transcript Extraction & Emotion Recognition | *Aug 2025*

Built an end-to-end subtitle generation pipeline for a freelance client, integrating HuggingFace speech-to-text and emotion recognition models to automatically transcribe podcast audio across multiple European languages. Generated emotionally-aware subtitles by mapping recognised emotional tone to a text-to-speech reader, producing localised, expressive output deployed on client infrastructure.

Breast Cancer Early Detection | *May 2023*

Academic project developing and evaluating VGG-based deep learning models for automated breast cancer detection from X-ray images. Applied transfer learning and data augmentation techniques to optimise model performance on a Kaggle dataset, achieving 71% accuracy. Conducted as part of a small team with findings presented as a comprehensive research study.

Demand Forecast Project | *Feb 2020 – Aug 2020*

Successfully delivered a two-phase demand forecasting engagement for a multinational consumer goods company. Phase one involved designing and implementing an ETL pipeline to process and load data into Azure SQL DB, enabling downstream analytics. Phase two delivered a cloud architecture design on MS Azure and Databricks to host and serve a demand forecasting ML model, handed off for production deployment.

Data Migration & Analysis | *Feb 2020 – Aug 2020*

Successfully migrated a mid-size Thai retail chain from a legacy data warehouse to MS Azure, designing and implementing an end-to-end ETL pipeline for data processing and storage. Enabled analytics and visualisation capabilities using Azure Analysis Services, modernising the client's data infrastructure and unlocking new business insights.

Databricks Workspace Manager | *Oct 2019*

Designed and proposed an internal tooling solution for automated Databricks cluster management and cross-workspace notebook synchronisation, architected for deployment on MS Azure. Presented as a formal proposal to address operational inefficiencies in managing distributed Databricks environments.

DQA: Data Quality Assessment Services | *Jan 2019 – Aug 2020*

Designed and built an internal data quality service adopted by analysts and integrated across data pipelines. Provided automated summary statistics and configurable data validation rules via a publish-subscribe architecture, exposing REST APIs and a Web UI for self-serve data quality monitoring.

SORA: A CRM System | *Aug 2015 – Dec 2016*

Designed and developed SORA, a full-featured in-house CRM system for Jirnexu, enabling telesales and customer service teams to track lead conversions and manage customer communications via VoIP, SMS, and email. Replaced manual tracking processes with a centralised platform supporting direct customer engagement workflows.

Nagios Monitoring System | *Jun 2012 – Sep 2012*

Led the installation and integration of a Nagios monitoring system for the billing department as part of a small team, configuring surveillance across 100+ processes and 12 servers. Implemented automated alerting and event-handling capabilities to detect and respond to unexpected incidents in real time.

OTHER WORKS

Online Course Instructor – Udemy, One Education, Almentor

Course: SQL for Newcomers – Mastery Course