

Emma Luk

Data Engineer

Watford • 07974 522 805 • emma_luk2@hotmail.com • [linkedin.com/in/emma-luk](https://www.linkedin.com/in/emma-luk) • emma-luk.github.io (Portfolio & GitHub)

Summary

Knowledgeable professional British citizen, with 12+ years of diversified industry hands-on experience in software development and data science, cloud migration, and application migration. Guided 50+ projects in global companies such as BAE, with project success of 96%. Holder of a Master of Science in Data Science and Analytics, Master of Science in Computer Science and a graduate from AiCore specialised in Data Engineering. Seeking a Data Engineering role in a reputable company. Open to remote work and relocation.

Education

AiCore (Data Engineering)

2023 – 2024

Certified in the practical application of AI & Data Engineering using industry-standard tools including:

- Software engineering (Git & GitHub, advanced Python, algorithms & data structures).
- Data engineering (SQL, data lakes, data warehousing, web scraping).
- Cloud engineering (cloud computing, designing and building APIs, Docker, Apache Airflow, AWS Serverless Stack)
- Coordinated 3 industry-relevant projects throughout.

MSc (Data Science and Analytics), Brunel University, London

Certified Scrum Master®, (Credential ID: 000691196) Scrum Alliance,

MSc (Computer Science), University of Bristol

Projects

AWS Data Engineering Pipeline ([GitHub link](#))

2024

- Implemented an end-to-end AWS-based data engineering pipeline that leveraged both batch and stream processing to analyse over 30k rows of data.
- Provisioned an API Gateway RESTful API to leverage a Kafka REST proxy & send data to 3 different Kafka topics on MSK.
- Distributed batch data from MSK to an AWS S3 data lake using MSK Connect plugin-connector pair.
- Created custom Spark transformations on Databricks, to clean and aggregate data stored in the S3 data lake.
- Automated Databricks Workloads to run daily using custom made DAG scripts on MWAA.
- Streamed data in real-time using 3 AWS Kinesis data streams & performed near real-time analysis using Spark Structured Streaming on Databricks.

Multinational Data Centralisation Pipeline ([GitHub link](#))

2023

- Developed a system that extracts sales data from 5 data sources: PDF documents, AWS RDS database, RESTful API, JSON & CSV files.
- Created custom Python scripts to clean and transform over 120k rows of data, before loading the data into a PostgreSQL database.
- Designed a star-schema database by joining 5 dimensions tables, resulting in data querying time with < 1 millisecond latency.

- Designed >10 complex SQL queries to extract business insights from the data, including velocity of sales, yearly revenue and regions with the most sales.

WORK EXPERIENCE

BAE Systems plc, Portsmouth

Data Scientist

2021 – 2023

- Spearheaded the implementation of a streamlined data pipeline, leading to a significant 30% reduction in data processing time and enhancing overall operational efficiency.
- Engineered a sophisticated data cleansing algorithm, enhancing data accuracy by 20% and ensuring reliable insights for decision-making processes.
- Developed interactive dashboards, revolutionising data visualisation techniques and facilitating improved understanding and interpretation of complex datasets.
- Executed the migration of 5TB of sensitive data to a secure cloud environment, achieving 99.99% data availability and ensuring compliance with regulatory requirements.

Massive Analytic Ltd, London

Data Scientist

2020 – 2020

- Pioneered the development and deployment of an LSTM-based sequence prediction model, boasting an impressive prediction accuracy of 95%, surpassing previous models by 15% and demonstrating a commitment to cutting-edge data science methodologies.
- Led a multidisciplinary team in the design and implementation of a control unit for a self-driving car, resulting in 20% reduction in accidents during rigorous testing.
- Recognised for designing intuitive user interfaces and crafting engaging visualisations, enhancing client demonstrations and contributing to heightened satisfaction levels, underscoring a talent for communicating complex concepts and solutions.

Public Health England (PHE), Department of Health, Cambridge

Software Developer

2016 – 2018

- Enhanced public health decision-making through the design, development, and support of analytical systems, demonstrating a commitment to leveraging data for societal benefit ([website](#)).
- Leveraged technical proficiency in SQL Server, SSIS, and T-SQL to construct and maintain robust data warehouses and marts, facilitating efficient data storage and retrieval processes crucial for informed decision-making.
- Spearheaded improvements in system performance and quality by influencing testing and implementation processes, underscoring a dedication to continuous optimisation and innovation in data-driven initiatives within the public health sector.

SKILLS & OTHER

Tools: Python, SQL, AWS, Spark, Databricks, Kafka, Airflow, git, GitHub, SSIS, Docker,

Certifications: Databricks Certified Data Engineering Associate (in progress)

Data Science Project Portfolio: Overview of Data Science projects: https://emma-luk.github.io/docs/emma_luk_samples_of_Data-Science_work.pdf

Big Data SQL Project Portfolio: Overview of Project with Google BigQuery: <https://emma-luk.github.io/docs/project-with-BigQuery-emma-luk.pdf>

Web Development Project Portfolio: Overview of Web Development projects: https://emma-luk.github.io/docs/emma_luk_samples_of_web_development-work.pdf