Duy Doan *Data Engineer*

✉ duydoanHel@gmail.com
☎ 0415720533
📍 Helsinki
github.com/DuyDoan190203
linkedin.com/in/duy-doan-dataengineer

Profile

Started my data engineering journey during my BIT studies in Finland, where I discovered my passion for turning messy data into meaningful insights. Through hands-on experience at Innate AI and Listeds, I've learned that the best data solutions aren't just technically sound—they solve real business problems.

My technical toolkit includes AWS (Glue, Redshift, S3, Kinesis, Lambda), Apache Airflow, Kafka, Spark, Terraform, Docker, Kubernetes, dbt, Great Expectations, and both OLTP/OLAP systems.

Balancing my Arcada Big Data Analytics studies with hands-on practice—because I believe the best data engineers never stop learning. Currently excited about integrating ML into real-time systems and exploring how AI can enhance (not replace) human decision-making

Education

2025 – 2026
Helsinki, Finland

Big Data Analytics Specialization (Master Level)
Arcada University of Applied Sciences
Part-time specialization studies in analytical solutions, machine learning and data engineering implementation. Program includes:

- Machine Learning & Data Mining with real-world datasets
- Cloud Computing & Data Engineering using GPU-accelerated models
- Deep Learning & Foundation Models for big data processing
- AI-assisted Analytical Service Design with industry capstone project
- Data Visualization & Business Intelligence communication

2025/01 – 2025/06

AWS Data Engineering Specialization Program
AWS & Deeplearning.AI
Key Achievements:
+ Completed comprehensive specialization program covering advanced data engineering concepts
+ Hands-on experience with AWS data services (Glue, Redshift, S3, RDS, Lambda)
+ Mastered modern data stack including Apache Airflow, dbt, Terraform, and Apache Iceberg
+ Built production-grade capstone project demonstrating end-to-end data pipeline development
+ Specialized in data quality engineering, DataOps practices, and cloud infrastructure automation

2021 – 2025
Lahti, Finland

Business Information Technology
LAB University of Applied Sciences

Bachelor's degree combining business and technical expertise in information systems development. Specialized in full-stack development and data engineering.
Notable: Cross-institutional studies at multiple Finnish universities, practical training (internships) completed at Listeds and Innate AI, strong academic performance with excellent grades in programming.

Experience

2025/04 – 2025/07
Tampere, Finland

Data Engineering Intern
Listeds (Brainbites)

Built the data foundation platform for data-driven decision making at a fast-growing Finnish startup:

- Designed medallion architecture data platform using GCP and Infrastructure as Code, enabling real-time analytics for board and management reporting
- Developed ETL pipelines, implementing incremental loading and automated validation that reduced manual data processing by 80%
- Collaborated with data collection team to ensure seamless integration between web scraping workflows and cloud data storage
- Implemented comprehensive data quality frameworks, establishing monitoring systems that catch data issues before they impact business reporting

2024/11 – 2025/02
Helsinki, Finland

Data Engineering Intern
Innate AI (Pharmaceutical AI)

Built the foundation and architected cloud data infrastructure for pharmaceutical AI applications in Helsinki's growing healthtech scene:

- Built AWS-based data platform processing 5M+ clinical records with 99.9% accuracy using S3, Glue, and Redshift in medallion architecture
- Developed ETL pipelines with Apache Airflow for real-time ML model training, reducing data preparation time from days to hours
- Created Python validation frameworks, achieving 95% reduction in data quality issues across all healthcare datasets
- Expanded ML training data by 30% through automated integration of external healthcare APIs while maintaining strict HIPAA compliance

2024/04 – 2024/12
Helsinki, Finland

Software Engineer (Back-end)
BlueWave Data Solutions

- Developed Python-based REST APIs for data integrations.
- Managed PostgreSQL databases with SQL for schema design and queries, integrating with AWS S3 and Google Cloud.
- Built scalable server logic with Git for version control.

2024/02 – 2024/04
Espoo, Helsinki Finland

Software Engineer
Lomado Oy

- Built REST APIs for client data management systems
- Developed automated data processing pipelines using Python and PostgreSQL
- Created real-time analytics dashboard for business intelligence reporting
- Implemented microservices architecture supporting multiple client applications
- Built ETL processes for integrating data from various client systems

Certificates

Employment Certificate - Herizon Internship Program with Listeds(Brainbites)

Data engineering internship

AWS & DeepLearning.AI Professional Data Engineer

AWS & DeepLearning.AI

- Develop a mental model for the field of data engineering as a whole, including the data engineering lifecycle and its undercurrents.
- Learn a framework for approaching any data engineering project I work on so I can effectively create business value with data.
- Build my skill in the five stages of the data engineering lifecycle; including generating, ingesting, storing, transforming, and serving data.
- Learn the principles of good data architecture and apply them to build data systems on the AWS cloud.

Employment Certificate - Herizon Internship Program with Innate AI

Data engineering internship

Projects

DeFtunes Data Pipeline

Key Components:

- Architected and implemented an end-to-end data pipeline for music streaming analytics
- Designed medallion architecture with landing, transformation, and serving zones in AWS S3
- Orchestrated daily incremental loads with Apache Airflow for optimal data processing
- Implemented data quality checks and built analytical views using dbt and Redshift

- **Technologies Used:** AWS (Glue, S3, Redshift, IAM), Terraform, Python, SQL, dbt, Apache Iceberg, Apache Airflow, Data Quality frameworks
- **Outcomes:** Successfully delivered a production-ready data pipeline that processes music streaming and purchase data, enabling the analytics team to gain insights through optimized data models and visualization dashboards.

ML-Driven Mobility Service Pipeline

Developed an end-to-end data pipeline supporting ML models for ride duration prediction across three mobility service vendors. Key components included:

- Implemented automated data ingestion, preprocessing, and validation workflows using Apache Airflow
- Built data quality checks with Great Expectations to ensure reliable model inputs
- Deployed infrastructure-as-code using Terraform for reproducible environments
- Created streaming data processing with AWS Kinesis for real-time data handling
- Designed monitoring dashboards to track model performance metrics

Fraud Detection Pipeline and Platform

- Developed a real-time fraud detection system using Kafka, Python, and machine learning
- Implemented streaming data pipelines processing 1000+ transactions per minute
- Created Grafana dashboards for monitoring fraud metrics and visualizing alerts
- Integrated data quality checks using Great Expectations for transaction validation

Skills

Data Engineering

• Data Processing: Apache Spark, PySpark, AWS Glue, Kafka, Streaming ETL

• Data Storage: S3, Redshift, PostgreSQL, DynamoDB, Apache Iceberg

• Orchestration & Monitoring: Airflow, Terraform, Prometheus, Grafana

• Data Quality: Great Expectations, AWS Glue Data Quality, Data validation frameworks

• Cloud & Infrastructure: AWS (Certified), Infrastructure as Code, Docker, GCP

Languages & Tools

- **Programming:** Python, SQL, NoSQL, Typescript, React, C#
- **Analytics & Visualization:** BI Tools, Looker Studio, Dashboarding