



# Tom Lous

## Freelance Data & Software Engineer

e: tomlous@gmail.com

w: <https://lous.info>

m: +31645528510

---

### Summary

Freelance data engineer with focus on Functional Programming with Scala, Spark, Kafka & Kubernetes

### Work Experience

#### Senior Data Engineer at Eneco

*Apr 2024 -*

Building streaming data products with Python, Quarkus, Snowflake and dbt on Azure

#### Senior Software Engineer at DHL

*Oct 2022 - Apr 2024*

Building streaming data products with ZIO 2, Scala 3, Kafka and Cassandra on Azure Kubernetes

#### Lead Data Engineer at Schiphol

*Nov 2021 - Oct 2022*

Leading the data factory team and implementing scalable data ingestion solutions for a data mesh architecture using Spark, Scala, ZIO, Databricks, Kafka and Kubernetes (OpenShift)

#### Principal Engineer at Nike

*Jun 2021 - Nov 2021*

Data Engineering & Architecture for Nike's EMEA In Season Optimization data & ML products

#### Lead Data Engineer at Shell

*Jun 2019 - Jun 2021*

Designing and building the data architecture for a brand new global agile data team at Shell's Agile Hub using Spark, Kafka, Kubernetes, Airflow and Hadoop on top of Azure and AWS

#### Trainer Data Engineering & Data Science at Young Mavericks

*Oct 2020 -*

Host occasional trainings for future data engineers & scientists: DevOps for Data Engineers, Functional Programming with Scala, Building & Deploying Spark Applications

#### Senior Data Engineer at VodafoneZiggo

*Apr 2019 - May 2019*

Part of the Advanced Analytics Platform (AAP) and Technical Passport (TP). (Py)Spark, Hive, Oozie & Hadoop development.

## **Big Data Engineer at eBay**

*Jun 2018 - Apr 2019*

Data ingestion as a service (Kafka, Hadoop, Kubernetes, Scala) @ eBay's PE (Platform Engineering) team. Spark, Scala, Flink, Hadoop, Cassandra & Machine Learning @ eBay's CData (Central Data) team.

## **Big Data & Machine Learning Engineer at USponsor me**

*May 2018 - Jan 2020*

Part-time remote freelance contract for building a scalable Spark data ingestion, cleaning & deduplication pipeline on AWS with Spark, Scala, SparkML, GraphX & MongoDB

## **Big Data Software Engineer at Datling**

*Apr 2016 - Jun 2018*

Developed & implemented big data pipelines with machine learning and continuous ingestion in the cloud, using Apache Spark, Scala and a plethora of other tools, like: Airflow, Hadoop, Docker, Kubernetes, Elasticsearch, MongoDB, etc.

## **Operations Manager (Data & IT) at Datling**

*Mar 2014 - Apr 2016*

Managed IT & BI department, maintaining and developing core products for Datling, like Salesmapp, Data Outlet, Location Data Hub, etc. Helped hands-on by developing quick search and other tools. Why I left management: <https://www.linkedin.com/pulse/re-becoming-developer-tom-lous/>

## **Manager Development at dpdk**

*May 2010 - Feb 2014*

Introduced agile workflow and managed developers across multiple multi disciplinary SCRUM teams. Led the way for new innovations and helped develop web & mobile applications, hands on, when needed.

## **Technical Lead / Sr. Web Developer / Software Engineer at Mindwarp Internet Solutions**

*Nov 2003 - Apr 2010*

Full-stack web development, mainly LAMP stack, for a range of clients across a range of platforms.

## **System Administrator & Support at TOPXS.nl**

*Nov 2003 - Apr 2010*

Built and maintained web hosting systems for a range of internal and external clients. Mainly Debian linux.

## **Owner at GraphIQ**

*Feb 2001 - Jan 2007*

Owner of web development company GraphIQ Smart Design.  
Developed many small scale websites.

## Software Engineer & System Administrator at HydroLogic

*Feb 2003 - Nov 2003*

VB6/ASP developer for in house software product HydroNet.

## Freelance Java Developer at DotMachine

*Jan 2003 - Feb 2003*

Freelance Java developer for feedback system

### Key Skills

Big Data	Kubernetes
Scala	Apache Hadoop
Python	Machine Learning
FP	ZIO
Apache Spark	Cats
Apache Kafka	Terraform

### Education

## Artificial Intelligence at Vrije Universiteit Amsterdam

*1999 - 2002*

Cognitive Science, Human Ambience, Intelligent Systems Design, Webscience from philosophy, logic and psychology, to information science, linguistic analysis and mathematics. To explore knowledge acquisition and modelling, multi-agent systems and techniques to make internet searches more efficient and effective.

## Evansville High School

*1996 - 1997 High School Diploma*

Evansville High School, Wisconsin, USA. Exchange program.

## Comenius College te Capelle a/d IJssel

*1993 - 1999 VWO Diploma*

Dutch high school.

### Projects

#### Data Factory Components

*@Schiphol*

- Setup modularized components building data pipelines on the fly. Moving away from costly Databricks jobs for simple data ingestion, to lightweight GraalVM, Scala & ZIO containers on Kubernetes, feeding into the Kafka pipelines

#### Workhorse (Spark & Airflow as a Service)

*@Shell*

- Build Spark & Airflow as a service capability on top of auto-scalable Kubernetes cluster in Azure. Based on low config CI/CD pipelines scala & python spark jobs are transformed into docker images that are referenced in auto generated helm charts, which can be (helm) deployed using Airflow KubePodOperator task. These helm installs create custom Spark clusters (based on definitions in the helm chart) and run the batch jobs until completion.

#### Data Ingestion as a Service (DIaaS)

<https://www.ebayclassifiedsgroup.com/>

@ebay

- Build a service that will allow ECG (eBay Classifieds Group) platforms to spin up managed data ingestion pipelines on the ECG cloud. These data ingestion pipelines will support AVRO, JSON or schemaless events through an HTTP interface and will provide validation, routing and anonymization services. A schema registry will be provided to help ECG platforms with their data schema management.
- Kafka Streams & Connect Pipelines (Scala & Java)
- Schema registry for JSON & AVRO schema's
- HTTP proxy (Scala & Akka) for event ingestion and posting on a Kafka Topic
- Scalable and fully managed Kafka cluster per data ingestion as a service instance on Kubernetes
- Event validation against the data schema the end user defines.
- Event anonymization that is compliant with GDPR regulations.
- Event routing to Kafka and HDFS data sinks.
- Configurable & parameterizable Kafka topic names and HDFS paths.
- JSON to AVRO conversion for Kafka Connect
- Monitoring using Prometheus

## Importer Pipeline

<https://usponsorme.com/en/>

@Usponsorme Ingesting huge XLSX files into a MongoDB, cleaning, merging and restructuring the data on the fly.

- Scalable Spark Ingestion of Excel files
- Cleaning & Matching data
- LSH & other deduplication algorithms
- Deployment on AWS
- MongoDB integration

## Scalable Geocode Quality Assurance

@Datlinq Datalabs

- Check if geocoded locations are within the geo boundaries of postal code area
- Read ESRI & geojson shape files in Spark Dataframe
- Join shapes with location dataset on postal code
- Do point-in-polygon and mark misses to be rechecked
- In parallel on Spark cluster

## Location API

@Datlinq Datalabs

- Functional scalable backend with Scala `http4s` webserver with `blaze` (very fast async NIO microframework and Http Parser) & `rho` (self documenting swagger DSL).
- Pure Functional MySQL database access via `doobie` and `cats`
- Google Cloud SQL backend
- RESTful API deployed via Docker
- Deployed auto scalable Kubernetes cluster with sql cloud proxy and Google Cloud Endpoints to manage in- and outbound connections to container.
- Authentication & Authorization set up via Auth0 non interactive Auth0 clients. Enforced via autogenerated Openapi in google cloud endpoints

## Spark Big Data Pipeline

@Datlinq Datalabs

- Continuously ingest data from many sources
- Preprocess & clean data via Spark jobs
- Cross Match data from multiple sources creating record links between sources, using Elasticsearch & Spark
- Combine sources and store data into Hadoop, MySQL & Elasticsearch in cloud
- Run these jobs idempotently every night using Airflow
- Monitor & Log using StackDriver

## Certifications

### Data Science Specialization | Johns Hopkins

<https://www.coursera.org/account/accomplishments/specialization/6KP9TYFCDYCN>

### Data Science Capstone | Johns Hopkins

<https://www.coursera.org/account/accomplishments/records/7QFTSWDLJ54Y>

## Scalable Microservices with Kubernetes

<https://www.udacity.com/course/scalable-microservices-with-kubernetes--ud615>

## Developing Data Products | Johns Hopkins

<https://www.coursera.org/account/accomplishments/records/HHLH8CKCDNTB>

## Introduction to Apache Spark | University of California, Berkeley

<https://courses.edx.org/certificates/4eba607c3a1046a296ea867cc1fe6402>

## Parallel Programming in Scala | École Polytechnique Fédérale de Lausanne

<https://www.coursera.org/account/accomplishments/records/NMHLPXLLMBKJ>

## Functional Programming Principles in Scala | École Polytechnique Fédérale de Lausanne

<https://www.coursera.org/account/accomplishments/records/WNZW9WYRMB4J>

## Functional Program Design in Scala | École Polytechnique Fédérale de Lausanne

<https://www.coursera.org/account/accomplishments/records/AYFZPPPZCBZU>

## Implementing Predictive Analytics with Spark in Azure HDInsight

<https://courses.edx.org/certificates/5394fca54e704c84991f7113f82613ad>

## Practical Machine Learning | Johns Hopkins

<https://www.coursera.org/account/accomplishments/records/37TBFKURE45U>

## Regression Models | Johns Hopkins

<https://www.coursera.org/account/accomplishments/records/7LYPKSQMDA2M>

## Introduction to Big Data | University of California, San Diego

<https://www.coursera.org/learn/intro-to-big-data>

## Implementing Real-Time Analytics with Hadoop in Azure HDInsight

<https://courses.edx.org/certificates/user/5863662/course/course-v1:Microsoft+DAT202.2x+2T2016>

## Machine Learning | Stanford University

<https://www.coursera.org/course/ml>

## Statistical Inference | Johns Hopkins

<https://www.coursera.org/account/accomplishments/certificate/UANQPEUHBUB>

## Reproducible Research | Johns Hopkins

<https://www.coursera.org/records/Pa7r3a6CnypvbDP6>

## Exploratory Data Analysis | Johns Hopkins

<https://www.coursera.org/records/qwEvC8db2J9pJgpW>

## Getting and Cleaning Data | Johns Hopkins

<https://www.coursera.org/records/BR6jvmQb7w32XVRN>

## The Data Scientist's Toolbox | Johns Hopkins

<https://www.coursera.org/records/pKKcqAF6Vp3rEJh6>

## Introduction to Operations Management | University of Pennsylvania

<https://www.coursera.org/records/CbGv6cvdkQVNBKUu>

## Computing for Data Analysis | Johns Hopkins

<https://www.coursera.org/signature/certificate/GZ4CZ2ZESQ>

## Oracle Certified Professional, MySQL 5 Developer

<https://www.dropbox.com/s/v8azkt6s4fc7wev/Oracle%20Certified%20Professional%2C%20MySQL%205%20Developer.p>

Titanium Certified Application Developer (TCAD)

<http://training.appcelerator.com/assets/datasheet/tcd-certification-objectives.pdf>

Professional Scrum Master I

<https://www.scrum.org/>

Zend Certified Engineer, PHP 5

<http://www.zend.com/en/services/certification/>

Object Orientated Foundation (OOF)

<https://www.dropbox.com/s/79rjeeyn9kmf1q8/EXIN%20Object%20Ori%C3%ABntatie%20Foundation%20%28OOF%29.p>

IT Management Foundation (ITMF)

<https://www.dropbox.com/s/fc5sii2prb46xm6/EXIN%20IT%20Management%20Foundation%20%28ITMF%29.pdf>

Infrastructure Management Foundation (IMF)

<https://www.dropbox.com/s/awp5fcbzguisr0b/EXIN%20Infrastructure%20Management%20Foundation%20%28IMF%29.p>

Big Data Analysis with Scala and Spark | École Polytechnique  
Fédérale de Lausanne

<https://www.coursera.org/account/accomplishments/records/K4FKMHRNP52M>

Cryptography I | Stanford University

<https://www.dropbox.com/s/0xc4gb4tuevjwqw/Coursera%20crypto%202017.pdf?dl=0>

Functional Programming in Scala Capstone | École Polytechnique  
Fédérale de Lausanne

<https://www.coursera.org/account/accomplishments/records/PSMS3GZWVRJS>

Functional Programming in Scala | Specialization | École  
Polytechnique Fédérale de Lausanne

<https://www.coursera.org/account/accomplishments/specialization/U6AVP3GNVJUM>

Google Cloud Platform Fundamentals: Core Infrastructure

<https://www.coursera.org/account/accomplishments/records/8VXYNDLQZNEK>

Neural Networks for Machine Learning | University of Toronto

<https://www.coursera.org/account/accomplishments/records/WQENQUSY4GJE>

## Courses

Building Distributed Pipelines for Data Science using Kafka,  
Spark, and Cassandra (@O'Reilly)

Scrum Training for Scrum Masters (PSM I) (@iSense)

Creative Scala Workshop (@underscore.io)

Microservices Masterclass (@Trivento)

Advanced: Exploring Wikipedia with Spark (@GoDataDriven /  
Spark Summit)

Understand and Apply Deep Learning with Keras, Tensorflow, and  
Apache Spark 2.x. (Spark Summit)

Google Cloud Fundamentals: Big Data & Machine Learning  
(@Google)