# Tom Lous

Freelance Big Data & Machine Learning Software Engineer | Scala | Spark | Airflow

e: tomlous@gmail.com
w: https://lous.info
m: +31645528510

## Summary

Spark & Scala, all day, every day.
I've been coding since I bought my first programming book in 1990 and have been learning new tools and languages continuously. I'm now looking for freelance contracts where I can add my years of knowledge & experience to exciting projects and learn new skills from teammates.

## Work Experience

### Big Data Engineer - Contract at eBay

*Jun 2018 - Present*

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.

### ZZP - Freelance Big Data & Machine Learning Software Engineer at GraphIQ

*May 2018 - Present*

Developing Scala software that runs on a Spark cluster, dockerise microservices to run on a Kubernetes cluster, ingest data via Kafka. Proficient with many tools concerning setting up a big data ingestion & processing pipeline in the cloud and deploying the results via a scalable API. Cleaning & analysing huge amounts of data, followed by training, validating & testing machine learning models and deploying them in production.

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

*May 2018 - Present*

Part-time remote freelance contract for building a scalable Spark data ingestion, cleaning & deduplication pipeline on AWS.

### Big Data Software Engineer at Datlinq

*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 Datlinq

*Mar 2014 - Apr 2016*

Managed IT & BI department, maintaining and developing core products for Datlinq, 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 Smart Design

*Feb 2001 - Jan 2007*

Owner of web development company. 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
Scala
Apache Spark
Machine Learning
Artificial Intelligence
Shell Scripting
Software Development
Scrum
IT Management
IT Strategy
Operations Management
Business Intelligence
Web Development
Mobile Applications
SEO
Hadoop
Google Cloud Platform
Docker

## 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 Ingestion As A Service (DIAAS)

*Jul 2018 - Present* 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 HFDS paths.*
- *JSON to AVRO conversion for Kafka Connect*
- *Monitoring using Prometheus*

### Cassandra Export Tool

*Jun 2018 - Present* https://www.ebayclassifiedsgroup.com/

*@ebay*
- *Export TB's of data out of Cassandra*
- *Store to HDFS using Spark.*
- *Generate/update associated Hive tables*
- *Daily run job via Jenkins, Luigi*
- *Cleanup job for retention policy compliant with GDPR*
- *Monitoring using Prometheus*
- *Security via Kerberos*

### Importer Pipeline

*May 2018 - Present* 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

*Feb 2018 - Present*

*@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*

### Annotation Tool

*Nov 2017 - Present*

*@Datlinq Datalabs*
- *UI built in AngularJs 4 on top of Scala http4s backend*
- *Manual validation of machine learning predictions*
- *Manual annotation of labels to data*
- *Automated configurable inter annotator agreement based on samples. (Some records are checked by multiple persons, if they disagree a third person is also consulted)*
- *OAuth authentication & authorization via Auth0 and custom scopes per account*
- *API documented in Swagger2 spec*
- *Data stored in Google Cloud SQL instance*

### Location API

*Sep 2017 - Present*

*@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*

### Scalafiniti OSS

*Aug 2017 - Present* https://github.com/datlinq/scalafiniti

*@Datlinq Datalabs*
- *OSS*
- *A Scala wrapper for Datafiniti API*

## Certifications

### Neural Networks for Machine Learning | University of Toronto

*Feb 2018* https://www.coursera.org/account/accomplishments/records/WQENQUSY4GJE

### Google Cloud Platform Fundamentals: Core Infrastructure

*Jul 2017* https://www.coursera.org/account/accomplishments/records/8VXYNDLQZNEK

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

*Apr 2017* https://www.coursera.org/account/accomplishments/records/PSMS3GZWVRJS

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

*Apr 2017* https://www.coursera.org/account/accomplishments/specialization/U6AVP3GNVJUM

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

*Mar 2017* https://www.coursera.org/account/accomplishments/records/K4FKMHRNP52M

### Data Science Specialization | Johns Hopkins

*Jan 2017* https://www.coursera.org/account/accomplishments/specialization/6KP9TYFCDYCN

### Data Science Capstone | Johns Hopkins

*Jan 2017* https://www.coursera.org/account/accomplishments/records/7QFTSWDLJ54Y

### Scalable Microservices with Kubernetes

*Aug 2016* https://www.udacity.com/course/scalable-microservices-with-kubernetes--ud615

### Developing Data Products | Johns Hopkins

*Jul 2016* https://www.coursera.org/account/accomplishments/records/HHLH8CKCDNTB

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

*Jul 2016* https://courses.edx.org/certificates/4eba607c3a1046a296ea867cc1fe6402

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

*Jun 2016* https://www.coursera.org/account/accomplishments/records/NMHLPXLLMBKJ

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

*Jun 2016* https://www.coursera.org/account/accomplishments/records/WNZW9WYRMB4J

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

*Jun 2016* https://www.coursera.org/account/accomplishments/records/AYFZPPPZCBZU

### Implementing Predictive Analytics with Spark in Azure HDInsight

*May 2016* https://courses.edx.org/certificates/5394fca54e704c84991f7113f82613ad

### Practical Machine Learning | Johns Hopkins

*May 2016* https://www.coursera.org/account/accomplishments/records/37TBFKURE45U

### Regression Models | Johns Hopkins

*May 2016* https://www.coursera.org/account/accomplishments/records/7LYPKSQMDA2M

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

*Dec 2015* https://www.coursera.org/learn/intro-to-big-data

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

*Dec 2015* https://courses.edx.org/certificates/user/5863662/course/course-v1:Microsoft+DAT202.2x+2T2016

### Cryptography I | Stanford University

*Jul 2015* https://www.dropbox.com/s/0xc4gb4tuevjwqw/Coursera%20crypto%202017.pdf?dl=0

### Machine Learning | Stanford University

*Apr 2015* https://www.coursera.org/course/ml

### Statistical Inference | Johns Hopkins

*Feb 2015* https://www.coursera.org/account/accomplishments/certificate/UANQPEUHBU

### Reproducible Research | Johns Hopkins

*Dec 2014* https://www.coursera.org/records/Pa7r3a6CnypvbDP6

### Exploratory Data Analysis | Johns Hopkins

*Aug 2014* https://www.coursera.org/records/qwEyC8db2J9pJgpW

### Getting and Cleaning Data | Johns Hopkins

*Aug 2014* https://www.coursera.org/records/BR6jvmQb7w32XVRN

### The Data Scientist’s Toolbox | Johns Hopkins

*Jul 2014* https://www.coursera.org/records/pKKcqAF6Vp3rEJh6

### Introduction to Operations Management | University of Pennsylvania

*May 2014* https://www.coursera.org/records/CbGv6cvdkQVNKCUu

### Computing for Data Analysis | Johns Hopkins

*Feb 2014* https://www.coursera.org/signature/certificate/GZ4CZ2ZESQ

### Oracle Certified Professional, MySQL 5 Developer

*Feb 2014*

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

### Titanium Certified Application Developer (TCAD)

*Oct 2013* http://training.appcelerator.com/assets/datasheet/tcd-certification-objectives.pdf

### Professional Scrum Master I

*Apr 2011* https://www.scrum.org/

### Zend Certified Engineer, PHP 5

*Oct 2010* http://www.zend.com/en/services/certification/

### Object Orientated Foundation (OOF)

*Apr 2008*

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

### IT Management Foundation (ITMF)

*Apr 2008*

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

### Infrastructure Management Foundation (IMF)

*Apr 2008*

*https://www.dropbox.com/s/awp5fcbzguisr0b/EXIN%20Infrastructure%20Management%20Foundation%20%28IMF%29.pdf*

## 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)