# DUSTIN SMITH

## Solutions Architect–Data Engineering

@ Dustin.William.Smith@gmail.com  +44 7788 76 2345  London, UK  linkedin.com/in/dustin-s-photo
github.com/dwsmith1983  https://www.dustinsmith.info

## EXPERIENCE

### Solutions Architect

**Databricks**

June 2023 – Present  London, UK

- Support customers within the Databricks ecosystm

### MLOps Manager

**Delivery Hero**

July 2022 – April 2023  Berlin, Germany

- Plan MLOps roadmap and OKRs
- Grow, lead, and mentor the MLOps team
- Manage, lead, mentor my team of engineers so they can grow in their careers
- Meet with stakeholders in data science, machine learning, site reliability, cloud partners, senior managers, and directors

### Associate Director–MLOps Lead

**True Digital Group**

March 2019 – June 2022  Bangkok, Thailand

- **Analytics & AI / Platform Operations  August 2021 to June 2022**
  - Working with our corporate client AIA and their McKinsey consultants to implement MLOps on our on premise k8s cluster to deploy AIA models on our hybrid framework once complete.
  - Supporting Ascend credit models for our cloud framework.
  - Leading the MLOps team and development for an on premise and cloud infrastructure utilizing Tensorflow MLMD, TFX, MLflow, and a framework agnostic design in order to be able to deploy Tensorflow, Spark, H2O, ... workflows.
  - Support internal and external data science on machine learning pipelines, productionization, code quality, deployment, and monitoring.
- **Analytics & AI / Analytics Platform  January 2020 to August 2021**
  - Transitioning to the newly formed MLOps team May 2021 and working on on premise and cloud based MLOps hybrid infrastructure.
  - Identified an optimized way to compact our HDFS data resulting in a $\sim 75\%$ decrease in storage space and decrease of $\sim 70 - 90\%$ in query run time on our large tables; actualized storage saving of 2.331 PB. This results in an estimate savings of ฿ $20$ to $45$ million per year at our current on premise storage costs.
  - Optimized a Kafka geo streaming campaign by moving cell site geo triggers up the stack resulting in increased triggering capacity to 630 locations compared to 50
- **Analytics Business / Analytics Product  March 2019 to January 2020**
  - Created a Docker container so Windows user could easily test their PySpark code with Python 3.6.6 or 3.7.5 and Spark 2.4.4
  - Worked with our strategic partner, Eureka AI, on the development of an automated customer segmentation engine based on direct packet inspection data
  - Developed a way to track users on the Bangkok Mass Transit System in the absence of triangulation

## SKILLS

- Python / PySpark ●●●●●
- Scala / Spark ●●●●○
- Bash ●●●○○
- SQL ●●●●●
- Markdown ●●●●●
- Linux / iOS / Windows ●●●●●
- Docker ●●●●○
- LaTeX ●●●●●
- Git ●●●●●
- Kubernetes ●●●○○
- Airflow ●●●○○

## CLIFTON STRENGTHS

Learner | Competition | Achiever | Deliberative | Consistency | Analytic | Intellection | Adaptability | Input and Command

## EDUCATION

### M.Sc. in Mechanical Engineering

**University of Vermont**

January 2015

### B.Sc. in Mathematics

**University of Tampa**

May 2011

### B.Sc. in Finance

**University of Tampa**

May 2011

# EXPERIENCE

## Lead Data Scientist

**AppSmiths Technology**

May 2018 – February 2019 Bangkok, Thailand

- While in Houston, Texas, I managed the software team deploying a web based version of WinGLUE
- Implemented agile software development practices
- Deployed a small Spark cluster to process the data
- Developed a digital oil field strategy

## Data Analyst–Contractor for Caterpillar Inc

**Tata Technology**

September 2017 – April 2018 Bangkok, Thailand

- Completed the development of high altitude engine tuning
- Abstracted the graphical user interface (GUI) in order for it manage multiple engine profiles
- Completed the development of fuel derate optimization for high altitude engine tuning

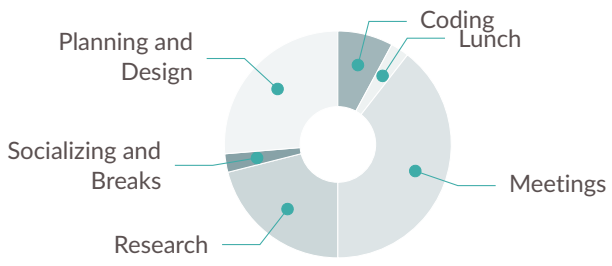
## Data Scientist III

**Caterpillar Inc**

January 2016 – September 2017 Champaign, Illinois

- **Optimization and Advanced Analytics June 2016 to September 2017**
  - Developed a Python based optimization GUI for high altitude engine tuning so the engineers were not relying on Excel; $75\%$ decrease in run time over Excel
  - Developed fast optimization for determine the highest altitude each engine speed could operate at without hardware failure
  - Developed a neural network for engine parameter prediction used in altitude optimization
  - Second developer on an integrate advanced design of experiments software utilizing Sandia Labs Dakota software; the alpha release provided an after tax risk adjusted net present value of $\$1.9$ million dollars for a single business unit
- **Information Analytics January 2016 to May 2016**
  - Analyzed drill performance for Martin Marietta Spec Aggregate Quarry when they switched from Atlas Copco to Caterpillar
  - Improved life cycle management tracking of medium track type tractors which lead to data issue identification in our Teradata and service oil data
  - Wrote a tutorial on using Google's reverse geocoding API to update the GPS location issue found during the life cycle management project

# AN AVERAGE WORKING DAY




# WRITING

**Data Optimization for Compacted Partitions**

Created a semi-linear and z ordering data optimization for compacting partitions which lead to a 2.311 PB storage savings as of 2021 July. We are currently still rolling this out on large data sources at True Digital Group. The blog post was published on both the company's tech blog and Towards Data Science on Medium.

- Data Optimization–True Analytics Tech Blog

**Configuration Files in Python using dataclasses**

A demonstration on how we at True Digital Group use dataclasses and dataconf to parse configuration. dataconf is similar to the library of pureconfig in Scala.

- Configuration Files in Python using dataclasses–True Analytics Tech Blog

**Capital Budgeting with Monte Carlo Simulations in Python**

Write up of our on premise to cloud capital budgeting exercises when we decided to move from on premise to cloud. Numbers are fictitious to demonstrate how we evaluated such as move.

- Monte Carlo Simulations–True Analytics Tech Blog

# CERTIFICATIONS

**Credentials**

Accredible Credential.net

# PROJECTS

**dataconf**

Working as a collaborator on dataconf where I added similar behavior to Scala case class parsing for Python dataclasses when parsing config files. As well as created a demo repository demo-dataconf.