Jakub Waller

reachout@jakubwaller.eu
https://www.jakubwaller.eu/
https://www.linkedin.com/in/jakubwaller/

Work experience

12/2022 – 08/2023	**Career Break** at World Trip
11/2017 – 11/2022	**Data Scientist/Engineer** at Qimia GmbH Main occupation: Working on various projects involving Data Science, Machine Learning and Data Engineering topics using a wide range of technologies.
02/2017 – 10/2017	**Research Scientist** at the Czech Technical University in Prague; Faculty of Information Technology Main occupation: Comparing various architectures of artificial neural networks on a set of time series classification data sets
03/2015 – 10/2016	**Java & JavaScript Developer** at Mibcon a.s. Main occupation: Programming SAP portal applications in Java, AngularJS and HTML for a web portal for ČEZ Distribuce, a. s.
04/2013 – 10/2017	**Network Administrator** at the Charles University in Prague; First Faculty of Medicine Main occupation: Maintenance of computers and other technical devices at the Institute of Immunology and Microbiology
07/2013 – 06/2014	**Bioinformatics Analyst** at the National Institute of Public Health Main occupation: Extracting information from rRNA using various bioinformatics software

Education

2014-2017	**Master of Science in Informatics** at the Czech Technical University in Prague; Faculty of Information Technology Study Field: Knowledge Engineering; Main Topics: Pattern Recognition, Data Mining Algorithms, Data Preprocessing, Enterprise Data Warehouse Systems, Parallel Algorithms Master's Thesis: "Time Series Classification with Artificial Neural Networks" (B/1.5)
05/2016 – 09/2016	**Exchange Semester** at the University of Waterloo, Canada Main topics: Artificial Intelligence, Forecasting
08/2014 – 01/2015	**Exchange Semester** at the Tallinn University of Technology, Estonia Main topics: Robotics, Malware, Analysis of Programming Languages
2011 – 2014	**Bachelor of Science in Informatics** at the Czech Technical University in Prague; Faculty of Information Technology Study Field: Computer Science; Main Topics: Programming Languages and Compilers, Algorithms, Operating Systems, Database Systems, Security, Artificial Intelligence Bachelor's Thesis: "Simulation of a Quantum Particle on a Twisted 2D Waveguide" (B/1.5)

Language Skills

Czech	Native speaker
English	C1 (fluent in spoken and written)
German	B2 (good intermediate knowledge)

Project Overview

Industry/Role/Date	Project Description	Assignments	Technologies
TV Data Scientist 01/2022-11/2022	Predicting users' age and gender based on their watching behaviour.	• Collect, process and analyse labelled data of several thousand households and their watching history in Sagemaker Notebooks and Athena. • Design features that would work both in the training set and the test set (different data sources) and iteratively train many machine learning models to find what features have the best predictive value. • Run hyperparameter optimisation on the models to improve accuracy. • Implement the training pipeline: Data processing, feature engineering, model training, model evaluation, using Glue, Batch, Metaflow and MLFlow. • Implement the inference pipeline: Data processing, feature engineering, model inference, prediction aggregation, using Glue, Batch, Metaflow and MLFlow. • Deploy both pipelines into production using code commit and code pipeline. • Monitor all jobs.	Python, PySpark, pandas, scikit-learn, xgboost, metaflow, mlflow, Athena, S3, Glue, Glue Crawler, Glue Catalog, Sagemaker, Batch, Code Commit, Code Pipeline, git, SQL, Jupyter notebook, pytest, mock, optuna, deepchecks
Logistics Data Engineer 11/2020-12/2021	On-prem Data Integration to Azure Data Lake via Talend; Event-driven Data Processing ETL Framework; Event Grid and Event Hubs stream processing using Spark Structured Streaming on Databricks; Semantic Data Model Warehouse Management Systems Data; Machine Learning models for Transport and Warehouse Optimization use-cases	• Investigate and improve a pipeline design initially based on Azure Function • Implement a Spark Structured Streaming job on Databricks connected to an Event Hub distributing the events between several batch jobs • Implement a Pyspark batch job on Databricks to process the data from an Azure data lake storage landing zone to output format (parquet) enabling data analytics via Synapse and Azure Data Explorer	Python 3.8, Apache Spark 3.1.1, Azure Databricks 8.1, Talend Cloud Data Integration 7.3, Talend Studio 7, Azure Data Lake Storage Gen2, Power BI, Azure Event Hubs, Azure Event Grid, Azure Functions, Azure SQL, Azure Python SDK, Azure Synapse Analytics, Synapse Serverless SQL Pool, Azure Data Explorer (Kusto), Azure Applications Insights, Azure Log Analytics, Azure Monitor, Azure DevOps, Terraform, Oracle DB, MySQL

		• Use Synapse and Azure Data Explorer to verify the results with the business users • Analyse the data and design and implement a semantic layer in Databricks enabling further data science • Design a machine learning pipeline which provides real-time predictions via an API • Monitor data migration using Power BI dashboards	
Data Analytics, Cloud SaaS Product Data Engineer 07/2020-10/2020	Cloud Big Data Migration Tool; Low Latency, High Throughput Multi Data-Source Data Migration; Reactive Non-Blocking IO Architecture	• Importing data From RDBMS (SQL Database) to Cloud Storage • Reactive Non-blocking parallel data Ingestion using Vert.x Async SQL Clients • Non-blocking async read-write from and to file using Vert.x Filesystem API • Design and Implementation of Configuration DSL as Java API and YAML • Time based Partitioning of Transactional data (tables) • Master Data, Dimensional data continuous Change-Capture implementation • Multi Cloud Storage: Local (POSIX) filesystem, AWS S3, Azure Blob Storage	Java 11, SQL, JDBC, Vert.x 3.9.3, Vert.x SQL Client, AWS, Azure, Google Cloud, PostgreSQL, MySQL, MariaDB, MS SQL, Parquet, Avro, JSON, CSV, S3, Azure Blob Storage, Google Cloud Storage, Reactive Programming, Non-blocking IO, Vert.x Async SQL Client, YAML, Builder Pattern, Picocli, SnakeYAML, Maven
FinTech Machine Learning Engineer 10/2019-06/2020	AWS Data Lake and DWH; Spark DAG ETL-Pipelines on EMR; Redshift DWH Development; Machine Learning Models and Inference API; Automation using CloudFormation and Python SDK	• Data Exploration and Feature-Engineering for Machine Learning Models • ETL Pipeline structure identification, establishment, and optimization • Data Warehousing with Spark ETL Jobs and Redshift Spectrum Scripts • Inference Classification Model feature engineering, training, evaluation, and deployment • Inference Regression Model feature engineering, training, evaluation, and deployment • Statistical analysis of existing historical data and new test data • Comprehensive data science training for new Client Staff Data Scientist	Python, xgboost, Scala, Apache Spark, Spark SQL, RDS Aurora, AWS Glue, Glue Crawler, Glue Catalog, AWS S3, AWS EMR, AWS Redshift, Redshift Spectrum, Jupyter Hub, MLeap, Spark ML, Play Framework, CloudFormation, CloudWatch, AWS IAM, Boto3, EC2, Hive, Livy, Hue

AdTech/Marketing Data Scientist 06/2019-09/2019	Azure Data Migration; Data Lake Creation, ETL, and Data Warehousing; Data Analytics	• Discovery, cleansing, and use case analysis of data from mixed formats and sources • Data preparation and integration into unified storage file format • Azure Data Lake integration and creation • Extract, Transform, Load processing featuring MySQL • Performance metric evaluation and analysis • Data Mart Creation with MS SQL and Power BI • Data analysis of campaign efficiency of various metrics • Discovery and presentation of future application integration and automation	Python, Pandas, Matplotlib, Jupyter, Scala, Apache Spark, Spark SQL, Azure Databricks, Azure SQL Data Warehouse, MySQL, MS SQL, IntelliJ, SBT, PowerBI
Logistics Software Engineer 04/2019-05/2019	Design and Implementation of an Optimisation Engine; Architecture and Implementation of a Server-Client Solution	• Exploration, analysis and benchmarking of optimisation frameworks • Implementation of a Java Spring Boot server side based on Kafka and PostgreSQL • Deploying services with docker-compose • Implementation and testing of React-based web and Android clients	Java, Spring Boot, Docker, PostgreSQL, OSRM, Kafka, Hasura, Maven, Gradle, Google Cloud
Electric Utility Machine Learning Engineer 02/2018-03/2019	Predictive Maintenance ML Models; Azure Cloud Data Lake Development; Azure Big Data Engineering	• Batch processing, ETL pipelines with Scala Spark and PySpark • Deploying jobs on a YARN cluster HDInsight using Gitlab CI/CD • Data processing and analysis with ArangoDB • Working with HDFS data on Azure data lake • Data exploration using Jupyter • ML model feature transformation pipeline using Spark and Pandas • ML model training using xgboost	Python, Jupyter, Pandas, xgboost, Azure HDInsight, Spark, PySpark, Spark SQL, Azure CLI, Scala, SQL, Hadoop HDFS, YARN, MSSQL, ArangoDB, Docker, Gitlab CI/CD
Post Data Engineer 11/2017-01/2018	Implementation of Cloud Real-time Tracking Platform on Kafka; Deployment of Micro-services on a Rancher Docker Platform	• Kafka-Streams Micro-service implementation using Avro Schemas and Confluent Schema-Registry	Java, Kafka, Confluent, Kafka-Streams, OrientDB, Elasticsearch, Spring Boot, Git, GitLab, Rancher, Docker, Docker-compose

		• Installation of Rancher and Docker • Confluent Enterprise Docker cluster with Zookeeper, Kafka Broker, (Avro) Schema-registry, Kafka-connector and Control-centre locally (Docker compose) and to cloud with Rancher • Spring Boot, Spring Data DAO implementation for Database connection • Use of OrientDB and OrientDB Graph as Persistence Storage	
Research Machine Learning Scientist 02/2017-10/2017	Time Series Classification with Artificial Neural Networks	• Studying and describing three state-of-the-art architectures of artificial neural networks • Identifying their theoretical differences • Designing and implementing an experimental procedure including an automatic optimization of hyper parameters • Generating time series classification data sets • Comparing the networks on these data sets	Python, Docker, Keras, Pandas, Jupyter
Agriculture Data Scientist 11/2016-01/2017	Data Analysis	• Data ingestion from Relation DB sources using MS SQL • Data cleaning • Data transformation using Pandas • Data analysis using Matplotlib	Python, Jupyter, Pandas, Matplotlib, SQL
Electric Utility Java & JavaScript Developer 03/2015-10/2016	Development of SAP portal applications for a web portal	• Development of SAP portal applications in Java (backend) and AngularJS and HTML (frontend) based on specifications from the customer • Deployment of the applications on the web portal • Testing of the applications • Communication with the customer's testers and incorporating their requests	Java, AngularJS, HTML, CSS, SAP NetWeaver Developer Studio, SAP GUI

Certificates

Name	Issuing organisation	Issue date	Certificate ID	Certificate URL
Data Visualization	Kaggle	11/2022		https://www.kaggle.com/learn/certification/jakubwaller/data-visualization
Deploying Machine Learning Models in Production	Coursera	07/2022	53REXCQ5PHHX	https://www.coursera.org/account/accomplishments/certificate/53REXCQ5PHHX
Machine Learning Engineering for Production (MLOps)	Coursera	07/2022	ZM4TB7LFHRFA	https://www.coursera.org/account/accomplishments/specialization/certificate/ZM4TB7LFHRFA
Machine Learning Modeling Pipelines in Production	Coursera	04/2022	XST476L3DT42	https://www.coursera.org/account/accomplishments/certificate/XST476L3DT42
Machine Learning Data Lifecycle in Production	Coursera	11/2021	5SGQK7P4T73Z	https://www.coursera.org/account/accomplishments/certificate/5SGQK7P4T73Z
Introduction to Machine Learning in Production	Coursera	08/2021	94AA8LVBEA3G	https://www.coursera.org/account/accomplishments/certificate/94AA8LVBEA3G
Microsoft Certified: Azure AI Fundamentals	Microsoft	05/2021	H828-3446	