My passion is working with data. By such, I mean ETL processes, data science, big data processing, and cloud computations. I'm a fan of football, used-cars market, and Polish cuisine. My big hobby is my dog and continuous growth of Big Data related knowledge.

Skills

- **Python** (environment management, testing, project flow, dev) ~ 6 years,
- **Apache Spark** (deployment, resources optimization, config optimization, applications efficiency monitoring, dev) ~ 4 years,
- **Hadoop ecosystem** (Hadoop on-premises and EMR, HDFS management, yarn, Sqoop, Hive, Impala) ~ 2 years,
- **Databricks** (dev, admin, integration, migration) ~ 4 years
- **SQL** (SQL Server, PostgreSQL) ~ 6 years,
- Microsoft Azure
- AWS
- Docker, git, bash, Jira, Linux,
- MS Office,
- Power BI,
- English – C1, German – A2,
- Scala and Java for Data Engineering

Experience

APRIL 2021 – present

Contractor

Advised on Data Lake maintenance and expansion (banking sector, EU, as Lead Data Engineer):

- Apache Spark / Airflow / AWS development (process / code / architecture) for Data Lake.
- Built analytical platform upon Databricks on AWS (resolving matters like scalability/data security in the cloud/ IaC automatization)
- Reduced prod AWS EMR processing costs by 25% and decreased downtime by 37%.

Environment: AWS (S3, IAM, Lambda, EC2, RDS, DynamoDB, Kinesis, Glue, EMR), Databricks
Tools: Airflow, Terraform, Python, Scala, bash, git, docker, GitHub, GitHub Actions, Apache Spark
Other: scrum methodology

Built a data processing framework for FHIR format compliant data (medical sector, US).

- Developed FHIR format – Azure – Databricks integration framework (also automated cucumber / pytest-bdd test framework)
- Troubleshot Delta Live Tables jobs.

Environment: Azure (ADLS, EventHubs, ACR), Databricks
Tools: Airflow, Python, git, docker, bitbucket, Jenkins, Apache Spark
Other: scrum methodology

Implemented a PoC for Azure Databricks-based Data Lake (e-commerce, PL).

- Designed ELT processes (pyspark, Databricks Workflows).
- Created CICD processes for schema migrations, workflows, cluster pools, etc.

Environment: Azure, Databricks
Tools: Python, git, Azure Repos, Azure Pipelines, Apache Spark
Other: scrum methodology

Designed Apache Airflow architecture for an MFT business case (energy sector, PL).

JULY 2020 – MARCH 2021

Big Data Developer – Lingaro

Developed custom Apache Spark listeners (FMCG)

- Led project.
- Gathered logs produced by Spark jobs on Databricks.
- Visualized and pointed out weak spots, cost generators, and suboptimal queries.

Environment: Azure (ADLS, EventHubs), Databricks
Tools: Python, Java, git, docker, bitbucket, Jenkins, Apache Spark, ELK, PowerBI, SQL Server
Other: scrum methodology

Master Data Engineering (FMCG)

- Migrated SAP-based ETL to Microsoft Azure.
- Built from scratch data processing engine (Databricks + Airflow + ADLS + Docker).
- Built REST APIs connecting the engine's components.

Environment: Azure (ADLS, Azure Functions), Databricks
Tools: Python, git, docker, Azure Repos, Azure Pipelines, Apache Spark
Other: scrum methodology

OCTOBER 2019 – JUNE 2020

Data Engineer (Senior Associate) – PwC Advisory (Data Analytics)

Big Data Engineering (Financial Services)

- Developed a solution responsible for orchestrating workflows from data vendors (public and private sources, both structured and unstructured) to a machine learning engine.
- Reviewed pull requests, distributed tasks to subordinates, and supervised them.
- Planned and executed data migration from HDFS to Azure Blob Storage.
- Optimized Apache Spark jobs and HDFS storage.

Environment: Azure (Azure Storage), on-premises
Tools: Python, Apache Spark, Scala, Hadoop, Hive, Kafka, Airflow
Other: scrum methodology

APRIL 2018 – SEPTEMBER 2019

Data Engineer (Associate) – PwC Advisory (Data Analytics)

Created store chain expansion model (Retail):

Designed and implemented a machine learning workflow responsible for the prediction of store income based on geographical and internal data.

Cost to Serve and SCM network optimization (Retail)

Cloudera Hadoop cluster administration:

- Configured nodes / roles, installed / updated software.
- Performance monitoring, and troubleshooting.
- Prepared and maintained a working environment for Data Scientists (JupyterHub, Cloudera Data Science Workbench, mlflow, RStudio Server, etc.)
- Completed Cloudera Administrator Training for Apache Hadoop

Environment: on-premises
Tools: Linux, Ansible, Hadoop, Apache Spark, Hive, Impala, Kafka, Nifi, Flume

MAY 2016 – MARCH 2018

Business Analyst – Creamfinance Poland

Developed KPIs tracking Shiny application

Developed a process responsible for handling loans assignment to external debt collectors.

Refactored an LGD calculation model from Excel based to a standalone Shiny dashboard.

JULY 2015 – SEPTEMBER 2015

Intern – Citi Service Center

Education

OCTOBER 2018 –

Big Data - MSc / Warsaw School of Economics

OCTOBER 2014 – JULY 2017

Econometrics - BSc / University of Warsaw

BSc thesis: *Analysis of dependencies between S&P 500, DAX and WIG20 changes.*

OCTOBER 2013 – SEPTEMBER 2016

Mathematics – BSc / University of Warsaw