ACHRAF BIRHRISSEN

Data Engineer | Big Data Engineer

achraf.birhrissen@uit.ac.ma 0767816287 linkedin.com/in/achraf-birhrissen
github.com/AchrafBir

PROFESSIONAL SUMMARY

Data Engineer with hands-on experience in building and optimizing data pipelines, ETL automation, and data infrastructure. Strong expertise in Python, PySpark, SQL, and cloud technologies. Proven track record of improving data processing efficiency by 40% through pipeline optimization. Skilled in designing scalable data solutions, data lake and data warehouse architectures, and collaborating with cross-functional teams. Master's degree in Big Data & Cloud Computing with practical experience in distributed systems and workflow orchestration.

KEY SKILLS

Data Engineering & Pipeline Development

- ETL/ELT pipeline design and optimization, data ingestion and transformation, workflow automation
- **Python** (Pandas, NumPy, PySpark), **Apache Spark**, **SQL** (MySQL, PostgreSQL), Airflow orchestration
- Data modeling and architecture, data quality assurance and governance, performance tuning

Big Data & Distributed Systems

- **Hadoop ecosystem** (HDFS, Hive), PySpark for distributed computing, large-scale data processing
- **Data lake and data warehouse** architectures, batch and real-time processing, **Apache Kafka** streaming
- Parquet, Avro, ORC columnar formats for optimized storage and query performance

Cloud Platforms & DevOps

- **Azure** (Data Factory, Data Lake, Synapse), **Docker** containerization, CI/CD (GitHub Actions)
- Infrastructure as Code, Git version control, automated deployment workflows

Machine Learning & Analytics

- ML pipeline development, TensorFlow, PyTorch, feature engineering, model deployment
- Power BI (DAX, data modeling), Tableau, SQL reporting, KPI design and dashboards

PROFESSIONAL EXPERIENCE

Specialist Analyst (Data & BI) *March 2025 - Present*

Paysera, Rabat, Morocco

- Developed and optimized automated **ETL pipelines** using Python, SQL, and PySpark, increasing data processing efficiency by 40% through robust error handling and validation
- Designed data transformation workflows with orchestration tools to support Power BI dashboards, ensuring data quality and consistency across multiple sources
- Built scalable **data ingestion processes** from CRM systems, databases, and APIs for downstream analytics
- Diagnosed and resolved **pipeline bottlenecks**, reducing reporting errors by 40% through validation rules and monitoring
- Implemented **data quality frameworks** with validation checks and anomaly detection for reliable data delivery
- Collaborated with stakeholders in **Agile environment** to translate business needs into technical solutions
- **Technologies**: Python, SQL, PySpark, Power BI, Tableau, ETL Automation, Azure, Airflow, Kafka

SELECTED PROJECTS

AI-Powered Python to PySpark Converter

- Developed automated conversion tool to migrate Python scripts to **PySpark** for distributed computing, with validation on **Parquet** datasets
- Implemented pipeline testing framework and containerized solution using **Docker** for reproducible execution
- Built **CI/CD pipeline** via GitHub Actions to automate testing, validation, and deployment
- **Technologies**: Python, PySpark, Parquet, Docker, GitHub Actions

Traffic Flow Prediction via Self-Supervised Learning

- Designed end-to-end **ML pipeline** for time-series forecasting, improving accuracy by 25% over baseline models
- Built data preprocessing workflows for temporal features, feature engineering, and data augmentation
- Optimized training pipeline, reducing training time by 30% through batch processing and parallel execution
- **Technologies**: Python, TensorFlow, SimCLR, Data Pipeline Development

Differentially Private Federated Learning for ICU Mortality Prediction (Master Thesis)

- Architected **distributed data pipeline** processing 97 clinical variables across federated nodes with privacy guarantees
- Implemented privacy-preserving ML pipeline achieving AUROC of 0.847 while ensuring data confidentiality
- Built automated hyperparameter optimization using **Optuna** to balance performance and privacy constraints
- Designed data validation and quality checks for consistency across distributed nodes
- **Technologies**: Python, PyTorch, Opacus (DP-SGD), Pandas, Scikit-learn, Optuna

EDUCATION

Master in Big Data & Cloud Computing (Bac+5) *2023 - 2025*
Ibn Tofail University, Faculty of Sciences, Kenitra, Morocco
Thesis: Differentially Private Federated Learning for ICU Mortality Prediction
Coursework: Distributed Systems, Big Data Analytics, Cloud Computing, Data Engineering, Machine Learning

Bachelor in Mathematics and Computer Science (SMI) (Bac+3) *2021 - 2023*
Ibn Zohr University, Faculty of Sciences, Agadir, Morocco

University Diploma in Technology (DUT) (Bac+2) *2019 - 2021*
Higher School of Technology, Guelmim, Morocco

CERTIFICATIONS

In Progress: Databricks Certified Data Engineer Associate | Microsoft Certified: Azure Data Engineer Associate (DP-203)

LANGUAGES

French: Professional working proficiency | English: Professional working proficiency | Arabic: Native