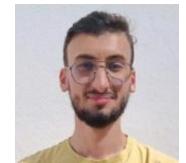


BEN TOUHAMI Mohamed Rida

Data Engineer | Master's Student in Data Science – Seeking End-of-Study Internship

Villeurbanne, France | +33 745-441-961 | bentouhamimohamedrida@gmail.com | Portfolio | LinkedIn



Profile

As a State-Certified Engineer in Data Engineering and a Master's student in Data Science, I am seeking a final-year internship to work on data pipelines and Big Data architectures.

EDUCATION

Claude Bernard University Lyon 1

2025 - 2026

Master 2 in Data Science | Relevant Coursework: Deep Learning, Big Data, Machine Learning, Cloud Computing.

National School of Applied Sciences, Abdelmalek Essaadi University

2020 - 2025

State Engineering Degree in Data Engineering | Relevant Coursework: Databases, Big Data, Machine Learning, Cloud.

PROFESSIONAL EXPERIENCE

SHIFTBRICKS COMPANY

Al Hoceima, Maroc

Data Engineering Intern

Jul. 2024 – Sept. 2024

- Developed a Medallion Big Data Architecture for Ingestion and Traceability of Arabic Legal Documents
- Transformed data and generated embedding vectors using OCR (Tesseract, EasyOCR) and LLM (Cohere).
- Automated the data pipeline with Airflow and containerized the solution using Docker.
- Built a user interface with the FARM stack (FastAPI, React) for human validation.

Tech Stack: Python, MongoDB, Airflow, FastAPI, React, Docker, Git, GitLab, OCR, GenAI.

DIGITAL PLACE COMPANY

Tangier, Maroc

Data Science & DevOps Intern

Feb. 2025 – July. 2025

- Designed a vision pipeline for automation, detection, segmentation, and tracking of semi-trailers.
- Trained YOLOv11 to detect license plates, achieving 97.8% accuracy.
- Trained EfficientNet-B2 for classification of Moroccan (Arabic) and European (Latin) license plates.
- Implemented PaddleOCR for license plate reading with 87.5% accuracy in Arabic and 93.75% in Latin.
- Applied DINOv2 to generate embeddings for truck components and enrich their visual DNA.
- Integrated the models into a FastAPI backend and a React frontend, containerized using Docker.
- Orchestrated the CI/CD pipeline using Bitbucket to automate build and continuous deployment on a VPS (DevOps).

Tech stack: YOLOv11, PaddleOCR, PyTorch, FastAPI, React.js, Docker, CI/CD (Bitbucket pipeline), Hostinger VPS.

TECHNICAL SKILLS

- **Languages:** Python, Java, SQL, PL/SQL, Spring Boot.
- **Big Data & BI Tools:** Spark, Hadoop, Kafka, Airflow, Power BI.
- **DataBases:** MySQL, Oracle, PostgreSQL, MongoDB, Cassandra, ClickHouse.
- **DevOps:** Git, GitLab, GitHub, Docker, Kubernetes.
- **Machine Learning & Deep Learning:** Classification, Clustering, Regression, scikit-learn, NLP, LLM, pyTorch, MLOps.
- **Cloud Services:** S3, Aws EC2, Databricks, Cosmos DB, Synapse Analytics, Azure Data Lake, Azure Data Factory, Azure Functions.

RECENT PROJECTS

Big Data Architecture Design for an AI Trends Chatbot Based on RAG – Azure

- Designed a Medallion Big Data architecture on Azure to automate the data pipeline (bronze, silver, gold) for a chatbot.
- Transformed data in Databricks with embedding creation, stored in Cosmos DB and AI Search for advanced search capabilities.
- Automated workflows with Data Factory and deployed the complete solution on Azure infrastructure via CI/CD (DevOps).

Tech Stack: Azure Functions, Cosmos DB, Databricks, AI Search, Data Factory, Container Apps, FastAPI, Cohere API, Next.js, BeautifulSoup.

Implementation of a Data Lakehouse Architecture and BI Solution on Azure Synapse

- Implemented a Data Lakehouse architecture on Azure Synapse, designing a star schema through an ETL pipeline and leveraging External Tables to virtualize data access without duplication.
- Designed interactive Power BI dashboards for decision-making, including KPI creation using DAX and dynamic filtering.

Tech Stack: Python, Pandas, Azure Synapse, SQL, Power BI.

Pipeline MLOps for Customer Churn Prediction in the Telecommunications Sector

- Implemented a real-time model monitoring system in production using Evidently, Prometheus, and Grafana.
- Automated MLOps orchestration for the model (retraining, evaluation, deployment) with MLflow and Airflow, leveraging customer data stored in Amazon RDS.
- Deployed the model to production on Amazon S3, containerized the application with Docker, and deployed it on an EC2 instance

Tech Stack: Python, Evidently, MLflow, Airflow, Prometheus, Grafana, Docker, AWS (RDS, S3, EC2)

Data Pipeline in Kappa Big Data Architecture – Real-time Apartment Data Processing

- Ingested and streamed data with Kafka, stored raw text in HDFS, and transformed in real-time with PySpark Streaming.
- Stored processed data in ClickHouse for fast queries and real-time visualizations via React.js and FastAPI.

Tech Stack: Hadoop HDFS, Apache Kafka, PySpark Streaming, Cohere API, ClickHouse, React.js, FastAPI.

CERTIFICATIONS

- Big Data with Spark and Hadoop Essentials (IBM).
- ETL and Data Pipelines with Shell, Airflow, and Kafka (IBM).
- Machine Learning with Apache Spark (IBM).
- Data Warehousing and BI Analytics Essentials (IBM).
- NoSQL Databases Essentials (IBM).

SOFT SKILLS

- Teamwork, methodological rigor, problem-solving skills, adaptability, clear communication, and technological curiosity.

LANGUAGES

- French (Advanced), English (Advanced), Arabic (Native)