Mohammed EL-KHOU

Data/Cloud Engineer | Big Data/Python Developer | R&D ML/AI

m.elkhou@hotmail.com	(+33) 06 13 43 51 06	France
Linkedin.com/in/m-elkhou	Github.com/m-elkhou	m-elkhou.github.io

COMPETENCES TECHNIQUES

- **Compétences principales :** Conception et gestion de bases de données, ETL, Big Data, Modélisation de données, Data Warehousing, Analyse des données, Programmation, Performance et optimisation, Programmation orientée objet, Gestion de versions et de code, Statistiques, Visualisation des données, Analyse quantitative, Machine Learning, Deep Learning.
- **Langage de programmation :** Python, Scala, Shell, Java, SQL, PL/SQL.
- **Bibliothèques Python :** PySpark, Pyarrow, Numpy, Pandas, SciPy, Matplotlib, Seaborn, PyQt5, Flask, SQLAlchemy, Scikit-Learn, TensorFlow, Keras, PyTorch, NLTK, SpaCy, Gensim, Requests, Selenium, boto3...
- **Big Data :** Spark, Hadoop (HDFS, Hive, Hbase, Pig, Storm)
- **Cloud :** GCP (Bigquery), AWS (Athena, S3, Lambda, API Gateway, EC2, ECR, Route 53, Glue, IAM, Lake Formation, AWS Bedrock, CloudFront, Secrets Manager, AWS Transfer Family…)
- **SQL Databases:** Oracle, PostgreSQL, MySQL, SQL Server, Salesforce…
- **NoSQL Databases:** Redis, Elasticsearch, Solr
- **Virtualisation :** Docker, Kubernetes, Vmware, Cloudera
- **Ordonnanceur :** Airflow, VTOM (Visual TOM)
- **Méthodologie :** Agile, Scrum
- **Conception :** UML, MERISE, Design Patterns
- **Visualisations :** Grafana, Kibana
- **BI :** Talend, Power BI
- **Format :** JSON, XML, XLT, CSV, EXCEL, YAML…
- **Systèmes :** Windows, Unix : (Debian, CentOs)
- **Outils :** Jupyter, VS-Code, Git, Colab, SSH, SAMBA, Slack, Ms Office

COMPETENCES FONCTIONNELLES

- Esprit d'analyse et force de proposition.
- Organisé, structuré et rigoureux.
- Empathie et bon relationnel client.
- Réactif avec un bon sens des priorités.
- Capacité d'apprentissage rapide.
- Problem Solving.
- Bonne capacité de communication et de travail en équipe.

EXPERIENCES PROFESSIONNELLES

Data Engineer / Cloud Engineer (AWS) RMC BFM ADS (Altice Media) - Paris, France	**01/2024 - présent**

- **Tâches :**
  - Maintenance des Bases de Données :
    - Suivi des performances des bases de données (AWS Athena, SQL Server, Oracle, PostgreSQL).
    - Résolution proactive des problèmes de performance, sécurité et fiabilité.
    - Optimisation des requêtes et des schémas pour améliorer l'efficacité des systèmes.
    - Traitement des incidents liés aux bases de données.

  - Gestion et Développement de Routines ETL (Talend) :
    - Gestion quotidienne des workflows ETL sur la plateforme Talend.
    - Debugging, amélioration et optimisation des routines existantes.
    - Création de nouvelles routines pour répondre aux besoins spécifiques de l'entreprise.
    - Intégration de techniques avancées (Data Science) pour améliorer la qualité des routines.

  - Pilotage de Projets Data Warehouse (Migration DWH):
    - Conception et déploiement d'une nouvelle architecture ETL entièrement en Python, remplaçant l'infrastructure legacy Talend.
    - Gestion de projets et coordination avec les cas d'usage et l'équipe SI.
    - Collaboration avec l'équipe infrastructure pour la création de VM Windows Server.
    - Développement de 14 modules ETL modulaires (core, db, helpers, tools) pour l'intégration multi-sources (Oracle, PostgreSQL, Salesforce, AWS Athena).
    - Développement, test et mise en production de nouvelles routines Python.
    - Réduction de 80 % des temps de traitement grâce à l'élimination des fichiers CSV temporaires et à l'adoption de la sérialisation via pickle.
    - Implémentation d'un système de logging structuré, de gestion d'erreurs robuste et de chargement incrémental configurable.
    - Élimination des coûts de licence Talend au profit d'une solution open-source performante et maintenable.

  - Plateforme automatisée de transcodage vidéo & génération VAST
    - Développement d'une solution serverless pour le traitement, la validation et la distribution de contenus publicitaires vidéo.
    - Création de fonctions AWS Lambda conteneurisées (Docker) orchestrant FFmpeg, OpenCV et MediaInfo pour le transcodage, la vérification des normes BTVS (1920×1080, 50 fps, -24 LUFS) et la correction de durée.
    - Génération dynamique de balises VAST XML avec intégration S3 et suivi des événements publicitaires.
    - Déploiement des API REST sécurisée avec AWS Cognito, Lambda@Edge, API Gateway et CloudFront OAC pour l'authentification et la protection des contenus.

  - Système de conversion audio → vidéo avec transcription IA
    - Automatisation de la création de contenus vidéo à partir de fichiers audio pour les campagnes médias.
    - Intégration d'AWS Transcribe pour la génération automatique de sous-titres en français.
    - Création de vidéos MP4 avec logo overlay, QR code dynamique (tracking de campagne) et synchronisation audio/visuelle via FFmpeg.
    - Stockage et distribution automatisés via S3 avec génération d'URLs publiques sécurisées.

- Automatisation de la facturation publicitaire (AAF)
  - Développement d'un moteur d'automatisation pour la réconciliation et la validation des montants facturés.
  - Traitement de factures PDF via AWS Textract (OCR) et extraction intelligente des champs (dates, montants, devises).
  - Intégration avec les APIs AppNexus et FreeWheel pour la collecte des données de livraison publicitaire.
  - Validation des chaînes de consentement IAB TCF (RGPD) avec décodage, hachage SHA256 des IPs et stockage conforme.

- Gestion des placements publicitaires FreeWheel en masse
  - Orchestration automatisée des campagnes publicitaires sur la plateforme FreeWheel.
  - Création de placements à partir de fichiers CSV avec ciblage géographique (codes postaux) et segmentation audience.
  - Gestion des créatifs via l'API V4, rendu multimédia et activation programmatique.
  - Mise en place de mécanismes de retry et de validation d'état pour garantir la fiabilité.

- Infrastructure Cloud & Sécurité
  - Conception d'une architecture AWS sécurisée, scalable et entièrement automatisée.
  - Utilisation de VPC, NAT Gateway, IAM roles et gestion des secrets via variables d'environnement.
  - Déploiement via CloudFormation et AWS CLI, avec versioning des fonctions Lambda@Edge.
  - Protection des buckets S3 via Origin Access Control (OAC) et cookies sécurisés (HttpOnly, Secure).

- Pipeline d'analyse des coûts AWS & monitoring CloudWatch
  - Mise en place d'un système d'observabilité cloud pour l'optimisation des coûts et la traçabilité des opérations.
  - Extraction, transformation et stockage des logs CloudWatch vers Amazon Athena au format Parquet partitionné.
  - Analyse de millions d'événements pour identifier les anomalies de consommation et générer des rapports de facturation automatisés.
  - Intégration avec des outils de reporting HTML et d'alerting par email.

- Pipeline Intelligence Audience & Matching IA (AWS Bedrock)
  - Conception et déploiement d'un pipeline end-to-end d'audience intelligence CTV pour 3 chaînes (RMC Story, RMC Life, RMC Découverte) : collecte des programmes TV en direct (EPG) → matching automatique → export des segments d'audience vers Mediarithmics (ciblage publicitaire programmatique).
  - Développement d'un service de matching sémantique par LLM via AWS Bedrock (Claude 3 Haiku / Claude Sonnet 4.5) entre les grilles EPG et le catalogue Replay/catch-up TV (PIXA) — architecture 2-pass : matching strict (score ≥ 80/100) + passe assouplie (score ≥ 60/100), avec scoring gradué 0–100 et raisonnement explicatif généré par le modèle.
  - Optimisation du prompt engineering : réduction de 82,5 % des tokens via index allégé (~350 chars/item), batching (5 EPG × 400 assets par appel Bedrock), chunking adaptatif et raffinement automatique des cas ambigus (top-2 scores à moins de 15 points d'écart).
  - Benchmarking multi-modèles (V1→V4) : taux de matching de 95,1 % atteint avec Claude 3 Haiku vs 55,8 % pour Claude Sonnet 4.5, à coût 92 % inférieur — recommandation et déploiement en production du modèle optimal pour ce use-case.
  - Optimisation des coûts AWS Bedrock : cache de résultats (−80 %), pré-filtrage temporel (−67 %), parallélisation ThreadPoolExecutor (3 workers) ; étude de migration vers inférence locale (Qwen 2.5, RTX 3060) pour élimination totale des coûts LLM.
  - Construction des tables Athena partitionnées (epg_programs, epg_serie_mapping, implicit_in_monitoring) via AWS Glue Catalog (awswrangler) ; matching EPG ↔ séries FreeWheel par algorithme 4 phases combinant fuzzy matching et normalisation NLP.

    - Export des segments d'affinité Implicit vers Mediarithmics (CTV programmatique) avec croisement du consentement RGPD (IAB TCF), génération de fichiers membership CSV et rapports HTML automatisés par email (AWS SES).
- **Environnement :** CentOS, Windows Server, Shell Scripting, Windows Task Scheduler comme orchestrateur, Talend Cloud Data Integration, Python 3.9+ (Vs-Code, Glue notebooks), SQL, PL/SQL, Oracle, PostgreSQL, SQL Server, AWS (Athena, S3, Lambda, API Gateway, EC2, ECR, Route 53, Glue, IAM, CloudFront, Cognito, CloudWatch, Transcribe, Textract, Bedrock), Salesforce, SFTP, CI/CD, GitHub, JIRA, Confluence.

Big Data Engineer BPCE-SI - Lille, France	**08/2022 – 01/2024**

- **Contexte :** Conception, développement et gestion des pipelines de Big Data de toutes les divisions Caisse d'Epargne Ile de France.
- **Tâches :**
    - Renforcement de l'équipe sur les technologies Big Data, en particulier sur Cloudera.
    - Développement, maintenance et automatisation de pipelines ETL pour mettre à jour le Datalake (cluster Hadoop ; Hive, HDFS…) de BPCE-SI à partir de diverses sources, y compris une base de données Oracle.
    - Collaboration étroite avec des data scientists et des analystes pour identifier des informations basées sur les données pour éclairer les décisions commerciales.
    - Garantie de la qualité et de l'exactitude des données par le développement et la mise en œuvre de politiques de gouvernance des données.
    - Automatisation des pipelines de données en utilisant Python et Shell, avec des développements de nouvelles fonctionnalités en SPARK/SCALA ou PySpark / Python.
    - Contribution à la formalisation des plans de tests, exécution des tests unitaires, résolution d'anomalies pendant les phases d'intégration, packaging, et déploiement en pré-production et production.
    - Surveillance de la production (le Run), maintenance du système actuel, et mise en œuvre de solutions de reprise en cas d'incident.
    - Optimisation des solutions Big Data en termes de performances, de scalabilité et de rentabilité.
    - Élaboration de documents de concept technique et de spécifications conformes aux meilleures pratiques et normes sur Confluence.
    - Intégration de la migration de Cloudera vers GCP avec un double RUN.
    - Travail au sein d'une équipe agile en appliquant Scrum et le développement agile.
- **Résultats :** Amélioration significative de la disponibilité des données, réduction des temps de traitement, augmentation de la qualité des données et la réduction des coûts d'infrastructure.
- **Environnement :** Unix/Linux, Shell scripting, Python (Jupyter, Vs-Code, Anaconda/Miniconda), Cloudera, Hadoop, Hive, HDFS, Spark, Scala, PySpark, SQL, PL/SQL, Oracle, GCP(BigQuery), VTOM (Visual TOM), CI/CD, GIT, Bitbucket, JIRA.

R&D Data Engineer / Machine Learning Engineer IMPERIUM, Casablanca, Maroc	**09/2021 - 08/2022**

- **Contexte :** Mise en prod des solutions d'IA dans le domaine du multimédia, traitant des données non structurées ou semi-structurées, implique la gestion du ressources et automatisation.
- **Contraintes :** Limitation des ressources matérielles, impliquant un recours fréquent au CPU plutôt qu'aux GPU. Gestion délicate d'une seule carte graphique sur un serveur.
- **Tâches :**
    - Mission 1 : Modernisation de l'Infrastructure :

- Mise en œuvre de microservices pour faciliter l'intégration de modèles d'IA encapsulés dans des images Docker.
- Développement de workflows ETL avec Python, impliquant la consommation de REST APIs pour intégrer des modèles d'IA.

- Mission 2 : Structuration du Pipeline de Training
  - Conversion d'un pipeline de training depuis des fichiers Jupyter vers un projet bien structuré, en respectant les normes de POO.
  - Intégration de fonctionnalités de multiprocessing, multithreading, et parallélisation GPU pour accélérer le processus de prédiction des modèles.
  - Automatisation du déploiement en production à l'aide d'Airflow et Kubernetes, assurant la gestion efficace des conteneurs Docker.
- Mission 3 : Création d'applications Desktop avec PyQt5 pour une utilisation interne, améliorant l'expérience utilisateur et l'efficacité opérationnelle.
- Mission 4 : End-to-End Web Mining des Sites de News
  - Élaborer une architecture microservices adaptée, en identifiant les composants nécessaires pour la collecte, le traitement et le stockage des données, tout en garantissant la scalabilité et la flexibilité du système
  - Implémenter des APIs distincts pour chaque étape du processus, notamment le scraping (Scrapy, request, selenium, BeautifulSoup…), crawling (test d'existence md5), la transformation des données (readability), et le stockage(PostgreSQL). Veiller à la modularité et à l'indépendance des services.
  - Intégrer des mécanismes de gestion des erreurs pour traiter les éventuelles interruptions ou modifications de la structure des sites web, assurant ainsi une collecte de données robuste et fiable.
  - Mettre en œuvre une planification cyclique pour chaque traitement en batch, garantissant une mise à jour régulière des données conformément aux besoins du client.
  - Intégrer des fonctionnalités de surveillance pour détecter les éventuels problèmes, et établir un plan de maintenance proactive pour assurer la stabilité du système sur le long terme.
- **Résultat :** Une structuration efficace du pipeline de training. Notamment, l'implémentation du multiprocessing pour les prédictions a permis d'atteindre une quasi-temps réel, avec des batchs s'exécutant de manière cyclique sans interruption malgré les ressources limitées, renforçant ainsi la dynamique et la réactivité du système.
- **Environnement :** Unix/Linux, Python(Jupyter, Vs-Code), Java(Eclipse), Docker, Docker-compose, Kubernetes, MySQL, PostgreSQL, SQL Server, Slack, Elasticsearch, Kibana, Airflow…

Développeur ML / AI 3W Media - Imperium, Casablanca, Maroc	**08/2020 à 09/2021**

- **Projet 1 :** Transcription
  - **Tâches :**
    - Transcription d'un discours prononcé par un locuteur en contenu textuel de manière automatisée.
    - En utilisant une reconnaissance vocale hors ligne, des modèles pré-entraînés pour 17 langues et dialectes - arabe, français, anglais, ...
  - **Environnement :** Python(Jupyter), Docker, multiprocessing, cronjob, Elasticsearch.

- **Projet 2 :** Identification et recherche de spots publicitaires à l'aide d'empreintes audio.
  - **Tâches :**
    - Mise en place d'une solution de reconnaissance audio : application sur les spots publicitaires ; pour localiser des instances de publicités TV ou RD connues dans une archive d'émissions de télévision enregistrées (ou d'autres enregistrements plus longs comme le flux radio) à l'aide de l'empreinte audio basée sur Landmark.
  - **Environnement :** Python(Jupyter), Docker, multiprocessing, numpy, pydub.AudioSegment, MySQL.

- **Projet 3 :** Détection du genre et du locuteur à partir de la parole (ASR).

- **Tâches :**
  - Concevoir un système de détection automatique du speaking et du genre pour la surveillance à grande échelle des flux audiovisuels bruts. Les estimations quantitatives du temps de parole des femmes aident à décrire l'évolution de l'égalité des sexes dans le temps, permettent aux chaînes de télévision et aux stations de radio de surveiller l'égalité des sexes dans leurs émissions.
- **Environnement :** Python(Jupyter), Tensorflow, Keras, KerasTuner, Docker, Kubernetes, Elasticsearch, Communication RESTful API.

- **Projet 4 :** Exploitation des réseaux sociaux pour les sujets d'actualités au Maroc (Topic detection) + comprendre l'opinion d'une personne sur un sujet particulier (Sentiment Analysis).
  - **Tâches :**
    - Analyser de grands ensembles de données hétérogènes (texte, données tabulaires, documents).
    - Extraire les données non structurées en temps réel. Une centaine d'articles qui proviennent de sources multiples, d'un comportement varié et de formats différents (textes, graphiques, tableaux) - Traitements Big Data.
    - Effectuer l'analyse des données, élaborer la conception de la base de données.
    - Développement et prototypage sur des Jupyter-Notebooks avec les packages AI/Data Science Python.
    - Entrainer le modèle d'apprentissage automatique pour la classification des sentiments et pour la détection de sujet d'actualité sur les énoncés en français (commentaires/publications).
    - Configuration et déploiements de kubernetes Pods avec des config YAML.
    - Documentation/ Formation.
  - **Environnement :** Python 3.8, API Twitter, Scrapy, NLTK, scikit-learn, word2vec, BOW, TF-IDF, Topic Modeling (Bert Topic), Docker, PostgreSQL.

- **Projet 5 :** Face AI
  - **Tâches :**
    - End_to_end pipeline de reconnaissance faciale et d'analyse des attributs faciaux (âge, genre, émotion et race) pour Python. Il s'agit d'un cadre hybride de reconnaissance des visages enveloppant des modèles pretrained de l'eta de l'art : VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib et SFace.
  - **Environnement :** Python(Jupyter), Tensorflow, Kerass, Docker, multiprocessing, GPU parallelization, cronjob, Elasticsearch.

- **Projet 6 :** Dictionnaire multilingue
  - **Tâches :**
    - MD est une alternative au dictionnaire alphabétique où la liste des mots est groupée selon leur signification et selon des hiérarchies structurelles familières.
  - **Environnement :** Python(Jupyter), Docker, SQL Server, Solr, WordNet (OMW).

Data Scientist 3W Media - Imperium, Casablanca, Maroc	**03/2020 à 08/2020**

- **Projet :** Détection et reconnaissance des objets (Marque, logo, forme ...) dans les écrans publicitaires sur un flux TV, les panneaux publicitaires et les journaux.
  - **Tâches :**
    - Etude comparative entre les algorithmes d'état-de-l'art de Détection d'objets (YOLO, RetinaNet, FasterR-CNN, Mask R-CNN, Cascade Mask R-CNN...)
    - End-to-end pipeline du scraping des données au déploiement en production avec Python.
    - Scraping de données (images, vidéos) à partir des médias sociaux.
    - Développer un pipeline pour le nettoyage des données (supprimer les images dupliquées).
    - Élaborer une nouvelle stratégie d'augmentation des données pour les problèmes de détection d'objets.

- Former l'équipe d'étiquetage.
- Entrainer des modèles pour la détection d'objets et la segmentation d'instance.
- Déployer une API Flask pour consommer les modèles.

- **Environnement :** Unix/Linux(CPU-GPU), Python(Jupyter, Vs-Code, PyTorch, Docker, Flask, Computer Vision, Detectron2, RetinaNet, Faster R-CNN, Mask R-CNN, Cascade Mask R-CNN, Selenium.

ETUDES

Master 2 : Sciences des Données et Intelligence Artificielle **2020**
Institut Galilei, Université de Sorbonne Paris Nord - Paris, France.

Master : Web Intelligence et Sciences Des Données **2020**
Formation approfondie de deux ans spécialisés dans le développement, le traitement big data, la science des données et l'Intelligence Artificielle.
Université Sidi Mohammed Ben Abdellah - Fès, Maroc

Licence Fondamentale : Sciences Mathématique et informatique **2018**
Université Sidi Mohammed Ben Abdellah - Fès, Maroc

Diplôme d'Études Universitaires Générales : Sciences Mathématique et informatique **2017**
Université Sidi Mohammed Ben Abdellah - Fès, Maroc

Baccalauréat : Sciences Mathématique - Série : B **2015**
Lycée EL ADARISSA - Fès, Maroc

LANGUES

- Arabe : Maternelle
- Français : Courant
- Anglais : Professionnel

CERTIFICATS

AWS Cloudfront: Serve content from multiple S3 buckets– Coursera, IBM **2024**

AWS S3 Basics– Coursera, IBM **2024**

Introduction to Big Data with Spark and Hadoop – Coursera, IBM **2024**

Deep Learning Specialization - Coursera, DeepLearning.AI **2022**

Hadoop Platform and Application Framework – Coursera, UCSanDiego **2022**

ETL and Data Pipelines with Shell, Airflow and Kafka – Coursera, IBM **2022**

Functional Programming Principles in Scala – Coursera, EPFL **2022**

Introduction to Big Data – Coursera, University of California, San Diego **2022**

Neural Networks and Deep Learning - Coursera, DeepLearning.AI **2020**

Machine Learning with Python - Level 1 - Fourni par IBM sur cclaim **2019**

Applied Data Science with Python - Level 2 - Fourni par IBM sur cclaim. **2019**

Data Analysis with Python - Fourni par CognitiveClass. **2019**

Data Visualization with Python - Fourni par CognitiveClass. **2019**

Python for Data Science - Fourni par IBM sur cclaim **2019**

Data Analysis Track - One Million Arab Coders in Udacity. **2019**