Hamza KAROUI Lead Data Engineer/Spark Expert

EXPERIENCE

BNP PARIBAS - CIB | Freelance | Lead Data Engineer
Mar 2021 – Fev 2024 (3 ans) | Paris, FR

Objectif: Développement de solutions pour la lute contre le blanchiment d'argent, générer une vue 360 des clients et entités associées, obtenir un aperçu complet de l'hérarchie du KYC.

- Ingestion de données depuis le Datahub, transformation, génération des entités associées, et création de Graphs clients.
- Implémentation de règles pour la découverte des risques cachés dans les réseaux des clients.
- Optimisation des jobs Spark et mise en place des bonnes pratiques
- Création et gestion des clusters Spark à l'aide de Ansible.
- Refonte de l'infrastructure vers du Spark sur Kubernetes
- Refonte du scheduler à l'aide de ArgoWorkflow
- Création de service d'inférence ML développé à l'aide de Spark + Deep Java Library + PyTorch, pour extraire et structurer les adresses dans les messages SWIFT

En tant que Lead :

- Review quotidien des PRs, assurance de la qualité de code et gestion des conflits techniques
- Mise en place des bonnes pratiques et les conventions de codage : Scala, Git, Spark
- Animation des workshops techniques et live programming
- Réalisation de POC: Spark sur kube, Argo, Cleversafe vs FlashBlade

Equipe: 1PO, 1 Lead et 3 Data Engineer

Stack: Kubernetes, Docker, Skaffold, kustomize, Scala/Python, Spark, SQL, GoLang, Argo, Jenkins, Ansible, Quantexa, Artifactory, Gradle, SBT

KERING | Consultant | Data Engineer (2 ans)
Avr 2019 – Fev 2021 | Paris, FR

Objectif: Centralisation des données des marques du groupe Kering dans le Datalake et développement des calcules spécifiques à la demande des marques.

- Ingestion de données depuis différents systemes: SalesForce Marketing Cloud, SAP, NFS, Bases de données Oracle.
- Concevoir, développer et planifier des jobs d'ingestion de données en Scala/Spark & PySpark
- Consolidation et enrichissement des données marketing, clients (10 TB / Jours)
- Génération de KPI Marketing: les modèles d'attribution & contribution
- Refonte des jobs d'analytics à l'aide de AWS Glue, Lambda et Athena pour accélérer le calcule
- Refonte du scheduler à l'aide de Airflow

Réalisations notables :

- Conception d'un job de calcul de l'attribution marketing (ATO-Engine) pour remplacer des anciens job en PySpark
- Réduction du temps de calcul (de 7h30 à 30min) et mémoire totale (de 1TB à 20GB)
- Conception d'un DSL (Domain Specific Language) en Scala, pour faciliter l'implémentation des règles business

Equipe: 2PO, 2ProxyPO, 2 Lead et 10 Data Engineer

Stack: AWS, Scala/Python, Spark, Spark Streaming, SQL, Nifi, Jenkins, Terraform.

LIENS

helkarou@gmail.com
Tel: 0630713550

Github:// **helkaroui**
LinkedIn:// **hamza-el-karoui**
Blog: **www.sharek.dev**

CERTIFICATS

- AWS Cloud Practitioner Associate
- Kubernetes and Cloud Native Associate

EDUCATION

TÉLÉCOM BRETAGNE
Ingénieur Généraliste
Sep 2016 - Sep 2018 | Brest, FR
Cum. GPA: 3.8 / 4.0

SUP'COM
Ingénieur Télécommunication
Sep 2014 | Août 2016 Tunis, TN
Distinction: Bourse pour double diplome à Télécom Bretagne

IPEIN
Classes Préparatoires
Sep 2012 - Août 2014| Tunis, TN
Filière: Math Physique
Rang national: 87 / 2500

SKILLS

PROGRAMMING
Scala, Python, Java, TypeScript, SQL, Shell

SCALA LIBRARIES
Akka, Akka Streams, Zio, Play, Slick, DeepLearning4j

PYTHON LIBRARIES
PySpark, Pandas, Flask, Fast API, PyTest

CI/CD
Jenkins, Ansible, Argo, Airflow, Terraform

BIG DATA
Spark, Kafka, ElasticSearch, Hadoop, Yarn, Nifi, Cassandra, Druid, Flink, Minio

CLOUD & CLOUD NATIVE
AWS: EC2,EMR,Athena,Glue,Lambda
Kubernetes, Docker, Kustomize, Skaffold, Helm

MAKE.ORG | Consultant | Junior Data Scientist
Jan 2019 – Mar 2019 (3 mois) | Paris, FR

Objectif: Make.org est une organisation neutre et indépendante dont la mission est de faire participer les citoyens et de mobiliser l'ensemble de la société civile pour transformer positivement la société.
L'objectif de la mission est de développer des modèles ML pour catégoriser les propositions citoyennes collectées lors des compagnes de Make.org.

- Création de modèles NLP, en 22 langues de l'union européen, pour la classification de textes courts
- Création d'un modèle de traduction de texte de 21 langues en Anglais.
- Création d'une API d'inférence en Scala (framework Akka) qui sert les modèles entraînés en Python

Stack: Spark, Scala, Python, Akka, SciKit Learn, PySpark, Word2vec, Superset, Flask, Docker, Gitlab CI/CD, Rundeck.

SG - KOMEOS | Projet en forfait | Junior Data Scientist
Nov 2018 – Dec 2018 (2 mois) | Paris, FR

Objectif: L'objectif est de numériser et rendre recherchable, les documents de référence annuels de la Société générale. Un autre objectif c'est de rendre les tableaux téléchargeable en format excel/csv.

- Extraction de textes, structures et paginations depuis des documents PDF en Python
- Indexation des textes de façon recherchable dans Elasticsearch
- Développement de modèles de localisation puis d'extraction de tableau (Deep learning) depuis les PDFs
- Création d'un outil web de navigation et recherche dans les documents grace à Flask et ElasticSearch. L'ui permet aussi de télécharger les tableaux extraits au préalable en format tabulaire

Stack: Python, Elasticsearch, Flask, AWS API GATEWAY / EC2 / S3, Jenkins

ORANGE LABS | Stage de fin d'études
Mar 2018 – Oct 2018 (6 mois) | Paris, FR

Objectif: Développement d'un modèle ML permettant de catégoriser les emails au sein d'une entreprise, ensuite extraire les processus métiers mis en pratique (Process Mining)

- Vectorization des textes et création d'un modèle XGBoost pour catégoriser les emails
- Proposition d'une approche statistique permettant de découvrir l'enchainement des actions dans un processus donné exemple: processus de recrutement

Stack: KERAS, Sci-kit learn, Spacy, NLTK ,Flask, PySpark