Khac Bao Anh **NGUYEN**

Data Engineer | Développeur Scala/Python/Spark

linkedin.com/in/nguyenkhacbaoanh github.com/nguyenkhacbaoanh
https://nguyenkhacbaoanh.github.io/
khacbaoanh.nguyen@gmail.com
+33 (0)6 02 73 41 12
5 Rue Christian Barnard, 77600 Bussy-Saint-Georges, France

Expérience Professionelle

Aujourd'hui
Avril 2024

Tech Lead, Equans, France

Data Enginner | Palantir Developpeur

- Proposer les architectures techniques adéquates pour assurer la fiabilité des flux de données et faciliter leur maintenance.
- Réaliser les demandes de développement d'intégration de données dans Palantir.
- Assurer la qualité technique des développements dans Palantir & Apache Spark.
- Participer au suivi quotidien de l'application en production et au support aux utilisateurs.
- Documenter les réalisations, présenter le résultat et faire le suivi des recettes techniques.
- Accompagner les Développeurs Junior.

Python | Palantir | Spark 3 | Azure Blob Storage | Git | SAP | Data Lineage | Monitoring | Communication | Leadership

Aujourd'hui
Mars 2024

Data Consultant, Talan, France

Data Enginner | DevOps Azure

Le traitement des AO est donc un processus chronophage qui représente un coût important pour les entreprises. Toutefois, ce coût pourrait être optimisé grâce aux dernières technologies d'IA que sont les IA génératives. L'objectif ce projet de recherche est d'étudier l'apport des IA génératives dans l'analyse et la réponse aux AO au sein d'un grand groupe tel que Talan

- Extraction et transformation des données à partir de divers formats de fichiers (PDF, DOCX, PPTX).
- Mise en œuvre de ChromaDB sur Azure Blob Storage pour l'embedding et les bases de données vectorielles.
- Création d'un workflow dans Databricks.
- Mise en œuvre de la recherche vectorielle dans Databricks : Unity Catalog, Endpoints, Indexing, et Delta Sync.
- Automatisation des tests, du Packaging, et du Deployment à l'aide d'un pipeline CI/CD dans un dépôt DevOps et un artefact.

Python | Databricks | LLM | Spark 3 | IA Génératives | ChatBot | Azure Blob Storage | ADF | Azure DevOps

Février 2024
Janvier 2021

Data Engineer | Big Data Engineer, Société Générale, France

Chapter Leader Big Data au sein de Quartier Finance Datalake

- Collaboration et partage des connaissance en facilitant les séances de partage de connaissances et d'expériences, où les membres peuvent présenter des projets, des défis et des solutions innovantes.
- Le mentorat et l'encadrement des membres de l'équipe sur des sujets transverses tels que la dette technique et les migrations.
- Le suivi de la consommation des ressources Big Data implique la révision de l'optimisation des requêtes du projet et la migration vers Spark 3.
- L'étude porte sur l'architecture Lakehouse, les formats de stockage Deltalake & Iceberg, et Trino/Starburst afin d'adapter l'utilisation des transactions ACID.

Leadership | Team-building | Communication | Spark 3 | Optimization | Azure Synapse Analytic | Azure Blob Storage | ADF

Migration du système comptable entre BDDF et Crédit du Nords

- Définition des modèles de données et optimisation des traitements de données.
- Mise en place de règles des métiers avec les BAs et Coordination des autres équipes.
- Proposition et développement une solution Cloud Serverless qui permet aux équipes métiers d'exploiter et de partager des données sans besoins d'accéder au data lake.
- Orchestration la chaîne des pipelines de données.

Scala/Python | Spark | Hadoop | Hive | Maven | AWS S3 | Lambda Function | CloudWatch | Talend | Ansible | Jenkins

Développement des chaînes CICD pour des pipelines de données développées sous Scala/Spark

- Conception, développement et amélioration la structuration de code d'applications.
- Développement des playbooks d'Ansible permettant aux déploiements aux Datalake via les Edges.
- Définition des pipelines de builds, test d'intégration, validation de code & déploiement.

Maven | Jenkins | Ansible | Ansible Tower | Nexus | Technical Leader

Pilotage de la migration de Presto vers Trino - une solution des requêtes distribuées

- Réalisation, Accompagnement & Documentation des étapes de la migration.
- Participation aux problèmes dans la migration avec l'équipe Trino.
- Mise en place la solution ODBC connecteur avec PowerBI Desktop et L'installation de DBeaver pour des utilisateurs.

Trino | API | ODBC | PowerBI | Management | Technical Leader

December 2020
Janvier 2020

Software Engineer | Développeur/Concepteur Python, Société Générale, France

Administrateur & Développeur Talend Data Management Plateforme

Talend est une nouvelle offre proposée par l'équipe MDW/APP/ETL de la SG. L'installation et Maintenance de service Talend.

- Accompagnement et Support des clients internes dans leurs développements.
- Mise en place les services Talend from scratch : TAC, JobServer, Artefact Répository, LogServer et aussi des migrations de version et projet.
- Développement des scripts de démarrage de service et maintenance de chaînes CICD.

Talend v7.1.1 & 7.3.1 | TAC | Jobserver | ELK-Beats | Nexus | Maven | Bash | Linux Redhat 7

Restitution des données de l'équipe ETL et Création des Dashboards MDW/APP

Middleware fourni des infrastructures aux autres équipes IT de la SG. Ce projet nous aide à évaluer l'utilisation de l'infrastructure de nos clients. Ensuite, ce projet effectue un rapprochement avec la facturation interne.

- Implémentation, analyse et livraison de correctifs de KPIs sur les outils ETL : Informatica, Datastage.
- Visualisation des KPIs en utilisant des données restituées au datalake.

Python 3 | SQLAchemy-ORM | Oracle | Grafana | Kafka | Kubernetes | ELK-Beats

Développement d'un portail d'API pour l'ETL Informatica

Suite aux besoins spécifiques de clients internes, l'équipe met en place une API qui donne la possibilité d'effectuer des arrêts/démarrages des Instances d'Informatica et/ou de consulter les nombres de workflow...

- Conception et développement d'applications API Restful sous l'architecture proposée.
- Développement des playbooks d'Ansible.
- Déploiement d'APIs pour plusieurs Plate-forme de SG (Docker avec Swarm ou Kubernetes).

Python 3 | Flask | Flask-restplus | Swagger UI | Oauth2 | Jira | SgConnect | Ansible | Docker | Kubernetes

L'automation des tâches périodiques

Airflow est l'outil d'ordonnancement standard à la SG, cet outil permet d'ordonnancer des jobs et d'effectuer des actions en tant d'Admin sur les VMs.

- Implémentation et Automation de tâches d'envoie du rapport mensuel.
- Implémentation et Automation de tâches de détection du panne du service.

Python 3 | Bash | Airflow | Docker | Github | Git | Cloud Privé | Terraform

Décembre 2019
Juin 2019

Développeur/Concepteur Python & Talend ETL, Société Générale, France

Développement d'un portail d'interface d'API pour l'ETL Talend

l'équipe Middleware implémente des API afin de mettre à la disponibilité des maîtrises d'oeuvre la possibilité de réaliser certaines opérations sur l'ETL Talend.

- Conception et développement d'applications API Restful adaptées aux normes de la SG.
- Consommation des commandes de Talend MetaServelet API.
- Déploiement d'APIs pour plusieurs Plate-forme de SG (Docker avec Swarm ou Kubernetes).

Python 3 | Flask | Flask-restplus | Talend Administrator Center | Talend Metaservlet | Docker | Kubernetes

Application de surveillance en temps réel
Suivi des états des plates-formes et des applications clientes pour chaque environnement : développement, homologation, production. Cette application permet aux administrateurs des plates-formes d'être prévenu en temps réel des incidents de production, d'avoir une meilleure proactivité et une détection rapide d'une ou des composants défectueux.

- Conception et développement d'applications web.
- Cahiers de charges, analyses des besoins de Service Manager et l'établissement des KPIs.
- Déploiement l'application aux différents environnements (DEV,UAT,PRD) de Serveur Cloud.

Python 3 | Flask | SQLAchemy-ORM | Dash/Plotly | Oracle | Javascript | HTML/CSS

L'intégration continu et déploiement continu (CICD) de Talend
Création d'un POC afin d'exécuter ses "workflow" dans un containeur de Docker ou de scheduler des jobs dans Kubernetes. Toutes les procédures de containairiser sont réalisées en CICD.

- Installation des outils et Préparation l'environment de CICD.
- Développement et déploiement un workflow ETL utilisant Talend Data Intégration.
- Préparation le demo de chaîne CICD et support aux clients.

Talend Open Studio v.7.1.1 | TAC | JRE 1.8 | Maven | Docker | Kubernetes | Jenkins | Github | Nexus

Compétences

Programmation & ETL Tools	Scala, Python, Talend Big Data, Airflow
Datalake	Hadoop, Hive, Spark, Palantir
Cloud	AWS, Azure, Docker EE, Kubernetes
Certification	Azure Fundamental (AZ-900), Azure Data Fundamental (DP-900)
Storage, Database	Object Storage (S3), PrestoSql/Trino, SqlServer, Oracle, PostgreSQL
CICD	Jenkins, Ansible, AWX, Nexus, GitHub
Monitoring	ELK, Grafana
Gestion de Projet	Jira, Agile

Formation & Certification

03/2024	**Certification Databricks** – Data Engineer Associate
11/2023	**Certification Databricks** – Lakehouse Fundamentals
02/2023	**Certification Microsoft** – Azure Data Fundamentals (AZ-900)
01/2023	**Certification Microsoft** – Azure Data Fundamentals (DP-900)
01/2021	**Certification Coursera** – Apache Spark SQL pour Data Analysts
07/2019	**Certification Talend** - Intégration Continue avec Talend
06/2019	**Certification Talend** - Data Intégration Basics & Avancé
2017 – 2019	**Master Big Data & Intelligence Artificielle** à l'école HETIC - Hautes études du numérique - Montreuil – France
2014 – 2017	**Master Monnaie - Banque - Finance - Assurance parcours Décision & Risque & Assurance** à l'Université Sorbonne Paris Nord – Villetaneuse – France