

Khac Bao Anh NGUYEN

Data Engineer | Développeur Scala/Python/Spark

[in linkedin.com/in/nguyenkhacbaoanh](https://www.linkedin.com/in/nguyenkhacbaoanh) github.com/nguyenkhacbaoanh
@ khacbaoanh.nguyen@gmail.com
+33 (0)6 02 73 41 12 | +33 (0)1 56 37 88 37
5 Rue Christian Barnard, 77600 Bussy-Saint-Georges, France



EXPÉRIENCE PROFESSIONNELLE

Aujourd'hui
Janvier 2021

Data Engineer | Big Data Engineer, SOCIÉTÉ GÉNÉRALE, France

Chapter Leader Big Data au sein de Quartier Finance Datalake

- Collaboration et partage des connaissances 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 Nord

- 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 d'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 CI/CD 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 Repository, LogServer et aussi des migrations de version et projet.
- Développement des scripts de démarrage de service et maintenance de chaînes CI/CD.

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 SQLAlchemy-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'envoi 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

Développeur/Concepteur Python & Talend ETL, SOCIÉTÉ GÉNÉRALE, France

Juin 2019

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 MetaServlet 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 SQLAlchemy-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 conteneur de Docker ou de scheduler des jobs dans Kubernetes. Toutes les procédures de containairiser sont réalisées en CICD.

- Installation des outils et Préparation l'environnement 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
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

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