

# Sophie Siake

94400, Vitry Sur Seine

0753854151 | siakesophie@gmail.com

Site Web : <https://www.linkedin.com/in/siake/>

## Profil Professionnel

Data Engineer spécialisée dans la gestion et l'intégration de données avec **Python**, **SQL**, **Airflow** et **Docker**. Compétences en développement de pipelines **ETL** et optimisation des workflows, avec une approche axée sur l'efficacité des processus et la qualité des données. Soucieuse d'apporter des solutions innovantes et de contribuer à des projets de data engineering.

## Compétences

- |              |          |
|--------------|----------|
| • Python     | • Docker |
| • SQL        | • Kafka  |
| • PostgreSQL | • Spark  |
| • Streamlit  | • Scala  |
| • DBT        | • Azure  |
| • Airflow    | • Git    |

## Parcours professionnel

### Apprentie Data Engineer

Septembre 2022 - Septembre 2024

**Orange Innovation** - Châtillon - Contrat d'apprentissage

**Résumé** : Mise à jour et intégration de données critiques de KPI, tels que les performances 4G et les sites Orange, à partir de fichiers Excel, pour optimiser le suivi et les prises de décisions stratégiques.

- Développement de pipelines ETL avec **Airflow** pour extraire, transformer et charger les données dans **PostgreSQL**.
- Automatisation du contrôle qualité des données avec des scripts **Python**.
- Conteneurisation des workflows de traitement avec **Docker** pour améliorer la portabilité.
- Requêtes complexes et optimisation des bases de données avec **SQL**.
- Gestion des versions et des collaborations via **Git**.

Outils : **Python**, **PostgreSQL**, **Airflow**, **Docker**, **SQL**, **Git**

## Projets réalisés

### Solution Streaming de Vote en Temps Réel

- Configuration des tables dans **PostgreSQL** et gestion des flux de données avec **Kafka**.
- Traitement et enrichissement des données en temps réel avec **Spark Streaming** et des scripts **Python**.
- Développement d'une application interactive avec **Streamlit** pour afficher les résultats en temps réel.
- Utilisation de **SQL** pour optimiser les requêtes sur les données agrégées.
- Gestion du versionnement et des workflows collaboratifs avec Git.

Outils : **Python**, **SQL**, **PostgreSQL**, **Kafka**, **Spark Streaming**, **Streamlit**, **Git**

### Optimisation des processus ETL avec Azure Data Factory

- Création des ressources nécessaires dans **Azure**, telles que les comptes de stockage **ADLS Gen2**.
- Utilisation de **Azure Data Factory** pour réaliser des processus ETL sans code.
- Ingestion des données depuis **Azure SQL Database** vers **Azure Data Lake Storage Gen2** (ADLS Gen2).
- Utilisation d'un **copy activity** pour effectuer des transformations et transférer les données.

Outils : **Azure Data Factory**, **Azure SQL Database**, **Azure Data Lake Storage Gen2**, **copy activity**

## Formation

Diplôme d'ingénieur : **Big Data & Machine Learning**, Septembre 2022 - Septembre 2024

EFREI - Paris

## Langues

Français : Langue maternelle

Anglais : Courant