IBRAHIMA DIATTARA

Email: ibrahima-diattara1992@hotmail.com

TEL:06 05 72 56 05

Certifié Microsoft Azure Data Engineer | Google Cloud Professional Data Engineer | Databricks Spark Scala

Parcours Professionnel

EDF depuis Octobre 2023

Rôle : Data Engineer

Projet : Migration d'une plateforme Data vers Google Cloud Platform (GCP)
Participation à un projet de migration d'une architecture data existante vers GCP, dans un objectif de modernisation, de montée en performance et de simplification de la maintenance des traitements de données.
Réalisations :

- ✓ Migration des données d'Amazon S3 vers Google Cloud Storage (GCS)
- ✓ Transformation des pipelines Apache NiFi en DAGs Airflow développés en PySpark
- ✓ Migration des tables external SQL Server vers des tables external BigQuery
- ✓ Réécriture et optimisation des requêtes SQL Server vers le langage SQL de BigQuery
- ✓ Mise en place de tests de validation (fonctionnels et techniques)
- ✓ Rédaction de la documentation technique des nouveaux workflows

Environnement technique :
Nifi, SQL Server, BigQyery, Google Cloud Storage, Airflow,PySpark

Projet : TEGG – Traitement de données piézométriques
Projet visant à récupérer, traiter et analyser des données issues de capteurs mesurant le niveau des nappes phréatiques (données piézométriques), afin de garantir leur qualité et produire des analyses fiables.
Réalisations :

- ✓ Optimisation des flux Nifi
- ✓ Devolopper des Custom Processors Java dans Apache NiFi
- ✓ Récupération automatique des fichiers depuis Amazon S3
- ✓ Développer des procédures stockées Neo4j en Java
- ✓ Intégration et orchestration des flux de données via Apache NiFi
- ✓ Contrôle qualité des données avec application de règles métiers spécifiques avec Nifi
- ✓ Mise en place d'une couche de traitement avancé avec Spark en Scala (agrégations, détection d'anomalies)
- ✓ Insertion des données enrichies Neo4J
- ✓ Indexer les logs applicatif dans elasticsearch
- ✓ Mise en place des visualisations de type Data Table dans kibana pour surveillller l'evolution des traitements

Environnement technique :
Nifi, Spark/Scala, Neo4j, Amazon S3, Java et groovy

SOCIETE GENERALE JANVIER 2021 – Octobre 2023

Rôle : Data Engineer/ Dev Ops Big Data

Projet Metroilogy :Développement d'une application Python/Flask avec une interface Swagger, permettant aux utilisateurs de créer leurs propres pipelines de données pour la collecte des métriques et des logs de leurs machines.
Réalisations :

- ✓ Automatisation de la création des flows Apache NiFi, des topics Kafka, des policies Ranger, et des répertoires HDFS.
- ✓ Création automatique des index patterns Elasticsearch pour permettre la visualisation et la création de dashboards dans Kibana.
- ✓ Gestion automatisée des index templates Elasticsearch pour structurer les données à l'ingestion.
- ✓ Mise en place des Column-Level Security dans Elasticsearch pour protéger l'accès aux champs sensibles.
- ✓ Configuration de l'Index Lifecycle Management (ILM) dans Elasticsearch pour gérer automatiquement le cycle de vie des index (rollover, suppression, etc.).

Projet : Montoring Metrology :L'objectif principal de ce projet est de surveiller en temps réel les clusters **NiFi** et **Kafka** à travers des tableaux de bord et des systèmes d'alerte, afin de prévenir toute perte de données et de garantir une gestion fluide des flux.
Realiations :

- ✓ Configuration de PrometheusReporting dans NiFi afin de collecter et surveiller les métriques du cluster (par exemple, les files d'attente et le nombre de threads).
- ✓ Mise en place de Telegraf pour collecter les métriques des machines hébergeant Kafka.
- ✓ Installation de Jolokia pour récupérer les métriques des brokers Kafka.
- ✓ Déploiement de Kafka Lag Exporter afin de collecter les retards des consommateurs Kafka et ainsi éviter la perte de données.
- ✓ Création de dashboards Grafana pour un monitoring complet des clusters NiFi et Kafka.
- ✓ Tuning des processors NiFi et des partitions Kafka pour optimiser les performances en cas d'alerte.

Environnement technique :
Nifi, Kafka, Jolokia, Kafka Lag Exporter, Grafana , Druid , Telegarfe, Prometheus,

Projet Upgrad HDF : L'objectif principal de ce projet est d'étudier les impacts de l'upgrade HDF (Hortonworks Data Platform)
Realisations :

- ✓ Étudier l'impact de l'upgrade sur les flux NiFi : Une analyse détaillée est effectuée pour identifier l'impact de l'upgrade HDF sur les flux NiFi, en particulier sur la performance, la compatibilité des processeurs
- ✓ Étudier le comportement des commsomateurs et producteurs Kafka apres l'upgrade
- ✓ Optimiser les flux de données après l'upgrade :

Environnement technique : Python, NIFI, Groovy, Kafka, Ranger, ElasticSearch/Kibana, HDFS, Fluentd, Telagraf

Projet DHR-Usage-Layer. L'objectif de ce projet est de rendre exploitable la BV (Business View) technique de la DRH
Réalisations :

- ✓ Automatiser la création ou la suppression des tables hive
- ✓ Mettre à jour les partitions des tables hive selon une source de données
- ✓ Collecter les données de Hive et les inscrire dans une base de données PostgreSQL
- ✓ Indexer les logs applicatifs dans Elasticsearch

Environnement technique :

Spark/Scala, HUE/hive, kafka, Elastcisearch/kibana, Jenkis, SonarQube et Nexus

SNCF FÉVRIER 2018 -JUIN 2020

Rôle : Devops Big Data / Data Engineer

Projet HDF : Dans le cadre du projet de migration Azure de la plateforme Big Data SNCF, il a été décidé de démarrer un chantier spécifique à l'ingestion de données en utilisant HDF (Hortonworks Data Platform)

Réalisations :

- ✓ Installation de NIFI et Kafka via Ambari Server
- ✓ Sécurisation de NIFI et Kafka en mode SSL et SASL_SSL respectivement avec keytool.
- ✓ Mise en place du système de connexion de NIFI via LDAP
- ✓ Gestion des ACLs des topics Kafka avec Apache Ranger
- ✓ Installation de NIFI Registry pour la gestion du versioning des flux NIFI.
- ✓ Faire le Tuning du cluster Nifi et Kafka
- ✓ Création et configuration de pipelines d'ingestion de données via NiFi, permettant de collecter, transformer et transférer efficacement les données vers les systèmes cibles dans la plateforme Big Data.
- ✓ Mise en place AmbariReportingTask pour le monitoring du cluster NIFI
- ✓ Mettre en place un système d'alerte et de monitoring avec Grafana et InfluxDB : flux de données

Environnement technique:

HDF(Nifi, Kafka, zookeeper, grafana, ambari), Groovy, influxDB, keytool

Projet DQM :L objective de ce projet est de développer une application transverse qui permet de transférer les données d'input vers raw
Réalisations

- Appliquer certaines vérifications sur les données (pas de colonne manquante…)□ Faire des transformations de format (CSV to JSON, JSON to parquet, XML to CSV, ...)
- Orchestrer les Jobs Spark/Scala via Rundeck en utilisant azure service bus comme système d'alerte

Environnement technique :

Spark, Scala,maven , Jenkins,Nexus, SonarQube Intellij ,HDFS, Azure service bus et Rundeck

Projet Sillon :
Estimer les coûts de péage des sillons selon la définition des différentes redevances, suivre et contrôler la facturation des sillons (infrastructure des trains)

Réalisations :

- Récolter les données avec NIFI
- Stocker les données dans HDFS
- Développer une application Spark/Scala pour nettoyer et agréger les données

Environnement technique :
Nifi, HDFS/ HADOOP, Spark,Scala ,Maven , Github Jenkis et Nexus

UNIVERSITE PARIS SORBONNE DEPUIS JUILLET 2021

Rôle : Professeur Data Engineering (Databricks, Google Cloud , Azure, Spark/Scala, Nifi, KAFKA , ELK, Airflow/PySpark)

RENAULT OCTOBRE 2016 - OCTOBRE 2017

Rôle: **APPRENTI ARCHITECTE BIG DATA**

Projet: Rebook **:** Optimiser les tâches des garagistes, pour que ces derniers passent moins de temps autour des véhicules lors descontrôles
Réalisations :

- ✓ Récolte des données(Nifi),
- ✓ Stockage données chaudes(ElasticSearch),
- ✓ Stockage données froides (HDFS/ HDOOP)
- ✓ Utiliser kibana pour la visualisation

Projet: Maintenance prédictive
Les robots reçoivent pour exécuter leurs tâches. Il arrive que ces robots s'arrêtent ou tombent en panne, bloquant ainsi toute la chaîne de production. Les besoins de l'équipe maîtrise d'œuvre manufacturing BI sont doubles :
Quantifier les arrêts sur panne par tranche de durée
Prédire les pannes
Réalisations :

- ✓ Récolte des données(Nifi)
- ✓ Stockage données (HDFS/ HDOOP)
- ✓ Etablir un modèle statistique pour détecter des séquences de pannes à l'aide de machine learning (Zeppelin/Spark MLib /Scala)
- ✓ Quantifier les arrêts sur panne par tranche de durée avec Hive en utilisant HUE

Environnement technique :
HDF(nifi, kafka, zookepper, Storm, ambari), HDP (HDFS, Spark, Zeppelin,Oozie, Sqoop, Ranger, Hive,HBase)
ELK(ElasticSearch, logstash &Kibana), HUE : Hadoop User Experience

SAFRAN AVRIL 2016-SEPTEMBRE 2016 :

Rôle: STAGE OUTILS SYSTÈME ET BASES DE DONNÉES
Réalisations:

- ✓ Analyser l'existant et développer un outil qui permet de consulter et d'administrer les bases de données Oracle
- ✓ Mettre en place un système de log, de cryptage et de décryptage, afin de superviser les administrateurs

Environnement technique : Java/swing et toad-

ALSTOM JUIN 2015 -SEPTEMBRE 2015 :

Rôle: Stage développement d'interface/base de données Réalisations:

- ✓ Modifier la structure de la base de données et développer des interfaces de saisie
- ✓ Compiler de nombreuses informations venant pour la plupart de fichier Excel, afinde les implanter dans la base et générer des fichiers excel pour la planification des risques

Environnement technique :
Bootstrap, Php, JavaScript, Jquery, dom et MySQL Workb

COMPÉTENCES TECHNIQUES

ETL	Nifi
Reporting	ELK/Grafana/DataDog
Base de données	Hbase, Cassandra, Redis, MongoDB, InfluxDB, Neo4J, MYSQL
Langages de développement	Scala/Java/Groovy/Python/Bash

COMPÉTENCES MÉTHODOLOGIQUES

Méthodes	Agile
Outil	Jira/Confluence

FORMATION ET CERTIFICATIONS

2021
- ➢ Databricks Certified Associate Developer for Apache Spark Scala
- ➢ Microsoft Certified Azure Data Engineer
- ➢ Google Cloud Professional Data Engineer

2016 Master2 Data Scientist (Université Paris Saclay)

2015 Master1 Business Intelligence (Université Paris Saclay)

2014 Licence3 Informatique (Université Paris Sud, Orsay