Je m'appelle Matin.
J'ai environs 5 ans d'expériences avec 2 ans en Big Data et 3 ans en Java orienté Web.

Les téchnologies que maitrise sont

- Spark, Spark JDBC, Sqoop, Oracle ODI pour l'ingestion des données dans un data lake.
- Hadoop cluster, Hive, HDFS, Cassandra, Amazon S3 pour le stockage des données
- Spark/Scala, Hadoop MapReduce pour le traitement des données
- Maven, Git, Nexus, Bash, IODA pour la gestion des projets
- Intellij IDEA, Apache Zeppelin, Cloudera HUE, Spark-shell en tant que l'environnement de travail

PROJET RCI - CMDM

- C'est un projet de RCI
- J'ai travaillé sur ce projet pendant 1 an et 7 moi
- Le besoin est de construire une vision globale, dite 360 degré des clients en se basant sur leurs contrats
- Nous avions un cluster Hadoop Kerberosé de distribution Cloudera et un cluster Cassandra de Datastax

Equipe

- Un project manager
- Un scrum-master
- Une personne pour la fronte / Angular (Mohamed Chaaban) - 6 mois
- Une personne pour la partie Datalake
- Une personne pour le moteur / clustering
- Une personne pour l'API Spring

Mon rôle

- J'ai utilisé d'abord Oracle ODI, puis Sqoop et finalement Spark JDBC pour l'ingestion des données dans des tables Hive dans le projet INGESTI.
- J'ai utilisé Spark/Scala, Spark SQL, Spark ML pour le traitement, transformation, filtrage, nettoyage des données dans le projet Datalake. J'ai souvent utilisé Cloudera HUE/Spark shell pour faire du HQL, des jobs simples et pour la consultation des donnés.
- J'ai utilisé Spark/JDBC pour exporter les données vers les sources dans le projet Exporter et Spark SQL / Scala pour la dénormalisation dans Cassandra.
- J'ai travaillé avec des fichiers txt, csv, xlsx pour le traitement des données de type transactions, taux-d'échanges, etc.
- J'ai utilisé Maven, Nexus et IODA pour le déploiement des projets en valid et en production.
- J'ai utilisé des script Kornshell (.ksh) pour orchestration des jobs via Autosys (Ordonnanceur)

PROJET ALSTOM - DataLab

- C'est un projet de ALSTOM
- J'ai travaillé sur ce projet pendant 6/7 mois
- Le besoin est de réaliser des statistiques sur les usages d'application métier et des serveurs d'applications en se basant sur les flux de connexions et de donnés de référentiels.
- Nous avions un cluster Amazon EC2 et S3 pour le stockage des données avec Kibana/Elasticsearch

Equipe

- Un project manager interne plus Guillaume Pinot
- 2 data engineers
- 2 data-scientists

Mon rôle

- Analyser le projet et faire la documentation technique et fonctionnel ⇒ L'architecture
- J'ai réalisé des reporting dans des notebooks Zeppelin avec Spark/Scala ⇒ Programmation
- Participation aux workshops pour la partage d'information avec des autres équipes de GE

Mes motivations

- D'après ce que vous avez racontez, je le trouve très intéressant et je suis partant !

Mes compétences techniques