

**Cahier des charges de la mise en place d’un système de déduplication de clients**

**Tierce Maintenance Applicatives**

**des applications Business Intelligence**

# Objectif du stage

L’objectif de ce cas d’usage est la mise en place d’un système permettant de regrouper les clients dupliqués en provenance des différents systèmes fournis par notre client.

Vous trouverez dans le document ci-dessous une description de l’existant ainsi que la solution cible à mettre en place.

**La date de remise du projet est le --/--/2019.** Tous les projets reçus après ce délai ne seront pas pris en compte.

# Presentation du contexte

Notre client est une banque qui vise à améliorer la connaissance de chacun de ses clients et de mettre à jour les fiches (clients) dont elle dispose pour mettre en place des campagnes marketing pertinentes, et personnalisées et de retrouver tous les contrats de ce client.

Actuellement, notre client dispose de plusieurs fiches clients qui peuvent faire référence à un même client.

# Description de la Mission

Le premier objectif de ce projet est de mettre en place une plateforme permettant de matcher les clients en se basant sur des clés de rapprochement (nom, prénom, date de naissance, adresse, ...) et de calculer un score de similitude avec un poids défini pour chaque champs, et de définir une limite qui permettra de grouper nos clients.

Exemple :

Notre seuil de tolérance : **80%** 🡺on ne groupe que les clients au-dessus de ce seuil

* [client 1, client 2, client 3, … client N] ont une similitude de **85%** => On considère alors que c’est un même client et on crée à partir de ce groupement client un Golden Client avec les données les plus fraiches.
* [client 1, client 2, client 3, … client N] ont une similitude de **79%** => On considère alors qu’il ne représente pas le même client et on les éclate.

L’idée, c’est de stocker les goldens clients dans Le DataLake et dans un moteur de recherche ou une base de données Nosql en temps réel afin de pouvoir les retrouver facilement.

L’architecture devra être nécessairement évolutive.

Plusieurs points doivent être abordés :

1. Choix de la plateforme Big Data (Hdp, Mapr Ou cloudera )
   * Choix des outils de chargement et de mise à disposition des données (Talend, Nifi, … )
   * Choix Base de données Nosql ou du moteur de recherche des données (Hbase,cassandra, mongodb, Neo4j, ElasticSearch, Solr, … )
   * Etc,
2. Modèle des données
   * Structuration des méta-informations des données stockées
   * Moteur de recherche de l’information
   * Etc,
3. Gestion des flux
   * Administration de la plateforme
   * Mise en œuvre de chargement de nouveaux flux
   * Mise en œuvre de nouveaux flux de mise à disposition de données pour traitement
4. Analyse des données

* Génération du reporting
* Génération des graphes