**Cahier de charges de la mise en place d’un système de collecte de métriques en temps réel**

**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 de de collecte de métriques en temps réel dans le domaine de l’industrie.

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 DERNIER JOUR DU PROJET FIL ROUGE.** Tous les projets reçus après ce délai ne seront pas pris en compte.

# Presentation du contexte

Notre client est un grand compte de l’industrie qui souhaite mettre en place un système de collecte de métriques provenant de ses produits, dans une approche IoT, afin de mieux cerner leur usage réel. Les métriques collectées serviront par la suite à la fois aux data-scientistes afin d’élaborer un modèle de garanties des produits, ainsi qu’aux ingénieurs de production qui s’en serviront pour éviter une sur-spécification des produits mis en vente.

Une sur-spécification des produits entraine des coûts supplémentaires lors de l’élaboration de ces derniers, ainsi notre client se retrouve avec des produits plus chers que ses concurrents engendrant quelquefois une disqualification lors des appels d’offres.

Actuellement, notre client n’a aucune visibilité quant à l’utilisation réelle de ses produits.

# Description de la Mission

Le premier objectif de ce projet est de mettre en place une plateforme récolte des métriques en temps réel capable de stocker un ensemble de métriques prédéfinies ainsi que les logs des sondes embarquées sur les produits.

Le principe est de commencer dans un premier temps avec des métriques fournies en mode batch via des fichiers plats de type CSV, le data-modèle de ces métriques seras quant à lui fournit sous format XML. Les données récoltées doivent être enrichies en les croisant avec des informations sur le produit (nom du produit, type du produit, composantes internes du produit,…) puis de mettre l’information finale dans des tables afin qu’elles puissent être exploitées.

Dans un second temps le mode batch sera remplacé par un bus de données dans lequel les informations vont transiter.

L’architecture devra être nécessairement évolutive.

Plusieurs points doivent être abordés :

1. Choix de la plateforme Big Data (Hadoop, Mapr, cloudera, HortonWorks )
   * 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, … )
   * Choix du bus de données (Kafka, Mapr-Streams, RabbitMQ,….)
   * Mise en place des outils de data-mining (R, Spark, Zeppelin, Spotfire, Tableaux…)
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