Date de rédaction : 9 Oct 2015
Nom du rédacteur du document : Denis Bouhineau
Spécialités : Informatique
==========

Cas d'étude Hubble: MoocAZ
Scénario hubble : Collecte sur UT, seulement, pour l'instant
Personnes impliquées pour la collecte : Denis Bouhineau, Tony Doat, Matthias, Matthieu
Période de la collecte : Juin-Sept 2015

Dispositif d'apprentissage (Etude de cas de Hubble)
[...]

Décrire en quelques mots la problématique posée :
[...]

Description du stockage des données:

Plateformes/outils utilisés: UnderTracks

Points forts de ces plateformes :	Points faibles :
- Accès aisé - Format simple	- En cours de développement

Production des données avant le traitement :

Décrire le processus de production des données brutes : [...]

Liste des variables initiales : Nom et Description

- il y en avait beaucoup (environ cent quarante), dans un premier temps je n'ai pas regardé ce qu'elle contenait (ni en terme de sémantique, ni en terme de remplissage, mais visiblement il y avait des données permettant d'identifier un temp, un utilisateur, un objet sur lequel une action avait été effectué, essentiellement des actions de récupération via le web) : type log de serveur web avec des informations supplémentaires

Plateformes/outils utilisés: [...]

Description des pré-traitements:

Objectifs des pré-traitements concernant les logs essentiellements (d'autres données/informations ont été également fournies et pourront être sauvegardés sur UnderTracks, travail prévu pour bientôt) :

- Choix sur la structure des données (voir ci-après)
- Transformation diverses (technique) pour rendre compatibles les données reçues avec les formats acceptés par UnderTracks.

Décrire le processus de pré-traitement:

- Avant téléchargement, la structure des logs de données (en JSON) était incompatible avec le format demandé par UnderTracks (UnderTracks demande un CSV) et 2 transformations possibles ont été testées :
  - Transformation privilégiant les attributs (variables) pour lesquels les données étaient régulièrement fournies et un attribut « vrac » pour les informations rarement fournies : objectif, avoir une table avec un nombre raisonnable d'attributs (variables)
  - Transformation conservant toutes les colonnes, même si la plupart seront quasi vide …
- Dans les deux cas, les transformations ont été de plusieurs types :
  - Réduction de la taille (de l'ordre du Go ou plus de données) à un taille compatible avec l'utilisation actuelle d'UnderTracks (de l'ordre de qlq dizaines de Mo). La sauvegarde de l'ensemble des données est possible, mais n'a pas été effectué à ce jour (prévu pour bientôt).
  - Transformation du JSON en CSV
  - Validation/Modification du CSV pour vérifier que le nombre de colonne est constant, que le csv est « simple », que le codage des caractères est en UTF8, …

Plateformes/outils utilisés:

- Divers éditeurs de texte/code acceptant les « gros fichiers »
- Commandes shell unix
- Excell

Points forts	Points faibles
- Potentiellement, peut « tout » faire	Demande une certaine compétence, parfois de la programmation

Description des analyses :

[...]

Description des données produites au cours du traitement

[...]

Description des Itérations

[...]