## Plan de gestion des données

INFORMATIONS GENERALES SUR LE PROJET

Renseignements administratifs:  
Acronyme:  
Code décision:  
Titre: Energy aware Federated Learning (FL) task scheduling on mobile phones  
Nom du coordinateur: Wapet  
Prénom du coordinateur: Patrick Lavoisier   
Affiliation: Laboratoire LIRIS, INSA Lyon  
Contact concernant le PGD: Vlad Nitu  
Version du PGD: Version 1  
Date: 23 Juin 2022

### 1. Description des données et collecte ou réutilisation de données existantes

1a. Les données sont collectées par l’utilisation des outils de mesure

- de la consommation énergétique sur les téléphones mobiles.

- D’une part par l’utilisation de l’API “dumpsys batterystat” d’Android

- D’autre part par l’utilisation d’un powermeter , le MONSOON Power monitor.

- De la charge totale de calcul effectué par une tâche sur le téléphone. Cette donnée est recueillie par calcul via une application que nous avons développée en interne.

Toutes ces données sont collectées au cours d’un procédé expérimental consistant à faire varier les

- le nombre de threads effectuant une même tâche sur le téléphone

- l’assignation des tâche aux différents types de coeurs sur le téléphone;

- la fréquence de ces coeurs

- la présence ou non des tâches interactives sur les sockets ou groupes de coeurs du téléphone

- la température externe du téléphone

Une fois obtenues ces données sont formatées pour être entrainées par un modèle dans le but de les interpréter.

La provenance des données est sensée être documentée dans la publication scientifique que nous prévoyons produire. Elle devra contenir avec beaucoup plus de détails les protocoles expérimentaux et les algorithmes utilisés pour l’interprétation de ces données. .

1b. Les données sont constituées:

- des fichiers textes décrivant les paramètres de chaque expérimentation, ainsi que les extraits des fichiers de logs contenant les resultats recherchés par l’expérimentateur (par exemple la workload calculée par les threads)

- Les graphiques (sous format d’image .png) construits sur la base des données extraites manuellement des fichiers précédents et décrivant les métriques comme l’efficacité énergétique en fonction de quelques une des combinaisions des paramètres d’expérimentations (qui sont en fait appelés configurations). La générations des graphiques se fait avec du code python utilisant la librairie matplotlib.

- Les fichiers textes ont été choisies pour journaliser les expérimentations car elles sont moins contraignantes que les formats exels et peuvent facilement contenir les logs des logiciels utilisés. De plus une attention a été mise sur la lisibilité des données, chaque bloc de texte décrivant une expérimentation.   
 - Le format des images (.png) facilite la visualisation des résultats d’expérimentation et permet aux parties prenantes du projet d’aisément émettre des avis. De plus le code source de génération de ces images est à proximité des données brutes au format txt, c’est à dire dans le même dossier.

- Il y’a aussi comme données générés les fichiers produits par le powermeter. Mais vu que ce qui nous intéresse ce sont les informations moyennes décrivant chaque expérimentation dans leur globalité, ces données sont concervées au cas où l’on aurait besoin d’informations granulaires sur l’expérimentation, c’est à dire des valeurs d’intensité ou d’Energie consommée à chaque unité de temps. Ils sont aussi disponibles aux formats .pt5 et .cd5, formats utilisé en lecture et en écriture par le logiciel powertool et sont associée à chacune des expérimentations, du powermeter sous windows.

- Au total les données générées jusqu’ici pèsent environ 2.5 GB. Avec pour l’instant 2 groupes de fichiers, l’un pour les expérimentations réalisée pour le google pixel 4A 5G, et l’autre pour le Samsung galaxy S8

### 2. Documentation et qualité des données

2a. Pour l’instant aucune méta-donnée n’a été choisie pour faciliter la recherche d’une information parmis les données collectées.   
 Cependant, il a paru jusqu’ici convenable au cours de l’évolution du projet d’organiser simplement les données collectées sous formes de dossiers.

Un dossier parent contient toutes les données collectées.

Ce dossier contient plusieurs sous dossiers et chacun des sous dossier est relatif à une série d’expérimentations partagant une même caractéristique descriptive. Comme exemple, on y trouve

- un dossiers décrivant les expérimentations [ réalisées sur le samsung galaxy en faisant varier le nombre de threads exécutant la tâche testée] – ce dossier s’appelle “samsung\_galaxy\_s8\_varying\_the\_number\_of\_threads”

- un dossier décrivant les expérimentations [ réalisées sur le samsung galaxy en faisant varier la fréquence des cores exécutant la tâche testée] – ce dossier s’appelle “samsung\_galaxy\_s8\_varying\_frequency”

Chacun de ses sous dossier (appelons les SD), contient

- les fichiers .cd5 et .pd5 générés par le powermeter

- un dossier nommé expe\_log\_file contenant l’historique des commandes saisies pour toutes les expérimentation du sous dossier SD et leur dates de saisie ainsi que les output obtenue pour ces commandes.

- un fichier nommé raw\_resut\_and\_powertool\_stats contenant aux format .txt les données brutes de chacune des expérimentations du dossier SD tel que décrit plus haut.

- des graphiques des résultats des résultats des expés décrivant chacun en fonction de chaque configuration enregistrée dans le dossier SD, l’énergie consommée par le téléphone, la puissance obtenue par le power-tool (qui peut ne pas être correcte) – la workload calculée par l’application exécutant la tâche – et le ratio energy/par workload.

- Du code python utilisé pour générer les graphiques.

- Du code python utilisé pour générer les data utilisables par un algorithme de machine learning devant aider à l’interprétation des données.

- le dossier machine\_learning\_datas contenant les données généré pour l’algorithme de machine learning devant aider à l’interprétation des donées.

Toutes ces données sont progressivement mises en ligne sur le serveur du laboratoire Liris, sur le compte github professionnel de l’agent contractuel chargé du projet.

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Pour l’instant la qualité des données collectées est assurée par:

- Le fait qu’une phase de calibration est faite au début des expérimentations, question de s’assurer de la rectitude des données.

- Le fait que les parties prenantes du projets s’assurent de la cohérence des résultats obtenues. Par exemple on s’attend à ce que la consommation énergétique de deux coeurs soit supérieure à celle d’un seul coeur. Cette cohérence est attendu à un niveau beaucoup plus fin, c’est à dire quelle doit épouser la tendance générale d’évolution des résultats suivant un paramètre donnée. Si une telle cohérence n’est pas observée l’expériementation est reprise et/ou alors ses résultats beaucoup plus investigées.

- Pour les résultats nécessitant une certaines discussion, comme le choix des paramètres en entrées ou l’utilisation d’une certaine formule pour le calcul d’un résultat. Toutes les informations nécessaure à leur compréhension sont directement contenues dans le fichier brut .txt, soit au début du fichier soit avant le résultat proprement dit.

- Nous ne nous sommes pas privés de réutiliser le vocabulaire connu dans le domaine de la mesure de consommation énergétique. Par exemple les termes utilisés par le powermeter sont les même qui ont été réutilisés pour la prise des résulats ( exemple : Ins Current (mA), Samples, Consumed Energy (mAh)).

### 3. Stockage et sauvegarde pendant le processus de recherche

3a. Pour l’instant les données et les métadonnées sont stockées :

- Sur le disque local de la machine de travail de l’agent collectant les données.

- Sur le serveur du laboratoire Liris. Sur le compte du même agent à l’intérieur de projets git-hub

Les données sont sauvegardées en moyenne journalièrement au cours de la prise de mesure, ou si jamais il y’a eu entre temps une modification des fichiers du projet.

3b. La sécurité des données est assurées par le fait que

- C’est la machine de travail de l’agent qui les héberge. Et elle est protégé au moins par un mot de pass.

- Le compte utilisateur github de l’agent est aussi protégé par un accès restraint puisqu’il fait partie de l’ensemble des ressource informatiques de l’établissement INSA.

En cas d’un incident sur la machine locale, les données pourront être retéléchargées et inversément si le serveur du LIRIS devient enaccessible, alors les informations sur la machines pourront être sauvegardées sur un autre serveur temporaire.

- pour l’instant les données sont en accès public pour être facilement accessibles par les collaborateurs non internes à l’INSA. Mais le lien de téléchargement ou de visualisation ne se partage qu’au cours des réunions au dans les comptes rendus de réunion. Et ceux qui y accèdent à réalité ne sont que les collaborateurs du projet.

- Les données ne sont pas à priori sensibles et pourront même être publiées en même temps que les éventuelles publications scientifiques qui seront associées aux projet.

- l’Insa pour l’instant a abauché un responsible de la sécurité des chargés d’informations qui se charge de la protection de données et qui nous a recommendé les bonnes pratique comme la bonne gestion des mots de pass et de nos poste de travail.

### 4. Exigences légales et éthiques, codes de conduite

4a. Aucune donnée à caractère personnelle n’est pour l’instant traitée.

4b. Pour l’instant les données générés sont la propriété de l’INSA, selon les disposition du Code de la Propriété intellectuelle, articles L.611-7 et R.611-11 à R.611-14. Et ce droit a été accordé par le contrat signé entre l’agent et l’organisme INSA Lyon.

Pour l’instant les données sont accessibles avec l’accord du supérieur héarchique direct, M Vlad Nitu.

Et aucune discussion n’a encore été fait dans le cas où plusieurs organismes devraient intervenir dans le traitement de ces données.

Les données peuvent être réutilisées par tout chercheur qui en aura accès au cas ou les publications relatives au projets sont accessibles sur interne et que le lien vers les données sera éventuellement partagé.   
 Note : au vu de la possibilité de clôture du compte de l’agent, des dispositions peuvent être prise avec l’accord du supérieur hierarchique, pour la migration de ces données vers un autre compte non forcément interne à l’INSA.

4c. L’aspéct éthique considéré ici est la justice, il est prévu que l’agent fasse de tout son possible pour respecter les engagements relatifs au données de recherches et qu’il a pris par son contrat d’embauche. Ensuite à chaque décision importante à prendre sur ces données il est prévu que son supérieur hiérarchique direct donne son avis.

Pour l’instant aucun comité d’éthique n’a été sollicité pour le traitement des données relatives à ce projet, et aucune sollicitation de ce type n’a été prévue.

### 5. Partage des données et conservation à long terme

5a. Il est prévu pour l’instant, que les données soient partagées dans le cadre d’une publication scientifique, nous prévoyons nous soumettre au canvas proposées par la conférence dans laquelle la publication va être soumise. IL est possible que certaines conférences demandent une copie des données de recherches et d’autres un lien de téléchargement pour tester sur ces données les logiciels éventuellement proposés dans la publication.   
 Au cas échéant, des dispositions peuvent être prises pour que les données de recherche leur soient accessibles.

Par exemple, uniquement les données demandées seront stockées sur un serveur créé en ligne spécialement pour l’occasion.

La date d’une telle mise à disposition des données dépendra déjà de l’évolution du travail de recherche et ensuite des échéance des dites conférence, et nous n’avons pas encore ces informations à notre portée.

De plus la durée de disponibilité de telles données pour ces conférence dépendent de la conférence en question, elles peuvent aller de quelques semaines, utiles à la validation de la publication ou carrément être à perpétuité pour la réuilitation de ces données par les pairs.

Les pairs qui devront avoir accès à ces données une fois publiée n’auront comme restriction que celles imposées par la conférence dans laquelle l’éventuelle publication aura été publié . Généralement ces derniers devront être inscrit dans une institution et faire partie d’un laboratoire de recherche.

Aucun plan de préservation de la données n’a encore été élaboré.

5b. Sur le long terme, il est préférable que toutes les données soient concervées. Aucune seléction n’est ici nécessaire. Aucun entrepôt de donnée n’a encore été sélectionné pour héberger les données à long terme.   
L’agent peut se charger de les stocker sur un disque dur externe, mais ces données risquent ne plus être publiquement accessibles.

Le plan de gestion de donnée à long terme ne fait pas encore partie de nos préoccupations puisque la données en soit n’a pas encore été exploitée de manière satisfaisant. Et elle est d’ailleurs incomplète.

Au cas où ces données devront être réutilisées. Ce pourra être pour:

- Vérifier les interprétations obtenues sur la base de ces données à l’aide d’autres outils d’analyses.

- Rechercher d’autre lessons learned obtenables par ces données.

5c. Pour utiliser les données, .cd5 et .pt5, il est nécessaire d’intaller le powertool du Monsoo power monitor sur un système d’explotation windows (> 10 de préférence).   
 Par contre la visualisation des graphes .png et la consultation du journal .txt ne nécessitent pas de logiciel spécialisé hors mis respectivement un lecteur d’image et de texte qui sont par défaut présents sur la majeure partie des systèmes d’exploitation.

Quant aux data, générés pour l’algorithme de machine learning, ces données sont aussi au format excel facilement compréhensibles et lisibles par n’importe quel langage de programmation.

Il est préférable que les données soient téléchargées avant d’être réutilisées. Mais comme mentionnés plus haut l’entrepot n’a pas encore été envisagé, les données étant pour l’instant téléchargeable depuis le compte git-hub de l’agent en question.

5d. Pour l’instant aucun identifiant unique de donnée ne sera attribué, et il n’est pas prévu que ce soit le cas. Si jamais les données doivent être consultées ou reférencées, elle devront être téléchargées ou dupliquées.

### 6. Responsabilités et ressources en matière de gestion des données

6a. Durant toute sa durée d’embauche c’est l’agent travaillant sur le projet , c’est à dire M. WAPET Patrick Lavoisier qui est responsable des données puisqu’il est les alimente et les modifie. De plus, L’INSA met à sa disposition le matériel de stockage.

C’est toujours le même agent qui élabore le PGD sous la validation de son supérieure hiérarchique M. Vlad Nitu.

Vu que le projet est prévu pour une durée d’ un an et que nous en sommes au quatrième mois sous le contrat post-doctoral à l’INSA, une dernière révision de ce présent PGD peut être faite à la fin du projet.

6b. A ma connaissance, hors mis la machine de travail et le compte git-hub hébergé sur les serveurs de l’INSA, aucune ressource n’a été allouée pour la gestion de la donnée. Le libre arbitre est laissé à l’agent quant à l’organisation des données, à sa mise en ligne et à son exploitation.

En effet une fois que les données sont collectées, elles sont immédiatement formattéee et enregistrées selon l’organisation décrite au 2.a. Puis sauvegardé sur le compte git-hub du serveur du LIRIS ( le LIRIS est un laboratoire de l’INSA).

Ce travail fait partie des tâches assignéee à l’agent dans le cadre de son contrat doctoral et il les réalise principalement durant son temps de travail.

S’il faille mettre ses données à disponibilité de la communauté internationale de recherche, et ce de manière perpétuelle, il faudra que la données soit migrées vers un serveur dédié à cela,   
 Mais vu le manque d’information que nous avons sur la question, nous ne savons pas encore quel serveur utiliser, encore moins quelles sont les possibilités nationales ou internationales pour le faire de manière assez contrôlée.