Evaluation (sur 12 points)

Exercice 1 – Sqoop (sur 4 points)

a) Importer les colonnes "order_id" et "order_status" de la table "orders" de la base de données "retail_db" en CSV dans le dossier HDFS "/user/votrelogin/ex1a" (2 points)

Pour rappel un fichier CSV est un fichier où les champs sont séparés par une virgule et les lignes par un retour à la ligne.

Le résultat doit ressembler à ça (partie tronquée) :

```
7187,PENDING_PAYMENT
17188,PENDING_PAYMENT
17189,PROCESSING
17190,COMPLETE
17191,PENDING_PAYMENT
```

Les infos de connexion sont les suivantes :

- host : jdbc:mysql://itversity.com
- table : orders
- base de données : retail_db
- username : retail_user
- password : itversity

b) Importer toutes les colonnes de la table order_items de la base de données "retail_db" en parquet dans HDFS "/user/votrelogin/ex1b" (2 points)

Il s'agit de n'importer que les order_items dont l'order_item_id est supérieur à 100.

Si on affichait ensuite (ce qui n'est pas demandé mais que vous pouvez faire pour valider votre résultat) le résultat avec Spark (en ne montrant que les 3 premiers éléments), on aurait ce résultat :

```
+-------------+-------------------+--------------------+-------------------+------
-------------+------------------------+
|order_item_id|order_item_order_id|order_item_product_id|order_item_quantity|order_
item_subtotal|order_item_product_price|
+-------------+-------------------+--------------------+-------------------+------
-------------+------------------------+
|       129174|              51692|                1014|                  4|
199.92|                   49.98|
|       129175|              51693|                1014|                  1|
49.98|                  49.98|
|       129176|              51693|                 797|                  1|
17.99|                  17.99|
+-------------+-------------------+--------------------+-------------------+------
```

```
-------------+------------------------+
```

Les infos de connexion sont les suivantes :

- host : jdbc:mysql://itversity.com
- base de données : retail_db
- table order_items
- username : retail_user
- password : itversity

=> Livrables : les 2 requêtes Sqoop

Exercice 2 HDFS - Spark (sur 5 points)

Nous souhaitons réaliser des statistiques sur une liste de diamants.

- Récupérer des sources parquet : un fichier qui contient des informations sur des diamants en lançant la commande :

```
wget -O diamonds.parquet
https://github.com/NastasiaSaby/data-sources/blob/master/diamonds.parquet?raw=true
```

- Mettre le fichier "diamonds.parquet" dans HDFS dans le dossier "/user/votrelogin/ex2"
- Nous voulons les diamants de plus de 1 carat et nous souhaitons savoir le prix max, le prix minimum et le prix moyen arrondi à l'unité par carat
- Enregistrer en parquet le résultat dans HDFS dans le dossier "/user/votrelogin/ex2result"

Si on affichait le résultat avec Spark (en ne montrant que les 3 premiers éléments), on aurait ce résultat :

```
+-----+---------+---------+-----------+
|carat|max_price|min_price|round_price|
+-----+---------+---------+-----------+
| 1.82|    15802|     5993|      12986|
|  2.4|    18541|    11988|      16362|
| 1.41|    17216|     4145|       9460|
+-----+---------+---------+-----------+
```

=> Livrables :

- La commande HDFS pour mettre les fichiers (1 point)
- Tout le contenu de ce que vous avez tapé dans spark-shell pour y parvenir (4 points)

Exercice 3 HDFS - Hive (sur 3 points)

- Récupérer des sources parquet : un fichier qui liste le nom et le salaire d'employés (il y a très peu de lignes dans ce fichier) en lançant la commande :

```
wget -O employee.parquet
https://github.com/NastasiaSaby/data-sources/blob/master/employee.parquet?raw=true
```

- Mettre le fichier "employee.parquet" dans HDFS dans le dossier "/user/votrelogin/ex3"
- Créer dans votre base de données une table "employee" Hive pour ça (vous aurez sans doute besoin de lire avec Spark pour savoir quel est le schéma des données)

=> Livrables :

- La commande HDFS pour mettre les fichiers (1 point)
- La commande pour créer la table Hive (2 points)