Année universitaire : 2021-2022
Resp. module : M.-L. Messai

TD 4 - TCSD

1 Objectifs

- L'utilisation de Sqoop « MySQL vers HDFS et HDFS vers MySQL »

2 Activité 1 : l'écosystème Hadoop

La commande suivante liste les différents composants Hadoop implémentés.

```
ls /usr/lib
```

Parmi les projets Apache autour de l'écosystème Hadoop, on trouve Sqoop (SQL + Hadoop). « *Sqoop est un outil conçu pour transférer des données entre Hadoop et les serveurs de bases de données relationnelles. Il est utilisé pour importer des données depuis des bases de données relationnelles telles que MySQL, Oracle vers Hadoop HDFS et exporter du système de fichiers Hadoop vers des bases de données relationnelles*.[1] ».

- Lancez la VM Cloudera puis démarrez le Framework Hadoop.
- Afin de Vérifier que les services ont bien démarré, tapez la commande : `sudo jps`.
- Assurez-vous que dans la liste figure : NameNode, NodeManager, DataNode, SecondaryNameNode, ResourceManager.

3 Activité 2 : Importer des données de MySQL vers HDFS

- Téléchargez le jeu de données `weblog_enteries.txt` depuis Moodle. Ce fichier présente les données de journalisation d'un site Web.
- Dans un terminal de la VM Cloudera, ouvrez un invité de commande et tapez la commande suivante :

```
mysql -u root -h localhost -p
```

- Le mot de passe (cloudera) vous sera demandé.
- Une fois connecté, vous pouvez entrer toutes les commandes SQL pour manipuler les bases de données.
- Créez une base de données nommée `logs`.
- Vérifiez que la base est bien créée. Rappel commandes SQL :
  - -> show databases; (lister les bases de données existantes)
  - -> use *database name* ; (sélectionner la base de données à manipuler)
  - -> show tables;
- Créez la table `weblogs` (md5 varchar(32), url varchar(65), request_date date, request_time time, ip varchar(15)) ;
- Vérifiez que la table est bien créée.
- Chargez les données du fichier `weblog_enteries.txt` l'aide de la commande "LOAD DATA INFILE" (Cherchez cette commande sur Internet).
- Vérifiez que le chargement est bien fait par l'interrogation de la base :

[1] https://getdoc.wiki/Sqoop-introduction

```
SELECT COUNT(*) From weblogs ;
```

- Pour vérifier que la connexion au serveur est valide et lister les tables importables :

```
sqoop list-tables --connect jdbc :mysql://127.0.0.1/3306/logs --username root -P
```

- Importez les données de la base dans HDFS :

```
sqoop import -m 1 --connect jdbc :mysql://127.0.0.1/3306/logs --username root -P --table weblogs  --targer-dir data/import
```

- Que signifie chaque option de cette commande ?
- Chaque ligne de la table est enregistrée dans un enregistrement séparé dans HDFS. Les enregistrements peuvent être stockés sous forme de fichiers texte ou en représentation binaire sous forme de fichiers Avro ou Sequence. Il existe 2 versions de sqoop: Sqoop1 et Sqoop2. Sqoop1 est l'outil largement accepté et recommandé pour les environnements de production[2].
- Vérifiez le contenu du dossier `data/import`.
- Visualisez le contenu du fichier importé dans HDFS :

```
hadoop fs -cat data/import/part-m-*
```

4 Activité 2 : Importer les résultats d'une requête sur des données de MySQL vers HDFS

- Une requête peut être utilisée à la place de la table dans l'opération d'importation. Importez que l'adresse ip, date et temps de consultation du site vers HDFS.
- Importez que les données de l'adresse ip 148.113.13.214.

5 Activité 3 : Exporter des données de HDFS vers MySQL

- Sur le SGBD MySQL, supprimez le contenu de la table `weblogs`. Vérifiez que la table est vide.
- Exportez le fichier `part-m-00000` depuis HDFS vers MySQL.
- Vérifiez que les données exportées ont bien été insérées dans la table.

6 Activité 4 : Dashboard

- Choisissez un jeu de données de votre choix.
- Créez la base de données puis la table qui va recevoir les données.
- Importez la table à l'aide de Sqoop.
- Cliquez sur le bouton Hue -> Query -> Dashboard
- Créez un index sur le fichier importé puis explorez la création de Dashbords.

[2] https://riptutorial.com
https://docs.cloudera.com/documentation/enterprise/5-4-x/topics/cdh_ig_sqoop_vs_sqoop2.html