Mettre en place un processus de capture et de transformation de données en python (ETL)

Moncef NECHE

Version 1.0

2018

1. Introduction :

Dans le cadre de la réalisation du travaille pratique, on aborde la partie extraction, transformation et chargement (ETL) de données, en utilisant le langage Python et ces librairies de connections base de données et traitements des ETLs.

En se basant sur l'environnement technique préparé dans le cours, les outils, Systèmes, IDEs, packages et leurs versions utilisés sont :

- Virtual Box : Version 5.2.8 r121009 (Qt5.6.2)
- Windows 8.1 Professionnel
- Python version : 3.6.5
- pycharm-community-2018.1.3
- PyMySQL : 0.8.1
- Petl : 1.1.1
- Pyodbc : 4.0.23
- Pip : 9.0.3
- Setuptools : 39.0.1

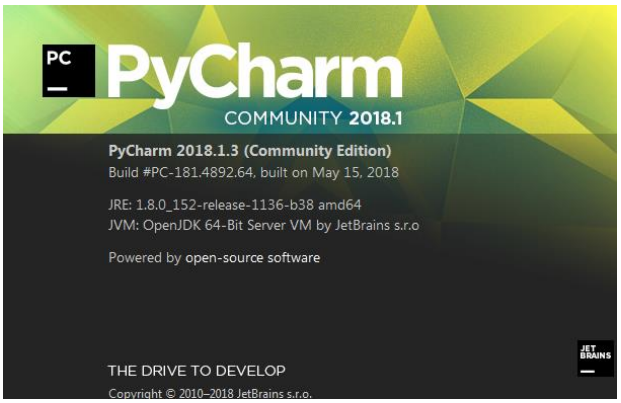

Figure 1: Environnement de développement

Ainsi, pour les données sources(SQL), on a opté pour SQL Server pour la création de la base de données sources, ce qui nous permis d'utiliser plus de packages de connexion et transformation. Mysql est la base de données de destination :

- Microsoft SQL Server Management Studio (Version 13.0.16106.4)
- SQL Server 2016
- MoWeS Portable II (Modular Webserver System Portable) : 2.2.3
- Apache HTTP Server (Version 2.2.11)
- MySQL Server (Version 5.1.35)
- PHP5 (Version 5.3.0)
- PHPMyAdmin (Version 3.1.2)

2. Étapes principales de mise en place du processus

A. Analyse des sources de données

Source de données CSV: week_cust.csv					
id	first_name	last_name	email	gender	ville
5214	Judy	Rose	jrose0@cnbc.com	Female	Michigan
5215	Andrea	Henderson	ahenderson1@blogs.com	Female	California

Source de données Json: cust_data.json					
[{"id":1	gender:"Male"	first_name:"Andrew"	last_name:"Scott"	email:"ascott0@shutterfly.com"	ville:"Connecticut"}
{"id":2	gender:"Female"	first_name:"Marilyn"	last_name:"Harper"	email:"mharper1@cornell.edu"	ville:"Massachusetts"}

Source de données SQLServer: client_DATA					
id	first_name	last_name	email	gender	ville
1	Stephanie	Andrews	sandrews0@google.cn	F	New York
2	Mildred	Duncan	mduncan1@bandcamp.com	F	Michigan

Notre dataset se compose de trois différentes sources de données, sous format de fichiers CSV, JSON et SQL comme présentées au-dessus. Après une analyse de ces sources de données on a constaté que :

- ✓ Les champs sont différents entre les trois sources (Ordre de colonnes, noms de colonnes…)
- ✓ Les données sont différentes en terme de format, type et parfois manquantes : Exemple : gender (F, M, Female, Male), id (1, '5467')

B. Création de base de données Destination

Suite à l'analyse et exploration des données, et en fonction des contraintes sur les données source, on a défini le schéma de la base de données de destination comme suit :

```
CREATE TABLE `customerdatabase`.`CustomerData` (
`Key` INT PRIMARY KEY,
`Id` INT NOT NULL ,
`First_name` VARCHAR( 20 ) NOT NULL ,
`Last_name` VARCHAR( 20 ) NOT NULL ,
`Email` VARCHAR( 50 ) NOT NULL ,
`Gender` CHAR( 1 ) NOT NULL DEFAULT 'N',
`City` VARCHAR( 20 )
) ENGINE = INNODB CHARACTER SET ASCII COLLATE ascii_bin;
```

Ce qui produit la structure de données suivante sur PhpMyAdmin :

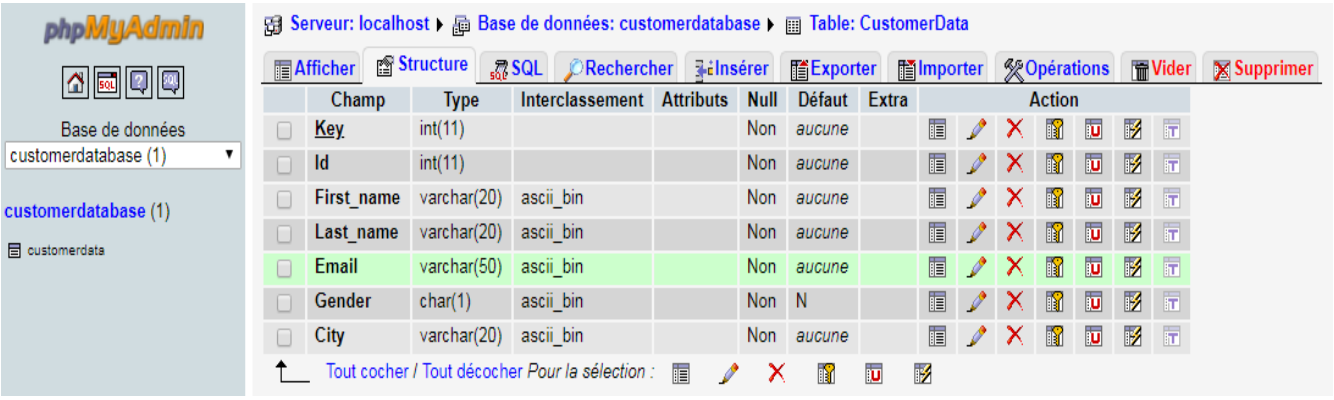

Figure 2: PhpMyAdmin - Structure de la table de destination

C. Extraction, Transformation et Chargement des données (ETL) :

Le processus ETL consiste à extraire les données à partir de différentes sources de données, puis les transformer et finalement charger les données formatées dans la base de destination.
Les données sont extraient des trois sources : Json, CSV et SQL Server, cette dernière a été créer et alimenter à partir d'un script SQL fourni. Des plugins Python sont utilisées la connexion aux bases de données, les extractions et les transformations. Les opérations de transformations manipulent les données (agrégations, typage, formatage et conversions…), et des mises en correspondance (mappages).

Une fois que les données préparées correspondent à la structure de la base de données de destination, on applique un chargement par le baies des packages Python. Le Schéma suivant (Figure 3) montre les types des sources de données et la destination.

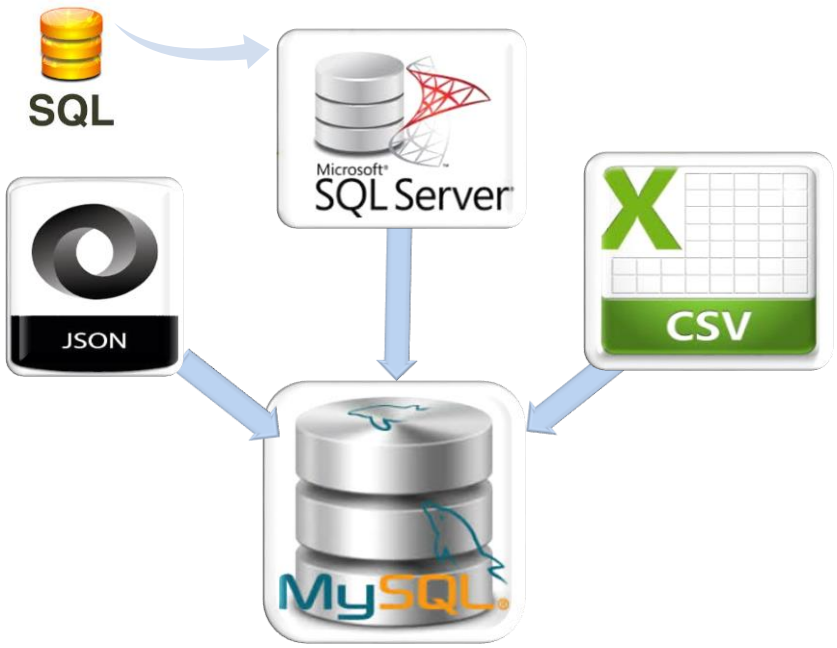

Figure 3 : Schéma du processus ETL

On peut décrire le processus à l'aide d'une notation et symboles de SSIS. Les deux figures suivantes présentent les taches et l'acheminement des traitements de flux de contrôle puis flux de données.

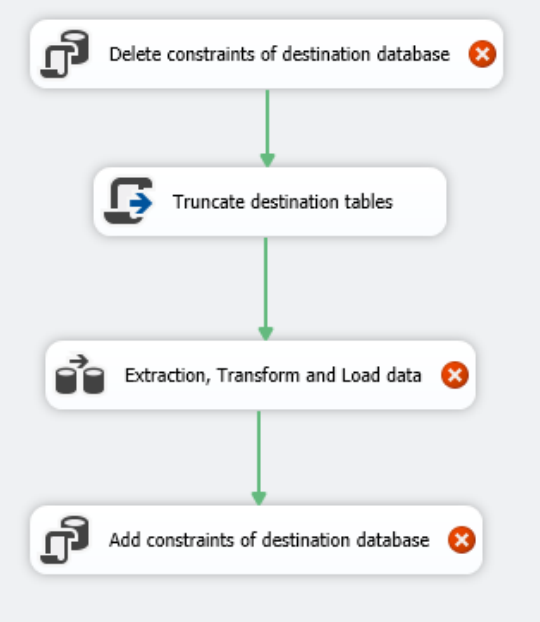

Figure 4: Schéma flux de contrôle du processus ETL (SSIS)

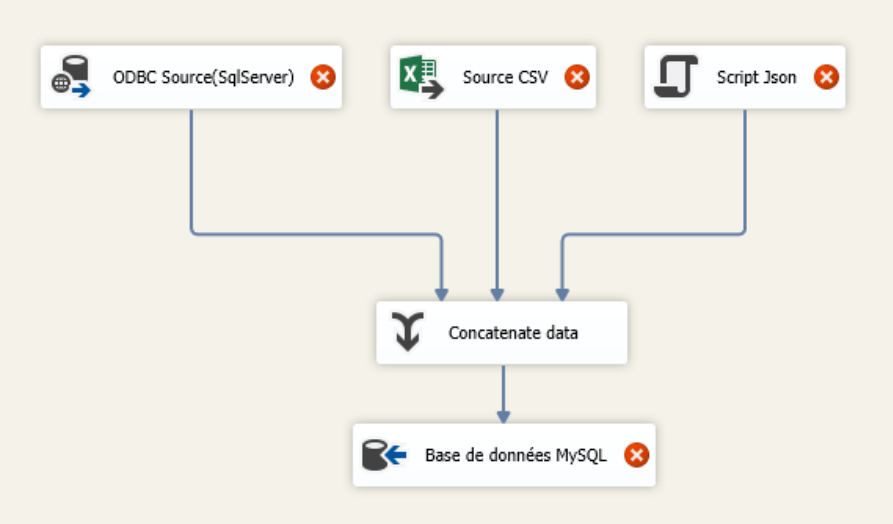

Figure 5: Schéma flux de données du processus ETL (SSIS)

3. Comment tester l'application

Notre processus ETL bien mis en place et prêt à être tester. Pour faire, on suit les étapes suivantes :

a- Vider la table de la base de données destination

b- Exécuter le script Python : etlCsvJsonSqlserverToMysql.py, qui contient les étapes :

1- Extraction des données et mappage des données provenant du fichier Json
2- Extraction des données et conversion de types à partir du fichier CSV
3- Connexion SQL Server et extraction des données et mappage
4- Concaténation et transformation des données issues des différentes sources dans un espace mémoire Appelé 'StagingArea'
5- Connexion à la base de données destination MySQL
6- Chargement de données dans la table de destination " CustomerData"

c- Tester et vérifier le bon chargement des données dans la table de destination, soit en consultant la table via PhpMyAdmin ou en exécutant le script 'databaseDestinationTest.py'

4. Conclusion

On a pu mettre en place un ETL python en utilisant les librairies pour le processus ETL, et un codage Python optimisé. Aussi, on a présenté le schéma général d'un ETL ainsi que l'adéquation et la puissance du python et la librairie PETL.

Les difficultés rencontrées ont été soulevé, a l'image des compatibilités entre les versions des packages et les mise a jours. Par ailleurs, au lancement de Mowes.exe sur la machine virtuelle, le serveur apache ne se lance pas. Ca été corrigé comme le montre la ***figure 7.***

Serveur Apache, PHP - MoWeS PORTABLE

Erreur : « Apache is not running »

« **Apache is not running** ». Le message d'erreur apparait au lancement de MoWeS :

L'erreur provient de ce que port « 80 » est déjà pris par un autre programme.

Il y a deux manières de corriger ce problème :
- Soit changer le port d'écoute **80** par 8080 dans la configuration du serveur Apache,
- Soit changer le port d'écoute **80** par 443 dans les autres logiciels (ex. Skype)

MoWeS Portable - Apache is not running - Changer port 80 par 8080

Dans le fichier de configuration d'Apache : « **apache2\conf\httpd.conf** » :
- Remplacer « **Listen 80** » par « Listen localhost:8080 ».
- Remplacer « **ServerName localhost:80** » par « ServerName localhost:8080 ».

Puis afficher la page : **http://localhost:8080/index.php**

Figure 6: Correction D'erreur 'Apache is not Running'