

Programmation Python

***Mini-Projet Python***

***« DataMigrator »***

Master Spécialisé Big Data & Cloud Computing

Imane JABAL

Mohamed AIT ABDERRAHMAN

Rachid REBIK

Encadré par:

M. Noreddine ABGHOUR

Année Universitaire 2015/2016

SOmmaire

[1 Introduction 2](#_Toc441918279)

[2 Modules utilisés 2](#_Toc441918280)

[3 Test 2](#_Toc441918281)

[4 Conclusion 2](#_Toc441918282)

# Introduction

Le langage Python permet d'accéder de manière très simple à la majorité des bases de données actuelles, y compris à des bases de données émergentes suivant la tendance actuelle du « NoSQL ».

Le but de ce projet et de réaliser une boite à outils qui permettra à un adminsitrateur de base de données d’extraire les données et les transformer en fichier binaire. Et aussin manipuler ces fichiers pour les convertir en diff2rent foròm q sqvoir JSON,CSV ..

# Problématique

A nos jours, les entreprises ont plusieurs sources de données et distribuer sur différents types de serveurs, ce qui rends difficile de les gérer sans unifier leur formats.

# Réalisation

## Outils

****

**PyCharm :**

PyCharm est un IDE spécialisé pour Python.

PyCharm existe en deux versions : une gratuit & open-source (Free Community) et une autre propriétaire et payante (Full-fledged Professional, 99$ pour un an de license).  
La version Free Community suffit amplement  
PyCharm dispose d'une interface sobre et moderne, très agréable !  
Quelques fonctionnalités :

* Auto-complétion du code (même par des variables d'un module importé !)
* Vérification du code
* Vérification PEP8
* Aide à la réalisation des docstrings
* Suggestions de correction automatique/optimisation du code
* Gestion efficace de projets
* Débuggeur intégré
* etc...

**Python :**

****

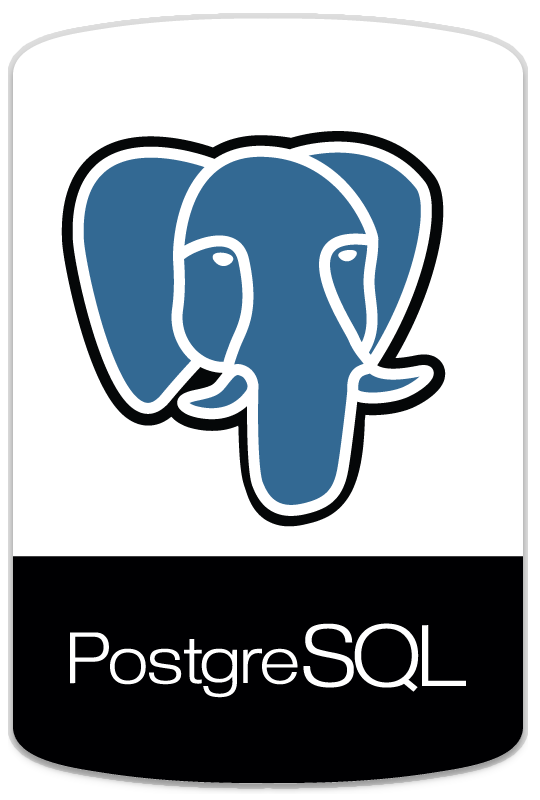
Python est un [langage de programmation](https://fr.wikipedia.org/wiki/Langage_de_programmation) [objet](https://fr.wikipedia.org/wiki/Programmation_orient%C3%A9e_objet), multi-[paradigme](https://fr.wikipedia.org/wiki/Paradigme_(programmation)) et [multiplateformes](https://fr.wikipedia.org/wiki/Plate-forme_(informatique)). Il favorise la [programmation impérative](https://fr.wikipedia.org/wiki/Programmation_imp%C3%A9rative) [structurée](https://fr.wikipedia.org/wiki/Programmation_structur%C3%A9e), [fonctionnelle](https://fr.wikipedia.org/wiki/Programmation_fonctionnelle) et [orientée objet](https://fr.wikipedia.org/wiki/Programmation_orient%C3%A9e_objet). Il est doté d'un [typage dynamique](https://fr.wikipedia.org/wiki/Typage_dynamique) [fort](https://fr.wikipedia.org/wiki/Typage_fort), d'une gestion automatique de la mémoire par [ramasse-miettes](https://fr.wikipedia.org/wiki/Ramasse-miettes_(informatique)) et d'un [système de gestion d'exceptions](https://fr.wikipedia.org/wiki/Syst%C3%A8me_de_gestion_d%27exceptions).

**MySQL :**

 MySQL est un [système de gestion de bases de données](https://fr.wikipedia.org/wiki/Syst%C3%A8me_de_gestion_de_base_de_donn%C3%A9es) relationnelles (SGBDR). Il est distribué sous une double licence [GPL](https://fr.wikipedia.org/wiki/Licence_publique_g%C3%A9n%C3%A9rale_GNU) et [propriétaire](https://fr.wikipedia.org/wiki/Logiciel_propri%C3%A9taire). Il fait partie des logiciels de gestion de [base de données](https://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es) les plus utilisés au monde, autant par le grand public (applications web principalement)

que par des professionnels, en concurrence avec [Oracle](https://fr.wikipedia.org/wiki/Oracle_Database), [Informix](https://fr.wikipedia.org/wiki/Informix" \o "Informix) et [Microsoft SQL Server](https://fr.wikipedia.org/wiki/Microsoft_SQL_Server).

**PostgreSQL :**

 PostgreSQL est un système de gestion de bases de données relationnelles objet (ORDBMS) basé sur [POSTGRES, Version 4.2](http://s2k-ftp.cs.berkeley.edu:8000/postgres/postgres.html), développé à l'université de Californie au département des sciences informatiques de Berkeley. POSTGRES a lancé de nombreux concepts rendus ensuite disponibles dans plusieurs systèmes de bases de données commerciales.

PostgreSQL est un descendant open-source du code original de Berkeley. Il supporte SQL92 et SQL99 tout en offrant de nombreuses fonctionnalités modernes :

* requêtes complexes ;
* clés étrangères ;
* triggers ;
* vues ;
* intégrité des transactions ;
* contrôle des accès concurrents (MVCC ou multiversion concurrency control).

De plus, PostgreSQL peut être étendu de plusieurs façons par l'utilisateur, par exemple en ajoutant de nouveaux

* types de données ;
* fonctions ;
* opérateurs ;
* fonctions d'agrégat ;
* méthodes d'indexage ;
* langages de procédure.

Et grâce à sa licence libérale, PostgreSQL peut être utilisé, modifié et distribué par tout le monde gratuitement quelque soit le but visé, qu'il soit privé, commercial ou académique.

**SQLite :**



SQLite est une bibliothèque écrite en C qui propose un moteur de base de données SQL. Contrairement aux serveurs de bases de données comme MySQL ou PostgreSQL, elle ne reproduit pas le schéma habituel client-serveur mais elle est directement intégrée aux programmes en utilisant des fichiers de base de données.

Elle implémente en grande partie le standard SQL92 et les Propriétés ACID.

**Oracle :**

****

*Oracle Database* est un [système de gestion de base de données](https://fr.wikipedia.org/wiki/Syst%C3%A8me_de_gestion_de_base_de_donn%C3%A9es) [relationnelle](https://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es_relationnelle) (SGBDR) qui depuis l'introduction du support du[modèle objet](https://fr.wikipedia.org/wiki/Programmation_orient%C3%A9e_objet) dans sa version 8 peut être aussi qualifié de [système de gestion de base de données relationnel-objet](https://fr.wikipedia.org/wiki/Syst%C3%A8me_de_gestion_de_base_de_donn%C3%A9es_relationnel-objet) (SGBDRO). Fourni par [Oracle Corporation](https://fr.wikipedia.org/wiki/Oracle_Corporation), il a été développé par [Larry Ellison](https://fr.wikipedia.org/wiki/Larry_Ellison), accompagné d'autres personnes telles que [Bob Miner](https://fr.wikipedia.org/wiki/Bob_Miner) et [Ed Oates](https://fr.wikipedia.org/w/index.php?title=Ed_Oates&action=edit&redlink=1).

**MS SQL Server**

Microsoft SQL Server est un [système de gestion de base de données](https://fr.wikipedia.org/wiki/Syst%C3%A8me_de_gestion_de_base_de_donn%C3%A9es) (abrégé en SGBD) incorporant entre autres un SGBDR (SGBD [relationnel](https://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es_relationnelle) ») développé et commercialisé par la société [Microsoft](https://fr.wikipedia.org/wiki/Microsoft). Il ne fonctionne que sous les OS Windows.

En fait MS SQL Server est une suite composée de cinq services principaux :

* Le moteur relationnel (OLTP) appelé SQL Server ;
* Le moteur décisionnel (OLAP) appelé SSAS (SQL Server Analysis Services) incluant un moteur de stockage pour les cubes, des algorithmes de forage (data mining) et différents outils de BI (Business Intelligence) ;
* Un ETL (Extract Transform and Load) appelé SSIS (SQL Server Integration Services) destiné à la mise en place de logiques de flux de données, notamment pour alimenter des entrepôts de données (data warehouse) ;
* Un outil de génération d'état appelé SSRS (SQL Server Reporting Services) permettant de produire des rapports sous différentes formes et exploitant les ressources du moteur décisionnel (bases "resportServer...") à la fois pour y stocker les rapports mais aussi y cacher les données de ces derniers afin de faire du "warmup" ;
* Un système de planification de travaux et de gestion d'alerte appelé Agent SQL qui utilise lui aussi les services du moteur SQL (base msdb).

D'autres services lui sont associés pour certains besoins :

* SQL Full-Text Filter Daemon launcher : pour charger certains documents (thésaurus et dll ifilter) externe nécessaires à la recherche "plain texte"
* SQL Browser : service de diffusion des services SQL destiné à faciliter la recherches des services SQL sur le réseau
* MS DTC (Data Transaction Coordinator) : coordinateur de transactions distribuées, utilisé pour des mises à jour en validation à 2 phases (2PC) combinant différents serveurs transactionnels, par forcément du monde Microsoft (XA, X/Open, TIP, WS...)

SQL Server inclut aussi de nombreux outils de développement :

* pour les bases transactionnelles, par le biais du SQL Server Management Studio (SSMS)
* pour les bases décisionnelles, le forage des données, le reporting et l'ETL, par le biais d'une surcouche de Visual Studio appelé BIDS (Business Intelligence Development Studio) puis SQL Server Data Tools depuis la version 2012.

## Modules utilisés

**Sqlacodegen :**

Automatic model code generator for SQLAlchemy : est module de génération du code automatique pour SQLAlchemy.

**SQLAlchemy :**

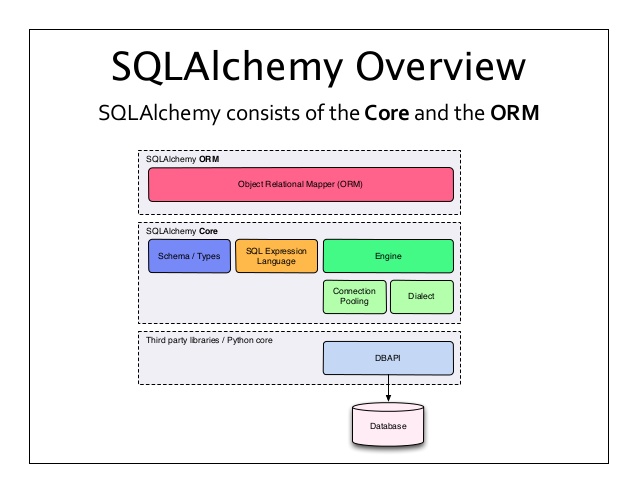
****

SQLAlchemy est un toolkit [open source](https://fr.wikipedia.org/wiki/Open_source) [SQL](https://fr.wikipedia.org/wiki/Structured_Query_Language) et un [mapping objet-relationnel](https://fr.wikipedia.org/wiki/Mapping_objet-relationnel" \o "Mapping objet-relationnel) (ORM) écrit en [Python](https://fr.wikipedia.org/wiki/Python_(langage)) et publié sous [licence MIT](https://fr.wikipedia.org/wiki/MIT_Licence).

SQLAlchemy a opté pour l'utilisation du pattern Data Mapper plutôt que l'[active record](https://fr.wikipedia.org/wiki/Active_record_(patron_de_conception)) utilisés par de nombreux autres ORM.

SQLAlchemy a été publié en [février](https://fr.wikipedia.org/wiki/F%C3%A9vrier_2006) [2006](https://fr.wikipedia.org/wiki/2006_en_informatique) et est rapidement devenu l'un des ORM les plus utilisés par la communauté Python.

**Structure de SQLAlchemy :**

****

**Pickle :**

Un pickle est un objet Python représentant une chaine d'octets. Ce processus est appelé Pickling. Donc nous avons converti avec succès notre objet en une chaîne d'octets, maintenant comment pouvons-nous revenir en arrière? Dépickler signifie reconstruire l'objet Python depuis la chaine picklée d'octets. Pour parler strictement, ce n'est pas une reconstruction dans un sens physique—cela signifie simplement que si nous avons pickler une liste, L, alors après dépicklage, nous retrouvons les contenus de la liste simplement en accédant à L.

Les termes « pickler » et « depickler » se rapportent respectivement à la sérialisation et à la désérialisation objet, qui sont des termes neutres par rapport au langage relatif à un processus qui transforme arbitrairement des objets complexes en représentations textuelles ou binaires de ces objets et inversement.

**ArgParser :**

Le module argparse est un "Command Line Parsing Module". En bon français, on dirait que le module argparse permet de parser correctement les arguments de la ligne de commande. Le module argparse est le successeur du module **optparse**. Ce module est inclus en standard dans python 2.7 et python 3.2. Vous pouvez sinon l'installer par un pip install argparse.

# Mis en place

## Spécifications

### db2bin

Extrait le schéma relation avec les données depuis une base de données fournit en paramètre

**Paramètres :**

**'-t'**, **'--type'**, help=**'Database server type'**, required=True)  
**'-s'**, **'--server'**, help=**'Database server type'**, default=**"localhost"**)  
**'-u'**, **'--username'**, help=**'Database server type'**, required=server\_param\_required)  
**'-p'**, **'--password'**, help=**'Database server type'**, required=server\_param\_required)  
**'-d'**, **'--database'**, help=**'Database name'**, required=server\_param\_required)  
**'-l'**, **'--location'**, help=**'Database path'**, required=**not** server\_param\_required)  
**'-o'**, **'--output'**, help=**'Output file'**, required=False)

### bin2json

Convertit un fichier binaire fournit en paramètre en format JSON.

**'Extract JSON, XML or CSV file from binary data file..'**  
**'-p'**, **'--path'**, help=**'Generated binary data file path'**, required=True)  
**'-o'**, **'--output'**, help=**'Output file path'**, required=False)

### bin2xml

Convertit un fichier binaire fournit en paramètre en format XML.

### bin2csv

Convertit un fichier binaire fournit en paramètre en format CSV.

# Conclusion

En définitive, ce projet nous a appris à programmer en python et nous a apporté quelques connaissances dans l’utilisation des modules python déjà prêts. Il a également été un bon entraînement dans la programmation événementielle. Au début, nous avons beaucoup réfléchie à la problématique de l'analyseur de modification redondante qui était notre première idée de projet mais, comme nous l’avons expliqué dans l'introduction, nousnous sommes plutôt dirigés vers de réutilisation des modules comme pickle, SQLAlchemy, argparse. Peut-être que nous le développerons un jour si nous en avons le temps.

A partir de cette base de script, on peut facilement et rapidement créer tous les outils que le veut, c'est l'un des avantages de python.