**بسمه تعالی**

**پایتون و استخراج، تبدیل و بارگذاری داده­ها**

اغلب کاربران نیاز به «استخراج» (Extract) ، «تبدیل» ( Transform) و «بارگذاری» (Load) داده‌ها از «پایگاه‌داده‌های» (databases)گوناگون به «انبارهای داده» (data warehouse) به منظور تجمیع داده‌ها جهت تحلیل‌های «هوش تجاری»( BI | Business Intelligence) و «تحلیل‌های داده» (Data Analysis) دارند. نرم‌افزارهای گوناگونی برای انجام فرآیند ETL تولید شده‌اند، ولی استفاده از آن‌ها برای کاربردهای ساده اغلب زیاده‌روی به نظر می‌رسد. در این مطلب چگونگی استخراج داده‌ها از SQL-Server ،MySQL و Firebird، تبدیل داده‌ها و بارگذاری آن‌ها در SQL-Server (به عنوان انبار داده) با استفاده از پایتون 3.۶ آموزش داده شده است. اکنون، کار با راه‌اندازی دایرکتوری پروژه آغاز می‌شود:



علاوه بر نصب SQL-Server ،MySQL و Firebird، نیاز به سه ماژول پایتونی که در زیر بیان شده و اجرای pip install [module\_name] وجود دارد.

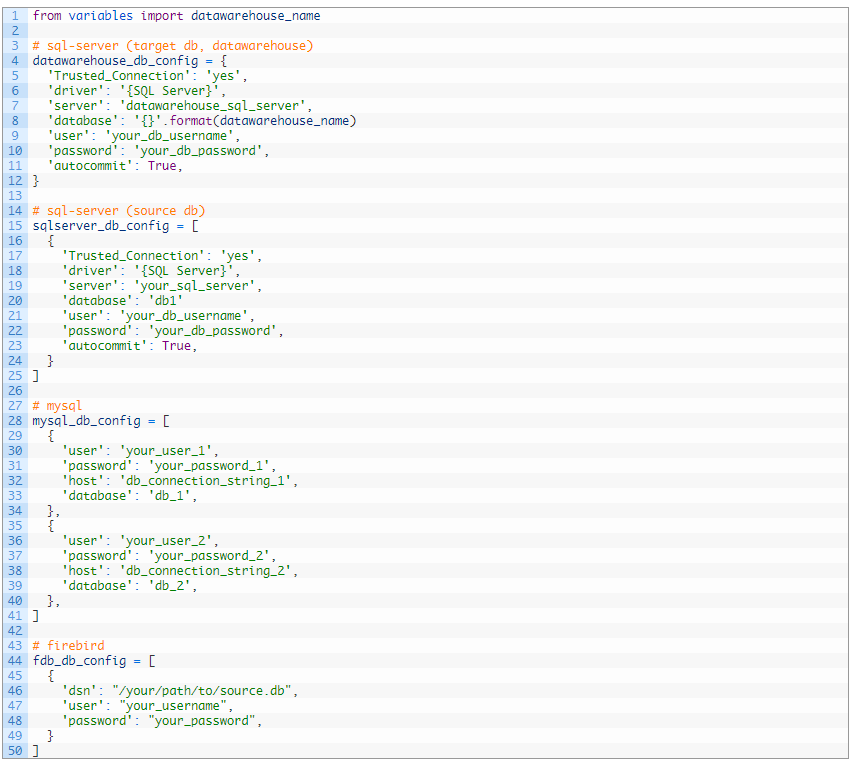
* mysql-connector-python : اتصال به MySQL
* pyodbc : اتصال به sql-server
* fdb : اتصال به Firebird

راه‌اندازی مجوزها و متغیرهای پایگاه داده

راه‌اندازی یک «متغیر» برای ذخیره‌سازی نام پایگاه‌داده انبار داده در variables.py با دستور زیر انجام می‌شود.

C:\Users\Nikan\Desktop\2.png

نصب همه پایگاه‌داده‌های منبع و هدف‌گیری رشته‌های اتصال پایگاه داده و مجوزها در db\_credentials.py به صورتی که در زیر نمایش داده شده قابل انجام است. سپس، کانفیگ باید به صورت list ذخیره شود تا بتوان بعدا آن را در پایگاه داده‌های زیادی تکرار کرد.



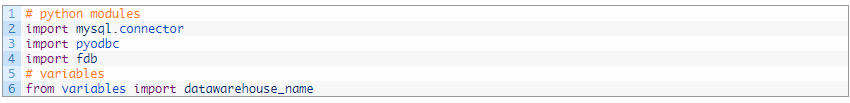
کوئری‌های SQL

sql\_queries.py جایی است که همه کوئری‌های SQL برای استخراج از پایگاه‌داده‌های منبع و بارگذاری در پایگاه‌داده هدف (انبار داده) ذخیره می‌شوند. با توجه به اینکه در این فرآیند با پلتفرم‌های گوناگونی کار می‌شود، می‌توان از «نحو» (syntax) گوناگونی برای هر سکوی داده با جداسازی کوئری‌ها مطابق با نوع پایگاه داده استفاده کرد.



استخراج، تبدیل، بارگذاری

در etl.py، ماژول‌ها و متغیرهای پایتون زیر باید بارگذاری شوند تا کار آغاز شود.



در اینجا دو متد ()etl و ()etl\_process وجود دارند. ()etl\_process متدی برای راه‌اندازی اتصال منبع پایگاه داده مطابق با پلتفرم پایگاه‌داه و فراخوانی متد ()etl است. در متد ()etl، ابتدا کوئری استخراج اجرا، داده‌های SQL در متغیر Data ذخیره و این متغیر در پایگاه‌داده هدف که در واقع انبار داده کاربر است «درج» می‌شود. تبدیل داده‌ها با دستکاری متغیر data که از نوع «تاپل» است صورت می‌پذیرد.



قرار دادن کلیه موارد در کنار هم

در main.py می‌توان از طریق مجوزها و انجام ETL برای همه پایگاه‌های داده حلقه زد. کد زیر همه متغیرها و متدهای مرتبط را ایمپورت می‌کند.



اکنون کافیست کاربر در ترمینال python main.py را وارد کند و از اسکریپت پایتونی که برای انجام فرآیند ETL نوشته شده استفاده کند.

منبع: https://codeburst.io/using-python-script-for-data-etl-53138c567906