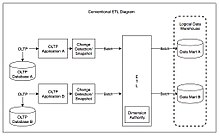
Student Name: Muhammad Shoaib Khan  
Seat Number: DS – 007 / 2021

**Implementing ETL Pipeline for Financial Data Sources**

ETL Process Description:

In [computing](https://en.wikipedia.org/wiki/Computing), **extract, transform, load** (**ETL**) is a three-phase process where data is extracted, transformed (cleaned, sanitized, scrubbed) and loaded into an output data container. The data can be collated from one or more sources and it can also be outputted to one or more destinations. ETL processing is typically executed using [software applications](https://en.wikipedia.org/wiki/Software_application) but it can also be done manually by [system operators](https://en.wikipedia.org/wiki/Sysop). ETL software typically automates the entire process and can be run manually or on reoccurring schedules either as single jobs or aggregated into a batch of jobs.

[](https://en.wikipedia.org/wiki/File:Conventional_ETL_Diagram.jpg)

Conventional ETL diagram[[1]](https://en.wikipedia.org/wiki/Extract,_transform,_load#cite_note-Kimball_2004-1)

A properly designed ETL system extracts data from source systems and enforces [data type](https://en.wikipedia.org/wiki/Data_type) and data validity standards and ensures it conforms structurally to the requirements of the output. Some ETL systems can also deliver data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Software Used in implementing this project:

1. Python
2. Microsoft VS CODE (Text Editor)
3. SQL Server 2019

Python Libraries used for Applying ETL Process:

1. OS Library:

The OS module in Python **provides functions for creating and removing a directory (folder), fetching its contents, changing and identifying the current directory, etc**. You first need to import the os module to interact with the underlying operating system.

1. Pyodbc library:

Pyodbc is an open source Python module that makes **accessing ODBC databases** simple. It implements the DB API 2.0 specification. Using pyodbc, you can easily connect Python applications to data sources with an ODBC driver.

1. Python Sys Library:

This module provides access to some variables used or maintained by the interpreter and to functions that interact strongly with the interpreter. It is always available.

1. Python petl Library:

petl is **a general purpose Python package for extracting, transforming and loading tables of data**.

1. Python ConfigParser Library:

The configparser module from Python's standard library **defines functionality for reading and writing configuration files** as used by Microsoft Windows OS. Such files usually have . INI extension.

1. Python Requests Library:

Requests is a HTTP library for the Python programming language. The goal of the project is to make HTTP requests simpler and more human-friendly. The current version is 2.28.0. Requests is released under the Apache License 2.0. Requests is one of the most popular Python libraries that is not included with Python.

1. Python DateTime Library:

The [datetime](https://docs.python.org/3/library/datetime.html#module-datetime) module supplies classes for manipulating dates and times.

1. Python json library:

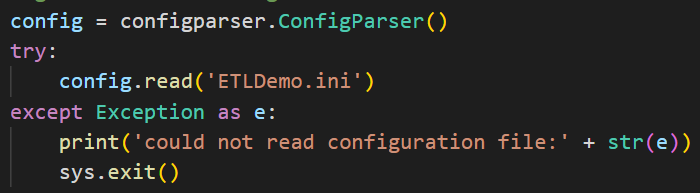
The json library **can parse JSON from strings or files**. The library parses JSON into a Python dictionary or list. It can also convert Python dictionaries or lists into JSON strings.

1. Python Decimal Library:

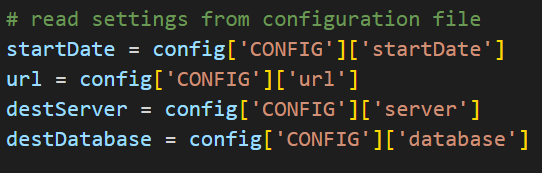
In Python, there is a module called Decimal, which is **used to do some decimal floating point related tasks**. This module provides correctly-rounded floating point arithmetic. To use it at first we need to import it the Decimal standard library module. import decimal.

**Python Coding for ETL Pipeline:**

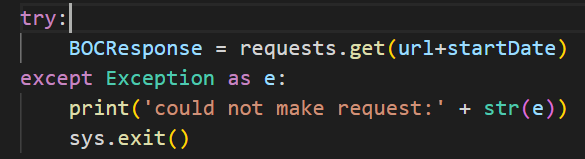
1. # get data from configuration file



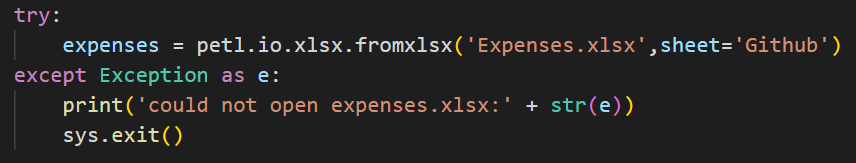
1. # read settings from configuration file



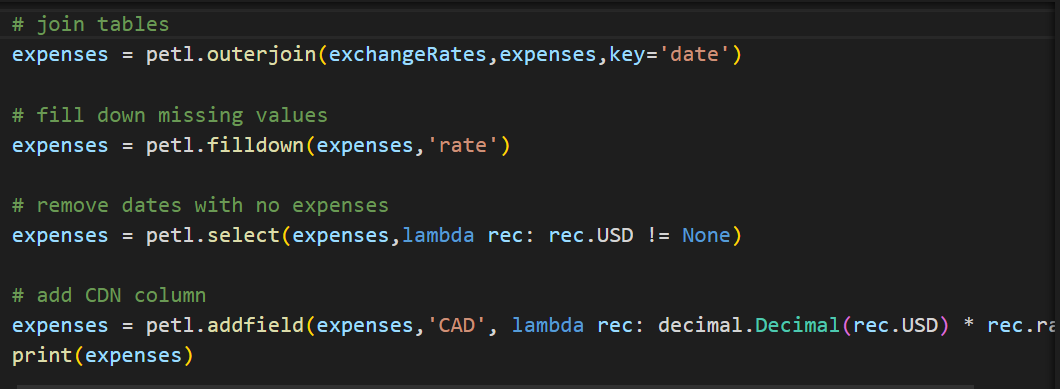
1. # request data from URL



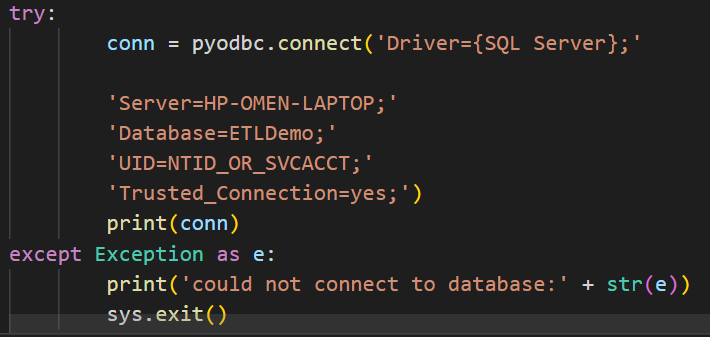
1. # load expense document



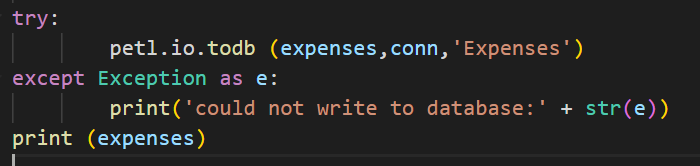
1. Data Transformation Process:



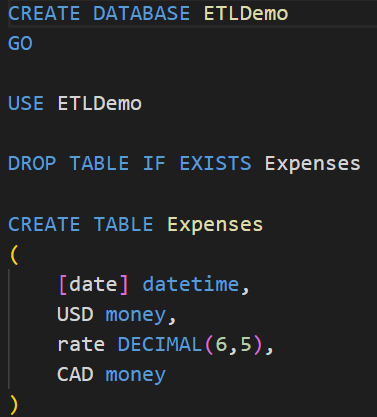
1. intialize database connection



1. # populate Expenses database table



**SQL Server Queries for Creating Schema and Table:**



Sources:

<https://en.wikipedia.org/wiki/Extract,_transform,_load>