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

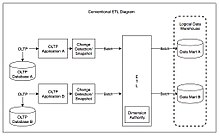
**Implementing ETL Pipeline for Financial Data Sources**

Course Name:

Datawarehouse (CT – 593)

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.

# The Importance of ETL Tools in Data Warehousing:

The importance of ETL in an organization is in direct proportion to how much the organization relies on data warehousing. ETL tools collect, read, and migrate large volumes of raw data from multiple data sources and across disparate platforms. They load that data into a single database, data store, or data warehouse for easy access. They process the data to make it meaningful with operations like sorting, joining, reformatting, filtering, merging, and aggregation. Finally, they include graphical interfaces for faster, easier results than traditional methods of moving data through hand-coded data pipelines.

ETL tools break down data silos and make it easy for your data scientists to access and analyze data, and turn it into business intelligence. In short, ETL tools are the first essential step in the data warehousing process that eventually lets you make more informed decisions in less time.

# The Importance of ETL For Financial Data:

As you read this, millions of dollars are changing hands on the global economy. To understand financial forecasts and formulate strategies, a data analyst must track the data with precision, security, and speed. The gathered raw data requires formatting as the AI and ML bots cannot read unstructured or siloed data. This is where the ETL tools come into action. ETL is the underlying foundation of financial data processing.

Data extraction tools collect both structured and unstructured data from the web and convert it into actionable insights based on your business requirements. With the ability to analyze diverse sets of data, financial companies can make effective decisions on improved customer care, fraud detection, better customer targeting, top channel results, and risk exposure assessment.

Software Used in implementing this project:

1. Python  
     
   Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming.
2. Microsoft VS CODE (Text Editor)

Visual Studio Code, also commonly referred to as VS Code, is a source-code editor made by Microsoft with the Electron Framework, for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

1. SQL Server 2019

Microsoft SQL Server is a relational database management system developed by Microsoft. As a database server, it is a software product with the primary function of storing and retrieving data as requested by other software applications—which may run either on the same computer or on another computer across a network.

1. Microsoft Power BI

Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence. It is part of the Microsoft Power Platform.

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.

**Data Source for ETL Project Pipeline:**

<https://www.bankofcanada.ca/valet/docs>

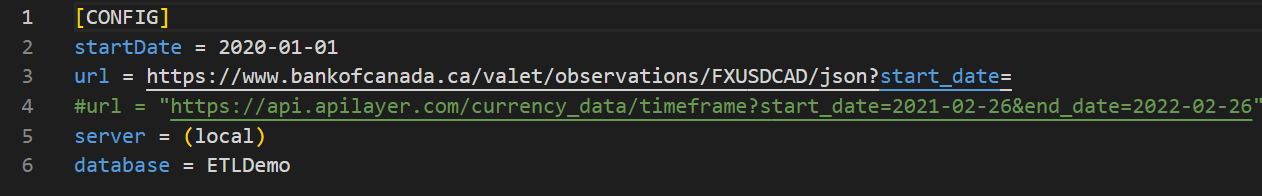
The Bank of Canada Valet Web Services offers programmatic access to global financial data. By using the Valet API, you can retrieve financial data and information from the Bank of Canada — such as daily exchange rates of the Canadian dollar against the European euro.

**Formats**

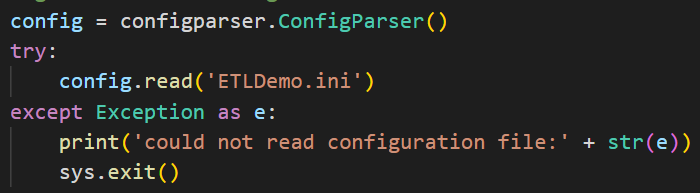
We provide data in JSON, XML, and CSV formats.

**Python Coding for ETL Pipeline:**

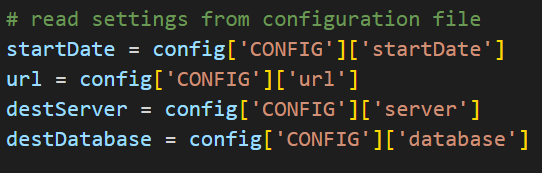
Create the configuration file:

****

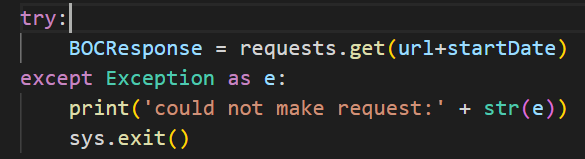
1. # get data from configuration file



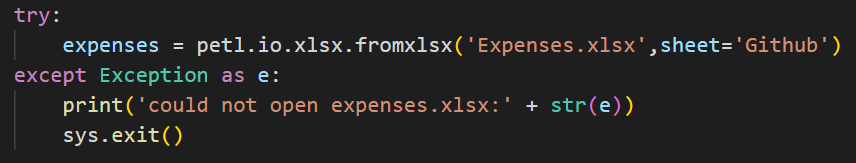
1. # read settings from configuration file



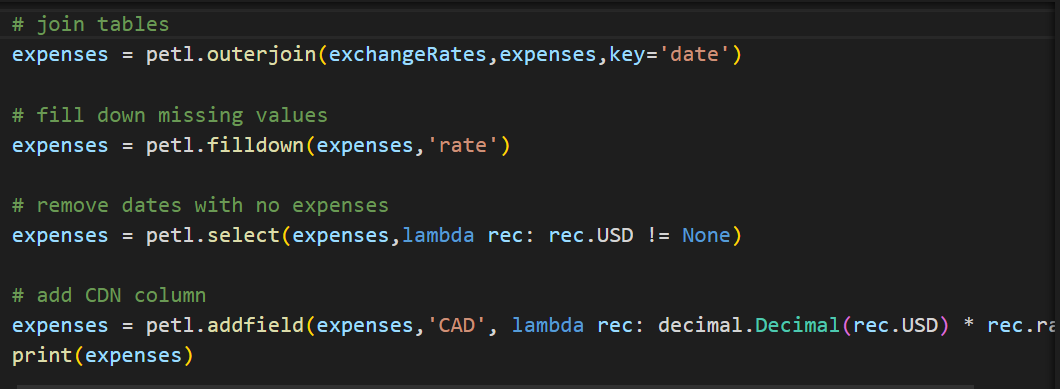
1. # request data from URL



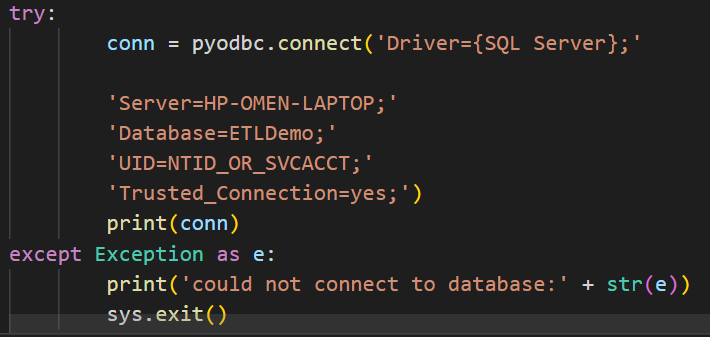
1. # load expense document



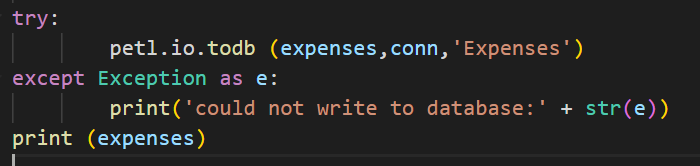
1. Data Transformation Process:



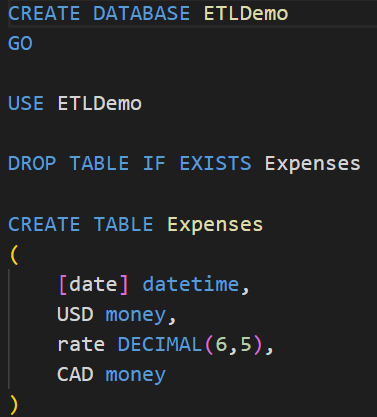
1. intialize database connection



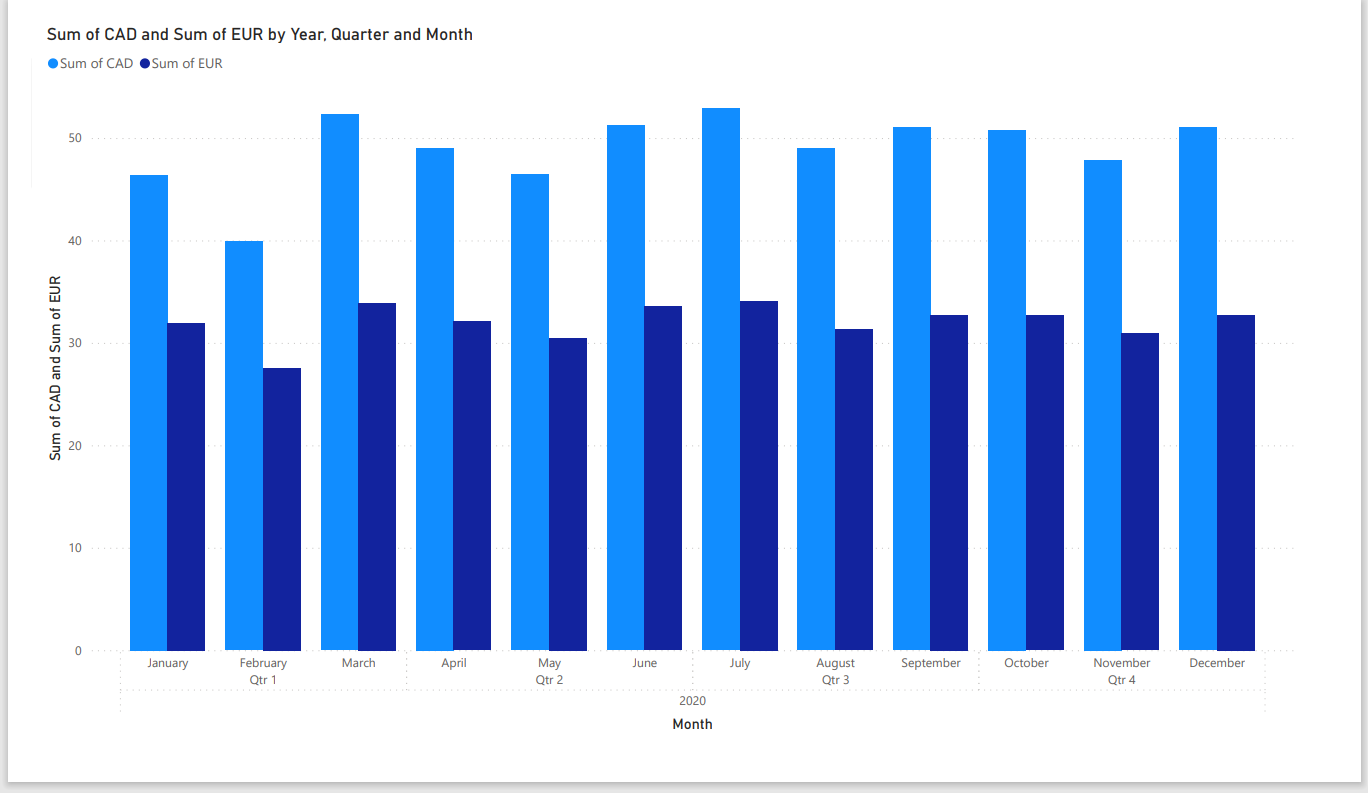
1. # populate Expenses database table

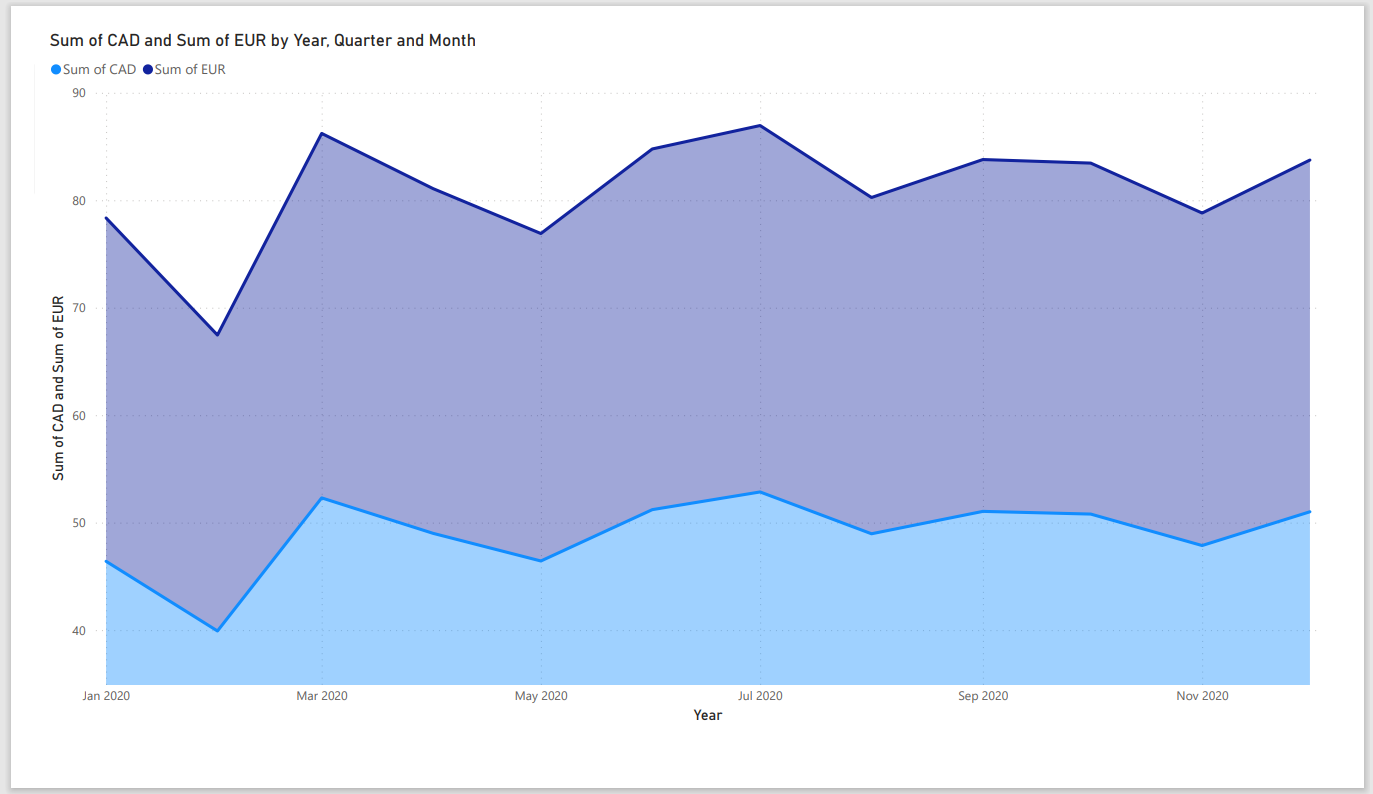


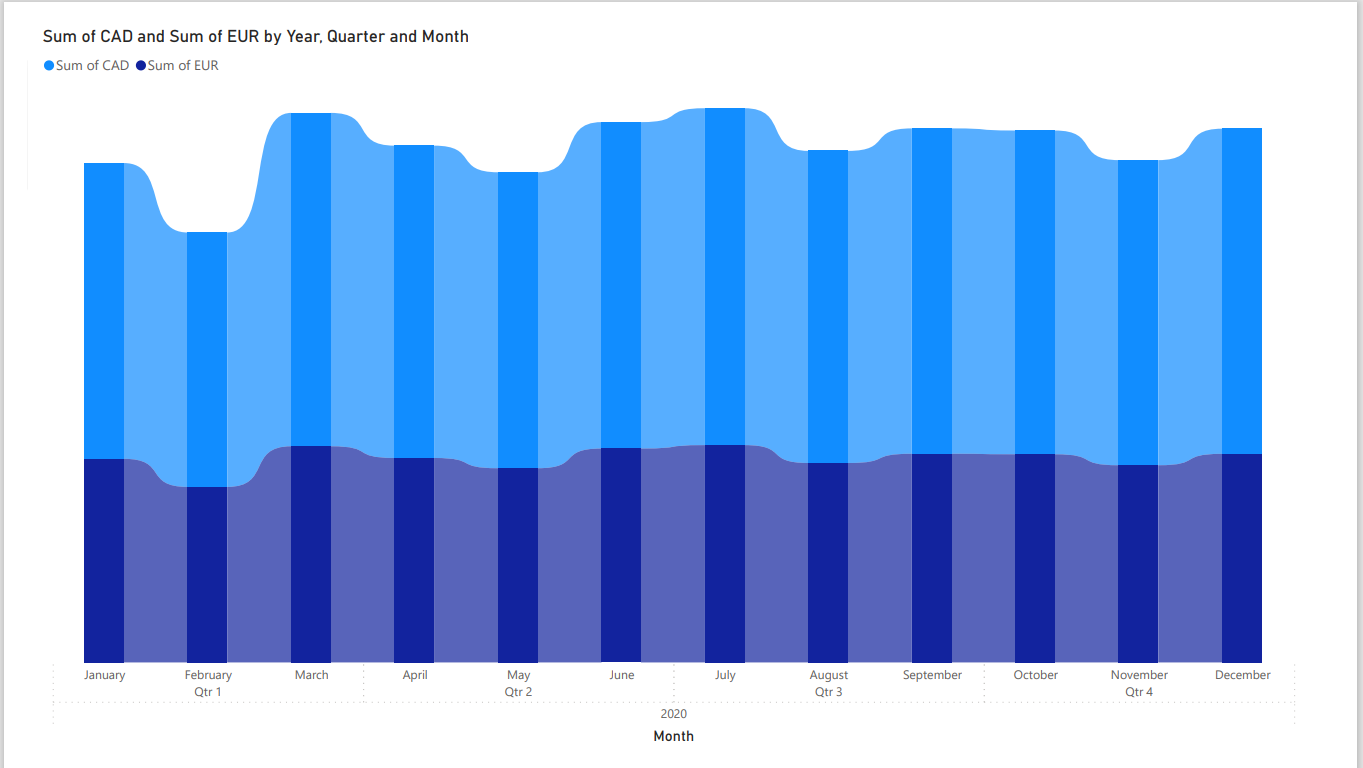
**SQL Server Queries for Creating Schema and Table:**



**Power BI Visualization:**







Sources:

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

<https://xtract.io/blog/why-are-etl-tools-necessary-for-financial-data-analysis-and-reporting/>

<https://www.matillion.com/resources/blog/the-importance-of-etl-tools-in-data-warehousing>

<https://en.wikipedia.org/wiki/Microsoft_Power_BI>

<https://en.wikipedia.org/wiki/Microsoft_SQL_Server>

<https://en.wikipedia.org/wiki/Visual_Studio_Code>