JSP	Jeff Pinegar		jeffpinegar1@gmail.com
	Assignment:	Project 2: ETL Proposal	717-982-0516
	Due Date:	Dec. 23, 2022	

Project 2: ETL Proposal

Source Files:

MUDDatabase

- Description: The data set contains configuration information for every IPC manufactured since 2008. The configuration information includes the Serial number, base article number, MAC addresses, CPU type, memory, screen size, etc. The total number of IPC in the data set is 103,099.
- Internal Link: http://nts-us-miappp01/
- File Name: 133156280355595299.csv
- Data as of Dec. 15, 2022
- Primary Key: The primary key for this file will be the serial number. The serial number uniquely identifies the IPC in this table.

2022 Europe IPC sales

- Description: This data set includes all IPC and HMI sales from Germany in 2022, up through Dec. 12, 2022. This data consists of the serial number, destination country, date of purchase, price, and currency, plus additional fields.
- File Source: Export from an SAP report ran at corporate headquarters in Germany. This SAP report is my only access to this type of information.
- File Name: "IPC shipped from Germany by serial number 2022 (until 5.12.2022).csv"
- Data as of Dec. 12, 2022
- Primary Key = Index
- Foreign Key = The Serial number link this table to the MUDDatabase
- The data set contains +15,000 transactions.

CMAT Codes to descriptions

- Description: This Excel sheet lists all the CMAT codes with an English-like description. There are approximately 450 unique CMAT codes. For example, a D47 is a 15.6" 1920x1080 LCD PTOUCH (FHD)
- File Name: CMAT Decoder.xlsx
- The primary key in this file will be the CMAT Code.

Project Tasks

Extract

1. Load resource files into Jupiter Notebooks
2. Prepare a Postgres Database

Transform

The sales data set and MUDDdatabase each need significant transformation. The transformation will include but is not limited to the following:

- Removal of defective rows (errors and omissions)
- Removal of irrelevant row (date outside of the time frame of interest)
- Changing the data type for several columns. For example, strings ➔ dates, strings ➔ floats
- Currency conversions
- Eliminating unnecessary columns
- Renaming, reordering reindexing columns

The CMAT Codes data set is much cleaner and smaller. Therefore, I will be cleaned by hand in Excel to remove some blank rows and transform some Excel tables to plain text before saving as a CSV file.

Load

Once the transformations are complete, the datasets will be loaded into a Postgres relational database. A relational database was selected because of the fixed structure of this data and the clear linkages via keys between the datasets.

Use Cases

Once ETL is complete, this data could be used to perform the following exploratory analysis of European sales patterns.

- Analysis of product mix by country
- Analysis of supply chain by country (mfg date to sell date)
- Analysis of product mix overall (which options are most popular)
- With the addition of cost analysis of margin by country