**Name**

ETL Project

**Key Deliverables**

This project focused on extract, transformation and loading skills. It requires the integration of a number of skills developed throughout the first half of the class. The project requirements included the following:

1. Extract data from two or more sources.
2. Transform the data so that it can be loaded into a relational or non-relational database. The transformation to include cleaning, joining, filtering, aggregating and other actions.
3. Place the final tables into a production database
4. Submit a technical report with the above information and steps required reproduce the ETL process

**Project Description**

For this project I chose to revisit a challenge from my prior work in an anesthesia practice. The practice struggled with, and spent a substantial amount of money, ensuring that clinicians met their licensing and health system facility credentialling requirements.

As background, the anesthesia needs are met by three clinician types: Anesthesiologist, Certified Registered Nurse Anesthetist (CRNA) and Anesthesiologist Assistants (AA). The CRNA licensing requirements are unique in that the CRNA must maintain an advanced practice nursing (APN) license AND a registered nurse (RN) license. To further complicate matters, the State of Colorado Department of Regulatory Affairs (DORA) will grant an APN license to a CRNA whose underlying RN license is maintained in states other than Colorado. The DORA (State of Colorado) does not verify the status of the underlying RN except at the time of license issue or renewal. As a result, it is possible that the underlying RN license will expire while the APN license show current. It is important for the health system and the subservient anesthesia practice to ensure that both licenses are always current.

As a prospective solution, I took the Professional Licensing File that the State of Colorado publishes daily and utilized it to create compliance and exception management reports that can be updated daily through a web API if desired.

**Data Sources & Extraction**

Two core data sources were utilized and staged for use as three project csv files = Anesthesia Licensing Data, Registered Nurse Licensing Data and a Group Practice Employee Roster. The Jupyter Notebook includes the scripts utilized to narrow the initial data set’s down to the desired anesthesia compliance related information. This notebook is titled extract.ipynb.

1. The core licensing file was extracted as a CSV file from Colorado.gov on a one-time basis for the purposes of this project. The same data is available through an API if desired.
2. The core licensing file was filter from approximately 1.2 million individuals to approximately 193K RN’s and 2.5K non-physician anesthesia clinicians. Plumbers, electricians, etc. were set aside. The core licensing file was not uploaded to github due to its size and github file limitations. Two sub files are included with this project repository.
3. A fictious employee roster was built based on publicly available information. The process can be replicated using an actual employee roster when appropriate.
4. Additional files were extracted with wage and additional license information but set aside as not required to achieve the projects compliance scope.

**Transform**

The three projects files are: license\_anesthesia, license\_nurse and employee-roster. These three files were imported into Jupyter Notebook and ‘worked-over’ using Python Pandas. Ultimately three clean files for loading into a PGAdmin/PostgreSQL database. These files were: load\_anesethesiaLicense, load\_nurselicense and load\_employeeRoster. The scripts are provided in the transform.ipynb notebook.

The transformation process required the following three key steps with each file:

1. A ‘matchName’ was created so that the files could be merged and joined in later activities
2. Excess information (columns) was removed to simplify and focus the effort
3. A sort and filtering effort were deployed to pinpoint the needed compliance and out of compliance clinicians
4. CSV files were exported to the project repository for importing into PGAdmin

**Load**

The three “load\_” files were imported into PGAdmin as three database tables. An ERD diagram is provided with this project repository. Key PGAdmin screens shots were provided in the project “pitchbook” to demonstrate completion.

**And …**

Four key reports were created to enable improved and timely management of the non-physician compliance process. Two reports are compliance reports establishing that timely due diligence was performed to ensure clinical license compliance. In addition, two exception reports were created to help the prospective compliance administrator focus on the key exceptions and prospective out of state license risks associated with a “compact” license process.