ETL Project

Data Related Job Market

Data Analytics Boot Camp

School of Continuing Studies

University of Toronto

Group Members:

May Ang, Kelvin Deng, Thao Hoang, Yijing Su

Date: December 19th, 2020

**TABLE OF CONTENT**

[1. INTRODUCTION 1](#_Toc59267562)

[2. METHODS 2](#_Toc59267563)

[2.1 Extraction 2](#_Toc59267564)

[2.2 Transformation 3](#_Toc59267565)

[2.3 Loading 5](#_Toc59267566)

[3. CONCLUSION 7](#_Toc59267567)

[4. FUTURE WORK 8](#_Toc59267568)

[REFERENCE 9](#_Toc59267569)

# INTRODUCTION

With the rapid advancement of technologies, data related jobs have become highly demanded in recent years. The computer technology boomed in the past 20 years has made computer usage much easier and cheaper for everyone, and this has led to the enormous increase in data generations and the demand for data organizations. With the easier access to data by users and the importance of data in decision making nowadays, data analysis has become significantly crucial in all industries. According to the LinkedIn Workforce Report, in the USA, the demand for data related professions has multiplied six times compared to five years ago, and this growth will continue for the next five years. Therefore, people with data organization and analysis skills are highly in demand in the work force and data related topics are current innovation focuses by many IT professionals.

With this background, this project focuses on the current job market for data related jobs for four different countries, USA, Canada, Australia, and Singapore. Using the ETL process, the project will extract all data related job postings from Indeed for these four countries, transform them and divide them into to Data Analyst, Data Scientist, Data Engineer and Machine Learning these four opportunities, and lastly, load them into a SQL database. Furthermore, this project also used ETL process to obtain the world’s university ranking data and mental health survey for further analysis on their corresponding correlations.

# METHODS

ETL refers to Data Extraction, Transformation, and Loading. It is the general procedure for data integration, the first step of data analysis. ETL process essentially is the strategy for extracting data from different sources and recombining them into new datasets for ease of analysis. Extraction is the first step in ELT. It is the step to extract data from different sources. Transformation is the step to transform data, including cleaning, summarization, selection, joining, filtering, and aggregating, to let all data have similar format that can be easily analyzed. Lastly, loading is to load transformed data into one database or one place for data visualization and data analysis after for problem solving.

## 2.1 Extraction

In this project, data were extracted for the current job market for data related jobs in USA, Canada, Australia, and Singapore. Specifically, data analysts, data scientists, data engineers, and machine learning are the four main job functionalities that this project focused on. Initially, several csv files related to this project were obtained from websites such as Kaggle but they still lacked some key information. Due to the inconsistency and noisiness of the csv data obtained from online, it is decided to use web scrapping to obtain all data related jobs from indeed. The web scrapping is performed using four separated Jupyter notebook programs stored in the Extraction Subfiles folder to represent four specific countries, and for each country, jobs that belong to the four job functionalities were scrapped. The web scrapping process is done by using the BeautifulSoup module, and the data is scrapped for every 10 pages among all the job listing pages. Firstly, API is requested using the requests class. Then BeautifulSoup class is used to retrieve the information needed that is stored in the page’s HTML. Furthermore, a for loop is used to go iterate through every div tags in the HTML to obtain the information needed for each individual job posting, job title, job id, company name and job location. Eventually, the information was stored into their corresponding lists created prior to the Data Frame construction. Same processes are performed for each job functionalities for all four countries. Besides, the world’s top 1000 universities’ ranking data from 2012 to 2015, and the Mental Health survey obtaining the world’s university ranking data and mental health survey were extracted for future analysis to find more possible correlations. These data were extracted into csv format from online sources and loaded into the Clean Data folder for the data transformation step.

The challenge imposed during the extraction step was the banning from Indeed during web scrapping. Indeed recognized the ongoing data extraction performed and banned the API request for jobs in Australia due to the block from the robot detection program, as well as the excessive amount of requests. To solve this problem, two approaches were conducted as shown in webscrape\_random.ipynb file. One approach is adding a random delay from 3 to 7 seconds in the loop. Another approach is the use of VPN. A VPN named Tunnel Bear is used in this project to change the user agent. This VPN installed a fake agent to provide a random user info and IP address for every 10 records to bypass the banning from Indeed. Another challenge imposed is the HTTM proxy being too slow. The solution used to solve this issue is to use library to get random time (Maybe you can expand the challenge we had for HTTM proxy part).

## 2.2 Transformation

Transformation is an important step of ETL. It uses all types of methods such as data cleaning to change all the data into the same format for ease of analysis afterwards. How data is transformed depends on the requirements of the project as well as what kind of analysis are needed for these data. For this project specifically, the goal was to transform the data from multiple sources, transform them into the same format and load them into a SQL database to establish relationship with each other. Since this ETL process involves the process of converting a NoSQL data to a SQL database, then data normalization must be a part of this data transformation. An ERD, entity relationship diagram, would be created after data transformation to visually display the relationship for each data that would be loaded into the SQL database.

In this project, for the four job market data extracted from four countries, a data cleaning was done on each to remove jobs that have more than one unique job title index and create country index column for data normalization and ERD creation later. Furthermore, these four transformed job market data were combined into one Data Frame with formatting on column names as well as splitting company location information into cities and states. As for the university ranking data and mental health data, a data cleaning was performed to identify and remove the NaN rows, as well as data formatting conducted to ensure the consistency of each column. For the data normalization of these two data, the country column was replaced with country index for data normalization purposes and the creation of ERD.

Besides the data transformation on the extracted data, several intermediate tables were created to establish a complete relationship between each data. The location summary table was created by splitting of company location information from the combined job market data, and the duplicated city and states were dropped to ensure the uniqueness of each city and state. The country table and the job title were created so that country index and job title index can be assigned to the job market data. Furthermore, to complete the location data, the latitude and longitude for each city was obtained by using an API from the open cage data website, and this complete location data is stored in the location coordinates table for the creation of ERD later.

The challenge encountered in data transformation and normalization is to identify which of the four main job functionalities does each job belong to. The best approach of this problem is by using machine learning. However, this is out of our scope and ability for this project. Therefore, the solution used in this project is to search each job functionality on indeed, web scrape those data, and based on that to assign corresponding job title index to each job. The issue with this approach is that this involves enormous amounts of duplicates, jobs appeared under one job functionality also appeared on another. To solve this, jobs that have more than one unique job title index would be dropped, and the last entry of that job would be kept.

May, could you please talk about the issue regarding average the university data at here, the issue you asked Laurel yesterday? I didn’t see a table with average university scores by country and I don’t know why we need to do that. Also I see you and Kelvin were discussing the API part of our project, the city\_location\_API.ipyng for quite a while, I’m not sure if you encountered any problems. If so, maybe you can address it here as well.

## 2.3 Loading

Loading is the final step of the ETL process, and it is to load all transformed data into a SQL database. After the data transformation and normalization, a complete relationship between each transformed table and intermediate table was established. SQLAlchemy in conjunction with Pandas was used to import data to the PostgresSQL database. PostgresSQL was chosen is due because of its easiness to allow users to control table structures, datatypes and relationships between each table within a certain database. To have a direct and clear visualization of the relationship between each table, a ERD was created. Based on the relationship displayed in the ERD, relational tables were created and imported directly using SQLAlchemy on Python. To accomplish this task, all csv files obtained from transformation step were loaded and transformed into data frames using Pandas. An engine was created to establish a route to allow access and query of database by using Python. With this engine created, classes could be instantiated by using the declarative base class to create relational tables in the SQL database using Metadata. Sessions are created to make connections between the program to the database which enables pandas.to\_sql() to commit dataframes into the database. Postgres SQL username and password were stored into the file named config.py. This file was not pushed to the Github to hide the confidential information from the public users.

Even though SQLAlchemy is a powerful module to interact with a database, users need to have a good understanding of what each step does and how they need to be performed to make it work. The following figure helped us to visualize the concept better.

(Not sure if we should include this as our challenge. If we didn’t encounter any challenges during loading, maybe we can just address we didn’t have any challenge? Or did we have any problems when we try to assign PK and FK and how to construct the ERD? maybe we can address that)

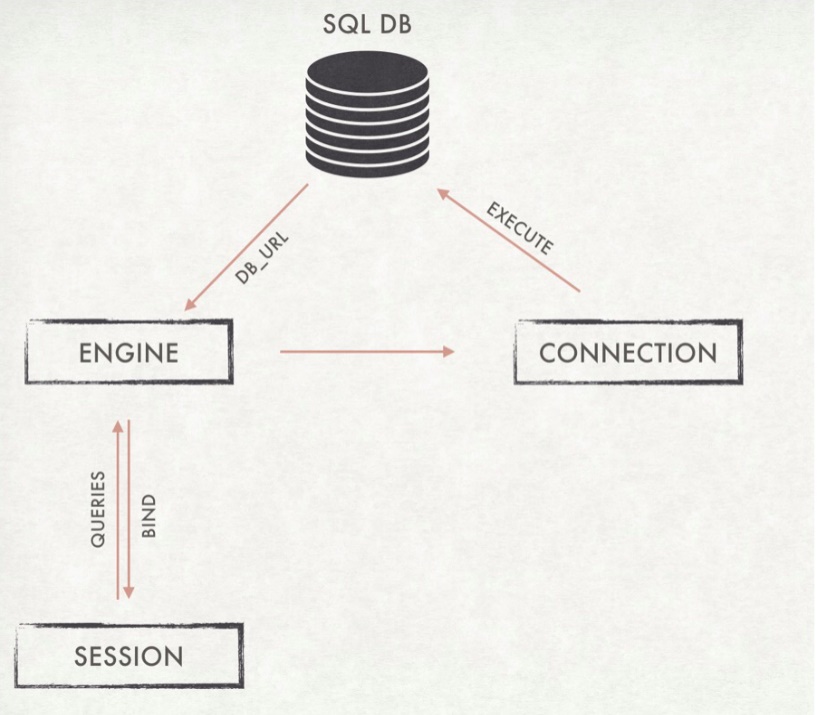


Figure 1. Database Connection

# CONCLUSION

With this background, this project focused on the current job market for data related jobs for four different countries such as USA, Canada, Australia, and Singapore. Using the ETL process, the project will extract all data related job postings from Indeed for these four countries, transform them and divide them into Data Analyst, Data Scientist, Data Engineer and Machine Learning, and lastly, load them into an SQL database. Furthermore, this project also used the ETL process to obtain the world’s university ranking data and mental health survey for further analysis on their corresponding correlations.

This project focused on the ETL process of the current job market data for data analysists, data scientists, data engineers and machine learning, these four functionalities, from four different countries, USA, Canada, Australia, and Singapore. The job market data of each country is extracted by web scrapping the Indeed website. The world’s university ranking data and the mental health survey were also extracted from online as csv files for the building of ERD and data analysis afterwards. Data transformation and normalization were performed on all extracted data and several intermediate tables were created for the construction of a complete relationship between each data, which would make data querying and analysis easier. Lastly, the transformed data and intermediate tables were loaded into PostgresSQL database based on their relational order, and a ERD was constructed to visualize the direct relationship between each table. By performing this ETL process to convert all topic related data extracted from different sources into one SQL database with a direct relationship for each table within, the data querying and analysis for the next step of data organization and analysis would be much easier (would be eased)

Lessons learned (not sure what should be include in here)

# FUTURE WORK

Based on the time constraint and the lack of knowledge in higher level of data analysis, there are some limitations and improvements that can be considered to improve the accuracy and completeness of the data.

During the process of data transformation, categorize each job into the four main functionalities was a big challenge. some data were dropped due to data normalization, while some job data would possibly been categorized into wrong groups. By using NLP or machine learning to include more data for analysis and categorizing, the accuracy and the validation of the data would be improved

Besides, due to time constraints and difficulties encountered, all transformed data were stored in the local host PostgresSQL database only, instead of cloud-based PostgresSQL-hosted database such as ElephantSQL. Cloud database has the benefits of storing data on the cloud to allow the ease of access by users through internet. It can also allow access through a web interface or vendor’s API, is easily scalable, and most importantly provide a security of the data through backups on remote servers.

(If there is more future work you can think of, please supplement at here)

# REFERENCE

<https://www.purdueglobal.edu/blog/information-technology/rise-of-data-analyst/>

<https://www.morningfuture.com/en/article/2018/02/21/data-analyst-data-scientist-big-data-work/235/>

<https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move>

<https://www.sas.com/en_ca/insights/data-management/what-is-etl.html>

(Here, I’m not sure what format of reference we need, and if any of you need to add more reference, please supplement at here)