**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**

[INTRODUCTION 1](#_Toc59146124)

[METHODS 2](#_Toc59146125)

[Extraction 2](#_Toc59146126)

[Transformation 4](#_Toc59146127)

[Loading 5](#_Toc59146128)

[CONCLUSION 6](#_Toc59146129)

[FUTURE WORK 6](#_Toc59146130)

[REFERENCE 7](#_Toc59146131)

# INTRODUCTION

With the rapid advancement in technologies, data related jobs have become the most in demand ones in recent years. The computer technology boom 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 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 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.

## 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).

## 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.

## Loading

xxxxxxx

# CONCLUSION

xxxxxxxxxxx

# FUTURE WORK

xxxxxx

# 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>