**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 become the most in demand job in recent years. The computer technology boom in the past 20 years has made computer use much easier and cheaper with everyone, and this leads to the enormous increase in data generation and the need for data organization. With the ease of access to data by every user and the importance of data in decision making nowadays, data analysis become significantly crucial in all industries. According to the LinkedIn Workforce Report, in the USA, 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 work force and data related topics are current innovation focuses by many IT professionals.

With this reason, this project will focus 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 uses 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 process. 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 (in) problem solving.

## Extraction

In this project, data is 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 is focusing on. Initially, several csv files related to this project were obtained from websites such as Kaggle but still lacking some key data. 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 notebooks stored in the Extraction Subfiles folder to respresent 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 class, 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 is 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 are extracted for future analysis to find more possible correlations. These data are 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 is 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 is 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

xxxxxx

## 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>