Initial Plan: Enhanced Company Data Compilation for Marketing Analysis

Author Name: Md Md Farid

Author Student Number: 21071339

Supervisor: Federico Liberatore

Project Description

This project entails the utilization of retrieved data from sources across the World Wide Web. Extracted data is then intended to be compiled and deployed as a valuable database for companies. With the database, companies can perform more effective marketing-related analysis to benefit themselves and gain an advantage in their respective industries.

A structured database or any other available source of information might have gaps in data that can render information application of to be ineffective, if not fruitless. To exemplify this, one database may contain instances of information such as the total number of employees and annual revenue per company but lacks the customer lifetime value, a key metric that measures a company's chances of retaining its customers (Jain 2023). Because of this missing piece of information, a complete marketing analysis and strategy cannot be produced.

A solution to this problem primarily requires data extraction, data cleansing and data integration. The main method of data extraction to be used is web scraping as it appears to be the most feasible solution to retrieve unstructured data from the World Wide Web. Data cleansing and data integration will allow several sources of data to be combined and assembled into a complete block of information.

Overall, the realm of marketing requires data to be varied and valuable to strategize a competitive edge in boosting the company's commercial objectives, and this can be augmented by developing a platform that provides a database intended to solve the outlined problem of information gaps. Although the solution sounds straightforward, the amount of time required to design, develop, and debug the program makes the project a challenge worth attempting to overcome. Several technologies need to be learned, including web scraping tools and various programming languages, to build the platform.

Aims and Objectives

The overall aim of this project is to develop a program that represents structured data as a database, with the following objectives to be used as building blocks for the program:

- Extract a sufficient variety of company data from the World Wide Web.
- Cleanse and compile data from extracted sources.
- Store collated data into a database.

For the objectives above, I will learn technologies like Python, a programming language that can be used to perform web scraping, and MongoDB, which offers well-structured databases.

In addition to the database, I also intend to enhance my knowledge of the marketing field. As this database is going to assist the marketeers in their strategies, I need to be aware of fundamental marketing techniques and principles. I will seek marketing knowledge to:

- Understand how marketeers think on an organizational basis.
- Relate the database to a typical marketing strategy.
- Keep the program in line with fair marketing principles.

Feasibility

Regarding ethical approval for this project, there does not appear to be a need for any. To avoid a long waiting line for approval and potential rejection, I intend to scrap or extract data from websites that permit web scraping or contain datasets public users are allowed to download. As companies reserve key data only for their own organizational usage, I will not seek data from any company's official website (e.g. The total revenue made from Toyota's main automotive website) as I also want to ensure that this project does not trespass the boundaries of restricted company data.

I want to use a wide range of data sources for this project. However, I am aware that there are limitations. Some websites may not have a variety of relevant data for marketing analysis, while others may lack a good quantity of instances (rows of information) to give my project a surmountable scale. Moreover, restricted terms and conditions hinder the opportunities of retrieving potentially mighty and beneficial data. I have formed some contingency plans, which are described in the 'Work Plan' section below.

No special computer hardware or software are required as well. With the use of generic web scraping tools and programming languages to collate data, I will use my own laptop to perform the technical operations and any available desktop computer from either the Abacws labs or Cardiff University libraries if necessary.

However, the biggest risk would be program test failures and the limited time to debug, which may occur during the evaluation stage near the final report submission due date. Two solutions may follow through depending on the available time left. Modify a feature within a day to get it to function.

- Delete a program upgrade which was intended to enhance the features of the database and bring the program to a previous state which works.

Another big risk would be permanent damage to my laptop which can lead to options as well:

- Send laptop for service.
- Replace laptop.

In both cases, the files and datasets may be lost, which is why I have prepared my physical backup drives to save the files, along with the use of OneDrive, an online storage system. OneDrive files can be retrieved on the Internet from any device with the use of my student account, hence, storing my project on OneDrive will be a big and flexible benefit.

Work Plan

Project Tracker

Start date:	29/01/2024
End date:	09/05/2024

Spring Term Week(s)	Start Date	End Date	Activity
1	29/01/2024	05/02/2024	Start with Initial Plan
2	05/02/2024	12/02/2024	Research & Gathering
3	12/02/2024	19/02/2024	Design & Analysis
4+5	19/02/2024	04/03/2024	Training & Implementation
6	04/03/2024	11/03/2024	Evaluation
7+8	11/03/2024	25/03/2024	Debugging & Re-evaluation
Easter break 1	25/03/2024	01/04/2024	Advanced Research
Easter break 2 + 3	01/04/2024	15/04/2024	Advanced Implementation
9	15/04/2024	22/04/2024	Evaluation 2
10	22/04/2024	29/04/2024	Debugging & Re-evaluation 2
11	29/04/2024	09/05/2024	Final Compilation

Meetings with Dr. Liberatore are flexible and can fit any day of each week in which both of us are available for 30 minutes. Starting off with week 1, I intend to complete this initial plan, which is the first milestone and deliverable, followed by commencing research on the marketing field to understand the context of the field's analysis and required technologies to develop the database.

One key technology to learn is a programming language to perform web scraping like Python, though I may change my language of choice during training and implementation. I will be comfortable switching languages because I would have fundamentally understood the common programming elements and syntax for web scraping, so typing different lines of code is workable.

Along with a programming language, I consider MongoDB to be a great tool for database structuring, so learning it is part of my technological gathering task too. Whilst seeking sources of data, I must pay attention to the terms and conditions. I will read them, ensure I am aware of the permissibility of scraping or extracting data from each site, and conduct data practices according to the legal requirements.

The data sources themselves have limitations, but enhancing extracted data is possible. One way is to append dummy data to the database via programming. Structure of dummy data, however, is going to be based on the schema of the originally available data. Another way is to modify the data model contingent on the progressive technical objectives of the project. It can be done by updating the schema of the database.

After completing 2 milestones of attaining both marketing and technological knowledge, a 'Design & Analysis' week to model the data and fit a database's syntactic criteria follows. UML (Unified Modelling Language) can be considered for that week's activity.

Each of the first 3 tasks can be finished within a week but applying technological knowledge (programming languages and web scraping tools) will require 2 weeks. I need at least a week to be comfortable using the technologies I learned and another week to assemble the first prototype database. Anticipating bugs and test failures, I estimate 3 weeks to be enough to mitigate them. 3 weeks comprise 1 week to evaluate the prototype program's capabilities and another 2 weeks to fix any program faults and document any extra functions. I aim to have a working prototype before the Easter Break.

3 weeks of Easter Break offer an opportunity for me to do advanced research to enhance my program to a higher caliber for marketing. I want to implement more technical features beyond the standard requirements, if possible. The last 3 weeks before the deadline are used for final evaluation of the program and touch ups.

Regarding the final report, I want to document my efforts each week, adding more content to the report after every week passes. It leaves report completion to be minimal and makes room for any final modifications before submission on 9 May 2024. I want to submit the report 1 day before the actual deadline shown on PATS, which is 10 May 2024, explaining why the project end date is stated to be 9 May 2024.

Milestones:

- Initial Plan, to be submitted on 5 February 2024.
- Gain fundamental knowledge of marketing by 12 February 2024, including, but not limited to:
    - Principles of marketing.
    - Marketing strategies.
    - Market segmentation.
- Gain fundamental knowledge of technologies required to develop the program by 12 February 2024, including, but not limited to:
    - Web scraping tools to extract data from online sources.
    - MongoDB, to design a structured database to store information.
    - Programming language, like Python, which can be used to perform web scraping.
- Prototype program before Easter Break (25 March 2024).
- Final program ready for deployment by 9 May 2024.
- Final report to be completed by 9 May 2024, containing elements including, but not limited to:
    - An introduction to the project.
    - Database research and development documentation of my efforts through the Spring term.
    - Evaluation of test case results from program testing and admitting any disbenefits that will occur.

Deliverables:

- Initial plan, which is to be submitted on 5 February 2024.
- Final report, which is to be submitted on 9 May 2024, one day before the deadline.
- A program consisting of a structured database with useful information for marketing analysis.

The program will be evaluated in 2 stages (basic and advanced), to give time and monitor all implemented features carefully.

References:

Clark, D. 2023. Business enterprises in the UK – Statistics & Facts | Statista. Available at: https://www.statista.com/topics/6848/businesses-in-the-uk/#topicOverview [Accessed: 30 January 2024].

Gaid, A. 2023. Marketing Principles: The Four Key Concepts To Understand. Available at: https://www.oberlo.com/blog/marketing-principles [Accessed: 1 February 2024].

Gillis, A. S. and Robinson, S. 2023. What are the 5 V's of Big Data? | Definition TechTarget. Available at: https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data [Accessed: 30 January 2024].

Ibm. 2024. What is Data Modeling? | IBM. Available at: https://www.ibm.com/topics/data-modeling [Accessed: 5 February 2024].

Jain, A. 2023. What is Customer Lifetime Value (CLV), and How to Calculate It. Available at: https://www.gartner.com/en/digital-markets/insights/what-is-customer-lifetime-value [Accessed: 30 January 2024].

Tarver, E. 2024. Market Segmentation: Definition, Example, Types, Benefits. Available at: https://www.investopedia.com/terms/m/marketsegmentation.asp [Accessed: 1 February 2024].

Vipinyadav15799. 2023. Data Science Process – GeeksforGeeks. Available at: https://www.geeksforgeeks.org/data-science-process/ [Accessed: 30 January 2024].