REAL-TIME CARBON NEUTRALITY MANAGEMENT AND OPTIMIZATION USING NATURAL LANGUAGE PROCESSING

Project ID: 2022-175

Project Proposal Report

Sathees Paskaran IT19052748

B.Sc. Special (Honors) Degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

February 2022

REAL-TIME CARBON NEUTRALITY MANAGEMENT AND OPTIMIZATION USING NATURAL LANGUAGE PROCESSING

Project ID: 2022-175

Project Proposal Report

Sathees Paskaran IT19052748

Supervisor Name: Ms. Anjali Gamage

B.Sc. Special (Honors) Degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

February 2022

Declaration

I declare that this is my own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or institute of higher learning, and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Project Title: Real-Time Carbon Neutrality Management and Optimization Using Natural Language Processing

Project ID: 2022-175

Name of the author

Student ID	**Student Name**	**Signature**
IT19052748	Sathees P.	

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor:………………………… Date:……………………

(Ms. Anjali Gamage)

Abstract

Carbon reporting is the process of calculating total emissions for a period and generating different types of reports to evaluate the overall carbon neutrality of organizations. A real-time carbon reporting system called Carbonis was proposed to provide timely emission status of organizations. For this purpose, emission records will be collected from employees of the organization as natural language inputs and necessary parts will be labeled. To calculate emission, a matching emission factor needs to be identified. However, finding emission factors using conventional methods would be cumbersome and can reduce the effectiveness of the real-time nature of the system. To overcome these issues an emission factor retrieval component was proposed as a part of Carbonis. This component will use term similarity and personalization with previous user history to rank emission factors for the emission technology, consumption unit (unit of measure), and year extracted by the emission collection component. This search component is expected to be tolerant to different variations of query terms as the inputs will be collected as natural language.

Keywords: Carbon Trading, Carbon Reporting, Emission Factors, Natural Language Processing, Information Retrieval.

Table of Contents

Declaration iii

Abstract iv

Table of Contents v

List of Figures vii

List of Tables viii

List of Abbreviation ix

List of Appendices x

List of Equations xi

1. INTRODUCTION 1

1.1. Background and Literature survey 1

1.2. Research Gap 6

1.3. Research Problem 8

2. OBJECTIVES 9

2.1. Main Objectives 9

2.2. Specific Objectives 9

3. METHODOLOGY 11

3.1. Complete System Architecture 11

3.2. Component Architecture 12

3.3. Technologies 16

3.4. Tasks And Sub-tasks 17

3.5. Material Support 18

3.6. Data Collection 19

3.7. Project Time Frame 19

3.8. Anticipated Conclusion 19

4. PROJECT REQUIREMENTS 20

4.1. Functional Requirements 20

4.2. User Requirements 20

4.3. System Requirements 20

4.4. Non-functional Requirements 21

5. WORK BREAKDOWN STRUCTURE AND GANTT CHART 22

6. COMMERCIALIZATION 24

7. BUDGET AND BUDGET JUSTIFICATION 25

8. REFERENCES 27

9. APPENDICES 31

9.1. Appendix – A: Sample of DEFRA Emission Factor Document 31

List of Figures

Figure 1.1: Survey summary about providing emission records 4

Figure 1.2: Survey summary regarding a good emission data collection approach 4

Figure 1.3: Survey summary on domain knowledge about carbon accounting 4

Figure 1.4: Survey summary on preferred lookup method 5

Figure 1.5: Survey summary on preferred search approach 5

Figure 1.6: Survey summary on personalized search systems 6

Figure 3.1: Simplified complete system architecture 11

Figure 3.2: Component architecture diagram of the emission factor search system 13

Figure 5.1: Work breakdown structure (WBS) 22

Figure 5.2: Project Gantt chart 23

Figure 7.1: Survey summary on a budget suggestion for a search feature 25

Figure 9.1: Screenshot of sample DEFRA 2021 emission factor dataset 31

List of Tables

Table 1.1: Previous research and products comparison 6
Table 3.1: Critical comparison of information retrieval approaches 15
Table 3.2: Technologies to be used in the component implementation 16
Table 4.1: User stories 20
Table 7.1: Budget for the component and shared features 25

List of Abbreviation

Abbreviation	Description
AWS	Amazon Web Services
B2B	Business to Business
BA	Business Analyst
BI	Business Intelligence
CDP	Carbon Disclosure Project
CO_2e	CO_2 equivalent
CRIS	Climate Registry Information System
DEFRA	Department for Environment, Food and Rural Affairs
EPA	Environmental Protection Agency
EU	European Union
GB	Gigabyte
GHG	Greenhouse Gas
IR	Information Retrieval
kWh	Kilo Watt-hour
MB	Megabyte
NDC	Nationally Determined Contributions
NGA	National Greenhouse Accounts
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
RAM	Random Access Memory
SEO	Search Engine Optimization
UI	User Interface
UK	United Kingdom
UN	United Nations
USA	United States of America
WBS	Work Breakdown Structure

List of Appendices

Appendix	Description	Page
Appendix - A	Sample of DEFRA Emission Factor Document	31

List of Equations

Equation 1.1: Greenhouse gas emission calculation formula 2

1. INTRODUCTION

1.1. Background and Literature survey

Concerns about the impacts of greenhouse gas (GHG) emissions on climate change have been growing significantly over the recent years [1]. Due to these concerns, many governments, and supra-national bodies such as United Nations (UN), and European Union (EU) have formed policies and agreements to encourage reducing GHG emissions [2]. Significant of these policies were Kyoto Protocol in 1997 and Paris Agreement in 2016 [1], [2]. Stricter measures were proposed by the Paris agreement which targets to limit global warming increase below 2 degrees Celsius, and preferably to 1.5 degrees Celsius [1]–[5]. It is required by the countries to provide their planned actions to contribute to the Paris agreement aim every five years, this is known as Nationally Determined Contributions (NDCs) [1]. As one of the parties of the Paris Agreement, Sri Lanka expects to achieve carbon neutrality by the year 2060 and has taken many steps to achieve it, such as the reduction of emissions related to electricity production by using renewable energy [6]. It is also noteworthy that GHG emissions from industries are a major contributor to global warming [7]. To accommodate NDCs, governments have implemented many national climate regulations methodologies such as carbon tax, carbon trade-off, carbon cap-and-trade, mandatory carbon reporting, and carbon emission disclosure (also known as Carbon Disclosure Project (CDP)) by the firms [8], [9]. These measures eventually lead to the birth of the carbon trading paradigm.

In carbon trading, firms purchase or obtain a certain amount of carbon credits. Carbon credits allow firms to emit a certain amount of GHG emissions. However, it is beneficial for the firms to stay within their available credit limit. Firms balance their total GHG emissions for a period with their available carbon credits [10]. This process is referred to as carbon accounting or carbon reporting. By staying within their credit limits firms achieve carbon neutrality [11]. These carbon trading implementations are expected to provide significant benefits such as encouraging consumers, companies, and managers to find sustainable alternatives even if they are expensive [12].

Moreover, it also promotes technological innovation and competitiveness in implementing sustainable alternatives [9].

Regardless of the carbon trading schemes used, the measurement of GHG emissions by the firm must be performed. GHG emission occurs when there is an emission activity (e.g., usage of 1kWh of electricity) is carried out in the firm [13]. The amount of GHG emissions is usually measured in Metric tons. During an emission activity, varieties of GHGs such as carbon dioxide, methane, nitrous oxide, hydrofluorocarbons, etc. are released and these GHGs have different global warming potentials [2], [14]. Global warming potential provides estimates on how much each GHG contribute to global warming compared to carbon dioxide [14]. However, conducting carbon reporting for all these gases would make the carbon accounting process complicated. Therefore, a standard measure known as **CO_2 equivalent** (CO_2e) is used to measure GHG emission for different emission sources [1], [13]. These CO_2e values for different emission sources are estimated and published in different document formats by different environmental entities (also known as **emission factors**) [13]. Firms usually adopt one of these emission factors as their standard for carbon reporting depending on their reporting jurisdiction. Some of those standards are published by,

- Department for Environment, Food and Rural Affairs (DEFRA) [15] – UK
- Climate Registry Information System (CRIS) [16] – USA and Canada
- Environmental Protection Agency (EPA) [17] – USA
- National Greenhouse Accounts (NGA) [18] – Australia

GHG emission for an emission activity can be calculated using equation 1.1 [13], [19]. For example, if we assume the emission factor for an average car is 0.1500 $kgCO_2e/km$ and we have traveled 4 km using this car. Calculated GHG emission = 4 km × 0.1500 kgCO2e/km = 0.6 kgCO2e.

$$GHG\ emission = consumption\ \times specific\ emission\ factor \qquad 1.1$$

Moreover, emission activities can be classified into three scopes such as Scope 1, Scope 2, and Scope 3. Scope 1 includes emissions from direct sources which are owned by the firm such as vehicles. Scope 2 includes emissions from indirectly purchased sources such as electricity. Scope 3 includes emissions from indirect sources which are not owned by the firm such as waste management [4], [13], [14].

Current approaches for managing carbon emissions within firms are done either manually by a business analyst (BA) or using a commercial emission calculator. During this literature survey, it was observed that there is an inadequacy in the research related to implementing corporate-level carbon emissions management. Additionally, there are some commercially available tools to calculate and manage emissions [20]–[22]. However, there are some gaps and issues observed with current carbon reporting implementations [1], [5]. During the discussion with the industry expert, it has become evident that most of the current systems expect emission data to be collected and provided by the BA of the firm. However, there can be many data collection issues in this approach. Moreover, emission data collection is usually done at the end of the reporting period, and this causes a delay in reporting. Due to this delay, there might be chances of unanticipated scenarios like emitting more than the desired targets to occur.

To solve the current issues of the carbon reporting system a real-time carbon neutrality management system was proposed. In this proposed system emission activity data is expected to be collected from the employees (refer to Figure 1.1) of the firm using a natural language input (refer to Figure 1.2) in real-time. The extraction component will be able to annotate and differentiate parts such as emission technology, date, consumption value, consumption unit, etc. of the emission activities provided by the employees. To proceed with emission calculation relevant emission factors for the emission technology (e.g., vehicle, generator, diesel, etc.) should be identified. However, for the accurate emission calculation, emission factors for each of these activities should be selected carefully [13]. It is time-consuming to refer to these emission factor files to find suitable emission factors and requires significant domain knowledge about carbon reporting (refer to Figure 1.3 and Figure 1.4). Therefore, there is a necessity for an emission factor search system to speed up the calculation process of the proposed system and make it suitable for novice users.

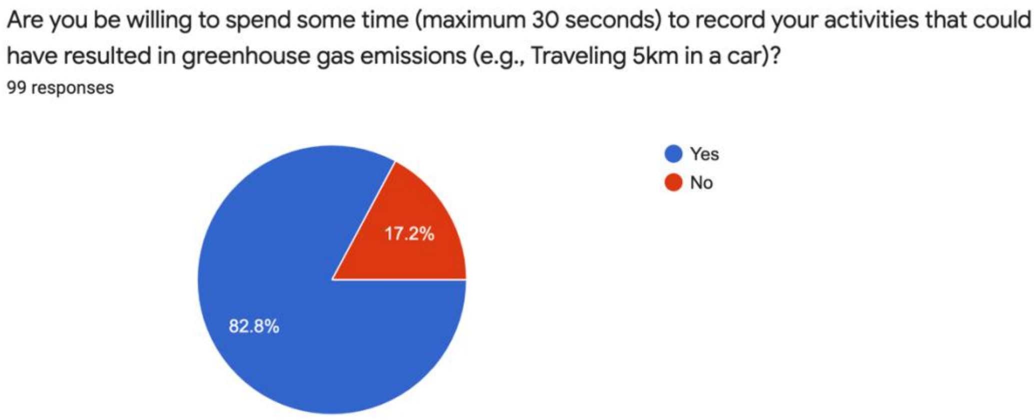

Figure 1.1: Survey summary about providing emission records

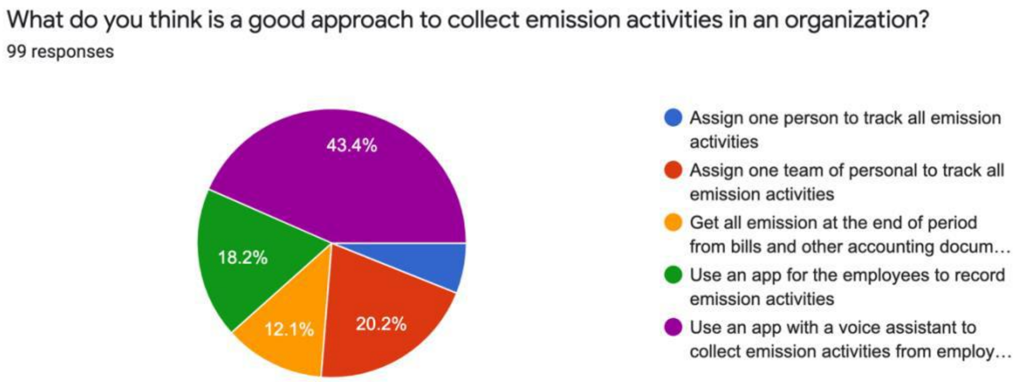

Figure 1.2: Survey summary regarding a good emission data collection approach

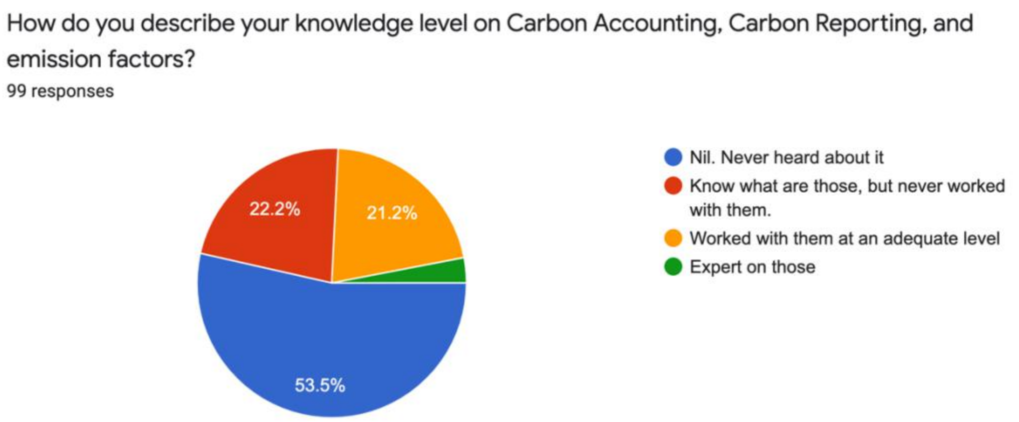

Figure 1.3: Survey summary on domain knowledge about carbon accounting

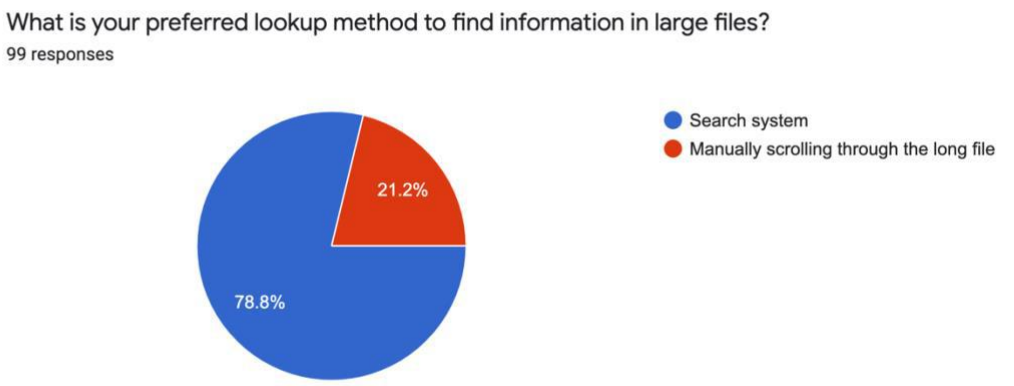

Figure 1.4: Survey summary on preferred lookup method

An efficient emission factor search component to retrieve relevant emission factors for the given emission technology (e.g., vehicle, generator, diesel, etc.), consumption unit, and year is considered suitable for the proposed system. Since these are gathered from natural language inputs, there will be no limit on the variation of the terms that users could use to describe emission technologies and consumption units. In this case, it is preferable to use an ad-hoc searching (refer to Figure 1.5) approach. The search system is expected to be tolerant to variation terms like synonyms. Moreover, employees usually tend to work with the same emission technologies regularly in their work environments. Therefore, having a personalization approach could be beneficial as it could give more weightage to the recently used emission factors in the results ranking (refer to Figure 1.6).

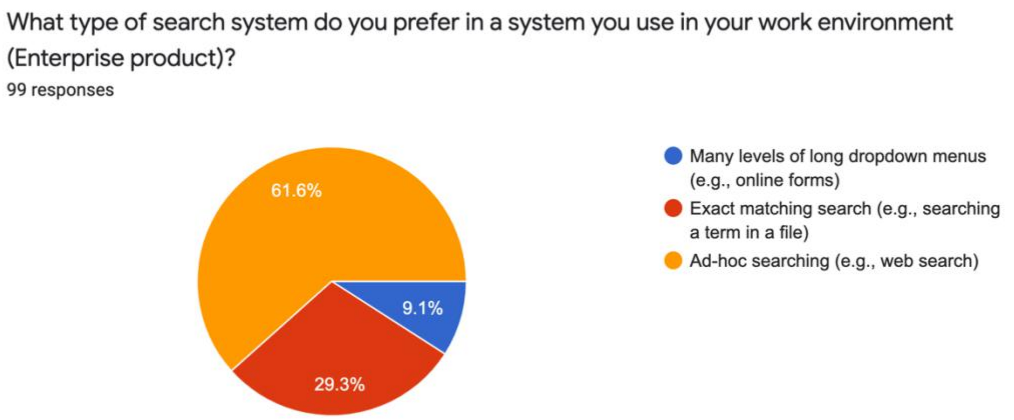

Figure 1.5: Survey summary on preferred search approach

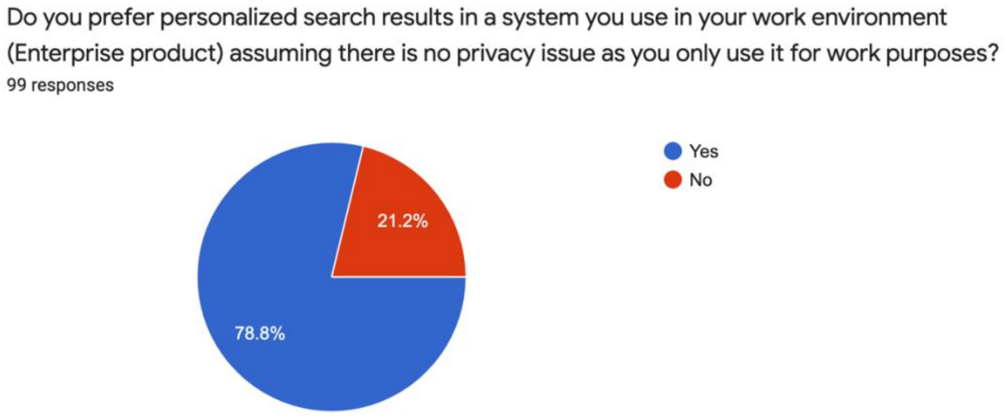

Figure 1.6: Survey summary on personalized search systems

1.2. Research Gap

It is observed that there is a necessity for more research in carbon reporting implementation at a corporate level and most of the available research concentrates on the effectiveness of various reporting schemes. In this Research A [23], the real-time carbon accounting method is only implemented for scope 2 emissions and does not focus on corporate-level emission management. This research does not feature an emission search system as well. However, there are many commercial systems implemented to solve the same issue such as Product A [20], Product B [21], and Product C [22]. All these products require getting emission data from a single person or group of personals of the firm using the user interface (UI) or file uploads in a specified file format. Moreover, these systems do not provide true real-time or timely emission status due to the delay in data collection. These products also do not feature an emission factor search system. All these commercial systems feature some sort of emission calculation.

Table 1.1: Previous research and products comparison

Research or Product	Emission calculation	Data collection from employees	Emission factor search system	Ad-hoc emission factor searching	Emission factor ranking using term similarity and personalization
Research A	✓	X	X	X	X

Product A	✓	X	X	X	X
Product B	✓	X	X	X	X
Product C	✓	X	X	X	X
Proposed System (Carbonis)	✓	✓	✓	✓	✓

From research and products comparison, the novelty of the proposed system (Carbonis) is justifiable as there were no previous works within this domain that feature emission data collection from employees, emission factor search system, and emission optimization. In the domain of carbon reporting technologies, it was observed there was no previous work on emission factor search systems and it proves the novelty of the component separately. It is noteworthy, that there can be many benefits of using an ad-hoc emission factor search system with personalization in any carbon reporting tool such as,

- Timesaving
  Searching can be faster than traversing through files or using dropdown menus. As in this case, emission factors must be found within thousands of emission factors in different documents using this search approach is expected to increase in performance of carbon reporting.
- Less fatigue
  Scrolling through files can cause fatigue to employees which could lead to a bad user experience in emission reporting. As searching can be less fatigue, employees will be more willing to report their emissions.
- Tolerant to various representations of the search queries
  Employees can give search queries for their information needs in different representations. Ad-hoc search approach can be considered tolerant to this issue.

The emission search system is relevant for the proposed system (Carbonis) because emission activity and consumption units can be in different representations. There is no other suitable retrieval approach that is practical in this scenario. The proposed system consists of many beneficial features compared to previous research and currently available products in the market. It is unique in the way of data collection from the employees using natural language input and it includes an emission factor searching component that is tolerant of various representations of terms and can rank results using term similarity and personalization.

1.3. Research Problem

Emission factors are published in various document formats and styles by the environmental entities and can be downloaded from their official websites. There are different editions of these documents every year. Emissions should be calculated using the relevant year's factor, even if there is the latest edition is available [24]. Usually, these emission technologies are represented in these documents using technical terms different from what a novice employee might call it. Moreover, these emission factor records can cause information overload for the users and might reduce the user experience of the whole reporting system. There is a necessity for an efficient retrieval system for emission factors to solve these issues.

Additionally, for practical reasons, an efficient emission factor searching component can be considered essential for the realization and success of the proposed system (Carbonis). According to the background survey, it is observed that most of the users prefer ad-hoc search systems which are tolerant to different query representations and have a ranking system.

Due to the above reasons, the research problem that this proposed component will be trying to solve is to implement an ad-hoc emission factor search system that is tolerant to various query representations and ranks results based on term similarities and personalization. The inputs of the component can include emission technology (e.g., vehicle, generator, electricity, etc.), consumption unit, activity year, and reporting standard. For personalization, it can keep track of user search histories and emission record histories. The output of the search system would be ranked emission factors results for the query and other inputs. This component approach is expected to be scalable to work with most of the emission factor standards.

2. OBJECTIVES

2.1. Main Objectives

The main objective of the whole proposed system would be to implement a cross-platform mobile application platform for organizations to manage and optimize their carbon emissions in real-time. Additionally, the objective of this proposal's component would be to implement a search feature that would provide ranked emission factor results using term closeness (or term similarity), and personalization weightage for the extracted query parameters. This objective is expected to be delivered within the course of this research project timeline shown as a Gantt chart in Figure 5.2.

2.2. Specific Objectives

For the successful accomplishment of this proposal component's objective of implementing a search system that would provide ranked emission factor results using term closeness (or term similarity), and personalization weightage for the extracted query parameters; sub-objectives need to be attained during the project timeline as follows,

1. Collect and process emission standard documents
   Emission standard documents of different types and formats will be processed to clean data, and necessary transformation will be carried out.
2. Create a common emission factor representation
   Design a common factor representation which will be a document unit of the search system and store processed emission factor data in this format.
3. Implement an emission factor search feature to provide ranked results of emission factors
   Design searching module with suitable algorithms which will use term closeness (or term similarity), term frequency, and personalization to rank emission factor results.
4. Calculate emission of the emission activities

Implement an emission calculation module that will calculate emission for the provided emission activity and selected emission factor. These records will be stored for further analysis and reporting.

3. METHODOLOGY

3.1. Complete System Architecture

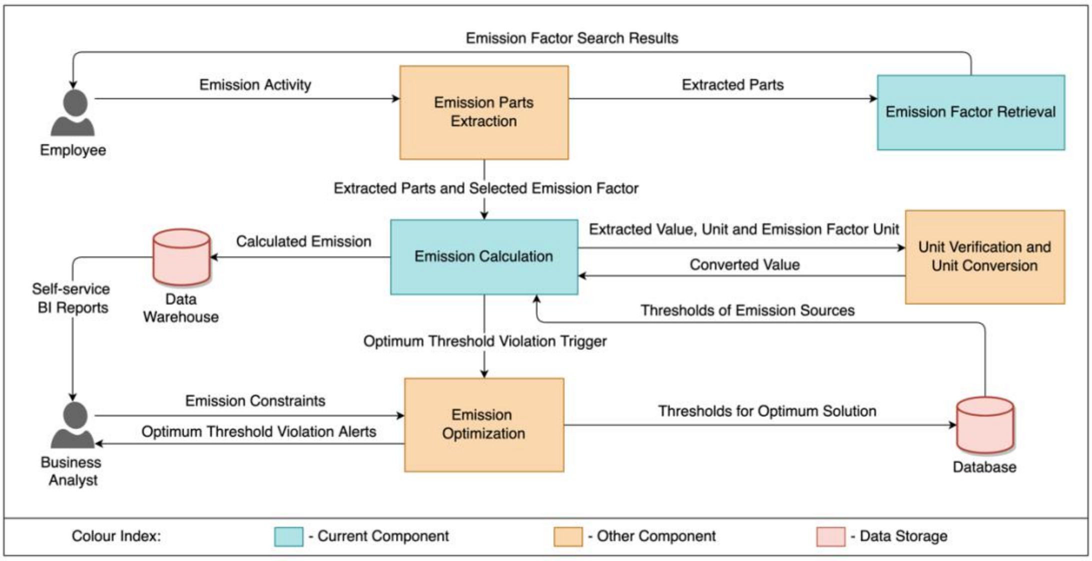

Figure 3.1: Simplified complete system architecture

As shown in Figure 3.1, in the proposed system (Carbonis) there are mainly four components as follows,

1. Emission records collection and parts extraction

   Emission records will be collected from employees as a natural language input. From these input data, necessary parts of emission activity such as emission technology, consumption's value, consumption's unit, date, etc. will be extracted. These converted parts will be provided to the emission factor retrieval component and the returned results will be shown to the user for selection. Once the employee chooses an emission factor for the activity, extracted parts and chosen emission factor will be provided to the emission calculation part.

2. Emission factor retrieval and emission calculation (focusing component of this proposal)

   For the extracted parts and other organizational parameters, the emission factor retrieval component will find and rank matching emission factors, and these results will be provided to the employee for selection. After the emission factor confirmation, extracted parts and chosen emission factor will be received for emission calculation. Consumption's value, consumption's unit, and emission factor's unit will be sent to unit verification and unit conversion component to ensure the units are matching. Once the unit conversion is completed, emission will be calculated and stored in a data warehouse (in a dimensional model) for self-service business intelligence (BI) access by BAs. On a new emission addition, optimal solution threshold violation will be checked and if there is any violation, alert triggers will be sent to the emission optimization component.

3. Unit verification and unit conversion
   The similarity of consumption's unit and emission factor's unit provided by the emission calculation module will be verified. If there is a mismatch in these units, consumption's values will be converted to match the emission factor's unit. This converted consumption's value will be provided for emission calculation.
4. Emission optimization module
   For the available emission credits and constraints on different emission sources of the firm, an optimized solution will be obtained and thresholds on different emission sources for this optimal solution will be stored. For a threshold violation trigger, warning alerts will be sent to responsible BAs.

3.2. Component Architecture

This component must complete the research objectives of implementing an emission factor search system with term similarity and personalization; and implementing emission calculation. The major focus of this research component will be given to the implementation of the emission factor search part. The emission calculation part is straightforward, and values will be calculated according to Equation 1.1. For the emission calculation, extracted consumption value from user input and selected emission factor will be used. Finally, calculated emissions will be stored in a data warehouse to be accessed by business analysts using self-service BI platforms.

The emission factor search part must retrieve ranked matching emission factors for the extracted emission technology, consumption unit (also called the unit of measure), and year parts. Additionally, the organization's preferred emission standard can be used for factor filtration. Previous employee history on emission factor selection will be used for the personalization feature.

Each emission factor value can be considered as a document unit for the search feature. The corpus of this document unit can be built by combining emission technology columns, column text, and units (refer to the sample emission factor dataset given in Appendix – A).

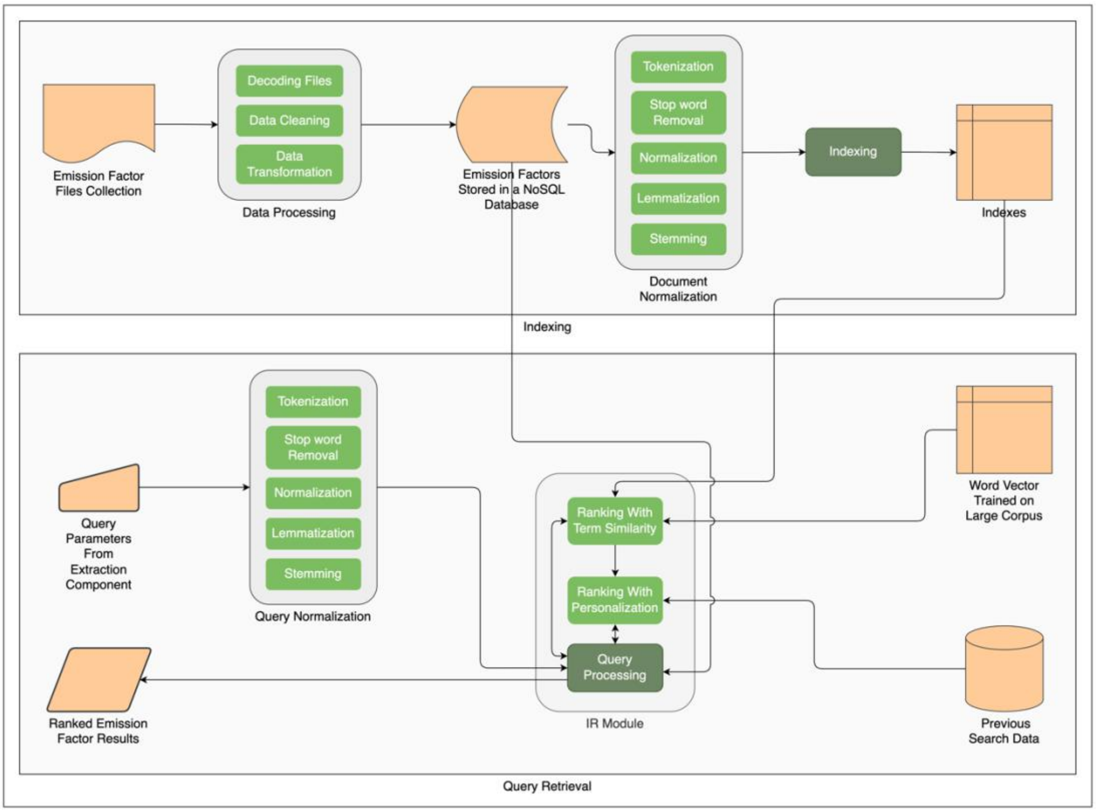

Figure 3.2: Component architecture diagram of the emission factor search system

As shown in Figure 3.2, this emission factor retrieval part consists of two main parts such as indexing and query retrieval. The indexing part is infrequent compared to query retrieval and only needs to be done when there is an addition in the emission factor files. The query retrieval part will take place for each query request.

In the indexing part, collected yearly emission factor files of the supported emission standards will be collected. Since different standards have different release formats such as spreadsheets, pdf, word, etc., and different documentation styles, these emission factor records need to be transformed into a standard representation that can be used to store emission factors of various emission factor standards. These transformations are custom for a certain emission standard and might not be suitable for another standard. These transformation approaches are expected to work with future releases of the same emission standard if there are no major changes in published file formats. The data preprocessing steps on these files include data decoding to read different file formats, data cleaning to remove unnecessary data and correct inconsistency issues, and data transformation to a standard representation. After this data processing, transformed emission documents units will be stored in a NoSQL database. Each document unit will contain fields such as corpus, emission factor value (CO_2e), year, standard, and scope.

Indexes should be created for the terms in these documents. This is done using an indexing algorithm. Depending on the requirements a necessary set of document normalization procedures such as tokenization, stop word removal, normalization, lemmatization, stemming, etc. can be applied before indexing. Tokenization splits phrases into separate words. In the stop word removal, common words (a, an, the, etc.) will be removed. Normalization changes terms into the same format e.g., U.S.A and USA are of the same term. Lemmatization variation forms to same base form e.g., am, is, are, belongs to be. Stemming reduces terms to their roots e.g., automate, automatic, automation to automat [25]. Created indexes will have terms and their locations within the documents. Indexes will be positional indexes with term frequency. These indexes can be created separately for each emission factor standard in a year-wise manner.

In the query retrieval part, matching emission factor documents will be retrieved in a ranked order for the given query parameters given by the parts extraction component. Query input parameters may include emission technology, consumption unit, and year. The same normalization techniques applied for document normalization before indexing should be applied to the query parameters such as emission technology and consumption unit before passing to query processing.

In the query processing part, there are two ranking parts. The first ranking will rank based on term similarity. For this part, a matching index of emission standard and year, and a word vector trained on a large corpus will be used. Word vectors are obtained from word embedding which uses algorithms such as word2vec and GloVe and trained on large corpus such as Wikipedia to identify the closeness of the terms. By using this word vector, even the query receives a non-indexed term, it could retrieve somewhat closer results. To identify closeness, measures such as the cosine similarity score can be used. There are many techniques available for information retrieval (as compared in Table 3.1), however, the usage of word embedding has been proving to achieve good results to retrieval with similarity [26]–[28].

Table 3.1: Critical comparison of information retrieval approaches

Technologies	**Ad-hoc Searching**	**Ranking with Term Similarity**	**Ranking with Personalization**	**Development Time**	**Startup Friendly (Cost, Need for Previous Data)**
Boolean model	X	X	X	Manageable	✓
Extended Boolean model	X	X	X	Manageable	✓
Vector space model	X	X	X	Manageable	✓
Probabilistic model	X	X	X	High	✓
Vector space classification	✓	X	X	High	X
Machine learning models	✓	X	X	High	X

Word embeddings	✓	✓	X	Manageable	✓

Even though word embedding ranking provides ranking with similarity, to achieve a personalization feature, another ranking with personalization will be applied to the results of the first ranking with similarity. For the personalization ranking part, the common approach is to use machine learning models, however, in this case, it would be unfeasible as there is no trainable data on hand. Therefore, an alternative personalization algorithm will be used to re-rank based on the user's previous search history. Here, more ranking weightage will be given to frequent and recently used emission factors. This ranking will use the user search history stored in the database. Finally ranked results will be provided to the employee to choose from. Once an employee chooses an emission factor, chosen factor value and other extracted parts will be passed for emission calculation.

3.3. Technologies

Technologies and tools expected to be used in this component implementation and their uses are discussed in Table 3.2.

Table 3.2: Technologies to be used in the component implementation

Technologies	Purposes
Languages	
Python	• Data processing • Algorithm development • Backend development
Libraries	
NumPy	• Used by Pandas library
Pandas	• Data processing
NLTK	• Natural language processing • IR system development

Jupyter	• Data processing • Algorithm development • Backend development
Cloud Services	
Cloud compute server (AWS EC2)	• IR system development
Serverless Backend service (AWS Lambda)	• Backend development
NoSQL database (AWS DocumentDB)	• Document unit storage
Relational database (AWS RDS) - PostgreSQL	• Application data storage • User history storage • Data warehouse implementation
Storage (AWS S3) - optional	• Factor files storage
Tools	
Visual Studio Code	• Data processing • Algorithm development
PyCharm	• Backend development
DBeaver or DataGrip	• Database management

3.4. Tasks And Sub-tasks

Tasks and subtasks necessary for this research component implementation, and other tasks allocated to Sathees P. (IT19052748) are as follows,

1. Tasks related to implementing emission factor search system
    a. Collection of emission factor data from the official factor websites [15]–[18].
    b. Perform necessary data processing steps on the acquired emission factor dataset files.
    c. Identify a common document structure and document unit to store different emission standards.

   d. Process the cleaned dataset into the common document structure and store these documents with associated fields such as year, standard name, scope, emission factor values, units, and corpus.
   e. Implement a suitable indexing algorithm and with the help of it create the index.
   f. Implement IR (Information Retrieval) modules with support for ranking based on term similarity and personalization.
   g. Using the previous user search history, implement a personalization feature.
2. Other implementation tasks
   a. Emission calculation for the given emission task and selected emission factor value.
   b. Store calculated in a data warehouse for further analysis.
3. Team tasks
   a. As a team leader manage the team and coordinate tasks within the team and with supervisors.

3.5. Material Support

Implementation of this component is completely software-based. Therefore, there is no hardware-related support is expected. However, it is essential to have cloud service support for this component implementation and deployment. More detail on the cloud services pricing can be found in section 7. Budget and Budget Justification.

Theoretical and technical knowledge about implementation technologies such as natural language processing (NLP), information retrieval (IR), and cloud computing (in this case Amazon Web Services (AWS)) is expected to be acquired from campus modules, journals, books, and third-party online courses. Guidance and insights about project tasks are expected from the project supervisors and support on domain knowledge is expected to be gained from the industry expert (external supervisor).

3.6. Data Collection

As mentioned in the introduction section necessary emission factor data will be collected from the official emission factors publication websites [15]–[18]. Word-vector trained on a large corpus of text can be obtained from public websites with permissible licenses.

3.7. Project Time Frame

Expected time frames for the component implementation have been discussed in detail in section 5. Work Breakdown Structure and Gantt Chart. It is noteworthy that major development part should be allocated for the development of emission search system.

3.8. Anticipated Conclusion

The proposed component is expected to retrieve emissions factors ranked by the term similarity, and personalization for the given emission technology, consumption unit, and year extracted from natural language input and organization's preferred emission standard. This implementation is anticipated to satisfy information retrieval metrics such as recall and precision along with satisfying non-functional requirements such as performance, speed, etc. as mentioned in section 4. Project Requirements.

4. PROJECT REQUIREMENTS

As this component proposed to implement a software solution, the requirements for this software component were collected and analyzed.

4.1. Functional Requirements

Functional requirements of this software component are listed in Table 4.1. It is observable that this component only tries to implement functional requirements involving the employee as the stakeholder.

Table 4.1: User stories

	As an…	**I want to…**	**So that…**
1.	Employee	retrieve ranked matches of the emission factors	I can save time when adding my emission data
2.	Employee	get personalized emission factor search results	I can get emission factors for my frequent activities faster
3.	Employee	calculate emission	I can save time

4.2. User Requirements

User requirements for this component that can be observed by the user (employee) include as follows,

1. The system would be able to give matching emission factors for the emission activity given as a natural language input.
2. The system would be able to give personalized emission factors results for emission activities given as a natural language input.
3. The system would be able to calculate and store emission records for the given emission activity.

4.3. System Requirements

Minimum system requirements of user end device for the client app to work is as follows,

1. Operating System: IOS or android

2. RAM: 1GB
3. Storage: 300MB free space
4. Internet connectivity

The minimum system requirement for a backend server is to have either Windows or Linux as the operating system, 8GB of RAM, and 30GB of storage. 200MB of storage size for databases is enough at the deployment phase. These system requirements are subject to increase while using and these resources should be scaled when that happens.

4.4. Non-functional Requirements

To ensure the usability of the system, following non-functional requirements needs to be achieved,

1. Speed

   For a search system response time is an essential requirement. This response time should be kept within an acceptable range (less than 20s).
2. Size

   As this search operation will happen frequently, resources needed for it should be minimized to be feasible.
3. Scalability

   This search component is expected to support future emission factor standards and other standards.
4. Ease of use

   As this component will be used by novice employees, this component should be usable by them without having to go through a training period.
5. Reliability

   Down-time in this component can result in the system being unusable. Therefore, higher uptime of this component is expected.

5. WORK BREAKDOWN STRUCTURE AND GANTT CHART

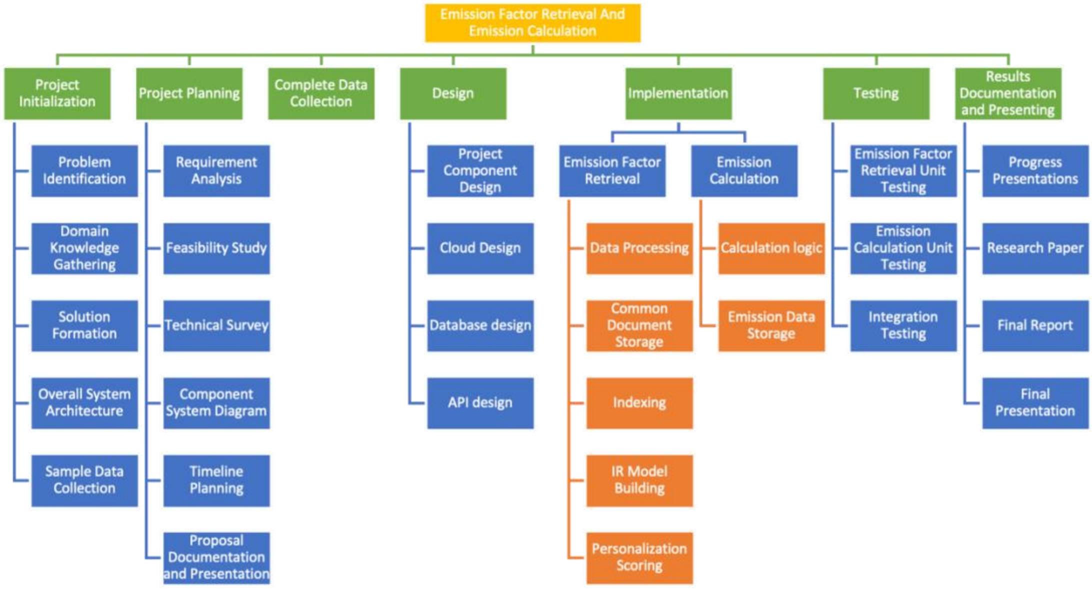

Figure 5.1: Work breakdown structure (WBS)

As illustrated in Figure 5.1, the work needed for the successful completion of the research component has been divided into seven major phases. In the project initialization phase, domain problem, solution, literature survey, and overall system architectures were created, and sample data were collected. As a result of the project planning phase, this proposal document was created, and a proposal presentation was conducted.

The next phases include complete data collection, coming up with necessary project designs, implementation of the component, testing the component, and finally documenting and presenting the results. More time will be allocated for the implementation part as it includes more workload. Within the implementation phase, more time will be given to developing emission factor retrieval.

Gantt chart of project timeline is depicted in Figure 5.2. In this Gantt chart also, it is observable that most time is given to developing and testing. The project timeline has been organized to allocate time for documentation and presentation tasks. According to this timeline, the project is expected to be completed at the end of November 2022.

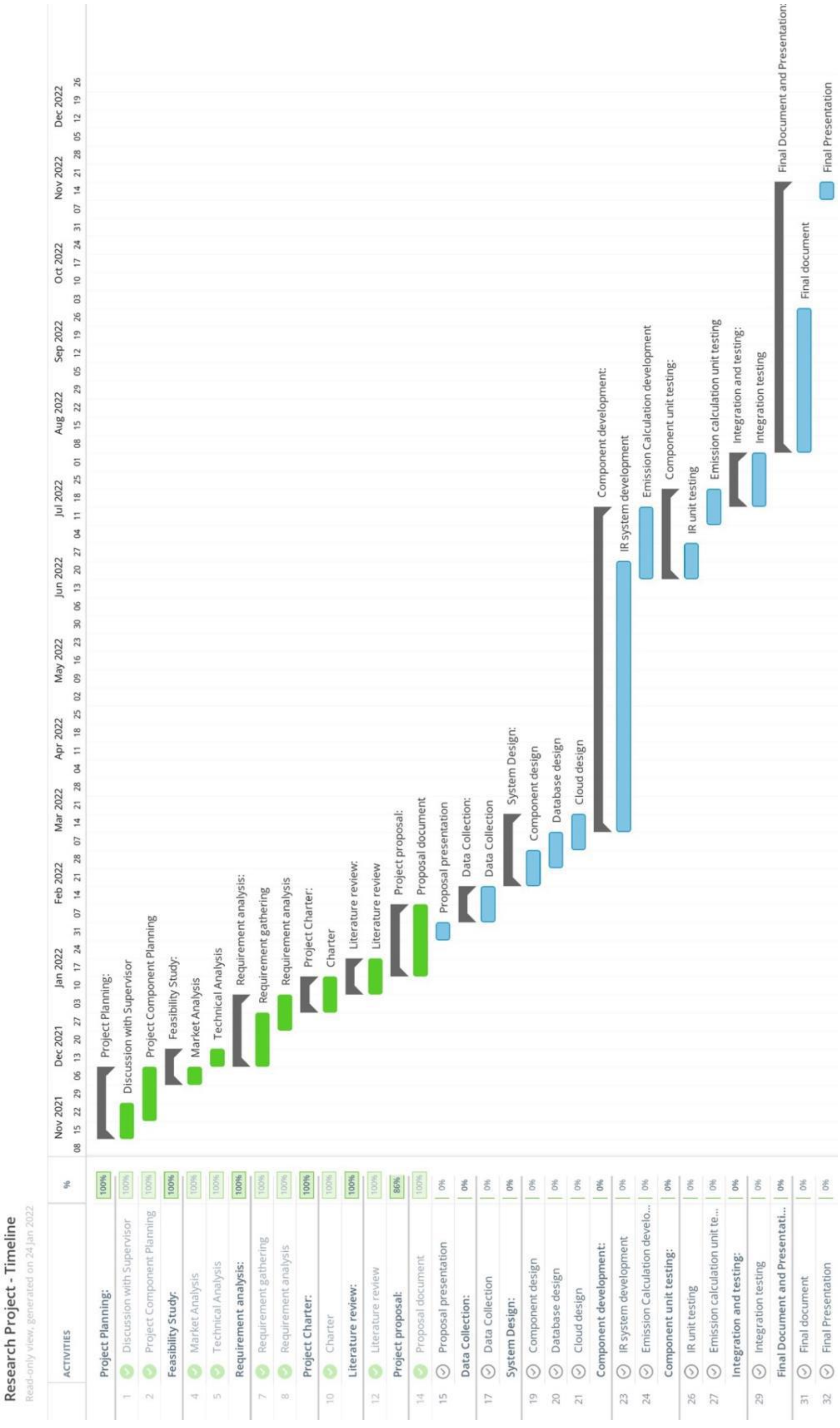

Figure 5.2: Project Gantt chart

6. COMMERCIALIZATION

As discussed in section 1.2. Research Gap, even though there are many commercial emission calculators available, most of them only focus on the emission calculation part without an emission factor search system or recommendation approach. Therefore, this component alone could be a commercially viable product or service that could be marketed to other systems as well. However, this component is a crucial part that makes the proposed system (Carbonis) practical.

Target clients of this component would be businesses and the suitable marketing paradigm would be B2B marketing. Some of the efficient marketing strategies that could be used in commercializing this component are,

1. Inbound marketing
   Creating online content about the domain (in this case carbon reporting) and introducing the product.
2. Social media marketing
   Making a social media presence for the system to reach potential clients.
3. Search engine optimization (SEO)
4. Search engine marketing (e.g., Google AdSense)
5. Industry events
6. Referral programs (affiliate programs)

Inbound marketing and social media marketing are expected to provide better results for this marketing paradigm and can be considered a good starting point in commercializing this component. Integration of this search feature would provide potential benefits such as faster emission recording and a better user experience to overall systems.

7. BUDGET AND BUDGET JUSTIFICATION

Results of the survey conducted (as shown in Figure 7.1) show that majority of respondents believe in paying more depending on the importance of the system. Table 7.1 describes the development cost and running cost of the search system. Some of the costs are shared among other components of this proposal (denoted as shared).

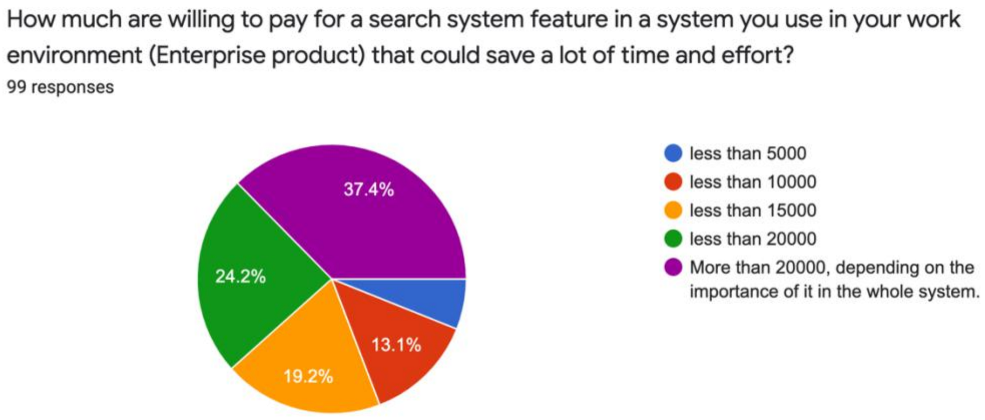

Figure 7.1: Survey summary on a budget suggestion for a search feature

Table 7.1: Budget for the component and shared features

Component	Cost (USD)	Cost (LKR)*
Development Expenses		
Cloud compute server (AWS EC2)	25.00	4950.00
Serverless Backend service (AWS Lambda) – free tier, shared	0.00	0.00
NoSQL document database (AWS DocumentDB) – free tier	0.00	0.00
Relational database (AWS RDS) – free tier, shared	0.00	0.00
Other cloud services - shared	30.00	5940.00
Total	55.00	10890.00
Operational Expenses		
Cloud compute server (AWS EC2)	30.00	5940.00
Serverless Backend service (AWS Lambda) –	20.00	3960.00

shared		
NoSQL document database (AWS DocumentDB)	25.00	4950.00
Relational database (AWS RDS) – shared	15.00	2970.00
Total	90.00	17820.00

*Used USD to LKR conversion rate of 198 Rs. on 3/2/2022.

8. REFERENCES

[1] M. Roelfsema *et al.*, "Taking stock of national climate policies to evaluate implementation of the Paris Agreement," *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–12, Apr. 2020, doi: 10.1038/s41467-020-15414-6.

[2] J. Bebbington, C. Larrinaga-González, C. Larrinaga-Gonza´lezgonza´, and G. Ãã, "Carbon Trading: Accounting and Reporting Issues," *http://dx.doi.org/10.1080/09638180802489162*, vol. 17, no. 4, pp. 697–717, 2008, doi: 10.1080/09638180802489162.

[3] F. Schreyer *et al.*, "Common but differentiated leadership: strategies and challenges for carbon neutrality by 2050 across industrialized economies," *Environmental Research Letters*, vol. 15, no. 11, p. 114016, Oct. 2020, doi: 10.1088/1748-9326/ABB852.

[4] K. Piper, J. Longhurst, M. Khare, and Z. P. Robinson, "Exploring corporate engagement with carbon management techniques," *Emerald Open Research 2021 3:9*, vol. 3, p. 9, May 2021, doi: 10.35241/emeraldopenres.14024.1.

[5] C. Quitmann, R. Sauerborn, and A. Herrmann, "Gaps in Reporting Greenhouse Gas Emissions by German Hospitals—A Systematic Grey Literature Review," *Sustainability 2021, Vol. 13, Page 1430*, vol. 13, no. 3, p. 1430, Jan. 2021, doi: 10.3390/SU13031430.

[6] "SRI LANKA UPDATED NATIONALLY DETERMINED CONTRIBUTIONS", 2021.

[7] H. Hashim *et al.*, "An Integrated Carbon Accounting and Mitigation Framework for Greening the Industry," *Energy Procedia*, vol. 75, pp. 2993–2998, Aug. 2015, doi: 10.1016/J.EGYPRO.2015.07.609.

[8] B. Sarkar, M. Omair, and S. B. Choi, "A Multi-Objective Optimization of Energy, Economic, and Carbon Emission in a Production Model under Sustainable Supply Chain Management," *Applied Sciences 2018, Vol. 8, Page 1744*, vol. 8, no. 10, p. 1744, Sep. 2018, doi: 10.3390/APP8101744.

[9] S. Tang and D. Demeritt, "Climate Change and Mandatory Carbon Reporting: Impacts on Business Process and Performance," *Business Strategy and the Environment*, vol. 27, no. 4, pp. 437–455, May 2018, doi: 10.1002/BSE.1985.

[10] C. L. Spash, "The Brave New World of Carbon Trading," 2009.

[11] F. Wang *et al.*, "Technologies and perspectives for achieving carbon neutrality," *The Innovation*, vol. 2, no. 4, p. 100180, Nov. 2021, doi: 10.1016/J.XINN.2021.100180.

[12] A. A. Rahman, Y. A. Aziz, and S. Sidek, "A Review on Drivers and Barriers towards Sustainable Supply Chain Practices Readiness of Malaysian Food-based Logistics Service Providers for Halal Practices View project Impact of food quality, food image and perceive value on tourist satisfaction and behavioral intentions. View project", doi: 10.7763/IJSSH.2015.V5.575.

[13] M. Brander, M. Gillenwater, and F. Ascui, "Creative accounting: A critical perspective on the market-based method for reporting purchased electricity (scope 2) emissions," *Energy Policy*, vol. 112, pp. 29–33, Jan. 2018, doi: 10.1016/J.ENPOL.2017.09.051.

[14] G. Shil, "Greenhouse Gas Emissions: A Case Study of Development of Data Collection Tool and Calculation of Emissions Outline GHG Emissions Inventory Development Process Reporting principles Organizational and operational boundaries Data collection and validation," 2007.

[15] “Government conversion factors for company reporting of greenhouse gas emissions - GOV.UK.” https://www.gov.uk/government/collections/government-conversion-factors-for-company-reporting (accessed Jan. 24, 2022).

[16] “CRIS Public Reports | The Climate Registry.” https://www.theclimateregistry.org/our-members/cris-public-reports/ (accessed Jan. 24, 2022).

[17] “GHG Emission Factors Hub | US EPA.” https://www.epa.gov/climateleadership/ghg-emission-factors-hub (accessed Jan. 24, 2022).

[18] “National Greenhouse Accounts Factors | Department of Industry, Science, Energy and Resources.” https://www.industry.gov.au/data-and-publications/national-greenhouse-accounts-factors (accessed Jan. 24, 2022).

[19] R. D. S. Jayathunga and M. H. N. K. T. Dulani, “A GUIDE for CARBON FOOTPRINT ASSESSMENT CLIMATE CHANGE SECRETARIAT MINISTRY OF MAHAWELI DEVELOPMENT AND ENVIRONMENT The Climate Change Secretariat Ministry of Mahaweli Development and Environment,” 2016.

[20] “CarbonView – Carbon reporting made easy.” https://carbon-view.com/ (accessed Jan. 24, 2022).

[21] “Simplified Carbon Reporting with Turbo Carbon™ | UL.” https://www.ul.com/services/digital-applications/simplified-co2-reporting (accessed Jan. 24, 2022).

[22] “Carbon Management & Reporting - Sphera.” https://sphera.com/carbon-management-reporting/ (accessed Jan. 24, 2022).

[23] B. Tranberg, O. Corradi, B. Lajoie, T. Gibon, I. Staffell, and G. B. Andresen, "Real-time carbon accounting method for the European electricity markets," *Energy Strategy Reviews*, vol. 26, p. 100367, Nov. 2019, doi: 10.1016/J.ESR.2019.100367.

[24] Defra, "Environmental Reporting Guidelines," 2019, Accessed: Jan. 23, 2022. [Online]. Available: www.nationalarchives.gov.uk/doc/open-government-licence/

[25] C. D. Manning, Prabhakar. Raghavan, and Hinrich. Schütze, "Introduction to information retrieval," p. 482, 2008.

[26] D. Hu *et al.*, "Recommending Similar Bug Reports: A Novel Approach Using Document Embedding Model," *Proceedings - Asia-Pacific Software Engineering Conference, APSEC*, vol. 2018-December, pp. 725–726, Jul. 2018, doi: 10.1109/APSEC.2018.00108.

[27] X. Yang, D. Lo, X. Xia, L. Bao, and J. Sun, "Combining Word Embedding with Information Retrieval to Recommend Similar Bug Reports," *Proceedings - International Symposium on Software Reliability Engineering, ISSRE*, pp. 127–137, Dec. 2016, doi: 10.1109/ISSRE.2016.33.

[28] D. Ganguly, D. Roy, M. Mitra, and G. J. F. Jones, "A word embedding based generalized language model for information retrieval," *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 795–798, Aug. 2015, doi: 10.1145/2766462.2767780.

9. APPENDICES

9.1. Appendix – A: Sample of DEFRA Emission Factor Document

UK Government GHG Conversion Factors for Company Reporting

Factors by Category

	Category							
Scope	Level 1	Level 2	Level 3	Level 4	Column Text	UOM	GHG	GHG Conversion Factor 2021
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Gross CV	kWh (Gross CV)	kg CO2e	0.22240
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Gross CV	kWh (Gross CV)	kg CH4	0.00017
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Gross CV	kWh (Gross CV)	kg CO2	0.22210
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Gross CV	kWh (Gross CV)	kg N2O	0.00013
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Net CV	kWh (Net CV)	kg CO2e	0.24106
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Net CV	kWh (Net CV)	kg CH4	0.00018
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Net CV	kWh (Net CV)	kg CO2	0.24074
Scope 1	Fuels	Gaseous fuels	Butane		Energy - Net CV	kWh (Net CV)	kg N2O	0.00014
Scope 1	Fuels	Gaseous fuels	Butane		Volume	litres	kg CO2e	1.74529
Scope 1	Fuels	Gaseous fuels	Butane		Volume	litres	kg CH4	0.00129
Scope 1	Fuels	Gaseous fuels	Butane		Volume	litres	kg CO2	1.74296
Scope 1	Fuels	Gaseous fuels	Butane		Volume	litres	kg N2O	0.00104
Scope 1	Fuels	Gaseous fuels	Butane		Tonnes	tonnes	kg CO2e	3033.32000
Scope 1	Fuels	Gaseous fuels	Butane		Tonnes	tonnes	kg CH4	2.25000
Scope 1	Fuels	Gaseous fuels	Butane		Tonnes	tonnes	kg CO2	3029.26000
Scope 1	Fuels	Gaseous fuels	Butane		Tonnes	tonnes	kg N2O	1.80000
Scope 1	Fuels	Gaseous fuels	CNG		Energy - Gross CV	kWh (Gross CV)	kg CO2e	0.18316
Scope 1	Fuels	Gaseous fuels	CNG		Energy - Gross CV	kWh (Gross CV)	kg CH4	0.00025
Scope 1	Fuels	Gaseous fuels	CNG		Energy - Gross CV	kWh (Gross CV)	kg CO2	0.18282
Scope 1	Fuels	Gaseous fuels	CNG		Energy - Gross CV	kWh (Gross CV)	kg N2O	0.00010
Scope 1	Fuels	Gaseous fuels	CNG		Energy - Net CV	kWh (Net CV)	kg CO2e	0.20297

Figure 9.1: Screenshot of sample DEFRA 2021 emission factor dataset

Source: https://www.gov.uk/government/collections/government-conversion-factors-for-company-reporting