LAB ASSIGNMENT I

SUBJECT: NATURAL LANGUAGE PROCESSING LAB (PMDS606P)

CLASS ID - VL2025260105170

Name: Soumyadeep Ganguly

Reg No: 24MDT0082

Course: M.Sc Data Science

In [42]:
```
import nltk
from nltk.corpus import stopwords
from nltk import bigrams
from collections import Counter
import string
import re
from wordcloud import WordCloud
import matplotlib.pyplot as plt

nltk.download('punkt_tab')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt_tab to
[nltk_data]     C:\Users\sambh\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\sambh\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[42]: True

1. Write-up

My name is Soumyadeep Ganguly, as a dedicated and analytical data science professional, I am currently expanding my expertise through an M.Sc. in Data Science at the Vellore Institute of Technology, where I hold a CGPA of 8.73. My academic foundation was built at Maulana Abul Kalam Azad University of Technology, where I earned a B.Sc. in Information Technology with a specialization in Data Science, graduating with a 9.19 CGPA. My coursework has given me a strong understanding of statistics, deep learning, machine learning, and data visualization. I have practical experience as a Backend Developer for Barracksbuddy, where I developed their website and a complete LMS system. This role enhanced my skills in version control and production management. My passion for applying data to solve real-world problems is evident in my projects. I designed an IoT-based health monitoring system for factory workers, using machine learning to predict health trends. Additionally, I conducted a performance analysis of Indian Railway Zones using MCDM approaches, which resulted in a publication. I am proficient in languages such as Python, SQL, and JavaScript, and skilled in technologies including Tensorflow, Tableau, Power BI, and Flask. I am eager to leverage my academic background and hands-on experience to contribute to impactful data-driven solutions.

In [18]:
```
corpas = """My name is Soumyadeep Ganguly, as a dedicated and analytical data science professional, I am currently expanding m
```

2. Converting into lowercase.

In [19]:
```
lower_corpas = corpas.lower()
print(lower_corpas)
```

```
my name is soumyadeep ganguly, as a dedicated and analytical data science professional, i am currently expanding my expertise t
hrough an m.sc. in data science at the vellore institute of technology, where i hold a cgpa of 8.73. my academic foundation was
built at maulana abul kalam azad university of technology, where i earned a b.sc. in information technology with a specializati
on in data science, graduating with a 9.19 cgpa. my coursework has given me a strong understanding of statistics, deep learnin
g, machine learning, and data visualization. i have practical experience as a backend developer for barracksbuddy, where i deve
loped their website and a complete lms system. this role enhanced my skills in version control and production management. my pa
ssion for applying data to solve real-world problems is evident in my projects. i designed an iot-based health monitoring syste
m for factory workers, using machine learning to predict health trends. additionally, i conducted a performance analysis of ind
ian railway zones using mcdm approaches, which resulted in a publication. i am proficient in languages such as python, sql, and
javascript, and skilled in technologies including tensorflow, tableau, power bi, and flask. i am eager to leverage my academic
background and hands-on experience to contribute to impactful data-driven solutions.
```

3. Removing Punctuations

In [20]:
```
tokenized_words = nltk.word_tokenize(lower_corpas)
words_without_punctuation = []
for word in tokenized_words:
    if word not in string.punctuation:
        words_without_punctuation.append(word)

filtered_corpus = ' '.join(words_without_punctuation)
print(filtered_corpus)
```

```
my name is soumyadeep ganguly as a dedicated and analytical data science professional i am currently expanding my expertise thr
ough an m.sc in data science at the vellore institute of technology where i hold a cgpa of 8.73. my academic foundation was bui
lt at maulana abul kalam azad university of technology where i earned a b.sc in information technology with a specialization in
data science graduating with a 9.19 cgpa my coursework has given me a strong understanding of statistics deep learning machine
learning and data visualization i have practical experience as a backend developer for barracksbuddy where i developed their we
bsite and a complete lms system this role enhanced my skills in version control and production management my passion for applyi
ng data to solve real-world problems is evident in my projects i designed an iot-based health monitoring system for factory wor
kers using machine learning to predict health trends additionally i conducted a performance analysis of indian railway zones us
ing mcdm approaches which resulted in a publication i am proficient in languages such as python sql and javascript and skilled
in technologies including tensorflow tableau power bi and flask i am eager to leverage my academic background and hands-on expe
rience to contribute to impactful data-driven solutions
```

4. Removing Numbers

In [21]:
```
corpus_without_numbers = re.sub(r'\d+', '', filtered_corpus)
clean_corpus = re.sub(r'\s+', ' ', corpus_without_numbers).strip()
print(clean_corpus)
```

```
my name is soumyadeep ganguly as a dedicated and analytical data science professional i am currently expanding my expertise thr
ough an m.sc in data science at the vellore institute of technology where i hold a cgpa of .. my academic foundation was built
at maulana abul kalam azad university of technology where i earned a b.sc in information technology with a specialization in da
ta science graduating with a . cgpa my coursework has given me a strong understanding of statistics deep learning machine learn
ing and data visualization i have practical experience as a backend developer for barracksbuddy where i developed their website
and a complete lms system this role enhanced my skills in version control and production management my passion for applying dat
a to solve real-world problems is evident in my projects i designed an iot-based health monitoring system for factory workers u
sing machine learning to predict health trends additionally i conducted a performance analysis of indian railway zones using mc
dm approaches which resulted in a publication i am proficient in languages such as python sql and javascript and skilled in tec
hnologies including tensorflow tableau power bi and flask i am eager to leverage my academic background and hands-on experience
to contribute to impactful data-driven solutions
```

In [22]:
```
#removing some other puctuations
pattern = r'[\.,]+'
clean_corpus = re.sub(pattern, '', clean_corpus)
clean_corpus = re.sub(r'\s+', ' ', clean_corpus).strip()
print(clean_corpus)
```

```
my name is soumyadeep ganguly as a dedicated and analytical data science professional i am currently expanding my expertise thr
ough an msc in data science at the vellore institute of technology where i hold a cgpa of my academic foundation was built at m
aulana abul kalam azad university of technology where i earned a bsc in information technology with a specialization in data sc
ience graduating with a cgpa my coursework has given me a strong understanding of statistics deep learning machine learning and
data visualization i have practical experience as a backend developer for barracksbuddy where i developed their website and a c
omplete lms system this role enhanced my skills in version control and production management my passion for applying data to so
lve real-world problems is evident in my projects i designed an iot-based health monitoring system for factory workers using ma
chine learning to predict health trends additionally i conducted a performance analysis of indian railway zones using mcdm appr
oaches which resulted in a publication i am proficient in languages such as python sql and javascript and skilled in technologi
es including tensorflow tableau power bi and flask i am eager to leverage my academic background and hands-on experience to con
tribute to impactful data-driven solutions
```

5. Removing special characters.

In [23]:
```
pattern = r'[^a-zA-Z0-9\s]'
final_corpus = re.sub(pattern, '', clean_corpus)
print(final_corpus)
```

```
my name is soumyadeep ganguly as a dedicated and analytical data science professional i am currently expanding my expertise thr
ough an msc in data science at the vellore institute of technology where i hold a cgpa of my academic foundation was built at m
aulana abul kalam azad university of technology where i earned a bsc in information technology with a specialization in data sc
ience graduating with a cgpa my coursework has given me a strong understanding of statistics deep learning machine learning and
data visualization i have practical experience as a backend developer for barracksbuddy where i developed their website and a c
omplete lms system this role enhanced my skills in version control and production management my passion for applying data to so
lve realworld problems is evident in my projects i designed an iotbased health monitoring system for factory workers using mach
ine learning to predict health trends additionally i conducted a performance analysis of indian railway zones using mcdm approa
ches which resulted in a publication i am proficient in languages such as python sql and javascript and skilled in technologies
including tensorflow tableau power bi and flask i am eager to leverage my academic background and handson experience to contrib
ute to impactful datadriven solutions
```

6. Removing English stopwords

In [ ]:
```
tokens = nltk.word_tokenize(final_corpus)
stop_words = set(stopwords.words('english'))
filtered_stopwords = [w for w in tokens if not w in stop_words]
```

In [30]:
```
print(filtered_stopwords)
```

```
['name', 'soumyadeep', 'ganguly', 'dedicated', 'analytical', 'data', 'science', 'professional', 'currently', 'expanding', 'expe
rtise', 'msc', 'data', 'science', 'vellore', 'institute', 'technology', 'hold', 'cgpa', 'academic', 'foundation', 'built', 'mau
lana', 'abul', 'kalam', 'azad', 'university', 'technology', 'earned', 'bsc', 'information', 'technology', 'specialization', 'da
ta', 'science', 'graduating', 'cgpa', 'coursework', 'given', 'strong', 'understanding', 'statistics', 'deep', 'learning', 'mach
ine', 'learning', 'data', 'visualization', 'practical', 'experience', 'backend', 'developer', 'barracksbuddy', 'developed', 'we
bsite', 'complete', 'lms', 'system', 'role', 'enhanced', 'skills', 'version', 'control', 'production', 'management', 'passion',
'applying', 'data', 'solve', 'realworld', 'problems', 'evident', 'projects', 'designed', 'iotbased', 'health', 'monitoring', 's
ystem', 'factory', 'workers', 'using', 'machine', 'learning', 'predict', 'health', 'trends', 'additionally', 'conducted', 'perf
ormance', 'analysis', 'indian', 'railway', 'zones', 'using', 'mcdm', 'approaches', 'resulted', 'publication', 'proficient', 'la
nguages', 'python', 'sql', 'javascript', 'skilled', 'technologies', 'including', 'tensorflow', 'tableau', 'power', 'bi', 'flas
k', 'eager', 'leverage', 'academic', 'background', 'handson', 'experience', 'contribute', 'impactful', 'datadriven', 'solution
s']
```

In [37]:
```
my_bigrams = list(bigrams(filtered_stopwords))
```

7. Top 50 Most common bigrams

In [41]:
```
bigram_counts = Counter(my_bigrams)
print("Index     Bigram              Frequency")
for i, (bigram, count) in enumerate(bigram_counts.most_common(50), 1):
  b_str = f"{bigram[0]} {bigram[1]}"
  print(f"{i:<6} {b_str:<25} {count:<10}")
```

Index	Bigram	Frequency
1	data science	3
2	machine learning	2
3	name soumyadeep	1
4	soumyadeep ganguly	1
5	ganguly dedicated	1
6	dedicated analytical	1
7	analytical data	1
8	science professional	1
9	professional currently	1
10	currently expanding	1
11	expanding expertise	1
12	expertise msc	1
13	msc data	1
14	science vellore	1
15	vellore institute	1
16	institute technology	1
17	technology hold	1
18	hold cgpa	1
19	cgpa academic	1
20	academic foundation	1
21	foundation built	1
22	built maulana	1
23	maulana abul	1
24	abul kalam	1
25	kalam azad	1
26	azad university	1
27	university technology	1
28	technology earned	1
29	earned bsc	1
30	bsc information	1
31	information technology	1
32	technology specialization	1
33	specialization data	1
34	science graduating	1
35	graduating cgpa	1
36	cgpa coursework	1
37	coursework given	1
38	given strong	1
39	strong understanding	1
40	understanding statistics	1

```
41     statistics deep           1
42     deep learning             1
43     learning machine          1
44     learning data             1
45     data visualization        1
46     visualization practical   1
47     practical experience      1
48     experience backend        1
49     backend developer         1
50     developer barracksbuddy   1
```

8. Word Cloud of Most Common Bigrams

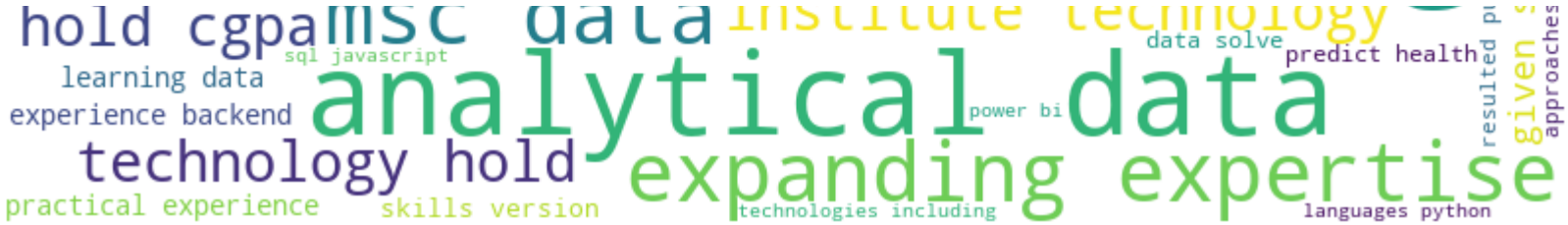
In [43]:
```
bigram_dict = {" ".join(bigram): count for bigram, count in bigram_counts.items()}


wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                min_font_size = 10).generate_from_frequencies(bigram_dict)

plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```


coursework given
realworld problems
built maulana
technology specialization
factory workers
evident projects
science professional
mcdm approaches
conducted performance
system factory
expertise msc
vellore institute
understanding statistics
projects designed
production management
deep learning
name soumyadeep
website complete
workers using
learning predict
science graduating
using machine
tableau power
system role
earned bsc
eager leverage
graduating cgpa
ganguly dedicated
cgpa coursework
background handson
university technology
kalam azad
health trends
handson experience
currently expanding
including tensorflow
data visualization
leverage academic
iotbased health
zones using
using mcdm
information technology
bsc information
learning machine
contribute impactful
designed iotbased
monitoring system
technology earned
data science
management passion
foundation built
academic foundation
health monitoring
tensorflow tableau
passion applying
visualization practical
professional currently
control production
barracksbuddy developed
trends additionally
strong understanding
bi flask
version control
science vellore
python sql
abul kalam
academic background
railway zones
complete lms
dedicated analytical
publication proficient
cgpa academic
azad university
maulana abul
javascript skilled
additionally conducted
developer barracksbuddy
soumyadeep ganguly
lms system
developed website
solve realworld
role enhanced
applying data
statistics deep
enhanced skills
specialization data
backend developer
analysis indian
performance analysis
machine learning
flask eager
problems evident
indian railway
proficient languages
experience contribute
skilled technologies


hold cgpa
msc data
institute technology
sql javascript
learning data
analytical data
data solve
predict health
experience backend
power bi
technology hold
expanding expertise
practical experience
skills version
technologies including
languages python
resulted
given
approaches

In [ ]: