Explanation of the Solution

In this project, I developed a high-performance, AI-driven web crawling tool that not only gathers data from websites efficiently but also processes it with advanced AI techniques to enhance its value. Here's a breakdown of the approach and technologies I used:

1. Optimized, Parallelized Crawling

- I built a `WebsiteCrawler` class using `ThreadPoolExecutor` for multi-threading, allowing simultaneous page crawls. This parallelization is key to speeding up the process, as multiple pages are fetched concurrently, significantly reducing overall runtime.
- Each thread works from a shared URL queue, with thread-safe locks to manage shared resources, ensuring data consistency while gathering information at scale.

2. Advanced Data Extraction and Parsing

- For each page, I used `BeautifulSoup` to parse HTML and extract meaningful content, specifically targeting `<p>` tags. This parsed content is stored in a dictionary keyed by URL, enabling easy access and future analysis.
- The crawler also dynamically discovers and queues internal links, expanding the scope to capture all relevant internal pages.

3. AI-Powered Processing with Google Vertex AI

- To transform raw data into actionable insights, I integrated Google Vertex AI with a grounding technique using Google Search. The grounding approach enriches the model’s output by leveraging real-time, contextual information from search results, which adds depth and relevance to the generated content.
- Additionally, I employed a custom prompt engineering strategy, tailoring prompts to ensure the model returns concise, relevant responses suited to the data analysis needs of the project. This prompt strategy was essential for directing the AI to produce outputs aligned with specific informational goals, making the results more actionable and informative.

4. Secure and Scalable API Configuration

- I set up secure handling of credentials using environment variables, decoding them from a base64 JSON file, which keeps sensitive information protected and simplifies cloud deployment.
- This structure ensures both data privacy and compatibility with cloud infrastructure, supporting secure and scalable usage.

By combining parallel processing and Vertex AI’s advanced techniques, including grounding and strategic prompt engineering, this solution can efficiently crawl, process, and enhance web data at scale. This project showcases expertise in both multi-threaded software engineering and the application of cutting-edge AI techniques for intelligent data processing and analysis.

Crawler

```
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from bs4 import BeautifulSoup
from urllib.parse import urljoin, urlparse
from concurrent.futures import ThreadPoolExecutor
import threading
import queue
import time

class WebsiteCrawler:
    def __init__(self, base_url, max_workers=5):
        self.base_url = base_url
        self.visited_urls = set()
        self.crawled_data = {}  # Dictionary to store data with URL as
key
        self.url_queue = queue.Queue()
        self.lock = threading.Lock()  # To handle thread-safe
operations on shared resources

        # Add the initial URL to the queue
        self.url_queue.put(base_url)

        # Set up Selenium WebDriver options for headless mode
        self.chrome_options = Options()
        self.chrome_options.add_argument("--headless")
        self.chrome_options.add_argument("--no-sandbox")
        self.chrome_options.add_argument("--disable-dev-shm-usage")
        self.chrome_options.add_argument(
            "user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Safari/537.36"
        )

        # Initialize the WebDriver service to be used by each thread
        self.service = Service()

        # Define the number of threads to use
        self.max_workers = max_workers

    def is_internal_link(self, link):
        parsed_base_url = urlparse(self.base_url)
        parsed_link = urlparse(link)
        return parsed_link.netloc == parsed_base_url.netloc or
parsed_link.netloc == ""

    def scrape_page(self, url):
        driver = webdriver.Chrome(service=self.service,
options=self.chrome_options)
```

```
        try:
            driver.get(url)
            time.sleep(2)  # Wait for JavaScript to load if necessary
            soup = BeautifulSoup(driver.page_source, "html.parser")
            return soup
        except Exception as e:
            print(f"Failed to fetch {url}: {e}")
            return None
        finally:
            driver.quit()

    def crawl_page(self):
        while not self.url_queue.empty():
            url = self.url_queue.get()
            if url in self.visited_urls:
                continue  # Skip if already visited

            # Mark as visited in a thread-safe way
            with self.lock:
                self.visited_urls.add(url)

            print(f"Visiting: {url}")
            soup = self.scrape_page(url)
            if soup is None:
                continue  # Skip if the page couldn't be fetched

            # Extract and store page content
            page_content = "\n".join([p.get_text() for p in
soup.find_all("p")])
            with self.lock:
                self.crawled_data[url] = page_content  # Store data
with URL as the key

            print("page_content:")
            print(f"{page_content[:200]} ...")

            # Find and enqueue internal links for crawling
            for link_tag in soup.find_all("a", href=True):
                link = urljoin(url, link_tag['href'])
                if link not in self.visited_urls and
self.is_internal_link(link) and not any(term in link.lower() for term
in ["blog", "blogs"]):
                    self.url_queue.put(link)

            self.url_queue.task_done()  # Mark the task as done

    def crawl(self):
        # Use ThreadPoolExecutor to manage a pool of threads
        with ThreadPoolExecutor(max_workers=self.max_workers) as
executor:
```

```
            # Submit the `crawl_page` function to each worker thread
            futures = [executor.submit(self.crawl_page) for _ in
range(self.max_workers)]
            # Wait for all threads to complete their tasks
            for future in futures:
                future.result()

    def get_crawled_data(self):
        return self.crawled_data

    def close(self):
        # No driver to close explicitly in this setup as each thread
manages its own
        pass

base_url = "https://growthx.com"

crawler = WebsiteCrawler(base_url, max_workers=5)
crawler.crawl()
crawled_data = crawler.get_crawled_data()
crawler.close()
```

```
Visiting: https://growthx.com
page_content:
We believe that the entrepreneur’s journey is not meant to be traveled
alone, and that hands-on sales help from passionate experts should be
accessible to everyone.
We understand that sales is dauntin ...
Visiting: https://growthx.com#content
page_content:
We believe that the entrepreneur’s journey is not meant to be traveled
alone, and that hands-on sales help from passionate experts should be
accessible to everyone.
We understand that sales is dauntin ...
Visiting: https://growthx.com/
page_content:
We believe that the entrepreneur’s journey is not meant to be traveled
alone, and that hands-on sales help from passionate experts should be
accessible to everyone.
We understand that sales is dauntin ...
Visiting: https://growthx.com/founders/
page_content:
We help B2B founders improve sales performance.
Whether it’s your first sale or you’re ready to scale, we are
committed to doing the real work of B2B sales with you.
That’s how we identify investment ...
Visiting: https://growthx.com/founders/revenue-accelerator/
page_content:
The Revenue Accelerator helps B2B founders improve sales performance
and get consistent results from the time and effort being invested
```

```
into go-to-market activities.
You’re assigned a dedicated expert ...
Visiting: https://growthx.com/founders/golisano-institute/
page_content:
The Revenue Accelerator at Golisano Institute for Business &
Entrepreneurship helps Western New York-based B2B tech companies
improve sales performance.
You’re assigned a dedicated expert sales coach  ...
Visiting: https://growthx.com/founders/alberta-founders/
page_content:
The Alberta Innovates Revenue Accelerator helps Alberta-based tech
startups improve sales performance and create consistent revenue
results.
You’re assigned a dedicated expert sales coach with proven  ...
Visiting: https://growthx.com/companies/
page_content:
companies
rising cities
tech sectors
nps score
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/investors/
page_content:
As experts, we help B2B founders improve sales performance and create
consistent results from the time and effort they are investing in go-
to-market activities.
As investors, we designed our program t ...
Visiting: https://growthx.com/accelerators/
page_content:
As an extension of your program, founders will work with one of our
dedicated expert sales coaches for 16 weeks as they use our proven
playbooks to create systematic revenue growth and secure outside  ...
Visiting: https://growthx.com/testimonials/
page_content:
"GrowthX and Golisano Institute helped us evolve the way that we talk
about our product - it's now less about what we do and more about what
we do for the customer. That's really carried a long way, i ...
Visiting: https://growthx.com/about-us/
page_content:
We love working in the weeds alongside great people. That’s where we
are happiest and most able to make the biggest impact.
We genuinely value the opportunity to help people. That’s how lasting
relati ...
Visiting: https://growthx.com/our-team/
page_content:
We are uniquely positioned to help founders solve their most difficult
sales challenges and reimagine venture investing.
“Our coach helped our company tremendously! One of the best we’ve had
```

```
in our jo ...
Visiting: https://growthx.com/contact/
page_content:
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com#helpFull
page_content:
We believe that the entrepreneur’s journey is not meant to be traveled
alone, and that hands-on sales help from passionate experts should be
accessible to everyone.
We understand that sales is dauntin ...
Visiting: https://growthx.com/revenue-accelerator/
page_content:
The Revenue Accelerator helps B2B founders improve sales performance
and get consistent results from the time and effort being invested
into go-to-market activities.
You’re assigned a dedicated expert ...
Visiting: https://growthx.com/privacy-policy/
page_content:
Effective Date: 22 June 2023
At GrowthX, we value your privacy and are committed to protecting your
personal information. This Privacy Policy outlines how we collect,
use, disclose, and protect the in ...
Visiting: https://growthx.com/#content
page_content:
We believe that the entrepreneur’s journey is not meant to be traveled
alone, and that hands-on sales help from passionate experts should be
accessible to everyone.
We understand that sales is dauntin ...
Visiting: https://growthx.com/#helpFull
page_content:
We believe that the entrepreneur’s journey is not meant to be traveled
alone, and that hands-on sales help from passionate experts should be
accessible to everyone.
We understand that sales is dauntin ...
Visiting: https://growthx.com/founders/#content
page_content:
We help B2B founders improve sales performance.
Whether it’s your first sale or you’re ready to scale, we are
committed to doing the real work of B2B sales with you.
That’s how we identify investment ...
Visiting: https://growthx.com/founders/revenue-accelerator/#content
page_content:
The Revenue Accelerator helps B2B founders improve sales performance
and get consistent results from the time and effort being invested
into go-to-market activities.
You’re assigned a dedicated expert ...
Visiting: https://growthx.com/revenue-accelerator-application/
page_content:
```

```
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/founders/revenue-accelerator/#testiCar
page_content:
The Revenue Accelerator helps B2B founders improve sales performance
and get consistent results from the time and effort being invested
into go-to-market activities.
You’re assigned a dedicated expert ...
Visiting: https://growthx.com/founders/golisano-institute/#content
page_content:
The Revenue Accelerator at Golisano Institute for Business &
Entrepreneurship helps Western New York-based B2B tech companies
improve sales performance.
You’re assigned a dedicated expert sales coach  ...
Visiting: https://growthx.com/revenue-accelerator-golisano-
application/
page_content:
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/founders/golisano-institute/#testiCar
page_content:
The Revenue Accelerator at Golisano Institute for Business &
Entrepreneurship helps Western New York-based B2B tech companies
improve sales performance.
You’re assigned a dedicated expert sales coach  ...
Visiting: https://growthx.com/alberta-revenue-accelerator-application/
page_content:
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/founders/alberta-founders/#content
page_content:
The Alberta Innovates Revenue Accelerator helps Alberta-based tech
startups improve sales performance and create consistent revenue
results.
You’re assigned a dedicated expert sales coach with proven  ...
Visiting: https://growthx.com/founders/alberta-founders/#testiCar
page_content:
The Alberta Innovates Revenue Accelerator helps Alberta-based tech
startups improve sales performance and create consistent revenue
results.
You’re assigned a dedicated expert sales coach with proven  ...
Visiting: https://growthx.com/companies/#content
page_content:
companies
rising cities
tech sectors
nps score
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
```

```
Visiting: https://growthx.com/investors/#content
page_content:
As experts, we help B2B founders improve sales performance and create
consistent results from the time and effort they are investing in go-
to-market activities.
As investors, we designed our program t ...
Visiting: https://growthx.com/investors/#trAccelerator
page_content:
As experts, we help B2B founders improve sales performance and create
consistent results from the time and effort they are investing in go-
to-market activities.
As investors, we designed our program t ...
Visiting: https://growthx.com/accelerators/#content
page_content:
As an extension of your program, founders will work with one of our
dedicated expert sales coaches for 16 weeks as they use our proven
playbooks to create systematic revenue growth and secure outside  ...
Visiting: https://growthx.com/accelerators/#oap
page_content:
As an extension of your program, founders will work with one of our
dedicated expert sales coaches for 16 weeks as they use our proven
playbooks to create systematic revenue growth and secure outside  ...
Visiting: https://growthx.com/testimonials/#content
page_content:
"GrowthX and Golisano Institute helped us evolve the way that we talk
about our product - it's now less about what we do and more about what
we do for the customer. That's really carried a long way, i ...
Visiting: https://growthx.com/about-us/#content
page_content:
We love working in the weeds alongside great people. That’s where we
are happiest and most able to make the biggest impact.
We genuinely value the opportunity to help people. That’s how lasting
relati ...
Visiting: https://growthx.com/growthx-capital/
page_content:
COMING SOON
GrowthX Capital is launching a new fund to invest in the rising city
founders across North America who we have had the privilege to build a
lasting relationship with by helping them get to ...
Visiting: https://growthx.com/founders/capital#pledge
page_content:
COMING SOON
GrowthX Capital is launching a new fund to invest in the rising city
founders across North America who we have had the privilege to build a
lasting relationship with by helping them get to ...
Visiting: https://growthx.com/our-team/#content
page_content:
We are uniquely positioned to help founders solve their most difficult
sales challenges and reimagine venture investing.
```

```
“Our coach helped our company tremendously! One of the best we’ve had
in our jo ...
Visiting: https://growthx.com/our-team/#oTeam
page_content:
We are uniquely positioned to help founders solve their most difficult
sales challenges and reimagine venture investing.
“Our coach helped our company tremendously! One of the best we’ve had
in our jo ...
Visiting: https://growthx.com/contact/#content
page_content:
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/revenue-accelerator/#content
page_content:
The Revenue Accelerator helps B2B founders improve sales performance
and get consistent results from the time and effort being invested
into go-to-market activities.
You’re assigned a dedicated expert ...
Visiting: https://growthx.com/revenue-accelerator/#testiCar
page_content:
The Revenue Accelerator helps B2B founders improve sales performance
and get consistent results from the time and effort being invested
into go-to-market activities.
You’re assigned a dedicated expert ...
Visiting: https://growthx.com/privacy-policy/#content
page_content:
Effective Date: 22 June 2023
At GrowthX, we value your privacy and are committed to protecting your
personal information. This Privacy Policy outlines how we collect,
use, disclose, and protect the in ...
Visiting: https://growthx.com/revenue-accelerator-application/#content
page_content:
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/revenue-accelerator-golisano-
application/#content
page_content:
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/alberta-revenue-accelerator-
application/#content
page_content:
© 2024 GrowthX | Privacy Policy
Best Ways To Say Hello: ...
Visiting: https://growthx.com/growthx-capital/#content
page_content:
COMING SOON
GrowthX Capital is launching a new fund to invest in the rising city
founders across North America who we have had the privilege to build a
```

```
lasting relationship with by helping them get to ...
Visiting: https://growthx.com/founders/capital#content
page_content:
COMING SOON
GrowthX Capital is launching a new fund to invest in the rising city
founders across North America who we have had the privilege to build a
lasting relationship with by helping them get to ...
Visiting: https://growthx.com/founders/capital
page_content:
COMING SOON
GrowthX Capital is launching a new fund to invest in the rising city
founders across North America who we have had the privilege to build a
lasting relationship with by helping them get to ...
```

Vextex AI with Gemini pro 1.5

```python
import os
from dotenv import load_dotenv
import base64
import json
from google.oauth2 import service_account

load_dotenv()

class Settings:
    def __init__(self):
        self.SERVICE_ACCOUNT_JSON_BASE64 =
os.getenv("SERVICE_ACCOUNT_JSON_BASE64")

        self.CREDENTIALS =
self._get_credentials_from_base64(self.SERVICE_ACCOUNT_JSON_BASE64)
        self.PROJECT_ID =
self._extract_project_info_from_base64(self.SERVICE_ACCOUNT_JSON_BASE6
4)

    def _get_credentials_from_base64(self,
service_account_json_base64: str):
        """
        Decodifica as credenciais da conta de serviço a partir do
base64.
        """
        decoded_credentials =
base64.b64decode(service_account_json_base64)
        credentials =
service_account.Credentials.from_service_account_info(
            json.loads(decoded_credentials)
        )
        return credentials
    def _extract_project_info_from_base64(self,
service_account_json_base64: str):
        """
        Extrai informações de projeto do JSON base64.
        """
        decoded_credentials =
base64.b64decode(service_account_json_base64)
        credentials_info = json.loads(decoded_credentials)

        project_id = credentials_info.get("project_id")
        if not project_id:
            raise ValueError("O campo 'project_id' não foi encontrado
nas credenciais.")

        return project_id

settings = Settings()

import base64
import json
import requests
from google.oauth2 import service_account
import vertexai
from vertexai.preview.generative_models import grounding
from vertexai.generative_models import (
    FunctionDeclaration,
    GenerationConfig,
    GenerativeModel,
    Part,
    Content,
    Tool,
    ToolConfig
)

class VertexAIProvider:
    def __init__(self, model_name, credentials, project_id, location:
str = 'us-central1'):
        vertexai.init(credentials=credentials, project=project_id,
location=location)
        self.model = GenerativeModel(model_name=model_name,

generation_config=GenerationConfig(temperature=0)
                                    )

    def call_llm(self, prompt: str, tools = []) -> str:
        """
        Calls the LLM with the provided prompt.
        """
        #print(f"Prompt: {prompt}")
        try:
            response = self.model.generate_content(
                prompt,
                generation_config=GenerationConfig(temperature=0),
                tools=tools
            )


            answer =
response.candidates[0].content.parts[0]._raw_part.text

            return answer

        except Exception as e:
            print(f"Error: {e}")
            raise e

vertexai_provider = VertexAIProvider(model_name="gemini-1.5-flash-
001",
                                     credentials=settings.CREDENTIALS,
                                     project_id=settings.PROJECT_ID
                                    )

PROMPT = f"""Using the provided dictionary data and the google search
tool, generate a concise,
informative article about the prospective company for the sales team.
Structure the article based on the following key sections:

Company Overview: Summarize the 'industry', 'products' or 'services',
and 'mission' fields from the dictionary to provide a clear snapshot
of what the company does.
Market Position and Differentiators: Use fields like 'target market',
'competitive advantages', or 'unique selling points' to highlight what
sets the company apart and their market focus.
Recent News or Highlights: Refer to 'recent achievements',
'partnerships', or 'expansions' to mention any current events or
strategic moves.
Key Contacts or Engagement Points: If the dictionary includes
'contacts' or 'departments', list any relevant points of contact that
may assist the sales team.
The output should be organized and brief, aimed at giving the sales
team a quick and clear understanding of the company’s profile and
potential engagement opportunities. Maintain a professional tone
throughout.

Data:
{crawled_data}
"""
# Use Google Search for grounding
tool =
Tool.from_google_search_retrieval(grounding.GoogleSearchRetrieval())
result = vertexai_provider.call_llm(prompt=PROMPT, tools=[tool])

print(result)
```

```
## GrowthX: Empowering B2B Founders to Scale Revenue

**Company Overview:**

GrowthX is a company dedicated to helping B2B founders achieve
sustainable revenue growth. They offer hands-on sales support and
expertise, believing that the entrepreneurial journey is best tackled
with a community of passionate experts. Their mission is to take the
guesswork out of go-to-market strategies and empower founders to build
successful businesses.

**Market Position and Differentiators:**

GrowthX focuses on the B2B tech startup market, particularly those in
the pre-seed and seed stages. They differentiate themselves by
offering a unique blend of expertise and hands-on support. Their
"Revenue Accelerator" program provides founders with dedicated expert
sales coaches and proven sales playbooks, enabling them to develop a
tailored go-to-market strategy and achieve real revenue results.

**Recent News or Highlights:**

GrowthX has recently secured a $1.5 million seed round, the largest
ever community-led seed round in India. This funding will fuel their
expansion and further support their mission of empowering founders.
They have also launched a new venture capital fund, GrowthX Capital,
to invest in promising B2B startups.

**Key Contacts or Engagement Points:**

The sales team can engage with GrowthX through their website, where
they can learn more about their programs and services. They can also
reach out to the GrowthX team directly through their contact page.

**Additional Information:**

GrowthX has a strong community of over 3,500 members, including
leaders from top internet companies in India. They offer a variety of
resources and programs, including live masterclasses, capstone
projects, and curated events. Their approach is highly collaborative
and emphasizes practical application of learned frameworks.
```

Result

GrowthX: Empowering B2B Founders to Scale Revenue

Company Overview:

GrowthX is a company dedicated to helping B2B founders achieve sustainable revenue growth. They offer hands-on sales support and expertise, believing that the entrepreneurial journey is best tackled with a community of passionate experts. Their mission is to take the guesswork out of go-to-market strategies and empower founders to build successful businesses.

Market Position and Differentiators:

GrowthX focuses on the B2B tech startup market, particularly those in the pre-seed and seed stages. They differentiate themselves by offering a unique blend of expertise and hands-on support. Their "Revenue Accelerator" program provides founders with dedicated expert sales coaches and proven sales playbooks, enabling them to develop a tailored go-to-market strategy and achieve real revenue results.

Recent News or Highlights:

GrowthX has recently secured a $1.5 million seed round, the largest ever community-led seed round in India. This funding will fuel their expansion and further support their mission of empowering founders. They have also launched a new venture capital fund, GrowthX Capital, to invest in promising B2B startups.

Key Contacts or Engagement Points:

The sales team can engage with GrowthX through their website, where they can learn more about their programs and services. They can also reach out to the GrowthX team directly through their contact page.

Additional Information:

GrowthX has a strong community of over 3,500 members, including leaders from top internet companies in India. They offer a variety of resources and programs, including live masterclasses, capstone projects, and curated events. Their approach is highly collaborative and emphasizes practical application of learned frameworks.