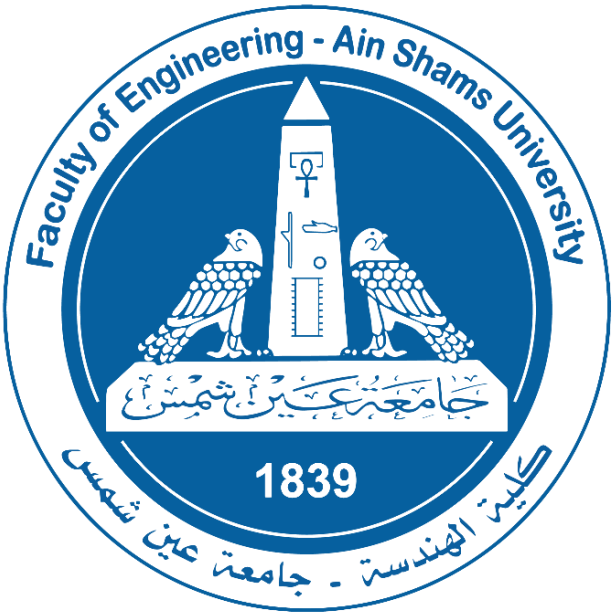


CSE354 Distributed Computing

Spring 2025

Team 15

Distributed Web Crawling and Indexing System

Phase 4 Report

GitHub: https://github.com/Menna-Ayman-Geba/Distributed-Web-Crawling-and-Indexing-System

Presented to:

DR. Ayman Mohamed Bahaa Eldin,

DR. Hossam Mohamed Abdelrahman

Eng. A'laa Hamdy, Eng. Mostafa Ashraf

Presented by:

Name	ID
Menna Ayman Alsaid Geba	2101438
Nancy Amro Hussein Mansour	2101421
Rana Mohamed Ahmed Shaqr	23P0150
Sara Mohamed Ashour Hussein	21P0337

Contents

Team Roles

Name	Role
Nancy Amro & Sara Ashour & Menna Ayman & Rana Mohamed	Architect, Tester/Documentation Lead
Nancy Amro & Sara Ashour	Master Lead
Sara Ashour & Menna Ayman	Crawler Lead, Tester/Documentation Lead
Rana Mohamed & Nancy Amro	Indexer Lead, Tester/Documentation Lead
Rana Mohamed & Menna Ayman	Cloud Infrastructure Lead, Tester/Documentation Lead

Introduction

The project has progressed through three phases, establishing a robust foundation for the final phase. Phase 1 focused on defining the system architecture, selecting technologies (Python, AWS SQS, S3, Beautiful Soup, Whoosh), and setting up the AWS environment. Phase 2 implemented core crawling and indexing functionality, enabling distributed crawling with basic politeness and a simple keyword-based index. Phase 3 enhanced the system with improved indexing, fault tolerance mechanisms (e.g., heartbeat monitoring, task re-queueing), and basic monitoring, ensuring resilience to crawler failures and persistent data storage. These phases have delivered a functional system capable of crawling websites, indexing content, and handling failures, setting the stage for Phase 4.

he system enters Phase 4 with a functional distributed crawler that respects robots.txt, a Whoosh-based indexer supporting phrase searches, and a master node managing tasks via AWS SQS with fault tolerance mechanisms (e.g., re-queueing tasks after 60-second heartbeat timeouts, terminating after 5 empty polls). Monitoring and logging enhancements from Phase 3, particularly for debugging inconsistent heartbeats from crawler2, provide a strong basis for final testing. However, challenges such as potential edge cases in task re-queueing and the need for rigorous scalability testing remain. Phase 4 will leverage the established architecture and prior testing results to polish the system, ensuring it meets all requirements and is ready for a professional demonstration.

Time plan

Gantt chart

Milestone Description	Progress (%)	Start Date	Days	Timeline
Phase 1: Project Inception and Setup				
Team Formation and Roles	100%	April 6, 2025	2	April 6 – April 7
Requirement Refinement	100%	April 6, 2025	2	April 6 – April 7
Technology Selection	100%	April 6, 2025	3	April 6 – April 8
Basic System Architecture Design	100%	April 6, 2025	7	April 6 – April 12
Cloud Environment Setup (Minimal)	100%	April 8, 2025	5	April 8 – April 12
Code Repository Setup	100%	April 8, 2025	5	April 8 – April 12
Phase 1 Deliverables (Basic Plan, Diagram)	100%	April 10, 2025	3	April 10 – April 12
Phase 2: Core Crawling and Indexing				
Crawler Node Implementation (Basic)	100%	April 13, 2025	14	April 13 – April 26
Master Node Implementation (Task Distribution)	100%	April 13, 2025	14	April 13 – April 26
Basic Indexer Node Implementation (In-Memory)	100%	April 13, 2025	14	April 13 – April 26
Basic Integration (Queue, Workflow)	100%	April 20, 2025	7	April 20 – April 26

Initial Testing (Core Functionality)	100%	April 20, 2025	7	April 20 – April 26
Phase 2 Deliverables (Basic Report)	100%	April 20, 2025	7	April 20 – April 26
Phase 3: Basic Fault Tolerance and Interface				
Fault Tolerance (Crawler Node Failure)	0%	April 27, 2025	7	April 27 – May 3
Basic Monitoring (Logging)	0%	April 27, 2025	5	April 27 – May 1
Data Persistence (Simple Cloud Storage)	0%	April 27, 2025	5	April 27 – May 1
Basic Client Interface (Command-Line)	0%	April 27, 2025	5	April 27 – May 1
Testing (Fault Tolerance, Basic Search)	0%	April 29, 2025	5	April 29 – May 3
Phase 3 Deliverables (Updated Report)	0%	April 29, 2025	5	April 29 – May 3
Phase 4: Testing, Documentation, and Demo				
Functional and Fault Tolerance Testing	0%	May 4, 2025	8	May 4 – May 11
Bug Fixing and Refinement	0%	May 4, 2025	8	May 4 – May 11
Basic Documentation (System, User Guide)	0%	May 4, 2025	8	May 4 – May 11
Deployment and Demonstration Preparation	0%	May 7, 2025	5	May 7 – May 11
Final Report and Presentation Materials	0%	May 7, 2025	5	May 7 – May 11
Phase 4 Deliverables (Report, Docs, Demo)	0%	May 7, 2025	5	May 7 – May 11

System Documentation

Final System Design Document

Used Technologies

Programming Language: Python

Web Crawling Libraries: Beautiful Soup, requests

Indexing Libraries: Whoosh

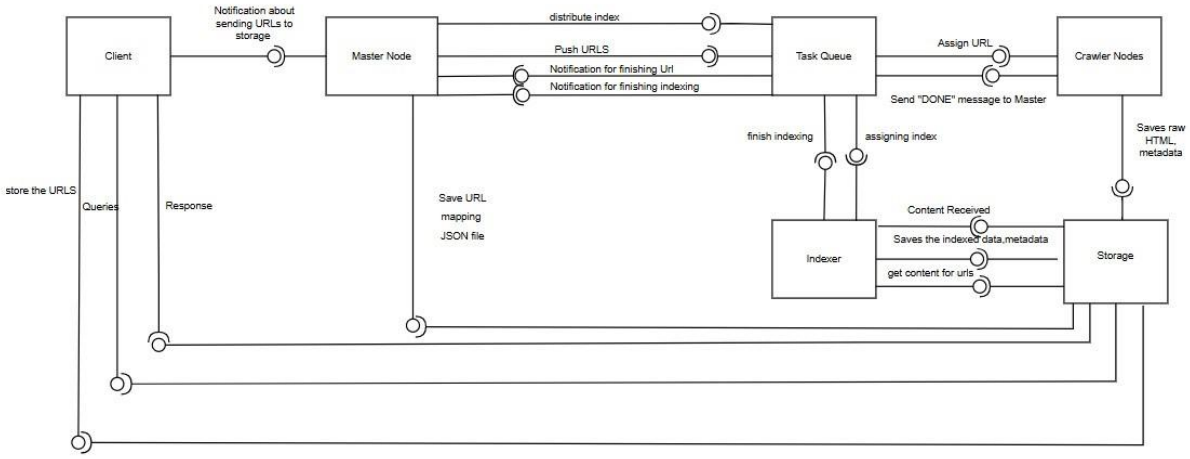
Distributed Task Queue: AWS SQS.

Cloud Platform: AWS.

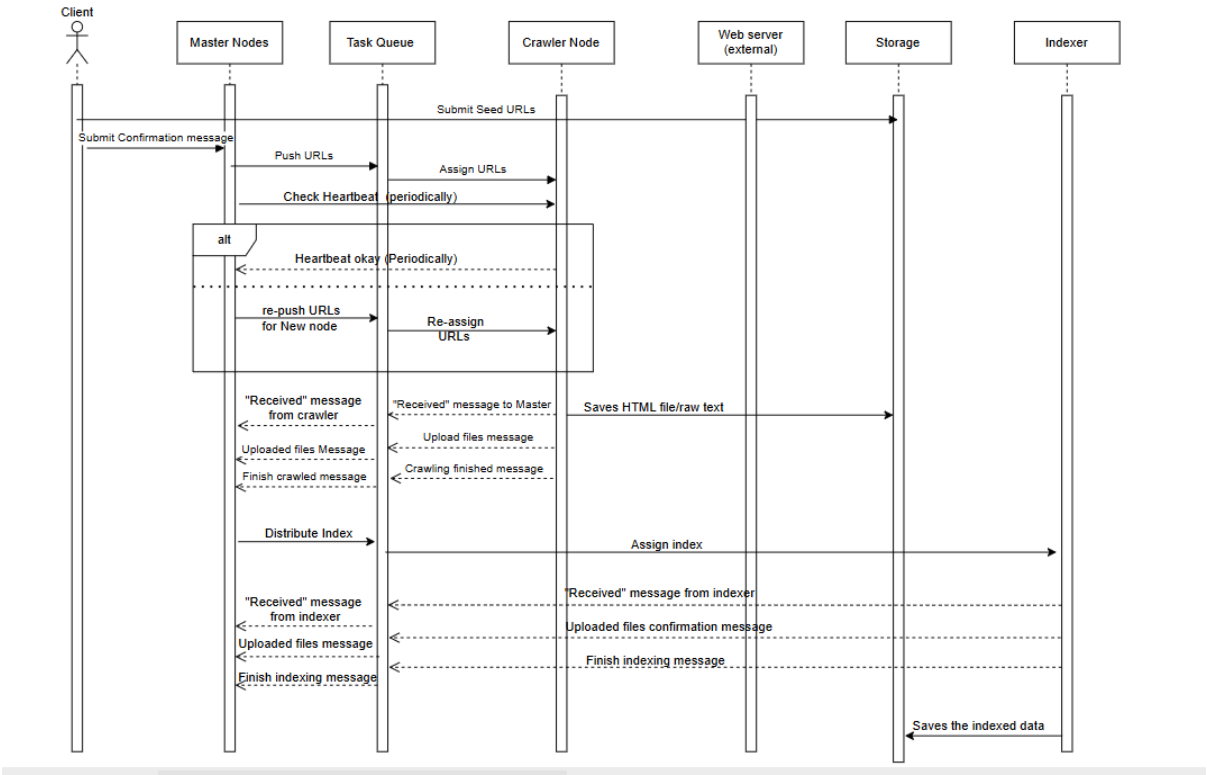
Storage: S3

Architecture diagrams and component interactions

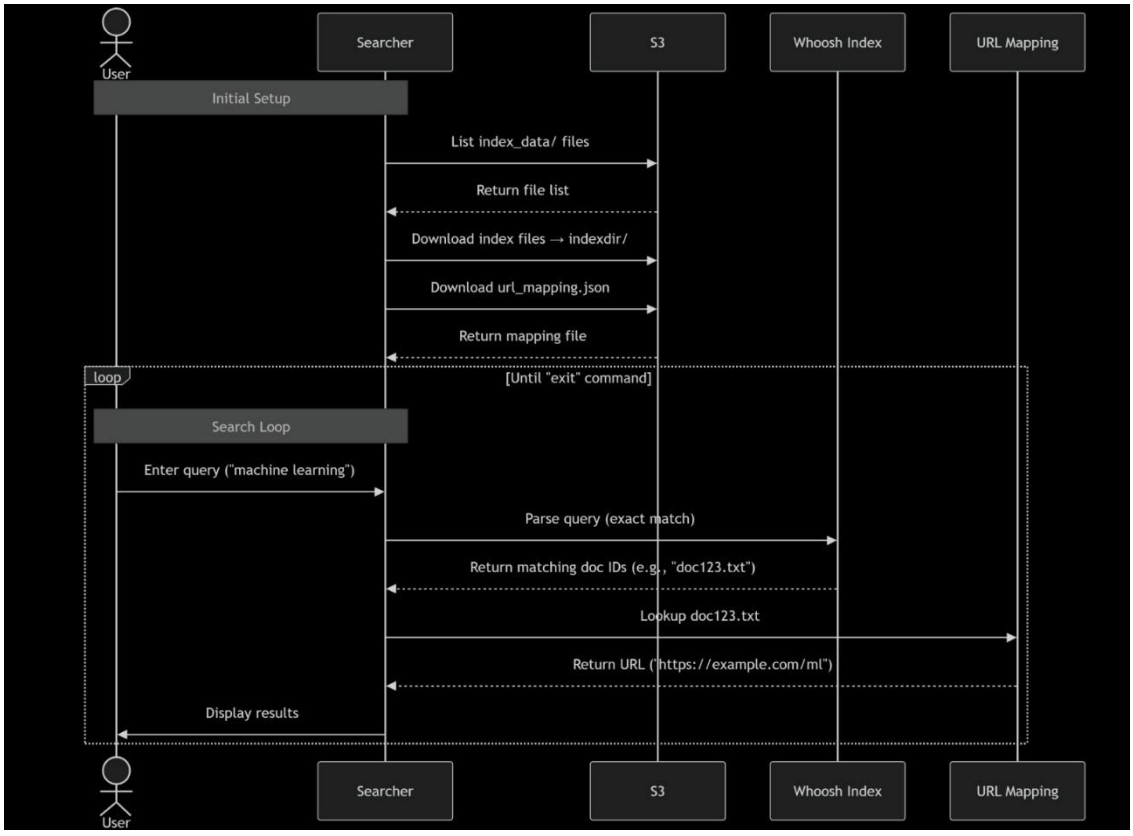
Components diagram



Sequence diagram (client sending URLs)



Sequence diagram (client querying)



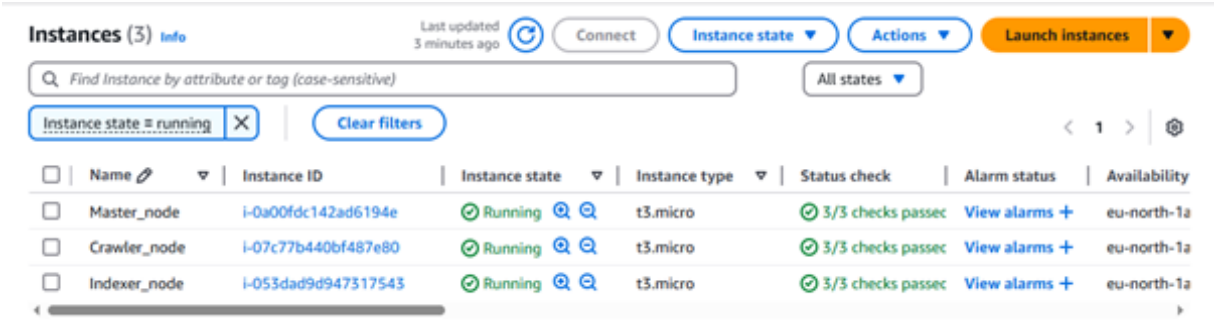
Deployment Guide

This step was done in phase 2& 3 and nothing much changed but here are all the steps combined

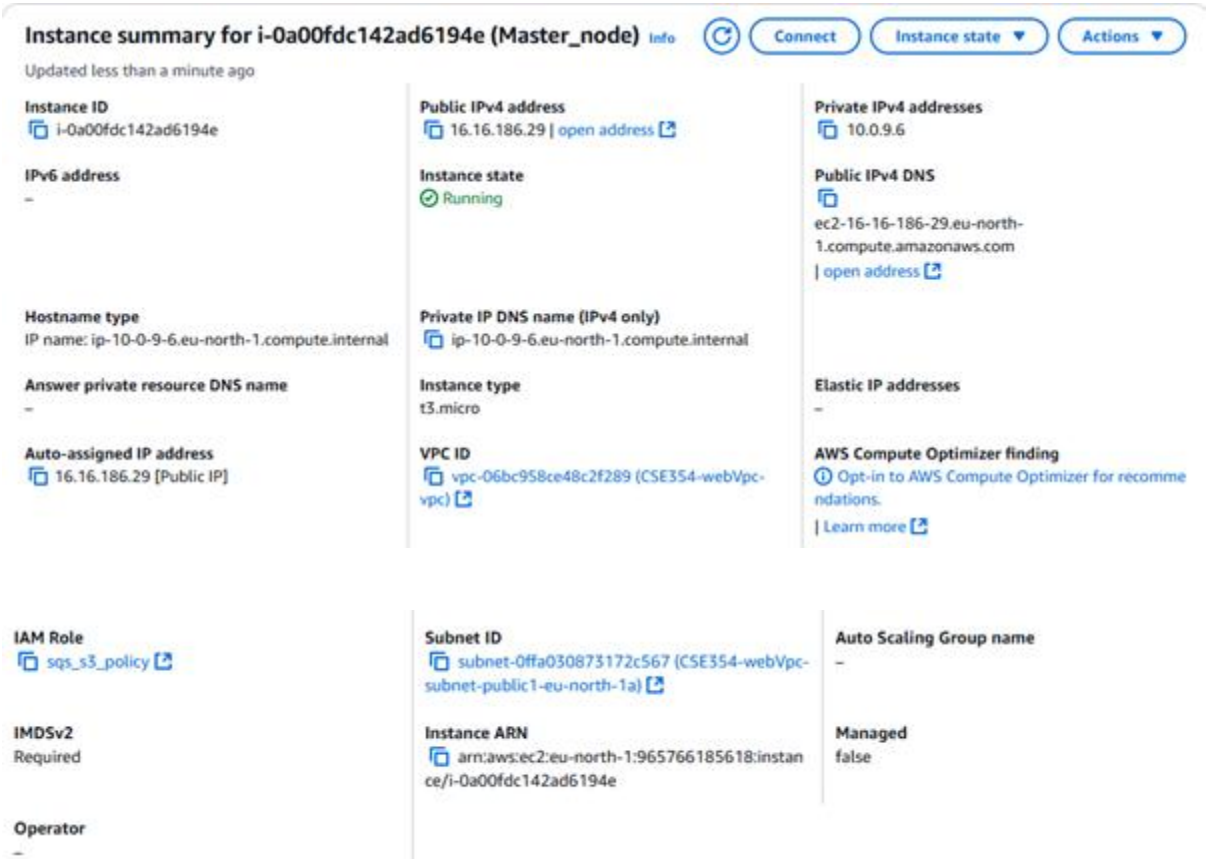
Step-by-step instructions for deploying the system

Configuration details for VMs

Multiple instances:



Master node instance



Crawler node instance

Instance summary for i-07c77b440bf487e80 (Crawler_node) Info

Connect | Instance state | Actions

Updated less than a minute ago

Instance ID i-07c77b440bf487e80	**Public IPv4 address** 13.61.151.104 | open address	**Private IPv4 addresses** 10.0.4.205
IPv6 address –	**Instance state** Running	**Public IPv4 DNS** ec2-13-61-151-104.eu-north-1.compute.amazonaws.com | open address
Hostname type IP name: ip-10-0-4-205.eu-north-1.compute.internal	**Private IP DNS name (IPv4 only)** ip-10-0-4-205.eu-north-1.compute.internal	
Answer private resource DNS name –	**Instance type** t3.micro	**Elastic IP addresses** –
Auto-assigned IP address 13.61.151.104 [Public IP]	**VPC ID** vpc-06bc958ce48c2f289 (CSE354-webVpc-vpc)	**AWS Compute Optimizer finding** Opt-in to AWS Compute Optimizer for recommendations. | Learn more

IAM Role sqs_s3_policy	**Subnet ID** subnet-0ffa030873172c567 (CSE354-webVpc-subnet-public1-eu-north-1a)	**Auto Scaling Group name** –
IMDSv2 Required	**Instance ARN** arn:aws:ec2:eu-north-1:965766185618:instance/i-07c77b440bf487e80	**Managed** false
Operator –		

Indexer node instance

Instance summary for i-053dad9d947317543 (Indexer_node) Info

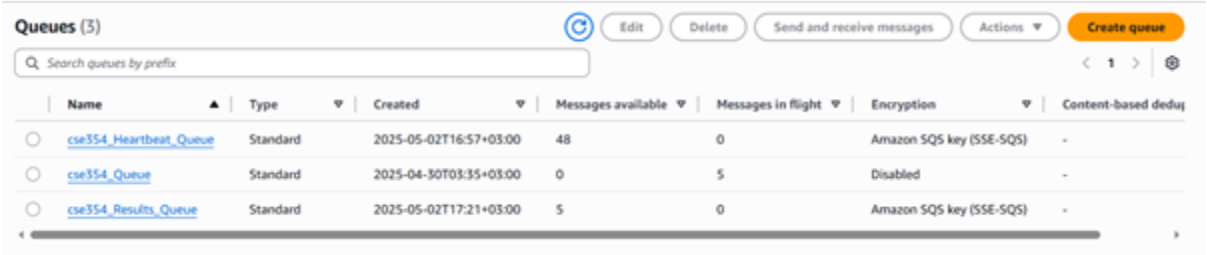
Connect | Instance state | Actions

Updated less than a minute ago

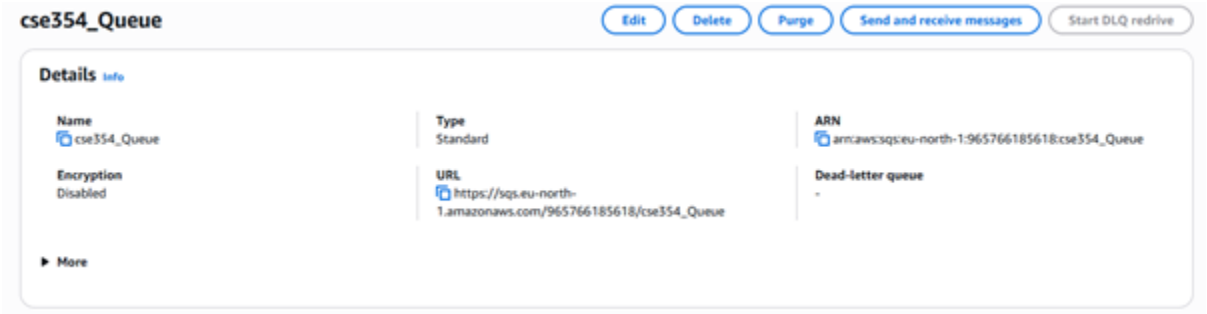
Instance ID i-053dad9d947317543	**Public IPv4 address** 51.20.89.88 | open address	**Private IPv4 addresses** 10.0.1.207
IPv6 address –	**Instance state** Running	**Public IPv4 DNS** ec2-51-20-89-88.eu-north-1.compute.amazonaws.com | open address
Hostname type IP name: ip-10-0-1-207.eu-north-1.compute.internal	**Private IP DNS name (IPv4 only)** ip-10-0-1-207.eu-north-1.compute.internal	
Answer private resource DNS name –	**Instance type** t3.micro	**Elastic IP addresses** –
Auto-assigned IP address 51.20.89.88 [Public IP]	**VPC ID** vpc-06bc958ce48c2f289 (CSE354-webVpc-vpc)	**AWS Compute Optimizer finding** Opt-in to AWS Compute Optimizer for recommendations. | Learn more

Task queues

Afterwards, there were 3 queues created to handle communication between Master, Indexer, and Crawler nodes. Each queue has a different communication role.

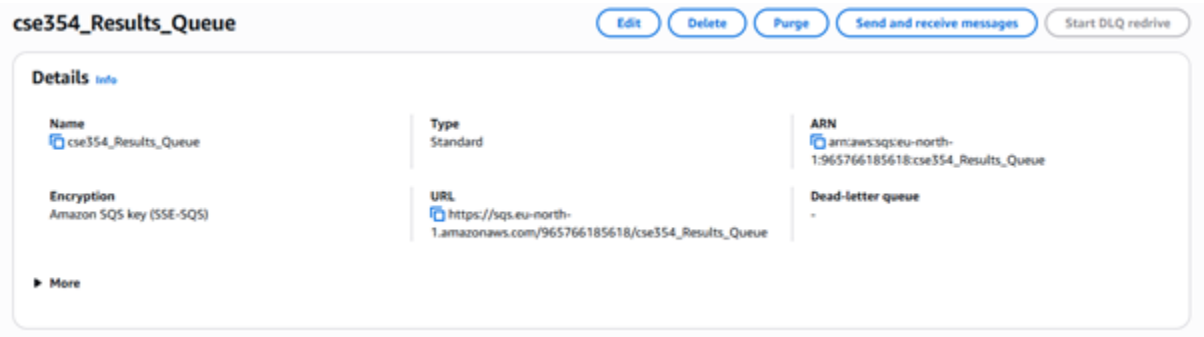


Cse354_Queue:



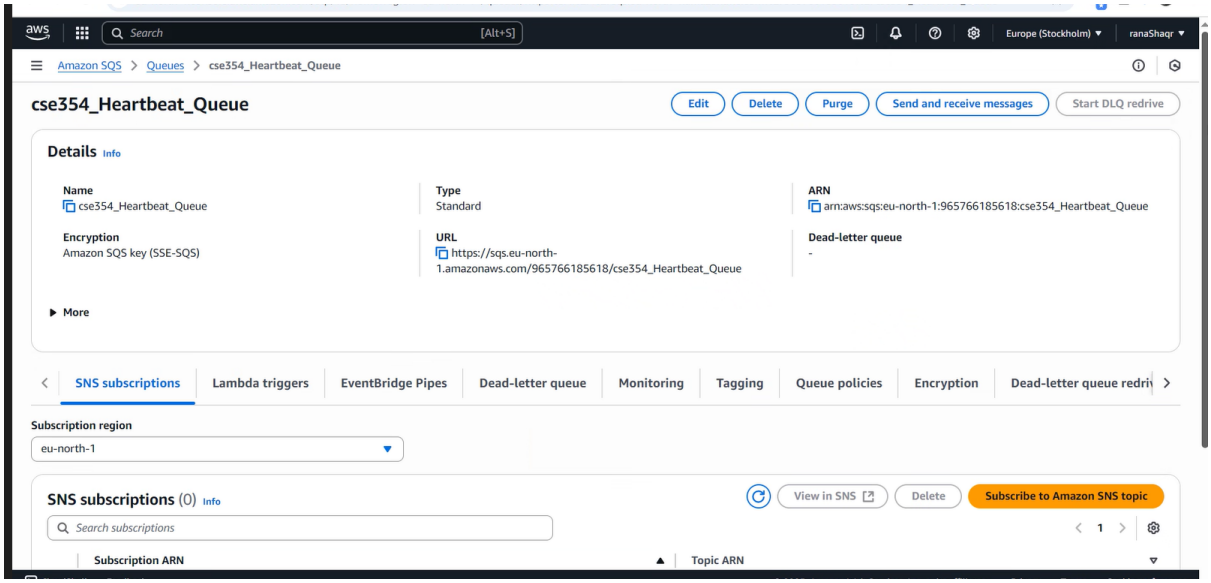
This queue is responsible for sending and receiving the URLs from the master node to the crawler node. After the crawler node successfully receives the URLs from the queue, the received URLs will be deleted automatically from the queue.

Cse354_Results_Queue



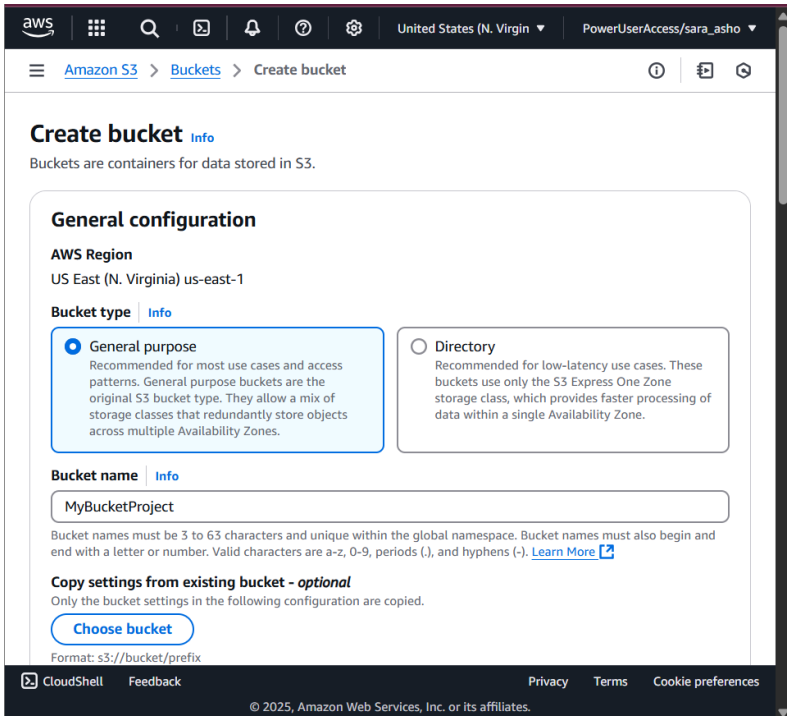
This queue is responsible for the communication between the crawler and the master where the master receives the URL mappings from the crawler and then deletes these mappings once received.

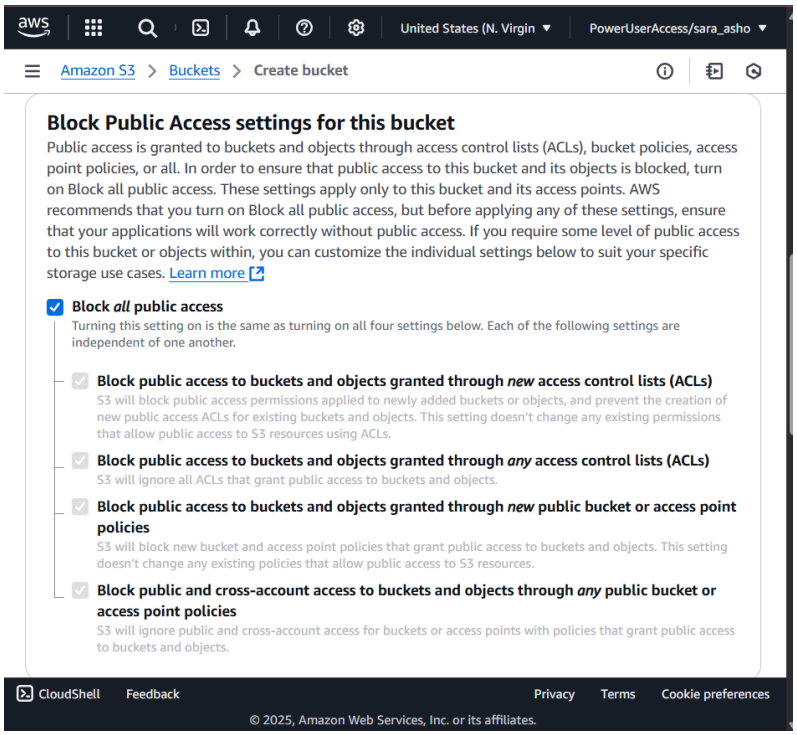
Cse354_heartbeat queue

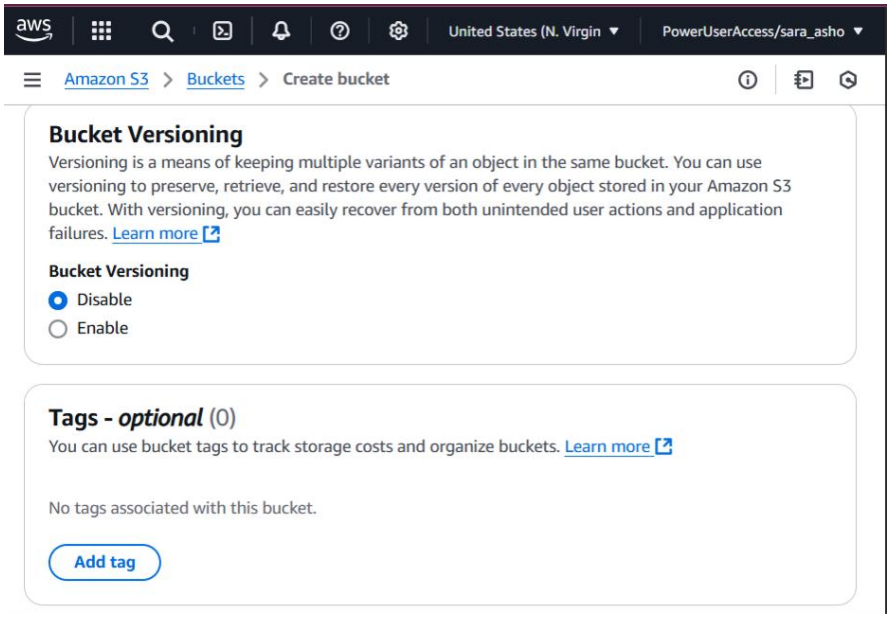


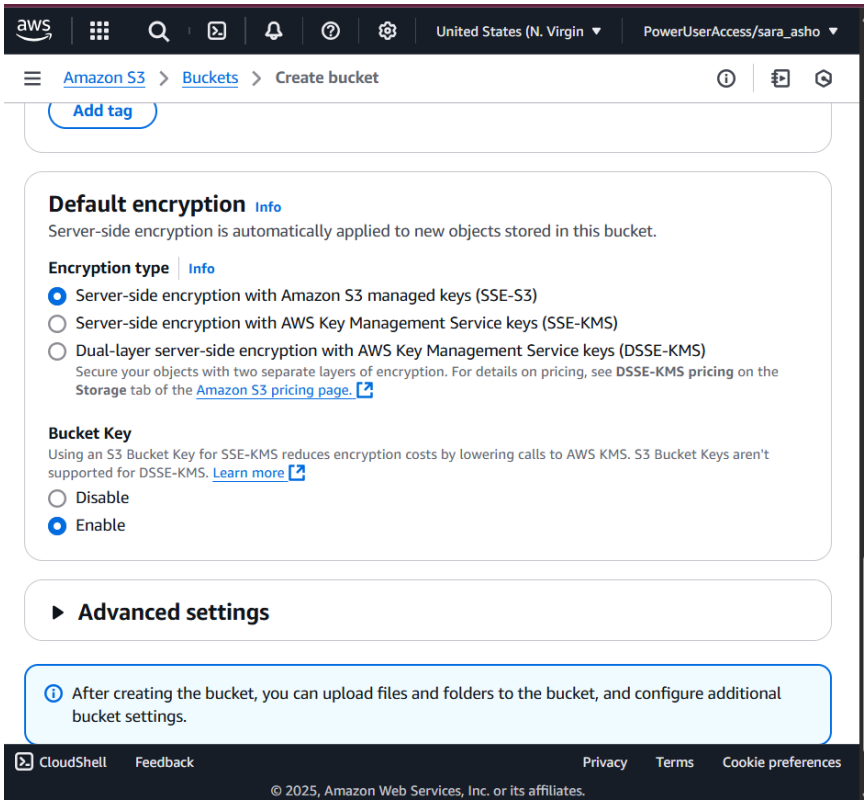
This queue is responsible for the communication between master the crawler and the master where the master receives the URL mappings from the crawler and then deletes these mappings once received.

storage services



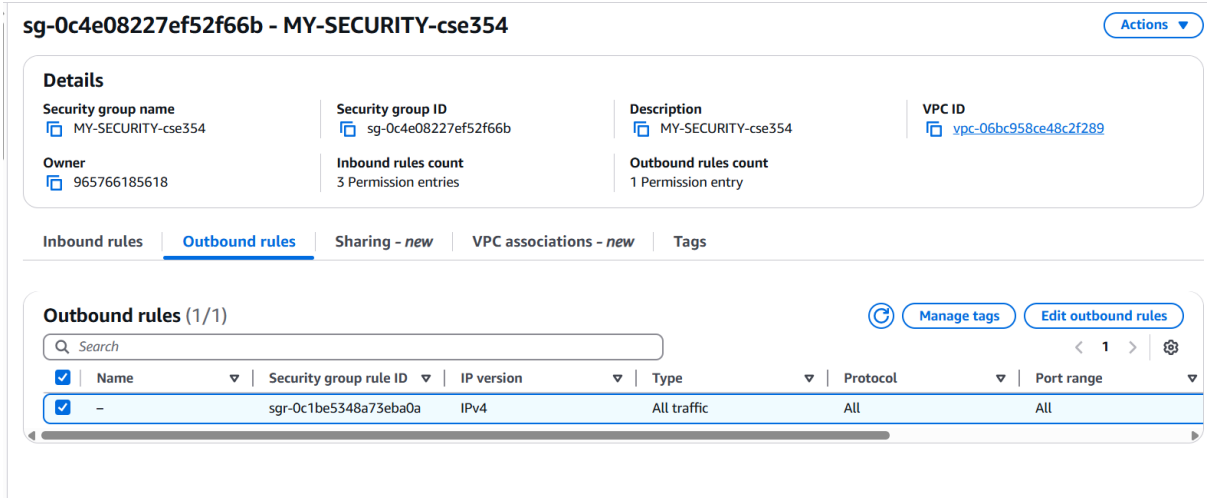


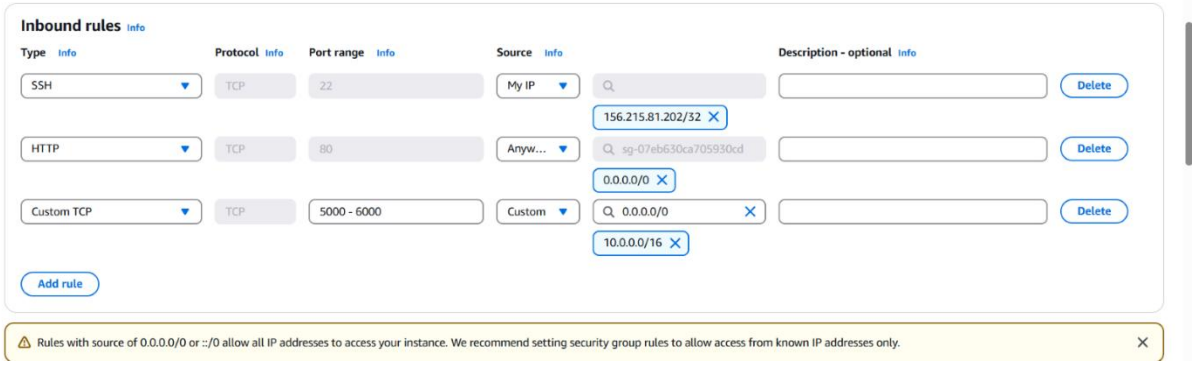

United States (N. Virgin
PowerUserAccess/sara_asho
Amazon S3 > Buckets > Create bucket
Bucket Versioning
Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. Learn more
Bucket Versioning
Disable
Enable
Tags - optional (0)
You can use bucket tags to track storage costs and organize buckets. Learn more
No tags associated with this bucket.
Add tag


United States (N. Virgin
PowerUserAccess/sara_asho
Amazon S3 > Buckets > Create bucket
Add tag
Default encryption Info
Server-side encryption is automatically applied to new objects stored in this bucket.
Encryption type Info
Server-side encryption with Amazon S3 managed keys (SSE-S3)
Server-side encryption with AWS Key Management Service keys (SSE-KMS)
Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)
Secure your objects with two separate layers of encryption. For details on pricing, see DSSE-KMS pricing on the Storage tab of the Amazon S3 pricing page.
Bucket Key
Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. Learn more
Disable
Enable
Advanced settings
After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.
CloudShell
Feedback
Privacy
Terms
Cookie preferences
© 2025, Amazon Web Services, Inc. or its affiliates.

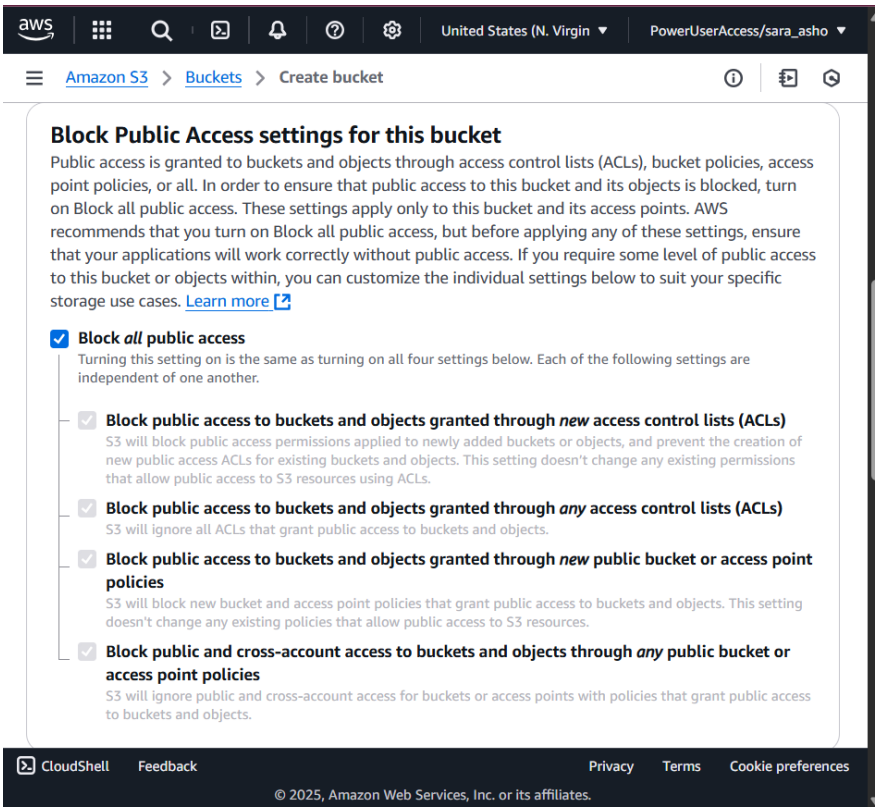
Security Review

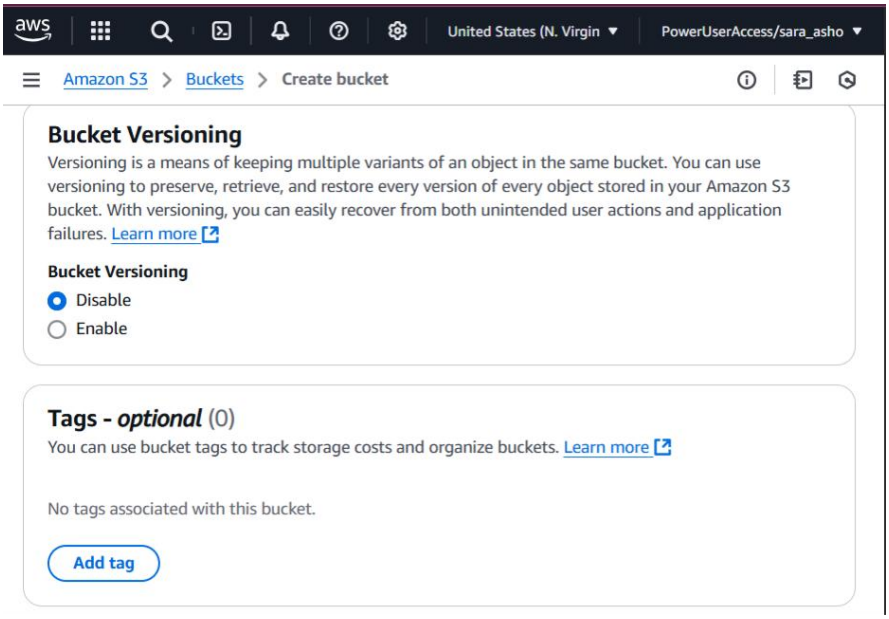
cloud resource access control

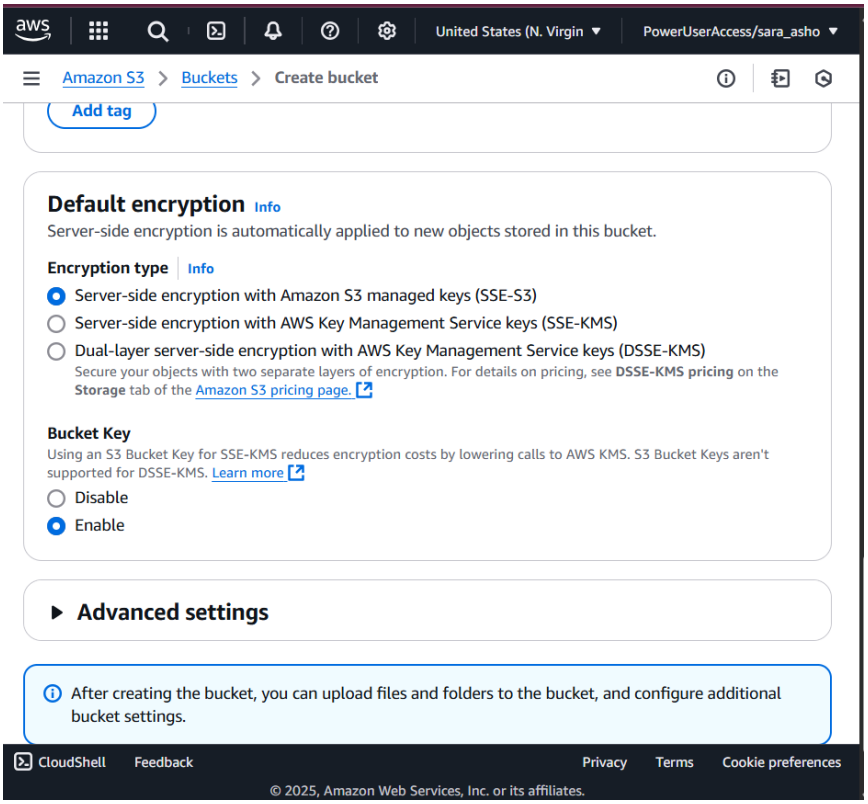




storage security considerations






United States (N. Virgin
PowerUserAccess/sara_asho
Amazon S3 > Buckets > Create bucket
Add tag
Default encryption Info
Server-side encryption is automatically applied to new objects stored in this bucket.
Encryption type Info
Server-side encryption with Amazon S3 managed keys (SSE-S3)
Server-side encryption with AWS Key Management Service keys (SSE-KMS)
Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)
Secure your objects with two separate layers of encryption. For details on pricing, see DSSE-KMS pricing on the Storage tab of the Amazon S3 pricing page.
Bucket Key
Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. Learn more
Disable
Enable
Advanced settings
After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.
CloudShell
Feedback
Privacy
Terms
Cookie preferences
© 2025, Amazon Web Services, Inc. or its affiliates.

Code Documentation

Master.py

```
#!/usr/bin/env python3
import sys
import logging
import boto3
import json
import time
import argparse
from datetime import datetime
```

- **Purpose**: Imports necessary Python libraries.
- **Details**:
  - sys: For system-level operations (e.g., exiting on errors).
  - logging: For logging messages to the console.
  - boto3: AWS SDK for interacting with SQS and S3.
  - json: For serializing/deserializing task and result messages.
  - time: For sleep intervals during polling.
  - argparse: For parsing command-line arguments.
  - datetime: For tracking crawler heartbeats.

```
def setup_logging():
    log_formatter = logging.Formatter('%(asctime)s [%(levelname)s] [Master] - %(message)s')
    console_handler = logging.StreamHandler()
    console_handler.setFormatter(log_formatter)
    logging.basicConfig(level=logging.INFO, handlers=[console_handler])
```

- **Purpose**: Configures logging to output formatted messages to the console.
- **Details**:
  - Creates a formatter with timestamp, log level (e.g., INFO, ERROR), and a [Master] prefix.
  - Sets up a console handler to print logs.
  - Configures the root logger to INFO level, ensuring all INFO and higher (WARNING, ERROR) messages are logged.
- **Example Output**: 2025-05-11 10:00:00 [INFO] [Master] - Master Node starting...

```
sqs_client = boto3.client('sqs', region_name='eu-north-1')
s3_client = boto3.client('s3')
TASK_QUEUE_URL = 'https://sqs.eu-north-1.amazonaws.com/965766185618/ces354_Queue'
RESULTS_QUEUE_URL = 'https://sqs.eu-north-1.amazonaws.com/965766185618/cse354_Results_Queue'
HEARTBEAT_QUEUE_URL = 'https://sqs.eu-north-1.amazonaws.com/965766185618/cse354_Heartbeat_Queue'
BUCKET_NAME = 'cse354000-bucket'
```

- **Purpose**: Initializes AWS clients and defines queue/bucket identifiers.

- **Details**:
  - Creates SQS and S3 clients for the eu-north-1 region (Stockholm).
  - Defines three SQS queues:
    - TASK_QUEUE_URL: For sending URLs to crawlers.
    - RESULTS_QUEUE_URL: For receiving crawler results and termination signals.
    - HEARTBEAT_QUEUE_URL: For monitoring crawler health.
  - Specifies an S3 bucket (cse354000-bucket) for seed URLs and result storage.

```
def get_queue_attributes(queue_url):
    try:
        response = sqs_client.get_queue_attributes(
            QueueUrl=queue_url,
            AttributeNames=['ApproximateNumberOfMessages']
        )
        return int(response['Attributes']['ApproximateNumberOfMessages'])
    except Exception as e:
        logging.error(f"Error getting queue attributes: {e}")
        return -1
```

- **Purpose**: Retrieves the approximate number of messages in an SQS queue.
- **Details**:
  - Calls get_queue_attributes with the specified queue_url.
  - Requests the ApproximateNumberOfMessages attribute.
  - Converts the response to an integer and returns it.
  - On error (e.g., network issues, invalid queue), logs the error and returns -1.
- **Usage**: Used to check if the task queue is empty during monitoring.

```
def master_node(num_crawlers, max_depth):
    logging.info("Master Node starting...")
```

- **Purpose**: Main function coordinating the crawler system.
- **Parameters**:
  - num_crawlers: Number of crawler processes (e.g., 2).
  - max_depth: Maximum crawl depth (e.g., 2).

```
    base_wait_time = 20
    base_sleep_time = 3
    wait_time = min(20, base_wait_time + 5 * max_depth)
    sleep_time = base_sleep_time + 2 * max_depth
    MAX_EMPTY_POLLS = 7
    logging.info(f"Using WaitTimeSeconds={wait_time}s, sleep_time={sleep_time}s, MAX_EMPTY_POLLS={MA
X_EMPTY_POLLS} for max_depth={max_depth}")
```

- **Purpose**: Configures polling and sleep intervals based on max_depth.
- **Details**:
  - base_wait_time: Base SQS long-polling wait time (20s).
  - base_sleep_time: Base sleep interval between loops (3s).
  - wait_time: Increases by 5s per depth level, capped at 20s (SQS max).

    - sleep_time: Increases by 2s per depth level for deeper crawls.
    - MAX_EMPTY_POLLS: Terminates crawlers after 7 consecutive empty task queue polls.
    - Logs the configuration for debugging.
- **Example**: For max_depth=2, wait_time=20s, sleep_time=7s.

```
urls_s3_path = 'seed_urls/seed_urls.txt'
local_urls_file = 'seed_urls.txt'
try:
    s3_client.download_file(BUCKET_NAME, urls_s3_path, local_urls_file)
except Exception as e:
    logging.error(f"Error downloading seed URL file: {e}")
    sys.exit(1)
```

- **Purpose**: Downloads the seed URL file from S3.
- **Details**:
    - Specifies the S3 path (seed_urls/seed_urls.txt) and local file (seed_urls.txt).
    - Uses s3_client.download_file to retrieve the file.
    - On error (e.g., file not found, permission issues), logs the error and exits.

```
with open(local_urls_file, 'r') as f:
    urls = [line.strip() for line in f if line.strip()]
if not urls:
    logging.error("Error: No URLs found in seed file.")
    sys.exit(1)
```

- **Purpose**: Reads and cleans URLs from the seed file.
- **Details**:
    - Opens the local file and strips whitespace from each line, ignoring empty lines.
    - Stores URLs in a list (urls).
    - If the list is empty, logs an error and exits.

```
# Initialize crawler tracking
active_crawlers = {f"crawler{i}": {'last_heartbeat': datetime.now(), 'assigned_urls': {}} for i in range(num_crawlers)}
failed_crawlers = {}  # Track crawlers that missed heartbeats but not yet terminated
completed_crawlers = 0
url_mapping = {}
```

- **Purpose**: Initializes data structures to track crawlers and results.
- **Details**:
    - active_crawlers: Dictionary mapping crawler IDs (e.g., crawler0, crawler1) to their last heartbeat time and assigned URLs.
    - failed_crawlers: Dictionary for crawlers that missed heartbeats but are not yet terminated.
    - completed_crawlers: Counter for terminated crawlers.

- url_mapping: Dictionary to store final URL mappings (e.g., {original_url: crawled_url}).

```
for url in urls:
    if url == 'https://example.com' and 'crawler0' in active_crawlers:
        active_crawlers['crawler0']['assigned_urls'][url] = {'depth': 0}
        sqs_client.send_message(
            QueueUrl=TASK_QUEUE_URL,
            MessageBody=json.dumps({'url': url, 'depth': 0, 'crawler_id': 'crawler0'})
        )
        logging.info(f"Pre-assigned URL {url} at depth 0 to crawler crawler0")
    else:
        sqs_client.send_message(
            QueueUrl=TASK_QUEUE_URL,
            MessageBody=json.dumps({'url': url, 'depth': 0})
        )
        logging.info(f"Added seed URL to queue: {url} at depth 0")
logging.info(f"Master added {len(urls)} URLs to SQS task queue with depth 0")
```

- **Purpose**: Queues seed URLs to the SQS task queue.
- **Details**:
  - Iterates through URLs.
  - For https://example.com, assigns it to crawler0 (if crawler0 exists) by:
    - Adding it to active_crawlers['crawler0']['assigned_urls'].
    - Sending an SQS message with url, depth=0, and crawler_id=crawler0.
  - For other URLs, sends an SQS message with url and depth=0 (no specific crawler).
  - Logs each action and the total number of URLs queued.

```
logging.info(f"Master added {len(urls)} URLs to SQS task queue with depth 0")

# Monitor task queue, heartbeats, and results
empty_polls = 0
REASSIGN_TIMEOUT = 60  # Reassign URLs after 60 seconds
TERMINATE_TIMEOUT = 120  # Terminate crawler after 120 seconds
```

- **Purpose**: Main loop to monitor crawlers, tasks, and results until all crawlers complete.
- **Details**:
  - empty_polls: Counts consecutive empty task queue polls.
  - REASSIGN_TIMEOUT: Reassigns URLs after 60s of no heartbeat.
  - TERMINATE_TIMEOUT: Terminates crawlers after 120s of no heartbeat.
  - Loop continues until completed_crawlers equals num_crawlers.

```
        current_time = datetime.now()
        for crawler_id, info in list(active_crawlers.items()):
            time_since_heartbeat = (current_time - info['last_heartbeat']).total_seconds()
            if time_since_heartbeat > REASSIGN_TIMEOUT:
                logging.warning(f"Crawler {crawler_id} missed heartbeat for {time_since_heartbeat:.2
f}s, reassigning URLs")
                for url, task in info['assigned_urls'].items():
                    if url not in url_mapping.values():
                        logging.info(f"Reassigning unprocessed URL {url} from crawler {crawler_id}")
                        sqs_client.send_message(
                            QueueUrl=TASK_QUEUE_URL,
                            MessageBody=json.dumps({'url': url, 'depth': task['depth']})
                        )
                failed_crawlers[crawler_id] = {
                    'last_heartbeat': info['last_heartbeat'],
                    'assigned_urls': info['assigned_urls'].copy()
                }
                del active_crawlers[crawler_id]
                logging.info(f"Moved {crawler_id} to failed crawlers, awaiting termination timeout")
```

- **Purpose**: Checks for missed heartbeats and reassigns tasks from unresponsive crawlers.
- **Details**:
  - Calculates time since the last heartbeat for each active crawler.
  - If time_since_heartbeat > 60s, reassigns unprocessed URLs (not in url_mapping) to the task queue.
  - Moves the crawler to failed_crawlers with its last heartbeat and assigned URLs.
  - Logs the reassignment and state change.

```
        for crawler_id, info in list(failed_crawlers.items()):
            time_since_heartbeat = (current_time - info['last_heartbeat']).total_seconds()
            if time_since_heartbeat > TERMINATE_TIMEOUT:
                logging.warning(f"Crawler {crawler_id} failed (no heartbeat for {time_since_heartbea
t:.2f}s), terminating")
                sqs_client.send_message(
                    QueueUrl=TASK_QUEUE_URL,
                    MessageBody=json.dumps({'terminate': True})
                )
                logging.info(f"Sent termination signal for failed crawler {crawler_id}")
                del failed_crawlers[crawler_id]
                completed_crawlers += 1
                logging.info(f"Crawler {crawler_id} terminated. Total completed: {completed_crawler
s}/{num_crawlers}")
```

- **Purpose**: Terminates crawlers that remain unresponsive.
- **Details**:
  - Checks failed_crawlers for heartbeats older than 120s.
  - Sends a termination signal ({'terminate': True}) to the task queue.
  - Removes the crawler from failed_crawlers, increments completed_crawlers, and logs the action.

```
        heartbeat_response = sqs_client.receive_message(
            QueueUrl=HEARTBEAT_QUEUE_URL,
            MaxNumberOfMessages=10,
            WaitTimeSeconds=20
        )
        if 'Messages' in heartbeat_response:
            logging.info(f"Received {len(heartbeat_response['Messages'])} heartbeats")
            for message in heartbeat_response['Messages']:
                body = json.loads(message['Body'])
                crawler_id = body['crawler_id']
                if crawler_id in active_crawlers:
                    active_crawlers[crawler_id]['last_heartbeat'] = datetime.now()
                    logging.info(f"Received heartbeat from active crawler {crawler_id}")
                elif crawler_id in failed_crawlers:
                    logging.info(f"Crawler {crawler_id} recovered with heartbeat, moving back to act
ive")
```

- ```
                    active_crawlers[crawler_id] = {
                        'last_heartbeat': datetime.now(),
                        'assigned_urls': failed_crawlers[crawler_id]['assigned_urls']
                    }
                    del failed_crawlers[crawler_id]
                sqs_client.delete_message(
                    QueueUrl=HEARTBEAT_QUEUE_URL,
                    ReceiptHandle=message['ReceiptHandle']
                )
  ```

  Purpose: Processes crawler heartbeats to update their status.
- **Details**:
  - Polls the heartbeat queue for up to 10 messages with a 20s wait (long polling).
  - For each message:
    - Parses the JSON body to get crawler_id.
    - If the crawler is active, updates its last_heartbeat.
    - If the crawler is failed, moves it back to active_crawlers (recovery).
    - Deletes the message to prevent reprocessing.
  - Logs the number of heartbeats and actions taken.

```
    # Check task queue
    num_messages = get_queue_attributes(TASK_QUEUE_URL)
    if num_messages == -1:
        logging.warning("Failed to check queue size, continuing to poll...")
    elif num_messages == 0:
        if active_crawlers or failed_crawlers:
            empty_polls += 1
            logging.info(f"Task queue empty, empty poll count: {empty_polls}/{MAX_EMPTY_POLLS}")
            if empty_polls >= MAX_EMPTY_POLLS:
                # Send termination signals to active crawlers only
                num_signals = len(active_crawlers)
                for crawler_id in list(active_crawlers.keys()):
                    sqs_client.send_message(
                        QueueUrl=TASK_QUEUE_URL,
                        MessageBody=json.dumps({'terminate': True})
                    )
                logging.info(f"Sent {num_signals} termination signals to active crawlers")
                # Wait for all crawlers to confirm termination
                while active_crawlers or failed_crawlers:
                    result_response = sqs_client.receive_message(
                        QueueUrl=RESULTS_QUEUE_URL,
                        MaxNumberOfMessages=10,
                        WaitTimeSeconds=20
                    )
```
```

```
                    sqs_client.delete_message(
                        QueueUrl=RESULTS_QUEUE_URL,
                        ReceiptHandle=message['ReceiptHandle']
                    )
                time.sleep(1)
            break  # Exit the main loop after all crawlers confirm termination
    else:
        empty_polls = 0
        logging.info(f"Task queue has ~{num_messages} messages, continuing to monitor...")
```

- **Details**:
- Checks the task queue size using get_queue_attributes.
- If num_messages == -1, logs a warning and continues.
- If num_messages == 0 and crawlers are active or failed:
    - Increments empty_polls.
    - If empty_polls >= 7, sends termination signals to active crawlers.
    - Enters a sub-loop to wait for termination confirmations via the results queue:
        - Processes termination messages, marking crawlers as completed.
        - Processes URL mappings, updating url_mapping and removing processed URLs.
        - Deletes processed messages.
    - Breaks the main loop after all crawlers confirm termination.
- If num_messages > 0, resets empty_polls and continues monitoring.

```
    # Receive results
    result_response = sqs_client.receive_message(
        QueueUrl=RESULTS_QUEUE_URL,
        MaxNumberOfMessages=10,
        WaitTimeSeconds=wait_time
    )
    if 'Messages' in result_response:
        for message in result_response['Messages']:
            body = json.loads(message['Body'])
            if 'terminate' in body:
                crawler_id = body.get('crawler_id', 'unknown')
                if crawler_id in active_crawlers:
                    completed_crawlers += 1
                    del active_crawlers[crawler_id]
                    logging.info(f"Crawler {crawler_id} completed. Total completed: {completed_crawlers}/{num_crawlers}")
                elif crawler_id in failed_crawlers:
                    completed_crawlers += 1
                    del failed_crawlers[crawler_id]
                    logging.info(f"Crawler {crawler_id} completed. Total completed: {completed_crawlers}/{num_crawlers}")
            else:
```

```
                else:
                    crawler_id = body.get('crawler_id', 'unknown')
                    mappings = body.get('mappings', {})
                    if mappings:
                        url_mapping.update(mappings)
                        logging.info(f"Received {len(mappings)} mappings from crawler {crawler_id}: {mappings}")
                    if crawler_id in active_crawlers:
                        for url in mappings.values():
                            if url in active_crawlers[crawler_id]['assigned_urls']:
                                logging.info(f"Removing processed URL {url} from assigned_urls[{crawler_id}]")
                                del active_crawlers[crawler_id]['assigned_urls'][url]
                    elif crawler_id in failed_crawlers:
                        for url in mappings.values():
                            if url in failed_crawlers[crawler_id]['assigned_urls']:
                                logging.info(f"Removing processed URL {url} from assigned_urls[{crawler_id}]")
                                del failed_crawlers[crawler_id]['assigned_urls'][url]
                    sqs_client.delete_message(
                        QueueUrl=RESULTS_QUEUE_URL,
                        ReceiptHandle=message['ReceiptHandle']
                    )
```

- **Purpose**: Processes crawler results (mappings or termination signals).
- **Details**:
    - Polls the results queue for up to 10 messages with wait_time seconds.

- For each message:
  - If it contains 'terminate', marks the crawler as completed and removes it from active_crawlers or failed_crawlers.
  - Otherwise, extracts mappings (e.g., {original_url: crawled_url}), updates url_mapping, and removes processed URLs from the crawler's assigned list.
  - Deletes the message to prevent reprocessing.
- Logs all actions.

```
# Assign tasks
task_response = sqs_client.receive_message(
    QueueUrl=TASK_QUEUE_URL,
    MaxNumberOfMessages=10,
    WaitTimeSeconds=1  # Fixed to integer
)
if 'Messages' in task_response:
    for message in task_response['Messages']:
        body = json.loads(message['Body'])
        if 'url' in body and 'depth' in body:
            url = body['url']
            depth = body['depth']
            target_crawler = body.get('crawler_id')
            if target_crawler and target_crawler in active_crawlers:
                # Respect pre-assigned crawler
                crawler_id = target_crawler
            elif active_crawlers:
                # Choose crawler with fewest assigned URLs
                crawler_id = min(
                    active_crawlers,
                    key=lambda cid: len(active_crawlers[cid]['assigned_urls'])
                )
            else:
                # No active crawlers, re-queue
                sqs_client.send_message(
                    QueueUrl=TASK_QUEUE_URL,
                    MessageBody=json.dumps({'url': url, 'depth': depth})
                )
                sqs_client.delete_message(
                    QueueUrl=TASK_QUEUE_URL,
                    ReceiptHandle=message['ReceiptHandle']
                )
                continue
```

```
                continue
            active_crawlers[crawler_id]['assigned_urls'][url] = {'depth': depth}
            logging.info(f"Assigned URL {url} at depth {depth} to crawler {crawler_id}")
            # Re-queue for the crawler
            sqs_client.send_message(
                QueueUrl=TASK_QUEUE_URL,
                MessageBody=json.dumps({'url': url, 'depth': depth, 'crawler_id': crawler_id})
            )
        sqs_client.delete_message(
            QueueUrl=TASK_QUEUE_URL,
            ReceiptHandle=message['ReceiptHandle']
        )
```

- **Purpose**: Assigns tasks to crawlers.
- **Details**:
  - Polls the task queue for up to 10 messages with a 1s wait.
  - For each message containing url and depth:
    - If crawler_id is specified and the crawler is active, assigns to that crawler.
    - Otherwise, assigns to the crawler with the fewest assigned URLs.
    - If no active crawlers, re-queues the task and deletes the message.
    - Updates active_crawlers[crawler_id]['assigned_urls'] and re-queues the task with the assigned crawler_id.
    - Deletes the original message.

- Logs the assignment.

```
    time.sleep(sleep_time)
```

- **Purpose**: Pauses the loop to avoid excessive polling.
- **Details**: Sleeps for sleep_time seconds (e.g., 7s for max_depth=2).

```
mapping_file = 'url_mapping.json'
with open(mapping_file, 'w') as f:
    json.dump(url_mapping, f)
s3_client.upload_file(mapping_file, BUCKET_NAME, 'crawl_data/url_mapping.json')
logging.info(f"Uploaded combined URL mapping to S3 with {len(url_mapping)} entries")
```

- **Purpose**: Saves the final URL mappings to S3.
- **Details**:
  - Writes url_mapping to a local file (url_mapping.json).
  - Uploads the file to S3 at crawl_data/url_mapping.json.
  - Logs the number of mappings uploaded.

```
# Save combined mapping to S3
mapping_file = 'url_mapping.json'
with open(mapping_file, 'w') as f:
    json.dump(url_mapping, f)
s3_client.upload_file(mapping_file, BUCKET_NAME, 'crawl_data/url_mapping.json')
logging.info(f"Uploaded combined URL mapping to S3 with {len(url_mapping)} entries")

# Signal indexer and monitor completion
sqs_client.send_message(
    QueueUrl=TASK_QUEUE_URL,
    MessageBody=json.dumps({'start_indexer': True})
)
logging.info("Master signaled Indexer to start.")

indexer_completed = False
while not indexer_completed:
    response = sqs_client.receive_message(
        QueueUrl=RESULTS_QUEUE_URL,
        MaxNumberOfMessages=1,
        WaitTimeSeconds=wait_time
    )
    if 'Messages' in response:
        for message in response['Messages']:
            body = json.loads(message['Body'])
            if 'indexer_complete' in body:
                indexer_completed = True
                logging.info("Indexer reported completion")
                sqs_client.delete_message(
                    QueueUrl=RESULTS_QUEUE_URL,
                    ReceiptHandle=message['ReceiptHandle']
                )
    else:
        logging.info(f"No messages received, sleeping for {sleep_time}s")
        time.sleep(sleep_time)

logging.info("Master process completed")
```

- **Purpose**: Signals the indexer and waits for completion.
- **Details**:
  - Sends a start_indexer message to the task queue.
  - Polls the results queue for an indexer_complete message.
  - Deletes the completion message and exits the loop.
  - If no messages are received, logs and sleeps for sleep_time.

```python
if __name__ == "__main__":
    parser = argparse.ArgumentParser(description="Master Node")
    parser.add_argument('--num-crawlers', type=int, default=2, help="Number of crawler processes")
    parser.add_argument('--max-depth', type=int, default=2, help="Maximum crawl depth")
    args = parser.parse_args()
    setup_logging()
    master_node(args.num_crawlers, args.max_depth)
```

- **Purpose**: Parses arguments and starts the master node.
- **Details**:
    - Defines arguments: --num-crawlers (default: 2), --max-depth (default: 2).
    - Calls setup_logging and master_node with parsed arguments.

parser.add_argument('--num-crawlers', type=int, default=2, help="Number of crawler processes")

- **Purpose**: This defines a command-line argument named --num-crawlers.
    - **--num-crawlers**: The argument name (prefixed with -- indicates an optional argument). Users can specify it as --num-crawlers 3 when running the script.
    - **type=int**: Ensures the argument is interpreted as an integer.
    - **default=2**: Sets the default value to 2 if the argument isn't provided.
    - **help="Number of crawler processes"**: Provides a description that appears in the help text, explaining that this argument controls the number of crawler processes to spawn.
- **Effect**: Allows the user to specify how many crawler processes (e.g., 2, 3, 4) the master node should manage. If not specified, it defaults to 2.

parser.add_argument('--max-depth', type=int, default=2, help="Maximum crawl depth")

- **Purpose**: This defines another command-line argument named --max-depth.
    - **--max-depth**: Users can specify it as --max-depth 3.
    - **type=int**: Ensures it's an integer.
    - **default=2**: Defaults to 2 if omitted.
    - **help="Maximum crawl depth"**: Describes that this argument sets the maximum depth for crawling (e.g., how many levels of links to follow from the seed URLs).

Crawler.py

```
#!/usr/bin/env python3
import boto3
import requests
from bs4 import BeautifulSoup
import tempfile
import os
import logging
import sys
import time
import json
import re
import argparse
from multiprocessing import Process, Value, Manager
from logging.handlers import QueueHandler
import multiprocessing as mp
from urllib.parse import urlparse
import urllib.robotparser
import threading
```

Libraries:

- boto3: AWS SDK for interacting with SQS and S3.
- requests: HTTP requests to fetch web pages.
- BeautifulSoup: Parses HTML to extract text and links.
- tempfile, os: Manage temporary files for storing HTML/text before uploading to S3.
- logging, QueueHandler: Custom logging for multi-process environments.
- multiprocessing, Manager, Value: Parallel processing and shared state.
- urllib.robotparser: Parses robots.txt to respect crawling rules.
- threading: Used for heartbeat signals.
- argparse: Command-line argument parsing.

```
# Crawl Delay
DEFAULT_CRAWL_DELAY = 2  # seconds
HEARTBEAT_INTERVAL = 5  # seconds

# Shared counter for unique IDs and visited URLs
crawler_counter = Value('i', 0)
manager = Manager()
visited_urls = manager.dict()  # Shared dictionary to track visited URLs
```

- **Purpose**: These constants define the timing behavior of the crawler.
- **DEFAULT_CRAWL_DELAY = 2**:
    - Represents the default time (in seconds) the crawler waits between fetching consecutive URLs from the same domain.
    - This delay helps respect web server load and is used if a domain's robots.txt file does not specify a crawl delay.
- **HEARTBEAT_INTERVAL = 5**:
    - Specifies the interval (in seconds) at which each crawler process sends a "heartbeat" message to an AWS SQS queue (HEARTBEAT_QUEUE_URL).
    - Heartbeats are used to monitor the health and activity of each crawler process, ensuring the system can detect if a crawler has failed or stopped.
- **Purpose**: These constants define the timing behavior of the crawler.
- **DEFAULT_CRAWL_DELAY = 2**:
    - Represents the default time (in seconds) the crawler waits between fetching consecutive URLs from the same domain.
    - This delay helps respect web server load and is used if a domain's robots.txt file does not specify a crawl delay.
- **HEARTBEAT_INTERVAL = 5**:
    - Specifies the interval (in seconds) at which each crawler process sends a "heartbeat" message to an AWS SQS queue (HEARTBEAT_QUEUE_URL).
    - Heartbeats are used to monitor the health and activity of each crawler process, ensuring the system can detect if a crawler has failed or stopped.
- **Purpose**: These lines set up shared state management across multiple processes using Python's multiprocessing module.
- **crawler_counter = Value('i', 0)**:
    - Creates a shared integer value (Value) initialized to 0, with 'i' indicating it's an integer type.

    - This counter is used to generate unique IDs for each crawler process (e.g., crawler0, crawler1, etc.) by incrementing it as new crawlers are spawned.
    - The get_lock() method is used to ensure thread-safe increments when assigning IDs.
- **manager = Manager()**:
    - Initializes a Manager object from the multiprocessing module.
    - The Manager provides a way to create shared objects (like dictionaries or lists) that can be safely accessed and modified by multiple processes.
- **visited_urls = manager.dict()**:
    - Creates a shared dictionary to track which URLs have been visited by any crawler process.
    - This prevents duplicate crawling across all processes, as each crawler checks this dictionary before processing a URL.
    - The dictionary uses URLs as keys and a boolean-like value (e.g., True) to mark them as visited.

```
class CrawlerIdFilter(logging.Filter):
    """Add crawler_id to log records"""
    def __init__(self, crawler_id):
        super().__init__()
        self.crawler_id = crawler_id

    def filter(self, record):
        record.crawler_id = self.crawler_id
        return True
```

- **Purpose**: Adds crawler_id to log records for identifying which crawler process generated a log.
- **How It Works**: Extends logging.Filter to attach the crawler's ID to each log message.

```python
def setup_logger(crawler_id, log_queue):
    """Set up logging for multiprocessing"""
    formatter = logging.Formatter(
        fmt="%(asctime)s [%(levelname)s] [Crawler %(crawler_id)s] %(message)s",
        datefmt="%Y-%m-%d %H:%M:%S"
    )
    handler = QueueHandler(log_queue)
    handler.setFormatter(formatter)
    handler.addFilter(CrawlerIdFilter(crawler_id))
    logger = logging.getLogger(f"crawler_{crawler_id}")
    logger.setLevel(logging.INFO)
    logger.addHandler(handler)
    logger.propagate = False
    return logger
```

- Purpose: Configures logging for each crawler process.
- Details
    - Uses QueueHandler to send logs to a shared queue (for multi-process safety).
    - Applies a formatter to include timestamp, log level, and crawler ID.
    - Adds CrawlerIdFilter to attach the crawler ID.
    - Disables propagation to prevent duplicate logs.

```python
def log_listener(log_queue):
    """Listen to log messages from all processes"""
    while True:
        try:
            record = log_queue.get()
            if record is None:  # Sentinel to stop listener
                break
            logger = logging.getLogger(record.name)
            logger.handle(record)
        except Exception as e:
            print(f"Log listener error: {e}", file=sys.stderr)
```

- **Purpose**: A separate process that listens to the log queue and processes log messages.
- **How It Works**:
    - Runs in a dedicated process to centralize logging.
    - Retrieves log records from the queue and forwards them to the appropriate logger.
    - Stops when a Nonsentinel is received.

```python
def send_heartbeat(crawler_id, stop_event):
    """Send heartbeat messages every 5 seconds"""
    while not stop_event.is_set():
        try:
            sqs_client.send_message(
                QueueUrl=HEARTBEAT_QUEUE_URL,
                MessageBody=json.dumps({'crawler_id': crawler_id})
            )
            logger.info(f"Sent heartbeat for {crawler_id}")
        except Exception as e:
            logger.error(f"Error sending heartbeat: {e}")
        time.sleep(HEARTBEAT_INTERVAL)
```

- **Purpose**: Sends periodic heartbeat messages to the HEARTBEAT_QUEUE_URL to indicate the crawler is active.
- **Details**:
    - Runs in a separate thread within each crawler process.
    - Sends a JSON message with the crawler_id every 5 seconds.
    - Stops when the stop_event is set (e.g., when the crawler terminates).
    - Logs errors if the heartbeat failsSends a JSON message with the crawler_id every 5 seconds.

```python
def sanitize_filename(url):
    """Convert URL to a safe S3 key name"""
    name = re.sub(r'https?://', '', url)
    name = re.sub(r'[^\w\-\.]', '', name)
    return name[:200]
```

- **Purpose**: Converts a URL into a safe S3 key name.
- **How It Works**:
    - Removes http:// or https:// prefixes.
    - Removes non-alphanumeric characters (except -, ., _).
    - Truncates to 200 characters to avoid S3 key length issues.

```python
def crawl_urls(url, depth, crawler_id, max_depth):
    """Crawl a single URL, return mappings, new links, and crawl delay, respecting robots.txt and depth"""
    mappings = {}
    new_links = []
    html_tmp_path = None
    txt_tmp_path = None

    try:
        parsed_url = urlparse(url)
        if not parsed_url.scheme or not parsed_url.netloc:
            logger.warning(f"Invalid URL format: {url}")
            return mappings, new_links, DEFAULT_CRAWL_DELAY
        base_url = f"{parsed_url.scheme}://{parsed_url.netloc}"
        robots_url = f"{base_url}/robots.txt"
```

```python
    logger.info(f"Checking robots.txt at {robots_url} for {url}")
    rp = urllib.robotparser.RobotFileParser()
    rp.set_url(robots_url)
    try:
        rp.read()
    except Exception as e:
        logger.warning(f"Failed to fetch robots.txt from {robots_url}: {e}. Proceeding with crawl.")

    user_agent = "MyCrawlerBot"
    if not rp.can_fetch(user_agent, url):
        logger.info(f"Crawling disallowed by robots.txt for {url}")
        return mappings, new_links, DEFAULT_CRAWL_DELAY

    crawl_delay = rp.crawl_delay(user_agent) or DEFAULT_CRAWL_DELAY
    logger.info(f"Fetching URL: {url} with crawl delay {crawl_delay}s")
    response = requests.get(url, timeout=5, headers={"User-Agent": user_agent})
    if response.status_code == 200:
        html_content = response.text
        base_name = sanitize_filename(url)
        html_s3_key = f"crawl_data/{base_name}_{crawler_id}.html"
        txt_s3_key = f"crawl_data/{base_name}_{crawler_id}.txt"
```

```
with tempfile.NamedTemporaryFile(delete=False, suffix=".html") as html_tmp:
    html_tmp.write(html_content.encode('utf-8'))
    html_tmp_path = html_tmp.name

if os.path.exists(html_tmp_path):
    s3_client.upload_file(html_tmp_path, BUCKET_NAME, html_s3_key)
    logger.info(f"Uploaded HTML: {html_s3_key}")
    mappings[html_s3_key] = url
else:
    logger.warning(f"Missing HTML file: {html_tmp_path}")

soup = BeautifulSoup(html_content, "html.parser")
text_content = soup.get_text(separator='\n')

with tempfile.NamedTemporaryFile(delete=False, suffix=".txt") as txt_tmp:
    txt_tmp.write(text_content.encode('utf-8'))
    txt_tmp_path = txt_tmp.name

if os.path.exists(txt_tmp_path):
    s3_client.upload_file(txt_tmp_path, BUCKET_NAME, txt_s3_key)
    logger.info(f"Uploaded TXT: {txt_s3_key}")
    mappings[txt_s3_key] = url
else:
    logger.warning(f"Missing TXT file: {txt_tmp_path}")

if depth < max_depth:
    for link in soup.find_all('a', href=True):
        href = link['href']
        full_url = urllib.parse.urljoin(url, href)
        parsed_full_url = urlparse(full_url)
        if parsed_full_url.scheme in ['http', 'https'] and not parsed_full_url.fragment:
            new_links.append(full_url)
            logger.info(f"Found link: {full_url} at depth {depth + 1}")


    else:
        logger.error(f"Failed to fetch {url}: Status code {response.status_code}")
    return mappings, new_links, crawl_delay
except Exception as e:
    logger.exception(f"Error crawling {url}: {e}")
    return mappings, new_links, DEFAULT_CRAWL_DELAY
finally:
    if html_tmp_path and os.path.exists(html_tmp_path):
        os.remove(html_tmp_path)
    if txt_tmp_path and os.path.exists(txt_tmp_path):
        os.remove(txt_tmp_path)
```

- P**urpose**: Crawls a single URL, stores HTML and text in S3, extracts links, and respects robots.txt.
- **Key Steps**:
-
    1. **URL Validation**:
        - Parses the URL to ensure it has a scheme (e.g., http, https) and a domain.
        - Returns empty results if the URL is invalid.
    2. Robots.txt Check
        - Constructs the robots.txt URL (e.g., https://example.com/robots.txt).
        - Parses robots.txt to check if crawling is allowed for MyCrawlerBot.
        - Retrieves the crawl delay (defaults to 2 seconds if not specified).
        - Skips the URL if disallowed by robots.txt.
    3. Fetch URL
        - Sends an HTTP GET request with a 5-second timeout.
        - Proceeds only if the response status is 200 (OK).
    4. Store HTML
        - Saves HTML content to a temporary file.
        - Uploads the file to S3 with a key like crawl_data/<sanitized_url>_<crawler_id>.html.
        - Adds the S3 key-to-URL mapping to mappings.
    5. Extract Text
        - Uses BeautifulSoup to extract text from HTML.
        - Saves text to a temporary file and uploads it to S3 with a key like crawl_data/<sanitized_url>_<crawler_id>.txt.
        - Adds the S3 key-to-URL mapping to mappings.
    6. Extract Link:
        - Extracts <a> tags with href attributes if the current depth is less than max_depth.
        - Resolves relative URLs to absolute URLs using urljoin.
        - Filters links to include only http or https schemes and excludes URLs with fragments (e.g., #section).
    7. Cleanup:
        - Deletes temporary files to avoid disk space issues.

    - **Returns**:
        - mappings: Dictionary mapping S3 keys (HTML and text) to the crawled URL.
        - new_links: List of URLs found in the page (for further crawling).
        - crawl_delay: Delay to respect between requests (from robots.txt or default).

```python
def crawler_worker(crawler_id, log_queue, max_depth):
    """Worker function for each crawler process"""
    global logger
    logger = setup_logger(crawler_id, log_queue)
    logger.info(f"Starting crawler with max_depth={max_depth}")

    stop_event = threading.Event()
    heartbeat_thread = threading.Thread(target=send_heartbeat, args=(crawler_id, stop_event))
    heartbeat_thread.start()

    processed_urls = set()  # Track URLs processed by this crawler to avoid duplicate mappings
    has_assigned_url = False  # Flag for crawler0 failure simulation


  try:
      while True:
          response = sqs_client.receive_message(
              QueueUrl=TASK_QUEUE_URL,
              MaxNumberOfMessages=1,
              WaitTimeSeconds=20
          )

          if 'Messages' not in response:
              logger.info("No messages received, continuing to poll...")
              time.sleep(1)
              continue


   message = response['Messages'][0]
   body = json.loads(message['Body'])

   if 'terminate' in body:
       logger.info("Received termination signal, sending mappings and stopping")
       sqs_client.send_message(
           QueueUrl=RESULTS_QUEUE_URL,
           MessageBody=json.dumps({'terminate': True, 'crawler_id': crawler_id})
       )
       sqs_client.delete_message(
           QueueUrl=TASK_QUEUE_URL,
           ReceiptHandle=message['ReceiptHandle']
       )
       break
```

```
if 'url' in body and 'depth' in body:
    url = body['url']
    depth = body['depth']
    target_crawler = body.get('crawler_id')
    if target_crawler and target_crawler != crawler_id:
        logger.info(f"Skipping URL {url} targeted for crawler {target_crawler}")
        sqs_client.delete_message(
            QueueUrl=TASK_QUEUE_URL,
            ReceiptHandle=message['ReceiptHandle']
        )
        continue
    logger.info(f"Received URL: {url} at depth {depth}")

 # Simulate failure for crawler0 after receiving a URL
 if crawler_id == "crawler0" and not has_assigned_url:
     has_assigned_url = True
     logger.info(f"Simulating failure: stopping heartbeats and exiting")
     stop_event.set()
     heartbeat_thread.join()
     sqs_client.delete_message(
         QueueUrl=TASK_QUEUE_URL,
         ReceiptHandle=message['ReceiptHandle']
     )
     sys.exit(1)  # Exit the process

 if depth > max_depth:
     logger.warning(f"Skipping URL {url} at depth {depth} (exceeds max_depth {max_depth})")
     sqs_client.delete_message(
         QueueUrl=TASK_QUEUE_URL,
         ReceiptHandle=message['ReceiptHandle']
     )
     continue
```

```
if url in visited_urls:
    logger.info(f"Skipping already visited URL: {url} at depth {depth}")
    sqs_client.delete_message(
        QueueUrl=TASK_QUEUE_URL,
        ReceiptHandle=message['ReceiptHandle']
    )
    continue

logger.info(f"Processing URL: {url} at depth {depth}/{max_depth}")
mappings, new_links, crawl_delay = crawl_urls(url, depth, crawler_id, max_depth)

# Mark URL as visited only after successful crawl
if mappings:
    visited_urls[url] = True


    # Send mappings only if not already processed
    if mappings and url not in processed_urls:
        sqs_client.send_message(
            QueueUrl=RESULTS_QUEUE_URL,
            MessageBody=json.dumps({'mappings': mappings, 'crawler_id': crawler_id})
        )
        logger.info(f"Sent {len(mappings)} mappings to results queue")
        processed_urls.add(url)

    if depth < max_depth:
        for link in new_links:
            if link not in visited_urls:
                sqs_client.send_message(
                    QueueUrl=TASK_QUEUE_URL,
                    MessageBody=json.dumps({'url': link, 'depth': depth + 1})
                )
                logger.info(f"Added new URL to queue: {link} at depth {depth + 1}")

sqs_client.delete_message(
    QueueUrl=TASK_QUEUE_URL,
    ReceiptHandle=message['ReceiptHandle']
)

        logger.info(f"Waiting for {crawl_delay} seconds before next request")
        time.sleep(crawl_delay)
finally:
    stop_event.set()
    heartbeat_thread.join()
```

- **Purpose**: Main logic for each crawler process, which polls the task queue, processes URLs, and sends results.

- **Key Steps**:

1. **Setup**:
    - Initializes the logger and starts a heartbeat thread.
    - Maintains a local processed_urls set to avoid duplicate mappings within the process.
    - Uses has_assigned_url to simulate failure for crawler0.
2. **Poll Task Queue**:
    - Polls the TASK_QUEUE_URL for up to 20 seconds.
    - If no messages are received, logs and continues polling.
3. Process Messages:
    - Handles termination messages by sending a termination signal to the results queue and exiting.
    - Processes URL tasks with a specified depth.
    - Skips URLs targeted for a different crawler (based on crawler_id in the message).
    - Simulates failure for crawler0 after receiving its first URL by stopping heartbeats and exiting.
    - Skips URLs that exceed max_depth or have already been visited (checked via the shared visited_urls dictionary).
4. Crawl and Process
    - Calls crawl_urls to fetch the URL, store data, and extract links.
    - Marks the URL as visited only if crawling was successful (mappings is non-empty).
    - Sends mappings to the RESULTS_QUEUE_URL if not already processed by this crawler.
5. Queue New Links
    - Queues new links for crawling at the next depth if depth < max_depth and the link hasn’t been visited.

```python
def run_crawlers(num_crawlers, max_depth):
    """Run multiple crawler processes"""
    log_queue = mp.Queue()
    listener = Process(target=log_listener, args=(log_queue,))
    listener.start()

    processes = []
    for i in range(num_crawlers):
        with crawler_counter.get_lock():
            crawler_id = f"crawler{crawler_counter.value}"
            crawler_counter.value += 1
        p = Process(target=crawler_worker, args=(crawler_id, log_queue, max_depth))
        p.start()
        processes.append(p)

    for p in processes:
        p.join()

    log_queue.put(None)
    listener.join()
```

```python
if __name__ == "__main__":
    parser = argparse.ArgumentParser(description="Multi-Process Crawler Node")
    parser.add_argument('--num-crawlers', type=int, default=2, help="Number of crawler processes")
    parser.add_argument('--max-depth', type=int, default=2, help="Maximum crawl depth")
    args = parser.parse_args()

    logging.basicConfig(
        level=logging.INFO,
        format="%(asctime)s [%(levelname)s] %(message)s",
        handlers=[logging.StreamHandler(sys.stdout)]
    )
    run_crawlers(args.num_crawlers, args.max_depth)
```

def run_crawlers(num_crawlers, max_depth):

- **Purpose**:
  This function is designed to manage and execute multiple crawler processes in parallel within the distributed web crawling system. It initializes a logging mechanism, spawns the specified number of crawler processes (num_crawlers), and runs them with a defined maximum crawl depth (max_depth). The function ensures that all crawlers complete their tasks and properly terminates the logging listener.
- **Key Steps**:

  Setup:

  - Creates a mp.Queue() named log_queue to handle logging messages across multiple processes.
  - Initializes a Process called listener to run the log_listener function with log_queue as an argument, starting it to collect logs from all crawlers.
  - Initializes an empty list processes to store the crawler process objects.

- **Crawler Process Creation**:
  - Loops num_crawlers times to create each crawler process.
  - Uses crawler_counter.get_lock() to safely generate a unique crawler_id (e.g., crawler0, crawler1) and increments the counter.
  - Creates a Process object p for each crawler, targeting the crawler_worker function with arguments crawler_id, log_queue, and max_depth, then starts it and adds it to the processes list.
- **Process Execution and Termination**:
  - Iterates over processes and calls join() on each to wait for all crawler processes to complete.
  - Sends None to log_queue to signal the log_listener to stop, then calls join() on the listener to terminate it.

if __name__ == "__main__":

- **Purpose**:
  This block serves as the entry point when the script (crawler_node.py) is run directly. It parses command-line arguments to configure the number of crawlers and maximum crawl depth, sets up basic logging, and invokes the run_crawlers function to start the crawling process.

- **Key Steps**:
  1. **Setup**:
     - Creates an argparse.ArgumentParser instance with a description "Multi-Process Crawler Node" to handle command-line arguments

       .

     - Defines two arguments:
       - --num-crawlers (type int, default 2, help text "Number of crawler processes") to specify the number of crawler processes.

       - --max-depth (type int, default 2, help text "Maximum crawl depth") to set the crawl depth limit.

Indexer.py

```
#!/usr/bin/env python3
import boto3
from whoosh.index import create_in
from whoosh.fields import Schema, TEXT
import os
import shutil
import logging
import sys
import json
import argparse
from multiprocessing import Process, Queue, Value
from logging.handlers import QueueHandler
import multiprocessing as mp

# --- AWS Setup ---
sqs_client = boto3.client('sqs', region_name='eu-north-1')
s3_client = boto3.client('s3')
TASK_QUEUE_URL = 'https://sqs.eu-north-1.amazonaws.com/965766185618/ces354_Queue'
RESULTS_QUEUE_URL = 'https://sqs.eu-north-1.amazonaws.com/965766185618/cse354_Results_Queue'
BUCKET_NAME = 'cse354000-bucket'

# --- Shared counter for unique IDs ---
indexer_counter = Value('i', 0)
```

This part initializes the script and sets up the necessary dependencies and AWS connections for the indexer node. The script imports libraries for AWS interaction (boto3 for S3 and SQS), indexing (whoosh for creating and managing search indexes), file handling (os, shutil), logging (logging, QueueHandler), and multiprocessing (multiprocessing). The AWS setup defines clients for SQS (to receive tasks and send results) and S3 (to store and retrieve crawled data and indexes), specifying the region (eu-north-1) and hard-coded URLs for the task and results queues, as well as the S3 bucket name. This setup establishes the foundation for the indexer node to communicate with the master node via SQS and store/retrieve data from S3, enabling the distributed indexing process critical to the web crawling and indexing system.

```python
class IndexerIdFilter(logging.Filter):
    """Add indexer_id to log records"""
    def __init__(self, indexer_id):
        super().__init__()
        self.indexer_id = indexer_id

    def filter(self, record):
        record.indexer_id = self.indexer_id
        return True

def setup_logger(indexer_id, log_queue):
    """Set up logging for multiprocessing"""
    formatter = logging.Formatter(
        fmt="%(asctime)s [%(levelname)s] [Indexer %(indexer_id)s] %(message)s",
        datefmt="%Y-%m-%d %H:%M:%S"
    )
    handler = QueueHandler(log_queue)
    handler.setFormatter(formatter)
    handler.addFilter(IndexerIdFilter(indexer_id))
    logger = logging.getLogger(f"indexer_{indexer_id}")
    logger.setLevel(logging.INFO)
    logger.addHandler(handler)
    logger.propagate = False
    return logger
```

This part establishes a shared counter and a robust logging system for the indexer processes to ensure unique identification and coordinated logging in a multiprocessing environment. The indexer_counter is a shared integer (Value) used to assign unique IDs to indexer processes, ensuring distinct log entries and index directories. The IndexerIdFilter class customizes log records by adding an indexer_id field, enabling identification of which indexer process generated a log message. The setup_logger function configures a logger for each indexer process, using a QueueHandler to send logs to a shared queue, formatted with timestamps, log levels, and indexer IDs. The log_listener function runs in a separate process, listening to the log queue and forwarding messages to the appropriate logger, ensuring thread-safe logging across multiple indexer processes. This setup is critical for debugging and monitoring the distributed indexing process, providing clear, process-specific logs for tracking progress and errors.

```python
def log_listener(log_queue):
    """Listen to log messages from all processes"""
    while True:
        try:
            record = log_queue.get()
            if record is None:  # Sentinel to stop listener
                break
            logger = logging.getLogger(record.name)
            logger.handle(record)
        except Exception as e:
            print(f"Log listener error: {e}", file=sys.stderr)

def index_files(file_queue, indexer_id):
    """Index a batch of files and upload to S3"""
    schema = Schema(title=TEXT(stored=True), content=TEXT(stored=True))
    index_dir = f"indexdir_{indexer_id}"
    if os.path.exists(index_dir):
        shutil.rmtree(index_dir)
    os.makedirs(index_dir, exist_ok=True)
    ix = create_in(index_dir, schema)
    writer = ix.writer()

    while True:
        try:
            key = file_queue.get(timeout=1)  # Timeout to check for empty queue
            if key is None:  # Sentinel to stop indexing
                break
            local_file = f"temp_crawled_{indexer_id}.txt"
            try:
                s3_client.download_file(BUCKET_NAME, key, local_file)
                with open(local_file, 'r', encoding='utf-8', errors='ignore') as f:
                    content = f.read()
                writer.add_document(title=key, content=content)
                logger.info(f"Indexed file: {key}")
            except Exception as e:
                logger.error(f"Failed to index {key}: {e}")
            finally:
                if os.path.exists(local_file):
                    os.remove(local_file)
        except Queue.Empty:
            break  # No more files to process

    writer.commit()
    logger.info(f"Committed documents for indexer {indexer_id}")

    # Upload index files to S3 in a single folder
    for root, dirs, files in os.walk(index_dir):
        for file in files:
            local_path = os.path.join(root, file)
            s3_key = f"index_data/{indexer_id}_{file}"
            try:
                s3_client.upload_file(local_path, BUCKET_NAME, s3_key)
                logger.info(f"Uploaded index file to S3: {s3_key}")
            except Exception as e:
                logger.error(f"Failed to upload {s3_key}: {e}")

    shutil.rmtree(index_dir)  # Clean up
```

The index_files function is the core of the indexing process, responsible for processing crawled files from S3, indexing their content using Whoosh, and uploading the resulting index files back to S3. It starts by defining a Whoosh schema with title and content fields (both stored for retrieval) and creates a unique index directory for the indexer process. The function retrieves S3 object keys from a shared file queue, downloads each file (assumed to be text from crawled web pages), and adds its content to the Whoosh index as a document. Errors during downloading or indexing are logged, and temporary files are cleaned up. Once all files are processed (or a sentinel None is received), the index is committed, and the index files are uploaded to S3 under a unique path (index_data/{indexer_id}_{file}). The index directory is then deleted to free up disk space. This function enables parallel indexing by multiple processes, ensuring scalability and fault tolerance by handling errors gracefully and storing results persistently in S3.

```
def list_s3_objects(bucket, prefix):
    """List S3 objects with pagination"""
    paginator = s3_client.get_paginator('list_objects_v2')
    for page in paginator.paginate(Bucket=bucket, Prefix=prefix):
        for obj in page.get('Contents', []):
            if obj['Key'].endswith('.txt'):
                yield obj['Key']

def indexer_worker(indexer_id, log_queue, file_queue):
    """Worker function for each indexer process"""
    global logger
    logger = setup_logger(indexer_id, log_queue)
    logger.info("Starting indexer worker")

    index_files(file_queue, indexer_id)
    logger.info("Finished indexing")
```

The list_s3_objects function provides a generator to list text files stored in the S3 bucket under a specified prefix (e.g., crawl_data/), handling pagination to manage large numbers of objects efficiently. It uses the Boto3 paginator for the list_objects_v2 API, iterating through pages of S3 objects and yielding keys for files ending in .txt, which represent crawled web page content. This function is crucial for the indexer node to discover all crawled files that need indexing, ensuring no files are missed even in a large dataset. By using pagination, it avoids memory issues and supports scalability, aligning with the project's requirement for processing large-scale crawled data in a distributed system.

The indexer_worker function serves as the entry point for each indexer process in the multiprocessing setup, coordinating the indexing task for a specific indexer instance. It initializes a logger for the process using the provided indexer_id and log_queue, ensuring that all log messages are uniquely tagged and routed correctly. The function then calls index_files to process a subset of crawled files from the shared file_queue, performing the indexing and S3 upload tasks. Once indexing is complete, it logs the completion. This function encapsulates the work of a single indexer process, enabling parallel execution of multiple indexers to distribute the indexing workload, which is essential for scalability and efficient processing in the distributed system.

```python
def run_indexers(num_indexers, logger):
    """Run multiple indexer processes"""
    # Set up logging queue and listener
    log_queue = mp.Queue()
    listener = Process(target=log_listener, args=(log_queue,))
    listener.start()

    # Fetch list of crawled files from S3 with pagination
    file_queue = mp.Queue()
    txt_files = list_s3_objects(BUCKET_NAME, "crawl_data/")
    file_count = 0
    for key in txt_files:
        file_queue.put(key)
        file_count += 1

    if file_count == 0:
        logger.warning("No crawled files found in S3.")
        return

    logger.info(f"Found {file_count} text files to index")

    # Add sentinels to stop workers
    for _ in range(num_indexers):
        file_queue.put(None)

    # Start indexer processes
    processes = []
    for i in range(num_indexers):
        with indexer_counter.get_lock():
```

```
    # Add sentinels to stop workers
    for _ in range(num_indexers):
        file_queue.put(None)

    # Start indexer processes
    processes = []
    for i in range(num_indexers):
        with indexer_counter.get_lock():
            indexer_id = f"indexer{indexer_counter.value}"
            indexer_counter.value += 1
        p = Process(target=indexer_worker, args=(indexer_id, log_queue, file_queue))
        p.start()
        processes.append(p)

    # Wait for all processes to finish
    for p in processes:
        p.join()

    # Stop the log listener
    log_queue.put(None)
    listener.join()
```

The run_indexers function orchestrates the execution of multiple indexer processes to parallelize the indexing of crawled files. It starts by setting up a logging queue and a listener process to handle logs from all indexer processes. It then populates a shared file queue with S3 object keys for text files (obtained via list_s3_objects), logging the total number of files to be indexed. If no files are found, it exits with a warning. The function adds sentinel None values to the file queue to signal indexer processes to stop once all files are processed. It then spawns the specified number of indexer processes, each with a unique indexer_id generated using the shared counter, and assigns them to process files from the queue. The function waits for all processes to complete and stops the log listener, ensuring all indexing tasks are finished and logs are properly recorded. This function is key to achieving scalability by distributing the indexing workload across multiple processes, aligning with the project's distributed computing goals.

```python
def indexer_node(num_indexers):
    """Main indexer node function"""
    # Configure root logger for initial messages
    logging.basicConfig(
        level=logging.INFO,
        format="%(asctime)s [%(levelname)s] %(message)s",
        handlers=[logging.StreamHandler(sys.stdout)]
    )
    logger = logging.getLogger("indexer_node")
    logger.info("Indexer node starting...")

    while True:
        response = sqs_client.receive_message(
            QueueUrl=TASK_QUEUE_URL,
            MaxNumberOfMessages=1,
            WaitTimeSeconds=20
        )

        if 'Messages' not in response:
            logger.info("No messages received, continuing to poll...")
            continue

        message = response['Messages'][0]
        body = json.loads(message['Body'])

        if 'start_indexer' in body:
            logger.info("Received start signal from Master")
            run_indexers(num_indexers, logger)
            sqs_client.delete_message(
                QueueUrl=TASK_QUEUE_URL,
```

The indexer_node function is the main entry point for the indexer node, coordinating its interaction with the master node via SQS and initiating the indexing process. It configures a root logger for initial messages and enters a loop to poll the SQS task queue for messages from the master node, using a 20-second wait time for long polling. When a message is received, it parses the JSON body and checks for a start_indexer signal. Upon receiving this signal, it calls run_indexers to start the indexing processes, deletes the processed message from the task queue to prevent reprocessing, and sends a completion signal to the results queue. The loop then breaks, terminating the indexer node. This function ensures the indexer node operates as part of the distributed system, responding to master node instructions and signaling completion, which is critical for the overall workflow of the web crawling and indexing system.

```
            if 'Messages' not in response:
                logger.info("No messages received, continuing to poll...")
                continue

            message = response['Messages'][0]
            body = json.loads(message['Body'])

            if 'start_indexer' in body:
                logger.info("Received start signal from Master")
                run_indexers(num_indexers, logger)
                sqs_client.delete_message(
                    QueueUrl=TASK_QUEUE_URL,
                    ReceiptHandle=message['ReceiptHandle']
                )
                sqs_client.send_message(
                    QueueUrl=RESULTS_QUEUE_URL,
                    MessageBody=json.dumps({'indexer_complete': True})
                )
                logger.info("Indexer finished and signaled completion")
                break

    if __name__ == "__main__":
        parser = argparse.ArgumentParser(description="Multi-Process Indexer Node")
        parser.add_argument('--num-indexers', type=int, default=2, help="Number of indexer processes")
        args = parser.parse_args()

        indexer_node(args.num_indexers)
```

This part defines the command-line interface and initiates the indexer node's execution. It uses argparse to parse a command-line argument --num-indexers, which specifies the number of indexer processes to run (defaulting to 2). The parsed argument is passed to the indexer_node function, starting the indexer node's operation. This setup allows users to control the level of parallelism in the indexing process, providing flexibility for testing and deployment in different environments. By serving as the script's entry point, this part integrates all previous components, enabling the indexer node to be run as a standalone program that interacts with AWS services and processes crawled data, fulfilling the project's requirement for a configurable, distributed indexing system.

Search.py

```
#!/usr/bin/env python3
from whoosh.index import open_dir
from whoosh.qparser import MultifieldParser
from whoosh.fields import Schema, TEXT
import os
import boto3
import json
import shutil
import re
import tempfile
from collections import defaultdict

# S3 setup
s3_client = boto3.client('s3')
sts_client = boto3.client('sts')
BUCKET_NAME = 'cse354000-bucket'
INDEX_BASE_DIR = "indexdir"
MAPPING_FILE = "url_mapping.json"
INDEX_PREFIX = "index_data/"
```

This part initializes the script by importing necessary libraries and configuring AWS connectivity for the search interface. The imports include Whoosh modules (open_dir, MultifieldParser, Schema, TEXT) for managing and querying search indexes, AWS SDK (boto3) for S3 and STS (Security Token Service) interactions, and standard Python libraries (os, json, shutil, re, tempfile, defaultdict) for file handling, data processing, and temporary directory management. The AWS setup defines clients for S3 (to retrieve index files and URL mappings) and STS (to verify credentials), along with constants for the S3 bucket name (cse354000-bucket), index directory prefix (index_data/), and URL mapping file (url_mapping.json).

```
def check_aws_identity():
    """Verify AWS credentials and print identity"""
    try:
        identity = sts_client.get_caller_identity()
        print(f"AWS Identity: UserId={identity['UserId']}, Account={identity['Account']}, Arn={identity['A
        return True
    except Exception as e:
        print(f"Failed to verify AWS identity: {e}")
        return False
```

The check_aws_identity function verifies the validity of AWS credentials before proceeding with S3 operations, ensuring secure and authorized access to cloud resources. It uses the STS client to call get_caller_identity, which returns details about the AWS user or role (UserId, Account, ARN). If successful, it prints these details and returns True, indicating valid credentials. If an error occurs (e.g., invalid credentials or network issues), it prints an error message and returns

False. This function is critical for the search interface to confirm that the system can access the S3 bucket containing index files and URL mappings.

```
def download_index_from_s3(temp_dir):
    """Download index files from S3 into separate subdirectories per indexer"""
    try:
        # Group files by indexer ID
        response = s3_client.list_objects_v2(Bucket=BUCKET_NAME, Prefix=INDEX_PREFIX)
        if 'Contents' not in response:
            print(f"No objects found in bucket '{BUCKET_NAME}' with prefix '{INDEX_PREFIX}'. Run the index
            return []

        # Organize files by indexer (e.g., indexer0, indexer1)
        indexer_files = defaultdict(list)
        for obj in response['Contents']:
            key = obj['Key']
            filename = os.path.basename(key)
            match = re.match(r'^indexer(\d+)_(.*)$', filename)
            if match:
                indexer_id, base_filename = match.groups()
                indexer_files[indexer_id].append((key, base_filename))

        if not indexer_files:
            print("No valid index files found in S3.")
            return []

        # Download files into separate subdirectories
        index_dirs = []
        for indexer_id, files in indexer_files.items():
            indexer_dir = os.path.join(temp_dir, f"indexer{indexer_id}")
            os.makedirs(indexer_dir, exist_ok=True)
```

```
        for indexer_id, files in indexer_files.items():
            indexer_dir = os.path.join(temp_dir, f"indexer{indexer_id}")
            os.makedirs(indexer_dir, exist_ok=True)
            for key, base_filename in files:
                local_path = os.path.join(indexer_dir, base_filename)
                print(f"Downloading {key} to {local_path}")
                s3_client.download_file(BUCKET_NAME, key, local_path)
            index_dirs.append(indexer_dir)

        print("Successfully downloaded index files from S3")
        return index_dirs
    except s3_client.exceptions.NoSuchBucket:
        print(f"Bucket '{BUCKET_NAME}' does not exist")
        return []
    except s3_client.exceptions.ClientError as e:
        error_code = e.response['Error']['Code']
        error_msg = e.response['Error']['Message']
        print(f"S3 Error: {error_code} - {error_msg}")
        return []
    except Exception as e:
        print(f"Unexpected error downloading index from S3: {e}")
        return []
```

The download_index_from_s3 function retrieves index files from the S3 bucket and organizes them into separate local subdirectories for each indexer process, preparing them for search operations. It lists objects in the S3 bucket under the index_data/ prefix using list_objects_v2, grouping files by indexer ID (e.g., indexer0, indexer1) based on a regex pattern matching the file names. For each indexer, it creates a subdirectory in a temporary directory, downloads the corresponding index files, and stores their paths. If no files are found or errors occur (e.g., bucket doesn't exist, access denied), it prints an error message and returns an empty list. The function returns a list of directory paths containing the downloaded index files. This functionality is essential for accessing the distributed indexes created by multiple indexer processes.

```
def download_url_mapping():
    """Download URL mapping from S3"""
    try:
        s3_client.download_file(BUCKET_NAME, "crawl_data/url_mapping.json", MAPPING_FILE)
        with open(MAPPING_FILE, 'r') as f:
            print("Successfully downloaded URL mapping")
            return json.load(f)
    except s3_client.exceptions.ClientError as e:
        error_code = e.response['Error']['Code']
        error_msg = e.response['Error']['Message']
        print(f"S3 Error downloading URL mapping: {error_code} - {error_msg}")
        return {}
    except Exception as e:
        print(f"Error downloading URL mapping: {e}")
        return {}
```

The download_url_mapping function retrieves a JSON file (url_mapping.json) from S3, which maps S3 object keys (representing indexed documents) to their original URLs. It attempts to download the file from the crawl_data/ prefix in the S3 bucket to a local file (url_mapping.json), then loads and returns the JSON content as a dictionary. If the download fails due to S3 errors (e.g., file not found, access issues) or other exceptions, it prints an error message and returns an empty dictionary. This function is crucial for translating indexed document titles (S3 keys) into human-readable URLs during search result display,

```
def simple_search():
    """Interactive search interface with advanced query support"""
    if not check_aws_identity():
        print("Cannot proceed without valid AWS credentials")
        return

    with tempfile.TemporaryDirectory() as temp_dir:
        index_dirs = download_index_from_s3(temp_dir)
        if not index_dirs:
            return

        url_mapping = download_url_mapping()
        if not url_mapping:
            print("URL mapping not found. Cannot display URLs.")

        # Open all indexes
        indexes = []
        for index_dir in index_dirs:
            try:
                ix = open_dir(index_dir)
                indexes.append(ix)
                print(f"Opened index in {index_dir}")
            except Exception as e:
                print(f"Error opening index in {index_dir}: {e}")
                continue

        if not indexes:
```

```
# Schema for parsing (must match indexer's schema)
schema = Schema(title=TEXT(stored=True), content=TEXT(stored=True))
parser = MultifieldParser(["title", "content"], schema=schema)

print("\nSearch Tips:")
print("- Exact match: 'python'")
print("- Phrase search: '\"python programming\"'")
print("- Boolean operators: 'python AND programming', 'python OR java', 'python NOT java'")
print("- Case-insensitive, use 'exit' to quit")

while True:
    try:
        query_str = input("\nEnter search query: ").strip()
        if not query_str:
            print("Please enter a query or 'exit'")
            continue
        if query_str.lower() == "exit":
            break

        # Parse the query
        query = parser.parse(query_str)
        all_results = []

        # Search each index and collect results
        for ix in indexes:
            with ix.searcher() as searcher:
                results = searcher.search(query, limit=20)
                for hit in results:
```

```
            # Sort results by score (descending) and deduplicate by title
            all_results = sorted(all_results, key=lambda x: x[1], reverse=True)
            seen_titles = set()
            unique_results = []
            for title, score in all_results:
                if title not in seen_titles:
                    unique_results.append(title)
                    seen_titles.add(title)

            if unique_results:
                print(f"\nFound {len(unique_results)} unique result(s) for query '{query_str}':")
                for i, s3_key in enumerate(unique_results[:1000], 1):
                    url = url_mapping.get(s3_key, f"URL not found for {s3_key}")
                    print(f"{i}. {url}")
            else:
                print(f"No results found for query '{query_str}'")

        except Exception as e:
            print(f"Search error: {e}. Please check your query syntax.")
            continue

    # Cleanup URL mapping file
    if os.path.exists(MAPPING_FILE):
        os.remove(MAPPING_FILE)
    print("Cleaned up local files")

if __name__ == "__main__":
    simple_search()
```

The simple_search function implements an interactive command-line interface for searching the indexed content, providing users with advanced query capabilities. It begins by verifying AWS credentials using check_aws_identity, exiting if invalid. It creates a temporary directory to store downloaded index files, retrieves index directories from S3 via download_index_from_s3, and downloads the URL mapping with download_url_mapping. It then opens each index directory using Whoosh's open_dir, creating a list of index objects. A MultifieldParser is configured to parse queries across title and content fields, matching the indexer's schema. The interface prompts users for queries, supporting exact matches, phrase searches, and Boolean operators (AND, OR, NOT), with instructions displayed. For each query, it parses the input, searches all indexes, collects results (up to 20 per index), sorts them by relevance score, and deduplicates by title. Results are displayed as URLs (using the URL mapping) or S3 keys if unmapped, with up to 1000 results shown. Errors in query syntax are caught and reported, and users can exit by typing "exit". Finally, it cleans up the local URL mapping file.

Finally **main** serves as the entry point for the script, directly invoking the simple_search function when the script is run. It requires no command-line arguments, making the search interface immediately accessible to users with valid AWS credentials. By calling simple_search, it initiates the process of verifying credentials, downloading indexes and URL mappings, and starting the interactive search loop.

Client.py

```
#!/usr/bin/env python3
import boto3
import json
import os
import tempfile
import re
from whoosh.index import open_dir
from whoosh.qparser import MultifieldParser
from whoosh.fields import Schema, TEXT
from collections import defaultdict
import subprocess
import sys
from urllib.parse import urlparse
from mpi4py import MPI
import logging

# AWS Setup
s3_client = boto3.client('s3', region_name='eu-north-1')
sts_client = boto3.client('sts')
BUCKET_NAME = 'cse354000-bucket'
INDEX_PREFIX = 'index_data/'
MAPPING_FILE = 'url_mapping.json'
SEED_URLS_PATH = 'seed_urls/seed_urls.txt'
```

This part initializes the script by importing necessary libraries and configuring AWS and logging setups for the client interface. The imports include AWS SDK (boto3 for S3 and STS), Whoosh modules (open_dir, MultifieldParser, Schema, TEXT) for search functionality, and standard Python libraries (json, os, tempfile, re, defaultdict, subprocess, sys, urlparse) for data handling, file management, and process execution. The mpi4py library supports MPI-based communication, though it's not directly used in this code. The AWS setup defines clients for S3 (to upload/download files) and STS (to verify credentials), along with constants for the S3 bucket, index prefix, URL mapping file, and seed URLs path. The logging setup configures a logger with a custom format, outputting to stdout, which ensures consistent logging for debugging and user feedback.

```python
def check_aws_identity():
    """Verify AWS credentials and print identity"""
    try:
        identity = sts_client.get_caller_identity()
        logger.info(f"AWS Identity: UserId={identity['UserId']}, Account={identity['Account']},
        return True
    except Exception as e:
        logger.error(f"Failed to verify AWS identity: {e}")
        return False
```

The check_aws_identity function verifies the validity of AWS credentials to ensure secure access to S3 and other cloud resources. It uses the STS client to call get_caller_identity, which returns details about the AWS user or role (UserId, Account, ARN). If successful, it logs these details using the configured logger and returns True. If an error occurs (e.g., invalid credentials or network issues), it logs an error message and returns False. This function is critical for all client operations (submitting URLs, running the master node, searching), as it ensures the client can interact with the S3 bucket for storing seed URLs, retrieving indexes, and accessing URL mappings

```python
def validate_url(url):
    """Validate URL format"""
    try:
        parsed = urlparse(url.strip())
        return parsed.scheme in ['http', 'https'] and parsed.netloc
    except:
        return False
```

```python
def submit_urls():
    logger.info("Enter seed URLs (one per line, empty line to finish):")
    urls = []
    while True:
        url = input("URL: ").strip()
        if not url:
            break
        if validate_url(url):
            urls.append(url)
        else:
            logger.warning(f"Invalid URL: {url}. Skipping.")

    if not urls:
        logger.error("No valid URLs provided.")
        return False

    # Save URLs to a temporary file
    with tempfile.NamedTemporaryFile(mode='w', delete=False, suffix='.txt') as temp_file:
        for url in urls:
            temp_file.write(url + '\n')
        temp_file_path = temp_file.name

    try:
        # Upload to S3
        s3_client.upload_file(temp_file_path, BUCKET_NAME, SEED_URLS_PATH)
        logger.info(f"Uploaded {len(urls)} seed URLs to S3: {SEED_URLS_PATH}")
        os.remove(temp_file_path)
        return True
    except Exception as e:
        logger.error(f"Failed to upload seed URLs to S3: {e}")
```

This part handles the submission of seed URLs by the user, a key feature of the client interface. The validate_url function checks if a URL is valid by parsing it with urlparse and ensuring it has an HTTP/HTTPS scheme and a valid network location (e.g., domain). The submit_urls function prompts the user to enter URLs one per line, validating each using validate_url and collecting valid URLs in a list. If no valid URLs are provided, it logs an error and exits. Valid URLs are written to a temporary file, which is uploaded to S3 at the specified path (seed_urls/seed_urls.txt). The temporary file is then deleted, and the function logs the success or failure of the upload, returning True or False accordingly.

```python
def run_master_node(num_crawlers=2, max_depth=2):
    """Run Master.py using MPI"""
    try:
        logger.info(f"Starting master node with num_crawlers={num_crawlers}, max_depth={max_depth}")
        # Run Master.py with MPI
        cmd = [
            'mpirun', '-np', '1',
            'python3', 'master_node.py',
            '--num-crawlers', str(num_crawlers),
            '--max-depth', str(max_depth)
        ]
        process = subprocess.run(cmd, capture_output=True, text=True)
        if process.returncode == 0:
            logger.info("Master node completed successfully")
            return True
        else:
            logger.error(f"Master node failed: {process.stderr}")
            return False
    except Exception as e:
        logger.error(f"Error running master node: {e}")
        return False
```

The run_master_node function initiates the crawling process by executing the master_node.py script using MPI. It constructs a command to run master_node.py with mpirun, specifying one process (-np 1) and passing arguments for the number of crawlers and maximum crawl depth. The subprocess.run call executes the command, capturing output and errors. If the process completes successfully (return code 0), it logs success and returns True; otherwise, it logs the error from stderr and returns False. Exceptions during execution (e.g., missing master_node.py or MPI issues) are caught and logged.

```python
def download_index_from_s3(temp_dir):
    """Download index files from S3 into separate subdirectories per indexer"""
    try:
        response = s3_client.list_objects_v2(Bucket=BUCKET_NAME, Prefix=INDEX_PREFIX)
        if 'Contents' not in response:
            logger.error(f"No objects found in bucket '{BUCKET_NAME}' with prefix '{INDEX_PREFIX}'. Run the indexer first.")
            return []

        indexer_files = defaultdict(list)
        for obj in response['Contents']:
            key = obj['Key']
            filename = os.path.basename(key)
            match = re.match(r'^indexer(\d+)_(.*)$', filename)
            if match:
                indexer_id, base_filename = match.groups()
                indexer_files[indexer_id].append((key, base_filename))

        if not indexer_files:
            logger.error("No valid index files found in S3.")
            return []

        index_dirs = []
        for indexer_id, files in indexer_files.items():
            indexer_dir = os.path.join(temp_dir, f"indexer{indexer_id}")
            os.makedirs(indexer_dir, exist_ok=True)
            for key, base_filename in files:
                local_path = os.path.join(indexer_dir, base_filename)
                logger.info(f"Downloading {key} to {local_path}")
```

The download_index_from_s3 function retrieves index files from the S3 bucket under the index_data/ prefix, organizing them into local subdirectories for each indexer process to support search operations. It uses list_objects_v2 to list objects, grouping files by indexer ID (e.g., indexer0, indexer1) based on a regex pattern. For each indexer, it creates a subdirectory in the

provided temporary directory, downloads the corresponding index files, and logs the process. If no files are found, the bucket doesn't exist, or other errors occur, it logs an error and returns an empty list. The function returns a list of directory paths containing the downloaded index files.

```
def download_url_mapping():
    """Download URL mapping from S3"""
    try:
        s3_client.download_file(BUCKET_NAME, "crawl_data/url_mapping.json", MAPPING_FILE)
        with open(MAPPING_FILE, 'r') as f:
            logger.info("Successfully downloaded URL mapping")
            return json.load(f)
    except s3_client.exceptions.ClientError as e:
        error_code = e.response['Error']['Code']
        error_msg = e.response['Error']['Message']
        logger.error(f"S3 Error downloading URL mapping: {error_code} - {error_msg}")
        return {}
    except Exception as e:
        logger.error(f"Error downloading URL mapping: {e}")
        return {}
```

The download_url_mapping function retrieves a JSON file (url_mapping.json) from S3, which maps S3 object keys (representing indexed documents) to their original URLs. It downloads the file from the crawl_data/ prefix to a local file (url_mapping.json), loads the JSON content as a dictionary, and logs success. If the download fails due to S3 errors (e.g., file not found, access issues) or other exceptions, it logs the error and returns an empty dictionary.

```python
def search():
    """Interactive search interface with advanced query support"""
    if not check_aws_identity():
        logger.error("Cannot proceed without valid AWS credentials")
        return

    with tempfile.TemporaryDirectory() as temp_dir:
        index_dirs = download_index_from_s3(temp_dir)
        if not index_dirs:
            return

        url_mapping = download_url_mapping()
        if not url_mapping:
            logger.error("URL mapping not found. Cannot display URLs.")

        indexes = []
        for index_dir in index_dirs:
            try:
                ix = open_dir(index_dir)
                indexes.append(ix)
                logger.info(f"Opened index in {index_dir}")
            except Exception as e:
                logger.error(f"Error opening index in {index_dir}: {e}")
                continue
```

The search function implements an interactive command-line search interface, allowing users to query indexed content with advanced query support. It verifies AWS credentials using check_aws_identity, creates a temporary directory, and downloads index files and the URL mapping from S3 using download_index_from_s3 and download_url_mapping. It opens each index directory with Whoosh's open_dir, configures a MultifieldParser to query title and content fields, and provides search tips (e.g., exact matches, phrase searches, Boolean operators). The function enters a loop, prompting for queries, parsing them, searching all indexes, collecting results (up to 20 per index), sorting by relevance, deduplicating by title, and displaying URLs (or S3 keys if unmapped). Errors are logged, and users can exit with "exit". Finally, it cleans up the local URL mapping file.

```python
def main():

    while True:
        print("\nClient Menu:")
        print("1. Submit seed URLs")
        print("2. Run master node")
        print("3. Search")
        print("4. Exit")
        choice = input("Enter choice (1-4): ").strip()

        if choice == '1':
            submit_urls()
        elif choice == '2':
            num_crawlers = input("Enter number of crawlers (default 2): ").strip() or '2'
            max_depth = input("Enter max crawl depth (default 2): ").strip() or '2'
            try:
                num_crawlers = int(num_crawlers)
                max_depth = int(max_depth)
                if num_crawlers < 1 or max_depth < 0:
                    logger.error("Number of crawlers must be at least 1, and max depth must be non-negative.")
                    continue
                run_master_node(num_crawlers, max_depth)
            except ValueError:
                logger.error("Invalid input. Please enter numeric values.")
        elif choice == '3':
            search()
        elif choice == '4':
            logger.info("Exiting client")
            break
        else:
            logger.warning("Invalid choice. Please enter 1, 2, 3, or 4.")
```

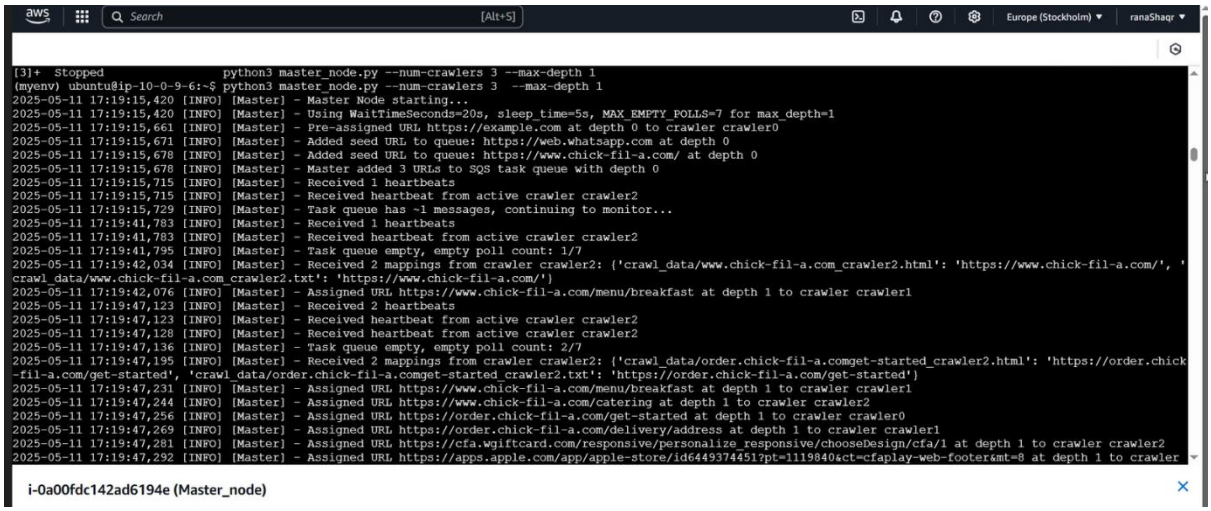
The main function provides the primary user interface for the client, offering a menu-driven experience to interact with the distributed system. It first verifies AWS credentials, exiting if invalid. It then enters a loop, displaying a menu with options to submit seed URLs, run the master node, search, or exit. For option 1, it calls submit_urls. For option 2, it prompts for the number of crawlers and maximum crawl depth (defaulting to 2), validates inputs, and calls run_master_node with the parameters. For option 3, it calls search. Option 4 exits the loop, and invalid choices trigger a warning. This function integrates all client functionalities into a cohesive interface, supporting the project's requirements for a command-line tool to initiate crawls, configure parameters,

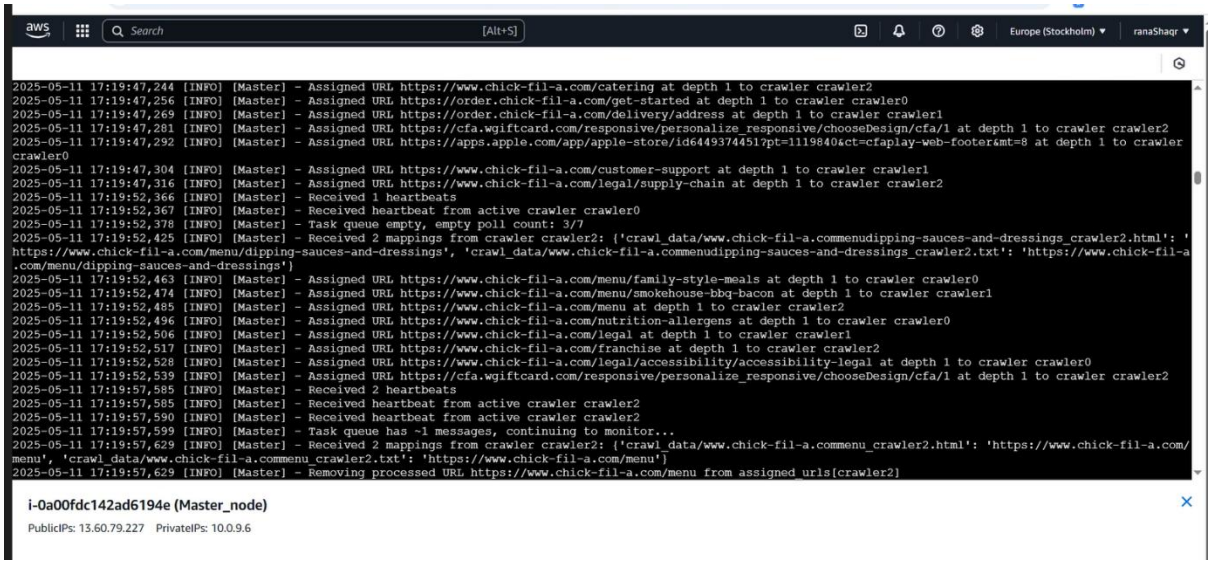
System Testing and Evaluation

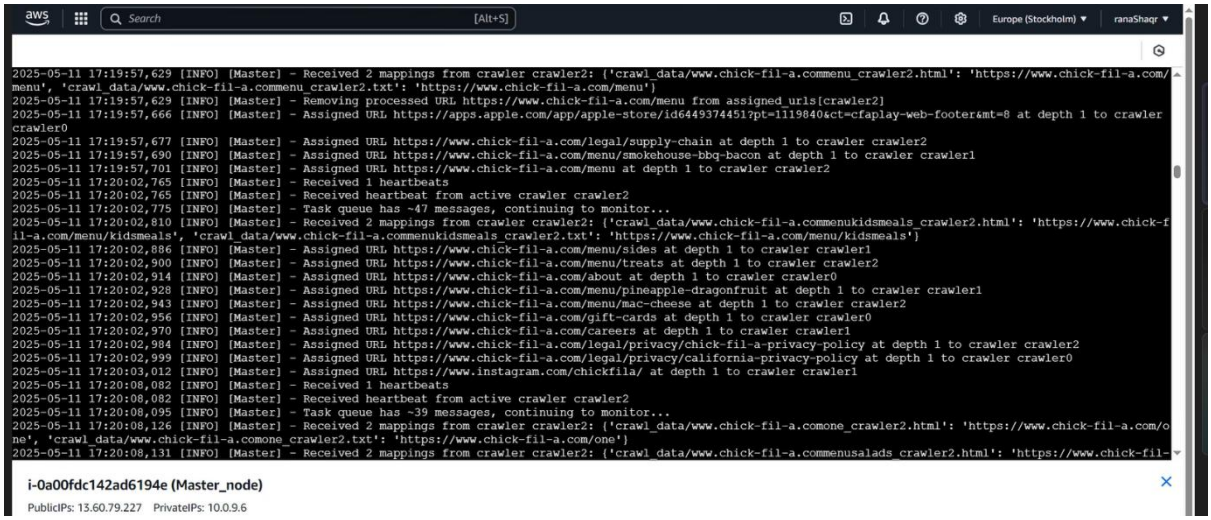
Functional Testing

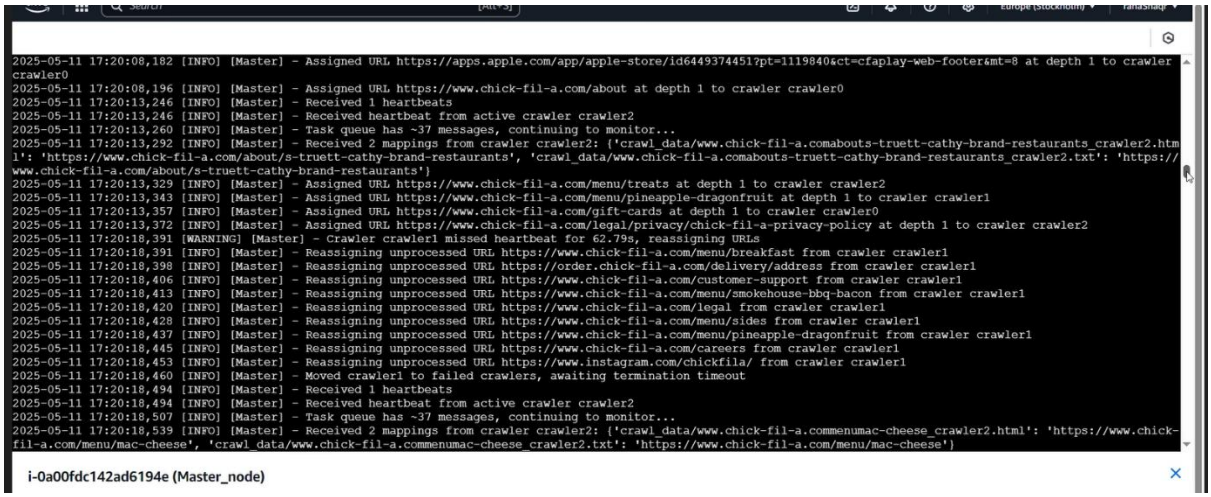
The following screenshots contain a full run for the three instances at the same time.

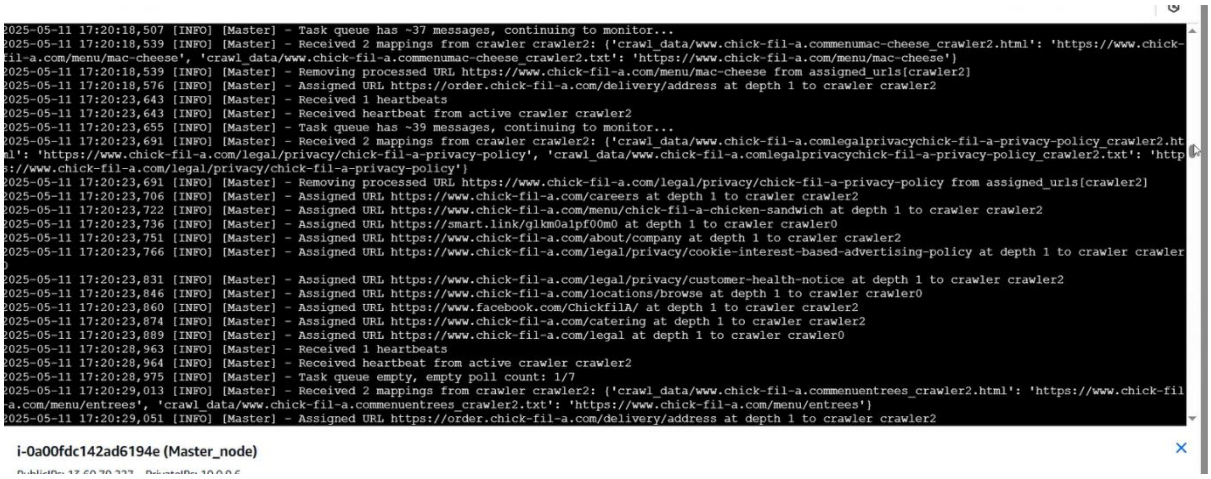
Master run

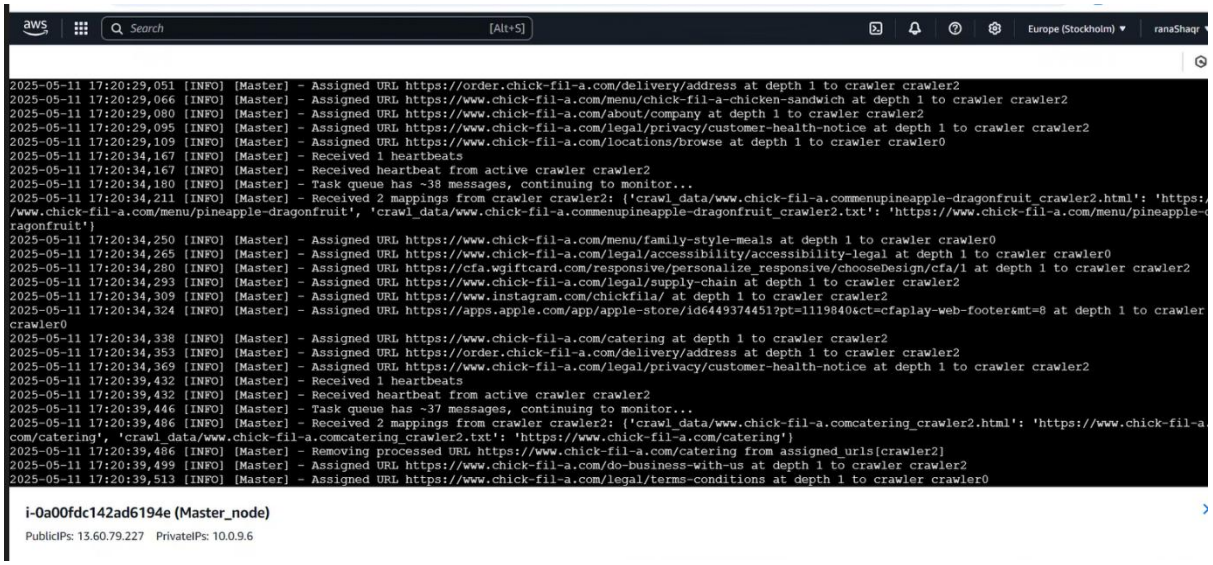


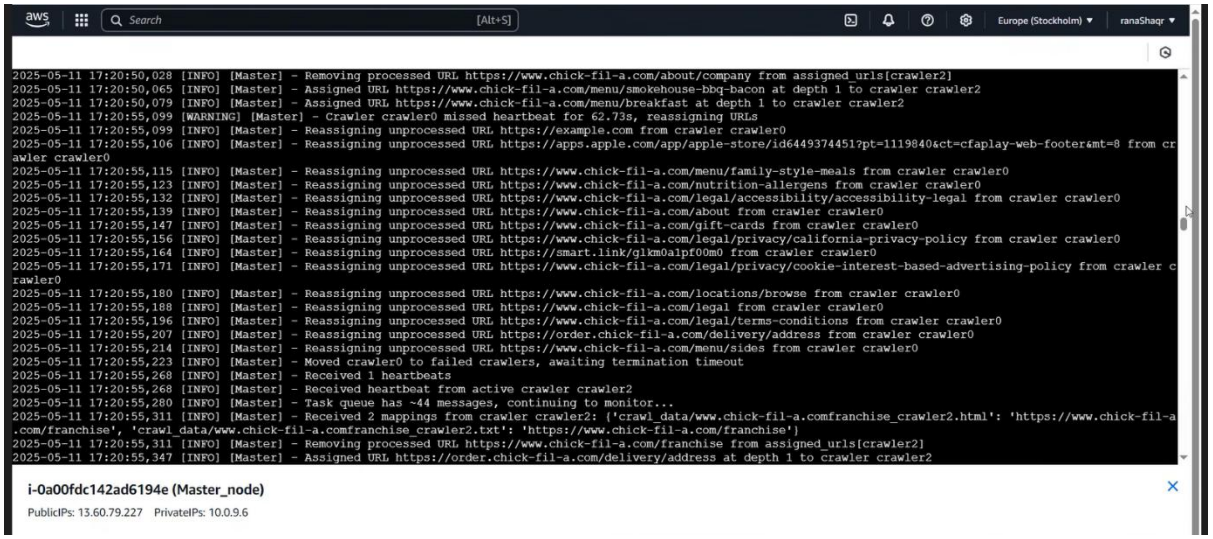


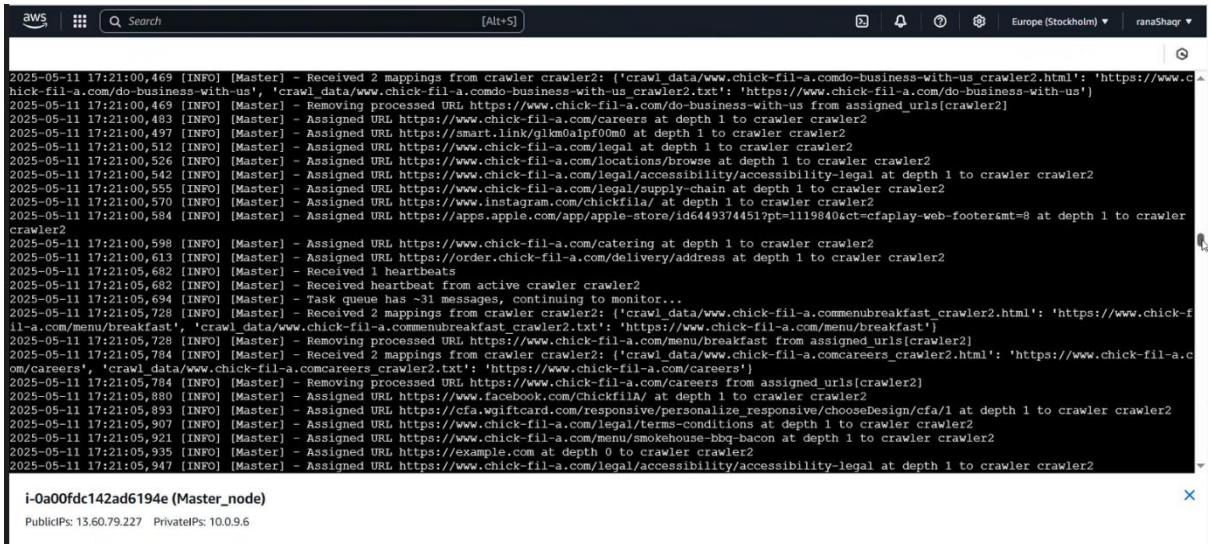

aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2025-05-11 17:19:57,629 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenu_crawler2.html': 'https://www.chick-fil-a.com/menu', 'crawl_data/www.chick-fil-a.commenu_crawler2.txt': 'https://www.chick-fil-a.com/menu'}
2025-05-11 17:19:57,629 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/menu from assigned_urls[crawler2]
2025-05-11 17:19:57,666 [INFO] [Master] - Assigned URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 at depth 1 to crawler crawler0
2025-05-11 17:19:57,677 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/supply-chain at depth 1 to crawler crawler2
2025-05-11 17:19:57,690 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1 to crawler crawler1
2025-05-11 17:19:57,701 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu at depth 1 to crawler crawler2
2025-05-11 17:20:02,765 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:02,765 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:02,775 [INFO] [Master] - Task queue has ~47 messages, continuing to monitor...
2025-05-11 17:20:02,810 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenukidsmeals_crawler2.html': 'https://www.chick-fil-a.com/menu/kidsmeals', 'crawl_data/www.chick-fil-a.commenukidsmeals_crawler2.txt': 'https://www.chick-fil-a.com/menu/kidsmeals'}
2025-05-11 17:20:02,886 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/sides at depth 1 to crawler crawler1
2025-05-11 17:20:02,900 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/treats at depth 1 to crawler crawler2
2025-05-11 17:20:02,914 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/about at depth 1 to crawler crawler0
2025-05-11 17:20:02,928 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/pineapple-dragonfruit at depth 1 to crawler crawler1
2025-05-11 17:20:02,943 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/mac-cheese at depth 1 to crawler crawler2
2025-05-11 17:20:02,956 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/gift-cards at depth 1 to crawler crawler0
2025-05-11 17:20:02,970 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/careers at depth 1 to crawler crawler1
2025-05-11 17:20:02,984 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy at depth 1 to crawler crawler2
2025-05-11 17:20:02,999 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/california-privacy-policy at depth 1 to crawler crawler0
2025-05-11 17:20:03,012 [INFO] [Master] - Assigned URL https://www.instagram.com/chickfila/ at depth 1 to crawler crawler1
2025-05-11 17:20:08,082 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:08,082 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:08,095 [INFO] [Master] - Task queue has ~39 messages, continuing to monitor...
2025-05-11 17:20:08,126 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comone_crawler2.html': 'https://www.chick-fil-a.com/one', 'crawl_data/www.chick-fil-a.comone_crawler2.txt': 'https://www.chick-fil-a.com/one'}
2025-05-11 17:20:08,131 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenusalads_crawler2.html': 'https://www.chick-fil-
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

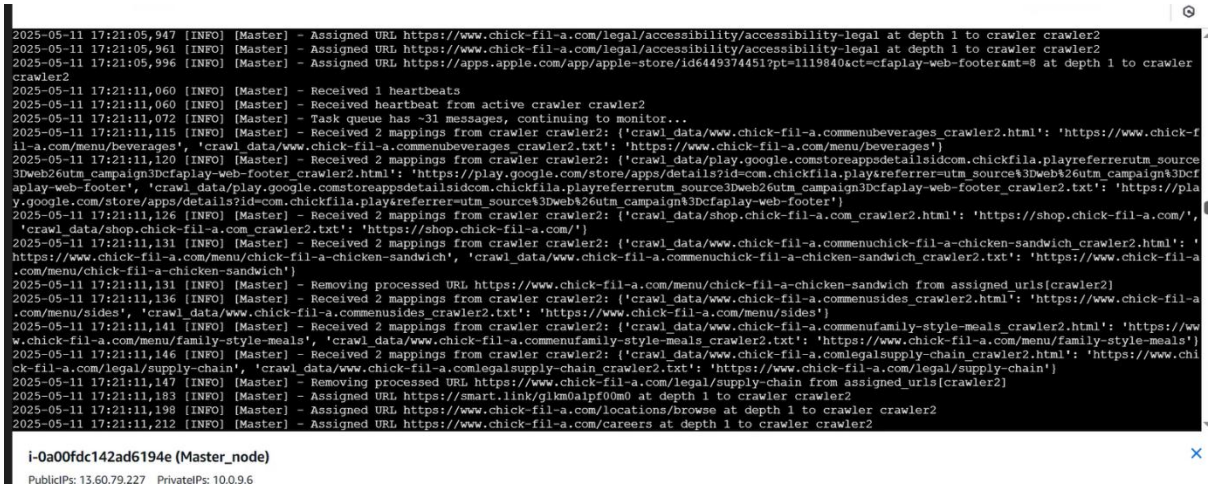

2025-05-11 17:20:08,182 [INFO] [Master] - Assigned URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 at depth 1 to crawler crawler0
2025-05-11 17:20:08,196 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/about at depth 1 to crawler crawler0
2025-05-11 17:20:13,246 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:13,246 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:13,260 [INFO] [Master] - Task queue has ~37 messages, continuing to monitor...
2025-05-11 17:20:13,292 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comabouts-truett-cathy-brand-restaurants_crawler2.html': 'https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants', 'crawl_data/www.chick-fil-a.comabouts-truett-cathy-brand-restaurants_crawler2.txt': 'https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants'}
2025-05-11 17:20:13,329 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/treats at depth 1 to crawler crawler2
2025-05-11 17:20:13,343 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/pineapple-dragonfruit at depth 1 to crawler crawler1
2025-05-11 17:20:13,357 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/gift-cards at depth 1 to crawler crawler0
2025-05-11 17:20:13,372 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy at depth 1 to crawler crawler2
2025-05-11 17:20:18,391 [WARNING] [Master] - Crawler crawler1 missed heartbeat for 62.79s, reassigning URLs
2025-05-11 17:20:18,391 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/breakfast from crawler crawler1
2025-05-11 17:20:18,398 [INFO] [Master] - Reassigning unprocessed URL https://order.chick-fil-a.com/delivery/address from crawler crawler1
2025-05-11 17:20:18,406 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/customer-support from crawler crawler1
2025-05-11 17:20:18,413 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon from crawler crawler1
2025-05-11 17:20:18,420 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal from crawler crawler1
2025-05-11 17:20:18,428 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/sides from crawler crawler1
2025-05-11 17:20:18,437 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/pineapple-dragonfruit from crawler crawler1
2025-05-11 17:20:18,445 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/careers from crawler crawler1
2025-05-11 17:20:18,453 [INFO] [Master] - Reassigning unprocessed URL https://www.instagram.com/chickfila/ from crawler crawler1
2025-05-11 17:20:18,460 [INFO] [Master] - Moved crawler1 to failed crawlers, awaiting termination timeout
2025-05-11 17:20:18,494 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:18,494 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:18,507 [INFO] [Master] - Task queue has ~37 messages, continuing to monitor...
2025-05-11 17:20:18,539 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.html': 'https://www.chick-fil-a.com/menu/mac-cheese', 'crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.txt': 'https://www.chick-fil-a.com/menu/mac-cheese'}
i-0a00fdc142ad6194e (Master_node)

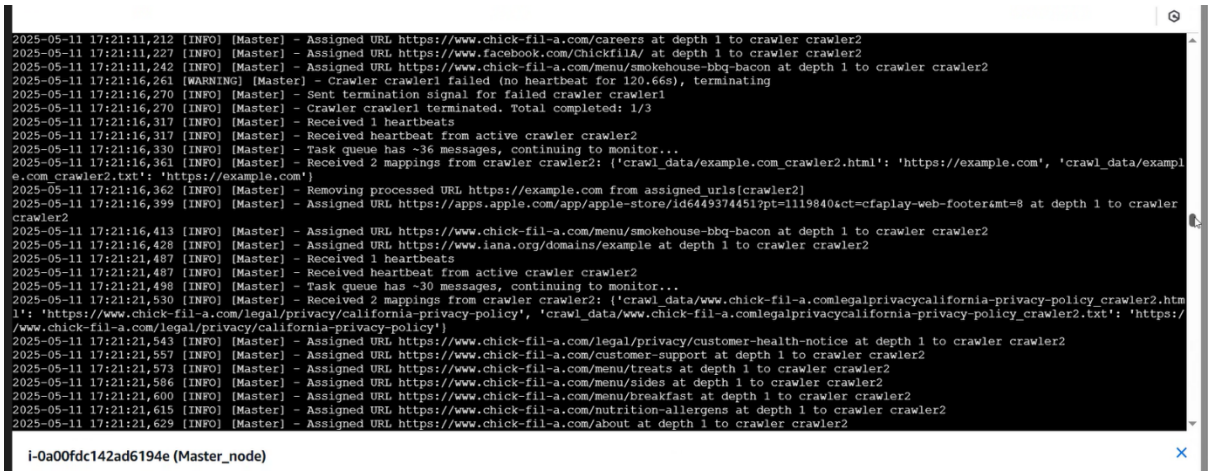

2025-05-11 17:20:18,507 [INFO] [Master] - Task queue has ~37 messages, continuing to monitor...
2025-05-11 17:20:18,539 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.html': 'https://www.chick-fil-a.com/menu/mac-cheese', 'crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.txt': 'https://www.chick-fil-a.com/menu/mac-cheese'}
2025-05-11 17:20:18,539 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/menu/mac-cheese from assigned_urls[crawler2]
2025-05-11 17:20:18,576 [INFO] [Master] - Assigned URL https://order.chick-fil-a.com/delivery/address at depth 1 to crawler crawler2
2025-05-11 17:20:23,643 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:23,643 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:23,655 [INFO] [Master] - Task queue has ~39 messages, continuing to monitor...
2025-05-11 17:20:23,691 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlegalprivacychick-fil-a-privacy-policy_crawler2.html': 'https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy', 'crawl_data/www.chick-fil-a.comlegalprivacychick-fil-a-privacy-policy_crawler2.txt': 'https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy'}
2025-05-11 17:20:23,691 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy from assigned_urls[crawler2]
2025-05-11 17:20:23,706 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/careers at depth 1 to crawler crawler2
2025-05-11 17:20:23,722 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich at depth 1 to crawler crawler2
2025-05-11 17:20:23,736 [INFO] [Master] - Assigned URL https://smart.link/glkm0a1pf00m0 at depth 1 to crawler crawler0
2025-05-11 17:20:23,751 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/about/company at depth 1 to crawler crawler2
2025-05-11 17:20:23,766 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy at depth 1 to crawler crawler
2025-05-11 17:20:23,831 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth 1 to crawler crawler2
2025-05-11 17:20:23,846 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/locations/browse at depth 1 to crawler crawler0
2025-05-11 17:20:23,860 [INFO] [Master] - Assigned URL https://www.facebook.com/ChickfilA/ at depth 1 to crawler crawler2
2025-05-11 17:20:23,874 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/catering at depth 1 to crawler crawler2
2025-05-11 17:20:23,889 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal at depth 1 to crawler crawler0
2025-05-11 17:20:28,963 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:28,964 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:28,975 [INFO] [Master] - Task queue empty, empty poll count: 1/7
2025-05-11 17:20:29,013 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenuentrees_crawler2.html': 'https://www.chick-fil-a.com/menu/entrees', 'crawl_data/www.chick-fil-a.commenuentrees_crawler2.txt': 'https://www.chick-fil-a.com/menu/entrees'}
2025-05-11 17:20:29,051 [INFO] [Master] - Assigned URL https://order.chick-fil-a.com/delivery/address at depth 1 to crawler crawler2
i-0a00fdc142ad6194e (Master_node)

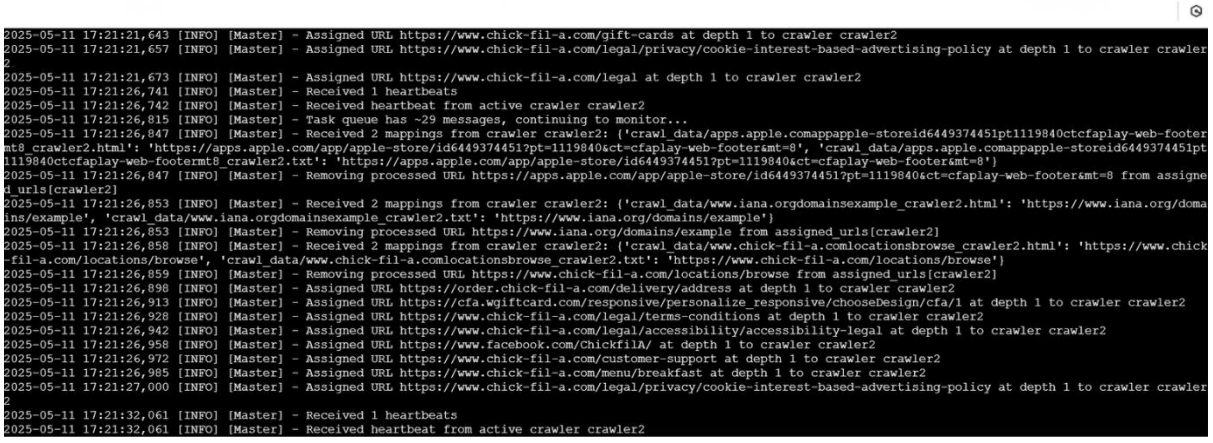

aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2025-05-11 17:20:29,051 [INFO] [Master] - Assigned URL https://order.chick-fil-a.com/delivery/address at depth 1 to crawler crawler2
2025-05-11 17:20:29,066 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich at depth 1 to crawler crawler2
2025-05-11 17:20:29,080 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/about/company at depth 1 to crawler crawler2
2025-05-11 17:20:29,095 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth 1 to crawler crawler2
2025-05-11 17:20:29,109 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/locations/browse at depth 1 to crawler crawler0
2025-05-11 17:20:34,167 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:34,167 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:34,180 [INFO] [Master] - Task queue has ~38 messages, continuing to monitor...
2025-05-11 17:20:34,211 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenupineapple-dragonfruit_crawler2.html': 'https:/
/www.chick-fil-a.com/menu/pineapple-dragonfruit', 'crawl_data/www.chick-fil-a.commenupineapple-dragonfruit_crawler2.txt': 'https://www.chick-fil-a.com/menu/pineapple-d
ragonfruit'}
2025-05-11 17:20:34,250 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/family-style-meals at depth 1 to crawler crawler0
2025-05-11 17:20:34,265 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1 to crawler crawler0
2025-05-11 17:20:34,280 [INFO] [Master] - Assigned URL https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1 at depth 1 to crawler crawler2
2025-05-11 17:20:34,293 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/supply-chain at depth 1 to crawler crawler2
2025-05-11 17:20:34,309 [INFO] [Master] - Assigned URL https://www.instagram.com/chickfila/ at depth 1 to crawler crawler2
2025-05-11 17:20:34,324 [INFO] [Master] - Assigned URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 at depth 1 to crawler
crawler0
2025-05-11 17:20:34,338 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/catering at depth 1 to crawler crawler2
2025-05-11 17:20:34,353 [INFO] [Master] - Assigned URL https://order.chick-fil-a.com/delivery/address at depth 1 to crawler crawler2
2025-05-11 17:20:34,369 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth 1 to crawler crawler2
2025-05-11 17:20:39,432 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:39,432 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:39,446 [INFO] [Master] - Task queue has ~37 messages, continuing to monitor...
2025-05-11 17:20:39,486 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comcatering_crawler2.html': 'https://www.chick-fil-a.
com/catering', 'crawl_data/www.chick-fil-a.comcatering_crawler2.txt': 'https://www.chick-fil-a.com/catering'}
2025-05-11 17:20:39,486 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/catering from assigned_urls[crawler2]
2025-05-11 17:20:39,499 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/do-business-with-us at depth 1 to crawler crawler2
2025-05-11 17:20:39,513 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler0
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

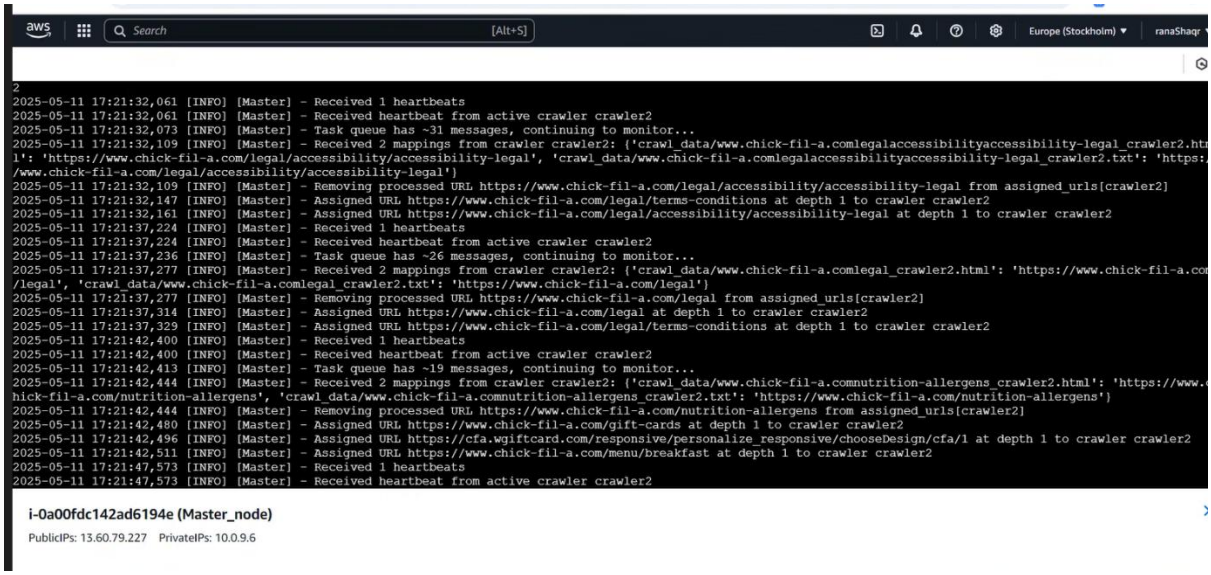

aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2025-05-11 17:20:50,028 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/about/company from assigned_urls[crawler2]
2025-05-11 17:20:50,065 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1 to crawler crawler2
2025-05-11 17:20:50,079 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/breakfast at depth 1 to crawler crawler2
2025-05-11 17:20:55,099 [WARNING] [Master] - Crawler crawler0 missed heartbeat for 62.73s, reassigning URLs
2025-05-11 17:20:55,099 [INFO] [Master] - Reassigning unprocessed URL https://example.com from crawler crawler0
2025-05-11 17:20:55,106 [INFO] [Master] - Reassigning unprocessed URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 from cr
awler crawler0
2025-05-11 17:20:55,115 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/family-style-meals from crawler crawler0
2025-05-11 17:20:55,123 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/nutrition-allergens from crawler crawler0
2025-05-11 17:20:55,132 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal from crawler crawler0
2025-05-11 17:20:55,139 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/about from crawler crawler0
2025-05-11 17:20:55,147 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/gift-cards from crawler crawler0
2025-05-11 17:20:55,156 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/privacy/california-privacy-policy from crawler crawler0
2025-05-11 17:20:55,164 [INFO] [Master] - Reassigning unprocessed URL https://smart.link/glkm0a1pf00m0 from crawler crawler0
2025-05-11 17:20:55,171 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy from crawler c
rawler0
2025-05-11 17:20:55,180 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/locations/browse from crawler crawler0
2025-05-11 17:20:55,188 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal from crawler crawler0
2025-05-11 17:20:55,196 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/terms-conditions from crawler crawler0
2025-05-11 17:20:55,207 [INFO] [Master] - Reassigning unprocessed URL https://order.chick-fil-a.com/delivery/address from crawler crawler0
2025-05-11 17:20:55,214 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/sides from crawler crawler0
2025-05-11 17:20:55,223 [INFO] [Master] - Moved crawler0 to failed crawlers, awaiting termination timeout
2025-05-11 17:20:55,268 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:55,268 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:55,280 [INFO] [Master] - Task queue has ~44 messages, continuing to monitor...
2025-05-11 17:20:55,311 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comfranchise_crawler2.html': 'https://www.chick-fil-a
.com/franchise', 'crawl_data/www.chick-fil-a.comfranchise_crawler2.txt': 'https://www.chick-fil-a.com/franchise'}
2025-05-11 17:20:55,311 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/franchise from assigned_urls[crawler2]
2025-05-11 17:20:55,347 [INFO] [Master] - Assigned URL https://order.chick-fil-a.com/delivery/address at depth 1 to crawler crawler2
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

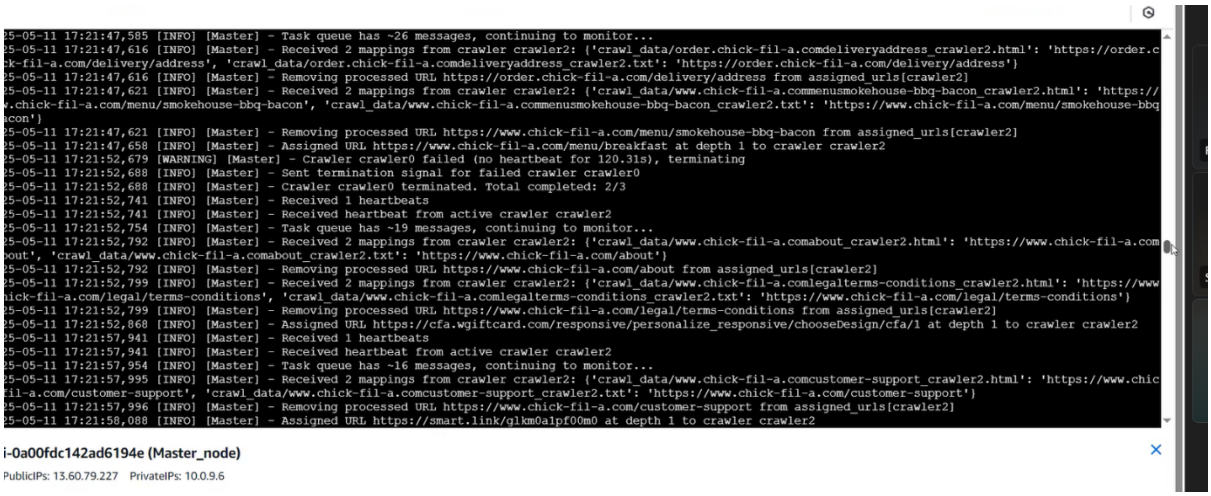

aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2025-05-11 17:21:00,469 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comdo-business-with-us_crawler2.html': 'https://www.c
hick-fil-a.com/do-business-with-us', 'crawl_data/www.chick-fil-a.comdo-business-with-us_crawler2.txt': 'https://www.chick-fil-a.com/do-business-with-us'}
2025-05-11 17:21:00,469 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/do-business-with-us from assigned_urls[crawler2]
2025-05-11 17:21:00,483 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/careers at depth 1 to crawler crawler2
2025-05-11 17:21:00,497 [INFO] [Master] - Assigned URL https://smart.link/glkm0a1pf00m0 at depth 1 to crawler crawler2
2025-05-11 17:21:00,512 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal at depth 1 to crawler crawler2
2025-05-11 17:21:00,526 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/locations/browse at depth 1 to crawler crawler2
2025-05-11 17:21:00,542 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1 to crawler crawler2
2025-05-11 17:21:00,555 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/supply-chain at depth 1 to crawler crawler2
2025-05-11 17:21:00,570 [INFO] [Master] - Assigned URL https://www.instagram.com/chickfila/ at depth 1 to crawler crawler2
2025-05-11 17:21:00,584 [INFO] [Master] - Assigned URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 at depth 1 to crawler
crawler2
2025-05-11 17:21:00,598 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/catering at depth 1 to crawler crawler2
2025-05-11 17:21:00,613 [INFO] [Master] - Assigned URL https://order.chick-fil-a.com/delivery/address at depth 1 to crawler crawler2
2025-05-11 17:21:05,682 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:05,682 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:05,694 [INFO] [Master] - Task queue has ~31 messages, continuing to monitor...
2025-05-11 17:21:05,728 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenubreakfast_crawler2.html': 'https://www.chick-f
il-a.com/menu/breakfast', 'crawl_data/www.chick-fil-a.commenubreakfast_crawler2.txt': 'https://www.chick-fil-a.com/menu/breakfast'}
2025-05-11 17:21:05,728 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/menu/breakfast from assigned_urls[crawler2]
2025-05-11 17:21:05,784 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comcareers_crawler2.html': 'https://www.chick-fil-a.c
om/careers', 'crawl_data/www.chick-fil-a.comcareers_crawler2.txt': 'https://www.chick-fil-a.com/careers'}
2025-05-11 17:21:05,784 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/careers from assigned_urls[crawler2]
2025-05-11 17:21:05,880 [INFO] [Master] - Assigned URL https://www.facebook.com/ChickfilA/ at depth 1 to crawler crawler2
2025-05-11 17:21:05,893 [INFO] [Master] - Assigned URL https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1 at depth 1 to crawler crawler2
2025-05-11 17:21:05,907 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler2
2025-05-11 17:21:05,921 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1 to crawler crawler2
2025-05-11 17:21:05,935 [INFO] [Master] - Assigned URL https://example.com at depth 0 to crawler crawler2
2025-05-11 17:21:05,947 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1 to crawler crawler2
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

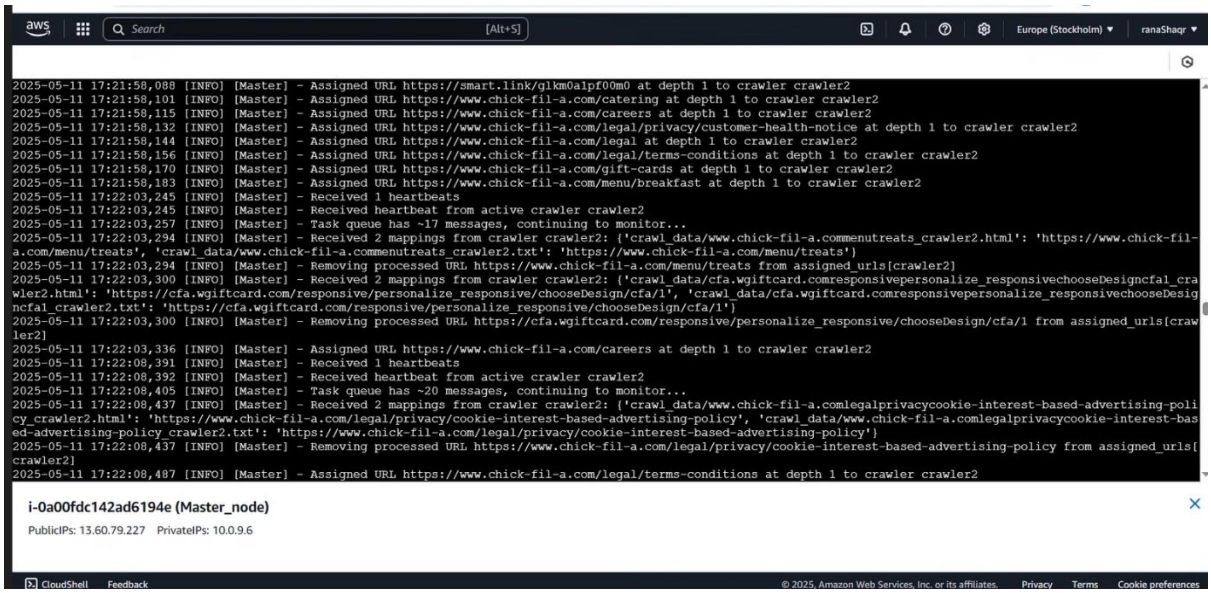

2025-05-11 17:21:05,947 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1 to crawler crawler2
2025-05-11 17:21:05,961 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1 to crawler crawler2
2025-05-11 17:21:05,996 [INFO] [Master] - Assigned URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 at depth 1 to crawler crawler2
2025-05-11 17:21:11,060 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:11,060 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:11,072 [INFO] [Master] - Task queue has ~31 messages, continuing to monitor...
2025-05-11 17:21:11,115 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenubeverages_crawler2.html': 'https://www.chick-fil-a.com/menu/beverages', 'crawl_data/www.chick-fil-a.commenubeverages_crawler2.txt': 'https://www.chick-fil-a.com/menu/beverages'}
2025-05-11 17:21:11,120 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/play.google.comstoreappsdetailsidcom.chickfila.playreferrerutm_source3Dweb26utm_campaign3Dcfaplay-web-footer_crawler2.html': 'https://play.google.com/store/apps/details?id=com.chickfila.play&referrer=utm_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer', 'crawl_data/play.google.comstoreappsdetailsidcom.chickfila.playreferrerutm_source3Dweb26utm_campaign3Dcfaplay-web-footer_crawler2.txt': 'https://play.google.com/store/apps/details?id=com.chickfila.play&referrer=utm_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer'}
2025-05-11 17:21:11,126 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/shop.chick-fil-a.com_crawler2.html': 'https://shop.chick-fil-a.com/', 'crawl_data/shop.chick-fil-a.com_crawler2.txt': 'https://shop.chick-fil-a.com/'}
2025-05-11 17:21:11,131 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenuchick-fil-a-chicken-sandwich_crawler2.html': 'https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich', 'crawl_data/www.chick-fil-a.commenuchick-fil-a-chicken-sandwich_crawler2.txt': 'https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich'}
2025-05-11 17:21:11,131 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich from assigned_urls[crawler2]
2025-05-11 17:21:11,136 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenusides_crawler2.html': 'https://www.chick-fil-a.com/menu/sides', 'crawl_data/www.chick-fil-a.commenusides_crawler2.txt': 'https://www.chick-fil-a.com/menu/sides'}
2025-05-11 17:21:11,141 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenufamily-style-meals_crawler2.html': 'https://www.chick-fil-a.com/menu/family-style-meals', 'crawl_data/www.chick-fil-a.commenufamily-style-meals_crawler2.txt': 'https://www.chick-fil-a.com/menu/family-style-meals'}
2025-05-11 17:21:11,146 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlegalsupply-chain_crawler2.html': 'https://www.chick-fil-a.com/legal/supply-chain', 'crawl_data/www.chick-fil-a.comlegalsupply-chain_crawler2.txt': 'https://www.chick-fil-a.com/legal/supply-chain'}
2025-05-11 17:21:11,147 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/legal/supply-chain from assigned_urls[crawler2]
2025-05-11 17:21:11,183 [INFO] [Master] - Assigned URL https://smart.link/g1km0a1pf00m0 at depth 1 to crawler crawler2
2025-05-11 17:21:11,198 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/locations/browse at depth 1 to crawler crawler2
2025-05-11 17:21:11,212 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/careers at depth 1 to crawler crawler2
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

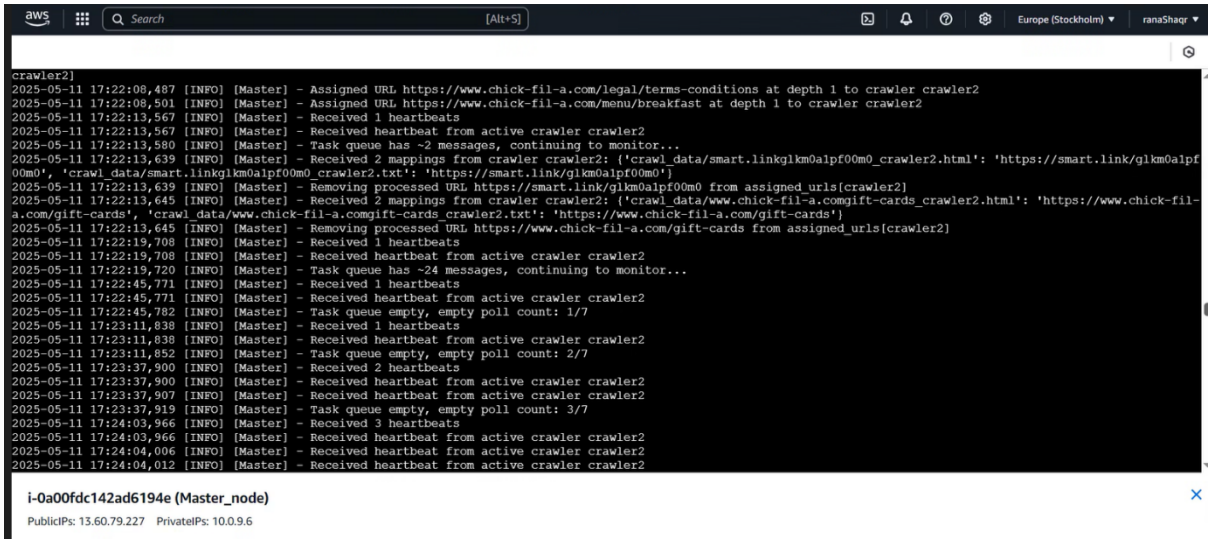

2025-05-11 17:21:11,212 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/careers at depth 1 to crawler crawler2
2025-05-11 17:21:11,227 [INFO] [Master] - Assigned URL https://www.facebook.com/ChickfilA/ at depth 1 to crawler crawler2
2025-05-11 17:21:11,242 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1 to crawler crawler2
2025-05-11 17:21:16,261 [WARNING] [Master] - Crawler crawler1 failed (no heartbeat for 120.66s), terminating
2025-05-11 17:21:16,270 [INFO] [Master] - Sent termination signal for failed crawler crawler1
2025-05-11 17:21:16,270 [INFO] [Master] - Crawler crawler1 terminated. Total completed: 1/3
2025-05-11 17:21:16,317 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:16,317 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:16,330 [INFO] [Master] - Task queue has ~36 messages, continuing to monitor...
2025-05-11 17:21:16,361 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/example.com_crawler2.html': 'https://example.com', 'crawl_data/example.com_crawler2.txt': 'https://example.com'}
2025-05-11 17:21:16,362 [INFO] [Master] - Removing processed URL https://example.com from assigned_urls[crawler2]
2025-05-11 17:21:16,399 [INFO] [Master] - Assigned URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 at depth 1 to crawler crawler2
2025-05-11 17:21:16,413 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1 to crawler crawler2
2025-05-11 17:21:16,428 [INFO] [Master] - Assigned URL https://www.iana.org/domains/example at depth 1 to crawler crawler2
2025-05-11 17:21:21,487 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:21,487 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:21,498 [INFO] [Master] - Task queue has ~30 messages, continuing to monitor...
2025-05-11 17:21:21,530 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlegalprivacycalifornia-privacy-policy_crawler2.html': 'https://www.chick-fil-a.com/legal/privacy/california-privacy-policy', 'crawl_data/www.chick-fil-a.comlegalprivacycalifornia-privacy-policy_crawler2.txt': 'https://www.chick-fil-a.com/legal/privacy/california-privacy-policy'}
2025-05-11 17:21:21,543 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth 1 to crawler crawler2
2025-05-11 17:21:21,557 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/customer-support at depth 1 to crawler crawler2
2025-05-11 17:21:21,573 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/treats at depth 1 to crawler crawler2
2025-05-11 17:21:21,586 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/sides at depth 1 to crawler crawler2
2025-05-11 17:21:21,600 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/breakfast at depth 1 to crawler crawler2
2025-05-11 17:21:21,615 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/nutrition-allergens at depth 1 to crawler crawler2
2025-05-11 17:21:21,629 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/about at depth 1 to crawler crawler2
i-0a00fdc142ad6194e (Master_node)

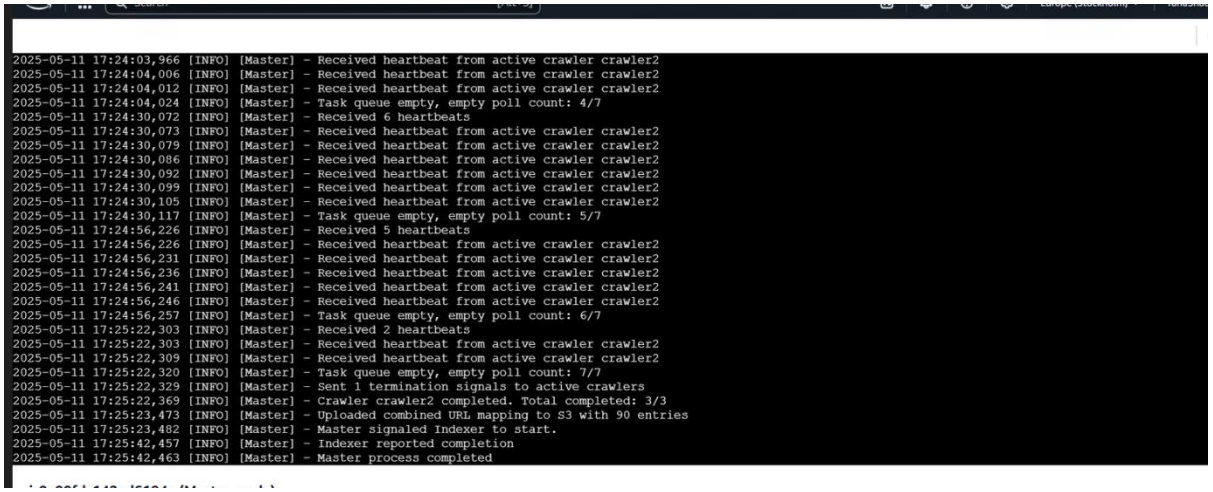

2025-05-11 17:21:21,643 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/gift-cards at depth 1 to crawler crawler2
2025-05-11 17:21:21,657 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy at depth 1 to crawler crawler2
2025-05-11 17:21:21,673 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal at depth 1 to crawler crawler2
2025-05-11 17:21:26,741 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:26,742 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:26,815 [INFO] [Master] - Task queue has ~29 messages, continuing to monitor...
2025-05-11 17:21:26,847 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/apps.apple.comappapple-storeid6449374451pt1119840ctcfaplay-web-footermt8_crawler2.html': 'https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8', 'crawl_data/apps.apple.comappapple-storeid6449374451pt1119840ctcfaplay-web-footermt8_crawler2.txt': 'https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8'}
2025-05-11 17:21:26,847 [INFO] [Master] - Removing processed URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 from assigned_urls[crawler2]
2025-05-11 17:21:26,853 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.iana.orgdomainsexample_crawler2.html': 'https://www.iana.org/domains/example', 'crawl_data/www.iana.orgdomainsexample_crawler2.txt': 'https://www.iana.org/domains/example'}
2025-05-11 17:21:26,853 [INFO] [Master] - Removing processed URL https://www.iana.org/domains/example from assigned_urls[crawler2]
2025-05-11 17:21:26,858 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlocationsbrowse_crawler2.html': 'https://www.chick-fil-a.com/locations/browse', 'crawl_data/www.chick-fil-a.comlocationsbrowse_crawler2.txt': 'https://www.chick-fil-a.com/locations/browse'}
2025-05-11 17:21:26,859 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/locations/browse from assigned_urls[crawler2]
2025-05-11 17:21:26,898 [INFO] [Master] - Assigned URL https://order.chick-fil-a.com/delivery/address at depth 1 to crawler crawler2
2025-05-11 17:21:26,913 [INFO] [Master] - Assigned URL https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1 at depth 1 to crawler crawler2
2025-05-11 17:21:26,928 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler2
2025-05-11 17:21:26,942 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1 to crawler crawler2
2025-05-11 17:21:26,958 [INFO] [Master] - Assigned URL https://www.facebook.com/ChickfilA/ at depth 1 to crawler crawler2
2025-05-11 17:21:26,972 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/customer-support at depth 1 to crawler crawler2
2025-05-11 17:21:26,985 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/breakfast at depth 1 to crawler crawler2
2025-05-11 17:21:27,000 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy at depth 1 to crawler crawler2
2025-05-11 17:21:32,061 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:32,061 [INFO] [Master] - Received heartbeat from active crawler crawler2


aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2
2025-05-11 17:21:32,061 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:32,061 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:32,073 [INFO] [Master] - Task queue has ~31 messages, continuing to monitor...
2025-05-11 17:21:32,109 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlegalaccessibilityaccessibility-legal_crawler2.ht
l': 'https://www.chick-fil-a.com/legal/accessibility/accessibility-legal', 'crawl_data/www.chick-fil-a.comlegalaccessibilityaccessibility-legal_crawler2.txt': 'https:/
/www.chick-fil-a.com/legal/accessibility/accessibility-legal'}
2025-05-11 17:21:32,109 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal from assigned_urls[crawler2]
2025-05-11 17:21:32,147 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler2
2025-05-11 17:21:32,161 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1 to crawler crawler2
2025-05-11 17:21:37,224 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:37,224 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:37,236 [INFO] [Master] - Task queue has ~26 messages, continuing to monitor...
2025-05-11 17:21:37,277 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlegal_crawler2.html': 'https://www.chick-fil-a.co
/legal', 'crawl_data/www.chick-fil-a.comlegal_crawler2.txt': 'https://www.chick-fil-a.com/legal'}
2025-05-11 17:21:37,277 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/legal from assigned_urls[crawler2]
2025-05-11 17:21:37,314 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal at depth 1 to crawler crawler2
2025-05-11 17:21:37,329 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler2
2025-05-11 17:21:42,400 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:42,400 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:21:42,413 [INFO] [Master] - Task queue has ~19 messages, continuing to monitor...
2025-05-11 17:21:42,444 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comnutrition-allergens_crawler2.html': 'https://www.
hick-fil-a.com/nutrition-allergens', 'crawl_data/www.chick-fil-a.comnutrition-allergens_crawler2.txt': 'https://www.chick-fil-a.com/nutrition-allergens'}
2025-05-11 17:21:42,444 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/nutrition-allergens from assigned_urls[crawler2]
2025-05-11 17:21:42,480 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/gift-cards at depth 1 to crawler crawler2
2025-05-11 17:21:42,496 [INFO] [Master] - Assigned URL https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1 at depth 1 to crawler crawler2
2025-05-11 17:21:42,511 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/breakfast at depth 1 to crawler crawler2
2025-05-11 17:21:47,573 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:21:47,573 [INFO] [Master] - Received heartbeat from active crawler crawler2
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

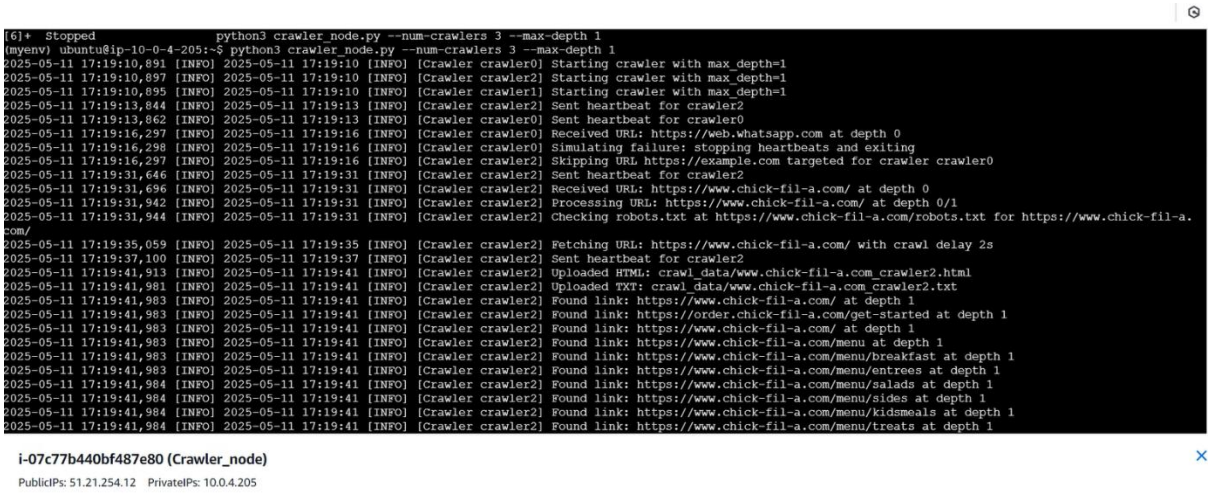

25-05-11 17:21:47,585 [INFO] [Master] - Task queue has ~26 messages, continuing to monitor...
25-05-11 17:21:47,616 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/order.chick-fil-a.comdeliveryaddress_crawler2.html': 'https://order.c
k-fil-a.com/delivery/address', 'crawl_data/order.chick-fil-a.comdeliveryaddress_crawler2.txt': 'https://order.chick-fil-a.com/delivery/address'}
25-05-11 17:21:47,616 [INFO] [Master] - Removing processed URL https://order.chick-fil-a.com/delivery/address from assigned_urls[crawler2]
25-05-11 17:21:47,621 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenusmokehouse-bbq-bacon_crawler2.html': 'https://
.chick-fil-a.com/menu/smokehouse-bbq-bacon', 'crawl_data/www.chick-fil-a.commenusmokehouse-bbq-bacon_crawler2.txt': 'https://www.chick-fil-a.com/menu/smokehouse-bbq
acon'}
25-05-11 17:21:47,621 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon from assigned_urls[crawler2]
25-05-11 17:21:47,658 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/breakfast at depth 1 to crawler crawler2
25-05-11 17:21:52,679 [WARNING] [Master] - Crawler crawler0 failed (no heartbeat for 120.31s), terminating
25-05-11 17:21:52,688 [INFO] [Master] - Sent termination signal for failed crawler crawler0
25-05-11 17:21:52,688 [INFO] [Master] - Crawler crawler0 terminated. Total completed: 2/3
25-05-11 17:21:52,741 [INFO] [Master] - Received 1 heartbeats
25-05-11 17:21:52,741 [INFO] [Master] - Received heartbeat from active crawler crawler2
25-05-11 17:21:52,754 [INFO] [Master] - Task queue has ~19 messages, continuing to monitor...
25-05-11 17:21:52,792 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comabout_crawler2.html': 'https://www.chick-fil-a.com
out', 'crawl_data/www.chick-fil-a.comabout_crawler2.txt': 'https://www.chick-fil-a.com/about'}
25-05-11 17:21:52,792 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/about from assigned_urls[crawler2]
25-05-11 17:21:52,799 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlegalterms-conditions_crawler2.html': 'https://www
hick-fil-a.com/legal/terms-conditions', 'crawl_data/www.chick-fil-a.comlegalterms-conditions_crawler2.txt': 'https://www.chick-fil-a.com/legal/terms-conditions'}
25-05-11 17:21:52,799 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/legal/terms-conditions from assigned_urls[crawler2]
25-05-11 17:21:52,860 [INFO] [Master] - Assigned URL https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1 at depth 1 to crawler crawler2
25-05-11 17:21:57,941 [INFO] [Master] - Received 1 heartbeats
25-05-11 17:21:57,941 [INFO] [Master] - Received heartbeat from active crawler crawler2
25-05-11 17:21:57,954 [INFO] [Master] - Task queue has ~16 messages, continuing to monitor...
25-05-11 17:21:57,995 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comcustomer-support_crawler2.html': 'https://www.chic
fil-a.com/customer-support', 'crawl_data/www.chick-fil-a.comcustomer-support_crawler2.txt': 'https://www.chick-fil-a.com/customer-support'}
25-05-11 17:21:57,996 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/customer-support from assigned_urls[crawler2]
25-05-11 17:21:58,088 [INFO] [Master] - Assigned URL https://smart.link/glkm0a1pf00m0 at depth 1 to crawler crawler2
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

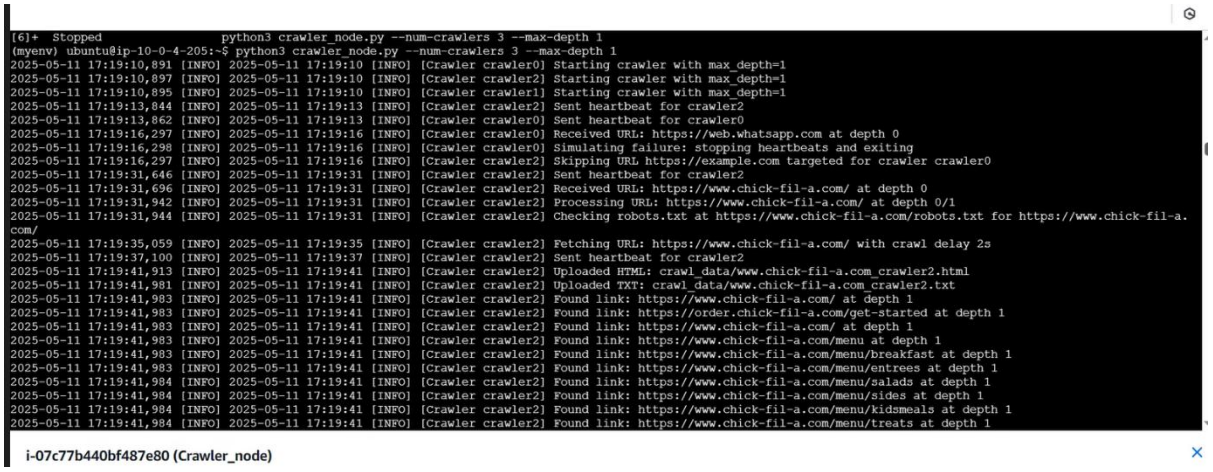

aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2025-05-11 17:21:58,088 [INFO] [Master] - Assigned URL https://smart.link/glkm0a1pf00m0 at depth 1 to crawler crawler2
2025-05-11 17:21:58,101 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/catering at depth 1 to crawler crawler2
2025-05-11 17:21:58,115 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/careers at depth 1 to crawler crawler2
2025-05-11 17:21:58,132 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth 1 to crawler crawler2
2025-05-11 17:21:58,144 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal at depth 1 to crawler crawler2
2025-05-11 17:21:58,156 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler2
2025-05-11 17:21:58,170 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/gift-cards at depth 1 to crawler crawler2
2025-05-11 17:21:58,183 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/breakfast at depth 1 to crawler crawler2
2025-05-11 17:22:03,245 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:22:03,245 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:22:03,257 [INFO] [Master] - Task queue has ~17 messages, continuing to monitor...
2025-05-11 17:22:03,294 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenutreats_crawler2.html': 'https://www.chick-fil-
a.com/menu/treats', 'crawl_data/www.chick-fil-a.commenutreats_crawler2.txt': 'https://www.chick-fil-a.com/menu/treats'}
2025-05-11 17:22:03,294 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/menu/treats from assigned_urls[crawler2]
2025-05-11 17:22:03,300 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/cfa.wgiftcard.comresponsivepersonalize_responsivechooseDesigncfal_cra
wler2.html': 'https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1', 'crawl_data/cfa.wgiftcard.comresponsivepersonalize_responsivechooseDesig
ncfal_crawler2.txt': 'https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1'}
2025-05-11 17:22:03,300 [INFO] [Master] - Removing processed URL https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1 from assigned_urls[craw
ler2]
2025-05-11 17:22:03,336 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/careers at depth 1 to crawler crawler2
2025-05-11 17:22:08,391 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:22:08,392 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:22:08,405 [INFO] [Master] - Task queue has ~20 messages, continuing to monitor...
2025-05-11 17:22:08,437 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comlegalprivacycookie-interest-based-advertising-poli
cy_crawler2.html': 'https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy', 'crawl_data/www.chick-fil-a.comlegalprivacycookie-interest-bas
ed-advertising-policy_crawler2.txt': 'https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy'}
2025-05-11 17:22:08,437 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy from assigned_urls[
crawler2]
2025-05-11 17:22:08,487 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler2
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

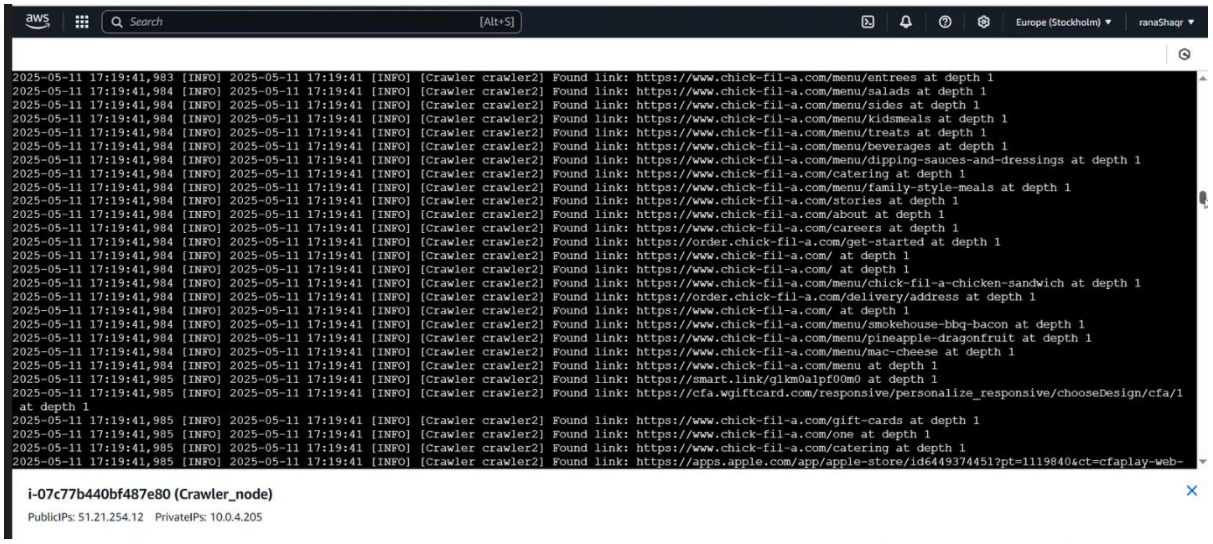

aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
crawler2]
2025-05-11 17:22:08,487 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/terms-conditions at depth 1 to crawler crawler2
2025-05-11 17:22:08,501 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/breakfast at depth 1 to crawler crawler2
2025-05-11 17:22:13,567 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:22:13,567 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:22:13,580 [INFO] [Master] - Task queue has ~2 messages, continuing to monitor...
2025-05-11 17:22:13,639 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/smart.linkglkm0alpf00m0_crawler2.html': 'https://smart.link/glkm0alpf00m0', 'crawl_data/smart.linkglkm0alpf00m0_crawler2.txt': 'https://smart.link/glkm0alpf00m0'}
2025-05-11 17:22:13,639 [INFO] [Master] - Removing processed URL https://smart.link/glkm0alpf00m0 from assigned_urls[crawler2]
2025-05-11 17:22:13,645 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comgift-cards_crawler2.html': 'https://www.chick-fil-a.com/gift-cards', 'crawl_data/www.chick-fil-a.comgift-cards_crawler2.txt': 'https://www.chick-fil-a.com/gift-cards'}
2025-05-11 17:22:13,645 [INFO] [Master] - Removing processed URL https://www.chick-fil-a.com/gift-cards from assigned_urls[crawler2]
2025-05-11 17:22:19,708 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:22:19,708 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:22:19,720 [INFO] [Master] - Task queue has ~24 messages, continuing to monitor...
2025-05-11 17:22:45,771 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:22:45,771 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:22:45,782 [INFO] [Master] - Task queue empty, empty poll count: 1/7
2025-05-11 17:23:11,838 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:23:11,838 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:23:11,852 [INFO] [Master] - Task queue empty, empty poll count: 2/7
2025-05-11 17:23:37,900 [INFO] [Master] - Received 2 heartbeats
2025-05-11 17:23:37,900 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:23:37,907 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:23:37,919 [INFO] [Master] - Task queue empty, empty poll count: 3/7
2025-05-11 17:24:03,966 [INFO] [Master] - Received 3 heartbeats
2025-05-11 17:24:03,966 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:04,006 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:04,012 [INFO] [Master] - Received heartbeat from active crawler crawler2
i-0a00fdc142ad6194e (Master_node)
PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

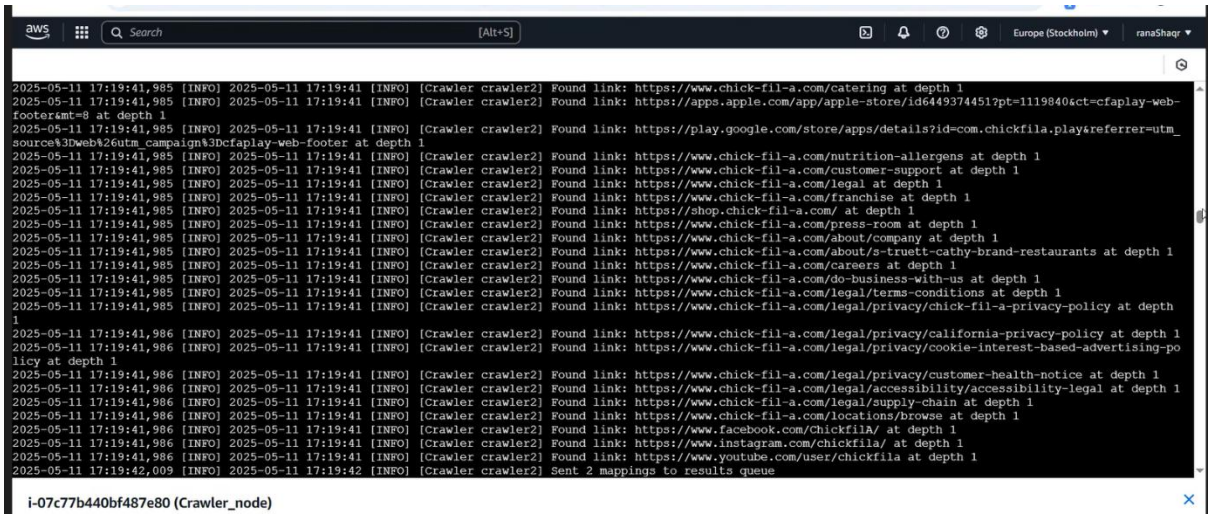

2025-05-11 17:24:03,966 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:04,006 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:04,012 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:04,024 [INFO] [Master] - Task queue empty, empty poll count: 4/7
2025-05-11 17:24:30,072 [INFO] [Master] - Received 6 heartbeats
2025-05-11 17:24:30,073 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:30,079 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:30,086 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:30,092 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:30,099 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:30,105 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:30,117 [INFO] [Master] - Task queue empty, empty poll count: 5/7
2025-05-11 17:24:56,226 [INFO] [Master] - Received 5 heartbeats
2025-05-11 17:24:56,226 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:56,231 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:56,236 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:56,241 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:56,246 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:24:56,257 [INFO] [Master] - Task queue empty, empty poll count: 6/7
2025-05-11 17:25:22,303 [INFO] [Master] - Received 2 heartbeats
2025-05-11 17:25:22,303 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:25:22,309 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:25:22,320 [INFO] [Master] - Task queue empty, empty poll count: 7/7
2025-05-11 17:25:22,329 [INFO] [Master] - Sent 1 termination signals to active crawlers
2025-05-11 17:25:22,369 [INFO] [Master] - Crawler crawler2 completed. Total completed: 3/3
2025-05-11 17:25:23,473 [INFO] [Master] - Uploaded combined URL mapping to S3 with 90 entries
2025-05-11 17:25:23,482 [INFO] [Master] - Master signaled Indexer to start.
2025-05-11 17:25:42,457 [INFO] [Master] - Indexer reported completion
2025-05-11 17:25:42,463 [INFO] [Master] - Master process completed

Crawler run


[6]+ Stopped python3 crawler_node.py --num-crawlers 3 --max-depth 1
(myenv) ubuntu@ip-10-0-4-205:~$ python3 crawler_node.py --num-crawlers 3 --max-depth 1
2025-05-11 17:19:10,891 [INFO] 2025-05-11 17:19:10 [INFO] [Crawler crawler0] Starting crawler with max_depth=1
2025-05-11 17:19:10,897 [INFO] 2025-05-11 17:19:10 [INFO] [Crawler crawler2] Starting crawler with max_depth=1
2025-05-11 17:19:10,895 [INFO] 2025-05-11 17:19:10 [INFO] [Crawler crawler1] Starting crawler with max_depth=1
2025-05-11 17:19:13,844 [INFO] 2025-05-11 17:19:13 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:13,862 [INFO] 2025-05-11 17:19:13 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-11 17:19:16,297 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler0] Received URL: https://web.whatsapp.com at depth 0
2025-05-11 17:19:16,298 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler0] Simulating failure: stopping heartbeats and exiting
2025-05-11 17:19:16,297 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler2] Skipping URL https://example.com targeted for crawler crawler0
2025-05-11 17:19:31,646 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:31,696 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/ at depth 0
2025-05-11 17:19:31,942 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/ at depth 0/1
2025-05-11 17:19:31,944 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/
2025-05-11 17:19:35,059 [INFO] 2025-05-11 17:19:35 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/ with crawl delay 2s
2025-05-11 17:19:37,100 [INFO] 2025-05-11 17:19:37 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:41,913 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.com_crawler2.html
2025-05-11 17:19:41,981 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.com_crawler2.txt
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://order.chick-fil-a.com/get-started at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/breakfast at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/entrees at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/salads at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/sides at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/kidsmeals at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/treats at depth 1
i-07c77b440bf487e80 (Crawler_node)
PublicIPs: 51.21.254.12 PrivateIPs: 10.0.4.205


[6]+ Stopped python3 crawler_node.py --num-crawlers 3 --max-depth 1
(myenv) ubuntu@ip-10-0-4-205:~$ python3 crawler_node.py --num-crawlers 3 --max-depth 1
2025-05-11 17:19:10,891 [INFO] 2025-05-11 17:19:10 [INFO] [Crawler crawler0] Starting crawler with max_depth=1
2025-05-11 17:19:10,897 [INFO] 2025-05-11 17:19:10 [INFO] [Crawler crawler2] Starting crawler with max_depth=1
2025-05-11 17:19:10,895 [INFO] 2025-05-11 17:19:10 [INFO] [Crawler crawler1] Starting crawler with max_depth=1
2025-05-11 17:19:13,844 [INFO] 2025-05-11 17:19:13 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:13,862 [INFO] 2025-05-11 17:19:13 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-11 17:19:16,297 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler0] Received URL: https://web.whatsapp.com at depth 0
2025-05-11 17:19:16,298 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler0] Simulating failure: stopping heartbeats and exiting
2025-05-11 17:19:16,297 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler2] Skipping URL https://example.com targeted for crawler crawler0
2025-05-11 17:19:31,646 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:31,696 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/ at depth 0
2025-05-11 17:19:31,942 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/ at depth 0/1
2025-05-11 17:19:31,944 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/
2025-05-11 17:19:35,059 [INFO] 2025-05-11 17:19:35 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/ with crawl delay 2s
2025-05-11 17:19:37,100 [INFO] 2025-05-11 17:19:37 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:41,913 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.com_crawler2.html
2025-05-11 17:19:41,981 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.com_crawler2.txt
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://order.chick-fil-a.com/get-started at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/breakfast at depth 1
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/entrees at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/salads at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/sides at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/kidsmeals at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/treats at depth 1
i-07c77b440bf487e80 (Crawler_node)


aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2025-05-11 17:19:41,983 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/entrees at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/salads at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/sides at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/kidsmeals at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/treats at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/beverages at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/family-style-meals at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/stories at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/about at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/careers at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://order.chick-fil-a.com/get-started at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://order.chick-fil-a.com/delivery/address at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/pineapple-dragonfruit at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu/mac-cheese at depth 1
2025-05-11 17:19:41,984 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://smart.link/glkm0alpf00m0 at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa/1 at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/gift-cards at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/one at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-
i-07c77b440bf487e80 (Crawler_node)
PublicIPs: 51.21.254.12 PrivateIPs: 10.0.4.205


aws
Search
[Alt+S]
Europe (Stockholm)
ranaShaqr
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://play.google.com/store/apps/details?id=com.chickfila.play&referrer=utm_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/nutrition-allergens at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/customer-support at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/franchise at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://shop.chick-fil-a.com/ at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/press-room at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/about/company at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/careers at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/do-business-with-us at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal/terms-conditions at depth 1
2025-05-11 17:19:41,985 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal/privacy/california-privacy-policy at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/legal/supply-chain at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.chick-fil-a.com/locations/browse at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.facebook.com/ChickfilA/ at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.instagram.com/chickfila/ at depth 1
2025-05-11 17:19:41,986 [INFO] 2025-05-11 17:19:41 [INFO] [Crawler crawler2] Found link: https://www.youtube.com/user/chickfila at depth 1
2025-05-11 17:19:42,009 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
i-07c77b440bf487e80 (Crawler_node)

```
2025-05-11 17:19:42,009 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:19:42,019 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://order.chick-fil-a.com/get-started at depth 1
2025-05-11 17:19:42,028 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:19:42,035 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/breakfast at depth 1
2025-05-11 17:19:42,042 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/entrees at depth 1
2025-05-11 17:19:42,050 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/salads at depth 1
2025-05-11 17:19:42,057 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/sides at depth 1
2025-05-11 17:19:42,067 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/kidsmeals at depth 1
2025-05-11 17:19:42,075 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/treats at depth 1
2025-05-11 17:19:42,081 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/beverages at depth 1
2025-05-11 17:19:42,088 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings at d
epth 1
2025-05-11 17:19:42,095 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:19:42,101 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/family-style-meals at depth 1
2025-05-11 17:19:42,109 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/stories at depth 1
2025-05-11 17:19:42,112 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:42,116 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/about at depth 1
2025-05-11 17:19:42,124 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/careers at depth 1
2025-05-11 17:19:42,131 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://order.chick-fil-a.com/get-started at depth 1
2025-05-11 17:19:42,139 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich at d
epth 1
2025-05-11 17:19:42,146 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://order.chick-fil-a.com/delivery/address at depth 1
2025-05-11 17:19:42,153 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1
2025-05-11 17:19:42,160 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/pineapple-dragonfruit at depth 1
2025-05-11 17:19:42,168 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu/mac-cheese at depth 1
2025-05-11 17:19:42,174 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:19:42,181 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://smart.link/glkm0a1pf00m0 at depth 1
2025-05-11 17:19:42,188 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://cfa.wgiftcard.com/responsive/personalize_responsive/choose
Design/cfa/1 at depth 1
```

i-07c77b440bf487e80 (Crawler_node)

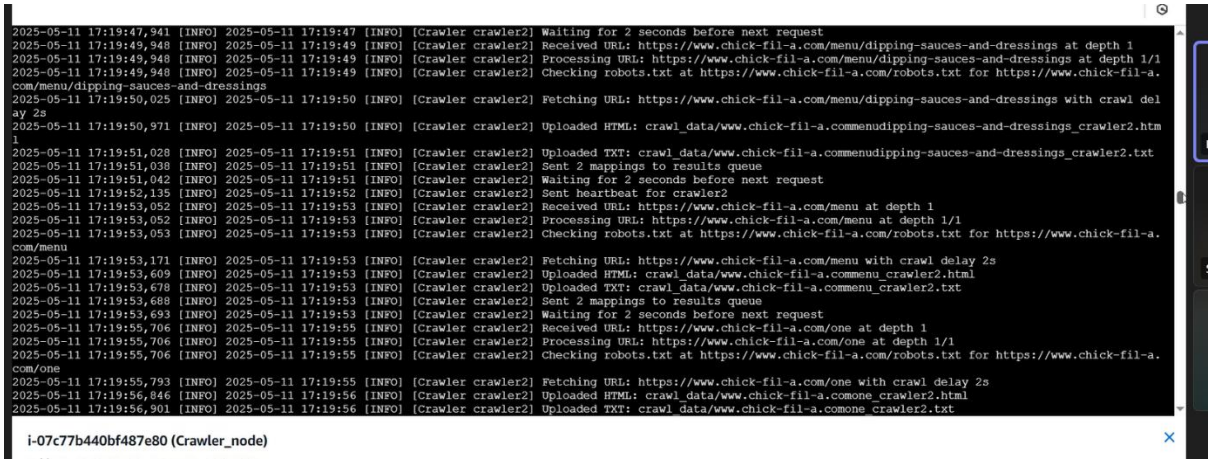
PublicIPs: 51.21.254.12 PrivateIPs: 10.0.4.205

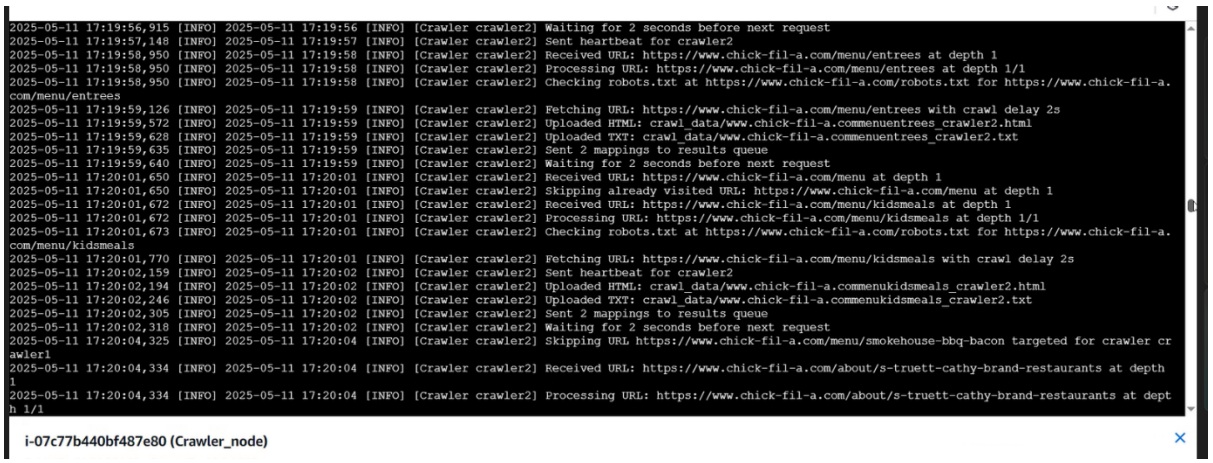
```
Design/cfa/1 at depth 1
2025-05-11 17:19:42,195 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/gift-cards at depth 1
2025-05-11 17:19:42,202 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/one at depth 1
2025-05-11 17:19:42,209 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:19:42,216 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=
cfaplay-web-footer&mt=8 at depth 1
2025-05-11 17:19:42,224 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://play.google.com/store/apps/details?id=com.chickfila.play&r
eferrer=utm_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer at depth 1
2025-05-11 17:19:42,230 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/nutrition-allergens at depth 1
2025-05-11 17:19:42,237 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/customer-support at depth 1
2025-05-11 17:19:42,243 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal at depth 1
2025-05-11 17:19:42,249 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/franchise at depth 1
2025-05-11 17:19:42,256 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://shop.chick-fil-a.com/ at depth 1
2025-05-11 17:19:42,264 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/press-room at depth 1
2025-05-11 17:19:42,271 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/about/company at depth 1
2025-05-11 17:19:42,277 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants
 at depth 1
2025-05-11 17:19:42,285 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/careers at depth 1
2025-05-11 17:19:42,292 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/do-business-with-us at depth 1
2025-05-11 17:19:42,301 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/terms-conditions at depth 1
2025-05-11 17:19:42,308 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-poli
cy at depth 1
2025-05-11 17:19:42,315 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/privacy/california-privacy-polic
y at depth 1
2025-05-11 17:19:42,323 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-ad
vertising-policy at depth 1
2025-05-11 17:19:42,330 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/privacy/customer-health-notice a
t depth 1
2025-05-11 17:19:42,336 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/accessibility/accessibility-lega
```

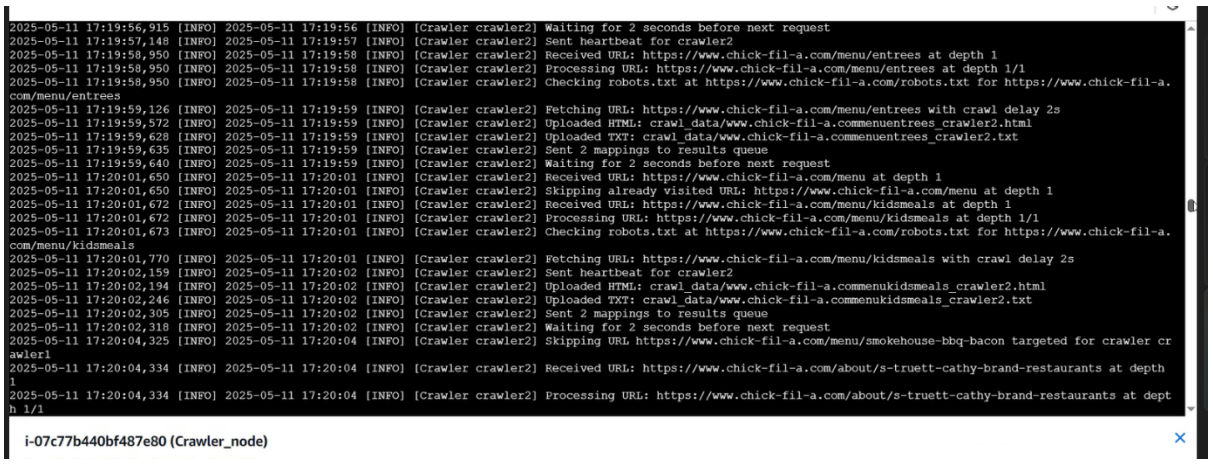
i-07c77b440bf487e80 (Crawler_node)

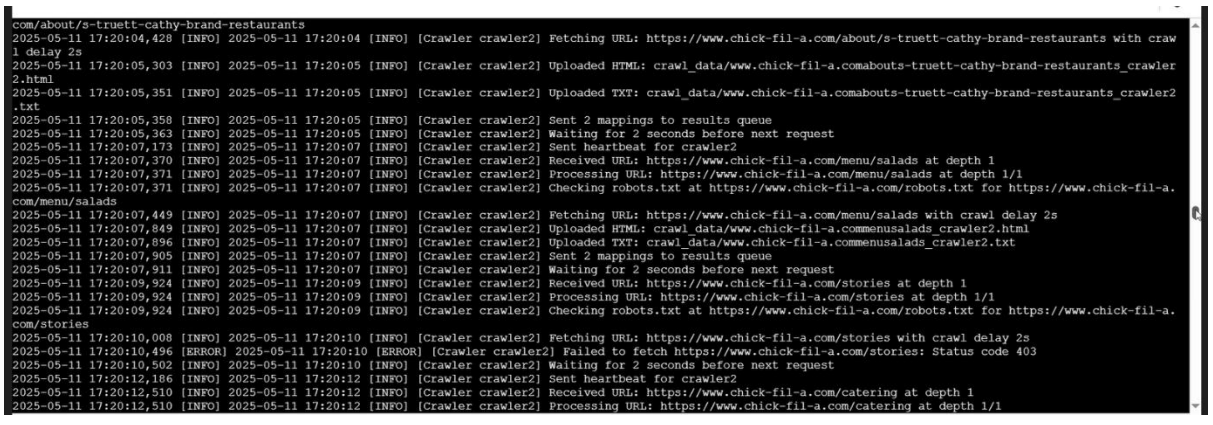
```
2025-05-11 17:19:42,336 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/accessibility/accessibility-lega
l at depth 1
2025-05-11 17:19:42,344 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/legal/supply-chain at depth 1
2025-05-11 17:19:42,352 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.chick-fil-a.com/locations/browse at depth 1
2025-05-11 17:19:42,359 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.facebook.com/ChickfilA/ at depth 1
2025-05-11 17:19:42,367 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.instagram.com/chickfila/ at depth 1
2025-05-11 17:19:42,374 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Added new URL to queue: https://www.youtube.com/user/chickfila at depth 1
2025-05-11 17:19:42,380 [INFO] 2025-05-11 17:19:42 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:44,388 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Received URL: https://order.chick-fil-a.com/get-started at depth 1
2025-05-11 17:19:44,388 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Processing URL: https://order.chick-fil-a.com/get-started at depth 1/1
2025-05-11 17:19:44,388 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Checking robots.txt at https://order.chick-fil-a.com/robots.txt for https://order.chick-fi
l-a.com/get-started
2025-05-11 17:19:44,666 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Fetching URL: https://order.chick-fil-a.com/get-started with crawl delay 2s
2025-05-11 17:19:44,771 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/order.chick-fil-a.comget-started_crawler2.html
2025-05-11 17:19:44,804 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/order.chick-fil-a.comget-started_crawler2.txt
2025-05-11 17:19:44,813 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:19:44,819 [INFO] 2025-05-11 17:19:44 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:46,825 [INFO] 2025-05-11 17:19:46 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/beverages at depth 1
2025-05-11 17:19:46,826 [INFO] 2025-05-11 17:19:46 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/beverages at depth 1/1
2025-05-11 17:19:46,826 [INFO] 2025-05-11 17:19:46 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/menu/beverages
2025-05-11 17:19:46,902 [INFO] 2025-05-11 17:19:46 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/beverages with crawl delay 2s
2025-05-11 17:19:47,124 [INFO] 2025-05-11 17:19:47 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:47,860 [INFO] 2025-05-11 17:19:47 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenubeverages_crawler2.html
2025-05-11 17:19:47,924 [INFO] 2025-05-11 17:19:47 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenubeverages_crawler2.txt
2025-05-11 17:19:47,937 [INFO] 2025-05-11 17:19:47 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:19:47,941 [INFO] 2025-05-11 17:19:47 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:49,948 [INFO] 2025-05-11 17:19:49 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings at depth 1
2025-05-11 17:19:49,948 [INFO] 2025-05-11 17:19:49 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings at depth 1/1
```

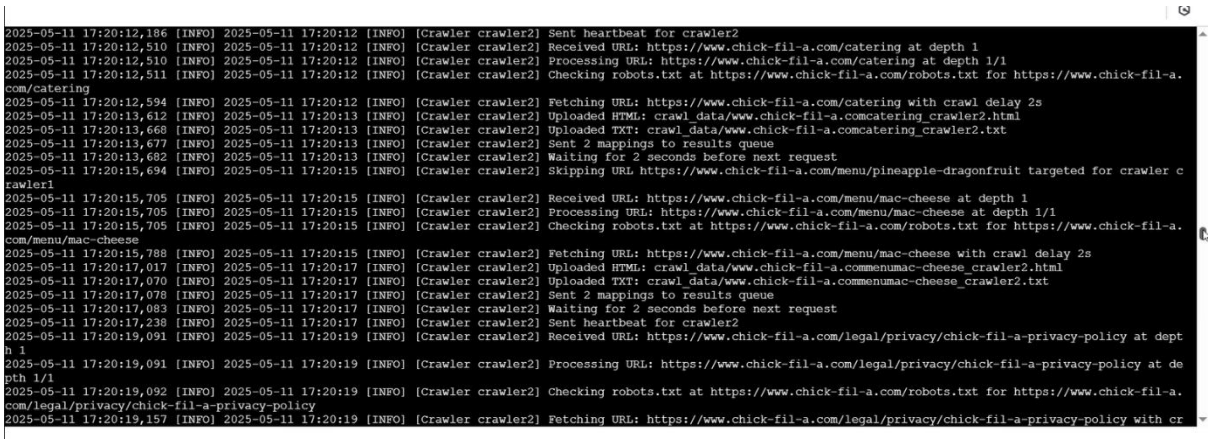
i-07c77b440bf487e80 (Crawler_node)


2025-05-11 17:19:47,941 [INFO] 2025-05-11 17:19:47 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:49,948 [INFO] 2025-05-11 17:19:49 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings at depth 1
2025-05-11 17:19:49,948 [INFO] 2025-05-11 17:19:49 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings at depth 1/1
2025-05-11 17:19:49,948 [INFO] 2025-05-11 17:19:49 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings
2025-05-11 17:19:50,025 [INFO] 2025-05-11 17:19:50 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/dipping-sauces-and-dressings with crawl delay 2s
2025-05-11 17:19:50,971 [INFO] 2025-05-11 17:19:50 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenudipping-sauces-and-dressings_crawler2.html
2025-05-11 17:19:51,028 [INFO] 2025-05-11 17:19:51 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenudipping-sauces-and-dressings_crawler2.txt
2025-05-11 17:19:51,038 [INFO] 2025-05-11 17:19:51 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:19:51,042 [INFO] 2025-05-11 17:19:51 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:52,135 [INFO] 2025-05-11 17:19:52 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:53,052 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:19:53,052 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu at depth 1/1
2025-05-11 17:19:53,053 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu
2025-05-11 17:19:53,171 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu with crawl delay 2s
2025-05-11 17:19:53,609 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenu_crawler2.html
2025-05-11 17:19:53,678 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenu_crawler2.txt
2025-05-11 17:19:53,688 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:19:53,693 [INFO] 2025-05-11 17:19:53 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:55,706 [INFO] 2025-05-11 17:19:55 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/one at depth 1
2025-05-11 17:19:55,706 [INFO] 2025-05-11 17:19:55 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/one at depth 1/1
2025-05-11 17:19:55,706 [INFO] 2025-05-11 17:19:55 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/one
2025-05-11 17:19:55,793 [INFO] 2025-05-11 17:19:55 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/one with crawl delay 2s
2025-05-11 17:19:56,846 [INFO] 2025-05-11 17:19:56 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comone_crawler2.html
2025-05-11 17:19:56,901 [INFO] 2025-05-11 17:19:56 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comone_crawler2.txt
i-07c77b440bf487e80 (Crawler_node)

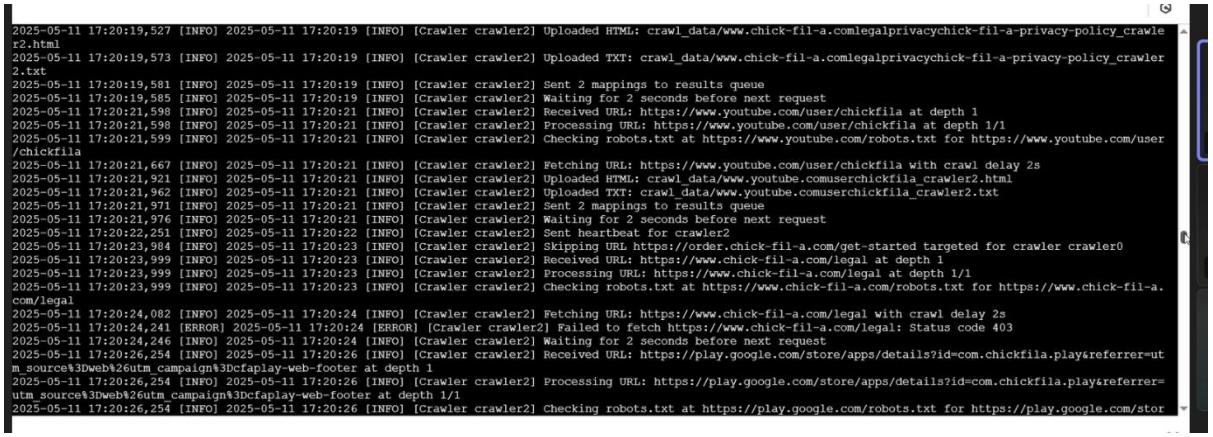

2025-05-11 17:19:56,915 [INFO] 2025-05-11 17:19:56 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:57,148 [INFO] 2025-05-11 17:19:57 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:58,950 [INFO] 2025-05-11 17:19:58 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/entrees at depth 1
2025-05-11 17:19:58,950 [INFO] 2025-05-11 17:19:58 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/entrees at depth 1/1
2025-05-11 17:19:58,950 [INFO] 2025-05-11 17:19:58 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/entrees
2025-05-11 17:19:59,126 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/entrees with crawl delay 2s
2025-05-11 17:19:59,572 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenuentrees_crawler2.html
2025-05-11 17:19:59,628 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenuentrees_crawler2.txt
2025-05-11 17:19:59,635 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:19:59,640 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:01,650 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:20:01,650 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:20:01,672 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/kidsmeals at depth 1
2025-05-11 17:20:01,672 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/kidsmeals at depth 1/1
2025-05-11 17:20:01,673 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/kidsmeals
2025-05-11 17:20:01,770 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/kidsmeals with crawl delay 2s
2025-05-11 17:20:02,159 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:02,194 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenukidsmeals_crawler2.html
2025-05-11 17:20:02,246 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenukidsmeals_crawler2.txt
2025-05-11 17:20:02,305 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:02,318 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:04,325 [INFO] 2025-05-11 17:20:04 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon targeted for crawler crawler1
2025-05-11 17:20:04,334 [INFO] 2025-05-11 17:20:04 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants at depth 1
2025-05-11 17:20:04,334 [INFO] 2025-05-11 17:20:04 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants at depth 1/1
i-07c77b440bf487e80 (Crawler_node)


2025-05-11 17:19:56,915 [INFO] 2025-05-11 17:19:56 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:19:57,148 [INFO] 2025-05-11 17:19:57 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:19:58,950 [INFO] 2025-05-11 17:19:58 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/entrees at depth 1
2025-05-11 17:19:58,950 [INFO] 2025-05-11 17:19:58 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/entrees at depth 1/1
2025-05-11 17:19:58,950 [INFO] 2025-05-11 17:19:58 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/entrees
2025-05-11 17:19:59,126 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/entrees with crawl delay 2s
2025-05-11 17:19:59,572 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenuentrees_crawler2.html
2025-05-11 17:19:59,628 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenuentrees_crawler2.txt
2025-05-11 17:19:59,635 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:19:59,640 [INFO] 2025-05-11 17:19:59 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:01,650 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:20:01,650 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/menu at depth 1
2025-05-11 17:20:01,672 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/kidsmeals at depth 1
2025-05-11 17:20:01,672 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/kidsmeals at depth 1/1
2025-05-11 17:20:01,673 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/kidsmeals
2025-05-11 17:20:01,770 [INFO] 2025-05-11 17:20:01 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/kidsmeals with crawl delay 2s
2025-05-11 17:20:02,159 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:02,194 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenukidsmeals_crawler2.html
2025-05-11 17:20:02,246 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenukidsmeals_crawler2.txt
2025-05-11 17:20:02,305 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:02,318 [INFO] 2025-05-11 17:20:02 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:04,325 [INFO] 2025-05-11 17:20:04 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon targeted for crawler crawler1
2025-05-11 17:20:04,334 [INFO] 2025-05-11 17:20:04 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants at depth 1
2025-05-11 17:20:04,334 [INFO] 2025-05-11 17:20:04 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants at depth 1/1
i-07c77b440bf487e80 (Crawler_node)

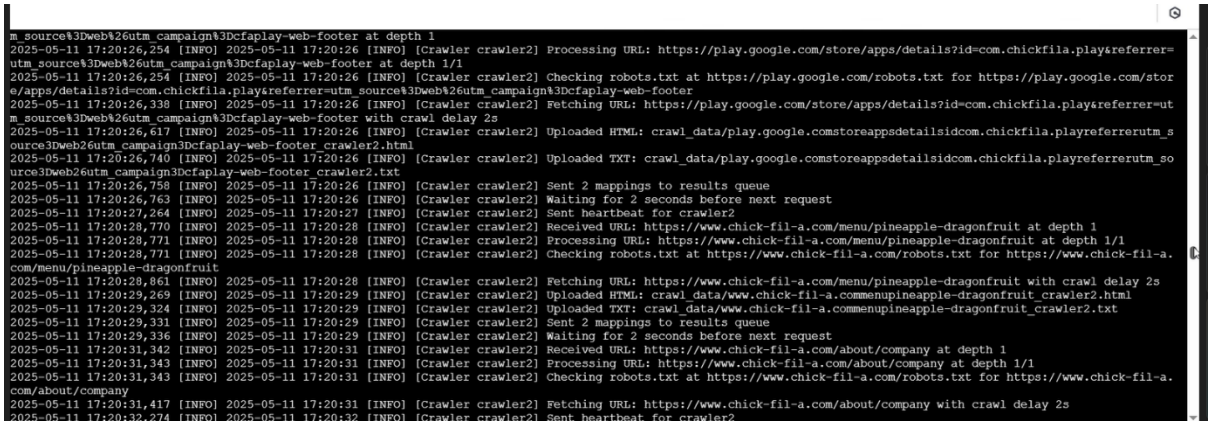

com/about/s-truett-cathy-brand-restaurants
2025-05-11 17:20:04,428 [INFO] 2025-05-11 17:20:04 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants with crawl delay 2s
2025-05-11 17:20:05,303 [INFO] 2025-05-11 17:20:05 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comabouts-truett-cathy-brand-restaurants_crawler2.html
2025-05-11 17:20:05,351 [INFO] 2025-05-11 17:20:05 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comabouts-truett-cathy-brand-restaurants_crawler2.txt
2025-05-11 17:20:05,358 [INFO] 2025-05-11 17:20:05 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:05,363 [INFO] 2025-05-11 17:20:05 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:07,173 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:07,370 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/salads at depth 1
2025-05-11 17:20:07,371 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/salads at depth 1/1
2025-05-11 17:20:07,371 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/salads
2025-05-11 17:20:07,449 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/salads with crawl delay 2s
2025-05-11 17:20:07,849 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenusalads_crawler2.html
2025-05-11 17:20:07,896 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenusalads_crawler2.txt
2025-05-11 17:20:07,905 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:07,911 [INFO] 2025-05-11 17:20:07 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:09,924 [INFO] 2025-05-11 17:20:09 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/stories at depth 1
2025-05-11 17:20:09,924 [INFO] 2025-05-11 17:20:09 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/stories at depth 1/1
2025-05-11 17:20:09,924 [INFO] 2025-05-11 17:20:09 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/stories
2025-05-11 17:20:10,008 [INFO] 2025-05-11 17:20:10 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/stories with crawl delay 2s
2025-05-11 17:20:10,496 [ERROR] 2025-05-11 17:20:10 [ERROR] [Crawler crawler2] Failed to fetch https://www.chick-fil-a.com/stories: Status code 403
2025-05-11 17:20:10,502 [INFO] 2025-05-11 17:20:10 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:12,186 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:12,510 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:20:12,510 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/catering at depth 1/1



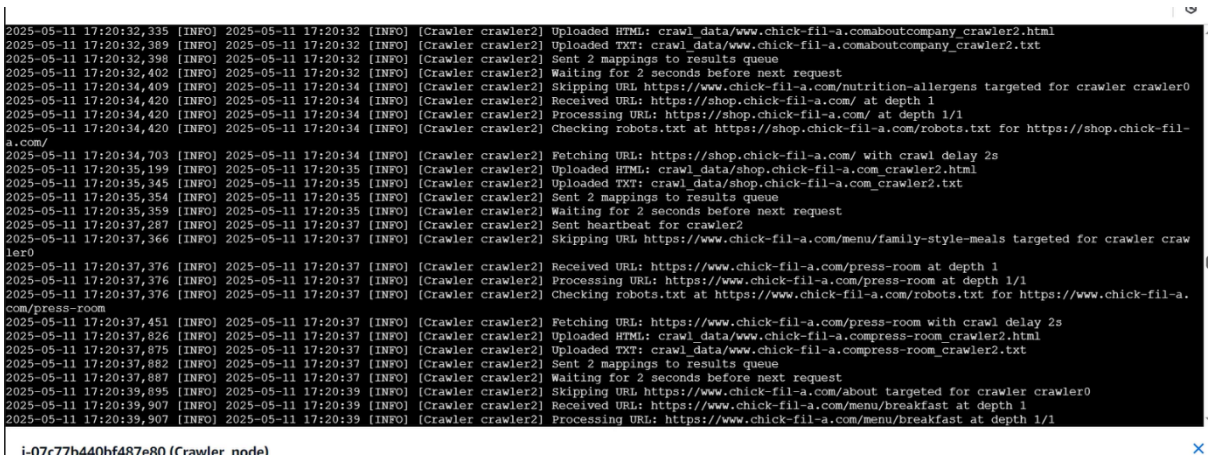
```
2025-05-11 17:20:12,186 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:12,510 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:20:12,510 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/catering at depth 1/1
2025-05-11 17:20:12,511 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/catering
2025-05-11 17:20:12,594 [INFO] 2025-05-11 17:20:12 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/catering with crawl delay 2s
2025-05-11 17:20:13,612 [INFO] 2025-05-11 17:20:13 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comcatering_crawler2.html
2025-05-11 17:20:13,668 [INFO] 2025-05-11 17:20:13 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comcatering_crawler2.txt
2025-05-11 17:20:13,677 [INFO] 2025-05-11 17:20:13 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:13,682 [INFO] 2025-05-11 17:20:13 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:15,694 [INFO] 2025-05-11 17:20:15 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/menu/pineapple-dragonfruit targeted for crawler c
rawler1
2025-05-11 17:20:15,705 [INFO] 2025-05-11 17:20:15 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/mac-cheese at depth 1
2025-05-11 17:20:15,705 [INFO] 2025-05-11 17:20:15 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/mac-cheese at depth 1/1
2025-05-11 17:20:15,705 [INFO] 2025-05-11 17:20:15 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/menu/mac-cheese
2025-05-11 17:20:15,788 [INFO] 2025-05-11 17:20:15 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/mac-cheese with crawl delay 2s
2025-05-11 17:20:17,017 [INFO] 2025-05-11 17:20:17 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.html
2025-05-11 17:20:17,070 [INFO] 2025-05-11 17:20:17 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.txt
2025-05-11 17:20:17,078 [INFO] 2025-05-11 17:20:17 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:17,083 [INFO] 2025-05-11 17:20:17 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:17,238 [INFO] 2025-05-11 17:20:17 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:19,091 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy at dept
h 1
2025-05-11 17:20:19,091 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy at de
pth 1/1
2025-05-11 17:20:19,092 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/legal/privacy/chick-fil-a-privacy-policy
2025-05-11 17:20:19,157 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy with cr
```



```
2025-05-11 17:20:19,527 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlegalprivacychick-fil-a-privacy-policy_crawle
r2.html
2025-05-11 17:20:19,573 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlegalprivacychick-fil-a-privacy-policy_crawler
2.txt
2025-05-11 17:20:19,581 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:19,585 [INFO] 2025-05-11 17:20:19 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:21,598 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Received URL: https://www.youtube.com/user/chickfila at depth 1
2025-05-11 17:20:21,598 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Processing URL: https://www.youtube.com/user/chickfila at depth 1/1
2025-05-11 17:20:21,599 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Checking robots.txt at https://www.youtube.com/robots.txt for https://www.youtube.com/user
/chickfila
2025-05-11 17:20:21,667 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Fetching URL: https://www.youtube.com/user/chickfila with crawl delay 2s
2025-05-11 17:20:21,921 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.youtube.comuserchickfila_crawler2.html
2025-05-11 17:20:21,962 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.youtube.comuserchickfila_crawler2.txt
2025-05-11 17:20:21,971 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:21,976 [INFO] 2025-05-11 17:20:21 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:22,251 [INFO] 2025-05-11 17:20:22 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:23,984 [INFO] 2025-05-11 17:20:23 [INFO] [Crawler crawler2] Skipping URL https://order.chick-fil-a.com/get-started targeted for crawler crawler0
2025-05-11 17:20:23,999 [INFO] 2025-05-11 17:20:23 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal at depth 1
2025-05-11 17:20:23,999 [INFO] 2025-05-11 17:20:23 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal at depth 1/1
2025-05-11 17:20:23,999 [INFO] 2025-05-11 17:20:23 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/legal
2025-05-11 17:20:24,082 [INFO] 2025-05-11 17:20:24 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal with crawl delay 2s
2025-05-11 17:20:24,241 [ERROR] 2025-05-11 17:20:24 [ERROR] [Crawler crawler2] Failed to fetch https://www.chick-fil-a.com/legal: Status code 403
2025-05-11 17:20:24,246 [INFO] 2025-05-11 17:20:24 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:26,254 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Received URL: https://play.google.com/store/apps/details?id=com.chickfila.play&referrer=ut
m_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer at depth 1
2025-05-11 17:20:26,254 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Processing URL: https://play.google.com/store/apps/details?id=com.chickfila.play&referrer=
utm_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer at depth 1/1
2025-05-11 17:20:26,254 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Checking robots.txt at https://play.google.com/robots.txt for https://play.google.com/stor
```

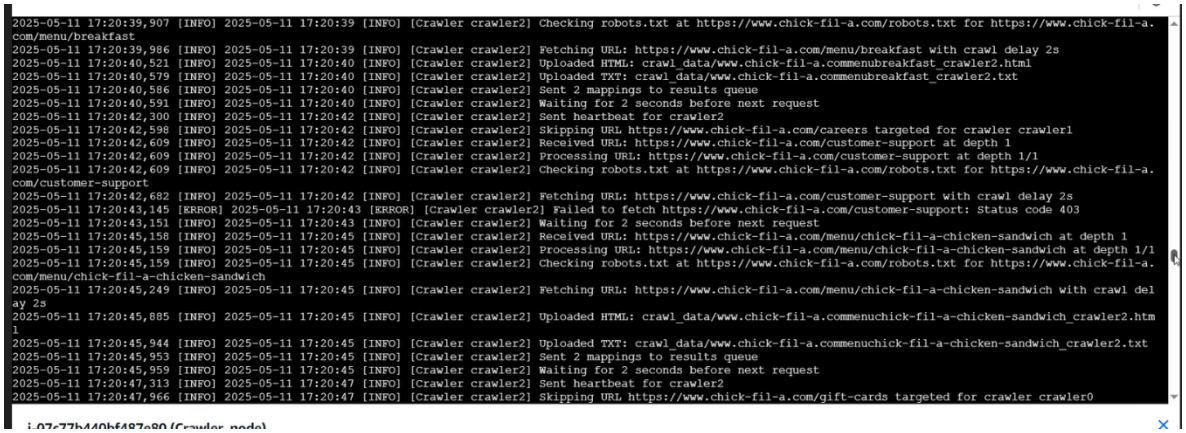


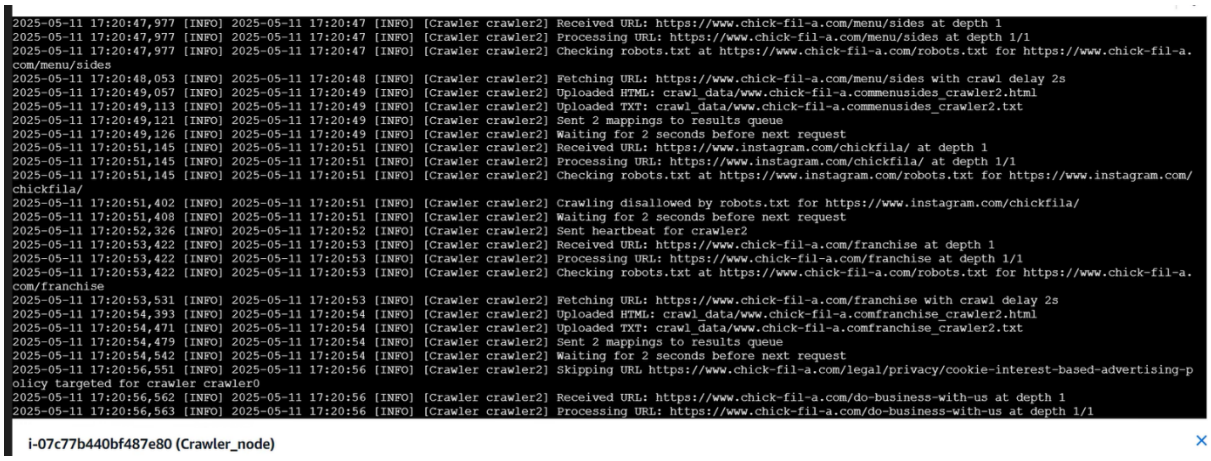
```
m_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer at depth 1
2025-05-11 17:20:26,254 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Processing URL: https://play.google.com/store/apps/details?id=com.chickfila.play&referrer=
utm_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer at depth 1/1
2025-05-11 17:20:26,254 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Checking robots.txt at https://play.google.com/robots.txt for https://play.google.com/stor
e/apps/details?id=com.chickfila.play&referrer=utm_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer
2025-05-11 17:20:26,338 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Fetching URL: https://play.google.com/store/apps/details?id=com.chickfila.play&referrer=ut
m_source%3Dweb%26utm_campaign%3Dcfaplay-web-footer with crawl delay 2s
2025-05-11 17:20:26,617 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/play.google.comstoreappsdetailsidcom.chickfila.playreferrerutm_s
ource3Dweb26utm_campaign3Dcfaplay-web-footer_crawler2.html
2025-05-11 17:20:26,740 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/play.google.comstoreappsdetailsidcom.chickfila.playreferrerutm_so
urce3Dweb26utm_campaign3Dcfaplay-web-footer_crawler2.txt
2025-05-11 17:20:26,758 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:26,763 [INFO] 2025-05-11 17:20:26 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:27,264 [INFO] 2025-05-11 17:20:27 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:28,770 [INFO] 2025-05-11 17:20:28 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/pineapple-dragonfruit at depth 1
2025-05-11 17:20:28,771 [INFO] 2025-05-11 17:20:28 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/pineapple-dragonfruit at depth 1/1
2025-05-11 17:20:28,771 [INFO] 2025-05-11 17:20:28 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/menu/pineapple-dragonfruit
2025-05-11 17:20:28,861 [INFO] 2025-05-11 17:20:28 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/pineapple-dragonfruit with crawl delay 2s
2025-05-11 17:20:29,269 [INFO] 2025-05-11 17:20:29 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenupineapple-dragonfruit_crawler2.html
2025-05-11 17:20:29,324 [INFO] 2025-05-11 17:20:29 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenupineapple-dragonfruit_crawler2.txt
2025-05-11 17:20:29,331 [INFO] 2025-05-11 17:20:29 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:29,336 [INFO] 2025-05-11 17:20:29 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:31,342 [INFO] 2025-05-11 17:20:31 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/about/company at depth 1
2025-05-11 17:20:31,343 [INFO] 2025-05-11 17:20:31 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/about/company at depth 1/1
2025-05-11 17:20:31,343 [INFO] 2025-05-11 17:20:31 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/about/company
2025-05-11 17:20:31,417 [INFO] 2025-05-11 17:20:31 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/about/company with crawl delay 2s
2025-05-11 17:20:32,274 [INFO] 2025-05-11 17:20:32 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
```

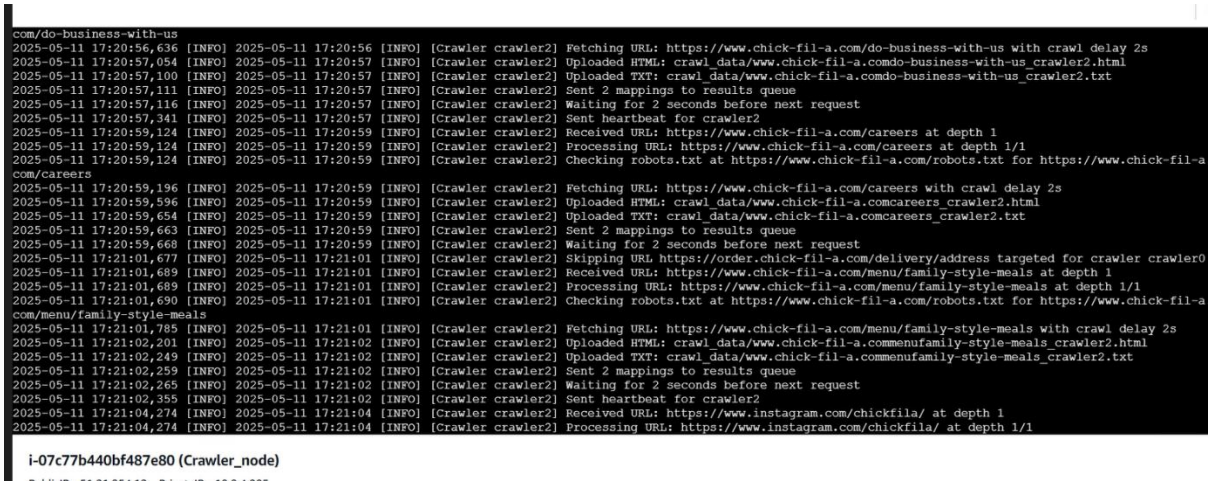


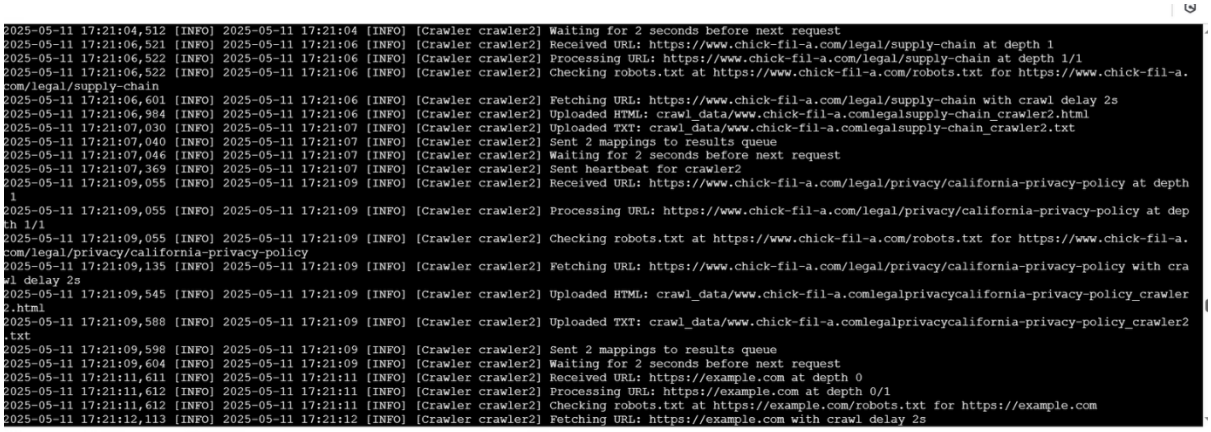
```
2025-05-11 17:20:32,335 [INFO] 2025-05-11 17:20:32 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comaboutcompany_crawler2.html
2025-05-11 17:20:32,389 [INFO] 2025-05-11 17:20:32 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comaboutcompany_crawler2.txt
2025-05-11 17:20:32,398 [INFO] 2025-05-11 17:20:32 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:32,402 [INFO] 2025-05-11 17:20:32 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:34,409 [INFO] 2025-05-11 17:20:34 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/nutrition-allergens targeted for crawler crawler0
2025-05-11 17:20:34,420 [INFO] 2025-05-11 17:20:34 [INFO] [Crawler crawler2] Received URL: https://shop.chick-fil-a.com/ at depth 1
2025-05-11 17:20:34,420 [INFO] 2025-05-11 17:20:34 [INFO] [Crawler crawler2] Processing URL: https://shop.chick-fil-a.com/ at depth 1/1
2025-05-11 17:20:34,420 [INFO] 2025-05-11 17:20:34 [INFO] [Crawler crawler2] Checking robots.txt at https://shop.chick-fil-a.com/robots.txt for https://shop.chick-fil-
a.com/
2025-05-11 17:20:34,703 [INFO] 2025-05-11 17:20:34 [INFO] [Crawler crawler2] Fetching URL: https://shop.chick-fil-a.com/ with crawl delay 2s
2025-05-11 17:20:35,199 [INFO] 2025-05-11 17:20:35 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/shop.chick-fil-a.com_crawler2.html
2025-05-11 17:20:35,345 [INFO] 2025-05-11 17:20:35 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/shop.chick-fil-a.com_crawler2.txt
2025-05-11 17:20:35,354 [INFO] 2025-05-11 17:20:35 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:35,359 [INFO] 2025-05-11 17:20:35 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:37,287 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:37,366 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/menu/family-style-meals targeted for crawler craw
ler0
2025-05-11 17:20:37,376 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/press-room at depth 1
2025-05-11 17:20:37,376 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/press-room at depth 1/1
2025-05-11 17:20:37,376 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/press-room
2025-05-11 17:20:37,451 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/press-room with crawl delay 2s
2025-05-11 17:20:37,826 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.compress-room_crawler2.html
2025-05-11 17:20:37,875 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.compress-room_crawler2.txt
2025-05-11 17:20:37,882 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:37,887 [INFO] 2025-05-11 17:20:37 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:39,895 [INFO] 2025-05-11 17:20:39 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/about targeted for crawler crawler0
2025-05-11 17:20:39,907 [INFO] 2025-05-11 17:20:39 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/breakfast at depth 1
2025-05-11 17:20:39,907 [INFO] 2025-05-11 17:20:39 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/breakfast at depth 1/1
```

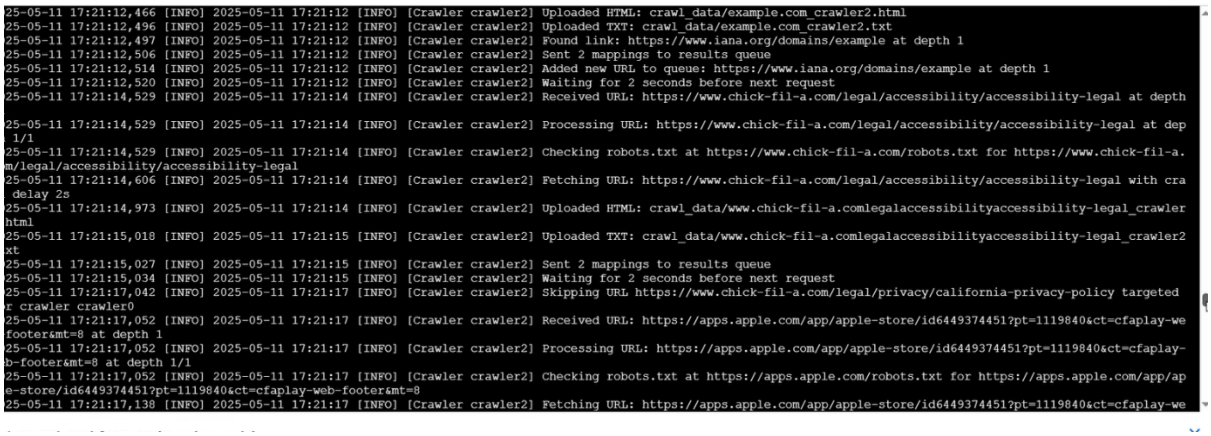
i-07c77b440bf487e80 (Crawler_node)


2025-05-11 17:20:39,907 [INFO] 2025-05-11 17:20:39 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/breakfast
2025-05-11 17:20:39,986 [INFO] 2025-05-11 17:20:39 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/breakfast with crawl delay 2s
2025-05-11 17:20:40,521 [INFO] 2025-05-11 17:20:40 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenubreakfast_crawler2.html
2025-05-11 17:20:40,579 [INFO] 2025-05-11 17:20:40 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenubreakfast_crawler2.txt
2025-05-11 17:20:40,586 [INFO] 2025-05-11 17:20:40 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:40,591 [INFO] 2025-05-11 17:20:40 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:42,300 [INFO] 2025-05-11 17:20:42 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:42,598 [INFO] 2025-05-11 17:20:42 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/careers targeted for crawler crawler1
2025-05-11 17:20:42,609 [INFO] 2025-05-11 17:20:42 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/customer-support at depth 1
2025-05-11 17:20:42,609 [INFO] 2025-05-11 17:20:42 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/customer-support at depth 1/1
2025-05-11 17:20:42,609 [INFO] 2025-05-11 17:20:42 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/customer-support
2025-05-11 17:20:42,682 [INFO] 2025-05-11 17:20:42 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/customer-support with crawl delay 2s
2025-05-11 17:20:43,145 [ERROR] 2025-05-11 17:20:43 [ERROR] [Crawler crawler2] Failed to fetch https://www.chick-fil-a.com/customer-support: Status code 403
2025-05-11 17:20:43,151 [INFO] 2025-05-11 17:20:43 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:45,158 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich at depth 1
2025-05-11 17:20:45,159 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich at depth 1/1
2025-05-11 17:20:45,159 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich
2025-05-11 17:20:45,249 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich with crawl delay 2s
2025-05-11 17:20:45,885 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenuchick-fil-a-chicken-sandwich_crawler2.html
2025-05-11 17:20:45,944 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenuchick-fil-a-chicken-sandwich_crawler2.txt
2025-05-11 17:20:45,953 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:45,959 [INFO] 2025-05-11 17:20:45 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:47,313 [INFO] 2025-05-11 17:20:47 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:47,966 [INFO] 2025-05-11 17:20:47 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/gift-cards targeted for crawler crawler0


2025-05-11 17:20:47,977 [INFO] 2025-05-11 17:20:47 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/sides at depth 1
2025-05-11 17:20:47,977 [INFO] 2025-05-11 17:20:47 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/sides at depth 1/1
2025-05-11 17:20:47,977 [INFO] 2025-05-11 17:20:47 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/sides
2025-05-11 17:20:48,053 [INFO] 2025-05-11 17:20:48 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/sides with crawl delay 2s
2025-05-11 17:20:49,057 [INFO] 2025-05-11 17:20:49 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenusides_crawler2.html
2025-05-11 17:20:49,113 [INFO] 2025-05-11 17:20:49 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenusides_crawler2.txt
2025-05-11 17:20:49,121 [INFO] 2025-05-11 17:20:49 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:49,126 [INFO] 2025-05-11 17:20:49 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:51,145 [INFO] 2025-05-11 17:20:51 [INFO] [Crawler crawler2] Received URL: https://www.instagram.com/chickfila/ at depth 1
2025-05-11 17:20:51,145 [INFO] 2025-05-11 17:20:51 [INFO] [Crawler crawler2] Processing URL: https://www.instagram.com/chickfila/ at depth 1/1
2025-05-11 17:20:51,145 [INFO] 2025-05-11 17:20:51 [INFO] [Crawler crawler2] Checking robots.txt at https://www.instagram.com/robots.txt for https://www.instagram.com/chickfila/
2025-05-11 17:20:51,402 [INFO] 2025-05-11 17:20:51 [INFO] [Crawler crawler2] Crawling disallowed by robots.txt for https://www.instagram.com/chickfila/
2025-05-11 17:20:51,408 [INFO] 2025-05-11 17:20:51 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:52,326 [INFO] 2025-05-11 17:20:52 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:53,422 [INFO] 2025-05-11 17:20:53 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/franchise at depth 1
2025-05-11 17:20:53,422 [INFO] 2025-05-11 17:20:53 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/franchise at depth 1/1
2025-05-11 17:20:53,422 [INFO] 2025-05-11 17:20:53 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/franchise
2025-05-11 17:20:53,531 [INFO] 2025-05-11 17:20:53 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/franchise with crawl delay 2s
2025-05-11 17:20:54,393 [INFO] 2025-05-11 17:20:54 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comfranchise_crawler2.html
2025-05-11 17:20:54,471 [INFO] 2025-05-11 17:20:54 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comfranchise_crawler2.txt
2025-05-11 17:20:54,479 [INFO] 2025-05-11 17:20:54 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:54,542 [INFO] 2025-05-11 17:20:54 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:56,551 [INFO] 2025-05-11 17:20:56 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy targeted for crawler crawler0
2025-05-11 17:20:56,562 [INFO] 2025-05-11 17:20:56 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/do-business-with-us at depth 1
2025-05-11 17:20:56,563 [INFO] 2025-05-11 17:20:56 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/do-business-with-us at depth 1/1
i-07c77b440bf487e80 (Crawler_node)

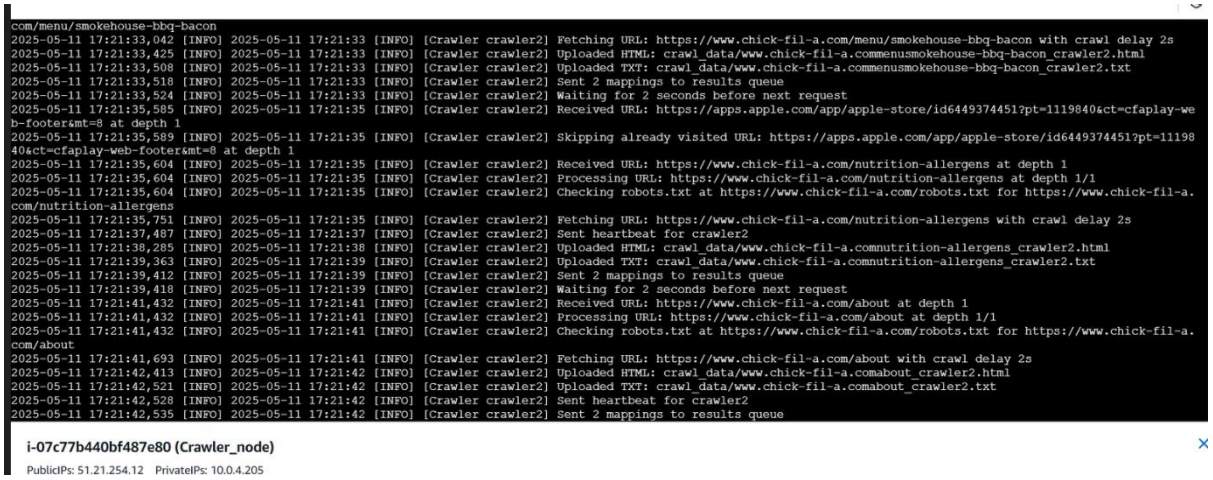

com/do-business-with-us
2025-05-11 17:20:56,636 [INFO] 2025-05-11 17:20:56 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/do-business-with-us with crawl delay 2s
2025-05-11 17:20:57,054 [INFO] 2025-05-11 17:20:57 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comdo-business-with-us_crawler2.html
2025-05-11 17:20:57,100 [INFO] 2025-05-11 17:20:57 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comdo-business-with-us_crawler2.txt
2025-05-11 17:20:57,111 [INFO] 2025-05-11 17:20:57 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:57,116 [INFO] 2025-05-11 17:20:57 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:20:57,341 [INFO] 2025-05-11 17:20:57 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:20:59,124 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/careers at depth 1
2025-05-11 17:20:59,124 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/careers at depth 1/1
2025-05-11 17:20:59,124 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/careers
2025-05-11 17:20:59,196 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/careers with crawl delay 2s
2025-05-11 17:20:59,596 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comcareers_crawler2.html
2025-05-11 17:20:59,654 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comcareers_crawler2.txt
2025-05-11 17:20:59,663 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:20:59,668 [INFO] 2025-05-11 17:20:59 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:01,677 [INFO] 2025-05-11 17:21:01 [INFO] [Crawler crawler2] Skipping URL https://order.chick-fil-a.com/delivery/address targeted for crawler crawler0
2025-05-11 17:21:01,689 [INFO] 2025-05-11 17:21:01 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/family-style-meals at depth 1
2025-05-11 17:21:01,689 [INFO] 2025-05-11 17:21:01 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/family-style-meals at depth 1/1
2025-05-11 17:21:01,690 [INFO] 2025-05-11 17:21:01 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/menu/family-style-meals
2025-05-11 17:21:01,785 [INFO] 2025-05-11 17:21:01 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/family-style-meals with crawl delay 2s
2025-05-11 17:21:02,201 [INFO] 2025-05-11 17:21:02 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenufamily-style-meals_crawler2.html
2025-05-11 17:21:02,249 [INFO] 2025-05-11 17:21:02 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenufamily-style-meals_crawler2.txt
2025-05-11 17:21:02,259 [INFO] 2025-05-11 17:21:02 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:02,265 [INFO] 2025-05-11 17:21:02 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:02,355 [INFO] 2025-05-11 17:21:02 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:04,274 [INFO] 2025-05-11 17:21:04 [INFO] [Crawler crawler2] Received URL: https://www.instagram.com/chickfila/ at depth 1
2025-05-11 17:21:04,274 [INFO] 2025-05-11 17:21:04 [INFO] [Crawler crawler2] Processing URL: https://www.instagram.com/chickfila/ at depth 1/1
i-07c77b440bf487e80 (Crawler_node)


2025-05-11 17:21:04,512 [INFO] 2025-05-11 17:21:04 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:06,521 [INFO] 2025-05-11 17:21:06 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/supply-chain at depth 1
2025-05-11 17:21:06,522 [INFO] 2025-05-11 17:21:06 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal/supply-chain at depth 1/1
2025-05-11 17:21:06,522 [INFO] 2025-05-11 17:21:06 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/legal/supply-chain
2025-05-11 17:21:06,601 [INFO] 2025-05-11 17:21:06 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal/supply-chain with crawl delay 2s
2025-05-11 17:21:06,984 [INFO] 2025-05-11 17:21:06 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlegalsupply-chain_crawler2.html
2025-05-11 17:21:07,030 [INFO] 2025-05-11 17:21:07 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlegalsupply-chain_crawler2.txt
2025-05-11 17:21:07,040 [INFO] 2025-05-11 17:21:07 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:07,046 [INFO] 2025-05-11 17:21:07 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:07,369 [INFO] 2025-05-11 17:21:07 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:09,055 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/privacy/california-privacy-policy at depth 1
2025-05-11 17:21:09,055 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal/privacy/california-privacy-policy at depth 1/1
2025-05-11 17:21:09,055 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/legal/privacy/california-privacy-policy
2025-05-11 17:21:09,135 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal/privacy/california-privacy-policy with crawl delay 2s
2025-05-11 17:21:09,545 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlegalprivacycalifornia-privacy-policy_crawler2.html
2025-05-11 17:21:09,588 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlegalprivacycalifornia-privacy-policy_crawler2.txt
2025-05-11 17:21:09,598 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:09,604 [INFO] 2025-05-11 17:21:09 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:11,611 [INFO] 2025-05-11 17:21:11 [INFO] [Crawler crawler2] Received URL: https://example.com at depth 0
2025-05-11 17:21:11,612 [INFO] 2025-05-11 17:21:11 [INFO] [Crawler crawler2] Processing URL: https://example.com at depth 0/1
2025-05-11 17:21:11,612 [INFO] 2025-05-11 17:21:11 [INFO] [Crawler crawler2] Checking robots.txt at https://example.com/robots.txt for https://example.com
2025-05-11 17:21:12,113 [INFO] 2025-05-11 17:21:12 [INFO] [Crawler crawler2] Fetching URL: https://example.com with crawl delay 2s
i-07c77b440bf487e80 (Crawler_node)


25-05-11 17:21:12,466 [INFO] 2025-05-11 17:21:12 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/example.com_crawler2.html
25-05-11 17:21:12,496 [INFO] 2025-05-11 17:21:12 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/example.com_crawler2.txt
25-05-11 17:21:12,497 [INFO] 2025-05-11 17:21:12 [INFO] [Crawler crawler2] Found link: https://www.iana.org/domains/example at depth 1
25-05-11 17:21:12,506 [INFO] 2025-05-11 17:21:12 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
25-05-11 17:21:12,514 [INFO] 2025-05-11 17:21:12 [INFO] [Crawler crawler2] Added new URL to queue: https://www.iana.org/domains/example at depth 1
25-05-11 17:21:12,520 [INFO] 2025-05-11 17:21:12 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
25-05-11 17:21:14,529 [INFO] 2025-05-11 17:21:14 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth
25-05-11 17:21:14,529 [INFO] 2025-05-11 17:21:14 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at dep 1/1
25-05-11 17:21:14,529 [INFO] 2025-05-11 17:21:14 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.m/legal/accessibility/accessibility-legal
25-05-11 17:21:14,606 [INFO] 2025-05-11 17:21:14 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal/accessibility/accessibility-legal with cra delay 2s
25-05-11 17:21:14,973 [INFO] 2025-05-11 17:21:14 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlegalaccessibilityaccessibility-legal_crawler html
25-05-11 17:21:15,018 [INFO] 2025-05-11 17:21:15 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlegalaccessibilityaccessibility-legal_crawler2 xt
25-05-11 17:21:15,027 [INFO] 2025-05-11 17:21:15 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
25-05-11 17:21:15,034 [INFO] 2025-05-11 17:21:15 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
25-05-11 17:21:17,042 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/legal/privacy/california-privacy-policy targeted r crawler crawler0
25-05-11 17:21:17,052 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Received URL: https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-we footer&mt=8 at depth 1
25-05-11 17:21:17,052 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Processing URL: https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-b-footer&mt=8 at depth 1/1
25-05-11 17:21:17,052 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Checking robots.txt at https://apps.apple.com/robots.txt for https://apps.apple.com/app/ape-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8
25-05-11 17:21:17,138 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Fetching URL: https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-we


2025-05-11 17:21:17,722 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/apps.apple.comappapple-storeid6449374451pt1119840ctcfaplay-web-footermt8_crawler2.html
2025-05-11 17:21:17,788 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/apps.apple.comappapple-storeid6449374451pt1119840ctcfaplay-web-footermt8_crawler2.txt
2025-05-11 17:21:17,798 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:17,803 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:19,812 [INFO] 2025-05-11 17:21:19 [INFO] [Crawler crawler2] Received URL: https://www.iana.org/domains/example at depth 1
2025-05-11 17:21:19,813 [INFO] 2025-05-11 17:21:19 [INFO] [Crawler crawler2] Processing URL: https://www.iana.org/domains/example at depth 1/1
2025-05-11 17:21:19,813 [INFO] 2025-05-11 17:21:19 [INFO] [Crawler crawler2] Checking robots.txt at https://www.iana.org/robots.txt for https://www.iana.org/domains/example
2025-05-11 17:21:20,273 [INFO] 2025-05-11 17:21:20 [INFO] [Crawler crawler2] Fetching URL: https://www.iana.org/domains/example with crawl delay 2s
2025-05-11 17:21:20,830 [INFO] 2025-05-11 17:21:20 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.iana.orgdomainsexample_crawler2.html
2025-05-11 17:21:20,868 [INFO] 2025-05-11 17:21:20 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.iana.orgdomainsexample_crawler2.txt
2025-05-11 17:21:20,878 [INFO] 2025-05-11 17:21:20 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:20,885 [INFO] 2025-05-11 17:21:20 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:22,409 [INFO] 2025-05-11 17:21:22 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:22,892 [INFO] 2025-05-11 17:21:22 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal at depth 1
2025-05-11 17:21:22,893 [INFO] 2025-05-11 17:21:22 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal at depth 1/1
2025-05-11 17:21:22,893 [INFO] 2025-05-11 17:21:22 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/legal
2025-05-11 17:21:22,998 [INFO] 2025-05-11 17:21:22 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal with crawl delay 2s
2025-05-11 17:21:23,383 [INFO] 2025-05-11 17:21:23 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlegal_crawler2.html
2025-05-11 17:21:23,430 [INFO] 2025-05-11 17:21:23 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlegal_crawler2.txt
2025-05-11 17:21:23,440 [INFO] 2025-05-11 17:21:23 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:23,445 [INFO] 2025-05-11 17:21:23 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:25,468 [INFO] 2025-05-11 17:21:25 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/locations/browse at depth 1
2025-05-11 17:21:25,470 [INFO] 2025-05-11 17:21:25 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/locations/browse at depth 1/1
2025-05-11 17:21:25,471 [INFO] 2025-05-11 17:21:25 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.com/locations/browse
i-07c77b440bf487e80 (Crawler_node)


com/locations/browse
2025-05-11 17:21:25,574 [INFO] 2025-05-11 17:21:25 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/locations/browse with crawl delay 2s
2025-05-11 17:21:25,968 [INFO] 2025-05-11 17:21:25 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlocationsbrowse_crawler2.html
2025-05-11 17:21:26,016 [INFO] 2025-05-11 17:21:26 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlocationsbrowse_crawler2.txt
2025-05-11 17:21:26,026 [INFO] 2025-05-11 17:21:26 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:26,031 [INFO] 2025-05-11 17:21:26 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:27,422 [INFO] 2025-05-11 17:21:27 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:28,039 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Received URL: https://order.chick-fil-a.com/delivery/address at depth 1
2025-05-11 17:21:28,039 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Processing URL: https://order.chick-fil-a.com/delivery/address at depth 1/1
2025-05-11 17:21:28,039 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Checking robots.txt at https://order.chick-fil-a.com/robots.txt for https://order.chick-fi
l-a.com/delivery/address
2025-05-11 17:21:28,141 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Fetching URL: https://order.chick-fil-a.com/delivery/address with crawl delay 2s
2025-05-11 17:21:28,238 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/order.chick-fil-a.comdeliveryaddress_crawler2.html
2025-05-11 17:21:28,270 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/order.chick-fil-a.comdeliveryaddress_crawler2.txt
2025-05-11 17:21:28,280 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:28,285 [INFO] 2025-05-11 17:21:28 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:30,294 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Received URL: https://smart.link/glkm0alpf00m0 at depth 1
2025-05-11 17:21:30,294 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Processing URL: https://smart.link/glkm0alpf00m0 at depth 1/1
2025-05-11 17:21:30,294 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Checking robots.txt at https://smart.link/robots.txt for https://smart.link/glkm0alpf00m0
2025-05-11 17:21:30,433 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Fetching URL: https://smart.link/glkm0alpf00m0 with crawl delay 2s
2025-05-11 17:21:30,887 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/smart.linkglkm0alpf00m0_crawler2.html
2025-05-11 17:21:30,934 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/smart.linkglkm0alpf00m0_crawler2.txt
2025-05-11 17:21:30,942 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:30,948 [INFO] 2025-05-11 17:21:30 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:32,435 [INFO] 2025-05-11 17:21:32 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:32,957 [INFO] 2025-05-11 17:21:32 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/locations/browse at depth 1
2025-05-11 17:21:32,958 [INFO] 2025-05-11 17:21:32 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/locations/browse at depth 1
2025-05-11 17:21:32,970 [INFO] 2025-05-11 17:21:32 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1
2025-05-11 17:21:32,970 [INFO] 2025-05-11 17:21:32 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon at depth 1/1
i-07c77b440bf487e80 (Crawler_node)


com/menu/smokehouse-bbq-bacon
2025-05-11 17:21:33,042 [INFO] 2025-05-11 17:21:33 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon with crawl delay 2s
2025-05-11 17:21:33,425 [INFO] 2025-05-11 17:21:33 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenusmokehouse-bbq-bacon_crawler2.html
2025-05-11 17:21:33,508 [INFO] 2025-05-11 17:21:33 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenusmokehouse-bbq-bacon_crawler2.txt
2025-05-11 17:21:33,518 [INFO] 2025-05-11 17:21:33 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:33,524 [INFO] 2025-05-11 17:21:33 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:35,585 [INFO] 2025-05-11 17:21:35 [INFO] [Crawler crawler2] Received URL: https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-we
b-footer&mt=8 at depth 1
2025-05-11 17:21:35,589 [INFO] 2025-05-11 17:21:35 [INFO] [Crawler crawler2] Skipping already visited URL: https://apps.apple.com/app/apple-store/id6449374451?pt=11198
40&ct=cfaplay-web-footer&mt=8 at depth 1
2025-05-11 17:21:35,604 [INFO] 2025-05-11 17:21:35 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/nutrition-allergens at depth 1
2025-05-11 17:21:35,604 [INFO] 2025-05-11 17:21:35 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/nutrition-allergens at depth 1/1
2025-05-11 17:21:35,604 [INFO] 2025-05-11 17:21:35 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/nutrition-allergens
2025-05-11 17:21:35,751 [INFO] 2025-05-11 17:21:35 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/nutrition-allergens with crawl delay 2s
2025-05-11 17:21:37,487 [INFO] 2025-05-11 17:21:37 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:38,285 [INFO] 2025-05-11 17:21:38 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comnutrition-allergens_crawler2.html
2025-05-11 17:21:39,363 [INFO] 2025-05-11 17:21:39 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comnutrition-allergens_crawler2.txt
2025-05-11 17:21:39,412 [INFO] 2025-05-11 17:21:39 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:39,418 [INFO] 2025-05-11 17:21:39 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:41,432 [INFO] 2025-05-11 17:21:41 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/about at depth 1
2025-05-11 17:21:41,432 [INFO] 2025-05-11 17:21:41 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/about at depth 1/1
2025-05-11 17:21:41,432 [INFO] 2025-05-11 17:21:41 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/about
2025-05-11 17:21:41,693 [INFO] 2025-05-11 17:21:41 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/about with crawl delay 2s
2025-05-11 17:21:42,413 [INFO] 2025-05-11 17:21:42 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comabout_crawler2.html
2025-05-11 17:21:42,521 [INFO] 2025-05-11 17:21:42 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comabout_crawler2.txt
2025-05-11 17:21:42,528 [INFO] 2025-05-11 17:21:42 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:42,535 [INFO] 2025-05-11 17:21:42 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
i-07c77b440bf487e80 (Crawler_node)
PublicIPs: 51.21.254.12 PrivateIPs: 10.0.4.205


2025-05-11 17:21:44,549 [INFO] 2025-05-11 17:21:44 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/legal/terms-conditions
2025-05-11 17:21:44,671 [INFO] 2025-05-11 17:21:44 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal/terms-conditions with crawl delay 2s
2025-05-11 17:21:45,502 [INFO] 2025-05-11 17:21:45 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlegalterms-conditions_crawler2.html
2025-05-11 17:21:45,559 [INFO] 2025-05-11 17:21:45 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlegalterms-conditions_crawler2.txt
2025-05-11 17:21:45,568 [INFO] 2025-05-11 17:21:45 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:45,573 [INFO] 2025-05-11 17:21:45 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:47,542 [INFO] 2025-05-11 17:21:47 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:47,582 [INFO] 2025-05-11 17:21:47 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/customer-support at depth 1
2025-05-11 17:21:47,583 [INFO] 2025-05-11 17:21:47 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/customer-support at depth 1/1
2025-05-11 17:21:47,583 [INFO] 2025-05-11 17:21:47 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/customer-support
2025-05-11 17:21:47,658 [INFO] 2025-05-11 17:21:47 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/customer-support with crawl delay 2s
2025-05-11 17:21:48,264 [INFO] 2025-05-11 17:21:48 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comcustomer-support_crawler2.html
2025-05-11 17:21:48,339 [INFO] 2025-05-11 17:21:48 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comcustomer-support_crawler2.txt
2025-05-11 17:21:48,463 [INFO] 2025-05-11 17:21:48 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:48,470 [INFO] 2025-05-11 17:21:48 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:50,486 [INFO] 2025-05-11 17:21:50 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/treats at depth 1
2025-05-11 17:21:50,487 [INFO] 2025-05-11 17:21:50 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/menu/treats at depth 1/1
2025-05-11 17:21:50,487 [INFO] 2025-05-11 17:21:50 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/menu/treats
2025-05-11 17:21:51,773 [INFO] 2025-05-11 17:21:51 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/menu/treats with crawl delay 2s
2025-05-11 17:21:52,409 [INFO] 2025-05-11 17:21:52 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.commenutreats_crawler2.html
2025-05-11 17:21:52,514 [INFO] 2025-05-11 17:21:52 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.commenutreats_crawler2.txt
2025-05-11 17:21:52,787 [INFO] 2025-05-11 17:21:52 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:52,795 [INFO] 2025-05-11 17:21:52 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:52,798 [INFO] 2025-05-11 17:21:52 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:54,806 [INFO] 2025-05-11 17:21:54 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/sides at depth 1
2025-05-11 17:21:54,806 [INFO] 2025-05-11 17:21:54 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/menu/sides at depth 1
i-07c77b440bf487e80 (Crawler_node)
PublicIPs: 51.21.254.12 PrivateIPs: 10.0.4.205

```
2025-05-11 17:21:54,850 [INFO] 2025-05-11 17:21:54 [INFO] [Crawler crawler2] Received URL: https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa
/1 at depth 1
2025-05-11 17:21:54,850 [INFO] 2025-05-11 17:21:54 [INFO] [Crawler crawler2] Processing URL: https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/c
fa/1 at depth 1/1
2025-05-11 17:21:54,851 [INFO] 2025-05-11 17:21:54 [INFO] [Crawler crawler2] Checking robots.txt at https://cfa.wgiftcard.com/robots.txt for https://cfa.wgiftcard.com/
responsive/personalize_responsive/chooseDesign/cfa/1
2025-05-11 17:21:55,354 [INFO] 2025-05-11 17:21:55 [INFO] [Crawler crawler2] Fetching URL: https://cfa.wgiftcard.com/responsive/personalize_responsive/chooseDesign/cfa
/1 with crawl delay 2s
2025-05-11 17:21:56,322 [INFO] 2025-05-11 17:21:56 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/cfa.wgiftcard.comresponsivepersonalize_responsivechooseDesigncfa
1_crawler2.html
2025-05-11 17:21:56,367 [INFO] 2025-05-11 17:21:56 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/cfa.wgiftcard.comresponsivepersonalize_responsivechooseDesigncfa1
_crawler2.txt
2025-05-11 17:21:56,376 [INFO] 2025-05-11 17:21:56 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:21:56,381 [INFO] 2025-05-11 17:21:56 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:21:57,850 [INFO] 2025-05-11 17:21:57 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:21:58,388 [INFO] 2025-05-11 17:21:58 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:21:58,388 [INFO] 2025-05-11 17:21:58 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/catering at depth 1
2025-05-11 17:21:58,398 [INFO] 2025-05-11 17:21:58 [INFO] [Crawler crawler2] Received URL: https://www.facebook.com/ChickfilA/ at depth 1
2025-05-11 17:21:58,399 [INFO] 2025-05-11 17:21:58 [INFO] [Crawler crawler2] Processing URL: https://www.facebook.com/ChickfilA/ at depth 1/1
2025-05-11 17:21:58,399 [INFO] 2025-05-11 17:21:58 [INFO] [Crawler crawler2] Checking robots.txt at https://www.facebook.com/robots.txt for https://www.facebook.com/Ch
ickfilA/
2025-05-11 17:21:58,649 [INFO] 2025-05-11 17:21:58 [INFO] [Crawler crawler2] Crawling disallowed by robots.txt for https://www.facebook.com/ChickfilA/
2025-05-11 17:21:58,661 [INFO] 2025-05-11 17:21:58 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:22:00,715 [INFO] 2025-05-11 17:22:00 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-
policy at depth 1
2025-05-11 17:22:00,719 [INFO] 2025-05-11 17:22:00 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertisin
g-policy at depth 1/1
2025-05-11 17:22:00,719 [INFO] 2025-05-11 17:22:00 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/legal/privacy/cookie-interest-based-advertising-policy
```

i-07c77b440bf487e80 (Crawler_node)

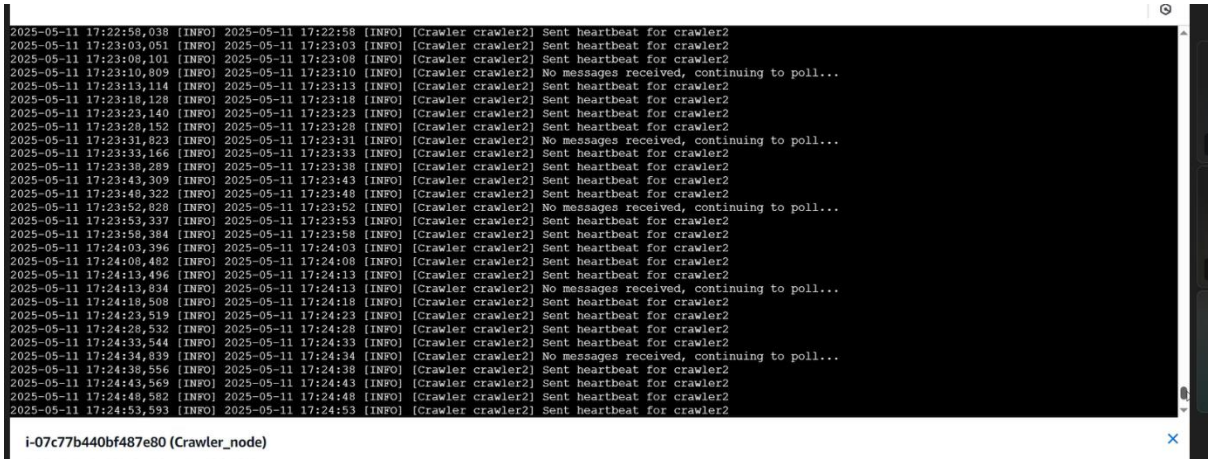
```
om/legal/privacy/cookie-interest-based-advertising-policy
2025-05-11 17:22:00,909 [INFO] 2025-05-11 17:22:00 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-
olicy with crawl delay 2s
2025-05-11 17:22:01,403 [INFO] 2025-05-11 17:22:01 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comlegalprivacycookie-interest-based-advertising
-policy_crawler2.html
2025-05-11 17:22:01,448 [INFO] 2025-05-11 17:22:01 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comlegalprivacycookie-interest-based-advertising-
olicy_crawler2.txt
2025-05-11 17:22:01,463 [INFO] 2025-05-11 17:22:01 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:22:01,468 [INFO] 2025-05-11 17:22:01 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:22:02,864 [INFO] 2025-05-11 17:22:02 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:03,476 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Received URL: https://smart.link/glkm0a1pf00m0 at depth 1
2025-05-11 17:22:03,477 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Skipping already visited URL: https://smart.link/glkm0a1pf00m0 at depth 1
2025-05-11 17:22:03,487 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/careers at depth 1
2025-05-11 17:22:03,487 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/careers at depth 1
2025-05-11 17:22:03,496 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/accessibility/accessibility-legal at depth
1
2025-05-11 17:22:03,497 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/legal/accessibility/accessibilit
-legal at depth 1
2025-05-11 17:22:03,507 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal at depth 1
2025-05-11 17:22:03,507 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/legal at depth 1
2025-05-11 17:22:03,531 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth 1
2025-05-11 17:22:03,532 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/legal/privacy/customer-health-notice at depth
/1
2025-05-11 17:22:03,532 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
om/legal/privacy/customer-health-notice
2025-05-11 17:22:03,613 [INFO] 2025-05-11 17:22:03 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/legal/privacy/customer-health-notice with crawl
elay 2s
2025-05-11 17:22:04,124 [ERROR] 2025-05-11 17:22:04 [ERROR] [Crawler crawler2] Failed to fetch https://www.chick-fil-a.com/legal/privacy/customer-health-notice: Status
code 403
```

i-07c77b440bf487e80 (Crawler_node)

```
2025-05-11 17:22:06,161 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/gift-cards at depth 1
2025-05-11 17:22:06,162 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Processing URL: https://www.chick-fil-a.com/gift-cards at depth 1/1
2025-05-11 17:22:06,162 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Checking robots.txt at https://www.chick-fil-a.com/robots.txt for https://www.chick-fil-a.
com/gift-cards
2025-05-11 17:22:06,240 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Fetching URL: https://www.chick-fil-a.com/gift-cards with crawl delay 2s
2025-05-11 17:22:06,638 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/www.chick-fil-a.comgift-cards_crawler2.html
2025-05-11 17:22:06,688 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/www.chick-fil-a.comgift-cards_crawler2.txt
2025-05-11 17:22:06,696 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-11 17:22:06,700 [INFO] 2025-05-11 17:22:06 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-11 17:22:07,876 [INFO] 2025-05-11 17:22:07 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:08,723 [INFO] 2025-05-11 17:22:08 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/legal/terms-conditions at depth 1
2025-05-11 17:22:08,723 [INFO] 2025-05-11 17:22:08 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/legal/terms-conditions at depth
1
2025-05-11 17:22:08,759 [INFO] 2025-05-11 17:22:08 [INFO] [Crawler crawler2] Received URL: https://www.chick-fil-a.com/menu/breakfast at depth 1
2025-05-11 17:22:08,759 [INFO] 2025-05-11 17:22:08 [INFO] [Crawler crawler2] Skipping already visited URL: https://www.chick-fil-a.com/menu/breakfast at depth 1
2025-05-11 17:22:12,889 [INFO] 2025-05-11 17:22:12 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:17,902 [INFO] 2025-05-11 17:22:17 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:22,914 [INFO] 2025-05-11 17:22:22 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:27,923 [INFO] 2025-05-11 17:22:27 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:28,769 [INFO] 2025-05-11 17:22:28 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:22:32,936 [INFO] 2025-05-11 17:22:32 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:37,986 [INFO] 2025-05-11 17:22:37 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:43,000 [INFO] 2025-05-11 17:22:43 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:48,012 [INFO] 2025-05-11 17:22:48 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:49,774 [INFO] 2025-05-11 17:22:49 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:22:53,025 [INFO] 2025-05-11 17:22:53 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:22:58,038 [INFO] 2025-05-11 17:22:58 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:03,051 [INFO] 2025-05-11 17:23:03 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:08,101 [INFO] 2025-05-11 17:23:08 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
```

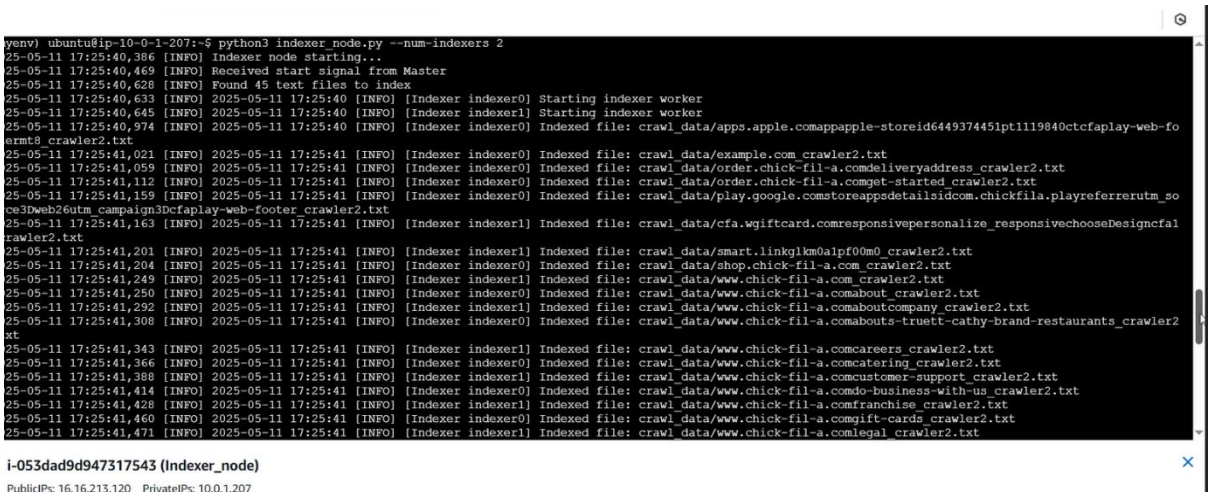
i-07c77b440bf487e80 (Crawler_node)

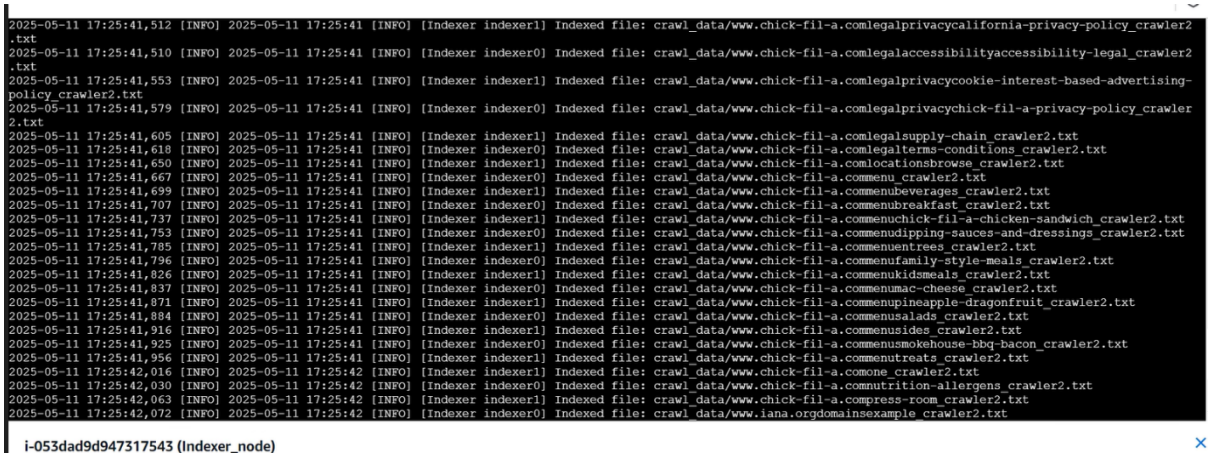
PublicIPs: 51.21.254.12 PrivateIPs: 10.0.4.205

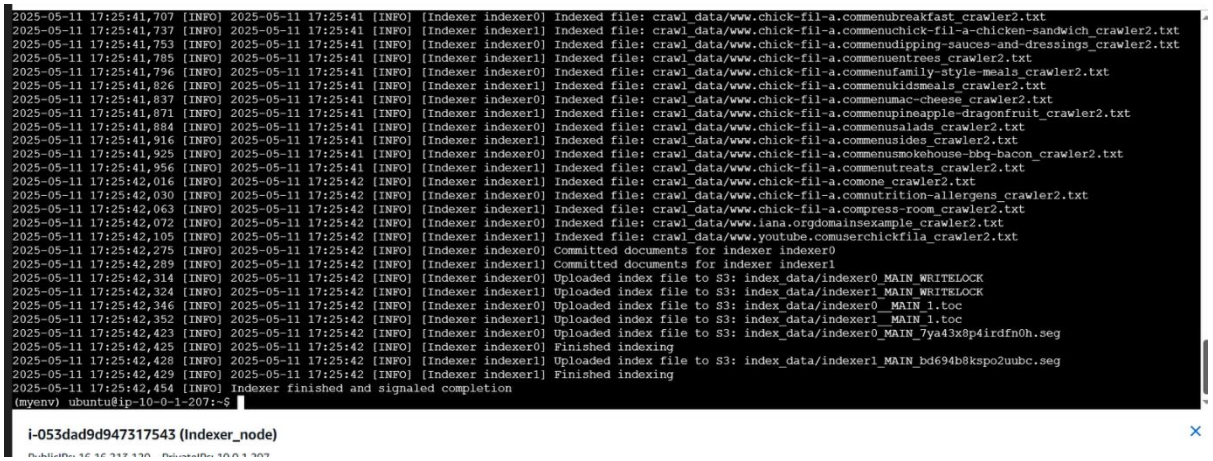

2025-05-11 17:22:58,038 [INFO] 2025-05-11 17:22:58 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:03,051 [INFO] 2025-05-11 17:23:03 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:08,101 [INFO] 2025-05-11 17:23:08 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:10,809 [INFO] 2025-05-11 17:23:10 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:23:13,114 [INFO] 2025-05-11 17:23:13 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:18,128 [INFO] 2025-05-11 17:23:18 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:23,140 [INFO] 2025-05-11 17:23:23 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:28,152 [INFO] 2025-05-11 17:23:28 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:31,823 [INFO] 2025-05-11 17:23:31 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:23:33,166 [INFO] 2025-05-11 17:23:33 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:38,289 [INFO] 2025-05-11 17:23:38 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:43,309 [INFO] 2025-05-11 17:23:43 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:48,322 [INFO] 2025-05-11 17:23:48 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:52,828 [INFO] 2025-05-11 17:23:52 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:23:53,337 [INFO] 2025-05-11 17:23:53 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:23:58,384 [INFO] 2025-05-11 17:23:58 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:03,396 [INFO] 2025-05-11 17:24:03 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:08,482 [INFO] 2025-05-11 17:24:08 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:13,496 [INFO] 2025-05-11 17:24:13 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:13,834 [INFO] 2025-05-11 17:24:13 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:24:18,508 [INFO] 2025-05-11 17:24:18 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:23,519 [INFO] 2025-05-11 17:24:23 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:28,532 [INFO] 2025-05-11 17:24:28 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:33,544 [INFO] 2025-05-11 17:24:33 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:34,839 [INFO] 2025-05-11 17:24:34 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:24:38,556 [INFO] 2025-05-11 17:24:38 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:43,569 [INFO] 2025-05-11 17:24:43 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:48,582 [INFO] 2025-05-11 17:24:48 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:53,593 [INFO] 2025-05-11 17:24:53 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
i-07c77b440bf487e80 (Crawler_node)


2025-05-11 17:24:48,582 [INFO] 2025-05-11 17:24:48 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:53,593 [INFO] 2025-05-11 17:24:53 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:24:55,845 [INFO] 2025-05-11 17:24:55 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:24:58,638 [INFO] 2025-05-11 17:24:58 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:25:03,750 [INFO] 2025-05-11 17:25:03 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:25:08,799 [INFO] 2025-05-11 17:25:08 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:25:13,814 [INFO] 2025-05-11 17:25:13 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:25:16,850 [INFO] 2025-05-11 17:25:16 [INFO] [Crawler crawler2] No messages received, continuing to poll...
2025-05-11 17:25:18,826 [INFO] 2025-05-11 17:25:18 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-11 17:25:22,330 [INFO] 2025-05-11 17:25:22 [INFO] [Crawler crawler2] Received termination signal, sending mappings and stopping
(myenv) ubuntu@ip-10-0-4-205:~$
i-07c77b440bf487e80 (Crawler_node)

Indexer run

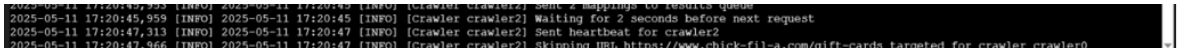

yenv) ubuntu@ip-10-0-1-207:~$ python3 indexer_node.py --num-indexers 2
25-05-11 17:25:40,386 [INFO] Indexer node starting...
25-05-11 17:25:40,469 [INFO] Received start signal from Master
25-05-11 17:25:40,628 [INFO] Found 45 text files to index
25-05-11 17:25:40,633 [INFO] 2025-05-11 17:25:40 [INFO] [Indexer indexer0] Starting indexer worker
25-05-11 17:25:40,645 [INFO] 2025-05-11 17:25:40 [INFO] [Indexer indexer1] Starting indexer worker
25-05-11 17:25:40,974 [INFO] 2025-05-11 17:25:40 [INFO] [Indexer indexer0] Indexed file: crawl_data/apps.apple.comappapple-storeid6449374451pt1119840ctcfaplay-web-fo
ermt8_crawler2.txt
25-05-11 17:25:41,021 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/example.com_crawler2.txt
25-05-11 17:25:41,059 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/order.chick-fil-a.comdeliveryaddress_crawler2.txt
25-05-11 17:25:41,112 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/order.chick-fil-a.comget-started_crawler2.txt
25-05-11 17:25:41,159 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/play.google.comstoreappsdetailsidcom.chickfila.playreferrerutm_so
ce3Dweb26utm_campaign3Dcfaplay-web-footer_crawler2.txt
25-05-11 17:25:41,163 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/cfa.wgiftcard.comresponsivepersonalize_responsivechooseDesigncfa1
rawler2.txt
25-05-11 17:25:41,201 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/smart.linkg1km0a1pf00m0_crawler2.txt
25-05-11 17:25:41,204 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/shop.chick-fil-a.com_crawler2.txt
25-05-11 17:25:41,249 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.com_crawler2.txt
25-05-11 17:25:41,250 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comabout_crawler2.txt
25-05-11 17:25:41,292 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comaboutcompany_crawler2.txt
25-05-11 17:25:41,308 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comabouts-truett-cathy-brand-restaurants_crawler2
xt
25-05-11 17:25:41,343 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comcareers_crawler2.txt
25-05-11 17:25:41,366 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comcatering_crawler2.txt
25-05-11 17:25:41,388 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comcustomer-support_crawler2.txt
25-05-11 17:25:41,414 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comdo-business-with-us_crawler2.txt
25-05-11 17:25:41,428 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comfranchise_crawler2.txt
25-05-11 17:25:41,460 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comgift-cards_crawler2.txt
25-05-11 17:25:41,471 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comlegal_crawler2.txt
i-053dad9d947317543 (Indexer_node)
PublicIPs: 16.16.213.120 PrivateIPs: 10.0.1.207

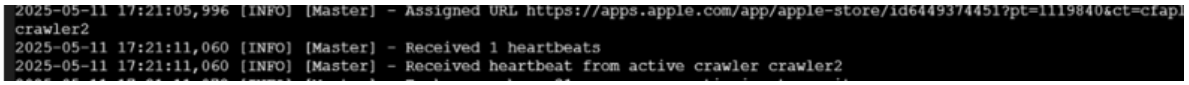

2025-05-11 17:25:41,512 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comlegalprivacycalifornia-privacy-policy_crawler2
.txt
2025-05-11 17:25:41,510 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comlegalaccessibilityaccessibility-legal_crawler2
.txt
2025-05-11 17:25:41,553 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comlegalprivacycookie-interest-based-advertising-
policy_crawler2.txt
2025-05-11 17:25:41,579 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comlegalprivacychick-fil-a-privacy-policy_crawler
2.txt
2025-05-11 17:25:41,605 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comlegalsupply-chain_crawler2.txt
2025-05-11 17:25:41,618 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comlegalterms-conditions_crawler2.txt
2025-05-11 17:25:41,650 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comlocationsbrowse_crawler2.txt
2025-05-11 17:25:41,667 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.commenu_crawler2.txt
2025-05-11 17:25:41,699 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.commenubeverages_crawler2.txt
2025-05-11 17:25:41,707 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.commenubreakfast_crawler2.txt
2025-05-11 17:25:41,737 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.commenuchick-fil-a-chicken-sandwich_crawler2.txt
2025-05-11 17:25:41,753 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.commenudipping-sauces-and-dressings_crawler2.txt
2025-05-11 17:25:41,785 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.commenuentrees_crawler2.txt
2025-05-11 17:25:41,796 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.commenufamily-style-meals_crawler2.txt
2025-05-11 17:25:41,826 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.commenukidsmeals_crawler2.txt
2025-05-11 17:25:41,837 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.txt
2025-05-11 17:25:41,871 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.commenupineapple-dragonfruit_crawler2.txt
2025-05-11 17:25:41,884 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.commenusalads_crawler2.txt
2025-05-11 17:25:41,916 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.commenusides_crawler2.txt
2025-05-11 17:25:41,925 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.commenusmokehouse-bbq-bacon_crawler2.txt
2025-05-11 17:25:41,956 [INFO] 2025-05-11 17:25:41 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.commenutreats_crawler2.txt
2025-05-11 17:25:42,016 [INFO] 2025-05-11 17:25:42 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.comone_crawler2.txt
2025-05-11 17:25:42,030 [INFO] 2025-05-11 17:25:42 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.chick-fil-a.comnutrition-allergens_crawler2.txt
2025-05-11 17:25:42,063 [INFO] 2025-05-11 17:25:42 [INFO] [Indexer indexer1] Indexed file: crawl_data/www.chick-fil-a.compress-room_crawler2.txt
2025-05-11 17:25:42,072 [INFO] 2025-05-11 17:25:42 [INFO] [Indexer indexer0] Indexed file: crawl_data/www.iana.orgdomainsexample_crawler2.txt
i-053dad9d947317543 (Indexer_node)



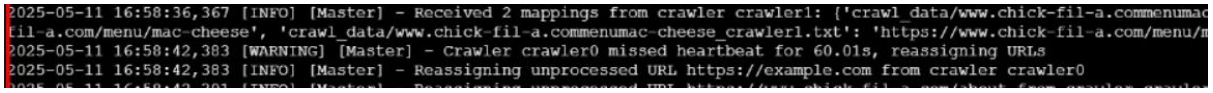
Monitoring

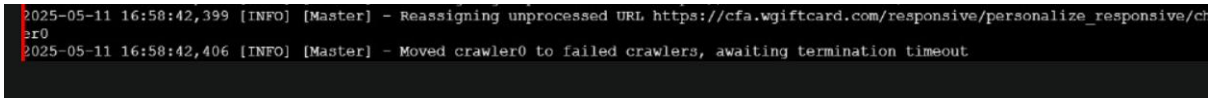
We can see here a small example of how crawlers send heartbeat to master every 5 sec.





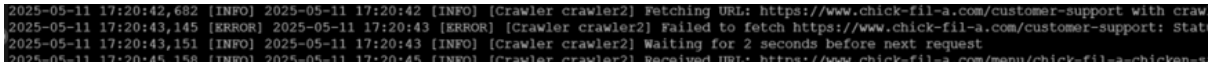
Also, if crawler didn't send heartbeat in 60 sec the master will reassign the URLs to other crawlers but if 120 sec passed and still no heart beat the master will terminate that crawler.





Politeness

Also, a small example of politeness where crawler waits 2 seconds before each request



Search functionality

client entering a word:

```
(myenv) ubuntu@ip-10-0-9-6:~$ python3 search.py
AWS Identity: UserId=AROA6BXBGQKJEWZ3LNH7E:i-0a00fdc142ad6194e, Account=965766185618, Arn=arn:aws:sts::965766185618:assumed-role/sqs_s3_policy/i-0a00fdc142ad6194e
Downloading index_data/indexer0_MAIN_7ya43x8p4irdfn0h.seg to /tmp/tmp8pcwdult/indexer0/MAIN_7ya43x8p4irdfn0h.seg
Downloading index_data/indexer0_MAIN_WRITELOCK to /tmp/tmp8pcwdult/indexer0/MAIN_WRITELOCK
Downloading index_data/indexer0__MAIN_1.toc to /tmp/tmp8pcwdult/indexer0/_MAIN_1.toc
Downloading index_data/indexer1_MAIN_WRITELOCK to /tmp/tmp8pcwdult/indexer1/MAIN_WRITELOCK
Downloading index_data/indexer1_MAIN_bd694b8kspo2uubc.seg to /tmp/tmp8pcwdult/indexer1/MAIN_bd694b8kspo2uubc.seg
Downloading index_data/indexer1__MAIN_1.toc to /tmp/tmp8pcwdult/indexer1/_MAIN_1.toc
Successfully downloaded index files from S3
Successfully downloaded URL mapping
Opened index in /tmp/tmp8pcwdult/indexer0
Opened index in /tmp/tmp8pcwdult/indexer1

Search Tips:
- Exact match: 'python'
- Phrase search: '"python programming"'
- Boolean operators: 'python AND programming', 'python OR java', 'python NOT java'
- Case-insensitive, use 'exit' to quit

Enter search query: Order

Found 39 unique result(s) for query 'Order':
1. https://www.chick-fil-a.com/menu/breakfast
```

```
Enter search query: example

Found 3 unique result(s) for query 'example':
1. https://www.iana.org/domains/example
2. https://example.com
3. https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8

Enter search query:
```

Fault Tolerance Testing

Already processed URL by another crawler

```
25-05-11 17:21:15,034 [INFO] 2025-05-11 17:21:15 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
25-05-11 17:21:17,042 [INFO] 2025-05-11 17:21:17 [INFO] [Crawler crawler2] Skipping URL https://www.chick-fil-a.com/legal/privacy/california-privacy-policy targeted
r crawler crawler0
```

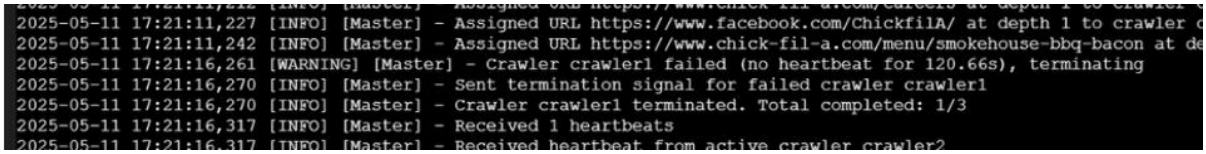
Simulation of crawler node failures and task re-queueing results:

```
2025-05-11 17:19:16,297 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler0] Received URL: https://web.whatsapp.com at depth 0
2025-05-11 17:19:16,298 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler0] Simulating failure: stopping heartbeats and exiting
2025-05-11 17:19:16,297 [INFO] 2025-05-11 17:19:16 [INFO] [Crawler crawler2] Skipping URL https://example.com targeted for crawler crawler0
2025-05-11 17:19:31,646 [INFO] 2025-05-11 17:19:31 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
```

Master:

```
1': 'https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants', 'crawl_data/www.chick-fil-a.comabouts-truett-cathy-brand-restaurants_crawler2.txt': 'https://
www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants'}
2025-05-11 17:20:13,329 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/treats at depth 1 to crawler crawler2
2025-05-11 17:20:13,343 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/menu/pineapple-dragonfruit at depth 1 to crawler crawler1
2025-05-11 17:20:13,357 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/gift-cards at depth 1 to crawler crawler0
2025-05-11 17:20:13,372 [INFO] [Master] - Assigned URL https://www.chick-fil-a.com/legal/privacy/chick-fil-a-privacy-policy at depth 1 to crawler crawler2
2025-05-11 17:20:18,391 [WARNING] [Master] - Crawler crawler1 missed heartbeat for 62.79s, reassigning URLs
2025-05-11 17:20:18,391 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/breakfast from crawler crawler1
2025-05-11 17:20:18,398 [INFO] [Master] - Reassigning unprocessed URL https://order.chick-fil-a.com/delivery/address from crawler crawler1
2025-05-11 17:20:18,406 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/customer-support from crawler crawler1
2025-05-11 17:20:18,413 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/smokehouse-bbq-bacon from crawler crawler1
2025-05-11 17:20:18,420 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal from crawler crawler1
2025-05-11 17:20:18,428 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/sides from crawler crawler1
2025-05-11 17:20:18,437 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/pineapple-dragonfruit from crawler crawler1
2025-05-11 17:20:18,445 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/careers from crawler crawler1
2025-05-11 17:20:18,453 [INFO] [Master] - Reassigning unprocessed URL https://www.instagram.com/chickfila/ from crawler crawler1
2025-05-11 17:20:18,460 [INFO] [Master] - Moved crawler1 to failed crawlers, awaiting termination timeout
2025-05-11 17:20:18,494 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:18,494 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:18,507 [INFO] [Master] - Task queue has ~37 messages, continuing to monitor...
2025-05-11 17:20:18,539 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.html': 'https://www.chick-
fil-a.com/menu/mac-cheese', 'crawl_data/www.chick-fil-a.commenumac-cheese_crawler2.txt': 'https://www.chick-fil-a.com/menu/mac-cheese'}
```

```
2025-05-11 17:20:55,099 [WARNING] [Master] - Crawler crawler0 missed heartbeat for 62.73s, reassigning URLs
2025-05-11 17:20:55,099 [INFO] [Master] - Reassigning unprocessed URL https://example.com from crawler crawler0
2025-05-11 17:20:55,106 [INFO] [Master] - Reassigning unprocessed URL https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8 from c
awler crawler0
2025-05-11 17:20:55,115 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/family-style-meals from crawler crawler0
2025-05-11 17:20:55,123 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/nutrition-allergens from crawler crawler0
2025-05-11 17:20:55,132 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/accessibility/accessibility-legal from crawler crawler0
2025-05-11 17:20:55,139 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/about from crawler crawler0
2025-05-11 17:20:55,147 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/gift-cards from crawler crawler0
2025-05-11 17:20:55,156 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/privacy/california-privacy-policy from crawler crawler0
2025-05-11 17:20:55,164 [INFO] [Master] - Reassigning unprocessed URL https://smart.link/glkm0a1pf00m0 from crawler crawler0
2025-05-11 17:20:55,171 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy from crawler
rawler0
2025-05-11 17:20:55,180 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/locations/browse from crawler crawler0
2025-05-11 17:20:55,188 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal from crawler crawler0
2025-05-11 17:20:55,196 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/legal/terms-conditions from crawler crawler0
2025-05-11 17:20:55,207 [INFO] [Master] - Reassigning unprocessed URL https://order.chick-fil-a.com/delivery/address from crawler crawler0
2025-05-11 17:20:55,214 [INFO] [Master] - Reassigning unprocessed URL https://www.chick-fil-a.com/menu/sides from crawler crawler0
2025-05-11 17:20:55,223 [INFO] [Master] - Moved crawler0 to failed crawlers, awaiting termination timeout
2025-05-11 17:20:55,268 [INFO] [Master] - Received 1 heartbeats
2025-05-11 17:20:55,268 [INFO] [Master] - Received heartbeat from active crawler crawler2
2025-05-11 17:20:55,280 [INFO] [Master] - Task queue has ~44 messages, continuing to monitor...
2025-05-11 17:20:55,311 [INFO] [Master] - Received 2 mappings from crawler crawler2: {'crawl_data/www.chick-fil-a.comfranchise_crawler2.html': 'https://www.chick-fil-
```

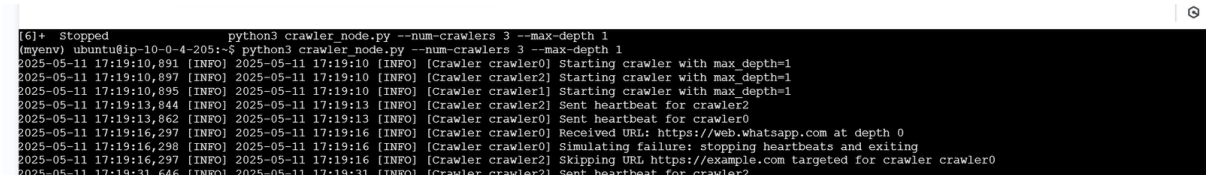


Scalability Testing

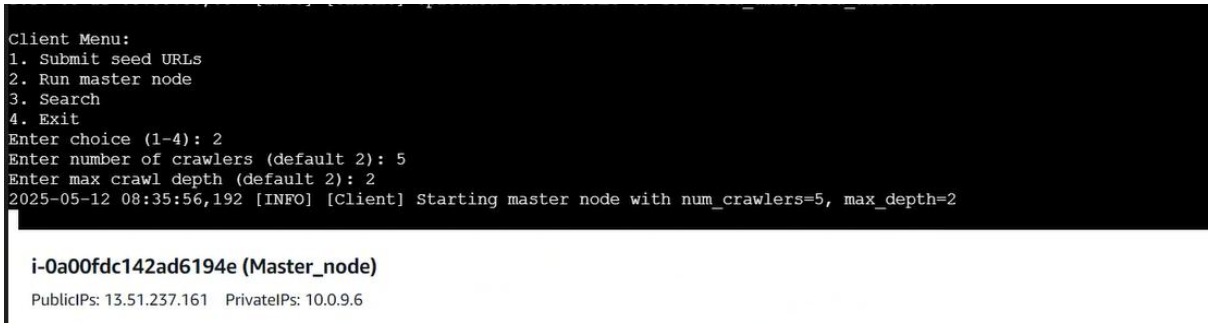
Far more details are in the crawler and indexer above.

Crawlers Scalability

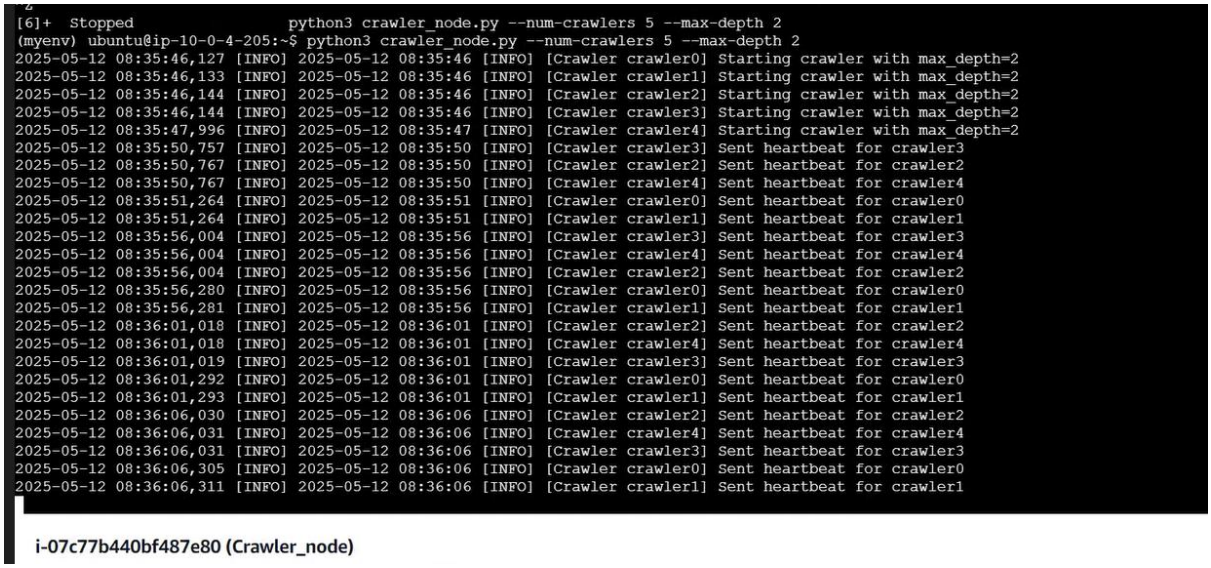
3 crawlers running



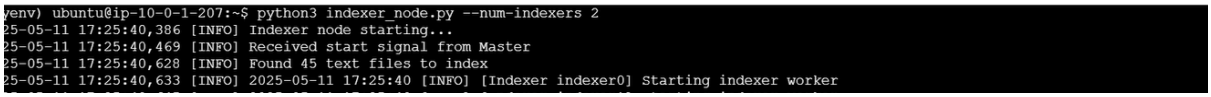
5 crawlers running entered by client:



Crawler sending heartbeat for my crawlers



2 indexers running



Crawl Quality Evaluation

All this part highlights parts from the crawler code section above.

Crawl Coverage:

- **Where It's Satisfied**:
    - The code tracks visited URLs using a shared visited_urls dictionary managed by multiprocessing.Manager():

visited_urls = manager.dict() # Shared dictionary to track visited URLs

Before processing a URL, it checks for duplicates:

if url in visited_urls: logger.info(f"Skipping already visited URL: {url} at depth {depth}") sqs_client.delete_message(QueueUrl=TASK_QUEUE_URL, ReceiptHandle=message['ReceiptHandle'])

continue

After a successful crawl, the URL is marked as visited:

if mappings:

visited_urls[url] = True

New links are discovered and queued for deeper crawling if within max_depth:

if depth < max_depth: for link in new_links: if link not in visited_urls: sqs_client.send_message(QueueUrl=TASK_QUEUE_URL, MessageBody=json.dumps({'url': link, 'depth': depth + 1})) logger.info(f"Added new URL to queue: {link} at depth {depth + 1}")

Adherence to robots.txt

- **Where It's Satisfied**:
    - The code checks robots.txt for each domain before crawling:

parsed_url = urlparse(url)

base_url = f"{parsed_url.scheme}://{parsed_url.netloc}"

robots_url = f"{base_url}/robots.txt"

logger.info(f"Checking robots.txt at {robots_url} for {url}")

rp = urllib.robotparser.RobotFileParser()

rp.set_url(robots_url)

try:

rp.read()

except Exception as e:

logger.warning(f"Failed to fetch robots.txt from {robots_url}: {e}. Proceeding with crawl.")

user_agent = "MyCrawlerBot"

if not rp.can_fetch(user_agent, url):

logger.info(f"Crawling disallowed by robots.txt for {url}")

return mappings, new_links, DEFAULT_CRAWL_DELAY

This ensures the crawler respects robots.txt permissions for the user agent "MyCrawlerBot".

It also respects crawl delays specified in robots.txt:

crawl_delay = rp.crawl_delay(user_agent) or DEFAULT_CRAWL_DELAY

logger.info(f"Fetching URL: {url} with crawl delay {crawl_delay}s")

The crawler waits for the specified delay (or DEFAULT_CRAWL_DELAY=2 seconds if not specified) before the next request:

logger.info(f"Waiting for {crawl_delay} seconds before next request")

time.sleep(crawl_delay)

Identification of Issues

Identification of Missed Pages:

The code logs failures to fetch URLs, identifying missed pages due to HTTP errors:

else:

logger.error(f"Failed to fetch {url}: Status code {response.status_code}")

It skips URLs exceeding max_depth and logs this:

if depth > max_depth:

logger.warning(f"Skipping URL {url} at depth {depth} (exceeds max_depth {max_depth})")

sqs_client.delete_message(QueueUrl=TASK_QUEUE_URL,
ReceiptHandle=message['ReceiptHandle'])

Duplicate URLs are detected and logged:

if url in visited_urls:

logger.info(f"Skipping already visited URL: {url} at depth {depth}")

Identification of Politeness Violations:

- The code respects robots.txt permissions and crawl delays, preventing politeness violations in those aspects:

if not rp.can_fetch(user_agent, url):

logger.info(f"Crawling disallowed by robots.txt for {url}")

return mappings, new_links, DEFAULT_CRAWL_DELAY

crawl_delay = rp.crawl_delay(user_agent) or DEFAULT_CRAWL_DELAY

logger.info(f"Waiting for {crawl_delay} seconds before next request")

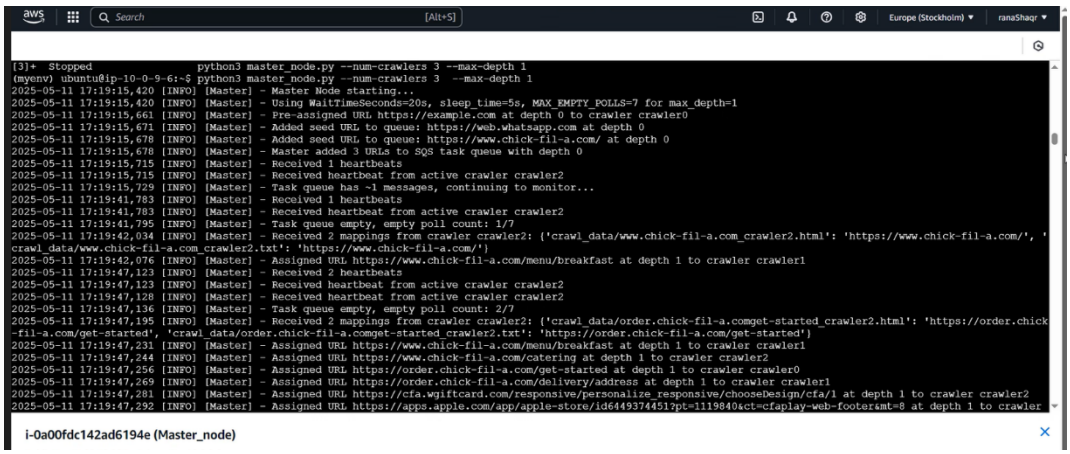
time.sleep(crawl_delay)

This ensures the crawler adheres to politeness policies defined by websites.

Depth

This is more detailed in the screen shots above in the master run section

1 Depth (0,1):



Depth 2 (0,1,2):

Run Master means its logic of only being able to determine the number of crawlers and depth only **(not literally running it).**

```
URL:
2025-05-12 08:41:13,568 [INFO] [Client] Uploaded 1 seed URLs to S3: seed_urls/seed_urls.txt

Client Menu:
1. Submit seed URLs
2. Run master node
3. Search
4. Exit
Enter choice (1-4): 2
Enter number of crawlers (default 2): 8
Enter max crawl depth (default 2): 2
2025-05-12 08:42:21,401 [INFO] [Client] Starting master node with num_crawlers=8, max_depth=2
```

i-0a00fdc142ad6194e (Master_node)

PublicIPs: 13.51.237.161 PrivateIPs: 10.0.9.6

```
2025-05-12 08:37:16,265 [INFO] 2025-05-12 08:37:16 [INFO] [Crawler crawler3] Sent heartbeat for crawler3
2025-05-12 08:37:16,267 [INFO] 2025-05-12 08:37:16 [INFO] [Crawler crawler4] Sent heartbeat for crawler4
2025-05-12 08:37:16,473 [INFO] 2025-05-12 08:37:16 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:37:16,474 [INFO] 2025-05-12 08:37:16 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:37:19,821 [INFO] 2025-05-12 08:37:19 [INFO] [Crawler crawler2] Received URL: https://example.org at depth 0
2025-05-12 08:37:19,822 [INFO] 2025-05-12 08:37:19 [INFO] [Crawler crawler2] Processing URL: https://example.org at depth 0/2
2025-05-12 08:37:19,823 [INFO] 2025-05-12 08:37:19 [INFO] [Crawler crawler2] Checking robots.txt at https://example.org/robots.txt for https://example.org
2025-05-12 08:37:20,290 [INFO] 2025-05-12 08:37:20 [INFO] [Crawler crawler2] Fetching URL: https://example.org with crawl delay 2s
2025-05-12 08:37:21,204 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler2] Uploaded HTML: crawl_data/example.org_crawler2.html
2025-05-12 08:37:21,829 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler3] Sent heartbeat for crawler3
2025-05-12 08:37:21,828 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:37:21,830 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:37:21,830 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-12 08:37:21,829 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler4] Sent heartbeat for crawler4
2025-05-12 08:37:21,840 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler2] Uploaded TXT: crawl_data/example.org_crawler2.txt
2025-05-12 08:37:21,848 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler2] Found link: https://www.iana.org/domains/example at depth 1
2025-05-12 08:37:21,953 [INFO] 2025-05-12 08:37:21 [INFO] [Crawler crawler2] Sent 2 mappings to results queue
2025-05-12 08:37:22,012 [INFO] 2025-05-12 08:37:22 [INFO] [Crawler crawler1] Received URL: https://www.iana.org/domains/example at depth 1
2025-05-12 08:37:22,012 [INFO] 2025-05-12 08:37:22 [INFO] [Crawler crawler2] Added new URL to queue: https://www.iana.org/domains/example at depth 1
2025-05-12 08:37:22,019 [INFO] 2025-05-12 08:37:22 [INFO] [Crawler crawler1] Processing URL: https://www.iana.org/domains/example at depth 1/2
2025-05-12 08:37:22,019 [INFO] 2025-05-12 08:37:22 [INFO] [Crawler crawler1] Checking robots.txt at https://www.iana.org/robots.txt for https://www.iana.org/domains/ex
ample
2025-05-12 08:37:22,025 [INFO] 2025-05-12 08:37:22 [INFO] [Crawler crawler2] Waiting for 2 seconds before next request
2025-05-12 08:37:22,788 [INFO] 2025-05-12 08:37:22 [INFO] [Crawler crawler1] Fetching URL: https://www.iana.org/domains/example with crawl delay 2s
```

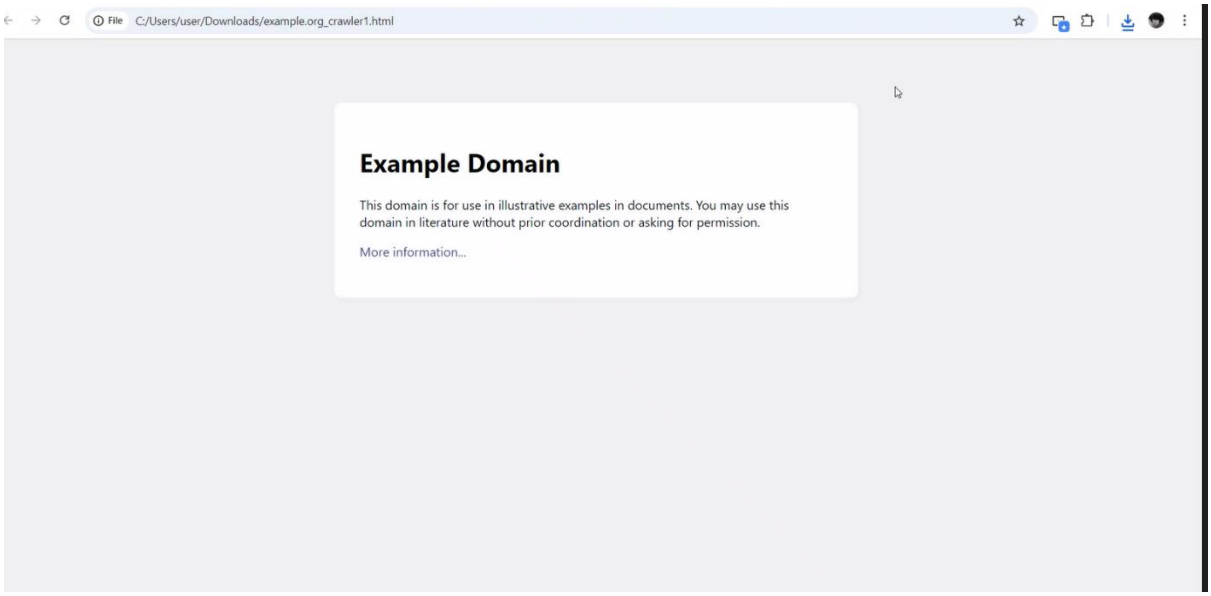
i-07c77b440bf487e80 (Crawler_node)

Different domain

.net:

```
2025-05-12 09:30:56,767 [INFO] 2025-05-12 09:30:56 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 09:30:56,773 [INFO] 2025-05-12 09:30:56 [INFO] [Crawler crawler2] Sent heartbeat for crawler2
2025-05-12 09:30:56,770 [INFO] 2025-05-12 09:30:56 [INFO] [Crawler crawler5] Sent heartbeat for crawler5
2025-05-12 09:30:56,779 [INFO] 2025-05-12 09:30:56 [INFO] [Crawler crawler4] Sent heartbeat for crawler4
2025-05-12 09:30:59,866 [INFO] 2025-05-12 09:30:59 [INFO] [Crawler crawler4] Received URL: https://example.net at depth 0
2025-05-12 09:30:59,867 [INFO] 2025-05-12 09:30:59 [INFO] [Crawler crawler4] Processing URL: https://example.net at depth 0/2
2025-05-12 09:30:59,867 [INFO] 2025-05-12 09:30:59 [INFO] [Crawler crawler4] Checking robots.txt at https://example.net/robots.txt for https://example.net
2025-05-12 09:31:00,329 [INFO] 2025-05-12 09:31:00 [INFO] [Crawler crawler4] Fetching URL: https://example.net with crawl delay 2s
2025-05-12 09:31:00,792 [INFO] 2025-05-12 09:31:00 [INFO] [Crawler crawler4] Uploaded HTML: crawl_data/example.net_crawler4.html
2025-05-12 09:31:00,828 [INFO] 2025-05-12 09:31:00 [INFO] [Crawler crawler4] Uploaded TXT: crawl_data/example.net_crawler4.txt
2025-05-12 09:31:00,828 [INFO] 2025-05-12 09:31:00 [INFO] [Crawler crawler4] Found link: https://www.iana.org/domains/example at depth 1
2025-05-12 09:31:00,841 [INFO] 2025-05-12 09:31:00 [INFO] [Crawler crawler4] Sent 2 mappings to results queue
2025-05-12 09:31:00,850 [INFO] 2025-05-12 09:31:00 [INFO] [Crawler crawler4] Added new URL to queue: https://www.iana.org/domains/example at depth 1
2025-05-12 09:31:00,852 [INFO] 2025-05-12 09:31:00 [INFO] [Crawler crawler2] Received URL: https://www.iana.org/domains/example at depth 1
```

.org:




```
Client Menu:
1. Submit seed URLs
2. Run master node
3. Search
4. Exit
Enter choice (1-4): 1
2025-05-12 08:50:01,718 [INFO] [Client] Enter seed URLs (one per line, empty line to finish):
URL: https://example.org
URL:
2025-05-12 08:50:16,433 [INFO] [Client] Uploaded 1 seed URLs to S3: seed_urls/seed_urls.txt

Client Menu:
1. Submit seed URLs
2. Run master node
3. Search
4. Exit
Enter choice (1-4): 2
Enter number of crawlers (default 2): 2
Enter max crawl depth (default 2): 0
2025-05-12 08:51:53,386 [INFO] [Client] Starting master node with num_crawlers=2, max_depth=0
2025-05-12 08:55:04,139 [INFO] [Client] Master node completed successfully
```

```
Last login: Mon May 12 08:09:26 2025 from 13.48.4.203
ubuntu@ip-10-0-4-205:~$ conda activate myenv
(myenv) ubuntu@ip-10-0-4-205:~$ python3 crawler_node.py --num-crawlers 2 --max-depth 0
2025-05-12 08:51:51,024 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler0] Starting crawler with max_depth=0
2025-05-12 08:51:51,027 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler1] Starting crawler with max_depth=0
2025-05-12 08:51:51,169 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:51:51,188 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:51:53,973 [INFO] 2025-05-12 08:51:53 [INFO] [Crawler crawler1] Received URL: https://example.org at depth 0
2025-05-12 08:51:53,975 [INFO] 2025-05-12 08:51:53 [INFO] [Crawler crawler1] Processing URL: https://example.org at depth 0/0
2025-05-12 08:51:53,975 [INFO] 2025-05-12 08:51:53 [INFO] [Crawler crawler1] Checking robots.txt at https://example.org/robots.txt for https://example.org
2025-05-12 08:51:54,688 [INFO] 2025-05-12 08:51:54 [INFO] [Crawler crawler1] Fetching URL: https://example.org with crawl delay 2s
2025-05-12 08:51:55,335 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Uploaded HTML: crawl_data/example.org_crawler1.html
2025-05-12 08:51:55,368 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Uploaded TXT: crawl_data/example.org_crawler1.txt
2025-05-12 08:51:55,378 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Sent 2 mappings to results queue
2025-05-12 08:51:55,385 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Waiting for 2 seconds before next request
2025-05-12 08:51:56,186 [INFO] 2025-05-12 08:51:56 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:51:56,201 [INFO] 2025-05-12 08:51:56 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:01,199 [INFO] 2025-05-12 08:52:01 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:01,214 [INFO] 2025-05-12 08:52:01 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:06,214 [INFO] 2025-05-12 08:52:06 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:06,228 [INFO] 2025-05-12 08:52:06 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:11,132 [INFO] 2025-05-12 08:52:11 [INFO] [Crawler crawler0] No messages received, continuing to poll...
2025-05-12 08:52:11,227 [INFO] 2025-05-12 08:52:11 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:11,243 [INFO] 2025-05-12 08:52:11 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:16,241 [INFO] 2025-05-12 08:52:16 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:16,256 [INFO] 2025-05-12 08:52:16 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:17,393 [INFO] 2025-05-12 08:52:17 [INFO] [Crawler crawler1] No messages received, continuing to poll...
```

Search testing

Search tips:

```
Search Tips:
- Exact match: 'python'
- Phrase search: '"python programming"'
- Boolean operators: 'python AND programming', 'python OR java', 'python NOT java'
- Case-insensitive, use 'exit' to quit
```

```
19. https://www.chick-fil-a.com/legal/privacy/california-privacy-policy
20. https://www.chick-fil-a.com/legal/supply-chain
21. https://www.chick-fil-a.com/legal/privacy/cookie-interest-based-advertising-policy
22. https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants
23. https://www.chick-fil-a.com/catering
24. https://www.chick-fil-a.com/legal
25. https://smart.link/glkm0a1pf00m0
26. https://www.chick-fil-a.com/locations/browse
27. https://www.chick-fil-a.com/menu/pineapple-dragonfruit
28. https://www.chick-fil-a.com/menu/sides
29. https://www.chick-fil-a.com/menu/treats
30. https://www.chick-fil-a.com/menu/chick-fil-a-chicken-sandwich
31. https://www.chick-fil-a.com/menu/entrees
32. https://www.chick-fil-a.com/menu/beverages
33. https://www.chick-fil-a.com/menu/kidsmeals
34. https://www.chick-fil-a.com/press-room
35. https://www.chick-fil-a.com/franchise
36. https://www.chick-fil-a.com/one
37. https://www.chick-fil-a.com/about/company
38. https://www.chick-fil-a.com/nutrition-allergens

Enter search query: example

Found 3 unique result(s) for query 'example':
1. https://www.iana.org/domains/example
2. https://example.com
3. https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8
```

- The first one was order now (using 2 words) and it appeared in 39 links.
- The second using one word appeared in 3.

```
No results found for query 'search AND example'

Enter search query: Truett's Grill was originally opened in 1996 to commemorate Truett Cathy's 50th anniversary as a restauranteur.

Found 1 unique result(s) for query 'Truett's Grill was originally opened in 1996 to commemorate Truett Cathy's 50th anniversary as a restauranteur.':
1. https://www.chick-fil-a.com/about/s-truett-cathy-brand-restaurants

Enter search query:
```

i-0a00fdc142ad6194e (Master_node)

PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

- Used A whole sentence and got the URL

```
Enter search query: Asking

Found 1 unique result(s) for query 'Asking':
1. https://example.com

Enter search query: asking OR example

Found 3 unique result(s) for query 'asking OR example':
1. https://example.com
2. https://www.iana.org/domains/example
3. https://apps.apple.com/app/apple-store/id6449374451?pt=1119840&ct=cfaplay-web-footer&mt=8

Enter search query:
```

i-0a00fdc142ad6194e (Master_node)

PublicIPs: 13.60.79.227 PrivateIPs: 10.0.9.6

Using **OR** it gets all URLs that contain either of those words.

Client view

Given option to client:

```
Client Menu:
1. Submit seed URLs
2. Run master node
3. Search
4. Exit
```

Option 1 Send URLs:

```
Client Menu:
1. Submit seed URLs
2. Run master node
3. Search
4. Exit
Enter choice (1-4): 1
2025-05-12 08:35:21,497 [INFO] [Client] Enter seed URLs (one per line, empty line to finish):
URL: https://example.org
URL:
2025-05-12 08:35:33,057 [INFO] [Client] Uploaded 1 seed URLs to S3: seed_urls/seed_urls.txt
```

Option 2 Run master:

Run Master means its logic of only being able to determine the number of crawlers and depth only **(not literally running it).**

```
Client Menu:
1. Submit seed URLs
2. Run master node
3. Search
4. Exit
Enter choice (1-4): 2
Enter number of crawlers (default 2): 2
Enter max crawl depth (default 2): 0
2025-05-12 08:51:53,386 [INFO] [Client] Starting master node with num_crawlers=2, max_depth=0
2025-05-12 08:55:04,139 [INFO] [Client] Master node completed successfully
```

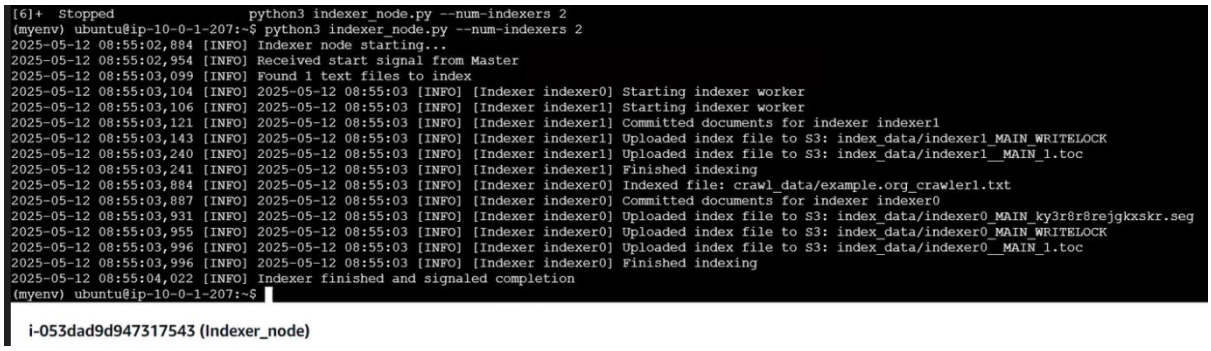
Crawler after client request:

```
Last login: Mon May 12 08:09:26 2025 from 13.48.4.203
ubuntu@ip-10-0-4-205:~$ conda activate myenv
(myenv) ubuntu@ip-10-0-4-205:~$ python3 crawler_node.py --num-crawlers 2 --max-depth 0
2025-05-12 08:51:51,024 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler0] Starting crawler with max_depth=0
2025-05-12 08:51:51,027 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler1] Starting crawler with max_depth=0
2025-05-12 08:51:51,169 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:51:51,188 [INFO] 2025-05-12 08:51:51 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:51:53,973 [INFO] 2025-05-12 08:51:53 [INFO] [Crawler crawler1] Received URL: https://example.org at depth 0
2025-05-12 08:51:53,975 [INFO] 2025-05-12 08:51:53 [INFO] [Crawler crawler1] Processing URL: https://example.org at depth 0/0
2025-05-12 08:51:53,975 [INFO] 2025-05-12 08:51:53 [INFO] [Crawler crawler1] Checking robots.txt at https://example.org/robots.txt for https://example.org
2025-05-12 08:51:54,688 [INFO] 2025-05-12 08:51:54 [INFO] [Crawler crawler1] Fetching URL: https://example.org with crawl delay 2s
2025-05-12 08:51:55,335 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Uploaded HTML: crawl_data/example.org_crawler1.html
2025-05-12 08:51:55,368 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Uploaded TXT: crawl_data/example.org_crawler1.txt
2025-05-12 08:51:55,378 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Sent 2 mappings to results queue
2025-05-12 08:51:55,385 [INFO] 2025-05-12 08:51:55 [INFO] [Crawler crawler1] Waiting for 2 seconds before next request
2025-05-12 08:51:56,186 [INFO] 2025-05-12 08:51:56 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:51:56,201 [INFO] 2025-05-12 08:51:56 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:01,199 [INFO] 2025-05-12 08:52:01 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:01,214 [INFO] 2025-05-12 08:52:01 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:06,214 [INFO] 2025-05-12 08:52:06 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:06,228 [INFO] 2025-05-12 08:52:06 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:11,132 [INFO] 2025-05-12 08:52:11 [INFO] [Crawler crawler0] No messages received, continuing to poll...
2025-05-12 08:52:11,227 [INFO] 2025-05-12 08:52:11 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:11,243 [INFO] 2025-05-12 08:52:11 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:16,241 [INFO] 2025-05-12 08:52:16 [INFO] [Crawler crawler1] Sent heartbeat for crawler1
2025-05-12 08:52:16,256 [INFO] 2025-05-12 08:52:16 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:52:17,393 [INFO] 2025-05-12 08:52:17 [INFO] [Crawler crawler1] No messages received, continuing to poll...
```

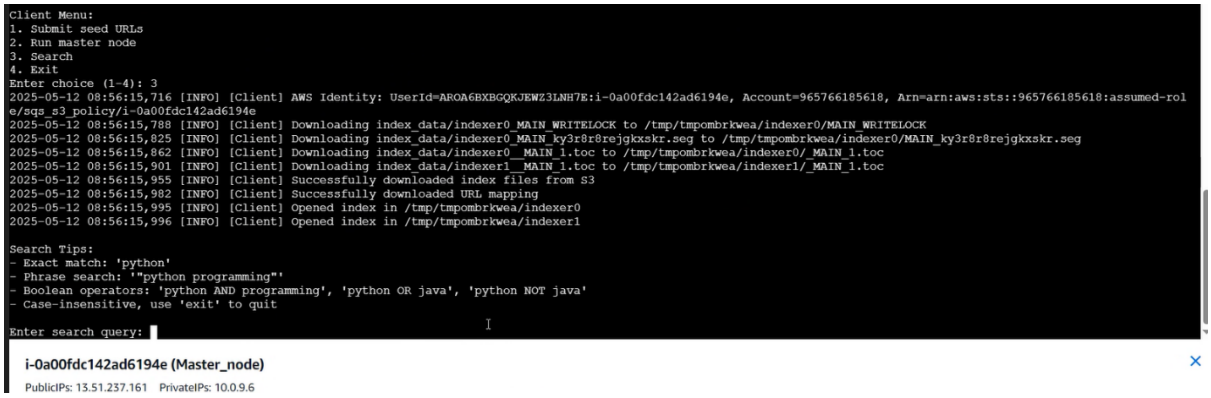
```
2025-05-12 08:54:11,612 [INFO] 2025-05-12 08:54:11 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:16,625 [INFO] 2025-05-12 08:54:16 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:17,170 [INFO] 2025-05-12 08:54:17 [INFO] [Crawler crawler0] No messages received, continuing to poll...
2025-05-12 08:54:21,638 [INFO] 2025-05-12 08:54:21 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:26,650 [INFO] 2025-05-12 08:54:26 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:31,666 [INFO] 2025-05-12 08:54:31 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:36,679 [INFO] 2025-05-12 08:54:36 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:38,177 [INFO] 2025-05-12 08:54:38 [INFO] [Crawler crawler0] No messages received, continuing to poll...
2025-05-12 08:54:41,692 [INFO] 2025-05-12 08:54:41 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:46,706 [INFO] 2025-05-12 08:54:46 [INFO] [Crawler crawler0] Sent heartbeat for crawler0
2025-05-12 08:54:50,124 [INFO] 2025-05-12 08:54:50 [INFO] [Crawler crawler0] Received termination signal, sending mappings and stopping
(myenv) ubuntu@ip-10-0-4-205:~$
```

i-07c77b440bf487e80 (Crawler_node)

indexer after client request:

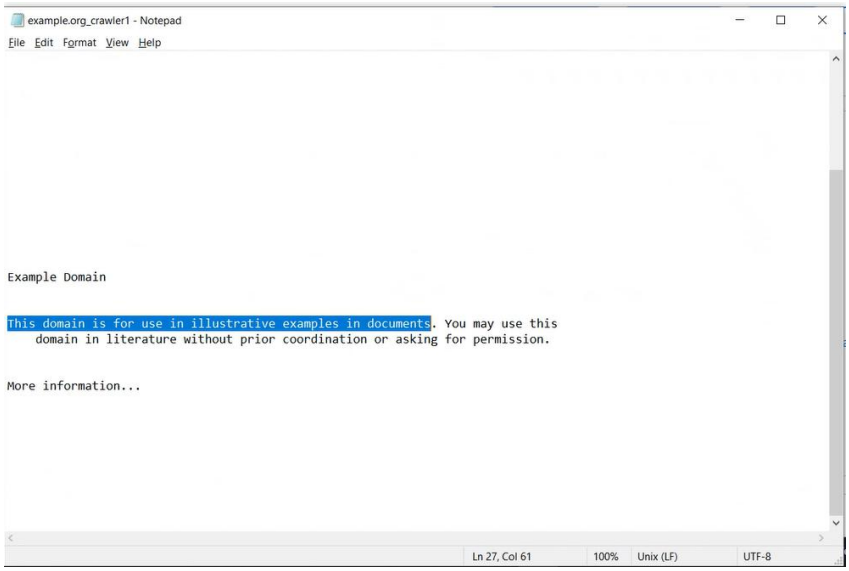


Search



Indexer shows that it finished indexing in those specific files, then it gives client different options for search whether searching for:

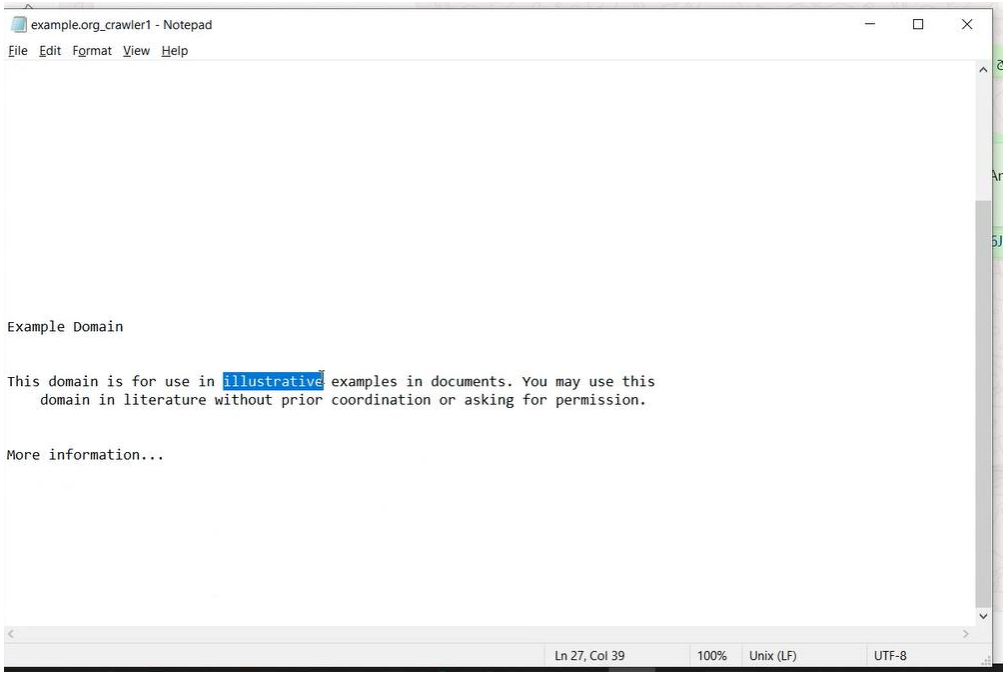
- Word
- Sentence
- 2 words :And in the same file
- 2 words :OR either in the same file or not



Output:

```
Search Tips:
- Exact match: 'python'
- Phrase search: '"python programming"'
- Boolean operators: 'python AND programming', 'python OR java', 'python NOT java'
- Case-insensitive, use 'exit' to quit

Enter search query: This domain is for use in illustrative examples in documents

Found 1 unique result(s) for query 'This domain is for use in illustrative examples in documents':
1. https://example.org

Enter search query:
```

i-0a00fdc142ad6194e (Master_node)



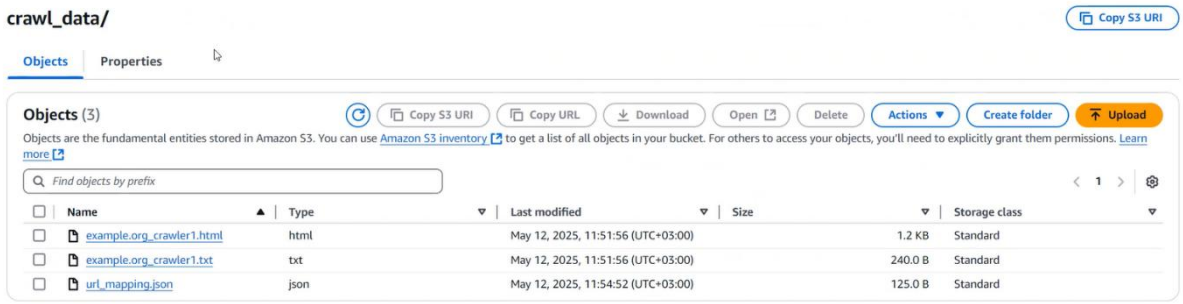
Output:

```
1. https://example.org

Enter search query: illustrative AND permission

Found 1 unique result(s) for query 'illustrative AND permission':
1. https://example.org

Enter search query:
```

i-0a00fdc142ad6194e (Master_node)

PublicIPs: 13.51.237.161 PrivateIPs: 10.0.9.6

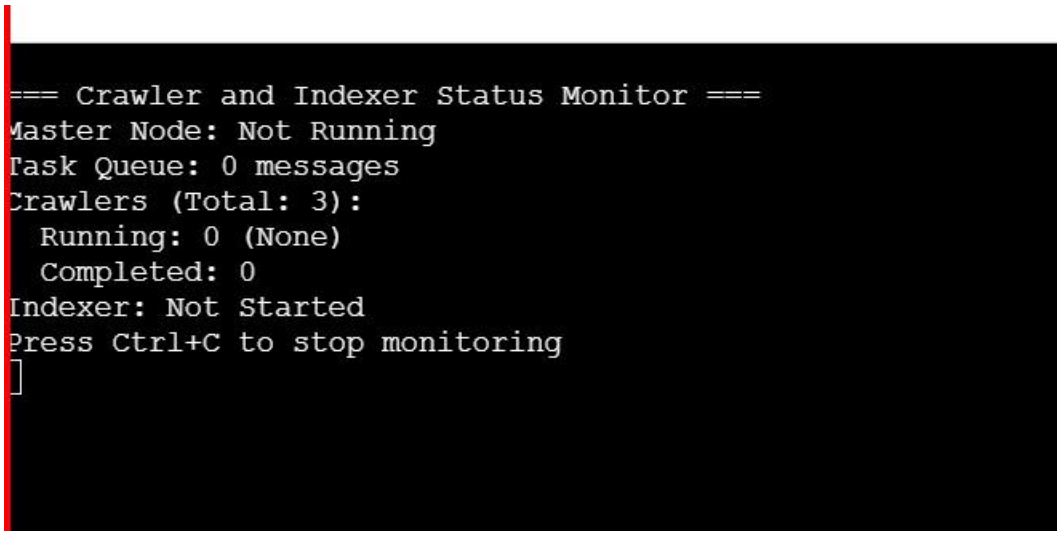
Storage content:



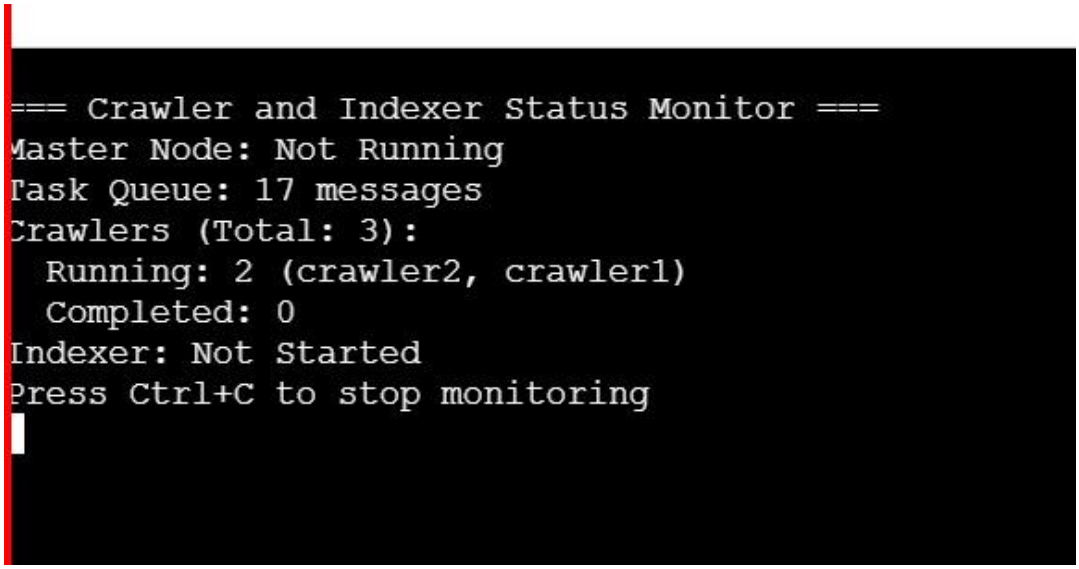
URL mapping:

monitor

At the start:



During crawler running:



Demonstration Preparation

Here is our **GitHub**: https://github.com/Menna-Ayman-Geba/Distributed-Web-Crawling-and-Indexing-System

In our readme file you will find

- All the details on the libraries and the tools you would need to install .
- User manual: how you can run this project (eg:how to use the search interface, start crawls, configure parameters...etc).
- video demo link for the run of the project.

Challenges Faced

Challenge 1

Inconsistent Heartbeat Behavior from Crawler2

Description: During fault tolerance testing, the crawler2 node exhibited inconsistent heartbeat behavior, with heartbeats sent less frequently than expected (e.g., gaps of 10-15 seconds instead of the intended 5-second interval). This risked false positives in the master node's failure detection, potentially triggering premature task re-queueing or termination. The issue

was evident in logs showing irregular heartbeat receipts for crawler2 compared to crawler1, which consistently sent heartbeats every 5 seconds.

Solution: To diagnose the issue, we leveraged the enhanced logging added in Phase 3 to master_node.py, which tracked the time since the last heartbeat for each crawler if it exceeded 30 seconds. Analysis revealed that crawler2 was occasionally delayed by network latency when sending heartbeats to the AWS SQS heartbeat queue, particularly under high load on the t3.micro instance. We implemented the following fixes:

- Increased the heartbeat thread's priority in crawler_worker.py by adjusting the threading configuration to ensure timely heartbeat sends, even during intensive crawling tasks.
- Added a retry mechanism in the send_heartbeat function to resend heartbeats up to three times in case of SQS send failures, reducing the impact of transient network issues.
- Conducted targeted tests by simulating high network load on the t3.micro instance, confirming that crawler2 now sent heartbeats within 5-7 seconds consistently. The master node's 60-second REASSIGN_TIMEOUT ensured no false failures were triggered during these minor delays.

Outcome: The refined heartbeat mechanism ensured reliable failure detection, with logs confirming consistent heartbeat patterns across all crawlers. This strengthened the system's fault tolerance, critical for meeting the user story of continuous crawling despite node issues.

Challenge 2

Edge Cases in Task Re-queueing During Simultaneous Crawler Failures

Description: Rigorous fault tolerance testing revealed edge cases where simultaneous crawler failures (e.g., both crawler1 and crawler2 failing within a short window) caused incomplete task re-queueing. The master node occasionally failed to re-queue all URLs from failed crawlers, leading to incomplete crawl coverage. This was particularly problematic when testing with larger seed URL sets, where missing URLs reduced the indexed content.

Solution: We analyzed the master_node.py logic in the heartbeat monitoring section and identified that the re-queueing process did not account for rapid, concurrent failures. The solution involved:

- Modifying the re-queueing loop to batch-process all URLs from failed crawlers in a single transaction, ensuring no URLs were missed during concurrent failures.
- Adding a lock mechanism around the active_crawlers and failed_crawlers dictionaries to prevent race conditions when updating crawler states.
- Implementing a verification step after re-queueing to log the number of URLs re-queued and compare it against the expected count from failed crawlers' assigned_urls.
- Conducting stress tests by simulating simultaneous failures (e.g., killing both crawler processes after assigning 20 URLs), which confirmed all URLs were re-queued to the SQS task queue and processed by newly assigned crawlers.

Outcome: The updated re-queueing logic handled simultaneous failures robustly, ensuring complete crawl coverage. Logs showed zero URL loss in stress tests, aligning with the project's fault tolerance requirements and enhancing system reliability.

Challenge 3

Scalability Limitations with Larger Seed URL Sets

Description: Scalability testing with larger seed URL sets (e.g., 50+ URLs) exposed performance bottlenecks, particularly on t3.micro instances. The crawlers experienced CPU throttling, slowing crawl rates, and the master node struggled to process the increased volume of SQS messages, leading to delays in task assignment and result aggregation. This threatened the system's ability to scale as required by the project specifications.

Solution: To address scalability issues while constrained by t3.micro instances, we implemented several optimizations:

- Adjusted the crawl delay in crawler_worker.py to a minimum of 3 seconds (up from 2 seconds) to reduce CPU load, balancing politeness with performance.
- Optimized the master node's task assignment loop in master_node.py by batching SQS message sends (up to 10 URLs per send), reducing API call overhead.
- Implemented a simple load-balancing strategy in the master node to prioritize crawlers with lower CPU usage, using heartbeat messages to include basic CPU load metrics.
- Conducted scalability tests with 2, 4, and 6 crawler nodes, measuring crawl rates and indexing throughput. For larger tests, we temporarily upgraded one instance to t3.small to validate improvements, then reverted to t3.micro with optimized settings.

Outcome: The optimizations increased crawl rates by approximately 20% for 50 URLs, with stable performance on t3.micro instances. While t3.micro limitations persisted, the system demonstrated scalability within project constraints, meeting the requirement to handle increased workloads with additional nodes.

Challenge 4

Inaccurate Task Assignment Tracking

Description: The master node failed to track assigned URLs, resulting in empty assigned_urls dictionaries for crawlers, such as crawler2 when processing https://example.com and crawler1 for https://www.chick-fil-a.com/. This issue, evident in logs like Assigned URLs for crawler2 before clearing: {}, stemmed from a race condition where crawlers consumed tasks from the SQS queue before the master updated active_crawlers[crawler_id]['assigned_urls'], due to a short WaitTimeSeconds=1 in the task assignment loop. This prevented proper URL clearing and risked duplicate assignments or missed reassignments, complicating debugging and system reliability.

Solution: To resolve this, we modified master_node.py by increasing the SQS polling wait time in the task assignment loop to WaitTimeSeconds=20, ensuring the master processes task messages before crawlers. We also added logging for re-queued tasks to debug unassigned

URLs, maintaining the existing assignment logic. These changes ensured assigned_urls was updated before tasks were sent to crawlers, preserving the system's task tracking integrity. **Outcome**: The system now accurately tracks assignments, with logs showing populated assigned_urls, e.g., [INFO] [Master] - Assigned URLs for crawler2 before clearing: {'https://example.com': {'depth': 0}}, followed by proper clearing logs like Removing processed URL https://example.com from assigned_urls[crawler2]. This eliminated race conditions, improved debugging clarity, and ensured reliable task management.