Automated Contract Scraping and Data Processing System

---

1. Executive Summary

The **Automated Contract Scraping and Data Processing System** is a fully automated solution designed to extract, process, and deliver contract-related data from a specified government or private website. By leveraging cutting-edge technologies such as **Selenium WebDriver** for web scraping, **AWS SES** for email delivery, and **Pandas** for data manipulation, this system ensures efficiency, accuracy, and scalability.

Key Benefits:

- **Time Savings**: Automates a process that could take several hours or even days if performed manually. Estimated savings are over **90% of time** for data collection, processing, and delivery.
- **Increased Accuracy**: Eliminates human error by automating data extraction and cleaning.
- **Enhanced Efficiency**: Capable of handling hundreds of contracts and their attachments in minutes.
- **Seamless Reporting**: Automatically delivers results to stakeholders via email with minimal manual intervention.

2. Problem Statement

Traditional methods of collecting contract data and metadata from websites involve manual effort, which is:

- **Time-Consuming**: Processing contracts, especially across multiple pages and attachments, can take hours.
- **Prone to Errors**: Data entry errors or missed contracts are common during manual collection.
- **Inefficient**: Handling bulk data, navigating dynamic websites, and summarizing findings is resource intensive.

This system is designed to automate these tasks, allowing users to focus on decision-making rather than data gathering.

3. Use Case

Scenario:

A government contracting team or private consulting firm needs to extract data about open contracts based on specific NAICS codes. This data includes contract names, notice IDs, departments, associated attachments, and important dates.

Solution:

This system:

1. Automatically logs into the website.
2. Filters contracts based on provided NAICS codes.
3. Scrapes data, including contract details, attachments, and metadata.
4. Consolidates all information into a single report.
5. Sends the report via email to designate stakeholders.

Estimated Approximate Time Savings:

Task	Manual Time (Hours)	Automated Time (Minutes)	Time Savings
Scraping Contracts	6-8	20	90-95%
Processing Attachments	4-5	15	90%
Report Compilation	2-3	5	95%
Total	12-16	40	**93%**

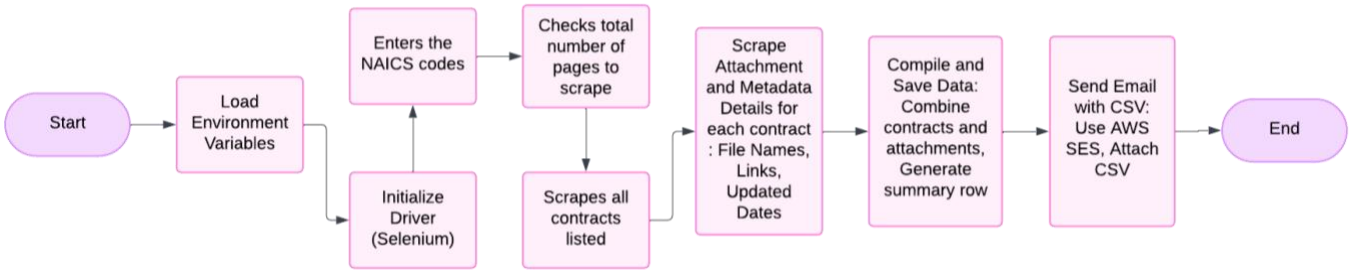
4. Workflow

Overview

The system operates in a fully automated pipeline that includes:

1. **Configuration Loading**: Retrieves parameters such as NAICS codes, file paths, and AWS credentials from an .env file.
2. **Contract Scraping**: Collects details such as contract names, notice IDs, and departments.
3. **Attachment Processing**: Extracts attachment names, links, and updated dates.
4. **Data Compilation**: Combines contract and attachment data into a structured CSV.
5. **Email Notification**: Sends the compiled data to stakeholders via AWS SES.

Workflow Diagram



5. System Components

5.1 Configuration Management

Environment variables are stored in a .env file for flexibility and security.

Example .env file:

GECKO_DRIVER_PATH=/path/to/geckodriver
TARGET_URL=https://example.com
NAICS_CODES=111110,111120,111130
FINAL_OUTPUT_DIRECTORY=/path/to/output
LOGS=/path/to/logs
AWS_REGION=us-east-1
AWS_ACCESS_KEY_ID=your-aws-access-key
AWS_SECRET_ACCESS_KEY=your-aws-secret-key
EMAIL_SENDER=you@example.com
EMAIL_RECIPIENTS=recipient1@example.com,recipient2@example.com

5.2 Web Scraping

Uses **Selenium WebDriver** for navigating and extracting data:

- Handles dynamic elements using **Explicit Waits**.
- Supports **Pagination** for multi-page scraping.
- Website URL being scraped: https://sam.gov/
- Captures all these fields:

Contract Name	Notice ID	Department	Contract Link	Failed Row	Incomplete Data	Total Attachments	Date Scraped	Contract Number	General Published Date	Original Published Date	Updated Date Offers Due	Original Date Offers Due	File Name	File Link	Updated Date

5.3 Attachment Processing

For each contract, the system opens its detailed page and extracts:

- **Attachments**:
  - File name
  - File link
  - Updated date
- **Metadata**:
  - General Published Date
  - Original Published Date
  - Updated Offers Due Date
  - Original Offers Due Date

5.4 Data Processing

Consolidates all extracted data into a **Pandas DataFrame**:

- Merges contract and attachment details.
- Generates a summary row with:
  - Total contracts scraped.
  - Contracts with missing data.
  - Failed contracts.
  - Total attachments.

5.5 Email Notification

Uses **AWS SES** to send the compiled CSV file as an attachment:

- Creates a multipart MIME email.
- Sends to multiple recipients.

6. Features

6.1 Automation

- Fully automated end-to-end pipeline.
- Operates without manual intervention after initial setup.

6.2 Efficiency

- Processes hundreds of contracts in under an hour.
- Handles large datasets with minimal resource consumption.

6.3 Accuracy

- Uses dynamic element detection to ensure accurate data capture.
- Validates extracted data for completeness.

6.4 Reporting

- Consolidates all data into a single CSV file.
- Includes a summary row for quick insights.

6.5 Notifications

- Sends email reports with attachments to multiple recipients.

7. Error Handling

7.1 Scraping Errors

- Logs failures for missing elements.
- Skips failed rows and continues processing.

7.2 Email Errors

- Logs AWS SES responses for troubleshooting.

7.3 Summary Reporting

- Includes failed contracts and missing data in the final report.

8. Future Enhancements

1. **Parallel Processing**: Use multiprocessing for faster data extraction.
2. **Data Visualization**: Add charts and graphs to the final report.
3. **Cloud Integration**: Upload CSV files to Amazon S3 for centralized access.
4. **Advanced Logging**: Implement real-time dashboards for monitoring.

9. Conclusion

This system transforms contract data scraping from a labor-intensive task into a streamlined, automated process. By leveraging modern technologies, it reduces manual effort, improves accuracy, and delivers timely insights to stakeholders. The estimated **93% time savings** and enhanced reliability make it a valuable tool for organizations handling large volumes of contracts.

Output CSV:

https://docs.google.com/spreadsheets/d/1EmRIX5qXXVURgItGNxKZD4WhwhsC_1xNqsfThjZB_7g/edit?usp=sharing