CFPB Web Scraper Requirements

We aim to build an automated web scraping system to scrape and store the enforcement actions listed on the Consumer Financial Protection Bureau website.

We want to scrape each enforcement action listed on the CFPB website and store it for future access through our Neo4j database. Our system will create nodes containing the details of each enforcement action in our Neo4j database, and upload the documents referenced in the enforcement action to an S3 bucket. Links to the documents uploaded to S3 will be included in the graph database nodes for reference.

The scraper should be deployed to run periodically. On each run, it should look for any new enforcement actions that are added to the CFPB website, as well as compare the enforcement actions already scraped for any changes.

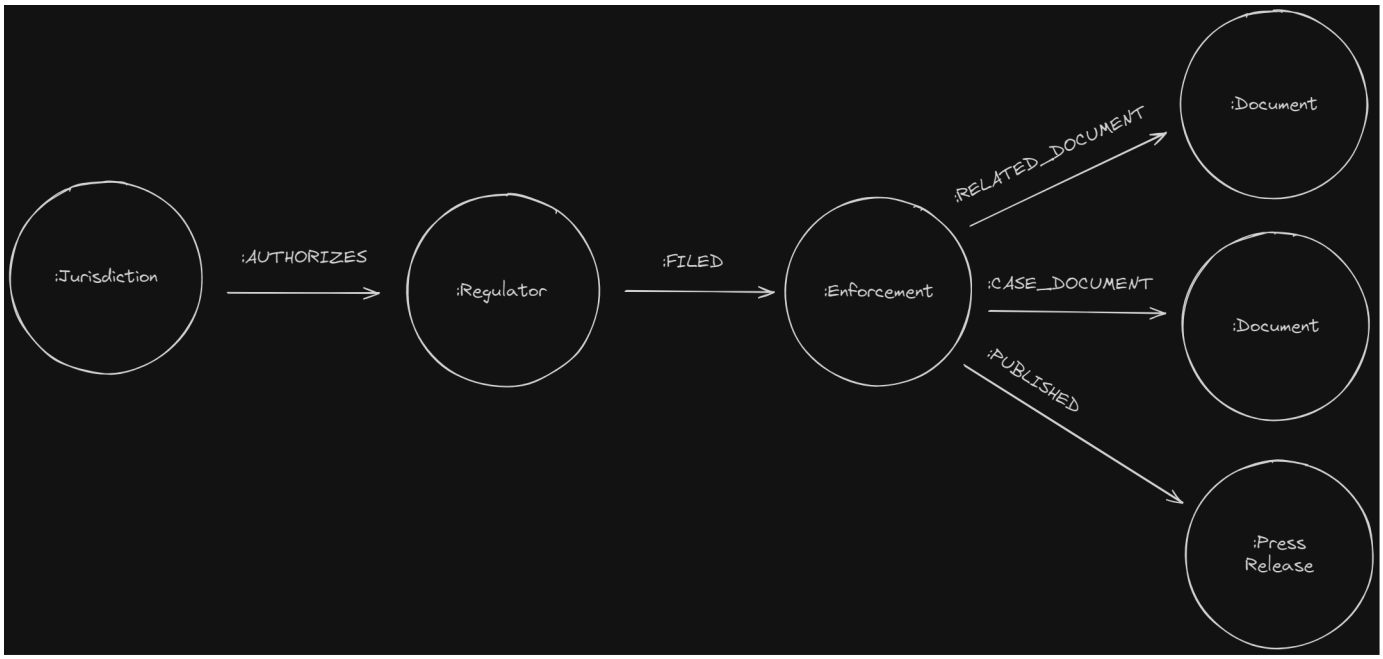
The activity of the scraper should be recorded in logs. The logs should include information such as:

- what information was scraped
- if any part of the scraping process failed or was incomplete
- if there were any changes to enforcement actions already scraped
- if a new enforcement action was added

We want to have a UI to view the scrapers activity in real time. The interface should display a list of the logs with pagination, and have basic search and filtering features for sorting the logs.

Database Outline

We will store the enforcement action information in a schema like:



Since for now we are just working with data scraped from the CFPB website, all the scraped enforcement actions will be attached to a "Federal" Jurisdiction node and "CFPB" Regulator node.

The Enforcement node will contain information like:

- Name (name of enforcement action)
- Date Filed
- Squib (Shortened description of enforcement action)
- Enforcement Body: (Body text of Enforcement Action)
- Press Release Link: (link to press release)
- Case Docket Link: (link to case docket)
- Forum
- Docket Number
- Initial Filing Date
- Status
- Body Links (links included in Enforcement Action body text)
- Products (products associated with the enforcement action)

The Press Release node will contain information like:

- Name (name of press release)
- Sub Heading (sub heading text of press release)
- Date Published
- Press Release Body: (Body text of press release)
- Body Links: (links included in the press release body text)
- Topics: (topics associated with the press release)

The Document node(s) will contain information like:

- Name (name of document)
- Path (path to s3 object)
- Date Filed
- Short Description
- Docket Number

For example, for this enforcement action we would create nodes that look like:

Enforcement Node


```
{
Name: "TD Bank, N.A",
URL: "https://www.consumerfinance.gov/enforcement/actions/td-bank-na-furnishing-2024/",
DateFiled: "Sep 11, 2024",
Squib: "On September 11, 2024, the Bureau issued an order against TD Bank, N.A...",
EnforcementBody: "On September 11, 2024, the Bureau issued an order against TD Bank, N.A, a
national bank headquartered in Cherry Hill, New Jersey...(full body text)",
PressReleaseLink: "https://www.consumerfinance.gov/about-us/newsroom/cfpb-orders-td-bank-to-
pay-28-million-for-breakdowns-that-illegally-tarnished-consumer-credit-reports/",
CaseDocketLink: "https://www.consumerfinance.gov/administrative-adjudication-
proceedings/administrative-adjudication-docket/td-bank-na-furnishing-2024/",
Forum: "Administrative Proceeding",
DocketNumber: "2024-CFPB-0009",
```

```
Status: "Post Order/Post Judgment",
BodyLinks: [],
Products: ["CREDIT CARDS", "DEPOSITS", "FURNISHING"]
}
```


Press Release Node


```
{
Name: "CFPB Orders TD Bank to Pay $28 Million for Breakdowns that Illegally Tarnished Consumer Credit Reports",
URL: "https://www.consumerfinance.gov/about-us/newsroom/cfpb-orders-td-bank-to-pay-28-million-for-breakdowns-that-illegally-tarnished-consumer-credit-reports/",
SubHeading: "TD Bank's illegal behavior threatened the ability of tens of thousands of consumers to access credit, housing, and employment",
DatePublished: "Sep 11, 2024",
PressReleaseBody: "**WASHINGTON, D.C.** - Today, the Consumer Financial Protection Bureau (CFPB) ordered TD Bank to pay $7.76 million to tens of thousands of victims of the bank's illegal actions. (full body of text)",
BodyLinks: ["https://www.consumerfinance.gov/enforcement/payments-harmed-consumers/civil-penalty-fund/", "https://www.consumerfinance.gov/enforcement/actions/td-bank-na-furnishing-2024/", "https://www.consumerfinance.gov/about-us/newsroom/cfpb-announces-settlement-td-bank-illegal-overdraft-practices/", "https://www.consumerfinance.gov/consumer-tools/credit-reports-and-scores/", "https://www.consumerfinance.gov/about-us/newsroom/cfpb-finds-violations-of-credit-report-accuracy-requirements-including-for-survivors-of-human-trafficking/", "https://www.consumerfinance.gov/data-research/consumer-complaints/search/?date_received_max=2024-08-23&date_received_min=2011-12-01&page=1&searchField=all&searchText=credit%20furnishing%20td%20bank&size=25&sort=created_date_desc&tab=List", "https://www.consumerfinance.gov/complaint/", "https://www.consumerfinance.gov/enforcement/information-industry-whistleblowers/"],
Topics: ["FINANCIAL SERVICE PROVIDERS", "CHECKING ACCOUNT", "BANKING", "CREDIT CARDS", "CREDIT REPORTS AND SCORES", "ENFORCEMENT"]
}
```

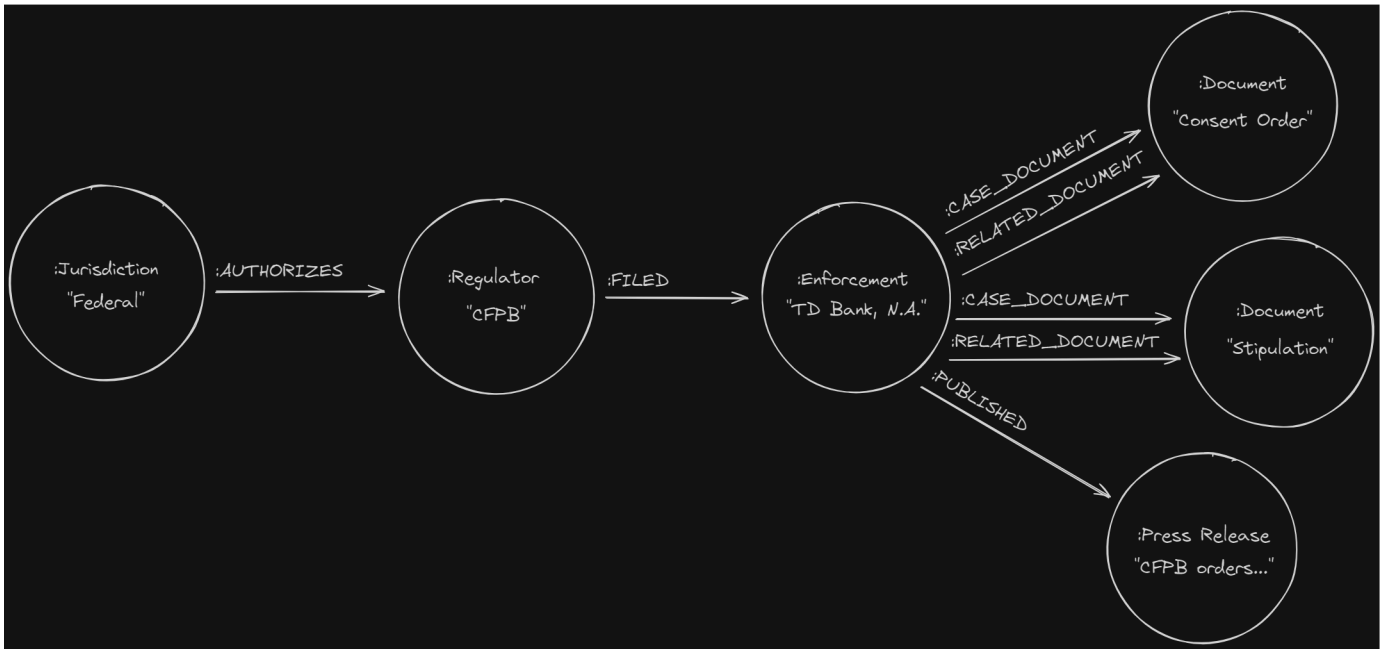

Document Nodes


```
{
Name: "Consent Order",
PathToS3: "(s3 endpoint)/Federal/CFPB/2024/September/cfpb_td-bank-na-consent-order_2024_09.pdf",
DateFiled: "09/11/2024",
ShortDescription: "Consent Order (Filed by Consumer Financial Protection Bureau)"
}
```



```
{
Name: "Stipulation",
PathToS3: "(s3 endpoint)/Federal/CFPB/2024/September/cfpb_td-bank-na-stipulation_2024_09.pdf",
DateFiled: "09/11/2024",
ShortDescription: "Stipulation (Filed by Consumer Financial Protection Bureau)"
}
```


These nodes will come together in a graph that looks like:



Information stored in the graph database will follow a normalized structure. In the enforcement action used for this example, the case docket documents are also included in the enforcement actions "Related Documents" section. The result is both a :CASE_DOCUMENT relationship and :RELATED_DOCUMENT relationship pointing to each Document node from the Enforcement action node.