


EDINET Data Ingestion Pipeline - Engineering Challenge

Background

EDINET is the Japanese Financial Services Agency's electronic disclosure system where all publicly listed Japanese companies publish regulated filings and disclosures. Reliable access to, and processing of, this data is crucial for analytical, compliance, and research applications.

Objective

Design and implement a robust, scalable, and fully automated ETL (Extract, Transform, Load) pipeline in Python to extract, process, and store all public filing documents for every listed company on EDINET (<https://disclosure2.edinet-fsa.go.jp/week0020.aspx>)

 EDINET Electronic Disclosure for Investors' Network	Font Size M L Japanese
<div>Top Page</div> <div>Detailed Document Search</div> <div>EDINET TAXONOMY & CODE LIST Download</div>	
<div>TopMenu</div> <div>INFORMATION SPEC OPERATION GUIDES</div> <div>SYSTEM MAINTENANCE INFORMATION</div> <div>None</div>	<div>Simple document search</div> <div> Document submitter / Securities issuer / Fund information / Stock Code <input type="text"/> <div>Search</div> <input type="checkbox"/> Including Funds </div> <div> Type of document <ul style="list-style-type: none"> <input checked="" type="checkbox"/> Annual Securities Report / Semiannual Securities Report / Quarterly Securities Report <input checked="" type="checkbox"/> Report of Possession of Large Volume <input checked="" type="checkbox"/> Extraordinary Report <input checked="" type="checkbox"/> Other types of documents (Including amended version of each report) </div> <div> Submission period <input type="text" value="In the entire period"/> </div>

Requirements

- **Data Extraction:** Automatically retrieve all publicly available filing documents for every listed Japanese company from the EDINET website using HTTP requests (preferred approach).
- **Data Storage:** Store extracted data in a MongoDB database, organizing it for efficient retrieval and association between companies and their filings.
- **Automation:** The pipeline should support end-to-end automation, including scheduling, queuing of extraction tasks, and progress tracking.
- **Performance & Scalability:** Ensure the scraper is high-performance (concurrent requests or task queues), and built for reliability and fault tolerance (automatic retries, robust error handling).
- **Statistics & Monitoring:** Collect and expose scraping and pipeline statistics (e.g., number of companies/documents processed, errors, processing times) for monitoring and debugging.

- **Extensibility:** Your design should facilitate easy integration with other services that may consume individual filings, metadata, or statistics from the pipeline.
- **Documentation:** Provide a concise README explaining your architecture, handling of edge cases, and instructions for running the pipeline.

Tips

- Consider efficient ways to structure company and document relationships in MongoDB for real-world query needs.
- Plan for large data volumes: thousands of companies and potentially hundreds of thousands of documents.
- Reliability, speed, and service readiness are highly valued.

Simple document search

Document submitter /
Securities issuer /
Fund information /
Stock Code

tyo

Search

☐ Including Funds

Type of document

☒ Annual Securities Report / Semiannual Securities Report / Quarterly Securities Report

☒ Report of Possession of Large Volume

☒ Extraordinary Report

☒ Other types of documents

(Including amended version of each report)

Submission period

In the entire period

Date & time of submission	Submitted document	Code	Submitter / Fund	Issue / Subject / Subsidiary	PDF	XBRL	CSV	Remarks
2025/06/13 15:40	Confirmation Letter	E05353	CARE TWENTYONE CORPORATION		PDF			
2025/06/13 15:40	semi-annual securities report-32th Period(2024.11.01-2025.10.31)	E05353	CARE TWENTYONE CORPORATION		PDF	XBRL	CSV	
2025/02/26 17:01	Extraordinary Report	E05353	CARE TWENTYONE CORPORATION		PDF	XBRL	CSV	Art 19 para 2 item 3
2025/01/30 15:45	Extraordinary Report	E05353	CARE TWENTYONE CORPORATION		PDF	XBRL	CSV	Art 19 para 2 item 9-2
2025/01/30 15:34	Internal Control Report-31th Period(2023.11.01-2024.10.31)	E05353	CARE TWENTYONE CORPORATION		PDF	XBRL	CSV	
2025/01/30 15:33	Confirmation Letter	E05353	CARE TWENTYONE CORPORATION		PDF			

Deliverables

- Python codebase implementing your solution, ready to run with minimal setup.
- Instructions for setup, configuration, and usage (e.g., in a README).
- Statistics into how many publicly listed companies were scraped and the distribution of different files and folders across the companies.
- A one-page doc summarizing how you have built this pipeline end-to-end.
- Any additional files (e.g., requirements.txt, Dockerfiles, sample configs) needed to use your pipeline

