Interview Question

Owner	Shubho Shaha
Tags	

Take-Home Interview Question: Implement a Generic Web Crawling System

Overview

Design & Build a production-ready, scalable web crawling system that can crawl both static and dynamic websites, extract structured data, and handle various edge cases at scale.

Focus: 70% Implementation, 30% Design

Language: Python (perfered). But you can choose any language you are comfortable with. You can use any tools/frameworks you want as long as your solution works at scale in produciton system.

Problem Statement:

Implement a generic web crawler that can:

1. Crawl both static HTML and JavaScript-rendered (dynamic) websites
2. Follow links intelligently with depth and domain control
3. Extract and structure content from crawled pages
4. Handle rate limiting, retries, and errors gracefully
5. Run distributed across multiple workers
6. Store crawled data efficiently

What Should Be Delivered:

1. GitHub Repository containing:

- Required Code Files:
    - `src/crawler/` - All core crawler implementations
    - `tests/` - Unit and integration tests (>70% coverage)
    - `docker-compose.yml` - Local development setup
    - `Dockerfile` - Container image
    - `examples/` - At least 3 working example scripts
    - `requirements.txt` - Python dependencies
- Required Documentation:
    - `README.md` - Complete setup instructions, usage examples, design overview

1. Submission Format:
    - Public GitHub repository (or shared with interviewer)
    - Repository must be clone-able and runnable following README instructions
    - All tests must pass
    - Docker Compose must start successfully