System Design - Crawler

Design a web crawling system that can visit millions of web pages per day, extract structured data, and store it for later analysis. Consider scalability, fault tolerance, and security.

Functional

1. Accept a list of URLs to crawl
2. Download HTML and relevant assets
3. Parse and extract structured data
4. Avoid visiting duplicate URLs
5. Store extracted data

Non-Functional

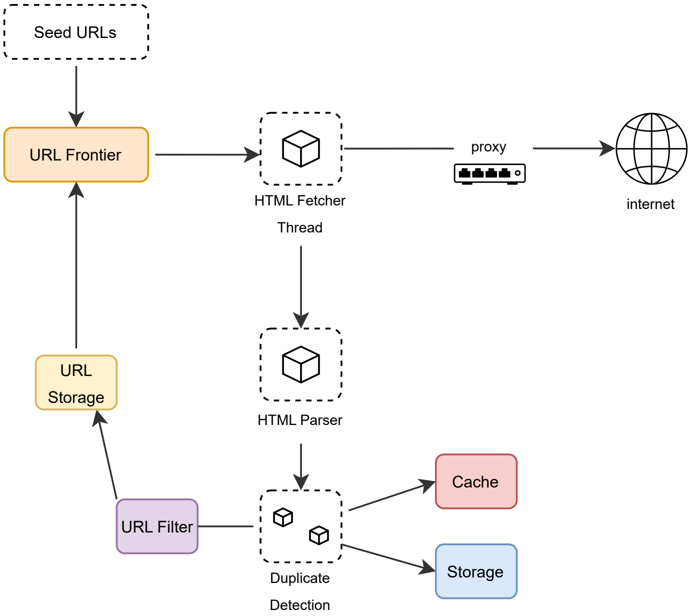
1. High concurrency
2. Rate limiting
3. Fault tolerance and retry logic
4. Security

Storage:

2 KB (per website) * 100,000,000 (sites month) = 200,000,000 KB = 200,000 GB = 200 TB

Band width:

4,500,000 (sites day) * 10 KB (HTML) = 45,000,000 KB = 45,000 GB = 45 TB/day * 30 = 1350 TB/day


Seed URLs
URL Frontier
HTML Fetcher Thread
proxy
internet
HTML Parser
URL Storage
URL Filter
Duplicate Detection
Cache
Storage