# Requirements and constraints

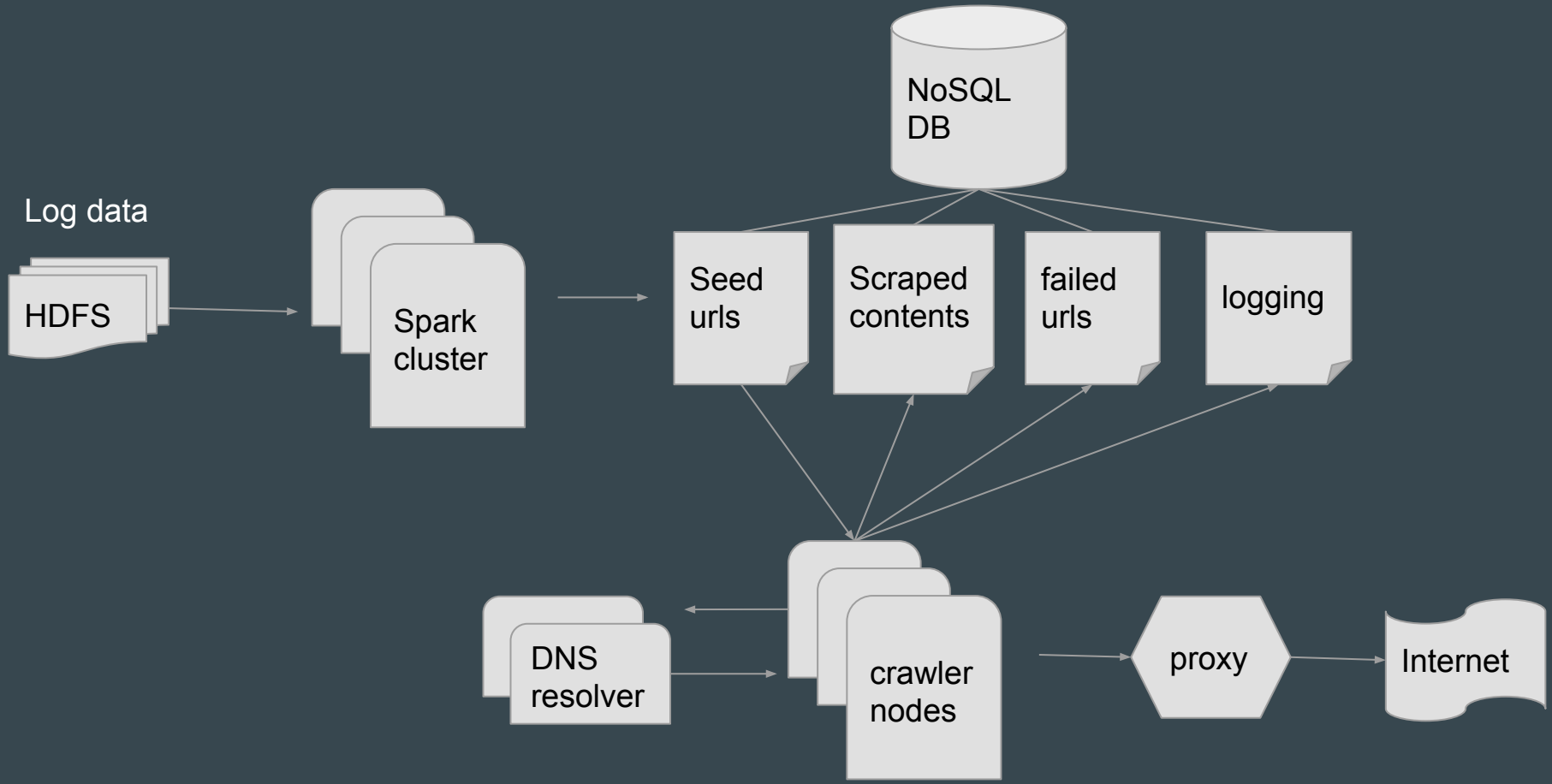
Receive 5M urls per day

Storage: 5KB per web page; 6T per year

No need to get result in real time; batching processing

No need to extract other urls from the scraped page

On-prem servers


NoSQL
DB
Log data
HDFS
Spark
cluster
Seed
urls
Scraped
contents
failed
urls
logging
DNS
resolver
crawler
nodes
proxy
Internet

# Data validation

Filter user types

Whitelist / blacklist

Discard urls that have been scraped

# Url parsing

Extract domain

English domains: net, com, ca, org, uk

Remove query parameters *https://www.domain.com/page?key1=value1&key2=value2*

Remove duplicate urls

# Seed urls

Table partitioned by domain, time

Urls with same domain are ranked by time order

Each time crawler worker fetches certain number of urls.

| url | domain | timestamp |
| --- | --- | --- |
| www.nn.com/gg/ba/ | www.nn.com | 2022-09-01 19:20:05 |
| www.aa.net/jdf/nnns | www.aa.net | 2022-09-01 19:23:05 |
| www.jj.ca/kk/nhn/mn | www.jj.ca | 2022-09-01 19:26:05 |

## Distributed Crawler - Scrapy

DNS caching

Multiple nodes; multi-threading

Fetched urls are kept in memory.

Timout

Concurrent_Requests_per_Domain;

AutoThrottle; download_delay

## HTTP request

accept-language = en

Rotate user-agent;

Disable cookies;

# Content parsing

Texts only

Only select certain html tags

Filter out non-English texts

Size limit

## DNS resolver

Internal DNS server

Caching

## Proxy

Rotate IP

# Scraped contents

| url | title | content |
|---|---|---|
| www.nn.com/gg/ba/ | good | Www hhh nnn mmm |
| www.aa.net/jdf/nnns | great | Www hhh 2432 er |
| www.jj.ca/kk/nhn/mn | wonderful | 0ew fj 442 |

# Fail-to-Scrape urls

| url | attempt | timestamp |
|---|---|---|
| www.nn.com/gg/ba/ | 1 | 2022-09-01 19:20:05 |
| www.aa.net/jdf/nnns | 1 | 2022-09-01 19:23:05 |
| www.jj.ca/kk/nhn/mn | 2 | 2022-09-01 19:26:05 |

# Summary

Availability

Reliability

Scalability

Performance

Politeness

# references

https://www.enjoyalgorithms.com/blog/web-crawler

https://medium.com/analytics-vidhya/web-scrapping-html-parsing-and-json-api-using-python-spider-scrapy-1bc68142a49d

https://www.scrapehero.com/how-to-prevent-getting-blacklisted-while-scraping/