System Design Web Crawler

FR : 1) The web crawler crawls pages and indexes the content and links to other pages.

2) The crawler should not crawl same page twice

3) The users can manually provide seed URLs which should be taken on priority

4) Analytics and performance monitoring.

NFR: 1) Highly available

2) Fast. Should be able to crawl large no of pages quickly.

Capacity estimation

1) Bandwidth - assume 1B new URLs generated every month. each page size is 100 KB. => 1/(30*24*60) ~= 23k URLs per minute => 23k x 100KB ~= 2 GB data to fetch per min

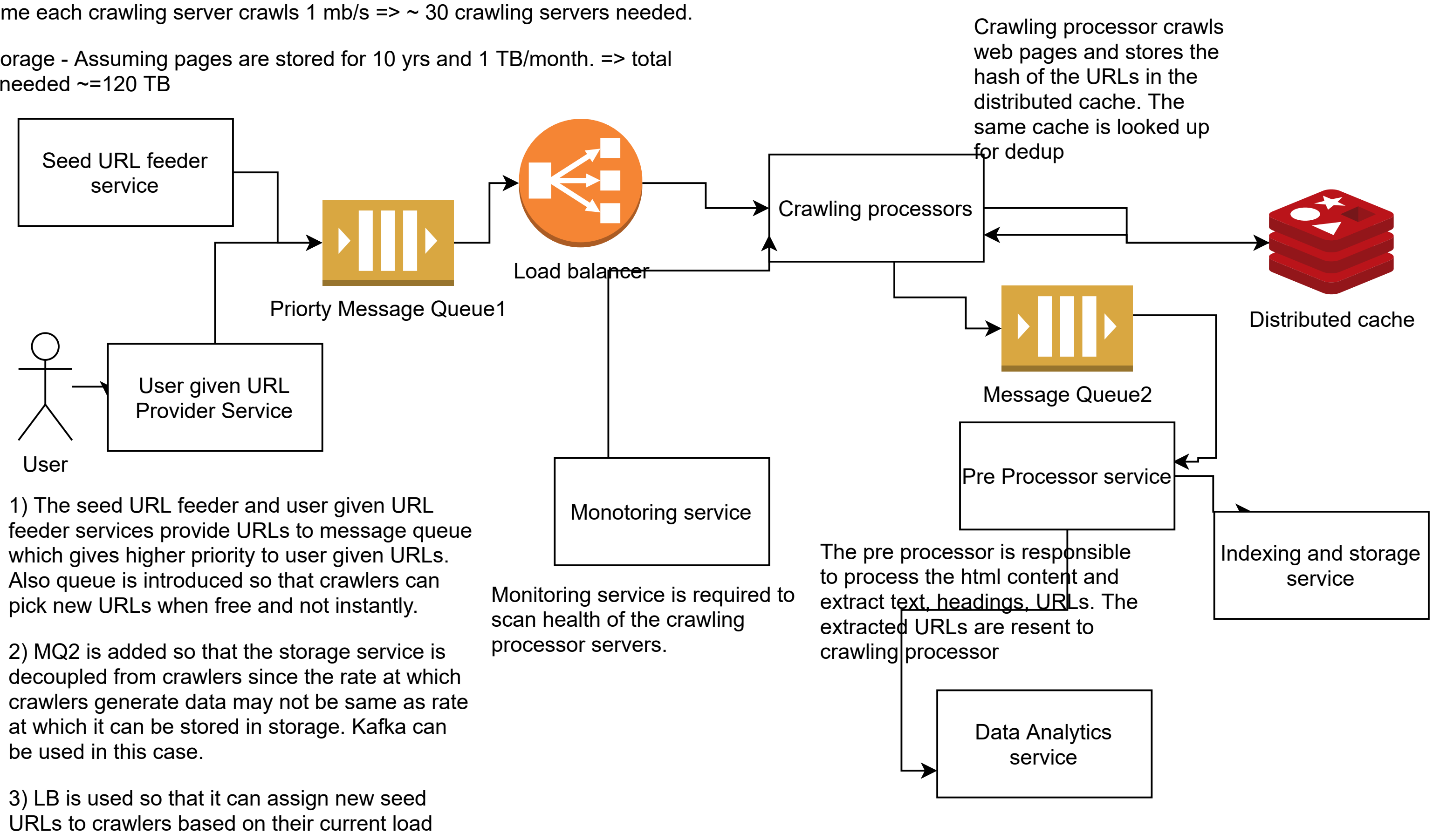
assume each crawling server crawls 1 mb/s => ~ 30 crawling servers needed.

2) Storage - Assuming pages are stored for 10 yrs and 1 TB/month. => total size needed ~=120 TB

Data storage

Preferably high available nosql since there is no relational schema needed. The system just needs to store K-V pairs where key is URL and value is the page meta data. Also the NoSQL is easily scalable which is needed considering size of incoming data.

The distributed cache needs to store hashed values of URLs to save space. It is used since it is fast which is needed for lookup of URL existence quickly



Crawling processor crawls web pages and stores the hash of the URLs in the distributed cache. The same cache is looked up for dedup

1) The seed URL feeder and user given URL feeder services provide URLs to message queue which gives higher priority to user given URLs. Also queue is introduced so that crawlers can pick new URLs when free and not instantly.

2) MQ2 is added so that the storage service is decoupled from crawlers since the rate at which crawlers generate data may not be same as rate at which it can be stored in storage. Kafka can be used in this case.

3) LB is used so that it can assign new seed URLs to crawlers based on their current load

Monitoring service is required to scan health of the crawling processor servers.

The pre processor is responsible to process the html content and extract text, headings, URLs. The extracted URLs are resent to crawling processor