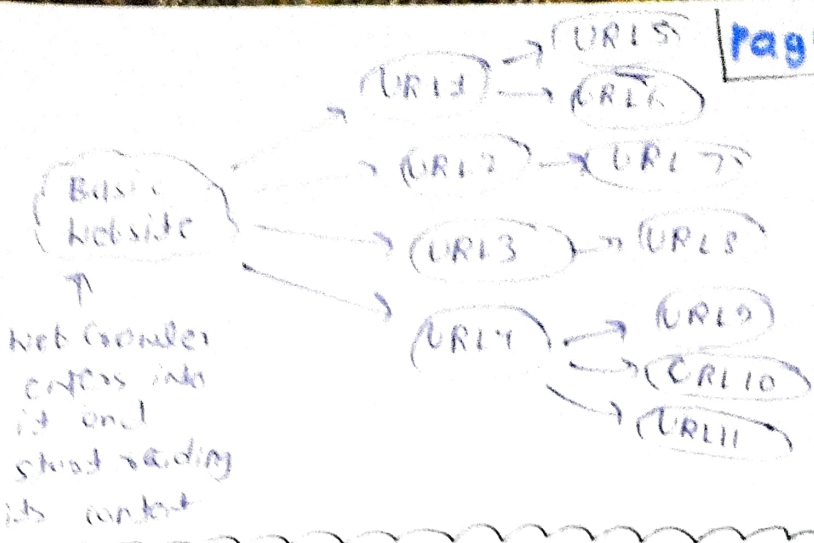
① Design a Web-Crawler

When you try to search any webpage on the Google, So Google does not do crawling at that time because it does not start crawling at time of search, if it do so, Web-crawler will take a lot of time and you normally see that google returns the result in hardly ±-2 second.

So, Google runs Web-crawler in Back-End continuous in order to provide the fast result.

In December 2023, Around 1.2 billion websites are there on Google.

Web-crawler pick one website, crawl over all URL present in it and try to find out URL from the website by which it enters into new website.



So, normally, web-crawler starts with some basic website links, these URLs are called as **Seed URLs**

How to decide Seed-URLs

Suppose I want to make an application that beat all the e-commerce web applications, Now, we can beat all the e-commerce web-applications if we lower the price of the item by comparing it with all other e-commerce web-application. As customer will use that e-commerce web-application that offer the item in lowest price.

So, typically we need to use web-crawler in order to do the analysis of the price of the item in which they are sold.

Hence, I use web-crawler and seed-URLs that I will use typically will be of E-commerce web apps so that it can crawl over almost all the E-commerce web application and returns the result.

Where Web-Crawler is used

1. Used in Search-Engines like Google, Bing etc.
2. When you make projects in college, there are many software tools that will detect copyright violation detection / ensures identifying the copying of reports not done
   - copying → Plagiarism
3. We use web-crawling in the field of Data science in order to do data fetching
4. Keyword-based search we require web-crawling.

(5) In malware detection, web-crawling is also used.

Steps to build a system

(1) Try to understand what the system actually means which you are designing, how does it works

(2) Do functional Requirement Gathering of that system

Functional Requirements :-

(a) Scan and store data for all the websites.

(b) Donot store duplicate records.

Why websites most of the time allow them crawled

Typically, if a website does not want itself to be crawled, we can make something use of the flag called as Robots.txt, it is a file that tell search engine crawlers which URLs the crawler can access on your site.

But disadvantage here is that Google will not index you because if you do not allow google you to scan so you connot come up in the initial results.

Captcha you do not usually allow to be crawled.

Non-Functional Requirements :-

(a) Security should be higher since mostly web-crawler system are internal system so it is easy to achieve.

(b) It should be scalable so as to take more load.

(c) It should be Reliable system

(d) It should to fast.

(e) It should be polite on website because crawlers try to hit URLs and read their content

Explain of Polite :-

As a crawler, your intention is only to read its content, but since if you make your crawler more aggressive that you bombarded the website by impacting the performance of the website as you have not gone slow and easy on the website and you started competing with your own customers, you never want that your customer experience is decreased by increasing response time of the website, you cannot do that you are using multiple web-crawlers reading different pages of the website at the same time, this might can decrease users experience.

Note:- How to Avoid the website Hacking by Web-Crawling. Especially if your website is public read only content.

If your website has not public content, if someone is hitting a URL regularly from a specific zone/ specific IP Address, we will try to make it backlist by using Authentication/Authorization in the action i.e access control so that you can blacklist few of the users based on the hostname/ IP Address.

But if your website is public content & it is open, you normally don't have much control over who is trying to hit you because you are not storing that so basically you have the provision to understand that from which IP you are getting the request and blocked those IP.

(3) Get Estimations

Estimate the scale of the Problem

(a.) Total number of Website ≈ 1.2 Billion
We check that how many websites are active from 1.2 billion websites, many of the website will be government so noone is watching those websites.

(b.) Let's try to make assumption that 60% of website are active.

Because many of the website might have nude / illegal content like pornography, so in these kinds of data, we are usually not interested in, we are only interested in only valuable data providing website

60% of 1.2 Billion = 720 Million websites

Sometimes, Some of the website might contain only 1 page, some website might contain 100 pages Ex:- many websites like facebook / Wikipedia can have millions of pages.

Whenever you are doing these kind of assumptions, always speak loud that what assumptions you are doing and always keep your interviewer in conference.

So, lets assume that number of webpages / website = 100

So, Total number of Webpages / URLs = 720 Million × 100
= 72000 Million
= 72 Billion

Now, since 72 Billion URLs so this is a huge number which my crawler has to crawl so, my crawler cannot be done by the single machine, it must use a distributed system.

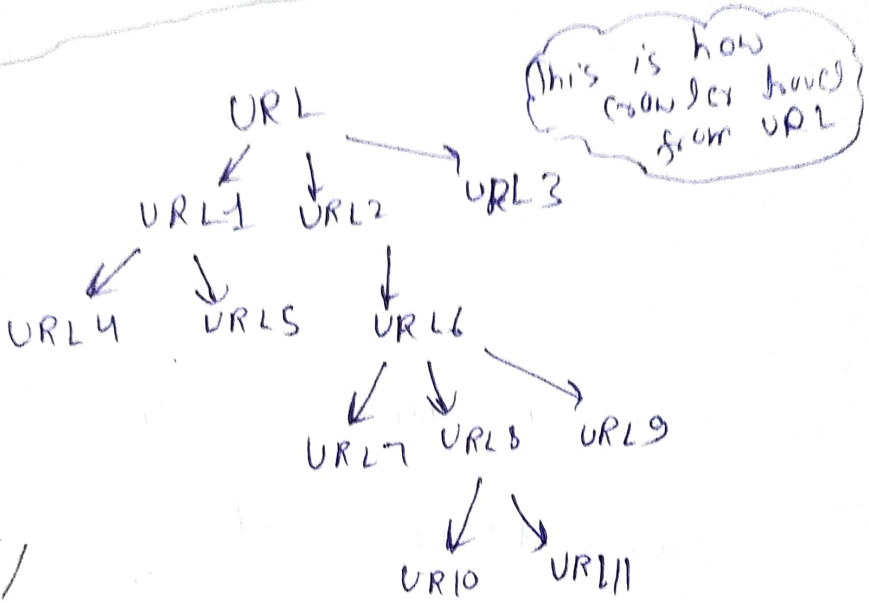
(d) Let's say that I am only storing the text data so crawler is not storing the images / video / gif files etc. except text content of the website

(e) Let Avg size of the Web page = 100 KB
No. of pages = 72 Billion
∴ Total size of Webpage to be stored
= 72 Billion × 100 KB
= 7.2 PBs
↓
Petabytes

If I apply some compression Algorithm let 50% compression is done → 3.6 PB of Data my crawler storage will store

(-) So, web-crawler is working at Big-scale. When you crawl multiple times in order to update the data stored

Note:- Caching has no limit if your RAM is 1TB, it can hold 1TB of data, Caching limit will be 1TB, all these things are determined by RAM



Complete graph / Undirected graph

↳ Graph traversal we typically have BFS / DFS

Where to use BFS / DFS?

Both of them return all possibilities eventually

BFS always give priority to the neighbours, when you try to find the details around you, you can use BFS Algorithm, basically BFS is used to find out the **shortest distance.**

So, Basically if you want to the result at further distance, you typically make use of DFS Algo.

Since we are designing the web-crawler that provide comprehensive coverage of the website, so BFS will also make use of a lot of sense here.

Note:- If I have to go search on specific topic and want to go into deep, we prefer DFS in this case.

so in scanning / crawling all the websites, the BFS approach is used

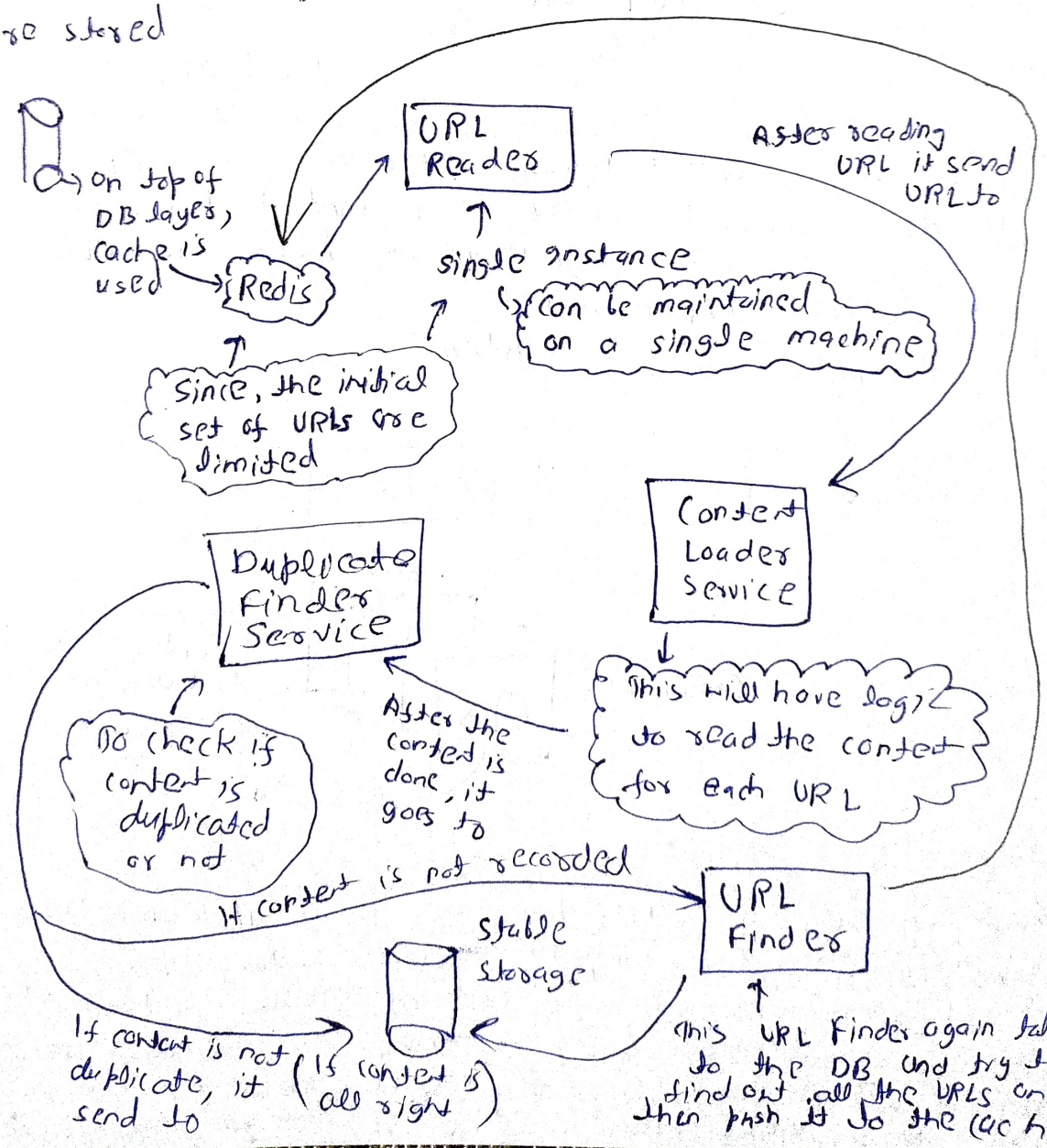
In maze problem where result is further from the starting point, we make use of DFS algorithm

When you do the google search your results usually come from the indexed results, and this indexed result come from stored data and this stored data come from web-crawler. So in google search no web-crawling is done Result are maintained in cache/storage by web-crawler already

(4) Components identification

(a) We want seed URLs, they can be in form of file / database

(b) URL Reader :- There should be a component which read from the database/file where URLs are stored



3.6 PB of data is usually stored in stable storage.

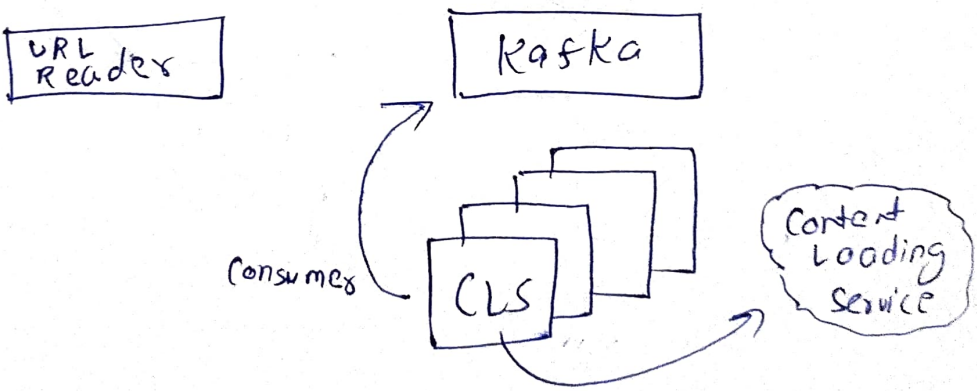
As the thing get heat up, when load increase many URLs came up in cache i.e Redis, and once many URL is read by URL readers and send to content loader service, this start scaling issue, this is because the rate which URL reader will be getting up the URL is much higher than the rate at which content reader will read all the content from that URL.

So, we need to scale up Content Loader service. Note:- Hitting the URL, Fetching the Data and Reading the content is the slow operation.

Since, URL reader is working at much faster rate and Content loading service at slower rate and both system are working synchronously this leads to cascading failure.

Instead of synchronous b/w URL reader and Content loading service i.e KAFKA is Message Queue and do Horizontal Scaling of content loading service, we bring asynchronousity here so CLS now read from KAFKA service

Changes done in system:-



Lambda / Serverless

You typically use this when you have intermittent operation or low compute services. In CLS, it is not staying vacant because of scale at which it need to operate so it does not have time to rest. You typically use Lambda when your server has time to take rest, when your server is not getting the continuous

you get request once too a while and then you have nothing to do so typically in all those things you make use lambda / serverless Architecture.

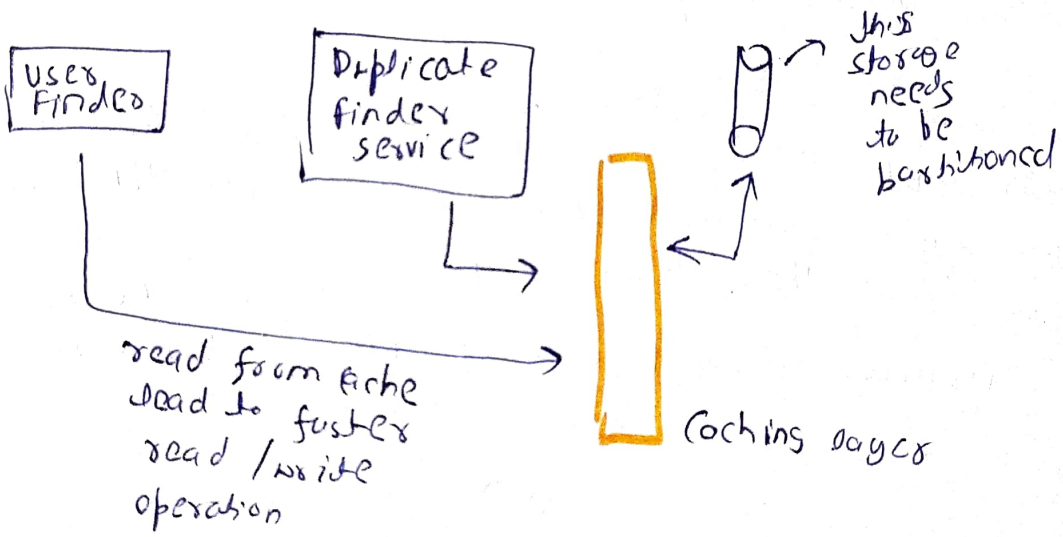
How to store 3-6 PB of Data?

We can use Distributed storage like HDFS / Cloud Storage / Data lake, in these system due to disk i/o, this make read/write operation slow Due to this, At distributed data layer, we need to use some strategy:-

(a) Partitioning

(b) Caching → used to store hot data / latest data, so redis is used as cache here.

So, Changes done in system:-



So, you use caching to speed up the read.

Caching Type

In-Memory

Ex- Memcache

You lose data after every restart / failure

Consistent / Stable

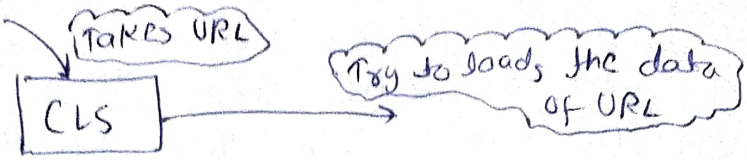
Ex:- Redis provide you both stable and in-memory data so you have strong hold on data.

You do not lose data even after restart

It keeps on making a copy in stable storage also

(5) Checking scope of improvement

Biggest player in this system is CLS



For getting the data of URL

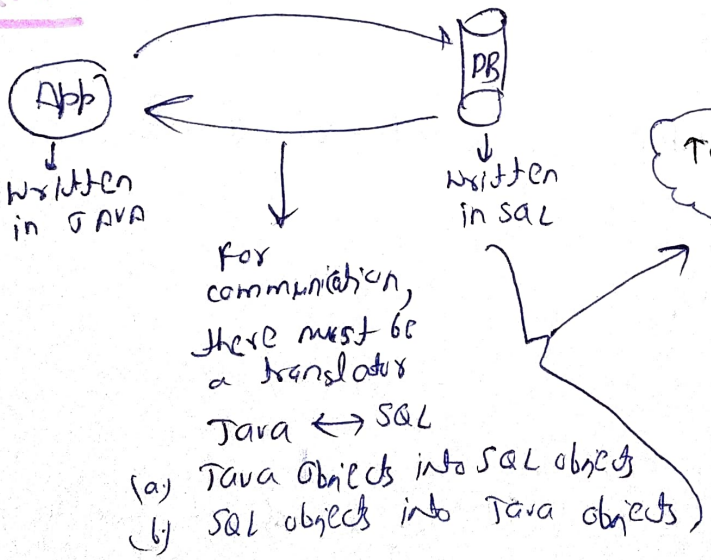
① It has to do DNS Lookup, where it find the corresponding IP Address of URL. Now to boost CLS action, I have to put DNS lookup in Local Cache, instead of telling control loader service to go to the DNS server and find out what is the corresponding IP address of the URL, if I manage it in local cache, so everytime I get a URL, I know the corresponding address from cache easily.

② This improve the speed of the system

Noted Points

In HLD, if you have to decide cloud & On-Promise you usually not decide these things as per HLD level, You decide these thing as per Org level. So SDE usually do Org Level Design followed by HLD and then LLD.

ORM :-



Tool technique used is ORM for doing the translation

ORM makes the life of developer easy but it does not play any role in the optimization