design a web crawler

contacts at school, people
if i knew all the websites

10 pages / school
500schools ⇒ 10k pages to crawl → 10B unique pages
seed → how are we gonna know the schools
listings

url,content,priority,
robots – (url, robots) if UH says crawl me every hour
global-index: URL, hash (100B, 16B) → 1000KB = 1MB ← cache
(redis-cache-disk)
data-table: name, phone-number, URL → 10K * 100 * 100B = 1GB

frontier: urls
workers: for each url, they will hash it, to know if it has been
crawled

BFS/DFS
—---

queue
100B – 4B
4B * 10K == 10KB

multiple directories
user
out of scope would be like to crawl online phon