Bloom Filters

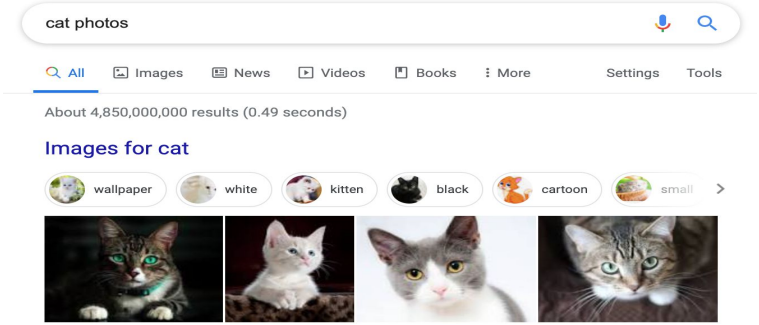
Learning Objectives

1. Understand when Bloom Filters are used

Memory Constrained Environments

What method would you use to build a search index on a collection of objects *in a memory-constrained environment*?

Constrained by Big Data (Large N)



Memory unit	Description
Kilo Byte	1 KB = 1024 Bytes
Mega Byte	1 MB = 1024 KB
Giga Byte	1 GB = 1024 MB
Tera Byte	1 TB = 1024 GB
Peta Byte	1 PB = 1024 TB
Hexa Byte	1 EB = 1024 PB
Zetta Byte	1 ZB = 1024 EB
Yotta Byte	1 YB =1024 ZB
Bronto Byte	1 Bronto Byte = 1024 YB
Geop Byte	1 Geo Byte = 1024 Bronto Bytes

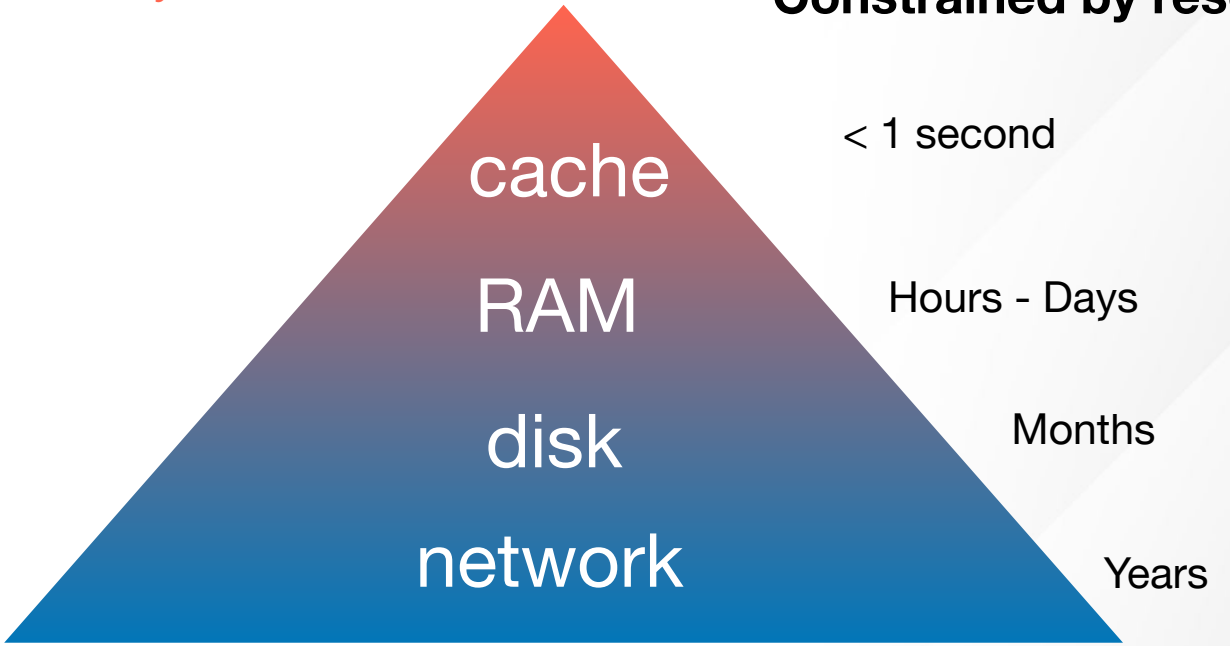
Google Index Estimate: > 400 billion webpages

Size of Internet (2022): 175 Zetta bytes

Speed in Different Spaces

What method would you use to build a search index on a collection of objects *in a memory-constrained environment*?

Constrained by resource limitations



< 1 second

Hours - Days

Months

Years

(Estimates are Time x 1 billion courtesy of https://gist.github.com/hellerbarde/2843375)

Reducing Storage Costs

1. Compression

2. Don’t store information you don’t need

Bloom Filters

1. Probabilistic Data Structure
2. Space Efficient
3. Tests whether an element is in a set
    a. Helps prevent looking for files that don’t exist in databases
    b. Identify if website are potentially malicious
    c. Minimizes caches misses for browsers