Arquivo.pt image search **2020 → 2021**

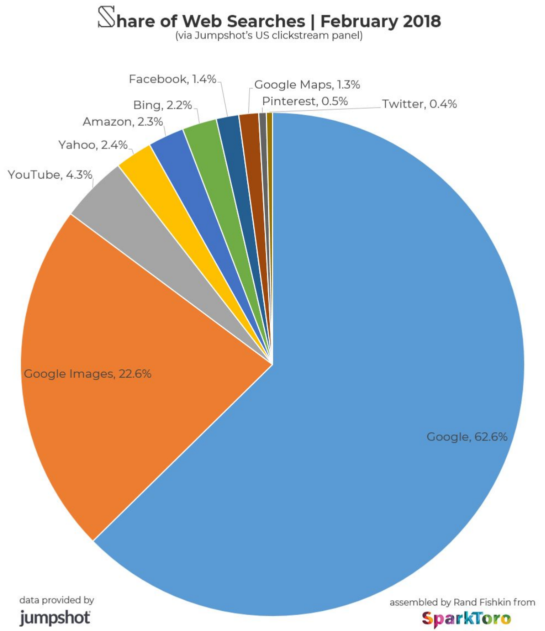
February 9th 2021

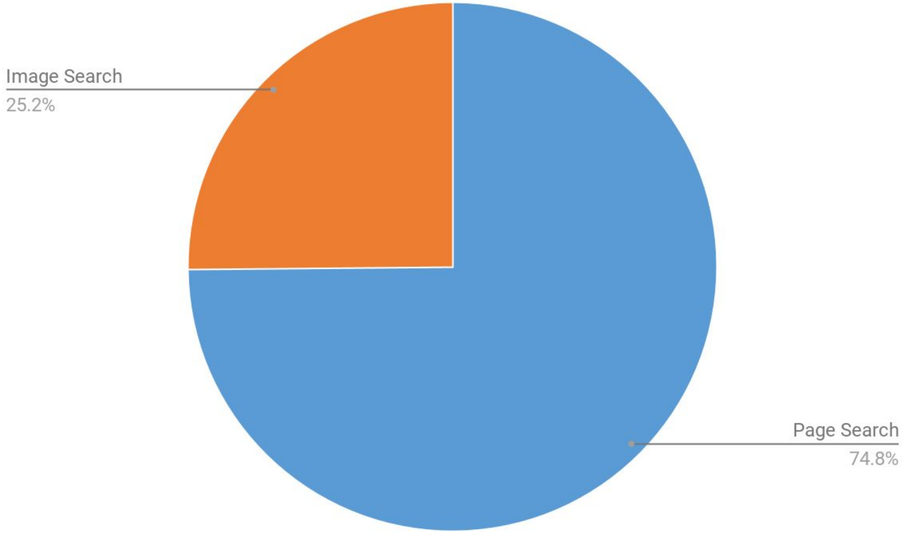
André Mourão
R&D engineer
andre.mourao@fccn.pt

ARQUIVO.PT

FCT
Fundação
para a Ciência
e a Tecnologia

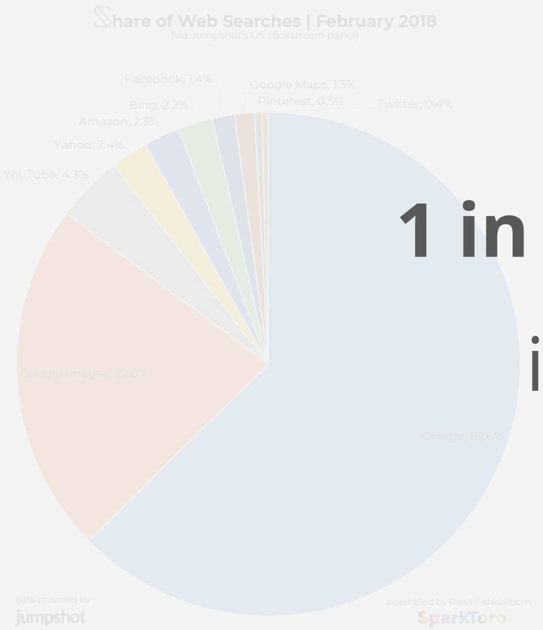
Why does image search matter?





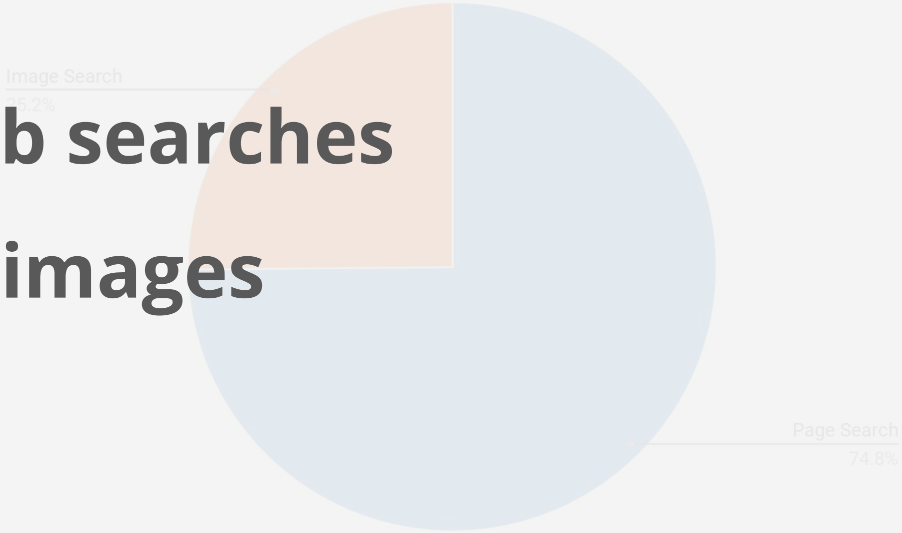
sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Why does image search matter?



1 in 4 web searches

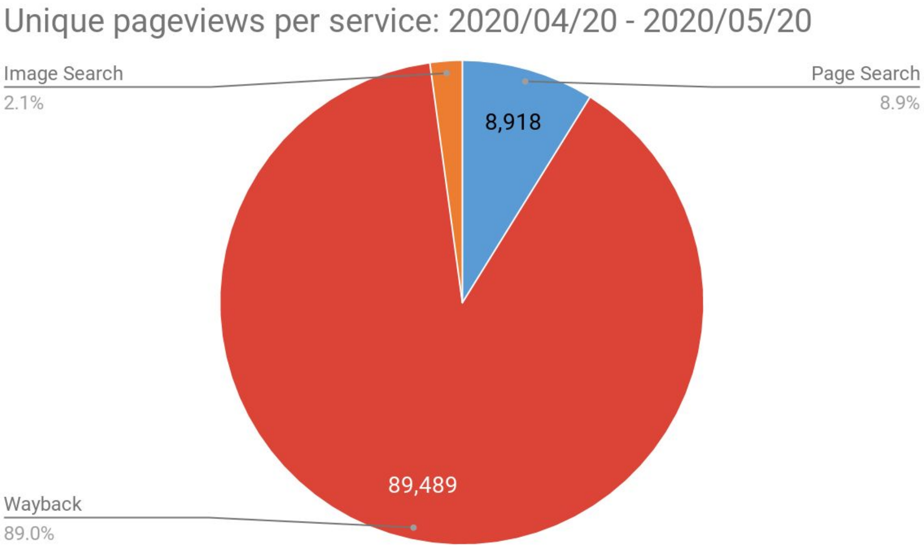
is for **images**

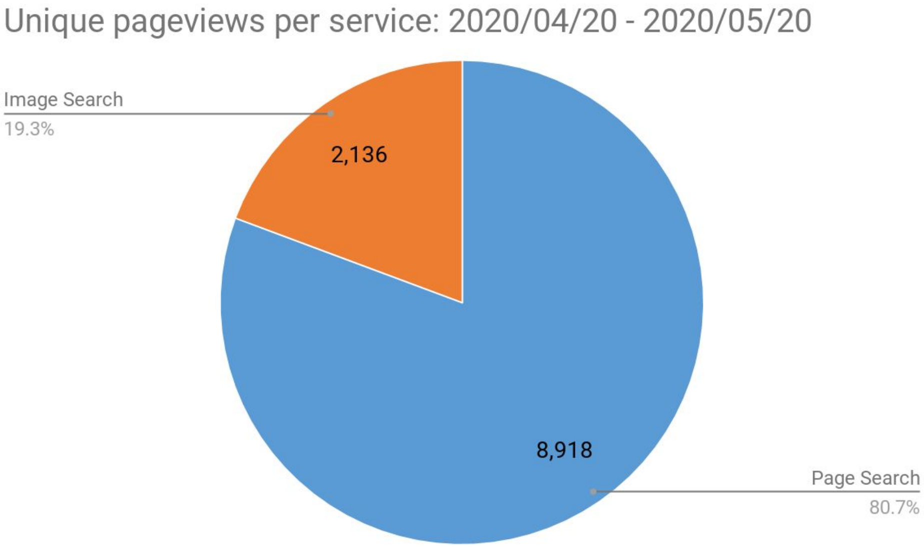


1 in 4 web searches is for images

sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

What about Arquivo.pt?



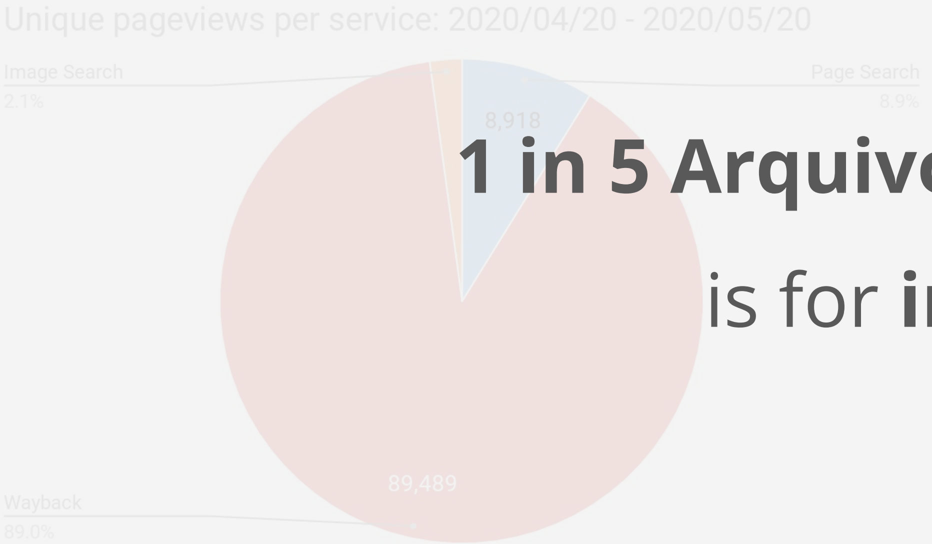


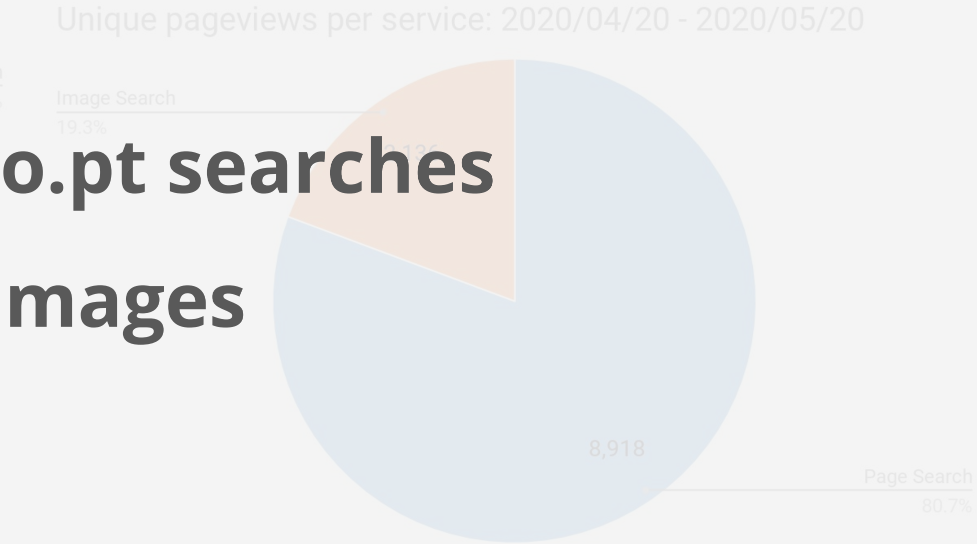
Arquivo.pt Google Analytics

What about Arquivo.pt?

1 in 5 Arquivo.pt searches

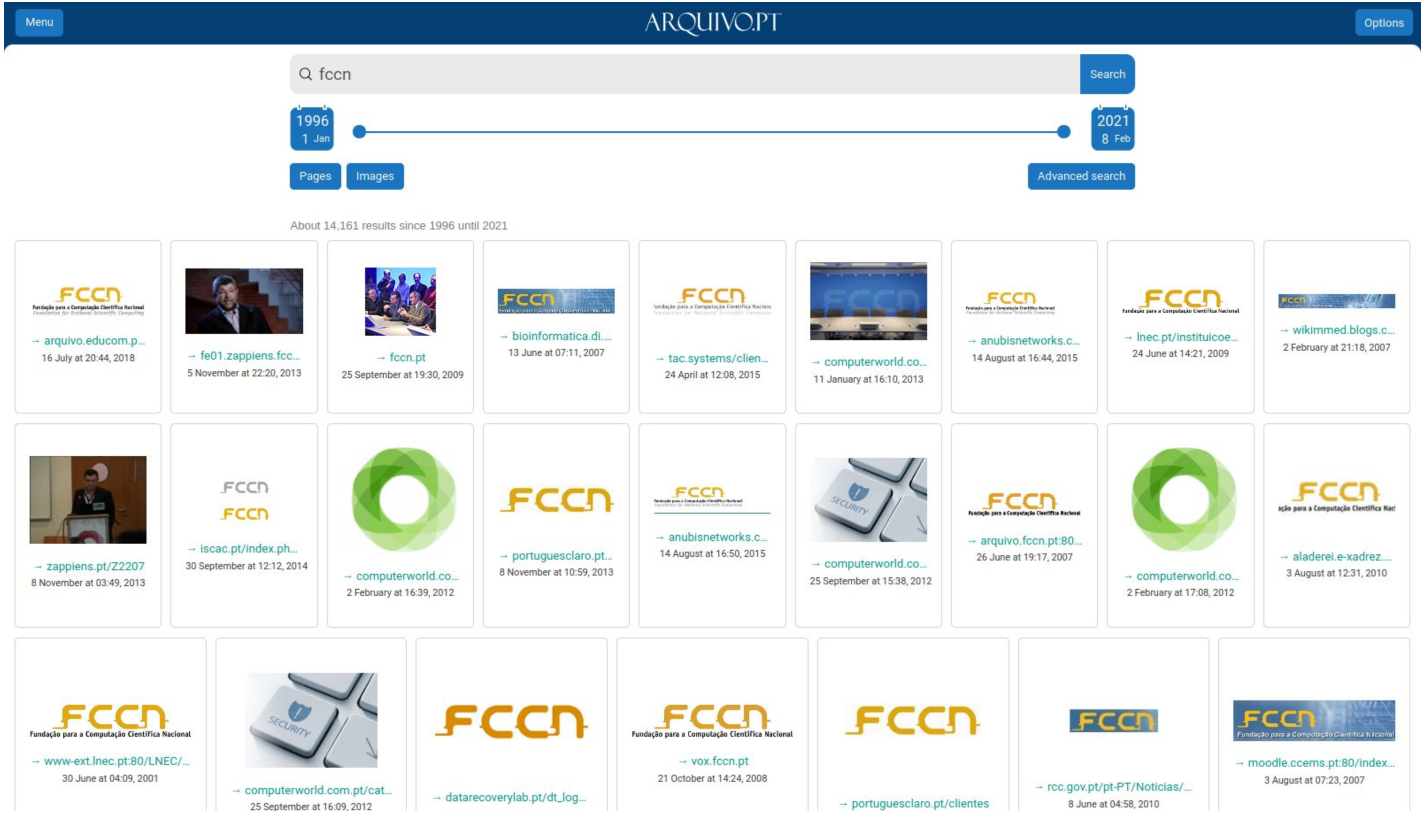
is for **images**





Arquivo.pt Google Analytics

Arquivo.pt Image Search (as of Jan 2020)



About 14,161 results since 1996 until 2021

→ arquivo.educom.p...
16 July at 20:44, 2018

→ fe01.zappiens.fcc...
5 November at 22:20, 2013

→ fccn.pt
25 September at 19:30, 2009

→ bioinformatica.di....
13 June at 07:11, 2007

→ tac.systems/clien...
24 April at 12:08, 2015

→ computerworld.co...
11 January at 16:10, 2013

→ anubisnetworks.c...
14 August at 16:44, 2015

→ lnec.pt/instituicoe...
24 June at 14:21, 2009

→ wikimmed.blogs.c...
2 February at 21:18, 2007

→ zappiens.pt/Z2207
8 November at 03:49, 2013

→ iscac.pt/index.ph...
30 September at 12:12, 2014

→ computerworld.co...
2 February at 16:39, 2012

→ portuguesclaro.pt...
8 November at 10:59, 2013

→ anubisnetworks.c...
14 August at 16:50, 2015

→ computerworld.co...
25 September at 15:38, 2012

→ arquivo.fccn.pt:80...
26 June at 19:17, 2007

→ computerworld.co...
2 February at 17:08, 2012

→ aladerei.e-xadrez....
3 August at 12:31, 2010

→ www-ext.lnec.pt:80/LNEC/...
30 June at 04:09, 2001

→ computerworld.com.pt/cat...
25 September at 16:09, 2012

→ datarecoverylab.pt/dt_log...

→ vox.fccn.pt
21 October at 14:24, 2008

→ portuguesclaro.pt/clientes

→ rcc.gov.pt/pt-PT/Noticias/...
8 June at 04:58, 2010

→ moodle.ccems.pt:80/index...
3 August at 07:23, 2007

Arquivo.pt APIs

- Arquivo.pt makes **8,000+ milion pages** e **22*+ million images** available for visualization and search:
  - Archived web pages -> **Text Search API**/Memento/CDX Server
  - Text and metadata search -> **Text Search API**
  - Image search -> **Image Search API**
- Available for the general public
- Open Source
- https://github.com/arquivo/pwa-technologies/wiki/APIs

Arquivo.pt Image Search (as of Jan 2020)

Indexed images	**22 million**
Collection count	90
(W)ARCs	3 million
(W)ARC sizes	334 TB
Total collected files	6,000 million
Total collected images	**1,602 million**
Oldest image date	15/04/1994
Newest image date	14/11/2019
Daily page views	~87

Opportunities for improvement

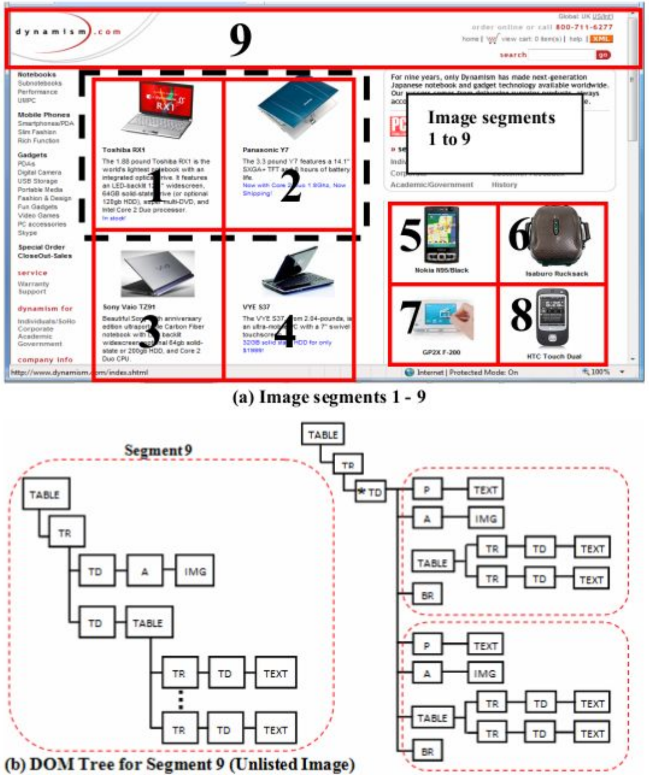
- Lack of image specific metadata
  - 43% (10,163,080 images) without imgAlt or imgTitle
- Why is the difference between collected and indexed so large?
- Only the oldest page per image is indexed
- Search result ranking does not take image popularity into account

Finding images in pages results

- <img> tag attributes
- <a> tag attributes
- Inline CSS background images
- Inline base64 images
- Images set by JS
- <figure>, <picture>

	Percentage of references
<img>	90.6%
<a>	8.7%
css	0.7%
Normal images	99.9%
base64	0.1%

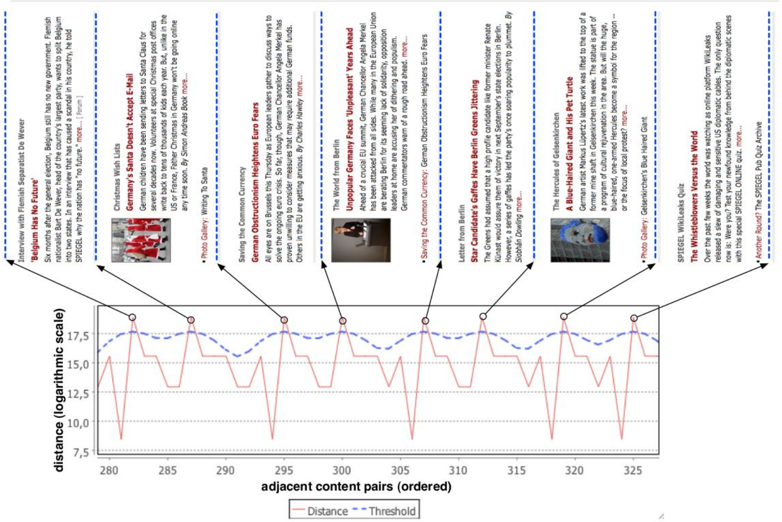
Finding an image caption



(a) Image segments 1 - 9

(b) DOM Tree for Segment 9 (Unlisted Image)

Fauzi, Fariza & Hong, Jer Lang & Belkhatir, Mohammed. (2009). Webpage segmentation for extracting images and their surrounding contextual information. 649-652. 10.1145/1631272.1631379.



Sadet, Alcic & Conrad, Stefan. (2011). A Clustering-based Approach to Web Image Context Extraction. MMEDIA - International Conferences on Advances in Multimedia.

Image caption extraction

I arrived at the following method

First parent with text

- Default method
- Works well for images in boxes or *reasonably* structured pages

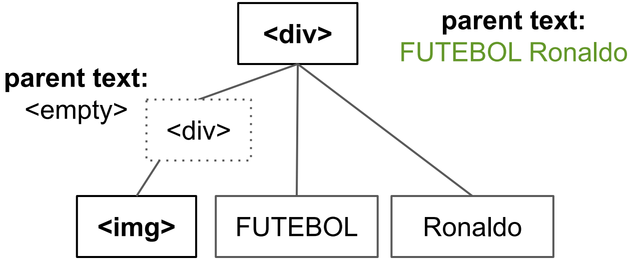
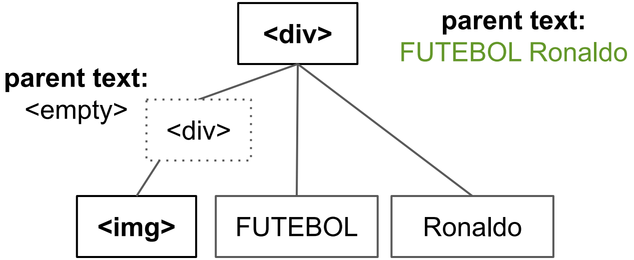


Image caption extraction

I arrived at the following method

First parent with text

- Default method
- Works well for images in boxes or *reasonably* structured pages



Previous and next node text

- Used if the first parent with text is at the level of the page with more siblings
- List of images as in a blog

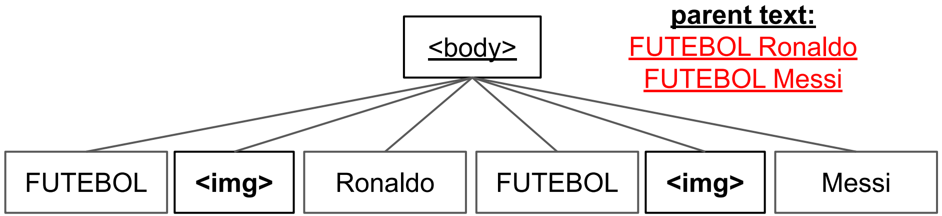
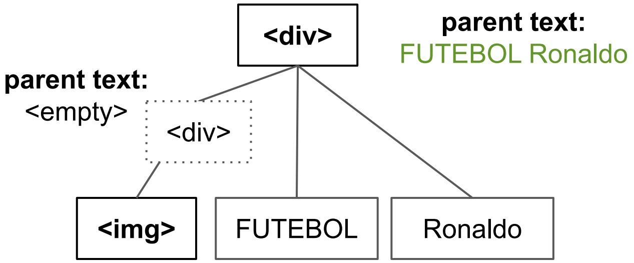


Image caption extraction

I arrived at the following method

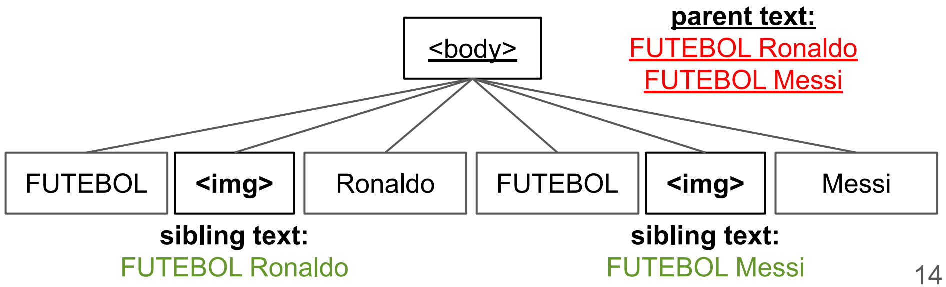
First parent with text

- Default method
- Works well for images in boxes or *reasonably* structured pages



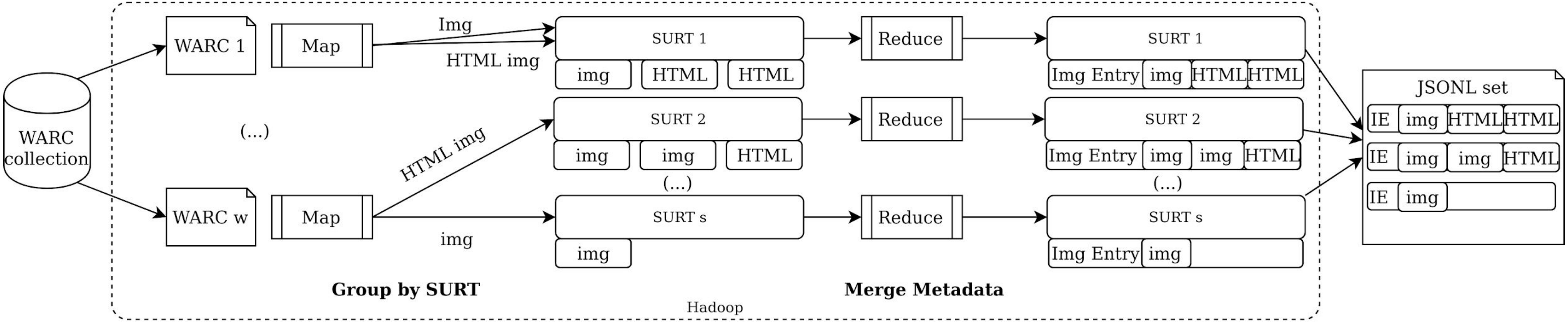
Previous and next node text

- Used if the first parent with text is at the level of the page with more siblings
- List of images as in a blog



Indexing Architecture

Map Reduce: Extract images and metadata



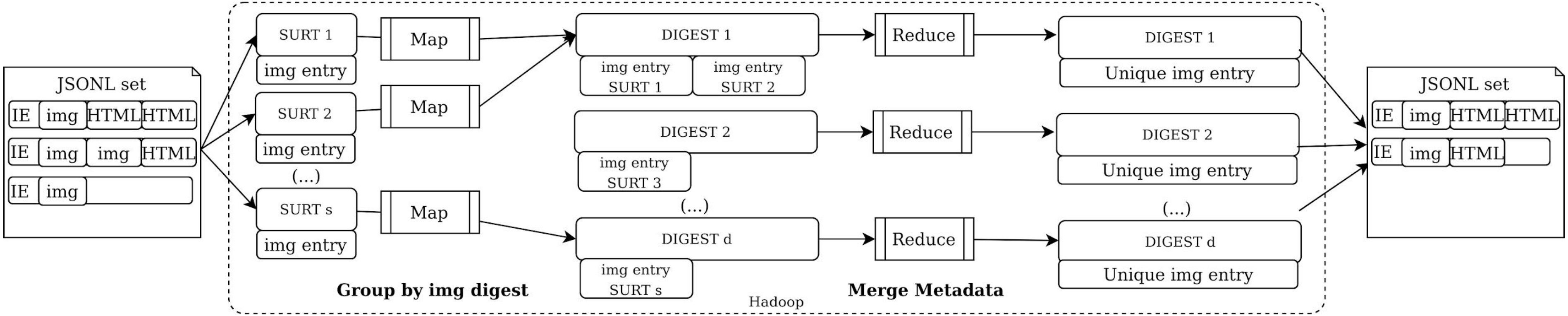
How to deal with duplicate information?

- The amount of data produced by this step is huge!
- Generating a lot of documents for indexing
- But most of this information is duplicate
    - Images and pages that were crawled at different times but have not changed
    - References to the images that have the same caption/metadata

Deduplication selected solution

- After careful examination, we arrived at the 3 deduplication scenarios:
  a. every page-image pair is a document
  b. the oldest page that references the image is the canonical document
  c. **oldest page information and image specific information from all pages**
     - keep reference to oldest page
     - Add all new image specific information (title, alt, caption) to the document
     - replace oldest page reference if a new oldest document shows up

Map Reduce: Group by digest



Duplicates across collections

- Hadoop processing is performed across per collection
    - To better manage computing resources (e.g. HDFS disk space)
    - Thus, deduplication is only performed on a per-collection basis
- We added an extra “group by digest” step when sending docs to Solr

My predictions in May 2020

Arquivo.pt Image Search (as of Jan 2020)

Indexed images	**22 million**
Collection count	90
(W)ARCs	3 million
(W)ARC sizes	334 TB
Total collected files	6,000 million
Total collected images	**1,602 million**
Oldest image date	15/04/1994
Newest image date	14/11/2019
Daily page views	~87

Tested collections - number of images

Collection	Old Parser	New Parser	Diff New to Current	Ratio vs New
AWP24	865,589	14,133,997	+13,268,408	16.33
AWP15	552,275	26,127,269	+25,574,994	47.31
FAWP26	213,527	1,562,617	+1,349,090	7.32
Tomba	169,308	1,076,967	+907,659	6.36
BlogsSapo2018	71,668	752,679	+681,011	10.50
Weblog	6,336	87,252	+80,916	13.77
DinisAlves2018	1,215	1,216	+1	1.00
DEM-IST	191	360	+169	1.88
BlocoEsquerda	15	16	+1	1.07

Takeways

~200-650 million images

1,880,124 -> 43,742,373

23,589,395 -> ??,???,???

~9-28x more images

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	225,868	3,414,742	6[illegible]569	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

~400-1,300 million pages (2/image)

~18-56x more pages

Takeways

~200-650 million images

1,880,124 -> 43,742,373

~9-28x more images

23,589,395 -> [illegible]

~400-1,300 million pages (2/image)

~18-56x more pages

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	[illegible]	[illegible]	[illegible]	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

Takeways

654 million images

1,880,124 -> 43,742,373

29x more images

23,589,395 -> ??.???.???

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	[illegible]	[illegible]	[illegible]0,669	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

1,252 million pages (1.91/image)

55x more pages

Takeways

+ 317 million images in one year (2019)

48% growth

+ 610 million pages in one year (2019)

49% growth

1,880,124 -> 43,742,373

23,589,395 -> ?????????

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	225,368	3,414,742	661,310,699	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

Takeways

971 million images

1,880,124 -> 43,742,373

42x more images

23,589,395 -> ??,???,???

1,862 million pages (1.91/image)

81x more pages

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	225,968	[illegible]	[illegible]	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

Impact of deduplication

	Number of documents
a	1,862 million image-page pair documents
b	584 million unique documents (971 million before deduplication across collections)
c	**584 million** documents, containing information from all 1,862 million image-page pairs

How will we index these **584 million** documents?

Current Solr indexing architecture

Current image index has **31 million** documents

(22 million plus some special crawls we added in 2020)

on one 20 core, 40 thread server with 512 GB RAM

(one server per branch, two redundant branches)

running Solr 6.3 with a 735 GB index

What to do with new data?

Our indexing process resulted in

584 million documents

(expected index size: ~720GB)

Where will we fit all this data?

Arquivo.pt response time guidelines

The 355 rule

- **3 responses per second**
- With an average query time **below 5 seconds**
- For **5 concurrent users**

- We are currently performing these experiments

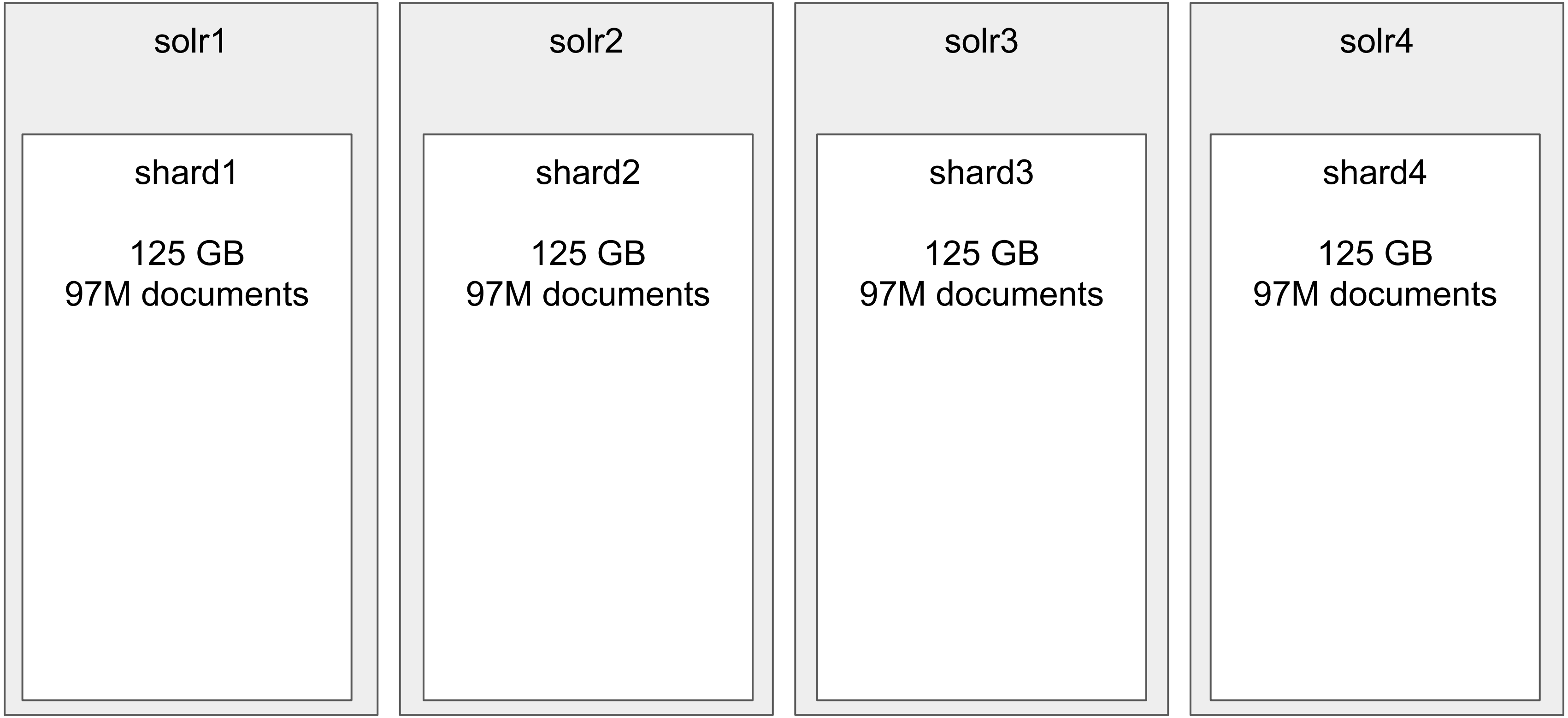
Planning SolrCloud resource allocation

- Expected index size: **~720GB**

- SolrCloud servers:
    - 8 servers, 4 per branch
        - **512GB**: p87, p91 (20/40 cores/threads)
        - **256GB**: p82, p83 (12/24 c/t), p93, p94, p98, p99 (20/40 c/t)
    - **2560GB** total, **1280GB** per branch

- No SSD, only HDD, but we have more RAM than indexed data

How we configured SolrCloud? - Try 1



solr1

shard1

125 GB
97M documents

solr2

shard2

125 GB
97M documents

solr3

shard3

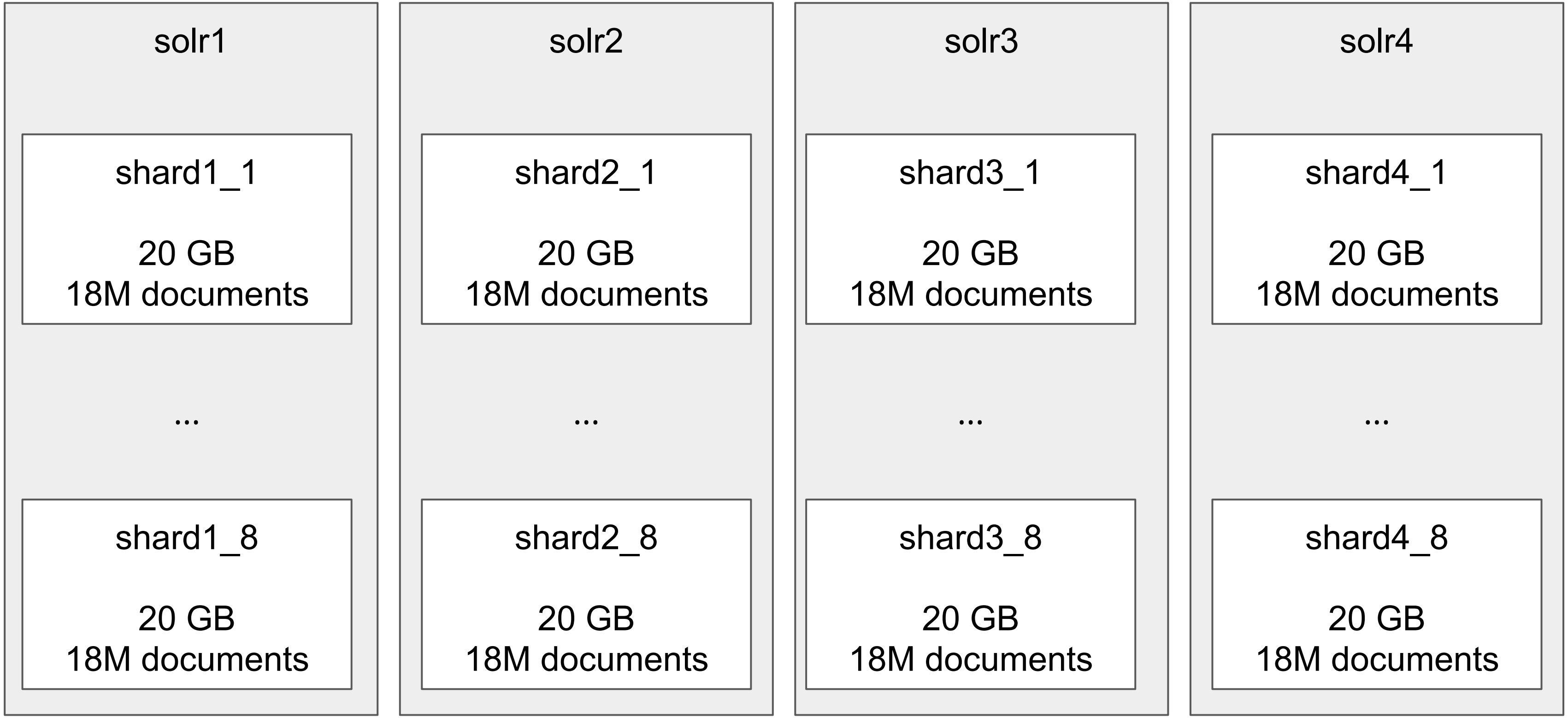
125 GB
97M documents

solr4

shard4

125 GB
97M documents

How we configured SolrCloud? - Try 2



solr1

shard1_1

20 GB
18M documents

...

shard1_8

20 GB
18M documents

solr2

shard2_1

20 GB
18M documents

...

shard2_8

20 GB
18M documents

solr3

shard3_1

20 GB
18M documents

...

shard3_8

20 GB
18M documents

solr4

shard4_1

20 GB
18M documents

...

shard4_8

20 GB
18M documents

How to test?

- Search with increasing concurrent users
  - 1, 3, 5, 10, 20, 50 concurrent users

- For a set period of time
  - 5 minutes

How to select realistic queries?

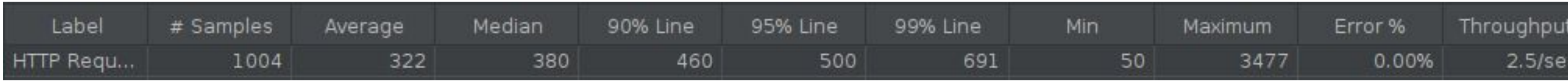
- Two sets of queries:
  - User queries extracted from logs
  - Random pairs of Portuguese words

- Warmup the index using 50 queries

- Query for 5 minutes and parse the results

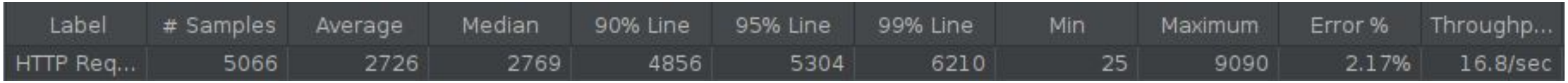
(Fresh off the press) results

Single user, random queries (pairs of portuguese words)



Label	# Samples	Average	Median	90% Line	95% Line	99% Line	Min	Maximum	Error %	Throughpu
HTTP Requ...	1004	322	380	460	500	691	50	3477	0.00%	2.5/se

50 users, random queries (pairs of portuguese words)

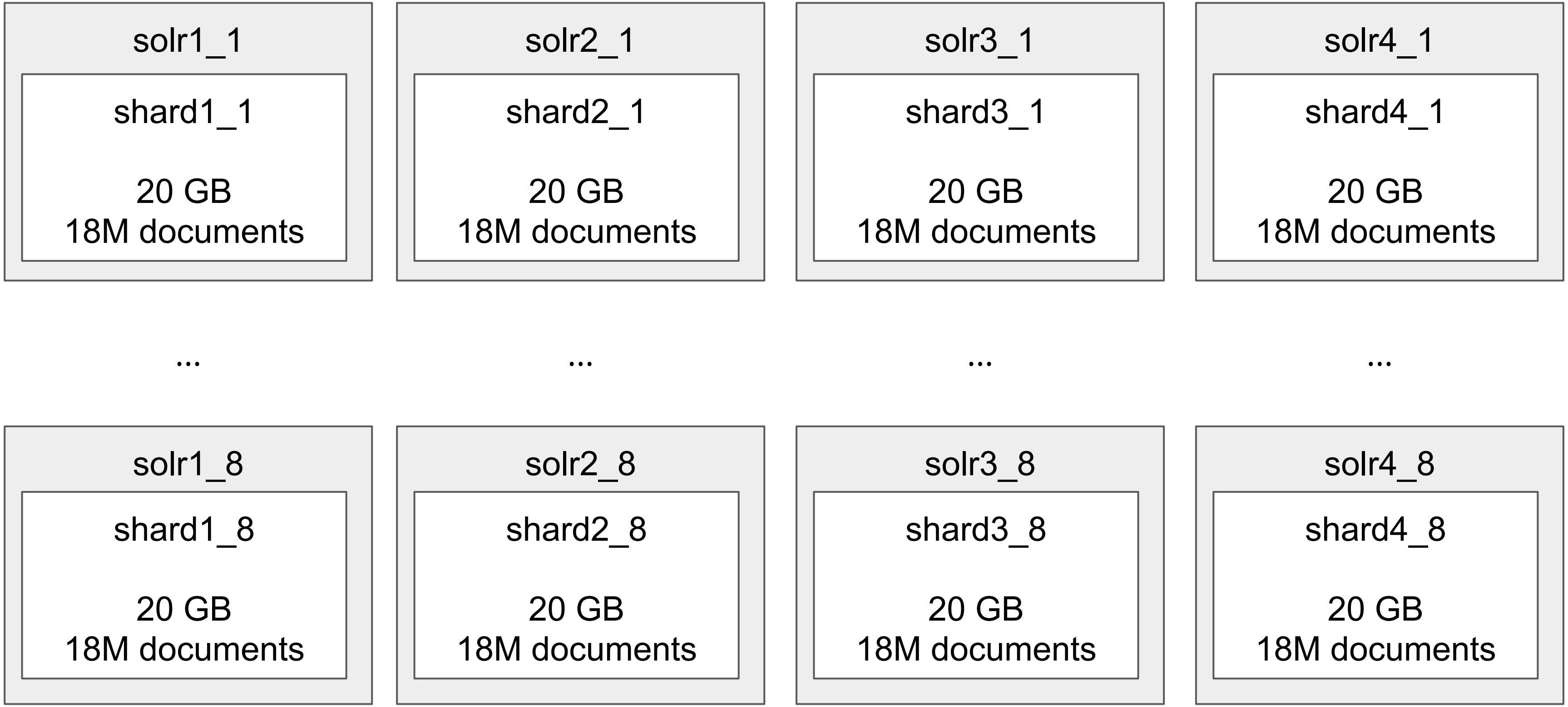


Label	# Samples	Average	Median	90% Line	95% Line	99% Line	Min	Maximum	Error %	Throughp...
HTTP Req...	5066	2726	2769	4856	5304	6210	25	9090	2.17%	16.8/sec

Tips and parameters

- vmtouch tool to force OS to keep index files in RAM
- Heap size: 31GB
  - Smaller sizes made Solr crash on parallel query situations
  - Larger sizes means Java can’t use compressed pointers https://lucene.apache.org/solr/guide/8_7/taking-solr-to-production.html#running-multiple-solr-nodes-per-host

How we configured SolrCloud? - Plan



solr1_1

shard1_1

20 GB
18M documents

solr2_1

shard2_1

20 GB
18M documents

solr3_1

shard3_1

20 GB
18M documents

solr4_1

shard4_1

20 GB
18M documents

... ...

solr1_8

shard1_8

20 GB
18M documents

solr2_8

shard2_8

20 GB
18M documents

solr3_8

shard3_8

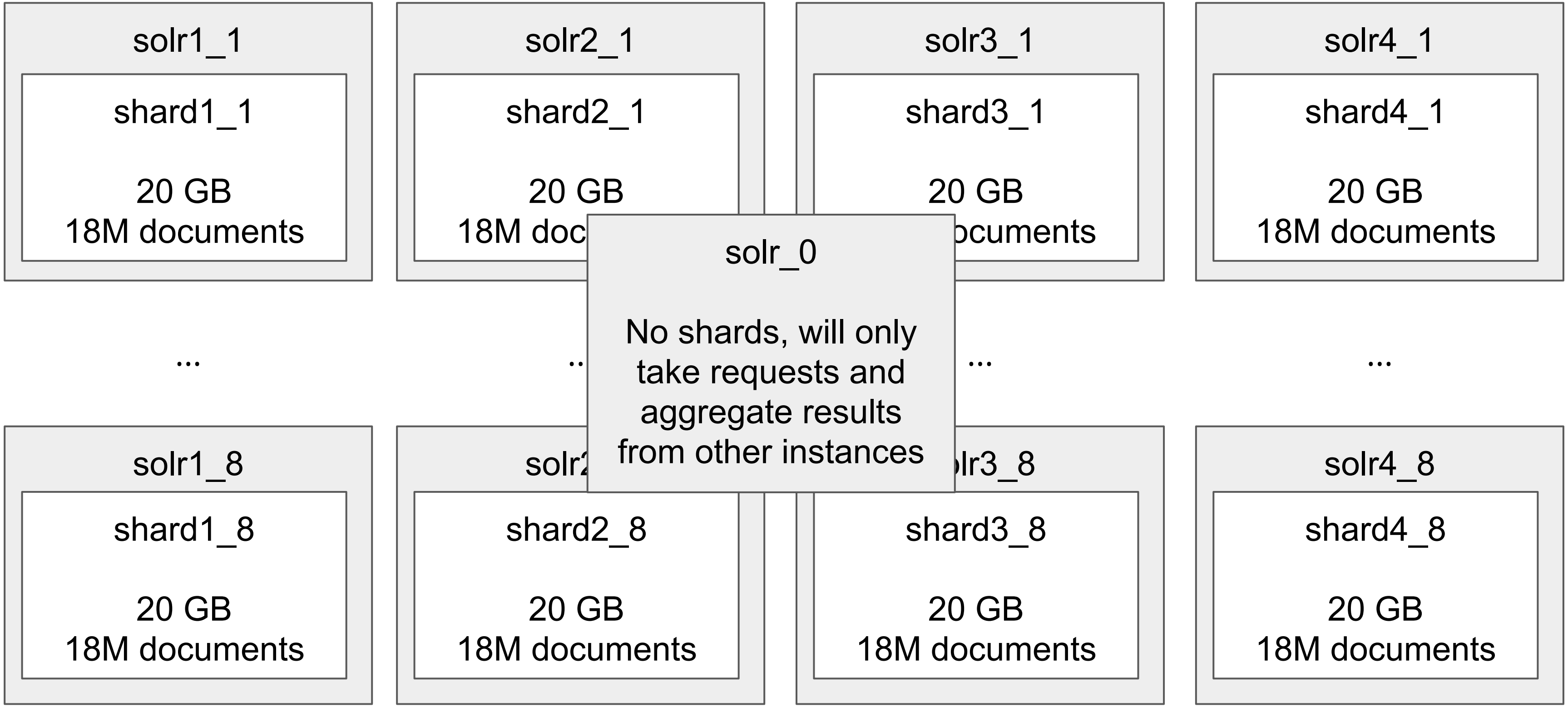
20 GB
18M documents

solr4_8

shard4_8

20 GB
18M documents

How we configured SolrCloud? - Plan



solr1_1

shard1_1

20 GB
18M documents

solr2_1

shard2_1

20 GB
18M doc

solr3_1

shard3_1

20 GB
ocuments

solr4_1

shard4_1

20 GB
18M documents

solr_0

No shards, will only take requests and aggregate results from other instances

...

..

...

...

solr1_8

shard1_8

20 GB
18M documents

solr2

shard2_8

20 GB
18M documents

lr3_8

shard3_8

20 GB
18M documents

solr4_8

shard4_8

20 GB
18M documents

Future problems: Migrate page search to SolrCloud

- Currently, we have an highly customized version of Lucene optimized not to search the full posting lists

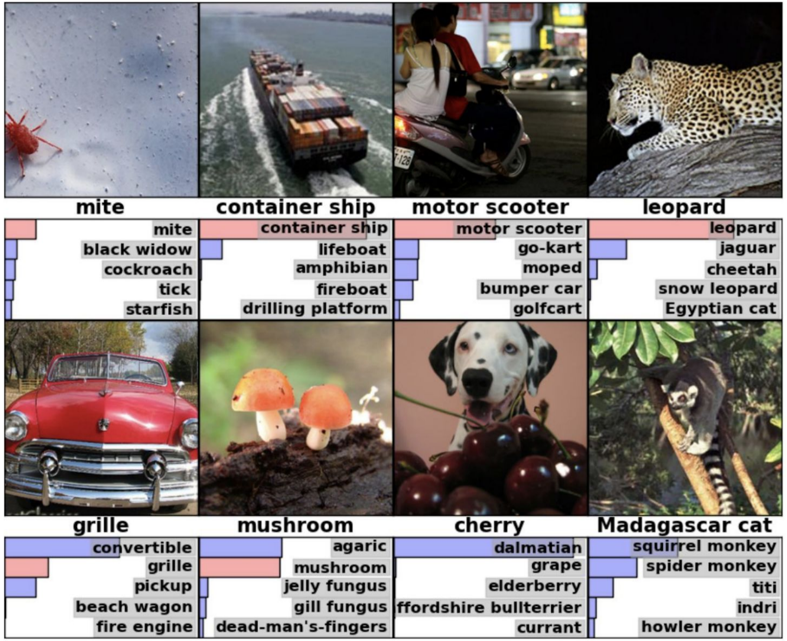
- Scale
  - 6-7,000 million documents
  - 5 servers with 4.5TB of RAM in total

Summary of what changed in 2020?

- More metadata per image
  - All pages that mention the image are parsed
  - Heuristic extraction of image captions from the HTML page structure
  - Additional features extracted from the HTML and images
- Improved NSFW image processing
  - 7x faster processing (40 -> 280 images per second)
  - Returns more image information for ranking (e.g. drawing vs. photo)
- Improved indexing architecture and processing
  - Removed MongoDB dependency
  - Ensure all archived images and pages are parsed
  - Find images in <a> links, CSS and JS code
- Distributed search index
  - Transition from single node Solr to distributed SolrCloud architecture
  - Improved schema so that the index only grows by 32% when covering 81x more images

Plan for the future

- Deal with images that have **no metadata**
  - Cannot find pages for 300+ million images
  - Deep Image classification, **tag extraction**
- Content based hashes
  - Similar images show up all over the place (different resolutions and formats)
  - Find and deduplicate **near duplicates**
- Improve Solr **ranking**
  - Use the newly extracted popularity features



Ranking features for 2021

imgCaption

- portion of the HTML page text that is closest to the image

matchingImages

- number of times the image was crawled (by image content digest)

matchingPages

- number of times the image was referenced on <*img*> tags, css or JS

imagesInOriginalPage

- number of images in the oldest page

imageMetadataChanges

- number of times that the image metadata (alt, title or caption) changes

pageMetadataChanges

- number of times that the page metadata (title) changes

drawing/photo

- whether the image is a drawing or a photo