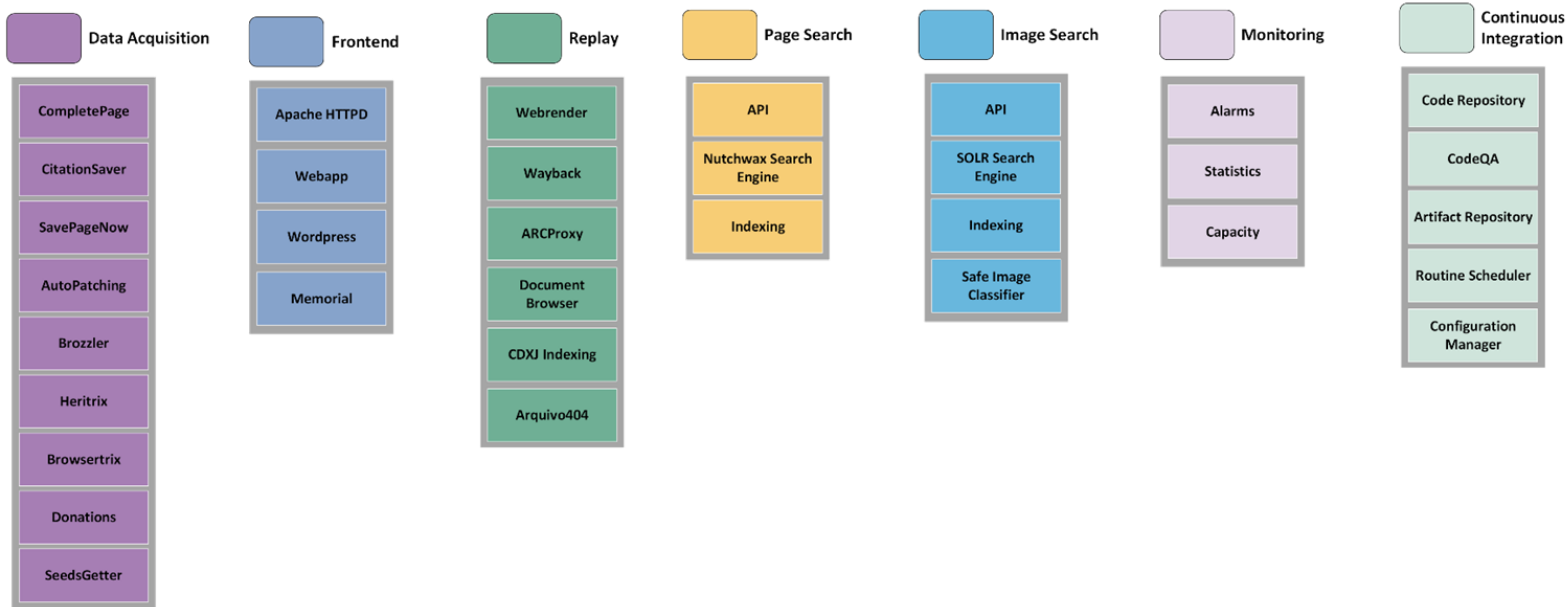


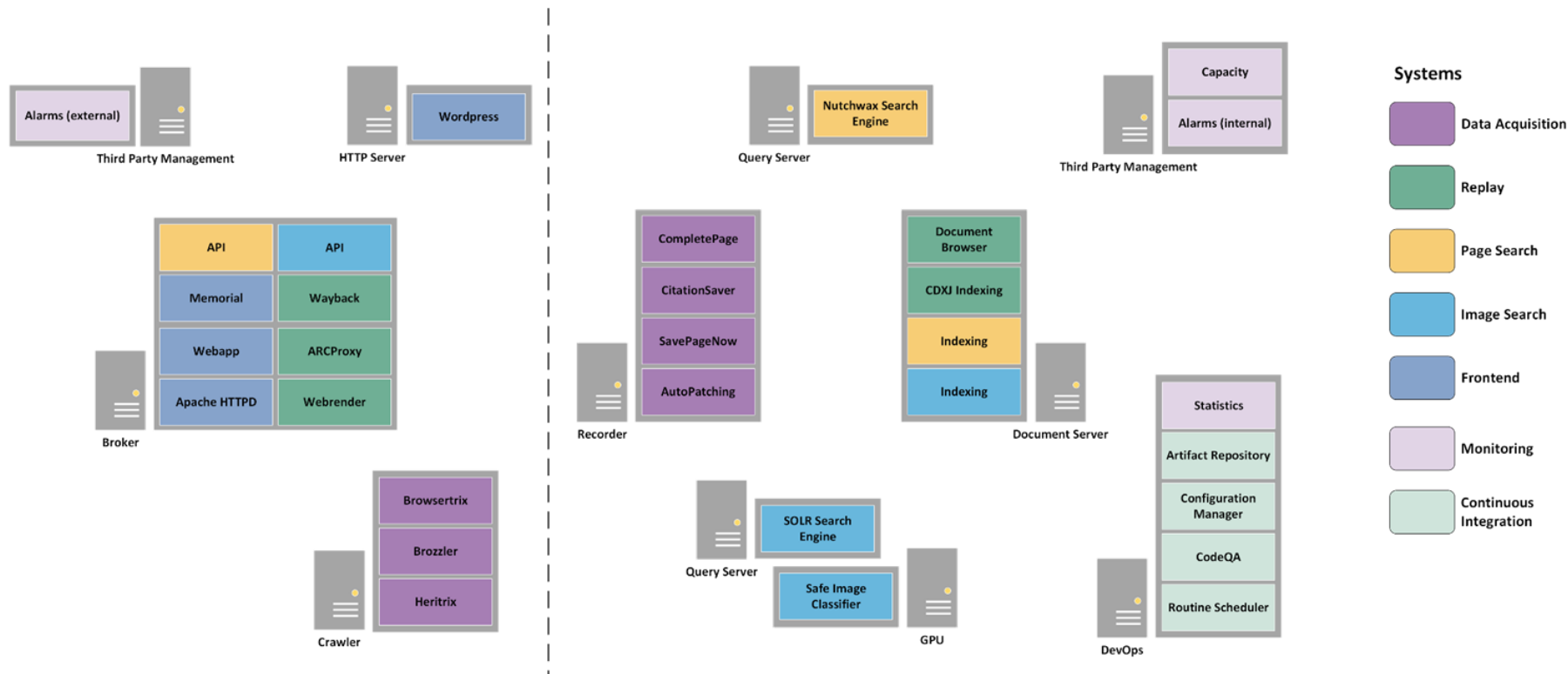
A arquitetura de um sistema de big data, o Arquivo.pt

francisco.esteveira@fccn.pt

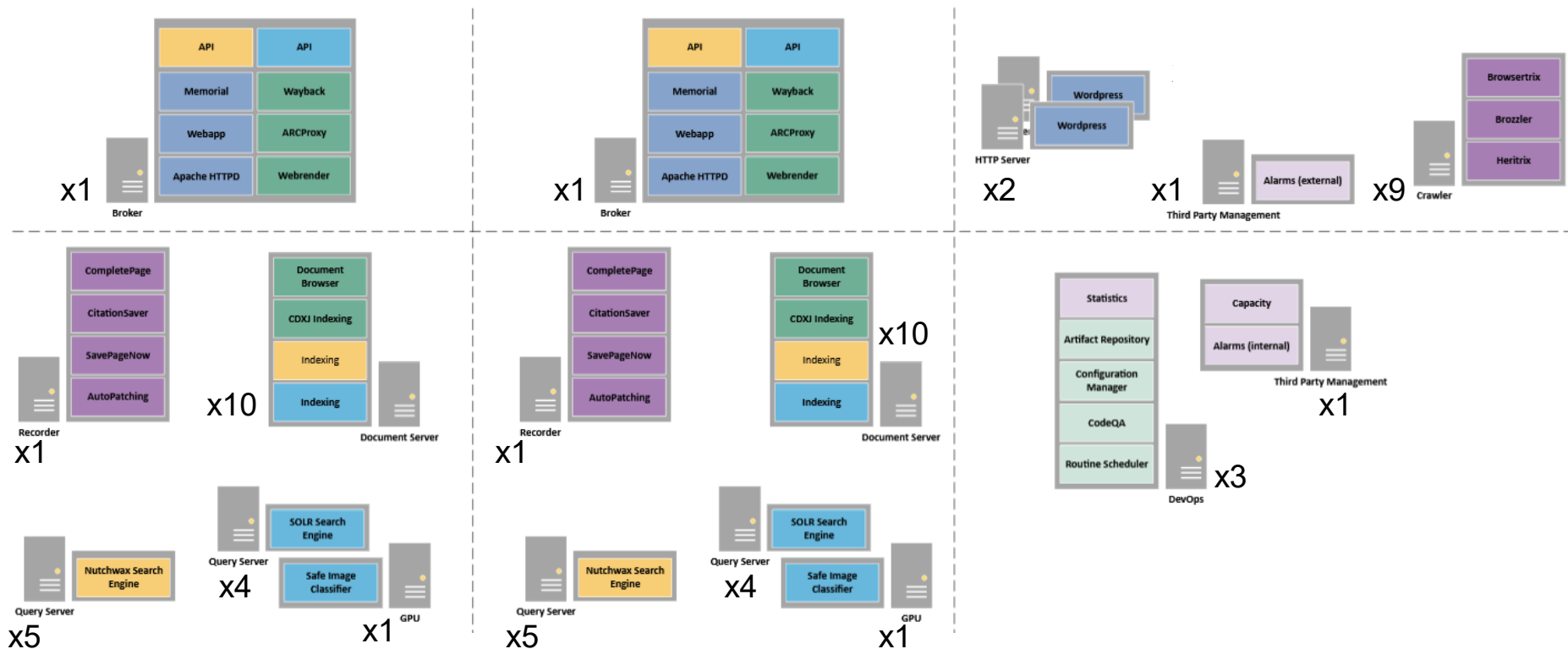
Arquitectura de Software



Arquitectura de Hardware



Ramais



[A Marinha](#)[O que fazemos](#)[Meios](#)[Serviços](#)[Junta-te a nós](#)[Institucional](#)[Media Center](#)[Contactos](#)

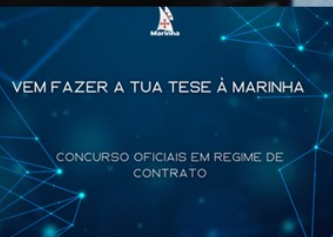
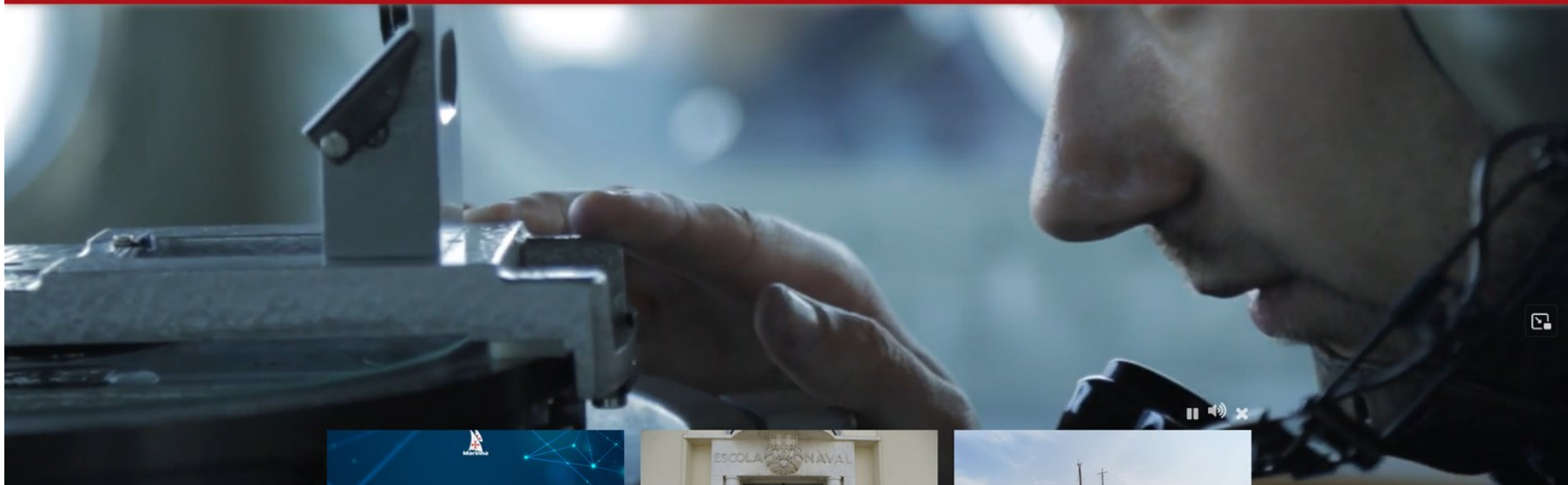
EMERGÊNCIA
NO MAR
214 401 919



COOKIES NO PORTAL DA MARINHA

Atenção este site utiliza cookies.
Ao navegar no site estará a sua utilização.

► SABER MAIS



És aluno das áreas das Engenharias,
Matemática ou Química? Esta
oportunidade é para ti!



Prepara o teu futuro: O concurso para a
Escola Naval já tem o edital publicado

77 8481 7070 10-06



NRP Sagres inicia viagem com rumo à
América do Sul

10 1116 7070 11-00

Aquisição de Dados



Data Acquisition

CompletePage

CitationSaver

SavePageNow

AutoPatching

Brozzler

Heritrix

Browsertrix

Donations

SeedsGetter

Collection ID prefix

AWP

MAWP

FAWP

EAWP

SAWP

RAQ

PATCHING

Varied prefixes (Roteiro, Dinis, Weblog, etc.)

Description. AWP means "arquivo da web portuguesa" (archive of the Portuguese web)

Trimestral: a list of all the domains registered under the .PT domain is collected every 3 months using a conventional crawler due to the high volume of data involved.

Monthly crawls: a list of selected national and international online publications is collected every month using a combination of browser-based crawlers.

Daily crawls: a list of selected online national publications is collected every day using a combination of browser-based and conventional crawlers.

Special crawls: selected pages about a given theme or event crawled with varied frequency.

Save Page Now crawls: pages web-archived by users in high-quality using the savepagenow service.

High-quality crawls: a list of selected websites that were carefully archived and curated with the highest-quality possible using the best combination of technologies available.

Web-archived resources crawled by CompletePage service after user interaction.

Donated collections

Aquisição de Dados

Data Acquisition

CompletePage

CitationSaver

SavePageNow

AutoPatching

Brozzler

Heritrix

Browsertrix

Donations

SeedsGetter

WARC/1.0

WARC-Type: response

WARC-Record-ID: <urn:uuid:b8138dc0-5370-4437-9aba-d0b2d30f8f71>

WARC-Date: 2023-03-21T17:00:20Z

WARC-Target-URI: <http://www.akademaii.com/content/16h828g2u476t7q0/fulltext.pdf>

WARC-IP-Address: 213.253.193.151

Content-Type: application/http;msgtype=response

WARC-Payload-Digest: sha1:EICAVLXYGCBISIVUHIRZSAHNRJCTWAFN

Content-Length: 441

WARC-Block-Digest: sha1:WDAPTQ2BXX2N3NHMOSWLGA7L34DQ33LR

HTTP/1.1 302 Found

Date: Tue, 21 Mar 2023 17:00:20 GMT

Location: <https://akjournals.com/content/16h828g2u476t7q0/fulltext.pdf>

Content-Length: 244

Content-Type: text/html; charset=iso-8859-1

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 2.0//EN">
```

```
<html><head>
```

```
<title>302 Found</title>
```

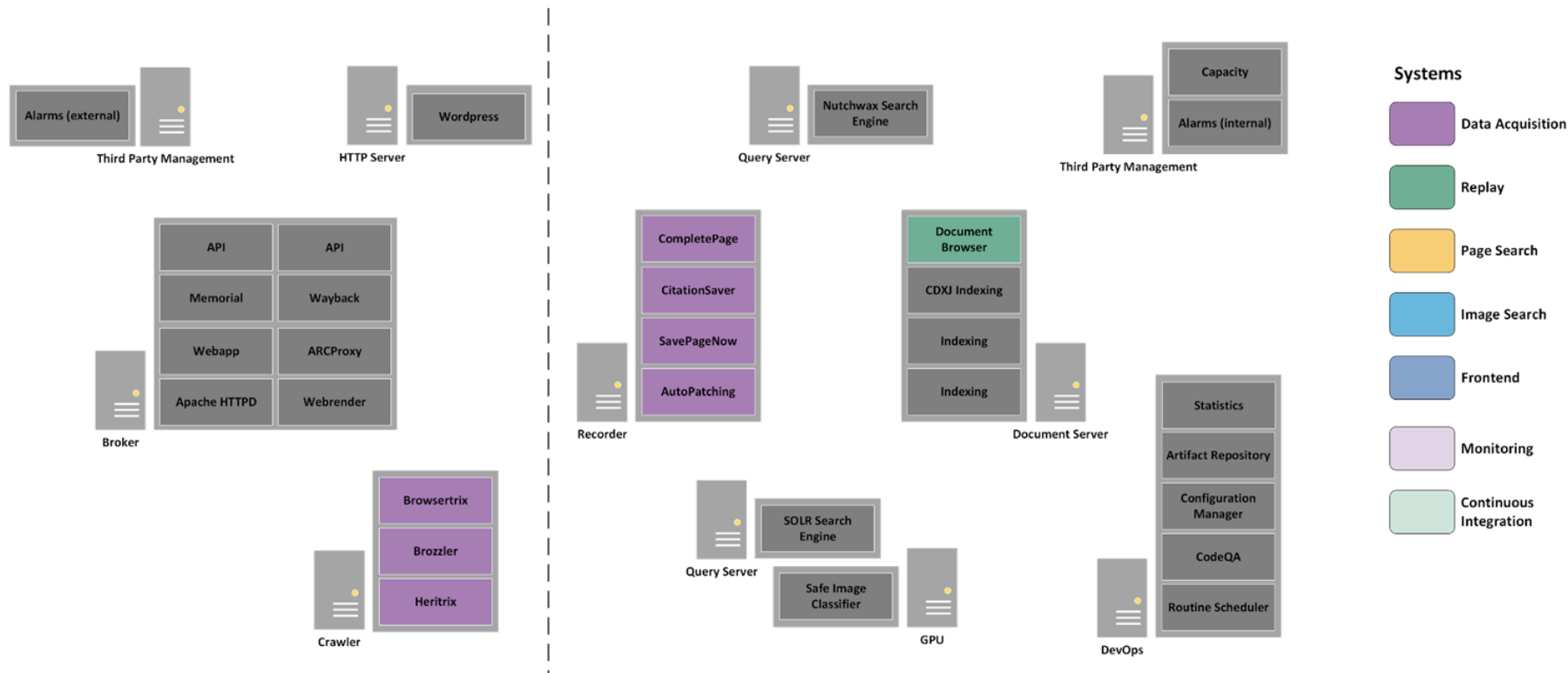
```
</head><body>
```

```
<h1>Found</h1>
```

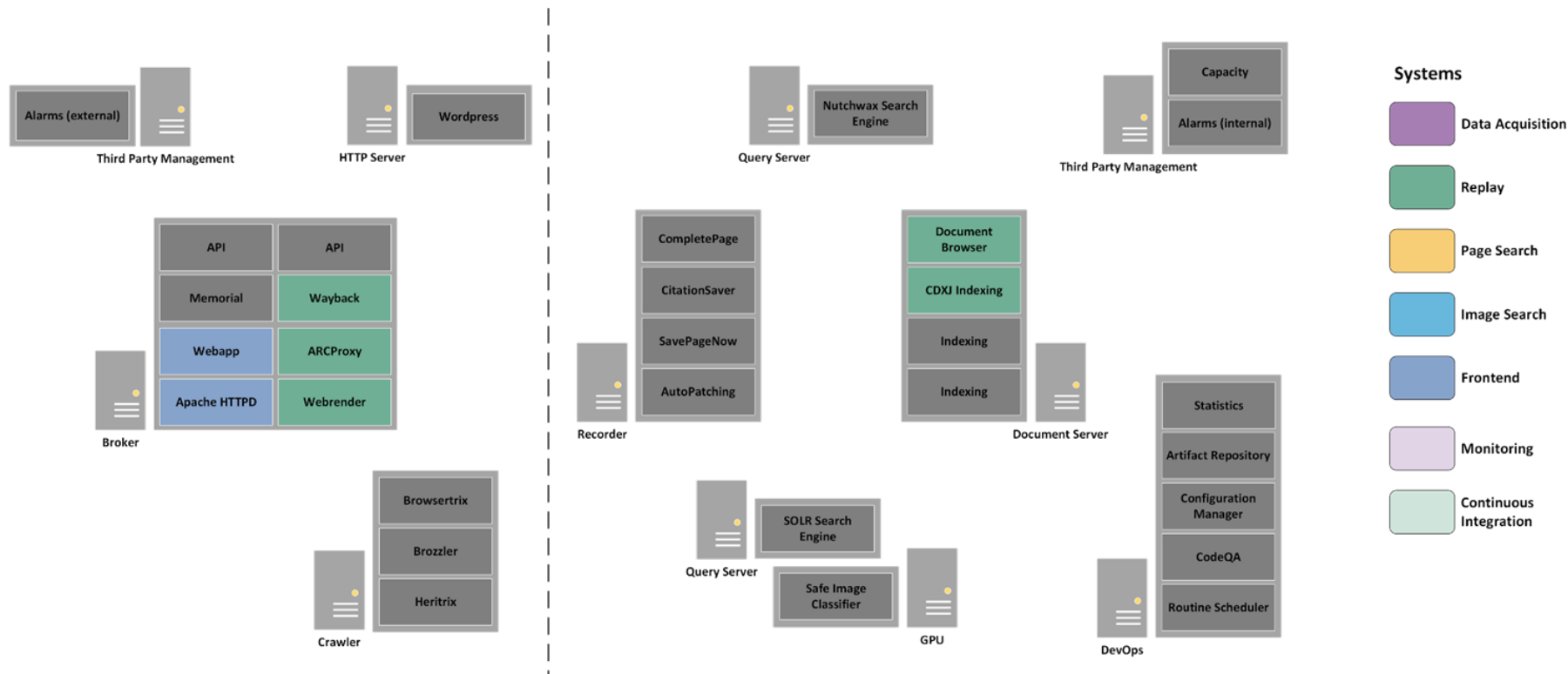
```
<p>The document has moved <a href="https://akjournals.com/content/16h828g2u476t7q0/fulltext.pdf">here</a>.</p>
```

```
</body></html>
```

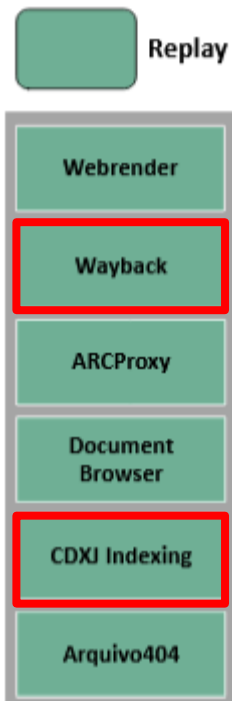
Arquitectura de Hardware



Arquitectura de Hardware



Reprodução



pt,marinha)/pt 20230321170203

```
{  
  "url": "https://www.marinha.pt/pt",  
  "mime": "text/html",  
  "status": "200",  
  "digest": "QFUHMA4PBGH3UC4TDTM4M6FIGCJV63BW",  
  "length": "49640",  
  "offset": "42908982",  
  "filename": "rec125-raq-fcul-memorial-  
20230321170015964-00025-pwz8bf2x.warc.gz",  
  "collection": "RAQ2023"  
}
```

Reprodução



Replay

Webrender

Wayback

ARCProxy

Document
Browser

CDXJ Indexing

Arquivo404

Q marinha.pt

Pesquisar

1991 6 Ago 2023 26 Jun

Páginas Imagens Narrativa

Pesquisa avançada

Cerca de 504 resultados desde 1991 até 2023

Tabela Lista	
1999	5 versões
2000	12 versões
2001	8 versões
2002	6 versões
2003	11 versões
2004	25 versões
2005	37 versões
2006	16 versões
2007	24 versões
2008	3 versões
2009	6 versões

Reprodução

Replay

Webrender

Wayback

ARCProxy

Document Browser

CDXJ Indexing

Arquivo404

Menu

ARQUIVO.PT

Options

marinha.pt/pt 21 March at 17h02, 2023

Marinha

A Marinha O que fazemos Meios Serviços Junta-te a nós Institucional Media Center Contactos



Table

2021

2022

2023

February

March

14 March 12h08, 2023

18 March 22h36, 2023

21 March 17h02, 2023

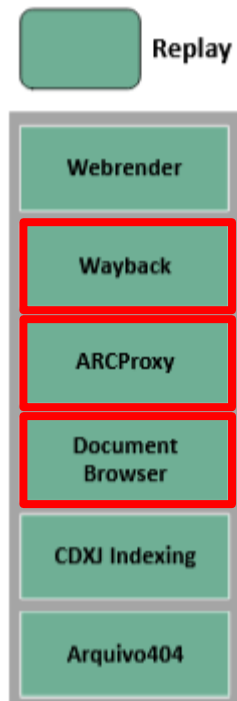
21 March 17h02, 2023

21 March 17h02, 2023

EFEMÉRIDE

NRP Centauro foi aumentado ao efetivo da Marinha

Reprodução



"length": "49640"

"offset": "42908982"

"filename": "rec125-raq-fcul-memorial-20230321170015964-00025-pwz8bf2x.warc.gz"

<http://document-server/RAQ2023/rec125-raq-fcul-memorial-20230321170015964-00025-pwz8bf2x.warc.gz>

Reprodução

Replay

Webrender

Wayback

ARCProxy

Document
Browser

CDXJ Indexing

Arquivo404

Menu

Tat

2021

2022

2023

February

March

14 March

18 March

21 March

21 March

21 March

```
<li class="menu-link">
  <a href="/pt/a-marinha/Paginas/default.aspx" class="pm-toplink-main">A Marinha<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
  <div class="subnavbar">
    <div class="pm-center">
      <ul class="nav navbar-subnav">
        <li class="menu-sublink"><a href="/pt/a-marinha/Paginas/cema-amn.aspx" target="_self" aria-label="CEMA e AMN" class="pm-toplink-sub">CEMA e AMN<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/discursos/Paginas/default.aspx" target="_self" aria-label="Discursos" class="pm-toplink-sub">Discursos<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/entrevistas/Paginas/default.aspx" target="_self" aria-label="Entrevistas" class="pm-toplink-sub">Entrevistas<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/Paginas/estrutura.aspx" target="_self" aria-label="Estrutura" class="pm-toplink-sub">Estrutura<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/Paginas/missao.aspx" target="_self" aria-label="A Missão" class="pm-toplink-sub">A Missão<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/Paginas/postos.aspx" target="_self" aria-label="Os Postos" class="pm-toplink-sub">Os Postos<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/historia/Paginas/default.aspx" target="_self" aria-label="História" class="pm-toplink-sub">História<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/simbolos-tradicoes/Paginas/default.aspx" target="_self" aria-label="Símbolos e Tradições" class="pm-toplink-sub">Símbolos e Tradições<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/pt/a-marinha/estudos-e-reflexoes/Paginas/default.aspx" target="_self" aria-label="Estudos e Reflexões" class="pm-toplink-sub">Estudos e Reflexões<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
      </ul>
    </div>
  </div>
</li>
```

```
<li class="menu-link">
  <a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/Paginas/default.aspx" class="pm-toplink-main">A Marinha<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
  <div class="subnavbar">
    <div class="pm-center">
      <ul class="nav navbar-subnav">
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/Paginas/cema-amn.aspx" target="_self" aria-label="CEMA e AMN" class="pm-toplink-sub">CEMA e AMN<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/discursos/Paginas/default.aspx" target="_self" aria-label="Discursos" class="pm-toplink-sub">Discursos<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/entrevistas/Paginas/default.aspx" target="_self" aria-label="Entrevistas" class="pm-toplink-sub">Entrevistas<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/Paginas/estrutura.aspx" target="_self" aria-label="Estrutura" class="pm-toplink-sub">Estrutura<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/Paginas/missao.aspx" target="_self" aria-label="A Missão" class="pm-toplink-sub">A Missão<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/Paginas/postos.aspx" target="_self" aria-label="Os Postos" class="pm-toplink-sub">Os Postos<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/historia/Paginas/default.aspx" target="_self" aria-label="História" class="pm-toplink-sub">História<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/simbolos-tradicoes/Paginas/default.aspx" target="_self" aria-label="Símbolos e Tradições" class="pm-toplink-sub">Símbolos e Tradições<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
        <li class="menu-sublink"><a href="/noFrame/replay/20230321170203/https://www.marinha.pt/pt/a-marinha/estudos-e-reflexoes/Paginas/default.aspx" target="_self" aria-label="Estudos e Reflexões" class="pm-toplink-sub">Estudos e Reflexões<span class="fa fa-chevron-right" aria-hidden="true"></span></a>
      </ul>
    </div>
  </div>
</li>
```

 Replay

Webrender

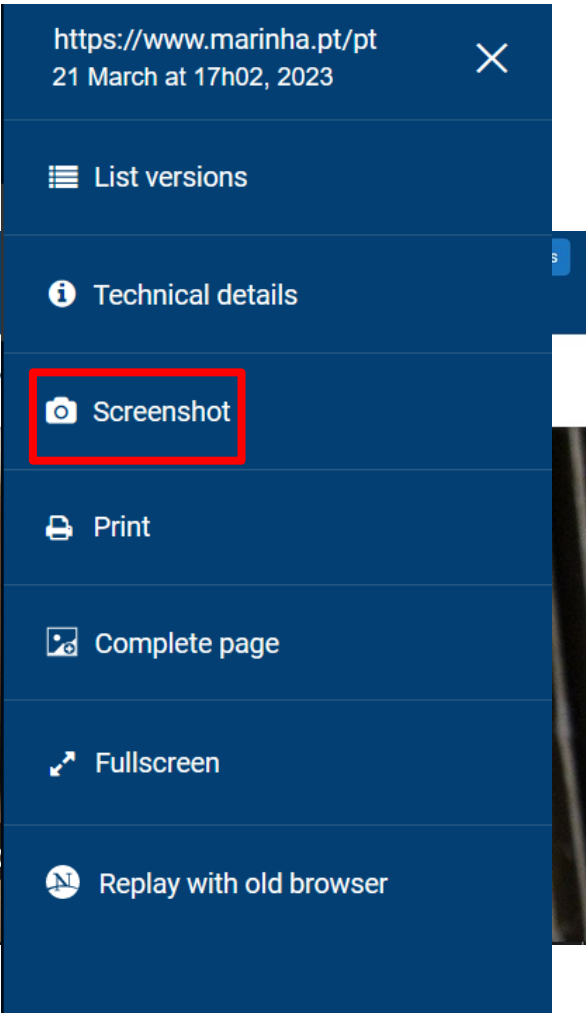
Wayback

ARCProxy

Document
Browser

CDXJ Indexing

Arquivo404



List versions

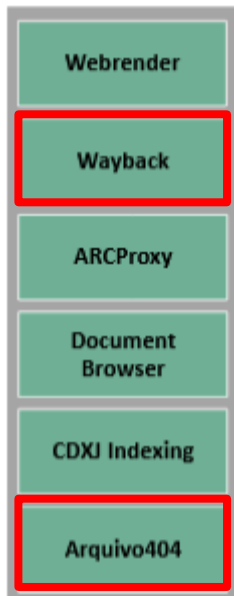
Technical details

 Screenshot Print Complete page

 Fullscreen

 [Replay with old browser](#)

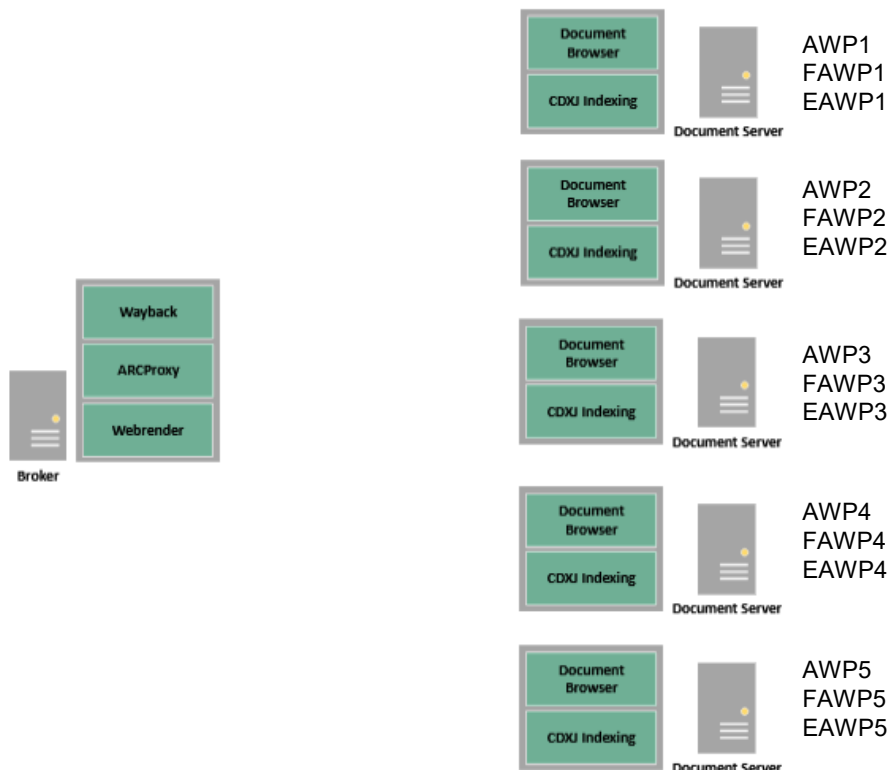
Reprodução



Página não encontrada.

[Visite uma versão anterior desta página no Arquivo.pt.](#)

Arquitetura de Hardware

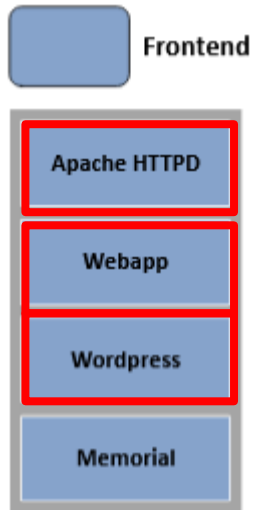


x10

Document Server:
256GB Memória RAM
24 vCPU
120TB espaço em disco RAID6

Broker:
256GB Memória RAM
80vCPU
20TB armazenamento SSD RAID6

Frontend

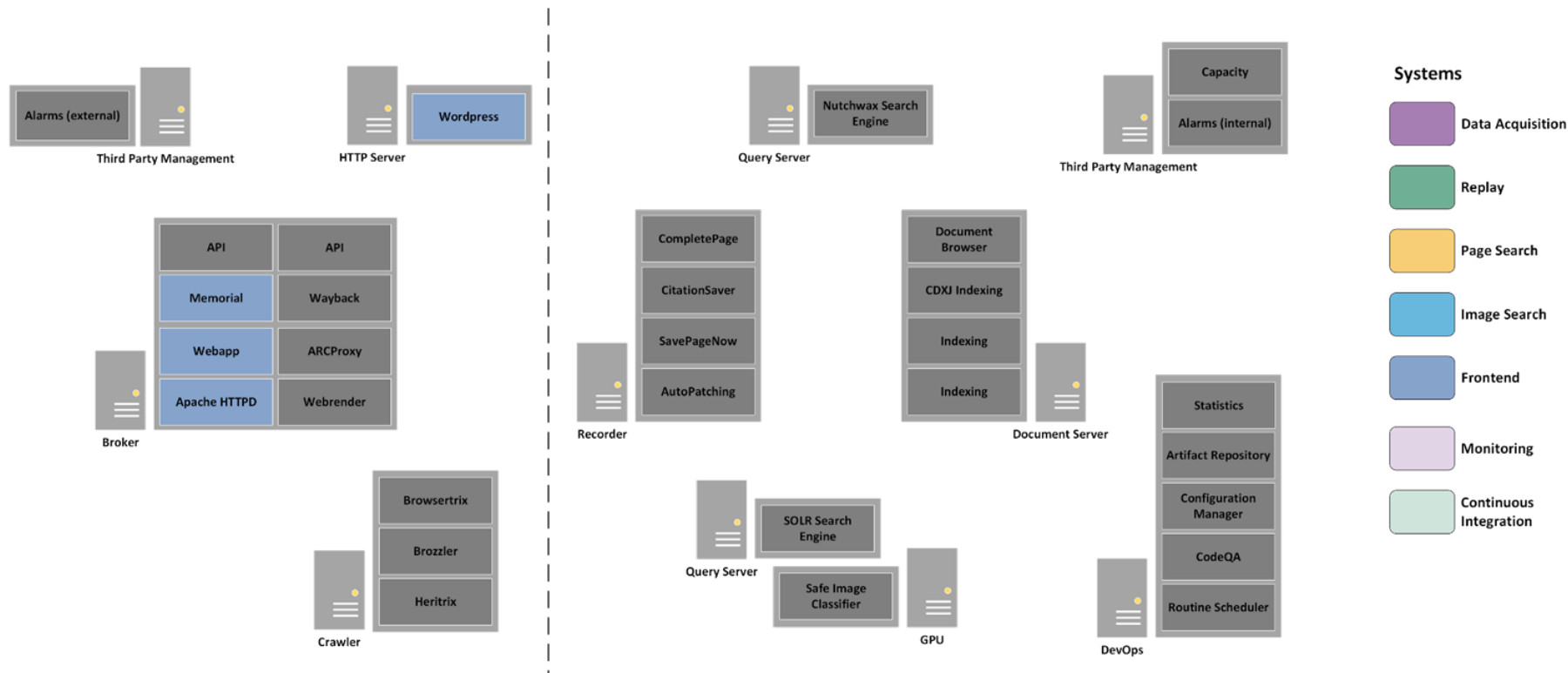


The screenshot displays the Arquivo.pt website. At the top, there is a navigation bar with the Arquivo.pt logo, a search bar, and a menu with links: Arquivo.pt, Notícias, Exemplos, Publicações, Colabore, Acerca, Perguntas frequentes, Mapa do sítio, and Contacto. Below the navigation bar, the main content area features a section titled 'Informações acerca do serviço Arquivo.pt' with a sub-header 'Última atualização em 21 de Abril de 2023 às 9:58'. The text states that Arquivo.pt preserves millions of files collected from the web since the 1990s and provides a public search service. A video player is embedded, showing a thumbnail for 'Arquivo.pt: pesquise páginas do passado (vídeo oficial 2021)' with 'Watch Later' and 'Share' buttons. Below the video player, there is a search interface with a search bar, a 'Pesquisar' button, and a timeline slider from 1991 to 2023. The search results section shows a video titled 'Prémio Arquivo.pt 2023' with the text 'VIAGEM NO TEMPO e ganhe 10.000€'. Below the video, there are two sections: 'O que é o Arquivo.pt?' and 'Testemunhos de utilizadores', each containing a video player. On the right side of the page, there is a 'Notícias' section with several news items, including 'Participação do Arquivo.pt em eventos internacionais sobre arquivos da Web', 'Arquivo.pt nas Jornadas FCCN 2023: inscreva-se já!', 'Museu Virtual do Turismo criou coleção de Websites preservados', 'Acesso em bloco a conteúdos arquivados', and 'CitationSaver preserva citações para conteúdos online'.

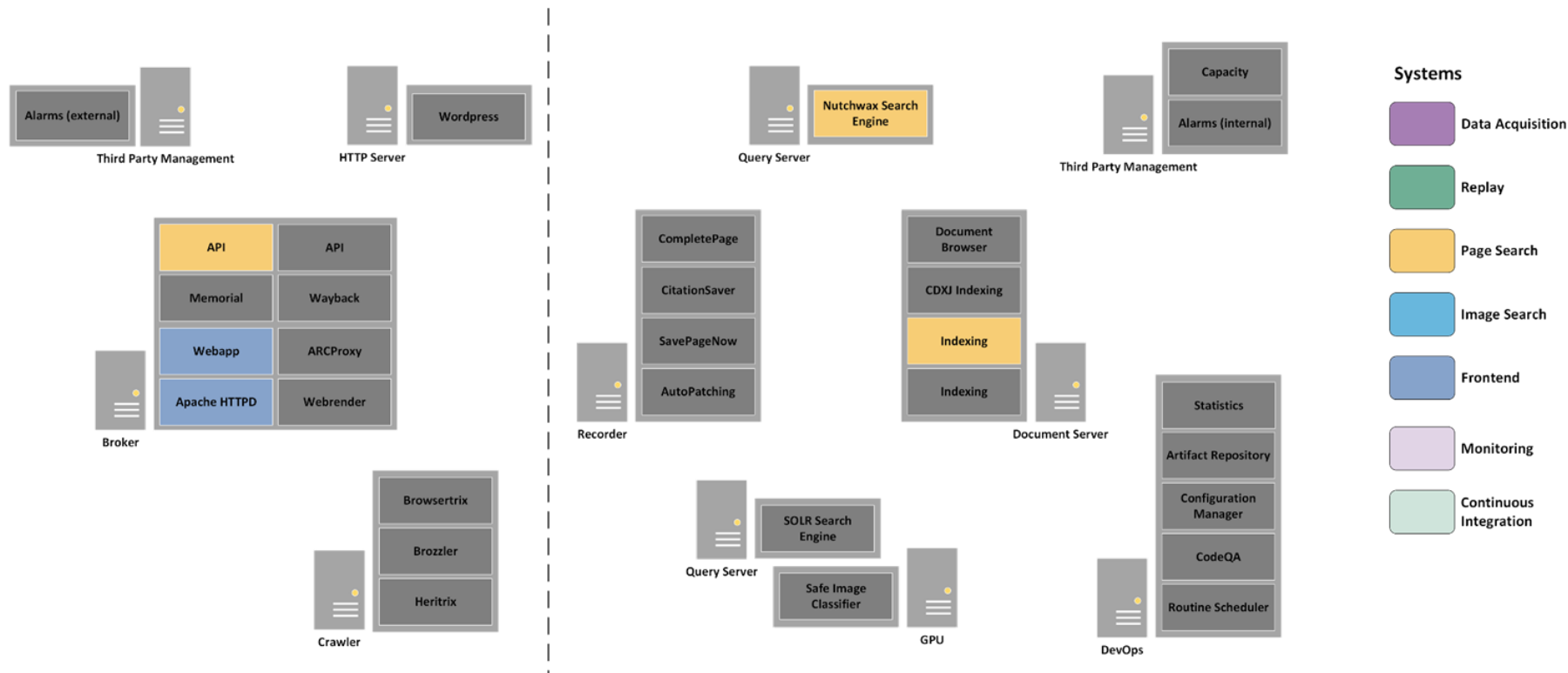
Frontend



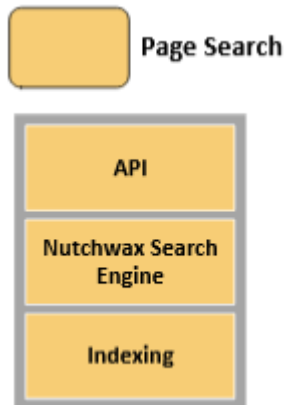
Arquitectura de Hardware



Arquitectura de Hardware



Pesquisa de Texto



```
serviceName: "Arquivo.pt - the Portuguese web-archive"
linkToService: "https://arquivo.pt"
next_page: "https://arquivo.pt/textsearch?q=marinha&offset=50"
estimated_nr_results: 54069352
▼ request_parameters:
  offset: 0
  dedupValue: 2
  dedupField: "site"
  q: "marinha"
  maxItems: 50
▼ response_items:
  ▼ 0:
    title: "Document Moved"
    originalURL: "http://www.marinha.pt/"
    ▼ linkToArchive:
      tstamp: "20091218103350"
      contentLength: 169
      digest: "bc9d5cb940fb07692161bda8a8304ab"
      mimeType: "text/html"
      encoding: "iso-8859-1"
      date: "1261132430"
    ▼ linkToScreenshot: "https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2FrepLay%2F20091218103350%2Fhttp%3A%2F%2Fwww.marinha.pt%2F"
    ▼ linkToNoFrame: "https://arquivo.pt/noFrame/replay/20091218103350/http://www.marinha.pt/"
    ▼ linkToExtractedText: "https://arquivo.pt/textextracted?m=http%3A%2F%2Fwww.marinha.pt%2F%2F20091218103350"
    ▼ linkToMetadata: "https://arquivo.pt/textsearch?metadata=http%3A%2F%2Fwww.marinha.pt%2F%2F20091218103350"
    ▼ linkToOriginalFile: "https://arquivo.pt/noFrame/replay/20091218103350/id_/http://www.marinha.pt/"
    ▼ snippet: 'Document Moved Object Moved This document may be found here<span class="ellipsis"> ... </span>'
      fileName: "IAH-20091218102919-08151-awp01.fccn.pt"
      collection: "AwP6"
      offset: 32745700
  ▼ 1:
    ▼ title: "DOMÍNIOS CO . PT - Serviço comercial de registo de domínios."
    originalURL: "http://marinha.co.pt/"
    ▼ linkToArchive: "https://arquivo.pt/wayback/20200407015331/http://marinha.co.pt/"
    tstamp: "20200407015331"
    contentLength: 47154
    digest: "5e6ad7e2690a925d5fb1af3f89ec0e03"
    mimeType: "text/html"
    encoding: "windows-1252"
    date: "1586224411"
    ▼ linkToScreenshot: "https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2FrepLay%2F20200407015331%2Fhttp%3A%2F%2Fmarinha.co.pt%2F"
    ▼ linkToNoFrame: "https://arquivo.pt/noFrame/replay/20200407015331/http://marinha.co.pt/"
    ▼ linkToExtractedText: "https://arquivo.pt/textextracted?m=http%3A%2F%2Fmarinha.co.pt%2F%2F20200407015331"
```

Pesquisa de Texto



Pesquisar

1991
6 Ago

2023
26 Jun

Páginas Imagens Narrativa Pesquisa avançada

Cerca de 54.069.352 resultados desde 1991 até 2023

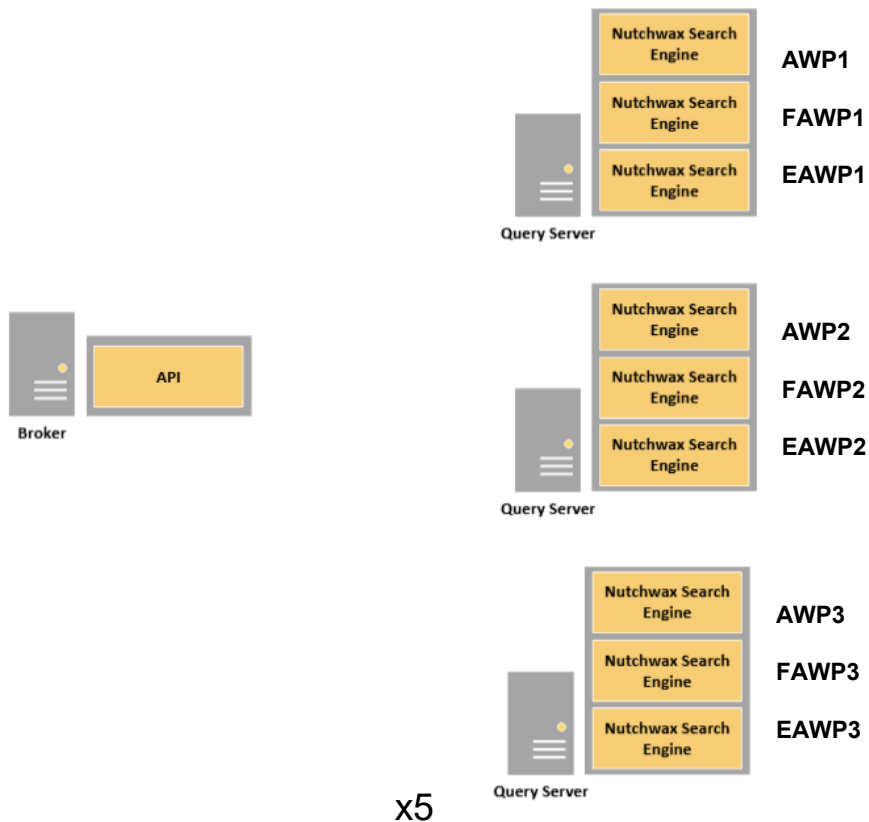
www.marinha.pt
Document Moved
18 Dezembro 2009
Document Moved Object Moved This document may be found here ...

marinha.co.pt
DOMÍNIOS CO . PT - Serviço comercial de registo de domínios.
7 Abril 2020
DOMÍNIOS CO . PT - Serviço comercial de registo de domínios. Sites Web Notícias Vídeos Emprego ...

www.marinha.pt
Document Moved
6 Junho 2010
Document Moved Object Moved This document may be found here ...

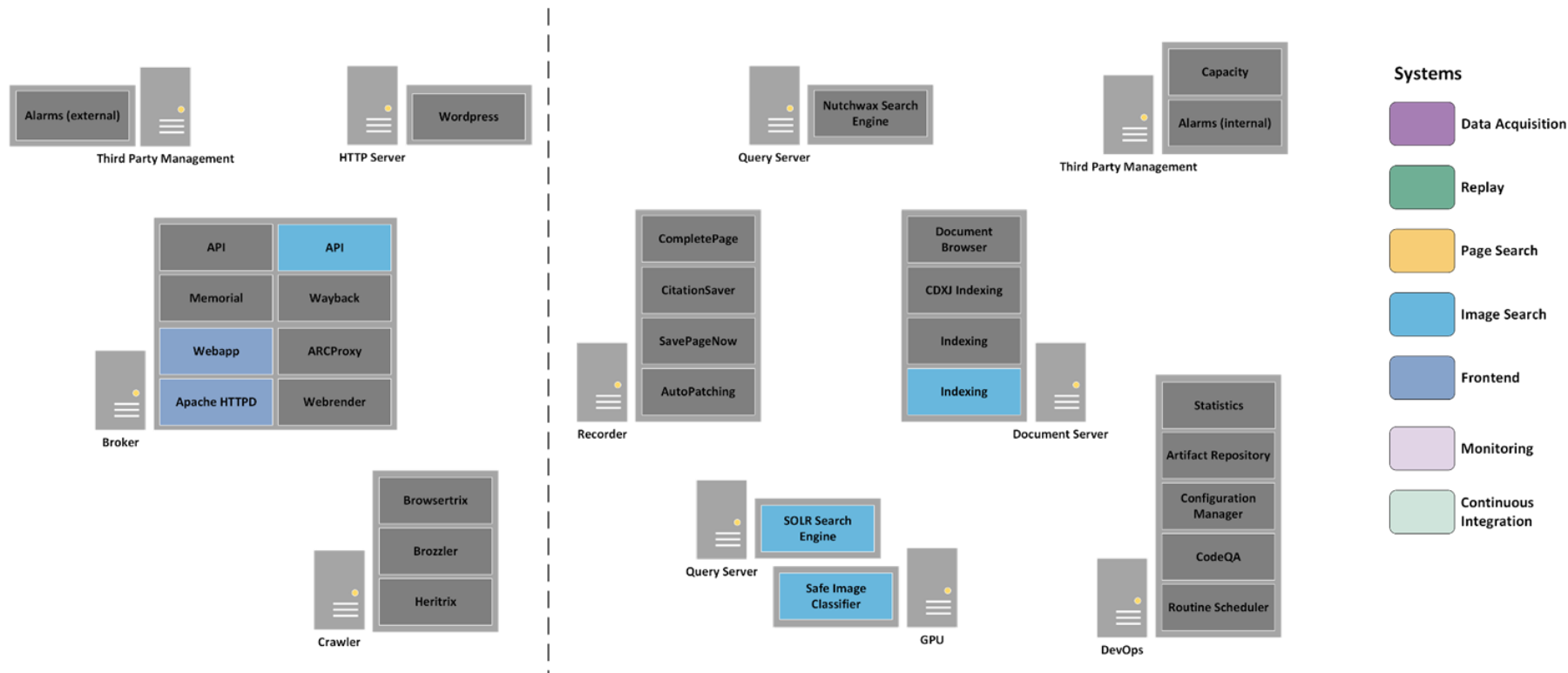
www.edificio-santa-marinha-sermax-5652.empresendimentos.com.pt
Empreendimento Edifício Santa **Marinha - SerMax**
13 Abril 2018
Empreendimento Edifício Santa **Marinha** - SerMax Início Empreendimento Localização Contactos Empreendimento EDIFÍCIO SANTA **MARINHA** - SERMAX Localização Porto - Vila Nova de Gaia [ver mais] Edifício NOVO c várias Tipologias JUNTO A COIMBRÕES Edifício constituído por habitações T2 e T3 com áreas ...

Arquitetura de Hardware



512GB memória RAM
32 vCPU
25TB de armazenamento

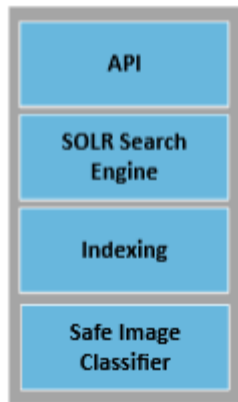
Arquitectura de Hardware



Pesquisa de Imagens

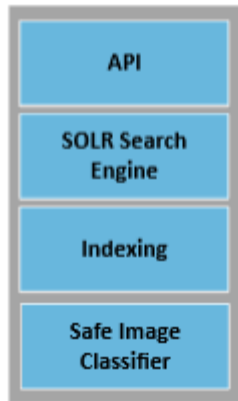


Image Search



```
serviceName: "Arquivo.pt - image search service."
linkToService: "https://arquivo.pt/images.jsp"
▼ linkToMoreFields: "https://arquivo.pt/imagesearch?q=marinha&more=pageHost,matchingImages,safe"
nextPage: "https://arquivo.pt/imagesearch?q=marinha&offset=50"
previousPage: "https://arquivo.pt/imagesearch?q=marinha&offset=0"
totalItems: 878792
numberOfResponseItems: 50
offset: 0
▼ responseItems:
  ▼ 0:
    ▼ imgDigest: "ff48ef46b21a56aead21c464e17e5f448ffa7ed871a882bdf2db1e74c9f634b8"
    ▼ imgSrc: "https://www.medi.pt/fileadmin/local/processed/en/8/c/csm_medi-colours-treatment-of-oedemas-navy-m-96568_9e6cfb44d0.jpg"
      pageTitle: "Cores de peças mediven para tratamento de edemas"
      imgHeight: 599
      imgWidth: 599
      imgMimeType: "image/jpeg"
    ▼ pageURL: "https://www.medi.pt/produtos/vantagens-exclusivas/cores/extras-colours-treatment-oedema/"
      imgTstamp: "20171214042008"
      pageTstamp: "20171214041447"
    ▼ imgCaption:
      0: "Marinha"
    ▼ imgAlt:
      0: "Marinha"
      1: "Marinha - Marinha"
    ▼ imgTitle:
      0: "Marinha "
```

Pesquisa de Imagens



serviceName: "Arquivo.pt - image search service."
linkToService: "https://arquivo.pt/images.jsp"
linkToMoreFields: "https://arquivo.pt/imagesearch?q=marinha&more=pageHost,matchingImages,safe"

ne:

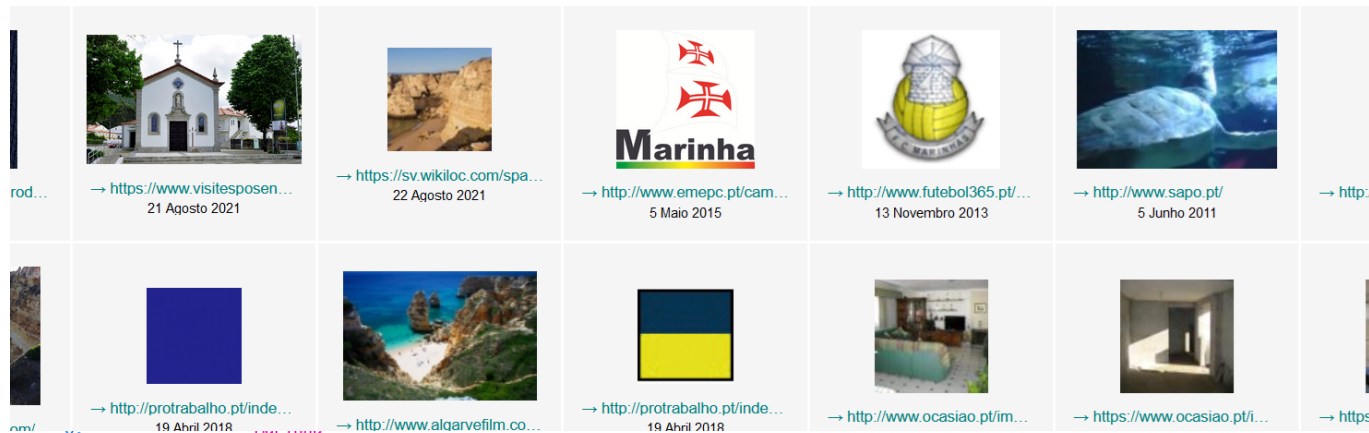
pr:

to:

of:

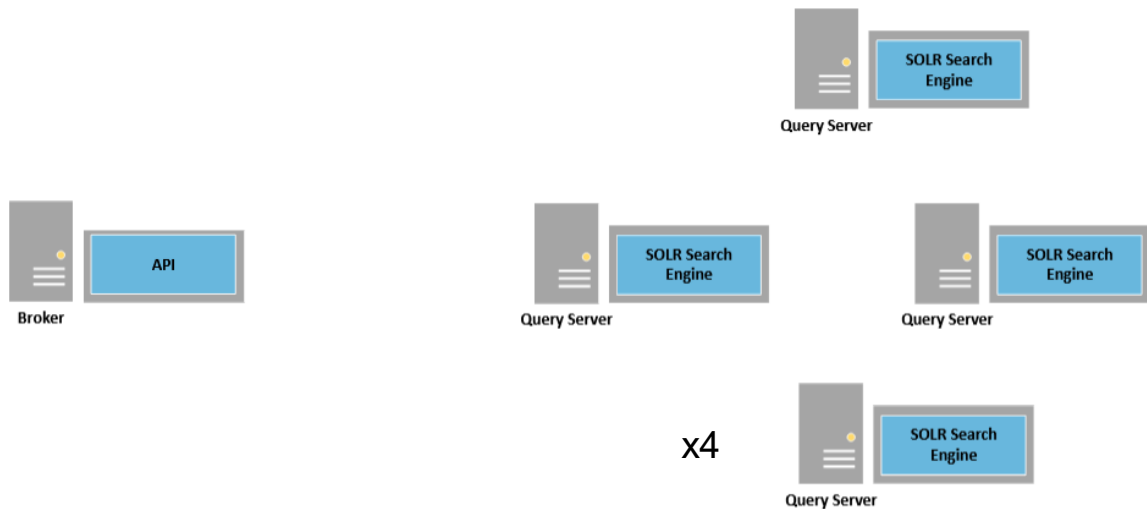
re:

Cerca de 878.792 resultados desde 1991 até 2023



ipa"

Arquitetura de Hardware



256GB memória RAM
40 vCPU
40TB de armazenamento RAID6

Monitorização

Monitoring

Alarms

Statistics

Capacity

Summary					
Reported period	Month Jun 2023				
First visit	01 Jun 2023 - 00:00				
Last visit	27 Jun 2023 - 00:59				
	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Viewed traffic *	122,479	216,470 (1.76 visits/visitor)	8,409,736 (38.84 Pages/Visit)	12,970,987 (59.92 Hits/Visit)	53.45 TB (265138.56 KB/Visit)
Not viewed traffic *			23,125,366	29,009,985	497.51 GB

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

Monthly history					
Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2023	130,740	302,643	5,910,059	25,291,562	3.02 TB
Feb 2023	116,734	179,549	4,793,444	13,472,686	4.05 TB
Mar 2023	98,246	150,686	9,722,793	24,786,334	9.20 TB
Apr 2023	123,773	217,300	6,630,714	12,026,785	17.87 TB
May 2023	122,479	216,470	8,409,736	12,970,987	53.45 TB
Jun 2023	122,479	216,470	8,409,736	12,970,987	53.45 TB
Jul 2023					
Aug 2023					
Sep 2023					
Oct 2023					
Nov 2023					
Dec 2023					

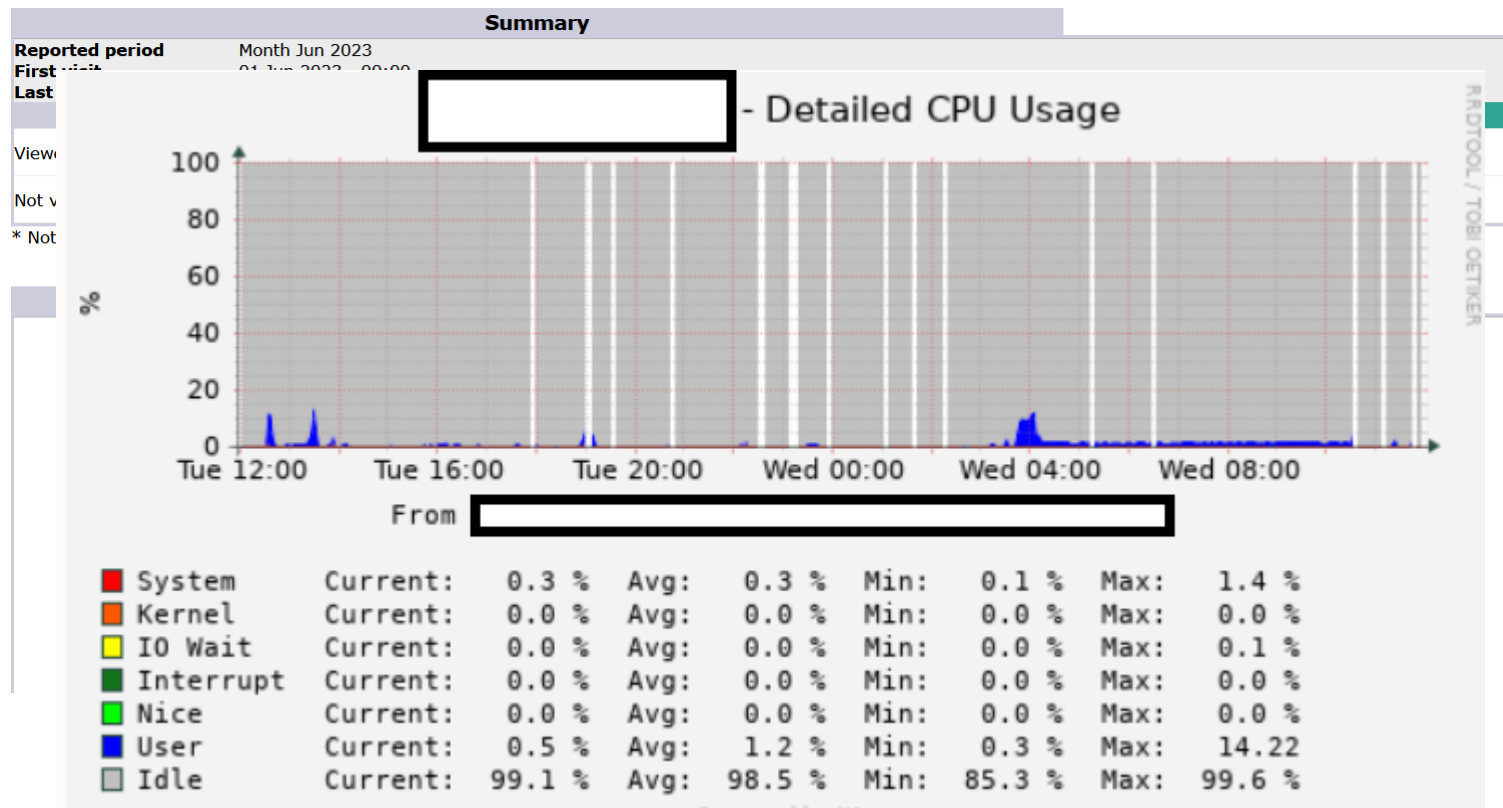
Monitorização

Monitoring

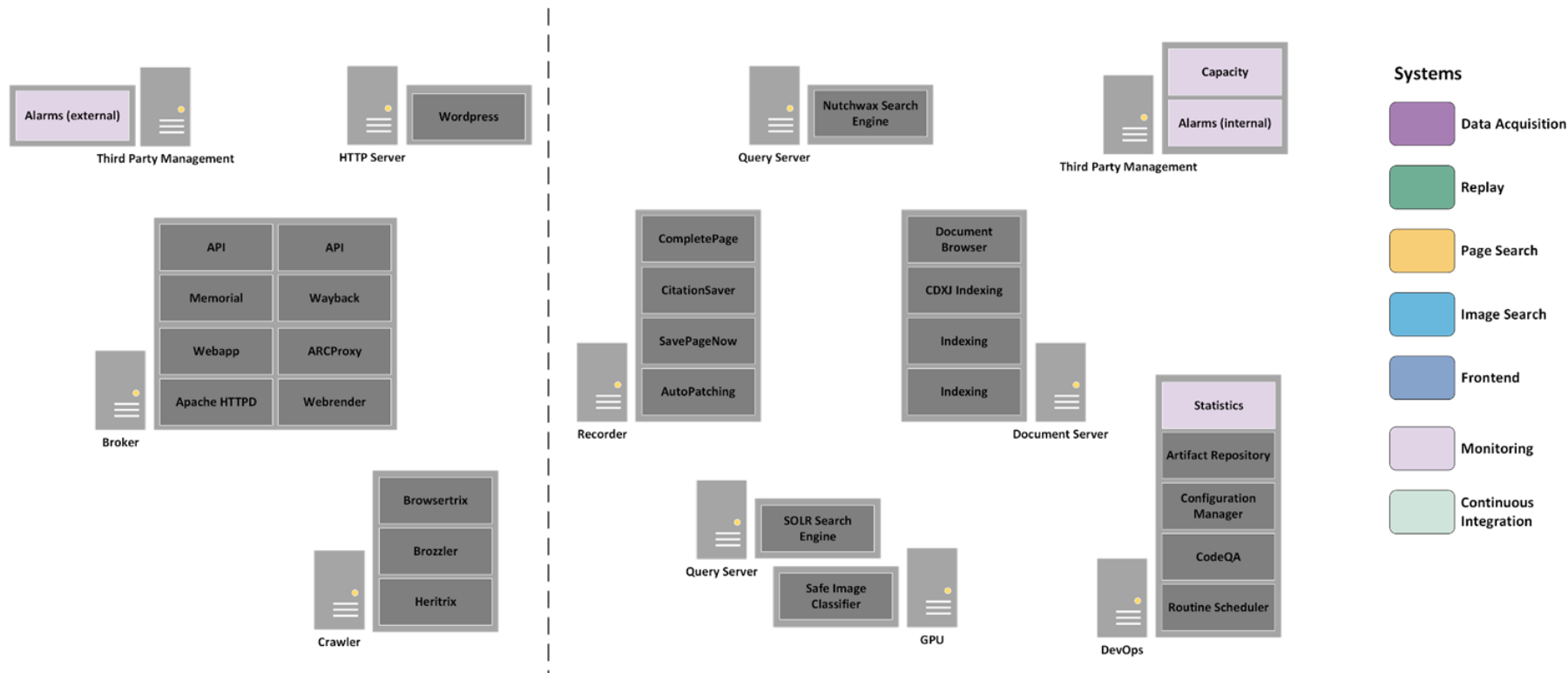
Alarms

Statistics

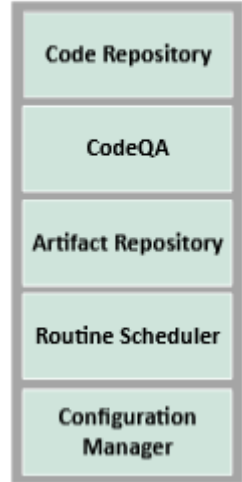
Capacity



Arquitectura de Hardware



CI/CD



Testes Funcionais

Deploys de Software

Configuração de Servidores

Rotinas de Operação (cópia de ficheiros, indexações)