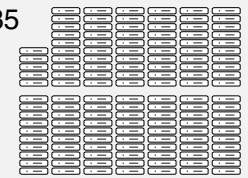
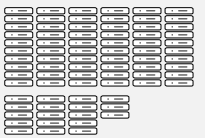
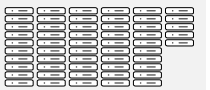
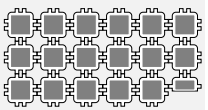
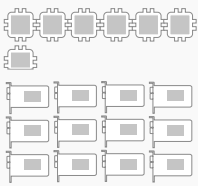
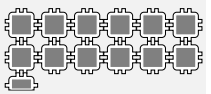
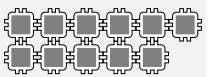
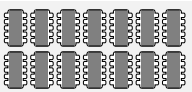
FastWARC: Optimizing Large-Scale Web Archive Analytics

Janek Bevendorff, Martin Potthast, Benno Stein

Webis Group
Bauhaus-Universität Weimar, Leipzig University
webis.de | resiliparse.chatnoir.eu

OSSYM 2021 – 3rd International Open Search Symposium, October 12, 2021

Webis BigData Cluster Infrastructure

	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	135 	9 	78 	55 
Disk [PB]	4.1 	0.08 	12 	0.1 
Cores	1,740  ≅ 67.4 TFLOPs	672 + 227,328  ≅ 8 PFLOPs	1,248  ≅ 119.8 TFLOPs	1,100  ≅ 44 TFLOPs
RAM [TB]	28 	7.5 	10 	7 

Webis Web Archive Analytics Stack

	Task stack	Technology stack	Vendor stack
Data Consumption Layer	· Query and explore · Visualize and interact · Explain and justify	· Visual analytics · Immersive technologies · Intelligent agents	
Data Analytics Layer	· Diagnose and reason · Structure identification · Structure verification	· Distributed learning · State-space search · Symbolic inference	
Data Management + Hardware Layer	· Provenance tracking · Normalization · Cleansing · Monitoring · Replication	· Key-value store · RDF triple store · Graph store · Object store · Orchestration · Parallelization · Virtualization	
Data Acquisition Layer	· Replay · Collect · Log	· Distant supervision · Crowdsourcing · Crawling and archiving	

Webis Web Archive Analytics Stack

The ChatNoir Web Search Engine

- Indexes Clueweb09, Clueweb12, Common Crawl 11/2015, 04/2017
- Approx. 5 billion documents
- 87 TiB of index data
- 1 920 shards
- Service provided free of charge to the IR community.

Webis Web Archive Analytics Stack

	Task stack	Technology stack	Vendor stack
Data Consumption Layer	· Query and explore · Visualize and interact · Explain and justify	· Visual analytics · Immersive technologies · Intelligent agents	
Data Analytics Layer	· Diagnose and reason · Structure identification · Structure verification	· Distributed learning · State-space search · Symbolic inference	
Data Management + Hardware Layer	· Provenance tracking · Normalization · Cleansing · Monitoring · Replication	· Key-value store · RDF triple store · Graph store · Object store · Orchestration · Parallelization · Virtualization	
Data Acquisition Layer	· Replay · Collect · Log	· Distant supervision · Crowdsourcing · Crawling and archiving	

Webis Web Archive Analytics Stack

	Task stack	Technology stack	Vendor stack
Data Consumption Layer	· Query and explore · Visualize and interact · Explain and justify	· Visual analytics · Immersive technologies · Intelligent agents	
Data Analytics Layer	· Diagnose and reason · Structure identification · Structure verification	· Distributed learning · State-space search · Symbolic inference	
Data Management + Hardware Layer	· Provenance tracking · Normalization · Cleansing · Monitoring · Replication	· Key-value store · RDF triple store · Graph store · Object store · Orchestration · Parallelization · Virtualization	
Data Acquisition Layer	· Replay · Collect · Log	· Distant supervision · Crowdsourcing · Crawling and archiving	

WARC Basics

```
WARC/1.0
WARC-Type:  response
WARC-Date:  2021-10-08T17:27:41Z
WARC-Record-ID: <urn:uuid:25a91250-dffe-4cab-a56b-2ffa07c21ec7>
Content-Length:  36578
Content-Type:  application/http; msgtype=response
WARC-Warcinfo-ID: <urn:uuid:dba95eaf-b2ef-4453-82b8-21d106ec3463>
WARC-Concurrent-To:  <urn:uuid:3ea5bff2-8f7f-4d80-9c11-320393c638ad>
WARC-Target-URI: https://example.com
WARC-Payload-Digest:  sha1:CUJAG44UJDP64LTXHBIFKPXD54UHBP4S
WARC-Block-Digest:  sha1:QXLZPJ4UUKOBZFWZWSAGHPHBT7X7UD2Q
WARC-Identified-Payload-Type:  text/html

HTTP/1.1 200 OK
Content-Type:  text/html
X-Crawler-Content-Encoding:  gzip
Vary:  Accept-Encoding
Server:  Microsoft-IIS/8.5
X-Powered-By:  PHP/5.2.17

...
```

Sample Calculation

- A somewhat recent Common Crawl[1]:

  64 000 WARC files, totalling **62.5 TiB** GZip-compressed

- WARCIO[2] needs approx. **19.5 s** / **file**.

[1] https://commoncrawl.org/
[2] https://github.com/webrecorder/warcio

Sample Calculation

- A somewhat recent Common Crawl[1]:

  64 000 WARC files, totalling **62.5 TiB** GZip-compressed

- WARCIO[2] needs approx. **19.5 s** / **file**.
- That's **345 hours** of compute! (5.5 hours / TiB)

[1] https://commoncrawl.org/
[2] https://github.com/webrecorder/warcio

Sample Calculation

- A somewhat recent Common Crawl[1]:

  64 000 WARC files, totalling **62.5 TiB** GZip-compressed

- WARCIO[2] needs approx. **19.5 s** / **file**.
- That's **345 hours** of compute! (5.5 hours / TiB)
- Time per PiB: **5 670 hours**

[1] https://commoncrawl.org/
[2] https://github.com/webrecorder/warcio

Sample Calculation

- A somewhat recent Common Crawl[1]:

  64 000 WARC files, totalling **62.5 TiB** GZip-compressed

- WARCIO[2] needs approx. **19.5 s** / **file**.
- That's **345 hours** of compute! (5.5 hours / TiB)
- Time per PiB: **5 670 hours**

Can we speed this up?

[1] https://commoncrawl.org/
[2] https://github.com/webrecorder/warcio

Sample Calculation

- A somewhat recent Common Crawl[1]:

  64 000 WARC files, totalling **62.5 TiB** GZip-compressed

- WARCIO[2] needs approx. **19.5 s** / **file**.
- That's **345 hours** of compute! (5.5 hours / TiB)
- Time per PiB: **5 670 hours**

Can we speed this up?

A reduction to only **12 s** / **file** (speedup: 1.6x) would save. . .

[1] https://commoncrawl.org/
[2] https://github.com/webrecorder/warcio

Sample Calculation

- A somewhat recent Common Crawl[1]:

  64 000 WARC files, totalling **62.5 TiB** GZip-compressed

- WARCIO[2] needs approx. **19.5 s** / **file**.
- That's **345 hours** of compute! (5.5 hours / TiB)
- Time per PiB: **5 670 hours**

Can we speed this up?

A reduction to only **12 s** / **file** (speedup: 1.6x) would save. . .

- **132 hours** on a Common Crawl,
- **2 157 hours per PiB!**

[1] https://commoncrawl.org/
[2] https://github.com/webrecorder/warcio

Well. . . We can!

We introduce: **FastWARC**, a new WARC parsing library:

- Written from the ground up in optimized C/C++ using Cython
- API familiar to WARCIO users
- Compatible with Python 3.7+
- Pre-built binaries for Linux, macOS, Windows on PyPi
- Open Source under the Apache 2.0 license

```
from fastwarc.warc import ArchiveIterator

for record in ArchiveIterator(open('warcfile.warc.gz', 'rb')):
    pass
```

FastWARC Benchmarks (Then)

Comp.	Parser	Records/s	Speedup
AMD Ryzen Threadripper 2920X (NVMe SSD)			
None	WARCIO	13 971.5	–
None	FastWARC	64 698.0	4.6
None	WARCIO+HTTP	13 570.2	–
None	FastWARC+HTTP	58 354.0	4.3
None	WARCIO+HTTP+Checksum	7 890.9	–
None	FastWARC+HTTP+Checksum	11 528.6	1.5
GZip	WARCIO	5 898.8	–
GZip	FastWARC	8 899.1	1.5
GZip	WARCIO+HTTP	5 986.1	–
GZip	FastWARC+HTTP	8 659.0	1.4
GZip	WARCIO+HTTP+Checksum	4 544.7	–
GZip	FastWARC+HTTP+Checksum	5 022.6	1.1
LZ4	FastWARC	36 862.8	6.2*
LZ4	FastWARC+HTTP	36 327.9	6.2*
LZ4	FastWARC+HTTP+Checksum	10 110.0	2.2*
Intel(R) Xeon(R) CPU E5-2620 v2 (remote Ceph storage)			
None	WARCIO	7 865.7	–
None	FastWARC	29 307.7	3.7
GZip	WARCIO	3 438.4	–
GZip	FastWARC	4 583.3	1.3
LZ4	FastWARC	18 337.0	5.3*

*LZ4 speedup is over WARCIO with GZip, since WARCIO does not support LZ4.

FastWARC Benchmarks (Now)

Comp.	Parser	Records/s	Speedup
AMD Ryzen Threadripper 2920X (NVMe SSD)			
None	WARCIO	16 945.5	–
None	FastWARC	108 488.0	6.4 (4.6)
None	WARCIO+HTTP	11 661.6	–
None	FastWARC+HTTP	79 297.0	6.8 (4.3)
None	WARCIO+HTTP+Checksum	6 986.7	–
None	FastWARC+HTTP+Checksum	21 320.9	3.1 (1.5)
GZip	WARCIO	6 460.1	–
GZip	FastWARC	10 413.4	1.6 (1.5)
GZip	WARCIO+HTTP	5 435.6	–
GZip	FastWARC+HTTP	10 101.5	1.9 (1.4)
GZip	WARCIO+HTTP+Checksum	4 121.6	–
GZip	FastWARC+HTTP+Checksum	7 433.0	1.8 (1.1)
LZ4	FastWARC	49 825.4	7.7 (6.2)*
LZ4	FastWARC+HTTP	42 394.5	7.8 (6.2)*
LZ4	FastWARC+HTTP+Checksum	16 992.2	4.1 (2.2)*
Intel(R) Xeon(R) CPU E5-2620 v2 (remote Ceph storage)			
None	WARCIO	7 969.1	–
None	FastWARC	49 396.5	6.2 (3.7)
GZip	WARCIO	3 555.7	–
GZip	FastWARC	6 335.1	1.8 (1.3)
LZ4	FastWARC	28 313.8	8.0 (5.5)*

*LZ4 speedup is over WARCIO with GZip, since WARCIO does not support LZ4.

FastWARC Benchmarks (Now)

Comp.	Parser	Records/s	Speedup
AMD Ryzen Threadripper 2920X (NVMe SSD)			
None	WARCIO	16 945.5	–
None	FastWARC	108 488.0	6.4 (4.6)
None	WARCIO+HTTP	11 661.6	–
None	FastWARC+HTTP	79 297.0	6.8 (4.3)
None	WARCIO+HTTP+Checksum	6 986.7	–
None	FastWARC+HTTP+Checksum	21 320.9	3.1 (1.5)
GZip	WARCIO	6 460.1	–
GZip	FastWARC	10 413.4	1.6 (1.5)
GZip	WARCIO+HTTP	5 435.6	–
GZip	FastWARC+HTTP	10 101.5	1.9 (1.4)
GZip	WARCIO+HTTP+Checksum	4 121.6	–
GZip	FastWARC+HTTP+Checksum	7 433.0	1.8 (1.1)
LZ4	FastWARC	49 825.4	7.7 (6.2)*
LZ4	FastWARC+HTTP	42 394.5	7.8 (6.2)*
LZ4	FastWARC+HTTP+Checksum	16 992.2	4.1 (2.2)*
Intel(R) Xeon(R) CPU E5-2620 v2 (remote Ceph storage)			
None	WARCIO	7 969.1	–
None	FastWARC	49 396.5	6.2 (3.7)
GZip	WARCIO	3 555.7	–
GZip	FastWARC	6 335.1	1.8 (1.3)
LZ4	FastWARC	28 313.8	8.0 (5.5)*

*LZ4 speedup is over WARCIO with GZip, since WARCIO does not support LZ4.

FastWARC Benchmarks (Now)

Comp.	Parser	Records/s	Speedup
AMD Ryzen Threadripper 2920X (NVMe SSD)			
None	WARCIO	16 945.5	–
None	FastWARC	108 488.0	6.4 (4.6)
None	WARCIO+HTTP	11 661.6	–
None	FastWARC+HTTP	79 297.0	6.8 (4.3)
None	WARCIO+HTTP+Checksum	6 986.7	–
None	FastWARC+HTTP+Checksum	21 320.9	3.1 (1.5)
GZip	WARCIO	6 460.1	–
GZip	FastWARC	10 413.4	1.6 (1.5)
GZip	WARCIO+HTTP	5 435.6	–
GZip	FastWARC+HTTP	10 101.5	1.9 (1.4)
GZip	WARCIO+HTTP+Checksum	4 121.6	–
GZip	FastWARC+HTTP+Checksum	7 433.0	1.8 (1.1)
LZ4	FastWARC	49 825.4	7.7 (6.2)*
LZ4	FastWARC+HTTP	42 394.5	7.8 (6.2)*
LZ4	FastWARC+HTTP+Checksum	16 992.2	4.1 (2.2)*
Intel(R) Xeon(R) CPU E5-2620 v2 (remote Ceph storage)			
None	WARCIO	7 969.1	–
None	FastWARC	49 396.5	6.2 (3.7)
GZip	WARCIO	3 555.7	–
GZip	FastWARC	6 335.1	1.8 (1.3)
LZ4	FastWARC	28 313.8	8.0 (5.5)*

*LZ4 speedup is over WARCIO with GZip, since WARCIO does not support LZ4.

FastWARC Benchmarks (Now)

Comp.	Parser	Records/s	Speedup
AMD Ryzen Threadripper 2920X (NVMe SSD)			
None	WARCIO	16 945.5	–
None	FastWARC	108 488.0	6.4 (4.6)
None	WARCIO+HTTP	11 661.6	–
None	FastWARC+HTTP	79 297.0	6.8 (4.3)
None	WARCIO+HTTP+Checksum	6 986.7	–
None	FastWARC+HTTP+Checksum	21 320.9	3.1 (1.5)
GZip	WARCIO	6 460.1	–
GZip	FastWARC	10 413.4	1.6 (1.5)
GZip	WARCIO+HTTP	5 435.6	–
GZip	FastWARC+HTTP	10 101.5	1.9 (1.4)
GZip	WARCIO+HTTP+Checksum	4 121.6	–
GZip	FastWARC+HTTP+Checksum	7 433.0	1.8 (1.1)
LZ4	FastWARC	49 825.4	7.7 (6.2)*
LZ4	FastWARC+HTTP	42 394.5	7.8 (6.2)*
LZ4	FastWARC+HTTP+Checksum	16 992.2	4.1 (2.2)*
Intel(R) Xeon(R) CPU E5-2620 v2 (remote Ceph storage)			
None	WARCIO	7 969.1	–
None	FastWARC	49 396.5	6.2 (3.7)
GZip	WARCIO	3 555.7	–
GZip	FastWARC	6 335.1	1.8 (1.3)
LZ4	FastWARC	28 313.8	8.0 (5.5)*

*LZ4 speedup is over WARCIO with GZip, since WARCIO does not support LZ4.

FastWARC Benchmarks (Now)

Comp.	Parser	Records/s	Speedup
AMD Ryzen Threadripper 2920X (NVMe SSD)			
None	WARCIO	16 945.5	–
None	FastWARC	108 488.0	6.4 (4.6)
None	WARCIO+HTTP	11 661.6	–
None	FastWARC+HTTP	79 297.0	6.8 (4.3)
None	WARCIO+HTTP+Checksum	6 986.7	–
None	FastWARC+HTTP+Checksum	21 320.9	3.1 (1.5)
GZip	WARCIO	6 460.1	–
GZip	FastWARC	10 413.4	1.6 (1.5)
GZip	WARCIO+HTTP	5 435.6	–
GZip	FastWARC+HTTP	10 101.5	1.9 (1.4)
GZip	WARCIO+HTTP+Checksum	4 121.6	–
GZip	FastWARC+HTTP+Checksum	7 433.0	1.8 (1.1)
LZ4	FastWARC	49 825.4	7.7 (6.2)*
LZ4	FastWARC+HTTP	42 394.5	7.8 (6.2)*
LZ4	FastWARC+HTTP+Checksum	16 992.2	4.1 (2.2)*
Intel(R) Xeon(R) CPU E5-2620 v2 (remote Ceph storage)			
None	WARCIO	7 969.1	–
None	FastWARC	49 396.5	6.2 (3.7)
GZip	WARCIO	3 555.7	–
GZip	FastWARC	6 335.1	1.8 (1.3)
LZ4	FastWARC	28 313.8	8.0 (5.5)*

*LZ4 speedup is over WARCIO with GZip, since WARCIO does not support LZ4.

ChatNoir Resiliparse

FastWARC is part of the **ChatNoir Resiliparse** web archive analytics library:

-  ChatNoir: **https://chatnoir.eu**
-  GitHub: **https://git.io/resiliparse**
-  (Extensive) documentation: **https://resiliparse.chatnoir.eu**
-  PyPi package: `pip install fastwarc`

ChatNoir Resiliparse

FastWARC is part of the **ChatNoir Resiliparse** web archive analytics library:

-  ChatNoir: **https://chatnoir.eu**
-  GitHub: **https://git.io/resiliparse**
-  (Extensive) documentation: **https://resiliparse.chatnoir.eu**
-  PyPi package: `pip install fastwarc`

Thank you!