# StormCrawler at Common Crawl and other use cases

IIPC 2019 Zagreb

CameraForensics

# Part 1

Julien Nioche - CameraForensics

Bristol, UK

*Introduction to StormCrawler*

*Use cases*

CameraForensics

# Part 2

## Sebastian Nagel - CommonCrawl

*News crawler*

CameraForensics

# STORMCRAWLER

CameraForensics

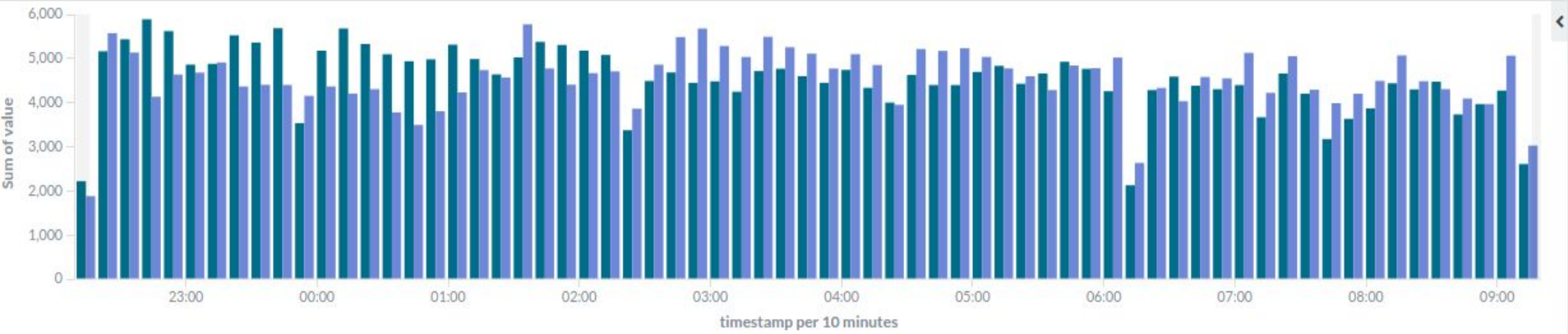
- Written in Java
- Open Source (Apache licence)
- Distributed / Scalable
- Robust
- Efficient yet polite
- Low latency
- Modular
- Lightweight
- Versatile

# Ecosystem

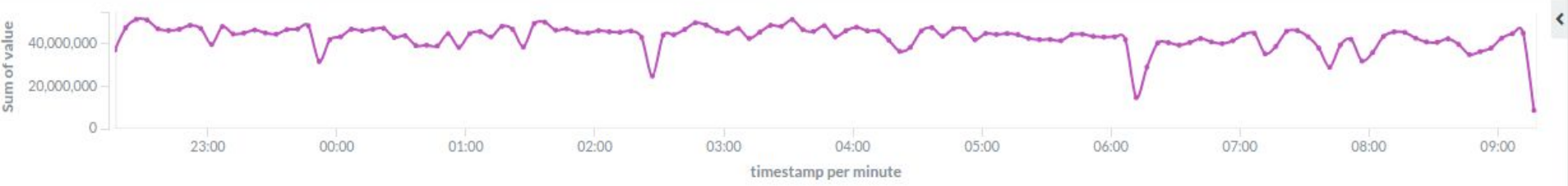
- Rich set of features
  - Extraction
  - Filtering
  - Sitemaps
  - Robots.txt
- Modules
  - including WARC
- Monitoring
- Metrics
  - Kibana
  - Grafana

CameraForensics

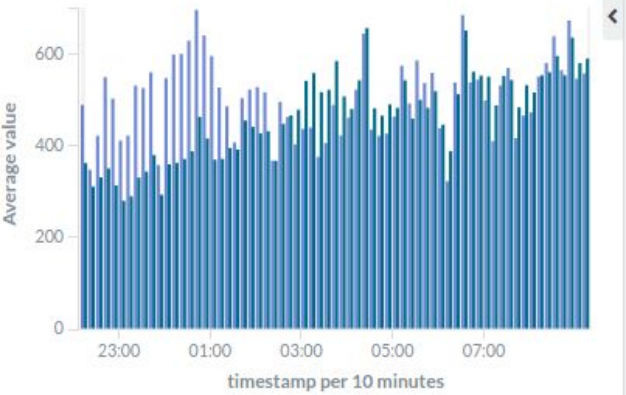
Fetcher : pages fetched



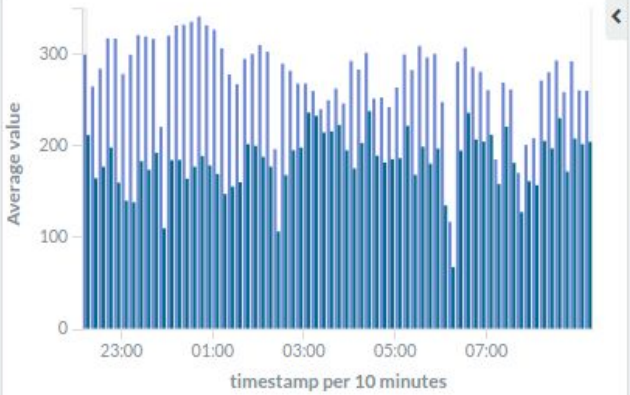

Total bytes fetched



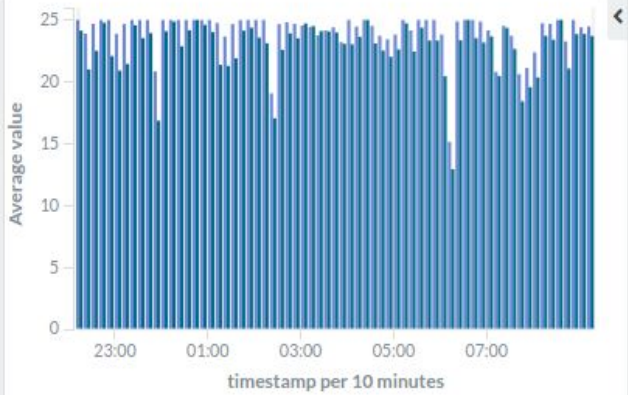

Fetcher : URLs waiting in queues




Fetcher : num queues




Fetcher : # active threads

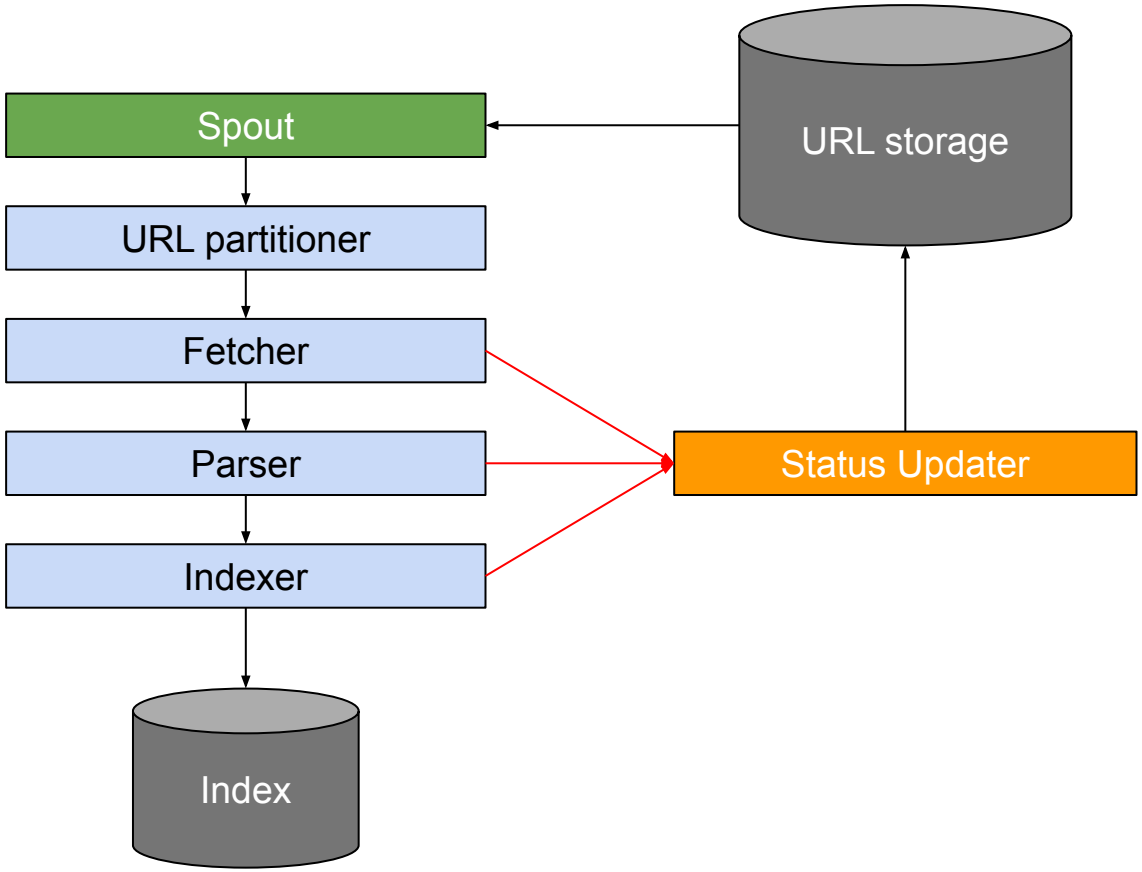


# Storage agnostic

- Elasticsearch
- SOLR
- SQL
- Distributed Queues
- ...

CameraForensics

# Basic Dataflow



# Sample of users

https://github.com/DigitalPebble/storm-crawler/wiki/Powered-By

# Internet Advertisement

- Streaming URLs
- Distributed queue
- Simple architecture
- Low latency

CameraForensics

# Government of Northwest Territories

## Search Engine

- Replace Google Search Appliance
- Limited number of sites
- Index to Elasticsearch

CameraForensics

# Search

"search engine"

60 results found in 28ms

Filters

Domain: gov.nt.ca | x

Website
- Education, Culture, and Employment 15
- Finance 12
- Industry, Tourism, and Investment 3
- Justice 3
- Environment and Natural Resources 2

View all

Domain
- **gov.nt.ca** 60

Document Format
- pdf 44
- html 14
- word 2

Type
- Information 12

## Top 5 Takeaways from the 2019/20 NWT Tourism Marketing ...

*https://www.iti.gov.nt.ca/fr/node/4651*

Using **Search Engine** Optimization (SEO) and **Search Engine** Marketing (SEM), NWTT will pull more of the right people to read and interact with the NWT stories and content.

## Top 5 Takeaways from the 2019/20 NWT Tourism Marketing ...

*https://www.iti.gov.nt.ca/en/newsroom/top-5-takeaways-201920-nwt-tourism-ma...*

Using **Search Engine** Optimization (SEO) and **Search Engine** Marketing (SEM), NWTT will pull more of the right people to read and interact with the NWT stories and content.

## Mark this Page!

*https://www.gov.nt.ca/en/newsroom/nwt-residents-get-new-direct-information-...*

RSS readers can be found using your favourite **search engine**. For more information contact: James MacKenzie Media Liaison Office of the Premier and Executive Council Government of the Northwest Territories Tel: (867) 669-2302 Fax: (867) 873-0111

## Mark this Page!

*https://www.gov.nt.ca/fr/newsroom/nwt-residents-get-new-direct-information-...*

RSS readers can be found using your favourite **search engine**. For more information contact: James MacKenzie Media Liaison Office of the Premier and Executive Council Government of the Northwest Territories Tel: (867) 669-2302 Fax: (867) 873-0111

## Digital Literacy - Teacher Resource

*https://www.ece.gov.nt.ca/en/services/resources-current-teachersprincipals/...*

Its five modules focus on social networking and privacy; tricks of the trade of online advertisers; authentication skills and judging reliability through a mock **search**

# Image tracking

- Image Rights Protection
- Different types of crawls
  - total 200M pages / day
  - up to 10B URLs
  - 10’s millions domains
- StormCrawler + Elasticsearch
- http://digitalpebble.blogspot.com/2019/02/meet-stormcrawler-users-q-with-pixray.html

CameraForensics

CameraForensics

# Child protection

- Crawl photos to identify cameras
- Over 2 billion images
- Elasticsearch storage
- Heavily extended
  - Custom analysis

# http://stormcrawler.net

# StormCrawler at **Common Crawl** and other use cases

## Part II – News Crawler

Sebastian Nagel

sebastian@commoncrawl.org

IIPC Web Archiving Conference, 6–7 June 2019, Zagreb, Croatia

# About Common Crawl

- we're a non-profit that makes web data accessible to programmers and data scientists
- for natural language processing, web science, …
- hosted as Open Data set on Amazon Web Services
- main crawl data released monthly:
  3 billion page captures, 50 TiB WARC files
- 4 Petabytes of data in total (2008 – 2019)

Why a news dataset?

- monthly crawl and release schedule (as for main crawls) is not optimal for the news genre
- continuously release freshly crawled data
- broad and global sample
- news feeds and sitemaps are used as seeds
  - well-maintained and clean (almost no spam, few ads)
  - only recent news articles crawled using the publication date announced in feeds and sitemaps
- small collection, for users easy to work with
- as usual: HTML pages (news articles) archived in WARC files

# The News Data Set – Basic numbers

- started autumn 2016
- 5.5 TiB WARC files, 220 million pages in total (May 2019)
- currently: 8 – 10 million pages per month
- from 12,000 domains
- data from 200 top-level domains and over 100 languages

# The News Data Set – languages and top-level domains

| % | % cumul. | TLD |
|---|---|---|
| 44.4 | 44.4 | com |
| 6.9 | 51.3 | de |
| 4.0 | 55.3 | it |
| 3.3 | 58.6 | ru |
| 3.0 | 61.7 | uk |
| 2.8 | 64.4 | es |
| 2.4 | 66.9 | co |
| 2.0 | 68.8 | net |
| 1.8 | 70.6 | fr |
| 1.6 | 72.2 | gr |
| 1.4 | 73.6 | tr |
| 1.4 | 75.0 | ua |
| 1.3 | 76.3 | ro |
| 1.2 | 77.5 | au |
| 1.2 | 78.7 | se |
| 1.0 | 79.7 | ch |
| 1.0 | 80.7 | nl |
| 0.9 | 81.6 | mx |
| 0.9 | 82.5 | org |
| 0.7 | 83.3 | cz |
| 0.7 | 83.9 | pl |
| 0.7 | 84.6 | ca |
| 0.7 | 86.0 | hu |
| 0.7 | 86.6 | br |
| 0.7 | 87.3 | at |

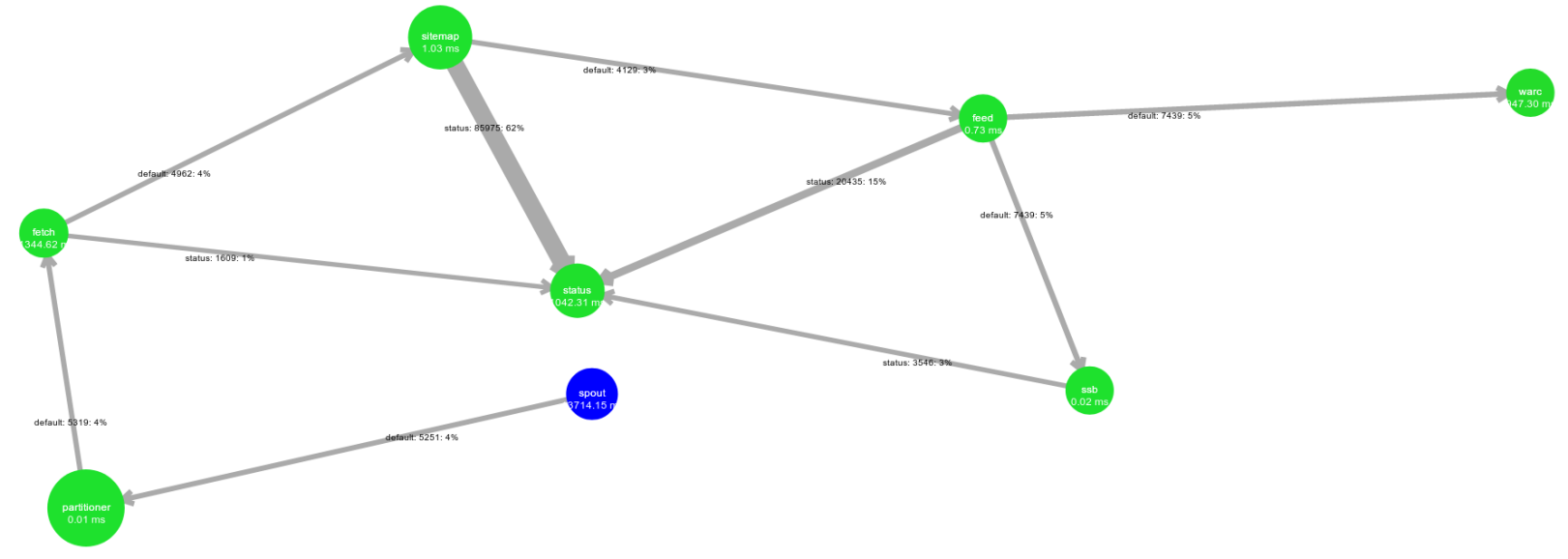
| % | % cumul. | language |
|---|---|---|
| 34.8 | 34.8 | English |
| 10.3 | 45.1 | Spanish |
| 8.9 | 54.0 | German |
| 5.7 | 59.6 | Russian |
| 5.5 | 65.2 | Italian |
| 4.6 | 69.7 | Arabic |
| 4.0 | 73.7 | Turkish |
| 3.7 | 77.4 | French |
| 2.3 | 79.7 | Hindi |
| 1.7 | 81.4 | Greek |
| 1.7 | 83.1 | Portuguese |
| 1.5 | 84.6 | Romanian |
| 1.3 | 85.9 | Swedish |
| 1.3 | 87.2 | Dutch |
| 0.8 | 88.0 | Chinese |
| 0.8 | 88.8 | Korean |
| 0.8 | 89.5 | Hungarian |
| 0.8 | 90.3 | Polish |
| 0.7 | 91.0 | Czech |
| 0.7 | 91.7 | Ukrainian |
| 0.6 | 92.3 | Japanese |
| 0.6 | 93.0 | Tamil |
| 0.6 | 93.6 | Indonesian |
| 0.6 | 94.1 | Bulgarian |
| 0.5 | 94.6 | Vietnamese |

# Why StormCrawler?

- streaming approach allows to follow links very quickly

  (Note: for the main crawls we use Nutch, a batch-based crawler)
- flexible topology creation
- Elasticsearch index allows at any time to instantly
  - add news feeds or sitemaps
  - get metrics and statistics
- runs with little supervision over months
- on a single dual core machine with 32 MB RAM

# News crawler topology




(Storm topology visualization)

- Spout reads Elasticsearch status index and emits tuples <URL,metadata> “ripe” for (re)fetch
- tuples are partitioned and passed to …
- Fetcher bolts fetch URLs and
  - send fetch status to Status Updater bolt
  - pass content to

# News crawler topology

- Sitemap Parser bolt
  - checks whether content is a news sitemap or sitemap index
  - if yes, parse sitemap and send links (subsitemaps or news articles) to Status Updater bolt
  - it not, pass forward to
- Feed Parser bolt
  - checks whether content is a RSS or Atom feed
  - if yes, parse and send links to Status Updater bolt
  - if not, pass forward to
- WARC writer and
- DummyIndexer which passes the tuple forward to
- Status Updater bolt
  - updates the fetch status in resp.
  - adds tuples to Elasticsearch status index

# Detection of feeds and sitemaps

- the robots.txt is checked for announced sitemaps
- news sitemaps are detected by format (other sitemaps are ignored)
- bootstrap topology to detect feeds from alternate links in HTML pages

  (needs supervision to remove comment and multi-media feeds)

# Re-fetch scheduling of feeds and sitemaps

- the crawler knows about 2 million feeds and sitemaps:
  60k feeds, 800k news sitemaps, 20k sitemap indexes, 1.2 million other sitemaps
- need to focus on frequently changing feeds/sitemaps
  - "forgotten" feeds not maintained anymore
  - some sites provide a separate sitemaps every day
  - or provide hundreds of feeds/sitemaps (by topic or location), most of them infrequently updated
- scheduling is adapted to the change frequency (detected by signature comparison)
- and may range between 90 min. and 90 days

# References and questions?

References:

- data prefix on Amazon S3: `s3://commoncrawl/crawl-data/CC-NEWS/`
- https://commoncrawl.org/2016/10/news-dataset-available/
- https://github.com/commoncrawl/news-crawl
- https://groups.google.com/forum/#!forum/common-crawl