WARC-DL: Scalable Web Archive Processing for Deep Learning

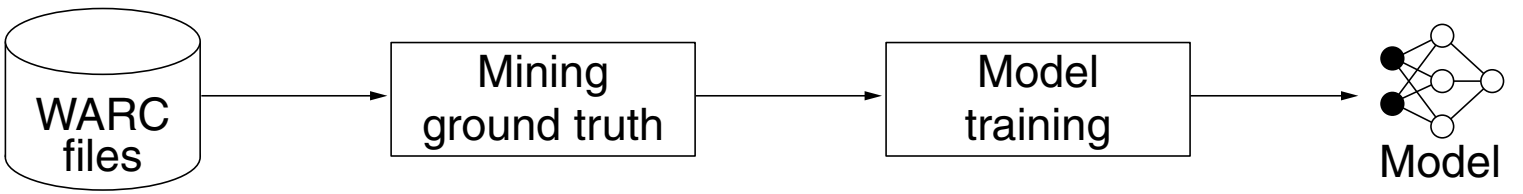
Niklas Deckers, Martin Potthast

Leipzig University
Webis Group
webis.de

OSSYM 2022 – 4th International Open Search Symposium, October 10, 2022

Web Archive Processing

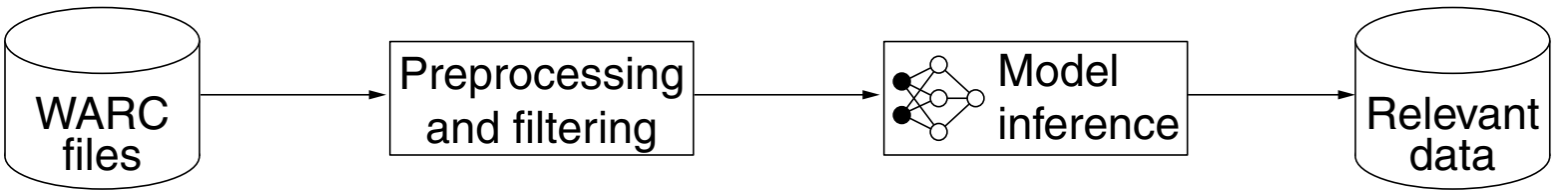
Model Training



- Given a learning task and ground truth within WARC files, train a model.
  Only a fraction of the records within the WARC files are ground truth.
- Goal: Training at web scale (billions of WARC files)

Web Archive Processing

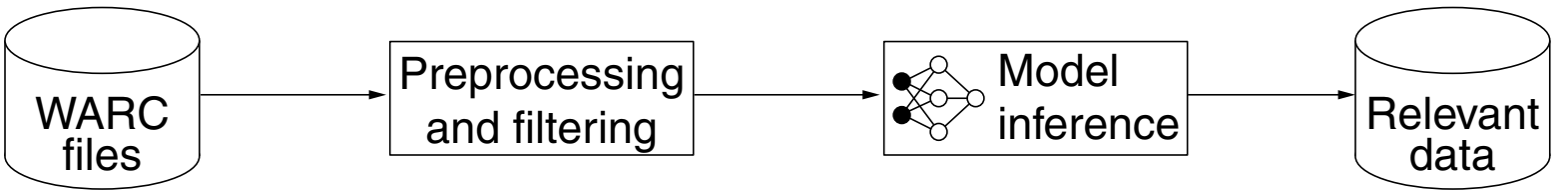
Mining



- Given a mining task and a trained (classification) model, collect relevant data.
  Only a fraction of the records within the WARC files are relevant.
- Goal: Mining at web scale (billions of WARC files)

Web Archive Processing

Mining

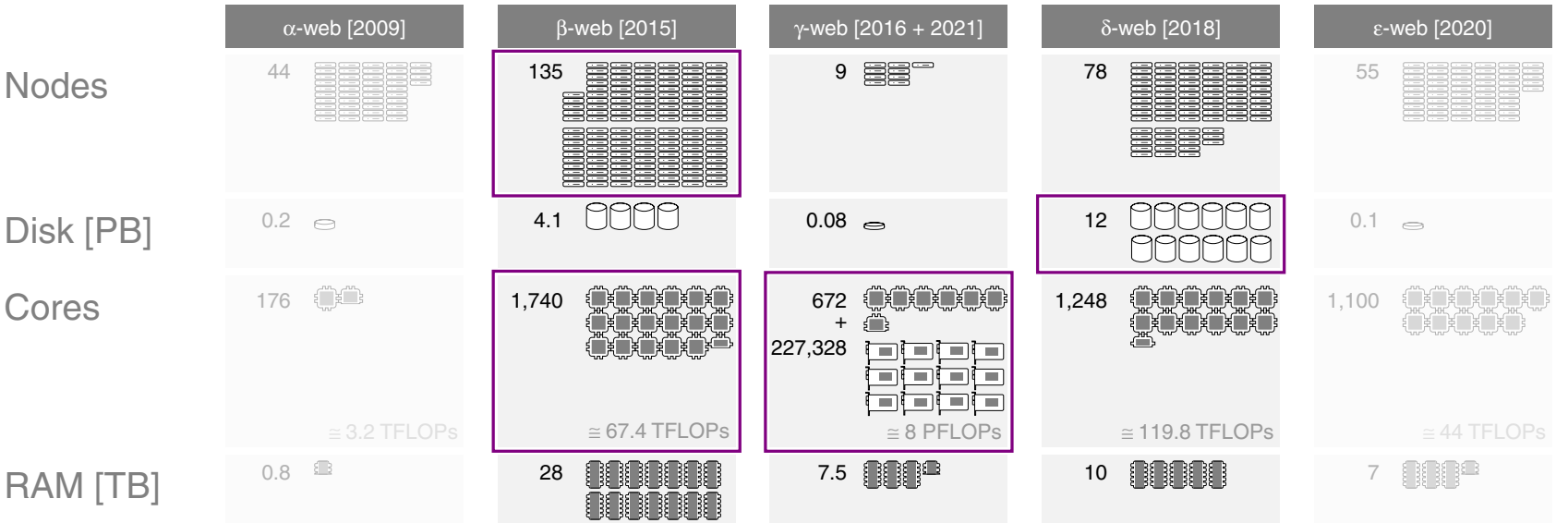


- Given a mining task and a trained (classification) model, collect relevant data.
  Only a fraction of the records within the WARC files are relevant.
- Goal: Mining at web scale (billions of WARC files)

Observations:

- Mining / filtering WARC files is “embarrassingly parallel”.
- Decompressing WARC files, and processing WARC records are CPU bound.
- The preprocessing step results in a variable data flow.
- Training of neural networks is GPU bound and presumes constant data flow.
- WARC storage, parallel processing, and GPU bound processing are on separate clusters.

Webis Data Center (Digital Bauhaus Lab)

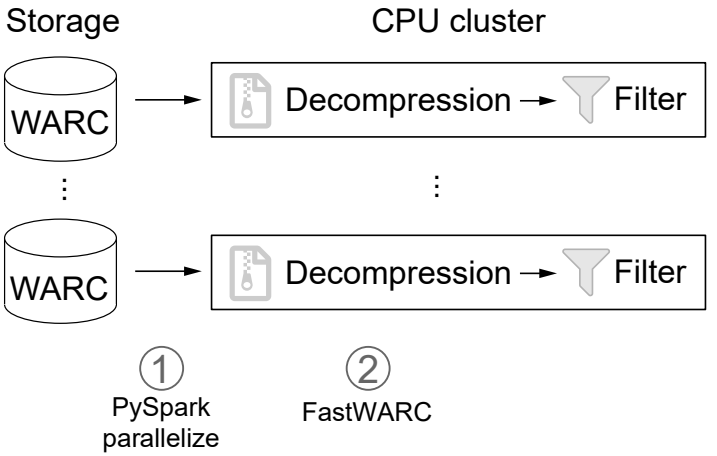


Typical research:

- α-Web. Teaching, Staging environment
- β-Web. Web mining (map reduce), CPU parallelization
- γ-Web. Machine learning (embedding, deep learning), Language modeling
- δ-Web. Web archive storage (10 PB from Internet Archive and Common Crawl)
- ϵ-Web. Search index construction, Argument search

Web Archive Processing

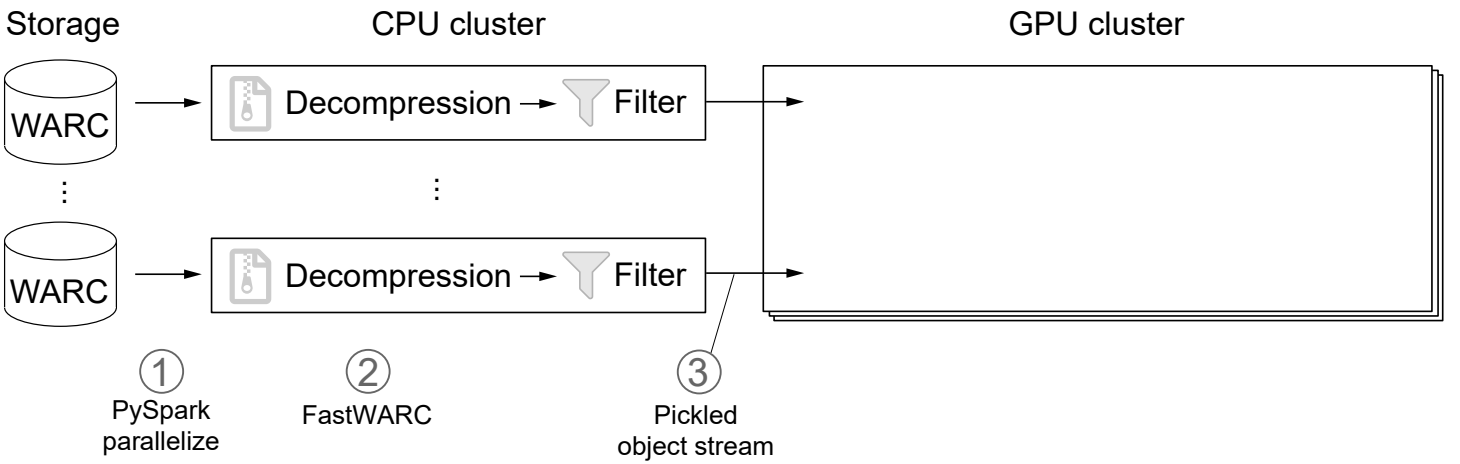
WARC-DL: Pipeline for Processing at Petabyte Scale



1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization

Web Archive Processing

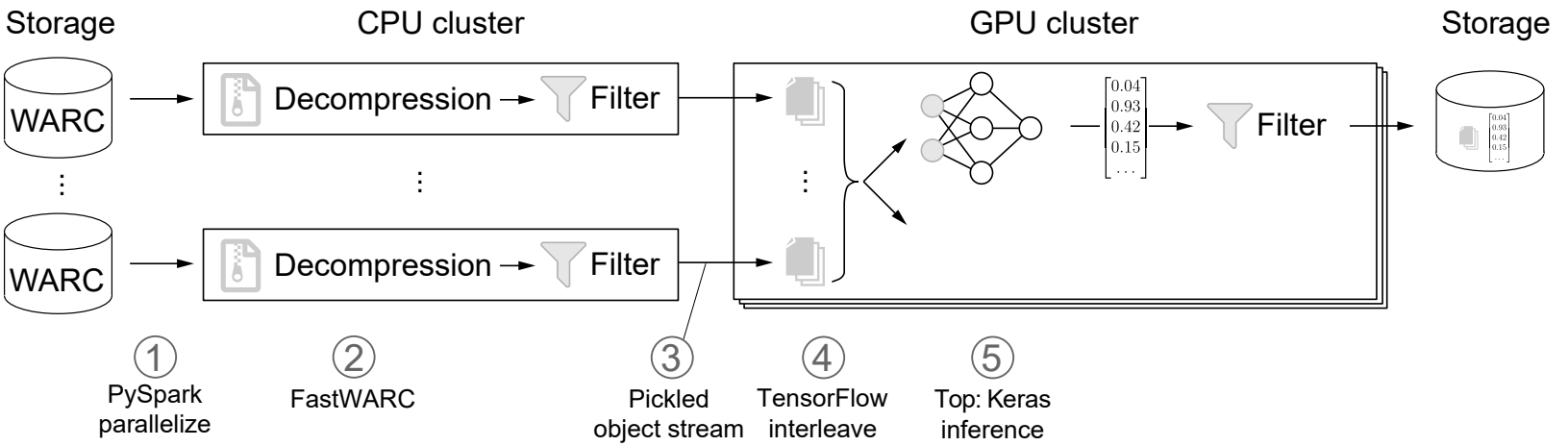
WARC-DL: Pipeline for Processing at Petabyte Scale



1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization
3. Pickled record streams

Web Archive Processing

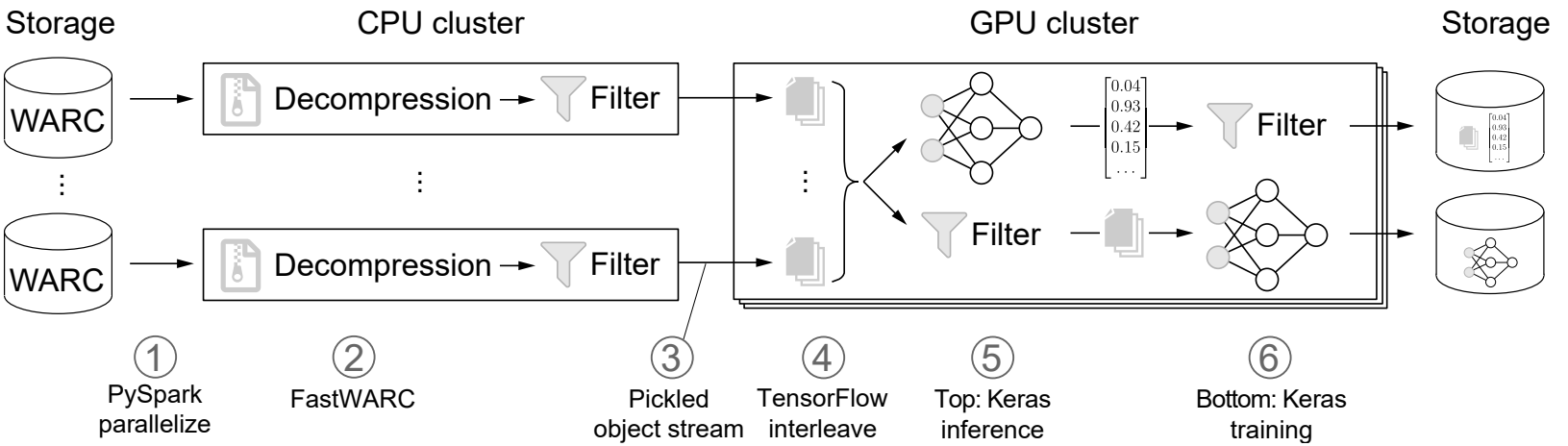
WARC-DL: Pipeline for Processing at Petabyte Scale



1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization
3. Pickled record streams
4. Conversion to TensorFlow datasets and source interleaving
5. Inference: Batched processing by a Keras model
   and second filtering based on classification results

Web Archive Processing

WARC-DL: Pipeline for Processing at Petabyte Scale



1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization
3. Pickled record streams
4. Conversion to TensorFlow datasets and source interleaving
5. Inference: Batched processing by a Keras model
   and second filtering based on classification results
6. Optional filtering (e.g., deduplication) and model training

Application: Building Large-Scale Multimodal Datasets

For Training Generative Text-To-Image Models

- CompVis group created the Latent Diffusion model
- LAION created a dataset of text-image pairs
  Consists of image urls and img alt attribute texts from Common Crawl
- Stability AI finetuned Latent Diffusion on this dataset to create Stable Diffusion

Image generated by Stable Diffusion with the prompt
“award-winning cake shaped like the Swiss Alps”

Application: Building Large-Scale Multimodal Datasets

For Training Generative Text-To-Image Models

- CompVis group created the Latent Diffusion model
- LAION created a dataset of text-image pairs
  Consists of image urls and img alt attribute texts from Common Crawl
- Stability AI finetuned Latent Diffusion on this dataset to create Stable Diffusion

- Next target together with LAION: Building a better multimodal dataset
- Obtaining such a dataset requires preprocessing,
  rule-based and DL-based filtering (e.g., NSFW filtering)
  Using the WARC-DL pipeline allows quick deployment on existing infrastructure
- Include text, images, videos and audio
- Extract more context from around the media links
  Will enable text-to-image models to work with more complex prompts

Conclusion

WARC-DL can be used for petascale web archive processing:

- Training and applying domain-specific models for web mining
- Dataset extraction
- (Multimodal) Search engines
  Will be applied in the upcoming Open Web Search project

Conclusion

WARC-DL can be used for petascale web archive processing:

- Training and applying domain-specific models for web mining
- Dataset extraction
- (Multimodal) Search engines
  Will be applied in the upcoming Open Web Search project

Thank you!