

Web Archive Analytics


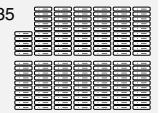









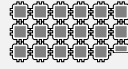








Infrastructure & Applications @ Webis

Michael Völske, Janek Bevendorff, Johannes Kiesel, Benno Stein
Bauhaus-Universität Weimar, Germany

Maik Fröbe, Matthias Hagen
Martin-Luther-Universität Halle-Wittenberg, Germany

Martin Potthast
Leipzig University, Germany

Hardware


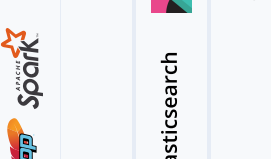
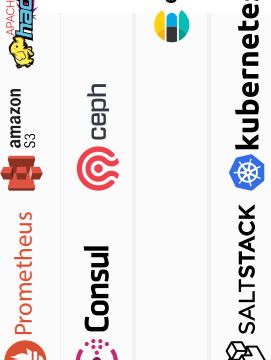
	α-web [2009]	β-web [2015]	γ-web [2016+2020]	δ-web [2018]	ε-web [2020]
Nodes	44 	135 	8 	78 	55 
Disk [PB]	0.2 	4.1 	0.1 	12 	0.1 
Cores	176  ≅ 3.2 TFLOPs	1,740  ≅ 67.4 TFLOPs	176 + 163,840  ≅ 504.7 TFLOPs	1,248  ≅ 119.8 TFLOPs	1,100  ≅ 44 TFLOPs
RAM [TB]	0.8 	28 	6.1 	10 	7 

Analytics

Task Stack

Technology stack

Vendor stack

Data Consumption Layer	<ul style="list-style-type: none"> - Query and explore - Visualize and interact - Explain and justify 	<ul style="list-style-type: none"> - Visual analytics - Immersive technologies - Intelligent agents 	
Data Analytics Layer	<ul style="list-style-type: none"> - Diagnose and reason - Structure identification - Structure verification 	<ul style="list-style-type: none"> - Distributed learning - State-space search - Symbolic inference 	
Data Management + Hardware Layer	<ul style="list-style-type: none"> - Provenance tracking - Normalization - Cleansing - Monitoring - Replication 	<ul style="list-style-type: none"> - Key-value store - RDF triple store - Graph store - Object store - Orchestration - Parallelization - Virtualization 	
Data Acquisition Layer	<ul style="list-style-type: none"> - Replay - Collect - Log 	<ul style="list-style-type: none"> - Distant supervision - Crowdsourcing - Crawling and archiving 	