WEB ARCHIVE ANALYTICS

Infrastructure & Applications @ Webis (extended abstract)

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

We describe our infrastructure to process petabytes of web data using two clusters with a total of 213 servers. A 78-node Ceph cluster will eventually host around 5 PB of web archive data from the Internet Archive and Common Crawl, with the goal of supplementing existing large scale web corpora and forming a non-biased subset of the 20 PB Internet Archive.

Figure 1 illustrates the infrastructure relevant to this effort: Two different computing clusters contribute to our web archive analytics efforts, out of five total currently operated by our research group. We enforce a rough separation of concerns between the β-web cluster (virtualization, compute) and the δ-web cluster (storage). The clusters comprise 135 and 78 Dell PowerEdge servers respectively, spread across two datacenters joined via a 400 GB/s interconnect. Each individual node is attached to a respective Cisco Nexus middle-of-row switch via a 10 GB/s link. The storage cluster maintains more than 12 petabytes of raw storage capacity across 1 248 physical spinning disks.

Both clusters are provisioned and orchestrated using the SaltStack IT automation software, which manages the deployment and configuration of the fundamental file-system- and infrastructure-level services on top of a network-booted Ubuntu Linux base image. We supervise the proper functioning of all system components with the help of Consul (service discovery) and Prometheus (event monitoring and alerting).

Our primary purpose for the δ-web cluster is a Ceph distributed storage system with one object storage daemon (OSD) per physical disk, as well as five redundant Monitor/Manager daemons. The bulk of the data payload is accessed via the RadosGW S3 API gateway service, which is provided by 7 redundant RadosGW daemons and backed by an erasure-coded storage pool. Smaller and more ephemeral datasets live in a CephFS distributed file system with 3 active MDS and 4 (hot) standbys.

On β-web, we maintain a Kubernetes cluster on top of which most of our internal and public-facing services are deployed. Kubernetes services are provided with persistent storage through a Rados block device (RBD) pool in the Ceph cluster. A Hadoop distributed file system (HDFS) holds temporary analytics data.

On top of Kubernetes, we operate a range of internal and public-facing services. The former include an Elasticsearch deployment powering our search engines, as well as a suite of data analytics tools including Hadoop Yarn, Spark, and Jupyterhub. Our public-facing services include the search engines args.me and chatnoir.eu, the netspeak.org writing assistant, the picapica.org text reuse detection system, and tira.io, our evaluation-as-a-service platform. The web archive-related services at archive.webis.de are still largely under construction and we envision allowing external collaborators to process our web archives on our infrastructure. As of July 2020, one petabyte of data—of which one third stems from the Internet Archive, the rest from the Common Crawl—has been downloaded and stored on our infrastructure so far.

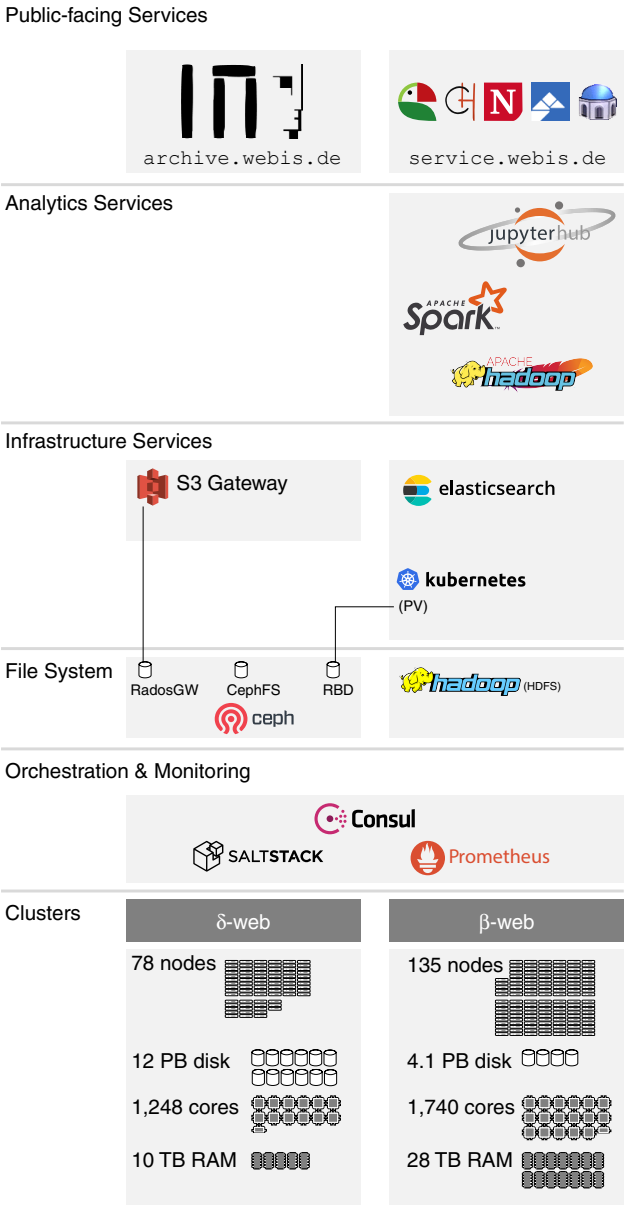

Figure 1: The Webis web archive analytics infrastructure stack.

[*] <first-name>.<last-name>@informatik.uni-halle.de
[†] <first-name>.<last-name>@uni-weimar.de
[‡] martin.potthast@uni-leipzig.de