DH 2013 Poster Proposal

**Architecture to enable large-scale computational analysis of millions of volumes**

Yiming Sun, Stacy Kowalczyk, Beth Plale, J. Stephen Downie, Loretta Auvil, Boris Capitanu, Kirk Hess, Zong Peng, Guangchen Ruan, Aaron Todd, Jiaan Zeng

The HathiTrust Research Center (HTRC) is a collaborative research center to provide Digital Humanities researchers access to not only millions of volumes from the HathiTrust (HT) digital library but also cutting-edge software tools and cyberinfrastructure to perform advanced computational analysis over the corpus at an unprecedented scale.

The corpus at the HTRC currently consists of over 3 million public domain volumes, and anticipates access to an additional 6 million in-copyright volumes. In their raw form at the HathiTrust, these volumes are stored as files on special hardware using an internal Pairtree structure. The internal HathiTrust structure is optimal for its primary function of the digital page image delivery to digital library patrons for viewing, however, it does not support well the large-scale computational analysis which is the primary function of the HTRC; navigating the Pairtree and uncompressing the text data would encounter major performance and scalability issues. While researchers from other scientific communities have been addressing aspects of the “Big Data” problem with success, the large corpus that HTRC hosts to support computational analysis presents a unique setting in that it consists of a massive number of small text-based files whereas most solutions from the scientific communities are tailored towards large files and non-text-based content. In this poster, we will present the approach the HTRC takes to solve this problem -- the HTRC keeps the Pairtree only for the purpose of synchronization with the HT, and processes and pushes the volume data from the local Pairtree to a NoSQL storage cluster using Apache Cassandra hosted on conventional hardware during the ingest process. In order to balance the data store and ingest workload, the developers at the HTRC and the HT also devised a very simple yet effective way to parallelize the rsync of the single source Pairtree at the HT on all Cassandra nodes by starting rsync at lower branches instead of at the root.

The use of a NoSQL cluster adds more complexity to the architecture than traditional file systems, but such complexity is transparent to the Digital Humanities researchers as most of the HTRC components with which user algorithms have interaction are RESTful web services, such as the Data API for accessing the data. The HTRC uses Blacklight, an open source bibliographic search and display interface, backed by a Solr index, to let users search for volumes for analysis and create collections. To apply analytical techniques to the data, a user may choose from a number of provided algorithms from the web portal, including SEASR/Meandre flows. In addition, the HTRC is actively researching and developing a secure computation environment (dubbed the Sloan Cloud) to support large-scale non-consumptive research over copyrighted volumes, and an experimental release is scheduled for end of March. This Sloan Cloud will allow researchers to deploy their own analysis algorithms against a corpus like the HT data, and to save intermediate data for later reuse, as well as to include custom worksets for the computation. We will present our early findings of the experimental Sloan Cloud and hope to get feedback from the digital humanities research community.