Title: Analyzing WARC on Serverless Computing

Since 2015, Virginia Tech Libraries uses Archive-It to preserve Virginia Tech's official web presence plus selected project web sites created and administered by Virginia Tech faculty, students, and staff. The crawl archive contains over a million web pages or 8 TiB of uncompressed content and stored in the WARC (Web ARChive) file format. To better understanding the content we have crawled or open research datasets such as Common Crawl data, we set up an on-premise Hadoop cluster with software installed to host, process, analyze, and visualize datasets.

Our usage for this cluster is Ad-hoc based, usually used when a new dataset has been crawled or an open research dataset is published for download. It becomes a heavy burden for maintaining such a cluster because it requires a dedicated system admin to ensure the cluster is on 24/7 with all the patches, security updates, and software versions remaining up to date. Moreover, the hardware could wear down and need to be replaced. Any of each action costs labor and money. As cloud computing becomes a promising place to host such a cluster, the server maintenance workload is not necessarily reduced but could instead increase using the same, traditional workflow. Provision a cluster with a set of instances always running for Ad-hoc jobs can be extremely costly and money and resources are wasted when the cluster is idle.

Ideally, we want a platform that is always on for accepting task requests but only provision the resources when it accepts the request and processes the task, scaled as needed, and cleans up the resources automatically when the job is completed. This platform can ensure near 100% resource utilization and thus comes to the idea of using serverless computing approach. To achieve our goal, we use AWS cloud-native services (AWS Lambda, Batch, ECS, etc) to design a serverless architecture that enables a resilient, scalable, and cost-effective platform. This serverless WARC processing platform costs nothing when it isn’t being used and scale as needed when processes large amounts of data. To test the scalability, cost, and performance of this platform, we use selected Common Crawl data stored in the AWS S3, extract the content, create derivatives, and measure the time spent and cost. Our analysis shows that the platform processed GBs of uncompressed data in just minutes. Furthermore, we now can have total control on how precisely the resources have been used and further optimize it.

This submission aims to present our serverless architecture design and implementations, elaborate the technical solution on integrating multiple AWS services with other techniques, and describe our streamlined and scalable approach to analyze extremely large WARC datasets. Our platform eliminates the need to manage underlying servers and delegate all the heavy lifting to AWS. We focus on implementing our business logic into microservices in AWS and construct this platform. We want to share our experience and humbly hope this work can open a new direction for institutions, libraries, and scholars on analyzing web archives.