UDaaS: A Cloud-based URL-Deduplication-as-a-Service for Big Datasets

Shams Zawoad, Ragib Hasan, Gary Warner, and Anthony Skjellum*

[zawoad, ragib, gar]@cis.uab.edu, skjellum@auburn.edu

Department of Computer & Information Sciences, University of Alabama at Birmingham, USA
*Department of Computer Science and Software Engineering, Auburn University, USA

Motivations

- Duplicate URLs introduce waste of computing and storage resources.
- The number of potential malicious URLs are too many to deduplicate using local resources.
- Leveraging the elastic nature of the cloud, we can deploy a highly scalable, parallel URL deduplication infrastructure.

Queue Manager

- Manage Fetcher Queues
- Create Fetcher Queues
- Load Queue List
- Clear Queue
- Clear All Queues
- Remove Queue

UDaaS Operation

- Create EC2 Instances
- Create SQS Queues
- Assign Queues to Instances
- Push URL Source
- Unique Website Repository

Features

- Instance Manager: Create, start, and stop instances, assign queues, view and clear logs, and change configurations.
- Queue Manager: Create, clear, and remove queues.

Applications

- UDaaS can increase a URL analyst's productivity by providing only unique content.
- Can improve the performance of phishing and other counterfeit websites detection rate.

Contribution

- We presented UDaaS, which can be used in academia and industry to easily deploy a highly scalable and distributed cloud-based infrastructure to deduplicate a big URL dataset.

References