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

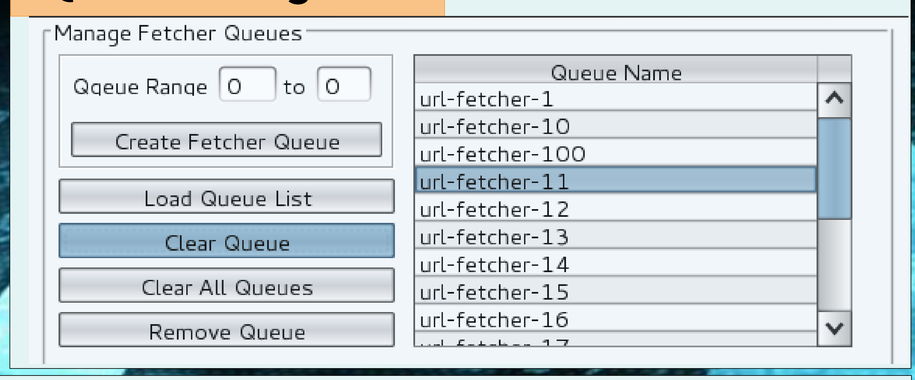
Department of Computer & Information Sciences, University of Alabama at Birmingham, USA

*Department of Computer Science and Software Engineering, Auburn University, USA

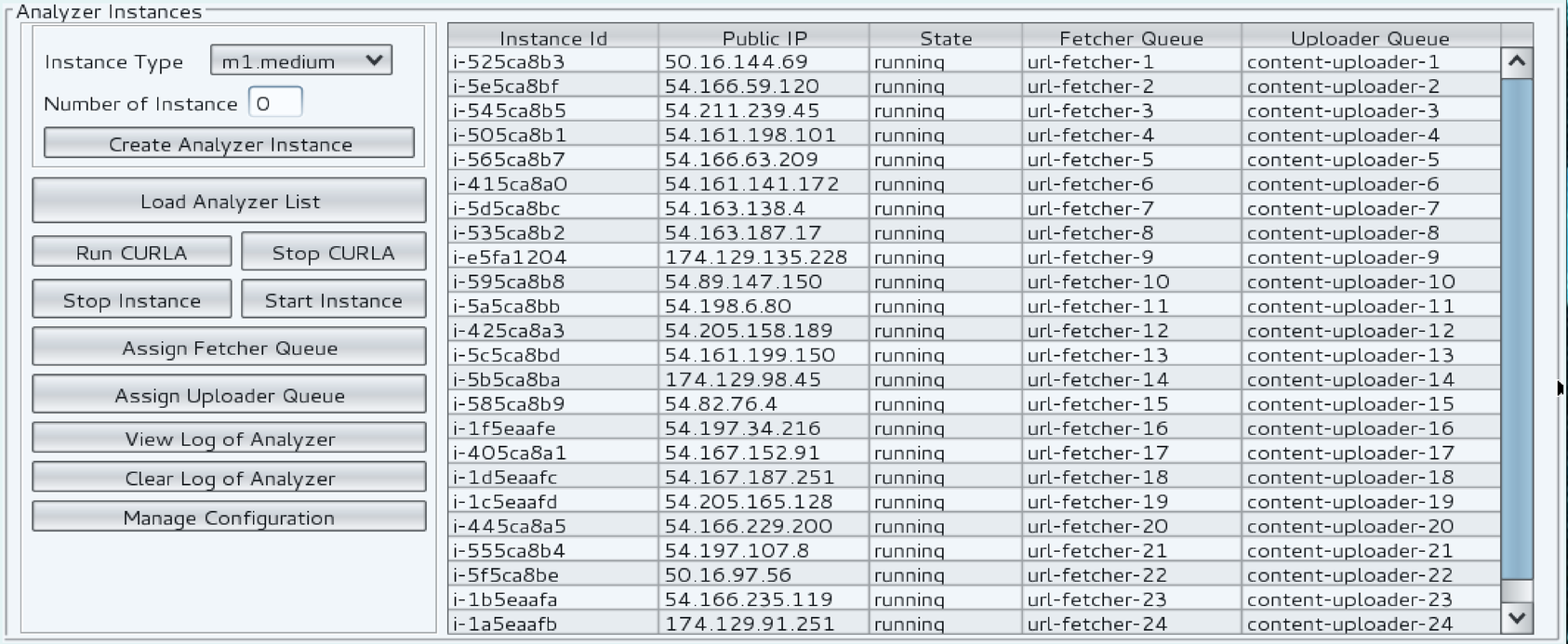
Motivations

- 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

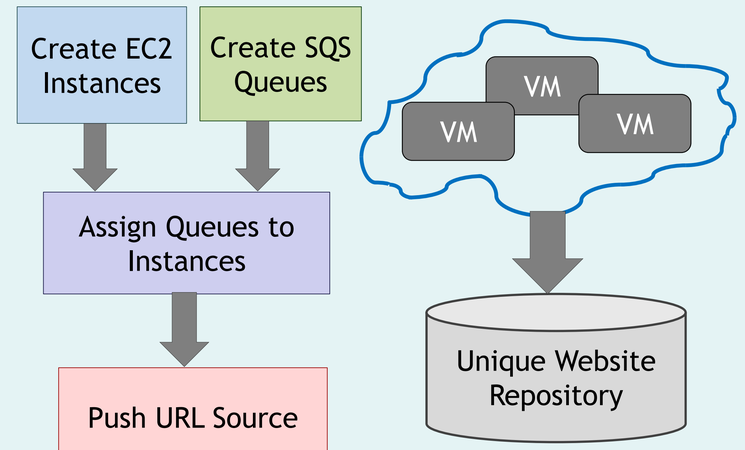


Instance Manager

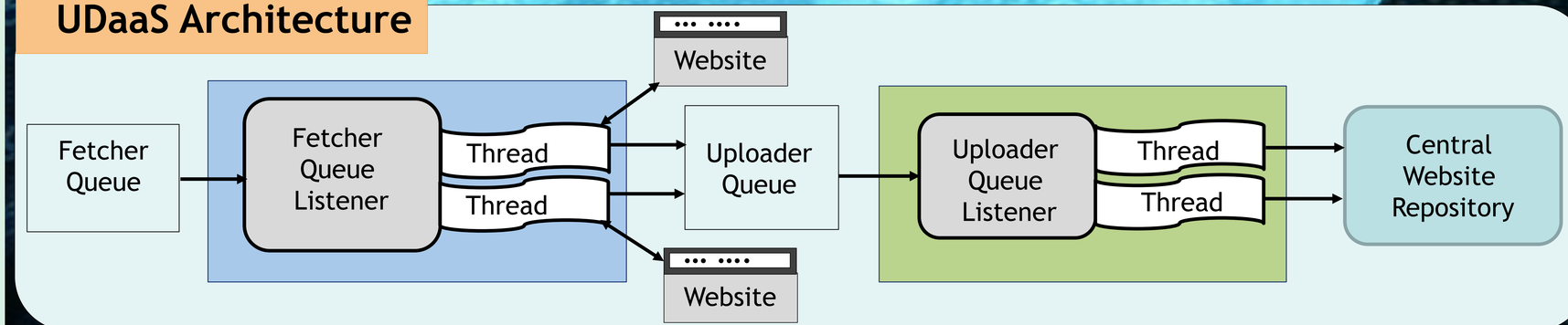


Instance Id	Public IP	State	Fetcher Queue	Uploader Queue
i-525ca8b3	50.16.144.69	running	url-fetcher-1	content-uploader-1
i-5e5ca8bf	54.166.59.120	running	url-fetcher-2	content-uploader-2
i-545ca8b5	54.211.239.45	running	url-fetcher-3	content-uploader-3
i-505ca8b1	54.161.198.101	running	url-fetcher-4	content-uploader-4
i-565ca8b7	54.166.63.209	running	url-fetcher-5	content-uploader-5
i-415ca8a0	54.161.141.172	running	url-fetcher-6	content-uploader-6
i-5d5ca8bc	54.163.138.4	running	url-fetcher-7	content-uploader-7
i-535ca8b2	54.163.187.17	running	url-fetcher-8	content-uploader-8
i-e5fa1204	174.129.135.228	running	url-fetcher-9	content-uploader-9
i-595ca8b8	54.89.147.150	running	url-fetcher-10	content-uploader-10
i-5a5ca8bb	54.198.6.80	running	url-fetcher-11	content-uploader-11
i-425ca8a3	54.205.158.189	running	url-fetcher-12	content-uploader-12
i-5c5ca8bd	54.161.199.150	running	url-fetcher-13	content-uploader-13
i-5b5ca8ba	174.129.98.45	running	url-fetcher-14	content-uploader-14
i-585ca8b9	54.82.76.4	running	url-fetcher-15	content-uploader-15
i-1f5eaafe	54.197.34.216	running	url-fetcher-16	content-uploader-16
i-405ca8a1	54.167.152.91	running	url-fetcher-17	content-uploader-17
i-1d5eaafc	54.167.187.251	running	url-fetcher-18	content-uploader-18
i-1c5eaafd	54.205.165.128	running	url-fetcher-19	content-uploader-19
i-445ca8a5	54.166.229.200	running	url-fetcher-20	content-uploader-20
i-555ca8b4	54.197.107.8	running	url-fetcher-21	content-uploader-21
i-5f5ca8be	50.16.97.56	running	url-fetcher-22	content-uploader-22
i-1b5eaafa	54.166.235.119	running	url-fetcher-23	content-uploader-23
i-1a5eaafb	174.129.91.251	running	url-fetcher-24	content-uploader-24

UDaaS Operation



UDaaS Architecture



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

[1] E. Ferguson, J. Weber, and R. Hasan, "Cloud based content fetching: Using cloud infrastructure to obfuscate phishing scam analysis," in IEEE SERVICES,2012, pp. 255-261.
[2] S. Zawoad, R. Hasan, M. M. Haque, and G. Warner, "CURLA: Cloud-based spam url analyzer for very large datasets," in IEEE Cloud, 2014.
[3] B. Wardman, T. Stallings, G. Warner, and A. Skjellum, "High-performance content-based phishing attack detection," in eCrime Researchers Summit, 2011. IEEE, 2011, pp. 1-9.
[4] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages" in NDSS, 2010.
[5] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in WWW 2007. ACM, 2007, pp. 639-648.