Comtrade Three-Sixty

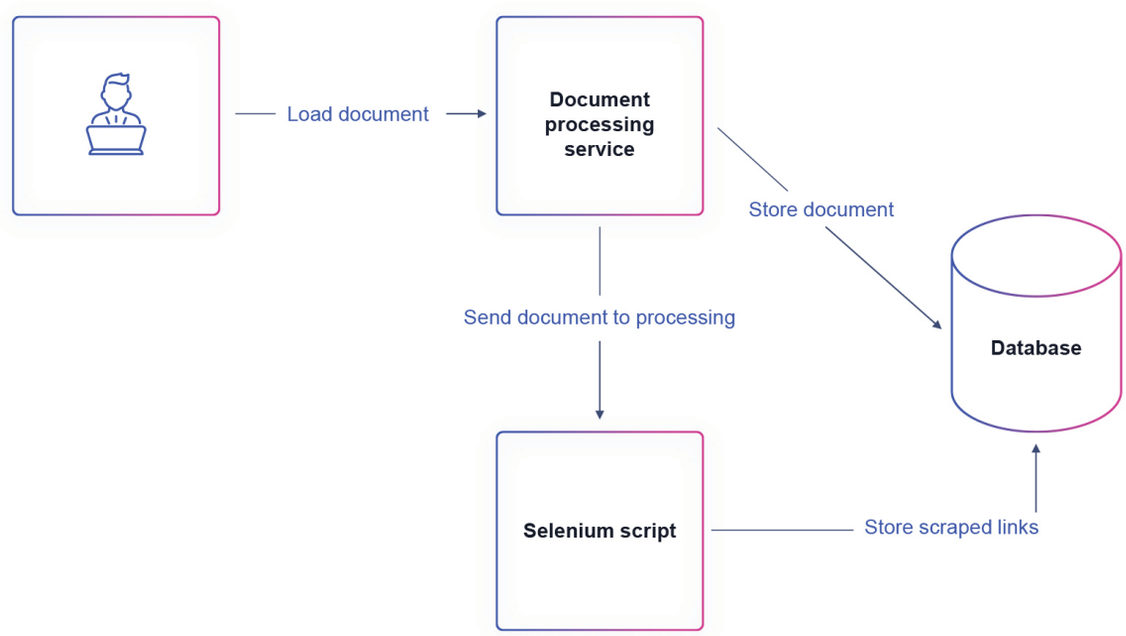
Comtrade 360 Collaboration with HPE

How We Enhanced Performance in Document Link Analysis by Streamlining Batching Process

For IT engineers, developing a system to extract, validate, and report links from thousands of documents can be challenging. The Comtrade 360 team, in collaboration with Hewlett Packard Enterprise (HPE), navigated through technological uncertainties to develop an agnostic document link gathering, validation, and reporting system tailored to their needs. A primary challenge for HPE, efficiently addressed by Comtrade 360, was to gather links from over 500,000 documents without spending excessive resources or impeding other functionalities of the application. Moreover, the system had to support multiple document types, including XML, HTML, and PDF, while being highly optimized.

To address these challenges for HPE, the Comtrade 360 team introduced a new step to gather all links from the document and store them in the database. However, validating multiple links from different files of different formats across multiple file storage locations in a performant manner was challenging. Existing solutions could not offer the required performance for the number of links seen within the database. Therefore, we needed to perform new types of activities to find an optimized solution.

Our team realized that there were no available solutions in the technological baseline that could reliably gather links from multiple different document types in a performant manner. Thus, we had to develop a highly optimized approach to scraping and validating links, ensuring that the process could be run on a weekly basis without incurring a large spend of resources or impeding other functionalities of the application.



To address these challenges, we devised a flexible system that could run in several different ways, on existing documentation, new documentation, and more. The diagram above shows the proposed high-level flow of the system architecture. Furthermore, in the event of failing links, our team ensured that the functionality of releasing the document and flagging it for later review was enabled. Overall, Comtrade 360 successfully developed a document link gathering, validation, and reporting system that met the project's stringent requirements while remaining flexible and efficient.

Solving bulk link collection bottleneck

The first challenge was to optimize the link gathering process by determining a robust methodology that would ensure speed and functionality while using a reasonable amount of resources. The team carried out this test semi-automatically by analyzing data on a sample of documents and finding ways to improve the time needed to extract links.

Initially, the team attempted to run it on all documents, but the code failed after a few hours and one of the documents failed. Therefore, the team realized they needed to split up the process into batches of documents to ensure that any failures would not compromise the entire process.
After running the process of collecting hyperlinks in batches of 1000 documents, the team found that they needed approximately 20 minutes for each batch, which was still a lot of time considering the large number of documents. They also discovered that processing documents with many subdocuments could take hours to be completely extracted. Hence, the team decided to split the list of documents into different batches consisting of subdocuments of the main documents.

This resulted in the Comtrade 360 team developing a new methodology to maximize document coverage in the presence of potential system failures. By splitting the document pool into batches, the team was able to optimize the link gathering process and maximize coverage while incurring only a small increase in processing time.

Batch number	Start time	End time	Document not available	OK	No link	Processing time	Sum per batch	Processing time
1	14:35:00	14:55:44	133	422	445	0:20:44	1000	0:00:01
2	14:55:44	15:20:54	365	159	476	0:25:10	1000	0:00:02
3	15:20:54	15:46:38	6	309	685	0:25:44	1000	0:00:02
4	15:46:38	16:16:22	0	266	734	0:29:44	1000	0:00:02
5	16:16:22	16:49:09	0	316	684	0:32:47	1000	0:00:02
6	16:49:09	17:09:36	0	292	708	0:20:27	1000	0:00:01
7	17:09:36	17:34:41	0	254	746	0:25:05	1000	0:00:02
8	17:34:41	17:56:32	0	185	815	0:21:51	1000	0:00:01
9	17:56:32	18:19:59	0	142	858	0:23:27	1000	0:00:01
10	18:19:59	18:49:50	0	160	840	0:29:51	1000	0:00:02
11	18:49:50	19:11:07	0	150	850	0:21:17	1000	0:00:01
12	19:11:07	19:31:44	0	212	788	0:20:37	1000	0:00:01
13	19:31:44	19:55:33	0	180	820	0:23:49	1000	0:00:01

Processing time for each Batch

Document size impacting processing times

This time, the Comtrade 360 team sought to understand the correlation between the number of links in a document and the processing time required for each document. The team wanted to ascertain if documents with more links took longer to process and if documents with no links took less time.
The team crafted this challenge to perform link extraction on two specific types of documents: HTML files with more links and XML files with fewer links, to examine the differences in processing times based on the number of links in each document.

After running the link extraction process on the Parts and Specifications documents, the team observed that documents with a higher number of links took longer to process than documents with fewer links. However, they also unearthed that the size of the document file influenced processing time, even if there were no links to extract. Through this challenge, the Comtrade 360 team developed a more profound comprehension of how the number of links and the size of the document file can impact processing time. These insights allowed the team to optimize their link extraction method and enhance their overall system performance.

After interpreting the results of the link extraction challenge on both types of documents, the team discerned that the presence or absence of links in a document did not significantly affect processing time. Documents with no links took approximately the same amount of time to process as those with links or that were not available. Therefore, the size of the document file is a critical factor in processing time, even in the absence of links.
This expanded understanding enabled Comtrade 360 to refine their link extraction process, significantly improving system performance for HPE.

Batch number	Start time	End time	Document not available	OK	No link	Processing time	Sum per batch	Processing time
1	14:35:00	14:55:44	133	422	445	0:20:44	1000	0:00:01
2	14:55:44	15:20:54	365	159	476	0:25:10	1000	0:00:02
3	15:20:54	15:46:38	6	309	685	0:25:44	1000	0:00:02
4	15:46:38	16:16:22	0	266	734	0:29:44	1000	0:00:02
5	16:16:22	16:49:09	0	316	684	0:32:47	1000	0:00:02
6	16:49:09	17:09:36	0	292	708	0:20:27	1000	0:00:01
7	17:09:36	17:34:41	0	254	746	0:25:05	1000	0:00:02
8	17:34:41	17:56:32	0	185	815	0:21:51	1000	0:00:01
9	17:56:32	18:19:59	0	142	858	0:23:27	1000	0:00:01
10	18:19:59	18:49:50	0	160	840	0:29:51	1000	0:00:02
11	18:49:50	19:11:07	0	150	850	0:21:17	1000	0:00:01
12	19:11:07	19:31:44	0	212	788	0:20:37	1000	0:00:01
13	19:31:44	19:55:33	0	180	820	0:23:49	1000	0:00:01

The results when link gathering has been run on HTML files with more links

Batch number	Start time	End time	OK	No link	Processing time	Sum per batch	Processing time
1	7:22:04	7:51:11	420	580	0:29:07	1000	0:00:02
2	7:51:11	8:14:20	368	632	0:23:09	1000	0:00:01
3	8:14:20	8:37:07	289	711	0:22:47	1000	0:00:01
4	8:37:07	8:56:37	337	663	0:19:30	1000	0:00:01
5	8:56:37	9:22:09	437	563	0:25:32	1000	0:00:02
6	9:22:09	9:53:27	358	642	0:31:18	1000	0:00:02
7	9:53:27	10:17:51	338	662	0:24:24	1000	0:00:01

The results when link gathering has been run on XML files with fewer links.

Preventing system jam with Subdocuments

The final obstacle was to establish a new method for dividing documents with a large number of subdocuments into batches. The team encountered an issue with the previous batch approach, where a single batch could get stuck in a loop, continually processing and blocking other batches in the queue. To solve this problem, the team needed to determine a new method for sizing batches that could accommodate the maximum number of subdocuments while preventing the looping phenomenon.

Two trials were conducted to find the optimal batch size. The first trial involved running the process of extracting URLs from documents in batches of 1000 documents, which had the best performance in terms of speed and resources. However, documents with more than 1000 subdocuments had issues processing the full list of subdocuments. The batch size was then increased to 4000 documents, and while the issues were reduced, documents with more than 4000 subdocuments still caused problems.

As we embarked on this challenge, we quickly realized the intricacies involved in handling documents containing a vast number of subdocuments. It became apparent that a more advanced approach to partitioning these documents into batches was necessary. Our goal, as the Comtrade 360 team, is to enhance the system's performance and deliver superior results to our clients by continuously exploring and refining our methods.

One potential solution we are contemplating is the storage of additional metadata, such as maintaining a list of IDs for subdocuments that have already been processed. We believe this could help us streamline our efforts and ensure a more efficient workflow.

For over 10 years, Comtrade 360 has been an invaluable partner to HPE. Firstly, supporting us as we separated HPE from HP and then, within our Services Division as we digital transformed how services are delivered to our customers. Comtrade 360's ability to deliver on accelerated targeted staffing, their low attrition, and expertise were critical success factors for us. Over a sustained period, their adaptability and extensive experience were critical to repeatedly delivering on time, high quality, innovative solutions that drove tangible value for our businesses.

Comtrade 360 consistently delivered on all commitments. Their unwavering dedication, collaborative spirit, and commitment to excellence have been essential in helping us achieve our strategic goals. They are a trusted partner.

Patrick Medley

Vice President HPE Hybrid Cloud
and Managing Director of HPE Galway

Conclusion

While our efforts have enabled us to optimize the process, we have recognized that working with batches containing 4000 documents or more could adversely affect the system's performance. Additionally, this approach may not entirely resolve the issue of becoming bogged down by large documents. As such, it is imperative to conduct further research to identify the optimal batch size that balances efficiency and performance.

By persisting in our pursuit to explore and fine-tune our methods, we are confident that we will successfully overcome these challenges. Our commitment to continual improvement at Comtrade 360 is driven by the goal of enhancing system performance, ensuring the satisfaction of our esteemed clients like HPE.

Comtrade
Three-Sixty

info@comtrade360.com
+1 617-546-7400
comtrade360.com