

Site-specific Travel Data Feeds

Introduction

The travel industry is ever changing with exa-bytes of data being updated on a daily (sometimes hourly) basis. For organizations operating in such a dynamic industry, with the whole world wide web acting as a database, access to accurate and structured data at the right time becomes of utmost importance to be able to take any strategic decisions to stay ahead of the competition. We have recently addressed one such issue. Read on to know more:

Who

A well-known budget hotel chain who wants to be updated about the hotel prices across India twice a day, to be able to strategize better in terms of pricing as well improve their offerings in general. With ever-growing competition, access to such data becomes a priority to stand out in the industry!

Crawl requirements:

The client had a specific set of requirements:

- Check-in Check out dates to be specified by them
- Such data was to be uploaded to a file sharing server and our crawlers were to pick them up at predefined times and process them. Possible situations:
 1. File available in the morning but not in the afternoon.
 2. File not available in the morning but available in the afternoon.
 3. File available both in the morning and afternoon.
 4. No file uploaded in the morning or afternoon.
- Fields for extraction were predefined and in a set order as specified by the client.
- Crawl frequency was twice in a day.
- Data crawled a particular day should be completed and delivered before 2359 hrs the same day.
- Hotels prices to be received in INR only.
- Inventory Crawls (Hotel details) to be carried twice a month to update.

This data was being used to strategically identify new hotels in the region to collaborate with them to increase the network across the country.

Target Sites and Approximate Monthly Volumes:

Most of the major OTAs (Online Travel Agents) in India with a volume of around 30 million records per month. Approximately **300,000 to 400,000 records per site per day** were delivered based on the files uploaded by the client.

Approaches Used to tackle the issues:

- We programmed the crawlers to search for the files on the sharing server at a pre-decided time and pick them up, if available. The crawler would check for files once in the morning and once in the afternoon.
- Additional scripts were written and additional resources were made use of to ensure data delivery happens before 2359 hrs of a particular day.
- The crawlers were programmed to detect if the currency was in INR. If not INR, the crawlers were programmed to change the currency to INR. This was important when servers from across the world were being used to crawl the data.
- The crawlers were also programmed to ensure not to hit the target servers very aggressively to avoid being blocked while, at the same time, ensuring that all the necessary data was captured before 2359 hrs.

Benefits delivered:

- Regular & flexible access to the required data as required. Since the crawler picks up data only if available, the client had the flexibility to upload the files only if required.
- Considering a dedicated team at the client side who were directly involved with this activity, a cost savings of about 23% was achieved by them.

With a bimonthly inventory crawl, they had access to updated data regularly (fortnightly) and were in a better position to increase their footprint across the country.

- With a low turnaround time, the data extracted could be used more effectively.
- Except for the initial onboarding period, the process was completely automated. Any disruption in service was also automatically updated to the support team to ensure that the crawls run in a smooth manner.

Result: While the setup was a challenging aspect of this use case, once implemented, the client has been successfully using the extracted data to improve their pricing strategy as well as their promotion strategy.