**Creation of real estate agent mailing database across US and Canada**

Client: Real estate construction company in US

Requirement: Client was looking for creating a database of real estate mailing lists for locating sales prospects in US/Canada and capturing highly targeted leads

Challenge:

The client wanted to build a database of realtors and real estate agents in US and Canada who were listed on various real estate franchises. The challenge was to collect more than 1 million records comprising real estate agent profile information ensuring low turnaround time, minimal duplicity and highest fill rate of contact information. This had to be done by gathering data from 8 websites (each having a different structure) with some websites requiring user to specifically search before listing the agents. The critical part of developing the database was to compare and identify the same agent across multiple websites and triangulate contact information.

Websites: REMAX(US & Canada), Corcoran, Royal lepage, Sotheby’s Realty, BHGRE, Century 21, Keller Williams, ERA Real Estate

Solution:

The key contact fields for the agents were identified across the websites including

Agency Name, Contact Name, Email, Address, City, State, Zip, Phone, Fax, Website, Specialization, Languages Known, Properties Price Range, Agent/Agency Summary, Listed from

8 data crawlers were setup for mass crawling that could collect the information related to the above fields in a structured format. The websites were crawled parallel to each other and 1M agent records were collect in a span of 1 week. These records were subsequently treated for duplication using agency and contact details and the final dataset was made available via an API.

Key Achievements:

1. All crawlers ran smoothly without any user involvement.
2. More than 500k records were uploaded ensuring 100% fill rate for email & name, 50% fill rate for phone/fax numbers and 10% fill rate for agent website.
3. Dynamic websites like Concoran (AJAX based loading of search results) and BHGRE, Royallepage (required search for getting agent listings) were crawled successfully
4. License/Joining information was included within ‘About Me’ section which required textual analysis to capture.
5. Agents were associated with multiple agencies across different time periods which was identified and such agent records were grouped.

**E-commerce product price comparison**

Client: Fortune 500 insurance carrier in US

Requirement: Collection of mobile selling price data from 17 sources across multiple geographies and collation of prices against a standard list of model names maintained by the client

Challenge: The client was looking for a data partner to collate mobile device price data from various countries so that their analytics team could focus on interesting analytics around determining competitive offer prices for consumers. The data extraction prcess was unstructured with some websites being crawled by a third party while the team manually gathered data from remaining web sources. This led to high efforts and data inconsistency issues with the third party unabke to ensure consistent record count and missing out on key high end device models. Even with the manual effort, structuring of the data in order to import it into their database was a challenge. Client was in need of clean data that could be uploaded into the database directly and utilized for creating a price benchmarking solution.

Solution:

The client provided the list of source websites to be crawled. The data had to be extracted on a daily basis meaning fresh information had to be supplied every day. Our team set up the crawlers to fetch data points from the sources and crawl all fields including device model, manufacturer, condition, storage, carrier, payment type, location. Since every website in the source list had a different structure and design, site specific customized crawlers were developed. Post completion of crawler development, the results of first crawl were shared with the client and once we received approval on data format and fill-rate, the crawlers started delivering data daily in CSV format ensuring consistent record count. Each crawled device model was automatically mapped against the standard model list provided by the client. We are delivering about 60k records on a daily basis.

**Key Achievements:**

* The client didn’t have to deal with any of the technical aspects in the process
* The setup was completed in just 1.5 months and the data flow was consistent since then
* We setup a robust monitoring system to make sure that no data was missed and ensure zero duplicity
* Some websites were complicated having dynamic result generation or rate limiting methods in place. These cases were effortlessly handled by our tech stack using methods like proxy rotation and user behavior replication
* We deliver 4X record count crawling 3 times the number of websites in 1/4th of cost charged by earlier data partner
* Low turnaround time of data improved the ability to market client’s services and capabilities
* Higher data quality levels led to improved pricing decision and more competitive pricing offered by client
* In case any data inconsistency issues were found, they were rectified within 2-3 hours post client notifying us the same.