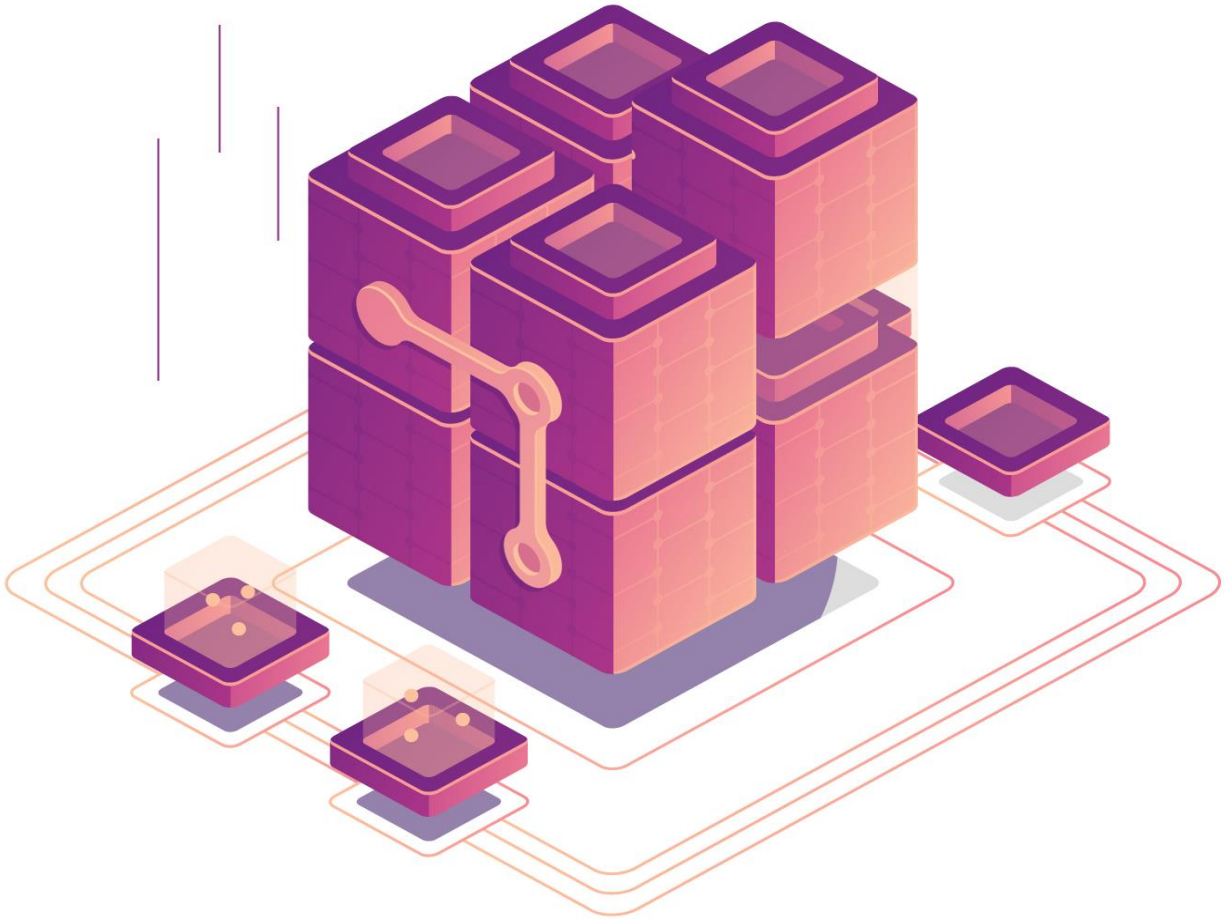
ProBytes software

Edsys

Education Institution Directory

school and college directory Web Scrapping



ProBytes software

www.probytes.net

Edsys Education institution directory – case Study

Requirement

The proposal was from Edsys – an educational software development company. They were planning to make an education institution directory which will have all the available details of schools and colleges across India. There were more than 50000+ pages to scrap from and more than 1.5 million data to be arranged, quality checked and to be transferred to a database.

Objective

After continuous contact with the client we came to a mutual conclusion that the prominence in this project was not for the data extraction infrastructure. In short, they were not interested in maintaining the crawlers or managing proxies. They were keen that they need reliable and clean data in a database format so that they can use it to list it in their website.

They were in hunt for Python developers to carry out the process from setting up servers, to check the data, convert them to the format they were hoping for and to transfer the data to a DB.

Our solution

We choose Scrapy to scrap data from websites. Scrapy is a robust Python based framework which requires less maintenance compared to other framework.

Scrapy helps to give extreme focus on data extraction using CSS selectors and choosing XPath expressions and not that much complicated compared to a conventional crawler. Apart from scrapy, multiple integrations were required to make the process efficient.

Even though we had the optimum tools, we dedicated 2 developers for the project and a team to assist them. Acquired data was cleansed and then arranged in requested order and was uploaded to client DB.

Hurdles in front

- There were around more than 50,000+ college websites to scrap from. The chances of architecture change in any one of this website was indeed a huge task.
- IP blocking was a major issue we faced or we expected to face we relied to advanced proxy servers to get nullify it
- There were websites with dynamic coding which were not crawler friendly
- Close monitoring was needed
- Duplication of data

- 1.5 million + Data was extracted! Since the quantity is humungous separate resources had to be dedicated in managing it.
- Even though every process was automated some required manual labour in sorting out and counter checking was required.

Resources

- 6-8 Months of time
- 2 full-time developers
- 2 Dedicated QA engineers

What we did?

- We were already told about the kind of data that was required from the website.
- Our team of engineers built a data extra extraction structure to feed data to the database
- Delivered first set of sample data to get the customer approval
- Along with the data extraction team a QA team was also set up to monitor the process and data quality
- A spider was set up and fed with the list/link of websites from which data has to be scraped
- Crawling process was automated since there were thousands of websites to crawl from
- Data storage mechanism and URL restriction was installed
- Acquired data has been added to a spreadsheet in DB
- Extraction was carried out daily
- Site specific crawl was chosen owing to the structural difference across websites

Benefits

- Our team handled all the technical aspects of the data extraction
- Client started receiving data within days of setting up the system
- Huge amount of data was handled efficiently
- A special side program was run by another system to track changes

Feel free to contact us

➔ *Pune*

➔ *Kalas road, Vishrantwadi, Pune, Maharashtra-411015,*

➔ *+91 81 1386 1000*

➔ *info@probytes.net*