zyte

Architecting a Scalable Web Scraping Solution

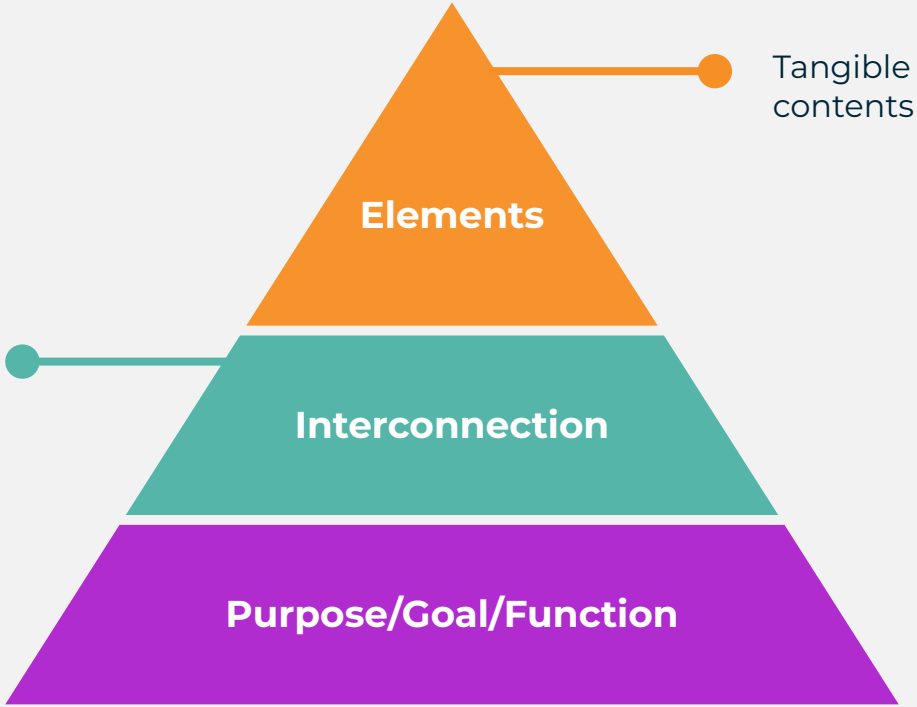
Speaker

Neha Setia Nagpal

Developer Advocate at Zyte

“ My role is to serve the web scraping developer’s community. ”

A System



Tangible and Intangible contents that work together.

How parts are interconnected.

8 steps to architecting a scalable web scraping solution.

Step 1: Clarify The Goal - Why?

Step 1: Clarify The Goal - Why?

The Purpose is to build ..... use case

that will help you to solve ..... problem

for which i need to extract ..... define the data schema to solve the problem

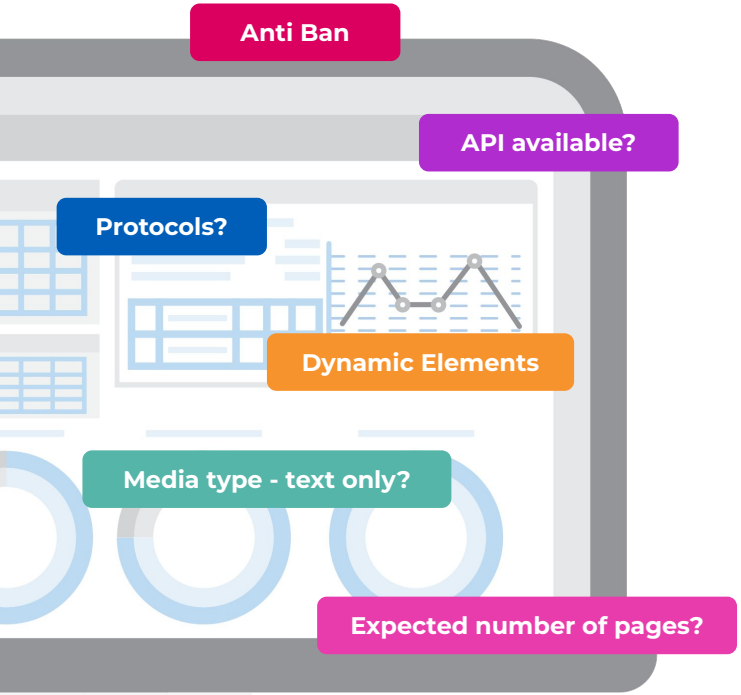
of ..... approx size - define how much

from ..... define Target Websites

In ..... time and

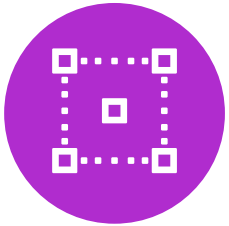
refresh in ..... days.

Step 2: Analyze The Website



1. Are there any APIs available?
2. Does the crawl involves dealing with dynamic elements on the web pages?
3. How powerful the anti-bot mechanisms are on the website?
4. Is it a crawler for text-only web pages only? Or should we fetch and store other types of media type as well?
5. What is the expected number of pages we will crawl?
6. What is the average size of the web page to be downloaded?

Step 3: Prioritize The Project Attributes

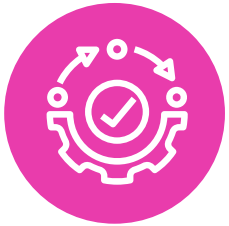


Scalability

Distributed Crawling.

Extensibility

Single Responsibility Principle.



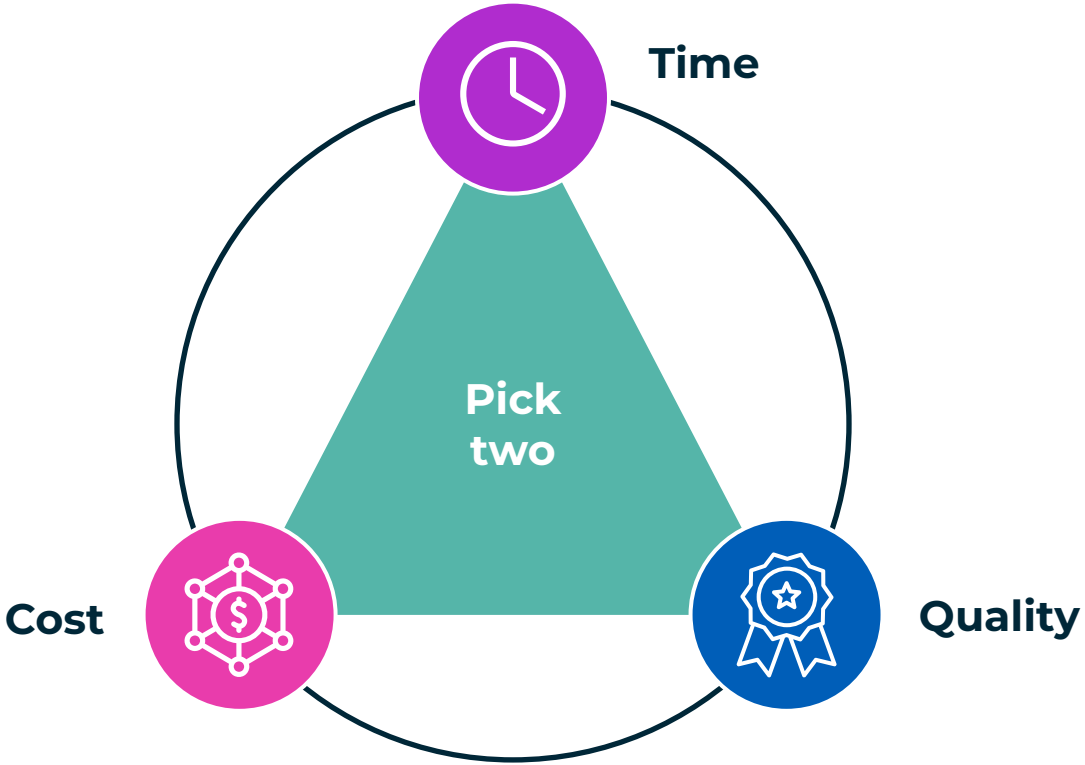
Availability

Low Latency and resilient monitoring systems.

Velocity

Strategize on the basis of the project needs.

Step 4: **Highlight** The Constraints



Step 5: Design The Crawl

How to crawl? Which algorithm to use for scanning the pages?
BFS/DFS or Path ascending crawls

How do we prevent requesting the same URL?

- Same Crawling Session- Scrapy handles it by default.
- Different Crawling Session- Frontier Service

How to make the crawler polite?
Fewer Requests

Validate the design for Single responsibility principle- look for “and”

Step 6: Plan The Quality Assurance Process

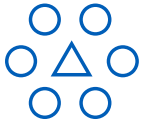
We have determined 6 data quality dimensions

Accuracy

Validity

Timeliness

Completeness

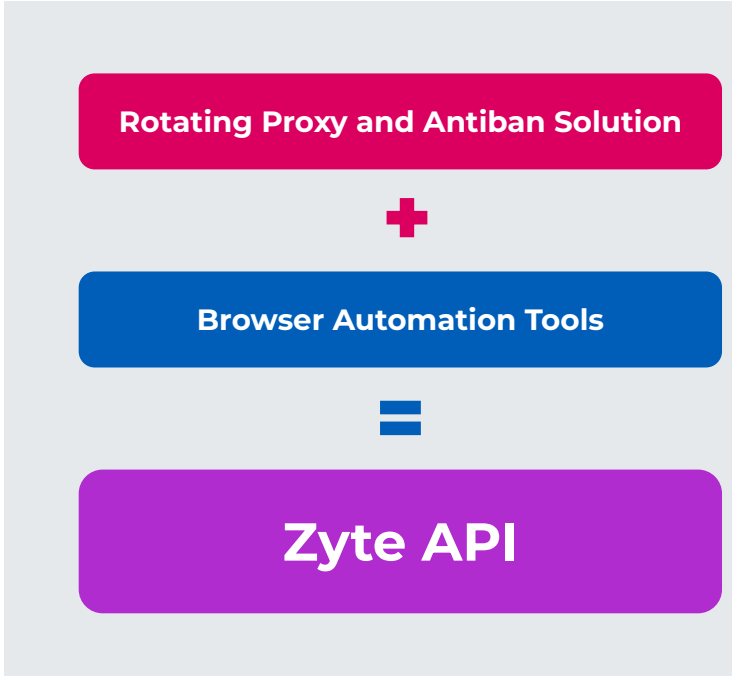


Uniqueness

Consistency

Step 7: Choose The Tech Stack

1. Base Technology/Framework- **Scrapy**
2. Deploy and Maintain Spiders- **Scrapy Cloud**
3. Rotating Proxy and Antiban Solution-Zyte API - **including Smart Proxy Manager/Smart Browser.**
4. Browser Automation Tools- **Headless Browser libraries**
5. Maintenance and Monitoring- **Spidermon and Matt**

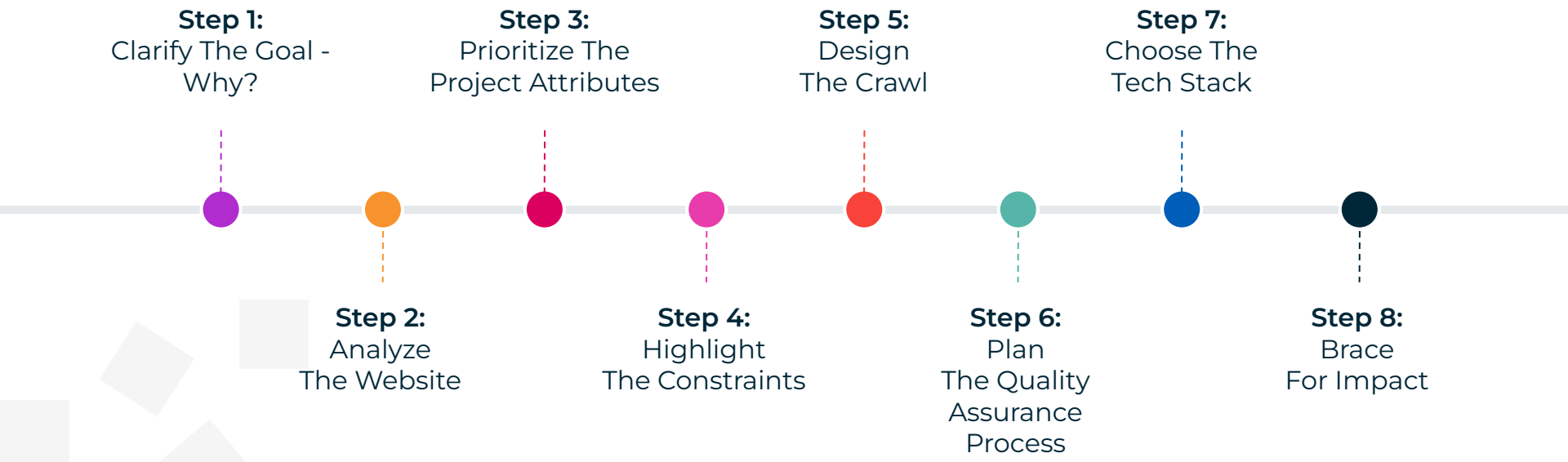


Step 8: Brace For Impact

Brace the uncertainty by preparing for the challenges in maintaining a large scale web scraping project by

Low latency Monitoring Tools like spidermon.

To Summarize



Thank you

zyte