zyte

Data Quality Assurance

A sneak peak inside Zyte quality assurance system

How Zyte ensures 99% data accuracy and coverage for our clients.

Introduction

When it comes to extracting data from the web, data quality is your #1 priority.

Without a consistent and high quality output of web data from your spiders, your web scraping projects are of little value and can even be detrimental to your business if they are consuming resources without delivering meaningful results.

Obviously, the first step to ensuring high levels of accuracy and coverage when scraping the web is developing highly robust spiders to extract the web data you need. However, no matter how good your spiders are their performance will inevitably degrade over time as websites are continuously making changes to their underlying structure that can break spiders.

As a result, a robust quality assurance process is critical if you want to ensure your spiders consistently deliver high quality data.

In this guide we're going to talk about data quality assurance for web scrapers, and give you a sneak peak into some of the tools and techniques Zyte has developed to ensure we can deliver our clients data with 99% accuracy and coverage.

These QA processes enable us to verify the quality of our clients' data at scale, and confidently give all our clients data quality and coverage guarantees.

QA System Overview

At a high level, the goal of a webscraping QA system is to assess the quality/ correctness of your data along with the coverage of the data you have scraped.

Data Quality and Correctness

- Verify that the correct data has been scraped (fields scraped are taken from the correct page elements).
- The field names match the intended field names stipulated by you.
- Where applicable, the data scraped has been postprocessed and presented in the format requested by you during the requirement collection phase (e.g. formatting, added/stripped characters, etc.).

Coverage

- **Item coverage** - verify that the all available items have been scraped (items are the individual products, articles, property listings, etc.).
- **Field coverage** - verify that all the available fields for each item have been scraped.

When scraping at any reasonable scale, to achieve these goals of obtaining high data quality and coverage from your web scraping it is very important that you incorporate a multi-layer QA process into your web scraping projects which combines both automatic and manual tests to verify data quality.

At Zyte we apply a **four-layer QA process** to our projects to ensure we are able to give our clients the highest quality data and coverage they require.

- **Layer 1** - Pipelines
- **Layer 2** - Spidermon
- **Layer 3** - Manually-Executed Automated QA
- **Layer 4** - Manual/Visual QA

The rest of this guide is dedicated to describing each layer of this QA process so you can replicate it with your own web scraping projects.

QA Layer 1: **Pipelines**

Item pipelines are rule-based Scrapy constructs built into the extraction spider that are designed to process data as it is being scraped.

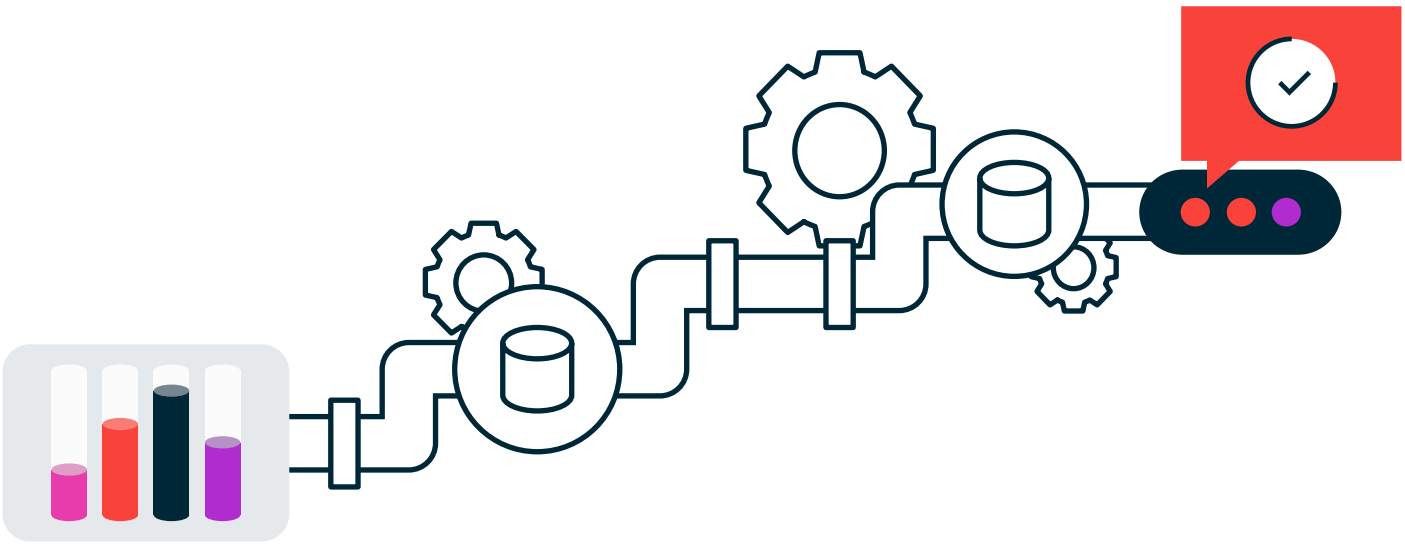
Imagine pipelines like an oil refinery. It starts with crude oil and it goes through a series of processing such as: desalter, hydrocracker, isomerization, etc. to turn that raw crude oil into usable products like gasoline or diesel. In the case of quality assurance, these pipelines are often used to:

- **Cleanse and normalise the data** (remove non-printable characters, convert to unicode, etc.).
- **Validate scraped data** (checking that the items contain pre-determined key fields, and dropping any that don't, if so desired).
- **Check for duplicates** (and dropping them if so desired).
- **Store the scraped items** in a database.

At Zyte, our developers make extensive use of pipelines to clean and validate the data as it is being scraped.

They act as a very robust first stage filter to remove any data quality issues before the data even reaches your databases.

Pipelines are designed and developed as part of the spider development process and run in real-time as the data is being scraped.



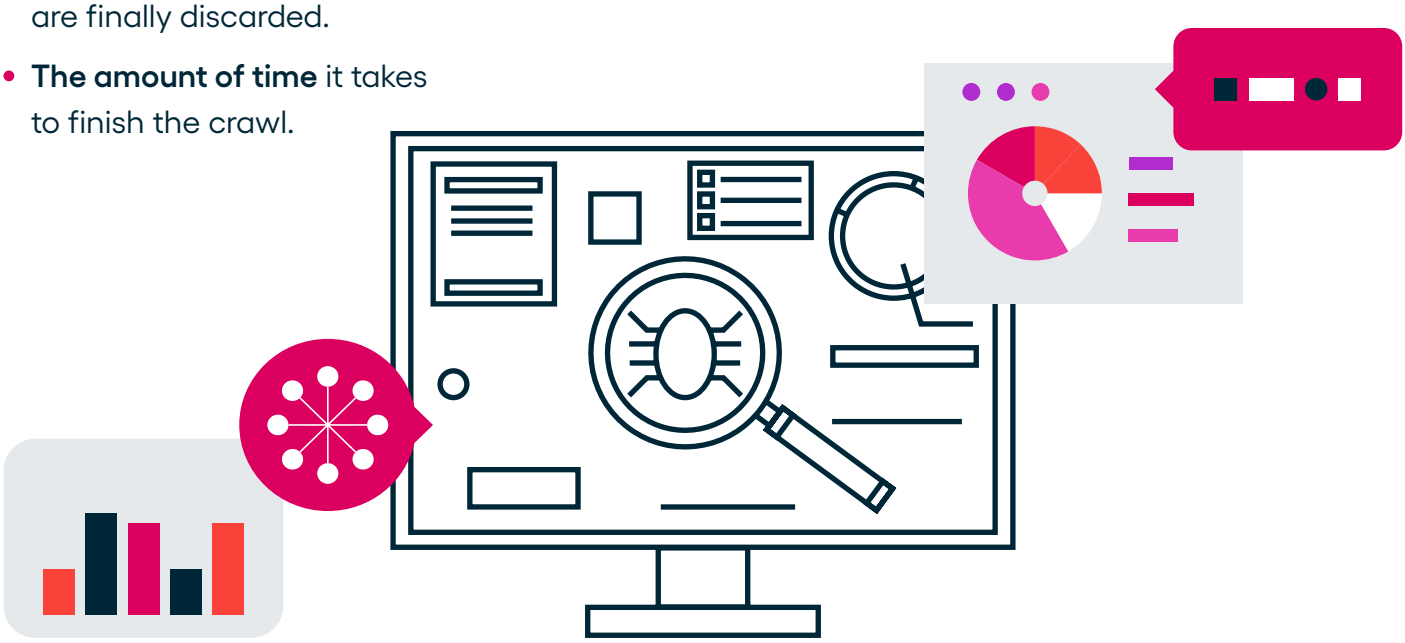
QA Layer 2: **Spidermon**

Spidermon is a framework for monitoring for Scrapy spiders. Spidermon can verify the information provided by Scrapy through stats functionality, which can also include custom stats to handle specific requirements depending on the project.

Additionally, when Spidermon is used to perform data validation with Scrapy (on top of applying basic monitors for other stats), it also uses these stats to pass information to the final monitors. The following are some of the usual checks ("monitors") that Spidermon is configured to perform:

- **The amount of items** extracted by the spider.
- **The amount of successful responses** received by the spider.
- **The amount of failed responses** (server-side errors, network errors, proxy errors, etc.).
- **The amount of requests** that reach the maximum amount of retries and are finally discarded.
- **The amount of time** it takes to finish the crawl.
- **The amount of errors in the log** (spider errors, generic errors detected by Scrapy, etc.).
- **The amount of bans.**
- **The job outcome** (if it finishes without major issues or if it is closed prematurely because it detects too many bans, for example).
- **The amount of items** with validation errors (missing required fields, incorrect format, values that don't match a specific regular expression, strings that are too long/short, numeric values that are too high/low, etc.).

Again, Spidermon is a development activity, however, this time there is often QA input, with QA Engineers regularly defining the validation to be performed and often implementing it themselves.



QA Layer 3: **Automatic QA**

The third component of Zyte's QA process are Python-based automated tests our dedicated QA team develops and executes.

Once the output of a spider has passed layers 1 and 2, the dataset is then sent to QA for final validation prior to delivering it to the customer.

During this stage, datasets are analysed to identify any potential sources of data corruption. If any issues are found, these are then manually inspected by the QA engineer.

Each dataset is validated using the following checks:

Accuracy

Using an inferred schema (compliant with the JSON Schema standard), the tests validate all fields to ensure that the correct field naming, value type and field formatting is being used, as defined by the project requirements.

Errors and warnings

Each dataset is checked to see if there were any errors or warnings triggered during the course of the data extraction.

Spider Execution Outcome

If the final status of the spider is anything other than "finished", this can be a sign that there might be a data inaccuracy or coverage issue.

Completeness (item coverage)

Using estimations of expected item count, the tests will determine the completeness of the dataset by comparing the number of scraped items with the estimate.

Historical Spider Execution Comparison

Compare the current dataset with previous test jobs and/or production jobs to verify that there are no abnormal changes in data accuracy or coverage:

- Compared to the the previous jobs, are there any items or fields missing from the current job?
- Are there any abnormal changes in the values scraped in particular fields? e.g. For price related fields, for example, warn if the price for the same product exhibits a greater than 50%(configurable) change between jobs.
- Are there any abnormal changes in the field coverage? Warn if the coverage per field decreases by more than a specified threshold.
- Are there any abnormal changes in the coverage per category? Warn if the coverage per category decreases by more than a specified threshold.

Rules-Based Field Checks

- Do the scraped items contain any unintentionally scraped data e.g.any HTML, CSS, or JavaScript.
- Are there are superfluous items i.e. duplicated items in the scraped data (items with exact same values - name, url, etc).
- Were any of the mandatory fields left empty?
- Did the spider execution conform to expectations?

QA Layer 4: **Manual QA**

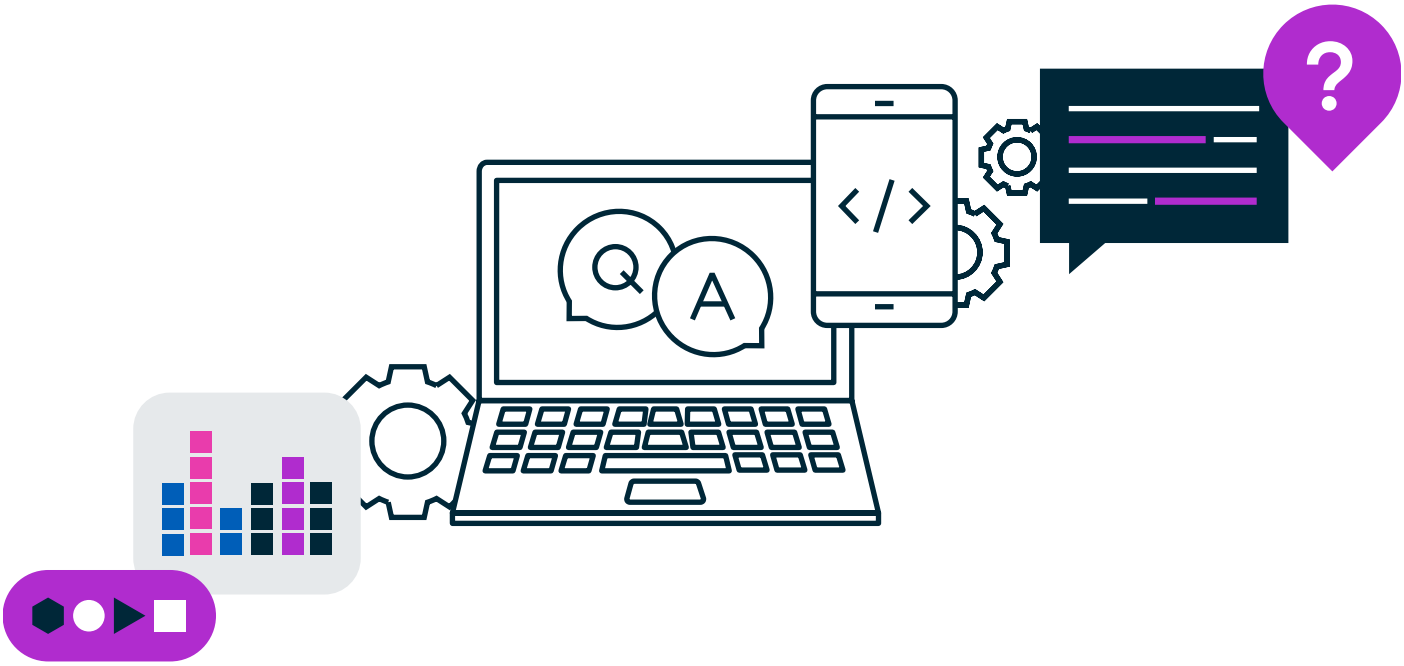
The final component of Zyte's QA process is the manual visual inspection of the data. As described in the previous section, any issues flagged by the automated QA tests are always manually inspected by one of our QA engineers. However, the assigned QA engineer carries out additional manual spot checks of the data to validate that the automated QA steps haven't missed any data issues.

Although automated QA processes are very powerful and greatly reduce the workload of validating the accuracy of web scraped data, manual visual spot checking of the data remains hugely valuable. It serves as a way of identifying data quality issues that automated QA isn't able to detect with perfect accuracy e.g. semantics.

A sample of the scraped data is then visually compared to the data on the web sites, with a view to answering questions such as:

- Did we scrape **the right thing?**
- Did we **fail to scrape** anything we should have scraped?
- Did we scrape **the right thing from the correct category?**
- Are the warnings generated by the automated tests **false alarms, or legitimate issues?**

Only after passing through all four of these layers is the dataset then delivered to the client.



Conclusion

As we have seen, there can be quite a bit of work involved in building a data quality assurance process that can give you a consistent feed of high quality web data.

Using all (or parts) of the 4 layer system outlined above will greatly improve the quality of data you can achieve from your web scraping projects.

It is the result of this exact QA process that Zyte is able to confidently give all our clients accuracy and coverage guarantees, and ensure all projects achieve a accuracy/coverage of 95-99%.

For those of you who are interested in scraping the web at scale but are wrestling with the decision of whether or not you should build up a dedicated web scraping team in-house or outsource it to a dedicated web scraping firm then be sure to check out our other guide, Enterprise Web Scraping: A Guide to Scraping the Web at Scale.

At Zyte we specialize in turning unstructured web data into structured data. If you would like to learn more about how you can use web scraped data in your business then feel free to contact our Sales team, who will talk you through the services we offer startups right through to Fortune 100 companies.

At Zyte we turn websites into data with industry leading technology and services.

Our solutions include:

- **Data Extraction Service**
  Let our web scraping experts build and manage the bespoke data extraction solution for your business needs.
- **Smart Proxy Manager (formerly Crawlera)**
  Forget about proxy lists. We manage hundreds of thousands of proxies, so you don't have to.
- **Automatic Extraction powered by AI**
  Instantly access accurate web data through our user-friendly interface or various Extraction APIs and save time getting the data you need.
- **Data extraction platform**
  Access developer tools, data extraction APIs and documentation, built and maintained by our world-leading team of over 100 extraction experts.

zyte

It's yours. The web data you need.

Access clean, valuable data with web scraping services that drive your business forward.

Talk to us

www.zyte.com

Cuil Greine House, Ballincollig Commercial Park Link Road, Ballincollig / Co. Cork, Ireland