



5 questions to ask

about where your data comes from.

A discussion guide

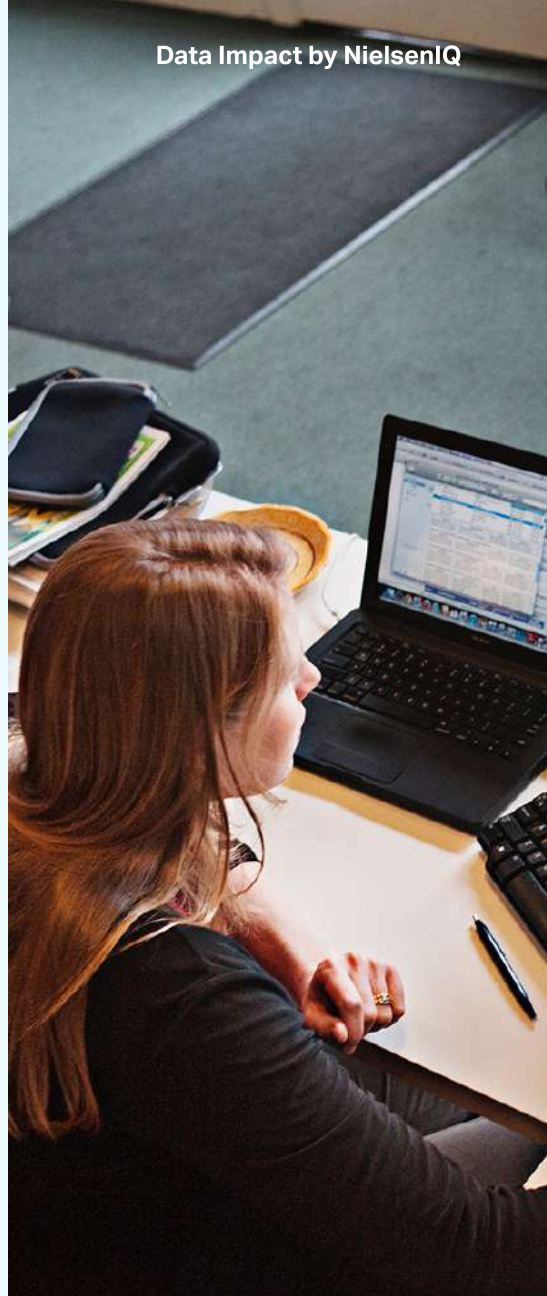
Collecting data is the first step in the process of creating a digital shelf performance analysis.

It's also known as **data acquisition**, or **harvesting**, **gathering**, or **scraping**, and it simply means importing data from websites to files. It's done differently by various service providers, manufacturers and third party outsourcing companies. Since so much depends on it – like accuracy and actionability--we've prepared five crucial questions you can ask about how data is processed. If the answers you get to these questions aren't satisfactory, you know your data isn't either.

💡 **Faulty data produces many negative consequences including:**

- inability to understand performance
- losing the trust of stakeholders, retailers and partners
- the inability to leverage data and thus optimize performance

For the CPG ecommerce industry, a primary source of data is retailer websites. Each retailer can have hundreds or thousands of locations with varying assortments, prices, search results, etc. This information changes constantly. With that in mind, let's get to the five questions that make all the difference in data collection.



5 questions that make all the difference in data collection:

1 Is your data collected by your service provider or do they outsource to a third party?

2 Is your data taken from a sample of online locations or is it exhaustive?

3 Is the data being collected from the URLs of retailer sites, a selection of shelves or the entire category?

4 Is all of your competitor's data being collected?

5 What data quality assurance do you have?

Question #1

Is your data collected by your service provider or do they outsource to a third party?



Context:

Scraping retailer sites is an increasingly complex task. More locations are coming online, more KPIs are being tracked and security measures are being enhanced. Some data providers outsource data collection, a practice known as “scraping as a service”. Here’s a comparison of key aspects of both:

In-house vs. external data collection:

	Service provider scrapes in-house	Service provider outsources scraping
Data quality	✔ High: purpose-built tools for different retailers that work in varying regions. Recurrent data quality monitoring	✗ Low: limited spider monitoring and data checks. One size fits all approach to websites
Coverage	✔ Extensive: complex retailers, apps & websites	✗ Often blocked on sophisticated sites
Responsiveness	✔ Ability to prioritize products & categories	✗ No ability to prioritize or direct scraping
Cost	✗ Higher	✔ Usually lower

Conclusion:

There are major benefits to hiring a provider that does their own scraping, from KPI accuracy to problem resolution to actionability. Simply asking a service provider if they outsource their data collection will indicate how reliable the final data analyses will be.

Scraping is complex and requires an experienced team. For example, Data Impact has approximately 50 in-house engineers tasked with collection.

Question #2

Is your data taken from a sample of online locations or is it exhaustive?

Context:

The majority of data service providers collect data from only a selection of retailer sites. This data is dangerously non-representative.

Even with **70% of locations** scraped, the accuracy rate of sample data remains **very low**.

Sample data is simply false data.
Only census or geo-complete data is reliable.

Sampling vs. exhaustive data collection:

	All locations	Sample of locations
Data accuracy	✔ Complete and reliable	✗ False analysis of key sales drivers (OOS, price...)
Actionability	✔ Full visibility on problem SKUs and locations	✗ No visibility on where issues occur
Usefulness to retailer	✔ Granularity brings value-added discussions retailer benefits from	✗ No trust in data and no insight to offer
Business performance	✔ Performance analytics . Possible to cross-reference sales with KPIs/business levers	✗ Impossible to link causal data to sales. E.g. OOS in a location affects sales in same location only

Conclusion:

All locations need to be scraped. It's the only way to get an accurate view on all KPIs. Geo-complete data is crucial. Local tastes, local supply issues, local pricing, local visibility variation—it must be taken into account for a CPG manufacturer to improve its performance.

For a sample to be representative, **all retailer locations** would have to be scraped for at least a year to render the sample **'intelligent'**. This is not currently provided by other providers due to their limited tech stack and inability to track all locations.



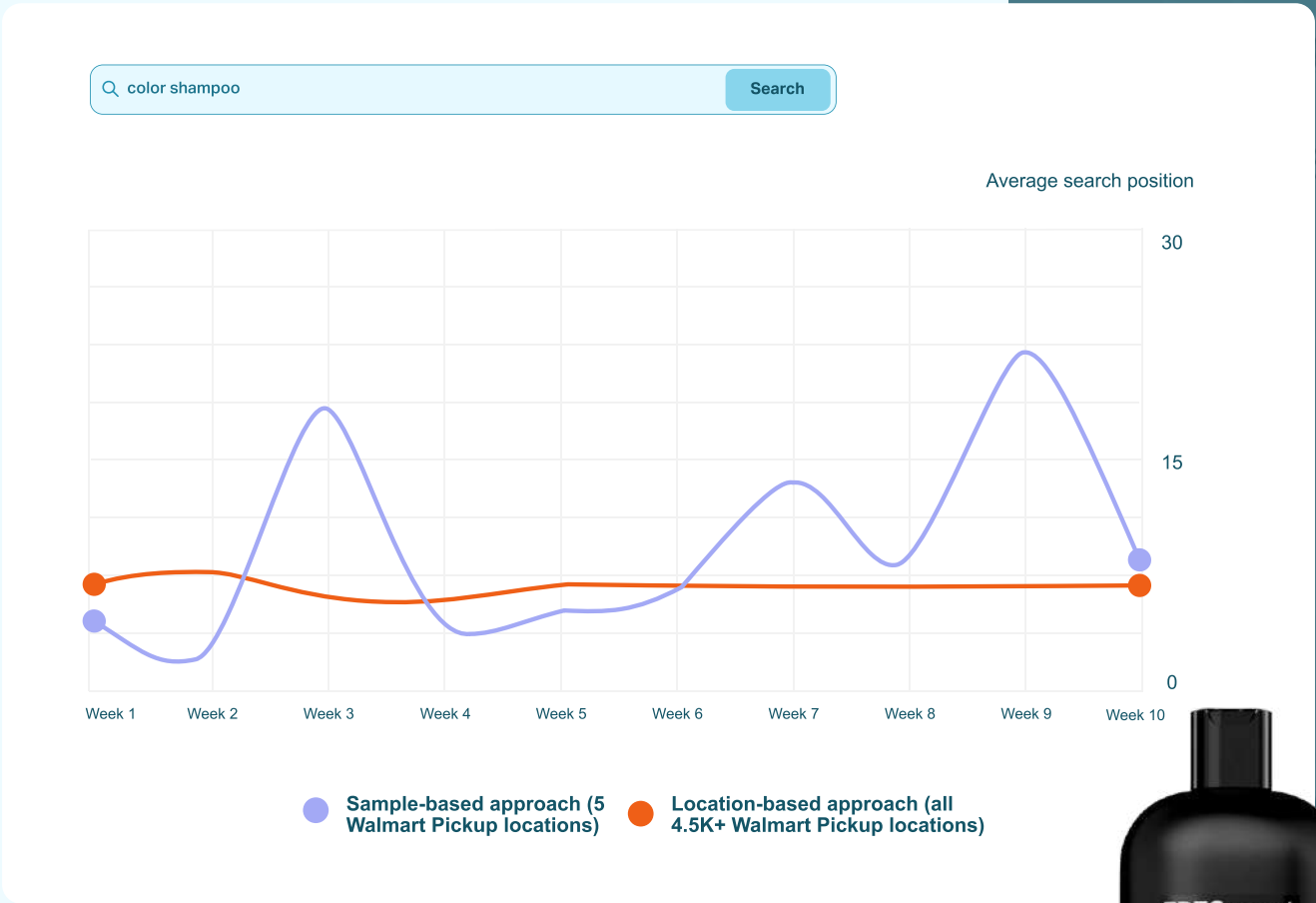
Data accuracy rate.
Sample dataset of
50 locations at
Tesco UK.

*Source: Data Impact by NielsenIQ
Note: Results are based on analysing OOS rates for a soda brand across 10,000 samples of random store combinations from February 8 - 14 2021 at online Tesco stores.

The problems a data sample creates:

Some service providers claim they take 'intelligent samples'. This is still problematic. The following graph illustrates how gathering from a sample makes search ranking appear much more volatile than it actually is. A CPG following the sampled data would spend lots of time and money fruitlessly trying to optimize for search.

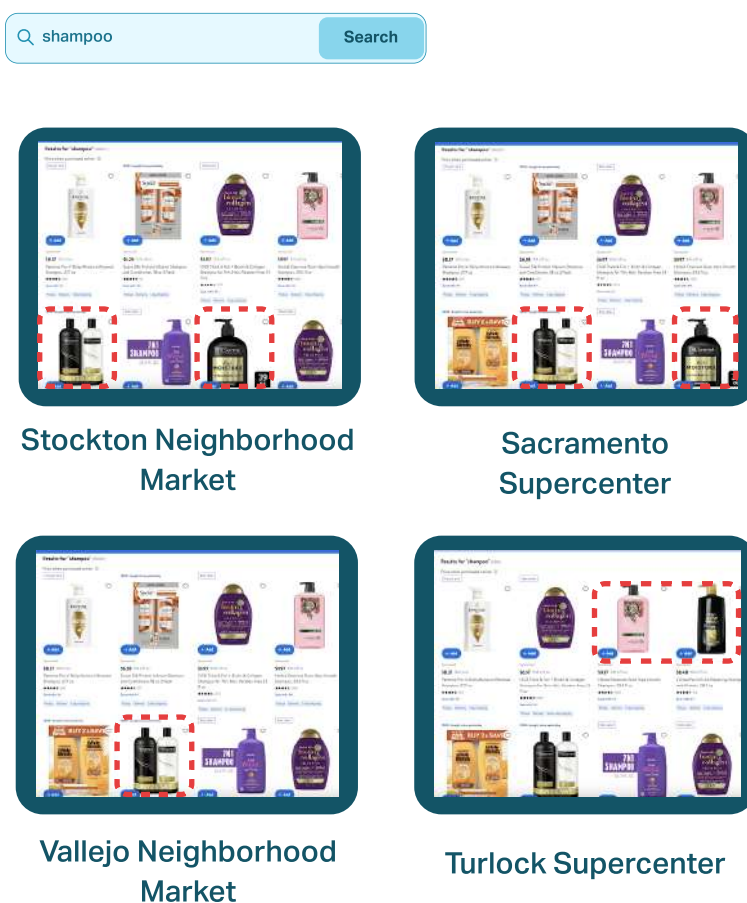
Evolution of the average search position of a TRESemmé SKU at Walmart



Sample data often misses search variations

To illustrate the variations in search results at different locations, notice how the search results for the keyword 'shampoo' differ at four different Walmart Pickup locations, all within a small radius in California.

Search results for 'shampoo' keyword at four Walmart Pickup locations in CA.



One last point on sample vs. complete data collection: cross-referencing digital shelf data with other datasets can only really be done if all locations are scraped. Take performance analysis, which pairs sales data with the digital shelf causal data to identify the digital shelf levers that impact sales--this can only be done when both datasets are equally granular and robust.

If out of stock data is only scraped from a sample of locations, it's not possible to get **an accurate analysis of how OOS affects sales.**

Question #3

Is the data being collected from the URLs of retailer sites, a selection of shelves or the entire category?



Context:

Most service providers and outsourcers either crawl retailer’s sites using the retailer websites of the product page, or just crawl select shelves. Crawling is the process of using an internet bot or spider to visit retailer websites and collect data.

URL vs. partial vs. complete category data collection:

	Entire category	Selection of shelves	URLs
Data maintenance	✔ No maintenance required	⚠ List of category names required. Categories change and products moved so must be updated	✗ High- URL of each new product needs to be constantly updated
Data quality	✔ High- all SKUs and locations collected	⚠ Often not every product in a category or aisle is collected, resulting in incomplete data	✗ Low- URLs change, products go OOS, new products get added to existing pages, promoted SKUs get moved
Reliability	✔ High	⚠ Partial-product info on additional promo pages, etc, missed	✗ Low- URLs change, products go OOS, new products get added to existing pages, promoted SKUs get moved
Data history	✔ High- Full history from service provider agreement onwards	✗ No history	✗ None- no evolution of pricing or distribution, etc

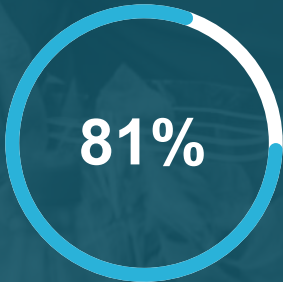
Conclusion:

The retailer’s entire site needs to be crawled, product by product, using SKU references such as UPC and EAN to identify each item.

When a service provider crawls the entire taxonomy of a retailer’s site **there are significant benefits to a CPG:**

- they don't need to maintain a SKU list
- there is no need to send product list links to their service providers
- they can automatically keep track of and manage their portfolio while it's visible to shoppers online

The service provider will be perfectly up to date on all the site information. For example, their spiders will automatically know when products are delisted or when new products are added.



of manufacturers don’t trust their data enough to take it into retailer negotiations*

*Source: DIBYNIQ webinar poll, 2022

Question #4

Benchmarking:

Is all of your competitor's
data being collected?



Context:

Most service providers only crawl top SKUs and top competitors, or some shelves. Not collecting all SKUs prevents CPGs from accessing a full category read and identifying market trends in their categories. It also limits their ability to benchmark their performance. What's more, it requires work in the form of data maintenance. Let's look at each briefly:

Consequences of not tracking competitors:

Complete suite of KPIs

When only a selection of SKUs are being scraped,

- benchmarking is impossible. There's no way to know how all competitors are performing and how the category is performing as a whole.

No market insight

Without tracking all products in a category, several blind spots occur. SKUs that are delisted may still be tracked, or, SKUs that are visible to shoppers may not be tracked. Basically, competitor innovation and market share evolution awareness is poor without this kind of benchmarking.

Laborious, ongoing data maintenance

- When all competitor product data is collected, no data maintenance is required. Product lists don't have to be updated and URLs don't have to be monitored for outdated information.

A limited view, while costing less as a service, often **costs more in the end** because market trends, challenger brand innovations and shopper behaviors are missed. There are usually other hidden costs, too, which are impossible to anticipate.

Conclusion:

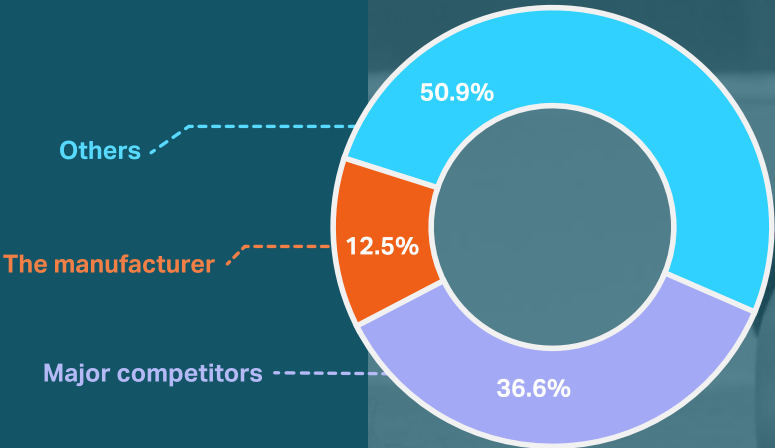
Category-wide collection is crucial and products need to be set to specific hierarchies. **This has two benefits:**

1. It's the best way to track SKUs as the digital shelf information attached to each of them changes.
2. It enables comparison of data across KPIs, enabling cross-referencing sales data with causal digital shelf data.



Competitor view

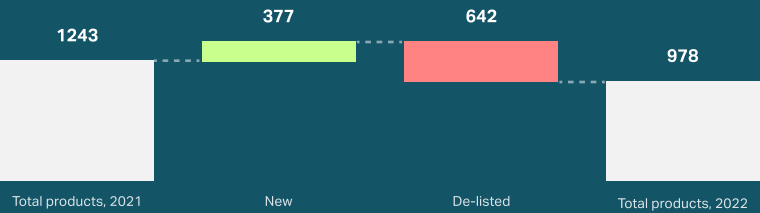
Here, the competitor view reveals that small brands collectively dominate the share of assortment. This provides yet another reason for tracking all competitors/ the entire category.



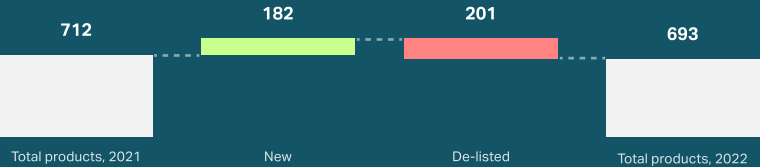
Assortment changes in the chocolate category

The following graph provides an example in the chocolate category. A full-category view reveals significant numbers of new and delisted SKUs. Monitoring this activity gives companies insight into product development and a view on retailer category changes and competitor innovation strategies.

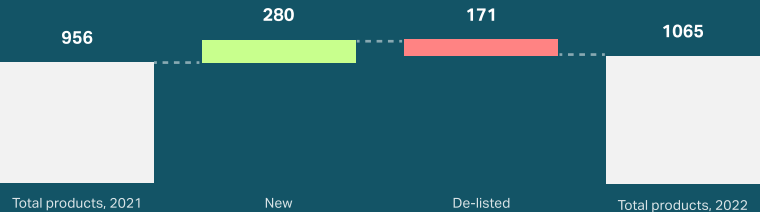
Walmart



Target



Kroger



Question #5

What data quality assurance do you have?



Context:

Ensuring collected data is of good quality is critical and most data providers rely just on AI, which is insufficient.

Reality:

Reliable scraping is a three phase process that includes human verification:

Tools

- A customized tech stack for each retailer
- Granular store level and aggregated view
- Spiders deployed quickly, maintained to avoid blockage
- Proprietary service provider tech

People

- A dedicated team that
 1. Monitors the tools
 2. Prioritizes tasks
 3. Assigns tasks
 4. Takes decisions
- Expertise to detect and know how to resolve blockages, security issues, etc.
- Real time monitoring of spiders that crawl retailer sites

Targeting

- All spiders for all stores are tested for errors and total product scraping
- A method in place to address different issues
- Data from each location is analyzed to confirm the number of products and site changes
- Automatic layer of alerts if issues occur

A reliable scraping process can only be instituted if a service provider **gathers data themselves.**

When there are problems with data acquisition (which occurs regularly due to product changes, new site tech, etc.) the problems will either not be addressed by third party vendors or won't be addressed quickly. CPGs have no control over the 'scraping as a service' providers their own service providers outsource to.

Conclusion

So those are the five essential questions that determine whether or not your data collection is reliable. Given that data acquisition is followed by data cleaning, maintenance and finally, analysis, this first step in the process is critical. Once a robust data harvesting operation is in place, a secondary system must be enabled for security. If a problem occurs, interruptions in data reporting—such as no availability data for days or weeks—can be costly for CPGs.

Put these five questions to your digital shelf service provider.

We know data collection is a complex, tech-heavy process. The purpose of this discussion guide is to lead you through a fruitful discussion with your data partner.

If you have any questions, please don't hesitate to contact us, we're here to offer support.

Let's talk! →



contact@dataimpact.io



<https://dataimpact.io>

