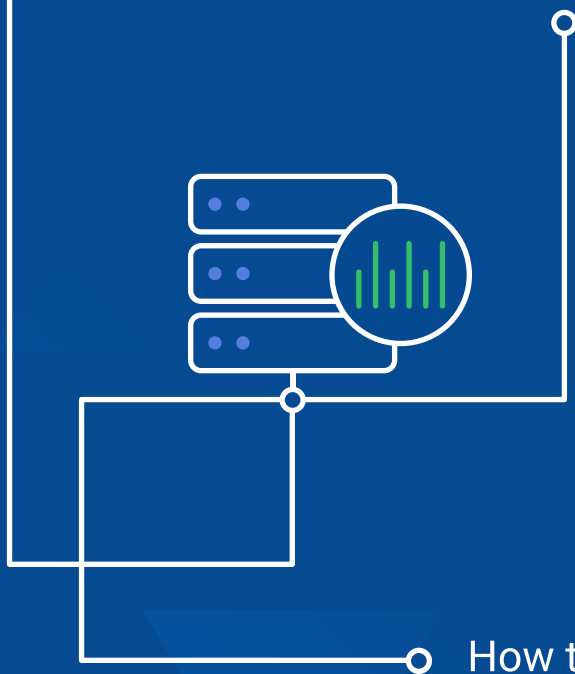
competera | Survey

Data Quality



How the Quality of Competitive Data Affects the Key Business Indicators of a Retailer

Table of Contents:

Description of the study

The problem of data quality

The amount of information that retailers have to collect about competitors from various sources only competes with the amount of data that is generated by the online store itself. Today, working with lots of data is the norm if you want to work effectively. At the same time, the excess of information that we get and its low quality, often causes retailers not to make the right conclusions as well as not make quick and informed decisions.

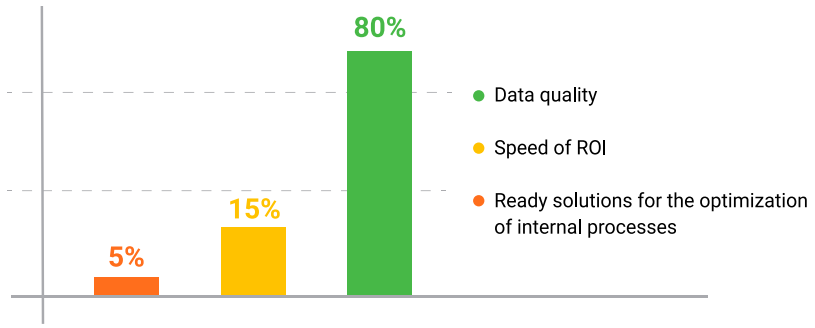
It's the quality of data that determines the economic effect for the business: high-quality, clean and timely data is the basis for having quality analytics and as the retailer, make effective strategic and tactical decisions. The quality of data directly affects pricing, since it underlies the rational revaluation:

- Collection and integration of data
- Data processing and configuration
- Visualization of reports: regularities, deviations, anomalies
- Decision making and change management
- Measuring the effectiveness of decisions
- Automation of decision-making and change management
- Business optimization strategy
- A mechanism that predicts the results of decisions made and the implemented strategy

In this sequence, the first phase, which is the collection and consolidation of data, is the foundation. It affects the final result. Poor quality, also known as "raw materials", lead to erroneous decisions, lost profits and a deterioration in your market position.

Today, 200 retailers from 18 countries use Competera products. According to Competera's internal survey, 80% of the company's customers consider the quality of a competitor's data to be the key factor that guides them when they are choosing a partner for the introduction, development and improvement of pricing processes.

Parameters that influence the selection of the contractor based on the delivery of data



Purpose of the study

The main purpose of this study is to standardize the notion of “quality data” so that retail representatives can use benchmarks to assess the quality of the data received from their suppliers.

As part of the study:

- The concept “quality of competitive data” is studied.
- Key indicators that affect the quality of the data are identified.
- A study of the “industrial standard” is conducted, which is where we look at the quality of data that retail leaders receive throughout the world.
- Standard values for each indicator are formulated and we talk about their impact on the retailer’s business results.

The study answers to these and other questions about data in retail:

1) Which settings affect the quality of the data?

2) What is the industry standard of the data?

3) How do you change the business indicators when the quality of the data changes?

4) How do you asses the data quality of a retailer?

Research Methodology

1. The study was conducted on the basis of data from Competera customers from six countries (Great Britain, USA, Netherlands, Malaysia, Russia, Ukraine).
2. All of the data is impersonal and confidential
3. To participate in the study, the companies that track and analyze more than 50 thousand commodity items were selected
4. The study was conducted in three industries: Consumer Electronics, Hypermarkets, Health & Beauty

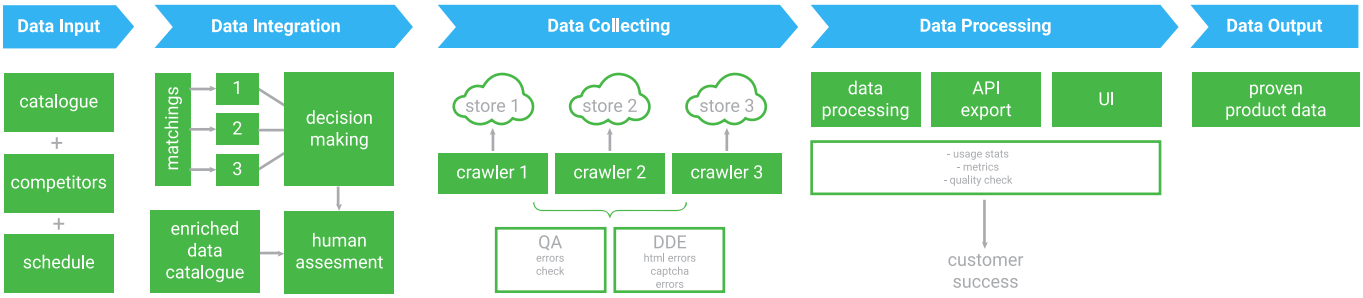
6 countries: Great Britain, USA, Netherlands, Malaysia, Russia, Ukraine

3 industries: Consumer Electronics, Hypermarkets, Health & Beauty

Assortment: from 50,000 products

Key elements of data quality

The quality indicators in the study are relevant for a properly tuned process for collecting and processing data with multiple quality checks for errors and anomalies:



Incoming data

For input data, the catalog of the retailer's goods, the list of competitors and the required schedule for scanning competitors (refer to "freshness of data") is used. Incoming data is updated according to a specified schedule.

Data processing

After processing incoming data and comparing the retailer's goods with the competitor's goods, we get an enriched catalog of positions, which will be used to collect data from competitors' sites (refer to "percentage of comparisons").

Data collection

Data is collected from selected websites according to a predefined scan schedule. The list of websites for scanning can be adjusted after several iterations as long as the scanned competitors are not key (read the article on how to choose products and competitors for monitoring).

Checking and structuring of data

The data collected on competitor's sites is checked for quality (refer to "the number of non-zero prices" and "the percentage of errors") and is structured and visualized in the suitable form for the retailer.

Data transfer to the ERP system of the retailer

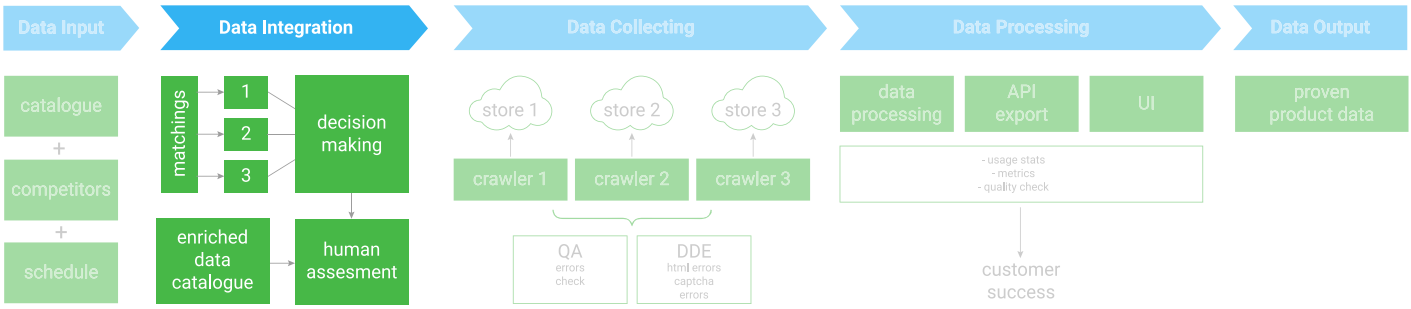
This step may be the last, but not least, in the entire sequence (refer to the corresponding paragraph).

If the data acquisition process is properly configured, you can proceed to assess the quality of the data received. As a result of the research, the five main indicators of the data quality were identified.

Indicator # 1: Percentage of comparisons

The number and accuracy of the obtained product comparisons with similar parameters is the first parameter that helps you understand how good the data is. Additionally, if the number of comparisons can be measured quickly using the percentage of comparisons, then their accuracy is affected by the type of comparisons as well as their depth..

Depth is the second most important parameter, but it is often overlooked when calculating the percentage of comparisons. It takes into account each parameter option of the selected product such as its colors, technical characteristics and other parameters. They are often not available on the main product card.



The ideal situation is when the retailer gets the most profound and accurately mixed comparisons. In this case, the indicator, “percentage of comparisons”, will be informative to its maximum.

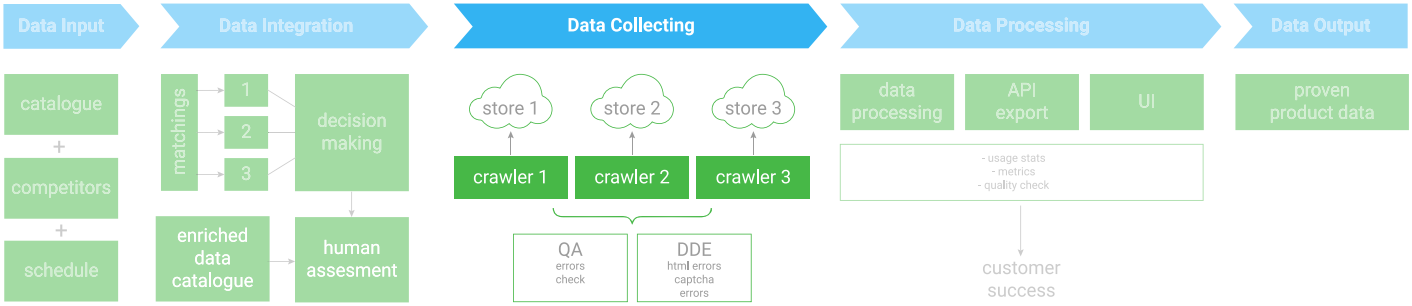
There are three different types of comparisons:

- **Hand.** They allow you to accurately compare goods not only by text parameters (name, price), but also by appearance (which is especially important for fashion retail). Disadvantages - low speed, high cost and a large number of errors caused by human factors.
- **Automatic.** They help you quickly collect large amounts of data sets, but do not always do it correctly, especially when the layout on the competitor’s site has changed or security is turned on.
- **Mixed.** With proper use, you can gain the advantages of both approaches - the speed and accuracy of automatic comparisons and the quality of manual comparisons.

Indicator # 2: Number of non-zero prices

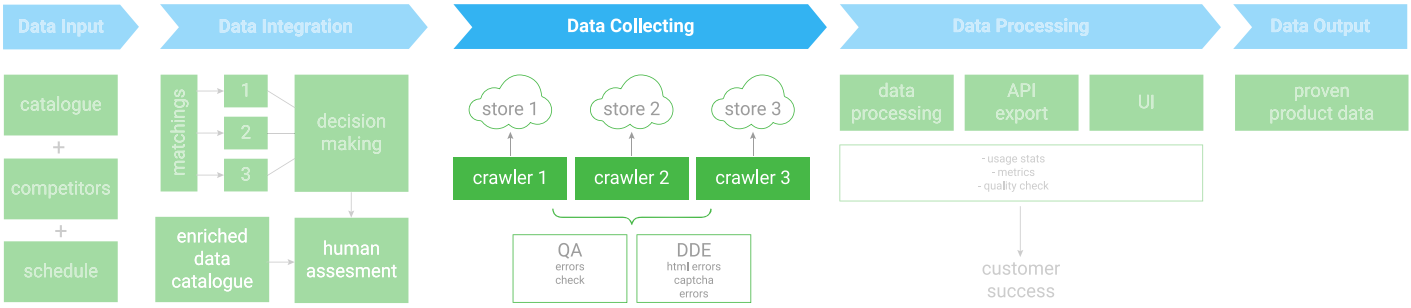
This indicator shows the percentage of prices found on the competitor's site. Sometimes, due to various reasons, a certain amount of zero prices appear during scanning:

- At the time of the data collection, the competitor ran out of goods. In this case, the collector receives information claiming that the goods are not available.
- The competitor has problems on the site, so the collector can't "take" the price.
- The competitor does not have that product. This situation is possible when the retailer is not properly informed about either the assortment of competitors or when, during monitoring, the key competitors were not selected.



Indicator # 3: Freshness of data

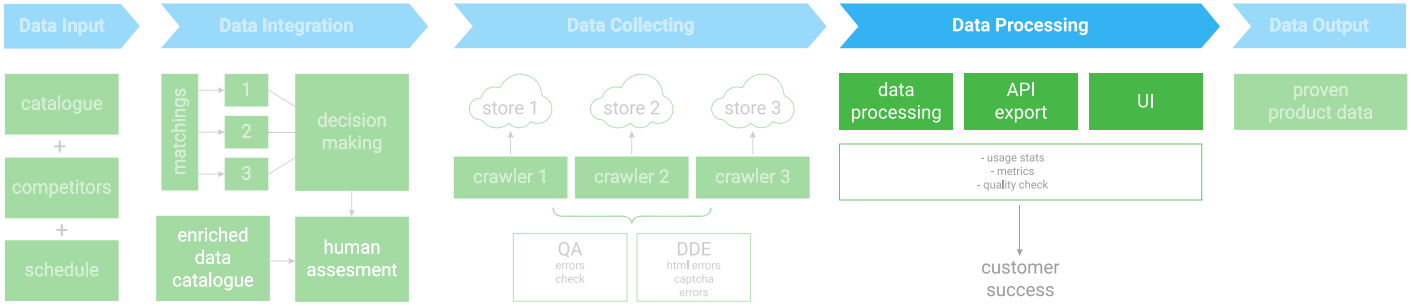
Some retailers, when re-evaluating products, drive off of those that were based on data that was collected a day or more before the revaluation. Naturally, that kind of data can't be a qualitative basis for analysis and decision-making. Retailers need to get the most relevant data at the time decisions are made on price changes, so it's important to take into account another indicator of data quality — the percentage of "fresh prices" collected in the two-hour range prior to revaluation.



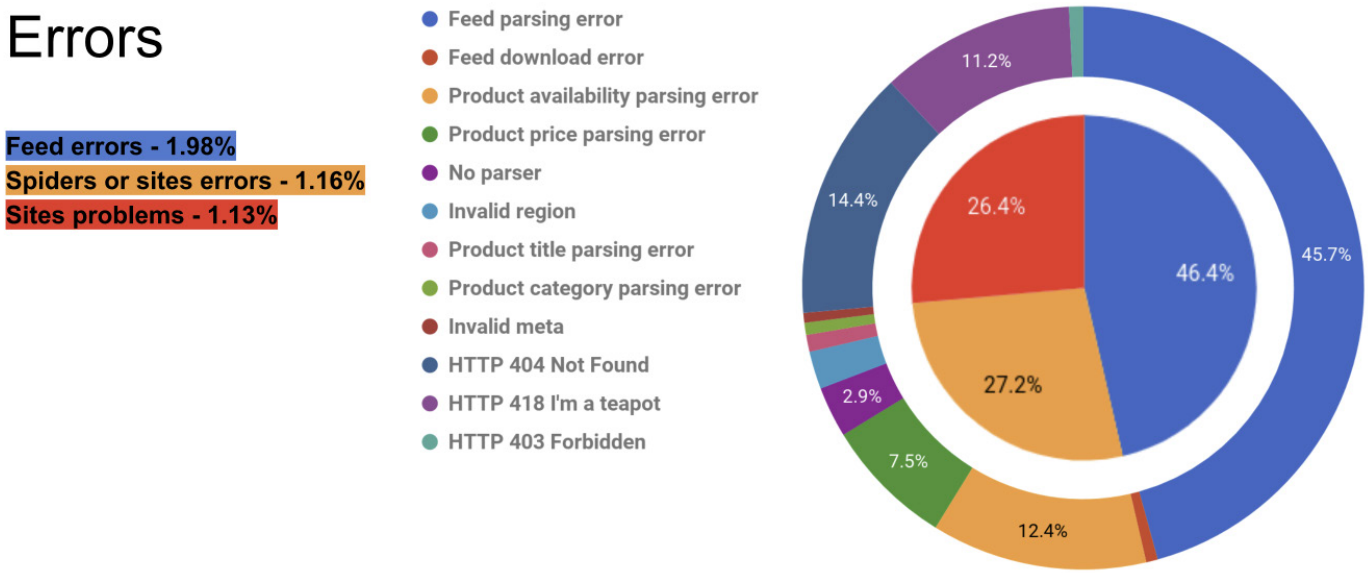
Important: This indicator directly depends on what technology the data provider uses to collect it. If the supplier collects data on the entire catalog of the competitor, from the moment you receive the first price to the moment you receive the last one by the retailer, up to 72 hours could have passed. If the collection occurs with a link to specific URLs of the goods, then the percentage of fresh data will be much higher, and the data itself will be more accurate.

Indicator # 4: Percentage of errors in the collected data

Despite the high accuracy, data collectors (parsers, crawlers - programs that collect data) are not immune to errors. The ideal variant is when the program of the collector is checked by either another program and/or person.



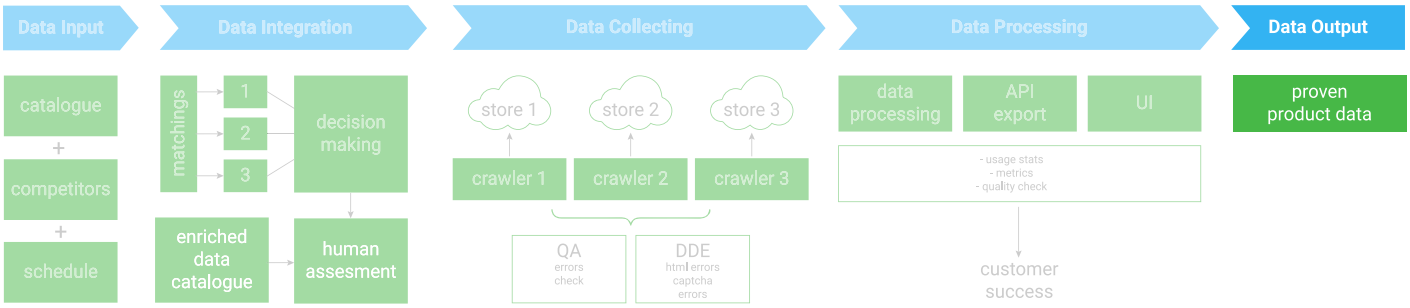
In this case, the competitor data collection system not only effectively collects the necessary data and checks for errors, but also, even identifies every possible problem with the collection in order to notify the user what exactly caused the data to be received with errors.



Data can be considered smart only if its quality is verified by scientific methods using special algorithms.

Indicator # 5: Data delivery time

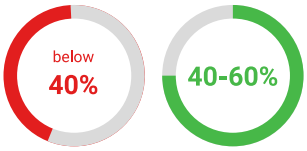
The less time that the data needs in order to get into the ERP system of the retailer after collection and processing, all of the processes of the retailer in the analysis and use of the data obtained are simpler and more flexible. As a result, the interaction of every business element is more effective.



The time it takes to deliver the data to the internal system of the retailer ***should be minimal***.

Results of the study

The reference for percentage comparison

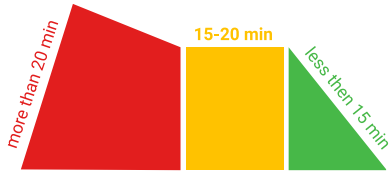


The level of the comparison coverage of a variety of stores depends on:

1. The complexity of the industry
2. The quality of the process or data delivery partner by product comparisons

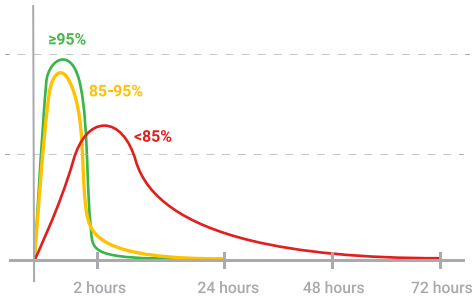
The reference time of comparison

The coverage time for 1 product by 10 competitors



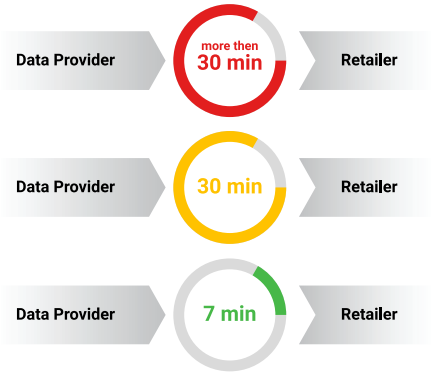
The maximum error indicator of tracking systems up to 1%

The relevance reference of data



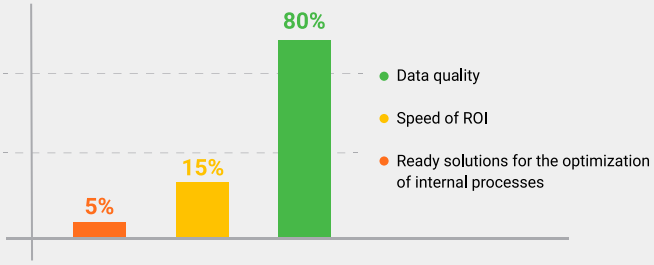
To conduct a correct revaluation, the relevance data indicator obtained in the two-hour range prior to the revaluation must fluctuate in the range of 95-98% of the positions of the entire assortment.

Reference of the delivery time in the ERP-system

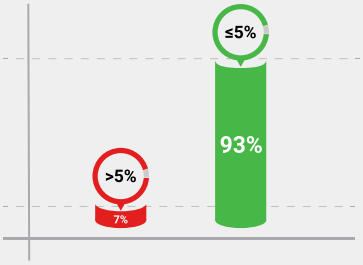


Question: the criteria of retailers when selecting a provider based on the delivered data

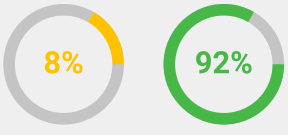
Parameters that impact the choice of provider based on the delivered data



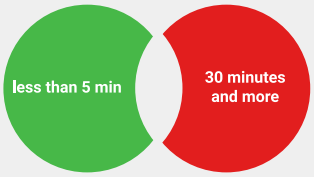
Expectations for the number of errors in the delivered data



Proactive data quality and error checking



Support Response Time



Modeling Results

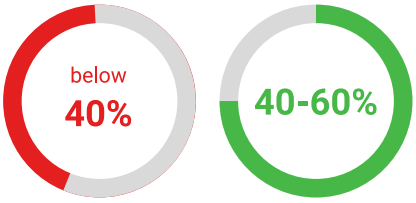
Now that we've examined the factors that affect the quality of the data and have seen the figures of the top retailers from six countries collected in the results of the Competera study, it's possible to simulate the effect caused by the deviation of these indicators from the norm.

For example, let's take a conventional retailer that's monitoring 50,000 items from five competitors. In this case, the maximum possible number of units that you can monitor is 250,000.

Percentage of comparisons

This indicator depends solely on the retailer.

Among the companies selected for research, the percentage of comparisons runs from 40% to 60%.



The level of the comparison coverage of a variety of stores depends on:

1. The complexity of the industry
2. The quality of the process or dara delivery partner

by product comparisons

In other words, for the current model, there can be, respectively, 150 thousand, 225 thousand and 37.5 thousand positions. For all further calculations, take the average of 60%, since it corresponds to 150 thousand positions.

The quality of the remaining indicators is provided by the data provider (if the retailer doesn't work with its own system of information gathering about competitors).

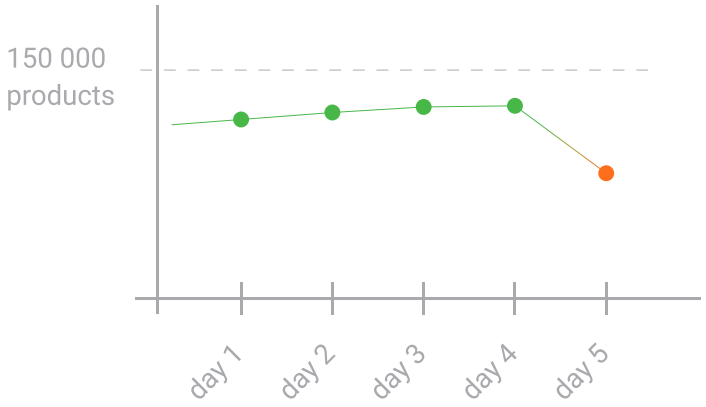
The number of non-zero prices

Each retailer independently determines how many non-zero prices to choose as the standard. It is important that this indicator is stable or increases, but it should not decrease.

The appearance of a large number of "zero prices" creates a "blind zone" for the category manager and for the revaluation of algorithms, so the goods with such data are not repriced, and as a result, they're not sold.

In our example, for instance, suppose that the number of non-zero prices will fluctuate at the level of 150 thousand, but at some point, it will drop sharply to ninety. In this case, the "blind zone" of the retailer will increase by 60 thousand positions in one day.

The influence of the number of comparisons for repricing

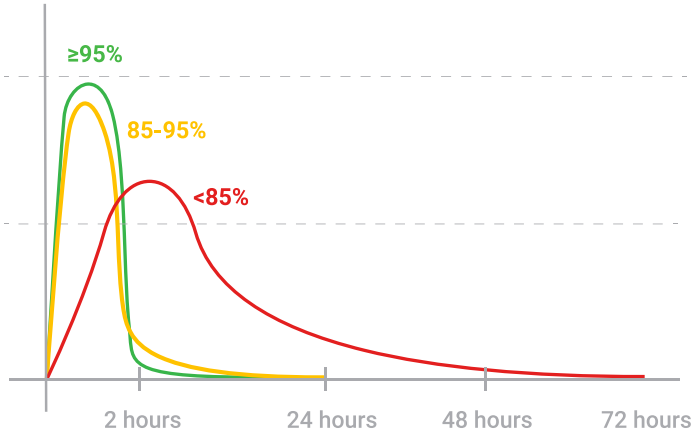


Freshness of data

As we noted above, it's important that the maximum amount of prices are collected by the retailer in the two-hour range prior to the revaluation of goods. This parameter determines what percentage of prices for analysis and decision-making are delivered to the "critical" period of time, which is necessary for reassessment.

In order to conduct a correct reassessment, the fresh data indicator should fluctuate in a range of 95-98% of the positions of the entire assortment. In our example, out of all of the matched items, up to 147 thousand pieces of data should be delivered in a two-hour range before repricing. This quantity is an indicator of the relevancy of the prices.

The systematic presentation of data with a fixed frequency and schedule is one of the most important factors affecting the optimization of pricing. In this case, the retailer receives the exact data that can be used for revaluation. Otherwise, it can operate with obsolete or "yesterday's" data, however its sales will fall. This option is similar to the previous example, but the damage from it to the business is slightly lower.

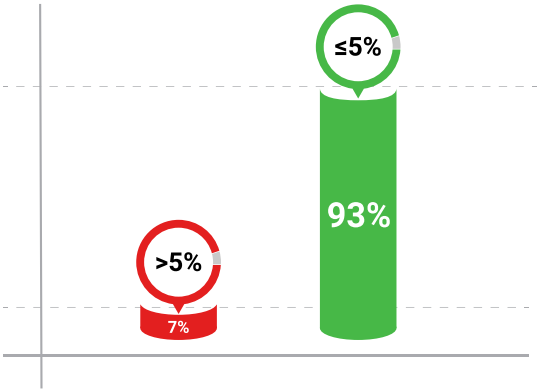


In most cases, the category manager knows the frequency and time of the repricing of its competitors, especially if the Price Index is used in the work. Based on this information, they choose the time to reassess their range, therefore the specific time interval for the appearance of data is critically important: outdated data overlooks possible changes in the competitors.

Percentage of errors in the collected data

Even the minimal percentage of errors leads to the fact that the retailer incorrectly reprices their goods. That means that you need to strive for a minimum score of erroneous data — from 5% to 2%. If the indicator is higher, in our example (147 thousand goods in a two-hour range), over 8,800 items will be repriced with errors.

The data provider, regardless of whether it is internal or external, must check the quality of its data, evaluate them for errors, give transparent access to the retailer about this information, and quickly respond to errors.

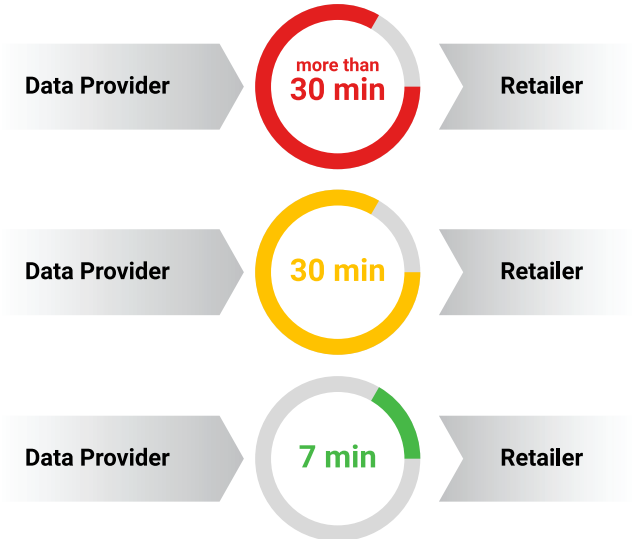


Data delivery time

The time it takes to deliver data to the ERP system of the retailer is not as critical as other indicators. At the same time, it affects the overall performance of category managers:

> ***If the manager does not have the task of monitoring and setting up the process of obtaining data and he is solely engaged in their processing, then this reduces their time spent on making repricing decisions.***

The industry standard for delivering data from the supplier to the ERP is up to 30 minutes (depending on the data transfer method and type of ERP system.) Typically, Competera clients receive new data in their ERP systems in 3-7 minutes.



Checklist for checking data quality

Now that you have read the results of the data quality study from Competera, you can check the quality of your data with the checklist below.

If your data quality metrics are in the green zone, then everything is fine. If some indicator is in the orange or red zone, you should consider talking to your data provider.

Checklist for checking the quality of the data

What comparison percentage are you getting?

What is happening to the number of non-zero prices?

What data percentage do you receive within 2 hours prior to repricing?

What is the percentage of errors in your collected data?

How long is the data delivery in your ERP-system?

The results of the data quality check

Compare the pattern that you received

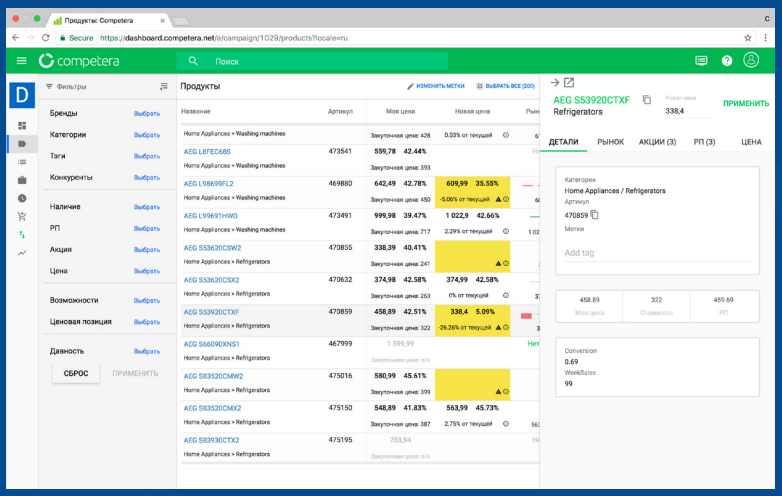
qualitative data

average quality

not quality data

Competera
Competitive Data

Get proven real-time competitive data in the right place



Full Stack Coverage
Historic and current pricing, promotion, stock data, tracked weekly, daily, and hourly from any brand

Built Errorless
Self-learning algorithms for data collection with 360 degree view on crawling errors, SLA - 98%

Built Flexible
API, high frequency and custom schedule with data streaming to ERP and more

Read more at competera.net