BIG DATA

Marketing Data Cleansing: Stages, Strategies, and Tools Explained

OLEKSANDR SHYKOLOVYCH OCTOBER 13, 2022

Today, we continue our series of articles on marketing data cleansing. Here are the key points we'll cover in this blog post:

- Marketing data cleansing stages
- Data cleansing strategies and best practices
- Marketing data cleansing tools

Check out our previous article if you want to review the fundamentals of marketing data cleansing.

Marketing data cleansing stages

Even with the most advanced and well-thought-out data schemas, some inconsistencies are inevitable. That's why it's important to set up a data cleansing process to fix any data inconsistencies on an ongoing basis. The data cleansing process usually consists of five essential steps. Let's consider each of them in detail.

Step #1: Data validation

Data validation is the process of verifying the accuracy and quality of the extracted data before taking any further action with it.

It would be unwise to assume that all data about your leads (names, emails, company names, etc.) has been entered correctly in the database. There are numerous ways to spoil data, ranging from data source malfunctions to human errors. This is particularly true for manually entered records.

That's why data validation is the first and most essential stage on the path to data hygiene.

There are five main types of data validation:

- Data type validation. Confirms that the entered data is the correct data type (numeric, string, etc.).
- Code validation. Ensures that the given data field is selected from an existing list of values and adheres to a specific list of formatting rules. For example, if you gather postal codes, you need to match your records against a list of valid codes and make sure they are stored in the correct format.
- Data range validation. Verifies whether given data falls within a certain range. For example, marketers conduct surveys using a scale that ranges from 0 to 10 to calculate the Net Promoter Score (NPS). Any values that fall outside of this range are invalid.
- Data format validation. Ensures that data is stored in a proper format. For example, date columns often use a format like ""YYYY-MM-DD" or "DD-MM-YYYY". To avoid errors, incoming data should be stored in the specified format.
- Consistency validation. Verifies the logical sequence of stored data. For example, the purchase date should follow the registration date, not the other way around.
- Uniqueness validation. Some records, such as IDs, emails, etc. are unique. This type of validation ensures that there are no duplicates of unique items.

Step #2: Aligning data formats

The second step in marketing data cleansing is to bring all metrics together in a unified form. The problem of disparate naming conventions is one of the most common in marketing data. We've already explained that the same metric on different platforms may have different names.

When using tens of different marketing and sales tools that don't talk to each other, marketers waste hours figuring out what's working and what's not. Keeping an eye on each metric's name and mapping all sources manually is a real pain in the neck.

For the sake of time and data granularity, marketers use automated solutions to align all metrics. For example, Improvado's MCDM (Marketing Common Data Model) automatically normalizes disparate naming conventions and supplies analysts with analysis-ready data.

The platform frees up to 20% of marketing analysts' time by automatically matching all data fields and normalizing heterogeneous data. With its help, marketing analysts can fast-forward routine data manipulations and dive straight into the analysis process.

Step #3: Getting rid of duplicates

With standardized data formats, the next step is to check your dataset for duplicates that were missed during the previous stages.

Duplicate entries are dangerous for multiple reasons. First off, if the same entry appears several times, the quality of the whole dataset deteriorates. You can no longer determine how effective your campaign is if the metrics don't match up.

Furthermore, duplicates become a real problem for companies dealing with predictive and prescriptive analytics.

Learn what prescriptive analytics is and how to reach analytics maturity with our guide.

When machine learning models receive false data, they will output unexpected results. This leads to biased performance estimates and disappointing results in future marketing campaigns.

In fact, duplicate cleansing isn't the hardest thing to do if analysts have some basic technical expertise. For example, in SQL, a primary data cleansing operation can be done with a few simple queries. You can:

- Find duplicates using the GROUP BY or ROW_NUMBER() functions.
- Use the DELETE statement to eliminate duplicate data rows.

However, this basic example is only true for small databases with a small number of records. The more data columns you have, the more actions you have to take with your data. Besides, processing large amounts of data requires significant computing power, which the average marketer usually doesn't have at their disposal.

Instead of SQL, marketers can also use automated data transformation tools to get rid of duplicates without SQL queries. Let's review the automated data transformation process using the example of Improvado.

Improvado's DataPrep allows marketers to automate the data transformation process and convert raw data into actionable insights in a no-code environment. Marketing and sales specialists can work in a familiar spreadsheet-like UI with drag-and-drop functionalities. The tool offers ready-made transformation recipes that help marketers merge disparate data in a single table.

The platform also supports custom transformation to meet the varying needs of marketing and sales analysts. Furthermore, built-in decision trees help teams enrich datasets and optimize data to achieve better analysis outcomes. By using the clustering feature, marketers can also find non-obvious data groups and identify similarities within a dataset.

Step #4: Normalizing missing and incomplete data

After getting rid of duplicates, the next step is to fill in the missing data and fix inaccurate values. Incomplete data lowers the overall quality of the database and doesn't allow for a precise analysis.

Incomplete or incorrect data often happens when different tools collect information in different ways. For example, some tools may record the city as "New York" while others may use "NY".

Let's assume that we have a database with company addresses that includes the following parameters:

- State
- Town
- ZIP code
- District
- Street

Some records might have missing data rows that don't allow the complete address of the company to be identified and the records to be sorted for further analysis. There are two ways for analysts to solve this problem:

1. Delete all records with empty values in any of the fields. This is the fastest option, but it also leads to the loss of significant amounts of data.
2. Fill in the missing data. Some data, such as state or ZIP code, is easy to find when you already have information about the town and street. Even though this approach takes more time, analysts can save lots of valuable data. If it's not

possible to fill in the missing information, the data row should be deleted completely.

The majority of marketing analysts fill in their incomplete data manually, wasting a lot of time on data collection and entry. However, companies with advanced analytics use proficient tools to automate this process.

For example, Edwin Tan, a data science specialist with over eight years of experience in different companies, shares his experience with Pandas in this guide. He explains how Pandas (the Python data science library) can be used to deal with missing data. The guide covers many approaches, such as:

- Fill with Constant Value
- Fill with Mean
- Forward Fill
- Back Fill
- Interpolation
- Etc.

Step #5: Identifying conflicts in the database

The final step of the marketing data cleansing process is conflict detection. Conflicting data are insights that contradict or exclude each other. At this stage, analysts' main goal is to identify contradicting data and eliminate it.

Let's get back to our example from Step #4. It appears that the state doesn't match the ZIP code or the town mentioned in the dataset.

This type of mistake is difficult to fix because there's no way to tell what exactly is incorrect: the ZIP code or the town. The best option here is to double-check the source or contact the person who entered this information into the database to figure out the details.

If that's impossible, the record should be marked in the database. In this way, analysts will know that this information is unreliable and can omit it during the analysis.

Marketing data cleansing: best practices and strategies

Now that we've covered the basics of marketing data cleansing, it's time to move on to the best practices and strategies.

Apply SQL

SQL-based data cleansing is one of the most common and at the same time complicated approaches to data hygiene. It allows raw data to be altered without any third-party tools and provides the most extensive functionalities for manipulations with the dataset.

The main problem with this approach is that marketers must know SQL and have a solid engineering background, which is a rarity for the majority of marketing analysts.

Of course, marketers can get comfortable with basic SQL queries to perform high-level data transformations. For example, Sayak Paul, a machine learning engineer at Carted, explains common data cleansing techniques in his guide for DataCamp. The guide proves that knowing several queries can help with cleansing messy data and deduplicating your dataset.

However, datasets differ, and there's no one-size-fits-all approach to marketing data cleansing. A dataset with millions of records requires a lot more transformations than a dataset with a hundred records. That's why analysts have to master SQL to deal with large databases and achieve more granular insights.

Choose an automated data cleansing tool

Apart from SQL, there are a lot of automated tools and programming languages that accelerate the data cleansing process and take the burden of routine work off analysts.

Let's consider some of the most popular programming languages first.

Python

Python is a universal tool for data scientists and analysts. The language is versatile and has an intuitive syntax, and it also offers a number of libraries for data scientists and engineers.

For example, Pandas is an open-source Python library for data analysis. It was specifically designed for data wrangling and pre-processing. The library takes data in CSV and TSK file formats. Moreover, it can process a whole SQL database and create Python objects based on it.

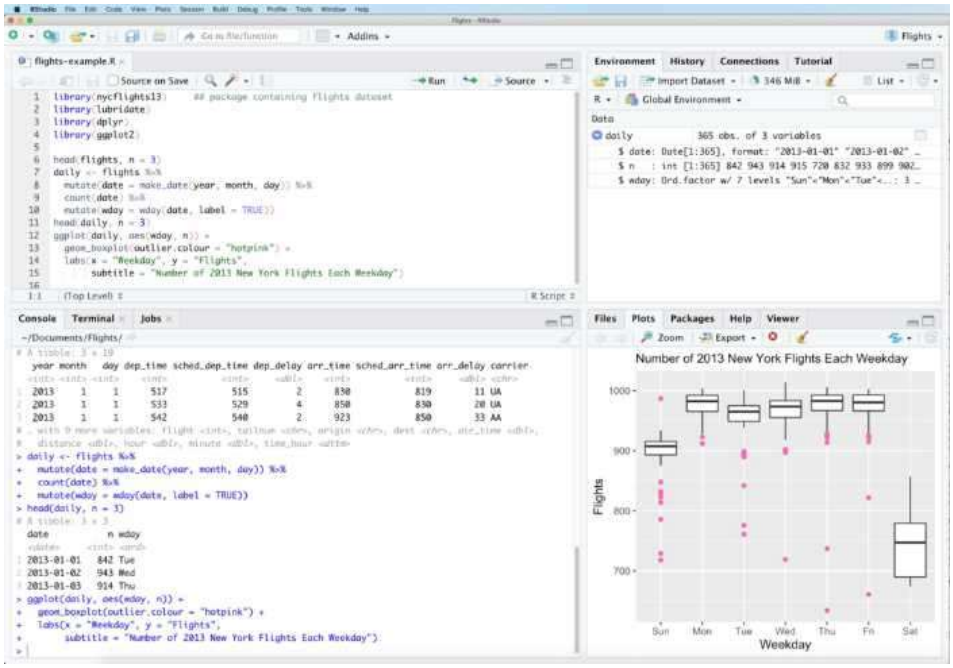
The library allows you to manipulate, merge, and sort data, as well as plot your datasets, input incomplete data, and more. These features make Pandas a fundamental library for data science and analytics.

Another well-known package is NumPy, an array-processing library that allows users to work with arrays. NumPy's main object is a multidimensional array that represents a table of values, indexed by a tuple of integers. Marketing analysts use NumPy to process arrays that store metrics and apply advanced array operations, such as stacking, splitting into sections, and broadcasting.

You can find even more advanced Python libraries in this guide on the top 10 Python libraries for data science from Rashi Desai, a data analyst at Blue Cross and Blue Shield of Illinois.

R

R is a free, open-source programming language and software environment for statistical computing and data plots. The language may be used to clean, analyze, and graph both raw and structured data. Proficient marketing analysts use it to measure the effectiveness of advertising efforts and visualize their performance.



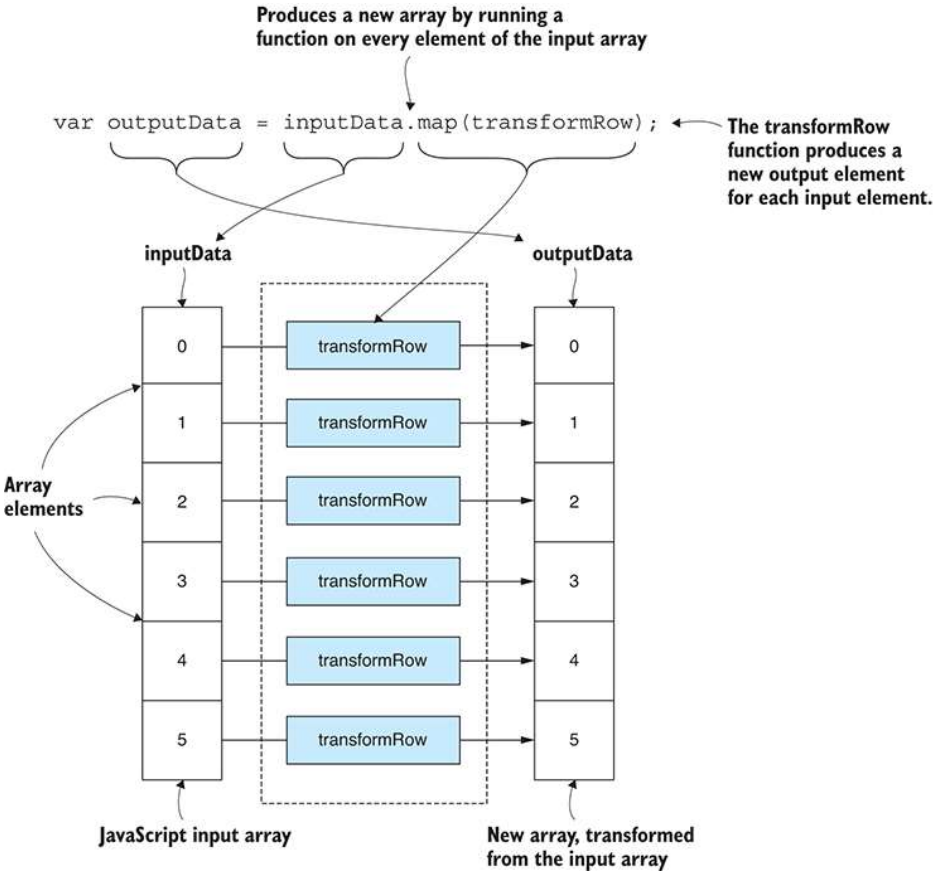
R studio UI

JavaScript

Despite being less popular among data analysts, JavaScript also has several libraries for data processing in the stash. For example, Data-Forge, a library

inspired by Pandas, offers a whole set of features for data cleansing. It helps analysts with:

- Rewriting false data rows
- Filtering data rows
- Filtering columns
- Data aggregation
- Etc.



Data-Forge map function for data transformation

Still, programming languages aren't the only thing to accelerate marketing data cleansing. Third-party tools stand alongside them and generate outstanding results.

Dbt (H4)

Dbt helps analysts transform their data in data warehouses using SQL's select statements. The platform helps them to create complex models, run tests, cleanse and transform data, etc.

As the dbt team says, their product represents the T in ELT (Extract, Load, Transform). Since it's an all-around tool for data transformation, it can clean up strings, change data types, modify values in the database, and apply different business logic for various marketing metrics.

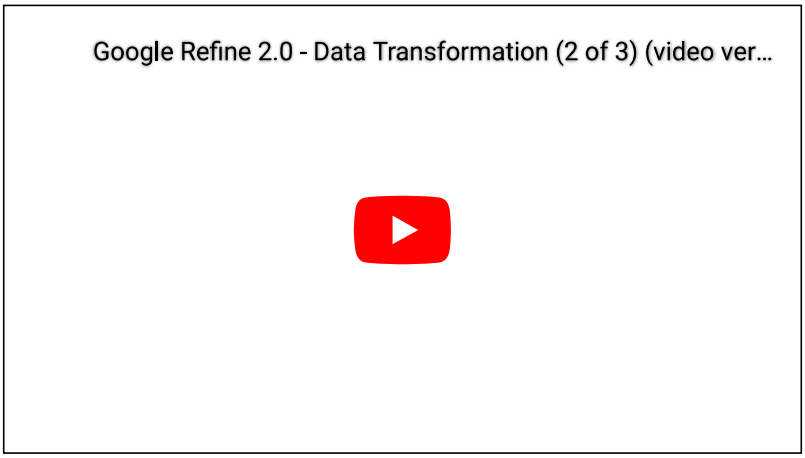
Dbt became the foundation for the modern data stack (a suite of tools for data integration). In combination with data extraction, loading, and visualization platforms, dbt is shaping the future of marketing analytics.

OpenRefine

OpenRefine lets marketing analysts clean, correct, modify, and expand data without effort. The platform was originally known as Google Refine, since it was actively supported by Google. Later, the tool became open-source and the name was changed to OpenRefine.

With the help of this tool, you can automatically fix typos, convert data to the right format, duplicate datasets, and complete many other actions.

Here's a video that briefly explains OpenRefine's data transformation capabilities:



Integrate an ETL system into your data infrastructure

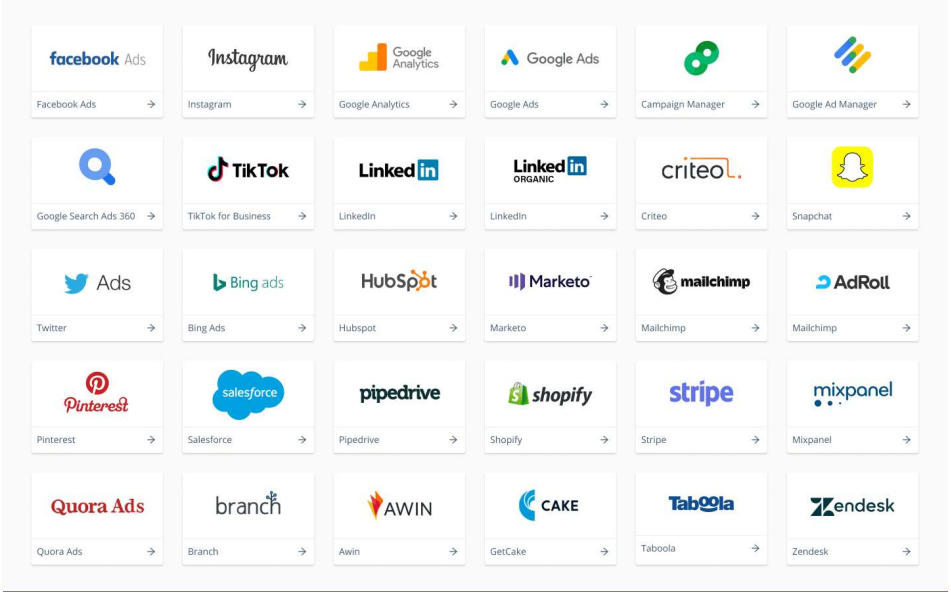
If you don't want to dig through complex SQL queries and test tens of different libraries, a marketing ETL platform is something you should try out.

ETL stands for extract, transform, load. An ETL platform is an all-in-one tool for all operations with data, from data cleansing to visualization.

⚙ Learn more about each stage of the ETL process and how it benefits marketers in our guide. ⚙

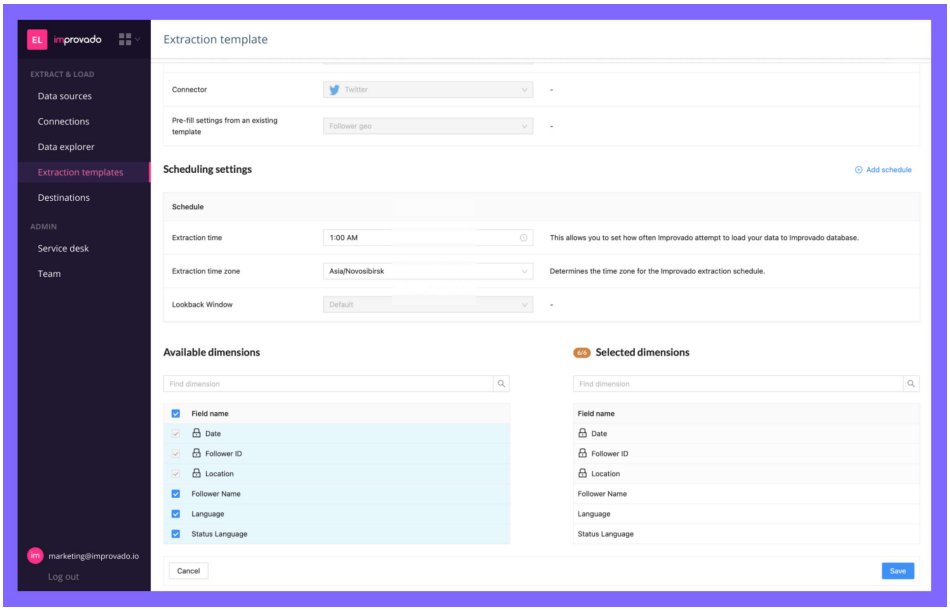
If you're not familiar with ETL platforms, we'll explain all of the details using the example of Improvado.

Improvado is a revenue ETL platform that handles a full cycle of marketing and sales data operations. At its core, it's a data pipeline that connects to 300+ marketing and sales data sources to extract performance insights.



Some of Improvado's data sources

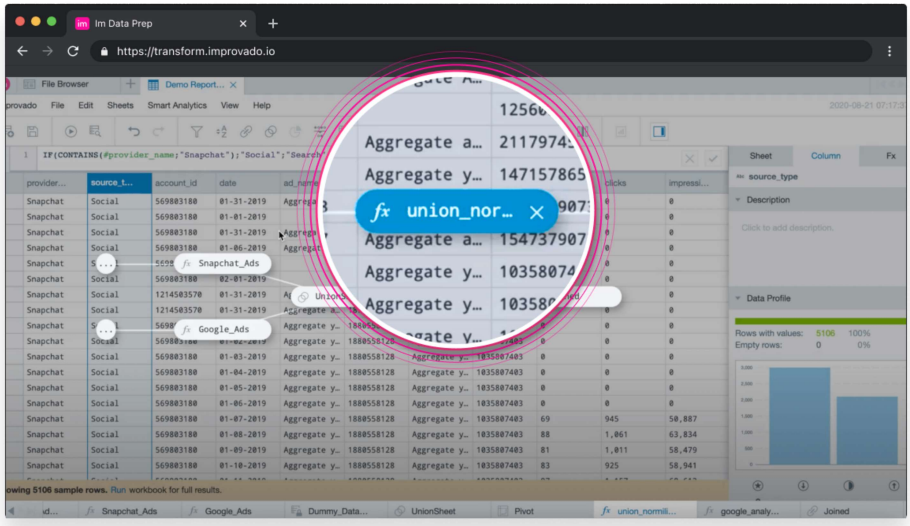
One of Improvado's distinctive features is that analysts don't need to deal with code. Usually, engineers have to create specific queries to trigger the data source's API and extract insights. Improvado has prepared everything beforehand. With predefined extraction patterns, marketers can connect all data sources automatically. It takes just a few clicks and doesn't require any additional actions.



Improvado's extraction workflow

All marketing data sources transfer raw, unstructured data that should be transformed into a digestible format. That's where marketing data cleansing comes into action. Improvado's DataPrep module offers new opportunities for data transformation and cleansing.

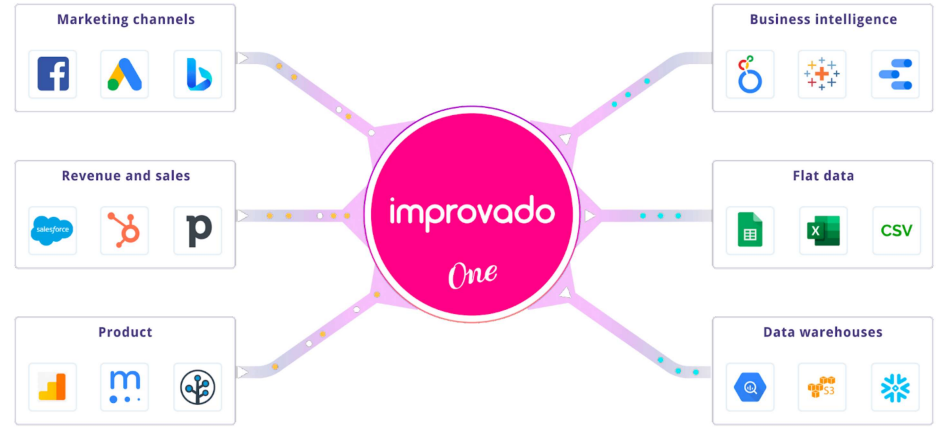
With the help of prebuilt transformation recipes, marketers can clean their data faster and get the right answers. Instead of building SQL queries, marketers can work with data in a traditional spreadsheet-like UI. Improvado also mitigates the risk of human error, since all actions are automated and follow a predefined pattern.



DataPrep UI

After all of the marketing data cleansing stages are complete, Improvado loads insights into a data warehouse. From there, analysts can access the data with just a few clicks. Improvado streamlines real-time insights to 15+ visualization tools, so analysts can build a holistic dashboard and get a full report on marketing performance in a single tab.

An ETL platform is the best way to stop juggling tens of analytics tools and gather all of them under the same roof. Improvado itself replaces dbt, data warehouses, the SQL environment, tens of APIs, spreadsheet software, and saves hundreds of hours on manual reporting.



A single tool for your marketing analytics

Automate marketing data cleansing with Improvado

Data cleansing is a lengthy and dull process when done manually. The larger your datasets, the more time analysts waste tidying them up. In the marketing world, where you need immediate results and insights, manual cleansing is simply not an option.

Improvado can help you automate your marketing data cleansing and gather granular marketing insights in one place. Schedule a consultation to learn more about us.

Learn how a revenue ETL platform can help you exceed your marketing goals and save time your analysts' time.

Ali Flynn
VP of Customer Relationship

CONTACT US

Oleksandr Shykolovych
Editor-in-Chief at Improvado

Oleksandr Shykolovych is an Editor-in-Chief at Improvado blog. A strong desire to share information with people and a keen eye for details help Oleksandr create engaging content and be on the same page with readers.

San Diego | Headquarters

3919 30th St, San Diego, CA 92104

-

San Francisco

2800 Leavenworth St, Suite 250, San Francisco, CA 94133

-