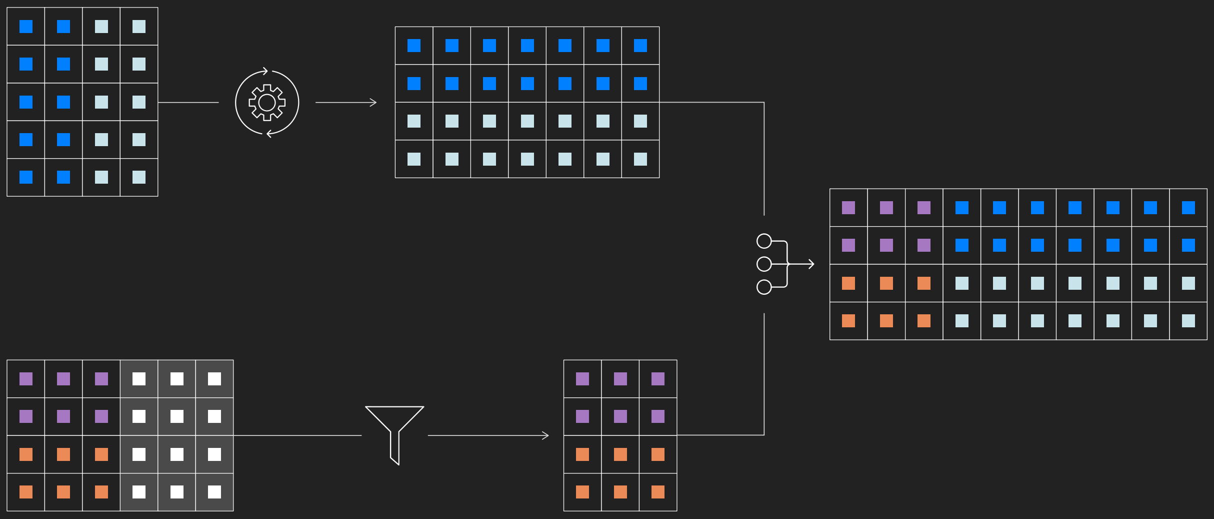
The Ultimate Guide to Data Transformation

How data transformation can supercharge your analytics efforts and enable your data professionals to do their best work



Fivetran

Executive Summary

Today, there is hardly any company in the world that doesn't recognize the value of data. Research shows data-driven companies are 20 times more likely to acquire new customers and six times more likely to keep them.

Data analytics is becoming critical in every industry, helping business leaders explore data in meaningful ways and make smarter decisions about everything from the products they deliver and what markets they should target to transforming supply chain management and more.

And data transformation plays a key role in converting data into actionable insights. In this ebook, we will cover the following topics:

1. **What data transformation is and why it matters** – Data transformation is essential to building a mature data practice as raw data is impractical for analytics or machine learning.
2. **Examples of data transformation** – There are many ways to transform data into usable data models. We will discuss common examples.
3. **ELT vs. ETL** – Why the future of data integration is ELT.
4. **Getting started with data transformation** – Best practices to help you get the most value out of your transformations.

1 What Is Data Transformation and Why Does It Matter?

Data transformation refers to any of the operations – revising, computing, separating and combining – involved in turning raw data into analysis-ready data models. Data models are representations of reality that can be readily turned into metrics, reports and dashboards to help users accomplish specific goals. In particular, businesses need KPIs and other metrics in order to quantify and understand what and how they are doing.

Why Transform Data?

Transformation prepares data for the full range of use cases. This includes:

- Analytics
- Machine learning
- Regulatory compliance

Analytics

Basic, fundamental analytics for supporting decisions needs to start with metrics.
Sometimes, metrics can be computed from a single source and only need a modest amount of transformation. Other times, the only way to compute a metric is to combine data from a wide range of sources and then aggregate it.

Some important metrics for your business can include:

Revenue

- Annual recurring revenue (ARR)
- Net revenue retention (NRR)
- Unit economics: e.g., customer acquisition cost, sales efficiency

Sales and marketing

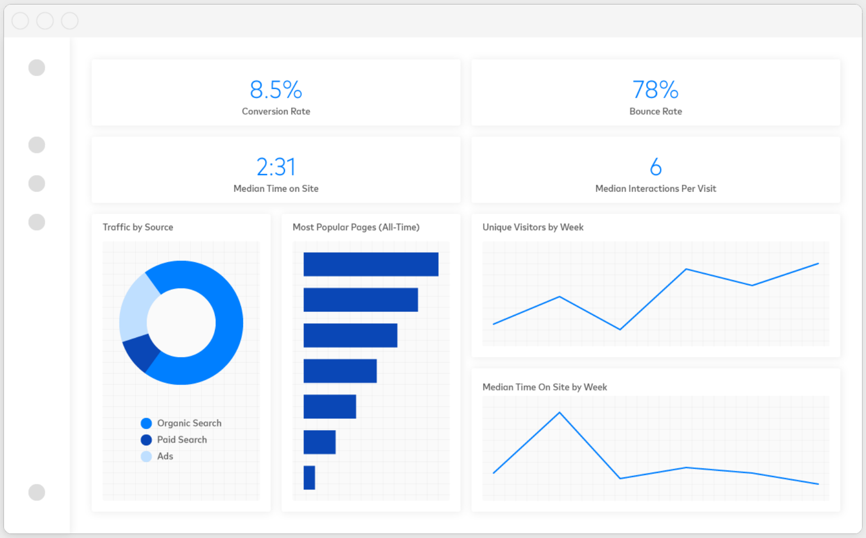
- Customer growth and churn rate
- Month-over-month revenue growth
- Marketing-qualified lead and conversion metrics

Product

- Daily, weekly, monthly active users
- Customer journey
- Feature adoption and usage
- Net promoter score

Other important business metrics can include those concerning supply chains, quality control, financial performance, recruiting and more. The purpose of metrics is to give your organization measurable, actionable goals and the ability to measure performance as well as identify strengths, weaknesses, opportunities and threats.

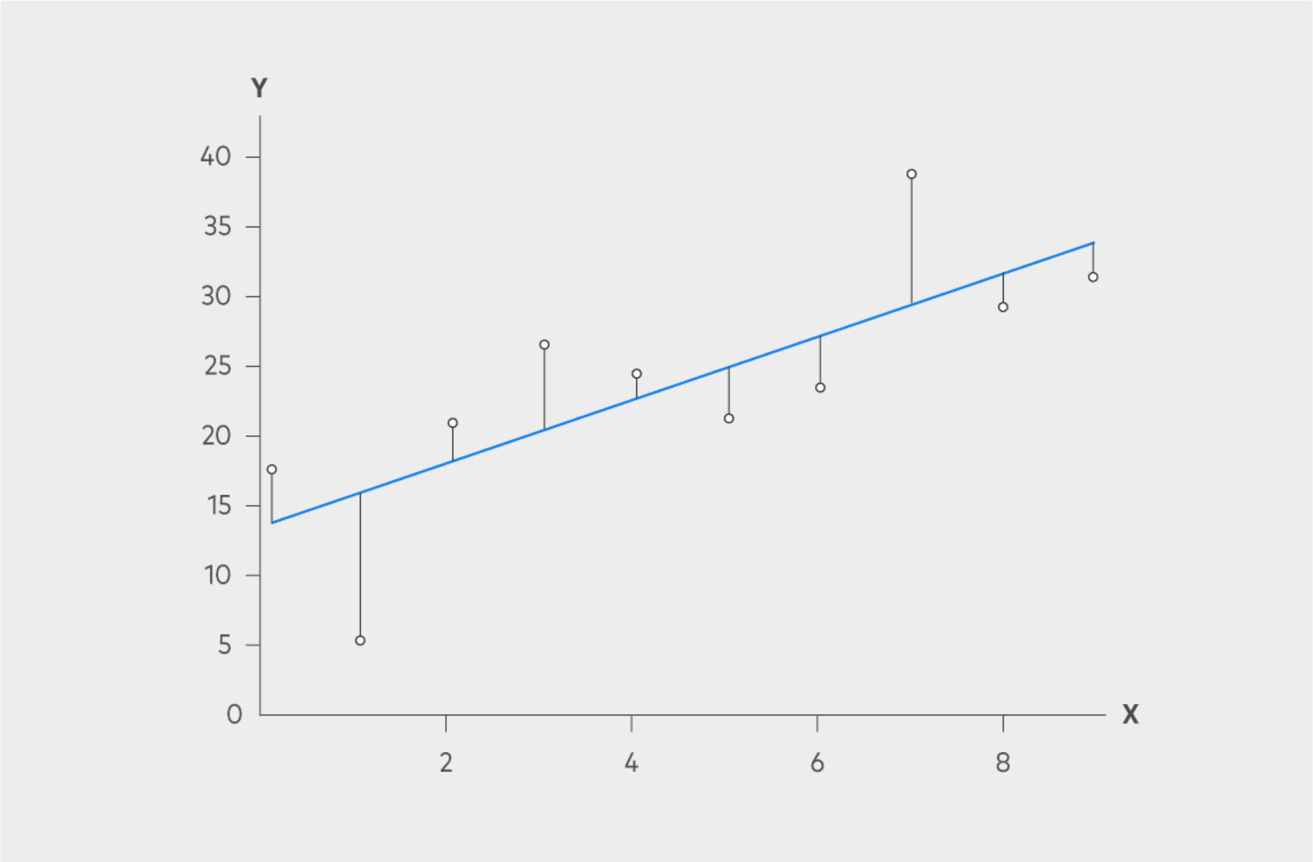
Metrics can be turned into a number of visualizations representing proportions, tables, rankings and trends over time and other important concepts. Visualizations can then be assembled into dashboards and reports to quickly summarize important findings to key stakeholders in your organization.



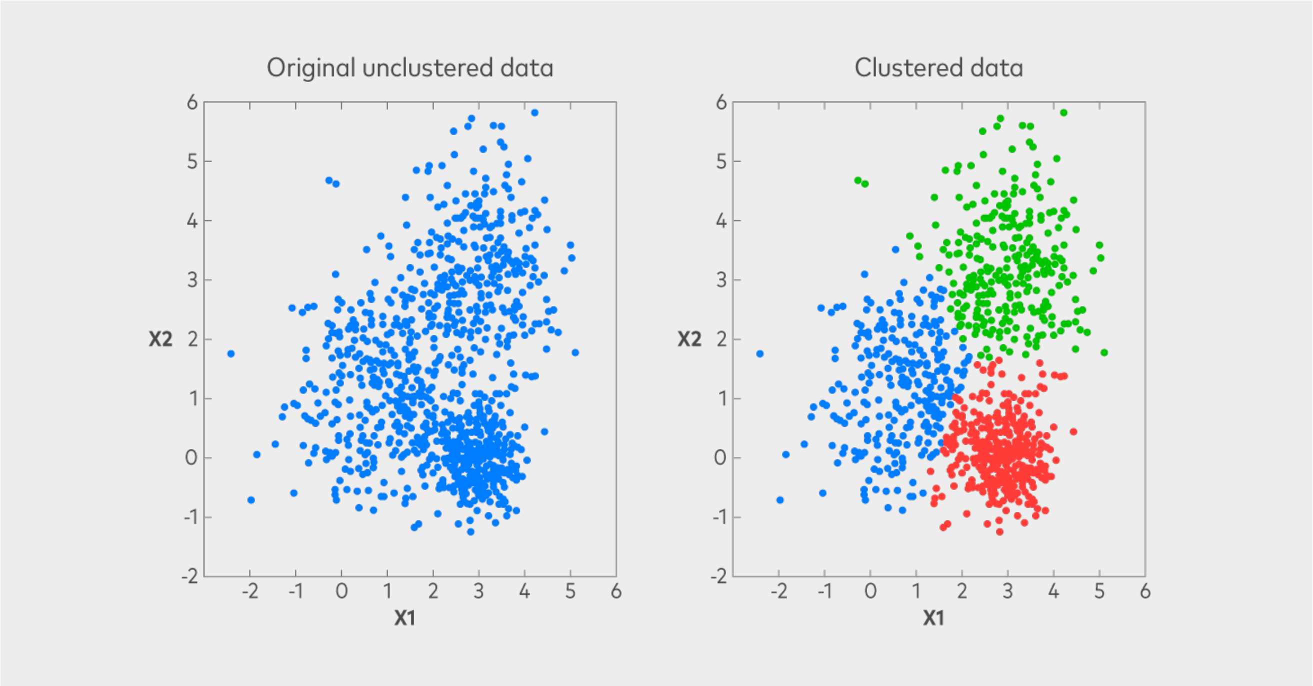
Machine Learning

Once you have met your basic analytics needs, you can tackle advanced applications of data such as machine learning and artificial intelligence. Machine learning can broadly be divided into the following categories: Supervised learning, unsupervised learning and reinforcement learning.

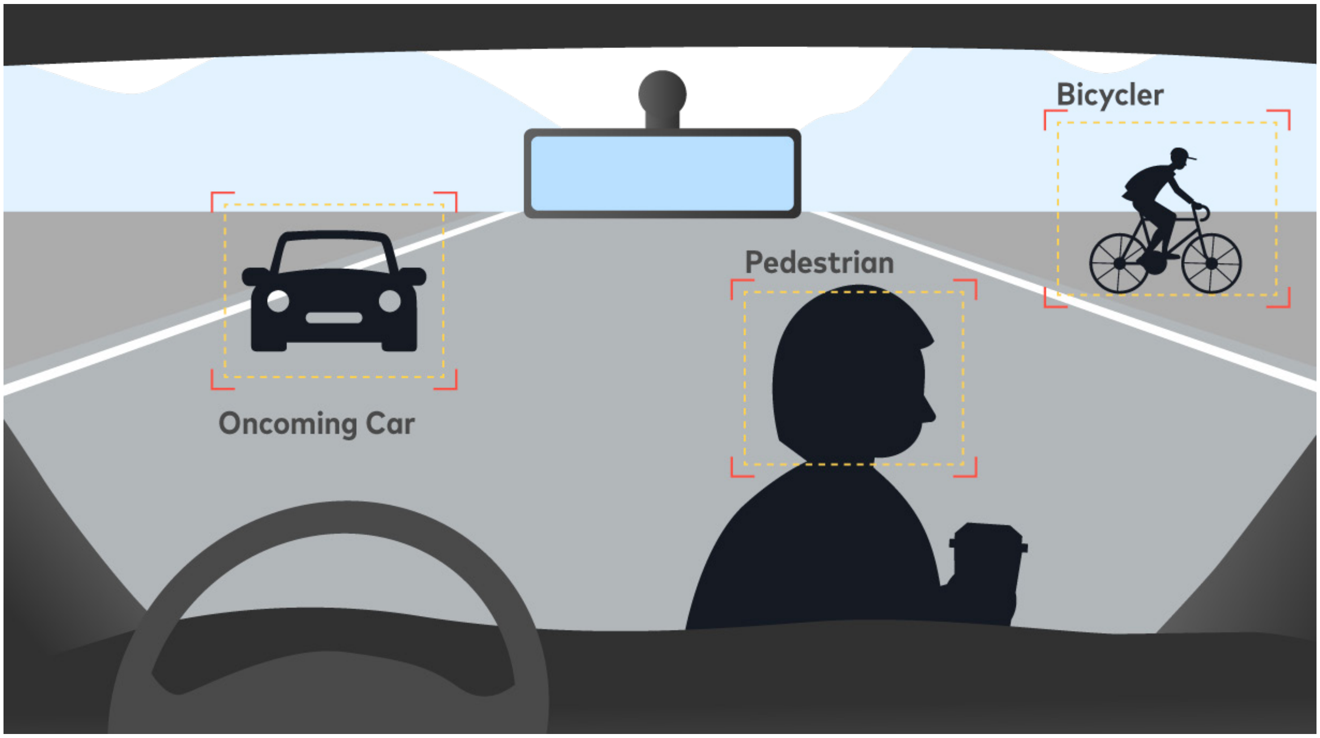
Supervised learning – Data scientists use a training set of known outputs and inputs to produce a predictive model. A simple example is drawing a regression line through some points on a graph. The equation that describes the regression line can be used to predict future values based on known inputs.



Unsupervised learning – Data scientists uncover patterns within a data set using a pattern-recognition algorithm without any previously known outputs and inputs. A common example is dividing data points into clusters based on similar characteristics.



Reinforcement learning – An artificial agent gradually improves its ability to act intelligently through trial and error. Self-driving vehicles are a well-known example (don't worry, the initial training is conducted in simulations, not real traffic), as are game-playing bots such as AlphaStar and AlphaGo.



Business applications of machine learning include revenue and profit projections, predictive modeling to describe the potential tradeoffs and benefits of major decisions, systems to recommend products for customers and all kinds of business process automation.

Regulatory Compliance

A simple but essential transformation is to simply omit or obscure data so that it can't easily be attributed to a person. Needlessly storing personal identifiable information (PII) leaves sensitive data vulnerable to a range of accidental and malicious data breaches, compromising the privacy of your data and creating serious problems both for you and your customers. Privacy is protected by a range of regulations and protocols in different jurisdictions and industries. Examples include HIPAA, SOC2, GDPR and more.

Top 3 Benefits of Getting Data Transformation Right

Reduce latency and time to analysis

An extremely important benefit of a robust data transformation program is reduced time between when something happens, when it is detected and when it is acted upon. An automated data pipeline combined with data transformation means a reduction of time between new data loads and transformations to turn that data into analytics-ready models that power visualizations and reports.

This allows your organization to identify new developments and pivot rapidly in response to them, opening up new opportunities such as new revenue streams. The time savings for analysts and engineers also mean more time spent on higher value business activities, such as building predictive models and products.

Foster a data-driven culture

Building a data-driven culture and democratizing the use of and access to data can empower everyone to make better decisions. There are three actions your organization can take to build a more data-literate culture:

- **Increase the ROI of your data** – As your organization sees results from investing in data, it will foster trust and spur a virtuous cycle in which companies routinely make data-backed decisions.
- **Accelerate analytics** – By reducing the turnaround time between when data is produced and acted upon, you will enable teams to make timely decisions with the freshest possible data, not to mention consistently meet your SLAs.
- **Establish trust in data** – People will trust your data more the more transparent its provenance is. The more easily analysts and non-technical users can follow how the data has been changed and how values have been calculated or derived, the more they will use it.

Magnify the impact of your data team

Data literacy fundamentally enables your organization to determine the relationship between actions and outcomes, and follow beneficial actions while avoiding harmful ones.

An analytics-centric data infrastructure that makes tools easily accessible to analysts and non-technical users means broader-based ownership of analytics workloads, reducing the burden and bottlenecks on engineering teams. In particular, the use of modular, off-the-shelf tools for transformation will make reporting even faster, in some cases enabling analysts to instantly stand up reporting for new sources. Good examples of such tools are Fivetran's pre-built data models (built using dbt Core by dbt Labs) which are designed to create analytics-ready data sets from raw data.

As your organization becomes more agile and data-driven, the common perception of the data team as a cost center will shift to that of an essential and trusted source of guidance and sound decision-making. In the longer-term, it will enable you to pursue more complex and higher value uses of data, such as using machine learning to build automated business processes and smart products.

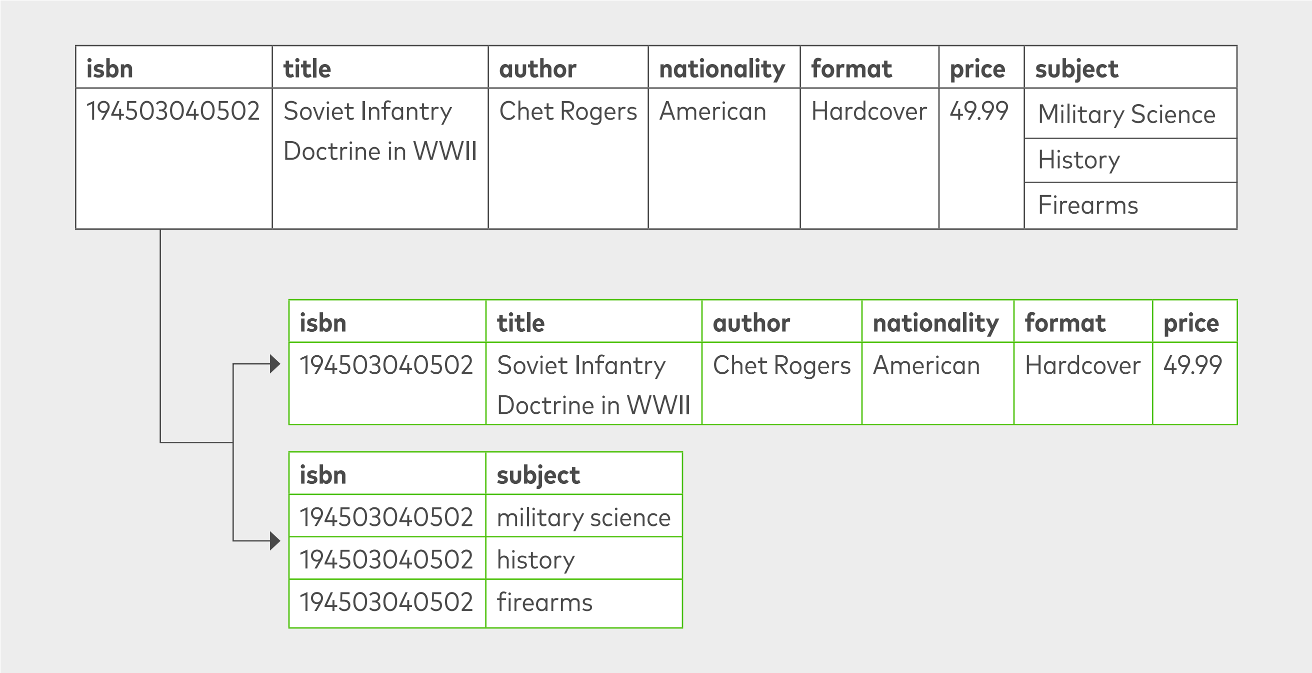
2 Examples of Data Transformation

Raw data is seldom structured or formatted in a way that is conducive to analysis. In this next section, we dive into some common examples of transformations that make data more readily useful. The examples listed below illustrate how additional processing is always required to turn raw data into usable data models.

Revising

Revising data ensures that values are correct and organized in a manner that supports their intended use.

Database normalization is one form of revising data by reducing a data model to a "normal" form without redundancies or one-to-many values in a column. Normalization reduces storage needs and makes a data model more concise and more legible to analysts. However, it is very labor-intensive, requiring a great deal of investigation, reverse engineering and critical thinking.



isbn	title	author	nationality	format	price	subject
194503040502	Soviet Infantry Doctrine in WWII	Chet Rogers	American	Hardcover	49.99	Military Science History Firearms

isbn	title	author	nationality	format	price
194503040502	Soviet Infantry Doctrine in WWII	Chet Rogers	American	Hardcover	49.99

isbn	subject
194503040502	military science
194503040502	history
194503040502	firearms

Data cleansing converts data values for formatting compatibility.

name	breed
Maisie	"NULL"

→

name	breed
Maisie	*NULL*

Format revision/conversion replaces incompatible characters, converting units, converting date formatting and otherwise changing data types.

name	adoption_fee
Maisie	"380"

→

name	adoption_fee
Maisie	380.00

Key restructuring creates generic identifiers out of values with built-in meanings, so they can be used as fixed, unique keys across tables.

name	adoption_id
Maisie	"978-16-1484"

→

name	adoption_id
Maisie	"158bef228a7c6b94aff235faf9d968b8"

Deduplication means identifying and removing duplicate records.

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47
maisie	pibble	07/14/2017	brown	47

→

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47

Data validation evaluates the validity of a record by the completeness of the data, usually by excluding incomplete records.

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47
NULL	*NULL*	*NULL*	merle	62

→

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47

Removing unused and repeated columns allows you to select the fields you want to use as features, i.e. the input variables to a predictive model. It can also improve the performance and overall legibility of a model.

name	breed	color	weight_lbs	weight_kilos
Maisie	pitbull	brown	47	21.4

→

name	breed	color	weight_kilos
Maisie	pitbull	brown	21.4

Computing

A common use case for computing new data values from existing data is to calculate rates, proportions, summary statistics and other important figures. Another is to turn unstructured data, such as from media files, into structured data that can be interpreted by a machine learning algorithm.

Derivation includes simple cross-column calculations.

admissions	applications
345	14556

→

admissions	applications	acceptance_rate
345	14556	0.0237

Summarization consists of using aggregate functions to produce summary values.

student_id	cum_sat
4321	1350
2534	1220
6633	1600
7787	1550
1235	1440
5432	1410
5155	1040
3151	800
6675	930
4515	880
5151	650
5167	610
5566	820
1423	780
6677	680
8897	800

→

statistic	value
max	1600
percentile_75	1365
average	1035
median	905
percentile_25	795
min	610

Pivoting turns row values into columns and vice-versa.

time	activity
1/1/2020	purchase
1/1/2020	return
1/1/2020	purchase
1/3/2020	return
1/3/2020	purchase
1/3/2020	purchase
1/4/2020	purchase
1/4/2020	return
1/4/2020	purchase
1/4/2020	purchase
1/5/2020	purchase
1/5/2020	purchase
1/5/2020	purchase
1/5/2020	purchase
…	…

time	count_purchase	count_return
1/1/2020	2	1
1/2/2020	0	1
1/3/2020	2	0
1/4/2020	3	1
1/5/2020	4	0
…	…	…

Sorting, ordering and indexing organize records in some ordinal manner to improve search performance.

student_id	first_name	last_name
4321	Archibald	Barry
2534	Brittany	Columbus
6633	Chad	Daniels
7787	Desmond	Ephram
1235	Eleanor	Fox
5432	Florence	Graham
5155	Grant	Hammond
3151	Helen	Ines
6675	Isabelle	Jackson
4515	Janet	King
5151	Katya	Luther
5167	Lance	Mondale
5566	Martin	Newman
1423	Nestor	Osbourne
6677	Olivia	Partridge
8897	Peyton	Quinn

student_id	first_name	last_name
1235	Eleanor	Fox
1423	Nestor	Osbourne
2534	Brittany	Columbus
3151	Helen	Ines
4321	Archibald	Barry
4515	Janet	King
5151	Katya	Luther
5155	Grant	Hammond
5167	Lance	Mondale
5432	Florence	Graham
5566	Martin	Newman
6633	Chad	Daniels
6675	Isabelle	Jackson
6677	Olivia	Partridge
7787	Desmond	Ephram
8897	Peyton	Quinn

Scaling, standardization and normalization put numbers on a consistent scale, such as fractions of a standard deviation in Z-score normalization. This allows dissimilar numbers to be compared with each other.

student_id	cum_sat
4321	1350
2534	1220
6633	1600
7787	1550
1235	1440
5432	1410
5155	1040
3151	800
6675	930
4515	880
5151	650
5167	610
5566	820
1423	780
6677	680
8897	800

student_id	cum_sat	sat_z_score	sat_min_max_scaling
4321	1350	0.926417163	0.7474747475
2534	1220	0.544086270	0.6161616162
6633	1600	1.66166888	1
7787	1550	1.514618537	0.9494949495
1235	1440	1.191107781	0.8383838384
5432	1410	1.102877575	0.8080808081
5155	1040	0.01470503434	0.4343434343
3151	800	-0.6911366138	0.1919191919
6675	930	-0.3088057211	0.3232323232
4515	880	-0.4558560644	0.2727272727
5151	650	-1.132287644	0.0404040404
5167	610	-1.249927919	0
5566	820	-0.6323164765	0.2121212121
1423	780	-0.7499567511	0.1717171717
6677	680	-1.044057438	0.07070707071
8897	800	-0.6911366138	0.1919191919

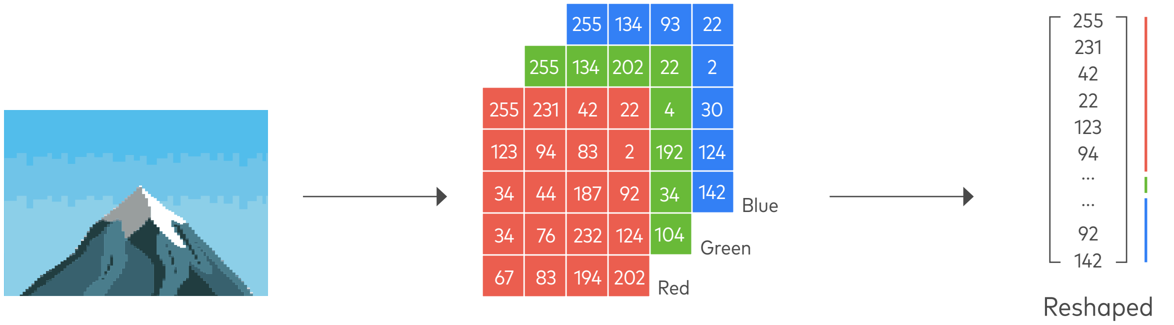
Vectorization converts non-numerical data into arrays of numbers. There are many machine learning applications of these transformations, such as natural language processing (NLP) and image recognition.

One example of vectorization is converting song lyrics into a “bag of words,” or a series of records consisting of word counts.

About the bird, the bird, bird bird bird
You heard about the bird
The bird is the word

phrase	about	bird	heard	is	the	word	you
"About the bird, the bird, bird bird bird"	1	5	0	0	2	0	0
"You heard about the bird"	1	1	1	0	1	0	1
"The bird is the word"	0	1	0	1	2	1	0

Another example is converting an image into a matrix of RGB values that represent the color values of the pixels in the image.



Pixel image

3-channel matrix

Reshaped image vector

Separating

Separating consists of dividing values into their constituent parts. Data values are often combined within the same field because of idiosyncrasies in data collection, but may need to be separated to perform more granular analysis.

Splitting a single column into multiple columns is often used for fields containing delimited values, or for converting a column with multiple possible categorical values into dummy variables for regression analysis.

name	breed_mix
Maisie	pitbull | australian shepherd | labrador retriever | australian cattle dog
Tacoma	husky | pitbull | australian shepherd | australian cattle dog

↓

name	australian_cattle_dog	australian_shepherd	husky	labrador_retriever	pitbull
Maisie	1	1	0	1	1
Tacoma	1	1	1	0	1

Filtering excludes data on the basis of certain row values or columns.

time	activity	location
1/1/2020	purchase	New York, NY
1/1/2020	return	Chicago, IL
1/1/2020	purchase	Atlanta, GA
1/2/2020	return	Atlanta, GA
1/3/2020	purchase	New York, NY
1/3/2020	purchase	New York, NY
1/4/2020	purchase	New York, NY
1/4/2020	return	New York, NY
1/4/2020	purchase	New York, NY
1/4/2020	purchase	Washington, DC
1/5/2020	purchase	Washington, DC
1/5/2020	purchase	Washington, DC
1/5/2020	purchase	San Francisco, CA
1/5/2020	purchase	Chicago, IL
...	...	...

time	activity	location
1/1/2020	purchase	New York, NY
1/3/2020	purchase	New York, NY
1/3/2020	purchase	New York, NY
1/4/2020	purchase	New York, NY
1/4/2020	return	New York, NY
1/4/2020	purchase	New York, NY

Combining

A common and important task in analytics is to combine records from across different tables and sources in order to build a full picture of an organization's activities.

Joining is the act of linking data across tables.

id	name	city
1337	Elite Academy	New York
8455	Lakeside Academy	Chicago
4377	Armitage High	Saint Louis
8088	Mountaintop High	Denver

id	acceptance_rate
1337	0.67
8455	0.23
4377	0.45
8088	0.56

id	name	city	acceptance_rate
1337	Elite Academy	New York	0.67
8455	Lakeside Academy	Chicago	0.23
4377	Armitage High	Saint Louis	0.45
8088	Mountaintop High	Denver	0.56

Merging, also known as appending or union, combines records from multiple tables. By blending the two tables using a common column, such as "email" in the example below, you can assemble parts of the sales and marketing funnel. This is also an example of integration, which consists of reconciling names and values for the same data element across different tables.

marketing_id	email	first_touchpoint
pgie13373	alastair@gothcorp.com	9/17/2021

sales_id	email	first_touchpoint
vboi4044	alastair@gothcorp.com	10/2/2021

customer_id	email	first_touchpoint	platform
pgie13373	alastair@gothcorp.com	9/17/2021	marketing
pgie13373	alastair@gothcorp.com	10/2/2021	sales

Transformations themselves are a key part of a broader process called data integration, without which analytics and data science are impossible.

3 How Data Integration and Data Transformation Enable Analytics

Data integration is an all-encompassing term for the process that starts with moving raw data from source to destination and ends with a unified, actionable view of an organization's data. Data transformation refers to the part of the process that turns raw data into models suitable for dashboards, visualizations and training sets. Data transformation is essential to data integration. Without end-to-end, synced data integration and transformation workflows, organizations can't build a mature, sustained analytics practice.

The modern data stack (MDS) is a suite of tools used for data integration. These tools make up three layers:

1. **Data pipeline** – to move data from sources to a destination. Data pipelines often include a data transformation tool to turn raw data into usable data models.

2. **Data warehouse** – to serve as a central repository for data.

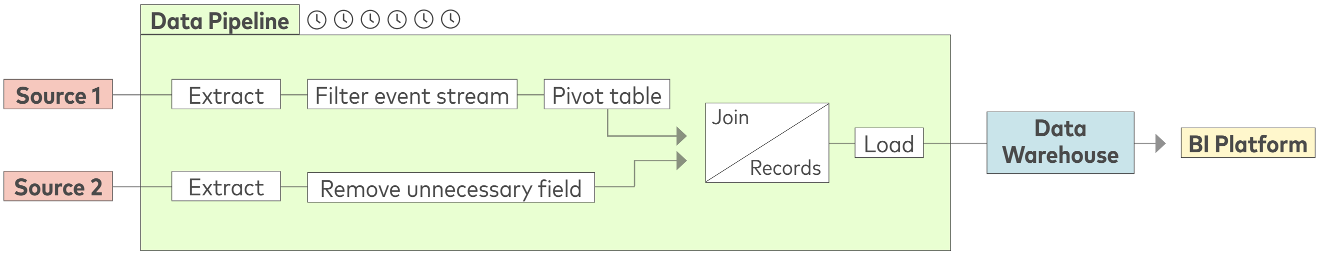
3. **Business intelligence platform** – to create visualizations and dashboards so that data is easily presented to people.

There are two major data integration architectures – extract-transform-load (ETL) and extract-load-transform (ELT). These architectures reflect the fact that data can be transformed at two stages of the data integration process. Transformation can take place **before** the data is loaded to its destination (ETL) or **after** (ELT). The destination is typically a data warehouse.

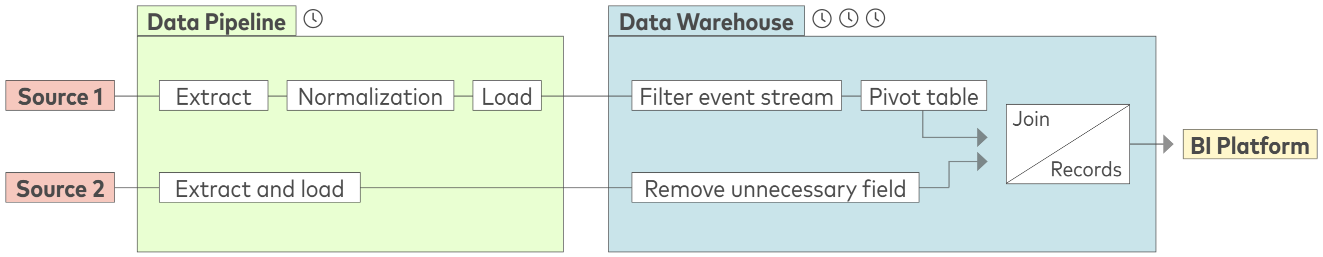
In traditional ETL, data is transformed into analysis-ready data models before it is loaded to the destination. Combining transformation with loading into the same step can preserve storage and compute resources but introduces a great deal of brittleness into the data engineering

workflow. It also circumscribes what the data can be used for, preventing it from being used to answer new or ad hoc questions and requiring new pipelines to answer new questions.
ETL makes transformation an engineering-intensive process, performed by software written in scripting languages such as Python and Java. In addition, since transformations often involve carefully sequenced, scheduled and coordinated workflows to work properly, transformations in ETL may require a great deal of complex orchestration performed within the pipeline.

In the diagram below for ETL orchestration, the transformation process takes place in a separate environment from either the source or the destination. Just two sources can require quite a bit of massaging in order to be combined into a usable data model. You can easily imagine how, as more sources are added and the data model becomes more comprehensive and complex, the path dependencies start to stack up. These workflows must be built using scripting languages and constantly tested for reliability and performance. As a result, transformations under ETL are bespoke, hand-built solutions that require the involvement of expert users, specifically engineers and data scientists.



By contrast, in the modern ELT workflow, raw data is transformed into analysis-ready data models within the data warehouse environment, after data is loaded. This means that transformations can be performed using SQL, the common language of relational databases, instead of scripting languages. SQL-based transformations make data modeling broadly accessible to analysts and other SQL-literate members of your organization, rather than only engineers, data scientists and other people with serious coding chops. Specifically, it makes data modeling accessible to stakeholders in your organization without a huge development effort.



In addition, transformations performed within the data warehouse can be represented as views, which are tables generated on-the-fly from queries, or materialized views, which are tables prepopulated with the results of a query and physically stored on disk. Since both the raw and processed data, as well as the relevant queries, are accessible on the same platform, it's easy to track the provenance or lineage of every data model. This makes it easier to reproduce analyses as well as foster general trust in the data team's analysis.

4 Getting Started With Data Transformation

Data transformation is a complicated task that calls for a thoughtful and systematic approach. With the right people, processes and technology, data transformation will make your data integration workflow smoother and empower your analysts, data engineers and data scientists to do their best work. Here are some best practices that will help set your data integration efforts on a sustainable path to success.

Keep Transformations Data Warehouse-Based and Modular

Staging transformations in the data warehouse allows raw data to be moved directly from source to destination and places control over transformations squarely in the hands of analysts. Specifically, this means:

1. **Compatibility with ELT architecture** – transformations can be performed within the data warehouse so that they're decoupled from the data engineering process and can be performed by analysts. This turns transformation from an engineering- or IT-centric activity into an analyst-centric one.

2. **SQL-based data modeling** – SQL is the universal language of relational databases and analysts. The alternatives are scripting languages like Python and R, which require a much higher barrier to entry, or drag-and-drop GUIs, which lack granularity and create vendor lock-in.

Transforming your data in SQL gives you considerable flexibility, but SQL can be slow to write, especially for beginners. Many metrics from common data sources are known and solved problems for which off-the-shelf, pre-built solutions exist. Off-the-shelf, pre-built solutions can be used and reused over and over, saving development resources while freeing up more time for analysis. The ability to instantly generate reports can tremendously accelerate your analytics.

Transformations should also be portable between other elements of the modern data stack. To avoid brittle workflows and vendor lock-in, your transformation tool should not be closely coupled with any other technologies and should support:

1. **Software development best practices** – Transformation tools should specifically support collaboration, testing and version control. A collaborative, repeatable process introduces some permanence and legibility to your transformations.

2. **Separation from a dedicated BI or visualization tool** – While BI tools may support transformations in a number of ways, they are seldom ideal and are often proprietary to the platform.

3. **Plug-and-play packages** – Pre-built transformations that you can use off the shelf can speed up development and reduce time to analysis and time to value. If they are open-source, they can also leverage the experience and expertise of many other people. This is related to the SQL-based data modeling above — the universality of SQL as a query language makes it especially useful in this regard.

4. **Automated scheduling and workflow orchestration** – You should be able to design relatively complex orchestrations in transformations separately from the engineering workflow.

Documentation, Documentation, Documentation

Data engineering time is scarce, yet data engineers spend far too much time and effort responding to data questions that can easily be answered by well-maintained, up-to-date data lineage and documentation. Data lineage and documentation can provide data analysts with a comprehensive view of how the data progresses from its raw form to its analytics-ready state, while improving data literacy across the organization. This enables data engineers to spend less time explaining their logic and allows data analysts to easily understand the source and dependencies of the data they are working with.

Join a Community of Experts

Your data team may be extremely lean. Whether you work in an enterprise or startup, you can find yourself and a few data professionals responsible for a mammoth project. We highly recommend finding a community of experts so you can lean on your peers for additional support and best practices. You can join Slack Communities such as those for dbt and Locally Optimistic to ask questions and share your best practices.

Fivetran Transformations accelerates data analytics with simple pre-built packages to model your data in minutes - not days or weeks. To start a free trial with Fivetran Transformations and see a full list of all data models supported by Fivetran, visit www.fivetran.com/transformations.

Fivetran is the global leader in modern data integration. Our mission is to make access to data as simple and reliable as electricity. Built for the cloud, Fivetran enables data teams to effortlessly centralize and transform data from hundreds of SaaS and on-prem data sources into high-performance cloud destinations. Fast-moving startups to the world's largest companies use Fivetran to accelerate modern analytics and operational efficiency, fueling data-driven business growth. For more info, visit Fivetran.com.