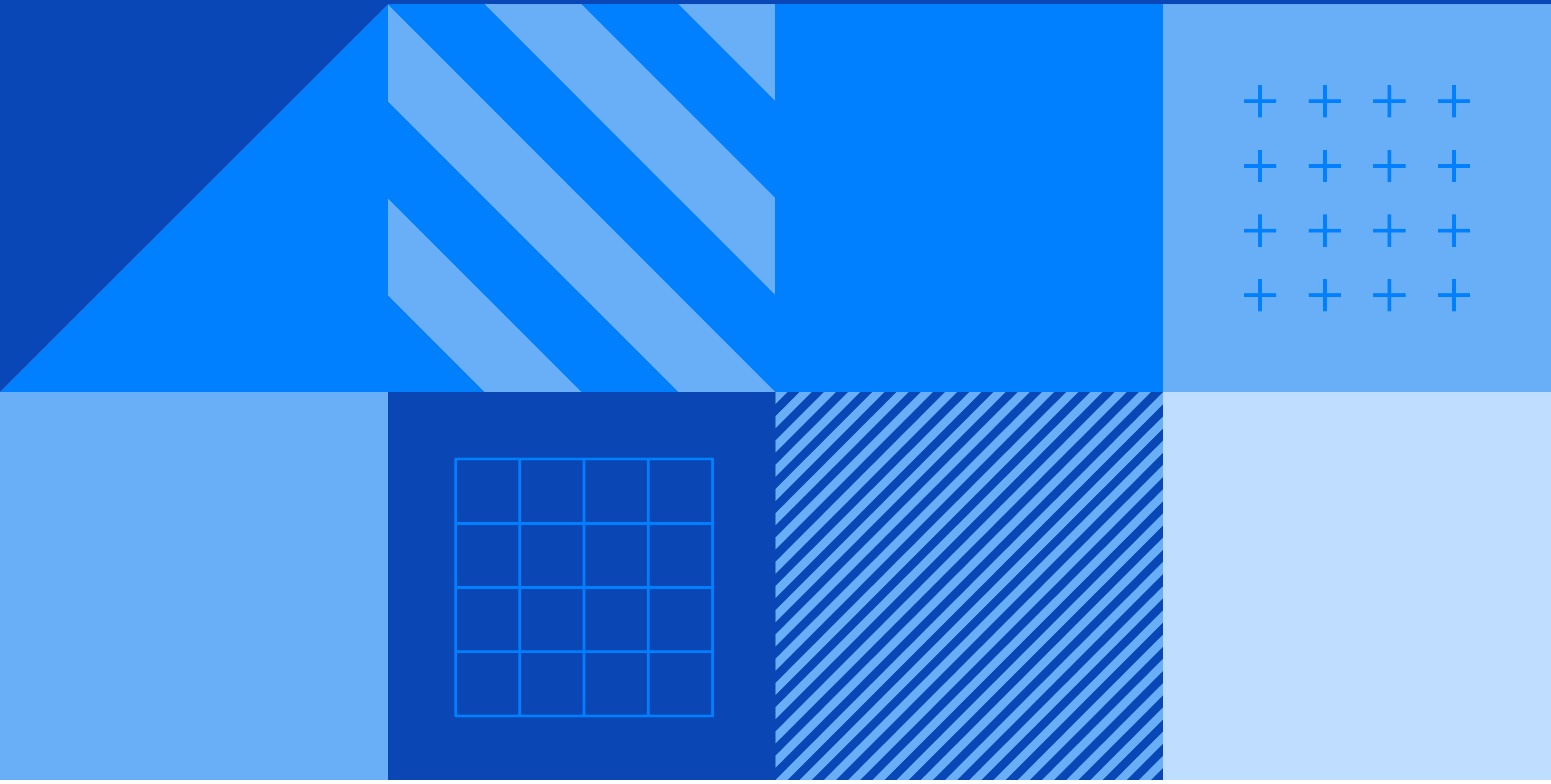
The Data Professional's Guide to Data Integration: How to Build a Modern Data Stack

Mastery over data gives you a serious competitive advantage. Here's how to bring any organization to the forefront of data science.



Fivetran

Highly competitive companies are fueled by data. Modern businesses use data to rapidly adjust to the market and create new products. Companies that are data-driven also give their customers personalized experiences and make digital interactions more fulfilling for everyone, including employees.

But building a data-driven company requires a number of internal processes and a suite of tools. What comes first?

Executive Summary

You should understand the following points after reading this guide:

1. **Why You Need Data Science and Data Integration** – Data science can help your organization make better decisions, discover new opportunities, and eventually build smart, data-driven systems.

   The first step is data integration – a solid foundation of data collection and modeling.

2. **Three Essential Steps to Getting Started With Data Integration** – To build this foundation, you must:

   a. Establish internal systems of record to resolve conflicting data.

   b. Use an automated data pipeline and cloud data warehouse to consolidate the data in a single platform.

   c. Create data models to make sense of your business's operations and support the creation of reports and dashboards.

   This requires a suite of tools called the modern data stack.

3. **How to Build a Modern Data Stack** – To build a modern data stack, you will need a cloud data warehouse, data pipeline, a means to transform data, and business intelligence tool.

   a. **Cloud data warehouses** tend to be similar in price and performance for most common use cases. The real choice is between configurability and ease of use.

   b. **Data pipelines** should follow an ELT (extract-load-transform) architecture and be easy to use. Choose one that supports transformations within the data warehouse.

   b. **Transformation** are often featured in data pipelines or business intelligence platforms. They are essential to massaging data into models that represent important business metrics.

d. **Business intelligence platforms** allow you to turn data models into reports and dashboards

4. **Your First Analytics Project** – Understanding your customers is fundamental. Start with marketing and sales analytics. Then move on to analytics for product, operations, engineering, and so on.

5. **Productionizing/Operationalizing Analytics Data** – As your organization's data capabilities continue to mature, you will need to move analytics data back into operational systems to build predictive models, smart products, and personalized customer experiences.

Let's discuss these topics in further detail.

1.

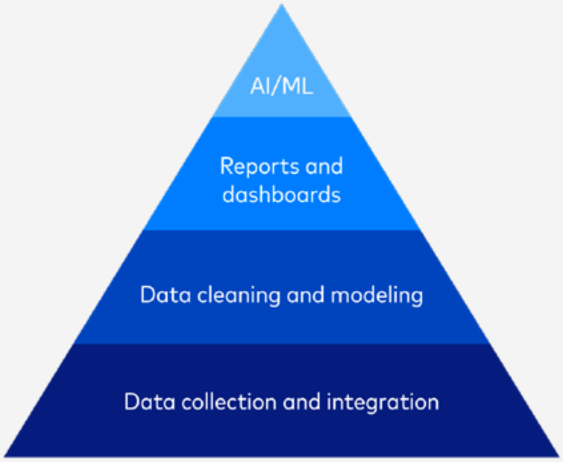
Why You Need Data Science and Data Integration

The purpose of **data science** is to use data to support better decision-making — and ultimately to build systems that can combine data-driven decisions with automation. **Data integration** consists of all the activities that support data science, specifically moving data from sources to a central repository and readying it for analysis.

Leaders in every business are excited to bring cutting-edge data science to their industry, and in the last decade, data science has regularly topped lists of most desirable careers.

But many businesses hire data scientists before they are ready to work with them. These highly paid team members end up spending their time doing data integration (which is better left to data engineers) and reporting (which is better left to analysts).

What does it take to be ready to hire data scientists? To answer this question, we need to understand the **data hierarchy of needs**:



You must fulfill lower levels of the data hierarchy of needs in order to access higher ones.

Data Collection and Integration

The base of the pyramid requires a solid enterprise **data warehouse** — all of your data should be in one place. A modern cloud-based data warehouse ought to contain a faithful, up-to-date replica of all the data in your business systems. Taking this replication-first approach gives you a “single source of truth” that can support all your data questions, today and in the future.

The tools and processes for moving data from your sources to your data warehouse make up a data pipeline. Building a **data pipeline** is a complicated, engineering-intensive activity that you should outsource or automate as much as possible. With the labor savings from outsourcing and automation, your data team can focus on data modeling, analysis, and visualization, which has a far greater potential to add value to your business.

Data Cleaning and Modeling

The next level of the pyramid depends on a clean, curated view of all your data. Real-world data is messy, and the same concepts are often duplicated among systems. For example, you may have customer data in both your CRM system and your accounting system — and there may be small contradictions between these systems. You'll have to decide which system is the system of record for each concept.

Your analysts will use SQL queries to resolve these contradictions and convert the raw tables delivered by your data pipeline into data models, or a simplified view of your data that provides the foundation for everything else you do.

Reports and Dashboards

The data models your analysts create enable classic business intelligence and analytics. These are the spreadsheets, reports, and dashboards that provide day-to-day decision support for your managers and leaders. Data science may get more attention, but traditional business intelligence is still the foundation of using data to make decisions. This type of work is also done by analysts, whose primary tools are SQL and data visualization platforms, such as Tableau or Looker.

Predictive Modeling, Machine Learning, and AI

Data scientists should be used for the tip of the pyramid. If you've hired well and done a good job building the lower levels of the pyramid, your data scientists can use their specialized skills in advanced statistics to build predictive models and machine learning products, leveraging the data integration and cleanup that has already been done by your analysts. These efforts will eventually culminate in machine learning and artificial intelligence.

Data scientists are the star athletes of your data team. Like the starting lineup of the Golden State Warriors, they have highly specialized skills and are supported by a much larger cast of teammates. A typical NBA basketball team will have five starters, 10 other players on the roster and thousands of other employees throughout the organization. You wouldn't want Steph Curry answering phones in the front office — nor should you have your data scientists doing traditional analyst or data engineering work. They're perfectly capable of doing this work, but it's not the right way to use your most valuable players.

So before you hire a team of star data scientists, make sure you build the foundation that will make them productive.

2.

Building a Foundation for Data Science: Three Essential Steps

Longtime data practitioners have developed a set of best practices to fulfill the first two levels of the hierarchy of needs: data collection and integration, and data cleaning and modeling. These practices apply to every company and consist of three main principles:

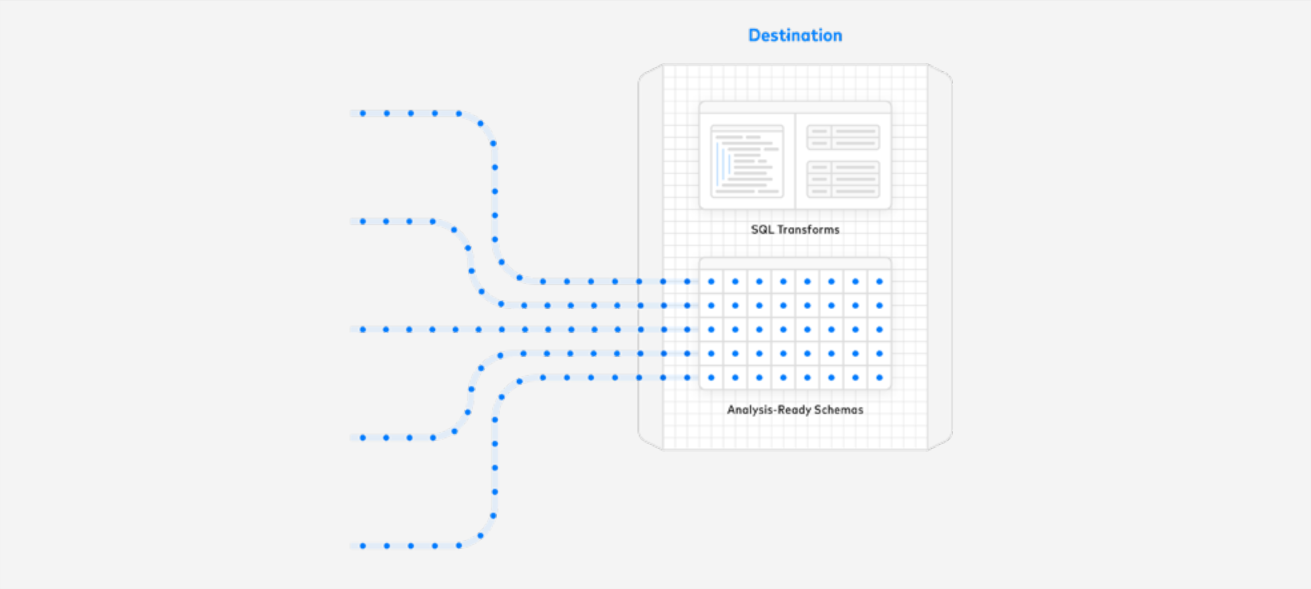
Agree on an internal system of record

A complex business will inevitably have multiple copies of the same data stored in different systems. For example, sales, marketing and support departments all collect and store customer data, and some of it is redundant.

The key to avoiding an explosion of complexity, due to disagreement among different "sources of truth," is to designate a single location as the "system of record" for each concept. For example, you might designate Salesforce as the system of record for the definition of a customer. If the data in Zendesk disagrees with Salesforce about whether someone is a customer, Salesforce is right and Zendesk is wrong. By making it clear where "home" is for each concept, you provide clear direction for your team to resolve discrepancies and fix bugs.

You must fulfill lower levels of the data hierarchy of needs in order to access higher ones.

Consolidate data in a cloud data warehouse



A data warehouse is a relational database that serves as a central repository and "single source of truth" for a company or business unit.

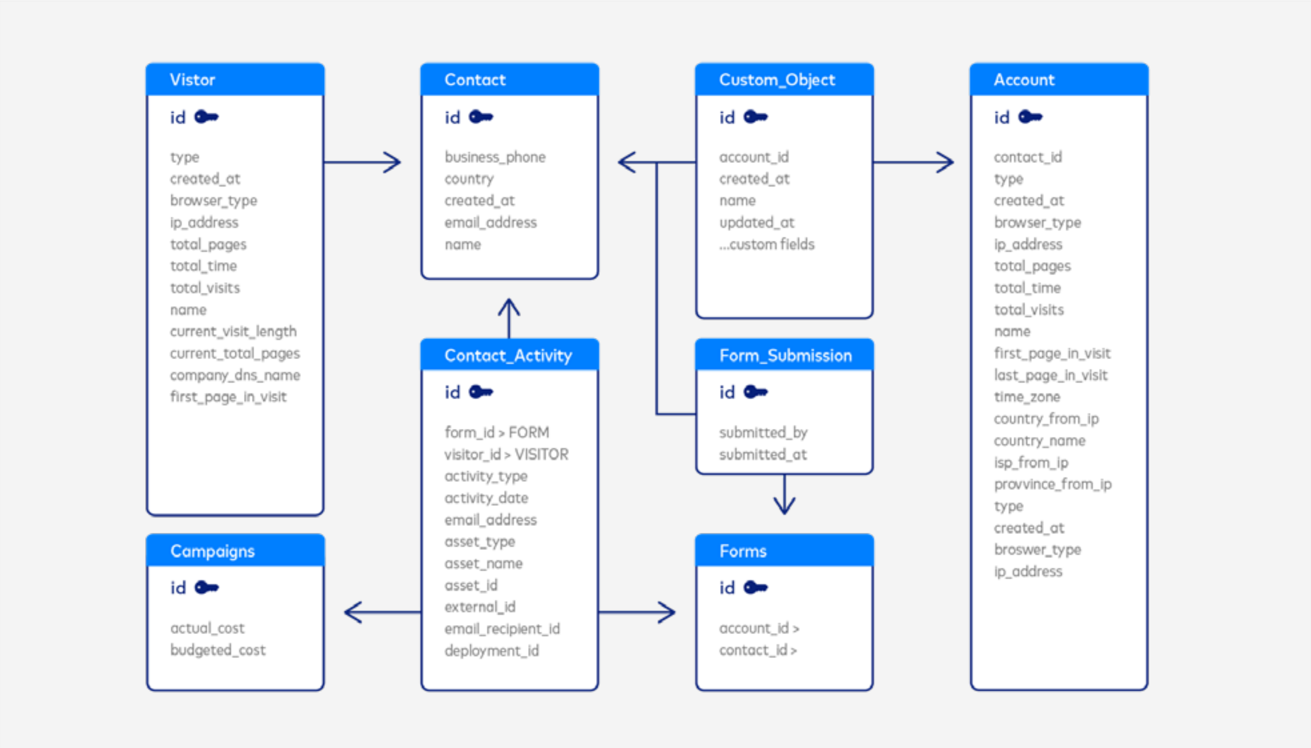
Before you can use data to make decisions, you will need to consolidate that data in a single location. That location should be a **cloud-based data warehouse**. Data warehouses are the Swiss army knife of data. They excel at the most common analytics tasks: querying structured, tabular data. They are also good enough at the less-common workflows — you will be able to solve your daily data problems with a minimum of complexity.

You will use an **automated data pipeline** to move data from sources such as SaaS apps, operational databases, event trackers, files, and data feeds. A data pipeline is a platform that uses connectors to move data from your data sources to your data warehouse. The most basic function of a data pipeline is to extract and load the data; it may also clean and normalize the data to make it more manageable for analysts.

Create a solid dimensional schema

For smaller companies with simple data problems, steps 1 and 2 are enough to answer most questions you will ask while you make data-driven decisions. But larger companies will have enough complexity within their data model that the data warehouse will begin to turn into a mess. There is no magic bullet to solve this problem, but the tried-and-true approach is to build a dimensional schema.

A dimensional schema is a simplified view of all your data. Any large, longstanding business will have many systems of record, each of which has many tables. The resulting data warehouse will have an intimidating level of complexity. You will mitigate this problem by designing a dimensional schema that has fewer tables but can still support the questions you need to answer.



A dimensional schema is a company-wide data model that resolves conflicting information and organizes data into tables that are easily represented to stakeholders as reports or dashboards.

For example, analysts in several departments may be interested in analyzing your customers for a variety of reasons. Information about customers comes from several different sources and must be merged and cleaned before it is ready to analyze. You don't want multiple analysts to duplicate their work and come up with different representations of a customer. Instead, you want the merging and cleaning to be done once by a core team of data analysts, resulting in a common customer table or view that the rest of the analysts in the company can use as a starting point.

A dimensional schema is in some ways a work of fiction. By simplifying the data models of your systems of record, you are erasing some of the real-world complexities of your business. There will be some questions that can’t be answered by this simplified view. The key to designing a good dimensional schema is to erase as much complexity as possible, while still supporting the queries you actually need. You should expect your dimensional schema to evolve as you change the questions you need to ask of the data.

The dimensional schema exists in addition to raw data from the source. Make sure you retain the original data in its original schema so it will be available in the future.

The most reliable way to produce a dimensional schema, as well as all subsequent data models, is to use a SQL-based **transformation tool** that sits on top of the data warehouse and supports collaboration and version control.

Once you have a handle on data collection and integration as well as data cleaning and modeling, you will be prepared to ascend to the next rung of the data hierarchy of needs: reports and dashboards. To make your dimensional schema comprehensible to analysts and stakeholders, you will need a **business intelligence platform** to visualize your data.

You Will Need a Modern Data Stack

In the course of this chapter, we have mentioned in passing the tools necessary to build a solid foundation for data science. Put together, this suite of tools is called the **modern data stack**, which consists of:

1. Automated data pipeline
2. Cloud data warehouse
3. Transformation tool
4. Business intelligence platform

These tools can generally be purchased off-the-shelf, sparing your company the engineering effort of building and maintaining them.

3.

How to Build a Modern Data Stack

While high-tech data engineering may be fashionable, 99 percent of the time it is not necessary. What most companies need is a modern data stack, and a leadership team that is ready to change their opinions in the face of new data. The modern data stack can be assembled using several off-the-shelf technologies.

Key Business Considerations

For each element of your data stack, you should consider the following business factors:

1. Pricing and costs – how much, and is it based on usage, or flat by time interval? This includes the cost of ongoing maintenance, as well as of infrastructure and other capital.
2. Fit to team's skills and future plans – does your team have the skills to use this tool?
3. Vendor lock-in and future-proofing – is the tool continuously upgraded? Can it be easily replaced if it becomes unavailable or obsolete?

These considerations have more to do with how you plan to grow and sustain your organization in the future than the technical characteristics of each tool.

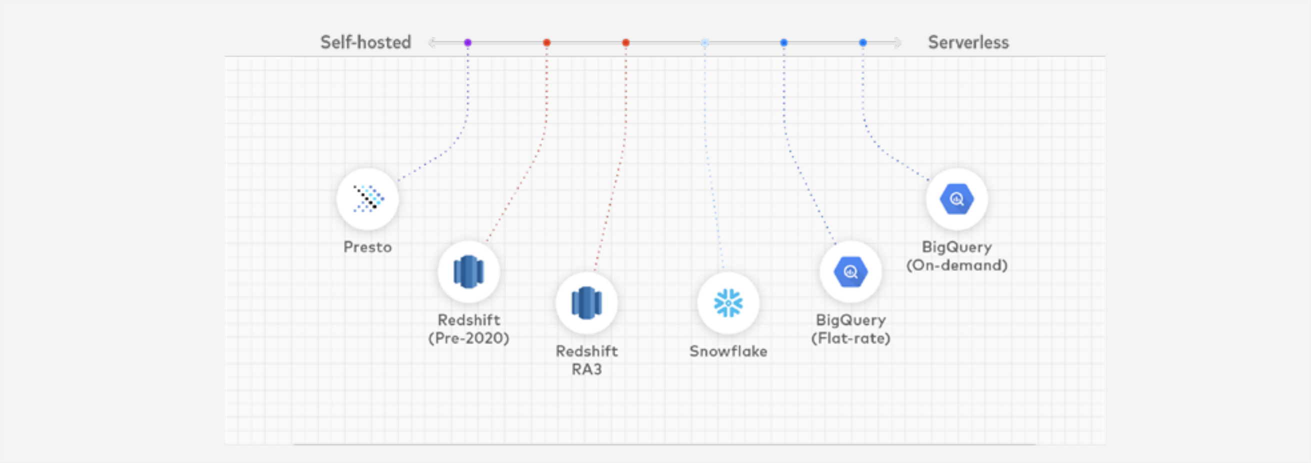
How to Choose a Cloud Data Warehouse

At Fivetran, we are often asked, “What data warehouse should I choose?” The key is to choose a cloud-based data warehouse. According to our benchmarks, common modern cloud data warehouses such as Amazon Redshift, Snowflake, Presto, and Google BigQuery are virtually tied for cost and performance.

Based on tests that represent typical Fivetran use cases, we have found that these common cloud-based warehouses all have excellent execution speed and are basically similar in price and performance. This isn’t surprising: Fast columnar data warehouses are now a mature technology. Be skeptical of claims that one data warehouse is dramatically faster than another.

How Are the Warehouses Different?

Warehouses do differ by user experience and pricing models. We can place them along a spectrum:



These common data warehouses represent the range of configurability. Typically, configurability is traded off against ease of use.

On the "self-hosted" end of the spectrum is Presto, where the user is responsible for provisioning servers and detailed configuration of the Presto cluster. Presto is open-source, unlike the other commercial systems in this benchmark, which is important to some users.

The other end of the spectrum is BigQuery on-demand, a pure serverless model, where Google spins up storage and compute resources as needed and the user pays per query.

The most important differences among warehouses are the qualitative differences caused by their design choices: Some warehouses emphasize tunability, others ease of use. Tunability often costs more labor and less money; ease of use often costs more money but less labor. If you're evaluating data warehouses, you should demo multiple systems, and choose the one that strikes the right balance for you.

What About Data Lakes?

A data lake is similar in purpose to a data warehouse, but they are designed to accommodate both structured (i.e. tabular) and unstructured data, such as sensor feeds, media files, and other data that isn’t easily represented in tables.

IT consultants and vendors often recommend a two-tier system, in which a data lake is a storage repository and staging area for a data warehouse. Don’t take such a recommendation uncritically; you may be able to simplify your architecture by using a data warehouse as a multifunctional **data** "**lakehouse**."

Modern cloud data warehouses often offer unlimited storage capacity and scalability, increasingly support semi-structured data like JSON and XML, and are more easily governed than a two-tiered system.

That said, you may still need to store truly unstructured data (images, audio, and video) in object storage (in S3, for example). A best practice is to store the unstructured files in object storage and then store metadata in a data warehouse, with references to the locations of the files in your data warehouse tables.

At the end of the day, it's worth your time to run the numbers before you follow a consultant's recommendation for a data lake. In most cases, combining your data lake and data warehouse will give you a simpler system, with lower end-to-end latency, for less money.

How to Choose a Data Pipeline

A data pipeline is a process that routes raw data from a source to a destination. Off-the-shelf automated data pipelines can radically simplify your data integration workflow. To this end, when you're in the market for a data pipeline, the most important considerations include:

- **ELT architecture** – A longer discussion about the differences between ETL (extract-transform-load) and ELT (extract-load-transform) is beyond the scope of this guide, but in short, ELT is a more modern approach that leverages advances in cloud technology. Transformations and data modeling should generally be performed within the data warehouse after the data has already been extracted and loaded.

- **Automation and ease of use** – Using the tool should involve as little coding and configuration as possible.

- **Connectors supported** – Does the data pipeline support the data sources you use? The entire purpose of an automated data pipeline is to save you the hassle of building and maintaining a complex engineering product.
- **Completeness of data** – Make sure the data pipeline captures all of the data models from your sources.
- **Support for transformations** – How does the data pipeline allow you to turn data into dimensional schemas and other data models for analytics? The ideal solution is a data pipeline that either features or easily interfaces with a SQL-based transformation tool that transforms data within the data warehouse.

There are a number of additional technical considerations, but you can leave them to your data team.

How to Choose a Business Intelligence Tool

Business intelligence tools offer your analysts an easy, replicable way to construct reports and dashboards. They can vary widely in terms of feature sets, capabilities, and pricing. Look for the following characteristics:

- **Seamless integration with cloud data warehouses** – Your BI tool must be compatible with your cloud data warehouse.
- **Modeling layer with collaboration and version control** – It should be easy for your analysts to share code and data models, and iteratively develop off of each others' work.

- **Ease of use and drag-and-drop interfaces** – You want to promote organization-wide data literacy, so make sure your BI tools are easy to learn.
- **Speed, performance, and responsiveness** – Your analysts will be happier if dashboards, visualizations, and data models load and refresh quickly.
- **Automated reporting and notifications** – Writing reports by hand is tedious. Does the BI tool allow you to schedule reports to publish automatically? What about alerting users when the data changes?
- **Ability to ingest and export data files** – Your analysts and data scientists may sometimes want to explore data on an ad hoc basis without the overhead of having to go through a data warehouse first.
- **Support for visualizations** – Pie charts, column charts, trendlines and other basic visualizations can only take you so far. Does the BI tool feature more specialized visualizations like heat maps or radar charts? Does it allow you to build your own custom visualizations?

How to Choose a Business Intelligence Tool

The ability to transform data will usually be featured either in a BI platform or by your data pipeline provider. Make sure it's present somewhere in your stack, and that your analysts are comfortable using it. Some important features to consider include:

1. SQL-based modeling. SQL is the default language of analysts, and
2. Transforming data within the cloud data warehouse rather than as part of the extraction and loading workflow.

3. Version control and collaboration between analysts.

4. Automated scheduling and workflow orchestration

The more of the data orchestration and transformation process the tool automates, the better.

For all elements of the modern data stack, it pays to carefully review a range of perspectives on different tools so that you can independently evaluate the tradeoffs involved. Research firms such as Gartner often aggregate this information. Read before you leap!

Migrating From a Legacy System (or Not)

There is a good chance that your organization already has a legacy data stack. You may even use on-premise infrastructure and maintain your own data center.

Data pipeline providers should be able to help you migrate data from old infrastructure to your new data stack, but the task is often a hassle because of the complexity of data and the need to keep operations running while the migration is in progress. You may opt to start from scratch instead, depending on how important historical data is to you.

Regardless, excepting some very specific use cases, you should adopt a modern, automated, cloud-based infrastructure. Moving to the cloud offers:

- **Improved scalability** – Cloud-based infrastructure doesn't require you to purchase, maintain, and spin up (and down) machines as needed.

- **Ease of use** – Cloud-based technologies are generally built to be usable off-the-shelf, saving you the engineering time of building, configuring, and maintaining bespoke systems.

- **Accessibility** – Cloud-based infrastructure can be accessed from anywhere, not just your private intranet, leading to less downtime.

- **Cost savings** – Since you won't have to spend on physical equipment, utilities, maintenance, and other pertinent expenses, your operating costs will be significantly minimized. Also, when your business resources are in the cloud, if there is a need for extra hardware capacity, resources can be easily spun up as needed without having to procure new hardware that would sit idle during times of less demand.

Cloud migration is more than just a matter of moving data from one place to another. You will also have to move data models and stored procedures. Moreover, minor incompatibilities and other idiosyncrasies between different platforms can add up. Consider enlisting the services of a systems integrator to help you.

4.

Your First Analytics Project

Once you have determined your internal systems of record and chosen a data warehouse, you should use a data pipeline to connect data from your sources to it. Your analysts can then produce data models to support dashboards and reports for various analytical pursuits.

Sales and marketing are the most obvious areas to begin with analytics. With the vast number of brand choices presented to consumers today, customer loyalty can be elusive. But the more you know about your customers and their buying motives, the more you can anticipate their needs and influence their purchase decisions. To get the information you need, you must collect and analyze a lot of data on your customers — data that's typically stored in line-of-business applications across your company.

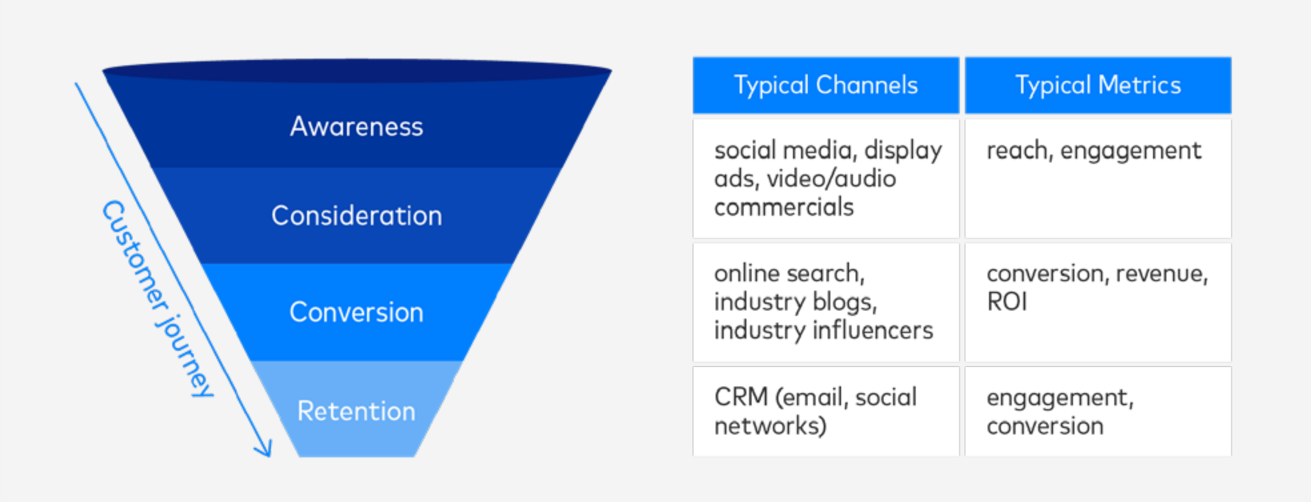
Salesforce, NetSuite, Zendesk, Shopify, Asana and thousands of other applications each contain data on your customers. Some of that data is shared across applications, and some is unique to individual applications. When you aggregate and analyze such varied data, you get a customer 360 view that allows you to grow your relationships with your customers and generate more revenue.

Some important metrics you should track early include:

- Revenue metrics
    - 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
    - Customer segmentation

Marketing analytics can also help you quantify the various steps in your marketing and sales funnel.



Typical Channels	Typical Metrics
social media, display ads, video/audio commercials	reach, engagement
online search, industry blogs, industry influencers	conversion, revenue, ROI
CRM (email, social networks)	engagement, conversion

The marketing and sales funnel is a common model for understanding how prospects turn into customers. You should be able to quantify every stage of the funnel.

These metrics should be turned into reports and dashboards that are accessible to all executives, sales, and marketing people within your organization. With the business climate and customer demand becoming increasingly dynamic, competitive, and unpredictable, you need to generate actionable insights that let your company change course quickly.

Beyond Marketing and Sales

While sales and marketing analytics are bread-and-butter use cases for many companies, there are other common applications as well, such as:

- Product analytics
- Operational analytics
- Financial analytics
- Support analytics
- People analytics
- Engineering analytics

As your company continues to grow in data literacy and capability, you will be able to use data to power better decisions across a wide variety of teams and business units.

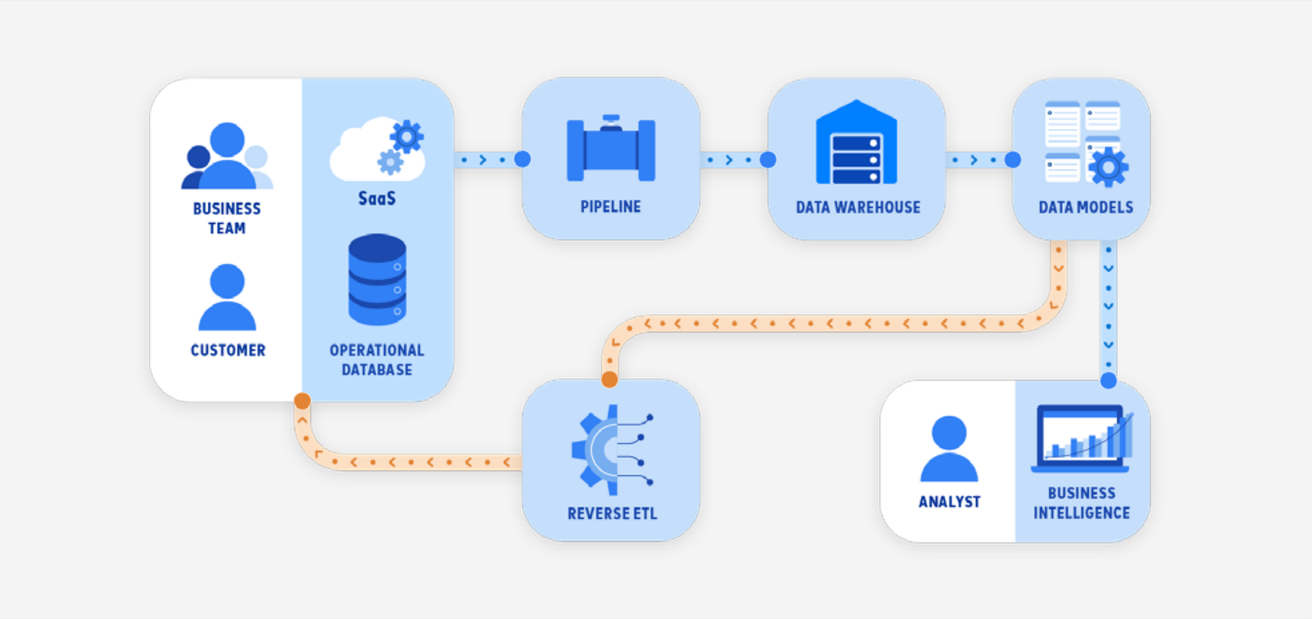
5.

Productionizing/Operationalizing Analytics Data

With your infrastructure in place and your reporting and dashboards capabilities maturing, you will be ready for actual data science.

The easiest way to make your analytics data useful to the rest of the company is to expose it. It can be as simple as an account manager dashboard that shows you what accounts your account managers should call today. These actions are valuable, and we see them proliferating as they become more possible with modern data warehouses.

A more sophisticated use of analytics data is to pipe it back into customer-facing operational systems so that you can offer personalized experiences. This requires a technology called **Reverse ETL**.



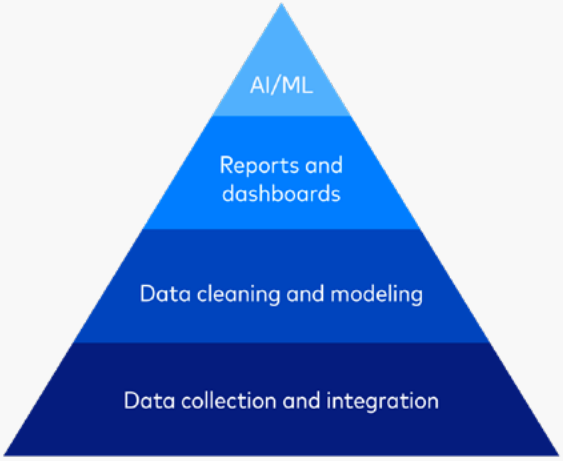
Reverse ETL moves data from analytics systems back into operational systems, making it usable to your team as well as customers.

By operationalizing analytics data, you can use it as training sets for machine learning. This will enable your data scientists to turn predictive models into automated, data-driven systems, leading to smarter customer experiences and automated data-driven decisions.

In the longer term, keep a lookout for continued evolution in the data warehouse space. While data warehouses were traditionally designed for analytical (rather than operational) workloads, newer developments involve taking on more and more characteristics of operational systems so that they can accommodate hybrid analytics and operational workloads.

Revisiting the Data Hierarchy of Needs

Everything your organization pursues in terms of data integration and data science will correspond to a stage on the data hierarchy of needs:



Your end goal is to build a data-driven company that can support all of its decisions with data, build smart products, and offer personalized experiences to customers.

That end goal requires a solid foundation of data integration infrastructure and cloud-based tools, a strong team of analysts — and, ultimately, data scientists and tools for operationalizing analytics data. That goal will also require you to determine a number of internal processes, weigh tradeoffs for different tools, and establish an initial set of metrics. It will also require you to stay mindful of the changing capabilities of new tools and technologies.

All of that is worth it to build an innovative, bleeding-edge organization.

Fivetran can help you take the first step toward building a modern data stack. We support a wide range of common destinations, hundreds of common data sources, and transformations.

Organizations have used Fivetran to jumpstart their way to the modern data stack in a number of ways:

“Instead of hiring Data Engineers, Fivetran allows us to focus on business value, hiring analysts, dashboard builders, people who are experts in web analytics and paid media. Our infrastructure is a lot broader and more advanced than it was a year or two ago.” Chris Klaczynski, Marketing Analytics Manager, Databricks

“With Fivetran, you point and click on a connector, enter a couple of logins, and you've got data replicated. I don't have to think about anything else. I don’t have to think about pushing schema changes. I don’t have to think about maintaining a thing.” Joe Nowicki, VP of Data and Insights, HOMER

Begin a free trial at fivetran.com/signup, or see a demo at get.fivetran.com/demo

Good luck!

About Fivetran: Shaped by the real-world needs of data analysts, Fivetran technology is the smartest, fastest way to replicate your applications, databases, events, and files into a high-performance cloud warehouse. Fivetran connectors deploy in minutes, require zero maintenance, and automatically adjust to source changes — so your data team can stop worrying about engineering and focus on driving insights. Learn more at **Fivetran.com**.