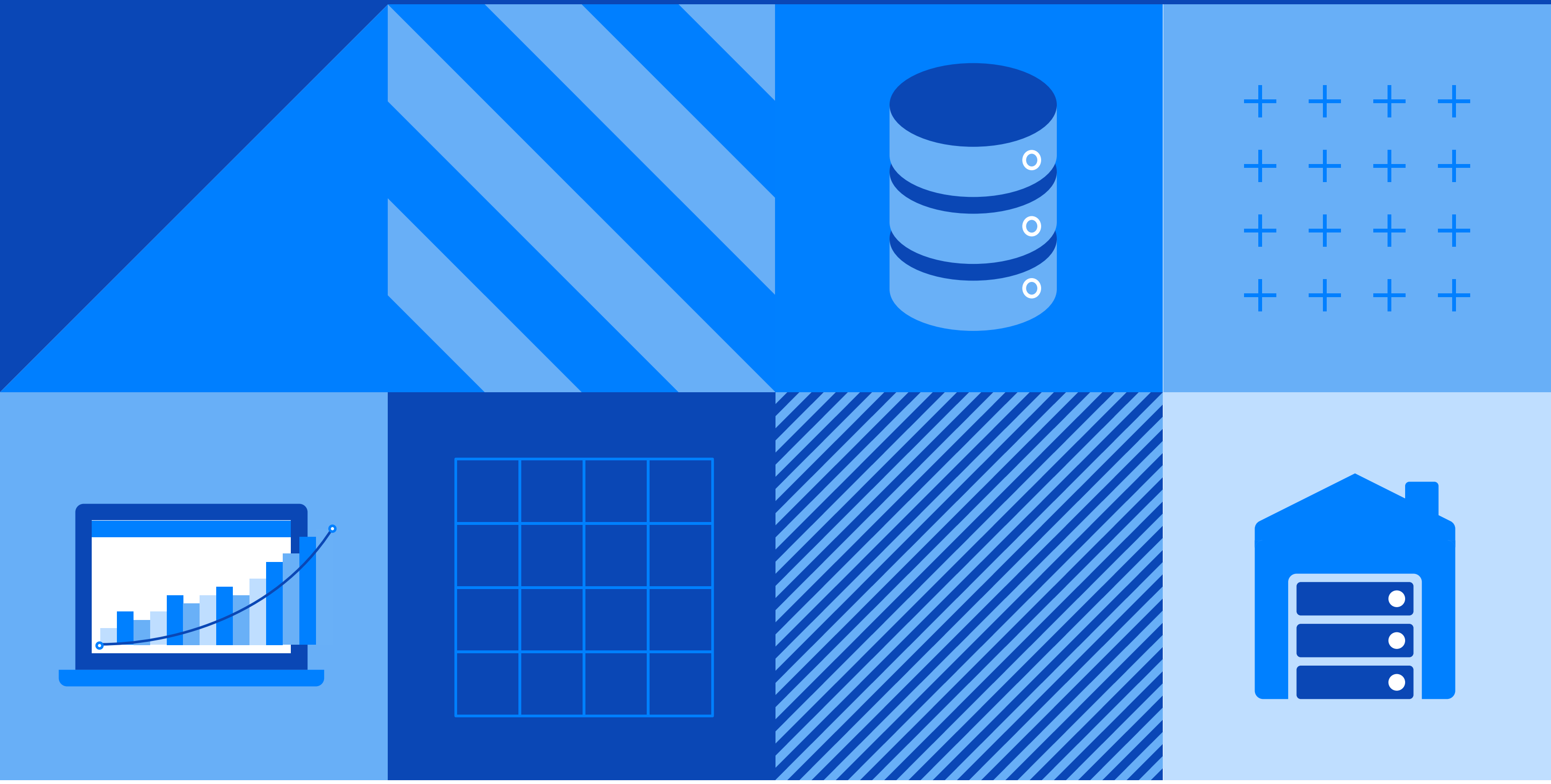
7 Insider Tips for Building Modern Data Stacks

Data management best practices we've learned from building analytics stacks at J.P. Morgan, Fivetran and more.

Veronica Zhai *and Charles Wang*



Fivetran

Executive Summary

This is a guide to building a modern data stack and other data management best practices. You will read about:

Preface

I started my career as an options trader at J.P. Morgan and went on to build their first modern data stack. At first, I was struck by the sophistication of the financial system and the quantitative nature of day-to-day work. It is data-driven in a very direct and literal sense: Traders work with eight monitors, each filled with hundreds of numbers and moving indicators. If a trader doesn't manage an option that expires at a particular strike price correctly as the markets move, they can literally make or lose hundreds of thousands of dollars in a few minutes. Data is money, and technology and data can be an organization's biggest competitive assets.

Despite the transformative potential of data, companies of all sizes often struggle to make data useful. Building an effective data infrastructure can involve a multitude of methods and tools, and the data profession contains many competing perspectives and approaches. The right solutions are not obvious, even for behemoths like J.P. Morgan whose work is highly sensitive and quantitative in nature.

Currently, I am a principal analytics leader at Fivetran, a market leader in data integration. I first encountered Fivetran when I began researching different technologies to enable data engineering. The more I learned about Fivetran, the more I fell in love with what the company stands for.

Throughout my career, I've had to learn how to build complex information systems the hard way. Today, I want to share some best practices I've picked up that will help you succeed.

1. How to Build a Modern Data Stack

A data stack is a suite of tools that enable data integration. The modern data stack consists of cloud-native data tools centered around automation, lowered costs and ease-of-use to end users throughout the lifecycle of data management. With the growth of the cloud and cloud-based data platforms throughout the 2000s, companies today can easily jumpstart their data integration efforts using a suite of cloud-native data tools.

Legacy companies may experience a slow and painful process moving to the cloud from on-premise technologies with high, inflexible fixed costs and big data-related performance issues. Slower procurement cycles, larger data volumes and higher risks can all complicate and slow down the process. Newer companies, by contrast, can start fresh with cloud-native data tools.

In both cases, building a modern data stack follows this framework:

1. **Data warehouse –** First, set up a cloud-based data warehouse. Different data warehouses will offer different scalability, pricing models, dialects of SQL and other features. Notable examples of cloud-native data warehouses include BigQuery, Snowflake and Redshift.

2. **Business intelligence** – Then, connect a cloud-native BI tool with your data warehouse. Different BI tools offer varying levels of visualization power, user-friendliness, collaboration and other features. Notable examples of BI tools include Looker, Tableau, Qlik and Mode.

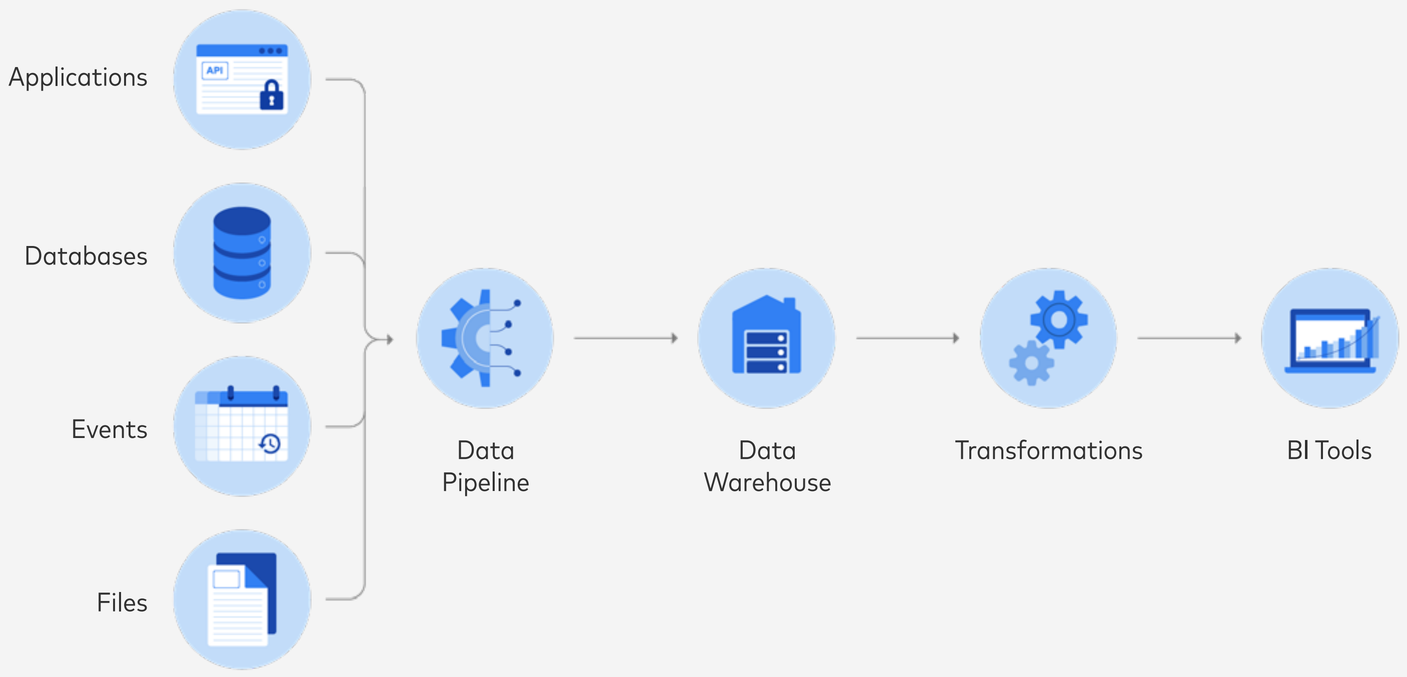
3. **Data pipeline** – You will need a tool to extract data from your applications and operational systems and load it to the central data warehouse. Different pipeline vendors have different approaches to ease-of-use, configurability, security and customer service. Examples include Fivetran, Stitch, Xplenty and Matillion. Data

pipelines are often colloquially referred to using the acronyms "ETL" (extract-transform-load) or "ELT" (extract-load-transform), based on the sequence of actions involved in moving data from a source to a destination. The most important consideration in choosing a data pipeline is the degree to which it automates and saves data engineering time.

4. **Data transformation** – Finally, you will need tools to transform data into models for reporting and predictive modeling. Many data pipelines and business intelligence tools include transformation tools; an example is Fivetran Transformations.

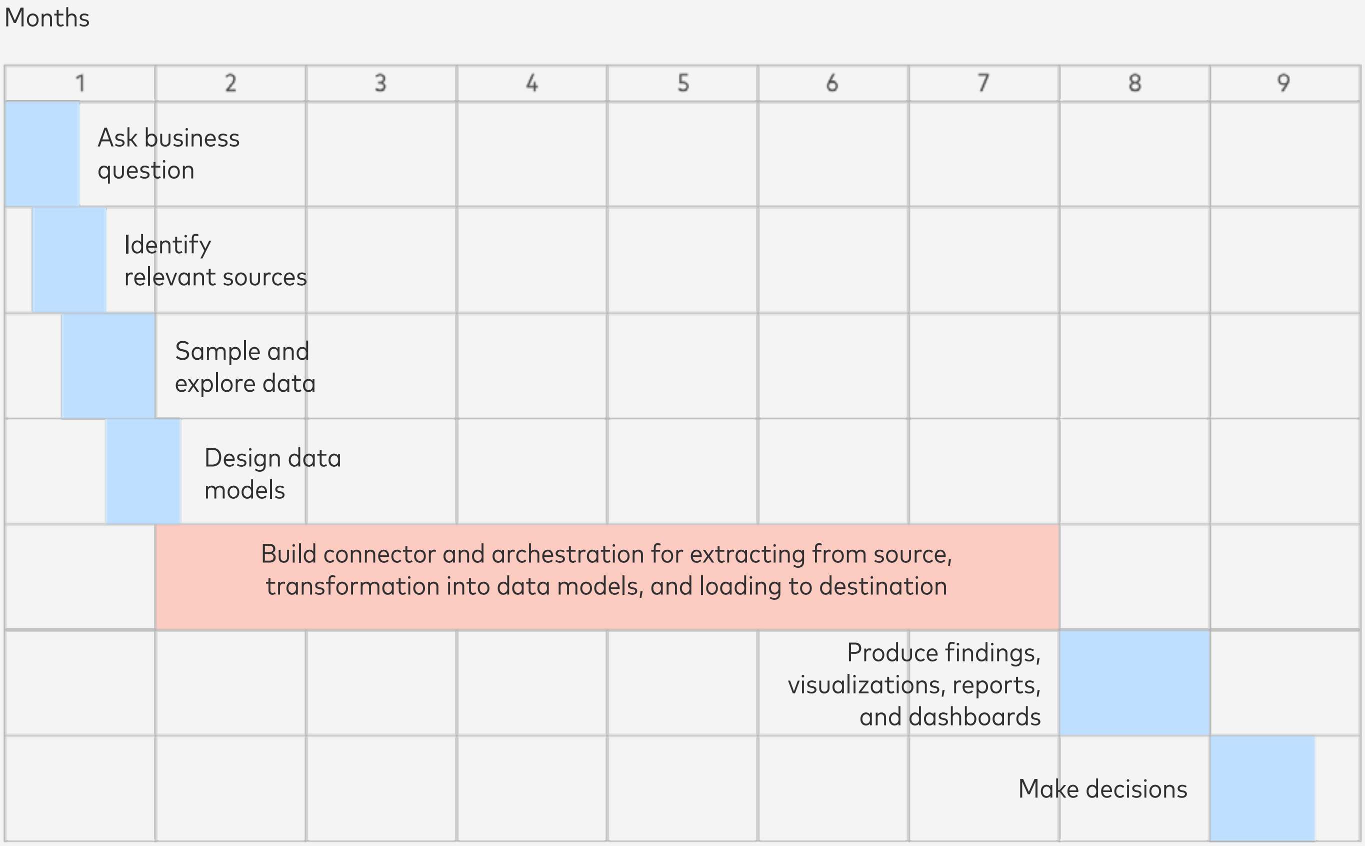
Most vendors offer free trials. For detailed comparisons, consult industry publications like Gartner.

The exact flow, from left to right, is illustrated below:



The data engineering required to build this workflow from scratch can be a major hurdle for organizations of all sizes. Production and maintenance can be time-consuming and expensive even with a large, well-resourced team.

Gantt charts and workflow diagrams can show you where business processes cause downtime. For instance, the engineering-heavy nature of ETL often creates downtime for analysts:



AUTODESK.

Take Autodesk, for example: When VP of Data Platforms and Insights Jesse Pederson inherited Autodesk's data stack, data ingestion was a major problem.

"The modern data stack has really helped to free up our teams, frankly, to work on problems that they'd rather be working on. I don't get emails anymore saying, 'Urgent — Pipelines are broken.' I get emails now that are like, 'Hey, when can I put my data in.'"

The team had been importing data from Salesforce, SAP, Siebel and Autodesk's own products such as AutoCAD, Revit and Maya into an S3 data lake.

To simplify the process, Jesse brought in an automated data pipeline and cloud data warehouse — and clearly delineated a process for data ingestion. Today, Autodesk has bifurcated its data pipelines, with two routes for data ingestion and storage:

- If the source is a structured data store, Jesse's team uses Fivetran for ingestion. Structured data is stored in Snowflake.
- If the source is an unstructured data store – for example, usage metrics from Autodesk's own products and software – Autodesk uses AWS Kinesis for high-volume ingestion into S3.
- Data is replicated between the two repositories as necessary. Product usage data is promoted to Snowflake for easier visualization and analytics, subject to Autodesk's privacy controls, and snapshots of Snowflake data are persisted into S3 for historical reference and machine learning purposes.

With the engineering resources freed up from data integration, Autodesk was able to build an early-warning system to predict customer churn, allowing Autodesk to better direct the efforts of the customer support team.

2. Pitfalls to Avoid While Building a Modern Data Stack

In the previous section, we covered the “dos” of building a modern data stack. There are also a number of important “don’ts.” These mistakes can hobble your organization’s ability to use data effectively:

Choosing legacy infrastructure instead of migrating to the cloud. Many companies maintain legacy infrastructure in expensive, maintenance-intensive physical data centers.

On-premise data stacks have the following disadvantages:

- You have to estimate costs for hardware and build enough excess capacity to account for peak time usage. This leaves a lot of slack during non-peak time, and is more costly in general.

- A custom, on-premise setup requires lots of configuration and tuning. Configuration and tuning requires specialized talent and poses a very high barrier to entry. This approach is only accessible to large, well-resourced organizations, and, because of its slow, labor-intensive and generally difficult nature, is arguably not a good idea even when possible.
- The performance of your data stack is ultimately limited by the constraints posed by your existing hardware. This contrasts unfavorably with a cloud-based stack, where additional compute and storage resources can be spun up and deactivated as needed. Scaling an on-premise system for intermittently high activity, as well as future growth, is difficult.

An outsourced, cloud-based data infrastructure offers many benefits, including better scalability, ease of use, accessibility and cost. It can radically simplify your organization's workflow.

Migrating to the cloud can be a huge hassle because of the inherent complexity of data, as well as the need to keep operations running while the migration is in progress.

Consider enlisting the help of an automated data pipeline, like Fivetran, to make the process easier. Oldcastle and Copyright used the modern data stack to migrate from on-premise to the cloud and continue integrating data.

DIYing your data pipeline. Data integration is more than a matter of moving records from a source to a central location. It involves serious engineering challenges and design considerations, such as the ability to incrementally read and update data, robustness to failure, normalized schemas, schema migrations, parallelization, pipelining, and more.

In addition, the time-consuming and labor-intensive nature of a DIY data pipeline diverts engineering resources from other product or infrastructure duties, and creates downtime for analysts. The solution to this pitfall is to leverage outsourcing and automation whenever possible in order to abstract away as much of the complexity involved as possible.

Choosing ETL instead of ELT. The traditional approach to data integration, ETL, is so ubiquitous that it is practically synonymous with data integration. Unfortunately, ETL, with its tight coupling between extraction and transformation and its reliance on bespoke engineering, is a far more brittle workflow. It is no longer the best approach for most organizations, and ELT, which enables data to be automatically loaded in a near-raw state before being modeled by analysts, is now a more practical option.

A more detailed discussion of ETL and ELT is beyond the scope of this white paper, but the main benefit offered by modern ELT is saving labor. This is the result of technological trends, specifically the plummeting costs of storage, computation and network bandwidth. Fundamentally, ETL conserves technological resources at the expense of labor, while ELT leverages technological capabilities to save labor.

Lack of data management discipline. Over time, your data team will grow and your data integration efforts will expand. You will adopt new tools and give more people access to tools to create data models, dashboards and other data assets. You will regularly pivot, leaving some of these data assets obsolete. Growth introduces the danger of creating data assets that are disorganized and difficult to find or correctly interpret, creating a form of technical debt.

You must construct guardrails by periodically auditing and discarding data assets that are no longer useful. As with gardens, the solution to data-related technical debt is pruning and curating. When it comes to clearly communicating actionable, data-backed insights, less is often more.

Ritual

The story of Ritual, a subscription wellness brand, illustrates the brittleness of DIY data pipelines. Ritual's customized data pipeline regularly failed or ran behind schedule. Their analytics depended on nightly snapshots that the pipeline provided with stale data, or not at all, resulting in missing retention data in the company's daily Looker report. Brett Trani, Director of Data and Analytics, explains:

> "With the failures and data gaps, people lost trust in the data and would go out and find their own sources – spreadsheets, ad platforms, random notes – and end up with different numbers for the same metric. There was no single source of truth for retention."

With the help of a fully automated data pipeline, Ritual was able to build a single source of truth and experienced a 95% reduction in data pipeline issues. This led to a 75% reduction in query times, and tripled the data team's productivity.

3. How to Make the First Few Data Hires

One hard problem we've experienced at Fivetran as we grow is how to hire good talent fast. Having hired over a dozen talented data professionals last year, I want to share a proven framework that will help you level up.

1. **First screen: test for advanced technical skills, particularly advanced SQL skill.** You can use testing software such as HackerRank to increase screening efficiency.

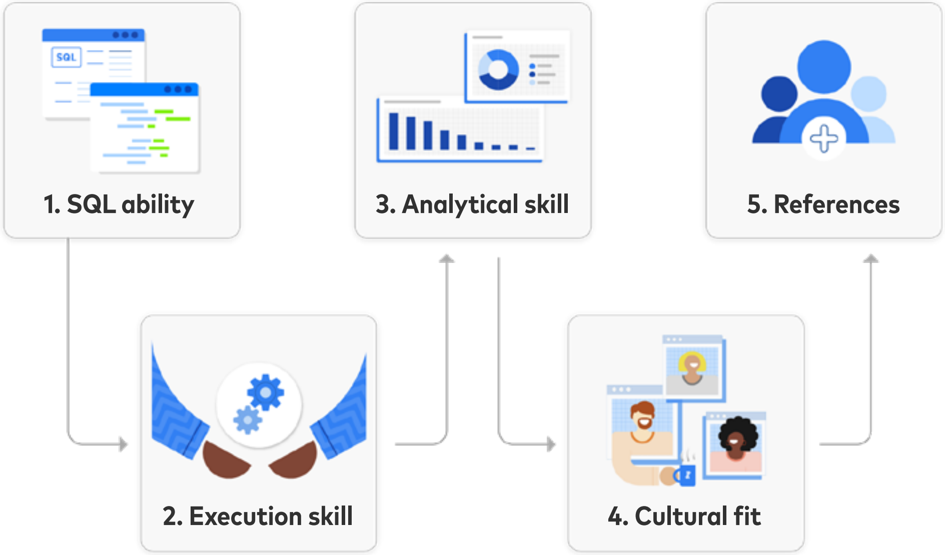
2. **Second screen: test for execution skill.** If this is the first hire, have the candidates write a 30-60-90 day plan and evaluate their business strategy. If this is not the first hire, have them work through existing code featuring business logic. This is a practical test that mimics their typical duties on the job.

3. **Next, screen for analytical skill:** Ask your candidates to present insights and visualizations from a complex sample data set, with implications for business strategy and decisions. Alternatively, use a timeboxed case study involving a relevant use case to evaluate their analytical skill. This will test your candidate's ability to analyze data and articulate their ideas.

4. **Superb cultural fit is a must:** We have consistently improved talent density by pursuing cultural fit as alignment with company values. Prioritize new hires whose traits and abilities complement your team.

5. **When in doubt, make reference calls:** Investing an hour up front can prevent you from making a costly hiring mistake.

6. **Keep hiring analysts!** Modern data tools dramatically lower the barriers of entry to data integration and enable your data team to use SQL for both modelling and transformation. This obviates the need for scripting in Python or Java. You can put off hiring data engineers for some time. Instead, keep building a team of analysts. At some point, consider hiring a data architect to optimize the overall system.



Last but not least, lean heavily on networking and referrals to improve hiring speed and quality. Rallying your employees for referrals and pairing this with referral incentives goes a long way.

You may not have the resources to hire analysts who meet all the criteria listed above. That's okay – when in doubt, find people who are hungry and adaptable. Technical skills can be learned on the job.

Data professionals are active in a number of online data communities and job boards. Try talking to people in places such as:

- Locally Optimistic (locallyoptimistic.com)
- Outer Join (outerjoin.us)

With the help of software, you can take the grunt work out of data engineering, allowing your analysts to focus on analytics and your engineers to focus on product and infrastructure improvements.

Automating your data ingestion early can help you to prioritize in the early stages of hiring, focusing more on analytics than engineering – especially when you need to demonstrate your growing department's value to the wider organization.

databricks

Take Chris Klaczynski, a Marketing Analytics Manager and modern data stack power user at Databricks.

Chris was the first analytics hire at Databricks. As Databricks rapidly expanded, Chris recognized the importance of a scalable data integration solution:

> "Our new hires can hit the ground running and start building dashboards immediately. They don't have to build pipelines or get familiar with notebooks or writing code. They can have immediate access to data so they can focus on insights and building relationships with their stakeholders."

Automated data integration immediately stood out as a solution to Chris' data pipeline needs:

> “Instead of hiring the kind of traditional engineering headcount, automated data integration allowed 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."

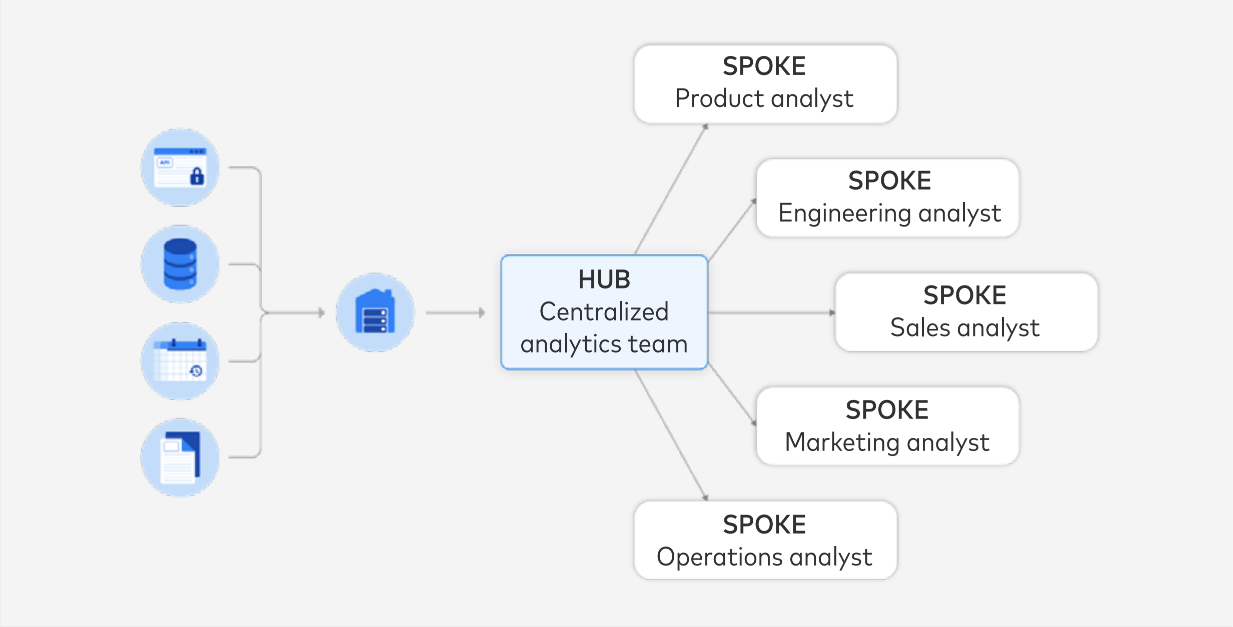
Instead of building data connectors or pipelines, Databricks was able to focus on analytics, BI and information architecture. Their new hiring focused far more on analysts than data engineers.

4. What to Do in the First 180 Days

Your first six months are crucial to setting the stage for your company's analytics efforts. Here is a framework that will help you build a solid foundation for early success.

1. **Design a centralized data team:** A centralized team with a "hub-and-spoke" structure is a superior model for the majority of companies, because it helps align strategy and execution. The analytics team (hub) should report directly to the CEO or a technical executive, and pods (spokes) that specialize in particular business domains should be functionally aligned with their respective departments. This model has worked well both at J.P. Morgan, where the team must support businesses on a massive scale, and at Fivetran, where the company needs to scale aggressively.

   The hub-and-spoke structure allows your analysts to develop close working relationships and expertise with specific functional units of your organization,



while ensuring that you have a core team of analysts who can coordinate and address tasks that must be centralized.

2. **Work with peer teams:** First, identify other teams that are already using analytics, and how. Build alliances by helping them automate their data integration and avoid duplicate work. Second, determine the team's scope, particularly identifying tasks that are out of scope to improve focus and execution.

3. **Align foundational metrics with the leadership:** Your organization's leadership should ensure the BI layer is an integral part of business strategy because "what gets measured gets managed." Here is a simple framework of important early KPIs for a SaaS company:
   - 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
   - Product
     - Daily, weekly, monthly active users
     - Customer journey
     - Feature usage
     - Net promoter score

Beware of vanity metrics that look good without influencing outcomes that matter to the organization. The goal of setting metrics is to provide clear direction and align the incentives of everyone within your organization. Since your organization's leadership sees more pieces of the puzzle than other people in the organization, it behooves your leaders to take extraordinary care about setting metrics.

Maravai LifeSciences is a leading life sciences company. Leadership wanted more holistic insight and reporting on all of Maravai LifeSciences, beginning with financial planning and analysis (FP&A). With multiple, autonomous subsidiaries, analytics was beset by challenges such as:

- Disparate data sources required manual blending, significant cleanup and transformation to create a full data set
- Information was not readily available and required digging to uncover
- No "single source of truth " system with access to all key decision-makers

Using an automated data pipeline, cloud data warehouse and cloud BI tool, the business can now answer all of its financial questions, including:

- How is revenue performing compared to the same time last year?
- What is one subsidiary's average daily sales for the past quarter? What about the past year?
- How has a subsidiary company's margin increased or decreased since last year?
- Who are the top five customers for each subsidiary?
- Which customer region contributes most to total revenue?

Maravai LifeSciences now drives business decisions using dashboards for critical finance, customer and sales analytics data, and plans to add dashboards for product analytics, sales analytics, inventory analytics, and profit and loss analytics.

5. How to Run Your Data Team as an R&D Team

Traditionally, a data team is simply viewed as a support team, an engineering team, or, increasingly, a product team. However, I believe that a data team is a combination of the three.

- **Build with a product lens:** Erik Jones, Director of Analytics at New Relic, summarizes this aspect brilliantly: "A successful analytics team must also be good at gathering requirements, defining scope, managing expectations, marketing and rollout, training end users and ultimately driving adoption of what is being built." The data team should center around enabling self-service. I recommend using usage adoption and NPS as North Star metrics for the team.

The North Star metrics of a data team are usage and customer satisfaction. A data team's main objective is to help a company make better decisions. This means that your data team must enable and promote usage of reports, dashboards and other assets. How often members of your organization use reports, dashboards and other analytics assets is an important KPI and sanity check in its own right, as it demonstrates that people actually use your team's assets. Consider building a dashboard that measures usage of your other dashboards. Ultimately, you should strive for universal adoption of data-based decision-making.

- **Operate with engineering principles:** A successful analytics team should also invest at least 25% of resources in building an easily navigable and scalable data infrastructure. Leveraging these engineering principles will improve operational efficiency: user requests logs, bi-weekly sprints, implementing code review and quality assurance (QA) processes, continuous automation, and extensive documentation.

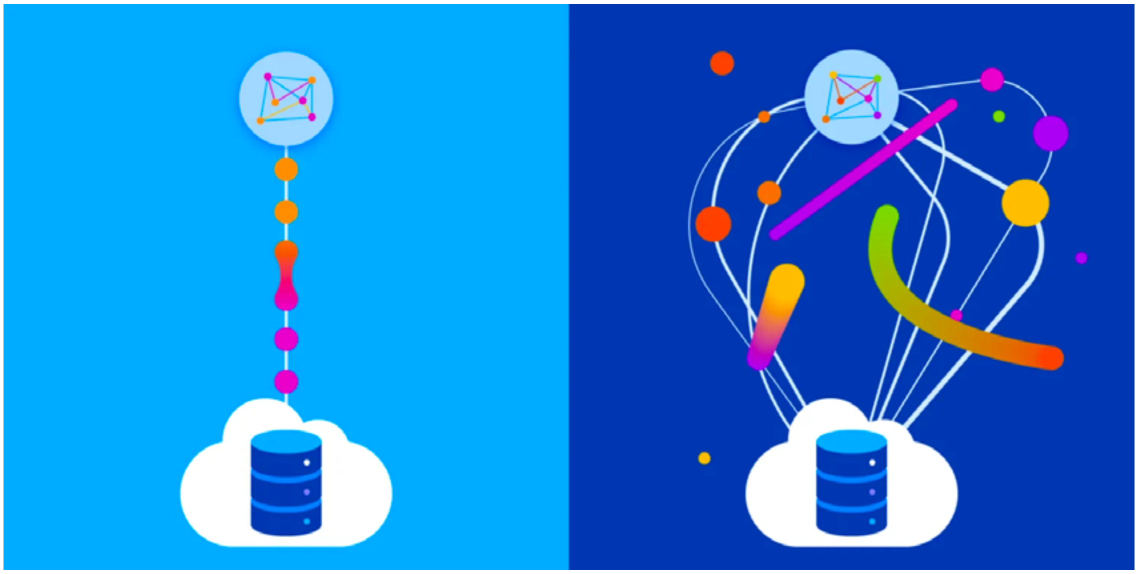
- **Service with a customer-centric mentality:** The analytics team should build in a technical customer success function that provides onboarding and ongoing support, responds to issues escalated by users, works with partner teams to resolve production issues, and develops end-user training materials. These functions can be assigned to dedicated technical customer success roles as the team scales.

6. Systems Thinking Part I: Optimizing Data Integration Workflows

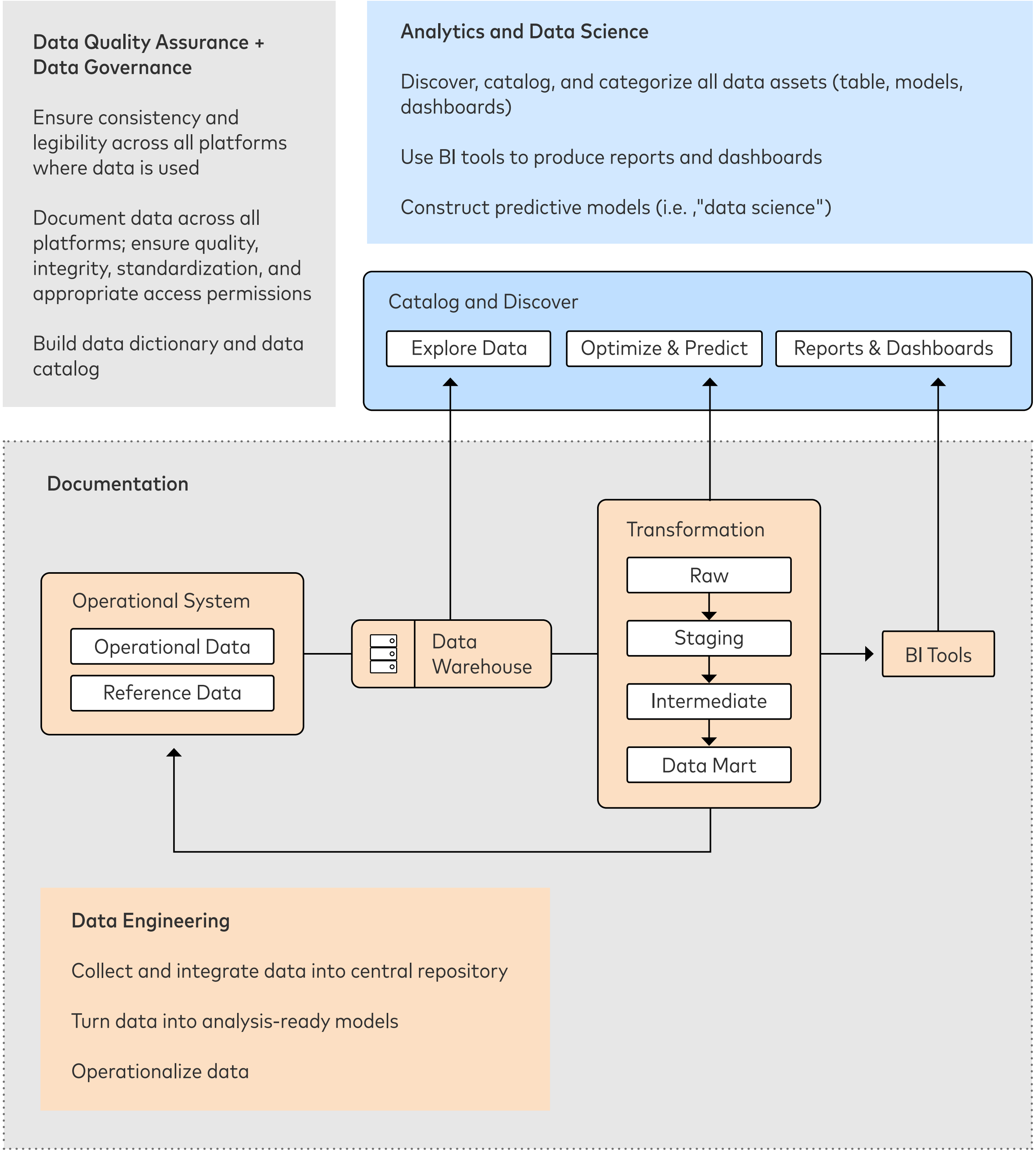
Approach your data modernization efforts with curiosity and patience. You will need to gain a strong grasp of the moving parts involved in data integration, identify the bottlenecks, and seek opportunities to abstract away complexity whenever possible.

Enterprise Data Is Complex and Chaotic

The complexity of enterprise data also complicates the workflows involved in integrating that data. Like many large enterprises, J.P. Morgan has gone through many mergers and acquisitions over the past century, and has had to integrate many different systems that produce data. This complexity is amplified through the lifecycle of data management: analysis and reconciliation work, system and data integration, data monitoring and governance, standardization across businesses, complex permissioning, and foundational performance issues involved in moving, transforming and interacting with petabytes of data.



The following workflow diagram illustrates the complexity of managing enterprise data across its full lifecycle:



Data Quality Assurance + Data Governance

Ensure consistency and legibility across all platforms where data is used

Document data across all platforms; ensure quality, integrity, standardization, and appropriate access permissions

Build data dictionary and data catalog

Analytics and Data Science

Discover, catalog, and categorize all data assets (table, models, dashboards)

Use BI tools to produce reports and dashboards

Construct predictive models (i.e. ,"data science")

Data Engineering

Collect and integrate data into central repository

Turn data into analysis-ready models

Operationalize data

Systems Thinking Can Help You Achieve Success

I stumbled upon success while managing data for the financing business, which handles over $1 trillion worth of capital. Salespeople, traders and more ceaselessly made data requests. However, most of these requests were time-consuming to execute, involved small, local optimizations and made only limited impact. I discovered that strategically placing trades on a macro level to match assets and liabilities would potentially free up hundreds of millions of dollars worth of capital, increasing the firm's competitive edge. Such actions were beyond the mandate of individual teams, because no single team could access all of the data.

This meant that our first task was to break down silos and centralize the data. We then explored machine learning algorithms to analyze how to allocate capital to clients to maximize ROI. Because it was extremely complex to automate the production of any system-level metric in a large enterprise, both providing real-time visibility to the macro metric and optimizing for it on a global scale proved to be of limitless value.

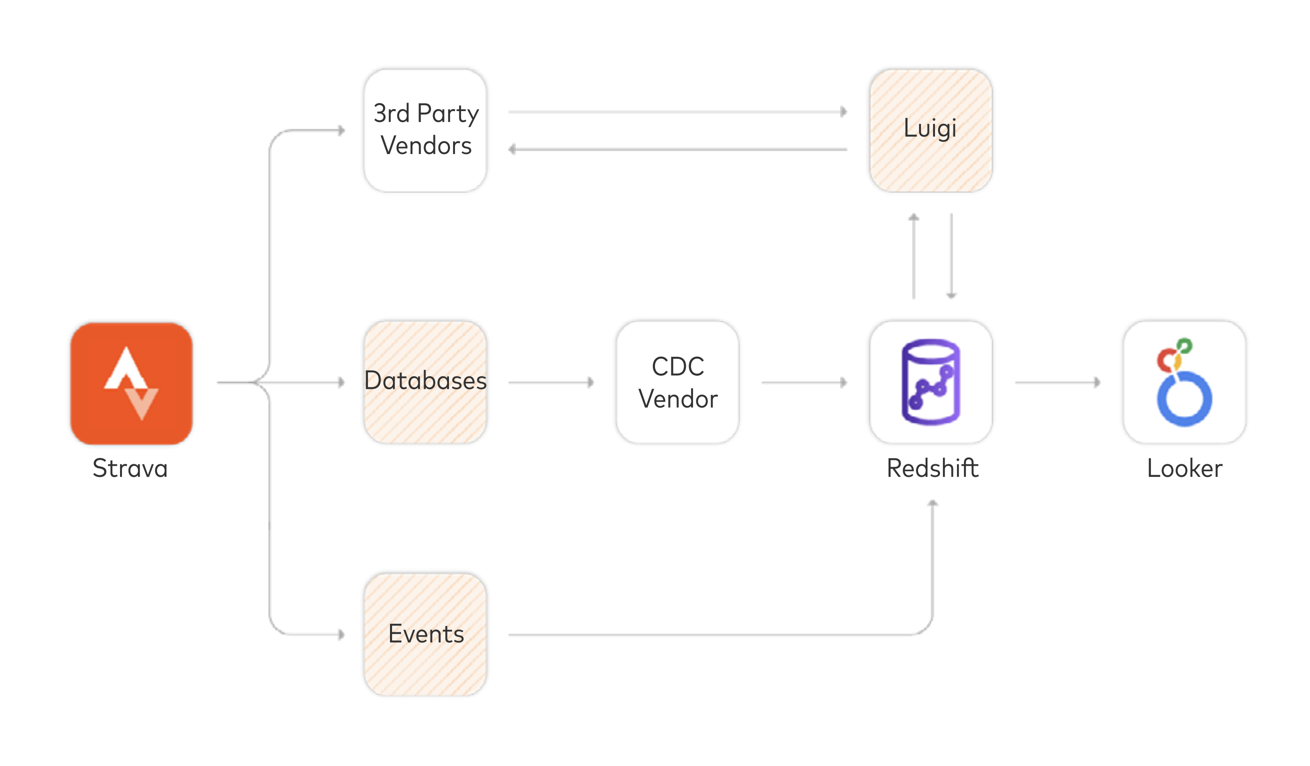
.

STRAVA | LABS

Strava Data Engineering Lead Daniel Huang's team used to race to keep up with the needs of the fast-growing, 12-year-old company.

Eventually, they were forced to consider the future of their company's data engineering culture. "It started with a shift toward a platform," Huang recalls. "Our role as data engineers should be to build the platform and guide people through it. Let the platform serve the data needs."

Strava's original data infrastructure is diagrammed below:

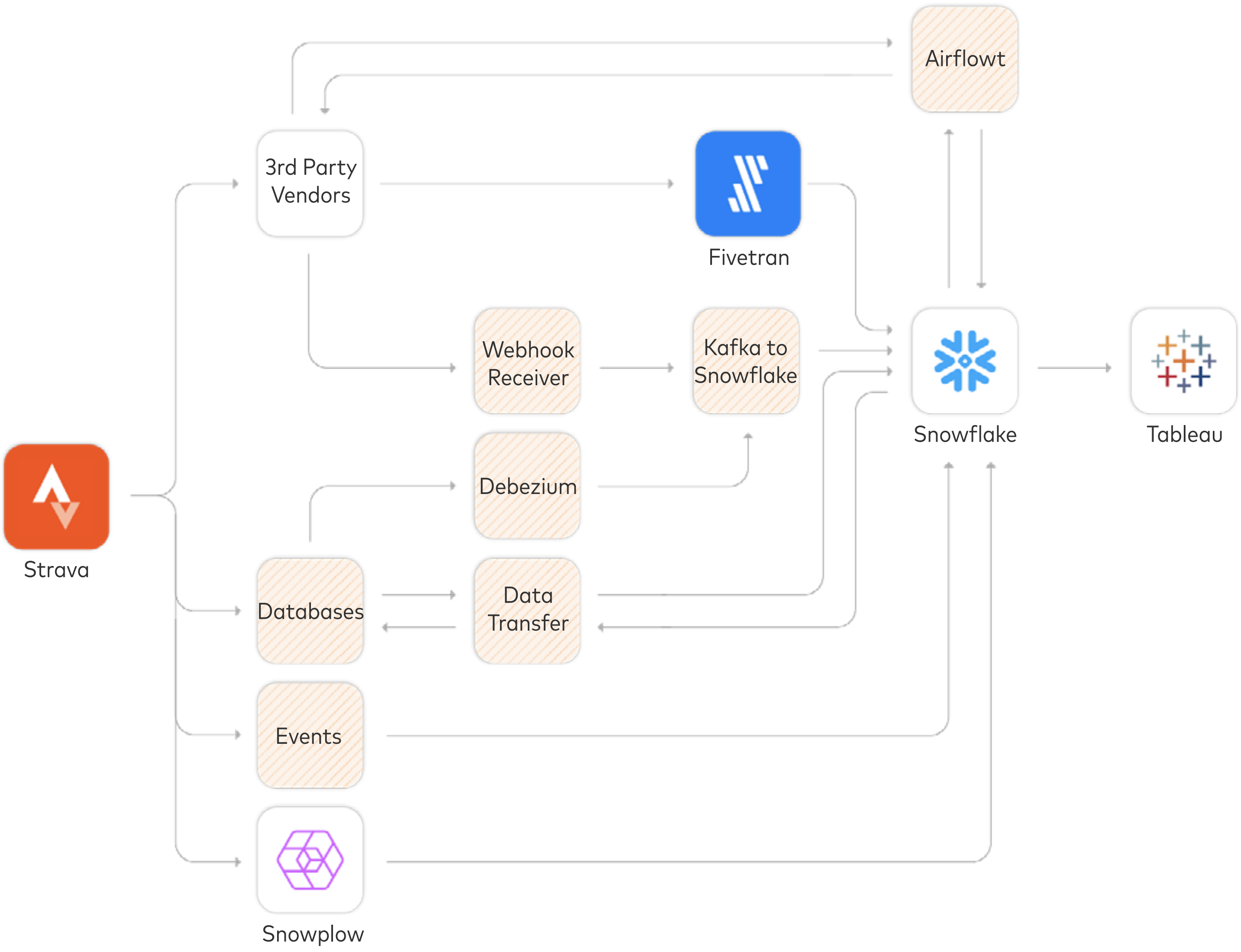


According to Dimensional Research, 63% of companies still rely on manual scripting, even though companies are moving more data faster than ever. In fact, 72% of organizations now need data to be moved daily, hourly or even every few seconds.

"In the beginning, all of our ETL jobs were authored by a couple of data engineers," says Huang, "and that meant we were maintaining all these jobs. We were on call to fix these jobs instead of building underlying infrastructure or services. Bottom line: We were becoming the interface for data more often than we liked, and we were becoming a bottleneck for the company."

With a new vision, Strava implemented a cloud data stack built on Snowflake as a data warehouse, Tableau as a BI tool, and Fivetran as its data pipeline provider to automatically land data in Snowflake from third-party vendors.

The following represents Strava's modern approach to data management, using the modern data stack:

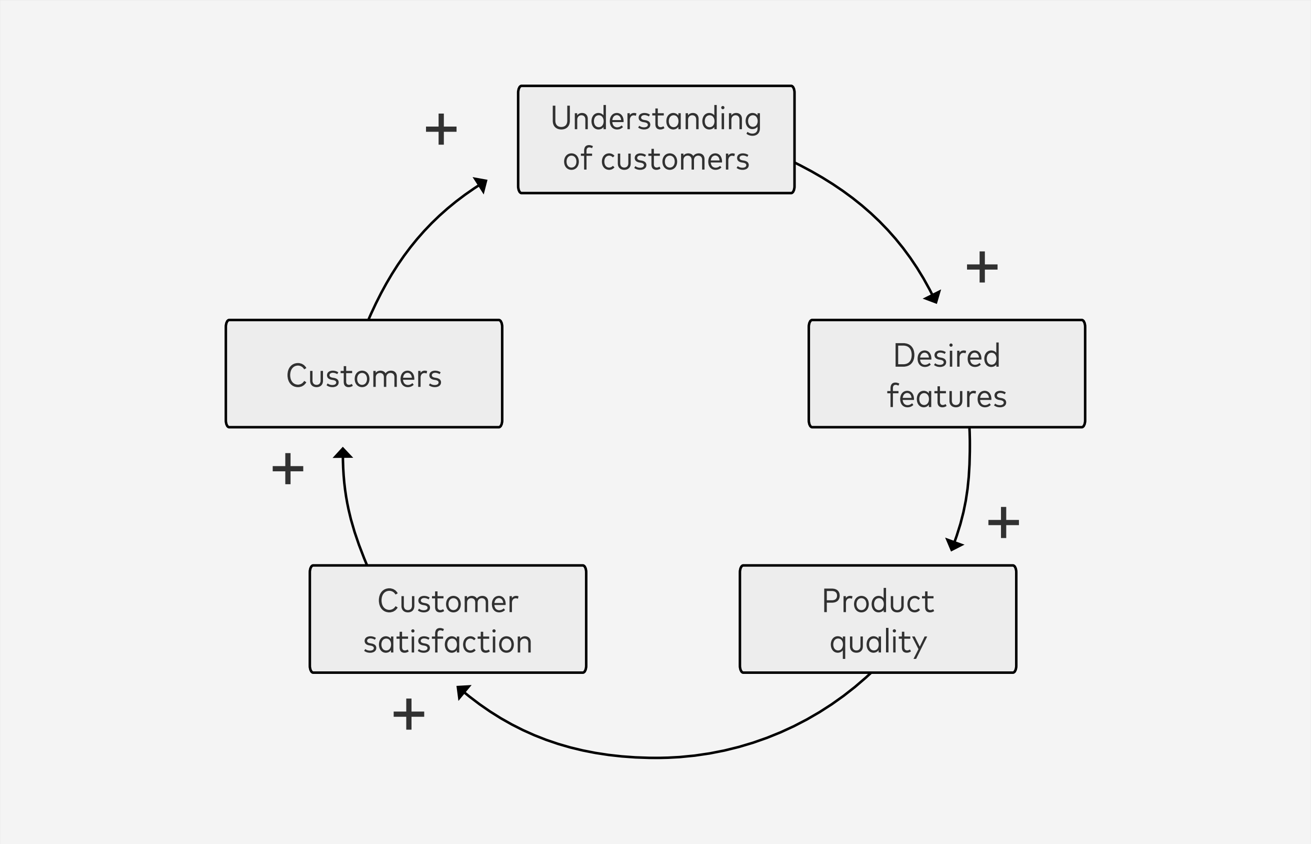


“We’re still a small team but we have made progress,” says Huang, reflecting today on the midpoint of their journey. “The ease of cloud-based tools has freed up our team to think about the company’s data culture as a whole, and we’ve set up a categorization of internal data users to better understand and meet their needs.”

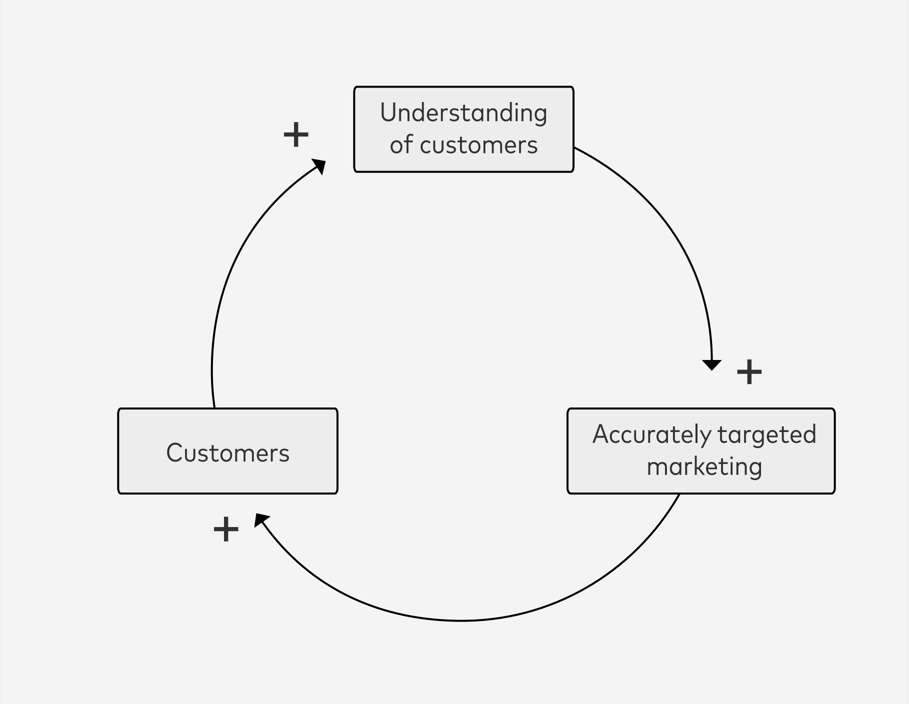
7. Systems Thinking Part II: Using Information as Leverage

Once you have built a well-functioning data stack, you can identify the virtuous (and vicious) cycles that impact your business. In order to make effective decisions, it is key for analysts and stakeholders alike to understand the organization's business model, constraints and incentives.

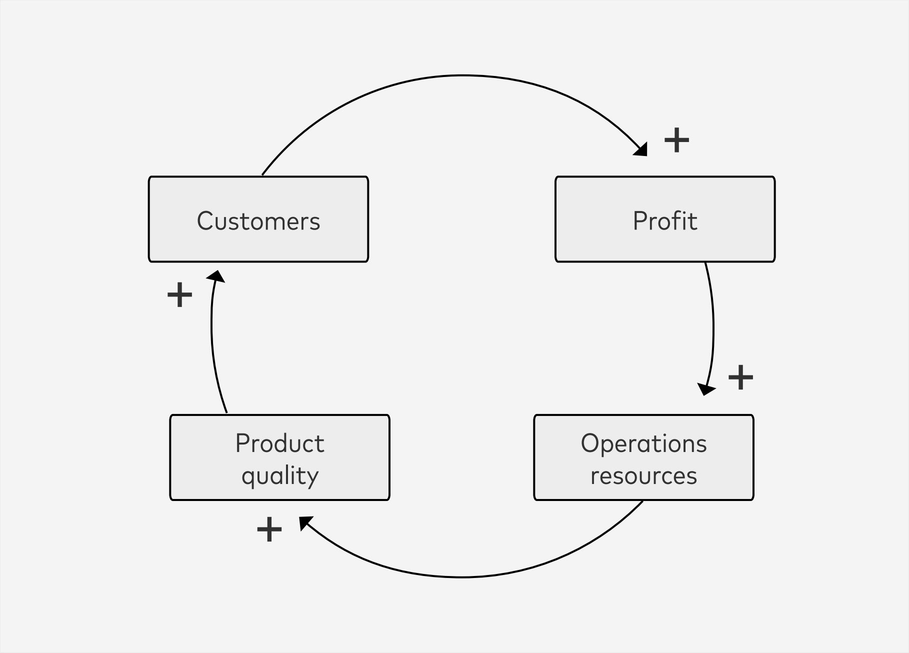
Identify the flywheels that make your business succeed. For instance, a keen understanding of your customers will lead to a superior product, promoting customer satisfaction and increasing your customer base:



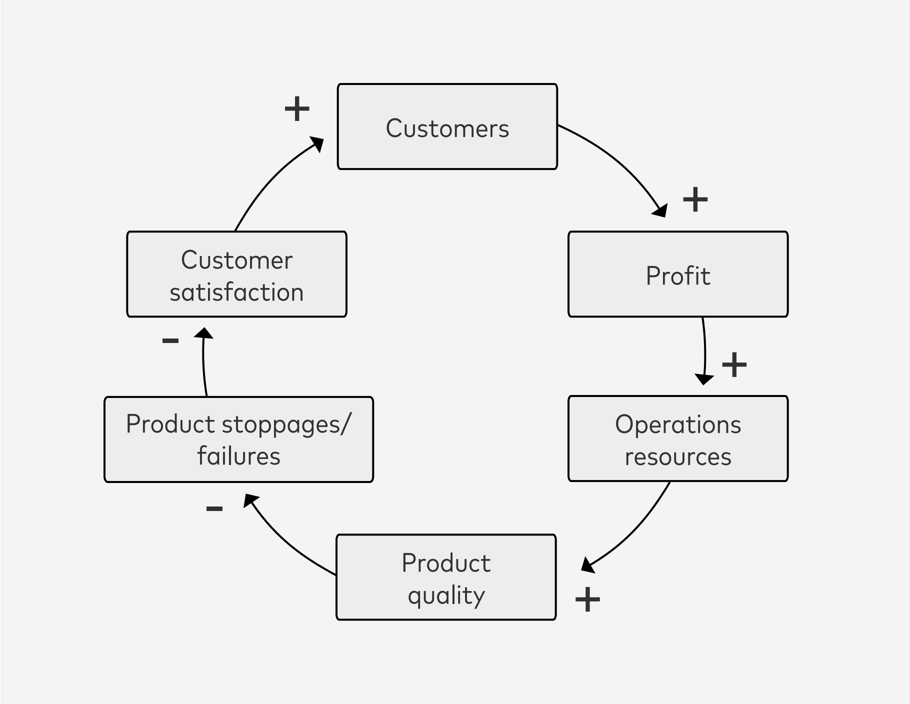
Not to mention improving your marketing efforts:



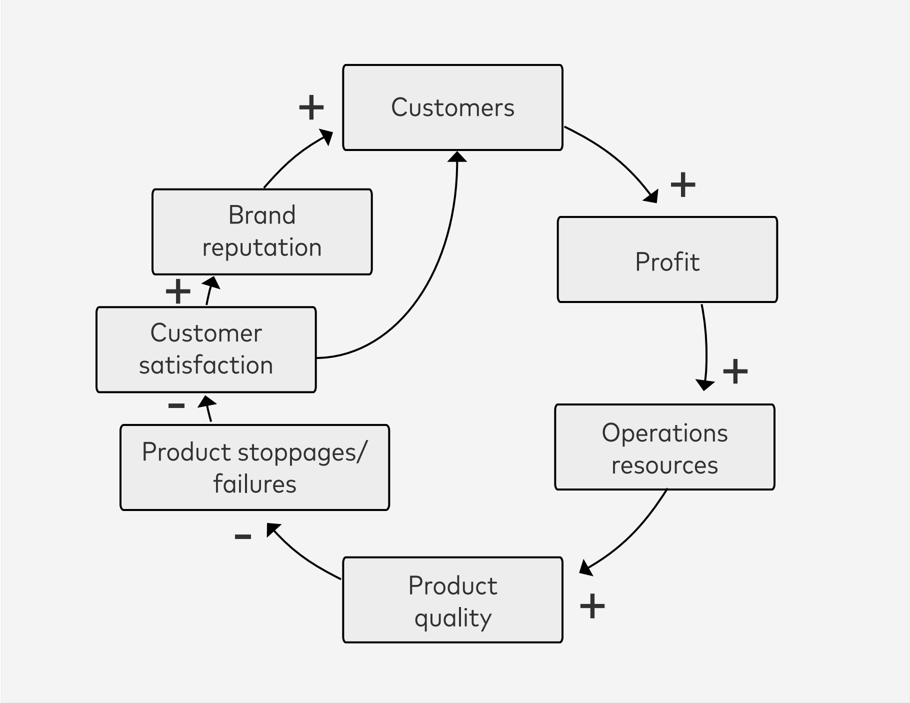
The better your product, the more customers, profit and resources are available to further improve your product:



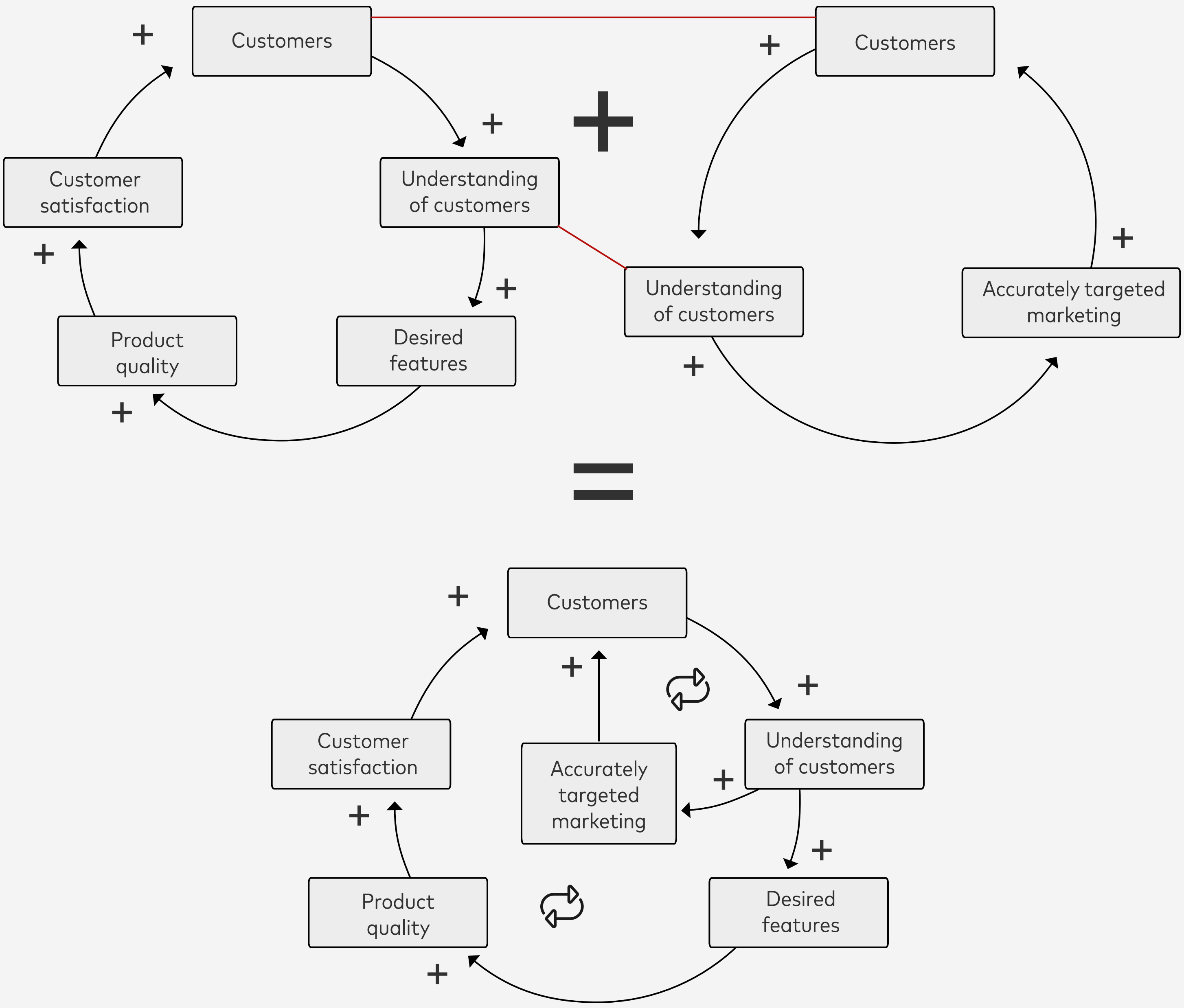
On the other hand, product stoppages and failures have an inverse relationship with customer satisfaction:

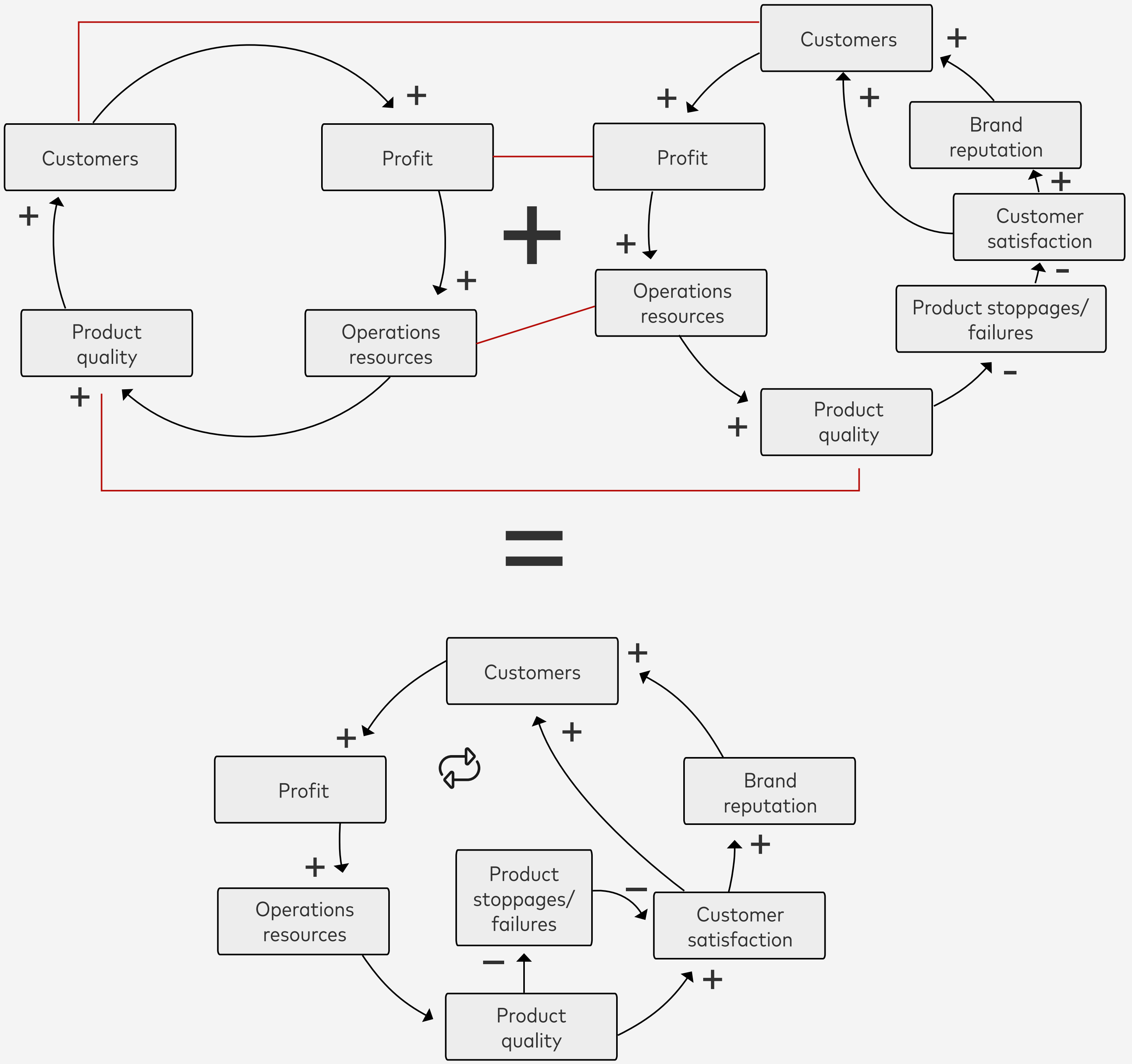


This has the same overall effect when mediated through brand reputation:

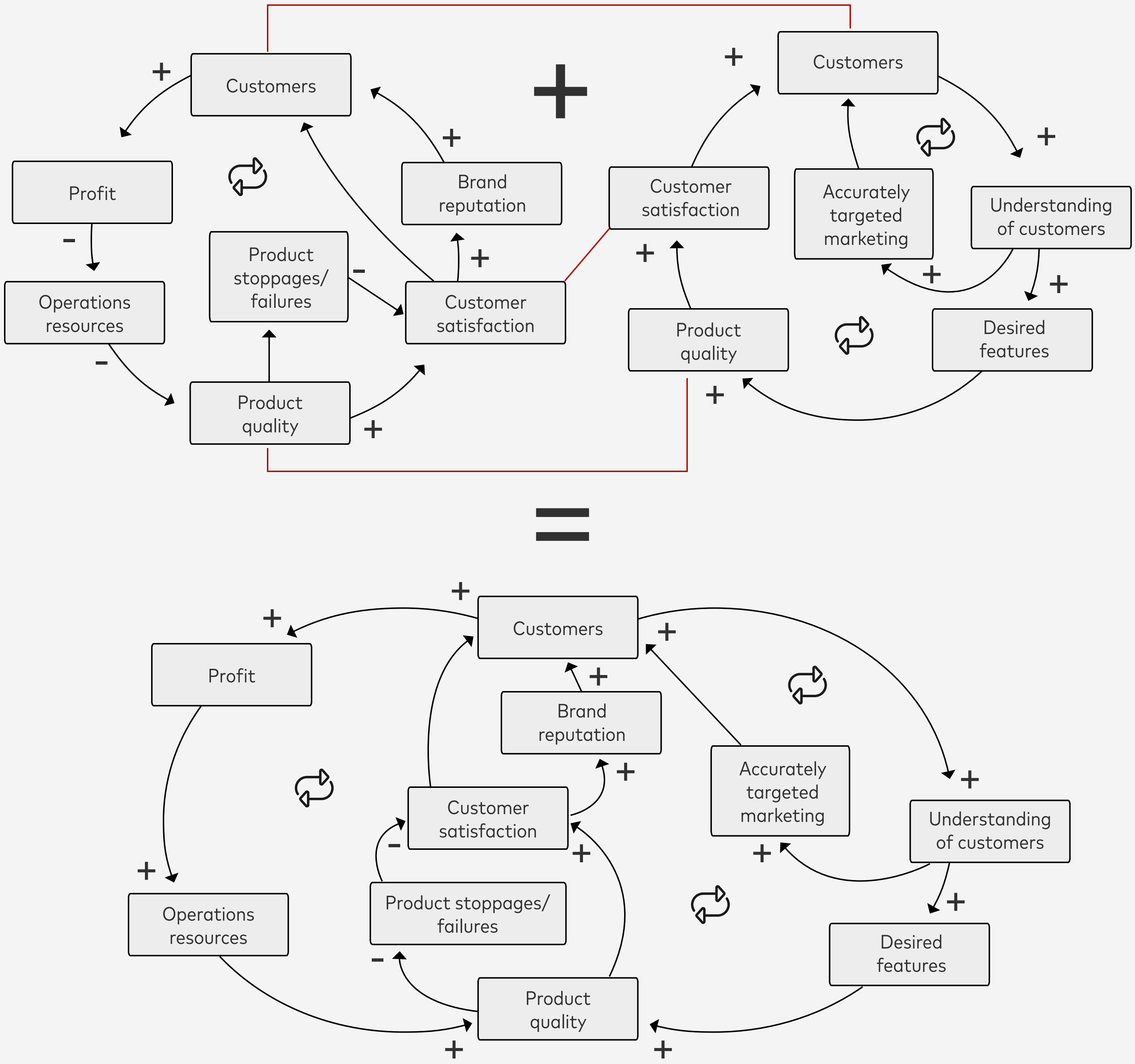


You can assemble these flywheels together into **causal loop diagrams** to more fully illustrate your business's operations. These flywheels have elements in common, allowing you to connect them into larger loop diagrams:


Customers
Understanding of customers
Desired features
Product quality
Customer satisfaction
Customers
Understanding of customers
Accurately targeted marketing
Customers
Understanding of customers
Desired features
Product quality
Customer satisfaction
Accurately targeted marketing


Customers
Profit
Product quality
Operations resources
+
Customers
Profit
Operations resources
Product quality
Brand reputation
Customer satisfaction
Product stoppages/ failures
=
Customers
Profit
Operations resources
Product quality
Product stoppages/ failures
Customer satisfaction
Brand reputation
+
–

The more comprehensive your diagram, the more you can see all of the moving parts in one place:



Causal loop diagrams allow you to lay out the moving parts that underlie your organization and its activities in one place. Specifically, you will be able to pinpoint areas where you can make a difference without trying to simultaneously solve every problem.

With the help of data, you can quantify the various elements of your operations. With data and systems thinking, you will be able to identify leverage points that allow you to best use your resources to promote the right outcomes.

As a data professional, I enjoy designing an elegant information system and believe that the right use of data can have a profoundly positive impact on the world and even help humanity achieve a higher level of consciousness. While many people may feel triumphant about the significant technological advances in data management so far, I believe this is just the beginning of a data revolution.