CONFLUENT

The Tech Exec Guide to Data Products

A Strategic Framework for Moving Past Analysis Paralysis and Ensuring Maximum Value at Every Stage of Your Adoption Journey

Travis Hoffman

Contents

Introduction

IN THIS E-BOOK, WE'RE KEEPING IT SIMPLE: You'll learn a straightforward, strategic framework for quickly choosing the most valuable data products from a large number of options—without getting bogged down in a long, costly analysis. Often, these options are presented in the form of a complex solution that requires significant analysis and rework. We'll show you a smarter (and more cost-effective) way to deal with that too.

When presented with a surplus of data products possibilities, organizations often suffer from "analysis paralysis"[1] trying to unravel a sprawling, complicated system just to figure out their next best action. The fundamental question stakeholders have to answer is, "How do I ensure I get the most value out of my investment in data products?" You've probably only ever considered the cost of your data—how much it costs to store it, move it, or work with it. We're going to show you how to quickly zero in on the data products that deliver the most impact by focusing on your data's value.

Whether you're starting from scratch, or rearchitecting an existing solution, this framework offers:

01 a practical strategy to analyze your problem quickly,

02 a fresh lens for rapid value assessment using heuristics, and

03 a quick-start solution to prove value in the most efficient way.

With this mechanism, you'll identify your best Proof Of Value (POV) projects that will quickly demonstrate the value of data products to stakeholders—no need for an exhaustive, pricey analysis of all options. You'll emerge with a prioritized list focused on maximizing your investment in data products. And if you're working with an existing solution, you'll gain new insights into untangling and optimizing your complex, spaghetti mess[2] architecture along the way.

LET'S GET STARTED.

---

[1] The state where analysis leads to confusion about where to begin, and often an inability to begin.

[2] Any sufficiently complex solution with many interconnections where the architecture diagram is generally overwhelming, indecipherable, and looks like a "bowl of spaghetti."

1 The Challenge
Defining Data Products

YOUR ORGANIZATION HAS a huge number of potential data products, but you're not sure where to start. Your stakeholders are eager to see a quick return on investment, and no one wants to throw money at the wrong thing. Your mission is clear: "Get value from data products" fast, and do so efficiently. But you've got limited resources, looming deadlines, and a need to show results before your stakeholders lose faith. It's a classic Triple Constraint Problem—you need to find the right balance of limited time, money, and resources.

THE HIDDEN HURDLE
Separating Cost From Value

To make things trickier, there is no clear analysis of the value of your existing data types. This is a common challenge for organizations that haven't yet shifted to thinking in terms of data products. Historically, data has been treated as a cost center—"just a cost of doing business,"—and distinct data types have been lumped together in one big cost bucket. There's been little to no detailed investigation into the actual value each distinct data type brings to your organization.

Your goal is to identify the data products that will deliver the most value to your organization. But right now, you only have a grasp of the costs, not the value. The question is, how do you quickly choose the best data products to prove ROI, especially with no existing framework to guide you?

THE BEST PROOF OF VALUE
Quick Wins, Big Impact

The best way to start is with a quick win—a small experiment that proves the value of data products and secures more investment. But how do you decide which data products to tackle first? The crux of the challenge is: "Which data products will deliver the highest value back to my organization, with the least time and money, so I can keep my stakeholders on board?"

Doing a deep dive on every potential data product would give you global insight into the landscape of choices, but it would also be costly and very time consuming, delaying your ability to begin quickly. On the flip side, skipping the analysis step entirely often leads to suboptimal choices that don't deliver the impact you need. No single person in your organization has a complete view of the data landscape. Everyone has their own best guesses or pet projects, but you need more than just a gut feeling. Organization-wide strategic objectives may offer some general guidance, but they don't provide enough detail for this level of decision making.

What you require is a fast, reliable mechanism for picking the best data products—without having to comb through every possible option.

THE CLASSIC TRIPLE CONSTRAINT PROBLEM

You need to find the right balance
of limited **time**, **money**, and **resources**

2 The Solution
Progress Over Perfection

SOLVING THIS CONUNDRUM starts with shifting how you think about data—not as a cost center but as a value driver. Then, you need to apply a few heuristics to quickly narrow down which data products are most likely to be your "best" choices. In short, you're going to make "good enough" selections, which value making progress over chasing perfection.

A NEW MINDSET
From Data Cost to Data Value

The key to this is a mindset shift—from thinking of data as an operational expense to seeing it as a value stream—where its continuous flow produces ongoing opportunities for revenue, insight, and impact. And while the costs of storing and moving data might be similar for all your data products, the value each data product delivers varies dramatically. It is critical to orient your thinking towards the truth that "the value of data is in the eye of the consumer."

To implement data products most efficiently, you need a strategy to prioritize data products in a way that makes sense to your stakeholders. This means evaluating data in terms of the value it continuously generates for your organization. In our work with customers, we've found that a data product's value strongly correlates with the number of reuses of that data; the more the data is reused, the higher its value to the organization. If a dataset is reused eight times, it's usually at least eight times more valuable than data which is only used once.

Additionally, we have found there is not a uniform distribution of connections among data sources. As it turns out, the vast majority of traditional data sources in most organizations have very few connections and low levels of reuse—typically just one or two. At the same time, network analysis finds that there are usually a small number of systems (typically 1-2 per domain) which have a much higher number of connections and are responsible for the highest amount of data reuse.

This tells us something critical: The 80/20 rule[1] applies to data products too. More importantly, it gives us a way to quickly compare the value of data products with a lightweight analysis.

What does this mean? You can almost always use the number of connections among data sources as a meaningful way to estimate their value and quickly prioritize which to convert into data products first.

DATA VALUE BY NUMBER OF CONNECTIONS

[1] The 80/20 rule, also known as the Pareto Principle, states that roughly 80% of effects come from 20% of causes. In business, it often means that a small proportion of inputs are responsible for the majority of results.
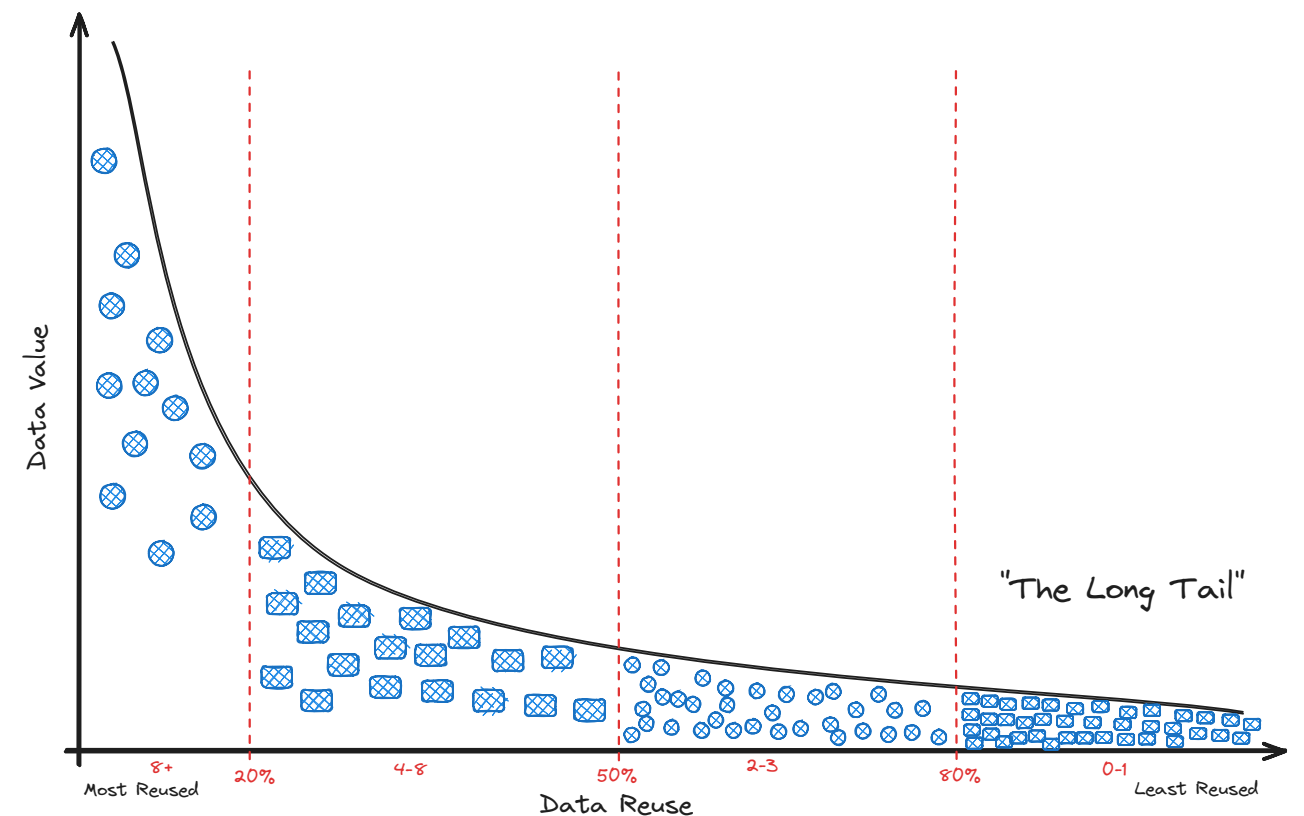
THE VALUE OF VALUE
Finding the True Worth of Your Data

ONCE YOU UNDERSTAND the value of each data product, you can start thinking like a product manager. Just as a product manager evaluates offerings based on their performance, you should assess your data products using three key factors: Value, Cost, and Risk.

One common approach to prioritizing products is the R.I.C.E framework, which is short for Reach, Impact, Confidence, and Effort. This approach can be important for justifying the work of a data product versus other kinds of products (e.g. frontend features)—but that requires a more detailed analysis of the objective value of data products. This kind of analysis can be done with data products which have extrinsic monetary value (e.g. bank transactions, trade confirmations) that relate directly to real-world exchanges of money.

However, this kind of analysis becomes very difficult when it comes to most types of data products (e.g. customer, addresses, etc.) in the operational estate. And it isn't much easier in the analytical estate. But these hurdles aren't as daunting as they seem; we've found that to prioritize, all you really need to do is focus on the Rate of Return of your data products. This is because all data has effectively the same cost basis—a direct, linear relationship between

quantity of data and the cost of serving that data. Furthermore, the cost of implementing a data product typically averages out across all data products with few outliers and is usually a one-time, upfront expense. Over time, this cost is outweighed by the data product's long-term rate of return.

It's tempting to focus only on the upfront costs of implementing a data product, treating it as a one-time project with immediate returns. But the real value comes over time. Because streaming data is a continuously generated resource, the longer you operate a data product, the more value it generates.

This is why Rate of Return becomes crucial for evaluating and comparing the longer-term impact of your data products. The greater the Rate of Return, the bigger the impact a data product makes on your portfolio's Rate of Return.

You don't need a perfect, exhaustive analysis to prioritize data products—a relative ranking of value over time will do. The key is consistency; as long as you evaluate data products the same way each time, you can make decisions that maximize returns without getting bogged down by analysis paralysis.

REAL VALUE COMES OVER TIME

Because streaming data is a continuously generated resource, the **longer you operate** a data product, the **more value** it generates.

Picking the Best Data Products for Maximum Impact

WHEN IT COMES to prioritizing data products, focus on those with the highest "reward slope" (Value divided by Cost). This gives your portfolio the biggest return on investment with each step forward. In the diagram to the right, the data products closest to the top-left corner have the highest rate of return and will demonstrate the greatest Proof of Value.

However, it should be noted that data products which show the greatest Proof of Value don't always demonstrate that value quickly. If speed is your utmost priority, then you may want to focus on a Proof of Concept instead. In that case, look for the data products with the lowest absolute cost to implement, which will let you demonstrate value as fast as possible. In the diagram below, the data products closest to

the left edge (regardless of where they land on Value) have the lowest cost to implement and will demonstrate the fastest Proof of Concept. Proof of Concept should only be used when time-to-value matters more than long-term ROI.

Sometimes, you'll find the value of several data products is very close, or approximately the same. Perhaps you've uncovered a 4-way tie for maximum data reuse. Or it may be difficult to decide among several similar options for other reasons. In this case we need to remember our primary goal: to get started and deliver value as quickly as possible. In such cases you may need to consider organizational, non-technical factors such as the budgets, timelines, and availability of resources of the group who will own each data product.

IN SUMMARY

You're assessing your portfolio of potential data products and identifying the choices which provide the greatest Value, lowest Cost, and least Risk. In the next section, we'll detail the process to map these three dimensions (Value, Cost, and Risk) to the best data products based on a RICE assessment, organizational goals and objectives, data product value, and implementation cost.


DATA PRODUCT VALUE VS. COST
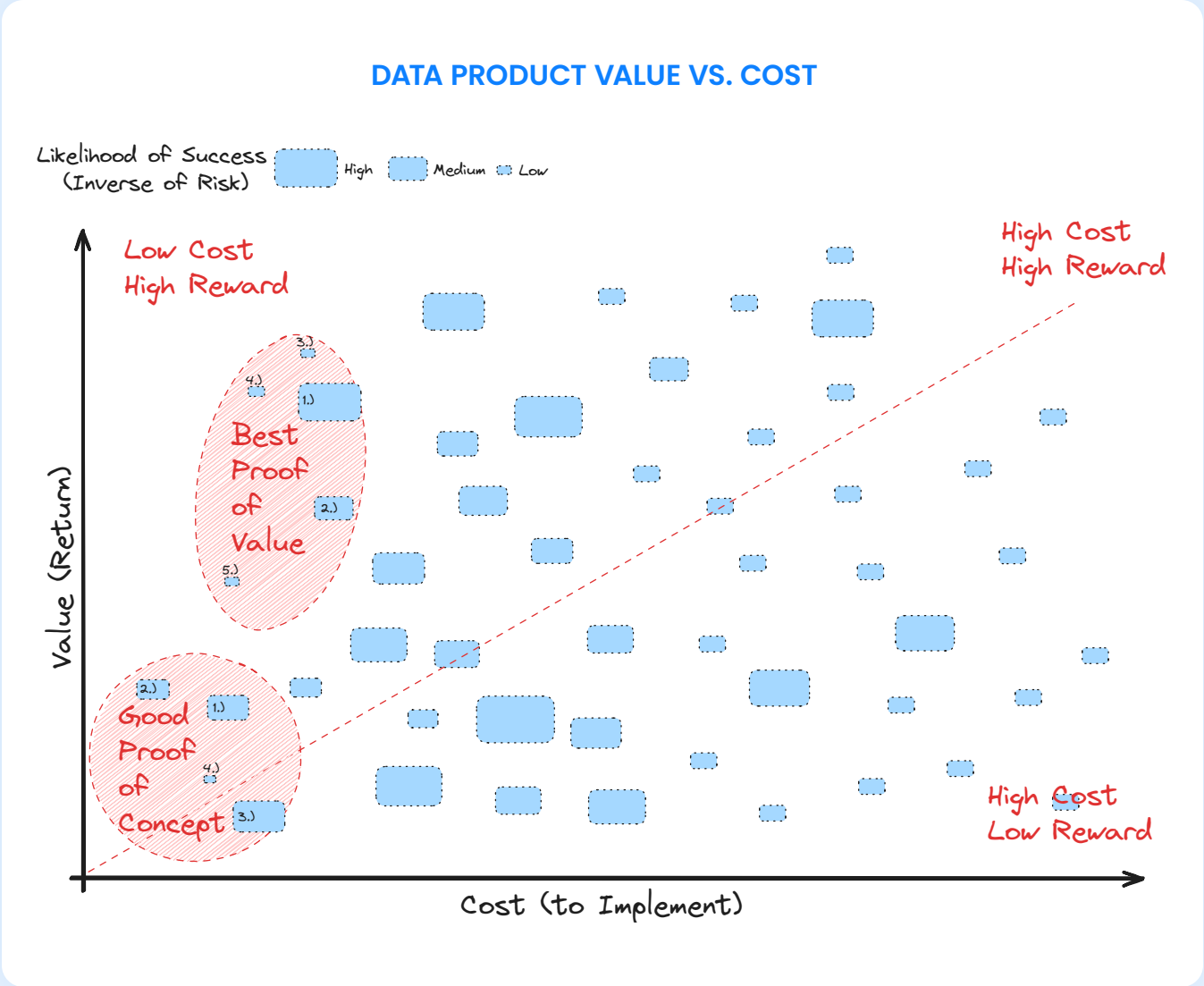
3 The Process

R.I.C.E. Prioritization

WE MENTIONED R.I.C.E. prioritization earlier; it is a popular approach for prioritizing a portfolio of products or features based on four factors: Reach, Impact, Confidence, and Effort. In short, Reach and Impact work together to determine how valuable a feature is.

For our customers, we've found that the R.I.C.E approach can feel overwhelming when starting with a large number of options. When evaluating data products for the first time, product owners often lack concrete examples to calculate Reach or Impact. As a result, attempting a R.I.C.E.-like analysis to find absolute answers for each data product tends to lead to analysis paralysis.

We've found that starting with simple heuristics is often enough to quickly estimate the Reach and Impact of data products, helping answer an easier question: "Which data products are most likely to be among the 'Top N' most valuable?" Our process is not a replacement for R.I.C.E.; it's a way to get started with a bias toward action.

Once you've identified your "Top N" data products,you can later conduct a more detailed analysis to calculate their R.I.C.E. scores. However, most of our customers prefer to stay focused on identifying the next "Top N" data products.

R x **I** x **C**

Reach   Impact   Confidence

E

Effort
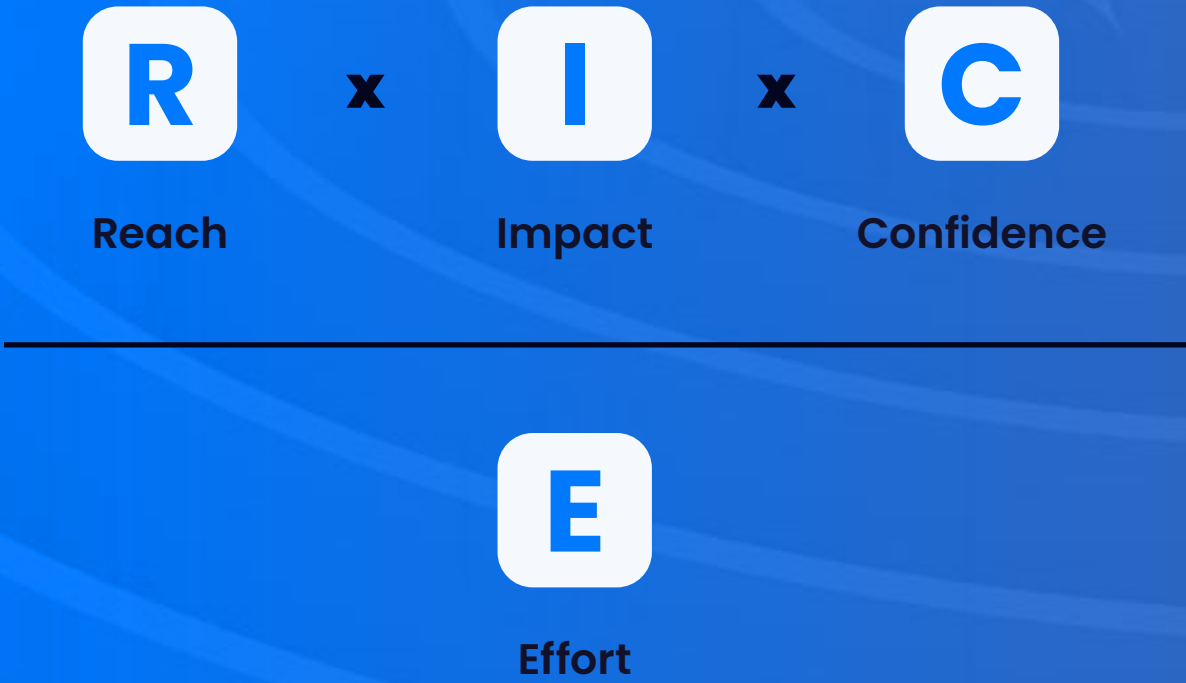
Getting Started

LET'S DIVE INTO the actual steps for narrowing down the right data products for your organization using a series of heuristics. These can be applied in the order presented or adjusted based on your assessment of how quickly they will reduce the search space. The process begins with shifting your mindset to focus on the value of data products.

Align to Strategic Objectives

A great way to reduce your data products search space is by aligning to your organization's strategic objectives. If your company has a clearly applicable goal, like focusing on a particular domain or subset of systems, you can often immediately (and sometimes dramatically) shrink your search space. This can quickly eliminate a large number of data products, saving you analysis time and energy.

For example, if you're evaluating 1000 data products across 20 domains (averaging 50 data products per domain), and your organization's focus is on one domain, you can immediately cut down your list to the 50 data products that are most relevant to your business goals. That's a 95% reduction in your search space!

However, it's important to be mindful that this heuristic can have downsides. If your organization's objectives are to have data "work better across domains," you may need to start with a different heuristic if you can't easily find examples of data products being used across domains. Another consideration is that your organization's strategic goals are not guaranteed to reduce down to the most efficient use of resources from a global perspective. You may end up selecting locally optimal "Top Three" data products that reflect the highest value in a particular domain, but would be considered suboptimal when evaluated on a global scale. Thankfully, there are other options to help address these scenarios. Read on for other heuristics to try.

Note: It may also be the case that there are multiple (often conflicting) priorities which organizations are tracking.

QUICK TIP

Your goal is to quickly get to your "Top N" data products, not to assess every data product. You don't need to exhaustively analyze every single data product with every single heuristic—you can just apply heuristics to the remaining options. This simple approach helps to rapidly reduce the search space with confidence that you're always focusing on the most important data products first.

NARROWING DOWN THE DATA PRODUCT SEARCH SPACE
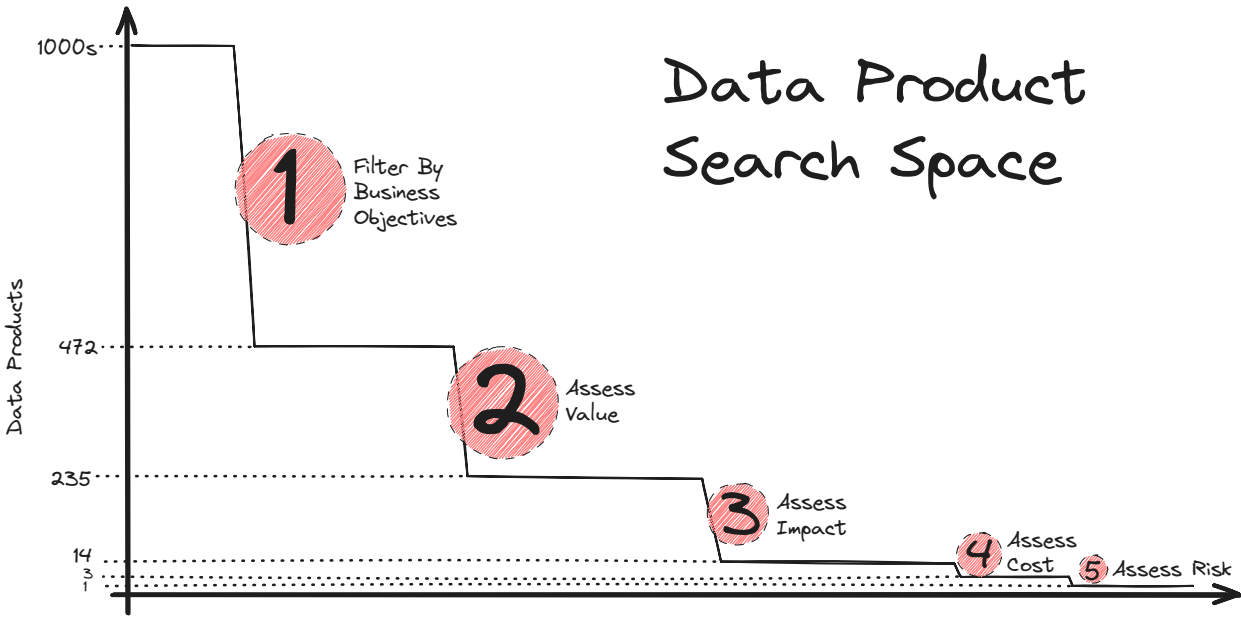
QUICK TIP
Apply heuristics in whatever order will reduce your options the fastest at each step.

Collect Wisdom

THE NEXT APPROACH is to use crowdsourcing to get an initial estimate of relative value. We recommend the "$100 Technique," an approach commonly used in agile planning. In this approach, you ask each of your data consumers to invest a limited budget on the data products, systems, integrations, or domains that they feel deserve investment. This wisdom-of-the-crowd approach does a good job of prioritizing a list of options based on a complex set of considerations.

When setting up the $100 Technique, there are a few key considerations. This technique works best with a sufficiently large number of participants. It is also important to evenly distribute your panel of participants so that they cover all key data domains. If a majority of your participants are from one domain, your answer will likely be biased toward that one domain's needs; do your best to draw equally from all domains.

Also, keep the options manageable. If you introduce too many choices to your participants, they might get overwhelmed. Consider a two-step approach, where you first poll participants for which domain(s) to focus on, and then dive deeper. Sometimes, it may be necessary to do some preliminary filtering based on another heuristic.

Another approach is to let participants identify the data products for themselves. This requires that you do a bit more pre-work to level-set with the audience you're polling. For example, if the organization's objective is to "better work across domains," then you may need to focus the audience to specify the most important data products from a different domain.

PRIORITIZING USING THE $100 TECHNIQUE
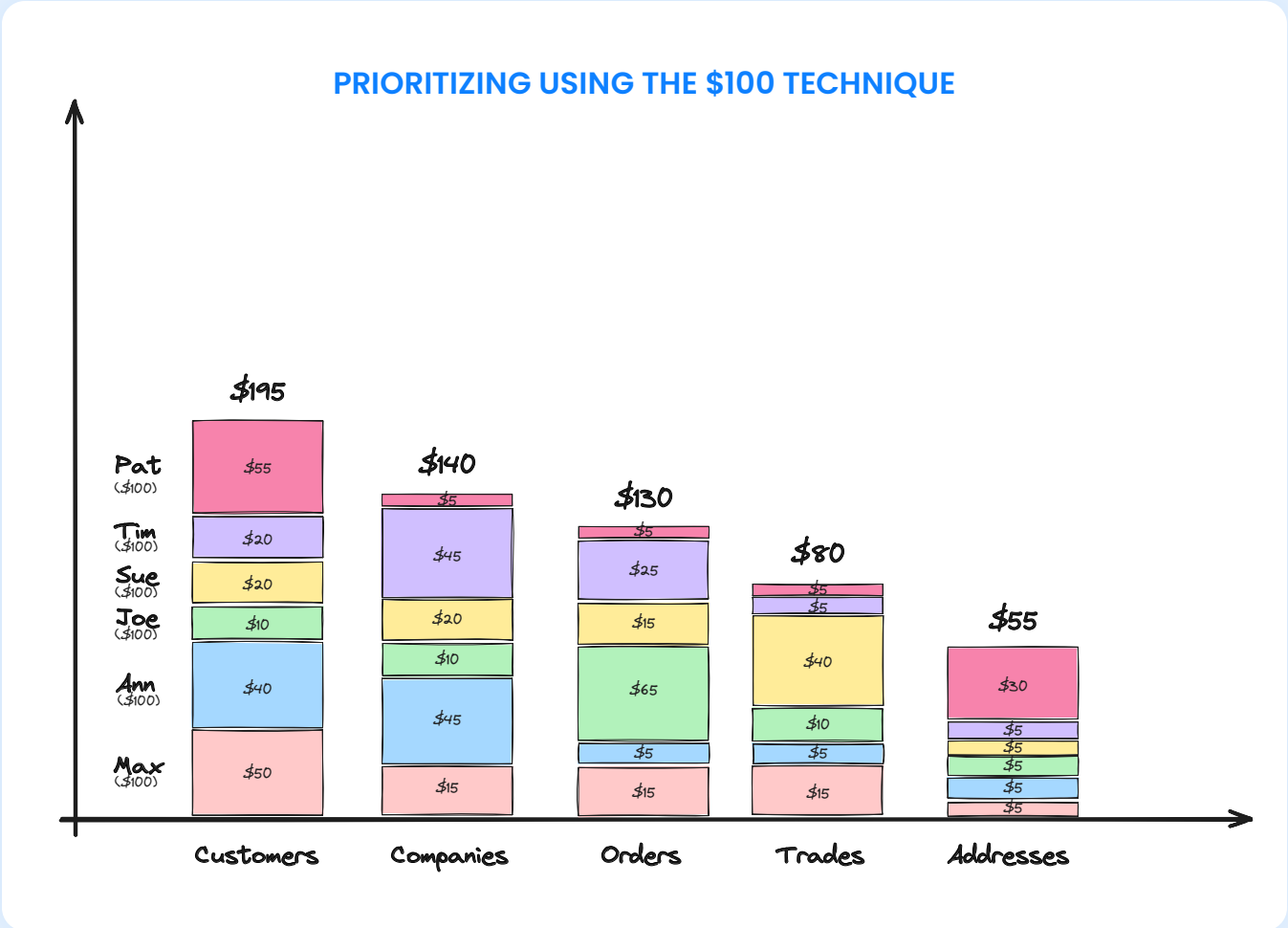
Identify Domain Drivers

NOW, LET'S LOOK at the architecture that supports your data products. We've found that in spaghetti mess architectures, about 20% of your systems will handle 80% of the integrations—the classic 80/20 rule. These are your "core" systems, and once you've identified them, you can create simplified architectural diagrams through a series of easy steps. We've found that it is almost always possible to focus on one (or two) key system(s) per domain.

To make the interactions clearer, we've created the Hoffman Diagram, which helps to better visualize the most "core" systems and their integrations.

The core systems are shown as colored rectangles on the left side of the diagram. Integrations between core systems are represented by one-way, vertical arrows between the dashed lines. This diagrammatic approach provides greater clarity on how an organization's most central systems are connected with each other.

The primary benefit of this approach is that it reduces (or can eliminate) the criss-crossing of integration lines, allowing more space for them to connect to core systems and align vertically. Secondarily, you can also represent the sequence by which the one-way interactions happen.

Once you've identified the "Core System" for each domain, you can draw one or more diagrams which emphasize your core system(s). We recommend doing one diagram per domain for legibility.

As you can see in the expanded Hoffman Diagram to the bottom right, the integrations between the Blue Domain core system and all other systems are much more readily traceable. Beyond improved readability, this diagrammatic approach allows you to document not just that there is a connection between systems, but also the full details of connection types and data exchanges, which are typically represented as a single line in a "spaghetti mess" integration diagram. The primary benefit of the Hoffman Diagram is its ability to visualize and inventory all data types exchanged between systems while keeping the diagram easy to understand.

HOFFMAN DIAGRAM
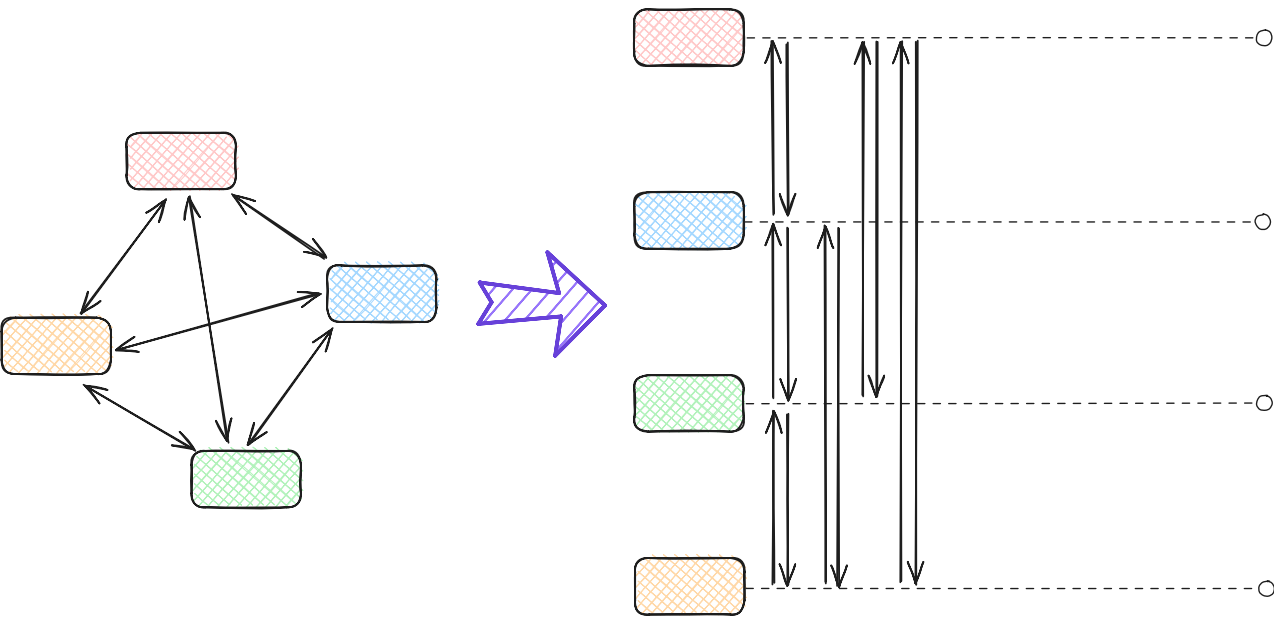
PER-DOMAIN HOFFMAN DIAGRAM

Focus on Blue Domain Core System:
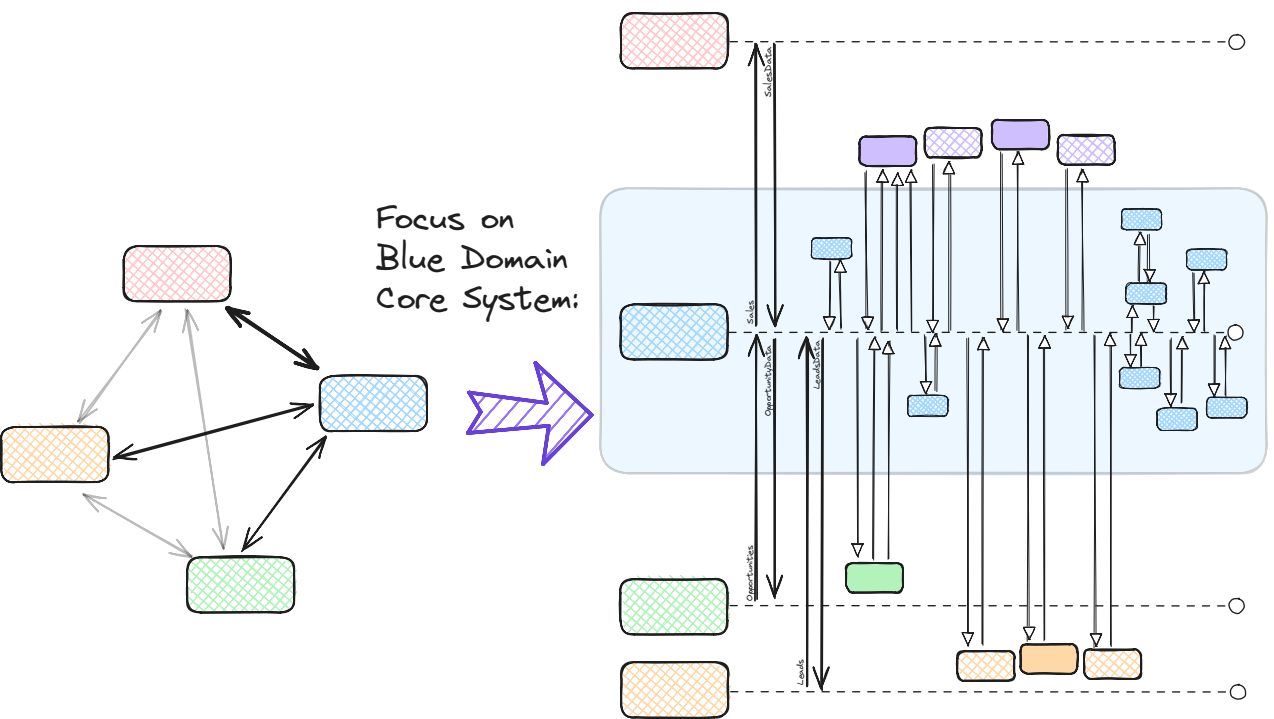
Inventorying the Data Mess

IN OUR EXPERIENCE, most "spaghetti mess" diagrams only show that there is an exchange of data happening, not what specifically is being exchanged. Unfortunately, because they're visually complex and confusing, they inevitably summarize multiple integrations down to just a single, double-ended arrow. This glosses over many details that we need to consider when determining data product prioritization.

To better refine your estimate of data product value, you'll want to create an inventory that breaks down every connection in the "spaghetti mess" architecture, which will help you better understand how often data products are reused. The more times a data product is reused, the more valuable it is. This is where the Hoffman Diagram shines, as it's still readable even when there are a large number of directional connections to represent.

It's critical to track the count of each data type that is exchanged between systems. The count acts as a multiplier on the intrinsic value of a data product. All things being equal, a data product that is reused 10 times is approximately 10 times more valuable than if it were used only once. Simple as that.
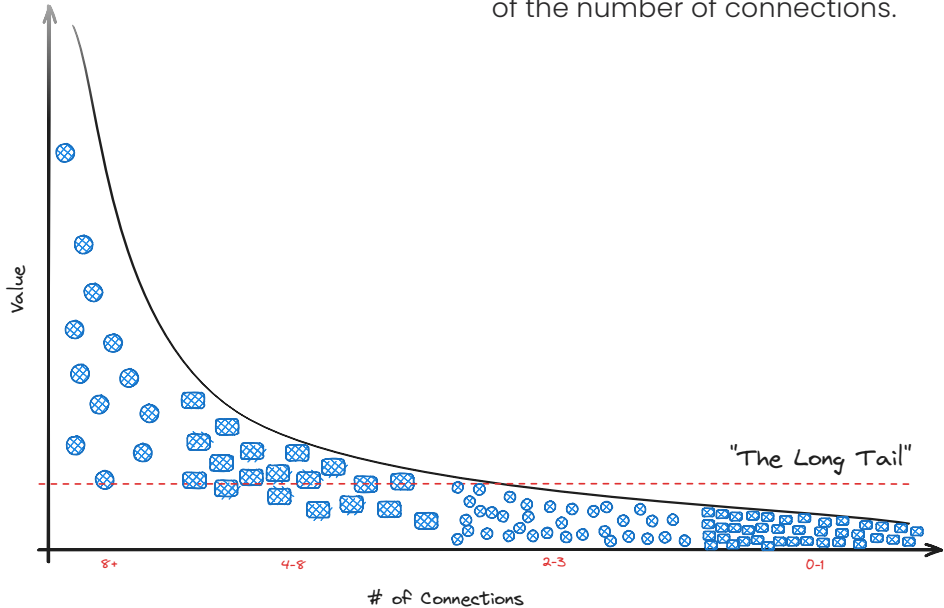
It's important to note that this inventory does not need to be exhaustive. Your end goal is to identify the N most important data products, and you only need to evaluate a large enough sample of data products to ensure that you're finding a globally optimal set of data products.

But how many is enough? Given that N is generally small, and you're making informed decisions (not random sampling), a good rule of thumb is to analyze 7-9x more data products than you need to end up with. You really only need to develop a detailed inventory of just enough systems to confidently identify your most valuable data product candidates.

For example, if you're trying to identify the top three data products from a portfolio of ~100 integrated systems, following the 80/20 rule, you can focus on the top 21-24 systems identified from the "$100 technique" exercise.

In our experience, it becomes evident pretty quickly which data types are the most commonly reused because the distribution is not linear. We find the distribution of data looks like a "long tail" distribution as shown below. This means that systems with the greatest value tend to be very obvious when looked at in terms of the number of connections.

PER-SYSTEM HOFFMAN DIAGRAM
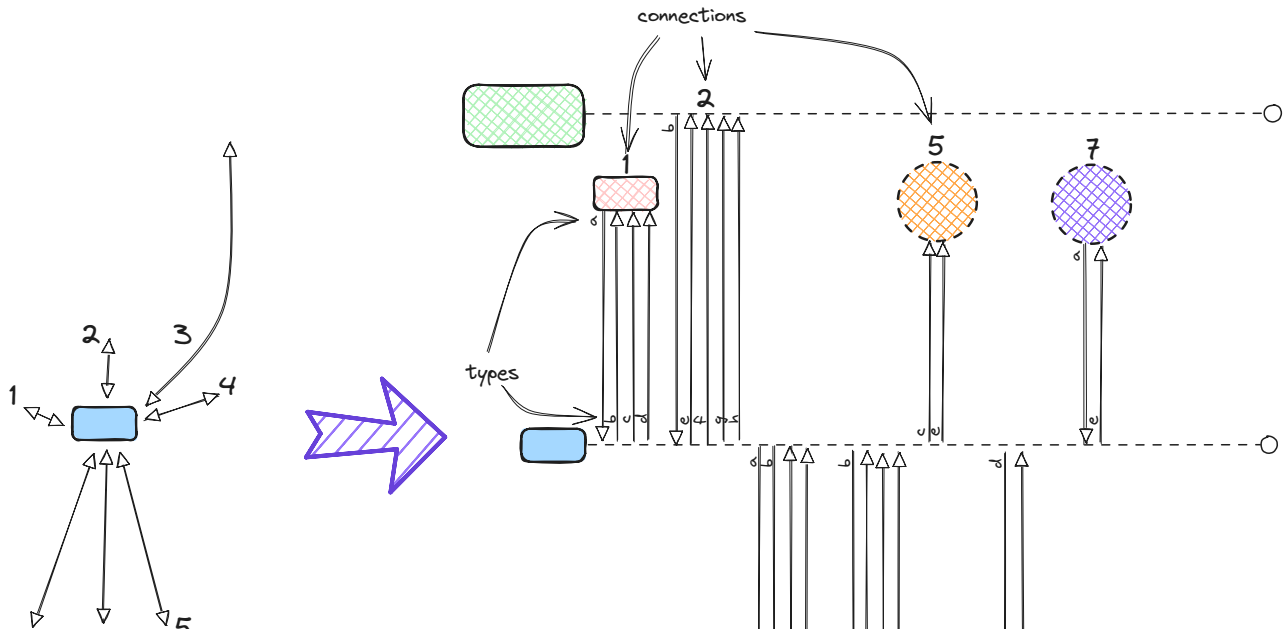
SOURCE DOMAIN	DATA TYPE	COUNT	VARIATION(S)	DESTINATION DOMAIN(S)
Sales	Leads	12	QualifiedLeads, EnrichedLead	Sales, Marketing
Sales	LeadData	10	EnrichedLeadData, FullyEnrichedLeadData	Marketing
Marketing	Company	9	EnrichedCompany, FullyEnrichedLeadData	Marketing
Sales	CompanyData	8	EnrichedCompanyData, FullyEnrichedCompanyData	Sales, Marketing
Revenue	Accounts	8	AccountData	Sales, Marketing
Revenue	Opportunities	7	OpportunityData	Sales, Marketing

ONCE YOU'VE COMPLETED your detailed analysis, prioritize system value based on how often the same "type" is reused. In our experience, it isn't necessary to split data types into minor variations. Instead, focus on what can be grouped together into a common data product.


Estimating Reach and Impact

At this point in the process, we have a rough estimate of the relative reach of a large number of potential data products based on the number of times data types are used. By tracking the Destination Domains and/or destination systems, we can also get a better understanding of the impact of the data products.

If you end up with a large number of ties, or no clear guidance on which domain(s) and system(s) are more impactful, it may be necessary to do a deeper-dive R.I.C.E. analysis on the "Top N" data types to come to a relative prioritization. You can also consider applying the $100 Technique again on this limited set of "Top N" data types.

Whichever approach you choose, to prevent this from causing analysis paralysis, focus on a small, limited, and fixed number of potential data products.
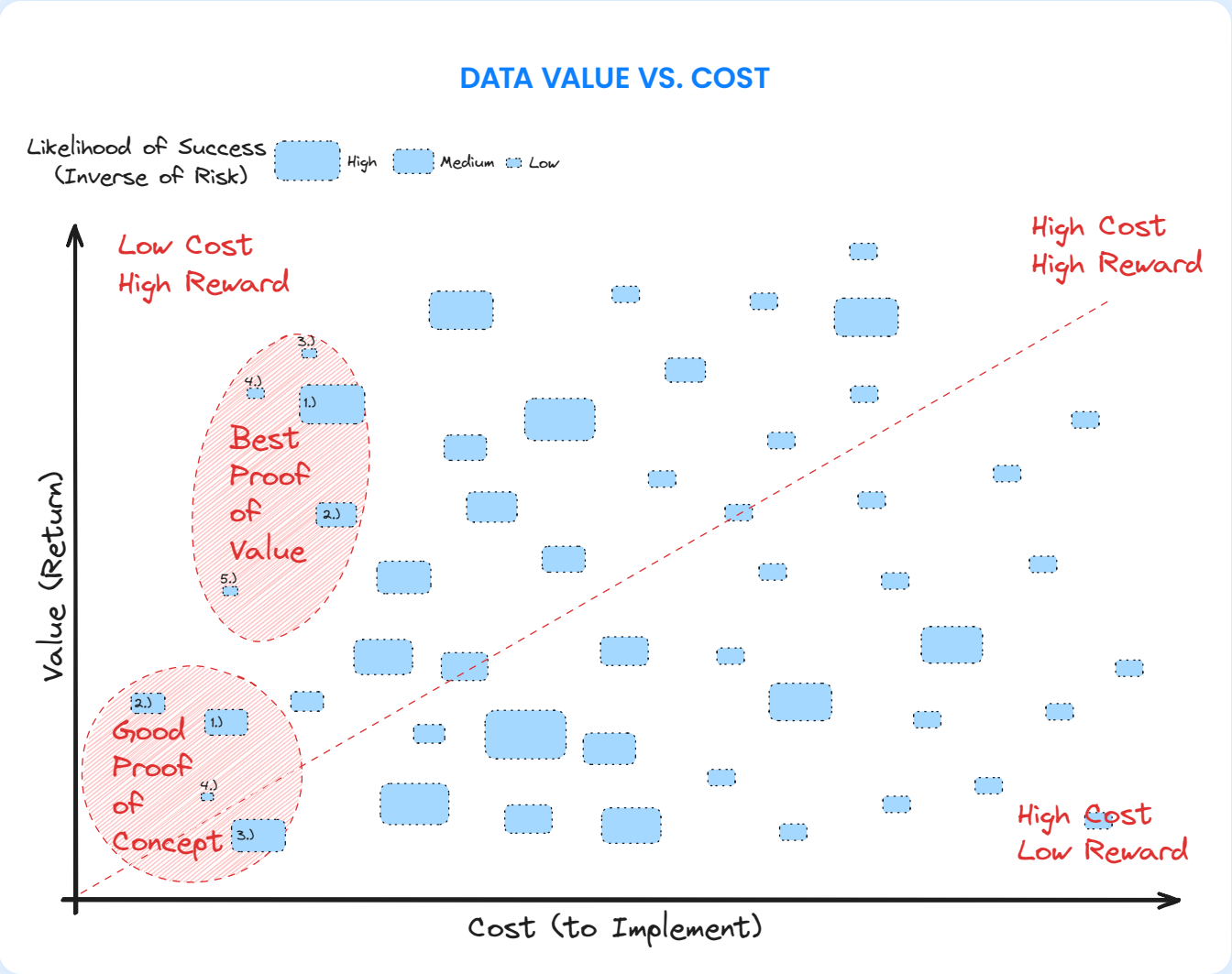
Focus On Best Return

The last step is to compare the "Expected Value" vs. the "Cost to Implement" of the "Top N" assessed data products. If an exhaustive analysis has been done, it would look something like the diagram above.

However, because we've been using our heuristics, we have eliminated most of the possible answers and are focusing solely on the "Best Proof of Value" and "Good Proof of Concept" options at the left side of the graph.

If it wasn't already clear, the chart above illustrates how much analysis our heuristics have enabled us to skip over. If you've applied the heuristics well, you should avoid more than 80% of a complete analysis of every data product.

4 Moving Forward

AND, THAT'S IT. It's time to stop reading about data products, and start doing them through the Confluent Data Streaming Platform. And, Confluent provides the resources to go beyond implementing data products.

CONFLUENT

The Data Streaming Organization

Driving Value & Competitive Advantage From Data Streaming

VICTOR GATUNE, TRAVIS HOFFMAN, JAKUB KORAB, MARK MCCALLA, MIKE OWENS, JON VINES, ERIC WINTER

You can leverage the **Data Streaming Organization** framework to equip your organization to meet and exceed any data goal with the right tools, mindset, and technology in partnership with Confluent's Executive Advisors.

Read it here

⚡ Additional Resources

Conquer Your Data Mess With Universal Data Products

Shift Left: Unifying Operations and Analytics With Data Products

Creating Reusable Data Products: Is Shifting Left the Key?

2024 Data Streaming Report

Tech Exec Resources

Confluent Data Streaming Platform Overview

CONFLUENT