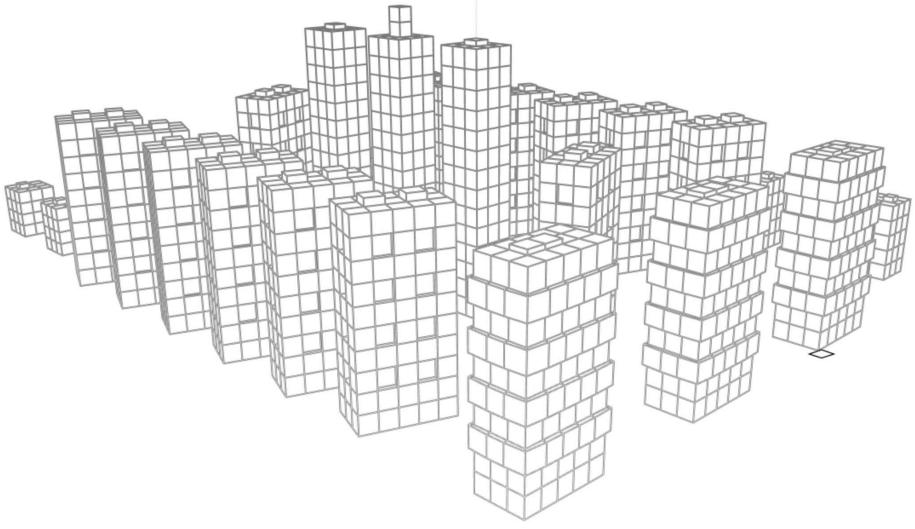
SPHERE BY CRUX

The True Costs of Data Integrations

The Comprehensive Guide to Understanding the Economics of External Data Integration



Executive Summary

Companies that rely solely on internal data, leave gaps in their decision making, missing latent opportunities and threats on the table. The exponential value of integrating internal and external (third-party) data sources looks like 1+2=10. The insight gleaned from combining a business's critical data with financial market data, sustainability metrics, benchmarking data, weather information, etc. empowers decision-makers to address the blind spots, better anticipate trends, mitigate risk and gain a competitive advantage.

Yet, this level of data integration has its challenges. The difficulties and expenses often arise when it comes to integrating, supporting, updating, and harmonizing external data throughout its lifecycle. This is not a download-and-forget-it scenario. Data integration costs are like an iceberg. The visible part includes the cost of purchasing and analyzing the data; however, the unexposed portion that greatly exceeds these costs is the ingestion and maintenance phases which cost more than the initial procurement. third-party data must continually be updated and maintained to provide the utmost value.

This report is designed for those planning or approving an external data integration initiative helping to assemble a complete picture of the financial requirements throughout the lifecycle of an external data integration project. It explores the many areas of data integration economics and the true costs associated with acquiring, ingesting, and maintaining external data which even the largest enterprises are underestimating by 35-70%. This is due to the unique challenges associated with cost attribution that leads to six critical costs that are often overlooked, the lack of a working framework, and inconsistency in the definition of a dataset across the industry that significantly magnifies the situation. After reading this report you will have a reference guide for developing an improved dataset economics business case and a framework to assess how to reduce those costs.

GLOBAL CRISIS DRIVING URGENCY FOR EXTERNAL DATA NEEDS

External (third-party) data has become critical to running a business. To stay competitive and best represent the needs of stakeholders, companies must proactively use data from outside sources to gain insights into new trends, uncover opportunities, gain market share, and mitigate risk. In addition, enterprises require greater transparency into who they are doing business with, which includes procedures to ensure that they conform with financial and environmental regulations and socially responsible sourcing.

> *"According to a recent MIT Sloan Management Review report, the companies making the most innovative use of data and analytics were more likely than others to leverage external data sources, including social, mobile, and publicly available data. A separate study found faster-growing companies were more likely to be considering the use of external data than those with lower growth rates."*
>
> Source: Deloitte

Before we explore the rising demand for external data, let's first clarify what external data is and where it comes from. Most enterprises have relied on their internal data (sales, operational capacity, etc.) for making business decisions. However, as businesses become more interconnected, and data becomes more readily available, the need for data inputs and insights outside an organization is much more critical. This is where external data comes in. External data can be anything such as competitive intelligence, weather information, financial market data, customer transaction, or geolocation data that provides a deeper insight into the world as well as identify opportunities and risks outside the organization. This data is especially valuable when combined and analyzed alongside internal data sources.

Organizations can access external data from thousands of sources, which include public sources (such as government entities or non-profit organizations), data vendors (such as Bloomberg or Nielsen), and alternative data suppliers (such as point of sale data panels or web scraping research companies). The diversity and fragmentation of the various external data types and sources make it challenging for companies to discover, understand, ingest, and analyze all these inputs.

The COVID-19 pandemic has exacerbated these challenges for businesses as a sharp rise in supply chain issues, border closures, and manufacturing disruptions force organizations to make difficult decisions in less time than ever before.

There is an urgent need for more external data to help meet these challenges and eliminate the unknown. Research recently conducted by Crux, found that 79% of data leaders in the United States consider external data "very valuable" to their work, while a separate report from Deloitte found that 92% of data analytics professionals say their organization needs to increase the use of external data to enrich its advanced analytics. And IDC reports global demand for external data spending is expected to increase 13% every year until 2025. Yet, many organizations struggle with the maturity of their data and to understand the full cost of their data integration efforts.

Deloitte reports 92% of data analytics professionals state their company needs to increase the use of external data to enrich their advanced analytics.

Source: Deloitte

But once you find it, what are the costs associated with onboarding, integrating, and monitoring new datasets? How about remediating changes? How many datasets can your business afford to ingest this year, next year, or the year after that?

The total costs of external data integration are significantly underestimated by even the largest enterprises. These 'hidden' costs are enormous and have a direct impact on an organization's operating costs, productivity, and profitability.

THE CHALLENGE OF COST ATTRIBUTION

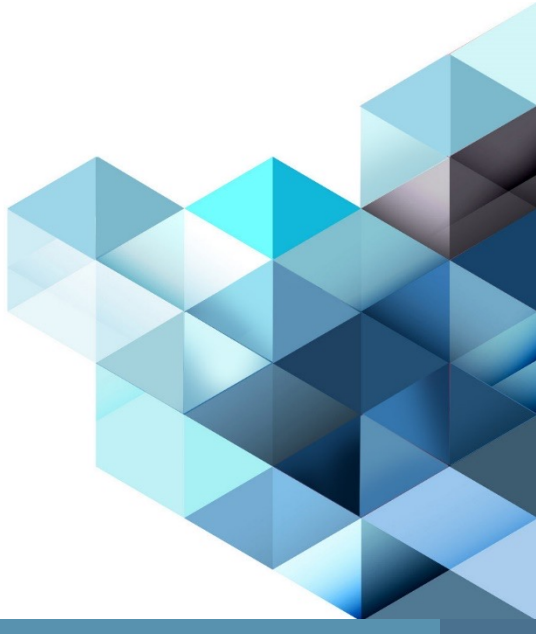
There are two major reasons why organizations struggle with the economics of external data ingestion.

First, most organizations don't have a centralized approach to managing their external data resources. With pipelines and datasets spread out across siloed departments or systems, even those who want to achieve a full attribution of their external data costs have a hard time doing so.

Second, many organizations lack a sufficient understanding of what constitutes a dataset, which makes accurately identifying the cost of integrating and supporting external data throughout its lifecycle extremely difficult.

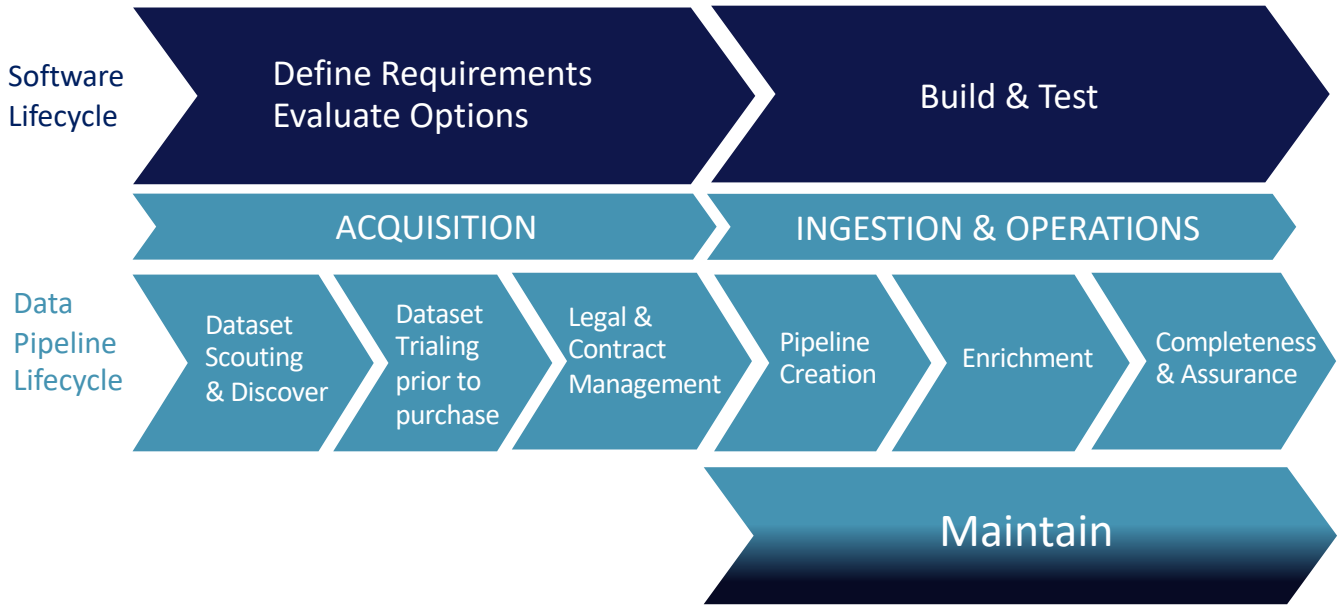
To an executive, the cost of a dataset often refers to a static business asset without technical characteristics such as engineer salary, database costs, and engineer productivity; whereas data engineers would likely define those costs by their resource, productivity, and licensing costs; in turn, infrastructure engineers will have a separate set of concerns that revolve around security, networks, storage, and availability of those datasets to end-users.

In each case, they are partially right about their economics and ongoing costs, but without a common understanding of what constitutes a dataset and its lifecycle, it's impossible to properly define the true economics of a dataset.



When you dissect the end to end activities associated with identifying, onboarding, curation, managing and maintaining datasets, it becomes obvious that there's no need to invent a new data lifecycle for datasets. The similarities between software development and support, 'config as code' and datasets, is stark.

Figure 1: Similarities between the life cycles of software and data pipeline development



DATA PIPELINE ECONOMICS PARALLELS SOFTWARE DEVELOPMENT

Organizations don't need to reinvent the wheel to estimate the costs of data pipelines. As shown in the upper portion of Figure 1, the software development lifecycle shares the same key characteristics of an external data pipeline, which makes it an excellent framework for the data pipeline development lifecycle and understanding the complete cost model.

The four main characteristics that software development and data pipeline development share are:

Identification, Evaluation, and Proof of Concept

Just as with software applications, the lifecycle of an external data pipeline begins with a phase of rigorous requirements gathering and analysis. Exploring the context, requirements, and desired outcomes of a data problem enables stakeholders to better understand their needs and identify the best available resources. The next step, data trialing, helps determine whether the chosen dataset is worthy of further investment, similar to how a proof of concept (POC) functions within the software development lifecycle.

Build Phase Includes Code and Technical Configuration

An external data pipeline is very similar to code. It consists of code and configuration that extracts raw data from a source, transforms it until it conforms to the required standard, and loads it into a destination.

In the software development model, a decision would be made to build or buy the solution. Although the same decision exists in the data engineering field, it's often overlooked in the world of data, though it has serious cost implications.

As with software development, there are direct operational costs associated with developing the code, including the hiring of data engineering resources and the productivity of those resources. The indirect costs include the infrastructure requirements to support the pipeline and the costs to manage the infrastructure to keep the datasets up and running.

Finally, data engineers must test both the underlying dataset and the data pipeline itself. This helps to remove null values, reveal incomplete data and identify other common data quality issues while ensuring that new data pipeline outputs conform to the requirements.

Maintenance is Often Overlooked

There's a tendency in both data pipeline and software development to orient team effort around instantiation, or when the product goes "live." Focusing on the launch of a software application or onboarding of a new dataset obscures the importance of day-to-day operational costs. These costs cannot be ignored if we want to achieve a full-cost attribution.

As per the software development process, effort doesn't stop after an application is delivered to end-users. Engineers spend a significant amount of time troubleshooting and maintaining the application after it's delivered to ensure that it achieves its desired goal and continues to work optimally.

Data pipelines are no different. Pipelines must be updated, analyzed, and maintained just like a software application to ensure that the data is delivered as expected in the format needed. Maintenance is often overlooked and is a significant cost driver not just in the first year, but each year after that. Drawing from popular software maintenance studies, developers should provision

between 15-25% of the total software development budget for annual maintenance and support. According to other research studies, the lifetime maintenance costs typically fall within 40% to 80% of the original build, with an average of 60%.

The type and frequency of maintenance will vary by dataset. For instance, even small changes to a dataset schema such as a new unit of measurement will require a period of analysis, adjustment, and testing. Luckily, large data suppliers don't frequently change the dataset of their schemas to avoid disrupting their large customer base, who rely on the data to inform critical operations.

Less established data providers, whose offerings are in a state of constant evolution, present a different set of challenges. Frequent and sometimes dramatic changes to the dataset schema, whether revealed to consumers or not, require a significant amount of both attention and maintenance work. Note: This level of maintenance wouldn't apply to static datasets.

*Budget between 15% - 25% of the build cost of an external data pipeline for maintenance.**
— Gartner

With the background on some of the factors that go into the economic model, you can perform a detailed cost analysis of the data pipeline lifecycle, uncovering "hidden" costs when using a comprehensive cost model.

CALCULATING TOTAL COST OF DATA PIPELINE OWNERSHIP

Financial leaders are looking for traditional ROI business cases on large-scale investments. The total cost of ownership (TCO) is a useful model to measure the investment required at the time of purchase (licensing) as well as other long-term and short-term expenses for supporting and maintaining a data pipeline. It's a critical metric to identify the total economic impact from the cost of the data sourcing to pipeline build and scale the operations over a 3-5 year period.

*Source: Robert L. Glass, "Frequently Forgotten Fundamental Facts about Software Engineering" (IEEE May/June 2001)

COSTS ASSOCIATED WITH DATA ACQUISITION, INGESTION, OPERATIONS, AND MAINTENANCE

The bulk of the costs associated with external datasets fall into four categories: acquisition, ingestion and operation, maintenance, and the human capital required at each of these phases.

ACQUISITION COSTS

The data acquisition process is a seminal phase of this external dataset process. It is where quantitative analysts, scientists, and data scouts not only procure potential datasets, but also trial them to ensure they will be a good fit and will satisfy the business's requirements. The costs in this pre-ingestion phase can be quite high, even though the actual integration hasn't even started yet.

+ Data Scouting

Finding the right datasets for your analytics goals is the first step. The process begins when a data acquisition specialist or "data hunter" starts to navigate all the available datasets for specific resources that align with your analytics goals. An informal Forrester survey found that nearly 33% of global organizations surveyed reported having a data hunter or equivalent role in their organization. Although there are costs to scouting whether you are doing it in-house or hiring an external resource, the costs need to be identified and budgeted. Oftentimes, this budget line item sits in the department requesting the dataset, and without a centralized model for sourcing this information, it is not uncommon for the same dataset to have been purchased multiple times across various departments with no insight into these duplicative expenditures. Centralized procuring or budget approvals help to eliminate this duplication.

> *1. Cost to license the dataset includes: entitlement, time, cost of in-house or external data hunter.*
>
> Source: Blog

+ Data Trialing Lacks Standardization

The costs associated with data trialing are unavoidable. Once an acquisition expert has identified the right assets, the data engineering team must implement the basic pipeline to trial a sample of the data to ensure it's the right match. The cost range here is dependent upon the state of the supplier's data and how similar it is to the target data environment.

For even a small dataset, the cleaning, wrangling, and testing of trial data will consume engineering cycles, while the trialing of complex datasets can easily become days of sunken cost. Often, this effort leads to a wasted effort if it is determined that the data isn't the right dataset. The costs are straightforward — the time it takes to get the trial up and running and the cost of the data engineer's time.

To make matters more complex, there is no standardized method of gauging the quality of new datasets. Curated trial data has often been scrubbed clean of error by the provider, obscuring the real effort needed to operationalize the dataset. The only way to accurately assess that workload is for your data engineers to onboard a sample of real historic and ongoing data, which of course involves significant cost.

2. Calculate the data engineer's salary/hour. and multiply it by the number of hours spent trialing the data.

INGESTION AND OPERATIONS

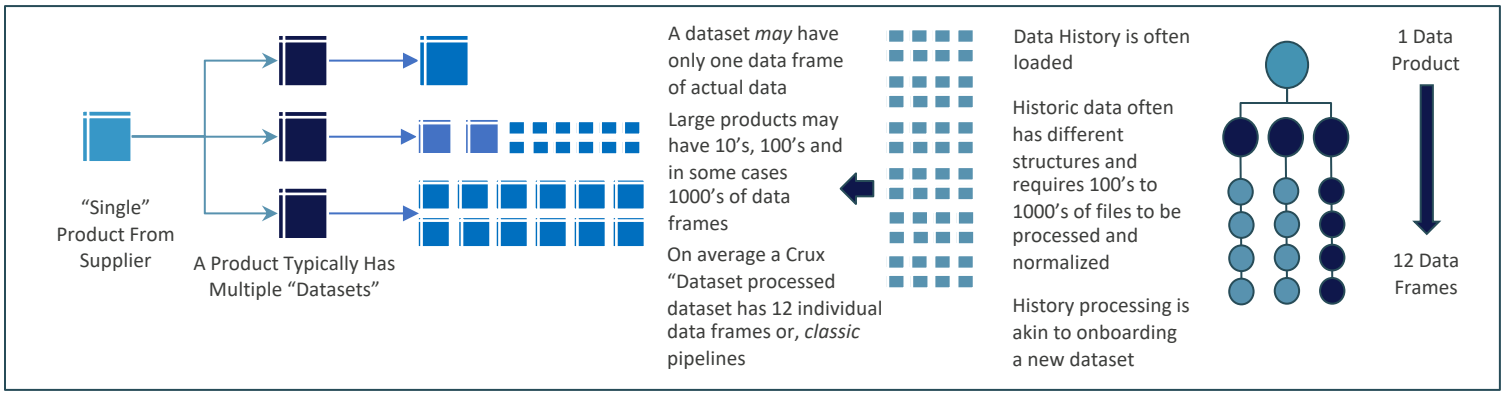
Once the third-party data is tried, proven to be worthy, and then acquired, it needs to be ingested into the organization. This is a highly complex phase as the new data isn't often designed in a manner that is a one-to-one fit with existing datasets. A certain level of unpackaging, normalization, quality checking, data platform development, and ongoing dataset operational work must occur - and this requires a significant time commitment from your data engineers. A good portion of their daily responsibilities must be allocated to this process to ensure program success.

+ Data Onboarding & Ingestion

How large and complex is the dataset that you're trying to onboard? How much history is available? How much do you need to load? Many organizations will encounter the term data product when buying datasets from vendors. This is confusing because a data product often consists of several — sometimes several hundred — related datasets packaged together, each of which have unique data engineering characteristics and requirements.

How many data frames does it have? What format is the data in? Some datasets have a few frames, while others have 10s to 100s of them, and they all must be cleansed, validated, and mapped. That means the engineering cost of onboarding a single dataset can be several times higher than what a cursory examination would reveal.

What is a Dataset?



As a baseline, assume that a skilled data engineer can onboard between two and four datasets a month – if focused solely on this task. This number coupled with the annual salary of a data engineer, will provide you with an estimate to start calculating your total cost of onboarding. When thinking about the capacity, assume that one engineer can onboard 24-48 datasets a year. And at the rate new external data is demanded, you will need many data engineers to handle the load, or you need to find a means to supplement the resources of your team with a service provider that specializes in these tasks.

The data ingestion process requires performing quality checks, schema changes, and other transformations on a new dataset to ensure that the new data matches the structure expected by its destination. How much does it cost for a data engineer to perform these tasks? Many organizations are not aware of the time it takes to complete this process and, unfortunately, it can vary significantly depending upon the dataset.

3. On average a data engineer can onboard 2-4 datasets a month.

+ Data Platform

There are significant costs implications on whether or not to build a data platform. Although there's an upfront expense to developing your own data platform, in the long term it is the most cost-effective and dependable way to ensure consistent data feeds. Organizations who choose to skip this step and perform on-the-fly maintenance will dramatically increase their costs as they struggle to maintain "snowflake" pipelines, a term borrowed from the server maintenance field for systems that have bespoke maintenance needs.

Enterprises that are committed to an external data integration strategy should include the cost of data platform build, maintenance, and time-to-value into a comprehensive cost accounting.

4. There are costs to ingest, decode, massage and process external data.

+ Dataset Operations

The costs related to dataset operations are one of the most overlooked aspects of the external data lifecycle. Like software applications, datasets aren't static elements that simply function after they've been onboarded. There are ongoing operational costs to monitoring the dataset and ensuring that it works the way it needs to.

Operational tasks include vendor delivery and management outreach to ensure that the data is received according to the service-level agreement (SLA), staying ahead of updates, ensuring that the feed is updating properly, and verifying that the right files have been uploaded.

For mission-critical data pipelines, these tasks are handled by a 24/7 DataOps team, which has a cost structure similar to other mission-critical enterprise systems. Less important datasets may be delegated to the specific department that uses the data.

A good starting point is to calculate data operations costs of at least 10% of your initial build cost per year and adjust upward or downward as needed.

5. At least 10% of the initial build cost per year needs to be assigned to data operations.

MAINTENANCE

third-party data is a living, breathing entity that can provide ample ongoing value if its integrity is maintained. An often forgotten phase of this entire process is the maintenance phase, where businesses must correct data fields, make minor schema changes, create new frames and merge historic structures with the new ones. This stage cannot be overlooked and can incur significant costs if the first two phases aren't handled correctly.

+ Dataset Maintenance

Along with operations, maintenance tasks are the most overlooked financial calculations surrounding external dataset management.

At least 20% of the build costs should be allocated to maintenance to help keep pace with the volatility of change. Maintenance includes correcting data fields or minor schema changes, as well as major changes where entirely new frames have been added to a dataset or handling major structural changes that require a serious commitment of time and resources. Maintenance also includes merging historical data structures with the new ones to provide a coherent view of the data.

The best data suppliers will help ease maintenance costs with documentation and adequate change notification, while in worst-case scenarios changes can silently undermine the accuracy of your data pipelines without any forewarning.

When calculating maintenance costs, don't underestimate the change management you need to keep data flowing. While all organizations bear these costs, working with poorly managed datasets or vendors who lack a consistent approach to change management will increase costs by forcing data engineers to locate and remediate issues reactively.

As illustrated in the model on page 16, in just a few years, maintenance costs can overshadow the costs of building new pipelines, especially in organizations that are scaling up their data-driven initiatives. To arrive at a complete cost accounting for maintenance, organizations need to adopt a multi-year view of their portfolio's growth that accounts for the compounding impact of maintenance.

6. At least 20% of build costs should be allocated to maintenance costs.

STRATEGIC APPROACH

The approach an organization takes to handle external data ingestion needs is a strategic one that has implications for both the short-term and long-term. Organizations need to weigh the best option from a wide range of solutions to meet their external data ingestion needs, from developing individual pipelines on the fly to comprehensive, enterprise-wide data platforms, and everything in between. The architecture of each of these varies from highly centralized to highly distributed solutions. Regardless of the route chosen, most enterprises are looking to improve one or more of the following:

- Resource constraints and capability
- Capacity and speed
- Data quality improvement
- Sophisticated data capabilities
- Cost efficiency

Broadly speaking, three strategic approaches can be taken:

Table 1: Strategic Approach and Business Impact Implications

Approach	Impact	Time To Market
In-house	Build using internal resources/expertise	Expensive, slowest
Hybrid	Use data tools + your own resources	Complex
Managed Service	<New> Turnkey service provider	Fastest

And there are two parts to the problem that needs to be considered when evaluating the best approach:

- Developing data pipelines with low TCO
- Implementing a platform to allow data pipelines to be built efficiently

Table 2 identifies the key considerations that need to be evaluated for any solution to have a positive impact on the TCO. Note that it is critical to integrate any solution into the existing

infrastructure and flows. At the core, the solution chosen needs to make new data feeds 'plug compatible' for ingestion.

Table 2: Key Considerations to Identify the Best Strategic Approach

Stage	Key Considerations
Scouting & Trials	Catalog of datasets to find and explore
	Ability to trial "real" data quickly
Build	Portfolio of pre-built third-party pipelines
	Connectors to third-party applications
	Ability to run supplier loading programs
	Event-based pipelines, not just regular 'batch'
	Broad range of input data formats supported
	Broad range of destinations supported
	Ability to transform data
	Ability to enrich data
Datasets	Enables capacity cost-efficiently
	Dataset curation
	Create datasets to deploy (effort/time)
Quality	Standard quality metrics for datasets
	Custom quality metrics for datasets
Operations	24/7 Monitoring of data and infrastructure
	Supplier data delivery issues escalation
	Data and infrastructure operational SLAs
Maintenance	Platform maintenance included
	Data changes from supplier included
Security	Certified solution (e.g. SOC II, Infosec)
Migration	Ease at which to configure a new platform to existing systems
	Easy to reconcile new or existing data feeds
Commercial	Avoid upfront investment needed
	TOC % better than the current solution

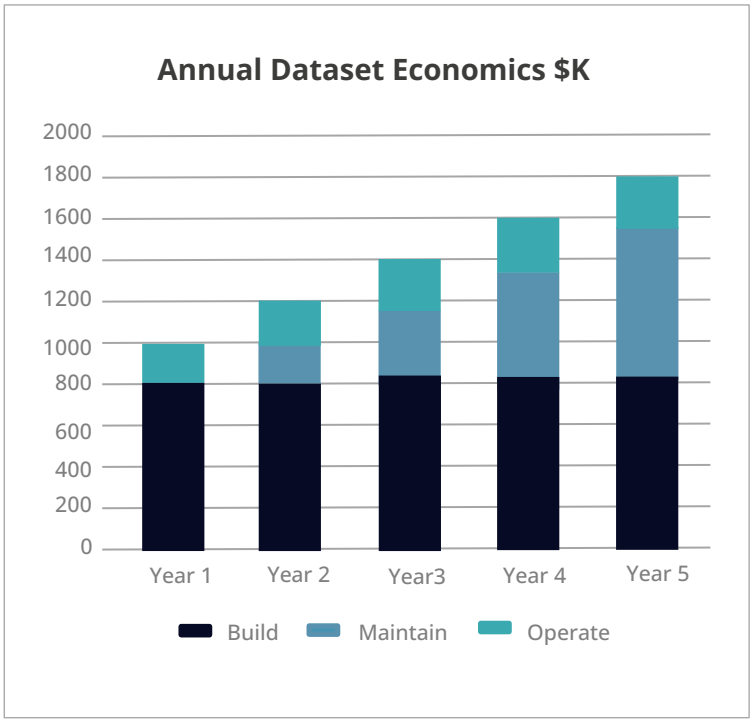
ILLUSTRATIVE TOTAL COST OF OWNERSHIP EXAMPLES

With a clear understanding of all the costs that impact the economics of dataset integration, it is easy to assemble a complete picture of the financial requirements throughout the lifecycle of an external data pipeline. We've created an example to illustrate the importance and impact of identifying these costs:

Data engineering cost/year:	$250,000	Data operations cost /year:	15% (1.1 days/dataset/year)
Target new datasets/year:	100	Dataset productivity:	2.5 datasets/month, 30 datasets/year
Maintenance on build/year:	20% (1.5 days/dataset/year)		

Assuming all other costs remain equal, the maintenance and operations cost to build 100 pipelines begin to dramatically increase each year.

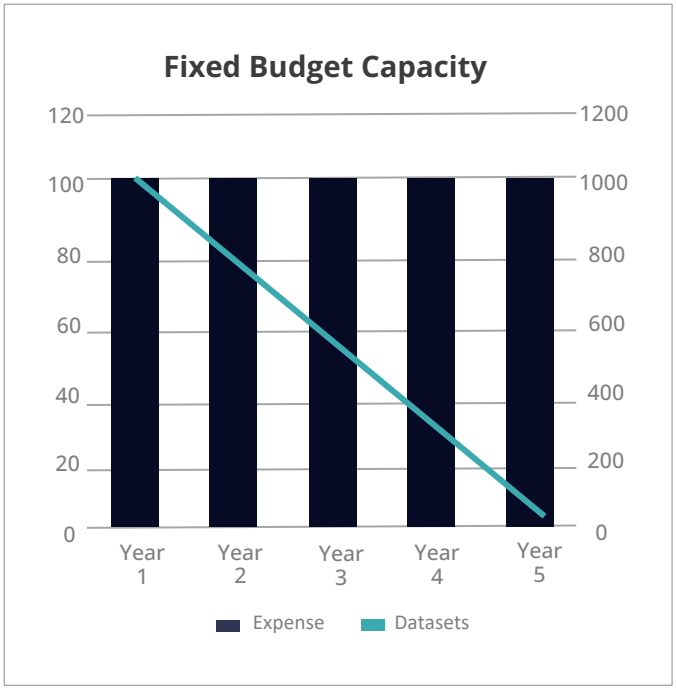
As illustrated in the adjacent bar chart, by the second year the same data engineering team needs to expand to operate the growing number of datasets and maintain the existing ones. This trend continues and by year 5, assuming that you've added no new headcount to your team at the same cost, the expense of maintaining, operating, and building 100 datasets a year is now nearly twice the year 1 cost.



This example excludes the costs of supporting the underlying infrastructure and the maintenance associated with the software platform. In addition, the costs to build the platform have been kept consistent in these models to highlight how significant the overlooked maintenance and operations costs can be. As a result, the costs in this example are extremely conservative.

In a second example, if we assume that the data engineering budget is fixed at $1M, then we can see efficiency drop significantly each year with the increased burden of dataset operations and maintenance.

The bar graph shows that the capacity available to build new datasets drops to effectively zero by year 5. If infrastructure and platform costs are factored in, which is typically the case, only the first three years end up being productive.



These illustrative models are conservative and simplistic, leaving out assumptions regarding salary increases, bonuses, cost of living increases, and attrition. Yet, even so, one can see how important it is to fully understand the economics of external datasets and the implications. While these figures aren't meant to be a universal representation of all organizations, they are a useful starting point to help decision-makers improve how to budget for external data pipelines.

AN ALTERNATE APPROACH - DRIVE COST SAVINGS IN YOUR EXTERNAL DATA STRATEGY

These cost projections can be daunting, especially as organizations rapidly expand their portfolios to integrate new ESG (environmental, social, and governance) data, mitigate supply chain disruption, and meet the challenge of increased competition.

Eventually, scaling in-house resources becomes too expensive, and at what risk? For most organizations, scaling data operations means either growing your team with new hires or outsourcing pieces of the process, resulting in a fractured and inefficient data supply chain. Many enterprises seek a service provider to supplement their data engineering staff to scale external data integration while offsetting the costs of building, operating, and maintaining their datasets.

	CRUX	Relative Cost	Stakeholder	Typical Organization Resources	Total Cost Over Time
ACQUISITION	Data Scouting & Discover	$	CDO, Quantitative Analyst	Quantitative Analyst, Scientist, Data Scout	5-10%
ACQUISITION	Data Trialing & Evaluation	$$	CDO, Quantitative Analyst	Quantitative Analyst, Scientist	+15%
INGEST, OPERATIONS, MAINTENANCE	Platform Development & Maintenance	$$ Allocation[1]	CTO, CDO	Application Developer	15-20%
INGEST, OPERATIONS, MAINTENANCE	Pipeline Build, Onboard & Ingestion	$$$	CTO, CDO	Data Engineer	20-25%
INGEST, OPERATIONS, MAINTENANCE	Operations (Infrastructure & Data)	$ Allocation[2]	CTO, Data Ops	Data Ops, Infrastructure	+10%
INGEST, OPERATIONS, MAINTENANCE	Data Maintenance & Enhancements	$$	CTO or Supplier	Data Engineer	+20%

1. Allocated cost if you have previously invested in a platform for data management. Typically, these large one off and ongoing maintenance costs, are allocated on a per dataset basis
2. Many organizations tax/allocate the cost of physical infrastructure and operations across the business

At Crux we believe data is the lifeblood of modern enterprises, fueling an enterprise's decision-making process by exposing blind spots, anticipating trends, and mitigating risk while gaining a competitive advantage. Enterprise decision-makers need more and better data, faster. Whether the data is already on board the Crux platform, purchased from other external sources, or scraped from public websites we can help you get access to that data quickly. Our mission is to help enterprises offset their external data lifecycle costs by delivering services to support their data engineering and operations teams. We can help you onboard new data sources fast, with higher data quality, and at a fraction of the overhead needed to build new data pipeline lifecycles.

"We've been using Crux for almost two years now and are happy with the company's ability to meet our high standards. We consume a large number of datasets from a wide range of sources through Crux, which simplifies our ingestion and operational burden dramatically, allowing us to focus our team on higher-value and more differentiated activities. We are impressed by Crux's customer focus, service level, and data inventory, as well as acceleration we see in dataset onboarding."

— One of the World's Largest Investment Banks

THE VALUE OF TCO ASSESSMENT

If your enterprise relies on external data from various data sources and you anticipate your needs growing, you need an affordable solution that will allow you to scale data while reducing your overall integration costs and accelerating access to new data sources to position your organization for greater agility and competitive advantage.

You can hire the best people, but if you can't arm them with the latest data sources they need when they need it, they won't be able to perform to your enterprises' potential. At Crux, we want to reduce the cost of getting access to new data so that your team can focus on what they do best, driving business value.

Want help calculating the TOC of your external data?

To get started, schedule a 45-minute meeting to discuss calculating the TCO of your external data and we can share ways to reduce your data pipeline costs while mitigating risk to your growth strategy.

Schedule a meeting

About Crux

Crux is the leading provider of external data integration solutions for organizations looking to integrate, support, and transform external datasets. The company's cloud-native data integration technology accelerates the ingestion, preparation, observability, and ongoing delivery of any external dataset into any destination, including Amazon, Google Cloud Platform, Microsoft Azure, and Snowflake. Using Crux, clients receive quality data in the right place, in the right format, and at the right time. The company combines best-of-category technology, world-class services, and deep-rooted expertise gained from over 25,000+ successful onboarded datasets. Our objective is to support the full scope of a customer's external data pipeline lifecycle in a cost-effective manner.