aws | AWS Data Exchange

A Practical Guide to Using Third-Party Data in the Cloud

Today's top data and IT leaders on why third-party data, data-sharing programs, and data valuation are the new best practices in cloud computing

Contents

Introduction 3

Leverage a Data-Sharing Culture to Make Smarter Decisions.... 4

- Data sharing in the age of digital transformation 5
- Three models of data sharing 6
- How to create a data sharing culture 8
- The data sharing mindset 9
- The data sharing information architecture 10
- Summary 11

Organizations That Leverage External Data Will Outperform By Double Digits 12

- Define earlier signals 14
- Re-envision sharing data across your ecosystem 17
- Why 2021 will be external data's break out year 17
- Summary 18

Speed of Data 19

- Summary 25

What's Your Organization's Data Innovation Strategy? 26

- What's your innovation strategy? 27
- Does your shadow data hide or reveal value? 28
- Missing opportunities for innovation 28
- Innovation and the last mile of analytics 28
- How can you get started on supporting your innovation strategy with technology? 29
- Technology to support the innovation strategy 30
- Summary 31

Expand the Power of Analytics with Third-Party Data 32

- Exploring value areas 34
- Challenges and concerns 34
- Summary 35

What Is Data Analysis and What Is Its Importance for Companies? 36

- How and where to use different types of data analysis 38
- The importance of machine learning for data analysis 40
- Machine learning use cases in data analysis 40
- Summary 41

The External Data Imperative and Valuation Approaches 42

- The Wide World of External Data 44
- Determining the Value of External Data 46
- Thriving in the The Data Economy 49
- Summary 50

Conclusion 51

Introduction

Digital transformation has made organizations more agile than ever, but increasingly companies are turning to third party datasets to create new, unique advantages over their peers. Gartner notes: "Organizations that offer users access to a curated catalog of internally and externally prepared data will realize 100% more business value from analytics investments than those that do not."[1] But to do so, organizations must first overcome their innate suspicion of sharing data and identify and locate trusted data sources.

However, the challenge of finding a trustworthy source for third party data is just one part of the equation. That data also needs to be widely accessible across the enterprise to build valued, deeper insights users desire. Businesses also need to deliver on *critical business outcomes*. That nuance has created a fundamental shift in how business decision makers view IT investments and capabilities and, conversely, can inform the role IT plays in supporting business decisions.

It is both a technical and a cultural challenge – but solving for it is crucial.

> *Markets move faster than companies, and organizations must adapt to survive the technological flip... Enterprises could try to understand reporting and data, but they may forget about third-party data sources. These sources are powerful opportunities to gather data about the human as a social animal; the data that we produce every day.*

Jennifer Stirrup,
Founder & CEO of Data Relish

Today, companies that establish capabilities to share both internal and external data across the value chain can move more quickly – move beyond forecasting to 'nowcasting'. Data marketplaces such as AWS Data Exchange can help build better decision making, drive efficiency and productivity, and create opportunity. In the commentaries that follow, we'll explore a range of approaches to building a third-party data strategy and overcoming common implementation challenges:

- **Modernize data infrastructure** – Developing a data ecosystem that connects analytics, business needs, IT and business teams, and legal departments while balancing speed and accuracy to enable better decision making
- **Customer insights** – Understanding and supporting your most valuable customers and recognizing (even forecasting) changing consumer demands to initiate appropriate pivots in your supply chain
- **Security** – Managing permissions and control and learning how the cloud could reduce the risk of shadow IT
- **Efficient data procurement and entitlement** – Building an effective third-party data strategy, removing friction associated with finding, licensing, and delivering data sets

As the C-suite becomes more involved in decision making over data modernization roadmap, it becomes critical to find ways to bridge the gap between technical capabilities and business outcomes. That is of crucial importance in a world where data is relied upon to predict future trends. Let's take a closer look at how organizations can leverage a wide range of data sources to create value, drive innovation, and transform their businesses.

[1]Laurence Goasduff. Gartner, "Data Sharing Is a Business Necessity to Accelerate Digital Business," May 20, 2021. https://www.gartner.com/smarterwithgartner/data-sharing-is-a-business-necessity-to-accelerate-digital-business/

Leverage a Data-Sharing Culture to Make Smarter Decisions

Dave Mariani – Founder and Chief Technology Officer, AtScale

Founded in 2013, AtScale is the Business Insight engine for the cloud-first enterprise, empowering analysts and data scientists to simplify and accelerate business insight programs globally. Enterprises like JPMC, Visa, Bloomberg, UnitedHealthcare, Cigna, Kohls, Home Depot, Wayfair, and Toyota all use AtScale to make data "analytics ready" for their employees and partners. Prior to AtScale, Dave ran data and analytics for Yahoo! where he pioneered the development of Hadoop for analytics and created the world's largest multi-dimensional analytics platform.

Data sharing in the age of digital transformation

Business runs at the speed of data and today analytics often make the difference between business success and failure. During the COVID-19 pandemic, we've seen this phenomenon work at warp speed, especially in retail. What differentiates retail winners and losers is clear: companies who have invested in a self-service, trust-based, data-driven culture outperform their cohorts.

As we speed down the digital transformation highway, the culture of data sharing is becoming the new differentiator. Gone are the days when IT can spoon carefully curated data to the business or where data is held close to the chest. According to Gartner, by 2023, organizations that promote data sharing will outperform their peers on most business value metrics.[2]

In this paper, we'll examine the ways in which organizations can create a data-sharing culture both within the organization and with external stakeholders. We'll also look at the technical requirements to promote and operationalize data sharing while encouraging technical stakeholders to align their interests with business outcomes.

[2] Goasduff. Gartner, "Data Sharing Is a Business Necessity."

Three models of data sharing

Data sharing is more than just sharing files, reports, and dashboards with your peers. There are different models of data sharing each with unique characteristics.

1 First-party data sharing

First-party data is the information your organization collects directly from your business activities and customers.

First-party data sharing is probably what comes to mind to most when you hear the term "data sharing." First-party data is internal, proprietary data not meant for external consumption.

At first glance, it should seem obvious that first-party data is shared freely within an organization. However, internal data tends to be siloed and stored in multiple, business process-specific data stores. For example, financial data may be locked away in an enterprise resource planning (ERP) system, sales data in a customer relationship management (CRM) system, and employee data in an HR platform.

In addition to the challenge of siloed data, many enterprises have empowered individual business units to drive the selection of their own tools and technology. This autonomy may accelerate progress in the early days. However, as the business scales, it often buckles under the growing deluge of data and lack of engineering expertise to manage the volume. This fragmentation creates barriers to sharing data and has a direct, negative impact on business performance, making organizations less agile.

By contrast, adopting a "data-as-a-service" strategy promotes autonomy and frees business users to focus on improving business performance rather than wrangling data.

Customer use case:

A home furnishings retailer migrated their analytics infrastructure to the cloud and created a self-serve data platform for more than 2,000 business users. As a result, it was able to scale its data infrastructure in response to changing market conditions due to COVID-19 and increase sales by almost 70 percent.

2 Second-Party Data Sharing

Second-party data is first-party data that two or more business partners decide to share on a "private" basis for mutual benefit.

In a modern enterprise, internal data sharing is table stakes – a given. That's especially true in retail where shared data among business partners has immediate and powerful impact. By sharing inventory data with their suppliers and vendors, retailers can enable their business partners to effectively manage inventory on their behalf. The two parties' incentives are aligned. The retailer wants shelves stocked with goods that will sell; the supplier wants to push as much volume as possible.

The marketing automation software industry is another great example of the power of second-party data sharing. Through e-commerce habits, consumers leave digital breadcrumbs across the WebSphere. This data is captured by dozens of third parties and stored in thousands of proprietary databases. Marketing software vendors often share data among themselves so their joint customers can construct consumer profiles and improve targeting and sales.

By sharing data with partners, organizations can scale their reach beyond the corporate walls to better serve customers and drive operational efficiency.

Customer use case:

A Fortune 500 home improvement retailer shared their inventory data with their 9,000 suppliers. By providing a real-time query interface into the retailer's inventory at the store and SKU level, the suppliers stocked the right products in the right locations and expanded sales by 20 percent during the COVID-19 pandemic.

3 Third-Party Data Sharing

Third-party data is information that is collected from a variety of websites and platforms and is aggregated by an outside entity.

Leaders in data and analytics can enhance their first-party data by integrating third-party datasets to create unique, competitive advantages. Gartner highlights this trend, predicting that by 2023, 85 percent of data sharing strategies that include external data sources will drive revenue generating digital business outcomes.[3]

During the time of COVID-19, we've seen a wide range of third-party data consumption via data marketplaces like AWS Data Exchange, making a difference in gauging demand and predicting sales. Data from retail foot traffic (from Safegraph) or Weather-Driven Demand for Ice Cream Novelties (from Planalytics), and other marketplaces makes it easy to browse available datasets like you would browse products on Amazon.com, with free and "try-before-you-buy" options available.

When organizations combine these "data enhancements" with their own data, they can use the additional data sources to train complex models using machine learning (ML) and artificial intelligence (AI) to create new, ever-more-accurate business insights.

Customer use case:

A Fortune 100 food processor combined foot-traffic data of third-party stores with their inventory and sales data. By forecasting COVID-19 conditions by market, they accelerated the switch from wholesale to retail by market, increasing sales and improving profits by 10 percent.

[3] Goasduff. Gartner, "Data Sharing Is a Business Necessity."

How to create a data sharing culture

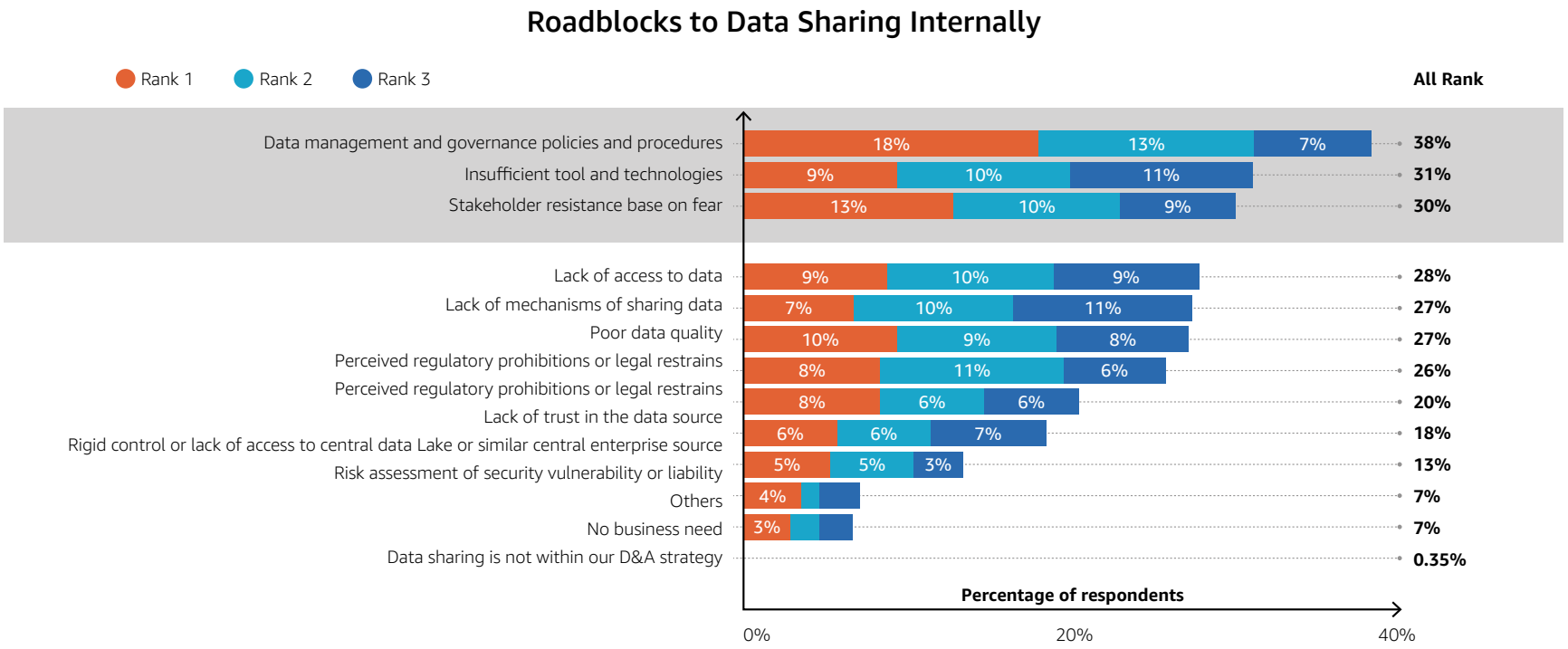
Several potential barriers may impact data sharing, according to Gartner.

These barriers fall into 2 main categories:

1— Data-sharing mindset

2— Data-sharing information infrastructure

Gartner Data Sharing Survey



n= 285 respondents who are sharing data internally (Q12), excluding "unsure"

Q: What do you see as the top challenges to sharing data internally at your organization? Please rank up to 3 challenges in order of importance.
Source: Gartner's Fifth Annual CDO Survey (2019)
Note: Values less than 3% not shown.

The data sharing mindset

There's a natural tendency for most organizations to be fearful and suspicious of sharing data. IT has traditionally protected access to data for security, privacy, and regulatory reasons. However, sharing data, whether it's interdepartmental, with strategic partners, or with data marketplaces, is becoming a key capability for creating enterprise value.

Breaking from that tendency to overprotect access, organizations must recast their data management practice from an IT function into a business function.

By making data management a key business capability, enterprises can transition into what Gartner calls a "Must Share Data Unless"3 data sharing model.[4]

In other words, organizations should assume exposure, either internally or externally, so proper safeguards and processes are put in place at data collection time. By creating a default, "share everything" mindset, the business can decide whether, what, and when to share data to create a data agile enterprise.[5]

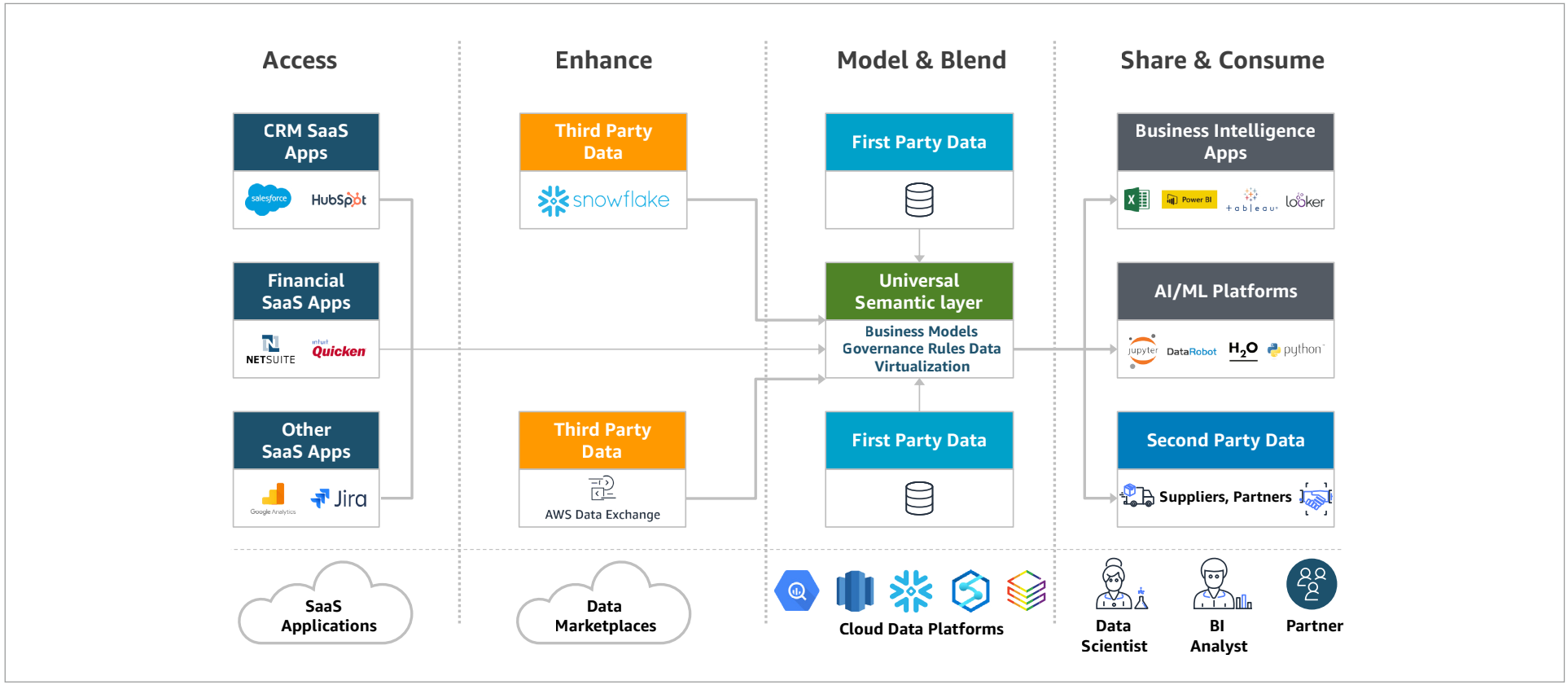
[4] Lydia Clougherty Jones. Gartner, "Flip 'Don't Share Data' Mantras—Introducing Gartner's 'Must Share Data Unless' Data Sharing Model," September 1, 2020. ID: G00727589.

[5] Goasduff. Gartner, "Data Sharing Is a Business Necessity."

The data sharing information architecture

The more obvious barrier to data sharing is a lack of technical infrastructure suitable to do so successfully. The critical component is to begin by abstracting the physical location and format of enterprise data so that internal and external analytical engines can access and blend data that spans multiple business processes. Data integration strategies like virtualization or creating a centralized data warehouse can be useful tools. Still, it's essential to create a business-friendly semantic layer with integrated security and governance.

The diagram below illustrates how a universal semantic layer can be a fundamental building block to enabling all three methods of data sharing.



Equally important is driving data literacy. Without a guide that maps the data's location and context, it's tough to share data with anyone. Tools like enterprise data catalogs can help drive data literacy by documenting data sets, standardizing on business terms, and creating a centralized glossary.

Taken together, a universal semantic layer backed by an enterprise data catalog can serve as the data sharing hub for the enterprise. With this single data control plane, data sharing with the requisite controls creates the confidence and trust needed to drive a data sharing culture.

Summary

As you can see, data sharing can take many forms and the benefits are transformational. Whatever mode of data sharing makes sense for your organization, it's imperative to create the right culture and infrastructure to support sharing. By treating data as a core asset and competitive differentiator, enterprises will thrive in difficult business environments like those experienced during the COVID-19 pandemic.

Organizations That Leverage External Data Will Outperform By Double Digits

Cindi Howson – Chief Data Strategy Officer, ThoughtSpot and Host, The Data Chief Podcast

Cindi Howson is the Chief Data Strategy Officer at ThoughtSpot and host of The Data Chief podcast. Cindi is an analytics and BI thought leader and expert with a flair for bridging business needs with technology. As Chief Data Strategy Officer at ThoughtSpot, she advises top clients on data strategy and best practices to becoming data-driven, influences ThoughtSpot's product strategy, and is the host of The Data Chief podcast.

Cindi was previously a Gartner research Vice President, as the lead author for the data and analytics maturity model and analytics and BI Magic Quadrant, and a popular keynote speaker. She introduced new research in data and AI for good, NLP/BI Search, and augmented analytics and brought both the BI bake offs and innovation panels to Gartner globally. She's rated a top 12 influencer in big data and analytics by Onalytca, Solutions Review, and Humans of Data.

Prior to joining Gartner, she was founder of BI Scorecard, a resource for in-depth product reviews based on exclusive hands-on testing, contributor to Information Week, and the author of several books including: Successful Business Intelligence: Unlock the Value of BI & Big Data and SAP BusinessObjects BI 4.0: The Complete Reference. She served as The Data Warehousing Institute (TDWI) faculty member for more than a decade. Prior to founding BI Scorecard, Howson was a manager at Deloitte & Touche and a global BI standards leader for Dow Chemical. She has an MBA from Rice University.

Data has never been more valuable. In 2020, the most-data driven companies outperformed organizations who were less effective at leveraging data. In 2021, those who capitalize on external data for more timely signals and comprehensive insights will further accelerate their competitive position.

Companies have often leveraged external data like weather, jobs reports, and social media to better understand customer behavior and supply chain planning. And yet, they did so inefficiently, with multiple departments and lines of business procuring external data sets, wasting millions of dollars in repeat purchases. Rarely did the Chief Data Officer have insight into all the external data purchased by each functional area.

Further, ingesting, preparing, and analyzing external data was often a laborious, manual process, made even more so because it was siloed and replicated throughout an organization.

The cloud changes the game. Not just in terms of data available, but also how that data can be shared. In a digital world, newer data sources like human mobility, credit card transactions, hiring trends, and air quality allow for far greater near-term analysis, or what I call "nowcasting."

Let's compare what that may look like. With traditional external data types, a supply chain manager may use internal data such as historical sales and orders to forecast future demand. Click stream analysis on the website may provide signals to an uptick or downturn.

Organizations have often used external data such as monthly government job reports to see if the economy is heating up or slowing down, factoring that into forecast. Nielsen ratings and reports revealed more trends within specific industries and brands. While this data is rich, it's often compiled weeks, sometimes months after an event has occurred.

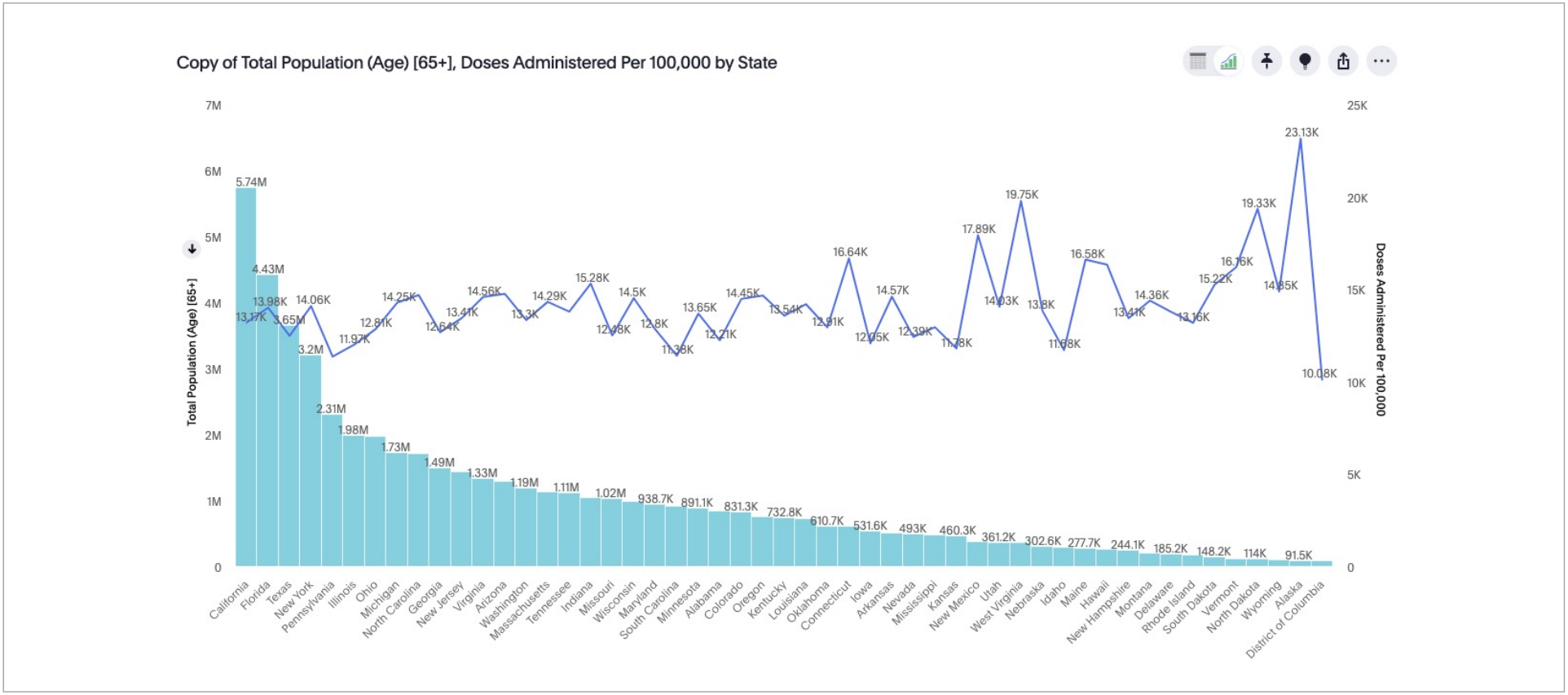
Improving Forecasting with External Data: Slow Moving

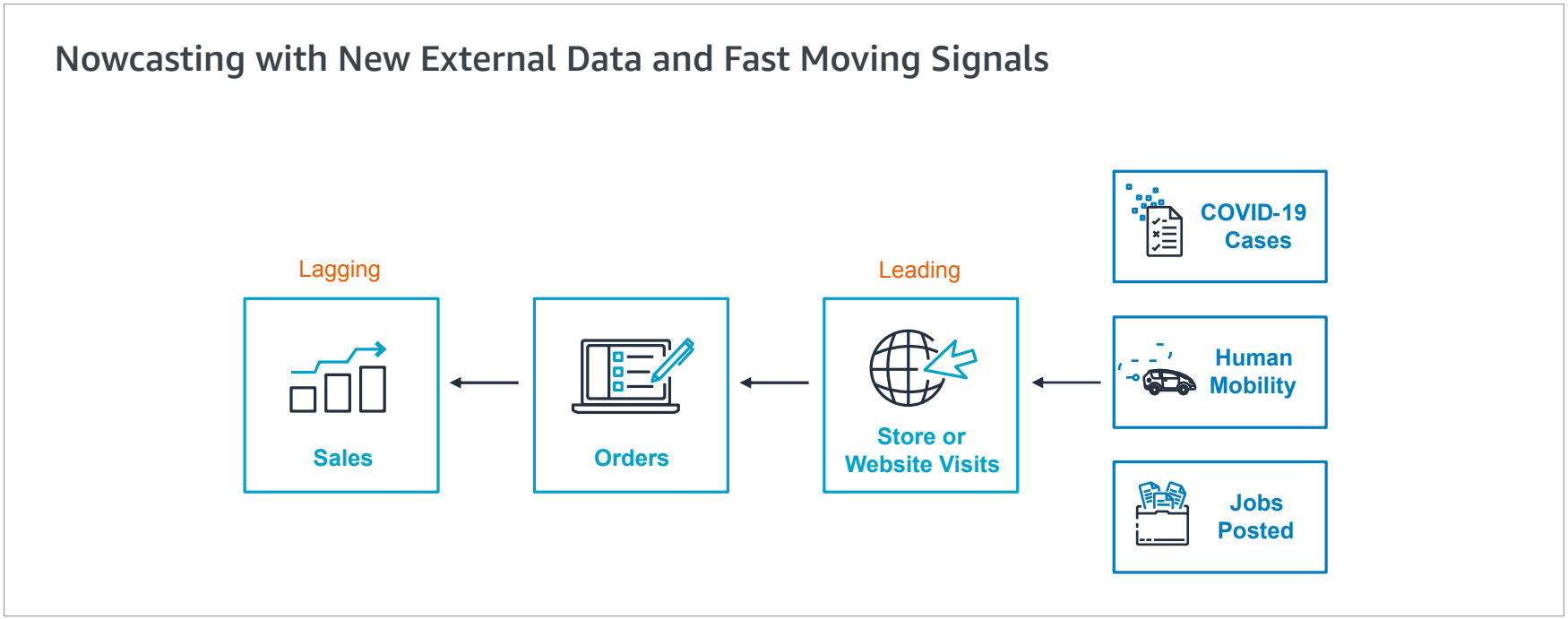
Define earlier signals

In our digital world, waiting weeks or months isn't tenable. Instead, leaders need to ask themselves what data sources are faster moving that might give you an earlier signal?

These faster moving data sets take us from forecasting to nowcasting. Human mobility has become a popular one in the last year, showing where people are walking, driving, generally out and about. Those month's old job reports can be replaced with real time job postings scraped from sites such as Indeed, LinkedIn, Monster, to show which sectors and cities are hiring faster. Lastly, COVID case counts are one of the most popular new data sets that teams are leveraging, both for demand planning but also employee safety.

Starschema (the company, not the data model) has been a key provider of this data, compiled from a range of reporting agencies. As Raj Gupta, senior IT director at Flex, described, "We really need to understand where our employees are and how we can potentially get them back safely. **Some of that data is internal, but a lot of it is external around some of the COVID-19 hotspots."**



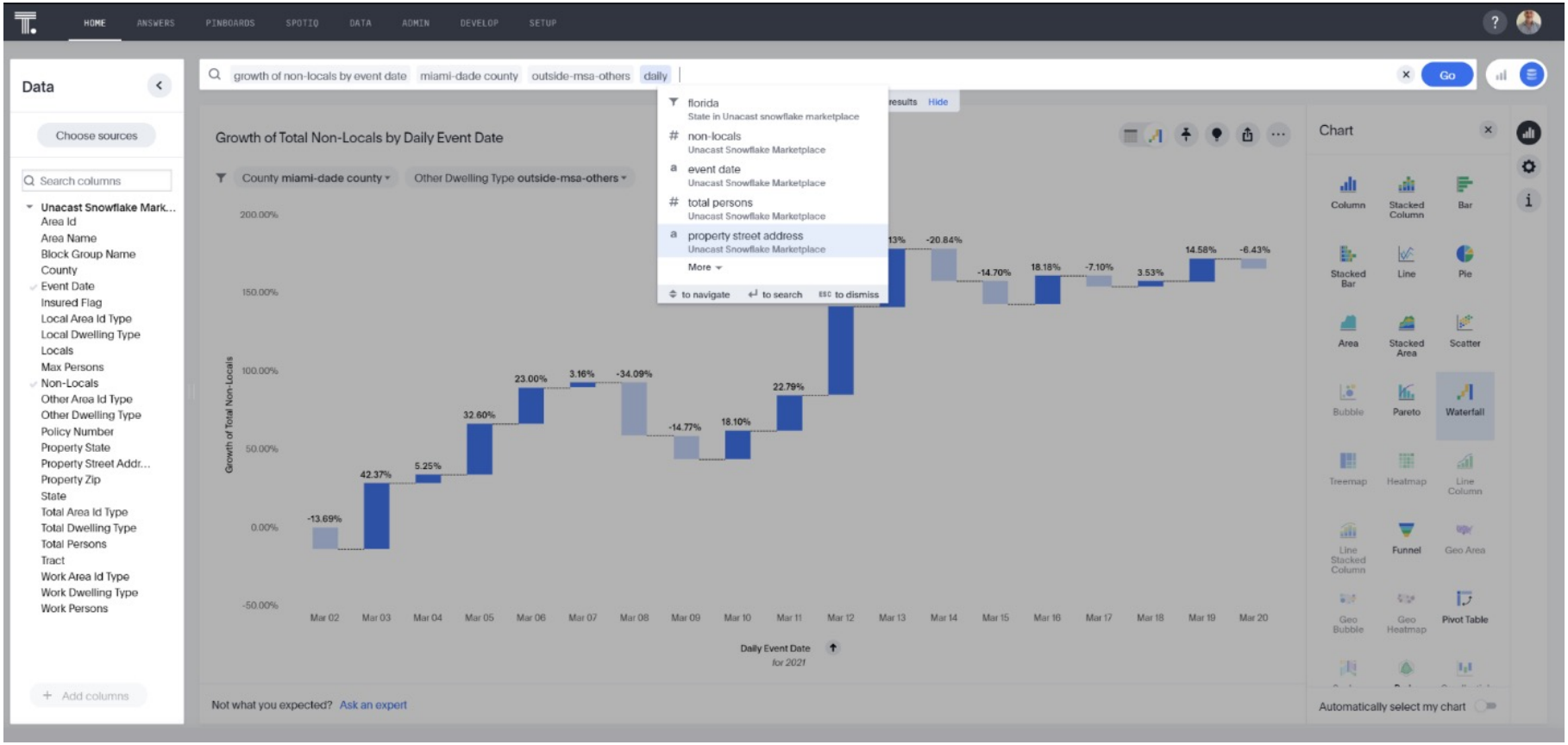


Likewise, in 2020, Hershey was better able to pivot their supply chain during lock downs where different chocolate was consumed at home, leading to a 5.5% increase in same-store sales. Similar, Bacardi changed its product mix, informed by faster-moving external data. Retail giant Canadian Tire used foot traffic, weather, traffic patterns, and shifts in demand for bicycles and outdoor furniture, contributing to a 19.1-percent increase in year-over-year retail sales through Q3. The Hartford uses hundreds of external data sources in assessing risk, underwriting, and claims processing.

As an example, an insurer could combine population counts, with human mobility, to assess the degree that certain properties exceed insured capacity limits. The following chart shows an example using U.S. Census data with Unacast human mobility data to see spikes in Miami, Florida during spring break 2021, leveraging ThoughtSpot.

Companies leveraging external data to find the early signals are outperforming competitors. Not just slightly, but in substantial, meaningful, long term ways.

These newer and external data sets are especially important when the past can no longer predict the future. As Tom Davenport, Professor of Babson College, and author of The *AI Advantage*, wrote, "Trying to model low-probability, highly disruptive events will require an increase in the amount of external data used to better account for how the world is changing. The right external data could provide an earlier warning signal than what can be provided by internal data."[6]



Source: ThoughtSpot

[6] Jeffrey D. Camm and Thomas H. Davenport. MIT Sloan Management Review, "Data Science, Quarantined," July 15, 2020. https://sloanreview.mit.edu/article/data-science-quarantined/

Re-envision sharing data across your ecosystem

Purchasing external data is one way to get started, but data and analytics leaders should think about the potential for data sharing across the full value chain. This may include setting up data sharing ecosystems within the Snowflake Data Cloud, for example, or using technologies like Blockchain. DataBricks also recently released its new Delta Sharing as part of its Delta Lake.

For example, Daimler is testing analytics on shared data with suppliers, without ever having to transfer that data or reveal details via Ocean Protocol's data exchange platform.

In battling the pandemic, competitors in pharmaceuticals turned to collaborators and allowed each other to train their machine learning algorithms on one another's data sets. MELLODDY is a data exchange running in AWS and leveraging blockchain technology to promote drug discovery. In this way, machine learning models benefit from broader data without the privacy restrictions that often prohibit sharing.

Car manufacturers are beginning to share data gathered from sensors in the car to warn about road hazards with other car manufacturers. Imagine you are driving a Ford and the anti-braking system kicks in because a tree blocked the road. A driver of the BMW a few miles behind you can be warned.

Hulu is using Snowflake on AWS and its data sharing capabilities with media partners to improve personalization, while maintaining privacy. Jeff Nemecek, Director of Architecture and Quality, describes this as "the end of ftp as we know it."

Why 2021 will be external data's break out year

Companies have long leveraged external data in their analyses but that data may have been slow moving and hard to analyze or combine with internal data. 2021 is shaping up to be a breakout year for data marketplaces, private data exchanges, and increased data sharing.

Some signs:

- Eagle Alpha, a pioneer in alternative, big data sources counted 1400 data sets at year end, estimated to increase to 2000 by end of 2021.
- A recent IDC report estimates more than half the data companies need for analytics is external.
- The rise of the Snowflake Data Cloud and its Data Marketplace (for consuming external data) and Data Exchange (for sharing) allow organizations to share data with no data movement, a game changer in all those messy, duplicate data movements.
- A steady rise of third party data available via AWS Data Exchange since its release in 2019, and with automatic update notification, an ability for CDOs to get better visibility into all external data purchases under one umbrella.
- Increased adoption of blockchain supports data sharing without data movement by which calculations can be run on shared data.

And yet, while there is a range of powerful third party data available to business leaders, value is not guaranteed. The number of data sources for human mobility alone is overwhelming.

Further, just as organizations often fail to align to business value in leveraging internal data, these same struggles can be magnified when it's external data you must purchase. One best practice: start with the business outcome or metric you are trying to improve, then work backward to identifying the data that would support that use case. Is it knowing the customer better? Predicting demand?

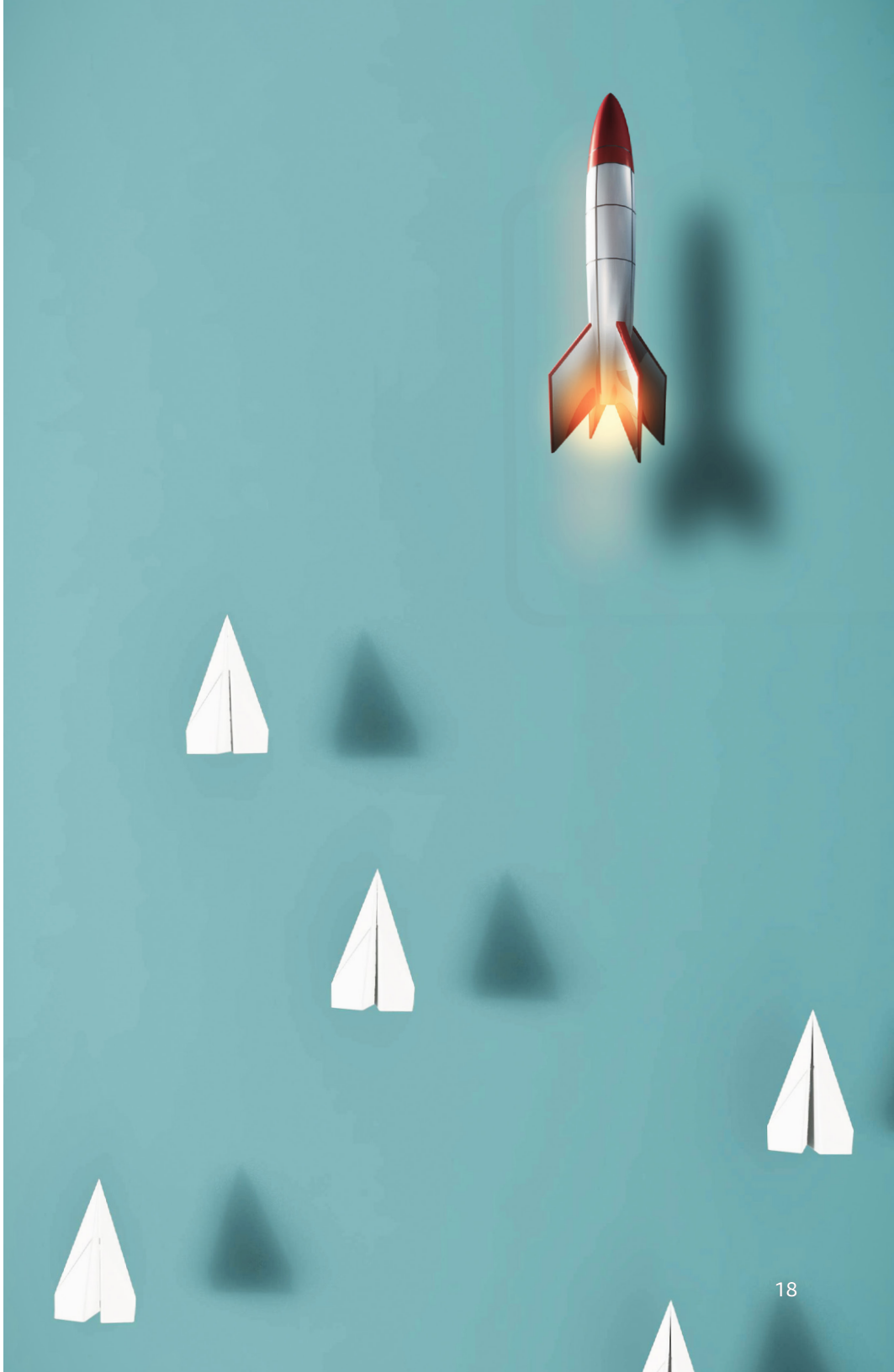
Summary

1. Use your imagination to think differently about leading indicators and new data sources that can provide better signals for customer behavior, supply chain, and rapidly changing business questions.

2. Consolidate external data purchases under the CDO to minimize costs and maximize leverage across the organization.

3. Be realistic about the data preparation required to merge external and internal data, but look for options that include concordance using augmented data preparation and crowd sourcing of master data.

4. As you look to monetize and share data across the whole value chain, get education on the range of options from large data marketplace vendors such as Snowflake and Amazon, but also from pure plays such as Harbr, Dawex, and Ocean Protocol.



Speed of Data

Robert Koch – Enterprise Architect, S&P Global

Robert is the Lead Architect at S&P Global and one of the community leaders of DeafintheCloud.com. He helps drive cloud-based architecture, blogs about migrating to the cloud and use of Lambdas and loves to talk data and event-driven systems. In a recent lightning talk, he gave an overview on Redshift's symbiotic relationship with PostgreSQL.

Robert is actively involved in the development community in Denver, often speaking at Denver Dev Day, a bi-annual mini-conference and at the AWS Denver Meetup. Robert's goal is to help the community understand the benefit of migrating to the cloud and illustrate the advantages of "serverless" applications and databases.

We often refer to data as the "single source of truth." When you engage in truthful endeavors, everyone benefits. From a data management perspective, accuracy is critical but so are speed and value.

I've been thinking about and blogging on the topic of data speed for some time, and I recognize that it's part of a broader discussion. As we know, data at rest has no intrinsic value. Raw, unverified data is not useful until after it's been reviewed (a sometimes slow, human process). Quality control can also slow data down and lessen its value. Even when data validation is completely automated, the process is designed, programmed, and tested by humans. Ultimately, the humans that use the process can't anticipate every potential scenario that requires validation.

Fortunately for the business, end-users don't see much raw data. Most has already been aggregated for better visualization. Aggregation of raw data helps us see trends or detect anomalies at the outset and over time. Visualizing aggregate data on, say, a dashboard also helps to highlight unexpected swings that can't be caught by the naked eye. Still, end users are human. Even with highly-tested AI helping to drive the car, minor data flaws can cause unanticipated accidents. Humans are constantly learning from that process and making our data better and more intelligent for the systems that consume it.

To increase our success in automating validation, I believe we need to create an independent, stand-alone system to digest and make sense of the data. (It's a given that AI is dumb until we make it smarter!) This is the value of exposing the data to APIs which can tap into the data source and independently verify it. When validation and programming are performed separate teams, we can mitigate the risk of human bias. It's the approach I took years ago, before I became a software developer, when I set up validation workflows. For example, if a programming team uses C#, validation should be crafted by another team using a different programming language. Taking that approach helps to avoid human error and remediate potential flaws in the datasets being analyzed.

This process of validation, however, gets us back to the issue of data speed. Validation can slow speeds across the system to an unacceptable rate. What's more, if a discrepancy emerges, it may slow the validation process (and workflow) still more, making the data less valuable with every passing minute. For a team that relies on obtaining timely data, that can be a significant problem.

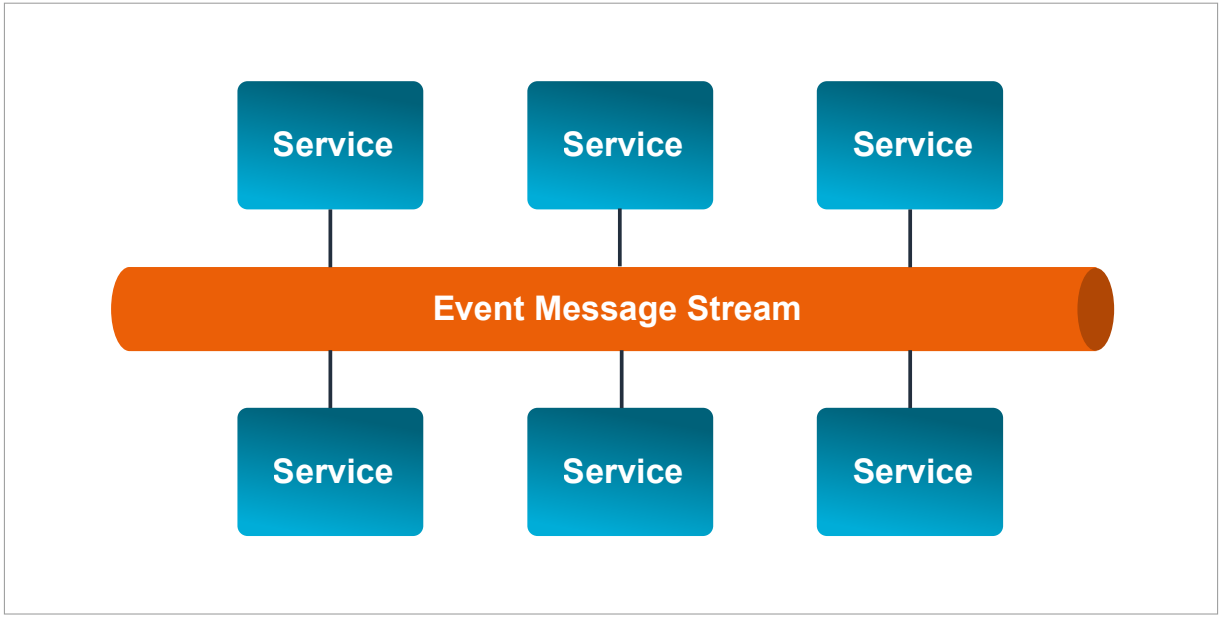
In light of those challenges – quality, validation, or embargoes—how can we speed up the movement of data through the enterprise? The approach I'd like to explore today is migrating to an event-driven system, which enables aggregated data to be validated on the fly.

Using algorithm-driven evaluations on granular data as it flows in (as opposed to a batch-driven system) speeds the validation process. We can do that using an enterprise message bus. If there is a new scenario we want to implement, we can plug the new algorithm into the message bus pipeline without adversely affecting performance or causing bottlenecks.

Cloud technologies make that process easier since we won't have to wait too long to provision services and with data traveling close to the speed of light. We can quickly orchestrate the best tool for the job without the limitations of servers (disk space, network bandwidth, proximity between services, and shared security.) Cloud-based tooling works well provided it's available through a common API or using transitional data points such as comma separated value (CSV) files or through the event bus. CSVs will slow the data velocity but it is a ubiquitous data format and many software tools have the capability to ingest it.

Event-driven architecture differs from batch-oriented workflows or command-and-control systems, which rely on timely data availability or risk manual processes that introduce the potential for human error. Although it's not a new concept, event-driven architecture is becoming a common (if not dominant) approach for enterprise integration patterns. Even IBM is using event messaging streams to supplement their product offerings.

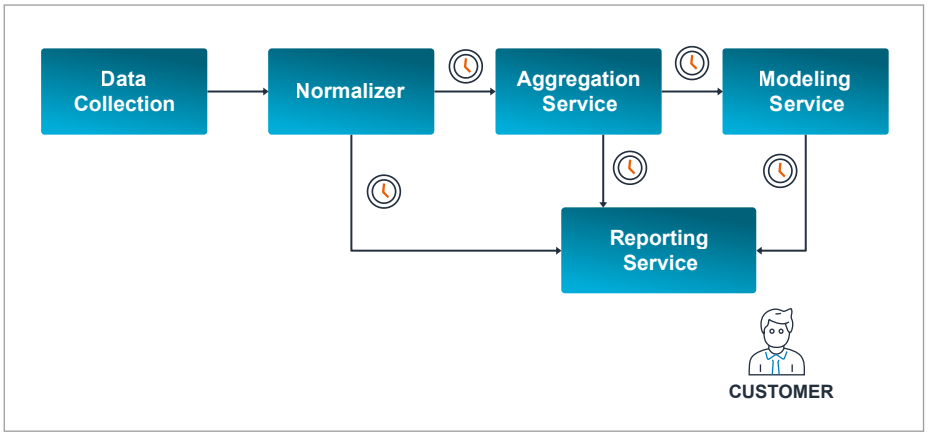
Event-driven architecture is designed to move at the rate we move our data whether rapidly or more deliberately. Another added benefit is the flexibility it offers when we integrate our system incrementally. If we need to add a component or a service, we can plug it into the bus without source mapping or configuration. The onus is then on the domain experts (as it should be) to consume the data in the right way.



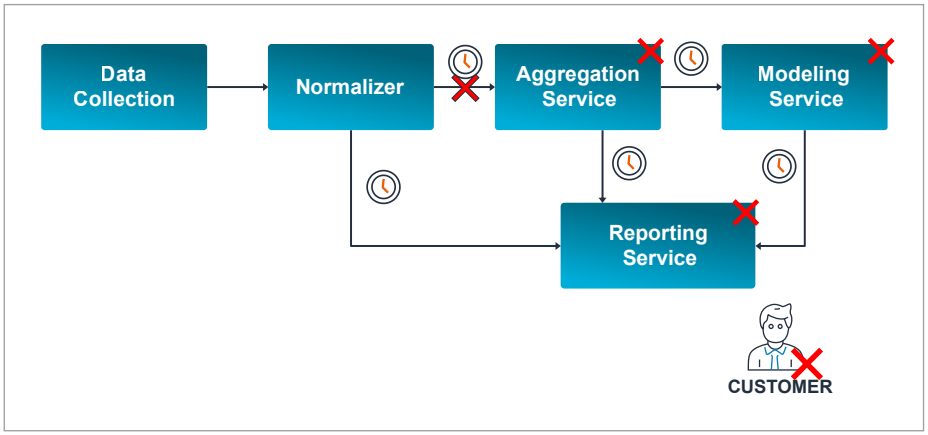
The series of diagrams that follow will hopefully convey the big picture of the benefits of event-driven architecture.

Let's start by looking at the diagram below which illustrates a batch-driven architecture. As is often the case with a legacy system, event-driven technology wasn't readily available and was probably cost-prohibitive. The designers of this approach had to build with the knowledge they had available at the time (on cost, time, skills, circumstances, etc.)

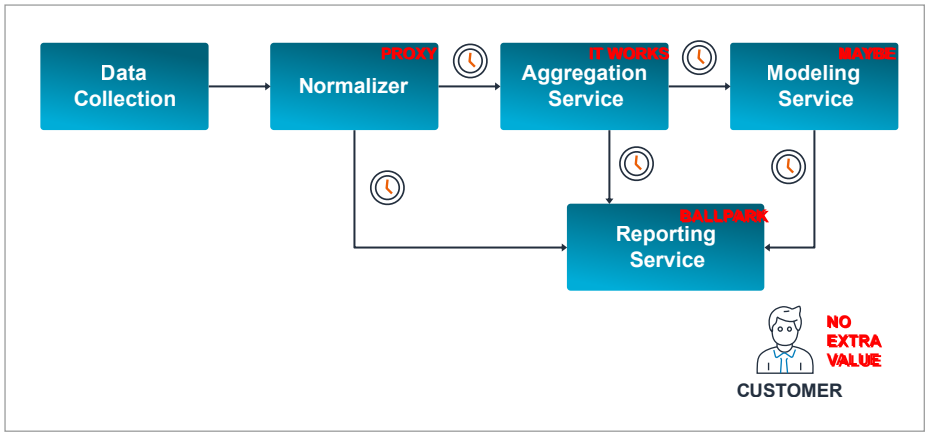
Pay special attention to all the clock icons in the workflow diagram.



This diagram shows that the process is batch driven (denoted by a clock). When any of the subsequent processes run, all we can do is hope that nothing goes wrong to upend the rest of the workflow. Here is an example of a situation where the data might not be present causing issues down the road, requiring immediate manual intervention.

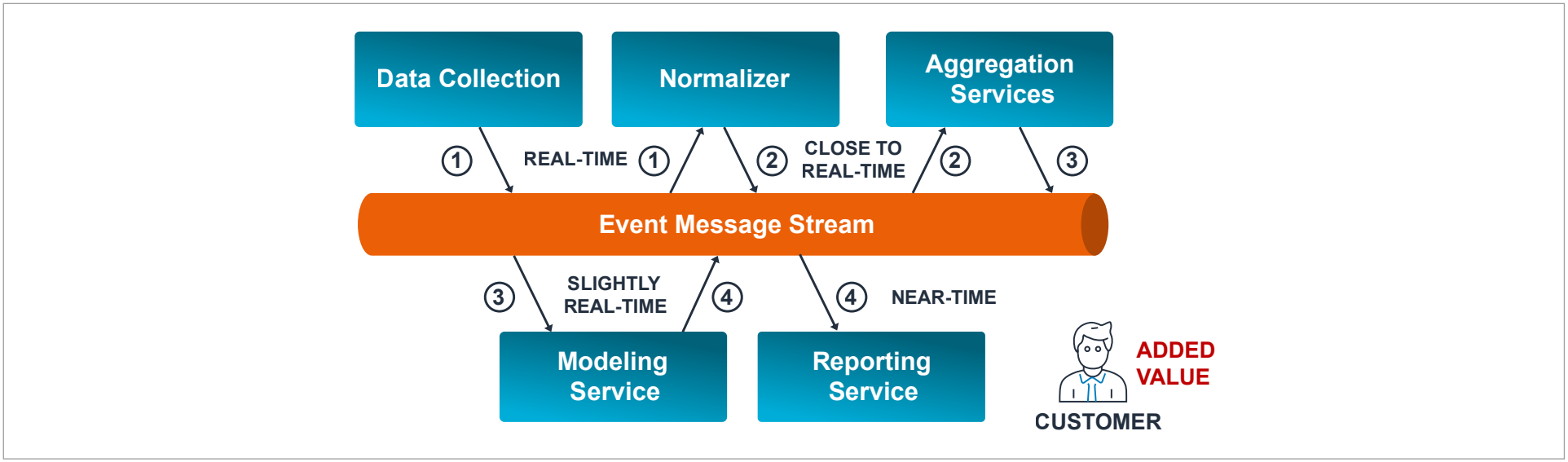


When there is no data available, our options are limited. We could make a workaround: proxy yesterday's data as a placeholder until the correct data comes in. Of course, that approach impacts modeling, and we won't be able to perform accurate forecasting. It would also impact the aggregation service because there's no new data to categorize. Let's see what that approach looks like in the figure below.



In this situation, we can't provide the data customer actionable and timely reporting because the proxied data has no additional value.

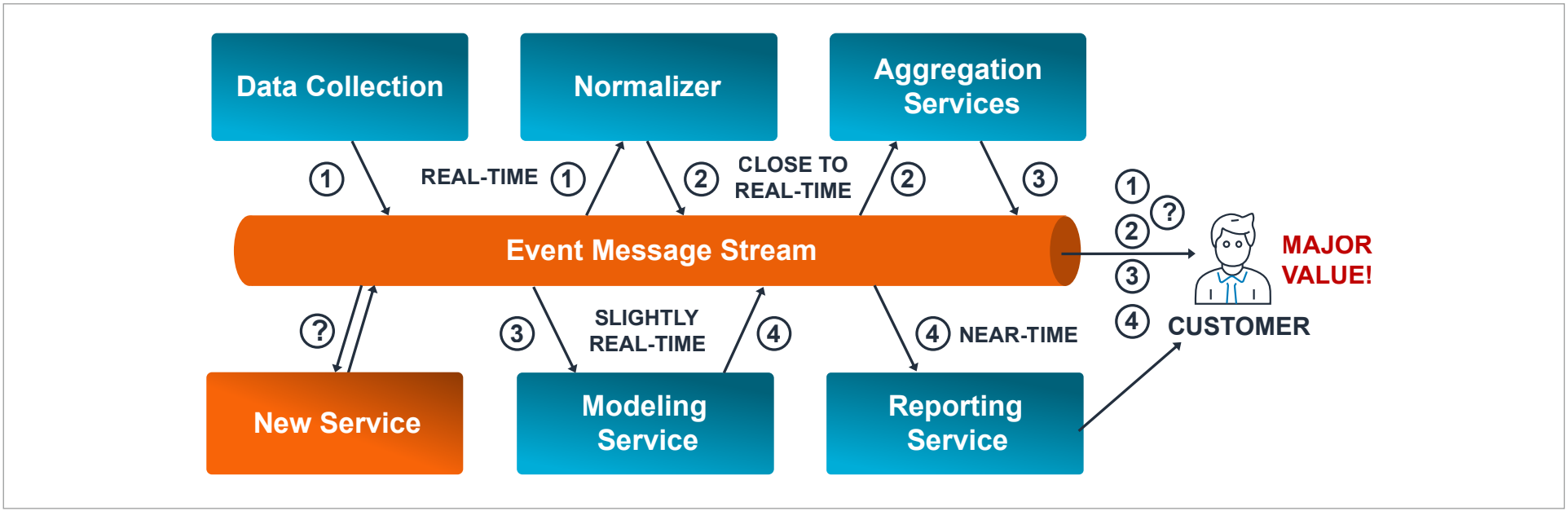
Now let's pause and take a look at a diagram of an event-driven architecture using the same systems we are using above.



The numbers you see are points in the workflow where the data may need to be normalized or we need to perform some aggregations. Other services will require some time to process the data and generate datasets for the workflow down the pipe. Note the diagram has been changed from a sequential left to right to a spoke-like system. You could even envision a spoke system where each of the services would process the data all at the same time!

With this approach, the customer would have the advantage of near real-time data availability.

We can also create new products for customers easily using the event message stream. Customers would be able to tap into the intelligence we have in our products to supplement their own learning systems. Whenever a customer asks for a product that we haven't created yet, we could deliver without too much orchestration involving the event message stream in nearly real-time. This is what it might look like:



In the example above, we plugged a 'New Service' (in orange) into the Event Message Stream and provided the customer with the value of that additional service. The '?' denotes that this service isn't bound to a particular step in the workflow. It could be something that goes in between ① and ② or at the end of the whole workstream. The new pluggable service could be something like a risk assessment service or a fraud detection service. Either would alert the customer when incoming data seems suspect—something of demonstrable value. In an event-driven system, tardy data causes only a slight delay to travel through the workflow and, once updated, automatically pushes out to the consumer. As a result, the customer walks away happy knowing we provided them the data as quickly as possible.

Containers are another good example of an event-driven system. We may fire up containers to isolate workflows and to minimize impact (or blast points). When a process needs to be run and there are memory constraints, containers can help. They do so either by spinning up another container to run the extra process or by increasing memory on the existing containers (via a restart).

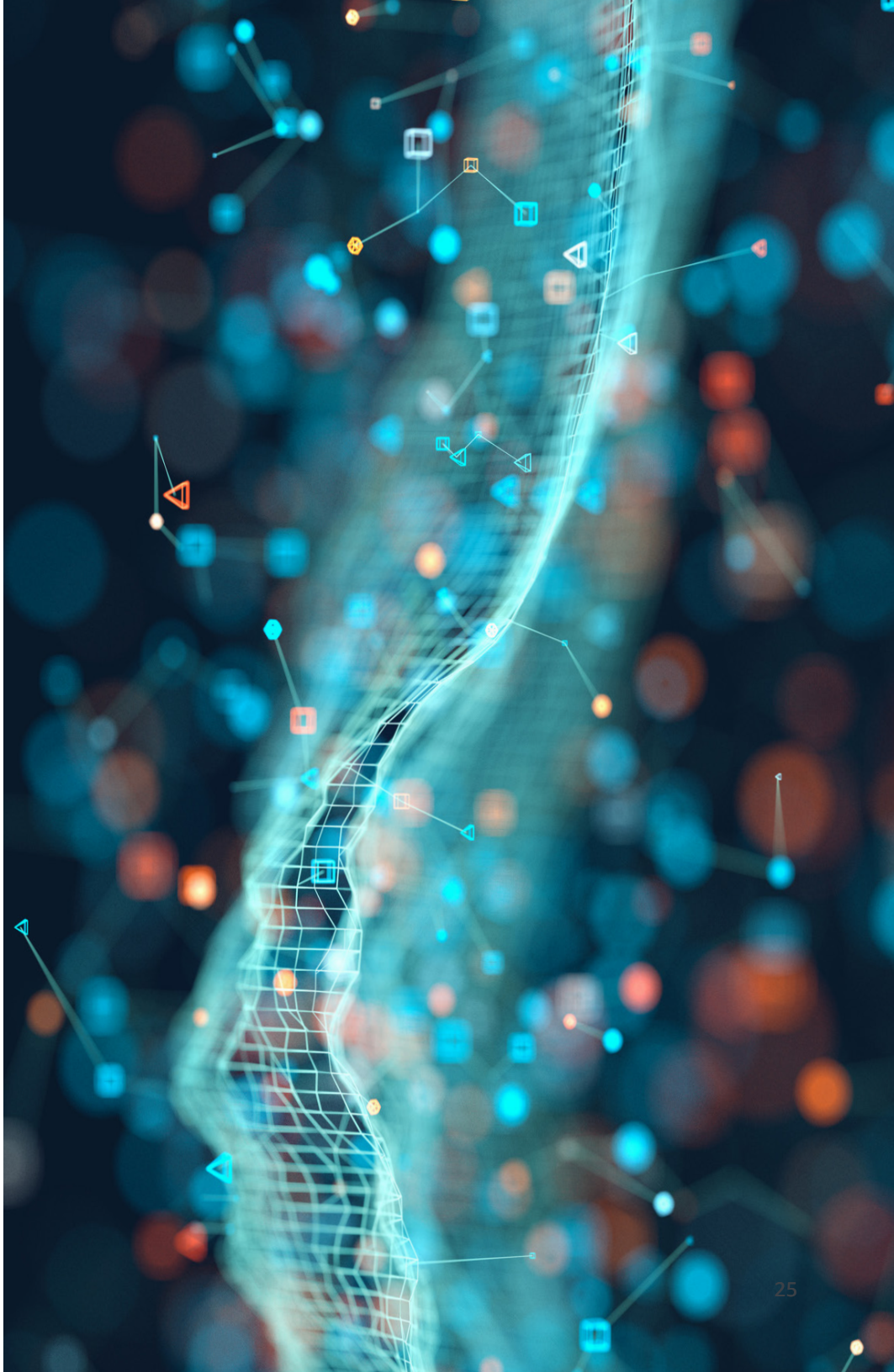
Containers can be a good way to scale as quickly as possible, increase data velocity, and thus increase the value of the data.

Going serverless (via the cloud) is another way to get increasing data velocity. Because serverless components offer more resiliency and it are scalable out of the box, using the cloud framework is a good way to implement your workflow. Implementation support from cloud providers has matured significantly and can provide an easy, quick path to migrate workloads and gain complete visualization.

Setting up an event-driven system is a complex undertaking and fraught with mistakes and stalls due to learning curves, organizational dynamics, and uncertain prioritization. Having an event messaging bus is a great approach to keeping the data moving and adding value as it traverses the workflow. In the short term, tools that ingest CSV files are less limited. Over the long-term, investing in messaging-based tooling to read events and to extract more value arguably wins a cost-benefit analysis.

Summary

I hope from reading this you can get some ideas on how you may increase the speed of data in your workflow while adding value to it. Your goal is to attain enlightenment with your data. Event-driven systems add speed to incoming data as it moves within your workflow, velocity that provides critical value to stakeholders and decision-makers within your organization. your workflow. Implementation support from cloud providers has matured significantly and can provide an easy, quick path to migrate workloads and gain complete visualization.



What's Your Organization's Data Innovation Strategy?

Jen Stirrup – Founder and CEO, Data Relish

Jennifer Stirrup is the Founder and CEO of Data Relish, a UK-based AI and Business Intelligence leadership boutique consultancy delivering data strategy and business-focused solutions. Jen is a recognized authority in AI and Business Intelligence Leadership and is a Fortune 100 global speaker, named among the most influential Top 50 Women in Technology worldwide. She has also been named as a Top 50 Global Data Visionaries, a Top Data Scientists to follow on Twitter.

Businesses have overcome the perception that the cloud is simply storage and are discovering that the cloud offers scalable, robust solutions for maximizing value from data. Organizations are finding out that the cloud provides opportunities to do so much more with their data, including the ability to use third-party data.

That said, businesses experience common obstacles in their journey toward deriving value from their data. It can be frustrating to demonstrate success when data visualization isn't as straightforward or quick as people think.

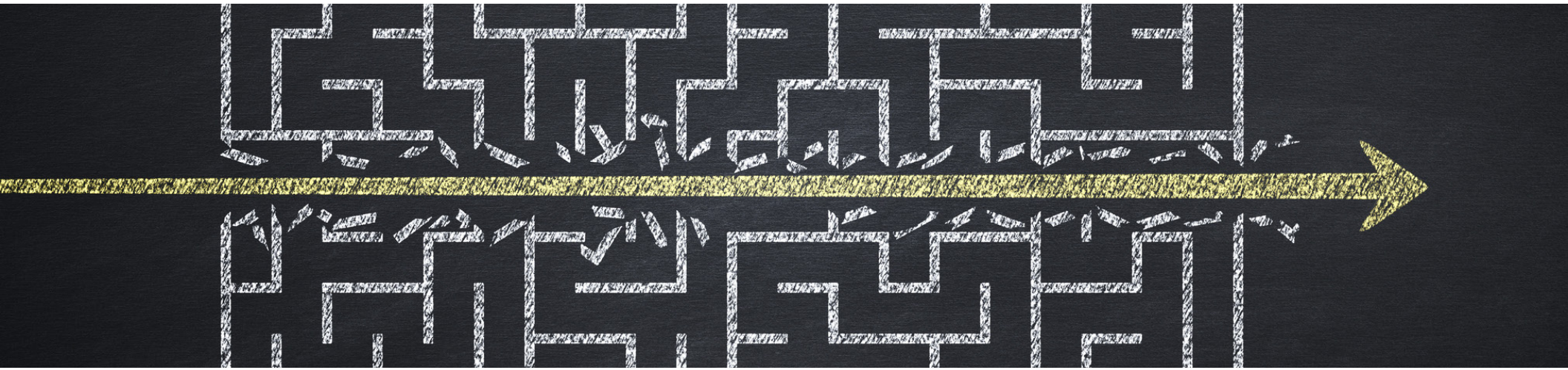
Business users want data that is easy to understand and quick to digest. Data forms part of business processes as well as being a product of those processes. This chapter offers insights into the challenges and solutions to adopting third-party data solutions. In addition, it provides a roadmap to achieve success with deriving business value with data visualization and third-party data in the cloud.

What's your innovation strategy?

The business landscape is changing. Customers want more than a product. Customers expect a meaningful relationship with their suppliers, and data can help organizations understand what is meaningful to them. In turn, this leads to innovation for customer-centered organizations who are determined to be their customers' best "go-to" solution. Data is most interesting and useful when set into context with other pieces of data.

Organizations often grapple with putting different sources of data together. Third-party data sources, along with data sources owned by the business, can help organizations hear the customer's voice in data, observing what is meaningful to customers.

Customers want a culture of innovation that speaks to them, and your data, along with third party data, can help them to achieve it.



Does your shadow data hide or reveal value?

Often to the dismay of IT departments, shadow IT and shadow data exist in many organizations. They are usually a just-in-time and popular method of solving a need for business insights. However, a shadow data strategy should not be in place for a sustainable business strategy to maximize data value. While technical teams often focus on big data, business teams frequently use "little data" in Excel worksheets, Tableau workbooks, or Google Sheets.

One key success criterion for little data is that it is often merged with third-party data sources, whether the sources are internal or external to the organization. There are many shadow data examples in a business's operational life, but it does not help the organization maximize value from the data horizontally across the organization.

Therefore, from the business perspective, little data can be as essential as big data because it helps quickly move the organization forward through a bottleneck. However, it is limited in terms of the opportunity to derive a longer-term, sustainable value that has a real impact on an innovation strategy.

Cloud storage helps settle the shadow IT and shadow data issues by de-risking data storage, securing data accessibility, and lessening human error in data engineering. Siloed data sources result in siloed analytics, which can introduce security issues from data fragmentation across departments, offices, and even countries. Third-party data can help to break down the silos if it is ingested well.

Cloud storage helps make data easy by securing data while freeing it to be accessible, thereby facilitating collaboration across business departments and IT teams. It is possible to eliminate shadow data by creating meaningful, business-relevant metrics that business users can access easily. Ideally, data engineering efforts could seamlessly align the metrics with a mixture of data assets and third-party data stores. Since cloud storage aligns data sources into one place, it adds value by reducing the organization's duplication of data sources while reducing shadow data reliance. Further, the scalability of cloud storage allows the business to scale and accommodate third-party data sources. In turn, the increased data consistency reduces error and time-wasting spent searching for the "right" version of the data.

Missing opportunities for innovation

In a competitive environment, organizations need to focus on tracking sales and finding new trends in data. In statistics, outliers often denote the start of a new trend, and they can be combined with useful third-party data sources. If your organization is not looking closely at third-party sources and data assets, it is easy to miss the outliers and take advantage of the potential opportunity they bring.

Innovation and the last mile of analytics

Automation can help the organization to maximize the benefits of innovation. Automation takes the value an organization gets from embracing cloud analytics to the next level. Automation streamlines business processes by removing the heavy management aspects which are generally associated with managing traditional business intelligence.

Automation offers an organization additional benefits by automatically including third party data sources in the data pipelines. Automation allows business intelligence and data science teams to reliably reproduce data-related assets by executing pipelines. Automating processes using DataOps principles means that the analytics solution is more likely to be uniform. Uniformity leads to a reduction in the likelihood of human error. In addition, it allows IT departments to include security, business intelligence, and analytics improvements, helping to imbue the organization with a privacy and security culture from the data upwards.

How can you get started on supporting your innovation strategy with technology?

Organizations can get stuck in the innovation 'Bermuda Triangle' because they do not know how to move forward. As a result, the organization can become an innovation graveyard where great ideas go to die. Here are some steps to help you to bring your innovation strategy into real life.

1 Customer as catalyst

The customer should be the catalyst for innovation in the organization. Being customer-centered means that it will be easier for the organization to find support for these initiatives.

2 Innovate while solving a problem

Offering a roadmap of initiatives will help garner broad support if the effort is directed towards helping to solve problems, such as data cleansing, as the program progresses. It can be difficult to get support for potentially risky initiatives. It is easier to persuade executive sponsors if you can demonstrate that the initiatives will also help to chip away at data or technical problems in the organization.

3 Start small but think big

Choose a business problem that adds value to your organization and automate it. If the organization is starting out on this journey, then it is recommended that the organization chooses a small problem that has visibility across the organization. Here is a breakdown of the types of problems that the organization could tackle first.

4 Measure the speed of the business

For some teams, innovation will be satisfied if the organization has succeeded to improve internal processes without necessarily changing the way that the organization functions, or the business works. For organizations like this, the innovation process will remain inside the organization, smoothing out processes to improve efficiency and customer satisfaction. An example of this activity could include handwriting transcription, for example, using functionality available in Amazon Comprehend.

For other organizations, innovation will only be satisfied if the customer sees a visible, external change in automation. An example of that may be interacting with a website through omnichannel chat such as Amazon Lex. Organizations can work towards this innovation by starting with smaller, internal automation projects first, before moving towards activity that is visible to the customer.

5 Flavors of automation in the organization

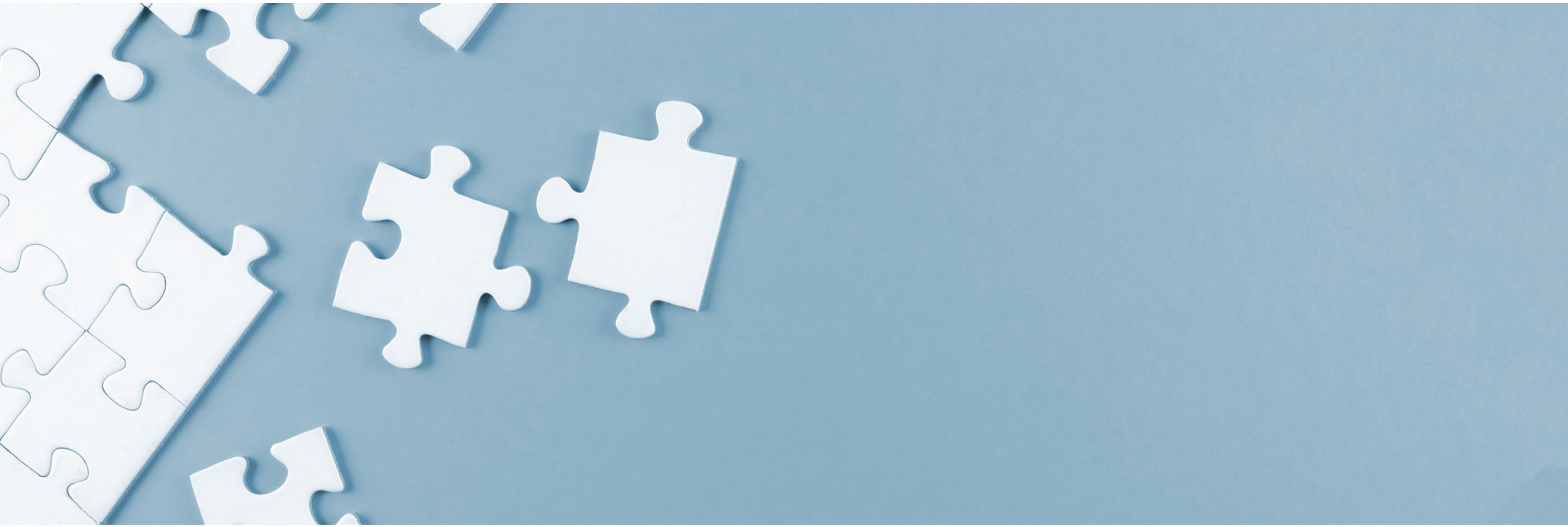
Organizations can start small, and then move forward to build on their successes, which will build trust. It's perfectly fine to go through a pilot or proof of concept project and deliver a prototype first.

Technology to support the innovation strategy

A strategy is not just a list of projects. The customer must be the focus of the success, driving innovation, which will help to form an overarching strategy. Once a proof-of-concept project is delivered, teams can rest on their laurels once a proof of concept is completed. The strategy needs to be big and bold if it is to escape from the Innovation Bermuda Triangle to address key business challenges and create value.

The automation strategy should align with the business strategy, but it may also challenge it. Executives must understand the benefits and value of the overall strategy and how it fits into such objectives as risk reduction, cost reduction, or improvements in customer experience. The question is this: What should the organization achieve? Then, the business teams can work backwards to determine the capabilities, technologies, RACI, program, and so on.

Automation can help to meet business goals such as reduction of risk and costs, and third-party data sources can be incorporated along the way to enrich business intelligence and analytics while supporting business decisions. Businesses are increasingly reliant on to meet customer needs, and there is little appetite for downtime or data loss.



Summary

Organizations can maximize investments using cloud technology and data in a smart, data-driven way that is open to innovation as well as transformation. Analytics can offer businesses great opportunities, and ideas are developing into reality at accelerated rates.

Newton's laws worked well until objects moved fast, and Einstein's insight was always thought to be that the speed of light is an absolute constant. Einstein's real insight, however, is that when we push our perceptions beyond our known world, our principles of operating change and we must rethink "normal." Enterprises need to rethink their normal, and how organizations function when humans cross the boundary where external clock speeds outrun our internal clock speeds.

We live in a culture of shock, a culture of now. Analytics, data visualization, and automation can help us to get there. However, we need to ensure that it is directed by an innovation strategy that is led by our customers.

Expand the Power of Analytics with Third-Party Data

Ronald van Loon – CEO, Principal Analyst at Intelligent World

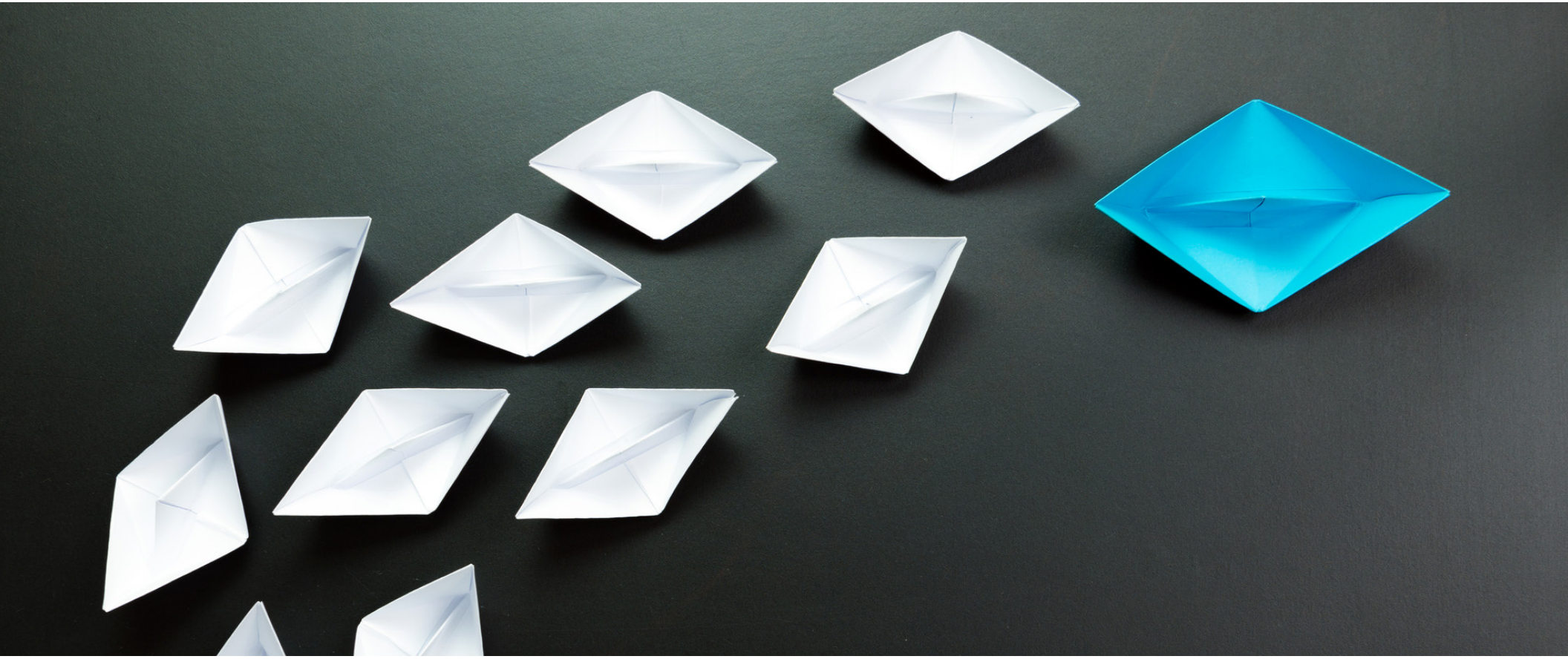
Ronald van Loon is CEO at Intelligent World, a network of thought leaders and analysts focused on enabling collaboration and content-sharing in the AI and tech space. His expertise spans big data, IoT, AI, machine learning, 5G, deep learning, predictive analytics, cloud, edge, and data science. He has been recognized as a global influencer and thought leader with a social media presence that reaches over 340,000 active data and analytics followers.

Expanding analytics to include third-party data enables organizations to grasp new competitive opportunities, understand shifting market factors, and changing customer behaviors. Additionally, companies can fill in gaps and address limitations and overlooked risks that potentially arise from exclusively using internally-produced insights and information.

Adopting a strategy to maximize the power of external data and effectively integrate it with internal data is crucial for success. It can help organizations drive faster data-driven decision making, optimize supply and demand, enhance data accessibility across the organization, and efficiently meet changing customer demands.

Leading analytics innovators use the full breadth of data sources to augment their business and data ecosystem. Doing so creates a robust analytics stack that improves the quality of predictive analytics and modeling capabilities.

This can't be accomplished with internal data alone. Businesses must take advantage of diverse data sources and types to better adapt. This collection of data helps organizations overcome challenges emerging from the pandemic, minimize risks, and investigate areas that can lead to more value and growth.



Exploring value areas

Third-party data sources can help organizations explore critical areas of value throughout diverse business functions:

- Build a 360-degree view of customers across interactions and supplier and partner environments
- Recognize shifting consumer sentiments and emerging events that impact product and service demand
- Generate new revenue streams from strategic product and service development and improvement
- Optimize organizational talent based on peer, competitor, and market benchmarks
- Mitigate operational, reputational, and supplier risks according to real-time analytics

Challenges and concerns

Organizations must incorporate responsible approaches to third-party data to avoid potential issues surrounding privacy and ethics. Third-party data can also present concerns over trust if the data is reliable, biased, or inaccurate.

Other challenges include:

- Obtaining and understanding what third-party data is available due to its rapid growth and fragmented nature
- Having the right skills and teams in place to support third-party data sourcing, including understanding how the data is collected, identifying possible use cases, and evaluating operational value
- Maintaining data privacy and security standards in increasingly complex regulatory environments
- Accounting for changes to infrastructure, data architecture, and systems when integrating external and internal data
- Verifying that external data sources are ethically sourced

Summary

Organizations must build a strategy that addresses challenges and allows them to ultimately operationalize and deploy third-party data at scale to drastically enhance business outcomes.

This should include implementing a continual process of identifying, evaluating, and integrating new third-party data sources with internal data. Develop a data ecosystem program that connects analytics, business needs, IT, business teams, and legal departments. Ensure that a platform is developed to effectively manage all factors revolving around data acquisition, integration, and management.

Augmenting analytics efforts with the right third-party data allows organizations to reimagine their operations and meet new urgency to innovate and respond to continual disruption and change.

What Is Data Analysis and What Is Its Importance for Companies?

Marcus Borba – Founder, Borba Consulting

Marcus Borba is a technology expert, consultant, entrepreneur, executive, speaker, and advisor who is passionate about innovation and digital transformation. He has more than three decades of in-depth business and IT experience, and more than 20 years of experience developing data-driven solutions for companies. Marcus is also the founder of Borba Consulting, which specializes in developing strategies to plan and implement innovative data-driven solutions.

Digital transformation and the evolution of computing power and data storage has resulted in increasing agility for organizations seeking information access for decision making. It has become clear that the data revolution is changing companies in a profound and unalterable way. Most large companies constantly collect data, but in its raw form the data has no value.

Data analysis is the process of analyzing raw data to extract patterns, trends, and insights that can generate meaningful insights to enable intelligent decisions. It helps us to unravel, compile, and gain insights from massive quantities of data. Doing so enables professionals to capture results from sales, marketing, customer relations, and virtually any other source to make better, rational decisions that impact aspects of business.

Technological innovation has increasingly changed the structure of our organizations, who is empowered to make decisions, and how we work. So, almost everyone from the C-suite across the enterprise plays some role in this process, whether to meet a specific need or provide context for decision makers.

There are four types of analysis (defined below) divided into two layers.

The first layer is the traditional one, capturing what has already happened without issuing judgments. It can also involve diagnostic analyses, which determines what motivated a given event based on the relationship between two or more variables. In this case, indicators are constructed, and the value judgment is adopted.

A second layer of analysis includes types of prescriptive and predictive analysis. This layer provides more in-depth analysis and is complementary to—not replacing—descriptive and diagnostic analyses. The combination gives the analyst a better understanding of the scenario as a whole.

In the following, we define four types of analyses:

1. **Predictive analysis.** Forecasting, or predictive, analysis combines powerful statistics and artificial intelligence to help you understand what might happen in the future. This well-known model makes it possible to make decisions in a more assertive and accurate way. Gartner defines predictive analytics as "a form of advanced analytics which examines data or content to answer the question 'What is going to happen?' or more precisely, 'What is likely to happen?'" It is characterized by techniques such as regression analysis, forecasting, multivariate statistics, pattern matching, predictive modeling, and forecasting. Predictive analysis includes a variety of statistical techniques that analyze current and historical facts to make predictions about the future.

   There are two main techniques used in descriptive analysis: data aggregation and data mining. Data aggregation is the process of collecting data and presenting it in summary form. Data mining is when the analyst explores the data to discover any patterns or trends.

2. **Descriptive analysis**: Descriptive analysis is a process that, through statistical techniques, can summarize or describe a data set. As one of the main types of data analysis, descriptive analysis is popular for its ability to generate accessible insights from data not otherwise interpreted. The purpose of this model is to allow the analyst to understand the events in real time. Descriptive analysis does not issue a judgment of value. Instead, it is intended to visualize the data and understand the impact on the present without creating any relationships to the past or the future.

3. **Prescriptive analysis**: Prescriptive analyses are aimed at projection, but with a greater focus on decisions that are made, helping to define of strategies to be developed and followed. It works to analyze a range of scenarios, predict the outcome of each, and decide what is the best course to take based on the findings. A goal is outlined and based on that, the paths are defined that must be followed to reach it.

4. **Diagnostic analysis:** Diagnostic analysis is a type of advanced investigation that analyzes content or data to answer the question "Why did this happen?" The idea is to be able to analyze the impact and scope of an action that was carried out. From this, strategies can be devised to improve the results.

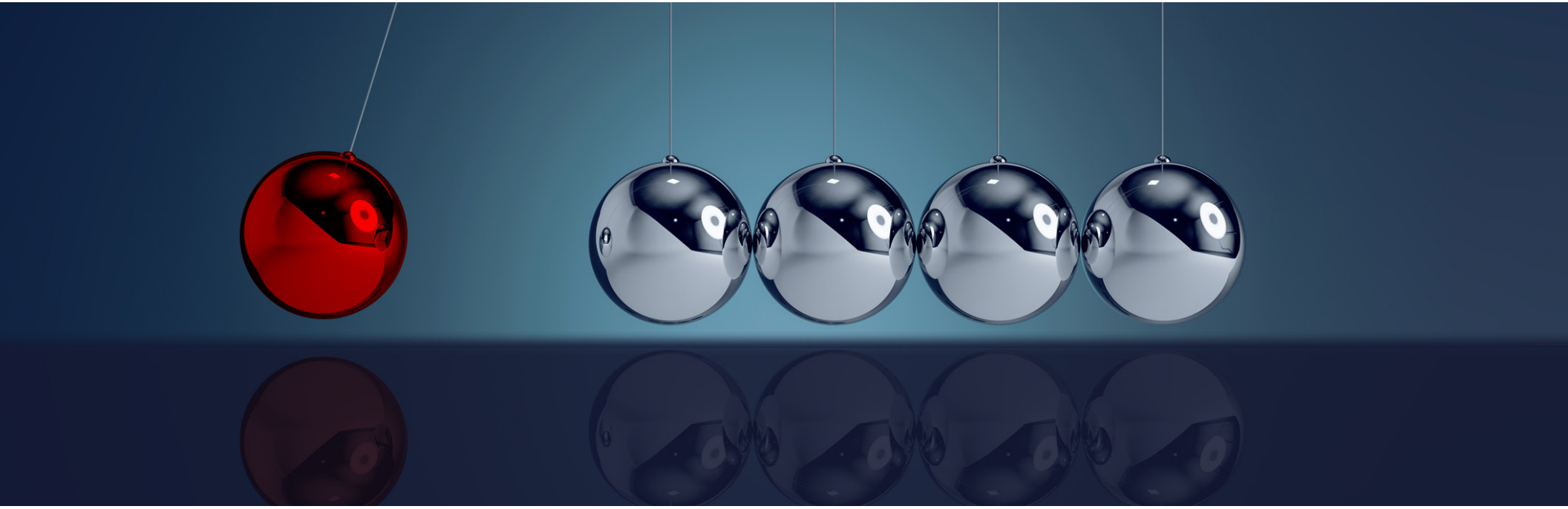
How and where to use different types of data analysis

Each type of data analysis serves a specific purpose. Predictive analysis is the most used because it allows companies to understand the impacts of some of the metrics they track. Predictive analytics is employed by companies like Amazon to help understand what products or services to recommend to their customers.

Descriptive analysis does not seek to make predictions about the future. It draws insights only from previous data, manipulating it in ways that make it more meaningful. Examples of descriptive data analysis may include reports, evaluation of results, application of metrics, and segmentation of customers. One of the main uses of descriptive analysis is to guide the construction of strategies.

Prescriptive analysis uses advanced tools and technologies, such as machine learning, business rules, and algorithms, which make it the most sophisticated to implement and manage. A prescriptive model considers all possible decision patterns or paths that a company can take and their likely results. For example, it can be used to predict whether a given topic will be popular with readers based on search data for related topics.

Diagnostic analysis allows you to better understand your data. It enables you assess the scope of an action taken, better understand the action's effects, and respond more quickly to critical business questions.



The main steps of a data analysis process are:

Collection of data requirements: Ask yourself why you are doing this analysis, what kind of data analysis you want to use, and what data you want to analyze.

Data collection: After capturing requirements, you will have a clear idea of what should be measured and some likely results. Data will be collected from a range of sources and must be processed or organized for analysis. You should keep a record of the date of collection and the source of the data.

Data cleaning: Not all the data you collect will be useful, so take the time to clean it. During this phase, you remove duplicate records, blanks, and other basic errors. Data cleaning is mandatory before submitting the information for analysis.

Data analysis: Once the data is collected, cleaned, and processed, it is ready for analysis. As you manipulate the data, you may find that you have the exact information you need, or you may need to collect more data.

Data interpretation: Now that you have your results, you need to interpret them and propose the best courses of action based on your findings.

Sharing your results: After finishing the analysis, you should share these insights. Sharing insights is more complex than simply sharing the results of raw data because you must contextualize the findings and present them in a way that is easier for everyone to understand. Since you will often present information to decision makers, it is very important that the insights presented are clear and unambiguous.

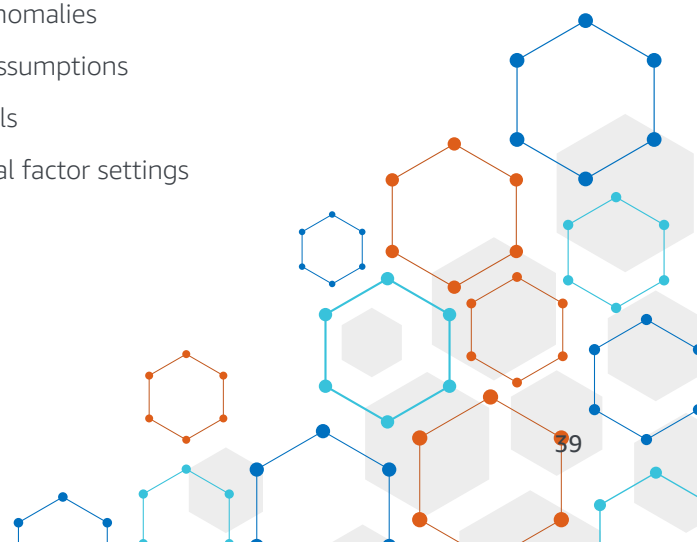
For this reason, data analysts often use dashboards (interactive visualizations) and reports to support their findings. How you interpret and present results often influences the direction of a business. Depending on what you share, your organization may decide to launch a new product or service, or even close business units.

That dependence is why it is very important to provide all the evidence you have gathered. Ensuring that you cover everything in a clear and concise manner will prove that your conclusions are scientifically correct and based on facts. Just as important, you should highlight any gaps in the data or signal any insights that may be open to interpretation. Honest communication during this process is critical to the business and to helping you to stand out in your work.

Data visualization is a simple yet sophisticated way to display your information graphically in a way that people can see and understand. You can use maps, markers, graphs, tables, or any range of other methods. But it's more than just a good way to share information—it also helps you to gain valuable insights as you compare data sets and observe relationships.

Exploratory data analysis (EDA): EDA is what data analysts do with large data sets to identify patterns and summarize the main characteristics of the data set, and what they learned in modeling and testing hypotheses. EDA is normally used to check assumptions, study the relationships between variables, check the data for errors, or observe trends. During the EDA process an analyst may:

- Maximize insights into a data set
- Discover underlying structures
- Extract important variables
- Detect outliers and anomalies
- Test the underlying assumptions
- Develop thrifty models
- Determine the optimal factor settings



The importance of machine learning for data analysis

Machine learning is used to help automate the construction of models for data analysis. Data analysis has been characterized by a trial-and-error approach and is thus considered impossible to use when there are heterogeneous and significant datasets in question.

The need for a manual touchpoint is one of the challenges of big data. The availability of more data is directly proportional to the difficulty of building precise, new predictive models. Traditional statistical analyses are focused on analysis of samples frozen in time, generating inaccurate and unreliable conclusions. Machine learning can help companies filter through a range of inputs to discover what variables impact the conclusion.

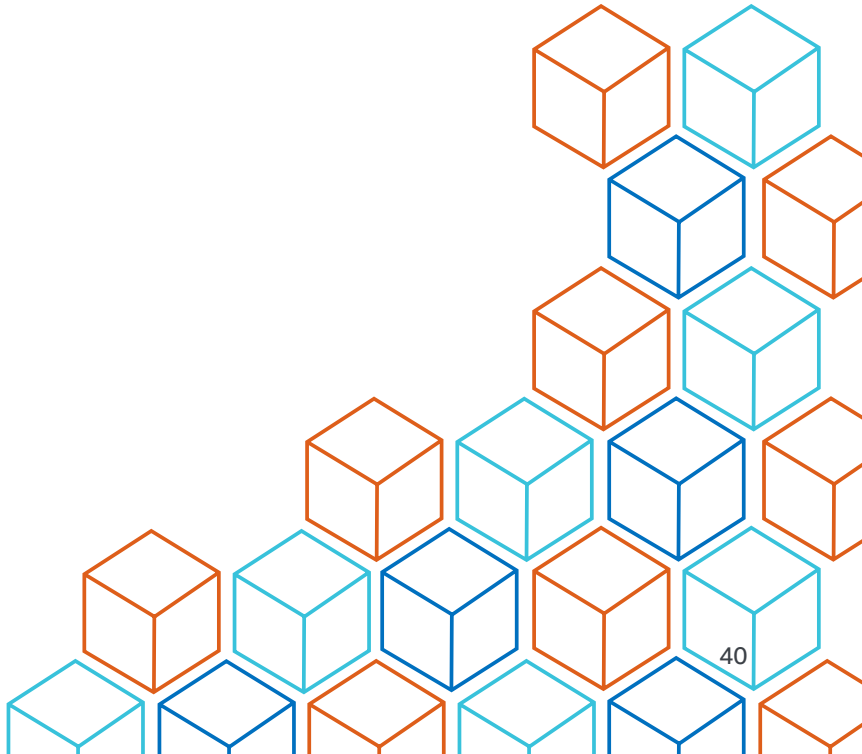
Machine learning use cases in data analysis

Healthcare: In healthcare, data analysts can use machine learning to automate drug research or predict epidemic outbreaks. In a diagnostic setting, it can help to personalize treatment or detect differences in cancerous and healthy tissues, among many other applications.

Financial services: Predictive analysis based on machine learning can help companies prepare revenue projections toward which objectives, goals, and cash flows can be planned. It can be used in such diverse areas as fraud detection and prevention, loan and/or insurance underwriting, algorithmic trading, and to recommend financial products.

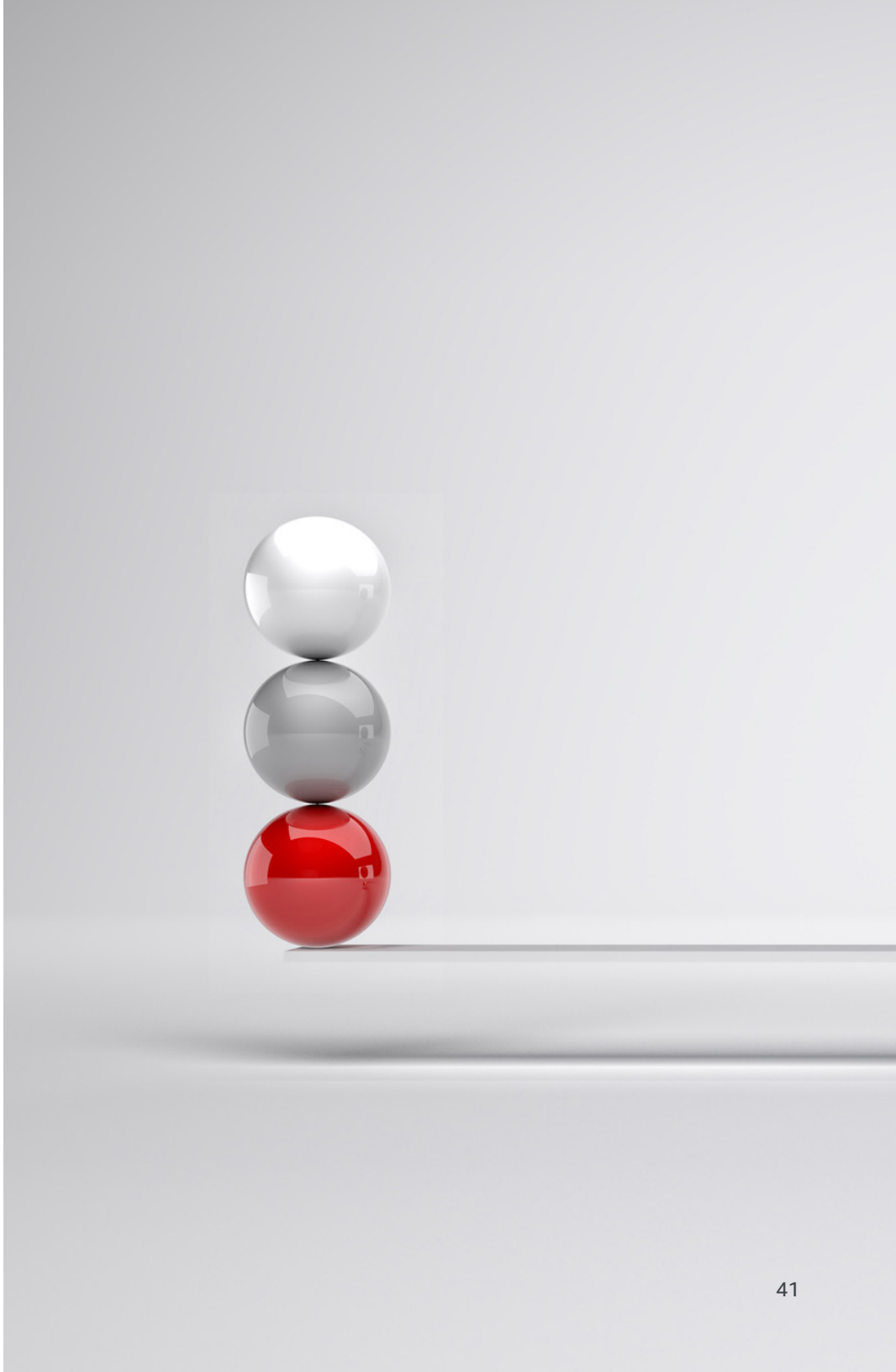
Sales and marketing: There are many examples of machine learning in sales and marketing. Predictive machine learning analysis can combine historical data points of customer behavior with market trends. Data analysis and machine learning are also used for marketing campaigns. This process uses algorithms to analyze lists of social media comments, as well as the most frequent positive and negative words and phrases. This data is used by marketers to make better decisions in campaigns and to define marketing strategies.

E-commerce: Predictive analytics supported by machine learning algorithms can help retailers understand customer behavior and preferences. Recommendation Engines are one of the main Machine Learning use cases in the online retail space, can be used to customize the experiences of your customer, and allow you to take the experience of other customers and recommend products that people who live close to the customer are buying. Customer segmentation separates customers into smaller groups that are properly targeted. The segments can be identified logically or using a machine learning algorithm. Predictive machine learning analysis also facilitates the management of supply chain processes.



Summary

All types of data analysis have a role within organizations. Each of them addresses a specific need that, applied to a given context, can help guide companies to make better data-based decisions.



The External Data Imperative and Valuation Approaches

Doug Laney – Innovation Fellow, Data & Analytics Strategy, Infomics

Doug Laney is the originator of the field of infonomics and author of the best-selling book, Infonomics: How to Monetize, Manage, and Measure Information for Competitive Advantage. Doug is the Data & Analytics Strategy Innovation fellow with the consultancy West Monroe. Formerly he was a vice president and distinguished analyst with Gartner's Chief Data Officer (CDO) research and advisory practice. He is an accomplished practitioner and recognized authority on data and analytics strategy, is a three-time recipient of Gartner's annual Thought Leadership Award, and is regularly considered one of the top global influencers these topics. In 2001, he coined the "3Vs" of volume, velocity and variety, now commonly used in defining Big Data. Doug is also a visiting professor at the University of Illinois Gies School of Business and the Carnegie Mellon Heinz College of Business, where he teaches graduate-level courses on analytics and infonomics, also available via Coursera.

Shifts in global and local economies due to the Covid-19 pandemic and the discrepant, ever-changing national and local responses have rendered the rudimentary analytic models of many organizations useless, or worse—detrimental to the business. Too many businesses rely on too many analytic models that rely on too much internal data.

The difference in seeing the economy as interconnected and interdependent versus insular and independent is the difference between trend-based and driver-based modeling. Trend based models rely on the company's own historical data, such as sales or production data, and sometimes macro-level industry data. These formulae expect trends to continue on a similar path and at a similar pace. Driver-based models, on the other hand, rely more on leading indicators of performance or business activity. They incorporate external data about situations or observations that highly correlate to and presage one's own business outcomes. This is like driving the business by looking in the rearview mirror versus peering ahead at various nearby and distant traffic patterns while noting the weather and even listening to what other drivers and pedestrians are thinking.

Although forecasts crunching historical data may work fine in normal times when markets are relatively stable, in anomalous times like today these trend-based models' falter. In times of turmoil, leading indicators prevail. Still, some organizations have decided not to wait for turmoil, but rather have advanced to relying on driver-based models, regardless.

Economists, business analysts and data scientists are finding that even historical data from the last several recessions is of little use. In those instances, the bubbles bursting predictably spread from one sector to another. It was easy to discern and establish standard sets of leading and lagging indicators. Not today. And not likely ever again.

Many company executives, for lack of a better approach, have simply reverted to using their intuition. Unfortunately, it takes an economic crisis to get business and IT executives to realize that new analytics techniques and data sources are required. Other executives remain paralyzed. A lot of executives just can't seem to wrap their heads around what a leading indicator is or where to find them. This speaks to another global epidemic, one throughout the business world: a deficiency of data literacy.

The Wide World of External Data

Organizations that don't know what data they have, or need, or could get their hands on, are unable to leverage it as an asset. As one CIO for a large insurance company once admitted to me, "It's silly that someone around here has an inventory of our office furniture, but nobody in the company has an inventory of our data." For most organizations, this includes external data as well.

So, the first job in being able to manage data as an asset, which in turn enables it to be sufficiently leveraged as one, is to know what information you have or could get, and where it is. The variety of sources of data are varied, and the number of data sets and elements even more so. In addition to the typical data collected and generated during the normal course of business, a wide world of external data exists with the potential to help an organization become even more data driven.

Operational Data

This is data about customers, suppliers, partners and employees that is readily accessible in online transaction processing and/or online analytical databases. It typically includes transactional data, contact data, process data and master data. Enterprises often have the opportunity to collect even more data during the course of business via sensors or process monitoring such as:

- Log data
- Smart meters
- Internet-connected devices (IoT)
- Voice
- Image
- Security cameras
- Radio frequency identification (RFID)

Dark Data

Then there is data collected during the course of business that remains in archives, is not generally accessible, or has only been used for a single operational purpose. It can include emails, contracts, documents, multimedia, system logs or other intellectual property. Parsing, tagging, linking or otherwise structuring or extracting usable information from these sources often is the greatest immediate opportunity for most businesses among all types of internal data.

External Stakeholder Data

More recently, many organizations have begun entering into data exchange agreements with external stakeholders such as their partners, suppliers, or even their customers. Business partners share information about prospects, sales, product support and service, and industry trends. With suppliers, companies share data on sales and production forecasts, shipping and distribution, supplier product usage and performance, and consumer information such as demographics and product/service preferences. And with customers, businesses can encourage them to share information about themselves, their likes/dislikes, buying intentions, and so forth. These data exchange agreements with external stakeholders can be bidirectional or unidirectional in return for cash or favorable commercial terms.

Syndicated Data

For many years, industry-specific data aggregators (D&B, Equifax, Nielsen, and Experian, for example) have offered syndicated credit, real estate, postal, household and other information by subscription. Today, marketplaces are emerging for almost any variety of available information from companies like Dawex, IOTA, Eagle Alpha, etc. And black markets for personal and commercial data are rampant. Even among business partners, data assets are being used in private bartering where once only financial and material assets were. Data and analytics leaders, along with business leaders and enterprise architects, all need to be aware of commercial information sources that relate to their market and assess their potential, as well as work with business partners to encourage the availability of their information.

Open Data

Many governments and NGOs also have begun opening their information coffers as a matter of principle or dictum. Open government initiatives for economic development and for health, welfare and citizen services are in various stages of implementation throughout the world. This data can also have real commercial value, especially when mashed with other sources, to understand and act on local or global market conditions, population trends and weather, for example.

Social Media Content

Individuals and businesses blogging, tweeting, yammering, Facebook and LinkedIn posting have created a fast-growing, potentially precious source of data about preferences, trends, attitudes, behavior, products and companies. Posts, trends and even usage patterns themselves increasingly are used to identify and forecast target customers and segments, market opportunities, competitive threats, business risks, and to select job applicants.

Web Content

Scraping content from competitor or partner or industry websites can turn the Web itself into your company's biggest database. Vendors such as Scrapestack, Apify, Mozenda, and Import.io specialize in harvesting this type of content. Increasingly, organizations keep tabs on their marketplace by harvesting information primarily from competitor websites such as press releases, product pricing and descriptions, job postings, and so forth.

Determining the Value of External Data

Curating, acquiring and integrating external data takes time and money. The rewards usually are substantial. My research compiling over 500 high-value use cases of data and analytics shows that upwards of 75 to 80 percent of them make use of externally obtained data. Regardless, this kind of expense should be formally justified, just as any other kind of purchase.

But how does an organization value data? Due to arcane and archaic accounting regulations dating back to the Great Depression, data is not a balance sheet asset. Quantifying the value of data is not something that businesses, their accountants, or their CFOs typically have any experience with. The good news is that standard valuation approaches can be readily adapted to valuing data, and even non-financial valuation methods can provide input into decisions about licensing or otherwise acquiring external data assets.

In my book "Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage," I shared details on the following financial and foundational models for measuring data's value which were developed in collaboration with valuation experts, accountants and clients.

First, however, organizations should be sure to understand the various quality characteristics of candidate data sources. This includes measuring up to a dozen different dimensions, including:

Objective Data Quality Metrics

Objective metrics are those which can be readily measured.

Validity: how well non-missing data accurately represents reality, irrespective of precision or format.

Completeness: the percentage of instances of data in the source versus the total available, and/or the percentage of missing fields in a record.

Integrity: the existence and correctness of linkages to and from a record, and often the legitimacy of relationships among attribute values within a record.

Consistency: the number of different forms, formats, and/or structures data takes when stored in multiple datasets or records.

Uniqueness: the percentage of alternate or duplicative forms of the data instance that exist.

Precision: the degree of exactitude of a value (or the level of detail for unstructured data).

Timeliness: how current or fresh is the data being harvested or licensed

Accessibility: the estimated number of business processes or personnel that can benefit from the data and that would be able to use it.

Subjective Data Quality Metrics

Subjective indicators of data quality may be more difficult to determine for external data sources, but should still be considered, some via surveying intended or potential users.

Existence: the degree to which business events, objects, and ideas of importance to the business are represented in corporate information assets.

Scarcity: the probability that other organizations (particularly competitors and partners) also have the same data. Typically, externally acquired data lacks scarcity, but there are exceptions.

Relevancy: the number of business processes that use, or could benefit from, this type of data.

Usability: the degree to which data is helpful in performing a business function.

Interpretability: the degree to which data has a unique meaning and is easy to understand.

Believability: the degree to which the data is trusted.

Objectivity: the degree to which the source of the data is believed to be impartial and unbiased.

Foundational Data Valuation Models

More so than just data quality characteristics, data's value can (and should) be assessed just like any other asset. The following fundamental valuation models are for organizations or departments that are not yet ready or have no pressing need to ascribe a monetary value to their information assets. These models are useful for assessing an information asset's quality and potential-versus-actual utility to help in improving identity access management (IAM) efforts. They may also be useful as leading indicators of an information asset's potential economic benefit.

Intrinsic Value of Information (IVI)

The intrinsic value of information is its presumptive benefit that enables broad comparisons across information classes regardless of how the information may currently be being used.

The IVI is a function of:

Validity. Percentage of records deemed to be correct.

Completeness. Percentage of total records versus the universe of potential or supposed records.

Scarcity. Percentage of your market or competitors that also likely have this same data.

Lifecycle. The reasonable usable length of utility for any given unit (record) of the information asset (e.g., in months)

$$IVI = Validity*Completeness*(1 - Scarcity)*Lifecycle$$

Where p = the number of business processes or functions.

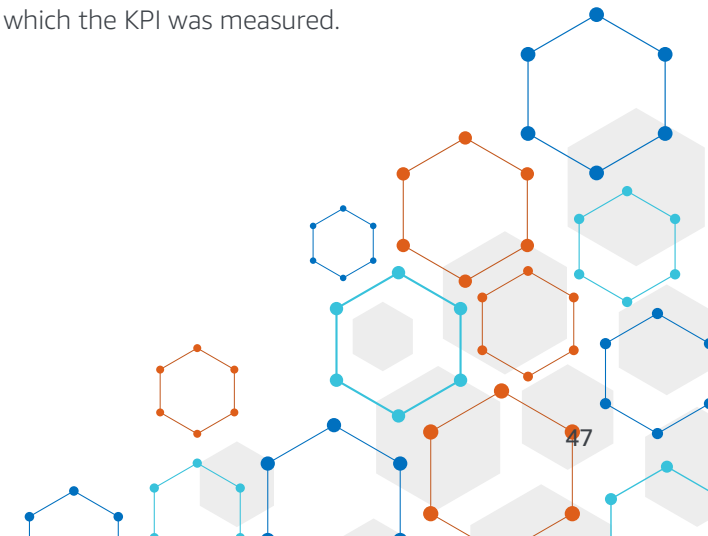
Performance Value of Information (PVI)

This approach looks at the realized (or estimated) impact of an information asset on a business objective that is represented as key performance indicators (KPIs). This answers the question: How much does having this information improve business performance? In short, it requires running a controlled experiment (or conjecturing one), but this method results in a definitive, empirical value measure. The PVI is a simple ratio that calculates KPI improvement by incorporating a given information asset, extrapolated over the usable life span of any given instance of data:

$$PVI = \left[\left(\frac{KPI_i}{KPI_c}\right) - 1\right] * T/t$$

Where:

- KPI(i) = Business process instances using the information asset (informed group).
- KPI(c) = Business process instances not using the information (control group).
- T = The average usable life span of any data instance.
- t = The duration over which the KPI was measured.



Financial Valuation Models

A financial information valuation model is useful to organizations that need to determine how information assets perform compared to other assets; what to invest in their collection, management, security, and deployment; and how to express their value when used in business transactions (for example, merger and acquisitions, data syndication, information bartering). These economic models are variants on established asset valuation models that are used by valuation experts and accountants to value traditional assets. However, these models have been adapted to accommodate one of the nuances of information's unique characteristics including that they are non-depletable, non-rivalrous, and more licensable than salable.

Cost Value of Information (CVI)

This method simply assesses an information asset as the financial expense required to generate, capture, or collect it.

$$CVI = \frac{ProcExp*Attrib*T}{t}$$

Where:

- ProcExp = The annualized cost of the process(es) involved in capturing the data.
- Attrib = The portion (percent) of process expense attributable to capturing the data.
- T = Average life span of any given instance of data.
- t = Time period over which the process expense is measured.

Market Value of Information (MVI)

This method looks at the potential or actual financial value of an information asset in an open marketplace. Typically, data monetization is transacted among trading partners in return for cash, goods, or services, or other considerations such as preferential contract terms and conditions. For licensed data assets, this model can be used to determine the value of comparable data.

$$MVI = \frac{Exclusive\ Price*Number\ of\ Licenses}{Premium}$$

Economic Value of Information (EVI)

This method generates the net financial value of an information asset by applying the traditional income approach for asset valuation, then subtracting the information's associated lifecycle expenses. Like the PVI, this method empirically calculates the information asset's actual value. As such, it is more of a trailing indicator than a leading indicator of information value—unless the first revenue term can be estimated adequately.

The EVI considers the realized change in revenue when a particular information asset is incorporated into one or more revenue generating processes. Then, the cost to acquire, administer, and apply the data is netted out:

$$EVI = [Revenue_i - Revenue_c - (AcqExp + AdmExp + AppExp)]*T/t$$

Where:

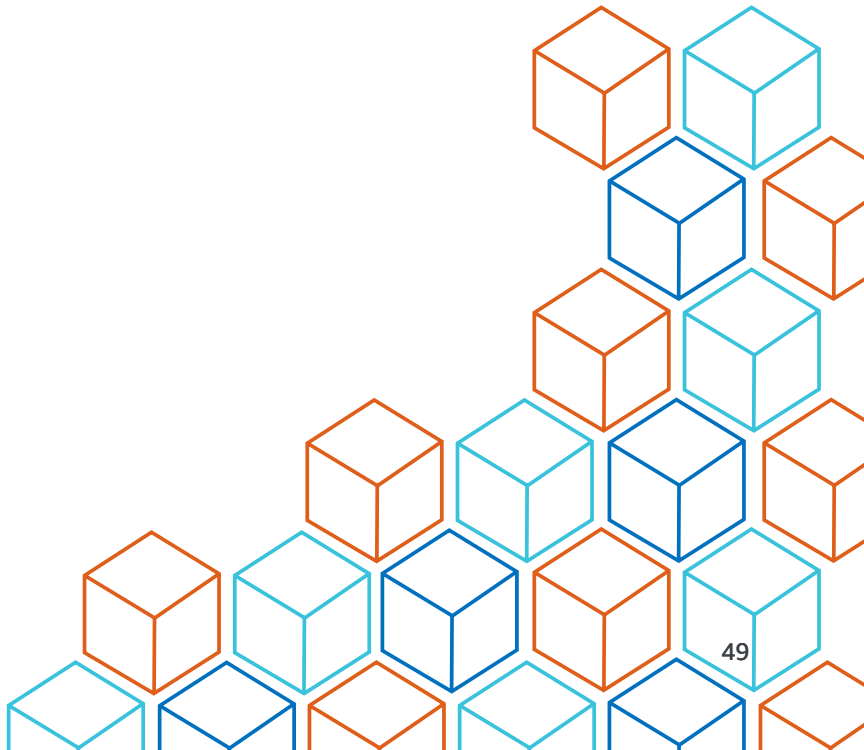
- Revenue(i) = The revenue generated using the information asset (informed group).
- Revenue(c) = The revenue generated without the information asset (control group).
- T = The average expected life span of any given information instance or record.
- t = The period of time during which the EVI experiment or trial was executed.

Thriving in the The Data Economy

Wielding data more broadly and continuously to drive performance and growth, to shrink costs and mitigate risks, and to digitize products and services requires a vision, an R&D approach to data, and strong foundational components. This is where the role of the Chief Data Officer (CDO) is a critical complement to the existing executive suite. A study of over organizations I led, recently published by MIT finds:

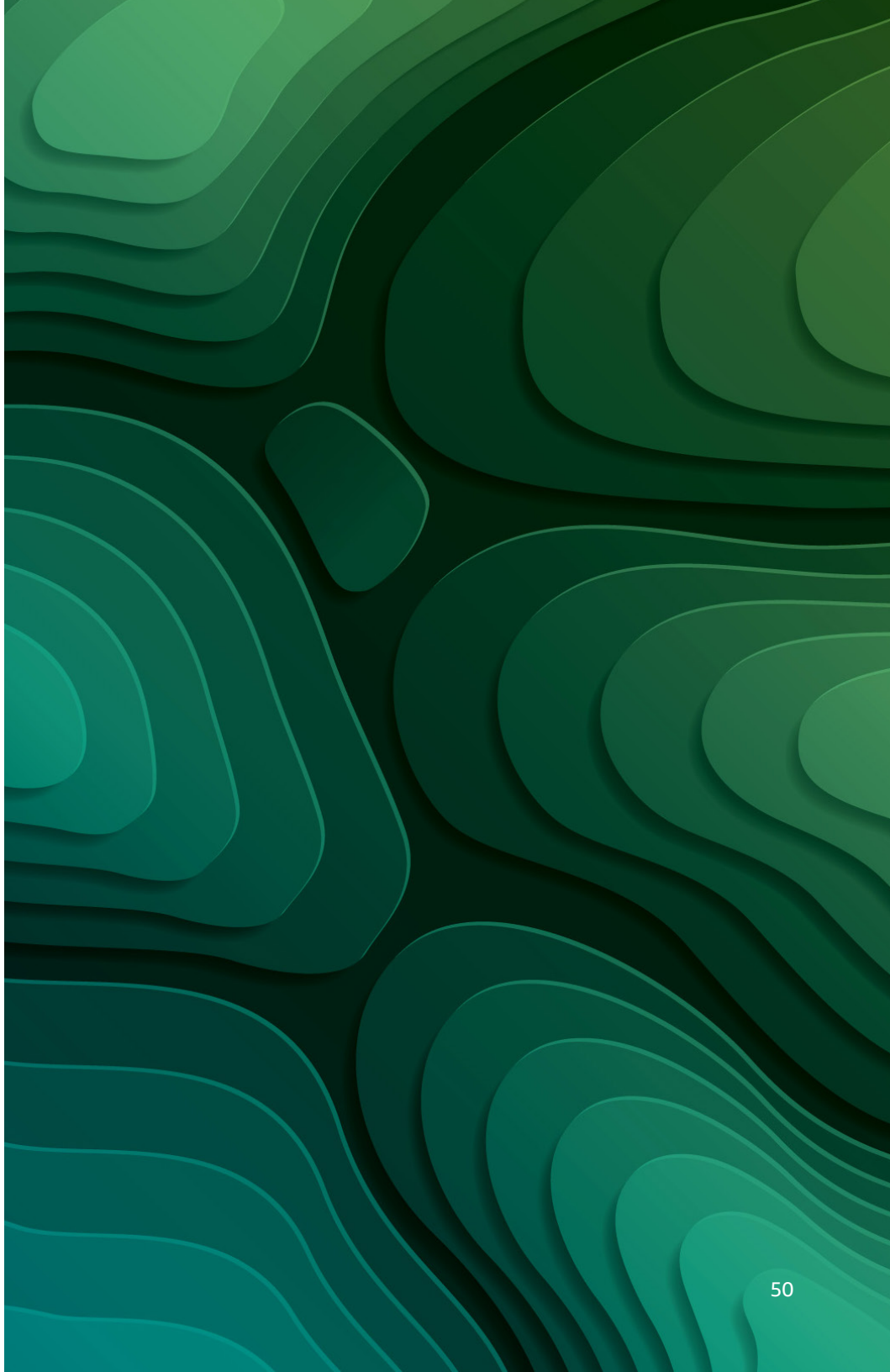
- Organizations with actual C-level CDOs are 4x more likely to be transforming their business with data.
- Organizations with only a CIO responsible for their data assets are only half as likely as those with a CDO to be benefitting from advanced analytics.
- Organizations with CDOs are 3x more likely to be sharing data freely among business units.
- Organizations with a C-level CDO are 7x more likely to be generating external monetary value from their data, and 3x more likely to be generating other forms of commercial value from their data
- And organizations with a C-level CDO are 3x more likely to formally value their data assets.

Moreover, investors have started favoring companies that treat data as an actual asset. A study I ran at Gartner and published in my Infonomics book found that public companies demonstrating data-driven behaviors have a market-to-book value ratio nearly two times higher than the market average. And companies that sell data, data derivatives, or digital products have a 300% higher market-to-book ratio. Even savvy private equity firms have started performing “data diligence” on target companies – assessing their data-related capabilities, architectures, quality, governance, and risks, in addition to opportunities for generating new data-driven value streams.



Summary

Business and technology leaders looking to thrive – not just survive – in today's and tomorrow's data or digital economy must evolve to measuring, managing, and monetizing it like one. Data's unique attributes make it a resource far superior to oil or other resources that deplete when consumed, can only be used one way at a time, and are not regenerative like data is. Today more than ever there's an imperative to treat data as an actual asset, not just talk about it as one.



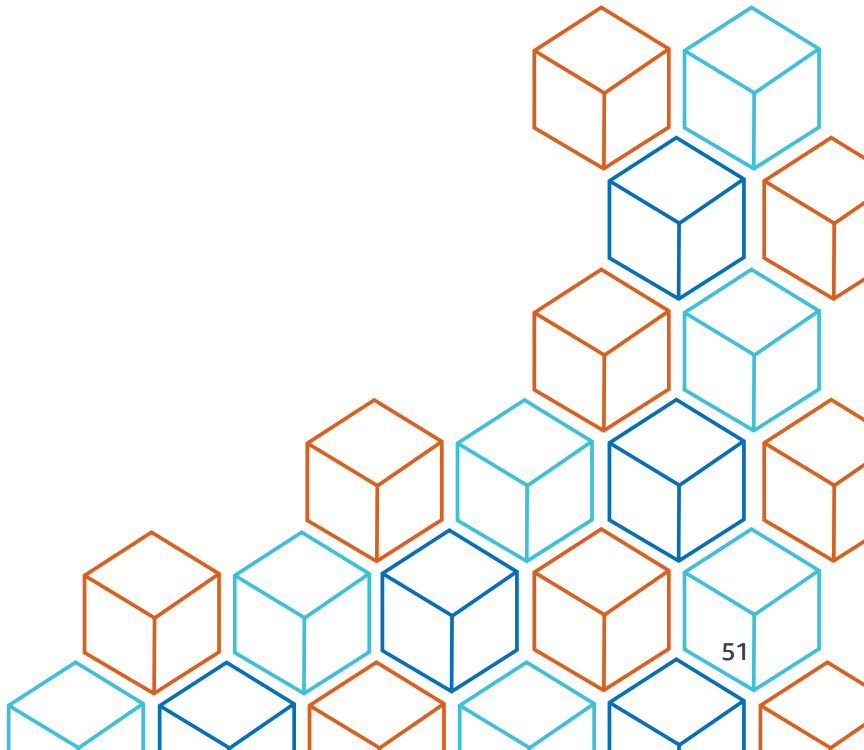
Conclusion

With AWS Data Exchange you can find, subscribe to, and use third-party data in the cloud to complement your internal data for improved decision-making.

AWS Data Exchange provides a broad portfolio of data sources and the tools and services that help streamline data licensing and delivery, enabling customers to extract value from their third-party data.

As you've seen firsthand, third-party data is being used in multiple ways to stay ahead of competition. Those who use third-party data with their first-party data will be light years ahead of their competitors in finding new opportunities and creating a frictionless data experience across their enterprise. The AWS Data Exchange approach is rooted in helping customers create that value as quickly and efficiently as possible.

Find the Right Data Products for You




aws
AWS Data Exchange